Modelli statistici lineari

Sergio Polini

19 gennaio 2010

Indice

Capitolo 1

Disegni sperimentali e modelli statistici parametrici

Uno studio empirico richiede sempre un'analisi statistica dei dati, siano essi provocati o solo osservati dal ricercatore. La sezione 1.1 descrive brevemente gli aspetti fondamentali di uno studio empirico, la sezione 1.2 illustra la struttura della matrice dei dati, distinguendo tra una *variabile risposta* (una variabile aleatoria) e una o più *variabili esplicative* (non aleatorie), e la scelta di un *modello campionario*, cioè di una famiglia parametrica cui si ipotizzino appartenere i risultati in quanto determinazioni della variabile risposta. La sezione 1.3 mostra come si possono reinterpretare i parametri sulla base delle variabili esplicative.

La sezione 1.4 entra con maggior dettaglio nell'esame del *modello lineare normale*, mostrando come vengono stimati i parametri e come vengono calcolati i valori teorici e gli scostamenti da questi dei valori osservati. Si illustra poi il *teorema di Cochran*, che consente di costruire un test di verifica della significatività del modello adottato mediante l'analisi della quota di variabilità del fenomeno che risulta spiegata da questo.

1.1 Il disegno sperimentale

Si conducono esperimenti per stabilire relazioni di causa-effetto tra diversi fenomeni (*studi sperimentali*, *experimental study* in inglese), si tenta di rilevare relazioni – tutte da interpretare – in un processo osservato (*studi osservazionali*, *observational study*).

In entrambi i casi si distingue tra *variabili esplicative* e *variabili risposta*, che vengono osservate su *unità sperimentali*.

Negli studi sperimentali, il ricercatore seleziona le unità sperimentali e applica loro un *trattamento* (ogni trattamento è definito da un insieme di particolari valori o livelli delle variabili esplicative) secondo un processo di *randomizzazione* (somministrazione casuale dei trattamenti alle unità sperimentali). Si possono stabilire relazioni di causa-effetto tra le variabili esplicative e la variabile risposta proprio perché i trattamenti sono scelti e controllati dal ricercatore, soprattutto se questi introduce un *trattamento di controllo* che consiste nel rilevare i valori della variabile risposta nel caso di assenza di trattamento o di trattamento standard.

Negli studi osservazionali, invece, il ricercatore osserva i valori o livelli di alcune variabili in alcune unità, senza poter assegnare ad esse casualmente i valori o livelli di

interesse. Inoltre, essendo la variabile risposta anch'essa osservata come quelle esplicative, non si possono stabilire relazioni di causa-effetto se non utilizzando evidenze di altro tipo (esempio classico in ambito economico: si può rilevare che la quantità di moneta aumenta quando aumentano i prezzi, ma da ciò non si può dedurre che uno dei due aumenti è causa dell'altro; non a caso, le opinioni degli economisti divergono da secoli al riguardo).

Il *disegno sperimentale* definisce la struttura logica sia di uno studio sperimentale che di uno studio osservazionale.

1.1.1 Le componenti di un esperimento

Studi sperimentali

Le componenti di uno *studio sperimentale* sono:

a) le variabili esplicative, dette anche fattori sperimentali;

b) i trattamenti;

c) le unità sperimentali;

d) il processo di assegnazione dei trattamenti alle unità sperimentali (randomizzazione);

e) la rilevazione dei dati, la loro analisi e la presentazione dei risultati.

I *fattori sperimentali* sono le variabili di cui si vuol rilevare l'effetto sulle unità sperimentali.

L'insieme dei *trattamenti* è determinato dai livelli di ciascun fattore: se vi è un solo fattore con tre livelli, sono possibili tre trattamenti; se vi sono più fattori, sono possibili tanti trattamenti quante sono le combinazioni dei loro livelli. È spesso utile prevedere un trattamento di *controllo*, ovvero un caso di assenza di trattamento o di trattamento standard col quale confrontare gli altri (ad esempio, somministrazione di un farmaco già in uso accanto a quella di un farmaco nuovo).

Le *unità sperimentali* sono le unità più piccole a cui può essere assegnato un trattamento. Il numero delle unità sperimentali dipende sia da valutazioni circa la potenza dei test statistici, sia da vincoli di costo o di tempo. In genere il numero delle unità è un multiplo del numero dei trattamenti e vi è lo stesso numero di unità per ciascun trattamento; si parla, in questi casi, di *esperimenti completi* (più unità per ciascun trattamento) *e bilanciati* (lo stesso numero di unità per tutti i trattamenti). La replicazione consente di stimare la variabilità dell'errore sperimentale, variabilità che non sarebbe valutabile se vi fosse una sola unità per ogni trattamento.

I trattamenti vengono assegnati alle unità sperimentali in modo casuale (*randomizzazione*), al fine di eliminare l'influenza di fattori fuori del controllo del ricercatore.

Si usa talvolta una *randomizzazione a blocchi*, che consiste nell'introdurre fattori, detti *fattori di disturbo* o *subsperimentali*, che non siano vere e proprie variabili esplicative (quelle di cui interessa verificare gli effetti), ma permettano di spiegare parte della variabilità.

Ad esempio, se si intende verificare l'ipotesi che la vitamina C aiuta a prevenire il raffreddore, si possono dividere le unità sperimentali in *blocchi* di uguale numerosità e omogenei rispetto all'età, al sesso, allo stato di salute generale, alle abitudini alimentari ecc. Immaginando di voler tenere conto solo del sesso, si effettua lo stesso esperimento sia sul blocco delle femmine che su quello dei maschi e si tiene conto poi dei risultati osservati nei due blocchi. Appare evidente che in una randomizzazione semplice non vi è

alcuna garanzia che tra le unità sperimentali vi siano tante femmine quanti maschi; con i due blocchi, invece, si è in grado di valutare sia se l'effetto della vitamina C varia secondo il sesso, sia – come risultato secondario – se vi sono differenze legate al sesso (uno dei due sessi prende più facilmente il raffreddore) anche nel caso la vitamina C risultasse non avere alcun effetto.

La rilevazione dei risultati dà luogo ad una *matrice dei dati*, che viene analizzata per verificare se si può rifiutare un'ipotesi nulla, la cui natura dipende dalle finalità dell'esperimento.

Studi osservazionali

Gli *studi osservazionali* differiscono da quelli sperimentali in quanto non è possibile assegnare casualmente i livelli dei fattori alle unità sperimentali; si possono pertanto stabilire *associazioni* tra i fattori e le variabili risposta, ma non relazioni di causa-effetto. Gli studi osservazionali sono stati classificati in molti modi; in prima istanza si possono distinguere studi *cross-section* (rilevazioni su una o più popolazioni in uno stesso istante o in uno stesso intervallo di tempo), studi *prospettici* (rilevazioni su uno o più gruppi nel corso del tempo, al fine di prevedere l'andamento della variabile di interesse) e studi *retrospettivi* (rilevazione dell'andamento passato di un fenomeno al fine di individuarne possibili cause).[1]

1.1.2 Disegni sperimentali standard

Vi sono molti possibili disegni sperimentali; i più usati sono:

a) *disegno completamente randomizzato* (DCR): si tratta della forma più semplice di disegno sperimentale, che viene usata quando vi è un solo fattore e le unità sperimentali sono relativamente omogenee; vi sono quindi tanti trattamenti quanti sono i livelli del fattore, più unità sperimentali per ciascun trattamento;

b) *disegno fattoriale* (DF): vi sono più fattori e interessa studiare non solo e non tanto l'effetto che hanno singolarmente sulla variabile risposta (a tal fine potrebbero essere studiati separatamente), ma soprattutto gli *effetti interattivi*, cioè gli ulteriori effetti combinati di due o più fattori; si tratta, come il precedente, di un disegno completo (con replicazione completa), in quanto vi sono almeno due unità statistiche per tutti i livelli dei fattori e per tutte le loro combinazioni;

c) *disegno a blocchi randomizzati* (DBR): si inseriscono fattori sub-sperimentali che contribuiscano a spiegare la variabilità, assumendo che non interagiscano con i fattori sperimentali; il DBR può essere completo oppure incompleto, vi possono cioè essere più unità statistiche per ciascuna combinazione dei livelli dei fattori sperimentali e subsperimentali, oppure può esservi replicazione solo per le combinazioni dei livelli dei fattori sperimentali.

[1] Gli studi prospettici e quelli retrospettivi sembrano simili, in quanto evidentemente basati entrambi su serie storiche; la principale differenza risiede nel fatto che i primi si basano necessariamente su gruppi ristretti che possano essere tenuti sotto osservazione e, inoltre, il tempo di osservazione è limitato; gli studi retrospettivi, invece, possono basarsi su rilevazioni condotte su gruppi molto più ampi e per tempi molti più lunghi.

Esempio 1.1. Il file `caffeina.csv`[2] contiene le 30 osservazioni relative ad un esperimento condotto secondo un *disegno completamente randomizzato*. Vi è un solo fattore, la caffeina, con tre livelli; vi sono quindi tre trattamenti (codificati con 1, 2 e 3) e per ciascuno 10 unità sperimentali (replicazione completa e bilanciata), su ciascuna delle quali è misurato il livello di ansietà, la variabile risposta.

Esempio 1.2. Il file `dietepec.csv`[3] contiene 40 osservazioni relative ad un esperimento condotto secondo un *disegno fattoriale*: si vuole studiare l'effetto sull'incremento del peso (variabile risposta) di 40 pecore (le unità sperimentali) del rame e del cobalto presenti nelle loro diete (i fattori). Interessa in particolare l'effetto interattivo del rame e del cobalto, ovvero l'effetto che rame e cobalto insieme hanno in più, rispetto a quello che hanno quando somministrati separatamente. I due fattori presentano due livelli, da intendere come assenza (1) e presenza (2). Vi sono quindi 4 diversi possibili trattamenti (1 e 1, 1 e 2, 2 e 1, 2 e 2), assegnati ciascuno a 10 pecore (ancora una replicazione completa e bilanciata).

Esempio 1.3. Il file `dietetop.csv`[4] contiene 40 osservazioni relative ad un esperimento condotto secondo un *disegno a blocchi*. Interessa l'effetto di 5 diverse diete, indicate con le lettere da "a" a "e", sulla variabile risposta. Si introduce un fattore subsperimentale, la nidiata (otto diverse nidiate indicate con i numeri da 1 a 8), nell'assunzione che possa contribuire a spiegare la variabilità del fenomeno ma che non interagisca con la dieta. Vi sono 8 diverse unità sperimentali per ciascuna dieta, una sola per ciascuna combinazione tra diete e nidiate; si ha quindi replicazione completa e bilanciata per la dieta (il fattore sperimentale, quello di cui si intende studiare l'effetto), assenza di replicazione per la combinazione diete/nidiate:

1.2 Dalla matrice dei dati al modello campionario

I risultati dell'esperimento vengono raccolti in una *matrice di dati* $\underset{n,k}{\mathbf{D}}$.

La matrice ha tante righe, n, quante sono le unità sperimentali. Quanto alle colonne, si hanno normalmente:

a) una colonna con gli identificativi delle unità sperimentali, che possono anche essere semplicemente i numeri da 1 a n (può essere utile per una interpretazione dei risultati, ma è superflua ai fini dell'analisi statistica in sé);

b) una colonna per la variabile risposta;

c) tante colonne quanti sono le variabili esplicative, quindi i fattori sperimentali e subsperimentali (di blocco) considerati.

Le colonne relative alle unità sperimentali ed alle variabili esplicative sono sotto il controllo del ricercatore, che sceglie le une e le altre. Si tratta quindi di *variabili matematiche* (non aleatorie). La variabile risposta è invece costituita dalle determinazioni di una *variabile aleatoria* che va studiata adottando un appropriato modello statistico.

[2] `http://web.mclink.it/MC1166/ModelliStatistici/caffeina.csv`.
[3] `http://web.mclink.it/MC1166/ModelliStatistici/dietepec.csv`.
[4] `http://web.mclink.it/MC1166/ModelliStatistici/dietetop.csv`.

Esempio 1.4. Si vuole studiare la relazione tra pressione sistolica ed età in un gruppo di 33 donne adulte. La matrice dei dati si presenta nel modo seguente:

$$\underset{n,k}{\mathbf{D}} = \begin{bmatrix} 1 & y_1 & x_1 \\ \vdots & \vdots & \vdots \\ i & y_i & x_i \\ \vdots & \vdots & \vdots \\ 33 & y_{33} & x_{33} \end{bmatrix} \qquad \begin{array}{l} y = \text{pressione sistolica (variabile risposta)} \\ x = \text{età (variabile esplicativa)} \end{array}$$

In generale, un modello statistico parametrico può essere formalizzato come segue:

$$M : \{\mathbf{Y}, \mathcal{Y}^n, p_n(\mathbf{Y} \mid \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$$

dove:

- $\underset{n}{\mathbf{Y}}$ è un vettore casuale (una variabile aleatoria multipla $Y_1, \ldots, Y_n$) di cui le osservazioni costituiscono una determinazione;
- $\mathcal{Y}$ è lo spazio dei risultati possibili, o spazio campionario, relativo ad un singolo elemento di $\mathbf{Y}$;
- $\mathcal{Y}^n$ è lo spazio campionario relativo al vettore $\mathbf{Y}$;
- $p_n(\mathbf{Y} \mid \boldsymbol{\theta})$ è una funzione di massa/densità di probabilità congiunta, dipendente da un vettore di parametri $\boldsymbol{\theta}$;
- Θ è lo spazio dei parametri, l'insieme dei valori che il vettore $\boldsymbol{\theta}$ può assumere.

Come noto, se $\mathbf{Y}$ fosse un campione casuale le $Y_1, \ldots, Y_n$ sarebbero variabili aleatorie indipendenti e identicamente distribuite; se si assumesse $Y_i \sim N(\mu, \sigma^2)$, vi sarebbero due soli parametri incogniti e il modello statistico sarebbe del tipo:

$$M : \left\{ \mathbf{Y}, \mathcal{Y}^n, p(\mathbf{y} \mid \mu, \sigma) = \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{y_i - \mu}{\sigma} \right)^2 \right\} \right), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^+ \right\}$$

In ambito sperimentale, ciò equivarrebbe ad ipotizzare che i trattamenti (i fattori e le loro combinazioni) non hanno alcun effetto: i valori osservati della variabile risposta non sarebbero altro che oscillazioni casuali intorno ad uno stesso valore medio.

È spesso questa la forma che assume l'ipotesi nulla, che si è interessati a rifiutare per poter accettare un'ipotesi alternativa secondo cui i trattamenti hanno invece effetto. A tale scopo si parte da assunzioni quasi opposte, secondo le quali il vettore $\mathbf{Y}$ è costituito da variabili aleatorie appartenenti ad una stessa *famiglia parametrica* (un insieme di funzioni di massa/densità di probabilità dello stesso tipo, diverse per i valori dei parametri), ma non identicamente distribuite, per poi ridurre il numero dei parametri.

Viene detto *modello campionario* un'assunzione circa le caratteristiche distribuzionali della variabile risposta. Ad esempio, assumendo la famiglia parametrica normale per n variabili aleatorie non indipendenti e non identicamente distribuite, si avrebbe un modello campionario del tipo: $\mathbf{Y} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, con $2n + n(n-1)/2$ parametri incogniti:

- n valori medi;
- n varianze;
- $n(n-1)/2$ covarianze.

Tale numero può essere ridotto assumendo:

– indipendenza a due a due, quindi covarianze nulle (si scende così a $2n$);
– omoschedasticità, ovvero varianze tutte uguali a σ^2.

Rimane così un modello con $n+1$ parametri:

$$M : \left\{ \mathbf{Y}, \mathcal{Y}^n, p_n(\mathbf{y} \mid \boldsymbol{\mu}, \sigma) = \prod_{i=1}^{n} \left(\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{y_i - \mu_i}{\sigma} \right)^2 \right\} \right), (\boldsymbol{\mu}, \sigma) \in \mathbb{R}^n \times \mathbb{R}^+ \right\}$$

Si riducono poi ulteriormente i parametri mediante *riparametrizzazione*.

1.3 Modelli di riparametrizzazione

Nell'operazione di *riparametrizzazione* il ricercatore lega i parametri della famiglia parametrica alle altre informazioni contenute nella matrice dei dati osservata.

Si può pensare, per un esempio, alla matrice dei dati contenuta nel file `caffeina.csv`[5]: tre trattamenti, somministrati ciascuno a 10 unità sperimentali per 30 osservazioni complessive. Le ipotesi di indipendenza e di omoschedasticità consentono di ridurre i parametri da $2 \cdot 30 + 30(30-1)/2 = 495$ a 31, 30 medie ed una varianza. Prescindendo dalla varianza, si avrebbe un modello con 30 parametri incogniti, $\mu_1, \ldots, \mu_{30}$, del tipo:

$$Y_j = \mu_j + \varepsilon_j, \qquad j = 1, \ldots, 30$$

dove Y_j è la j-esima variabile aleatoria, μ_j il suo valore medio e ε_j un *errore* dovuto alla variabilità (dipendente quindi dalla varianza $\sigma^2_{Y_j}$). Da notare che né μ_j né ε_j sono osservabili.

Il ricercatore ha però rilevato i 30 valori della variabile risposta dopo aver diviso le 30 unità in tre gruppi di uguale numerosità e dopo aver somministrato trattamenti diversi ai tre gruppi. Assume quindi che quei 30 valori dipendano a gruppi di 10 dai tre trattamenti e procede alla *riparametrizzazione*, sostituendo ai 30 parametri i seguenti quattro:

– μ: un effetto di riferimento;
– α_i, $i = 1, 2, 3$: effetti differenziali, rispetto a quello di riferimento, indotti dai tre trattamenti.

Ciò consente di ridefinire i parametri come segue:

$$\mu_{ir} = \mu + \alpha_i, \qquad i = 1, \ldots, 3 \quad r = 1, \ldots, 10$$

In forma matriciale,

$$\text{si passa da: } \mu_j = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_{10} \\ \mu_{11} \\ \vdots \\ \mu_{20} \\ \mu_{21} \\ \vdots \\ \mu_{30} \end{bmatrix} \quad \text{a: } \mu_{ir} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \mu + \alpha_1 \\ \vdots \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_2 \\ \mu + \alpha_3 \\ \vdots \\ \mu + \alpha_3 \end{bmatrix}$$

[5] `http://web.mclink.it/MC1166/ModelliStatistici/caffeina.csv`.

dove la prima matrice nella seconda espressione (quella con elementi 1 o 0) è la *matrice di riparametrizzazione*, **A**.

La matrice di riparametrizzazione presenta spesso colonne linearmente dipendenti (in quella sopra costruita, ad esempio, la prima colonna è chiaramente somma delle altre tre). Tale ridondanza deve essere eliminata per consentire l'interpretazione e la stima dei parametri (v. sez. 1.4.1) e ciò viene fatto introducendo dei vincoli; ad esempio:

a) si pone uno degli α_i uguale a 0, ad esempio $\alpha_1 = 0$, intendendo μ come l'effetto del primo trattamento, $\alpha_2 = \mu_2 - \mu$ e $\alpha_3 = \mu_2 - \mu$ come gli effetti differenziali, rispetto al primo, del secondo e del terzo trattamento; ciò risulta equivalente ad una sostituzione della seconda colonna della matrice **A** con un vettore di zeri e quindi ad eliminarla; si parla in questi casi di *riparametrizzazione corner point*: un effetto differenziale viene considerato nullo e costituisce così il "termine di riferimento" per valutare gli altri;

b) si pone $\sum_{i=1}^{3} a_i = 0$; in questo caso uno dei nuovi parametri può essere espresso in funzione degli altri, ad esempio $\alpha_3 = -\alpha_1 - \alpha_2$, con conseguente ristrutturazione della matrice di riparametrizzazione:

$$\mu_{ir} = \begin{bmatrix} 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & -1 & -1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \mu + \alpha_1 \\ \vdots \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_2 \\ \mu - \alpha_1 - \alpha_2 \\ \vdots \\ \mu - \alpha_1 - \alpha_2 \end{bmatrix} = \begin{bmatrix} \mu + \alpha_1 \\ \vdots \\ \mu + \alpha_1 \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_2 \\ \mu + \alpha_3 \\ \vdots \\ \mu + \alpha_3 \end{bmatrix}$$

In ogni caso, il modello di riparametrizzazione consente di pervenire ad un nuovo modello statistico. Si parte da un *modello campionario* relativo alla colonna della matrice dei dati che contiene la variabile risposta:

$$Y_j = \mu_j + \varepsilon_j, \qquad j = 1, \ldots, 30$$

si passa da un *modello di riparametrizzazione* che, pur utilizzando le altre colonne della matrice dei dati (le covariate), *interessa solo i parametri* e tiene conto delle replicazioni:

$$\mu_{ir} = \mu + \alpha_i, \qquad i = 1, \ldots, 3 \quad r = 1, \ldots, 10$$

e si arriva ad un nuovo modello statistico per la variabile risposta:

$$Y_{ir} = \mu + \alpha_i + \varepsilon_{ir}, \qquad i = 1, \ldots, 3 \quad r = 1, \ldots, 10$$

dove:

- Y_{ir} è la variabile aleatoria di cui è determinazione la r-esima osservazione (replica) nell'ambito dell'i-esimo trattamento;
- μ è il parametro, non osservabile, attribuito al livello generale del fenomeno;
- α_i è il parametro, non osservabile, corrispondente all'effetto dell'i-esimo trattamento;

– ε_{ir} è la variabile aleatoria *errore*, uno scostamenteo casuale di Y_{ir} dal valore atteso $\mu_{ir} = \mu + \alpha_i$.

Gli stimatori di μ e di α_i, $\hat{\mu}$ e $\hat{\alpha}_i$, consentono di definire sia una variabile aleatoria *valore teorico* che è loro combinazione lineare:

$$\hat{Y}_{ir} = \hat{\mu}_{ir} = \hat{\mu} + \hat{\alpha}_i$$

sia una variabile aleatoria *residuo* come differenza tra le variabili Y_{ir} e i valori teorici:

$$e_{ir} = Y_{ir} - \hat{Y}_{ir} = Y_{ir} - (\hat{\mu} + \hat{\alpha}_i)$$

A differenza della v.a. errore, la v.a. residuo è osservabile;[6] una volta stimati i parametri, i valori osservati (le determinazioni di Y_{ir}) potranno essere interpretati come segue:

$$y_{ir} = \hat{\mu} + \hat{\alpha}_i + e_{ir}$$

Nel nuovo modello l'ipotesi nulla consiste nell'assegnare ai parametri valori tali da configurare un qualche effetto dei trattamenti sulla variabile risposta. Nei casi più semplici, l'ipotesi nulla diventa: $\alpha_i = 0\ \forall i$, ovvero $\sum_{i=1}^{3} \alpha_i^2 = 0$. Ciò equivale a dire che, pur potendosi individuare tre gruppi, uno per ciascun trattamento, si ha $Y_{ir} = \mu + \varepsilon_{ir}$: esiste una sola media, intorno alla quale si distribuiscono casualmente i diversi possibili valori delle 30 variabili aleatorie, quindi il trattamento ha un effetto nullo sulla variabile risposta.

Esempio 1.5. In un esperimento si hanno, di norma, diversi trattamenti. Ci si può ricondurre ad una situazione più semplice considerando due soli trattamenti, ad esempio il primo e il terzo. In questo caso, tutto si riduce al familiare confronto tra due medie: per verificare che la differenza tra le medie di due gruppi non sono attribuibili al caso, si può effettuare un test t. Con R si usa la funzione `t.test()` con l'opzione `var.equal=TRUE` (omoschedasticità):

```
> caffeina <- read.csv("caffeina.csv")
> attach(caffeina)
> t.test(y[tr==1], y[tr==3], var.equal=TRUE)

        Two Sample t-test

data:  y[tr == 1] and y[tr == 3]
t = -3.3942, df = 18, p-value = 0.003233
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.66643 -1.33357
sample estimates:
mean of x mean of y
    244.8     248.3
```

Con SAS, dopo aver importato il file `caffeins.csv` con l'opzione Import Data del menù File, si possono usare i comandi:

[6] Per altre importanti differenze vedi sez. 1.4.3.

```
data caffeina2;
  set caffeina;
  if tr = 2 then delete;
run;
proc ttest data=caffeina2 method=pooled;
  class tr;
  var y;
run;
```

il cui output è:[7]

```
                              The TTEST Procedure

                                  Variable:  y
     tr              N        Mean     Std Dev     Std Err     Minimum     Maximum
     1              10       244.8      2.3944      0.7572       242.0       248.0
     3              10       248.3      2.2136      0.7000       245.0       252.0
     Diff (1-2)           -3.5000      2.3058      1.0312

 tr           Method              Mean      95% CL Mean       Std Dev     95% CL Std Dev
 1                                244.8      243.1    246.5     2.3944     1.6470   4.3713
 3                                248.3      246.7    249.9     2.2136     1.5226   4.0412
 Diff (1-2)   Pooled             -3.5000    -5.6664  -1.3336     2.3058     1.7423   3.4099
 Diff (1-2)   Satterthwaite      -3.5000    -5.6674  -1.3326

              Method           Variances        DF    t Value    Pr > |t|
              Pooled           Equal            18      -3.39      0.0032
              Satterthwaite    Unequal       17.89      -3.39      0.0033
```

Si può rifiutare l'ipotesi nulla, $H_0 : \bar{Y}_{1\cdot} = \bar{Y}_{3\cdot}$,[8] in quanto:

a) il valore della variabile aleatoria

$$t = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{3\cdot}}{S_p\sqrt{\frac{1}{10} + \frac{1}{10}}} = \frac{\bar{Y}_{1\cdot} - \bar{Y}_{3\cdot}}{S_p\sqrt{\frac{2}{10}}}$$

con $Sp = \sqrt{\frac{(10-1)S_1^2+(10-1)S_3^2}{10+10-2}} = \sqrt{\frac{9S_1^2+9S_3^2}{18}}$, è nettamente maggiore di 1 in valore assoluto, quindi la differenza tra le medie del primo e del terzo gruppo è nettamente maggiore della variabilità attribuibile all'accidentalità (radice quadrata della devianza divisa per i gradi di libertà);

b) il *p-value*, 0.003, è minore di qualsiasi valore ragionevole della probabilità dell'errore di primo tipo (rifutare l'ipotesi nulla quando è vera).

Se il numero dei trattamenti è maggiore di due, occorrono test più sofisticati.

[7] Il metodo `Pooled` assume omoschedasticità. Il metodo `Satterthwaite` assume varianze diverse nelle due replicazioni e corrisponde all'opzione di default `var.equal=FALSE` di R. L'output di SAS comprende anche una parte, qui omessa, sul confronto tra le varianze dei due gruppi, che si ottiene in R con la funzione `var.test()`.

[8] $\bar{Y}_{1\cdot}$ è la media delle Y_{1r} per $r = 1, \ldots, 10$.

1.4 Modelli statistici lineari

I modelli statistici lineari sono i più semplici e vengono utilizzati spesso. In generale, un modello statistico viene detto *lineare* se è *lineare nei parametri*, cioè se la variabile risposta può essere considerata come il risultato di una trasformazione lineare dei parametri, trasformazione alla quale è associata una matrice che altro non è che la matrice di riparametrizzazione appena vista.

Si dice quindi *modello lineare generale* un modello del tipo:

$$\mathbf{Y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$$

dove:

- $\mathbf{Y}$ è un vettore di n variabili aleatorie osservabili;
- $\mathbf{A}$ è una matrice di costanti note di ordine $n \times p$;
- $\boldsymbol{\theta}$ è un vettore di parametri incogniti e non osservabili di ordine p;
- $\boldsymbol{\varepsilon}$ è un vettore di errori casuali, cioè di variabili aleatorie non osservabili con media nulla e a due a due incorrelate.

Si dice invece *modello lineare normale* un modello lineare costruito mediante riparametrizzazione di un modello campionario basato sulla famiglia parametrica normale. Si assume quindi che $\mathbf{Y}$ sia, o risulti, una v.a. di distribuzione $MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ con componenti a due a due indipendenti e omoschedastiche ($\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, ovvero nella matrice di varianze e covarianze sono non nulle, e uguali tra loro, solo le varianze disposte lungo la diagonale principale). Vi sono, al riguardo, due possibili chiavi di lettura:[9]

a) secondo l'impostazione più tradizionale, si aggiunge al modello lineare generale l'assunto della multinormalità della variabile errore, traendo da ciò la multinormalità della variabile risposta:

$$\begin{cases} \mathbf{Y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\varepsilon} \sim MN(\mathbf{0}, \sigma^2\mathbf{I}) \end{cases} \qquad \Rightarrow \qquad \mathbf{Y} \sim MN(\mathbf{A}\boldsymbol{\theta}, \sigma^2\mathbf{I})$$

b) secondo un'impostazione più moderna, si assume che $\mathbf{Y}$ sia multinormale con $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ e che il suo valore atteso sia $\mathbf{A}\boldsymbol{\theta}$ (linearità) e da ciò si ricava la multinormalità della v.a. errore:

$$\begin{cases} \mathbf{Y} \sim MN(\boldsymbol{\mu}, \sigma^2\mathbf{I}) \\ \mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu} = \mathbf{A}\boldsymbol{\theta} \end{cases} \qquad \Rightarrow \qquad \mathbf{Y} - \mathbb{E}[\mathbf{Y}] = \boldsymbol{\varepsilon} \sim MN(\mathbf{0}, \sigma^2\mathbf{I})$$

Con la riparametrizzazione si sostituisce il vettore delle n medie $\boldsymbol{\mu}$ con il prodotto $\mathbf{A}\boldsymbol{\theta}$, dove $\mathbf{A}$ è una matrice funzione delle variabile esplicative (la sua i-esima riga descrive l'i-esima unità sperimentale rispetto alle variabili esplicative) e $\boldsymbol{\theta}$ è un vettore di p nuovi parametri incogniti, con $p \leq n$, che esprimono le relazioni tra la variabile risposta e le

[9] Le due chiavi di lettura sono equivalenti nel caso dei modelli lineari; non lo sono più quando si passa a modelli di altro tipo.

variabili esplicative. Si ha quindi:[10]

$$\begin{aligned}\mathbb{E}[\mathbf{Y}] &= \mathbf{A}\boldsymbol{\theta}\\ \boldsymbol{\Sigma}_Y &= \boldsymbol{\Sigma}_U = \sigma^2\mathbf{I}\\ f_{\mathbf{Y}}(Y) &= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{A}\boldsymbol{\theta})'(\mathbf{Y}-\mathbf{A}\boldsymbol{\theta})\right\}\end{aligned}$$

Si distinguono:

a) modelli *ANOVA* (*ANalysis Of VAriance*), per variabili esplicative anche qualitative, nei quali la matrice di riparametrizzazione viene generalmente indicata con $\mathbf{A}$ e il vettore dei nuovi parametri con $\boldsymbol{\eta}$;

b) modelli *di regressione*, per variabili esplicative quantitative, nei quali la matrice di riparametrizzazione viene generalmente indicata con $\mathbf{X}$ e il vettore dei nuovi parametri con $\boldsymbol{\beta}$;

c) modelli *ANCOVA* (*ANalysis Of COVAriance*), per variabili esplicative sia qualitative che quantitative.

Osservazione. Se si hanno variabili esplicative qualitative che possono essere lette anche come quantitative, si può passare da un modello ANOVA ad uno regressivo imponendo dei vincoli sui parametri del modello ANOVA. Nella matrice dei dati `caffeina` il trattamento ha tre modalità (1, 2 e 3); se queste possono essere interpretate come misure di diverse quantità di caffeina, si può passare dal modello ANOVA:

$$\mu_i = \mu + \alpha_i \qquad\qquad i = 1, 2, 3$$

in cui vi sono 4 parametri (di cui uno ridondante) ad un modello regressivo:

$$\mu_i = \alpha + \beta x_i \qquad\qquad x_i = \begin{cases} 1 & \text{quando } i = 1\\ 2 & \text{quando } i = 2\\ 3 & \text{quando } i = 3\end{cases}$$

con 2 parametri. Per ottenere ciò, si introduce nel modello ANOVA un vincolo, ad esempio $\alpha_1 = 0$, e si interpretano come segue i parametri:

a) modello ANOVA:
   - μ: ansietà con una dose di caffeina pari a 1;
   - α_2: effetto differenziale di una dose di caffeina pari a 2 rispetto alla dose pari a 1;
   - α_3: effetto differenziale di una dose di caffeina pari a 3 rispetto alla dose pari a 1;

[10] $\mathbb{E}[\mathbf{Y}] = \mathbf{A}\boldsymbol{\theta}$ vale in quanto $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, $\boldsymbol{\Sigma}_Y = \boldsymbol{\Sigma}_\varepsilon = \sigma^2\mathbf{I}$ per ipotesi. Quanto alla legge di probabilità della variabile aleatoria n-dimensionale $\mathbf{Y}$, essa, per l'ipotesi di indipendenza, è uguale al prodotto delle funzioni:

$$f_i(Y_i \mid \mathbf{A}^{(i)}\boldsymbol{\theta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{1}{2\sigma^2}(Y_i - \mathbf{A}^{(i)}\boldsymbol{\theta})^2\right\}$$

dove $\mathbf{A}^{(i)}$ è la i-esima riga della matrice di riparametrizzazione $\mathbf{A}$ e $\boldsymbol{\theta}$ è il vettore dei parametri (ad esempio, se il modello fosse quello a pag. 11, per $i = 1$ si avrebbe il prodotto scalare $(1, 1, 0)(\mu, \alpha_1, \alpha_2) = \mu + \alpha_1$); nel prodotto, la frazione $1/\sqrt{2\pi\sigma^2}$ viene elevata a n e gli esponenti si sommano: al variare di i, le differenze $Y_i - \mathbf{A}^{(i)}\boldsymbol{\theta}$ sono gli elementi del vettore $\mathbf{Y} - \mathbf{A}\boldsymbol{\theta}$ e la somma dei quadrati dei suoi elementi non è altro che il prodotto scalare standard del vettore per se stesso, quindi $(\mathbf{Y} - \mathbf{A}\boldsymbol{\theta})'(\mathbf{Y} - \mathbf{A}\boldsymbol{\theta})$.

Tabella 1.1. Relazioni tra modello ANOVA e modello regressivo, e i rispettivi parametri, nel caso dell'esperimento `caffeina`.

	Modello ANOVA	Modello regressivo
$\mu_1 =$	μ	$\alpha + \beta \cdot 1$
$\mu_2 =$	$\mu + \alpha_2$	$\alpha + \beta \cdot 2$
$\mu_3 =$	$\mu + \alpha_3$	$\alpha + \beta \cdot 3$

$$\begin{aligned} \mu &= \alpha + \beta \\ \mu + \alpha_2 = \alpha + \beta \cdot 2 \quad &\Rightarrow \quad \alpha_2 = \beta \\ \mu + \alpha_3 = \alpha + \beta \cdot 3 \quad &\Rightarrow \quad \alpha_3 = 2\beta \end{aligned}$$

b) modello regressivo:

- α: ansietà con una dose di caffeina pari a 0;
- β: incremento di ansietà dovuto all'aumento di una dose di caffeina.

In termini matriciali, rispettivamente:

$$\mu_i = \begin{bmatrix} 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 = 0 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} \mu \\ \vdots \\ \mu \\ \mu + \alpha_2 \\ \vdots \\ \mu + \alpha_2 \\ \mu + \alpha_3 \\ \vdots \\ \mu + \alpha_3 \end{bmatrix} \qquad \mu_i = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 2 \\ \vdots & \vdots \\ 1 & 2 \\ 1 & 3 \\ \vdots & \vdots \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \alpha + \beta \\ \vdots \\ \alpha + \beta \\ \alpha + 2\beta \\ \vdots \\ \alpha + 2\beta \\ \alpha + 3\beta \\ \vdots \\ \alpha + 3\beta \end{bmatrix}$$

Si stabiliscono quindi le relazioni tra i due modelli ed i rispettivi parametri esposte nella tabella 1.1.

1.4.1 Stima dei parametri

La stima dei parametri col metodo dei minimi quadrati si basa sulla minimizzazione della somma dei quadrati degli scarti dei valori osservati $\mathbf{y}$ dai valori attesi $\mathbb{E}[\mathbf{Y}] = \mathbf{A}\boldsymbol{\theta}$, ovvero, con i simboli generalmente usati per il modello ANOVA:

$$S(\boldsymbol{\eta}) = \sum_{i=1}^{n} (Y_i - \mathbf{A}^{(i)}\boldsymbol{\eta})^2 = (\mathbf{Y} - \mathbf{A}\boldsymbol{\eta})'(\mathbf{Y} - \mathbf{A}\boldsymbol{\eta}) = \mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{A}\boldsymbol{\eta} + \boldsymbol{\eta}'\mathbf{A}'\mathbf{A}\boldsymbol{\eta}$$

dove $\mathbf{A}^{(i)}$ è la i-esima riga della matrice $\mathbf{A}$ (sarebbe equivalente riferirsi più direttamente al modello regressivo, scrivendo $\sum_i (y_i - \mathbf{X}^{(i)}\boldsymbol{\beta})^2$).

Si tratta quindi di risolvere il sistema di equazioni, detto *sistema di equazioni normali*:

$$\frac{\partial S(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = -2\mathbf{A}'\mathbf{Y} + 2\mathbf{A}'\mathbf{A}\boldsymbol{\eta} = \mathbf{0} \quad \Rightarrow \quad \mathbf{A}'\mathbf{A}\boldsymbol{\eta} = \mathbf{A}'\mathbf{Y}$$

Se $\mathbf{A}$ è una matrice a rango pieno, lo è anche $\mathbf{A}'\mathbf{A}$, che è quindi invertibile.[11] Se $\mathbf{A}$ non è a rango pieno, si introducono dei vincoli sul vettore dei nuovi parametri $\boldsymbol{\eta}$; ciò equivale ad aggiungere un'equazione al sistema di equazioni normali, rendendo così possibile la stima di $\boldsymbol{\eta}$:

$$\hat{\boldsymbol{\eta}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y}$$

(per la stima vera e propria si usano ovviamente i valori osservati $\mathbf{y}$: $\hat{\boldsymbol{\eta}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y}$).

I parametri possono anche essere stimati col metodo della massima verosimiglianza. Data la funzione di densità congiunta:

$$f_{\mathbf{Y}}(Y) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{A}\boldsymbol{\eta})'(\mathbf{Y}-\mathbf{A}\boldsymbol{\eta})\right\}$$

La funzione di log-verosimiglianza è:

$$\ell(\boldsymbol{\eta},\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{(\mathbf{Y}-\mathbf{A}\boldsymbol{\eta})'(\mathbf{Y}-\mathbf{A}\boldsymbol{\eta})}{2\sigma^2}$$

Si vede che $\ell(\boldsymbol{\eta},\sigma^2)$, per qualsiasi valore di σ^2, è massimizzata dai valori di $\hat{\boldsymbol{\eta}}$ che minimizzano il numeratore dell'ultimo termine, che a sua volta altro non è che la quantità $S(\boldsymbol{\eta})$, minimizzata da $\hat{\boldsymbol{\eta}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y}$.

I due metodi portano quindi allo stesso risultato. Si deve tuttavia osservare che, mentre il metodo dei minimi quadrati può essere applicato solo nei modelli lineari normali, il metodo della massima verosimiglianza può essere usato anche con modelli di altro tipo.

Gli stimatori dei parametri $\boldsymbol{\eta}$ sono *stimatori corretti* (o *non distorti*); infatti:[12]

$$\mathbb{E}[\hat{\boldsymbol{\eta}}] = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbb{E}[\mathbf{Y}] = (\mathbf{A}'\mathbf{A})^{-1}(\mathbf{A}'\mathbf{A})\boldsymbol{\eta} = \boldsymbol{\eta}$$

inoltre:[13]

$$\mathrm{Cov}(\hat{\boldsymbol{\eta}}) = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\,\mathrm{Cov}(\mathbf{Y})\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} = (\mathbf{A}'\mathbf{A})^{-1}\sigma^2_{\mathbf{y}}$$

[11] Si dimostra che, date due matrici $\mathbf{A}$ e $\mathbf{B}$ moltiplicabili, $\mathrm{rk}(\mathbf{AB}) \le \min\{\mathrm{rk}(\mathbf{A}), \mathrm{rk}(\mathbf{B})\}$ (v. Appendice A, proposizione A.32). Si può quindi dimostrare che, se $\mathbf{B}$ è a rango pieno, allora $\mathrm{rk}(\mathbf{AB}) = \mathrm{rk}(\mathbf{A})$, cioè che la moltiplicazione per una matrice a rango pieno non cambia il rango di una matrice; infatti: $\mathrm{rk}(\mathbf{A}) \ge \mathrm{rk}(\mathbf{AB})$, ma, essendo $\mathbf{A} = \mathbf{ABB}^{-1}$, $\mathrm{rk}(\mathbf{AB}) \ge \mathrm{rk}((\mathbf{AB})\mathbf{B}^{-1}) = \mathrm{rk}(\mathbf{A})$.

[12] Il valore atteso di una variabile aleatoria multipla è il vettore dei valori attesi dei singoli elementi. Per la proprietà di linearità del valore atteso, se $Y = aX + b$ allora $\mathbb{E}[Y] = a\mathbb{E}[X] + b$. Ciascun elemento del vettore $\hat{\boldsymbol{\eta}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y}$ è dato dal prodotto scalare $\hat{\eta}_i = [(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}']^{(i)}\mathbf{Y}$, dove $[(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}']^{(i)}$ è la i-esima riga della matrice $p \times n$ $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$; essendo questa costante, si ha per la linearità:

$$\mathbb{E}[\hat{\eta}_i] = [(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}']^{(i)}\mathbb{E}[\mathbf{Y}]$$

e l'intero vettore $\mathbb{E}[\hat{\boldsymbol{\eta}}]$ risulta uguale a $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbb{E}[\mathbf{Y}]$.

[13] In generale, se $\mathbf{Y}$ è una variabile aleatoria multipla con matrice di varianza e covarianza $\mathrm{Cov}(\mathbf{Y})$ e $\underset{p,1}{\mathbf{Z}} = \underset{p,n}{\mathbf{C}}\ \underset{n,1}{\mathbf{Y}}$, $\mathbf{Z}$ avrà matrice di varianza e covarianza $\mathrm{Cov}(\mathbf{Z}) = \mathbf{C}\,\mathrm{Cov}(\mathbf{Y})\mathbf{C}'$. Infatti,

$$\begin{aligned}
\underset{p,p}{\mathrm{Cov}(\mathbf{Z})} &= \mathbb{E}[(\underset{p,1}{\mathbf{Z}} - \mathbb{E}[\mathbf{Z}])(\underset{1,p}{\mathbf{Z}} - \mathbb{E}[\mathbf{Z}])'] = \mathbb{E}[\,(\underset{p,n}{\mathbf{C}}\ \underset{n,1}{\mathbf{Y}} - \underset{p,n}{\mathbf{C}}\ \mathbb{E}[\underset{n,1}{\mathbf{Y}}])\,(\underset{p,n}{\mathbf{C}}\ \underset{n,1}{\mathbf{Y}} - \underset{p,n}{\mathbf{C}}\ \mathbb{E}[\underset{n,1}{\mathbf{Y}}])'\,] \\
&= \mathbb{E}[\,\underset{p,n}{\mathbf{C}}(\underset{n,1}{\mathbf{Y}} - \mathbb{E}[\mathbf{Y}])\,(\underset{1,n}{\mathbf{Y}} - \mathbb{E}[\mathbf{Y}])'\underset{n,p}{\mathbf{C}'}\,] = \underset{p,n}{\mathbf{C}}\ \mathbb{E}[(\underset{n,1}{\mathbf{Y}} - \mathbb{E}[\mathbf{Y}])\,(\underset{1,n}{\mathbf{Y}} - \mathbb{E}[\mathbf{Y}])']\ \underset{n,p}{\mathbf{C}'} \\
&= \underset{p,n}{\mathbf{C}}\ \underset{n,n}{\mathrm{Cov}(\mathbf{Y})}\ \underset{n,p}{\mathbf{C}'}
\end{aligned}$$

Inoltre, se $\mathrm{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}$, si ha:

$$\mathrm{Cov}(\mathbf{Z}) = \mathbf{C}\sigma^2\mathbf{I}\mathbf{C}' = \mathbf{C}\mathbf{C}'\sigma^2$$

Nel caso di $\hat{\boldsymbol{\eta}} = [(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}']\mathbf{Y}$, ricordando che $\mathbf{A}'\mathbf{A}$ è simmetrica (è infatti uguale alla sua trasposta:

dove $\text{Cov}(\hat{\boldsymbol{\eta}})$ e $\text{Cov}(\mathbf{Y})$ sono le matrici di varianza e covarianza di $\hat{\boldsymbol{\eta}}$ e di $\mathbf{Y}$ (indicate anche con $\text{Var}(\hat{\boldsymbol{\eta}})$ e con $\text{Var}(\mathbf{Y})$).

Da notare che *la struttura di varianza e covarianza delle stime dipende dalla matrice di riparametrizzazione*, che è sotto il controllo del ricercatore.

Esempio 1.6. Usando la matrice di dati `caffeina`:

```
> ## preparazione del dataframe
> caffeina <- read.csv("caffeina.csv")
> caffeina$tr <- as.factor(caffeina$tr)
> attach(caffeina)
> # modello
> mod <- lm(y ~ tr)
> # matrice di riparametrizzazione e sua trasposta
> A <- model.matrix(mod)
> At <- t(A)
> # stime parametri
> eta.hat <- solve(At %*% A) %*% At %*% y
> eta.hat
             [,1]
(Intercept) 244.8
tr2           1.6
tr3           3.5
```

1.4.2 Valori teorici

I *valori teorici* associati alle singole unità sperimentali si ottengono sostituendo, nel modello, i parametri con le loro stime:

$$\mathbf{Y} = \mathbf{A}\boldsymbol{\eta} \qquad \rightarrow \qquad \hat{\mathbf{Y}} = \mathbf{A}\hat{\boldsymbol{\eta}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$$

La matrice

$$\underset{n,n}{\mathbf{H}} = \underset{n,p}{\mathbf{A}}(\underset{p,n}{\mathbf{A}'}\underset{n,p}{\mathbf{A}})^{-1}\underset{p,n}{\mathbf{A}'}$$

che risulta quadrata e simmetrica, viene detta *matrice hat*, in quanto "mette il cappello" a $\mathbf{Y}$.

È un *operatore di proiezione ortogonale*, in quanto proietta $\mathbf{Y}$ sullo spazio individuato dalle colonne di $\mathbf{A}$; è quindi idempotente e ha rango uguale a quello di $\mathbf{A}$.[14]

$(\mathbf{A}'\mathbf{A})' = \mathbf{A}'(\mathbf{A}')' = \mathbf{A}'\mathbf{A}$):

$$\text{Cov}(\hat{\boldsymbol{\eta}}) = [(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'][(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}']'\sigma^2 = (\mathbf{A}'\mathbf{A})^{-1}(\mathbf{A}'\mathbf{A})(\mathbf{A}'\mathbf{A})^{-1}\sigma^2 = (\mathbf{A}'\mathbf{A})^{-1}\sigma^2$$

[14] Per la simmetria, basta dimostrare che $\mathbf{H} = \mathbf{H}'$; essendo $(\mathbf{A}'\mathbf{A})^{-1}$ simmetrica (perché è simmetrica $\mathbf{A}'\mathbf{A}$):

$$\mathbf{H}' = [\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}']' = (\mathbf{A}')'[(\mathbf{A}'\mathbf{A})^{-1}]'\mathbf{A}' = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' = \mathbf{H}$$

Per l'idempotenza:

$$\begin{aligned}\mathbf{H}\mathbf{H} &= [\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'][\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'] = \mathbf{A}[(\mathbf{A}'\mathbf{A})^{-1}(\mathbf{A}'\mathbf{A})](\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' = \mathbf{A}\mathbf{I}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' \\ &= \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' = \mathbf{H}\end{aligned}$$

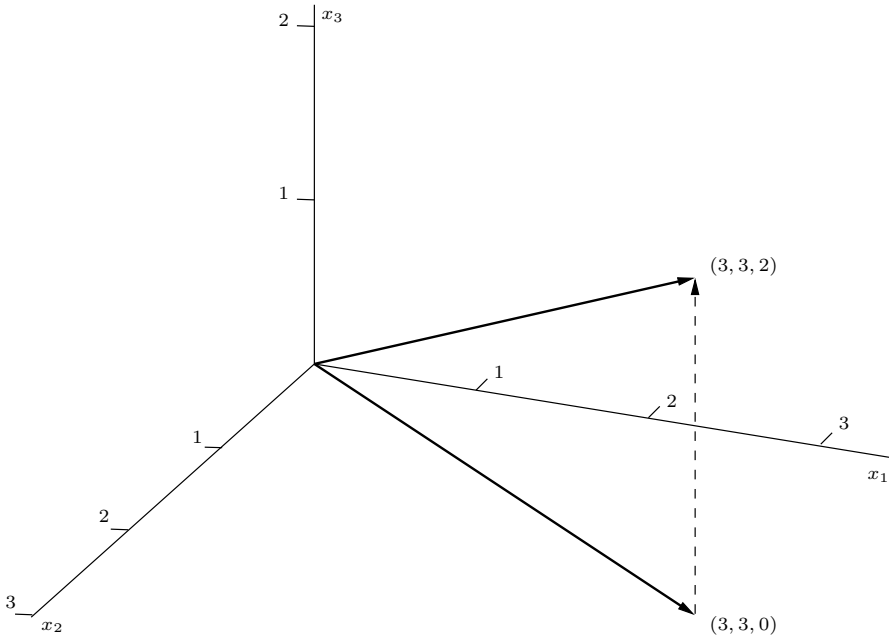


Figura 1.1. Proiezione da $\mathbb{R}^3$ sul piano $x_3 = 0$.

Esempio 1.7. Una matrice quadrata $\mathbf{P}$ è infatti una matrice di proiezione ortogonale se è idempotente, cioè se $\mathbf{P}^2 = \mathbf{PP} = \mathbf{P}$, e se è simmetrica (per un approfondimento, cfr. Appendice A). È tale, ad esempio, la matrice:

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{P}^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{P}$$

che trasforma un vettore di $\mathbb{R}^3$ nella sua proiezione ortogonale sul sottospazio $W = \{\mathbf{v} \in \mathbb{R}^3 : x_3 = 0\}$, cioè sul piano $x_3 = 0$ (figura 1.1). Si nota che $\mathbf{P}$ ha rango uguale alla dimensione di W, cioè 2 (l'immagine di una matrice, quindi dell'applicazione lineare associata, ha dimensione sempre uguale al rango della matrice, in quanto è l'insieme delle combinazioni lineari delle sue colonne). $\mathbf{P}$ è associata all'applicazione:

$$T : \mathbb{R}^3 \to \mathbb{R}^3 \qquad T\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix}$$

La trasformazione è idempotente in quanto, ovviamente, $T(x_1, x_2, 0) = (x_1, x_2, 0)$. In altri termini, una volta proiettato un vettore su un sottospazio, l'ulteriore proiezione

---

Quanto al rango, la moltiplicazione per una matrice a rango pieno non altera il rango (cfr. nota 11), quindi:

$$\mathrm{rk}(\mathbf{A}) = \mathrm{rk}[(\mathbf{A}'\mathbf{A})(\mathbf{A}'\mathbf{A})^{-1}(\mathbf{A}'\mathbf{A})]$$

in quanto il secondo termine è il rango del prodotto di matrici quadrate di ordine p tutte di rango p; ma il rango del prodotto di matrici è minore o uguale al minore dei loro ranghi, quindi:

$$\mathrm{rk}(\mathbf{A}) = \mathrm{rk}[\mathbf{A}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{A}] \leq \mathrm{rk}[\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'] \leq \mathrm{rk}(\mathbf{A})$$

Da $\mathrm{rk}(\mathbf{A}) \leq \mathrm{rk}[\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'] \leq \mathrm{rk}(\mathbf{A})$ segue $\mathrm{rk}(\mathbf{H}) = \mathrm{rk}(\mathbf{A})$.

della proiezione non cambia la proiezione (si proietta una volta sola; ulteriori proiezioni non hanno effetto).

Esempio 1.8. Proseguendo l'elaborazione iniziata nell'esempio 1.6:

```
> # matrice hat e valori teorici osservati
> H <- A %*% solve(At %*% A) %*% At
> y.hat <- H %*% y
```

I valori teorici osservati vengono comunque calcolati dalla funzione `lm()` e, se si assegna ad una variabile `mod` il risultato, si trovano in `mod$fitted.values`.

1.4.3 Variabile aleatoria "residuo"

La differenza tra la variabile risposta e i valori teorici è la variabile aleatoria *residuo*: Una volta calcolati i valori teorici, si possono osservare i *residui*, ovvero le determinazione della variabile aleatoria *residuo*, che è osservabile ed è definita come differenza tra le determinazioni di $\mathbf{Y}$ e i corrispondenti valori teorici:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{A}\hat{\boldsymbol{\eta}} = \mathbf{Y} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$$

Si tratta di una v.a. osservabile in quanto le sue determinazioni si ottengono sottraendo dai valori osservati $\mathbf{y}$ i valori teorici $\hat{\mathbf{y}}$ come calcolati sulla base delle stime.

Esempio 1.9. Proseguendo ancora l'elaborazione degli esempi 1.6 e 1.8, si possono calcolare i residui con:

```
> e <- y - y.hat
```

Comunque anche i residui sono calcolati dalla funzione `lm()` e si possono leggere in `mod$residuals`.

Si ha:

$$\mathbb{E}[\mathbf{e}] = \mathbf{0} \qquad \text{infatti:} \qquad \begin{aligned}\mathbb{E}[\mathbf{e}] &= \mathbb{E}[\mathbf{Y}] - \mathbf{A}\mathbb{E}[\hat{\boldsymbol{\eta}}] \\ &= \mathbb{E}[\mathbf{Y}] - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbb{E}[\mathbf{Y}] \\ &= \mathbf{A}\boldsymbol{\eta} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}(\mathbf{A}'\mathbf{A})\boldsymbol{\eta} = \mathbf{0}\end{aligned}$$

$$\mathrm{Cov}(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\sigma^2 \qquad \text{infatti:} \qquad \mathrm{Cov}(\mathbf{e}) = (\mathbf{I} - \mathbf{H})^2\sigma^2_{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\sigma^2_{\mathbf{y}}$$

in quanto anche la matrice $(\mathbf{I} - \mathbf{H})$ è idempotente.[15]

Quindi *la struttura di varianza e covarianza dei residui* ***non*** *riproduce l'indipendenza e l'omoshedasticita della variabile aleatoria errore*, ma dipende anch'essa, come quella delle stime, dalla matrice di riparametrizzazione.

[15] Infatti, essendo idempotenti sia $\mathbf{I}$ che $\mathbf{H}$:

$$(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H}) = \mathbf{I}^2 - \mathbf{I}\mathbf{H} - \mathbf{H}\mathbf{I} + \mathbf{H}^2 = \mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H} = \mathbf{I} - \mathbf{H}$$

Inoltre, *il vettore dei residui è incorrelato col vettore delle stime* (i due vettori sono ortogonali). Infatti:[16]

$$\begin{aligned}
\mathbf{e} &= (\mathbf{I}-\mathbf{H})\mathbf{Y} \\
\hat{\boldsymbol{\eta}} &= (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{Y} \\
\operatorname{Cov}(\mathbf{e},\hat{\boldsymbol{\eta}}) &= (\mathbf{I}-\mathbf{H})\operatorname{Cov}(\mathbf{Y})\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} = [\ (\mathbf{I}-\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}')\ (\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1})\ ]\sigma^2 \\
&= [\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} - \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}(\mathbf{A}'\mathbf{A})(\mathbf{A}'\mathbf{A})^{-1}]\sigma^2 = 0
\end{aligned}$$

Poiché i valori teorici sono funzione delle stime, *il vettore dei residui è incorrelato anche col vettore dei valori teorici*, come si verifica facilmente; ricordando che la matrice $\mathbf{H}$ è simmetrica e idempotente:

$$\begin{aligned}
\mathbf{e} &= (\mathbf{I}-\mathbf{H})\mathbf{Y} \\
\hat{\mathbf{Y}} &= \mathbf{H}\mathbf{Y} \\
\operatorname{Cov}(\mathbf{e},\hat{\mathbf{Y}}) &= (\mathbf{I}-\mathbf{H})\operatorname{Cov}(\mathbf{Y})\mathbf{H}' = [(\mathbf{I}-\mathbf{H})\mathbf{H}']\sigma^2 = (\mathbf{H}-\mathbf{H}^2)\sigma^2 = 0
\end{aligned}$$

Da un punto di vista grafico, si può notare che nella figura 1.1 il vettore del residuo, quello tratteggiato ($(3,3,2)-(3,3,0)=(0,0,2)$), è ortogonale al vettore del valore teorico, $(3,3,0)$.

L'incorrelazione è importante perché permette di utilizzare i residui per la critica e la validazione del modello.

1.4.4 Il teorema di Cochran e l'analisi della varianza

Una volta stimati i parametri e calcolati valori teorici e residui sulla base dei valori osservati della variabile risposta, si può procedere al calcolo delle seguenti quantità:

a) *devianza totale*, *SSTOT* (*total sum of squares*), è la somma dei quadrati degli scarti dei valori osservati dalla loro media:
   - in generale:

$$\begin{aligned}
SSTOT &= \sum_{i=1}^{n}(y_i-\bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\left(\frac{1}{n}\sum_{i=1} y_i\right)^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \\
&= \mathbf{y}'\mathbf{I}\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}\mathbf{y} = \mathbf{y}'\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{y}
\end{aligned}$$

     dove $\mathbf{J}$ è una matrice quadrata di ordine n i cui elementi sono tutti 1;[17]

[16] Quanto alla matrice di varianza e covarianza, ponendo $\mathbf{B}=\mathbf{I}-\mathbf{H}$ e $\mathbf{C}=(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ si ha:

$$\begin{aligned}
\operatorname{Cov}(\mathbf{e},\hat{\boldsymbol{\eta}}) &= \mathbb{E}[\ (\mathbf{e}-\mathbb{E}[\mathbf{e}])\ (\hat{\boldsymbol{\eta}}-\mathbb{E}[\hat{\boldsymbol{\eta}}])'\ ] = \mathbb{E}[\ (\mathbf{B}\mathbf{Y}-\mathbf{B}\mathbb{E}[\mathbf{Y}])\ (\mathbf{C}\mathbf{Y}-\mathbf{C}\mathbb{E}[\mathbf{Y}])'\ ] \\
&= \mathbb{E}[\ \mathbf{B}(\mathbf{Y}-\mathbb{E}[\mathbf{Y}])\ (\mathbf{Y}-\mathbb{E}[\mathbf{Y}])'\mathbf{C}'\ ] = \mathbf{B}\ \mathbb{E}[\ (\mathbf{Y}-\mathbb{E}[\mathbf{Y}])\ (\mathbf{Y}-\mathbb{E}[\mathbf{Y}])'\ ]\ \mathbf{C}' \\
&= \mathbf{B}\operatorname{Cov}(\mathbf{Y})\mathbf{C}' = (\mathbf{I}-\mathbf{H})\operatorname{Cov}(\mathbf{Y})\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}
\end{aligned}$$

[17] Il prodotto $\mathbf{y}'\mathbf{J}\mathbf{y}$ è il quadrato della somma degli elementi di $\mathbf{y}$; ad esempio, per $n=3$:

$$\begin{aligned}
\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} &= \begin{bmatrix} y_1+y_2+y_3 & y_1+y_2+y_3 & y_1+y_2+y_3 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \\
&= y_1(y_1+y_2+y_3)+y_2(y_1+y_2+y_3)+y_3(y_1+y_2+y_3) \\
&= (y_1+y_2+y_3)(y_1+y_2+y_3)
\end{aligned}$$

– se $\mathbf{y}$ è un vettore centrato, cioè se $\bar{y} = 0$:

$$SSTOT = \sum_{i=1}^{n} y_i^2 = \mathbf{y}'\mathbf{I}\mathbf{y}$$

b) la *devianza spiegata*, $SSMOD$ (*model sum of squares*), è la somma dei quadrati degli scarti tra i valori teorici e la media:

– in generale, essendo $\mathbf{H}$ simmetrica e idempotente:

$$\begin{aligned} SSMOD &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}\hat{y}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 \\ &= \mathbf{y}'\mathbf{H}'\mathbf{H}\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}\mathbf{y} = \mathbf{y}'\mathbf{H}\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}\mathbf{y} = \mathbf{y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{y} \end{aligned}$$

– se $\mathbf{y}$ è un vettore centrato:

$$SSMOD = \mathbf{y}'\mathbf{H}\mathbf{y}$$

c) la *devianza residua*, $SSRES$ (*residual sum of squares*, spesso detta un po' impropriamente SSE, *error sum of squares*),[18] è la somma dei quadrati dei residui, cioè degli scarti tra i valori osservati e quelli teorici; essendo anche $\mathbf{I} - \mathbf{H}$ simmetrica e idempotente:

$$SSRES = \mathbf{e}'\mathbf{e} = \mathbf{y}'(\mathbf{I} - \mathbf{H})'(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$$

Si nota che la devianza totale, spiegata o residua può essere espressa mediante una *forma quadratica*, ovvero un'applicazione del tipo $\mathbf{y}'\mathbf{M}\mathbf{y}$ dove $\mathbf{M}$ è una matrice simmetrica.

Si possono poi intendere le quantità osservate $\mathbf{y}'\mathbf{M}\mathbf{y}$ – dove $\mathbf{M}$ è una delle matrici $\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)$, $\mathbf{I}$, $\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)$, $\mathbf{H}$ e $\mathbf{I} - \mathbf{H}$, tutte simmetriche e idempotenti[19] – come determinazioni di variabili aleatorie del tipo $\mathbf{Y}'\mathbf{M}\mathbf{Y}$, alle quali può essere applicato il teorema di Cochran.

---

[18] La variabile aleatoria errore non è osservabile; gli scarti tra valori osservati e teorici sono determinazioni della v.a. residuo che, come visto, ha una diversa distribuzione; si usa comunque parlare di somme di quadrati dell'errore intendendo riferirsi alla devianza *attribuita* all'errore, cioè alla devianza dovuta alla componente accidentale del modello.

[19] La simmetria è evidente. Quanto all'idempotenza, si è già visto che $\mathbf{I} - \mathbf{H}$ lo è (nota 15). Per verificare che anche $\mathbf{I} - \frac{1}{n}\mathbf{J}$ è idempotente, basta osservare che lo è $\frac{1}{n}\mathbf{J}$: la matrice $\mathbf{J}$ è una matrice quadrata di ordine n i cui elementi sono tutti 1; il suo quadrato $\mathbf{J}^2$ è una matrice quadrata di ordine n i cui elementi sono tutti n, in quanto il suo generico elemento di riga r e colonna c, j_{rc}, è uguale al prodotto della r-esima riga per la c-esima colonna e questo è n. Quindi $\mathbf{J}$ *non è* idempotente. La matrice $\frac{1}{n}\mathbf{J}$ ha però come elementi tutti $\frac{1}{n}$ e il generico elemento del suo quadrato è ancora $n\frac{1}{n^2} = \frac{1}{n}$; ad esempio, per $n = 3$:

$$\begin{bmatrix}1 & 1 & 1\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix} = 1+1+1 = 3 \qquad \begin{bmatrix}1/3 & 1/3 & 1/3\end{bmatrix}\begin{bmatrix}1/3\\1/3\\1/3\end{bmatrix} = 1/9+1/9+1/9 = 3/9 = 1/3$$

Tenendo conto dell'idempotenza di $\mathbf{I}$ e di $\frac{1}{n}\mathbf{J}$:

$$\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)^2 = \mathbf{I}^2 - \frac{1}{n}\mathbf{I}\mathbf{J} - \frac{1}{n}\mathbf{J}\mathbf{I} + \left(\frac{1}{n}\mathbf{J}\right)^2 = \mathbf{I} - 2\frac{1}{n}\mathbf{J} + \frac{1}{n}\mathbf{J} = \mathbf{I} - \frac{1}{n}\mathbf{J}$$

Venendo a $\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)$, si deve considerare che se la matrice di riparametrizzazione, $\mathbf{A}$ o $\mathbf{X}$, ha tutti 1 nella prima colonna (come accade in quelle sopra considerate), allora:

Il Teorema di Cochran

Lemma. *Sia* $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ *una successione di matrici simmetriche di ordine* n *tali che* $\sum_{i=1}^{k} \mathbf{A}_i = \mathbf{A}$*, dove* $\mathbf{A}$ *sia una matrice idempotente di rango* g*. Le seguenti condizioni sono equivalenti (ciascuna implica le altre due):*

a) la somma dei ranghi delle matrici $\mathbf{A}_i$ *è* g*:* $\sum_{i=1}^{k} \mathrm{rk}(\mathbf{A}_i) = g$*;*

b) ciascuna matrice $\mathbf{A}_i$ *è idempotente:* $\mathbf{A}_i^2 = \mathbf{A}_i,\ i = 1, \ldots, k$*;*

c) il prodotto di due matrici distinte è la matrice nulla di ordine n*:* $\mathbf{A}_i'\mathbf{A}_j = \mathbf{O},\ i \neq j$*.*

Teorema di Cochran. *Sia* $\mathbf{Y}$ *una variabile aleatoria multinormale,* $\mathbf{Y} \sim MN(\boldsymbol{\mu}, \sigma^2\mathbf{I})$ *e sia* $\mathbf{A}_1, \ldots, \mathbf{A}_k$ *una successione di matrici simmetriche di rango* $g_1, \ldots, g_k$ *tali che* $\sum_{i=1}^{k} \mathbf{A}_i = \mathbf{A}$*, con* $\mathbf{A}$ *idempotente di rango* g*. Se vale una (e quindi tutte) le condizioni del lemma precedente, allora le forme quadratiche* $\mathbf{Y}'\mathbf{A}_i\mathbf{Y}$*, divise per* σ^2*, sono distribuite come Chi quadrati non centrati indipendenti:*

$$\frac{\mathbf{Y}'\mathbf{A}_i\mathbf{Y}}{\sigma^2} \sim \chi^2_{g_i,\lambda_i} \qquad\qquad \lambda_i = \frac{\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu}}{\sigma^2}$$

Esempio 1.10. In un modello regressivo le relazioni già viste si scrivono:

a) modello: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$;

b) stimatori dei parametri: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$;

c) valori teorici: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$;

d) residui: $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$.

Un esempio molto semplice potrebbe essere il seguente. Per $\mathbf{x} = (1, 2, 3)$ si osservano i valori $\mathbf{y} = (2.9, 5.2, 6.9)$; si costruisce pertanto il modello:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad\qquad \begin{bmatrix} 2.9 \\ 5.2 \\ 6.9 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \boldsymbol{\varepsilon}$$

La stima dei parametri conduce ai valori:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2.9 \\ 5.2 \\ 6.9 \end{bmatrix} = \begin{bmatrix} \hat{\alpha} = 1 \\ \hat{\beta} = 2 \end{bmatrix}$$

– da $\mathbf{H}\mathbf{A} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{A} = \mathbf{A}$ segue che le somme di riga di $\mathbf{H}$ sono tutte pari a 1; ad esempio, considerando la i-esima riga di una matrice $\mathbf{H}$ di ordine 3 e la prima colonna di una $\mathbf{A}$:

$$\begin{bmatrix} 0.5 & 0.3 & 0.2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 0.5 + 0.3 + 0.2 = 1$$

e quindi sarà 1 l'i-esimo elemento della prima colonna del prodotto, come deve essere;

– analogamente, da $\mathbf{A}'\mathbf{H} = \mathbf{A}'\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}' = \mathbf{A}'$ segue che le somme di colonna di $\mathbf{H}$ sono pari a 1.

Da ciò segue che $\mathbf{H}\left(\frac{1}{n}\mathbf{J}\right) = \frac{1}{n}\mathbf{J}$ in quanto l'elemento di indici ij del prodotto è uguale al prodotto di una riga di somma 1 e di una colonna di tutti $1/n$, quindi è uguale alla media degli n elementi della riga, che è appunto $1/n$. Valendo 1 anche le somme di colonna di $\mathbf{H}$, si ha anche $\left(\frac{1}{n}\mathbf{J}\right)\mathbf{H} = \frac{1}{n}\mathbf{J}$. Quindi:

$$\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)^2 = \mathbf{H}^2 - \mathbf{H}\left(\frac{1}{n}\mathbf{J}\right) - \left(\frac{1}{n}\mathbf{J}\right)\mathbf{H} + \left(\frac{1}{n}\mathbf{J}\right)^2 = \mathbf{H} - 2\frac{1}{n}\mathbf{J} + \frac{1}{n}\mathbf{J} = \mathbf{H} - \frac{1}{n}\mathbf{J}$$

da cui i valori teorici e i residui:

$$\begin{aligned} \hat{y}_1 &= \hat{\alpha} + \hat{\beta}x_1 = 1 + 2 \cdot 1 = 3 & e_1 &= y_1 - \hat{y}_1 = 2.9 - 3 = -0.1 \\ \hat{y}_2 &= \hat{\alpha} + \hat{\beta}x_2 = 1 + 2 \cdot 2 = 5 & e_2 &= y_2 - \hat{y}_2 = 5.2 - 5 = 0.2 \\ \hat{y}_3 &= \hat{\alpha} + \hat{\beta}x_3 = 1 + 2 \cdot 3 = 7 & e_3 &= y_3 - \hat{y}_3 = 6.9 - 7 = -0.1 \end{aligned}$$

La devianza totale è:

$$SSTOT = \sum_{i=1}^{3}(y_i - \bar{y})^2 = (2.9-5)^2 + (5.2-5)^2 + (6.9-5)^2 = 8.06$$

in forma matriciale:

$$\begin{aligned} SSTOT &= \mathbf{y}'\left(\mathbf{I} - \frac{1}{3}\mathbf{J}\right)\mathbf{y} = \mathbf{y}'\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}\right)\mathbf{y} \\ &= \begin{bmatrix} 2.9 & 5.2 & 6.9 \end{bmatrix} \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix} \begin{bmatrix} 2.9 \\ 5.2 \\ 6.9 \end{bmatrix} = 8.06 \end{aligned}$$

Si verifica facilmente che la matrice $\mathbf{A} = \left(\mathbf{I} - \frac{1}{3}\mathbf{J}\right) = \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix}$ è idempotente e che ha rango 2. Si calcolano analogamente la devianza spiegata:

$$\begin{aligned} SSMOD &= \mathbf{y}'\left(\mathbf{H} - \frac{1}{3}\mathbf{J}\right)\mathbf{y} = \mathbf{y}'\left(\begin{bmatrix} 5/6 & 1/3 & -1/6 \\ 1/3 & 1/3 & 1/3 \\ -1/6 & 1/3 & 5/6 \end{bmatrix} - \begin{bmatrix} 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{bmatrix}\right)\mathbf{y} \\ &= \begin{bmatrix} 2.9 & 5.2 & 6.9 \end{bmatrix} \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 2.9 \\ 5.2 \\ 6.9 \end{bmatrix} = 8 \end{aligned}$$

e la devianza residua:

$$SSRES = \begin{bmatrix} 2.9 & 5.2 & 6.9 \end{bmatrix} \begin{bmatrix} 1/6 & -1/3 & 1/6 \\ -1/3 & 2/3 & -1/3 \\ 1/6 & -1/3 & 1/6 \end{bmatrix} \begin{bmatrix} 2.9 \\ 5.2 \\ 6.9 \end{bmatrix} = 0.06$$

Si nota che $SSTOT = SSMOD + SSRES$ e si verifica facilmente che le matrici:

$$\mathbf{A}_1 = \left(\mathbf{H} - \frac{1}{3}\mathbf{J}\right) = \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \end{bmatrix} \qquad \mathbf{A}_2 = (\mathbf{I} - \mathbf{H}) = \begin{bmatrix} 1/6 & -1/3 & 1/6 \\ -1/3 & 2/3 & -1/3 \\ 1/6 & -1/3 & 1/6 \end{bmatrix}$$

- sommate insieme danno la matrice $\mathbf{A}$: $\mathbf{A}_1 + \mathbf{A}_2 = \mathbf{A}$;
- hanno entrambe rango 1, quindi la somma dei loro ranghi è uguale al rango di $\mathbf{A}$;
- sono entrambe idempotenti: $\mathbf{A}_1^2 = \mathbf{A}_1$ e $\mathbf{A}_2^2 = \mathbf{A}_2$;
- moltiplicate tra loro danno la matrice nulla: $\mathbf{A}_1\mathbf{A}_2 = \mathbf{O}$.

Ne segue, per il teorema di Cochran, che le variabili aleatorie $\mathbf{Y}'\mathbf{A}_1\mathbf{Y}$ e $\mathbf{Y}'\mathbf{A}_2\mathbf{Y}$ (le cui determinazioni sono, rispettivamente, $SSMOD$ e $SSRES$) sono indipendenti e distribuite come Chi quadrati non centrati.

Esempio 1.11. Si usa spesso "centrare" i dati, cioè sostituirli con i loro scarti dalla media aritmetica. Ciò può essere fatto moltiplicando un vettore di n valori per una *matrice di centratura*, che altro non è che la matrice $\mathbf{I} - \frac{1}{n}\mathbf{J}$ già vista:

$$\left(\mathbf{I} - \frac{1}{3}\mathbf{J}\right)\mathbf{y} = \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix} \begin{bmatrix} 2.9 \\ 5.2 \\ 6.9 \end{bmatrix} = \begin{bmatrix} -2.1 \\ 0.2 \\ 1.9 \end{bmatrix}$$

$$\mathbf{y}'\left(\mathbf{I} - \frac{1}{3}\mathbf{J}\right) = \begin{bmatrix} 2.9 & 5.2 & 6.9 \end{bmatrix} \begin{bmatrix} 2/3 & -1/3 & -1/3 \\ -1/3 & 2/3 & -1/3 \\ -1/3 & -1/3 & 2/3 \end{bmatrix} = \begin{bmatrix} -2.1 & 0.2 & 1.9 \end{bmatrix}$$

Poiché $\mathbf{I} - \frac{1}{n}\mathbf{J}$ è simmetrica e idempotente, la devianza totale come calcolata nell'esempio precedente può essere riformulata come segue:

$$SSTOT = \mathbf{y}'\left(\mathbf{I} - \frac{1}{3}\mathbf{J}\right)\mathbf{y} = \mathbf{y}'\left(\mathbf{I} - \frac{1}{3}\mathbf{J}\right)'\left(\mathbf{I} - \frac{1}{3}\mathbf{J}\right)\mathbf{y} = \bar{\mathbf{y}}'\bar{\mathbf{y}} = 8.06$$

dove $\bar{\mathbf{y}}$ è un vettore centrato, il vettore degli scarti dalla media dei valori osservati della variabile risposta. Si deve sottolineare che $\bar{\mathbf{y}}$ *non è una traslazione* di $\mathbf{y}$, ma una *proiezione* da uno spazio di dimensione 3 (quello cui appartiene $\mathbf{y}$) ad uno di dimensione 2 (perché 2 è il rango di $\mathbf{I} - \frac{1}{3}\mathbf{J}$).[20] Per verificarlo, è sufficiente calcolare una base dell'immagine di $\mathbf{I} - \frac{1}{3}\mathbf{J}$, che può essere $\{(1, 0, -1), (0, 1, -1)\}$, e si ha:

$$\begin{bmatrix} -2.1 \\ 0.2 \\ 1.9 \end{bmatrix} = -2.1 \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} + 0.2 \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix}$$

Analogamente per la devianza spiegata. La matrice $\mathbf{H} - \frac{1}{n}\mathbf{J}$, essendo simmetrica ed idempotente come $\mathbf{H}$ e $\frac{1}{n}\mathbf{J}$ ed essendo $\mathbf{H}\frac{1}{n}\mathbf{J} = \frac{1}{n}\mathbf{J}$ (v. nota 19), può essere vista come:

$$\mathbf{H} - \frac{1}{n}\mathbf{J} = \mathbf{H}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{H}$$

Si ha inoltre che le matrici $\mathbf{H}$ e $\frac{1}{n}\mathbf{J}$ commutano (il loro prodotto è commutativo).[21] Quindi:

$$\begin{aligned} SSMOD &= \mathbf{y}'\left(\mathbf{H} - \frac{1}{3}\mathbf{J}\right)\mathbf{y} = \mathbf{y}'\mathbf{H}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{H}\mathbf{y} \\ &= \left[\mathbf{y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)'\right]\mathbf{H}'\mathbf{H}\left[\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{y}\right] = \bar{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}} = 8 \end{aligned}$$

[20] Si avrebbe una traslazione se $\bar{\mathbf{y}}$ fosse un vettore di scarti da una costante data, ma la media dei valori di $\mathbf{y}$ è un valore calcolato a partire da questi stessi valori; è questo il motivo per cui si ha una riduzione dimensionale.

[21] Ciò avviene perché sono *simultaneamente diagonalizzabili*: esiste una matrice $\mathbf{P}$ tale che $\mathbf{P}^{-1}\mathbf{H}\mathbf{P}$ e $\mathbf{P}^{-1}\left(\frac{1}{n}\mathbf{J}\right)\mathbf{P}$ sono entrambe diagonali. La matrice $\mathbf{P}$ può essere ottenuta ortogonalizzando e normalizzando gli autovettori di $\mathbf{H}$, oppure di $\frac{1}{n}\mathbf{J}$.

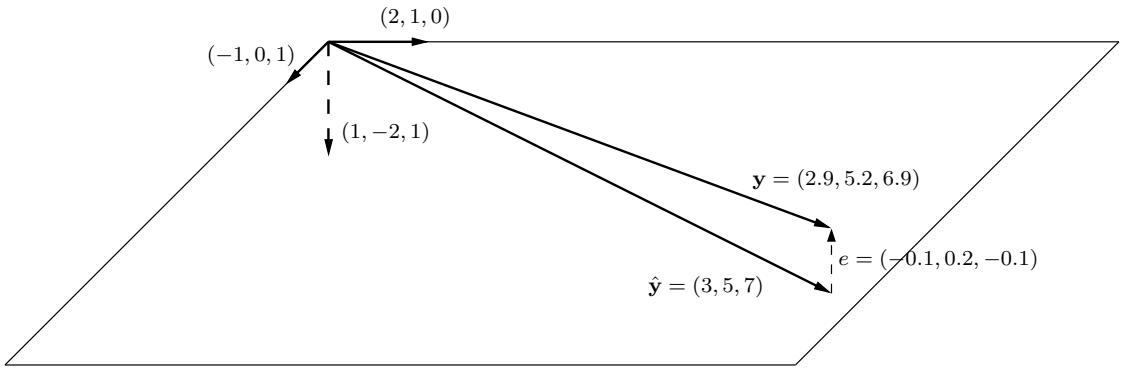


Figura 1.2. Interpretazione geometrica del teorema di Cochran.

Quanto alla devianza residua, si verifica facilmente che:

$$\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)' (\mathbf{I} - \mathbf{H}) \left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) = \mathbf{I} - \mathbf{H}$$

quindi

$$SSRES = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \bar{\mathbf{y}}'(\mathbf{I} - \mathbf{H})\bar{\mathbf{y}} = 0.06$$

Si hanno così le matrici $\mathbf{A} = \mathbf{I}$, $\mathbf{A}_1 = \mathbf{H}$ e $\mathbf{A}_2 = \mathbf{I} - \mathbf{H}$ e anche in questo caso le due matrici $\mathbf{A}_1$ e $\mathbf{A}_2$:

- sommate insieme danno la matrice $\mathbf{A}$: $\mathbf{H} + (\mathbf{I} - \mathbf{H}) = \mathbf{I}$;
- hanno rispettivamente rango 2 e rango 1, quindi la somma dei loro ranghi è uguale al rango di $\mathbf{A}$ (che ora è $\mathrm{rk}(\mathbf{I}) = 3$);
- sono entrambe idempotenti;
- moltiplicate tra loro danno la matrice nulla: $\mathbf{H}(\mathbf{I} - \mathbf{H}) = \mathbf{H} - \mathbf{H} = \mathbf{O}$.

Ne segue, per il teorema di Cochran, che le variabili aleatorie $\bar{\mathbf{Y}}'\mathbf{H}\bar{\mathbf{Y}}$ e $\bar{\mathbf{Y}}'(\mathbf{I} - \mathbf{H})\bar{\mathbf{Y}}$ (le cui determinazioni sono, rispettivamente, $SSMOD$ e $SSRES$) sono indipendenti e distribuite come Chi quadrati non centrati.

Osservazione. Il teorema di Cochran ha un'interessante interpretazione geometrica (cfr. figura 1.2). Tornando ai dati dell'esempio precedente, si può osservare che $\mathbf{y}$ (il vettore dei valori osservati della variabile risposta) appartiene allo spazio $\mathbb{R}^3$. Il vettore dei valori teorici, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, appartiene invece ad un sottospazio di $\mathbb{R}^3$ di dimensione 2, in quanto la matrice $\mathbf{H}$ è di rango 2. In particolare, una base dell'immagine di $\mathbf{H}$ è costituita dai vettori $(2, 1, 0)$ e $(-1, 0, 1)$. L'immagine di una matrice è una combinanzione lineare delle sue colonne; una base può quindi trovarsi individuando le colonne linearmente indipendenti. Dato però che $\mathbf{H}$ è una matrice simmetrica, è possibile e conveniente diagonalizzarla, pervenendo a $\mathbf{H} = \mathbf{M}\mathbf{\Lambda}\mathbf{M}^{-1}$:

$$\begin{bmatrix} 5/6 & 1/3 & -1/6 \\ 1/3 & 1/3 & 1/3 \\ -1/6 & 1/3 & 5/6 \end{bmatrix} = \begin{bmatrix} 2 & -1 & 1 \\ 1 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 2 & -1 & 1 \\ 1 & 0 & -2 \\ 0 & 1 & 1 \end{bmatrix}^{-1}$$

Si ottengono così tre autovettori (le colonne di $\mathbf{M}$), i primi due dei quali, essendo non nulli i relativi autovalori, costituiscono una base dell'immagine. Si nota anche che il terzo vettore

(una base del kernel) è ortogonale ai primi due, che generano il piano cui appartiene il vettore $\hat{\mathbf{y}}$:

$$\begin{bmatrix} 3 \\ 5 \\ 7 \end{bmatrix} = 5 \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} + 7 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

I residui appartengono invece allo spazio immagine della matrice $\mathbf{I} - \mathbf{H}$; diagonalizzando:

$$\begin{bmatrix} 1/6 & -1/3 & 1/6 \\ -1/3 & 2/3 & -1/3 \\ 1/6 & -1/3 & 1/6 \end{bmatrix} = \begin{bmatrix} 1 & 2 & -1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & -1 \\ -2 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}^{-1}$$

si ritrovano gli stessi autovettori, ma ora c'è un solo autovalore non nullo e il relativo autovettore, che costituisce una base dell'immagine, è ortogonale agli altri due. Si vede così che il vettore dei residui, $(-0.1, 0.2, -0.1) = -\frac{1}{10}(1, -2, 1)$, appartiene ad uno spazio ad una dimensione *ortogonale* a quello di cui è elemento il vettore delle stime. La scomposizione della devianza può quindi essere rappresentata come scomposizione dello spazio del fenomeno osservato in sottospazi ortogonali; si può dire che si parte da uno spazio $\mathbb{R}^3$ con base i tre autovettori e che questo viene scomposto in uno spazio di dimensione 2 (di cui è elemento il vettore dei valori teorici dati dal modello) ed in uno spazio di dimensione 1 (di cui è elemento il vettore dei residui). Inoltre, essendo i due spazi ortogonali, la devianza del modello e quella dei residui sono indipendenti (i motivi per cui $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ è una matrice di proiezione ortogonale sullo spazio generato dalle colonne di $\mathbf{X}$, mentre $\mathbf{I} - \mathbf{H}$ è una matrice di proiezione sul complemento ortogonale di quello spazio, sono illustrati nell'Appendice A).

L'analisi della varianza

La varianza, in ambito inferenziale, è data dalla devianza divisa per i gradi di libertà.

Se vi sono n unità sperimentali (la matrice dei dati ha n righe), i gradi di libertà della devianza totale sono $n - 1$, in quanto una volta dati $n - 1$ scarti dalla media $\bar{y}$ l'n-esimo scarto è univocamente determinato (la somma degli scarti dalla media è 0).

Se, a seguito della riparametrizzazione, il modello comprende p parametri, i gradi di libertà del modello sono $p - 1$. In un modello ANOVA, infatti, vi sono p trattamenti e altrettante medie di trattamento μ_i, ma, dati $p - 1$ loro scarti dalla media generale, il p-esimo scarto risulta univocamente determinato. In un modello regressivo, invece, ciascuna media μ_i è data dalla somma di un parametro α (detto *intercetta*) e di prodotti di parametri β per valori delle variabili esplicative, quindi i gradi di libertà sono tanti quante sono queste (nell'esempio precedente, $\mu_i = \alpha + \beta x_i$, quindi due parametri ma un solo grado di libertà).

I gradi di libertà della variabile residuo (quelli attribuibili all'errore, alla componente accidentale) sono $n - p$: $(n - 1) - (p - 1) = n - p$.

Come in parte anticipato nell'esempio precedente, i gradi di libertà coincidono con i ranghi delle matrici che intervengono nelle forme quadratiche che esprimono le devianze totale, spiegata e residua. Infatti la matrice $\mathbf{I} - \frac{1}{n}\mathbf{J}$ (devianza totale) ha rango $n - 1$, la matrice $\mathbf{H} - \frac{1}{n}\mathbf{J}$ (devianza spiegata) ha rango $p - 1$ e la matrice $\mathbf{I} - \mathbf{H}$ (devianza residua) ha rango $n - p$.[22]

[22] Per qualsiasi matrice idempotente il rango è uguale alla traccia (v. Appendice A, proposizione A.22). In

Si può quindi applicare il teorema di Cochran:

$$SSMOD = \mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} \sim \sigma^2\chi^2_{p-1,\lambda_m} \qquad \lambda_m = \frac{\boldsymbol{\mu}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\boldsymbol{\mu}}{\sigma^2}$$

$$SSRES = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \sim \sigma^2\chi^2_{n-p,\lambda_r} \qquad \lambda_r = \frac{\boldsymbol{\mu}'(\mathbf{I} - \mathbf{H})\boldsymbol{\mu}}{\sigma^2}$$

essendo $SSMOD$ e $SSRES$ *indipendenti*.

Si tratta di Chi quadrati non centrati con parametri di non centralità ignoti, in quanto non è nota σ^2. Non è noto neppure il vettore $\boldsymbol{\mu}$, ma la riparametrizzazione lo ha sostituito con $\mathbf{A}\boldsymbol{\eta}$ ($\mathbf{X}\boldsymbol{\beta}$).

L'analisi della varianza ha comunque lo scopo di sottoporre a verifica l'ipotesi nulla H_0 secondo la quale la variabilità del fenomeno è dovuta solo a fattori accidentali. Ciò vuol dire che, sia in modelli ANOVA del tipo $\mu_i = \mu + \alpha_i$, sia in modelli regressivi del tipo $\mu_i = \alpha + \beta x_i$ (ovviamente anche con più di due parametri), nell'ipotesi nulla si assume $\mu_i = \mu$, cioè che $\mathbf{A}\boldsymbol{\eta}$ ($\mathbf{X}\boldsymbol{\beta}$) sia un vettore di elementi tutti uguali tra loro.

In questo caso i parametri di non centralità si annullano (si annulla il loro numeratore), in quanto le matrici $\mathbf{H} - \frac{1}{n}\mathbf{J}$ e $\mathbf{I} - \mathbf{H}$ hanno somme di riga e di colonna pari a 0.[23]

Si ha così che, ai fini di una verifica dell'ipotesi nulla:

$$\frac{SSMOD}{\sigma^2} = \frac{\mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}}{\sigma^2} \sim \chi^2_{p-1}$$

$$\frac{SSRES}{\sigma^2} = \frac{\mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}}{\sigma^2} \sim \chi^2_{n-p}$$

---

particolare, poiché $\mathbf{H}$ ha rango p (il numero di colonne della matrice di riparametrizzazione, eventualmente resa a rango pieno mediante un vincolo) e $\frac{1}{n}\mathbf{J}$ ha rango 1 (ha righe e colonne tutte uguali), si ha:

$$\mathrm{tr}(\mathbf{I}) = \mathrm{rk}(\mathbf{I}) = n \qquad \mathrm{tr}(\mathbf{H}) = \mathrm{rk}(\mathbf{H}) = p \qquad \mathrm{tr}\left(\frac{1}{n}\mathbf{J}\right) = \mathrm{rk}\left(\frac{1}{n}\mathbf{J}\right) = 1$$

Dal momento che, in generale, $\mathrm{tr}(a\mathbf{A} + b\mathbf{B}) = a\,\mathrm{tr}(\mathbf{A}) + b\,\mathrm{tr}(\mathbf{B})$, si ha:

$$\mathrm{rk}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) = \mathrm{tr}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) = \mathrm{tr}(\mathbf{I}) - \mathrm{tr}\left(\frac{1}{n}\mathbf{J}\right) = n - 1$$

$$\mathrm{rk}\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right) = \mathrm{tr}\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right) = \mathrm{tr}(\mathbf{H}) - \mathrm{tr}\left(\frac{1}{n}\mathbf{J}\right) = p - 1$$

$$\mathrm{rk}(\mathbf{I} - \mathbf{H}) = \mathrm{tr}(\mathbf{I} - \mathbf{H}) = \mathrm{tr}(\mathbf{I}) - \mathrm{tr}(\mathbf{H}) = n - p$$

[23] Sia $\mathbf{H}$ che $\mathbf{I}$ hanno somme di riga e di colonna pari a 1 (per $\mathbf{H}$ v. nota 19), ma anche $\frac{1}{n}\mathbf{J}$; quindi le matrici differenza, anch'esse simmetriche, hanno somme di riga e di colonna pari a 0. Premoltiplicando e postmoltiplicando tali matrici per vettori con elementi tutti uguali si ottiene 0. Con i dati dell'esempio 1.10, ponendo per ipotesi nulla $\mu_i = \alpha = \mu$, quindi $\beta = 0$, si ha:

$$H_0 : \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \mu \\ 0 \end{bmatrix} = \begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix}$$

Il numeratore del parametro di non centralità per $SSMOD$ è quindi:

$$\begin{bmatrix} \mu & \mu & \mu \end{bmatrix} \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} \mu \\ \mu \\ \mu \end{bmatrix} = \begin{bmatrix} \mu & \mu & \mu \end{bmatrix} \begin{bmatrix} 1/2\mu + 0 - 1/2\mu \\ 0 \\ -1/2\mu + 0 + 1/2\mu \end{bmatrix} = \begin{bmatrix} \mu & \mu & \mu \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = 0$$

e si può costruire la statistica test:

$$\frac{\frac{SSMOD}{\sigma^2}/(p-1)}{\frac{SSRES}{\sigma^2}/(n-p)} = \frac{\frac{SSMOD}{p-1}}{\frac{SSRES}{n-p}} \sim F_{p-1,n-p}$$

Si vede che $\frac{SSMOD}{p-1}$ è la varianza del modello, indicata con $MSMOD$ (*mean square* invece di *sum of squares*) ovvero la devianza spiegata da ciascun grado di libertà del modello, mentre $\frac{SSRES}{n-p}$ è la varianza dei residui, $MSRES$, la devianza che compete a ciascun grado di libertà dei residui. Si può quindi scrivere:

$$\frac{MSMOD}{MSRES} \sim F_{p-1,n-p}$$

Se le due varianze non sono troppo diverse, in particolare se la prima non è troppo maggiore della seconda, si può concludere che la variabilità che si vorrebbe spiegata dal modello non è diversa da quella attribuibile al caso, quindi si accetta l'ipotesi nulla. Se invece la varianza del modello è significativamente maggiore di quella dei residui, si può rifiutare l'ipotesi nulla in favore dell'ipotesi alternativa: i diversi trattamenti (ANOVA) o i diversi valori delle variabili esplicative (regressione) hanno un effetto significativo sui valori della variabile risposta.

Il test basato sulla F di Snedecor consente di quantificare espressioni altrimenti vaghe come "significativamente maggiore", in modo simile ad un familiare confronto tra le varianze di due campioni.

Osservazione. Si può dimostrare che:

$$\mathbb{E}[MSRES] = \mathbb{E}\left[\frac{SSRES}{n-p}\right] = \mathbb{E}\left[\frac{\mathbf{e}'\mathbf{e}}{n-p}\right] = \sigma^2$$

Infatti, tenendo presente che $\mathbf{e}'\mathbf{e}$ è uno scalare, che la traccia di uno scalare è lo scalare stesso e che in generale, quale che sia $\mathbf{e}$, si ha $\mathrm{tr}(\mathbf{e}'\mathbf{e}) = \mathrm{tr}(\mathbf{e}\mathbf{e}')$,[24] si può scrivere (tenendo presente che la traccia non è altro che una somma):

$$\mathbb{E}[\mathbf{e}'\mathbf{e}] = \mathbb{E}[\mathrm{tr}(\mathbf{e}'\mathbf{e})] = \mathbb{E}[\mathrm{tr}(\mathbf{e}\mathbf{e}')] = \mathrm{tr}(\mathbb{E}[\mathbf{e}\mathbf{e}']) = \mathrm{tr}(\mathrm{Cov}(\mathbf{e})) = \mathrm{tr}(\mathbf{I}-\mathbf{H})\sigma^2 = (n-p)\sigma^2$$

in quanto la traccia di $\mathbf{I}-\mathbf{H}$ è uguale al suo rango, che è appunto $n-p$ (nota 22). Ne segue:

$$\mathbb{E}[MSRES] = \mathbb{E}\left[\frac{SSRES}{n-p}\right] = \sigma^2$$

[24] Ad esempio, se $\mathbf{e} = (a, b, c)$ si ha:

$$\mathbf{e}'\mathbf{e} = \begin{bmatrix} a & b & c \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = a^2 + b^2 + c^2 \qquad \mathrm{tr}(a^2 + b^2 + c^2) = a^2 + b^2 + c^2$$

ed anche:

$$\mathbf{e}\mathbf{e}' = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \begin{bmatrix} a & b & c \end{bmatrix} = \begin{bmatrix} a^2 & ab & ac \\ ba & b^2 & bc \\ ca & cb & c^2 \end{bmatrix} \qquad \mathrm{tr}\left(\begin{bmatrix} a^2 & ab & ac \\ ba & b^2 & bc \\ ca & cb & c^2 \end{bmatrix}\right) = a^2 + b^2 + c^2$$

ovvero che *la varianza dei residui è uno stimatore corretto di* σ^2. Analogamente, si può dimostrare che:[25]

$$\mathbb{E}[MSMOD] = \mathbb{E}\left[\frac{SSMOD}{p-1}\right] = \mathbb{E}\left[\frac{\mathbf{Y}'\left(\mathbf{H}-\frac{1}{n}\mathbf{J}\right)\mathbf{Y}}{p-1}\right] = \sigma^2 + \frac{\mathbf{Y}'\left(\mathbf{H}-\frac{1}{n}\mathbf{J}\right)\mathbf{Y}}{p-1}$$

La prima uguaglianza afferma che il valore atteso della varianza residua è uguale alla varianza dell'errore e ciò risulta intuitivamente ragionevole. Quanto alla seconda, basta ricordare che:

$$\mathbf{Y}'\left(\mathbf{H}-\frac{1}{n}\mathbf{J}\right)\mathbf{Y} = \mathbf{Y}'\mathbf{H}\mathbf{Y} - \frac{1}{n}\mathbf{Y}'\mathbf{J}\mathbf{Y}$$

non è altro che la somma dei quadrati degli scarti dei valori teorici dalla media; ne segue che il valore atteso della varianza spiegata è tanto maggiore di σ^2 quanto più i valori teorici sono diversi dalla media generale, cioè quanto più la diversità dei valori teorici dalla media spiega la variabilità del fenomeno.

[25] Se $\mathbf{Y}$ è un vettore centrato, $\mathbb{E}[MSMOD] = \sigma^2 + \frac{\mathbf{Y}'\mathbf{H}\mathbf{Y}}{p} = \sigma^2 + \frac{\dot{\boldsymbol{\eta}}'\mathbf{A}'\mathbf{A}\dot{\boldsymbol{\eta}}}{p}$.

Capitolo 2

Il modello ANOVA

Il modello ANOVA consente di analizzare i risultati di un esperimento quale che sia la natura delle variabili esplicative, in particolare anche quando queste sono qualitative.

La sezione 2.1 illustra il modello a un solo fattore, mostrando in dettaglio come si usa il teorema di Cochran per la verifica della significatività del modello, come si conduca analoga verifica per i singoli parametri e come si determinino i loro intervalli di confidenza.

La sezione 2.2 discute i modelli a due fattori relativi ad esperimenti completi e bilanciati, nei quali oltre agli effetti dei singoli fattori può essere presente un ulteriore *effetto interattivo* dovuto alla somministrazione simultanea dei due fattori. Si mostrano sia le differenze nella stima dei parametri e nell'analisi della varianza che ne conseguono, sia le azioni da intraprendere se l'effetto interattivo risulta non significativo. La sezione 2.3 tratta dei modelli a tre o più fattori; dato che non vi sono differenze sostanziali rispetto ai precedenti, si mostrano soprattutto le tecniche per la stima dei parametri e per la semplificazione del modello nel caso alcuni effetti interattivi risultassero non significativi.

La sezione 2.4 si occupa degli esperimenti a blocchi randomizzati, che tendono a depurare la varianza residua della quota di variabilità attribuibile alla eterogeneità delle unità sperimentali, rendendo così più affidabile il test di ipotesi sul modello.

La sezione 2.5 tratta, infine, degli esperimenti non bilanciati.

2.1 Esperimenti con un solo fattore

Vi sono t trattamenti, consistenti nella somministrazione di un unico fattore in t livelli, ciascuno contrassegnato da un indice $i = 1, \ldots, t$. Ciascun trattamento viene assegnato a n_i unità sperimentali, quindi $\sum_{i=1}^{t} n_i = n$ è il numero complessivo delle unità sperimentali. Nell'ambito di ciascun trattamento vi sono quindi n_i repliche, ciascuna contrassegnata da un indice $r = 1, \ldots, n_i$ (disegno completamente randomizzato).

Il modello ANOVA più immediato, detto modello *a medie di cella* (*cell means model*) è:

$$Y_{ir} = \mu_i + \varepsilon_{ir} \qquad i = 1, \ldots, t \qquad r = 1, \ldots, n_i$$

dove:

– Y_{ir} è il valore della variabile risposta nella r-esima replica per l'i-esimo trattamento (per l'i-esimo livello dell'unico fattore);

- μ_i sono i t parametri, da intendere come le t medie della variabile risposta corrispondenti ai t trattmenti;[1]
- ε_{ir} è una variabile aleatoria "errore"; le ε_{ir} hanno tutte distribuzione normale con varianza costante (omoschedasticità), $\varepsilon_{ir} \sim N(0, \sigma^2)$, e sono a due a due indipendenti; in altri termini, per la variabile aleatoria multipla "errore" si ha: $\boldsymbol{\varepsilon} \sim MN(0, \sigma^2\mathbf{I})$;
- per ogni Y_{ir} si ha $Y_{ir} \sim N(\mu_i, \sigma^2)$.

Esempio 2.1. Con riferimento alla matrice dei dati contenuta nel file `caffeina.csv`:[2]

- vi sono $n = 30$ unità sperimentali (la matrice ha 30 righe);
- vi sono $t = 3$ trattamenti (nella colonna `tr` compaiono le modalità 1, 2 e 3), quindi $i = 1, 2, 3$;
- ciascun trattamento (una dose di caffeina) viene somministrato a $n_1 = n_2 = n_3 = 10$ unità (disegno bilanciato); le unità cui viene somministrato uno stesso trattamento costituiscono un gruppo; in ogni gruppo vi sono 10 repliche, quindi $r = 1, \ldots, 10$;
- si ipotizza che ciascun gruppo di 10 unità abbia una propria media; le medie osservate (quindi anche le stime dei parametri μ_i) sono:[3]

$$\hat{\mu}_1 = 244.8 \qquad \hat{\mu}_2 = 246.4 \qquad \hat{\mu}_3 = 248.3$$

- si ipotizza quindi che la colonna della variabile risposta della matrice dei dati contenga le determinazioni di una variabile aleatoria normale multivariata $\mathbf{Y}$ e che per ciascuna Y_{ir} si abbia:

$$Y_{ir} = \mu_i + \varepsilon_{ir}, \qquad \varepsilon_{ir} \sim N(0, \sigma^2) \qquad Y_{ir} \sim N(\mu_i, \sigma^2)$$

  ovvero che, essendo la varianza σ^2 unica, vi siano tre funzioni di densità di probabilità che differiscano solo per la media (v. figura 2.1);
- si tratta di un modello lineare, in quanto può essere espresso nella forma $\mathbf{Y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$:

$$\mathbf{Y} = \begin{bmatrix} Y_{1,1} \\ \vdots \\ Y_{1,10} \\ Y_{2,1} \\ \vdots \\ Y_{2,10} \\ Y_{3,1} \\ \vdots \\ Y_{3,10} \end{bmatrix} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,10} \\ \varepsilon_{2,1} \\ \vdots \\ \varepsilon_{2,10} \\ \varepsilon_{3,1} \\ \vdots \\ \varepsilon_{3,10} \end{bmatrix}$$

- i valori osservati della variabile risposta, y_{ir} (colonna `y` della matrice dei dati), vengono

[1] Per questo il modello viene detto "a medie di cella". I parametri possono anche essere interpretati in modo diverso, $\mu_i = \mu + \alpha_i$, come si fa nei modelli *a effetti dei fattori* (sez. 2.1.5).

[2] `http://web.mclink.it/MC1166/ModelliStatistici/caffeina.csv`.

[3] Si ottengono in R con `by(caffeina$y, caffeina$tr, mean)`, in SAS con:
`proc means data=caffeina; by tr; run;`.

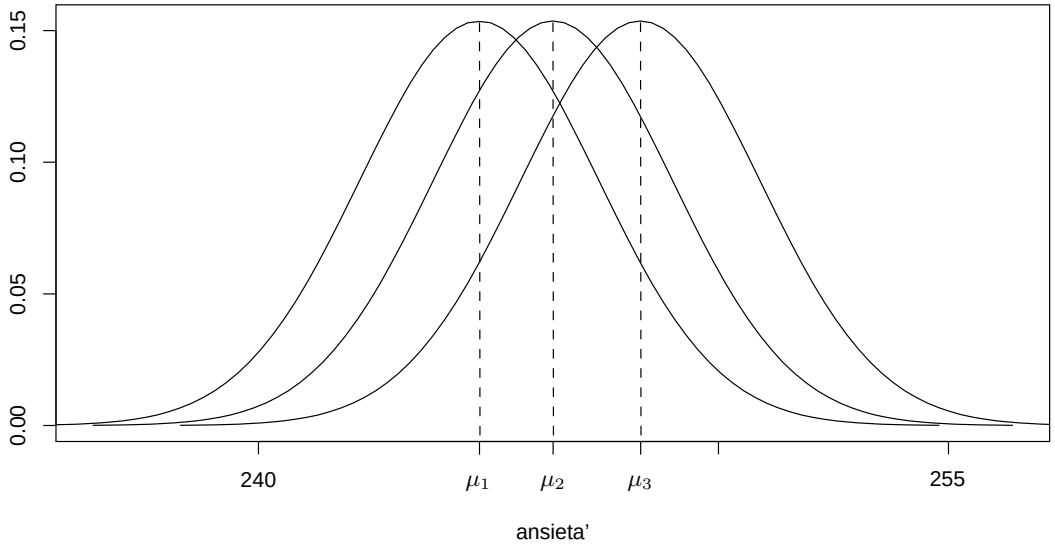


Figura 2.1. Modello ANOVA per la matrice di dati `caffeina`.

interpretati come segue:

$$\hat{y}_{ir} = \hat{\mu}_i = \begin{cases} \hat{\mu}_1 = 244.8 & \text{per } i = 1 \\ \hat{\mu}_2 = 246.4 & \text{per } i = 2 \\ \hat{\mu}_3 = 248.3 & \text{per } i = 3 \end{cases} \qquad y_{ir} = \hat{\mu}_1 + e_{ir} = \begin{cases} 244.8 + e_{1r} & \text{per } i = 1 \\ 246.4 + e_{2r} & \text{per } i = 2 \\ 248.3 + e_{3r} & \text{per } i = 3 \end{cases}$$

dove le e_{ir} sono residui (determinazioni della variabile aleatoria "residuo").

2.1.1 La stima dei parametri

I parametri del modello sono incogniti e vanno quindi stimati.

Secondo il criterio dei *minimi quadrati*, deve essere minimizzata la somma dei quadrati degli scarti tra le osservazioni e i loro valori attesi. Essendo per ciascun i (per ciascun trattamento) $\mathbb{E}[Y_{ir}] = \mu_i$, va minimizzata la quantità:

$$Q = \sum_{i}^{t} \sum_{r}^{n_i} (Y_{ir} - \mu_i)^2$$

Poiché tale quantità viene minimizzata dalla media aritmetica, si ha:

$$\hat{\mu}_i = \bar{y}_{i\cdot}$$

dove $\bar{y}_{i\cdot}$ è la media osservata per le repliche dell'i-esimo trattamento.

Come visto nel Capitolo 1, si ottiene lo stesso risultato adottando il criterio di *massimizzazione della verosimiglianza*.

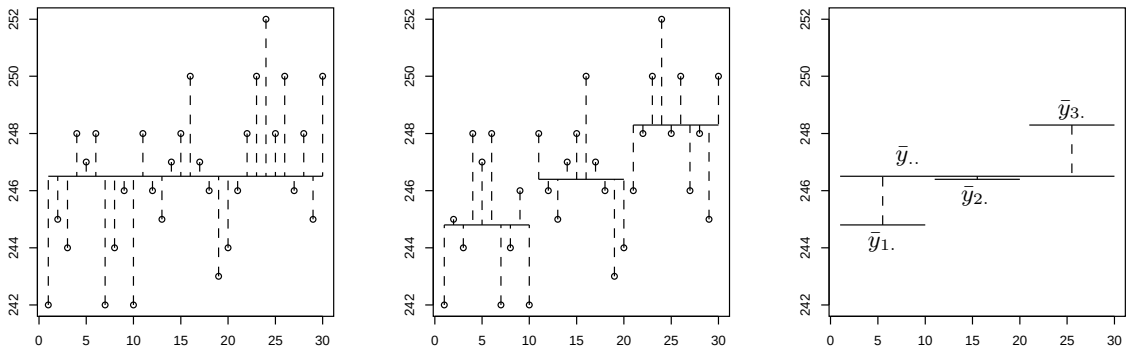


Figura 2.2. Scarti tra le osservazioni e la media generale, tra le osservazioni e le medie di trattamento, tra le medie di trattamento e la media generale (matrice di dati: `caffeina`).

2.1.2 L'analisi della varianza

Scomposizione della devianza

La *devianza totale* (*total sum of squares*) della variabile risposta è data da:

$$SSTOT = \sum_{i}^{t} \sum_{r}^{n_i} (y_{ir} - \bar{y}_{..})^2$$

La devianza totale può essere scomposta aggiungendo e sottraendo le medie osservate per ciascun trattamento, $\bar{y}_{i.}$; per ciascuna osservazione si ha:

$$y_{ir} - \bar{y}_{..} = (y_{ir} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})$$

dove $y_{ir} - \bar{y}_{..}$ è lo scarto tra l'osservazione y_{ir} e la media generale, $y_{ir} - \bar{y}_{i.}$ è lo scarto tra l'osservazione e la media per l'i-esimo trattamento, $\bar{y}_{i.} - \bar{y}_{..}$ è lo scarto tra la media di trattamento e la media generale (figura 2.2). Elevando al quadrato e sommando si ha:

$$\begin{aligned}\sum_i \sum_r (y_{ir} - \bar{y}_{..})^2 &= \sum_i \sum_r (y_{ir} - \bar{y}_{i.})^2 + 2 \sum_i \sum_r (y_{ir} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..}) + \sum_i \sum_r (\bar{y}_{i.} - \bar{y}_{..})^2 = \\ &= \sum_i \sum_r (y_{ir} - \bar{y}_{i.})^2 + \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2\end{aligned}$$

in quanto gli scarti non elevati al quadrato sono scarti dalla media, che hanno somme nulle. Inoltre $\sum_i \sum_r (\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2$ in quanto $r = 1, \ldots, n_i$, con $n_i = 10$ per ogni i essendo l'esperimento bilanciato.

Si ottiene così che la devianza totale è la somma di:

– la devianza delle medie di trattamento rispetto alla media generale (*model sum of squares*), ovvero la *devianza spiegata* dalle medie delle osservazioni per ciascun trattamento, che sono diverse proprio perché sono diversi i trattamenti, proprio perché ai t gruppi di unità sperimentali sono stati somministrati diversi livelli del fattore oggetto di studio:

$$SSMOD = \sum_i n_i (\bar{y}_{i.} - \bar{y}_{..})^2$$

– la devianza delle osservazioni dalle rispettive medie di trattamento, detta *devianza residua* (*residual sum of squares*) in quanto costituisce quella parte della variabilità complessiva del fenomeno che non è attribuibile ai trattamenti:

$$SSRES = \sum_i \sum_r (y_{ir} - \bar{y}_{i.})^2$$

Esempio 2.2. Svolgendo i calcoli con R sulla matrice di dati `caffeina` si ha:[4]

```
> caffeina <- read.csv("caffeina.csv")
> caffeina$tr <- as.factor(caffeina$tr)
> attach(caffeina)
> mu.gen <- mean(y)
> mu.tr <- by(y, tr, mean)
> SSTOT <- sum((y-mu.gen)^2)
> SSMOD <- sum((mu.tr - mu.gen)^2) * 10
> SSRES <- sum((y[tr==1]-mu.tr[1])^2) +
+ sum((y[tr==2]-mu.tr[2])^2) +
+ sum((y[tr==3]-mu.tr[3])^2)
> SSTOT; SSMOD; SSRES
[1] 195.5
[1] 61.4
[1] 134.1
```

Gradi di libertà

La devianza totale $SSTOT$ ha $n-1$ gradi di libertà; vi sono infatti n scarti dalla media generale, ma questi non sono indipendenti in quanto la loro somma deve essere nulla: $\sum_i \sum_r (y_{ir} - \bar{y}_{..}) = 0$.

La devianza spiegata $SSMOD$ ha $t-1$ gradi di libertà, in quanto vi sono t medie di trattamento ma la somma dei loro scarti dalla media generale deve essere nulla: $\sum_i n_i(\bar{y}_{i.} - \bar{y}_{..}) = 0$.

La devianza residua ha $n-t$ gradi di libertà. Per ogni trattamento, infatti, vi sono $n_i - 1$ gradi di libertà, in quanto vi sono n_i osservazioni ma la somma dei loro scarti dalla media di trattamento deve essere nulla. Per tutti i t trattamenti si ha quindi:

$$(n_1 - 1) + (n_2 - 1) + \cdots + (n_t - 1) = n - t$$

Esempio 2.3. Nell'esperimento `caffeina` i gradi di libertà sono:

a) $30 - 1 = 29$ per la devianza totale;

b) $3 - 1 = 2$ per la devianza spiegata (quindi per il modello);

c) $30 - 3 = 27$ per la devianza residua (quindi per l'errore).

Si può notare che, così come la devianza totale è la somma delle devianze spiegata e residua, anche i gradi di libertà della devianza totale sono la somma di quelli delle devianze spiegata e residua.

[4] R si presta meglio del SAS ad essere usato come calcolatrice. I valori qui calcolati si ritrovano comunque nell'output di SAS riprodotto nell'esempio 2.3.

Calcolo delle varianze

Le varianze vengono calcolate, come sempre in ambito inferenziale, dividendo le devianze per i rispettivi gradi di libertà. Interessano in particolare le varianze spiegata e residua:[5]

– varianza spiegata (*treatment mean square*):

$$MSMOD = \frac{SSMOD}{t-1}$$

– varianza residua (*residual mean square*):

$$MSRES = \frac{SSRES}{n-t}$$

Esempio 2.4. Nell'esperimento `caffeina` si ha:

a) varianza spiegata: $61.4/2 = 30.7$;

b) varianza residua: $134.1/27 = 4.9\bar{6}$.

2.1.3 Il test di ipotesi sul modello

Nel caso di esperimenti con un solo fattore (una sola variabile esplicativa), la tipica ipotesi nulla consiste nell'uguaglianza delle medie di trattamento:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$$

Se infatti tali medie fossero uguali, la variabile risposta si distribuirebbe come un campione casuale costituito da variabili aleatorie normali indipendenti e identicamente distribuite e la variabilità osservata andrebbe interpretata come effetto di oscillazioni accidentali, secondo una varianza σ^2, intorno ad un'unica media μ. In altri termini, i trattamenti non avrebbero alcun effetto, la variabile esplicativa non sarebbe la causa dei diversi valori osservati della variabile riposta.

Il *teorema di Cochran* consente di costruire un test per accettare o rifiutare l'ipotesi nulla usando la statistica test:

$$F^* = \frac{MSMOD}{MSRES} = \frac{SSMOD/(t-1)}{SSRES/(n-t)}$$

Se vale l'ipotesi nulla, allora si può applicare il teorema di Cochran e derivarne che:

– $\dfrac{SSMOD}{\sigma^2} \sim \chi^2_{t-1}$;

– $\dfrac{SSRES}{\sigma^2} \sim \chi^2_{n-t}$;

– le due variabili aleatorie sono indipendenti.

Da questo segue che $F^* = \dfrac{MSMOD}{MSRES}$ è distribuita come una variabile F di Snedecor con $t-1, n-t$ gradi di libertà:

$$F^* = \frac{MSMOD}{MSRES} \sim F_{t-1,n-t}$$

[5] Poiché le devianze spiegata e residua vengono divise per numeri diversi di gradi di libertà, la varianza totale *non è* uguale alla somma delle varianze spiegata e residua.

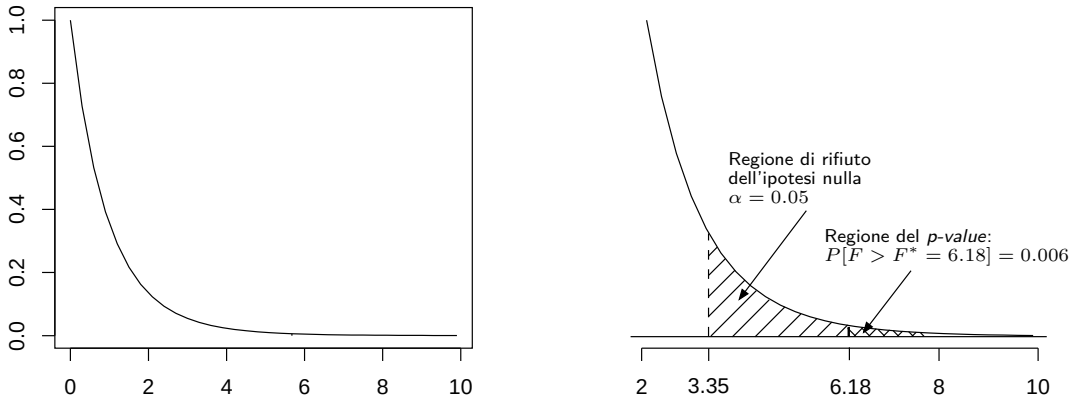


Figura 2.3. A sinistra la funzione di densità della v.a. $F_{2,27}$. A destra il suo tratto in un intervallo che contiene il 95° percentile e l'area del *p-value* (matrice di dati: `caffeina`).

Osservazione 2.5. Se vale l'ipotesi alternativa (diversità sistematica, non casuale, delle medie di trattamento), F^* si distribuisce come una F *non centrale*. Ciò ha tuttavia rilevanza per l'errore di II tipo (accettare l'ipotesi nulla quando è falsa), mentra interessa in prima istanza evitare l'errore di I tipo (rifiutare l'ipotesi nulla quando è vera).

Fissato un livello di significatività α, cioè una probabilità α di rifiutare l'ipotesi nulla quando è vera (errore di I tipo), si adotta la seguente regola:

$$\begin{aligned} F^* \leq F_{1-\alpha,t-1,n-t} \quad &\Rightarrow \quad \text{si accetta } H_0 \\ F^* > F_{1-\alpha,t-1,n-t} \quad &\Rightarrow \quad \text{si rifiuta } H_0 \end{aligned}$$

dove $F_{1-\alpha,t-1,n-t}$ è il $(1-\alpha) * 100$-esimo percentile della distribuzione $F_{t-1,n-t}$.

Il *p-value* aiuta a scegliere, in quanto è la probabilità $P[F_{t-1,n-t} > F^*]$, cioè che $F^* \sim F_{t-1,n-t}$ assuma un valore superiore a quello osservato, ovvero che si osservi un valore "estremo" in una determinazione della statistica test coerente con l'ipotesi nulla. Se $p > \alpha$, si accetta l'ipotesi nulla in quanto il valore osservato fa ritenere che le differenze tra le medie di trattamento e la media generale siano da attribuire al caso. Se invece $p < \alpha$, si rifiuta l'ipotesi nulla in quanto la probabilità che questa sia vera (la probabilità dell'errore di I tipo) è minore del livello di significatività.

Tali informazioni vengono sintetizzate nella cosiddetta *tabella ANOVA* (tabella 2.1).

Esempio 2.6. Nell'esperimento `caffeina`, $F^* = 6.18$; con R:

Tabella 2.1. Tabella ANOVA per l'esperimento `caffeina`.

	GdL	Devianza	Varianza corretta	F^*	*p-value*
Modello	$t-1$	$SSMOD = \sum_i n_i(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$	$MSMOD = \dfrac{SSMOD}{t-1}$	$\dfrac{MSMOD}{MSRES}$	$P[F_{t-1,n-t} > F^*]$
Errore	$n-t$	$SSRES = \sum_i \sum_r (y_{ir} - \bar{y}_{i\cdot})^2$	$MSRES = \dfrac{SSRES}{n-t}$		
Totale	$n-1$	$SSTOT = \sum_i \sum_r (y_{ir} - \bar{y}_{\cdot\cdot})^2$			

```
> Fstar <- (SSMOD/2) / (SSRES/27)
> Fstar
[1] 6.181208
```

Il 95° percentile della distribuzione $F_{2,27}$ è 3.35:

```
> qf(0.95,2,27)
[1] 3.354131
```

Essendo $F^* > F_{0.95,2,27}$ si rifiuta l'ipotesi nulla. Il *p-value*:

```
> pf(Fstar, 2, 27, lower.tail=FALSE)
[1] 0.006163214
```

consente di rifiutare l'ipotesi nulla con una probabilità di errore di I tipo molto bassa (poco superiore allo 0.6%). La tabella ANOVA può essere costruita manualmente, usando i dati calcolati in questo e negli esempi precedenti, oppure usando le funzioni `lm()` e poi `anova()` di R o la procedura `glm` di SAS. Con R:

```
> caffeina$tr <- as.factor(caffeina$tr)
> mod <- lm(y ~ tr, data=caffeina)
> anova(mod)
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value   Pr(>F)
tr         2   61.4 30.7000  6.1812 0.006163 **
Residuals 27  134.1  4.9667
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Ci si deve assicurare che il trattamento sia una variabile qualitativa, in quanto, in caso contrario, verrebbe effettuata un'analisi di regressione. Con SAS va usata l'opzione `class tr` per specificare che la variabile esplicativa `tr` serve solo a distinguere ("classificare", nel gergo di SAS) i diversi trattamenti; si può essere usare la procedura `glm` (*general linear model*) o `anova`:

```
proc glm data=caffeina;
  class tr;
  model y=tr;
run;
```

l'output:

```
                                The GLM Procedure

Dependent Variable: y
                                        Sum of
   Source                     DF         Squares     Mean Square    F Value    Pr > F
   Model                       2      61.4000000      30.7000000       6.18    0.0062
   Error                      27     134.1000000       4.9666667
   Corrected Total            29     195.5000000

                  R-Square     Coeff Var      Root MSE        y Mean
                  0.314066      0.904098      2.228602      246.5000
```

2.1.4 Confronti tra medie

SAS consente di includere nell'output di `proc glm` anche stime delle medie con il comando `means`.

Esempio 2.7. Dando il comando:

```
proc glm data=caffeina;
  class tr;
  model y=tr;
  means tr;
run;
```

dopo un'analisi della varianza uguale a quella appena vista vengono fornite le stime delle medie di trattamento e il loro scarto quadratico medio:

```
Level of            --------------y--------------
tr             N              Mean          Std Dev

1             10       244.800000       2.39443800
2             10       246.400000       2.06559112
3             10       248.300000       2.21359436
```

Soprattutto, col comando `contrast`, si ottengono stime e test dei *contrasti*, che sono confronti tra due o più medie. In generale, un contrasto L è una combinazione lineare di medie di fattore con coefficienti c_i a somma nulla; se il fattore presenta t livelli:

$$L = \sum_{i=1}^{t} c_i \mu_i \qquad \sum_{i=1}^{t} c_i = 0$$

Ad esempio, se interessa il confronto tra la prima e la terza media:

$$L = \mu_1 - \mu_3 \qquad c_1 = 1,\ c_2 = 0,\ c_3 = -1$$

e il comando `contrast` corrispondente è:

```
contrast 'tr1 vs tr3' tr 1 0 -1;
```

vanno quindi specificati una descrizione testuale del confronto, la colonna rispetto a cui si calcolano le medie e i coefficienti c_i.

Esempio 2.8. Aggiungendo comandi `contrast` all'istruzione data nell'esempio precedente:

```
proc glm data=caffeina;
  class tr;
  model y=tr;
  means tr;
  contrast 'tr1 vs tr2' tr 1 -1  0;
  contrast 'tr1 vs tr3' tr 1  0 -1;
  contrast 'tr2 vs tr3' tr 0  1 -1;
run;
```

si ottiene in coda all'output:

```
Contrast                 DF    Contrast SS    Mean Square   F Value   Pr > F

tr1 vs tr2                1    12.80000000    12.80000000      2.58   0.1200
tr1 vs tr3                1    61.25000000    61.25000000     12.33   0.0016
tr2 vs tr3                1    18.05000000    18.05000000      3.63   0.0673
```

La devianza per il confronto tra μ_1 e μ_2 è calcolata come il quadrato della differenza tra le due medie, $-1.6^2 = 2.56$, moltiplicato per il reciproco di $\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$, quindi per 5. Analogamente per gli altri due confronti. Si può notare che il confronto tra μ_1 e μ_3 risulta il più significativo.

2.1.5 Il modello a effetti dei fattori

Nel modello a medie di cella, appena visto, si usano i t parametri μ_i. È possibile riparametrizzare in modo diverso, distinguendo tra un *livello di riferimento* del fenomeno osservato e gli *effetti differenziali* dei diversi livelli del fattore sperimentale (*factor effects model*):

$$Y_{ir} = \mu + \alpha_i + \varepsilon_{ir}$$

In questo modo i parametri diventano $t + 1$ e ciò, come già notato (pag. 11), conduce ad una matrice di riparametrizzazione con colonne linearmente dipendenti, che viene pertanto rielaborata introducendo dei vincoli sui parametri.

Si può indendere μ come la media aritmetica generale del fenomeno, rispetto alla quale gli α_i sono scarti la cui somma è nulla:

$$\sum_{i}^{t} a_i = 0$$

In tal caso uno degli a_i può essere espresso in funzione degli altri:

$$\alpha_t = -\alpha_1 - \alpha_2 - \cdots - \alpha_{t-1}$$

Nel caso $t = 3$, la matrice di riparametrizzazione assume la forma già vista a pag. 11.

Si può invece intendere uno degli effetti differenziali come nullo. Ciò equivale a sostituire una colonna della matrice di riparametrizzazione con tutti zeri, quindi a eliminarla (riparametrizzazione *corner point*).

Esempio 2.9. Nel caso della matrice di dati `caffeina`, ponendo $\alpha_1 = 0$ si passa dal modello lineare visto nell'esempio 2.1 al seguente:

$$\mathbf{Y} = \begin{bmatrix} Y_{1,1} \\ \vdots \\ Y_{1,10} \\ Y_{2,1} \\ \vdots \\ Y_{2,10} \\ Y_{3,1} \\ \vdots \\ Y_{3,10} \end{bmatrix} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon} = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_2 \\ \alpha_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,1} \\ \vdots \\ \varepsilon_{1,10} \\ \varepsilon_{2,1} \\ \vdots \\ \varepsilon_{2,10} \\ \varepsilon_{3,1} \\ \vdots \\ \varepsilon_{3,10} \end{bmatrix}$$

e l'interpretazione dei valori osservati diventa:

$$\hat{y}_{ir} = \hat{\mu} + \hat{\alpha}_i = \begin{cases} \hat{\mu} = 244.8 & \text{per } i = 1 \\ \hat{\mu} + \hat{\alpha}_2 = 246.4 & \text{per } i = 2 \\ \hat{\mu} + \hat{\alpha}_3 = 248.3 & \text{per } i = 3 \end{cases} \qquad y_{ir} = \hat{y}_{ir} + e_{ir} = \begin{cases} 244.8 + e_{1r} & \text{per } i = 1 \\ 246.4 + e_{2r} & \text{per } i = 2 \\ 248.3 + e_{3r} & \text{per } i = 3 \end{cases}$$

dove le e_{ir} sono residui. Ovviamente, $\hat{\alpha}_2 = 1.6$ e $\hat{\alpha}_3 = 3.5$. R usa in effetti una matrice di questo tipo, come si può vedere con la funzione `model.matrix()`:

```
> mod <- lm(y ~ tr, data=caffeina)
> model.matrix(mod)
   (Intercept) tr2 tr3
1            1   0   0
2            1   0   0
...
11           1   1   0
12           1   1   0
...
29           1   0   1
30           1   0   1
```

SAS fissa invece come *corner point* l'ultimo parametro; nel caso di `caffeina`, pone quindi $\alpha_3 = 0$.

È importare notare che non vi è alcuna differenza tra i modelli a medie di cella e quelli a effetti dei fattori per quanto riguarda il test di ipotesi circa l'uguaglianza delle medie di trattamento; semplicemente si passa da:

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_t$$

a:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_t = 0 \qquad (\text{quindi } \mu_1 = \mu_2 = \cdots = \mu_t = \mu)$$

Cambiano solo la definizione dei parametri e le loro modalità di calcolo per la loro stima.

Esempio 2.10. La funzione `summary()` di R, quando le si passa il risultato di una chiamata della funzione `lm()`, fornisce informazioni sulla stima dei parametri:

```
> mod <- lm(y ~ tr, data=caffeina)
> summary(mod)

Call:
lm(formula = y ~ tr, data = caffeina)

Residuals:
   Min     1Q Median     3Q    Max
-3.400 -2.075 -0.300  1.675  3.700

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 244.8000     0.7047 347.359  < 2e-16 ***
tr2           1.6000     0.9967   1.605  0.12005
tr3           3.5000     0.9967   3.512  0.00158 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.229 on 27 degrees of freedom
Multiple R-squared: 0.3141,Adjusted R-squared: 0.2633
F-statistic: 6.181 on 2 and 27 DF,  p-value: 0.006163
```

I parametri vengono detti `Coefficients`. Con `(Intercept)` si indica il livello generale del fenomeno, quindi $\hat{\mu}$, uguale a $\hat{\mu} + \hat{\alpha}_1$ con $\hat{\alpha}_1 = 0$; si vede che il suo valore stimato è uguale a quello di $\hat{\mu}_1$ del modello a medie di cella come stimato col metodo dei minimi quadrati (sez. 2.1.1). Con `tr2` si indica $\hat{\alpha}_2$ che, sommato a $\hat{\mu} = \hat{\mu}_1$, permette di ottenere $\hat{\mu}_2$; analogamente, con `tr3` si indica $\hat{\alpha}_3$. SAS fornisce risultati apparentemente diversi, proprio in quanto usa come *corner point* l'ultimo parametro; chiamando la procedura `glm` con l'opzione `/solution`:

```
proc glm data=caffeina;
  class tr;
  model y=tr /solution;
run;
```

si ottiene la seguente stima dopo la tabella ANOVA:

Parameter		Estimate	Standard Error	t Value	Pr > |t|
Intercept		248.3000000 B	0.70474582	352.33	<.0001
tr	1	-3.5000000 B	0.99666109	-3.51	0.0016
tr	2	-1.9000000 B	0.99666109	-1.91	0.0673
tr	3	0.0000000 B	.	.	.

```
NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to
      solve the normal equations.  Terms whose estimates are followed by the letter 'B'
      are not uniquely estimable.
```

Si può notare che `Intercept` non è altro che $\hat{\mu}$ posto uguale a $\hat{\mu} + \hat{\alpha}_3$ con $\hat{\alpha}_3 = 0$ e che il suo valore coincide con quello di $\hat{\mu}_3$ del modello a medie di cella; analogamente, $\hat{\mu} - 3.5 = 244.8$ coincide con $\hat{\mu}_1$ e $\hat{\mu} - 1.9 = 246.4$ con $\hat{\mu}_2$. Quindi, nonostante la nota avverta che la stima dei parametri non è univoca, anche in questo caso si torna facilmente alle stime $\hat{\mu}_i$ calcolate nel modello a medie di cella (usando SAS, con il comando `means`). Per ottenere risultati uguali a quelli di R, inoltre, basta rendere la modalità del primo trattamento maggiore di quelle degli altri due; ad esempio:

```
data caffeina1;
  set caffeina;
  if tr=1 then tr=11;
run;
proc glm data=caffeina1;
  class tr;
  model y=tr /solution;
run;
```

Si deve notare, infine, che alla stima dei parametri si accompagnano, sia in R che in SAS, i test di ipotesi sui parametri.

2.1.6 I test di ipotesi sui parametri

Il test di ipotesi sul modello consente di scegliere se attribuire tutta la variabilità al caso (ipotesi nulla) oppure alla componente sistematica formalizzata nel modello. Anche se si rifiuta l'ipotesi nulla, tuttavia, da ciò non segue che tutti i parametri siano ugualmente significativi, né che siano stati stimati con uguale accuratezza. Occorrono quindi anche test sui parametri, che dipendono ovviamente dall'interpretazione che si dà dei parametri sulla base del modello di riparametrizzazione.

Sia SAS che R usano una riparametrizzazione *corner point*, nella quale il parametro μ viene interpretato come livello di riferimento del fenomeno (*intercept*) e gli altri come differenze da questo indotte dai trattamenti (dai livelli del fattore).

Nel primo caso si sottopone a verifica l'ipotesi nulla $H_0 : \mu = 0$, dove μ è la media della variabile risposta per il primo (R) o l'ultimo (SAS) trattamento, cioè una media il cui stimatore è:

$$\bar{Y}_{i\cdot} = \frac{1}{n_i} \sum_{r=1}^{n_i} Y_{ir} \qquad \text{per un dato } i$$

L'indice i vale 1 per R, t per SAS, ma in entrambi i casi si pone $\alpha_i = 0$; per tale i, quindi, il modello $Y_{ir} = \mu + \alpha_i + \varepsilon_{ir}$ diventa:

$$Y_{ir} = \mu + \varepsilon_{ir} \qquad\qquad \mathbb{E}[Y_{ir}] = \mu$$

$\bar{Y}_{i\cdot}$ è, come la media campionaria, una media di variabili aleatorie indipendenti e identicamente distribuite (si assume che siano tali nell'ambito di ciascun trattamento); il suo valore atteso e la sua varianza sono quindi:

$$\mathbb{E}[\bar{Y}_{i\cdot}] = \mathbb{E}\left[\frac{1}{n_i}\sum_{r=1}^{n_i} Y_{ir}\right] = \frac{1}{n_i}\sum_{r=1}^{n_i}\mathbb{E}[Y_{ir}] \overset{id}{=} \frac{1}{n_i}\, n_i\, \mathbb{E}[Y_{ir}] = \mathbb{E}[Y_{ir}] = \mu$$

$$\mathbb{V}[\bar{Y}_{i\cdot}] = \mathbb{V}\left[\frac{1}{n_i}\sum_{r=1}^{n_i} Y_{ir}\right] \overset{ind}{=} \frac{1}{n_i^2}\sum_{r=1}^{n_i}\mathbb{V}[Y_{ir}] \overset{id}{=} \frac{1}{n_i^2}\, n_i\, \mathbb{V}[Y_{ir}] = \frac{1}{n_i}\mathbb{V}[Y_{ir}] = \frac{\sigma^2}{n_i}$$

Dal momento che la varianza σ^2 non è nota e che un suo stimatore corretto è la varianza residua $MSRES$ (v. pag. 29), si usa la statistica test:

$$t^* = \frac{\bar{Y}_{i\cdot} - 0}{\sqrt{\frac{MSRES}{n_i}}} \sim t_{n-t}$$

dove $n - t$ sono i gradi di libertà della varianza residua.

Si tratta di un test a due code, quindi, fissato il livello di significatività α:

$$|t^*| \leq \left|t_{1-\frac{\alpha}{2},n-t}\right| \quad \Rightarrow \quad \text{si accetta } H_0$$

$$|t^*| > \left|t_{1-\frac{\alpha}{2},n-t}\right| \quad \Rightarrow \quad \text{si rifiuta } H_0$$

Si calcola inoltre il *p-value*, cioè la probabilità che $|t_{n-t}|$ assuma valori superiori al valore osservato di $|t^*|$.

Esempio 2.11. Proseguendo i calcoli iniziati nell'esempio 2.2:

```
> SSRES
[1] 134.1
> # dividendo SSRES per i suoi gradi di libertà:
> MSRES <- SSRES / 27
> MSRES
[1] 4.966667
> # media osservata per il primo trattamento:
> mu <- mean(y[tr==1])
> mu
[1] 244.8
> # radice quadrata della varianza stimata della media:
> StdError <- sqrt(MSRES / 10)
> StdError
[1] 0.7047458
> # statistica test:
> tstar <- mu / StdError
> tstar
[1] 347.3593
```

Si può notare che `StdError` e `tstar` hanno lo stesso valore che `Std. Error` e `t value` hanno nella riga `(Intercept)` dell'output di R riprodotto a pag. 41. Per il resto, `tstar` è talmente grande che il *p-value* non può che essere piccolissimo. Infatti il valore di $t_{1-\frac{0.05}{2},27} = t_{0.975,27}$ è molto minore di t^*:

```
> qt(0.975, 27)
[1] 2.051831
```

Quanto al *p-value*, questo è $P[|t| > |t^*|]$ e può essere calcolato così:

```
> pt(abs(tstar), 27, lower.tail=FALSE) + # P[t > |tstar|]
+ pt(-abs(tstar), 27)                   # P[t < -|tstar|]
[1] 8.004808e-51
```

Esempio 2.12. SAS, per default, usa come *corner point* il trattamento col valore maggiore, quindi 3. La media dei valori osservati per tale trattamento è 248.3. Poiché l'esperimento `caffeina` è bilanciato, la varianza e la sua stima sono le stesse, ma cambia `tstar`. Infatti:

```
> mu <- mean(y[tr==3])
> mu
[1] 248.3
> mu / StdError
[1] 352.3256
```

che è il valore che appare nell'output riprodotto a pag. 42.

Per i parametri relativi agli effetti differenziali si segue una logica analoga. In una riparametrizzazione come quella dell'esempio 2.9, si ha che le medie per i tre trattamenti

sono, rispettivamente, $\mu_1 = \mu$, $\mu_2 = \mu + \alpha_2$ e $\mu_3 = \mu + \alpha_3$. Un test su α_2 è quindi un test sulla differenza $\mu_2 - \mu$ (analogamente per α_3). Gli stimatori delle due medie sono $\bar{Y}_{1\cdot}$ e $\bar{Y}_{2\cdot}$ e uno stimatore della loro differenza è:

$$\hat{\alpha}_2 = \bar{Y}_{2\cdot} - \bar{Y}_{1\cdot}$$

il cui valore atteso è ovviamente $\mu_2 - \mu_1$. Poiché le due medie sono assunte indipedenti, la varianza di $\hat{\alpha}_2$ è:

$$\mathbb{V}[\hat{\alpha}_2] = \mathbb{V}[\bar{Y}_{2\cdot}] + \mathbb{V}[\bar{Y}_{1\cdot}] = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

che viene stimata da:

$$S^2_{\hat{\alpha}_2} = MSRES \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

L'ipotesi nulla è $H_0 : \alpha_2 = 0$, ovvero $H_0 : \mu_2 = \mu_1 = \mu$. Per verificarla si usa la statistica test:

$$t^* = \frac{\mu_2 - \mu}{\sqrt{MSRES \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{\hat{\alpha}_2}{\sqrt{MSRES \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim t_{n-t}$$

Analogamente per i parametri relativi agli altri effetti differenziali.

Esempio 2.13. Proseguendo ancora con R:

```
> StdError <- sqrt(MSRES * (1/10 + 1/10))
> StdError
[1] 0.996661
> alpha2 <- 1.6
> tstar <- alpha2 / StdError
> tstar
[1] 1.60536
> pt(abs(tstar), 27, lower.tail=FALSE) + # P[t > |tstar|]
+ pt(-abs(tstar), 27)                   # P[t < -|tstar|]
[1] 0.1200490
```

si ottengono gli stessi valori già visti nell'output riprodotto a pag. 41.

Esempio 2.14. Nella riparametrizzazione usata da SAS per default si ha:

$$\hat{y}_{ir} = \hat{\mu} + \hat{\alpha}_i = \begin{cases} \hat{\mu} + \hat{\alpha}_1 = 244.8 & (\alpha_1 = -3.5) \\ \hat{\mu} + \hat{\alpha}_2 = 246.4 & (\alpha_2 = -1.9) \\ \hat{\mu} \qquad = 248.3 & (\alpha_3 = \quad 0) \end{cases}$$

Il calcolo del valore osservato di t^* per il secondo trattamento e del relativo *p-value* è quindi:

```
> tstar <- -1.9 / StdError
> tstar
[1] -1.906365
> pt(abs(tstar), 27, lower.tail=FALSE) + # P[t > |tstar|]
+ pt(-abs(tstar), 27)                   # P[t < -|tstar|]
[1] 0.06729867
```

I valori così calcolati coincidono con quelli visti nell'output riprodotto a pag. 42.

Osservazione. Tra le variabili t di Student e F di Snedecor vale la relazione:

$$(t_\nu)^2 = F_{1,\nu}$$

I valori di t^* appena calcolati per il confronto tra μ_1 e μ_2 e per quello tra μ_2 e μ_3 sono, rispettivamente, 1.60536 e -1.906365, i cui quadrati sono 2.577181 e 3.634228. Sono infatti questi i valori di F^* calcolati con SAS per tali confronti con il comando `contrast`, come visto sopra nell'esempio 2.8.

2.1.7 Intervalli di confidenza dei parametri

Le quantità calcolate per i test di ipotesi possono essere utilizzate per calcolare gli intervalli di confidenza dei parametri. Infatti, indicando con i l'indice del parametro α posto uguale a zero, con $j \neq i$ gli indici degli altri parametri α:

$$\mu \in \left(\bar{Y}_{i\cdot} \pm t_{1-\frac{\alpha}{2},n-t} \sqrt{\frac{MSRES}{n_i}} \right) \quad \alpha_j \in \left((\bar{Y}_{j\cdot} - \bar{Y}_{i\cdot}) \pm t_{1-\frac{\alpha}{2},n-t} \sqrt{MSRES \left(\frac{1}{n_j} + \frac{1}{n_i} \right)} \right)$$

Esempio 2.15. Con SAS gli intervalli di confidenza dei parametri si ottengono usando l'opzione `clparm`, ad esempio:

```
proc glm data=caffeina;
  class tr;
  model y=tr /solution clparm;
run;
```

La parte finale dell'output, dopo la stima dei parametri già vista, è:

```
Parameter          95% Confidence Limits

Intercept        246.8539810  249.7460190
tr         1      -5.5449796   -1.4550204
tr         2      -3.9449796    0.1449796
tr         3           .             .
```

R fornisce invece gli intervalli di confidenza dei valori teorici, ma il calcolo di quelli dei parametri è semplice; per ottenere gli stessi valori forniti da SAS (con $\alpha_3 = 0$):

```
> # Intervallo di confidenza (alfa=0.05) per mu:
> estremo <- qt(0.975, 27) * sqrt(MSRES / 10)
> c(248.3 - estremo, 248.3 + estremo)
[1] 246.854 249.746
> # Intervallo di confidenza (alfa=0.05) per tr1 e tr2
> estremo <- qt(0.975, 27) * sqrt(MSRES * (1/10 + 1/10))
> c(-3.5 - estremo, -3.5 + estremo)
[1] -5.544980 -1.455020
> c(-1.9 - estremo, -1.9 + estremo)
[1] -3.9449796  0.1449796
```

2.2 Esperimenti completi e bilanciati con due fattori

Vi sono n unità sperimentali, cui vengono somministrati due fattori A e B. Il fattore A presenta a livelli, B ne presenta b. Vi sono quindi ab trattamenti, ciascuno dei quali viene somministrato a $n/(ab)$ unità, $n/(ab) > 1$ (disegno fattoriale), e altrettante *medie di trattamento* μ_{ij}, con $i = 1, \ldots, a$ e $j = 1, \ldots, b$.

Vi sono inoltre $a + b$ *medie di fattore*; la media della variabile risposta per le unità cui è stato somministrato l'i-esimo livello del fattore A è:

$$\mu_{i.} = \frac{\sum_{j=1}^{b} \mu_{ij}}{b}$$

mentre quella per il j-esimo livello del fattore B è:

$$\mu_{.j} = \frac{\sum_{i=1}^{a} \mu_{ij}}{a}$$

La media generale è quindi:

$$\mu_{..} = \frac{\sum_i \sum_j \mu_{ij}}{ab} = \frac{\sum_{i=1}^{a} \mu_{i.}}{a} = \frac{\sum_{j=1}^{b} \mu_{.j}}{b}$$

Su questa base si determinano facilmente gli effetti differenziali dei diversi livelli del primo fattore, α_i, e del secondo, β_i:

$$\alpha_i = \mu_{i.} - \mu_{..} \qquad \beta_j = \mu_{.j} - \mu_{..}$$

Esempio 2.16. Con riferimento alla matrice dei dati contenuta nel file `dietepec.csv`,[6] si può costruire la tabella 2.2. Il file contiene le osservazioni relative a 40 pecore (le unità sperimentali) cui sono stati somministrati 4 trattamenti (c'è una colonna `tratt` con valori da 1 a 4); le colonne `rame` e `cobalto` contengono 1 o 2 per indicare, rispettivamente, l'assenza o la presenza del metallo nella dieta. La variabile risposta, `incpeso`, registra l'incremento di peso di ciascuna pecora.

[6] `http://web.mclink.it/MC1166/ModelliStatistici/dietepec.csv`. Le medie di trattamento possono essere calcolate in R (dopo `attach(dietepec)`) con:

```
> by(incpeso, list(rame, cobalto), mean)
```

in SAS, dove i nomi delle variabili sono `y`, `t`, `a` e `b`, con:

```
proc means; varr y; by a b notsorted; run;
```

Le medie di fattore, ad esempio per il rame, in R con:

```
> by(incpeso, rame, mean)
```

ed in SAS con:

```
proc sort by a; proc means; var y; by a; run;
```

Tabella 2.2. Medie generale, di trattamento e di fattore; effetti differenziali e interattivi (matrice dei dati: `dietepec`).

Incremento medio di peso			
	Fattore B - Cobalto		
Fattore A - Rame	$j=1$: assenza	$j=2$: presenza	Medie di riga
$i=1$: assenza	$\mu_{11}=16.80$	$\mu_{12}=20.60$	$\mu_{1.}=18.70$
$i=2$: presenza	$\mu_{21}=15.30$	$\mu_{22}=21.10$	$\mu_{2.}=18.20$
Medie di colonna	$\mu_{.1}=16.05$	$\mu_{.2}=20.85$	$\mu_{..}=18.45$

Effetti del rame	Effetti del cobalto
$\alpha_1=\mu_{1.}-\mu_{..}=\ \ 0.25$	$\beta_1=\mu_{.1}-\mu_{..}=-2.40$
$\alpha_2=\mu_{2.}-\mu_{..}=-0.25$	$\beta_2=\mu_{.2}-\mu_{..}=\ \ 2.40$

Effetti interattivi	
$(\alpha\beta)_{11}=\mu_{11}-(\mu_{..}+\alpha_1+\beta_1)=\ \ 0.50$	$(\alpha\beta)_{12}=\mu_{12}-(\mu_{..}+\alpha_1+\beta_2)=-0.50$
$(\alpha\beta)_{21}=\mu_{21}-(\mu_{..}+\alpha_2+\beta_1)=-0.50$	$(\alpha\beta)_{22}=\mu_{22}-(\mu_{..}+\alpha_2+\beta_2)=\ \ 0.50$

2.2.1 Effetti interattivi

Si è già notato che i disegni fattoriali vengono utilizzati quando interessa non solo e non tanto l'effetto che i fattori hanno singolarmente sulla variabile risposta (potrebbero essere studiati separatamente), ma soprattutto gli *effetti interattivi*, cioè gli *ulteriori* effetti dovuti alla combinazione di due o più fattori. Nel caso di due fattori, l'assenza o presenza di effetti interattivi si rileva facilmente:

a) se ciascuna media di trattamento μ_{ij} è uguale alla somma della media generale e degli effetti differenziali del primo fattore al livello i e del secondo al livello j, ovvero se:

$$\mu_{11}=\mu_{..}+\alpha_1+\beta_1$$
$$\mu_{23}=\mu_{..}+\alpha_2+\beta_3$$

ecc., allora vi sono solo *effetti additivi*, ovvero l'effetto di due fattori, ciascuno considerato ad un suo dato livello, non è altro che la somma degli effetti singoli;

b) se invece tali uguaglianze non sussistono, i due fattori hanno un *effetto interattivo* che si somma, algebricamente, agli effetti differenziali singoli.

Esempio 2.17. I dati dell'esperimento `dietepec` mostrano che rame e cobalto hanno un effetto interattivo sull'incremento di peso delle pecore, infatti:

$$\begin{aligned}
\mu_{11}=16.80 &\neq \mu_{..}+\alpha_1+\beta_1=18.45+0.25-2.40=16.30\\
\mu_{12}=20.60 &\neq \mu_{..}+\alpha_1+\beta_2=18.45+0.25+2.40=21.10\\
\mu_{21}=15.30 &\neq \mu_{..}+\alpha_2+\beta_1=18.45-0.25-2.40=15.80\\
\mu_{22}=21.10 &\neq \mu_{..}+\alpha_2+\beta_2=18.45-0.25+2.40=20.60
\end{aligned}$$

La presenza di effetti interattivi risulta anche dai cosiddetti *grafici delle interazioni* (*treatment means plot* o *interaction plot*). Nella figura 2.4, il grafico a sinistra mostra l'incremento medio di peso dovuto al rame sia in assenza che in presenza di cobalto; la linea in basso congiunge le medie μ_{11} e μ_{21} (assenza di cobalto), quella in alto le medie μ_{12}

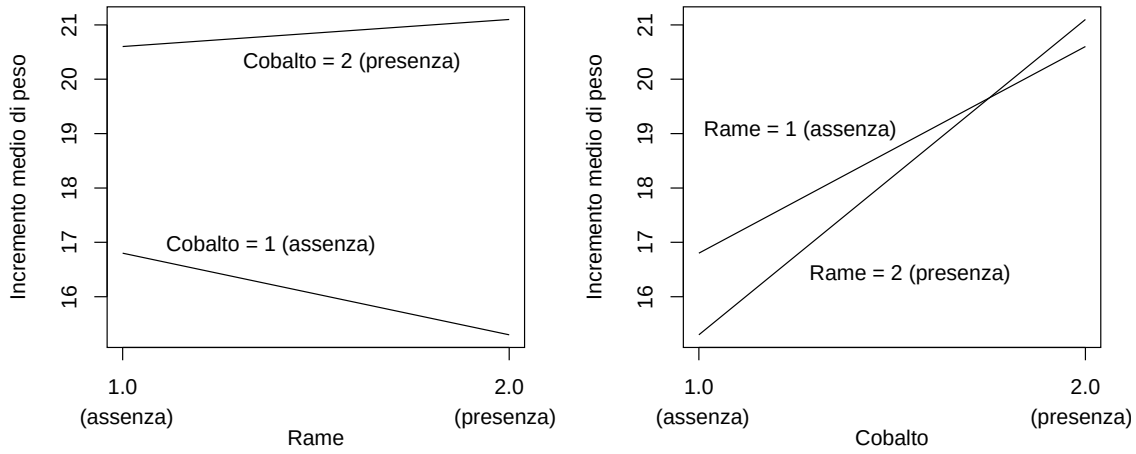


Figura 2.4. Grafici delle interazioni per l'esperimento `dietepec`.

e μ_{22} (presenza di cobalto). Si può notare che la somministrazione di rame comporta una diminuzione dell'incremento medio di peso in assenza di cobalto, ma un aumento quando nella dieta è presente anche il cobalto. Analogamente, il grafico a destra mostra che l'incremento medio di peso dovuto al cobalto risulta maggiore quando vi è anche il rame. Se non vi fossero effetti interattivi, in entrambi i grafici le due linee risulterebbero pressoché parallele.

2.2.2 Il modello a effetti dei fattori

Un modello a medie di cella per due fattori sarebbe analogo a quello già visto, per un solo fattore, nell'esempio 2.1:

$$Y_{ijr} = \mu_{ij} + \varepsilon_{ijr} \qquad \mathbb{E}[Y_{ijr}] = \mu_{ij}$$

La matrice $\mathbf{A}$ avrebbe ora ovviamente $t = ab$ colonne e verrebbe moltiplicata per un vettore di t medie di trattamento: $(\mu_{11} \ \ldots \ \mu_{ij} \ \ldots \ \mu_{ab})$.

Un equivalente modello a effetti dei fattori mette però meglio in evidenza gli effetti differenziali singoli e gli effetti interattivi:

$$Y_{ijr} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijr} \qquad \mathbb{E}[Y_{ijr}] = \mu_{ij} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} \tag{2.1}$$

dove si distinguono:

– la media generale del fenomeno:

$$\mu_{..} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}}{ab} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}}{t} \tag{2.2}$$

– l'effetto differenziale del fattore A al livello i, $i = 1, \ldots, a$:

$$\alpha_i = \mu_{i.} - \mu_{..} \tag{2.3}$$

– l'effetto differenziale del fattore B al livello j, $j = 1, \ldots, b$:

$$\beta_j = \mu_{.j} - \mu_{..} \tag{2.4}$$

$$\begin{bmatrix} \mu_{1,1,1} \\ \vdots \\ \mu_{1,1,10} \\ \mu_{1,2,1} \\ \vdots \\ \mu_{1,2,10} \\ \mu_{2,1,1} \\ \vdots \\ \mu_{2,1,10} \\ \mu_{2,2,1} \\ \vdots \\ \mu_{2,2,10} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_{..} \\ \alpha_1 \\ \alpha_2 \\ \beta_1 \\ \beta_2 \\ (\alpha\beta)_{11} \\ (\alpha\beta)_{12} \\ (\alpha\beta)_{21} \\ (\alpha\beta)_{22} \end{bmatrix} = \begin{bmatrix} \mu_{..} + \alpha_1 + \beta_1 + (\alpha\beta)_{11} \\ \vdots \\ \mu_{..} + \alpha_1 + \beta_1 + (\alpha\beta)_{11} \\ \mu_{..} + \alpha_1 + \beta_2 + (\alpha\beta)_{12} \\ \vdots \\ \mu_{..} + \alpha_1 + \beta_2 + (\alpha\beta)_{12} \\ \mu_{..} + \alpha_2 + \beta_1 + (\alpha\beta)_{21} \\ \vdots \\ \mu_{..} + \alpha_2 + \beta_1 + (\alpha\beta)_{21} \\ \mu_{..} + \alpha_2 + \beta_2 + (\alpha\beta)_{22} \\ \vdots \\ \mu_{..} + \alpha_2 + \beta_2 + (\alpha\beta)_{22} \end{bmatrix}$$

Figura 2.5. Forma matriciale del modello a effetti dei fattori per l'esperimento `dietepec`.

– l'effetto interattivo ulteriore della combinazione del fattore A al livello i e del fattore B al livello j:

$$\begin{aligned} (\alpha\beta)_{ij} &= \mu_{ij} - (\mu_{..} + \alpha_i + \beta_j) \\ &= \mu_{ij} - (\mu_{..} + \mu_{i.} - \mu_{..} + \mu_{.j} - \mu_{..}) \\ &= \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..} \end{aligned} \tag{2.5}$$

Inoltre, poiché gli effetti differenziali non sono altro che scostamenti delle medie di fattore dalla media generale, la loro somma è nulla; analogamente per l'effetto interattivo:

$$\sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \quad \sum_{i=1}^{a} (\alpha\beta)_{ij} = 0, \quad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0, \quad \sum_{i=1}^{a} \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0 \tag{2.6}$$

come si riscontra nella tabella 2.2.

2.2.3 La stima dei parametri

Se si esprime in forma matriciale un modello a effetti dei fattori, si ottiene una matrice con diverse colonne linearmente dipendenti (cfr. figura 2.5).

Sia R che SAS rimuovono la dipendenza lineare ponendo vincoli di tipo *corner point*. Nel caso dell'esperimento `dietepec`, R pone $\alpha_1 = 0$, $\beta_1 = 0$, $(\alpha\beta)_{11} = (\alpha\beta)_{12} = (\alpha\beta)_{21} = 0$, ma così facendo stima solo α_2, β_2 e $(\alpha\beta)_{22}$, SAS pone $\alpha_2 = 0$, $\beta_2 = 0$, $(\alpha\beta)_{12} = (\alpha\beta)_{21} = (\alpha\beta)_{22} = 0$, ma così facendo stima solo α_1, β_1 e $(\alpha\beta)_{11}$. La lettura delle stime di tutti i parametri ne risulta quindi poco agevole e appare preferibile, anche se equivalente dal punto di vista informativo, il sistema di vincoli detto "classico", coerente con le uguaglianze (2.6). In dettaglio, sempre restando a `dietepec`:

a) poiché $\sum_{i=1}^{a} \alpha_i = 0$, si pone $\alpha_2 = -\alpha_1$;

b) poiché $\sum_{j=1}^{b} \beta_j = 0$, si pone $\beta_2 = -\beta_1$;

$$\begin{bmatrix} \mu_{1,1,1} \\ \vdots \\ \mu_{1,1,10} \\ \mu_{1,2,1} \\ \vdots \\ \mu_{1,2,10} \\ \mu_{2,1,1} \\ \vdots \\ \mu_{2,1,10} \\ \mu_{2,2,1} \\ \vdots \\ \mu_{2,2,10} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} \mu_{..} \\ \alpha_1 \\ \beta_1 \\ (\alpha\beta)_{11} \end{bmatrix} = \begin{bmatrix} \mu_{..} + \alpha_1 + \beta_1 + (\alpha\beta)_{11} \\ \vdots \\ \mu_{..} + \alpha_1 + \beta_1 + (\alpha\beta)_{11} \\ \mu_{..} - \alpha_1 + \beta_1 - (\alpha\beta)_{11} = \mu_{..} + \alpha_2 + \beta_1 + (\alpha\beta)_{21} \\ \vdots \\ \mu_{..} - \alpha_1 + \beta_1 - (\alpha\beta)_{11} = \mu_{..} + \alpha_2 + \beta_1 + (\alpha\beta)_{21} \\ \mu_{..} + \alpha_1 - \beta_1 - (\alpha\beta)_{11} = \mu_{..} + \alpha_1 + \beta_1 + (\alpha\beta)_{12} \\ \vdots \\ \mu_{..} + \alpha_1 - \beta_1 - (\alpha\beta)_{11} = \mu_{..} + \alpha_1 + \beta_2 + (\alpha\beta)_{12} \\ \mu_{..} - \alpha_1 - \beta_1 + (\alpha\beta)_{11} = \mu_{..} + \alpha_2 + \beta_2 + (\alpha\beta)_{22} \\ \vdots \\ \mu_{..} - \alpha_1 - \beta_1 + (\alpha\beta)_{11} = \mu_{..} + \alpha_2 + \beta_2 + (\alpha\beta)_{22} \end{bmatrix}$$

Figura 2.6. Riparametrizzazione "classica" del modello a effetti dei fattori per l'esperimento `dietepec`. Nella seconda colonna si ha 1 quando il rame è assente (α_1), -1 quando è presente ($\alpha_2 = -\alpha_1$). Analogamente nella terza. La quarta colonna segue dalle due precedenti.

c) poiché $\sum_{i=1}^{a}(\alpha\beta)_{ij} = 0$, si pongono $(\alpha\beta)_{21} = -(\alpha\beta)_{11}$ e $(\alpha\beta)_{12} = -(\alpha\beta)_{22}$; poiché $\sum_{j=1}^{b}(\alpha\beta)_{ij} = 0$, si pongono $(\alpha\beta)_{12} = -(\alpha\beta)_{11}$ e $(\alpha\beta)_{21} = (\alpha\beta)_{22}$; in sostanza, si esprimono gli $(\alpha\beta)_{ij}$ in funzione di $(\alpha\beta)_{11}$:
   - $(\alpha\beta)_{12} = -(\alpha\beta)_{11}$;
   - $(\alpha\beta)_{21} = -(\alpha\beta)_{11}$;
   - $(\alpha\beta)_{22} = \quad (\alpha\beta)_{11}$;

(cfr. tabella 2.2 e figura 2.6).

Quanto alla stima dei parametri, sia con il metodo dei minimi quadrati che con quello della massimizzazione della verosimiglianza si tratta di minimizzare, per qualsiasi σ^2, la somma dei quadrati degli scarti delle osservazioni dai loro valori attesi, ovvero la quantità:

$$Q = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{r=1}^{n/(ab)}(Y_{ijr} - \mu_{ij})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{r=1}^{n/(ab)}[Y_{ijr} - \mu_{..} - \alpha_i - \beta_j - (\alpha\beta)_{ij}]^2$$

Grazie alla proprietà di invarianza funzionale della stima di massima verosimiglianza,[7] si possono tuttavia stimare i singoli parametri mediante funzioni lineari di stimatori più agevoli da individuare.

Quanto a $\mu_{..}$, si può partire dalla (2.1) e sommare rispetto agli indici i e j:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{..} + \sum_{i=1}^{a}\sum_{j=1}^{b}\alpha_i + \sum_{i=1}^{a}\sum_{j=1}^{b}\beta_j + \sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij}$$

Poiché $\sum_{i=1}^{a}\alpha_i = \sum_{j=1}^{b}\beta_j = \sum_{i=1}^{a}\sum_{j=1}^{b}(\alpha\beta)_{ij} = 0$:

$$\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{..} = ab\,\mu_{..} \quad \Rightarrow \quad \mu_{..} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ij}$$

[7] Se $\hat{\theta}$ è lo stimatore di massima verosimiglianza del parametro θ, allora lo stimatore di massima verosimiglianza di $\alpha = g(\theta)$ è $\hat{\alpha} = g(\hat{\theta})$, purché g sia biiettiva.

(nel caso di `dietepec`, $1/(ab) = 1/4$). Analogamente a quanto visto per il modello a un fattore, lo stimatore di una media di trattamento è la corrispondente media della variabile risposta, $\hat{\mu}_{ij} = \bar{Y}_{ij.}$. Si ha quindi:

$$\hat{\mu}_{..} = \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\bar{Y}_{ij.} = \bar{Y}_{...}$$

Sommando poi rispetto all'indice j, e tenendo ancora conto dei vincoli introdotti nella riparametrizzazione "classica", si ha:

$$\sum_{j=1}^{b}\mu_{ij} = \sum_{j=1}^{b}\mu_{..} + \sum_{j=1}^{b}\alpha_i + \sum_{j=1}^{b}\beta_j + \sum_{j=1}^{b}(\alpha\beta)_{ij} = b\,\mu_{..} + b\,\alpha_i$$

da cui, eliminando le somme nulle e sostituendo ai parametri i loro stimatori:

$$\sum_{j=1}^{b}\bar{Y}_{ij.} = b\,\bar{Y}_{...} + b\,\hat{\alpha}_i \quad \Rightarrow \quad \begin{cases} \hat{\alpha}_1 = \frac{\sum_{j=1}^{b}\bar{Y}_{1j.}}{b} - \bar{Y}_{...} = \bar{Y}_{1..} - \bar{Y}_{...} \\ \dots \\ \hat{\alpha}_a = \frac{\sum_{j=1}^{b}\bar{Y}_{aj.}}{b} - \bar{Y}_{...} = \bar{Y}_{a..} - \bar{Y}_{...} \end{cases}$$

Analogamente, sommando rispetto all'indice i:

$$\sum_{i=1}^{a}\bar{Y}_{ij.} = a\,\bar{Y}_{...} + a\,\hat{\beta}_j \quad \Rightarrow \quad \begin{cases} \hat{\beta}_1 = \frac{\sum_{i=1}^{a}\bar{Y}_{i1.}}{a} - \bar{Y}_{...} = \bar{Y}_{.1.} - \bar{Y}_{...} \\ \dots \\ \hat{\beta}_b = \frac{\sum_{1=1}^{a}\bar{Y}_{ib.}}{b} - \bar{Y}_{...} = \bar{Y}_{.b.} - \bar{Y}_{...} \end{cases}$$

Quanto agli effetti interattivi, sostituendo gli stimatori ai parametri nella (2.5):

$$\widehat{(\alpha\beta)}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$$

Riepilogando:

Parametro	Stimatore
$\mu_{..}$	$\hat{\mu}_{..} = \bar{Y}_{...}$
$\alpha_i = \mu_{i.} - \mu_{..}$	$\hat{\alpha}_i = \bar{Y}_{i..} - \bar{Y}_{...}$
$\beta_j = \mu_{.j} - \mu_{..}$	$\hat{\beta}_j = \bar{Y}_{.j.} - \bar{Y}_{...}$
$(\alpha\beta)_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$	$\widehat{(\alpha\beta)}_{ij} = \bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j.} + \bar{Y}_{...}$

Esempio 2.18. Nel caso di `dietepec` i parametri si stimano facilmente calcolando le medie osservate di trattamento e di fattore, come fatto nella tabella 2.2. SAS consente di ottenere gli stessi risultati con un comando `estimate` che risulta un po' complesso in questo caso, ma utile in esperimenti più articolati. Il comando stima i parametri α_i e β_j (effetti individuali dei fattori) mediante funzioni lineari delle medie di fattore, che consistono nella moltiplicazione di queste per un vettore (indicato con **L** nell'help di SAS). Ad esempio, per il parametro α_1 si ha:

$$\alpha_1 = \mu_{1.} - \mu_{..} = \mu_{1.} - \frac{1}{2}(\mu_{1.} + \mu_{2.}) = \frac{1}{2}\mu_{1.} - \frac{1}{2}\mu_{2.} = \begin{bmatrix} 1/2 & -1/2 & 0 & 0 \end{bmatrix}\begin{bmatrix} \mu_{1.} \\ \mu_{2.} \\ \mu_{.1} \\ \mu_{.2} \end{bmatrix}$$

mentre per β_2:

$$\beta_2 = \mu_{.2} - \mu_{..} = \mu_{.2} - \frac{1}{2}(\mu_{.1} + \mu_{.2}) = -\frac{1}{2}\mu_{.1} + \frac{1}{2}\mu_{.2} = \begin{bmatrix} 0 & 0 & -1/2 & 1/2 \end{bmatrix} \begin{bmatrix} \mu_{1.} \\ \mu_{2.} \\ \mu_{.1} \\ \mu_{.2} \end{bmatrix}$$

Le medie di fattore sono $a+b$, ma non è necessario tenere conto di tutte; basta indicare il fattore che interessa e gli elementi non nulli del vettore $\mathbf{L}$. Per α_1 si può quindi scrivere:

```
estimate 'effetto rame=no' a 0.5 -0.5;
```

Dopo il comando `estimate` compaiono una descrizione testuale (obbligatoria), il parametro come indicato nel modello (`model y = a b a*b`) e il vettore $\mathbf{L}$. Il parametro `a` indica a quali medie di fattore va applicato il vettore. È comunque disponibile l'opzione `divisor` che consente di usare solo interi:

```
estimate 'effetto rame=no' a 1 -1 / divisor=2;
```

Analogamente per gli effetti interattivi, che vengono però stimati in termini delle medie di trattamento. Ad esempio, per $(\alpha\beta)_{12}$ si ha:

$$\begin{aligned}
(\alpha\beta)_{12} &= \mu_{12} - \mu_{1.} - \mu_{.2} + \mu_{..} \\
&= \mu_{12} - \frac{1}{2}(\mu_{11} + \mu_{12}) - \frac{1}{2}(\mu_{12} + \mu_{22}) + \frac{1}{4}(\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22}) \\
&= -\frac{1}{4}\mu_{11} + \frac{1}{4}\mu_{12} + \frac{1}{4}\mu_{21} - \frac{1}{4}\mu_{22} \\
&= \begin{bmatrix} -1/4 & 1/4 & 1/4 & -1/4 \end{bmatrix} \begin{bmatrix} \mu_{11} \\ \mu_{12} \\ \mu_{21} \\ \mu_{22} \end{bmatrix}
\end{aligned}$$

La stima si ottiene quindi con:

```
estimate 'effetto rame=no cobalto=si' a*b -1 1 1 -1 / divisor=4;
```

Dando le seguenti istruzioni:

```
proc glm data=diete; class a b t; model y = a b a*b;
estimate 'effetto rame=no'             a    1 -1         /divisor=2;
estimate 'effetto rame=si'             a   -1  1         /divisor=2;
estimate 'effetto cobalto=no'          b    1 -1         /divisor=2;
estimate 'effetto cobalto=si'          b   -1  1         /divisor=2;
estimate 'effetto rame=no cobalto=no' a*b   1 -1 -1  1   /divisor=4;
estimate 'effetto rame=no cobalto=si' a*b  -1  1  1 -1   /divisor=4;
estimate 'effetto rame=si cobalto=no' a*b  -1  1  1 -1   /divisor=4;
estimate 'effetto rame=si cobalto=si' a*b   1 -1 -1  1   /divisor=4;
run;
```

dopo l'output dell'analisi della varianza (v. sezione successiva) si ottengono le seguenti stime:

```
Parameter                        Estimate          Error   t Value   Pr > |t|

effetto rame=no                0.25000000     0.80665634      0.31     0.7584
effetto rame=si               -0.25000000     0.80665634     -0.31     0.7584
effetto cobalto=no            -2.40000000     0.80665634     -2.98     0.0052
effetto cobalto=si             2.40000000     0.80665634      2.98     0.0052
effetto rame=no cobalto=no     0.50000000     0.80665634      0.62     0.5393
effetto rame=no cobalto=si    -0.50000000     0.80665634     -0.62     0.5393
effetto rame=si cobalto=no    -0.50000000     0.80665634     -0.62     0.5393
effetto rame=si cobalto=si     0.50000000     0.80665634      0.62     0.5393
```

Come si vede, le stime degli effetti singoli e interattivi dei fattori coincidono con quelli calcolati nella tabella 2.2.

2.2.4 L'analisi della varianza

L'analisi della varianza non presenta differenze sostanziali rispetto a quella condotta nel caso di un solo fattore. Si tratta in primo luogo di definire le devianze totale, spiegata e residua e i relativi gradi di libertà. Si ha evidentemente:

$$SSTOT = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{r=1}^{n/(ab)} (y_{ijr} - \bar{y}_{...})^2$$

$$SSMOD = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{r=1}^{n/(ab)} (\bar{y}_{ij.} - \bar{y}_{...})^2 = \frac{n}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{...})^2$$

$$SSRES = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{r=1}^{n/(ab)} (y_{ijr} - \bar{y}_{ij.})^2 = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{r=1}^{n/(ab)} e_{ijr}^2$$

Quanto ai gradi di libertà, questi sono $n-1$ per $SSTOT$ (il numero di osservazioni meno uno perché vi sono n scarti dalla media generale ma la loro somma è nulla), $ab-1$ per $SSMOD$ (il numero dei trattamenti meno uno in quanto la somma degli scarti delle medie di trattamento dalla media generale è nulla), $n-ab$ per $SSRES$ (per ogni trattamento vi sono $n/(ab)-1$ gradi di libertà, in quanto è nulla la somma degli scarti tra le osservazioni e la media di trattamento, e si moltiplica per ab in quanto tanti sono i trattamenti).

Si costruisce quindi agevolmente, come nel modello a un fattore, la statistica test:

$$F^* = \frac{MSMOD}{MSRES} = \frac{SSMOD/(ab-1)}{SSRES/(n-ab)} \sim F_{ab-1,n-ab}$$

Esempio 2.19. Eseguendo i calcoli con R sui dati dell'esperimento `dietepec`, le devianze risultano:

```
> attach(dietepec)
> media.gen <- mean(incpeso)
> medie.tratt <- by(incpeso, list(rame, cobalto), mean)
> SSTOT <- sum( (incpeso - media.gen)^2 )
> SSMOD <- 10 * sum((medie.tratt - media.gen)^2)
> SSRES <- sum((incpeso[tratt==1]-medie.tratt[1])^2) +
+   sum((incpeso[tratt==2]-medie.tratt[2])^2) +
```

```
+   sum((incpeso[tratt==3]-medie.tratt[3])^2) +
+   sum((incpeso[tratt==4]-medie.tratt[4])^2)
> SSTOT; SSMOD; SSRES
[1] 1179.9
[1] 242.9
[1] 937
```

I gradi di libertà sono 39 per $SSTOT$, 3 per $SSMOD$ e 36 per $SSRES$. Quindi per il test di ipotesi:

```
> Fstar <- (SSMOD / 3) / (SSRES / 36 )
> Fstar
[1] 3.110779
> p.value <- pf(Fstar, 3, 36, lower.tail=FALSE)
> p.value
[1] 0.03826003
```

Il comando SAS:

```
proc glm data = diete; class a b t; model y = a b a*b; run;
```

produce gli stessi risultati:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	242.900000	80.966667	3.11	0.0383
Error	36	937.000000	26.027778		
Corrected Total	39	1179.900000			

A differenza, tuttavia, di quanto accadeva con un solo fattore, ora interessa anche esaminare la devianza spiegata da ciascun fattore sia separamente (effetti singoli), sia congiuntamente (effetto interattivo). Si scompone quindi la devianza spiegata in tre componenti. Partendo da:

$$\bar{y}_{ij.} - \bar{y}_{...} = \underbrace{\bar{y}_{i..} - \bar{y}_{...}}_{\text{effetto di } A} + \underbrace{\bar{y}_{.j.} - \bar{y}_{...}}_{\text{effetto di } B} + \underbrace{\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}}_{\text{effetto interattivo}}$$

si scompone $SSMOD$ nella somma di SSA, SSB e $SSAB$, dove:

$$SSA = \frac{n}{a}\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2$$

$$SSB = \frac{n}{b}\sum_{j=1}^{b}(\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$SSAB = \frac{n}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

sono le devianze spiegate dal fattore A, dal fattore B e dall'effetto interattivo.[8] I rispettivi gradi di libertà sono $a-1$, $b-1$ e $(a-1)(b-1)$; il numero dei gradi di libertà di $SSAB$ può essere visto come la differenza tra quello di $SSMOD$ e la somma di quelli di SSA e SSB:

$$(ab-1)-(a-1)-(b-1) = ab-a-b+1 = (a-1)(b-1)$$

Esempio 2.20. Calcolando con R:

```
> medie.rame <- by(incpeso, rame, mean)
> medie.cobalto <- by(incpeso, cobalto, mean)
> SSA <- 20 * sum( (medie.rame - media.gen)^2 )
> SSB <- 20 * sum( (medie.cobalto - media.gen)^2 )
> SSAB <- 10 * (
+ sum((medie.tratt[1]-medie.rame[1]-medie.cobalto[1]+ media.gen)^2) +
+ sum((medie.tratt[2]-medie.rame[2]-medie.cobalto[1]+ media.gen)^2) +
+ sum((medie.tratt[3]-medie.rame[1]-medie.cobalto[2]+ media.gen)^2) +
+ sum((medie.tratt[4]-medie.rame[2]-medie.cobalto[2]+ media.gen)^2) )
> SSA; SSB; SSAB
[1] 2.5
[1] 230.4
[1] 10
```

Tenendo conto dei gradi di liberà – $(2-1)=1$, $(2-1)=1$ e $(2-1)(2-1)=1$ – si effettua facilmente il test di ipotesi. Ad esempio, per SSA:

```
> Fstar <- (SSA / 1) / (SSRES / 36)
> Fstar
[1] 0.09605123
> p.value <- pf(Fstar, 1, 36, lower.tail=FALSE)
> p.value
[1] 0.7584078
```

Proseguendo nella lettura dell'output del comando SAS:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
a	1	2.5000000	2.5000000	0.10	0.7584
b	1	230.4000000	230.4000000	8.85	0.0052
a*b	1	10.0000000	10.0000000	0.38	0.5393

Si può notare che solo il cobalto sembra avere un effetto significativo sull'incremento di peso delle pecore.[9]

[8] Per SSA e SSB la somma degli scarti tra le medie per ciascun livello del fattore e la media generale è moltiplicata per il numero delle loro occorrenze. Quanto a SSA, il fattore A è presente in a livelli, quindi vi sono a medie (e altrettante somme di scarti). Ognuna di queste medie è ripetuta per ciascuno dei livelli del fattore B, che sono b, e per ciascuna delle repliche di ciascun trattamento, che sono $n/(ab)$, quindi la somma degli scarti viene moltiplicata per $b\frac{n}{ab} = \frac{n}{a}$. Analogamente per SSB. Quanto alle somme degli scarti per gli effetti interattivi, queste vengono moltiplicate per le repliche di ciascun trattamento, che sono $n/(ab)$.

[9] Il significato di `Type I SS` verrà discusso nel capitolo 3, sez. 3.2.1.

Osservazione. I test d'ipotesi come sopra condotti sono possibili solo se $SSMOD$, SSA, SSB e $SSAB$, ciascuna divisa per σ^2, da un lato e $SSRES/\sigma^2$ dall'altro sono determinazioni di variabili aleatorie indipendenti e distribuite come Chi quadrati. Si è già visto nel capitolo 1 che ciò vale per $SSMOD$. Quanto alle devianze spiegate da singoli parametri, si possono semplificare i calcoli usando una matrice di dati con la variabile risposta centrata (al posto dei valori, i loro scarti dalla media), che, come mostra l'esempio 1.11, conduce agli stessi risultati. In questo caso, nella riparametrizzazione si ha $\mu_{..} = 0$ e ciò equivale ad eliminare la prima colonna della matrice mostrata nella figura 2.6; la matrice $\mathbf{A}$ che ne risulta può essere vista come ottenuta affiancando tre matrici di una sola colonna, ciascuna corrispondente ad un fattore:

$$\mathbf{A} = \underset{\alpha_i}{\mathbf{A}_1} \ : \ \underset{\beta_j}{\mathbf{A}_2} \ : \ \underset{(\alpha\beta)_{ij}}{\mathbf{A}_3}$$

Tali tre matrici (vettori) sono tra loro ortogonali. Le matrici:

$$\mathbf{H}_i = \mathbf{A}_i(\mathbf{A}_i'\mathbf{A}_i)^{-1}\mathbf{A}_i'$$

sono operatori di proiezione ortogonale sull'immagine di ciascuna matrice, cioè sullo spazio generato da ciascuna di esse; in altri termini, ciascun prodotto $\mathbf{H}_i\mathbf{Y}$ proietta il vettore $\mathbf{Y}$ sul sottospazio generato da $\mathbf{A}_i$. Essendo tra loro ortogonali le matrici $\mathbf{A}_i$, sono tali anche le matrici $\mathbf{H}_i$. Inoltre, la loro somma è la matrice di proiezione:

$$\mathbf{H} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$$

che proietta il vettore $\mathbf{Y}$ sul sottospazio generato da tutte le colonne della matrice di riparametrizzazione $\mathbf{A}$. Si ha quindi che:

$$\underset{SSTOT}{\mathbf{Y}'\mathbf{I}\mathbf{Y}} = \underset{SSMOD}{\mathbf{Y}'\mathbf{H}\mathbf{Y}} + \underset{SSRES}{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}} = \underset{SSA}{\mathbf{Y}'\mathbf{H}_1\mathbf{Y}} + \underset{SSB}{\mathbf{Y}'\mathbf{H}_2\mathbf{Y}} + \underset{SSAB}{\mathbf{Y}'\mathbf{H}_3\mathbf{Y}} + \underset{SSRES}{\mathbf{Y}'(\mathbf{I}-\mathbf{H})\mathbf{Y}}$$

e che valgono le condizioni del teorema di Cochran per le matrici $\mathbf{H}_i$ e $(\mathbf{I}-\mathbf{H})$.

2.2.5 I test di ipotesi sui parametri

L'output di SAS riprodotto nell'esempio 2.18 mostra anche i risultati dei test di ipotesi sui parametri. In ciascun test l'ipotesi nulla è H_0 : parametro $= 0$.

Per il parametro $\alpha_1 = \mu_{1.} - \mu_{..} = \frac{1}{2}\mu_{1.} - \frac{1}{2}\mu_{2.}$ si ha:[10]

$$\mathbb{V}[\hat{\alpha}_i] = \mathbb{V}\left[\frac{1}{2}\bar{Y}_{1..} - \frac{1}{2}\bar{Y}_{2..}\right] = \frac{1}{4}\left(\mathbb{V}[\bar{Y}_{1..}] + \mathbb{V}[\bar{Y}_{2..}]\right) = \frac{1}{4}\left(\frac{\sigma^2}{n/a} + \frac{\sigma^2}{n/a}\right) = \frac{1}{4}\frac{2\sigma^2}{n/a} = \frac{\sigma^2}{2n/a}$$

Analogamente per gli altri α_i e per i β_j. Per i parametri $(\alpha\beta)_{ij}$ si perviene in modo simile a:

$$\mathbb{V}[\widehat{(\alpha\beta)}_{ij}] = \frac{1}{16}\frac{4\sigma^2}{n/(ab)} = \frac{\sigma^2}{4n/(ab)}$$

[10] Ciascuna media $\bar{Y}_{i..}$ è media di n/a termini in quanto, essendo l'eseperimento completo e bilanciato, per ciascun livello del fattore A si hanno $n/(ab)$ osservazioni per ciascun livello del fattore B, quindi $b\frac{n}{ab} = \frac{n}{a}$ osservazioni.

Usando $MSRES$ come stima di σ^2, si possono costruire le statistiche test:

$$\frac{\hat{\alpha}_i - 0}{\sqrt{\frac{MSRES}{2n/a}}} \sim t_{n-ab} \qquad \frac{\hat{\beta}_j - 0}{\sqrt{\frac{MSRES}{2n/b}}} \sim t_{n-ab} \qquad \frac{\widehat{(\alpha\beta)}_{ij}}{\sqrt{\frac{MSRES}{4n/(ab)}}} \sim t_{n-ab}$$

dove $n - ab$ è il numero dei gradi di libertà della varianza residua.

Esempio 2.21. Nel caso di `dietepec` si ha $n = 40$, $a = b = 2$, da cui $2n/a = 2n/b = 4n/(ab) = 40$. Il denominatore è quindi lo stesso per tutte le statistiche test. La devianza residua è pari a 937, con 36 gradi di libertà; eseguendo i test con R, per α_1 si ha:

```
> MSRES <- 937 / 36
> StdError <- sqrt(MSRES / 40)
> StdError
[1] 0.8066563
> tstar <- 0.25 / StdError
> tstar
[1] 0.3099213
> p.value <- pt(abs(tstar), 36, lower.tail=FALSE) + # P[t > |tstar|]
+            pt(-abs(tstar), 36)                    # P[t < -|tstar|]
> p.value
[1] 0.7584078
```

I valori così calcolati coincidono con quelli dell'output di SAS visti nell'esempio 2.18.

In sostanza, i test di ipotesi sui parametri si conducono con le stesse modalità già viste per i modelli a un fattore. Non vi sono differenze di rilievo nemmeno per gli intervalli di confidenza.

2.2.6 Se l'effetto interattivo risulta non significativo

I modelli del tipo $Y_{ijr} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijr}$ (formulati in R con: `y ~ a+b+a:b`, in SAS con `y = a b a*b`) vengono detti *gerarchici*, in quanto si tiene conto di tutti i livelli "inferiori" (i parametri relativi a effetti dei singoli fattori) presenti nei livelli "superiori" (i parametri relativi agli effetti interattivi, uno solo nel caso di un modello a due fattori).

L'analisi della varianza potrebbe condurre alla conclusione che alcuni parametri non sono significativi e andrebbero esclusi dal modello. Va notato, però, che non sarebbe corretto eliminare più di un parametro alla volta, in quanto l'eliminazione di un parametro porta ad una nuova scomposizione della devianza (nel caso di esperimenti che non siano completi e bilanciati può portare anche ad una nuova stima dei parametri).

In generale, quindi, si deve procedere un passo alla volta, dai livelli "superiori" a quelli "inferiori". Nel caso di modelli a due fattori, si può eliminare una prima volta solo l'effetto interattivo, se non risulta significativo; solo dopo si può eventualmente eliminare il parametro relativo ad uno dei due fattori.

Esempio 2.22. L'esempio 2.20 aveva mostrato che, nel caso di `dietepec`, solo il cobalto sembra avere un effetto significativo; ad analoghe conclusioni sembra condurre anche il test di ipotesi sui parametri (v. l'output SAS nell'esempio 2.18). A rigore, tuttavia, si può solo concludere che l'effetto interattivo non risulta significativo, cioè che, per quanto

sembra presente dai calcoli effettuati nella tabella 2.2 e dai grafici della figura 2.4, va attribuito a fattori accidentali; si deve quindi ripetere l'analisi escludendolo dal modello. Il comando SAS:

```
proc glm data=diete; class a b t; model y = a b ;
estimate 'effetto rame=no'    a   1 -1  /divisor=2;
estimate 'effetto rame=si'    a  -1  1  /divisor=2;
estimate 'effetto cobalto=no' b   1 -1  /divisor=2;
estimate 'effetto cobalto=si' b  -1  1  /divisor=2;
run;
```

produce in primo luogo la nuova tabella ANOVA:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	232.900000	116.450000	4.55	0.0171
Error	37	947.000000	25.594595		
Corrected Total	39	1179.900000			

dalla quale risulta che la devianza prima "spiegata" dall'effetto interattivo è passata dalla devianza spiegata dal modello a quella residua; tuttavia, poiché sono diminuiti i gradi di libertà del modello (c'è un parametro in meno), il test d'ipotesi fornisce un *p-value* migliore. L'analisi della varianza per i parametri:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
a	1	2.5000000	2.5000000	0.10	0.7564
b	1	230.4000000	230.4000000	9.00	0.0048

e le stime dei parametri prodotte dal comando `estimate`:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
effetto rame=no	0.25000000	0.79991554	0.31	0.7564
effetto rame=si	-0.25000000	0.79991554	-0.31	0.7564
effetto cobalto=no	-2.40000000	0.79991554	-3.00	0.0048
effetto cobalto=si	2.40000000	0.79991554	3.00	0.0048

mostrano che, anche dopo aver eliminato l'effetto interattivo, solo il cobalto risulta avere un effetto significativo. Si procede quindi con un ultimo modello:

```
proc glm data=diete; class a b t; model y = b ;
estimate 'effetto cobalto=no' b    1 -1  /divisor=2;
estimate 'effetto cobalto=si' b   -1  1  /divisor=2;
run;
```

che risulta non solo più semplice dei precedenti, ma anche migliore per il *p-value*:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	230.400000	230.400000	9.22	0.0043
Error	38	949.500000	24.986842		
Corrected Total	39	1179.900000			

2.2.7 Se vi è una sola osservazione per trattamento

Se vi è una sola osservazione per trattamento, se cioè $n = ab$, il modello viene detto *saturo* e l'analisi della varianza non può essere condotta come appena visto.

Ad esempio, la matrice di dati seguente:[11]

```
data premium;
input premium size region;
cards;
140 1 1
100 1 2
210 2 1
180 2 2
220 3 1
200 3 2
;
run;
```

contiene i premi per l'assicurazione di un'automobile in 6 città di 3 dimensioni diverse e appartenenti a 2 regioni. Il fattore A (`size`) ha $a = 3$ livelli, il fattore B ne ha $b = 2$, vi sono in tutto ab trattamenti e $n = ab$ osservazioni, una per ciascuna combinazione dei due fattori.

Se si adottasse un modello con effetto interattivo, l'analisi della varianza non potrebbe essere condotta perché i gradi di libertà di $SSRES$ sarebbero $n - ab = 0$. Da altro punto di vista, $SSRES$ è la somma dei quadrati degli scarti tra le osservazioni Y_{ijr} e le medie di trattamento $\bar{Y}_{ij.}$, ma poiché vi è una sola osservazione per trattamento si ha $\bar{Y}_{ij.} = Y_{ijr}$, quindi $SSRES = 0$.

La soluzione più immediata consiste nell'assunzione che non vi sia effetto interattivo. In questo modo il modello diventa:

$$Y_{ij} = \mu_{..} + \alpha_i + \beta_j + \varepsilon_{ij}$$

la devianza che sarebbe stata spiegata da $SSAB$ diventa così $SSRES$ e si può procedere analogamente a quanto già visto.

In particolare, per la stima del parametro α_1 si ha:

$$\alpha_1 = \mu_{1.} - \mu_{..} = \mu_{1.} - \frac{1}{3}(\mu_{1.} + \mu_{2.} + \mu_{3.} = \frac{2}{3}\mu_{1.} - \frac{1}{3}\mu_{2.} - \frac{1}{3}\mu_{3.}$$

e si usa quindi il comando:

```
estimate 'size=1' size 2 -1 -1 /divisor=3;
```

Analogamente per α_2 e α_3. Per β_1 e β_2, essendo due i livelli, si usando comandi `estimate` uguali a quelli già visti per l'esperimento `dietepec`.

L'analisi si effettua quindi con:

[11] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 882 (file CH20TA02.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`).

```
proc glm data = premium; class size region; model premium = size region;
estimate 'size=1' size 2 -1 -1 /divisor=3;
estimate 'size=2' size -1 2 -1 /divisor=3;
estimate 'size=3' size -1 -1 2 /divisor=3;
estimate 'region=1' region 1 -1 /divisor=2;
estimate 'region=2' region -1 1 /divisor=2;
run;
```

2.3 Esperimenti completi e bilanciati con tre o più fattori

I modelli per esperimenti completi e bilanciati con tre più fattori sono una semplice estensione di quelli con due fattori. Si deve peraltro tenere conto di effetti interattivi più complessi, che possono risultare o meno significativi.

Nel caso di tre fattori A, B e B, con numeri di livelli rispettivamente a, b e c, vi sono $t = abc$ trattamenti, ciascuno dei quali viene somministrato a $n/(abc)$ unità, e altrettante medie di trattamento μ_{ijk}.

Vi sono poi $a + b + c$ medie di fattore; le medie della variabile risposta per le unità cui è stato somministrato l'i-esimo livello del fattore A, il j-esimo livello del fattore B e il k-esimo elemento del fattore C sono:

$$\mu_{i..} = \frac{\sum_{j=1}^{b}\sum_{k=1}^{c}\mu_{ijk}}{bc} \qquad \mu_{.j.} = \frac{\sum_{i=1}^{a}\sum_{k=1}^{c}\mu_{ijk}}{ac} \qquad \mu_{..k} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\mu_{ijk}}{ab}$$

La media generale è:

$$\mu_{...} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\mu_{ijk}}{abc}$$

Esempio 2.23. La matrice dei dati `surr`[12] contiene il peso delle ghiandole surrenali rilevato in 64 topi secondo il ceppo paterno ($a = 4$ livelli), il ceppo materno ($b = 4$ livelli) e il sesso ($c = 2$ livelli). Con SAS la media generale, 1.2832813, si ottiene con:

```
proc means data=peso mean; var y; run;
```

le 32 medie di trattamento con:

```
proc means data=peso mean; var y; by cepa cema sex notsorted; run;
```

Quanto alle medie di fattore, si possono visualizzare in modo sintentico con:

```
proc means data=peso maxdec=2 mean std;class cepa cema sex; ways 1;var y;
output out=outmeans1 mean=media_y std=std_y;
run;
proc print data=outmeans1; var cepa cema sex media_y; run;
```

che produce:

```
Obs    cepa    cema    sex    media_y

 1      .       .      1     0.80188
```

[12] Scaricabile da `http://web.mclink.it/MC1166/ModelliStatistici/surr.csv`.

```
 2     .     .     2    1.76469
 3     .     1     .    1.19625
 4     .     2     .    1.33313
 5     .     3     .    1.37563
 6     .     4     .    1.22813
 7     1     .     .    1.21375
 8     2     .     .    1.31563
 9     3     .     .    1.36875
10     4     .     .    1.23500
```

Già dall'esame delle medie di fattore può rilevarsi che il sesso sembra avere un effetto piuttosto netto e che i ceppi paterno e materno 2 e 3 sembrano avere maggiore effetto di quelli 1 e 4.

Gli effetti differenziali dei fattori sono:

$$\alpha_i = \mu_{i..} - \mu_{...} \qquad \beta_j = \mu_{.j.} - \mu_{...} \qquad \gamma_k = \mu_{..k} - \mu_{...}$$

Vi sono poi $ab + ac + bc$ effetti interattivi doppi; per $(\alpha\beta)_{ij}$ si ha:

$$(\alpha\beta)_{ij} = \mu_{ij.} - [\mu_{...} + \alpha_i + \beta_j] = \mu_{ij.} - \mu_{...} - \mu_{i..} + \mu_{...} - \mu_{.j.} + \mu_{...} = \mu_{ij.} - \mu_{i..} - \mu_{.j.} + \mu_{...}$$

Analogalente:

$$(\alpha\gamma)_{ik} = \mu_{i.k} - \mu_{i..} - \mu_{..k} + \mu_{...}, \qquad (\beta\gamma)_{jk} = \mu_{.jk} - \mu_{.j.} - \mu_{..k} + \mu_{...}$$

e abc effetti tripli, cioè ulteriori rispetto agli effetti singoli e interattivi doppi:

$$\begin{aligned}(\alpha\beta\gamma)_{ijk} &= \mu_{ijk} - [\mu_{...} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk}] \\ &= \mu_{ijk} - \mu_{...} - \mu_{i..} + \mu_{...} - \mu_{.j.} + \mu_{...} - \mu_{..k} + \mu_{...} \\ &\quad - \mu_{ij.} + \mu_{i..} + \mu_{.j.} - \mu_{...} \\ &\quad - \mu_{i.k} + \mu_{i..} + \mu_{..k} - \mu_{...} \\ &\quad - \mu_{.jk} + \mu_{.j.} + \mu_{..k} - \mu_{...} \\ &= \mu_{ijk} - \mu_{ij.} - \mu_{i.k} - \mu_{.jk} + \mu_{i..} + \mu_{.j.} + \mu_{..k} - \mu_{...}\end{aligned}$$

Ne risulta il modello a effetti dei fattori:

$$Y_{ijkr} = \mu_{...} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkr}$$

Analogamente a quanto visto per il modello a due fattori, si ha:

$$\sum_{i=1}^{a} \alpha_i = \sum_{j=1}^{b} \beta_j = \sum_{k=1}^{c} \gamma_k = 0$$

$$\sum_{i=1}^{a} (\alpha\beta)_{ij} = \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0 \quad \sum_{i=1}^{a} (\alpha\gamma)_{ik} = \sum_{k=1}^{c} (\alpha\gamma)_{ik} = 0 \quad \sum_{j=1}^{b} (\beta\gamma)_{jk} = \sum_{k=1}^{c} (\beta\gamma)_{jk} = 0$$

$$\sum_{i=1}^{a} (\alpha\beta\gamma)_{ijk} = \sum_{j=1}^{b} (\alpha\beta\gamma)_{ijk} = \sum_{k=1}^{c} (\alpha\beta\gamma)_{ijk}$$

2.3.1 La stima dei parametri

Come già visto nel caso di due fattori, in una riparametrizzazione *corner point* verrebbero posti uguali a zero molti parametri e la lettura delle stime risulterebbe ardua. Conviene quindi ricorrere, col SAS, al comando `estimate`.

Effetti singoli

Nel caso del `sesso` dell'esperimento `surr` si ha una situazione uguale a quella già vista nel caso di due fattori, in quanto, come i fattori di `dietepec`, `sesso` ha due soli livelli ed i relativi effetti differenziali vengono espressi in termine delle medie di fattore:

$$\gamma_1 = \mu_{..1} - \mu_{...} = \mu_{..1} - \frac{1}{2}(\mu_{..1} + \mu_{..2}) = \frac{1}{2}\mu_{..1} - \frac{1}{2}\mu_{..2}$$

quindi dopo:

```
proc glm data=peso; class cepa cema sex;
model y=cepa cema sex cepa*cema cepa*sex cema*sex cema*cepa*sex;
```

(le variabili che compaiono nei comandi `estimate` devono essere prima specificate nel `model`) si scrive:

```
estimate 'sex1' sex  1 -1 /divisor=2;
estimate 'sex2' sex -1  1 /divisor=2;
```

La situazione per `cepa` e `cema` è analoga; ad esempio, per l'effetto del primo livello di cepa:

$$\alpha_1 = \mu_{1..} - \mu_{...} = \mu_{1..} - \frac{1}{4}(\mu_{1..} + \mu_{2..} + \mu_{3..} + \mu_{4..}) = \frac{3}{4}\mu_{1..} - \frac{1}{4}\mu_{2..} - \frac{1}{4}\mu_{3..} - \frac{1}{4}4..$$

e per il terzo livello di `cema`:

$$\beta_1 = \mu_{.3.} - \mu_{...} = \mu_{.3.} - \frac{1}{4}(\mu_{.1.} + \mu_{.2.} + \mu_{3..} + \mu_{.4.}) = -\frac{1}{4}\mu_{.1.} - \frac{1}{4}\mu_{.2.} + \frac{3}{4}\mu_{.3.} - \frac{1}{4}\mu_{.4.}$$

quindi si scrive:

```
estimate 'cepa1' cepa   3 -1 -1 -1 /divisor=4;
estimate 'cepa2' cepa  -1  3 -1 -1 /divisor=4;
estimate 'cepa3' cepa  -1 -1  3 -1 /divisor=4;
estimate 'cepa4' cepa  -1 -1 -1  3 /divisor=4;

estimate 'cema1' cema   3 -1 -1 -1 /divisor=4;
estimate 'cema2' cema  -1  3 -1 -1 /divisor=4;
estimate 'cema3' cema  -1 -1  3 -1 /divisor=4;
estimate 'cema4' cema  -1 -1 -1  3 /divisor=4;
```

Effetti interattivi doppi

Nel caso degli effetti interattivi doppi, i parametri vanno espessi in termini delle medie $\mu_{ij.}, \mu_{i.k}$ e $\mu_{.jk}$. Ad esempio, per l'effetto interattivo del secondo livello di `cema` e del primo livello di `sex` si usano le medie $\mu_{.jk}$:

$$\begin{aligned}(\beta\gamma)_{21} &= \mu_{.21} - \mu_{.2.} - \mu_{..1} + \mu_{...} \\ &= \mu_{.21} - \frac{1}{2}(\mu_{.21} + \mu_{.22}) - \frac{1}{4}(\mu_{.11} + \mu_{.21} + \mu_{.31} + \mu_{.41}) \\ &\quad + \frac{1}{8}(\mu_{.11} + \mu_{.12} + \mu_{.21} + \mu_{.22} + \mu_{.31} + \mu_{.32} + \mu_{.41} + \mu_{.42})\end{aligned}$$

da cui, ordinando i termini:

$$(\beta\gamma)_{21} = -\frac{1}{8}\mu_{.11} + \frac{1}{8}\mu_{.12} + \frac{3}{8}\mu_{.21} - \frac{3}{8}\mu_{.22} - \frac{1}{8}\mu_{.31} + \frac{1}{8}\mu_{.32} - \frac{1}{8}\mu_{.41} + \frac{1}{8}\mu_{.42}$$

quindi per gli effetti interattivi di `cema` e `sex` si scrive:

```
estimate 'cema1-sex1'  cema*sex   3 -3 -1  1  -1  1 -1  1  /divisor=8;
estimate 'cema2-sex1'  cema*sex  -1  1  3 -3  -1  1 -1  1  /divisor=8;
estimate 'cema3-sex1'  cema*sex  -1  1 -1  1   3 -3 -1  1  /divisor=8;
estimate 'cema4-sex1'  cema*sex  -1  1 -1  1  -1  1  3 -3  /divisor=8;

estimate 'cema1-sex2'  cema*sex  -3  3  1 -1   1 -1  1 -1  /divisor=8;
estimate 'cema2-sex2'  cema*sex   1 -1 -3  3   1 -1  1 -1  /divisor=8;
estimate 'cema3-sex2'  cema*sex   1 -1  1 -1  -3  3  1 -1  /divisor=8;
estimate 'cema4-sex2'  cema*sex   1 -1  1 -1   1 -1 -3  3  /divisor=8;
```

Analogamente per gli effetti interattivi di `cepa` e `sex` (poiché sia `cepa` che `cema` sono a quattro livelli i coefficienti sono uguali).

Per l'effetto interattivo del primo livello di `cepa` e del secondo di `cema` si usano invece le medie $\mu_{ij.}$:

$$\begin{aligned}(\alpha\beta)_{12} &= \mu_{12.} - \mu_{1..} - \mu_{.2.} + \mu_{...} \\ &= \mu_{12.} - \frac{1}{4}(\mu_{11.} + \mu_{12.} + \mu_{13.} + \mu_{14.}) \\ &\quad - \frac{1}{4}(\mu_{12.} + \mu_{22.} + \mu_{32.} + \mu_{42.}) \\ &\quad + \frac{1}{16}(\mu_{11.} + \mu_{12.} + \mu_{13.} + \mu_{14.} + \mu_{21.} + \mu_{22.} + \mu_{23.} + \mu_{24.} \\ &\qquad + \mu_{31.} + \mu_{32.} + \mu_{33.} + \mu_{34.} + \mu_{41.} + \mu_{42.} + \mu_{43.} + \mu_{44.})\end{aligned}$$

da cui, ordinando i termini:

$$\begin{aligned}(\alpha\beta)_{12} = &-\frac{3}{16}\mu_{11.} + \frac{9}{16}\mu_{12.} - \frac{3}{16}\mu_{13.} - \frac{3}{16}\mu_{14.} + \frac{1}{16}\mu_{21.} - \frac{3}{16}\mu_{22.} + \frac{1}{16}\mu_{23.} + \frac{1}{16}\mu_{24.} \\ &+ \frac{1}{16}\mu_{31.} - \frac{3}{16}\mu_{32.} + \frac{1}{16}\mu_{33.} + \frac{1}{16}\mu_{34.} + \frac{1}{16}\mu_{41.} - \frac{3}{16}\mu_{42.} + \frac{1}{16}\mu_{43.} + \frac{1}{16}\mu_{44.}\end{aligned}$$

quindi si scrive:

```
estimate 'cepa1-cema1' cepa*cema   9 -3 -3 -3  -3  1  1  1  -3  1  1  1  -3  1  1  1  /divisor=16;
estimate 'cepa2-cema1' cepa*cema  -3  1  1  1   9 -3 -3 -3  -3  1  1  1  -3  1  1  1  /divisor=16;
estimate 'cepa3-cema1' cepa*cema  -3  1  1  1  -3  1  1  1   9 -3 -3 -3  -3  1  1  1  /divisor=16;
estimate 'cepa4-cema1' cepa*cema  -3  1  1  1  -3  1  1  1  -3  1  1  1   9 -3 -3 -3  /divisor=16;

estimate 'cepa1-cema2' cepa*cema  -3  9 -3 -3   1 -3  1  1   1 -3  1  1   1 -3  1  1  /divisor=16;
estimate 'cepa2-cema2' cepa*cema   1 -3  1  1  -3  9 -3 -3   1 -3  1  1   1 -3  1  1  /divisor=16;
estimate 'cepa3-cema2' cepa*cema   1 -3  1  1   1 -3  1  1  -3  9 -3 -3   1 -3  1  1  /divisor=16;
estimate 'cepa4-cema2' cepa*cema   1 -3  1  1   1 -3  1  1   1 -3  1  1  -3  9 -3 -3  /divisor=16;

estimate 'cepa1-cema3' cepa*cema  -3 -3  9 -3   1  1 -3  1   1  1 -3  1   1  1 -3  1  /divisor=16;
estimate 'cepa2-cema3' cepa*cema   1  1 -3  1  -3 -3  9 -3   1  1 -3  1   1  1 -3  1  /divisor=16;
estimate 'cepa3-cema3' cepa*cema   1  1 -3  1   1  1 -3  1  -3 -3  9 -3   1  1 -3  1  /divisor=16;
estimate 'cepa4-cema3' cepa*cema   1  1 -3  1   1  1 -3  1   1  1 -3  1  -3 -3  9 -3  /divisor=16;

estimate 'cepa1-cema4' cepa*cema  -3 -3 -3  9   1  1  1 -3   1  1  1 -3   1  1  1 -3  /divisor=16;
estimate 'cepa2-cema4' cepa*cema   1  1  1 -3  -3 -3 -3  9   1  1  1 -3   1  1  1 -3  /divisor=16;
estimate 'cepa3-cema4' cepa*cema   1  1  1 -3   1  1  1 -3  -3 -3 -3  9   1  1  1 -3  /divisor=16;
estimate 'cepa4-cema4' cepa*cema   1  1  1 -3   1  1  1 -3   1  1  1 -3  -3 -3 -3  9  /divisor=16;
```

Effetti interattivi tripli

Nel caso degi effetti interattivi tripli, i parametri vanno espressi in termini delle medie di trattamento ($abc = 32$ nel caso di `surr`). I comandi `estimate` risultano in tali casi di maggiore complessità; si parte infatti da:

$$(\alpha\beta\gamma)_{ijk} = \mu_{ijk} - \mu_{ij.} - \mu_{i.k} - \mu_{.jk} + \mu_{i..} + \mu_{.j.} + \mu_{..k} - \mu_{...}$$

e si sostituiscono i termini dopo il primo con le medie, rispettivamente, di 2, 4, 4, 8, 8, 16 e 32 medie di trattamento.

Fortunatamente ciò è raramente necessario, in quanto gli effetti interattivi tripli o più risultano spesso non significativi. Conviene quindi effettura per prima cosa un'analisi della varianza.

2.3.2 L'analisi della varianza

Il modello per esperimenti con tre (o più) fattori è un modello *gerarchico*, nel senso che si tiene conto espressamente di tutti gli effetti inclusi negli effetti interattivi. Ad esempio, se si vuole testare la significatività di un effetto $(\alpha\beta)_{ij}$, si includono espressamente nel modello gli effetti α_i e β_j; solo così, infatti, l'effetto interattivo può essere correttamente interpretato come effetto *ulteriore* rispetto a quelli singoli. Analogamente, se si vuole testare un effetto triplo, si includono non solo i tre effetti singoli, ma anche i tre effetti delle loro combinazioni a due a due.

Il primo obiettivo consiste nella verifica della significatività prima dell'effetto interattivo triplo, poi di quelli doppi, infine degli effetti singoli, allo scopo di semplificare il modello escludendo (uno alla volta, partendo dai livelli "superiori") gli effetti che risultassero non significativi.

Esempio 2.24. Nel caso di `surr`, eseguendo in SAS:

```
proc glm data=peso;
 class cepa cema sex;
 model y = cepa cema sex cepa*cema cepa*sex cema*sex cepa*cema*sex;
run;
```

si ottiene in primo luogo la tabella ANOVA:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	31	17.42386094	0.56206003	14.55	<.0001
Error	32	1.23595000	0.03862344		
Corrected Total	63	18.65981094			

I gradi di libertà totali sono $n-1 = 63$; poiché vi sono $t = abc = 32$ medie di trattamento, i gradi di libertà del modello sono $t - 1 = 31$, restano quindi 32 gradi di libertà per la componente accidentale (vi sono due osservazioni, quindi un grado di libertà, per ciascun trattamento). Il test d'ipotesi conferma la significatività del modello. Quanto ai parametri:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
cepa	3	0.24826719	0.08275573	2.14	0.1142
cema	3	0.34605469	0.11535156	2.99	0.0456
sex	1	14.83212656	14.83212656	384.02	<.0001
cepa*cema	9	1.30946406	0.14549601	3.77	0.0025
cepa*sex	3	0.02507969	0.00835990	0.22	0.8843
cema*sex	3	0.39949219	0.13316406	3.45	0.0280
cepa*cema*sex	9	0.26337656	0.02926406	0.76	0.6549

si può osservare che i gradi di libertà dei parametri relativi agli effetti interattivi sono il prodotto dei gradi di libertà dei parametri coinvolti. Importa comunque soprattutto notare che l'effetto interattivo triplo risulta non significativo. Si ripete quindi l'analisi escludendolo dal modello, e si ottiene:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	22	17.16048438	0.78002202	21.33	<.0001
Error	41	1.49932656	0.03656894		
Corrected Total	63	18.65981094			

I gradi di libertà e la devianza dell'effetto interattivo tripo sono ora passati dal modello alla componente accidentale e il valore di F^* è aumentato. In realtà non è possibile un confronto diretto tra i due F^* perché, al fine di confrontare i due test di ipotesi, si deve tenere conto del cambiamento dei gradi di libertà. Comunque, calcolando con R:

```
> Fstar1 <- 0.56206003 / 0.03862344
> Fstar1
[1] 14.55230
> Fstar2 <- 0.78002202 / 0.03656894
> Fstar2
[1] 21.33018
> p.value <- pf(Fstar1, 31, 32, lower.tail=FALSE)
> p.value
[1] 1.134918e-11
> p.value <- pf(Fstar2, 22, 41, lower.tail=FALSE)
> p.value
[1] 5.882806e-16
```

si vede che il *p-value* è nettamente diminuito. Passando poi ai test sui parametri:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
cepa	3	0.24826719	0.08275573	2.26	0.0955
cema	3	0.34605469	0.11535156	3.15	0.0349
sex	1	14.83212656	14.83212656	405.59	<.0001
cepa*cema	9	1.30946406	0.14549601	3.98	0.0010
cepa*sex	3	0.02507969	0.00835990	0.23	0.8759
cema*sex	3	0.39949219	0.13316406	3.64	0.0204

si nota che, mentre devianze e varianze corrette sono rimaste immutate (in quanto si tratta di un esperimento completo e bilanciato), i *p-value* sono leggermente diversi in quanto sono cambiati i gradi di libertà della componente accidentale. In ogni caso, l'effetto interattivo `cepa*sex`, ovvero $(\alpha\gamma)_{ik}$, viene confermato non significativo. Si ripete quindi l'analisi escludendolo dal modello e si ottengono, dopo la nuova tabella ANOVA, i nuovi test sui parametri:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
cepa	3	0.24826719	0.08275573	2.39	0.0816
cema	3	0.34605469	0.11535156	3.33	0.0280
sex	1	14.83212656	14.83212656	428.11	<.0001
cepa*cema	9	1.30946406	0.14549601	4.20	0.0006
cema*sex	3	0.39949219	0.13316406	3.84	0.0158

Ora gli effetti interattivi rimasti risultano entrambi significativi. Il parametro `cepa`, ovvero α_i, sembra avere un effetto singolo scarsamente significativo, ma non può essere rimosso dal modello; se così si facesse, infatti, la devianza spiegata da `cepa` confluirebbe in quella spiegata da `cepa*cema`, che non potrebbe essere più interpretato come effetto ulteriore rispetto agli effetti singoli.

Una volta semplificato il modello, si può passare alla stima dei parametri col comando `estimate`. Avendo ridotto i parametri da 7 a 5, ci si può limitare ai comandi elencati nella sezione precedente, evitando sia gli 8 relativi all'effetto `cepa*sex`, sia i 32 relativi agli effetti interattivi tripli.

2.4 Esperimenti a blocchi randomizzati

Se le unità sperimentali non sono omogenee tra loro, si usa raggruppare le unità in blocchi omogenei rispetto alla variabile risposta, per poi somministrare casualmente i trattamenti alle unità di ciascun blocco (*disegno a blocchi randomizzati*).

In questo modo si cerca sia di ridurre la variabilità accidentale, sia di aumentare la validità delle inferenze sugli effetti dei trattamenti.

Esempio 2.25. Nell'esperimento `dietetop` si intende verificare l'efficacia di 5 diete rilevando il peso di 40 topi. I topi appartengono però a 8 nidiate diverse e, pertanto, una parte della variabilità del peso potrebbe essere l'effetto di fattori genetici. Si scelgono quindi 5 topi per ciascuna nidiata e si somministrano loro casualmente le 5 diete.

Si deve notare che, mentre i trattamenti sono sotto il controllo del ricercatore e sono quindi fattori sperimentali a pieno titolo, la variabile di blocco è un fattore osservazionale; sarebbe quindi arduo ipotizzare relazioni di causa-effetto tra la variabile di blocco e la variabile risposta. D'altra parte, l'esperimento è finalizzato a studiare gli effetti dei trattamenti e si serve dei blocchi solo per ridurre la componente accidentale del modello;

ne segue, tra l'altro, che si assume *assenza di interazioni* tra i trattamenti e la variabile di blocco. Il modello è quindi del tipo:

$$Y_{ij} = \mu_{..} + \beta_i + \tau_j + \varepsilon_{ij}$$

dove:

- $\mu_{..}$ è una costante;
- β_i sono gli effetti dei b blocchi, con $i = 1, \ldots, b$ e $\sum_{i=1}^{b} \beta_i = 0$;
- τ_j sono gli effetti dei t trattamenti, con $j = 1, \ldots, t$ e $\sum_{j=1}^{t} \tau_j = 0$.

Non vi sono medie μ_{ij}, in quanto si ha una sola osservazione per ciascuna coppia blocco-trattamento. Vi sono comunque medie di blocco e di trattamento:

$$\mu_{i.} = \frac{\sum_{j=1}^{t} Y_{ij}}{t} \qquad \mu_{.j} = \frac{\sum_{i=1}^{b} Y_{ij}}{b}$$

i cui stimatori sono $\bar{Y}_{i.}$ e $\bar{Y}_{.j}$.

Gli stimatori dei parametri sono quindi:

$$\hat{\mu}_{..} = \bar{Y}_{..} \qquad \hat{\beta}_i = \bar{Y}_{i.} - \bar{Y}_{..} \qquad \hat{\tau}_j = \bar{Y}_{.j} - \bar{Y}_{..}$$

2.4.1 L'analisi della varianza

Obiettivo dell'analisi è verificare la significatività degli effetti del trattamento, mediante un confronto tra la variabilità spiegata da questo ed una variabilità accidentale *depurata degli effetti della variabile di blocco.*

La devianza spiegata dal modello non è altro che la somma delle devianze spiegate dai blocchi e dal trattamento:

$$SSMOD = SSBL + SSTR = t \sum_{i=1}^{b} (\bar{y}_{i.} - \bar{y}_{..})^2 + b \sum_{j=1}^{t} (\bar{y}_{.j} - \bar{y}_{..})^2$$

ed i gradi di libertà sono $(b-1) + (t-1)$.

La variabile aleatoria residuo è:

$$\begin{aligned} e_{ij} &= Y_{ij} - \hat{Y}_{ij} \\ &= Y_{ij} - [\hat{\mu}_{..} + \hat{\beta}_i + \hat{\tau}_j \\ &= Y_{ij} - \bar{Y}_{..} - (\bar{Y}_{i.} - \bar{Y}_{..}) - (\bar{Y}_{.j} - \bar{Y}_{..}) \\ &= Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..} \end{aligned}$$

quindi la devianza residua è:

$$SSRES = \sum_{i=1}^{b} \sum_{j=1}^{t} (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$$

con $(bt-1) - (b-1+t-1) = bt - b - t + 1 = (b-1)(t-1)$ gradi di libertà (essendo $bt-1$ quelli della devianza totale).

Esempio 2.26. Con SAS si può usare anche `proc anova`:

```
proc anova data=diete;
 class bl tr;
 model y = bl tr;
run;
```

e si ottiene:

```
                                 Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F

Model                      11    6446.340750     586.030977     14.42   <.0001
Error                      28    1137.729000      40.633179
Corrected Total            39    7584.069750

              R-Square     Coeff Var      Root MSE        y Mean

              0.849984      9.762113      6.374416      65.29750

Source                     DF       Anova SS    Mean Square   F Value   Pr > F

bl                          7    6099.469750     871.352821     21.44   <.0001
tr                          4     346.871000      86.717750      2.13   0.1029
```

L'output potrebbe essere riorganizzato come segue, per tenere conto delle specifiche finalità dell'analisi della varianza in caso di disegni a blocchi randomizzati:

```
                                 Sum of
Source                     DF       Squares    Mean Square   F Value   Pr > F

tr                          4     346.871000      86.717750      2.13   0.1029
bl                          7    6099.469750     871.352821     21.44   <.0001
Error                      28    1137.729000      40.633179
Corrected Total            39    7584.069750
```

in quanto ciò che interessa è la significatività del parametro τ_j, quindi la statistica test:

$$F^* = \frac{MSTR}{MSRES} = \frac{SSTR/(t-1)}{SSRES/[(b-1)(t-1)]} \sim F_{t-1,(b-1)(t-1)}$$

Si vede che il risultato del test porta ad accettare l'ipotesi nulla $H_0 : \sum_{j=1}^{t} \tau_j^2 = 0$.

Nell'esempio precedente si sarebbe ottenuto un *p-value* ancora maggiore (0.7935) se non si fosse inclusa nel modello la variabile di blocco. In tal caso, però, le conclusioni dell'analisi sarebbero comunque state viziate dalla mancata considerazione dell'eterogeneità delle unità sperimentali. Può essere utile un ulteriore esempio.

Esempio 2.27. La matrice dei dati `RiskPremium`[13] riporta i punteggi da 0 (minimo) a 20 (massimo) attribuiti da 15 manager a 3 metodi di valutazione del rischio usati nel determinare il premio che sono disposti a pagare per una polizza di assicurazione. I 15 manager sono divisi secondo l'età in 5 gruppi, da 1 (i più anziani) a 5 (i più giovani). Se si usa un modello senza fattore di blocco (`model y = method`), si ottiene l'output:

[13] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 896 (file CH21TA01.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`).

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	202.8000000	101.4000000	6.23	0.0139
Error	12	195.2000000	16.2666667		
Corrected Total	14	398.0000000			

Si può rifiutare l'ipotesi nulla, secondo la quale i punteggi attribuiti ai tre metodi sono uguali, con un livello di significatività 0.05, ma non con un livello 0.01. Se si usa l'età come fattore di blocco (`model y = age method`) si ottiene un output che, riorganizzato come nell'esempio precedente, risulta:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model					
age	4	171.3333333	42.8333333	14.36	0.0010
method	2	202.8000000	101.4000000	33.99	0.0001
Error	8	23.8666667	2.9833333		
Corrected Total	14	398.0000000			

Come si vede, dopo aver ridotto la variabilità accidentale da 195.2 a $23.8\bar{7}$ grazie alla separata considerazione della variabilità indotta dall'età, si ottiene un netto miglioramento del *p-value* nel test di ipotesi sul parametro `method`.

In entrambi gli esempi si ottengono *p-value* bassi anche per i test di ipotesi sulla variabile di blocco, ma questo, mentre conferma che le medie della varibile risposta sono diverse per i diversi blocchi, non autorizza ulteriori conclusioni. Ad esempio, nell'esperimento `RiskPremium` potrebbe risultare che i manager più giovani hanno ricevuto una formazione più orientata ai metodi quantitativi rispetto a quelli anziani e, quindi, sarebbe la formazione, non l'età, la vera variabile esplicativa.

2.5 Esperimenti non bilanciati

Finora si sono considerati esperimenti nei quali vi era lo stesso numero di osservazioni per ciascun trattamento (esperimenti bilanciati). Quanto ciò non avviene, la scomposizione della varianza non può più basarsi sull'ortogonalità delle colonne della matrice di riparametrizzazione.

Considendo un solo fattore con due livelli, si hanno due trattamenti; se l'esperimento è bilanciato, ad esempio se vi sono 2 osservazioni per ciascun trattamento, si può avere una matrice di riparametrizzazione come la $\mathbf{A}$ (versione semplificata della matrice mostrata nella figura 2.6); se invece vi sono una sola osservazione per il primo trattamento e 2 per il secondo, si ha una matrice come la $\mathbf{B}$:

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \end{bmatrix}$$

Come si vede, la seconda e la terza colonna sono ortogonali in $\mathbf{A}$ (il loro prodotto è $1 - 1 - 1 + 1 = 0$), ma non in $\mathbf{B}$ ($1 - 1 + 1 = 1$). Matrici come la $\mathbf{B}$ non rispettano le

Tabella 2.3. Medie generale, di trattamento e di fattore per la matrice dei dati `growhorm`. Con $\mu_{ij}(n)$ si indica la media della variabile risposta per il livello i-esimo del primo fattore e il livello j del secondo, calcolata su n unità sperimentali.

Variazione tasso di crescita				
	Fattore B - Ritardo nello sviluppo osseo			
Fattore A - Sesso	$j=1$: grave	$j=2$: medio	$j=3$: leggero	Medie di riga
$i=1$: maschi	$\mu_{11}(3)=2.00$	$\mu_{12}(2)=1.90$	$\mu_{13}(2)=0.90$	$\mu_{1.}(7)=1.657$
$i=2$: femmine	$\mu_{21}(1)=2.40$	$\mu_{22}(3)=2.10$	$\mu_{23}(3)=0.90$	$\mu_{2.}(7)=1.629$
Medie di colonna	$\mu_{.1}(4)=2.10$	$\mu_{.2}(5)=2.02$	$\mu_{.3}(5)=0.90$	$\mu_{..}(14)=1.643$

condizioni richiamate nell'Osservazione a pag. 57; ne segue che, se vi sono due fattori e si considera anche la loro interazione, non si ha più $SSMOD = SSA + SSB + SSAB$.

In esperimenti non bilanciati accade anche che, se si calcolano le medie come sopra fatto nella tabella 2.2, la media generale non è una media semplice delle medie di riga e di colonna, ma una media ponderata; conseguentemente, non è detto che le somme degli effetti interattivi così calcolati siano nulle.

Esempio 2.28. Si vogliono studiare gli effetti del sesso (fattore A con 2 livelli: 1 per i maschi e 2 per le femmine) e del ritardo nello sviluppo osseo (fattore B con 3 livelli: 1 per grave, 2 per medio, 3 per leggero) sulla somministrazione dell'ormone della crescita a bambini che ne sono carenti. Essendovi $a = 2$ livelli per un fattore e $b = 3$ per il secondo, vi sono $ab = 6$ trattamenti. Si scelgono a caso tre bambini per ciascun trattamento, 18 in totale, contando di osservarli per un anno; tuttavia, per vari motivi, non si riesce a mantenere il contatto con 4 di essi, quindi si dispone alla fine di sole 14 osservazioni, che vengono registrate nella matrice di dati `growhorm`.[14] Nella tabella 2.3 si sono calcolate le medie generale, di fattore e di trattamento. Si può notare che la media generale è media semplice delle medie di riga, in quanto vi sono tanti maschi quante femmine, ma non delle medie di colonna, in quanto:

$$\frac{\mu_{.1} + \mu_{.2} + \mu_{.3}}{3} \neq 1.643 = \frac{4\mu_{.1} + 5\mu_{.2} + 5\mu_{.3}}{14}$$

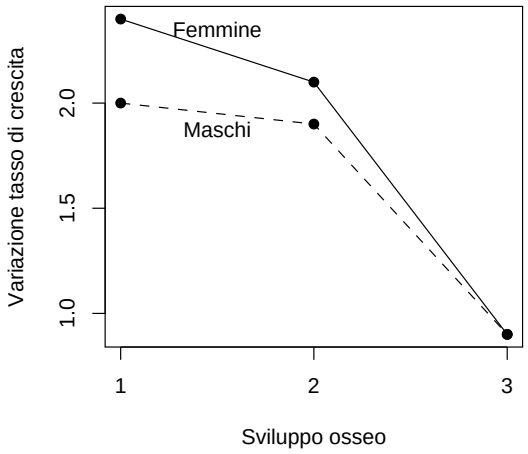
Ne segue che la somma degli effetti del fattore B, se fosse calcolata come nella tabella 2.2, non sarebbe nulla; si calcolerebbe:

$$\beta_1 = \mu_{.1} - \mu_{..} = 0.457 \qquad \beta_2 = \mu_{.2} - \mu_{..} = 0.377 \qquad \beta_3 = \mu_{.3} - \mu_{..} = -0.743$$

e si avrebbe: $\beta_1 + \beta_2 + \beta_3 = 0.091$.

Con esperimenti non bilanciati conviene adottare un modello regressivo, secondo le linee anticipate nel capitolo 1 (pag. 15). Nella pratica, come si vedrà, programmi come R e SAS consentono di pervenire direttamente ai risultati che interessano, ma può essere comunque utile vedere cosa accade dietro le quinte.

[14] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 956 (file CH21TA01.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`).



	y	a	b
1	1.4	1	1
2	2.4	1	1
3	2.2	1	1
4	2.1	1	2
5	1.7	1	2
6	0.7	1	3
7	1.1	1	3
8	2.4	2	1
9	2.5	2	2
10	1.8	2	2
11	2.0	2	2
12	0.5	2	3
13	0.9	2	3
14	1.3	2	3

Figura 2.7. Matrice dei dati e grafico delle medie di trattamento per `growhorm`.

2.5.1 Costruzione di un modello regressivo e test di ipotesi

In generale, si tratta di costruire un modello completo del tipo:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

per ottenere la devianza residua $SSRES(F)$ ("F" per *full model*). Per eseguire test del tipo:

$$H_0 : \beta_2 = 0 \qquad\qquad H_1 : \beta_2 \neq 0$$

si costruisce un *modello ridotto*:

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

e se ne calcola la devianza residua $SSRES(R)$ ("R" sta per *reduced model*). Si confronta quindi la riduzione della devianza residua con quella del modello completo, tenendo conto dei loro gradi di libertà, e si costruisce la statistica test:

$$F^* = \frac{\dfrac{SSRES(R) - SSRES(F)}{\mathrm{gdl}_R - \mathrm{gdl}_F}}{\dfrac{SSRES(F)}{\mathrm{gdl}_F}} \sim F_{\mathrm{gdl}_R - \mathrm{gdl}_F, \mathrm{gdl}_F}$$

Come si vedrà meglio nel capitolo 3, sez. 3.2.1, la differenza $SSRES(R) - SSRES(F)$ non è altro che la devianza spiegata dalla variabile X_2 quando aggiunta al modello ridotto, $SSMOD(X_2 \mid X_1)$, che a sua volta è indipendente da $SSRES(F)$, quindi sono rispettate le condizioni del teorema di Cochran.

Se $SSRES(R) - SSRES(F)$ è piccola in rapporto a $SSRES(F)$, l'aggiunta della variabile X_2 al modello ridotto non cambia sostanzialmente l'adattamento ai dati e si può accettare l'ipotesi nulla.

Esempio 2.29. Nel caso di `growhorm`, il grafico delle medie di fattore (figura 2.7)[15] sembra indicare la presenza di un effetto interattivo. In luogo di un modello a effetti dei fattori:

$$Y_{ijr} = \mu_{..} + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijr}$$

si procede alla definizione di variabili corrispondenti ai due fattori e ai loro livelli. Si muove dal sistema di vincoli "classico" (cfr. sez. 2.2.3):

$$\sum_{i=1}^{a} \alpha_i = 0, \qquad \sum_{j=1}^{b} \beta_j = 0, \qquad \sum_{i=1}^{a} (\alpha\beta)_{ij} = 0, \qquad \sum_{j=1}^{b} (\alpha\beta)_{ij} = 0$$

e si sostituiscono i parametri. Quanto a α_i, viene sostituito da un termine $\alpha_1 X_{ijr1}$, dove al variare di i la variabile X_{ijr1} assume tanti valori quanti sono i livelli del fattore A e valori tali che la loro somma sia nulla:

$$\alpha_i \qquad \rightarrow \qquad \alpha_1 X_{ijr1}, \quad \begin{cases} X_{1jr1} = 1 & \text{(Fattore } A\text{, primo livello)} \\ X_{2jr1} = -1 & \text{(Fattore } A\text{, secondo livello)} \end{cases}$$

quindi, per qualsiasi j e r:

$$\alpha_1 + \alpha_2 = 0 \quad \rightarrow \quad \alpha_1 X_{1jr1} + \alpha_1 X_{2jr1} = \alpha_1 - \alpha_1 = 0$$

Il parametro β_j, avendo tre livelli, viene sostituito da una coppia di termini $\beta_1 X_{ijr2} + \beta_2 X_{ijr3}$, in cui al variare di j sia X_{ijr2} che X_{ijr3} hanno tre valori a somma nulla:

$$\beta_j \qquad \rightarrow \qquad \beta_1 X_{ijr2} + \beta_2 X_{ijr3}, \quad \begin{cases} X_{i1r2} = 1,\ X_{i1r3} = 0 & \text{(Fattore } B\text{, 1° livello)} \\ X_{i2r2} = 0,\ X_{i2r3} = 1 & \text{(Fattore } B\text{, 2° livello)} \\ X_{i3r2} = -1,\ X_{i3r3} = -1 & \text{(Fattore } B\text{, 3° livello)} \end{cases}$$

quindi, per qualsiasi i e r:

$$\begin{aligned} \beta_1 + \beta_2 + \beta_3 = 0 \ \rightarrow & (\beta_1 X_{i1r2} + \beta_2 X_{i1r3}) + (\beta_1 X_{i2r2} + \beta_2 X_{i2r3}) + (\beta_1 X_{i3r2} + \beta_2 X_{i3r3}) \\ & = \beta_1 + \beta_2 - \beta_1 - \beta_2 = 0 \end{aligned}$$

Conseguentemente, l'unico parametro $(\alpha\beta)_{ij}$ per l'effetto interattivo viene sostituito dalla somma di due termini: $(\alpha\beta)_{11} X_{ijr1} X_{ijr2} + (\alpha\beta)_{12} X_{ijr1} X_{ijr3}$, e si ha, per qualsiasi j e qualsiasi r:

$$\begin{aligned} (\alpha\beta)_{1j} + (\alpha\beta)_{2j} = 0 \quad \rightarrow \quad & (\alpha\beta)_{11} X_{1jr1} X_{2jr2} + (\alpha\beta)_{12} X_{1jr1} X_{2jr3} + \\ & (\alpha\beta)_{11} \cdot 1 \cdot (-1) + (\alpha\beta)_{11} \end{aligned}$$

[15] Il grafico è stato creato con R. Per crearne uno analogo con SAS:

```
proc means data=growhorm mean; class a b; types a*b; var y;
 output out=ghmeans mean=media_y;
run;
symbol1 i=join c=black  v=plus   l=1;
symbol2 i=join c=black  v=plus   l=2;
proc gplot data=ghmeans;
 plot media_y*b=a;
run;
```

i	j	r
1	1	1
1	1	2
1	1	3
1	2	1
1	2	2
1	3	1
1	3	2
2	1	1
2	2	1
2	2	2
2	2	3
2	3	1
2	3	2
2	3	3

	X_1	X_2	X_3	X_1X_2	X_1X_3
1	1	1	0	1	0
1	1	1	0	1	0
1	1	1	0	1	0
1	1	0	1	0	1
1	1	0	1	0	1
1	1	−1	−1	−1	−1
1	1	−1	−1	−1	−1
1	−1	1	0	−1	0
1	−1	0	1	0	−1
1	−1	0	1	0	−1
1	−1	0	1	0	−1
1	−1	−1	−1	1	1
1	−1	−1	−1	1	1
1	−1	−1	−1	1	1

Figura 2.8. Matrice di riparametrizzazione del modello regressivo per `growhorm`.

Il modello diventa quindi:

$$Y_{ijr} = \mu_{..} + \underbrace{\alpha_1 X_{ijr1}}_{\text{Effetto } A} + \underbrace{\beta_1 X_{ijr2} + \beta_2 X_{ijr3}}_{\text{Effetto } B} + \underbrace{(\alpha\beta)_{11} X_{ijr1} X_{ijr2} + (\alpha\beta)_{12} X_{ijr1} X_{ijr3}}_{\text{Effetto interattivo}} + \varepsilon_{ijr}$$

a cui corrisponde la matrice mostrata nella figura 2.8. Si può notare che gli elementi della colonna X_1X_2 sono i prodotti dei corrispondenti elementi delle colonne X_1 e X_2, quelli della colonna X_1X_3 sono i prodotti dei corrispondenti elementi di X_1 e X_3. Dopo aver immesso la matrice in R:

```
> X <- matrix( c(rep(c(1, 1, 1, 0, 1, 0),3), rep(c(1, 1, 0, 1, 0, 1),2),
+                rep(c(1, 1,-1,-1,-1,-1),2),     c(1,-1, 1, 0,-1, 0),
+                rep(c(1,-1, 0, 1, 0,-1),3), rep(c(1,-1,-1,-1, 1, 1),3)),
+              nrow=14, byrow=TRUE)
```

si possono calcolare i coefficienti di regressione e le devianza totale, spiegata e residua:

```
> n <- nrow(X)
> I <- diag(1, n)
> J <- matrix(1, nrow=n, ncol=n)
> H <- X %*% solve(t(X) %*% X) %*% t(X)
> beta  <- as.vector( solve(t(X) %*% X) %*% t(X) %*% y )
> SSTOT <- as.vector( t(y) %*% (I - J/n) %*% y )
> SSMOD <- as.vector( t(y) %*% (H - J/n) %*% y )
> SSRES <- as.vector( t(y) %*% (I - H) %*% y )
> SSTOT; SSMOD; SSRES
[1] 5.774286
[1] 4.474286
[1] 1.3
```

I coefficienti risultano:

$$\mu_{..} = 1.5, \quad \alpha_1 = -1, \quad \beta_1 = 5, \quad \beta_2 = 3, \quad (\alpha\beta)_{11} = -1, \quad (\alpha\beta)_{12} = 0$$

Si può notare che si ottengono così valori teorici uguali alle medie di trattamento (che ne sono le stime); eseguendo infatti il prodotto **Hy** si ottengono i valori:

$$\hat{y}_{11r} = 2.0, \quad \hat{y}_{12r} = 1.9, \quad \hat{y}_{13r} = 0.9$$
$$\hat{y}_{21r} = 2.4, \quad \hat{y}_{22r} = 2.1, \quad \hat{y}_{23r} = 0.9$$

Con R si può ovviamente usare anche la funzione `lm()` dopo aver creato, come appena visto, la matrice `X`:

```
> mod <- lm(y ~ X[,2] + X[,3] + X[,4] + X[,5] + X[,6])
> anova(mod)
> mod$fitted.values
```

Con SAS, prima si modifica la matrice di dati aggiungendo le variabili X e i loro prodotti:

```
data growhormreg;
 set growhorm;
 if a eq 1 then x1 =  1;
 if a eq 2 then x1 = -1;
 if b eq 1 then do; x2 =  1; x3 =  0; end;
 if b eq 2 then do; x2 =  0; x3 =  1; end;
 if b eq 3 then do; x2 = -1; x3 = -1; end;
 x1x2 = x1 * x2;
 x1x3 = x1 * x3;
run;
```

poi si esegue una regressione con l'opzione `r`, che produce valori teorici e residui:

```
proc reg data = growhormreg;
 model y = x1 x2 x3 x1x2 x1x3 /r;
run;
```

Per sottoporre a verifica la significatività dell'effetto interattivo:

$$H_0 : (\alpha\beta)_{11} = (\alpha\beta)_{12} = 0, \qquad H_1 : (\alpha\beta)_{11} \neq 0 \vee (\alpha\beta)_{12} \neq 0$$

si esegue la regressione sul modello ridotto, quindi si ripetono i calcoli usando, al posto della matrice $\mathbf{X}$ una matrice $\mathbf{X}_R$ mancante delle ultime due colonne; si ottiene:

```
> SSTOTr; SSMODr; SSRESr
[1] 5.774286
[1] 4.398857
[1] 1.375429
```

Si nota che la devianza totale è ovviamente rimasta invariata, mentre la devianza residua è aumentata da 1.3 a 1.3754, quindi il contributo dato dall'effetto interattivo alla spiegazione della devianza, $SSMOD(R) = 0.0754$, è modesto. I gradi di libertà di $SSTOT$ sono $n - 1 = 13$, di cui $p - 1 = 5$ per $SSMOD(F)$ (p è il numero delle colonne della matrice `X`) e 8 per $SSRES(F)$, 3 per $SSMOD(R)$ e 10 per $SSRES(R)$. Eseguito il test:

```
> Fstar <- ( (SSRESr-SSRES)/(10-8) ) / (SSRES / 8)
> pf(Fstar, 2, 8, lower.tail=FALSE)
[1] 0.7980337
```

si accetta l'ipotesi nulla: l'effetto interattivo non è significativo. Si procede in modo analogo per i singoli fattori. Le ipotesi nulle del modello ANOVA:

$$H_0 : \alpha_1 = \alpha_2 = 0 \qquad H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

diventano:

$$H_0 : \alpha_1 = 0 \qquad H_0 : \beta_1 = \beta_2 = 0$$

che corrispondono ai modelli ridotti:

a) test sul fattore A (si prova con un modello ridotto che non lo comprende):

$$Y_{ijr} = \mu_{..} + \beta_1 X_{ijr2} + \beta_2 X_{ijr3} + (\alpha\beta)_{11} X_{ijr1} X_{ijr2} + (\alpha\beta)_{12} X_{ijr1} X_{ijr3} + \varepsilon_{ijr}$$

b) test sul fattore B:

$$Y_{ijr} = \mu_{..} + \alpha_1 X_{ijr1} + (\alpha\beta)_{11} X_{ijr1} X_{ijr2} + (\alpha\beta)_{12} X_{ijr1} X_{ijr3} + \varepsilon_{ijr}$$

Si costruiscono matrici quindi che non contengano, rispettivamente, la seconda oppure la terza e la quarta colonna (ma contengano le ultime due) e, procedendo come sopra, si perviene ai seguenti test:

$$F^* = \frac{(1.42 - 1.3)/(9 - 8)}{1.3/8} = \frac{0.12}{0.1625} = 0.74 \qquad \textit{p-value} = 0.415$$

$$F^* = \frac{(5.4897 - 1.3)(10 - 8)}{1.3/8} = \frac{2.0949}{0.1625} = 12.89 \qquad \textit{p-value} = 0.003$$

Se ne conclude che solo l'effetto del fattore B (ritardo nello sviluppo osseo) è significativo. Si potrebbe anche procedere in modo diverso: una volta trovato che l'effetto interattivo non è significativo, lo si potrebbe escludere dal modello completo, eliminando le ultime due colonne della matrice $\mathbf{X}$, e provare poi con i modelli ridotti. Per il test sul fattore A si escluderebbe ancora la seconda colonna, per il fattore B si escluderebbero la terza e la quarta, come sopra, ma la matrice non avrebbe più le colonne quinta e sesta. Si otterrebbero, rispettivamente, i *p-value* 0.4311 e 0.0008, che porterebbero alla stessa conclusione.

Il test di ipotesi sul modello è basato sul familiare confronto tra la devianza spiegata e quella residua, tenendo conto dei gradi di libertà. Il test di ipotesi su un parametro si basa sul confronto tra la devianza spiegata da questo, quando aggiunto ad un modello che contenga tutti gli altri, e la devianza residua del modello completo; tale devianza, come si vedrà meglio nel capitolo 3, viene detta *devianza di tipo III*.

Sia R che SAS consentono di pervenire direttamente ai test anche senza passare attraverso un modello regressivo. In R, dopo aver caricato la libreria `car`, si usa la procedura `Anova()` con l'opzione `type="III"`; in SAS si usa `proc glm` badando al prospetto in cui si mostra la `Type III SSS` (negli output di `proc glm` si era mostrato finora solo il prospetto con le devianze di tipo I, perché le devianze di tipo I e quelle di tipo III coincidono se vi è ortogonalità, ma in esperimenti non bilanciati le devianze dei due tipi sono diverse).

Esempio 2.30. Invece di usare `proc reg` con una matrice dei dati ristrutturata, si può usare `proc glm`:

```
proc glm data=growhorm;
 class a b;
 model y = a b a*b;
run;
```

La prima parte dell'output conferma la significatività del modello:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	4.47428571	0.89485714	5.51	0.0172
Error	8	1.30000000	0.16250000		
Corrected Total	13	5.77428571			

Seguono i prospetti con le devianze di tipo I e di tipo III:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
a	1	0.00285714	0.00285714	0.02	0.8978
b	2	4.39600000	2.19800000	13.53	0.0027
a*b	2	0.07542857	0.03771429	0.23	0.7980

Source	DF	Type III SS	Mean Square	F Value	Pr > F
a	1	0.12000000	0.12000000	0.74	0.4152
b	2	4.18971429	2.09485714	12.89	0.0031
a*b	2	0.07542857	0.03771429	0.23	0.7980

Le devianze di tipo I sono di tipo sequenziale: prima un modello con solo `a`, poi con `a` e `b`, poi con `a`, `b` e `a*b`. Le devianze di tipo III sono invece relative ai parametri quando aggiunti a modelli che contengano già tutti gli altri, quindi corrispondono a quelle calcolate nell'esempio precedente. Nonostante in questo caso sembri che i due gruppi di test conducano alle stesse conclusioni, i test corretti sono quelli basati sulle devianze di tipo III. Si potrebbe anche ora, come nell'esempio precedente, escludere l'effetto interattivo e ripetere `proc glm` con `model y = a b`; si otterrebbero ancora i `p-value` 0.4311 e 0.0008.

2.5.2 Stima e intervalli di confidenza dei parametri

Nell'esempio 2.28 si è visto che, quando i numeri di repliche dei trattamenti non sono uguali tra loro, le medie di fattore calcolate sui dati non sono medie semplici delle medie di trattamento. Tuttavia, per stimare le medie non si può fare altro che interpretare ciascuna media calcolata some stima e attribuire a ciascuna lo stesso peso, come se i numeri di repliche fossero uguali. In altri termini, per `growhorm` le medie dei livelli del primo fattore, i loro stimatori e le solo stime sono:

$$\mu_{1\cdot} = \frac{\sum_{j=1}^{b} \mu_{1j}}{b} \qquad \hat{\mu}_{1\cdot} = \frac{\sum_{j=1}^{b} \bar{Y}_{1j\cdot}}{b} \qquad \hat{\mu}_{1\cdot} = \frac{2.0 + 1.9 + 0.9}{3} = 1.6$$

$$\mu_{2\cdot} = \frac{\sum_{j=1}^{b} \mu_{2j}}{b} \qquad \hat{\mu}_{2\cdot} = \frac{\sum_{j=1}^{b} \bar{Y}_{2j\cdot}}{b} \qquad \hat{\mu}_{2\cdot} = \frac{2.4 + 2.1 + 0.9}{3} = 1.8$$

dove i numeri usati per le stime sono a loro volta le stime dei valori teorici (v. esempio 2.29). Analogamente per il secondo fattore:

$$\mu_{.1} = \frac{\sum_{i=1}^{a} \mu_{i1}}{a} \qquad \hat{\mu}_{.1} = \frac{\sum_{i=1}^{a} \bar{Y}_{i1.}}{a} \qquad \hat{\mu}_{.1} = \frac{2.0 + 2.4}{2} = 2.2$$
$$\mu_{.2} = \frac{\sum_{i=1}^{a} \mu_{i2}}{a} \qquad \hat{\mu}_{.2} = \frac{\sum_{i=1}^{a} \bar{Y}_{i2.}}{a} \qquad \hat{\mu}_{.2} = \frac{1.9 + 2.1}{2} = 2.0$$
$$\mu_{.3} = \frac{\sum_{i=1}^{a} \mu_{i3}}{a} \qquad \hat{\mu}_{.3} = \frac{\sum_{i=1}^{a} \bar{Y}_{i3.}}{a} \qquad \hat{\mu}_{.3} = \frac{0.9 + 0.9}{2} = 0.9$$

Poiché $\bar{Y}_{ij.} = \frac{\sum_{r=1}^{n_{ij}} Y_{ijr}}{n_{ij}}$ e le Y_{ijr} hanno varianza σ^2 e sono indipendenti, le varianze di tali stimatori sono:

$$\mathbb{V}[\hat{\mu}_{i.}] = \frac{1}{b^2} \sum_{j=1}^{b} \mathbb{V}[\bar{Y}_{ij.}] = \frac{1}{b^2} \sum_{j=1}^{b} \frac{\sigma^2}{n_{ij}} = \frac{\sigma^2}{b^2} \sum_{j=1}^{b} \frac{1}{n_{ij}} \tag{2.7}$$

$$\mathbb{V}[\hat{\mu}_{.j}] = \frac{1}{a^2} \sum_{i=1}^{a} \mathbb{V}[\bar{Y}_{ij.}] = \frac{1}{a^2} \sum_{i=1}^{a} \frac{\sigma^2}{n_{ij}} = \frac{\sigma^2}{a^2} \sum_{i=1}^{a} \frac{1}{n_{ij}} \tag{2.8}$$

Nei test σ^2, in quanto incognita, viene sostituita dal suo stimatore $MSRES$.

I singoli parametri possono essere stimati considerandoli come differenze tra una media di fattore e la media generale; ad esempio, nel caso di un parametro β_j relativo ad un fattore con tre livelli, lo stimatore di β_1 è:

$$\hat{\beta}_1 = \hat{\mu}_{.1} - \hat{\mu}_{..} = \hat{\mu}_{.1} + \frac{1}{3}(\hat{\mu}_{.1} + \hat{\mu}_{.2} + \hat{\mu}_{.3}) = \frac{2}{3}\hat{\mu}_{.1} - \frac{1}{3}\hat{\mu}_{.2} - \frac{1}{3}\hat{\mu}_{.3}$$

e la sua varianza è:

$$\mathbb{V}[\hat{\beta}_1] = \frac{4}{9}\mathbb{V}[\hat{\mu}_{.1}] + \frac{1}{9}\mathbb{V}[\hat{\mu}_{.2}] + \frac{1}{9}\mathbb{V}[\hat{\mu}_{.3}]$$

Esempio 2.31. Nel caso di `growhorm` si ha:

$$\hat{\beta}_1 = \frac{2}{3}2.2 - \frac{1}{3}2.0 - \frac{1}{3}0.9 = 0.5$$
$$\hat{\beta}_2 = -\frac{1}{3}2.2 + \frac{2}{3}2.0 - \frac{1}{3}0.9 = 0.3$$
$$\hat{\beta}_3 = -\frac{1}{3}2.2 - \frac{1}{3}2.0 + \frac{2}{3}0.9 = -0.8$$

Le varianze corrette, usando $MSRES$ come stimatore di σ^2, sono (ricordando che $a = 2$):

$$\begin{aligned}\mathbb{V}[\hat{\beta}_1] &= \frac{4}{9}\frac{MSRES}{a^2}\left(\frac{1}{3} + 1\right) + \frac{1}{9}\frac{MSRES}{a^2}\left(\frac{1}{2} + \frac{1}{2}\right) + \frac{1}{9}\frac{MSRES}{a^2}\left(\frac{1}{2} + \frac{1}{2}\right) \\ &= \frac{1}{9}\frac{MSRES}{a^2}\left(\frac{16}{3} + \frac{5}{6} + \frac{5}{6}\right) = \frac{7}{36}MSRES \\ \mathbb{V}[\hat{\beta}_2] = \mathbb{V}[\hat{\beta}_3] &= \frac{1}{9}\frac{MSRES}{a^2}\left(\frac{4}{3} + \frac{20}{6} + \frac{5}{6}\right) = \frac{11}{72}MSRES\end{aligned}$$

Per procedere al test t sulla base del modello completo, basta usare la devianza residua già calcolata, $SSRES = 1.3$, e dividerla per i suoi gradi di libertà:

```
> MSRES <- SSRES / 8
> beta <- c(0.5,0.3,-0.8)
> StdError <- c(sqrt(7/36*MSRES), rep(sqrt(11/72*MSRES),2))
> tstar <- beta / StdError
> round(tstar, 2)
[1]  2.81  1.90 -5.08
> p.value <- pt(abs(tstar), 8, lower.tail=FALSE) +
+            pt(-abs(tstar), 8)
> round(p.value, 4)
[1] 0.0227 0.0934 0.0010
```

Eseguendo `proc glm` con gli opportuni comandi `estimate`:

```
proc glm data=growhorm;
 class a b;
 model y = a b a*b /clparm;
 estimate 'b1' b 2 -1 -1 /divisor=3;
 estimate 'b2' b -1 2 -1 /divisor=3;
 estimate 'b3' b -1 -1 2 /divisor=3;
run;
```

si ottengono gli stessi risultati:

Parameter	Estimate	Standard Error	t Value	Pr > |t|
b1	0.50000000	0.17775608	2.81	0.0227
b2	0.30000000	0.15756392	1.90	0.0934
b3	-0.80000000	0.15756392	-5.08	0.0010

Quanto agli intervalli di confidenza, usando R come calcolatrice:

```
> estremi <- qt(0.975, 8) * StdError
> beta-estremi
[1]  0.09009376 -0.06334305 -1.16334305
> beta+estremi
[1]  0.9099062  0.6633431 -0.4366569
```

oppure con SAS, usando l'opzione `clparm`:

Parameter	95% Confidence Limits	
b1	0.09009376	0.90990624
b2	-0.06334305	0.66334305
b3	-1.16334305	-0.43665695

Come si vede, i primi due parametri (ritardo nello sviluppo osseo grave e medio) comportano un valore teorico della variabile risposta (variazione del tasso di crescita) superiore alla media, mentre il terzo (leggero ritardo nello sviluppo osseo) comporta un valore inferiore alla media. Dal momento che β_2 rientra in un intervallo con l'estremo inferiore negativo, si possono valutare le differenze tra gli effetti dei diversi livelli del fattore B aggiungendo comandi `contrast` come i seguenti a `proc glm`:

```
contrast 'b1 vs b2' b 1 -1 0;
contrast 'b1 vs b3' b 1 0 -1;
contrast 'b2 vs b3' b 0 1 -1;
```

Si ottiene:

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
b1 vs b2	1	0.07384615	0.07384615	0.45	0.5192
b1 vs b3	1	3.12000000	3.12000000	19.20	0.0023
b2 vs b3	1	2.90400000	2.90400000	17.87	0.0029

Se ne può concludere che la differenza tra i primi due livelli non è significativa, mentre lo è la differenza tra ciascuno dei primi due e il terzo, confermando le conclusioni già tratte. Eseguendo `proc glm` sul modello ridotto, `model y = a b`, si ottengono risultati analoghi.

Capitolo 3

La regressione lineare

Nella regressione lineare si tenta di stabilire una relazione funzionale lineare tra i valori di una o più variabili esplicative e i valori attesi della variabile risposta.

La sezione 3.1 tratta della regressione lineare semplice, in cui compare una sola variabile esplicativa.

La sezione 3.2 tratta della regressione lineare multipla, in cui compaiono più variabili esplicative, compreso il caso in cui la variabile esplicativa sia una sola ma compaia anche al quadrato (regressione polinomiale). Si mostrano i test di ipotesi resi possibili dalle devianze di tipo I, di tipo II e di tipo III, nonché le difficoltà indotte dalle correlazioni tra variabili esplicative (multicollinearità).

3.1 Regressione lineare semplice

Nella *regressione lineare semplice* vi è una sola variabile esplicativa e si adotta un modello del tipo:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i \tag{3.1}$$

dove:

- Y_i è l'i-esimo valore della variabile risposta, $i = 1, \ldots, n$;
- β_0 l'*intercetta*; se X può assumere il valore 0, β_0 è il valore atteso di Y_i quando $X = 0$;
- β_1 è il coefficiente angolare della *retta di regressione* ed esprime il cambiamento di $\mathbb{E}[Y]$ a seguito di un incremento unitario di X;

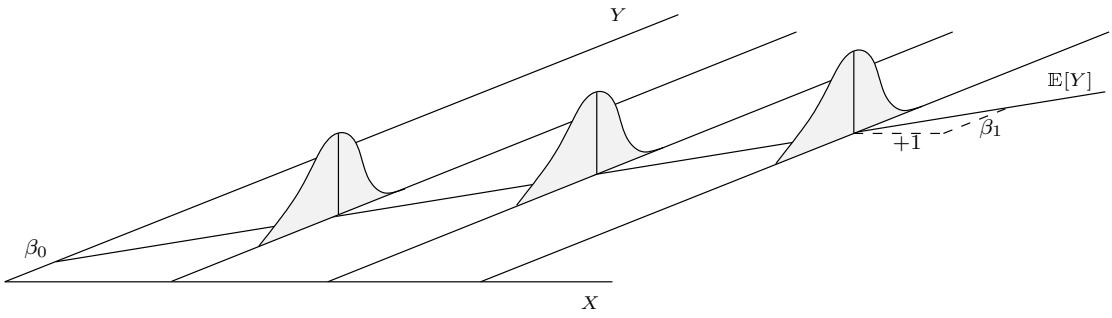


Figura 3.1. Rappresentazione geometrica di un modello regressivo lineare semplice.

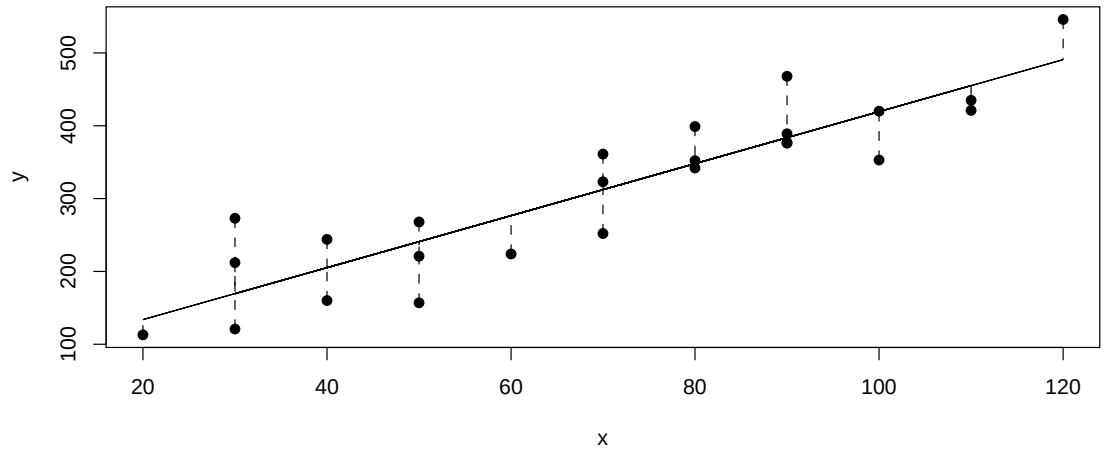


Figura 3.2. Lo *scatter plot* della matrice di dati `toluca`, con la retta di regressione e le distanze da essa dei valori osservati.

– ε_i è una variabile aleatoria errore di distribuzione $N(0, \sigma^2)$; ε_i e ε_j sono indipendenti, quindi $\sigma_{\varepsilon_i,\varepsilon_j} = 0$ per ogni i, j, $i \neq j$.

β_0 e β_1 vengono detti *coefficienti di regressione*.

Come si vede nella figura 3.1 (da confrontare con la figura 2.1), all'ipotesi che ai valori della variabile esplicativa X corrispondano valori significativamente diversi del valore atteso della variabile risposta Y si aggiunge l'ipotesi che esista una relazione lineare tra i primi e i secondi.

Esempio 3.1. L'azienda Toluca produce frigoriferi e parti di ricambio, una delle quali è stata prodotta in passato in lotti di dimensione diversa (da 20 a 120 unità). Dal momento che ogni volta si sono dovuti avviare appositi processi produttivi, comprendenti anche attività indipendenti dalla dimensione del lotto, l'azienda vuole studiare la relazione tra le unità prodotte e le ore di lavoro complessivamente necessarie. La relativa matrice di dati, `toluca`[1] contiene $n = 25$ osservazioni. Il relativo diagramma di dispersione (figura 3.2) mostra che, al crescere della dimensione dei lotti, x, aumentano anche le ore di lavoro necessario, y, e sembrano aumentare secondo una relazione lineare. Si usa quindi il modello lineare (3.1), la cui forma matriciale è:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{25} \end{bmatrix} = \begin{bmatrix} 1 & 80 \\ 1 & 30 \\ \vdots & \vdots \\ 1 & 70 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{25} \end{bmatrix} \qquad \mathbb{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

La matrice $\mathbf{X}$ contiene quindi due colonne, la prima di tutti 1 e la seconda con i valori della variabile esplicativa.

[1] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 19 (file CH01TA01.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/toluca.csv`).

3.1.1 La stima dei coefficienti di regressione e dei valori teorici

Sia che si usi il metodo dei minimi quadrati, sia che si usi quello di massima verosimiglianza, si tratta di minimizzare la quantità:

$$Q = \sum_{i=1}^{n} (Y_i - \mathbb{E}[Y_i])^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

e ciò si ottiene con:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Esempio 3.2. Con R, una volta caricato il file `toluca.csv`, si possono stimare i coefficienti con la funzione `lm()`; passando il risultato alla funzione `model.matrix()` si ottiene la matrice $\mathbf{X}$, che può essere usata per eseguire manualmente il calcolo:

```
> toluca <- read.csv("toluca.csv")
> mod <- lm(y ~ x, data=toluca)
> mod

Call:
lm(formula = y ~ x, data = toluca)

Coefficients:
(Intercept)            x
      62.37         3.57

> X <- model.matrix(mod)
> beta <- solve(t(X) %*% X) %*% t(X) %*% toluca$y
> beta
                 [,1]
(Intercept) 62.365859
x            3.570202
```

Nella regressione lineare semplice, in particolare, si ha:

$$\begin{aligned}
Q &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = \sum_{i=1}^{n} (Y_i^2 + \beta_0^2 + \beta_1^2 X_i^2 - 2\beta_0 Y_i - 2\beta_1 X_i Y_i + 2\beta_0\beta_1 X_i) \\
\frac{\partial Q}{\beta_0} &= \sum_{i=1}^{n} (2\beta_0 - 2Y_i + 2\beta_1 X_i) = -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) \\
\frac{\partial Q}{\beta_1} &= \sum_{i=1}^{n} (2\beta_1 X_i^2 - 2X_i Y_i + 2\beta_0 X_i) = -2\sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)
\end{aligned}$$

Uguagliando a zero la derivata rispetto a β_0:

$$\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum_{i=1}^{n} Y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} X_i = 0 \quad \Rightarrow \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Uguagliando a zero la derivata rispetto a β_1 e sostituendo $\hat{\beta}_0$:

$$\sum_{i=1}^{n} X_i(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = \sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0$$

$$\Rightarrow \quad \hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sigma_{XY}}{\sigma_X^2}$$

Esempio 3.3. Usando il dataframe `toluca` creato nell'esempio precedente, la stima dei coefficienti di regressione può anche essere ottenuta con:

```
> attach(toluca)
> beta1 <- cov(x,y) / var(x)
> beta1
[1] 3.570202
> beta0 <- mean(y) - beta1*mean(x)
> beta0
[1] 62.36586
```

Una volta stimati i coefficienti, i valori teorici si ottengono da:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

che è una stima della relazione lineare tra la variabile esplicativa e il valore atteso della variabile risposta.

Esempio 3.4. Nel caso di `toluca`, si ha:

$$\hat{Y}_i = 62.36586 + 3.570202 X_i$$

Ad esempio, per $X_i = 30$ si ha: $\hat{Y}_i = 169.4719$. Tali valori si trovano nella variabile `fitted.values` del risultato di `lm()`, precedentemente assegnato a `mod`:

```
> mod$fitted.values[x==30]
       2       17       21
169.4719 169.4719 169.4719
```

I valori osservati sono diversi:

```
> y[which(x==30)]
[1] 121 212 273
```

e le differenze sono determinazioni della variabile aleatoria residuo, che si trovano in `mod$residuals`:

```
> mod$residuals[x==30]
        2        17        21
-48.47192  42.52808 103.52808
```

Esempio 3.5. Con SAS, una volta caricato il dataset (basta un copia-incolla dal file CH01TA01.TXT), si può eseguire:

```
proc reg data=toluca;
 model y=x;
 output out=tolucareg predicted=y_hat residual=e_hat;
run;
```

La terza riga richiede la creazione di un dataset in cui siano presenti le colonne indicate dalle parole chiave `predicted` e `residual`, cui vengono assegnati i nomi `y_hat` e `e_hat`. L'output, che verrà esaminato con maggior dettaglio nelle sezioni successive, contiene anche la stima dei coefficienti:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	62.36586	26.17743	2.38	0.0259
x	1	3.57020	0.34697	10.29	<.0001

Il dataset `tolucareg` contiene le colonne dei valori teorici e dei residui osservati; dopo `proc print data=tolucareg; run;`:

Obs	x	y	y_hat	e_hat
1	80	399	347.982	51.018
2	30	121	169.472	-48.472
..	..	...	...	...
25	70	323	312.280	10.720

3.1.2 Il test di ipotesi sul modello e il coefficiente di determinazione

Analogamente a quanto già visto, la devianza può essere scomposta come segue:

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

$$SSTOT = SSMOD + SSRES$$

I rispettivi gradi di libertà sono:

- $SSTOT$: $n-1$, come di consueto;
- $SSMOD$: 1 solo grado di libertà, in quanto vi sono n scarti $\hat{Y}_i - \bar{Y}$, ma tutti gli $\hat{Y}_i$ giacciono sulla stessa retta (sono elementi di uno spazio vettoriale di dimensione 1; in generale, i gradi di libertà della devianza spiegata da un modello di regressione lineare sono tanti quante le variabili esplicative);
- $SSRES$: $n-2$, in quanto vi sono n scarti $Y_i - \hat{Y}_i$, ma i valori attesi $\hat{Y}_i$ sono funzioni delle stime dei due coefficienti di regressione, quindi si perdono due gradi di libertà.

Il numero dei gradi di libertà di $SSTOT$ è uguale alla somma di quelli di $SSMOD$ e di $SSRES$: $n-1 = 1 + (n-2)$.

Da punto di vista più generale, considerando le devianze come forme quadratiche e indicando con n e p, rispettivamente, il numero di righe e di colonne della matrice $\mathbf{X}$:

$$SSTOT = \mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} \qquad \mathrm{rk}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) = n - 1$$

$$SSMOD = \mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y} \qquad \mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad \mathrm{rk}(\mathbf{H}) = p$$

$$\mathrm{rk}\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right) = p - 1$$

$$SSRES = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y} \qquad \mathrm{rk}(\mathbf{I} - \mathbf{H}) = n - p$$

in quanto:

a) per qualsiasi coppia di matrici $\mathbf{A}, \mathbf{B}$:

$$\mathrm{tr}(\mathbf{A} + \mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$$

b) per qualsiasi matrice $\mathbf{A}$ idempotente (e tali sono $\mathbf{I}$, $\frac{1}{n}\mathbf{J}$, $\mathbf{I} - \frac{1}{n}\mathbf{J}$, $\mathbf{H}$, $\mathbf{H} - \frac{1}{n}\mathbf{J}$ e $\mathbf{I} - \mathbf{H}$)[2]:

$$\mathrm{rk}(\mathbf{A}) = \mathrm{tr}(\mathbf{A})$$

c) il rango di $\mathbf{I}$ è n, quello di $\mathbf{H}$ è p;

d) $\frac{1}{n}\mathbf{J}$ ha righe e colonne tutte uguali, il suo rango è 1 e la sua traccia è $n\frac{1}{n} = 1$, quindi:

$$\begin{aligned}
\mathrm{rk}\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right) &= \mathrm{rk}(\mathbf{I}) - \mathrm{rk}\left(\frac{1}{n}\mathbf{J}\right) = n - 1 \\
\mathrm{rk}\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right) &= \mathrm{rk}(\mathbf{H}) - \mathrm{rk}\left(\frac{1}{n}\mathbf{J}\right) = p - 1 \\
\mathrm{rk}(\mathbf{I} - \mathbf{H}) &= \mathrm{rk}(\mathbf{I}) - \mathrm{rk}(\mathbf{H}) = n - p
\end{aligned}$$

$\mathbf{I} - \frac{1}{n}\mathbf{J}$ è una matrice di proiezione ortogonale che proietta $\mathbf{Y}$ su uno spazio a $n - 1$ dimensioni, quello dei valori centrati di $\mathbf{Y}$ (cfr. cap. 1, esempio 1.11.).

$\mathbf{H} - \frac{1}{n}\mathbf{J}$ è una matrice di proiezione ortogonale che proietta $\mathbf{Y}$ su un sottospazio che ha tante dimensioni quante il rango della matrice, quindi $p - 1$ ($2 - 1 = 1$, una retta nel caso della regressione lineare semplice).

$\mathbf{I} - \mathbf{H}$ è una matrice ad essa ortogonale,[3] che proietta $\mathbf{Y}$ sul sottospazio ortogonale al precedente, che ha dimensione $(n - 1) - (p - 1) = n - p$.

In sostanza, come già visto nel caso dei modelli ANOVA, si può applicare il teorema di Cochran per verificare la significatività del modello, contro l'ipotesi nulla che le variabili Y_i abbiano la stessa media, cioè che $\beta_1 = 0$ (retta di regressione orizzontale).

Esempio 3.6. La procedura `proc reg` di SAS usata nell'esempio precedente fornisce in primo luogo un'analisi della varianza:

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	252378	252378	105.88	<.0001

[2] Cfr. capitolo 1, note 14, 15 e 19.

[3] $\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)(\mathbf{I} - \mathbf{H}) = \mathbf{H} - \mathbf{H} - \frac{1}{n}\mathbf{J} + \frac{1}{n}\mathbf{J} = \mathbf{O}$ in quanto $\frac{1}{n}\mathbf{J}\mathbf{H} = \frac{1}{n}\mathbf{J}$. Cfr. cap. 1, nota 19.

```
Error                    23          54825     2383.71562
Corrected Total          24         307203

              Root MSE            48.82331    R-Square     0.8215
              Dependent Mean     312.28000    Adj R-Sq     0.8138
              Coeff Var           15.63447
```

Il risultato del test consente di rifiutare l'ipotesi nulla in favore di quella alternativa, $H_1 : \beta_1 \neq 0$.

L'analisi della varianza in sé, tuttavia, non basta. Nei modelli ANOVA era sufficiente che le medie risultassero significativamente diverse da zero, ma nella regressione lineare semplice interessa che la retta individuata da un $\beta_1 \neq 0$ sia non solo la migliore possibile (quella che riduce al minimo la devianza residua), ma anche che si adatti bene ai dati.

Per valutare la bontà dell'adattamento della retta ai dati, si usa il *coefficiente di determinazione* R^2, definito da:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSMOD}{SSTOT} = 1 - \frac{SSRES}{SSTOT}$$

L'adattamento è tanto migliore quanto più R^2 è vicino a 1.

Esempio 3.7. Nel caso di `toluca`, il valore del coefficiente `R-square`, come risulta dall'output riprodotto nell'esempio precedente, è 0.8215. Infatti:

$$R^2 = \frac{252378}{307203} = 0.8215$$

Si tratta di un valore ragionevolmente vicino a 1.

Esempio 3.8. Si può usare la regressione anche con i dati dell'esperimento `caffeina` (esempio 2.1), interpretando la variabile `tr` come quantitativa (cfr. l'osservazione a pag. 15). L'output di `proc reg data=caffeina; model y=tr; run;` mostra un valore di R^2 nettamente più basso di quello ottenuto per `toluca`: 0.3133. Il peggior adattamento ai dati è ben messo in evidenza dalla figura 3.3, da confrontare con la figura 3.2.

3.1.3 I test di ipotesi sui coefficienti di regressione

Nel caso della regressione lineare semplice, l'ipotesi nulla per il test di ipotesi su β_1 coincide con quella per il modello: $H_0 : \beta_1 = 0$. Il test può essere comunque eseguito in un modo equivalente che vale la pena esaminare da vicino, sia per poter poi effettuare il test anche su β_0, sia per preparare il terreno ai test effetuati nei modelli di regressione multipla.

La variabile aleatoria $\hat{\boldsymbol{\beta}}$ ha distribuzione multinormale, perché è funzione lineare di $\mathbf{Y}$: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Inoltre:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta} \qquad \mathrm{Cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$$

Quindi:

$$\mathbb{E}[\hat{\beta}_0] = \beta_0 \qquad \mathbb{V}[\hat{\beta}_0] = a_{11}\sigma^2 \qquad \hat{\beta}_0 \sim N(\beta_0, a_{11}\sigma^2)$$
$$\mathbb{E}[\hat{\beta}_1] = \beta_1 \qquad \mathbb{V}[\hat{\beta}_1] = a_{22}\sigma^2 \qquad \hat{\beta}_1 \sim N(\beta_1, a_{22}\sigma^2)$$

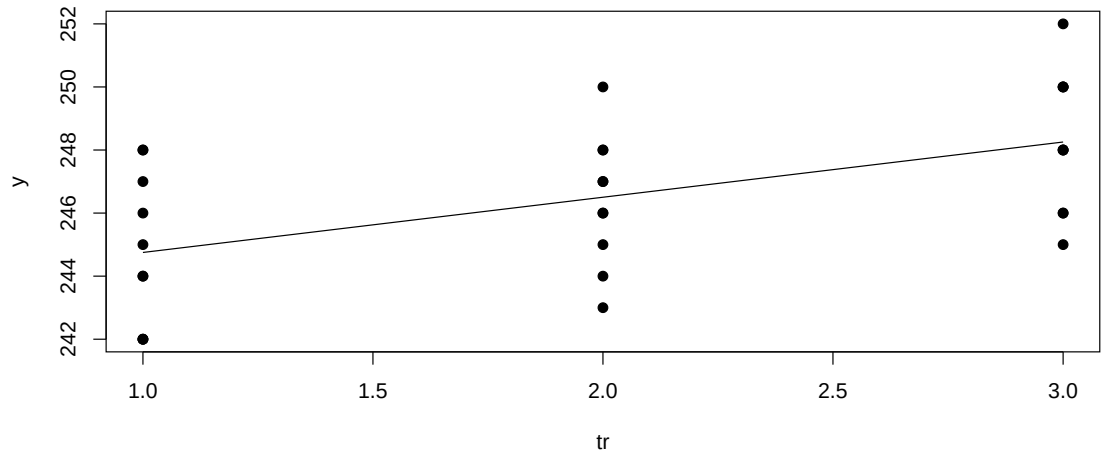


Figura 3.3. Lo *scatter plot* della matrice di dati caffeina, con la retta di regressione.

dove a_{11} e a_{22} sono i due elementi della diagonale principale di $(\mathbf{X}'\mathbf{X})^{-1}$.[4]

Dal momento che la varianza σ^2 non è nota e che un suo stimatore corretto è la varianza residua $MSRES = SSRES/(n-2)$ (v. osservazione a pag. 29), si usano le statistiche test:

$$t^* = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{a_{11} MSRES}} \sim t_{n-2} \qquad t^* = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{a_{22} MSRES}} \sim t_{n-2}$$

tenendo presente che nelle ipotesi nulle $\beta_0 = \beta_1 = 0$.

Esempio 3.9. Nel caso di `toluca`, volendo eseguire manualmente i calcoli con R, il valore di $MSRES$ può essere preso dall'output del SAS, oppure ricalcolato:

```
> ImenoHy <- (I - X %*% A %*% t(X)) %*% y
```

[4]Per comprendere meglio il significato degli elementi a_{11} e a_{22} della matrice $(\mathbf{X}'\mathbf{X})^{-1}$ si può partire da una matrice $\mathbf{X}$ molto semplice:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \qquad \mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix} \qquad (\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \begin{bmatrix} 2.\bar{3} & -1 \\ -1 & 0.5 \end{bmatrix}$$

Si vede così immediatamente che la matrice $\mathbf{X}'\mathbf{X}$, il suo determinante e la sua inversa (prodotto del reciproco del determinante per la matrice di cofattori) sono:

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \qquad n \sum x_i^2 - \left(\sum x_i\right)^2 \qquad \frac{1}{n \sum x_i^2 - \left(\sum x_i\right)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}$$

quindi, ricordando che $\sum (x_i - \bar{x})^2 / n = \sum x_i^2 / n - \bar{x}^2$:

$$a_{11} = \frac{\sum x_i^2}{n \sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\frac{\sum x_i^2}{n^2}}{\frac{\sum x_i^2}{n} - \bar{x}^2} = \frac{\frac{\sum x_i^2}{n^2}}{\frac{\sum (x_i - \bar{x})^2}{n}} = \frac{\frac{\sum (x_i - \bar{x})^2}{n} + \bar{x}^2}{\sum (x_i - \bar{x})^2} = \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}$$

$$a_{22} = \frac{n}{n \sum x_i^2 - \left(\sum x_i\right)^2} = \frac{\frac{1}{n}}{\frac{\sum x_i^2}{n} - \bar{x}^2} = \frac{\frac{1}{n}}{\frac{\sum (x_i - \bar{x})^2}{n}} = \frac{1}{\sum (x_i - \bar{x})^2}$$

```
> SSRES <- t(ImenoHy) %*% ImenoHy
> MSRES <- SSRES / (25 - 2)
> MSRES
         [,1]
[1,] 2383.716
```

Indicando poi con `A` la matrice $(\mathbf{X}'\mathbf{X})^{-1}$:

```
A <- solve(t(X) %*% X)
```

si ha tutto quanto occore per il test d'ipotesi su β_1:

```
> StdError <- sqrt(MSRES * A[2,2])
> StdError
          [,1]
[1,] 0.3469722
> tstar <- beta1 / StdError
> tstar
         [,1]
[1,] 10.28959
> p.value <- pt(abs(tstar), 23, lower.tail=FALSE)+ # P[t > |tstar|]
+            pt(-abs(tstar), 23)                   # P[t < -|tstar|]
> p.value
             [,1]
[1,] 4.448828e-10
```

Per il test su β_0:

```
> StdError <- sqrt(MSRES * A[1,1])
> StdError
         [,1]
[1,] 26.17743
> tstar <- beta0 / StdError
> tstar
         [,1]
[1,] 2.382428
> p.value <- pt(abs(tstar), 23, lower.tail=FALSE)+ # P[t > |tstar|]
+            pt(-abs(tstar), 23)                   # P[t < -|tstar|]
> p.value
           [,1]
[1,] 0.02585094
```

I valori coincidono con quelli prodotti dal SAS e contenuti nell'output riprodotto nell'esempio 3.5. Va ricordato che il test su $beta_0$ ha senso solo se la variabile esplicativa può assumere il valore $X = 0$.

Una volta definita la statistica test, si calcolano facilmente gli intervalli di confidenza. Quello di livello $1 - \alpha$ per β_1 è:

$$\beta_1 \in \left(\hat{\beta}_1 \pm t_{1-\frac{\alpha}{2},n-2}\sqrt{a_{22}MSRES}\right)$$

Analogamente per β_0.

Esempio 3.10. Calcolando manualmente con R ($\alpha = 0.5$):

```
> estremo <- qt(0.975, 23) * sqrt(MSRES * A[1,1])
> c(beta0-estremo, beta0+estremo)
[1]   8.21371 116.51801
> estremo <- qt(0.975, 23) * sqrt(MSRES * A[2,2])
> c(beta1-estremo, beta1+estremo)
[1] 2.852435 4.287969
```

Con SAS basta usare l'opzione `clb` dopo la specificazione del `model`; si ottengono gli stessi valori:

```
Variable     DF       95% Confidence Limits

Intercept     1        8.21371      116.51801
x             1        2.85244        4.28797
```

3.1.4 Le bande di confidenza

Può risultare interessante determinare una *banda di confidenza* per l'intera retta di regressione, ovvero la regione del piano entro la quale questa si colloca con un fissato livello di confidenza. Si tratta di determinare gli intervalli di confidenza per i valori teorici $\hat{Y}_i$.

Ragionando sui dati `toluca`, gli intervalli di confidenza per i coefficienti ci dicono che, con un livello di confidenza del 95%:

$$\beta_0 \in (\ L(\hat{\beta}_0),\ U(\hat{\beta}_0)\ ) = (8.21371, 116.51801)$$
$$\beta_1 \in (\ L(\hat{\beta}_1),\ U(\hat{\beta}_1)\ ) = (2.85244,\quad 4.28797)$$

Si potrebbe pensare che gli intervalli di confidenza per i valori teorici siano:

$$\hat{Y}_i \in (\ L(\hat{\beta}_0)+L(\hat{\beta}_1)X_i,\ U(\hat{\beta}_0)+U(\hat{\beta}_1)X_i\ ) = (8.21371+2.85244X_i, 116.51801+4.28797X_i)$$

e che, quindi, la retta di regressione sia collocata in una regione del piano delimitata da due rette, una "minima" una "massima". Sarebbe un errore. Gli intervalli di confidenza per i coefficienti, infatti, si basano sulle distribuzioni di ciascuno di essi considerato singolarmente, mentre quelli per i valori teorici devono basarsi su una distribuzione che li consideri *entrambi*. Si deve quindi ragionare in altro modo.

I valori teorici sono funzione lineare dei coefficienti stimati: $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Dato un singolo valore di X, indicato con X_h, si ha:

$$\hat{Y}_h = \mathbf{x}_h'\hat{\boldsymbol{\beta}} \qquad\qquad \mathbf{x} = \begin{bmatrix} 1 \\ X_h \end{bmatrix}$$

$\hat{Y}_h$ ha una distribuzione normale. Il valore atteso è $\mathbb{E}[\hat{Y}_h] = Y_h$. Quanto alla varianza, come visto nel capitolo 1, nota 13, da $\hat{Y}_h = \mathbf{x}_h'\hat{\boldsymbol{\beta}}$ segue:

$$\sigma^2_{\hat{Y}_h} = \mathbf{x}_h' \operatorname{Cov}(\hat{\boldsymbol{\beta}})\mathbf{x}_h = \mathbf{x}_h'[(\mathbf{X}'\mathbf{X})^{-1}\sigma^2_Y]\mathbf{x}_h = \sigma^2_Y[\mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h]$$

Sostituendo σ^2_Y con il suo stimatore $MSRES$, si può costruire la statistica test:

$$t^* = \frac{\hat{Y}_h - \mathbb{E}[Y_h]}{\sqrt{MSRES[\mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h]}} \sim t_{n-2}$$

L'intervallo di confidenza per $\mathbb{E}[\hat{Y}_h] = Y_h$ è quindi:

$$Y_h \in \left(\hat{Y}_h \pm t_{1-\frac{\alpha}{2},n-2}\sqrt{MSRES[\mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h]}\right)$$

Per determinare la regione del piano entro cui è compresa la retta di regressione (con un livello di confidenza $1-\alpha$), basta calcolare gli intervalli di confidenza degli Y_i per i diversi valori di X_i.

Esempio 3.11. Restando ai dati `toluca`, per $X_i = 30$ si ha (`MSRES` e la matrice `A` sono stati calcolati nell'esempio 3.9):

```
> beta.hat <- matrix(c(beta0, beta1), nrow=2)
> x.i <- matrix(c(1,30), nrow=2)
> y.hat <- t(x.i) %*% beta.hat
> y.hat
         [,1]
[1,] 169.4719
> estremo <- qt(0.975, 23) * sqrt(MSRES * (t(x.i) %*% A %*% x.i))
> y.hat - estremo; y.hat + estremo
         [,1]
[1,] 134.3673
         [,1]
[1,] 204.5765
```

Per $X_i = 100$:

```
> x.i <- matrix(c(1,100), nrow=2)
> y.hat <- t(x.i) %*% beta.hat
> y.hat
         [,1]
[1,] 419.3861
> estremo <- qt(0.975, 23) * sqrt(MSRES * (t(x.i) %*% A %*% x.i))
> y.hat - estremo; y.hat + estremo
         [,1]
[1,] 389.8615
         [,1]
[1,] 448.9106
```

E così via. R consente di ottenere gli intervalli di confidenza per tutti i valori di X con:

```
> predict(mod, interval="confidence")
```

Viene prodotta una matrice di 3 colonne, contenenti rispettivamente $\hat{y}_i$, $L(\hat{y}_i)$ e $U(\hat{y}_i)$, che può essere usata per tracciare un grafico come quello a sinistra nella figura 3.4. In SAS gli intervalli di confidenza per i valori teorici si ottengono usando l'opzione `clm` (*confidence limits* per la *media*, cioè il valore atteso, della variabile risposta). Con:

```
proc reg data=toluca;
 model y=x / clb clm;
 output out=tolucareg pred=y_hat residual=e_hat lclm=y_lwr uclm=y_upr;
run;
```

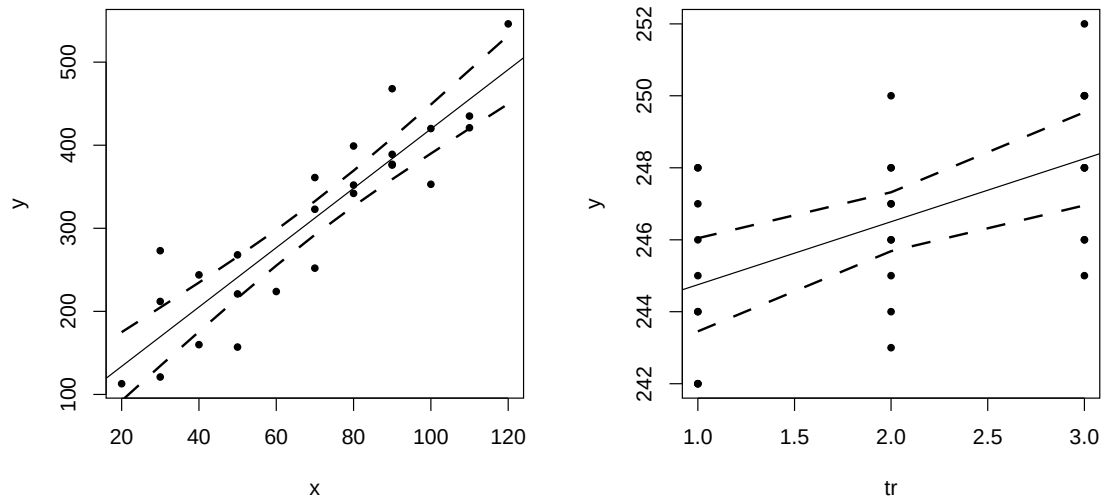


Figura 3.4. `Scatter plot`, retta di regressione e banda di confidenza per `toluca` (a sinistra) e per `caffeina` (a destra).

si ottiene un output che comprende gli intervalli i confidenza per i coefficienti β_i (`clb`) e per i valori teorici (`clm`). Viene inoltre creato un dataset con le colonne `x` e `y` (sempre presenti), `pred(icted)` e `residual` con i nomi `y_hat` e `e_hat`, `lclm` e `uclm` (*lower/upper confidence limit* per la *media* della variabile risposta) con i nomi `y_lwr` e `y_upr`.

Esempio 3.12. La retta di regressione su `caffeina` mostrava un minore adattamento ai dati, espresso da un R^2 più basso, ma evidenziato anche da una banda di confidenza più ampia che in `toluca`, come mostra il grafico a destra nella figura 3.4.

Osservazione. Per comprendere meglio il motivo per cui la banda di confidenza si allarga man mano che ci si allontana dal valore medio della variabile esplicativa, può essere utile determinare la varianza dei valori teorici sulla base di una diversa versione del modello, equivalente alla (3.1). Usando lo scarto $X_i - \bar{X}$ invece che X_i come variabile esplicativa, il modello diventa:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i - \bar{X}) + \beta_1\bar{X} + \varepsilon_i = (\beta_0 + \beta_1\bar{X}) + \beta_1(X_i - \bar{X}) + \varepsilon_i \\ &= \beta_0^* + \beta_1(X_i - \bar{X}) + \varepsilon_i \qquad\qquad \beta_0^* = \beta_0 + \beta_1\bar{X} \end{aligned}$$

Passando alle stime, e ricordando che $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$, si ha:

$$\hat{\beta}_0^* = \hat{\beta}_0 + \hat{\beta}_1\bar{X} = \bar{Y} \qquad\qquad \hat{Y}_i = \bar{Y} + \hat{\beta}_1(X_i - \bar{X})$$

Si rileva in primo luogo che, operando con i valori osservati, se $x_i = \bar{x}$ allora:

$$\hat{y}_i = \bar{y} + \hat{\beta}_1(x_i - \bar{x}) = \bar{y} + \hat{\beta}_1(\bar{x} - \bar{x}) = \bar{y}$$

ovvero: *la retta di regressione passa sempre per il punto* $(\bar{x}, \bar{y})$. Inoltre, tenendo presente

la nota 4, il calcolo della varianza di $\hat{Y}_h$ diventa:

$$\begin{aligned}\mathbb{V}[\hat{Y}_h] &= \mathbb{V}[\bar{Y} + \hat{\beta}_1(X_h - \bar{X})] = \mathbb{V}[\bar{Y}] + (X_h - \bar{X})\mathbb{V}[\hat{\beta}_1] \\ &= \frac{\sigma^2}{n} + (X_h - \bar{X})a_{22}\sigma^2 = \frac{\sigma^2}{n} + (X_h - \bar{X})\frac{\sigma^2}{\sum(X_i - \bar{X})^2} \\ &= \sigma^2\left(\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right)\end{aligned}$$

Si vede che la varianza dei valori teorici tende a diminuire all'aumentare di n (aumentano entrambi i denominatori), ma, per un dato n, aumenta quando X_h si allontana dalla media (e, ovviamente, tanto più quanto maggiore è σ^2).

3.2 Regressione lineare multipla

Nella *regressione lineare multipla* vi sono due o più variabili esplicative; i valori osservati vengono quindi proiettati non più su una retta, ma su un (iper)piano di dimensione pari al numero delle variabili esplicative. Si adotta un modello del tipo:

$$Y_i = \beta_0 + \sum_{j=1}^{p-1} \beta_j \mathbf{X}_{ij} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \tag{3.2}$$

dove p è il numero delle colonne della matrice di riparametrizzazione $\mathbf{X}$. In forma matriciale, infatti, il modello è:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1(p-1)} \\ \vdots & X_{21} & X_{22} & \dots & X_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{n(p-1)} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Inoltre, come visto nel capitolo 1:

- i coefficienti di regressione si stimano con $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$;
- i valori teorici con $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{H}\mathbf{Y}$;
- i residui con $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$;
- $SSTOT = \mathbf{Y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$, con $n-1$ gradi di libertà;
- $SSMOD = \mathbf{Y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{Y}$, con $p-1$ gradi di libertà (tanti quante le variabili esplicative);
- $SSRES = \mathbf{Y}'(\mathbf{I} - \mathbf{H})\mathbf{Y}$, con $n-p$ gradi di libertà.

Infine, generalizzando quanto visto nella sezione precedente:

- i test di ipotesi e gli intervalli di confidenza per i coefficienti si basano sulle loro varianze, che sono gli elementi della diagonale principale di $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$;
- la regione di confidenza, ovvero l'insieme degli intervalli di confidenza dei valori teorici, si basa sulla varianza di $\hat{Y}_h$, stimata da $MSERR[\mathbf{x}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_h]$.

Esempio 3.13. La Dwaine Studios Inc., che esegue ritratti di giovani in $n = 21$ città, è interessata alla relazione tra le vendite (`sales`) da un lato, la popolazione sotto i 16 anni (`targetpop`) e il reddito disponibile pro capite (`dispoinc`) dall'altro, al fine di decidere in quali altre città espandere la propria attività. I dati sono contenuti nel dataset `dwaine`.[5] Eseguendo in SAS:

```
proc reg data=dwaine;
 model sales = targtpop dispoinc / clb clm;
run;
```

si ottiene un output che conferma sia la relazione delle vendite con la popolazione giovanile e il reddito disponibile (*p-value* molto basso):

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	24015	12008	99.10	<.0001
Error	18	2180.92741	121.16263		
Corrected Total	20	26196			

sia un buon adattamento ai dati ($R^2 > 0.91$):

Root MSE	11.00739	R-Square	0.9167
Dependent Mean	181.90476	Adj R-Sq	0.9075
Coeff Var	6.05118		

Il test di ipotesi e gli intervalli di confidenza per i coefficienti mostrano che sia β_1 che β_2 sono significativamente diversi da 0 e che, con un livello di confidenza del 95%, cadono entrambi in intervalli con estremi positivi, quindi si può pensare che le vendite siano funzione crescente sia della popolazione giovanile che del reddito disponibile:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-68.85707	60.01695	-1.15	0.2663	-194.94801	57.23387
targtpop	1	1.45456	0.21178	6.87	<.0001	1.00962	1.89950
dispoinc	1	9.36550	4.06396	2.30	0.0333	0.82744	17.90356

Seguono i valori teorici e i relativi intervalli di confidenza:

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	174.4000	187.1841	3.8409	179.1146	195.2536	-12.7841
2	164.4000	154.2294	3.5558	146.7591	161.6998	10.1706
..	...	...	...	...	...	...
21	166.5000	157.0644	4.0792	148.4944	165.6344	9.4356

Il coefficiente R^2 migliora quando si aggiungono variabili esplicative. Ciò avviene perché la devianza spiegata (somma dei quadrati degli scarti tra i valori teorici $\hat{y}_i$ e la media $\bar{y}$) aumenta con l'aumentare del numero delle variabili esplicative. Ad esempio, nel caso di `dwaine`, R^2 vale 0.70 se l'unica variabile esplicativa è `dispoinc`, 0.89 se è `targtpop`, ma sale a oltre 0.91 se il modello le comprende entrambe.

[5] Tratto da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 237 (file CH06FI05.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/dwainestudios.csv`).

L'aumento del numero delle variabili esplicative migliora R^2, ma rende tendenzialmente meno agevole l'interpretazione del modello. Si usa quindi un R^2 *corretto*, spesso indicato con $\bar{R}^2$:

$$\bar{R}^2 = 1 - \frac{\frac{SSRES}{n-p}}{\frac{SSTOT}{n-1}} = 1 - \left(\frac{n-1}{n-p}\right)\frac{SSRES}{SSTOT}$$

Si può notare che aumentando il numero delle variabili esplicative aumenta p, quindi $n-p$ diminuisce e aumenta la quantità sottratta da 1, quindi $\bar{R}^2$ diminuisce. Nel caso di `dwaine`:

$$\bar{R}^2 = 1 - \left(\frac{20}{18}\right)\frac{2180.93}{26196} = 0.9075$$

3.2.1 Devianze di tipo I, II e III

Quando vi sono due o più variabili esplicative (quindi tre o più coefficienti di regressione), si usa scomporre la devianza spiegata in modo da tener conto del contributo che ciascuna variabile esplicativa apporta alla riduzione della devianza residua quando aggiunta al modello.

Se vi sono tre variabili esplicative, le devianze spiagate dai modelli che ne comprendono una sola, una coppia o tutte e tre si indicano con:

$$SSMOD(x_i) \qquad SSMOD(x_i, x_j) \qquad SSMOD(x_i, x_j, x_k)$$

Si può anche definire una devianza spiegata "marginale" $SSMOD(x_i \mid \dots)$, la devianza spiegata da una variabile quando viene aggiunta ad un modello, come differenza tra le devianze residue senza e con quella variabile:

$$\begin{aligned} SSMOD(x_i \mid x_j) &= SSRES(x_j) - SSRES(x_i, x_j) \\ SSMOD(x_i \mid x_j, x_k) &= SSRES(x_j, x_k) - SSRES(x_i, x_j, x_k) \end{aligned}$$

oppure anche, in modo equivalente, come l'incremento della devianza spiegata:

$$\begin{aligned} SSMOD(x_i \mid x_j) &= SSMOD(x_i, x_j) - SSMOD(x_j) \\ SSMOD(x_i \mid x_j, x_k) &= SSMOD(x_i, x_j, x_k) - SSMOD(x_j, x_k) \end{aligned}$$

Si può perfino definire una devianza spiegata da più variabili quando vengono aggiunte al modello:

$$\begin{aligned} SSMOD(x_j, x_k \mid x_i) &= SSMOD(x_i, x_j, x_k) - SSMOD(x_i) \\ &= [SSMOD(x_i, x_j, x_k) - SSMOD(x_i, x_j)] + [SSMOD(x_i, x_j) - SSMOD(x_i)] \\ &= SSMOD(x_j \mid x_i) + SSMOD(x_k \mid x_i, x_j) \end{aligned}$$

Esempio 3.14. La misura del grasso corporeo è onerosa, in quanto richiede l'immersione di una persona nell'acqua. Si cerca quindi di sostituirla con la più semplice rilevazione di tre fattori antropometrici: la spessore della plica tricipitale (`tst`, *triceps skinfold thickness*), la circonferenza della coscia (`tc`, *thigh circumference*) e la circonferenza del braccio

(`mac`, *midarm circumference*). Si usano le osservazioni su 20 donne, contenute nella matrice di dati `bodyfat`,[6] per verificare l'affidabilità di una stima del grasso corporeo basata sui tre parametri. Eseguendo la `proc reg` di SAS con diversi modelli si ottengono le seguenti analisi della varianza:

a) `model y = tst`:

Model	1	352.26980	352.26980	44.30	<.0001
Error	18	143.11970	7.95109		
Corrected Total	19	495.38950			

quindi $SSMOD(x_1) = 352.27$, $SSRES(x_1) = 143.12$;

b) `model y = tc`:

Model	1	381.96582	381.96582	60.62	<.0001
Error	18	113.42368	6.30132		
Corrected Total	19	495.38950			

quindi $SSMOD(x_2) = 381.97$, $SSRES(x_2) = 113.42$;

c) `model y = mac`:

Model	1	10.05160	10.05160	0.37	0.5491
Error	18	485.33790	26.96322		
Corrected Total	19	495.38950			

quindi $SSMOD(x_3) = 10.05$, $SSRES(x_3) = 485.34$;

d) `model y = tst tc`:

Model	2	385.43871	192.71935	29.80	<.0001
Error	17	109.95079	6.46769		
Corrected Total	19	495.38950			

quindi $SSMOD(x_1, x_2) = 385.44$, $SSRES(x_1, x_2) = 109.95$;

e) `model y = tst mac`:

Model	2	389.45533	194.72767	31.25	<.0001
Error	17	105.93417	6.23142		
Corrected Total	19	495.38950			

quindi $SSMOD(x_1, x_3) = 389.46$, $SSRES(x_1, x_3) = 105.93$;

f) `model y = tc mac`:

Model	2	384.27972	192.13986	29.40	<.0001
Error	17	111.10978	6.53587		
Corrected Total	19	495.38950			

quindi $SSMOD(x_2, x_3) = 384.28$, $SSRES(x_2, x_3) = 111.11$;

g) `model y = tst tc mac`:

Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

quindi $SSMOD(x_1, x_2, x_3) = 396.98$, $SSRES(x_1, x_2, x_3) = 98.41$.

Si possono verificare agevolmente le relazioni definite sopra; ad esempio:

$$\begin{aligned} SSMOD(x_2|x_1) &= SSRES(x_1) - SSRES(x_1, x_2) = 143.12 - 109.95 = 33.17 \\ &= SSMOD(x_1, x_2) - SSMOD(x_1) = 385.44 - 352.27 = 33.17 \end{aligned}$$

[6] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 257 (file CH07TA01.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/bodyfat.csv`).

oppure:

$$SSMOD(x_3 \mid x_1, x_2) = SSRES(x_1, x_2) - SSRES(x_1, x_2, x_3) = 109.95 - 98.41 = 11.54$$
$$= SSMOD(x_1, x_2, x_3) - SSMOD(x_1, x_2) = 396.98 - 385.44 = 11.54$$

e così via.

Come si vede, si possono calcolare molte devianze "marginali". Si sono comunque affermati tre approcci principali, che si basano sulla distinzione, introdotta dal SAS, tra devianza di tipo I, di tipo II e di tipo III.

Tipo I

Nella "devianza di tipo I" (*Type I SS* nel gergo di SAS), si calcolano i contributi alla devianza spiegata forniti da ciascuna variabile esplicativa man mano che viene aggiunta, seguendo l'ordine in cui compaiono nella definizione del modello. Con tre variabili, quindi, si calcolano nell'ordine:

- $SSMOD(X_1)$;
- $SSMOD(X_2 \mid X_1) = SSMOD(X_1, X_2) - SSMOD(X_1)$;
- $SSMOD(X_3 \mid X_1, X_2) = SSMOD(X_1, X_2, X_3) - SSMOD(X_1, X_2)$.

Esempio 3.15. Per avere le devianze di tipo I si deve eseguire la funzione `anova()` con R, `proc glm` con SAS. Ad esempio, usando con SAS i dati `bodyfat`, da:

```
proc glm data=bodyfat;
 model y = tst tc mac;
run;
```

si ottiene, dopo la tabella ANOVA:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tst	1	352.2697968	352.2697968	57.28	<.0001
tc	1	33.1689128	33.1689128	5.39	0.0337
mac	1	11.5459022	11.5459022	1.88	0.1896

Si può notare che per ogni variabile esplicativa la devianza ha un solo grado di libertà, quindi coincide con la varianza. Sulla base delle elaborazioni riprodotte nell'esempio 3.14:

- $SSMOD(x_1) = 352.27$;
- $SSMOD(x_2 \mid x_1) = SSMOD(x_2, x_1) - SSMOD(x_1) = 385.44 - 352.27 = 33.17$;
- $SSMOD(x_3 \mid x_1, x_2) = SSMOD(x_1, x_2, x_3) - SSMOD(x_1, x_2) = 396.98 - 385.44 = 11.54$.

Tipo II

Nella "devianza di tipo II" (*Type II SS*), si calcolano i contributi alla devianza spiegata forniti da ciascuna variabile esplicativa rispetto a quella che si ottiene considerando solo tutte le altre variabili.

Va precisato che anche nella regressione, come nei modelli ANOVA, possono includersi nel modello *effetti interattivi* (v. sez. 3.2.5). In questo caso, la devianza di tipo II con due variabili esplicative sarebbe:

– $SSMOD(x_1 \mid x_2)$;
– $SSMOD(x_2 \mid x_1)$;

e non verrebbero considerati i casi:

– $SSMOD(x_1 \mid x_2, x_1x_2)$;
– $SSMOD(x_2 \mid x_1, x_1x_2)$.

Ciò ha ovviamente senso solo se l'effetto interattivo risulta non significativo.

Sebbene alcuni preferiscano la devianza di tipo II a quella di tipo III, in generale con quest'ultima si ottengono gli stessi risultati esaminando prima un modello con effetti interattivi, poi nuovi modelli che li escludano se risultano non significativi.

Se il modello non considera effetti interattivi, la devianza di tipo II coincide con quella di tipo III.

Esempio 3.16. Restando a `bodyfat`, si può ottenere la devianza di tipo II in SAS aggiungendo l'opzione `ss2` dopo il `model`; ad esempio:

```
proc glm data=bodyfat;
 model y = tst tc mac / ss2;
run;
```

Con R si deve caricare la libreria `car` ed eseguire la funzione `Anova()`, con la "A" maiuscola, che calcola per default la devianza di tipo II:[7]

```
> mod <- lm(y ~ tst + tc + mac, data=bodyfat)
> library(car)
> Anova(mod)
Anova Table (Type II tests)

Response: y
          Sum Sq Df F value Pr(>F)
tst       12.705  1  2.0657 0.1699
tc         7.529  1  1.2242 0.2849
mac       11.546  1  1.8773 0.1896
Residuals 98.405 16
```

Tipo III

Nella "devianza di tipo III" (*Type III SS*) si calcola per ciascuna variabile esplicativa il contributo alla devianza spiegata dato da essa quando viene aggiunta alle altre nel modello:

– $SSMOD(x_1 \mid x_2, x_3)$;
– $SSMOD(x_2 \mid x_1, x_3)$;
– $SSMOD(x_3 \mid x_1, x_2)$.

Esempio 3.17. Con R, dopo aver caricato la libreria `car`, si usa `Anova()` con l'opzione `type="III"`. Con SAS basta usare `proc glm` e si ottiene, dopo la tabella ANOVA e la devianza di tipo I:

[7] Si può trovare in `cran.r-project.org/doc/contrib/Ricci-refcard-regression.pdf` un'utile raccolta di funzioni R per la regressione, con l'indicazione delle librerie in cui si trovano.

model	*SSMOD*	Type I SS		Type III SS		*SSRES*	*SSTOT*
y = a	2.50	a:	2.50	a:	2.50	1177.40	1179.90
y = b	230.40	b:	230.40	b:	230.40	949.50	1179.90
y = a b	232.90	a:	2.50	a:	2.50	947.00	1179.90
		b:	230.40	b:	230.40		
y = a b a*b	242.90	a:	2.50	a:	2.50	937.00	1179.90
		b:	230.40	b:	230.40		
		a*b:	10.00	a*b:	10.00		

Tabella 3.1. Devianze di tipo I e III nel caso di `dietepec`.

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tst	1	12.70489278	12.70489278	2.07	0.1699
tc	1	7.52927788	7.52927788	1.22	0.2849
mac	1	11.54590217	11.54590217	1.88	0.1896

(si può notare che, non comprendendo il modello alcun effetto interattivo, i risultati sono uguali a quelli ottenuti, con R ma anche con SAS, per la devianza di tipo II). Sulla base delle elaborazioni riprodotte nell'esempio 3.14:

- $SSMOD(x_1 \mid x_2, x_3) = SSMOD(x_1, x_2, x_3) - SSMOD(x_2, x_3) = 396.98 - 384.28 = 12.70$;
- $SSMOD(x_2 \mid x_1, x_3) = SSMOD(x_1, x_2, x_3) - SSMOD(x_1, x_3) = 396.98 - 389.46 = 7.52$;
- $SSMOD(x_3 \mid x_1, x_2) = SSMOD(x_1, x_2, x_3) - SSMOD(x_1, x_2) = 396.98 - 385.44 = 11.54$.

Osservazioni

Nel capitolo 2 si era riprodotto qualche output SAS con la devianza di tipo I. Se si fosse badato anche a quella di tipo III, questa sarebbe risultata uguale (esclusi, ovviamente, gli esperimenti non bilanciati trattati nella sez. 2.5). Nel caso dell'esperimento `dietepec`, eseguengo `proc glm` con `class t a b`,[8] variando il `model` si ottengono i risultati sintetizzati nella tabella 3.1. Come si vede, le devianze di tipo I e di tipo III sono uguali. Questo vuol dire, ad esempio, che `a` (il rame) dà lo stesso contributo alla devianza spiegata sia quando viene considerato come primo e unico fattore, sia quando viene aggiunto dopo `b` (il cobalto) oppure dopo `b` e `a*b`. Si tratta di capire perché, invece, le devianze di tipo I e di tipo III sono molto diverse nel caso di `bodyfat`.

Nel caso di `dietepec` ci sono due fattori, che diventano tre considerando l'effetto interattivo. Tre variabili esplicative in `bodyfat`.

[8] Quando si usa `class` per precisare che alcuni fattori sono di tipo qualitativo, l'ordine in cui essi vengono scritti rimane fissato anche se in `model` o in `estimate` vengono scritti in ordine diverso; ad esempio, se si scrive `class t a b`, anche scrivendo `b*a` in una riga `estimate` SAS intende che il primo fattore è `a` e il secondo è `b`, cioè legge `b*a` come se fosse `a*b`.

In `dietepec` la devianza di tipo I spiegata da `a` è (cfr. esempio 1.11):

$$\mathbf{y}'\left(\mathbf{H}_a - \frac{1}{n}\mathbf{J}\right)\mathbf{y} = \bar{\mathbf{y}}'\mathbf{H}_a\bar{\mathbf{y}}$$

dove $\mathbf{H}_a = \mathbf{A}_a(\mathbf{A}_a'\mathbf{A}_a)^{-1}\mathbf{A}_a'$ e $\mathbf{A}_a$ è una matrice avente solo le prime due colonne della matrice di riparametrizzazione illustrata nella figura 2.6, come se `a` fosse l'unico fattore del modello. La devianza di tipo III è invece:

$$\begin{aligned}\mathbf{y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right)\mathbf{y} - \mathbf{y}'\left(\mathbf{H}_{b,a*b} - \frac{1}{n}\mathbf{J}\right)\mathbf{y} &= \bar{\mathbf{y}}'\mathbf{H}\bar{\mathbf{y}} - \bar{\mathbf{y}}'\mathbf{H}_{b,a*b}\bar{\mathbf{y}} \\ &= \bar{\mathbf{y}}'\left(\mathbf{H} - \mathbf{H}_{b,a*b}\right)\bar{\mathbf{y}}\end{aligned}$$

dove $\mathbf{H}_{b,a*b}$ è ottenuta da $\mathbf{A}_{b,a*b}$, matrice con tutte le colonne della matrice di riparametrizzazione tranne la seconda.

Le due devianze sono uguali grazie alla scomposizione ortogonale della devianza spiegata vista nel capitolo 2, in particolare nell'osservazione a pag. 57:

$$\mathbf{H} = \mathbf{H}_a + \mathbf{H}_{b,a*b} \quad \Rightarrow \quad \mathbf{H}_a = \mathbf{H} - \mathbf{H}_{b,a*b}$$

Come si era visto, la scomposizione ortogonale è possibile perché le colonne della matrice di riparametrizzazione (esclusa la prima, composta di tutti 1) sono tra loro ortogonali, quindi diverse matrici $\mathbf{H}_i$ proiettano $\bar{\mathbf{y}}$ su sottospazi tra loro ortogonali.

Questo non accade nel caso di `bodyfat`, la cui matrice di riparametrizzazione è:

$$\mathbf{X} = \begin{bmatrix} 1 & 19.5 & 43.1 & 29.1 \\ 1 & 24.7 & 49.8 & 28.2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 22.7 & 48.2 & 27.1 \\ 1 & 25.2 & 51.0 & 27.5 \end{bmatrix}$$

Si può osservare, in compenso, che le devianze di tipo III sono sempre ortogonali alla devianza residua e che questo consente i eseguire i test di ipotesi F illustrati nella sezione 3.2.3.

Considerando infatti:

$$SSMOD(x_3 \mid x_1, x_2) = SSMOD(x_1, x_2, x_3) - SSMOD(x_1, x_2)$$

si ha che:

$$\begin{aligned}SSMOD(x_3 \mid x_1, x_2) &= \mathbf{y}'\left(\mathbf{H} - \frac{1}{n}\mathbf{J}\right) - \mathbf{y}'\left(\mathbf{H}_{x_1x_2} - \frac{1}{n}\mathbf{J}\right) = \mathbf{y}'(\mathbf{H} - \mathbf{H}_{x_1x_2})\mathbf{y} \\ SSRES &= \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

e inoltre:

$$(\mathbf{H} - \mathbf{H}_{x_1x_2})(\mathbf{I} - \mathbf{H}) = \mathbf{H} - \mathbf{H} - \mathbf{H}_{x_1x_2} + \mathbf{H}_{x_1x_2}\mathbf{H} = \mathbf{O}$$

quindi $SSMOD(x_3 \mid x_1, x_2)$ e *SSRES sono indipendenti*, in quanto $\mathbf{H}_{x_1x_2}\mathbf{H} = \mathbf{H}_{x_1x_2}$.

Tale ultima uguaglianza si spiega perché $\mathbf{H}$ proietta $\mathbf{y}$, elemento di uno spazio di dimensione n, su un sottospazio di dimensione 3 (le tre variabili esplicative), mentre $\mathbf{H}_{x_1,x_2}$ lo proietta su un sottospazio del precedente di dimensione 2. Ne segue:

$$\mathbf{H}_{x_1x_2}\mathbf{y} = \mathbf{H}_{x_1x_2}(\mathbf{H}\mathbf{y}) = (\mathbf{H}_{x_1x_2}\mathbf{H})\mathbf{y} \quad \Rightarrow \quad \mathbf{H}_{x_1x_2}\mathbf{H} = \mathbf{H}_{x_1x_2}$$

In altri termini, dato un vettore $\mathbf{y} \in \mathbb{R}^n$, la sua proiezione su un sottospazio di dimensione $p-2$ è uguale alla proiezione su di esso di una sua precedente proiezione su un sottospazio di dimensione $p-1$ che includa quello di dimensione $p-2$.

Un semplice esempio geometrico può aiutare a comprendere meglio l'uguaglianza. Siano:

$$\mathbf{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \in \mathbb{R}^3 \qquad \mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

La matrice $\mathbf{P}$ proietta $\mathbf{y}$ sul piano xy, $\mathbf{Q}$ lo proietta sull'asse x. La matrice $\mathbf{Q}$ dà lo stesso risultato sia se premoltiplicata per $\mathbf{y}$:

$$\mathbf{Qy} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

sia se premoltiplicata per la sua proiezione sul piano:

$$\mathbf{Py} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} \qquad \mathbf{Qy} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

quindi $\mathbf{Qy} = \mathbf{QPY}$, ovvero $\mathbf{Q} = \mathbf{QP}$:

$$\mathbf{QP} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \mathbf{Q}$$

3.2.2 I coefficienti di determinazione parziali

Un *coefficiente di determinazione parziale* misura il contributo a R^2, quindi alla spiegazione della variabilità di Y, fornito da ciascuna variabile esplicativa dopo che le altre sono state già comprese nel modello.

I coefficienti di determinazione parziale sono quindi calcolati sulla base delle devianze di tipo III (o anche di tipo II se non vi sono effetti interattivi); se vi sono tre variabili esplicative:

$$R_i^2 = \frac{SSMOD(x_i \mid x_j, x_k)}{SSRES(x_j, x_k)}$$

Esempio 3.18. Restando a `bodyfat`:

$$R^2_{tst} = \frac{SSMOD(x_1, x_2, x_3) - SSMOD(x_2, x_3)}{SSRES(x_2, x_3)} = \frac{396.9846 - 384.2797}{111.1098} = 0.1143$$

$$R^2_{tc} = \frac{SSMOD(x_1, x_2, x_3) - SSMOD(x_1, x_3)}{SSRES(x_1, x_3)} = \frac{396.9846 - 389.4553}{105.9342} = 0.0711$$

$$R^2_{mac} = \frac{SSMOD(x_1, x_2, x_3) - SSMOD(x_1, x_2)}{SSRES(x_1, x_2)} = \frac{396.9846 - 385.4387}{109.9508} = 0.1050$$

In SAS i coefficienti di determinazione parziale vengono indicati come quadrati dei coefficienti di correlazione parziale e si ottengono aggiungendo l'opzione `pcorr2` al `model`; si ottiene così:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Squared Partial Corr Type II
Intercept	1	117.08469	99.78240	1.17	0.2578	.
tst	1	4.33409	3.01551	1.44	0.1699	0.11435
tc	1	-2.85685	2.58202	-1.11	0.2849	0.07108
mac	1	-2.18606	1.59550	-1.37	0.1896	0.10501

3.2.3 I test di ipotesi sui coefficienti di regressione

In gerale, i test di ipotesi sui singoli coefficienti vengono effettuati in modo analogo a quanto già visto per la regressione semplice, usando statistiche test del tipo:

$$t^* = \frac{\hat{\beta}_i}{\sqrt{a_{i+1,i+1} MSRES}} \sim t_{n-p}$$

dove $a_{i+1,i+1}$ è l'elemento $i+1$ della diagonale principale di $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$.

Esempio 3.19. Restando a `bodyfat`, con R si possono preparare i test calcolando i coefficienti, la matrice $\mathbf{A}$ e $MSRES$:

```
> bodyfat <- read.csv("bodyfat.csv")
> attach(bodyfat)
> n <- nrow(bodyfat)
> mod <- lm(y ~ tst + tc + mac)
> X <- model.matrix(mod)
> XX <- t(X) %*% X
> A <- solve(XX)
> beta <- A %*% t(X) %*% y
> p <- length(beta)
> I <- diag(1, nrow=n)
> H <- X %*% A %*% t(X)
> SSRES <- t(y) %*% (I - H)  %*% y
> MSRES <- SSRES / (n-p)
> beta
                  [,1]
(Intercept) 117.084695
tst           4.334092
tc           -2.856848
mac          -2.186060
> MSRES
         [,1]
[1,] 6.150306
```

Si possono poi calcolare insieme tutte le statistiche test ed i relativi *p-value*:

```
> tstar <- beta / sqrt(diag(A) * MSRES)
> tstar
                 [,1]
```

```
(Intercept)  1.173400
tst          1.437266
tc          -1.106441
mac         -1.370142
> p.value <- pt(abs(tstar), n-p, lower.tail=FALSE) +
+            pt(-abs(tstar), n-p)
> p.value
                 [,1]
(Intercept) 0.2578078
tst         0.1699111
tc          0.2848944
mac         0.1895628
```

Con SAS si possono usare sia `proc reg` che `proc glm`; in entrambi i casi si ottiene:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	117.08469	99.78240	1.17	0.2578
tst	1	4.33409	3.01551	1.44	0.1699
tc	1	-2.85685	2.58202	-1.11	0.2849
mac	1	-2.18606	1.59550	-1.37	0.1896

I test T così condotti sono equivalenti ai test F basati sulle devianze di tipo III; ad esempio, in un modello con $p = 4$ coefficienti come `bodyfat`, per β_3:

$$F^* = \frac{\frac{SSMOD(x_3 \mid x_1, x_2)}{1}}{\frac{SSRES}{n-p}} \sim F_{1,n-p}$$

Si ha, infatti, $(t_\nu)^2 = F_{1,\nu}$.

Esempio 3.20. Usando R come calcolatrice:

```
> library(car)
> SSMODx <- Anova(mod, type="III")[2:4,1]
> SSMODx
[1] 12.704893  7.529278 11.545902
> Fstar <- SSMODx / MSRES
> Fstar
[1] 2.065734 1.224212 1.877289
> p.value <- pf(Fstar, 1, n-p, lower.tail=FALSE)
> p.value
[1] 0.1699111 0.2848944 0.1895628
```

Si può notare che i valori di `Fstar` sono uguali ai quadrati dei corrispondenti valori di `tstar` (non si calcola la devianza di tipo III per β_0), così come sono uguali i *p-value*, e che si ottengono gli stessi risultati forniti da SAS e già visti nell'esempio 3.17.

Vi sono invece situazioni nelle quali è possibile solo il test F. Per sottoporre a verifica l'ipotesi:

$$H_0 : \beta_2 = \beta_3 = 0$$

si può ricorrere a:

$$F^* = \frac{\frac{SSMOD(x_2, x_3 \mid x_1)}{2}}{\frac{SSRES}{n-p}} \sim F_{2,n-p}$$

ma non ad un test t.

Esempio 3.21. Dall'esempio 3.14:

$$SSMOD(x_2, x_3 \mid x_1) = SSMOD(x_1, x_2, x_3) - SSMOD(x_1) = 396.98461 - 352.26980$$

Calcolando con R:

```
> Fstar <- ( (396.98461-352.26980) / 2) / MSRES
> Fstar
        [,1]
[1,] 3.63517
> pf(Fstar, 2, n-p, lower.tail=FALSE)
          [,1]
[1,] 0.0499503
```

Osservazione. Dagli esempi precedenti, da una parte sembrerebbero non significativi tutti i coefficienti, dall'altra il test sulla coppia delle ultime due variabili esplicative non sembra consentire decisioni nette. Si deve anche notare che, se si escludesse la variabile `mac` (circonferenza del braccio) dai dati `bodyfat` e si eseguisse in SAS:

```
proc reg data=bodyfat;
 model y = tst tc;
run;
```

si otterrebbe un netto miglioramento del *p-value* per il coefficiente di `tc` (circonferenza coscia), che passerebbe da 0.2849 a 0.0369:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-19.17425	8.36064	-2.29	0.0348
tst	1	0.22235	0.30344	0.73	0.4737
tc	1	0.65942	0.29119	2.26	0.0369

Analogamente, se si escludesse `tc` si otterrebbe un netto miglioramento dei *p-value* per le altre due variabili:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.79163	4.48829	1.51	0.1486
tst	1	1.00058	0.12823	7.80	<.0001
mac	1	-0.43144	0.17662	-2.44	0.0258

Va notato che *cambierebbero anche le stime dei coefficienti*. Ciò accade per motivi che diverranno più chiari dopo aver discusso le conseguenze della non-ortogonalità delle colonne della matrice di riparametrizzazione (sez. 3.2.4).

3.2.4 La multicollinearità

In molti studi osservazionali, come `toluca` o `bodyfat`, le variabili esplicative tendono ad essere correlate tra loro e ciò crea problemi che verranno illustrati mediante tre esempi: uno di mancanza di correlazione, uno di perfetta correlazione ed uno intermedio.

Esempio 3.22. Si misura la produttività `y` di una squadra di lavoratori al variare del numero di lavoratori `x1`, 4 o 6, e dei premi aggiunti al salario `x2`, 2 o 3 dollari. Le osservazioni sono contenute nella matrice di dati `workcrew`.[9] Si verifica facilmente che le due variabili esplicative hanno correlazione nulla; ciò comporta che le rispettive colonne della matrice di riparametrizzazione sono ortogonali e che quindi, come visto nelle Osservazioni a pag. 99, le devianze di tipo I e di tipo III sono uguali. Infatti, eseguendo:

```
proc glm data=workcrew;
 model y = x1 x2;
run;
```

si ottiene:

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x1	1	231.1250000	231.1250000	65.57	0.0005
x2	1	171.1250000	171.1250000	48.55	0.0009

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x1	1	231.1250000	231.1250000	65.57	0.0005
x2	1	171.1250000	171.1250000	48.55	0.0009

Inoltre, le stime dei coefficienti delle variabili esplicative, β_1 e β_2, rimangono le stesse sia quando esse compaiono da sole nel modello, sia quando compaiono entrambe. Eseguendo `proc glm` prima con solo `x1`, poi con solo `x2`, infine con entrambe, si ottiene:

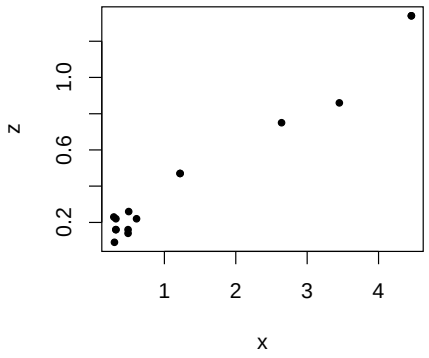
Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	23.50000000	10.11135912	2.32	0.0591
x1	5.37500000	1.98300067	2.71	0.0351
Intercept	27.25000000	11.60773808	2.35	0.0572
x2	9.25000000	4.55292946	2.03	0.0885
Intercept	0.375000000	4.74045093	0.08	0.9400
x1	5.375000000	0.66379590	8.10	0.0005
x2	9.250000000	1.32759180	6.97	0.0009

Ciò significa che gli effetti di ciascuna delle due variabili non cambiano se è presente o meno anche l'altra.

Esempio 3.23. La matrice di dati `perfectcorr`[10] è costruita in modo da avere perfetta correlazione tra tutte le variabili, al punto che si possono avere infinite funzioni che

[9] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 279 (file CH07TA06.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/workcrew.csv`).

[10] Adattata da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 281 e scaricabile da `http://web.mclink.it/MC1166/ModelliStatistici/perfectcorr.csv`.



	x	z
x	1.000	0.988
z	0.988	1.000

Tabella 3.2. Scatter plot e matrice di correlazione per le variabili esplicative della matrice di dati `inquina`.

consentono di ottenere i valori di `y` da quelli di `x1` e `x2`, ad esempio:

$$y = -87 + x_1 + 18x_2$$
$$y = -7 + 9x_1 + 2x_2$$

Ne segue che nessuna stima dei coefficienti è possibile.

L'esempio 3.22 rappresenta la soluzione cui si tende normalmente negli studi sperimentali, l'esempio 3.23 prospetta in modo estremo il rischio che si corre negli studi osservazionali, quando le variabili esplicative non sono sotto il controllo del ricercatore. Il prossimo esempio mostra gli effetti della multicollinearità come possono manifestarsi in concreto.

Esempio 3.24. Si sono registrati nel dataset `inquina`[11] il numero di decessi verificatosi nella contea di Londra dal 1° al 15 dicembre 1952, `y`, e due indicatori di inquinamento atmosferico: lo smog in mg/mc, `x`, e il diossido di solfuro in numero di particelle su un milione, `z`. Le due variabili esplicative sono fortemente correlate (tabella 3.2). Ne segue che eseguendo la regressione prima con la sola `x`, poi con la sola `y`, poi con entrambe, si ottengono coefficienti diversi. Operando con SAS:

```
                                Parameter       Standard
Variable    Label          DF    Estimate          Error   t Value   Pr > |t|

Intercept   Intercept       1   171.81881       31.43448      5.47     0.0001
x           Smog            1    63.76092       15.31226      4.16     0.0011

                                Parameter       Standard
Variable    Label          DF    Estimate          Error   t Value   Pr > |t|

Intercept   Intercept       1   144.11078       29.22749      4.93     0.0003
z           Diossido        1   256.23556       47.59353      5.38     0.0001

                                Parameter       Standard
Variable    Label          DF    Estimate          Error   t Value   Pr > |t|

Intercept   Intercept       1    89.51080       25.07824      3.57     0.0039
x           Smog            1  -220.32438       58.14315     -3.79     0.0026
z           Diossido        1  1051.81646      212.59597      4.95     0.0003
```

[11] Scaricabile da `http://web.mclink.it/MC1166/ModelliStatistici/inquina.csv`.

Si nota che la variazione arriva al punto che lo smog ha un coefficiente di regressione positivo (più smog → più decessi) quando considerato da solo, uno negativo (più smog → meno decessi) quando considerato insieme al diossido di solfuro. Aggiungendo l'opzione `clm` ai tre modelli si osserva che le deviazioni standard dei valori teorici diminuiscono leggermente, ma rimane l'impossibilità di stimare (e di sottoporre a test di ipotesi) i singoli coefficienti.

È ora opportuno tornare a `bodyfat` per mostrare un errore in cui si può facilmente incorrere nella valutazione dei test di ipotesi sui coefficienti di regressione.

Esempio 3.25. Si era visto nell'esempio 3.19 che i test t sui coefficienti di regressione per le tre variabili esplicative erano tali che si sarebbe potuto giudicare non significativo l'effetto di ciascuna delle tre. In realtà, ciò accadeva proprio perché, come mostrato in quell'esempio, i test t sui coefficienti di regressione sono equivalenti a test F basati sulla devianza di tipo III:

$$F^* = \frac{SSMOD(x_3 \mid x_1, x_2)}{MSRES} \sim F_{1,n-p}$$

sono cioè equivalenti a test in cui si misuri il contributo di ciascuna variabile esplicativa alla spiegazione della variabilità, *quando questa viene aggiunta a tutte le altre*. Infatti i *p-value* coincidono:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tst	1	12.70489278	12.70489278	2.07	0.1699
tc	1	7.52927788	7.52927788	1.22	0.2849
mac	1	11.54590217	11.54590217	1.88	0.1896

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	117.0846948	99.78240295	1.17	0.2578
tst	4.3340920	3.01551136	1.44	0.1699
tc	-2.8568479	2.58201527	-1.11	0.2849
mac	-2.1860603	1.59549900	-1.37	0.1896

Dovrebbe apparire evidente che, se le variabili esplicative sono tra loro correlate, una volta che una quota della variabilità complessiva sia stata spiegata da tutte meno una, l'ultima non può che aggiungere ben poco. Ne segue che, in presenza di correlazione, si può tenere conto solo dell'effetto di tutte le variabili esplicative; nel caso di `bodyfat` non se ne può escludere nessuna, sarebbe soprattutto sbagliato pensare di escluderne qualcuna sulla base dei test t, e ci si deve limitare a considerare che il modello che le comprende tutte e tre supera agevolmente il test di ipotesi F ($MSMOD = 132.33$, $MSRES = 6.15$, *p-value*< 0.0001) e presenta un buon adattamento ai dati ($R^2 = 0.80$).

La multicollinearità non è sempre di agevole rilevazione (una matrice di correlazione considera solo coppie di variabili) e, salvo il ricorso a soluzioni più sofisticate, ci si può avvalere di due accorgimenti:

a) centrare le variabili esplicative, sostituendole con gli scarti dalle rispettive medie, in modelli con effetti interattivi (sez. 3.2.5) o polinomiali (v. sez. 3.2.6);

b) limitarsi alla capacità predittiva del modello, cioè alla possibilità di calcolare valori teorici che siano funzione di nuovi valori delle variabili esplicative; ciò ha senso, peraltro,

solo se i nuovi valori rispettano lo schema di multicollinearità presente nella matrice dei dati (nel caso dell'esempio 3.23, ciò vuol dire che, essendo x_1 e x_2 legare dalla relazione $x_2 = 5 + 0.5x_1$, si può calcolare il valore teorico per $x_1 = 20$ e $x_2 = 5 + 0.5 \cdot 20 = 15$, non per $x_1 = 20$ e $x_2 = 30$).

3.2.5 Effetti interattivi

Quando in un modello compaiono più variabili esplicative, è possibile tenere conto di eventuali *effetti interattivi*, che vengono spesso espressi come prodotti; ad esempio, ad un modello del tipo $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$ si può aggiungere il termine $\beta_3 X_{i1} X_{i2}$, ottenendo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i \tag{3.3}$$

Si perviene così ad una regione di regressione curva e cambia l'interpretazione dei coefficienti. Dal momento che:

$$\frac{\partial Y_i}{\partial X_{i1}} = \beta_1 + \beta_3 X_{i2}$$

l'incremento di Y_i a seguito di un incremento unitario di X_{i1}, restando costante X_{i2}, non è più β_1, ma $\beta_1 + \beta_3 X_{i2}$.

Esempio 3.26. Nel caso di `bodyfat` il modello con effetti interattivi assume la forma:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3}$$

In SAS (ma anche con R) si deve prima creare un dataset che contenga le ulteriori colonne dei prodotti delle variabili esplicative. Si deve tenere presente che i prodotti rischiano di essere fortemente correlati sia tra di loro che con le singole variabili esplicative; è preferibile quindi controllare, eseguendo:

```
data bfint;
 set bodyfat;
 tst_tc = tst*tc;
 tst_mac = tst*mac;
 tc_mac = tc*mac;
run;
proc corr data=bfint noprob; run;
```

Esaminando la matrice di correlazione:

	tst	tc	mac	tst_tc	tst_mac	tc_mac
tst	1.00000	0.92384	0.45778	0.98878	0.90032	0.89071
tc	0.92384	1.00000	0.08467	0.96634	0.67197	0.65361
mac	0.45778	0.08467	1.00000	0.33239	0.78770	0.80641
tst_tc	0.98878	0.96634	0.33239	1.00000	0.83445	0.82186
tst_mac	0.90032	0.67197	0.78770	0.83445	1.00000	0.99836
tc_mac	0.89071	0.65361	0.80641	0.82186	0.99836	1.00000

si notano in effetti forti correlazioni, quali 0.989 tra X_1 e X_1X_2, 0.998 tra X_1X_3 e X_2X_3. Si procede quindi a centrare le variabili:

```
proc standard data=bodyfat mean=0 out=bfcen;
 var tst tc mac;
run;
data bfcenint;
 set bfcen;
 tst_tc = tst*tc;
 tst_mac = tst*mac;
 tc_mac = tc*mac;
run;
proc corr data=bfcenint noprob;
 var tst tc mac tst_tc tst_mac tc_mac;
run;
```

Si ottiene così una matrice di correlazione non ottimale, ma sicuramente migliore della precedente:

	tst	tc	mac	tst_tc	tst_mac	tc_mac
tst	1.00000	0.92384	0.45778	-0.47701	-0.17342	-0.22157
tc	0.92384	1.00000	0.08467	-0.42979	-0.17254	-0.14366
mac	0.45778	0.08467	1.00000	-0.21589	-0.03041	-0.23537
tst_tc	-0.47701	-0.42979	-0.21589	1.00000	0.23283	0.29191
tst_mac	-0.17342	-0.17254	-0.03041	0.23283	1.00000	0.89051
tc_mac	-0.22157	-0.14366	-0.23537	0.29191	0.89051	1.00000

Eseguendo la regressione con `proc glm` (per avere le devianze di tipo I e di tipo III):

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	407.6995001	67.9499167	10.07	0.0003
Error	13	87.6899999	6.7453846		
Corrected Total	19	495.3895000			

R-Square	Coeff Var	Root MSE	y Mean
0.822988	12.86055	2.597188	20.19500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tst	1	352.2697968	352.2697968	52.22	<.0001
tc	1	33.1689128	33.1689128	4.92	0.0450
mac	1	11.5459022	11.5459022	1.71	0.2134
tst_tc	1	1.4957180	1.4957180	0.22	0.6455
tst_mac	1	2.7043343	2.7043343	0.40	0.5376
tc_mac	1	6.5148360	6.5148360	0.97	0.3437

Si può sottoporre a verifica la significatività degli effetti interattivi, con ipotesi nulla:

$$H_0 : \beta_4 = \beta_5 = \beta_6 = 0$$

e con la statistica test:

$$F^* = \frac{\frac{SSMOD(x_1x_2, x_1x_3, x_2x_3 \mid x_1, x_2, x_3)}{3}}{MSRES} \sim F_{3,13}$$

Come visto a pag. 95, la devianza spiegata da più variabili quando aggiunte al modello è la somma delle loro devianze di tipo I, quindi:

$$F^* = \frac{\frac{1.496 + 2.794 + 6.515}{3}}{6.745} = 0.53 \qquad p\text{-}value = 0.67$$

In questo caso, quindi, si accetta l'ipotesi nulla.

3.2.6 La regressione polinomiale

Il modello (3.2) viene detto modello *del primo ordine* perché le variabili esplicative vi compaiono tutte con esponente 1; in altri termini, il modello è un polinomio di primo grado.

Si hanno anche modelli di ordine superiore. Ad esempio, un modello *del secondo ordine* con una sola variabile esplicativa è:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$$

Un modello del secondo ordine con due variabili esplicative può essere:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_{11} X_{i1}^2 + \beta_{22} X_{i2}^2 + \beta_{12} X_{i1} X_{i2} + \varepsilon_i$$

Va notato che, come già visto a proposito degli effetti interattivi, la presenza di potenze e di prodotti può comportare multicollinearità; conviene quindi centrare le variabili esplicative.

Esempio 3.27. Si rilevano l'età `x` e la massa muscolare `y` di 60 donne nella matrice di dati `musclemass`.[12] Volendo usare un modello del secondo ordine, si crea un nuovo dataset contenente una colonna `x2` con i quadrati delle età (variabile esplicativa) e si bada alla correlazione tra `x` e `x2`:

```
data mm2;
 set musclemass;
 x2 = x**2;
run;
proc corr data=mm2 noprob;
 var x x2;
run;
```

l'output mostra una correlazione pressoché perfetta:

	x	x2
x	1.00000	0.99609
x2	0.99609	1.00000

Si centra quindi la variabile esplicativa prima di elevarla al quadrato:

[12] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 36 (file CH01PR27.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/musclemass.csv`).

```
proc standard data=musclemass out=mmcen mean=0;
 var x;
run;
data mmcen2;
 set mmcen;
 x2 = x**2;
run;
proc corr data=mmcen2 noprob;
 var x x2;
run;
```

Si ottiene un netto miglioramento:

	x	x2
x	1.00000	-0.03836
x2	-0.03836	1.00000

Eseguendo la regressione con `proc reg`, si ottiene:

$$\hat{y}_i = 82.93575 - 1.183958(x_i - \bar{x}) + 0.0148405(x_i - \bar{x})^2, \qquad R^2 = 0.7632$$

inoltre:

$$F^* = \frac{MSMOD}{MSRES} = \frac{5915.31}{64.41} = 91.84 \qquad p\text{-}value < 0.0001$$

La verifica circa l'opportunità di un modello del secondo ordine, con ipotesi nulla $H_0 : \beta_{11} = 0$, può essere effettuata in vari modi. Se si scegliesse di usare la devianza di tipo III, si ripeterebbe l'analisi con `proc glm` e si otterrebbe:

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	11492.86575	11492.86575	178.44	<.0001
x2	1	203.13491	203.13491	3.15	0.0811

ovvero:

$$F^* = \frac{SSMOD(x2 \mid x)/1}{MSRES} = \frac{203.13}{64.41} = 3.15 \qquad p\text{-}value = 0.08$$

e si accetterebbe l'ipotesi nulla (in questo caso, quindi, sarebbe sufficiente un modello del primo ordine). Se si volessero comunque convertire i coefficienti $\hat{\boldsymbol{\beta}}'$ di:

$$\hat{y} = \hat{\beta}'_0 + \hat{\beta}'_1(x - \bar{x}) + \hat{\beta}'_{11}(x - \bar{x})^2$$

nei coefficienti $\hat{\boldsymbol{\beta}}$ di:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_{11} x^2$$

si procederebbe così:

$$\hat{\beta}_0 = \hat{\beta}'_0 - \hat{\beta}'_i \bar{x} + \hat{\beta}'_{11} \bar{x}^2 = 82.93575 + 1.183958 \cdot 59.98333 + 0.0148405 * 3598 = 207.3496$$
$$\hat{\beta}_1 = \hat{\beta}'_1 - 2\hat{\beta}'_{11} \bar{x} = -1.183958 - 2 \cdot 0.0148405 \cdot 59.98333 = -2.9643$$
$$\hat{\beta}_{11} = \hat{\beta}'_{11} = 0.0148405$$

quindi:

$$\hat{y}_i = 207.3496 - 2.9463\, x_i + 0.0148405\, x_i^2$$

3.2.7 La regressione con variabili esplicative qualitative

Si possono inserire in un modello di regressione anche variabili qualitative. Ciò si fa spesso ricorrendo alla loro *codifica disgiuntiva completa*: si scompone la variabile qualitativa in tante variabili con valori 0/1 quante sono le sue modalità. Tuttavia, così facendo si otterrebbe una matrice $\mathbf{X}$ con colonne non linearmente indipendenti, in quanto la somma delle colonne sarebbe uguale alla prima colonna costituita da tutti 1. Si risolve spesso il problema eliminando una colonna.[13] Ad esempio:

X			X_A	X_B	X_C			X_A	X_B
A			1	0	0			1	0
B	codifica		0	1	0	eliminazione		0	1
B	disgiuntiva	→	0	1	0	terza	→	0	1
C	completa		0	0	1	colonna		0	0
A			1	0	0			1	0

Se invece la variabile qualitativa ha due sole modalità, è sufficiente sostituirle con 0 e 1.

Esempio 3.28. Si vogliono confrontare 10 compagnie di assicurazione in forma mutua (gli assicurati ne sono i soci) e 10 costituite come società per azioni, per studiare la relazione tra la dimensione della compagnia `x`, in milioni di dollari, e il tempo in mesi `y` occorrente per l'introduzione di una innovazione. Le osservazioni vengono immesse in una matrice di dati `insurinn`,[14] in cui il tipo di compagnia è codificato con 0 se mutua, con 1 se SpA. Un primo modello potrebbe considerare l'interazione tra la dimensione e la forma societaria:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

dove X_{i1} è la dimensione della i-esima compagnia e X_{i2} la sua forma societaria (0 se mutua, 1 se per azioni). In SAS, aggiungendo prima la colonna del prodotto:

```
data iiint;
 set insuinn;
 x1x2 = x1*x2;
run;
proc glm data=iiint;
 model y = x1 x2 x1x2;
run;
```

si vede subito che la devianza di tipo III del prodotto (il suo contributo alla spiegazione della variabilità quando viene aggiunto al modello) è tale da consigliare di escluderlo (p-*value*=0.98):

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x1	1	667.0155931	667.0155931	60.51	<.0001
x2	1	54.5879744	54.5879744	4.95	0.0408
x1x2	1	0.0057084	0.0057084	0.00	0.9821

[13] Si fa così con le variabili *dummy* delle serie storiche.

[14] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 317 (file CH08TA02.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/insurinn.csv`).

Si adotta quindi un modello del primo ordine:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

Usando `proc glm` con l'opzione `clparm` per ottenere gli intervalli di confidenza dei coefficienti:

```
proc glm data=insuinn;
 model y = x1 x2 /clparm;
run;
```

si ottiene:

```
                                      Sum of
Source                        DF         Squares     Mean Square    F Value    Pr > F

Model                          2     1504.413335      752.206667      72.50    <.0001
Error                         17      176.386665       10.375686
Corrected Total               19     1680.800000

                  R-Square     Coeff Var      Root MSE        y Mean

                  0.895058      16.60377      3.221131      19.40000

Source                        DF       Type I SS     Mean Square    F Value    Pr > F

x1                             1     1188.167362     1188.167362     114.51    <.0001
x2                             1      316.245973      316.245973      30.48    <.0001

Source                        DF     Type III SS     Mean Square    F Value    Pr > F

x1                             1     1358.613335     1358.613335     130.94    <.0001
x2                             1      316.245973      316.245973      30.48    <.0001

                              Standard
Parameter        Estimate        Error   t Value   Pr > |t|     95% Confidence Limits

Intercept     33.87406904   1.81385830     18.68     <.0001    30.04716255  37.70097553
x1            -0.10174212   0.00889122    -11.44     <.0001    -0.12050094  -0.08298329
x2             8.05546921   1.45910570      5.52     <.0001     4.97702527  11.13391314
```

Si rileva che il modello supera il test di ipotesi (primo test F) e denota un buon adattamento ai dati ($R^2 = 0.895$). Risultano significative entrambe le variabili (i test F delle devianze di tipo I e di tipo III, i test t) e si ottiene un'espressione dei valori teorici del tipo:

$$\hat{y} = 33.874 - 0.102\, x_1 + 8.055\, x_2$$

Poiché x_2 assume solo i valori 0 e 1, si ha:

$$\begin{aligned} \text{compagnie in forma mutua:} \quad & \hat{y} = 33.874 - 0.102\, x_1 \\ \text{compagnie in forma di spa:} \quad & \hat{y} = 33.874 - 0.102\, x_1 + 8.055 \end{aligned}$$

si ottengono quindi due rette di regressione parallele, con una modesta inclinazione negativa, distanti 8.055 mesi. Poiché 8.055 è solo una stima, è più corretto concludere che, con un livello di confidenza del 95%, il tipo di società ha un effetto, in quanto sono positivi entrambi gli estremi dell'intervallo di confidenza per β_2 (il ritardo delle SpA nell'introdurre l'innovazione varia tra 5 e 11 mesi).

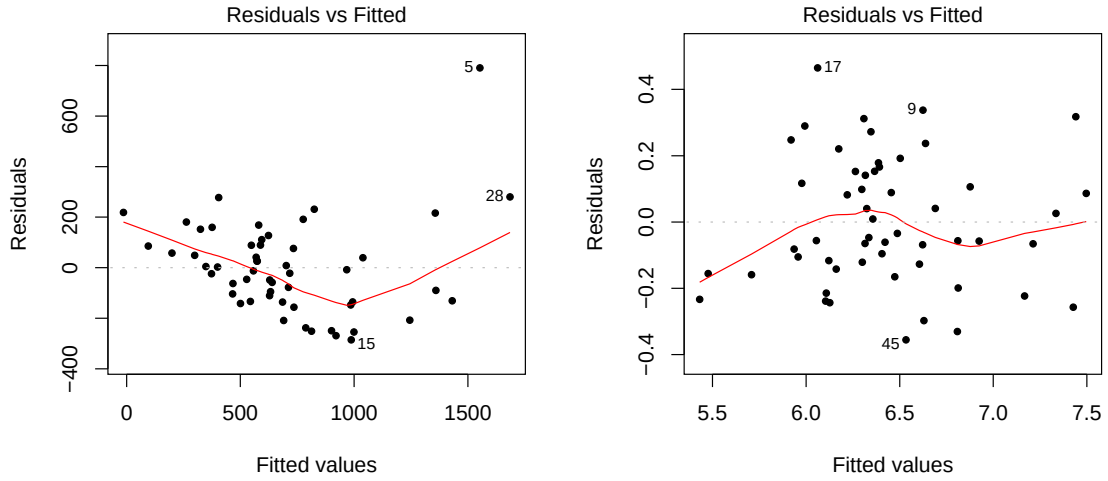


Figura 3.5. Residui e valori teorici usando come variabile risposta il tempo di sopravvivenza `y` (a sinistra) e il suo logaritmo `logy` (a sinistra).

3.2.8 Scelta delle variabili esplicative

Negli studi sperimentali le variabili esplicative sono sotto il controllo del ricercatore, ma negli studi osservazionali accade spesso che il ricercatore disponga di una lunga lista di variabili potenzialmente esplicative e deve quindi scegliere quali includere nel modello e quali tralasciare.

Esempio 3.29. In una unità chirurgica si cerca di determinare quali variabili spieghino meglio il tempo di sopravvivenze `y` dopo un particolare tipo di operazione al fegato. Si considerano le seguenti variabili (tra parentesi i campi di variazione dei valori osservati):

- `x1`: velocità di coagulazione del sangue (da 2.6 a 11.2);
- `x2`: indice prognostico (8 a 96);
- `x3`: test enzimatico (da 23 a 119);
- `x4`: test di funzionalità epatica (da 0.74 a 6.4);
- `x5`: età (da 30 a 70);
- `x6`: sesso (0 per maschio, 1 per femmina);
- `x7` e `x8`: uso di alcool secondo la codifica:

	x7	x8
Nessuno	0	0
Moderato	1	0
Eccessivo	0	1

La matrice di dati `surgunit`[15] contiene 54 osservazioni e 10 variabili: le 8 variabili esplicative, il tempo di sopravvivenza e il suo logaritmo. Il ricercatore, infatti, aveva iniziato tentando un modello del primo ordine con tutte le variabili esplicative, ma il grafico dei residui (a sinistra nella figura 3.5) mostrava una curvatura tale da suggerire di sostituire

[15] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 350 (file CH09TA01.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/surgunit.csv`).

la variabile risposta `y` con il suo logaritmo `logy`; si otteneva così un grafico migliore (a destra nella figura).[16] La matrice di correlazione:

	logy	x1	x2	x3	x4	x5	x6	x7	x8
logy	1.00	0.25	0.47	0.65	0.65	-0.14	0.23	-0.13	0.37
x1	0.25	1.00	0.09	-0.15	0.50	-0.02	0.04	-0.10	0.22
x2	0.47	0.09	1.00	-0.02	0.37	-0.05	0.12	0.13	-0.08
x3	0.65	-0.15	-0.02	1.00	0.42	-0.01	0.14	-0.09	0.12
x4	0.65	0.50	0.37	0.42	1.00	-0.21	0.30	-0.02	0.13
x5	-0.14	-0.02	-0.05	-0.01	-0.21	1.00	0.01	0.15	-0.11
x6	0.23	0.04	0.12	0.14	0.30	0.01	1.00	0.04	-0.06
x7	-0.13	-0.10	0.13	-0.09	-0.02	0.15	0.04	1.00	-0.51
x8	0.37	0.22	-0.08	0.12	0.13	-0.11	-0.06	-0.51	1.00

mostra che `logy` presenta una qualche correlazione lineare con le prime quattro variabili esplicative, soprattutto con `x3` e `x4`, ma anche che `x4` risulta correlata con altre variabili. Si può quindi tentare un modello del primo ordine senza effetti interattivi, ma resta da capire quante e quali variabili esplicative vanno incluse o escluse dal modello.

I criteri

Se vi sono $p-1$ potenziali variabili esplicative, i possibili modelli del primo ordine sono 2^{p-1} (256 nel caso di `surgunit`), dal modello senza alcuna variabile, $Y_i = \beta_0 + \varepsilon_i$, a quello che le comprende tutte. Sono tanti da rendere impraticabile un esame dettagliato di ciascuno. Si sono quindi sviluppate diverse procedure di scelta basate su un singolo indicatore calcolato su tutti i possibili modelli:

a) R^2_p: si sceglie un modello tale che l'aggiunta di altre variabili comporterebbe un miglioramento molto piccolo del coefficiente di determinazione;

b) $R^2_{a,p}$: analogo al precedente, ma si usa il coefficiente di determinazione corretto (che può diminuire aumentando il numero delle variabili);

c) C_p *di Mallow*: indicando con P il numero totale dei coefficienti disponibili, con p il numero di quelli compresi in un modello, è la quantità:

$$C_p = \frac{SSRES_p}{MSRES(x_1, \ldots, x_{P-1})} - (n - 2p)$$

in cui $SSRES_p$ e $MSRES(x_1, \ldots, x_{P-1})$ sono, rispettivamente, la devianza spiegata dal modello con p coefficienti compreso β_0, quindi con $p-1$ variabili esplicative, e la varianza spiegata usando tutte le $P-1$ variabili esplicative disponibili; il C_p è uno stimatore di:

$$\Gamma_p = \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(\mathbb{E}[\hat{Y}_i] - \mu_i^2) + \sum_{i=1}^{n}\sigma^2_{\hat{Y}_i}\right)$$

dove:

- $\hat{Y}_i$ è il valore teorico per l'i-esima osservazione;
- μ_i è il parametro media della distribuzione di $\hat{Y}_i$;
- $\hat{Y}_i - \mu_i = (\hat{Y}_i - \mathbb{E}[\hat{Y}_i]) + (\mathbb{E}[\hat{Y}_i] - \mu_i)$ è l'errore totale;

[16] I residui dovrebbero variare in modo casuale intorno ai valori teorici. Si tornerà sull'argomento nel capitolo 4.

- $\mathbb{E}[\hat{Y}_i] - \mu_i$ è una differenza nulla se il modello è corretto (componente sistematica dell'errore);
- $\hat{Y}_i - \mathbb{E}[\hat{Y}_i]$ è la differenza casuale tra il valore teorico e il suo valore atteso (componente accidentale dell'errore);
- $\sum_{i=1}^{n}(\mathbb{E}[\hat{Y}_i] - \mu_i^2) + \sum_{i=1}^{n}\sigma^2_{\hat{Y}_i}$ è la somma dei valori attesi dei quadrati dell'errore totale; per l'i-esima osservazione:

$$(\hat{Y}_i - \mu_i)^2 = [(\hat{Y}_i - \mathbb{E}[\hat{Y}_i]) + (\mathbb{E}[\hat{Y}_i] - \mu_i)]^2, \qquad \mathbb{E}[(\hat{Y}_i - \mu_i)^2] = (\mathbb{E}[\hat{Y}_i] - \mu_i^2) + \sigma^2_{\hat{Y}_i}$$

Va notato che se si usano tutte le variabili esplicative disponibili si ha per definizione:

$$\frac{SSRES_P}{MSRES(x_1, \ldots, x_{P-1})} = n - p \qquad\qquad C_p = n - p - n + 2p = p$$

Per il resto, se un modello con $p - 1 < P - 1$ variabili esplicative è corretto, se quindi $\mathbb{E}[\hat{Y}_i] = \mu_i$, il valore atteso di C_p è approssimato da p:

$$\mathbb{E}[\hat{Y}_i] = \mu_i \quad \Rightarrow \quad \mathbb{E}[C_p] \approx p$$

Valori di C_p sensibilmente maggiori di p mostrano che il modello non è adeguato (è elevata la componente sistematica dell'errore);

d) AIC_p (*Akaike Information Criterion*): l'indicatore è calcolato come:

$$AIC_p = n \ln SSRES_p - n \ln n + 2p$$

e si scelgono modelli che presentino valori bassi; $SSRES_p$ diminuisce all'aumentare di p, ma il termine $2p$ ovviamente aumenta penalizzando i modelli con molte variabili;

e) SBC_p (*Schwarz' Bayesian Criterion*): è analogo al precedente:

$$SBC_p = n \ln SSRES_p - n \ln n + p \ln n$$

ma penalizza maggiormente i modelli con molte variabili non appena n sia uguale o maggiore di 8, in quanto $\ln 8 = 2.079 > 2$;

f) $PRESS_p$ (*PREdiction Sum of Squares*): si tratta di un indicatore analogo a $SSRES = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, con la differenza che ogni valore teorico viene calcolato usando un modello elaborato escludendo la corrispondente osservazione dal dataset (si esclude la i-esima osservazione dal dataset, si stimano i coefficienti di regressione, si applicano quindi questi ai valori delle variabili esplicative che erano stati esclusi); indicando con $\hat{y}_{i(i)}$ un valore teorico così calcolato:

$$PRESS_p = \sum_{i=1}^{n}(y_i - \hat{y}_{i(i)})^2$$

Si sceglie quindi un modello con un valore basso dell'indicatore.

Esempio 3.30. Nel caso di `surgunit`, i risultati dei diversi criteri sono riportati nella tabella 3.3 per il caso in cui si volesse scegliere solo tra la prime quattro variabili esplicative. Si può rilevare che tutti i criteri porterebbero a scegliere il modello con le sole prime tre variabili, in quanto il coefficiente R^2 è quasi uguale a quello con quattro variabili, il coefficiente corretto è il maggiore, il C_p è il minore (si nota anche per che $p = 5$ l'indicatore vale anch'esso 5), gli altri tre indicatori sono al minimo.

Variabili	p	R^2_p	$R^2_{a,p}$	C_p	AIC_p	SBC_p	$PRESS_p$
Nessuna	1	0.000	0.000	151.498	-75.703	-73.714	13.296
X_1	2	0.061	0.043	141.164	-77.079	-73.101	13.512
X_2	2	0.221	0.206	108.556	-87.178	-83.200	10.744
X_3	2	0.428	0.417	66.489	-103.827	-99.849	8.327
X_4	2	0.422	0.410	67.715	-103.262	-99.284	8.025
X_1 X_2	3	0.263	0.234	102.031	-88.162	-82.195	11.062
X_1 X_3	3	0.549	0.531	43.852	-114.658	-108.691	6.988
X_1 X_4	3	0.430	0.408	67.972	-102.067	-96.100	8.472
X_2 X_3	3	0.663	0.650	20.520	-130.483	-124.516	5.065
X_2 X_4	3	0.483	0.463	57.215	-107.324	-101.357	7.476
X_3 X_4	3	0.599	0.584	33.504	-121.113	-115.146	6.121
X_1 X_2 X_3	4	0.757	0.743	3.391	-146.161	-138.205	3.914
X_1 X_2 X_4	4	0.487	0.456	58.392	-105.748	-97.792	7.903
X_1 X_3 X_4	4	0.612	0.589	32.932	-120.844	-112.888	6.207
X_2 X_3 X_4	4	0.718	0.701	11.424	-138.023	-130.067	4.597
X_1 X_2 X_3 X_4	5	0.759	0.740	5.000	-144.590	-134.645	4.069

Tabella 3.3. Indicatori per la selezione di variabili esplicative tra le prime quattro della matrice di dati `surgunit`.

Gli algoritmi *"best" subsets*

Se si provasse ad esplorare un modello con 8 variabili potenziali, la tabella 3.3 avrebbe 256 righe... Sono state quindi sviluppate procedure per la selezione automatica di un numero ridotto di modelli (*"best" subsets algorithms*). In SAS si possono utilizzare le opzioni `selection` e `best`: `selection` specifica i criteri di scelta del modello, `best` indica il numero massimo di modelli da valutare per ciascun numero di variabili esplicative.

Esempio 3.31. Volendo scegliere tra le 8 possibili variabili esplicative di `surgunit`, si può usare:

```
proc reg data=surgunit;
 model logy = x1-x8 / selection=rsquare adjrsq cp aic sbc best=2;
run;
```

Si usano così i criteri R^2_p, $R^2_{a,p}$, C_p, AIC_p e SBC_p (SAS non prevede il criterio $PRESS_p$ in questi tipi di analisi), chiedendo di vedere solo i due modelli migliori per ciascun numero di variabili esplicative. Si ottiene l'output riprodotto nella figura 3.6. Si può notare che, se si volesse scegliere un modello, la scelta dipenderebbe dal criterio:

a) R^2_p è ovviamente massimo con 8 variabili (0.8461), ma diminuisce di pochissimo con 7 se si esclude `x4` (0.8460);

b) $R^2_{a,p}$ è massimo (0.8234) con 6 variabili, escludendo `x4` e `x7`;

c) C_p è minimo (5.5406) con 5 variabili, escludendo `x4`, `x5` e `x7`;

d) AIC_p è minimo (-163.8343) con 6 variabili, escludendo `x4` e `x7` (concorda quindi con $R^2_{a,p}$);

e) SBC_p è minimo (-153.40643) con 4 variabili, `x1`, `x2`, `x3` e `x8`.

Number in Model	R-Square	Adjusted R-Square	C(p)	AIC	SBC	Variables in Model
1	0.4276	0.4166	117.4094	-103.8269	-99.84889	x3
1	0.4215	0.4104	119.1712	-103.2615	-99.28357	x4
2	0.6633	0.6501	50.4716	-130.4833	-124.51634	x2 x3
2	0.5995	0.5838	69.1318	-121.1126	-115.14561	x3 x4
3	0.7780	0.7647	18.9145	-150.9849	-143.02899	x2 x3 x8
3	0.7573	0.7427	24.9805	-146.1609	-138.20494	x1 x2 x3
4	0.8299	0.8160	5.7508	-163.3514	-153.40643	x1 x2 x3 x8
4	0.8144	0.7993	10.2670	-158.6593	-148.71434	x2 x3 x4 x8
5	0.8374	0.8205	5.5406	-163.8052	-151.87127	x1 x2 x3 x6 x8
5	0.8358	0.8187	6.0182	-163.2654	-151.33152	x1 x2 x3 x5 x8
6	0.8434	0.8234	5.7874	-163.8343	-149.91140	x1 x2 x3 x5 x6 x8
6	0.8392	0.8187	7.0295	-162.3890	-148.46607	x1 x2 x3 x6 x7 x8
7	0.8460	0.8226	7.0295	-162.7356	-146.82378	x1 x2 x3 x5 x6 x7 x8
7	0.8436	0.8198	7.7352	-161.8958	-145.98397	x1 x2 x3 x4 x5 x6 x8
8	0.8461	0.8188	9.0000	-160.7710	-142.87013	x1 x2 x3 x4 x5 x6 x7 x8

Figura 3.6. Output di una `proc reg` con opzione `selection=rsquare adjrsq cp aic sbc best=2`.

In realtà, tuttavia, il vero obiettivo di tali procedure non è la scelta di un modello, ma piuttosto la selezione di un numero ristretto di modelli "buoni" tra i 2^{p-1} possibili. I modelli selezionati vanno poi valutati con gli strumenti illustrati nel capitolo 4.

Gli algoritmi *stepwise*

Se le potenziali variabili esplicative sono nettamente più numerose (30 o più), la selezione di un sottoinsieme di modelli "buoni" non è più praticabile. Sono state quindi sviluppate anche procedure automatiche per la selezione di un singolo modello, che esaminano una variabile potenziale alla volta; in termini molto generali, si aggiungono al modello variabili per le quali il *p-value* corrispondente ad un test t o F sia minore di una soglia "di entrata", si escludono quelle per le quali il *p-value* sia maggiore di una soglia "di mantenimento".[17] Più in dettaglio, vi sono tre algoritmi (tra parentesi le soglie di default in SAS):

a) *Forward Stepwise Regression*: l'algoritmo inizia costruendo $P-1$ modelli, quante solo le potenziali variabili esplicative, e sceglie quello che risulta migliore sulla base del *p-value*; procede poi costruendo altri $P-2$ modelli con due variabili, aggiungendo una delle restanti a quella già inclusa, e sceglie quello per il quale il *p-value* è minore; il processo viene ripetuto fino a che non si sono considerate tutte le variabili, oppure fino a che il test presenta per tutte le variabili non ancora incluse un *p-value* superiore alla soglia "di entrata" (0.15), nel qual caso l'algoritmo si ferma; ad ogni passo, tuttavia,

[17] Dal momento che ad ogni passo si valuta una sola variabile, i test t e F sono equivalenti; cfr. esempi 3.19 e 3.25.

appena aggiunta una variabile a quelle già presenti, si ripete il test su queste ultime (si valuta cioè il contributo che darebbe ciascuna se fosse aggiunta ad un modello comprendente l'ultima variabile inclusa) e quelle che presentano un *p-value* maggiore della soglia "di mantenimento" (0.15) vengono escluse dal modello;

b) *Forward Selection*: si tratta di una versione semplificata del precedente, in quanto le variabili vengono solo aggiunte, fino a che il *p-value* è minore della soglia "di entrata" (0.50), senza verificare se una variabile già inclusa dovrebbe essere esclusa dopo l'aggiunta di altre;

c) *Backward Elimination*: è l'opposto del precedente; l'algoritmo inizia con un modello comprendente tutte le variabili e procede poi eliminando una alla volta le variabili col *p-value* più alto, se maggiore della soglia "di mantenimento" (0.10).

Esempio 3.32. Si possono eseguire i tre algoritmi in SAS usando le opzioni `stepwise`, `forward` e `backword`; si possono cambiare i valori di default delle soglie di entrata e di mantenimento per i *p-value* con le opzioni `slentry` e `slstay`. Usando il primo con `surgunit` e provando tutte le 8 variabili:

```
proc reg data=surgunit;
 model logy = x1-x8 / selection=stepwise;
run;
```

si ottiene un output che espone in dettaglio i singoli passi dell'algoritmo e termina con un prospetto di sintesi:

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x3		1	0.4276	0.4276	117.409	38.84	<.0001
2	x2		2	0.2357	0.6633	50.4716	35.70	<.0001
3	x8		3	0.1147	0.7780	18.9145	25.85	<.0001
4	x1		4	0.0519	0.8299	5.7508	14.93	0.0003
5	x6		5	0.0076	0.8374	5.5406	2.23	0.1418

Si può rilevare che, con un livello di significatività 0.15 (default sia per `slentry` che per `slstay`) viene scelto il modello con il migliore C_p; se si fosse aggiunto `slentry=0.05`, sarebbe stato scelto il modello con il migliore SBC_p. Dal momento che non si è avuta nessuna esclusione di variabili già immesse, si sarebbe ottenuto lo stesso risultato con `selection=forward slentry=0.15`.

Da notare che, quando si usa l'algoritmo di *Forward Stepwise Regression*, il *p-value* "di entrata" non deve essere maggiore di quello di "mantenimento"; se così fosse, infatti, una variabile potrebbe essere ciclicamente prima aggiunta, poi eliminata, poi nuovamente aggiunta ecc.

Esempio 3.33. Per eseguire su `surgunit` l'algoritmo di *Backward Elimination*:

```
proc reg data=surgunit;
 model logy = x1-x8 / selection=backward;
run;
```

Si ottiene in coda all'output il prospetto di sintesi:

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x4	7	0.0001	0.8460	7.0295	0.03	0.8645
2	x7	6	0.0026	0.8434	5.7874	0.77	0.3835
3	x5	5	0.0060	0.8374	5.5406	1.80	0.1862
4	x6	4	0.0076	0.8299	5.7508	2.23	0.1418

Rispetto all'esempio precedente, ora la variabile `x6` viene esclusa dal modello; ciò accade perché il livello di default di `slstay` è 0.10, mentre è 0.15 per l'opzione `stepwise`.

Negli esempi appena visti si raggiungono gli stessi risultati con tutti e tre gli algoritmi, ma ciò non accade sempre.

Soprattutto, si deve notare che a volte non ha molto senso includere una variabile ed escluderne un'altra; in `surgunit`, ad esempio, le due variabili `x7` (uso moderato, oppure no, di alcool) e `x8` (uso eccessivo, oppure no, di alcool), sono in realtà la codifica di un'unica variabile qualitativa con tre modalità (nessun uso di alcool, uso moderato, uso eccessivo). Le due variabili andrebbero quindi incluse o escluse insieme. Analogamente, nel caso di modelli di ordine superiore, se si includono effetti interattivi o potenze è preferibile che siano presenti anche i termini di primo grado. In SAS ciò si può ottenere "raggruppando" due o più variabili.

Esempio 3.34. Per includere o escludere insieme le variabili `x7` e `x8`, si usa la seguente sintassi:

```
proc reg data=surgunit;
 model logy = x1-x6 {x7 x8} / selection=stepwise
  groupnames='x1' 'x2' 'x3' 'x4' 'x5' 'x6' 'x7 x8';
run;
```

Si racchiudono tra parentesi graffe le variabili da raggruppare. Si aggiunge per comodità l'opzione `groupnames`, che assegna etichette alle variabili; in caso contrario, verrebbero mostrate tutte come `GROUPn`, con `n` variabile da 1 a 7, da 1 a 6 per i "gruppi" costituiti da una sola variabile, 7 per quello costituito da `x7` e `x8`. Si ottiene:

Summary of Stepwise Selection

Step	Group Entered	Group Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x3		1	0.4276	0.4276	117.409	38.84	<.0001
2	x2		2	0.2357	0.6633	50.4716	35.70	<.0001
3	x7 x8		4	0.1167	0.7800	20.3519	12.99	<.0001
4	x1		5	0.0517	0.8317	7.2269	14.75	0.0004
5	x6		6	0.0075	0.8392	7.0295	2.20	0.1450

Capitolo 4

L'analisi diagnostica

Come visto nei capitoli precedenti, nella costruzione di un modello lineare normale il ricercatore assume un modello campionario (la famiglia parametrica normale) e, basandosi sulle informazioni contenute nei dati, definisce un modello di riparametrizzazione; perviene così ad un modello del tipo:

$$\mathbf{Y} = \mathbf{A}\boldsymbol{\eta} + \boldsymbol{\varepsilon} \qquad\qquad \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

(a sinistra la formulazione usata per i modelli ANOVA, a destra quella per i modelli regressivi). Procede poi, sulla base dei valori osservati, alla stima dei parametri:

$$\hat{\boldsymbol{\eta}} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y} \qquad\qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

a quella dei valori teorici:

$$\hat{\mathbf{y}} = \mathbf{A}\hat{\boldsymbol{\eta}} = \mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{y} = \mathbf{H}\mathbf{y} \qquad\qquad \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

al calcolo dei residui:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Avvalendosi dell'analisi della varianza, effettua test di ipotesi sul modello nel suo complesso e sui singoli parametri, e calcola per questi anche intervalli di confidenza. Nei modelli regressivi calcola anche coefficienti di determinazione totali e parziali, effettua test e calcola intervalli anche per i valori teorici.

Tuttavia, nonostante l'esito apparentemente soddisfacente dei test, l'adeguatezza del modello risposa su alcuni assunti che potrebbero non essere, in realtà, soddisfatti. In particolare:

a) *componente parametrica* del modello: non solo può risultare opportuno, come in parte già visto, escludere alcune variabili o includerne di nuove, ma anche includere quelle già presenti in forma diversa (verifica del modello di riparametrizzazione);

b) *componente casuale* del modello: occorre verificare le ipotesi di omoschedasticità, di indipendenza e di normalità (verifica del modello campionario);

c) *qualità dei dati*: potrebbero esservi dati anomali.

Si richiede quindi un'attività diagnostica, che si basa prevalentemente sull'*analisi dei residui*,[1] a proposito della quale va richiamato quanto già evidenziato nel capitolo 1.

[1] Nella realtà, come in parte mostrato nell'esempio 3.29, si procede *prima* alla verifica del modello sulla

4.1 La variabile aleatoria "residuo"

Errore e residuo osservato sembrano molto simili; con riferimento ad un modello regressivo:

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta} \qquad \mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

Inoltre la media dei residui è nulla, come quella dell'errore, perché è nulla la loro somma in quanto $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, ma le somme di riga e di colonna di $\mathbf{I} - \mathbf{H}$ sono nulle:[2]

$$\boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}) \qquad \sum_{i=1}^{n} e_i = 0 \quad \Rightarrow \quad \bar{e} = 0$$

e la varianza dei residui, $MSRES$, è uno stimatore di σ^2 (se il modello è adeguato).

Sarebbe tuttavia *errato intendere il residuo come una determinazione della variabile aleatoria errore. Il residuo osservato è una determinazione della variabile aleatoria residuo, che ha distribuzione diversa da quella della v.a. errore.*

Un esempio molto semplice può aiutare a comprendere meglio la differenza. Si abbiano un modello campionario $Y_i = N(\mu_i, \sigma^2)$ ed un modello di riparametrizzazione $\mathbb{E}[Y_i] = \mu$ (si assume che le variabili risposta abbiano distribuzione normale, poi che abbiano in comune non solo la varianza, ma anche la media). Ne segue un modello:

$$Y_i = \mu + \varepsilon_i$$

in cui Y_i è una variabile aleatoria osservabile, μ è un parametro *incognito* (e destinato a rimanere tale), ε_i una variabile aleatoria *non osservabile*. La variabile aleatoria residuo è:

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\mu}_i = (\mu + \varepsilon_i) - \hat{\mu}_i = \varepsilon_i + (\mu - \hat{\mu}_i)$$

Se invece, a seguito di diversa riparametrizzazione, si usa il modello:

$$Y_i = \mu_i + \varepsilon_i = \alpha + \beta X_i + \varepsilon_i$$

la variabile aleatoria residuo diventa:

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\alpha} - \hat{\beta}x_i = \varepsilon_i + (\mu_i - \hat{\alpha} - \hat{\beta}x_i)$$

In entrambi i casi si ha una chiara differenza tra le variabili aleatorie errore e residuo, e *la differenza dipende dalla riparametrizzazione.*

---

base dei residui, *poi* ai test di ipotesi e al calcolo degli intervalli di confidenza. Si tratta di un processo di aggiustamenti successivi: si esamina un modello, se le attività diagnostiche non danno buon esito lo si adatta (trasformazioni della variabile risposta o delle variabili esplicative, inclusione/esclusione di variabili ecc.), si ripetono le attività diagnostiche fino a che il modello non appaia adeguato e solo a questo punto si traggono le inferenze che si interessano.

[2] Questo perché le somme di riga e di colonna di $\mathbf{H}$ valgono 1 (cfr. capitolo 1, nota 19). Si può anche considerare che in un modello regressivo semplice si ha:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

e la seconda somma non è altro la derivata rispetto a β_0, che vienne posta uguale a zero per minimizzare gli scarti tra valori osservati e valori teorici (cfr. sez. 3.1.1); più in generale, in un modello lineare si minimizza la quantità $(\mathbf{Y} - \mathbf{A}\boldsymbol{\eta})'(\mathbf{Y} - \mathbf{A}\boldsymbol{\eta})$ uguagliando a zero $-2\mathbf{A}'\mathbf{Y} + 2\mathbf{A}'\mathbf{A}\boldsymbol{\eta} = -2\mathbf{A}'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$, che è appunto la somma dei residui (cfr. sez. 1.4.1).

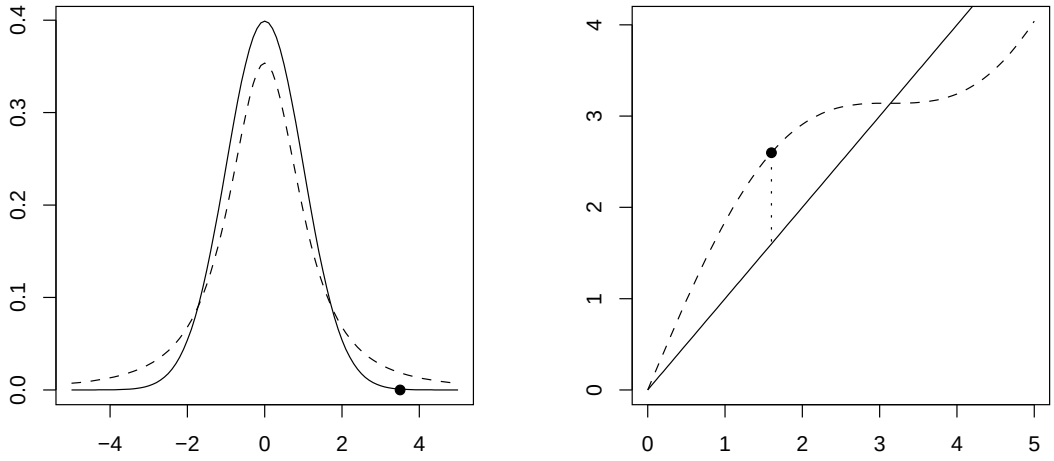


Figura 4.1. Un valore può apparire anomalo rispetto ad una distribuzione, ma non rispetto ad un'altra (a sinistra); un residuo può apparire grande secondo un modello, nullo secondo un altro (a destra).

Si era infatti visto nel capitolo 1 che:

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I} \qquad \mathrm{Cov}(\mathbf{e}) = (\mathbf{I}-\mathbf{H})\sigma^2$$

quindi che *la struttura di varianza e covarianza dei residui* ***non*** *riproduce l'indipendenza e l'omoschedasticità della variabile aleatoria errore, ma dipende dalla matrice di riparametrizzazione.*

Un primo effetto della diversa distribuzione è che quelli che sarebbero valori anomali rispetto alla distribuzione della v.a. errore, potrebbero non risultare tali rispetto ad un'altra distribuzione (figura 4.1 a sinistra). Può comunque anche accadere che un residuo appaia grande secondo un modello, piccolo o nullo secondo un altro (figura 4.1 a destra). In sostanza, non è possibile distingure in un residuo la parte dovuta all'errore casuale e quella dovuta ad una errata specificazione del modello.

Si usano comunque i residui (le determinazioni della variabile aleatoria errore) come "rappresentanti" dell'errore, in quanto esiste una relazione tra le due variabili aleatorie:[3]

$$\mathbf{e} = (\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}$$

e se ne possono trarre due considerazioni:

a) sia $\mathbf{Y}$ che $\boldsymbol{\varepsilon}$ sono vettori dello spazio $\mathbb{R}^n$; la matrice $(\mathbf{I}-\mathbf{H})$ è una matrice di proiezione ortogonale su uno spazio di $n-p$ dimensioni, supplementare a quello individuato dalle p colonne della matrice di riparametrizzazione; se $n \gg p$, la proiezione di $\boldsymbol{\varepsilon}$ ne fornisce una buona rappresentazione;

b) la dipendenza tra gli e_i discende dalla matrice di riparametrizzazione, che è di rango p; anche qui se $n \gg p$ le covarianze sono trascurabili (gli elementi di $\mathbf{I}-\mathbf{H}$ sulla diagonale principale sono nettamente maggiori degli altri).

[3] Infatti: $\mathbf{e} = (\mathbf{I}-\mathbf{H})\mathbf{Y} = (\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta} + (\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon} = [\mathbf{X} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})]\boldsymbol{\beta} + (\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon} = (\mathbf{X}-\mathbf{X})\boldsymbol{\beta} + (\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon} = (\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}$.

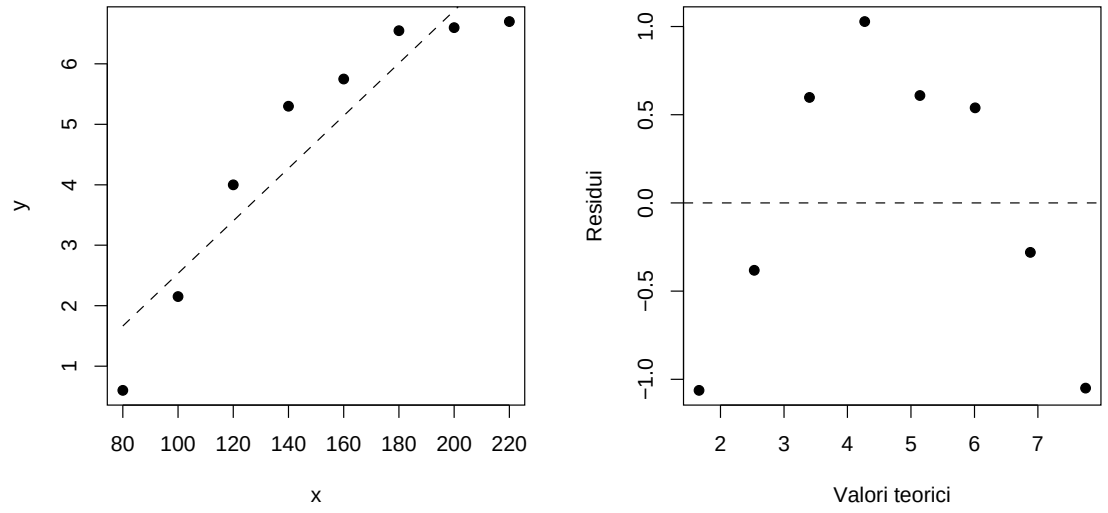


Figura 4.2. Grafici della variabile risposta contro la variabile esplicativa (*scatter plot*) e dei residui contro i valori teorici (*residual plot*) per il dataset `transit`.

4.2 Adeguatezza del modello

4.2.1 Verifica della linearità

I modelli lineari sono tali *nei parametri*, ma le variabili esplicative possono comparire in qualsiasi forma. Nei modelli regressivi, in particolare, si ipotizza che esista una relazione tra i valori delle variabili esplicative e i valori attesi della variabile risposta; si parte normalmente da relazioni espresse mediante modelli del primo ordine, in cui le variabili esplicative compaiono sempre in termini di primo grado, ma si deve verificare che tale assunzione iniziale sia corretta.

Per verificare l'adeguatezza di un modello del primo ordine, si possono utilizzare diagrammi di dispersione (*scatter plot*) della variabile risposta contro le variabili esplicative, oppure dei residui contro i valori teorici o le variabili esplicative (quando la variabile esplicativa è una sola, un grafico dei residui contro i valori teorici è equivalente ad uno contro i valori della variabile esplicativa, in quanto i primi sono funzione lineare dei secondi e, quindi, cambia solo la scala dell'asse delle ascisse).

Esempio 4.1. Si distribuiscono in $n = 8$ città delle cartine sui percorsi serviti da un trasporto pubblico e si rileva l'aumento del numero di persone che lo utilizzano. Il numero delle cartine, in migliaia, è la variabile esplicativa `x`; l'aumento dell'utilizzo, anch'esso in migliaia, è la variabile di risposta `y` (matrice di dati `transit`).[4] In SAS:

```
proc reg data=transit;
 model y = x;
 plot  y*x r.*p.;
run;
```

[4]Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 19 (file CH03TA01.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/transit.csv`).

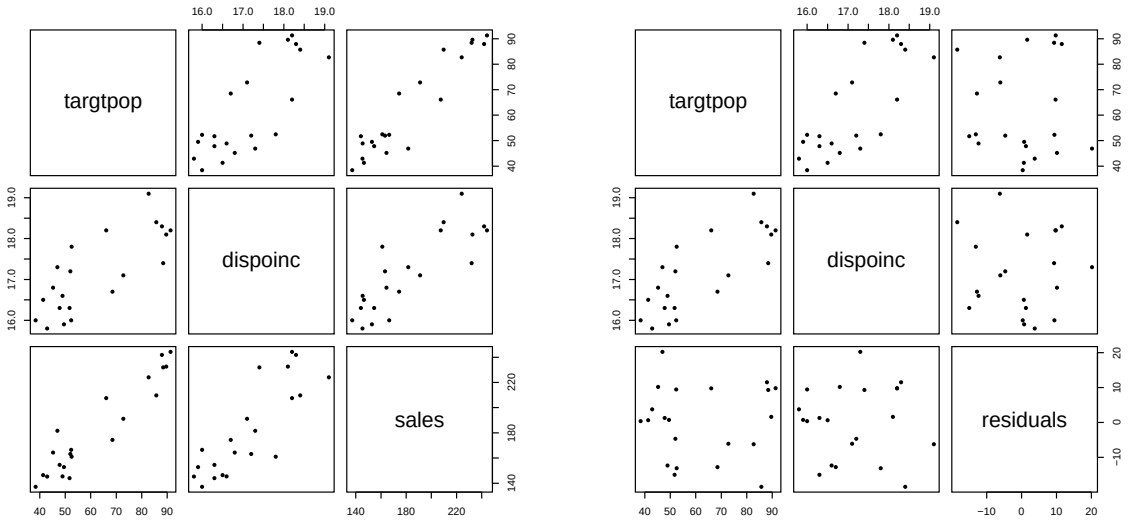


Figura 4.3. Scatter plot matrix della variabile risposta (a sinistra) e dei residui (a destra) contro le variabili esplicative (matrice di dati `dwaine`).

L'istruzione `plot y*x r.*p.` traccia il grafico della variabile risposta contro la variabile esplicativa, `y*x`, e quello dei residui contro i valori teorici, `r.*p.`, dove `r.` è un'abbreviazione per *residual* e `p.` per *predicted*. Si ottengono grafici analoghi a quelli riprodotti nella figura 4.2 per il modello $Y_i = -1.8161 + 0.0435x$; già il primo mostra l'inadeguatezza di una relazione lineare tra `y` e `x`, inadeguatezza che risulta ancora più evidente nel secondo.

Quando intervengono più variabili esplicative, i grafici dei residui contro i valori previsti rimangono analoghi a quelli appena visti, ma quelli della variabile risposta, o dei residui, contro le variabili esplicative non sono più semplici *scatter plot*, proprio perché le variabili esplicative sono più di una. In questi casi si ricorre a grafici detti *scatter plot matrix*.

Esempio 4.2. Nel caso di `dwaine` (esempio 3.13), in cui vi sono due variabili esplicative, un grafico di tipo *scatter plot matrix*[5] consente di visualizzare insieme le relazioni a coppie tra la variabile risposta `sales` e le variabili esplicative (figura 4.3 a sinistra), oppure tra i residui e le variabili esplicative (figura 4.3 a destra). Nel primo caso, il secondo e il terzo grafico della prima riga dall'alto della matrice mostrano, rispettivamente, i grafici di `targtpop` (sull'asse delle ordinate) contro `dispoinc` e `sales` (sull'asse delle ascisse); il primo e il secondo grafico dell'ultima colonna a destra mostrano, rispettivamente, i grafici di `targtpop` e di `dispoinc` (sull'asse delle ordinate) contro `sales` (sull'asse delle ascisse). Analogamente per gli altri grafici. Si può notare che `sales`, `targtpop` e `dispoinc` presentano una chiara correlazione, confermata da `proc corr`:

```
Pearson Correlation Coefficients, N = 21

                targtpop      dispoinc         sales
targtpop         1.00000       0.78130       0.94455
dispoinc         0.78130       1.00000       0.83580
sales            0.94455       0.83580       1.00000
```

[5] In R si possono usare varie funzioni, la più semplice delle quali è `pairs()`. In SAS si può usare la `proc sgscatter`, che però è disponibile solo a partire dalla versione 9.2.

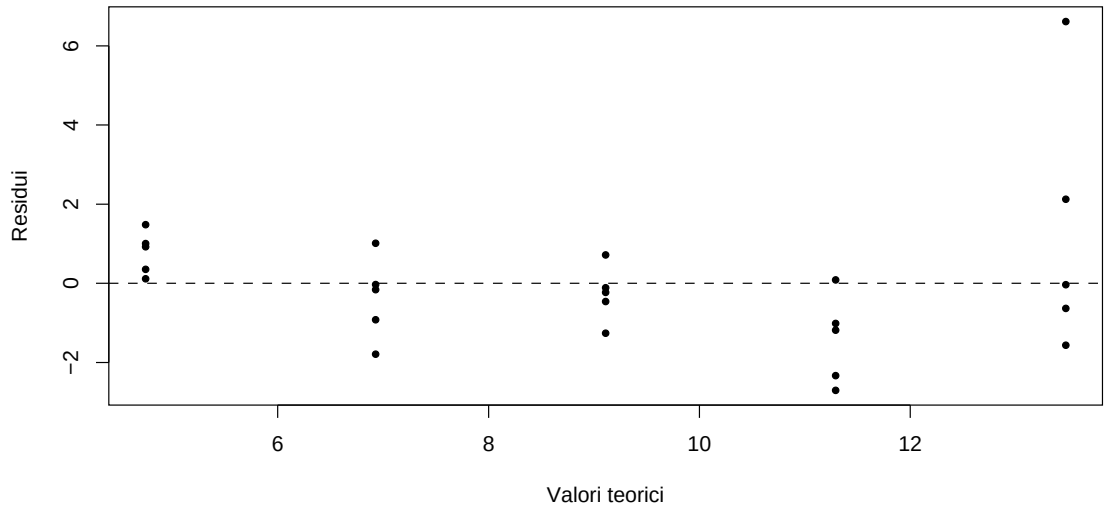


Figura 4.4. Grafico dei residui contro i valori teorici (*residual plot*) per il dataset `plasma`.

ma i residui non mostrano segni di non-linearità, né di varianza incostante. Si può quindi ritenere adeguato un modello del primo ordine del tipo $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$.

4.2.2 Verifica della costanza della varianza

Se il modello è corretto, la varianza dei residui deve essere costante in quanto stima di quella dell'errore (ipotesi di omoschedasticità). In tal caso, il grafico dei residui contro i valori teorici deve mostrare punti compresi entro una banda orizzontale centrata sulla retta $y = 0$, senza prevalenza di segni positivi o negativi; se così non è, si può dedurre che la varianza dei residui non è costante.

Esempio 4.3. Si misura la presenza di poliammine nel sangue di 25 bambini di età compresa tra 0 e 4 anni (matrice di dati `plasma`).[6] Il grafico dei residui contro i valori teorici, figura 4.4, mostra chiaramente sia un andamento curvilineo, che mette in dubbio l'ipotesi di linearità, sia un progressivo allontanamento dalla retta $y = 0$, sintomo di una varianza dei residui non costante.

4.2.3 Verifica dell'indipendenza

I residui dovrebbero anche essere indipendenti; in un grafico contro i valori teorici, quindi, dovrebbero disporsi in modo casuale intorno alla retta $y = 0$.

Non accade così né in `transit` né in `plasma`, ma l'anomalia è da imputare alla scelta del modello. Il problema della correlazione dei residui si presenta soprattutto nell'analisi delle serie storiche, dove è indice della presenza di fattori stagionali.

[6] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 133 (file CH03TA08.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/plasma.csv`).

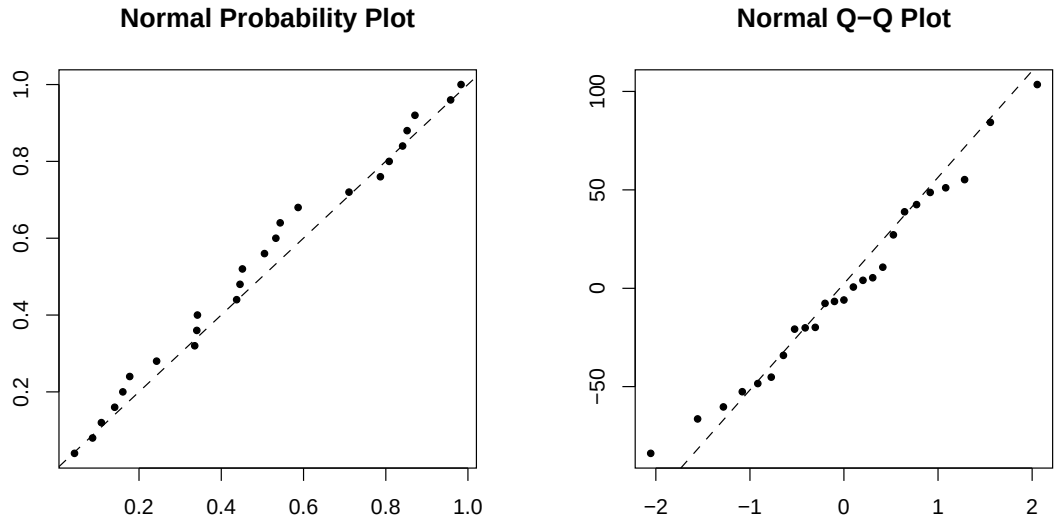


Figura 4.5. Normal Probability Plot e *Normal Q-Q plot* per `toluca`.

4.2.4 Verifica della normalità

I residui dovrebbero distribuirsi normalmente intorno ai valori teorici. Per la verifica si usano i `normal probability plot`, nei quali si confrontano i residui con i loro valori attesi secondo un'ipotesi di normalità.

Si usano a questo scopo due grafici equivalenti, che si basano entrambi sui residui ordinati dal più piccolo, $e_{(1)}$, al più grande, $e_{(n)}$:

a) *normal probability plot*: in ascissa vi sono i valori della funzione di ripartizione normale $\Phi\left(\dfrac{e_{(i)}-0}{\sqrt{MSRES}}\right)$, in ordinata i valori i/n; se i residui hanno una distribuzione normale, i punti si dispongono lungo la retta $y = x$;

b) *normal Q-Q plot*: in ascissa vi sono i quantili ottenuti invertendo una funzione di ripartizione normale; sembrerebbe di poter calcolare, per ogni i, i reciproci di $\Phi(i/n)$, ma così per l'ultimo termine di avrebbe $\Phi(1) = \infty$; si usano quindi espressioni corrette, del tipo $(i - 0.5)/n$ o $i/(n + 1)$.[7] In ordinata vi sono i residui osservati. Se i residui hanno una distribuzione normale, i punti si dispongono lungo una retta che passa per il primo e il terzo quartile.

Come nel caso dell'indipendenza, residui non normali possono presentarsi anche quando non sono soddisfatte le condizioni di linearità e di costanza della varianza, che quindi vanno esplorate per prime.

Esempio 4.4. In SAS i due grafici si ottengono usando le abbrevizioni `npp.` per il *normal probability plot*, `nqq.` per il *Q-Q plot*; ad esempio, per `toluca`, con:

```
proc reg data=toluca;
 model y = x;
 plot r.*npp. r.*nqq.;
run;
```

[7] SAS usa i reciproci di $\Phi\left(\dfrac{i-0.375}{n+0.25}\right)$.

si ottengono grafici simili a quelli riprodotti nella figura 4.5.

4.2.5 Azioni correttive

Occorre verificare in primo luogo la linearità e la costanza della varianza (che, se non soddisfatte, danno luogo ad apparente correlazione e/o non-normalità dei residui). Si devono distinguere due situazioni:

a) non-linearità, ma costanza della varianza: *si trasforma la variabile esplicativa*;

b) varianza non costante: *si trasforma la variabile risposta.*

Resta ovvio che, nel secondo caso, se alla varianza non costante si aggiunge anche la non-linearità e se la trasformazione della variabile risposta non risolve entrambi i problemi, si deve poi provare ad operare anche sulla variabile esplicativa.

Non-linearità, ma varianza dei residui costante

Se la verifica della linearità non ha dato buon esito, ma la varianza risulta costante, si deve intervenire *solo sulla variabile esplicativa*; se infatti si trasformasse la variabile risposta, si potrebbe indurre una varianza variabile dei residui.

Si possono provare diverse soluzioni, suggerite dalla forma del grafico della variabile risposta contro la variabile esplicativa; si tratta di trovare una funzione che approssimi al meglio il grafico. Ad esempio una funzione:

- crescente e concava verso il basso: $X' = \ln(X)$, $X' = \log_{10}(X)$, $X' = \sqrt{X}$;
- crescente e convessa verso il basso: $X' = X^2$, $X' = \exp(X)$
- decrescente e concava verso il basso: $X' = \sqrt{c - X}$;
- decrescente e convessa verso il basso: $X' = 1/X$, $X' = \exp(-X)$.

Dopo aver effettuato diversi tentativi, si sceglie la funzione che meglio soddisfa il requisito di linearità.

Esempio 4.5. La matrice di dati `salestraining`[8] registra l'efficacia di 10 venditori, misurata mediante l'attribuzione di un punteggio `y`, dopo un periodo di addestramento di `x` giorni. Un modello regressivo del primo ordine, $Y_i = \beta_0 + \beta_1 X_i$, non dà risultati soddisfacenti, in quanto i grafici variabile risposta / variabile esplicativa e residui / valori teorici (figura 4.6 in alto) mostrano un andamento curvilineo, concavo verso il basso. Da altro punto di vista, i residui sono sostanzialmente compresi entro una banda delimitata dalle rette $y = \pm 10$; non vi sono quindi sintomi di una varianza non costante (la non linearità è sufficiente a dar conto dell'alternanza di valori positivi e negativi, che tra l'altro sono di numero uguale). Si prova quindi a sostituire i valori della variabile esplicativa con i loro quadrati, $Y_i = \beta_0 + \beta_1 \sqrt{X_i}$. In SAS:

```
data salestrasqrt;
 set salestra;
 sqrx = sqrt(x);
run;
```

[8] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 130 (file CH03TA07.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/salestraining.csv`).

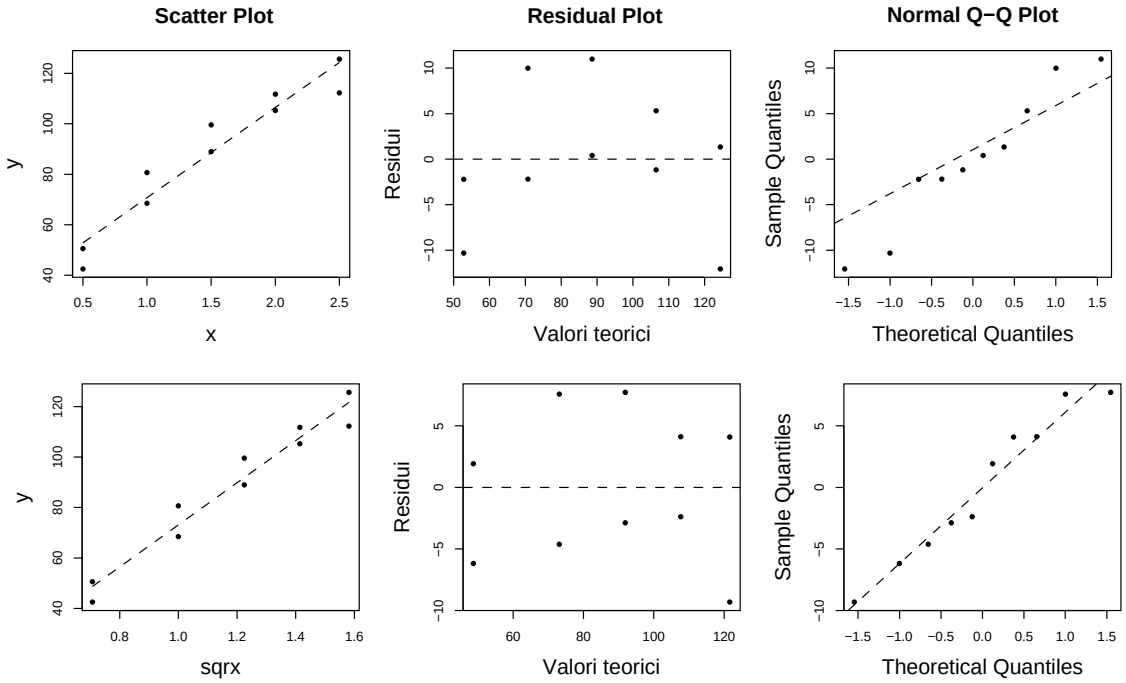


Figura 4.6. Analisi grafica di un modello $Y_i = \beta_0 + \beta_1 X_i$ (in alto) e di un modello $Y_i = \beta_0 + \beta_1 \sqrt{X_i}$ (in basso) per la matrice di dati `salestraining`.

```
proc reg data=salestrasqrt;
 model y=sqrx;
 plot y*sqrx r.*p. r.*nqq.;
run;
```

Si ottengono così grafici come quelli riprodotti nella figura 4.6 in basso, che mostrano un netto miglioramento (migliora anche R^2, che passa da 0.9256 a 0.9545).

Varianza dei residui con costante

Se la varianza non risulta costante, è necessario trasformare la variabile risposta in modo da ottenere valori teorici rispetto ai quali i residui siano meglio distribuiti. Ad esempio, se i residui mostrano un andamento crescente ed una variabilità anch'essa crescente, si può provare a trasformare Y in $1/Y$ se l'andamento è lineare, in $\sqrt{Y}$ se si rileva una concavità verso il basso; se l'andamento è decrescente con una convessità verso il basso si può provare con un logaritmo (spesso meglio in base 10). Anche in questo caso si devono provare diverse soluzioni e valutarle sulla base dei grafici.

Esempio 4.6. Ritornando alla matrice di dati `plasma` (figura 4.4), si nota che i residui hanno un andamento prima decrescente, poi crescente, con una convessità verso il basso. Si può quindi provare una trasformazione logaritmica della variabile risposta, usando le basi e e 10:

```
data plasmalog;
 set plasma;
```

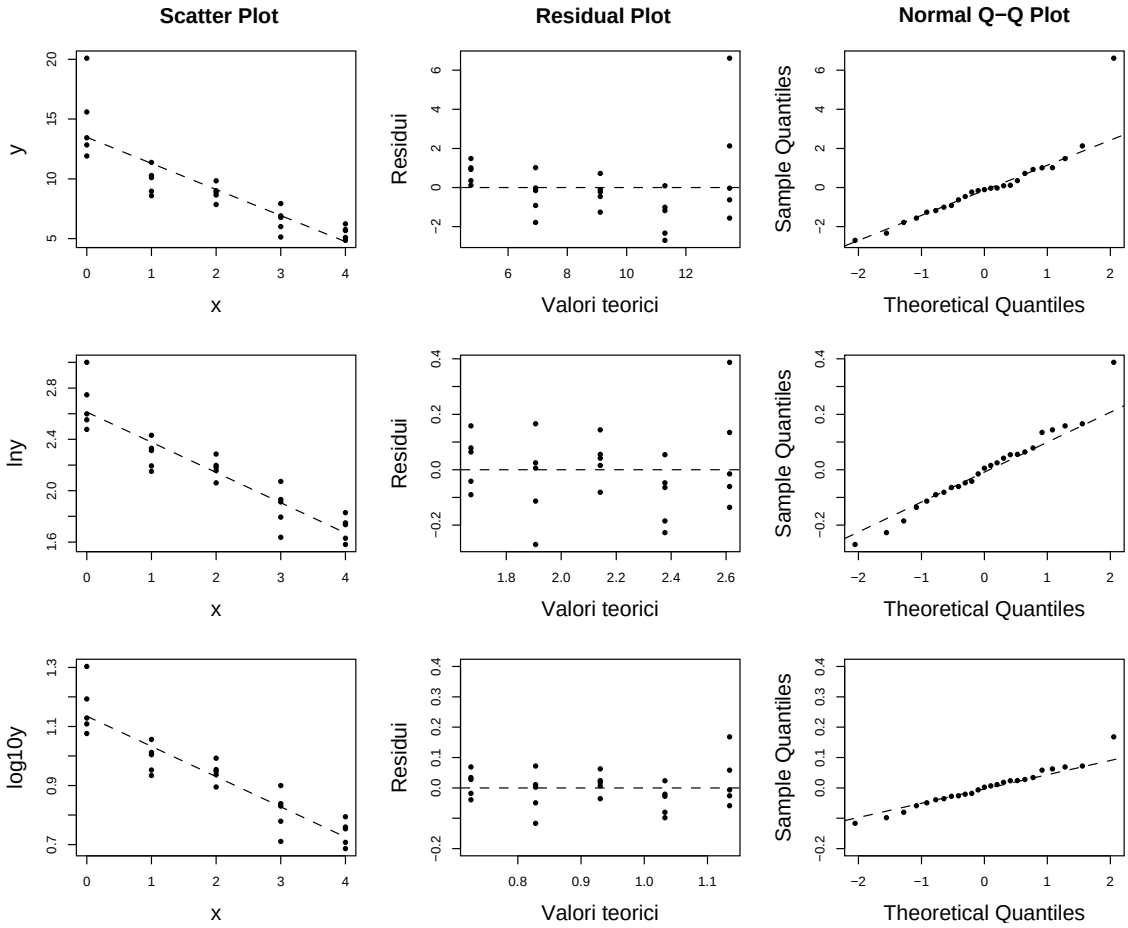


Figura 4.7. Analisi grafica dei modelli $Y_i = \beta_0 + \beta_1 X_i$ (in alto), $\ln(Y_i) = \beta_0 + \beta_1 X_i$ (al centro), $\log_{10}(Y_i) = \beta_0 + \beta_1 X_i$ (in basso) per la matrice di dati `plasma`.

```
 lny = log(y);
 log10y = log(y)/log(10);
run;
```

Usando poi i tre modelli `y = x`, `lny = x` e `log10y = x`, si vede che gli ultimi due presentano lo stesso miglioramento rispetto al primo sotto due aspetti:

– il test F assicura la significatività di tutti i modelli, ma F^* (il rapporto tra varianza spiegata e varianza residua) aumenta da 70.21 a 134.2;

– il coefficiente di determinazione R^2 aumenta da 0.75 a 0.85.

Tuttavia, l'analisi grafica (figura 4.7) mostra che con i logaritmi in base 10 i residui rispetto ai valori teorici oscillano entro una banda più ristretta.

Può risultare utile affrontare, con maggiore dettaglio, un esempio più completo: rilevazione di evidenti valori anomali nei dati, trasformazione della variabile risposta che risolve la varianza non costante ma svela un problema di non-linearità, conseguente trasformaziione anche della variabile esplicativa.

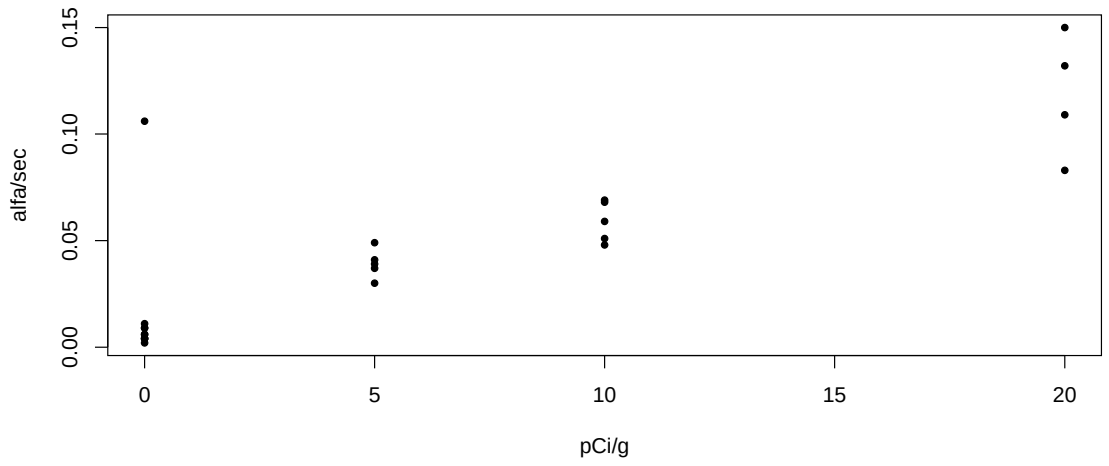


Figura 4.8. Scatter plot della matrice di dati `plutonio`.

Esempio 4.7. La matrice di dati `plutonio`[9] contiene i risultati di uno studio sperimentale teso a stabilire la relazione tra l'emissione di particelle alfa da parte di barrette di plutonio e la loro diversa attività radioattiva. L'attività radioattiva, `x`, è misurata in picocurie per grammo, l'emissione di particelle alfa, `y`, in numero di particelle al secondo. Vi sono barrette di plutonio di quattro tipi ($0, 5, 10$ e 20 picocurie per grammo). Un primo *scatter plot* dei dati (figura 4.8) mostra due aspetti interessanti:

– vi è emissione di particelle alfa anche da parte di barrette con attività radioattiva nulla, quindi il modello deve avere un'intercetta diversa da zero;[10]
– un'osservazione appare anomala, in quanto denota un'emissione di particelle alfa insolitamente alta per una barretta con attività radioattiva nulla.

Una breve indagine consente di appurare che il valore anomalo (osservazione num. 24) dipende da un errore nella registrazione dei dati. Si decide quindi di escluderlo e di procedere con un modello del primo ordine, `y = x`. Si ottiene:

```
                        Analysis of Variance

                                   Sum of           Mean
Source                     DF      Squares         Square    F Value    Pr > F

Model                       1      0.03619        0.03619     229.00    <.0001
Error                      21      0.00332     0.00015804
Corrected Total            22      0.03951

            Root MSE              0.01257    R-Square     0.9160
            Dependent Mean        0.04435    Adj R-Sq     0.9120
            Coeff Var            28.34708

                        Parameter Estimates
```

[9] Tratta da M.H. Kutner, C.J. Nachtsheim, J. Neter e W. Li, *Applied Linear Statistical Models*, McGraw-Hill, 2005, p. 141 (file CH03TA10.TXT scaricabile da `http://www.mhhe.com/kutnerALSM5e`, oppure da `http://web.mclink.it/MC1166/ModelliStatistici/plutonio.csv`).

[10] In caso contrario, si può escludere l'intercetta in R con formule del tipo `y ~ 0 + x`, in SAS con l'opzione `noint`.

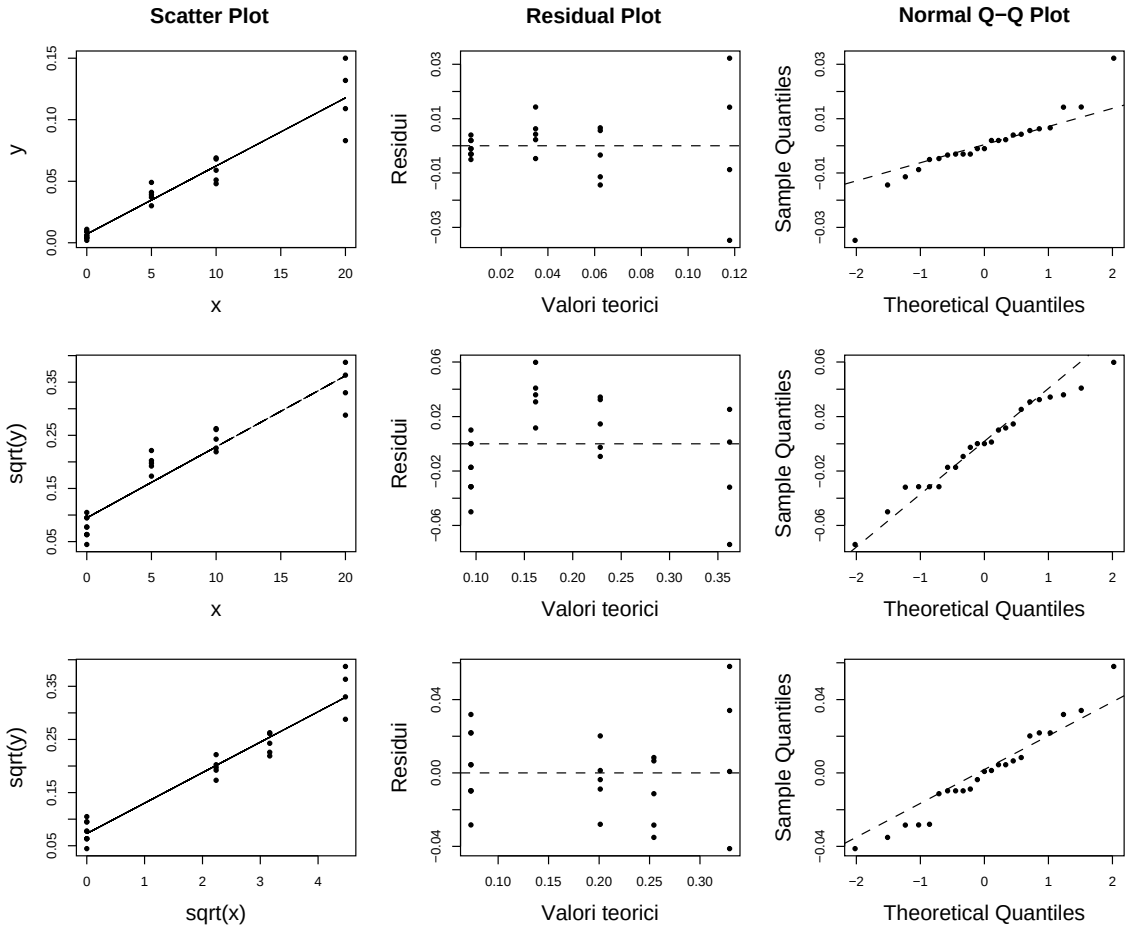


Figura 4.9. Analisi grafica dei modelli $Y_i = \beta_0 + \beta_1 X_i$ (in alto), $\sqrt{Y_i} = \beta_0 + \beta_1 X_i$ (al centro) e $\sqrt{Y_i} = \beta_0 + \beta_1 \sqrt{X_i}$ (in basso) per l'esperimento `plutonio`.

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.00703	0.00360	1.95	0.0641
x	1	0.00554	0.00036590	15.13	<.0001

I risultati sembrano incoraggianti: il test F conferma la significatività del modello, i test t quella del coefficiente β_1 (non altrettanto quella dell'intercetta β_0), un $R^2 = 0.916$ indica un buon adattamento ai dati. L'analisi grafica, tuttavia, mostra chiaramente una varianza dei residui crescente all'aumentare della variabile esplicativa (figura 4.9 in alto). Il *Q-Q plot* mostra un andamento sinusoidale che non si concilia con l'assunto di normalità, ma può risentire della varianza non costante. Si prova quindi a trasformare la variabile risposta sostituendola con la sua radice quadrata (il *residual plot* sembra mostrare una una variabilità crescente con qualche concavità verso il basso). Si ottiene così un R^2 un po' minore, ma una buona significatività anche per l'intercetta; migliorano anche il *residual plot* (maggiore stabilità della varianza) e il *Q-Q plot* mostra punti più vicini alla retta di riferimento (figura 4.9 al centro). Tuttavia, il chiaro andamento curvilineo del *residual plot* mostra che è intervenuto un problema di non-linearità. Notando che appare ancora

una concavità verso il basso, si prova a trasformare anche la variabile esplicativa nella sua radice quadrata, pervenendo al modello $\sqrt{Y_i} = \beta_0 + \beta_1\sqrt{X_i}$. Finalmente i residui appaiono meglio distribuiti e migliora anche il *Q-Q plot*.

4.3 Qualità dei dati

Eventuali valori anomali possono influire pesantemente sui coefficienti di un modello regressivo e vanno pertanto individuati.

Nel caso della variabile risposta, lo strumento più immediato per l'individuazione di valori anomali è l'analisi dei residui. Un residuo osservato, tuttavia, è la determinazione di una variabile aleatoria e, come tale, ha una variabilità che può aumentarne o diminuirne il valore per motivi puramente accidentali; si ricorre quindi ad una standardizzazione dei residui per affinare l'analisi.

Nel caso delle variabili esplicative, se queste sono tre o più risulta arduo individuare eventuali valori anomali mediante l'analisi grafica. Si vedrà che la *matrice hat* $\mathbf{H}$ fornisce uno strumento diagnostico più efficace.

In entrambi i casi, una volta individuati valori anomali si deve capire se, e in che misura, questi influenzano la stima del modello.

4.3.1 Individuazione di valori anomali della variabile risposta

La varianza dei residui, $MSRES$, è uno stimatore della varianza dell'errore σ^2. Si potrebbe quindi pensare di ricorrere ai cosiddetti *residui semistudentizzati*:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSRES}}$$

In realtà, tuttavia, la struttura di varianza e covarianza dei residui dipende dalla matrice di riparametrizzazione e, quindi, $MSRES$ è solo un'approssimazione alla varianza dell'i-esimo residuo.

Un primo miglioramento si ottiene partendo da $\text{Cov}(\mathbf{e}) = (\mathbf{I} - \mathbf{H})\sigma^2$; stimando σ^2 con $MSRES$, la varianza dell'i-esimo residuo è stimata da:

$$\hat{S}^2_{e_i} = MSRES(1 - h_{ii})$$

Dividendo ciascun residuo per la sua varianza stimata si ottengono i cosiddetti *residui standardizzati* (o *residui internamente studentizzati*, indicati da SAS con `student`):

$$r_i = \frac{e_i}{\sqrt{MSRES(1 - h_{ii})}}$$

A rigore, tuttavia, un valore anomalo della variabile risposta può far sì che la funzione di regressione (una retta, una curva, un piano, una superficie ecc.), in quanto calcolata tenendo conto di tutti i valori, "passi vicino" al valore anomalo riducendo il corrispondente residuo. Da altro punto di vista, va detto che il singolo residuo e_i e $MSRES$ non sono indipendenti, quindi r_i non può essere assimilato ad una variabile t di Student.

Si preferisce quindi calcolare l'i-esimo residuo come differenza tra il valore osservato della variabile risposta, y_i, e un valore teorico $\hat{y}_{i(i)}$ calcolato sulla base dei coefficienti stimati escludendo la i-esima osservazione dal dataset; in questo modo la funzione di

regressione non viene influenzata da y_i e, se questo è un valore anomalo, il residuo risulta più netto. La differenza:

$$d_i = y_i - \hat{y}_{i(i)}$$

viene detta *residuo cancellato* (*deleted*).[11]

Non è necessario calcolare tante funzioni di regressione quante sono le osservazioni, eslcudendole una per volta. Si ha infatti che:

$$d_i = \frac{e_i}{1 - h_{ii}}$$

dove e_i e h_{ii} sono, rispettivamente, l'i-esimo residuo e l'i-esimo elemento della diagonale principale della matrice $\mathbf{H}$ del modello stimato con tutte le osservazioni. Si può notare che il residuo cancellato d_i aumenta rispetto al residuo ordinario quando aumenta h_{ii}.

Per stimare la varianza di un residuo cancellato d_i si può procedere come segue. Il valore teorico corrispondente alla i-esima osservazione, la sua varianza e la stima di questa sono:[12]

$$\hat{Y}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}} \qquad \sigma^2_{\hat{Y}_i} = \mathbf{X}'_i \operatorname{Cov}(\hat{\boldsymbol{\beta}})\mathbf{X}_i = \sigma^2\mathbf{X}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i \qquad \hat{S}^2_{\hat{Y}_i} = MSRES[\mathbf{X}'_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i]$$

dove $\mathbf{X}_i$ è la i-esima riga della matrice di riparametrizzazione. Se $\hat{Y}_i$ viene stimato escludendo la i-esima osservazione, si ha:

$$\sigma^2_{\hat{Y}_{i(i)}} = \sigma^2\mathbf{X}'_i\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_i \qquad \hat{S}^2_{\hat{Y}_{i(i)}} = MSRES_{(i)}\left[\mathbf{X}'_i\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_i\right]$$

dove $\mathbf{X}_{(i)}$ è una matrice di riparametrizzazione mancante della i-esima osservazione e $MSRES_{(i)}$ la relativa varianza residua. La varianza stimata di $d_i = Y_i - \hat{Y_{i(i)}}$ è quindi:

$$\hat{S}^2_{d_i} = MSRES_{(i)} + \hat{S}^2_{\hat{Y}_{i(i)}} = MSRES_{(i)}\left[1 + \mathbf{X}'_i\left(\mathbf{X}'_{(i)}\mathbf{X}_{(i)}\right)^{-1}\mathbf{X}_i\right]$$

che si può riscrivere nella forma:

$$\hat{S}^2_{d_i} = \frac{MSRES_{(i)}}{1 - h_{ii}}$$

Poiché né $\mathbf{X}_{(i)}$ né $MSRES_{(i)}$ dipendono dalla i-esima osservazione, si ha che:

$$t_i = \frac{d_i}{\sqrt{\frac{MSRES_{(i)}}{1-h_{ii}}}} = \frac{e_i}{\sqrt{MSRES_{(i)}(1 - h_{ii})}} \sim t_{n-p-1}$$

ovvero che i *residui studentizzati* t_i (detti anche *residui esternamente studentizzati* e indicati da SAS con `rstudent`) si distribuiscono come t di Student con $(n-1)-p = n-p-1$ gradi di libertà.

Anche nel caso di $MSRES_{(i)}$ non è necessario calcolare tante funzioni di regressione quante sono le osservazioni, in quanto vi è una semplice relazione tra $MSRES$ e $MSRES_{(i)}$:

$$(n - p - 1)MSRES_{(i)} = (n - p)MSRES - \frac{e_i^2}{1 - h_{ii}}$$

[11] La somma dei quadrati dei residui cancellati viene detta $PRESS$; cfr. pag. 116.

[12] Cfr. pag. 17, nota 13.

Sostituendo nella relazione precedente, e ricordando che $(n-p)MSRES = SSRES$, si ottiene:

$$t_i = e_i \sqrt{\frac{n-p-1}{(1-h_{ii})SSRES - e_i^2}}$$

Esempio 4.8. Data la matrice `bodyfat`, che contiene $n = 20$ osservazioni, un modello che consideri le sole due prime variabili, `tst` e `tc`, presenta una devianza residua $SSRES = 109.95$ (cfr. esempio 3.14). Una volta elaborato il modello in R:

```
> bodyfat <- read.csv("bodyfat.csv")
> mod <- lm(y ~ tst + tc, data=bodyfat)
```

il valore teorico $\hat{y}_1$, il residuo e_1 e l'elemento della matrice hat h_{11} per la prima osservazione risultano:

```
> mod$fitted.values[1]
       1
13.58271
> mod$residuals[1]
        1
-1.682709
> hatvalues(mod)[1]
        1
0.2010125
```

Il residuo studentizzato è quindi:

```
> -1.682709 * sqrt( (20-3-1) / (109.95*(1-0.201)-(-1.682709)^2) )
[1] -0.729988
```

Per ottenere i residui standardizzata r_i e studentizzati t_i con SAS, si deve indicare un dataset di `output` e precisare le colonne che interessano; ad esempio:

```
proc reg data=bodyfat;
 model y = tst tc;
 output out=bfres p=y_hat r=e_i h=h_ii student=r_i rstudent=t_i;
run;
proc print data=bfres; run;
```

Nel comando `output`, `out=bfres` assegna un nome al dataset, le altre assegnazioni riguardano i valori teorici `p` (abbreviazione di `predicted`, i residui `r` (abbreviazione di `residual`), gli elementi della diagonale della matrice hat `h`, i residui standardizzati `student` e quelli studentizzati `rstudent`. Si ottiene:

Obs	tst	tc	mac	y	y_hat	e_i	r_i	h_ii	t_i
1	19.5	43.1	29.1	11.9	13.5827	-1.68271	-0.74023	0.20101	-0.72999
2	24.7	49.8	28.2	22.8	19.1571	3.64293	1.47658	0.05889	1.53425
3	30.7	51.9	37.0	18.7	21.8760	-3.17597	-1.57579	0.37193	-1.65433
4	29.8	54.3	31.1	20.1	23.2585	-3.15847	-1.31715	0.11094	-1.34848
5	19.1	42.2	30.9	12.9	12.9003	-0.00029	-0.00013	0.24801	-0.00013
6	25.6	53.9	23.7	21.7	22.0608	-0.36082	-0.15199	0.12862	-0.14755
7	31.4	58.5	27.6	27.1	26.3838	0.71620	0.30645	0.15552	0.29813

8	27.9	52.1	30.6	25.4	21.3853	4.01473	1.66061	0.09629	1.76009
9	22.1	49.9	23.2	21.3	18.6449	2.65511	1.10955	0.11464	1.11765
10	25.5	53.5	24.8	19.3	21.7748	-2.47481	-1.03165	0.11024	-1.03373
11	31.1	56.6	30.0	25.4	25.0642	0.33581	0.14078	0.12034	0.13666
12	30.4	56.7	28.3	27.2	24.9745	2.22551	0.92722	0.10927	0.92318
13	18.7	46.5	23.0	11.7	15.6469	-3.94686	-1.71215	0.17838	-1.82590
14	19.7	44.2	28.6	17.8	14.3525	3.44746	1.46861	0.14801	1.52476
15	14.6	42.7	21.3	12.8	12.2294	0.57059	0.27476	0.33321	0.26715
16	29.5	54.4	30.1	23.9	23.2577	0.64230	0.26552	0.09528	0.25813
17	27.7	55.3	25.7	22.6	23.4509	-0.85095	-0.35380	0.10559	-0.34451
18	30.2	58.6	24.6	25.4	26.1829	-0.78292	-0.34350	0.19679	-0.33441
19	22.7	48.2	27.1	14.8	17.6573	-2.85729	-1.16313	0.06695	-1.17617
20	25.2	51.0	27.5	21.1	20.0596	1.04045	0.41976	0.05009	0.40936

Si nota che i valori per la prima osservazione coincidono con quelli calcolati manualmente. Si nota, soprattutto, che i tre residui ordinari maggiori risultano quelli per le osservazioni 2, 8 e 13, ma i tre maggiori residui studentizzati sono quelli per le osservazioni 3, 8 e 13; è quindi in queste ultime che vanno individuati valori anomali della variabile risposta, quelli a proposito dei quali ci sarà da indagare quanto siano influenti.

4.3.2 Individuazione di valori anomali delle variabili esplicative

Gli elementi della diagonale principale della matrice hat, appena visti nel calcolo dei residui studentizzati, hanno proprietà che li rendono utili anche per l'individuazione di valori anomali delle variabili esplicative. Tali elementi h_{ii} sono sempre compresi tra 0 e 1, e la loro somma è sempre uguale a p, il numero di colonne della matrice di riparametrizzazione.

Soprattutto, però, gli elementi h_{ii} costituiscono una misura della distanza tra la i-esima osservazione (il vettore costituito dai valori delle $p-1$ variabili esplicative sulla i-esima riga della matrice dei dati) e il centroide di tutte le osservazioni.

Gli elementi h_{ii} sono anche detti *leverage* perché esprimono l'"effetto leva" della i-esima osservazione sulla funzione di regressione, che si manifesta in due modi:

a) i valori teorici sono funzione lineare dei valori osservati della variabile risposta, $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, e l'elemento h_{ii} è il peso del valore y_i nel determinare il valore teorico $\hat{y}_i$;

b) la varianza dell'i-esimo residuo è $\sigma^2_{e_i} = \sigma^2(1 - h_{ii})$, è quindi tanto minore quando maggiore è h_{ii}; ne segue che $\hat{y}_i$ è tanto più vicino a y_i quanto maggiore è h_{ii} (in altri termini, valori alti di h_{ii} tendono a far passare la funzione di regressione vicino al valore osservato y_i, e questo avviene tanto più quanto più i valori delle variabili esplicative per la i-esima osservazione sono lontani dal centroide).

Un *leverage* è considerato alto se è maggiore del doppio del leverage medio $\bar{h}$:

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}$$

Esempio 4.9. Restando a `bodyfat`, la figura 4.10 mostra un grafico delle due prime variabili esplicative nel quale si nota che due valori, quelli delle osservazioni 3 e 15, sono un po' lontani dalla nuvola di punti formata dagli altri. Una volta elaborato il modello con le due sole variabili, come nell'esempio precedente, si può usare la funzione `hatvalues()` per esaminare gli elementi della diagonale principale della matrice hat:

```
> mod <- lm(y ~ tst + tc, data=bodyfat)
```

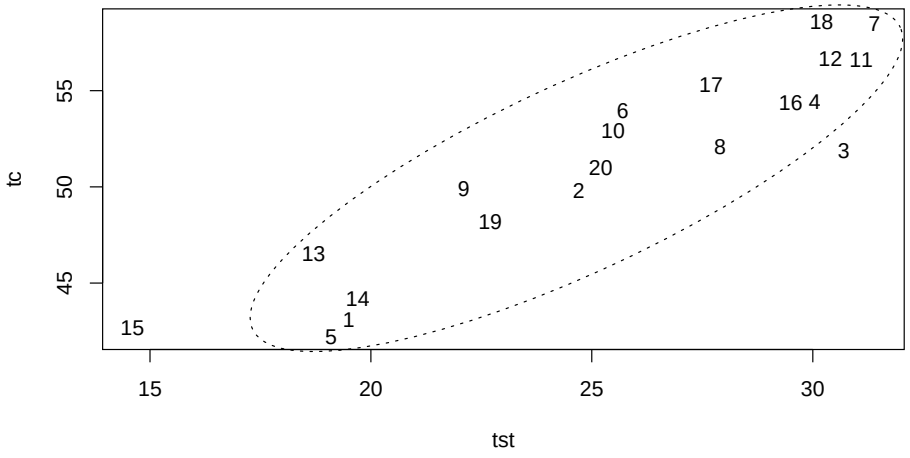


Figura 4.10. Diagramma di dispersione della variabile `tc` (circonferenza coscia) contro la variabile `tst` (plica tricipitale; matrice di dati `bodyfat`).

```
> round(hatvalues(mod),3)
    1     2     3     4     5     6     7     8     9    10
0.201 0.059 0.372 0.111 0.248 0.129 0.156 0.096 0.115 0.110
   11    12    13    14    15    16    17    18    19    20
0.120 0.109 0.178 0.148 0.333 0.095 0.106 0.197 0.067 0.050
```

Si può notare che gli elementi $h_{3,3}$ e $h_{15,15}$ sono in effetti i maggiori; inoltre, sono gli unici superiori a $2p/n = 2 \cdot 3/20 = 0.30$, che è il doppio della media $\bar{h}$. La 3 e la 15 sono quindi le osservazioni anomale, la cui influenza sul modello merita di essere indagata.

4.3.3 Individuazione dei casi influenti

Una volta individuati valori anomali nella variabile risposta o nelle variabili esplicative, si tratta di accertare se essi sono *influenti*, cioè se una loro esclusione dal modello comporterebbe un cambiamento sostanziale nella funzione di regressione.

Si usano allo scopo le *distanze di Cook*, D_i:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{p \cdot MSRES}$$

D_i viene calcolata per ogni osservazione. Al numeratore vi è la somma dei quadrati delle differenze tra i valori teorici calcolati sulla base di tutte le osservazioni e quelli calcolati escludendo la i-esima osservazione. La somma viene standardizzata dividendola per p volte la varianza residua $MSRES$.

Anche le distanze di Cook possono essere calcolate senza elaborare tante funzioni di regressione quante sono le osservazioni, in quanto la seguente espressione è equivalente alla precedente:

$$D_i = \frac{e_i^2}{p \cdot MSRES} \left[\frac{h_{ii}}{(1 - h_{ii})^2} \right]$$

Le distanze calcolate vengono interpretate rapportandole ad una distribuzione $F_{p,n-p}$ e determinando il corrispondente percentile: se questo è minore di 10 o 20, la corrispondente

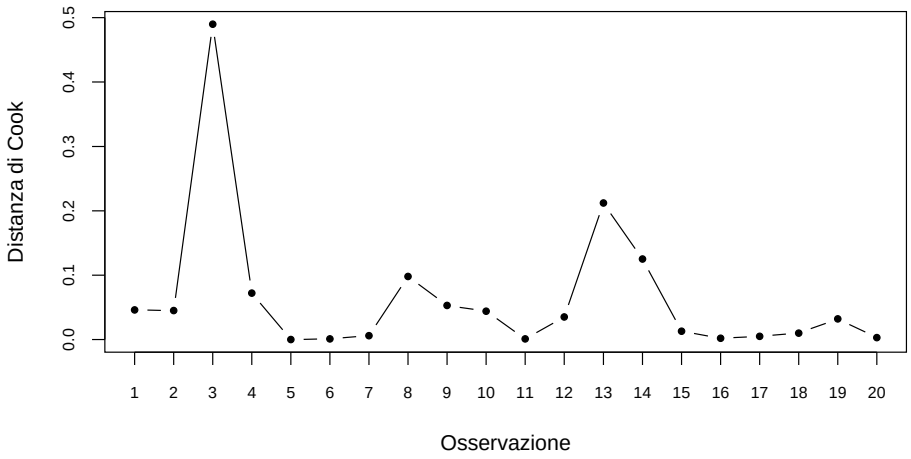


Figura 4.11. Distanze di Cook per le osservazioni di `bodyfat` con due variabili esplicative.

informazione viene considerata poco influente; influente in misura sostanziale, invece, se il percentile è vicino a 50 o maggiore.

Esempio 4.10. Restando ancora a `bodyfat` con due sole variabili esplicative, le distanze di Cook possono essere calcolate in R con la funzione `cook.distance()`:

```
> round(cooks.distance(mod),3)
    1     2     3     4     5     6     7     8     9    10
0.046 0.045 0.490 0.072 0.000 0.001 0.006 0.098 0.053 0.044
   11    12    13    14    15    16    17    18    19    20
0.001 0.035 0.212 0.125 0.013 0.002 0.005 0.010 0.032 0.003
```

In SAS si tratta di aggiungere una colonna `cookd=<nome>` al comando `output`, ad esempio:

```
proc reg data=bodyfat;
model y = tst tc;
output out=bfres p=y_hat r=e_i h=h_ii student=r_i rstudent=t_i cookd=D_i;
run;
```

Le distanze possono essere anche rappresentate in un grafico come quello riprodotto nella figura 4.11. Nell'esempio 4.8 si erano rilevati valori anomali della variabile risposta nelle osservazioni 3, 8 e 13; nell'esempio 4.9 si erano rilevati valori anomali delle variabili esplicative nelle osservazioni 3 e 15. Le distanze di Cook consentono di valutare che l'osservazione 15 è evidentemente ininfluente, mentre potrebbero esserlo la 3 e, in misura minore, la 13 e la 8. Tuttavia, il percentile di $F_{p,n-p} = F_{3,17}$ per 0.490, valore della distanza di Cook per la terza osservazione:

```
> pf(.490, 3, 17)
[1] 0.3061611
```

è il 30.6-esimo, quindi l'influenza dell'osservazione 3 (e a maggior ragione della 13 e della 8) risulta modesta, tanto da non richiedere azioni correttive.

4.3.4 Azioni correttive

Una volta individuati valori anomali influenti, si impone in primo luogo una verifica della corretta registrazione del dato; se questo risulta errato (come nel caso dell'esempio 4.7), può essere rettificato o eliminato.

Quando non è possibile stabilire con certezza la possibilità di rettificare o eliminare dati anomali, si può cercare di intervenire sul modello. Ad esempio, è possibile che le anomalie possano essere eliminate, o almeno ridotte sostanzialmente, mutando la forma funzionale del modello (da quadratica a esponenziale, ecc.), oppure aggiungendo variabili esplicative che si erano in un primo momento escluse. Inoltre, se alcuni valori anomali della variabile risposta sono associati a valori insolitamente alti o bassi di due (o più) variabili esplicative, può risultare opportuno aggiungere al modello un effetto interattivo. Nelle serie storiche, infine, possono essere intervenuti mutamenti strutturali tali da consigliare di esaminare distintamente i dati precedenti e successivi al mutamento. Se gli interventi sul modello non danno esito, si devono tentare altri approcci quali la *regressione robusta*, *non parametrica* ecc.

Appendice A

Complementi di algebra lineare

A.1 Matrici inverse e inverse generalizzate

Come noto, data una matrice quadrata $\mathbf{A}$ di ordine n a rango pieno, si dice sua *inversa* e si indica con $\mathbf{A}^{-1}$ una matrice tale che:

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

La definizione di inversa può essere tuttavia resa più generale e così applicabile anche a matrici non quadrate e/o non di rango pieno.

Definizione A.1. Data una matrice $\underset{m,n}{\mathbf{A}}$, si dicono sua *inversa destra* una matrice $\underset{n,m}{\mathbf{A}}^{-R}$, sua *inversa sinistra* una matrice $\underset{n,m}{\mathbf{A}}^{-L}$ tali che:

$$\underset{m,n}{\mathbf{A}}\ \underset{n,m}{\mathbf{A}}^{-R} = \underset{m,m}{\mathbf{I}} \qquad\qquad \underset{n,m}{\mathbf{A}}^{-L}\ \underset{m,n}{\mathbf{A}} = \underset{n,n}{\mathbf{I}}$$

Osservazione A.2. Un'inversa destra di $\underset{m,n}{\mathbf{A}}$ esiste solo se $m \leq n$ e $\mathrm{rk}(\mathbf{A}) = m$, un'inversa sinistra solo se $n \leq m$ e $\mathrm{rk}(\mathbf{A}) = n$. Ciò in quanto la moltiplicazione di una matrice per un'altra non può aumentarne il rango: $\mathrm{rk}(\mathbf{AB}) \leq \min\{\mathrm{rk}(\mathbf{A}), \mathrm{rk}(\mathbf{B})\}$ (v. proposizione A.32), ma il risultato di una moltiplicazione per un'inversa destra o sinistra è, per definizione, una matrice identità di rango, rispettivamente, m o n. Inoltre, se le inverse destra e sinistra esistono non sono uniche.

Esempio A.3. Date le seguenti tre matrici:

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 1 & 1 \\ -2 & -1 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} -5/18 & 1/9 & -13/18 \\ 1/2 & 0 & 1/2 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} -8/9 & 23/9 & -1/9 \\ 1 & -2 & 0 \end{bmatrix}$$

si verifica facilmente che $\mathbf{B}$ e $\mathbf{C}$ sono entrambe inverse sinistre di $\mathbf{A}$ e che le loro trasposte sono entrambe inverse destre della trasposta di $\mathbf{A}$:

$$\mathbf{BA} = \mathbf{CA} = \mathbf{I}_2 \qquad\qquad \mathbf{A}'\mathbf{B}' = \mathbf{A}'\mathbf{C}' = \mathbf{I}_2$$

Esempio A.4. In generale:

a) data una matrice $\underset{m,n}{\mathbf{A}}$ con $m > n$ e rango $r = n$, la matrice $\left(\underset{n,m}{\mathbf{A}}{}' \underset{m,n}{\mathbf{A}}\right)$ è una matrice simmetrica $n \times n$ di rango n, quindi è invertibile; un'inversa sinistra di $\mathbf{A}$ è:

$$(\underset{n,n}{\mathbf{A}}{}'\mathbf{A})^{-1}\underset{n,m}{\mathbf{A}}{}'$$

in quanto $[(\underset{n,m}{\mathbf{A}}{}'\mathbf{A})^{-1}\mathbf{A}']\underset{m,n}{\mathbf{A}} = \underset{n,n}{\mathbf{I}}$; nell'esempio precedente, infatti, la matrice $\mathbf{B}$ era stata ottenuta proprio in questo modo;

b) analogamente, data una matrice $\underset{m,n}{\mathbf{A}}$ con $m < n$ e rango $r = m$, un'inversa destra sarà $\underset{n,m}{\mathbf{A}}{}'(\underset{m,m}{\mathbf{A}}\mathbf{A}')^{-1}$, in quanto $\underset{m,n}{\mathbf{A}}[\mathbf{A}'(\underset{n,m}{\mathbf{A}}\mathbf{A}')^{-1}] = \underset{m,m}{\mathbf{I}}$.

Teorema A.5. *Se* $\mathbf{A}$ *è una matrice quadrata di rango pieno, le sue inverse destra e sinistra coincidono e sono uniche. La matrice* $\mathbf{A}^{-L} = \mathbf{A}^{-R} = \mathbf{A}^{-1}$ *viene detta l'*inversa *di* $\mathbf{A}$*.*

Definizione A.6. Data una matrice $\underset{m,n}{\mathbf{A}}$, si dice sua *inversa generalizzata* una matrice $\underset{n,m}{\mathbf{A}}{}^{-}$ tale che:

$$\underset{m,n}{\mathbf{A}}\,\underset{n,m}{\mathbf{A}}{}^{-}\,\underset{m,n}{\mathbf{A}} = \underset{m,n}{\mathbf{A}}$$

Osservazione A.7. Se $\mathbf{A}$ ha un'inversa destra o sinistra, questa è anche una sua inversa generalizzata; infatti:

$$\mathbf{A}\mathbf{A}^{-R}\mathbf{A} = \mathbf{I}\mathbf{A} = \mathbf{A} \qquad\qquad \mathbf{A}\mathbf{A}^{-L}\mathbf{A} = \mathbf{A}\mathbf{I} = \mathbf{A}$$

Ne segue che l'inversa generalizzata non è unica, a meno che $\mathbf{A}$ sia quadrata e di rango pieno; in tal caso, infatti, $\mathbf{A}^{-R} = \mathbf{A}^{-L} = \mathbf{A}^{-1}$ e $\mathbf{A}\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}$, oppure, se $\mathbf{A}$ non è quadrata o non è di rango pieno, che l'inversa generalizzata sia tale da soddisfare le proprietà esposte nella definizione che segue.

Definizione A.8. Data una matrice $\underset{m,n}{\mathbf{A}}$, un'inversa generalizzata $\underset{n,m}{\mathbf{A}}{}^{+}$ tale che:

a) $\mathbf{A}\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$;

b) $\mathbf{A}^{+}\mathbf{A}\mathbf{A}^{+} = \mathbf{A}^{+}$;

c) $\mathbf{A}\mathbf{A}^{+} = (\mathbf{A}\mathbf{A}^{+})'$;

d) $\mathbf{A}^{+}\mathbf{A} = (\mathbf{A}^{+}\mathbf{A})'$;

viene detta *pseudoinversa* (*di Moore-Penrose*).

Esempio A.9. La matrice $\mathbf{B}$ dell'esempio A.3 è la presudoinversa della matrice $\mathbf{A}$, come si verifica facilmente. Non lo è invece $\mathbf{C}$, in quanto $\mathbf{AC}$ non è simmetrica.

Una matrice può avere un'inversa destra o sinistra solo se è a rango pieno, ma si dimostra che ogni matrice ha una pseudo inversa di Moore-Penrose e, inoltre, che questa è unica.

Osservazione A.10. Per trovare la pseudoinversa di una matrice si può ricorrere alla *scomposizione ai valori singolari*, mediante la quale la matrice viene scomposta nel prodotto di tre matrici:

$$\underset{m,n}{\mathbf{A}} = \underset{m,m}{\mathbf{U}}\,\underset{m,n}{\mathbf{\Sigma}}\,\underset{n,n}{\mathbf{V}}{}'$$

dove:

– $\mathbf{U}$ è una matrice ortogonale le cui colonne sono autovettori di $\mathbf{AA}'$;
– $\mathbf{\Sigma}$ è una matrice "diagonale" (nel senso che $\sigma_{ij} = 0$ se $i \neq j$) i cui elementi σ_{ii} sono i *valori singolari* di $\mathbf{A}'\mathbf{A}$, cioè le radici quadrate dei suoi autovalori;
– $\mathbf{V}'$ è la trasposta di una matrice ortogonale $\mathbf{V}$ le cui colonne sono autovettori di $\mathbf{A}'\mathbf{A}$.

La pseudoinversa di $\mathbf{\Sigma}$ – in parole povere, la cosa più vicina che si può trovare ad una sua inversa – è una matrice che ha come unici elementi non nulli i reciproci degli elementi non nulli di $\mathbf{\Sigma}$:

$$\underset{m,n}{\mathbf{\Sigma}} = \begin{bmatrix} \begin{bmatrix} \sigma_{11} & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & \sigma_{rr} \end{bmatrix} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \quad \underset{n,m}{\mathbf{\Sigma}^+} = \begin{bmatrix} \begin{bmatrix} 1/\sigma_{11} & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & 1/\sigma_{rr} \end{bmatrix} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}$$

e si ha:

$$\mathbf{\Sigma\Sigma}^+ = \begin{bmatrix} \begin{bmatrix} 1 & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & 1 \end{bmatrix} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}_{m \times m} \quad \mathbf{\Sigma}^+\mathbf{\Sigma} = \begin{bmatrix} \begin{bmatrix} 1 & \dots & \dots \\ \dots & \dots & \dots \\ \dots & \dots & 1 \end{bmatrix} & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{bmatrix}_{n \times n}$$

Si vede facilmente che pre/post moltiplicando $\mathbf{\Sigma}$ per $\mathbf{\Sigma}^+$ si ottengono matrici simmetriche e che $\mathbf{\Sigma\Sigma}^+\mathbf{\Sigma} = \mathbf{\Sigma}$ e $\mathbf{\Sigma}^+\mathbf{\Sigma\Sigma}^+ = \mathbf{\Sigma}^+$. Ricordando che l'inversa di una matrice ortogonale è la sua trasposta, la pseudoinversa di $\mathbf{A} = \mathbf{U\Sigma V}'$ è $\mathbf{A}^+ = \mathbf{V\Sigma}^+\mathbf{U}'$, infatti:

$$\mathbf{AA}^+\mathbf{A} = \mathbf{U\Sigma V}'\mathbf{V\Sigma}^+\mathbf{U}'\mathbf{U\Sigma V}' = \mathbf{U\Sigma\Sigma}^+\mathbf{\Sigma V}' = \mathbf{U\Sigma V}' = \mathbf{A}$$

A.2 Matrici di proiezione

Come noto:

a) dato uno spazio vettoriale V, due suoi sottospazi U e W sono detti *ortogonali* se, comunque presi due vettori $\mathbf{u} \in U$ e $\mathbf{w} \in W$, si ha $\mathbf{u}'\mathbf{w} = \mathbf{w}'\mathbf{u} = 0$;
b) se $V = U \oplus W$, la somma diretta $U \oplus W$ viene detta *scomposizione ortogonale* di V, U viene scritto anche come $W^\perp$ e W come $U^\perp$, U e W vengono detti l'uno il *complemento ortogonale* dell'altro;
c) se U è un sottospazio di $\mathbb{R}^n$, $U \oplus U^\perp = \mathbb{R}^n$;
d) se i vettori di una base di uno spazio vettoriale sono tra loro a due a due ortogonali, la base viene detta *ortogonale*;
e) se i vettori di uba base di uno spazio vettoriale sono tra loro a due a due ortogonali e hanno norma unitaria, la base viene detta *ortonormale*.

Esempio A.11. Prima di procedere, potrebbe essere utile qualche esempio basato sui familiari spazi $\mathbb{R}^n$. Se $U \subset \mathbb{R}^2$ è uno spazio ad una dimensione, può essere l'insieme delle rette proporzionali al vettore unitario $\mathbf{e}_1 = (1, 0)$ (l'asse delle ascisse); il suo complemento ortogonale è il sottospazio W delle rette proporzionali al vettore $\mathbf{e}_2 = (0, 1)$; la somma diretta dei due sottospazi è il piano $\mathbb{R}^2$, con base ortonormale $\{(1, 0), (0, 1)\}$. Analogamente, se $U \subset \mathbb{R}^3$ è uno spazio a due dimensioni con base $\{\mathbf{e}_1 = (1, 0, 0), \mathbf{e}_2 = (0, 1, 0)\}$

può essere visto come il piano xy, i cui punti hanno ascissa x, ordinata y e quota nulla; il suo complemento ortogonale è il sottospazio W delle rette proporzionali al vettore $\mathbf{e}_3 = (0, 0, 1)$; la loro somma diretta è lo spazio tridimensionale $\mathbb{R}^3$ con base ortonormale $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$.

Definizione A.12. Dati lo spazio vettoriale $\mathbb{R}^n$ e una sua scomposizione ortogonale $\mathbb{R}^n = U \oplus U^\perp$, si dice *scomposizione ortogonale* di un vettore $\mathbf{v} \in \mathbb{R}^n$ la sua espressione come somma di due vettori $\mathbf{v}_1 \in U$ e $\mathbf{v}_2 \in U^\perp$:

$$\mathbf{v} = \mathbf{v}_1 + \mathbf{v}_2 \qquad \mathbf{v}_1 \in U,\ \mathbf{v}_2 \in U^\perp$$

Definizione A.13. Dati lo spazio vettoriale $\mathbb{R}^n$ e una sua scomposizione ortogonale $\mathbb{R}^n = U \oplus U^\perp$, si dice *matrice di proiezione* sullo spazio U una matrice quadrata $\mathbf{P}$ tale che:

a) $\mathbf{Pv} \in U$ per ogni $\mathbf{v} \in \mathbb{R}^n$;

b) $\mathbf{Pv} = \mathbf{v}$ per ogni $\mathbf{v} \in U$.

In altri termini, una matrice di proiezione trasforma qualsiasi vettore di $\mathbb{R}^n$ in un vettore di U e lascia immutato un vettore che già appartenga a U. È quadrata in quanto trasforma vettori di $\mathbb{R}^n$ in vettori di $\mathbb{R}^n$.

Osservazione A.14. Dalla definizione di matrice di proiezione segue che $\mathbf{PPv} = \mathbf{Pv}$ (da destra verso sinistra: $\mathbf{Pv}$ trasforma $\mathbf{v}$ in un vettore di U; la successiva moltiplicazione per $\mathbf{P}$ lascia immutato il risultato); segue cioè che una matrice di proiezione è una matrice *idempotente*: $\mathbf{P}^2 = \mathbf{P}$.

Osservazione A.15. La matrice identità $\mathbf{I}$ è chiaramente idempotente. Se $\mathbf{P}$ è una matrice idempotente, è tale anche $\mathbf{I} - \mathbf{P}$. Infatti:

$$(\mathbf{I} - \mathbf{P})^2 = \mathbf{I}^2 - \mathbf{IP} - \mathbf{PI} + \mathbf{P}^2 = \mathbf{I} - 2\mathbf{P} + \mathbf{P} = \mathbf{I} - \mathbf{P}$$

Definizione A.16. Se $\mathbf{P}$ è una matrice di proiezione su $U \subset \mathbb{R}^n$, $\mathbb{R}^n = U \oplus U^\perp$ e se $\mathbf{I} - \mathbf{P}$ è una matrice di proiezione su $u^\perp$, allora $\mathbf{P}$ viene detta *matrice di proiezione ortogonale* su U.

Osservazione A.17. Una matrice di proiezione ortogonale $\mathbf{P}$, oltre ad essere idempotente, è anche simmetrica. Infatti, per qualsiasi $\mathbf{v} \in \mathbb{R}^n = U \oplus U^\perp$, essendo $\mathbf{Pv} \in U$ e $(\mathbf{I} - \mathbf{P})\mathbf{v} \in U^\perp$ si deve avere:

$$(\mathbf{Pv})'(\mathbf{I} - \mathbf{P})\mathbf{v} = \mathbf{v}'\mathbf{P}'(\mathbf{I} - \mathbf{P})\mathbf{v} = 0$$

Potendo $\mathbf{v}$ essere un qualsiasi vettore di $\mathbb{R}^n$, deve risultare:

$$\mathbf{P}'(\mathbf{I} - \mathbf{P}) = \mathbf{P}' - \mathbf{P}'\mathbf{P} = \mathbf{O}$$

Ciò è possibile se e solo se $\mathbf{P}'\mathbf{P} = (\mathbf{P}')^2 = \mathbf{P}'$, cioè se e solo se $\mathbf{P} = \mathbf{P}'$.

Esempio A.18. Sia $\{\mathbf{u}_1 = (1, 0, 0), \mathbf{u}_2 = (1, 1, 0)\}$ una base di $U \subset \mathbb{R}^3$. Le matrici:

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{I} - \mathbf{P} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix}$$

sono entrambe idempotenti. $\mathbf{P}$ proietta qualsiasi vettore di $\mathbb{R}^3$ in U. Ad esempio, se $\mathbf{v} = (2, 1, 1)$, $\mathbf{Pv} = (3, 2, 0)$, che appartiene evidentemente a U: $\mathbf{Pv} = \mathbf{u}_1 + 2\mathbf{u}_2$. $\mathbf{I} - \mathbf{P}$ proietta invece $\mathbf{v}$ in uno spazio che non è ortogonale a U, infatti $(\mathbf{I} - \mathbf{P})\mathbf{v} = (-1, -1, 1)$ e $\mathbf{u}_1'\mathbf{v} = -1$, $\mathbf{u}_2'\mathbf{v} = -2$.

Esempio A.19. Sia $\{\mathbf{u}_1 = (1, 0, 0), \mathbf{u}_2 = (1, 1, 0)\}$ una base di $U \subset \mathbb{R}^3$. Le matrici:

$$\mathbf{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{I} - \mathbf{P} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

sono entrambe simmetriche oltre che idempotenti. $\mathbf{P}$ proietta qualsiasi vettore di $\mathbb{R}^3$ in U. Ad esempio, se $\mathbf{v} = (2, 1, 1)$, $\mathbf{Pv} = (2, 1, 0) = \mathbf{u}_1 + \mathbf{u}_2$. $\mathbf{I} - \mathbf{P}$ proietta $\mathbf{v}$ in uno spazio ortogonale a U, infatti $(\mathbf{I} - \mathbf{P})\mathbf{v} = (0, 0, 1)$ è ortogonale sia a $\mathbf{u}_1$ che a $\mathbf{u}_2$, quindi a tutte le loro combinazioni lineari (a tutti gli elementi di U). $\mathbf{P}$ è quindi una matrice di proiezione ortogonale.

Osservazione A.20. Dati uno spazio vettoriale V ed un suo sottospazio U, esistono molte matrici di proiezione su U, ma una sola matrice di proiezione ortogonale su U; esiste, cioè, una sola matrice di proiezione $\mathbf{P}$ tale che $\mathbf{I} - \mathbf{P}$ sia una matrice di proiezione su $U^\perp$.

Proposizione A.21. *Una matrice idempotente ha come autovalori solo 1 e/o 0.*

Dimostrazione. Sia $\mathbf{A}$ una matrice idempotente e sia $\mathbf{v}$ un vettore di tanti elementi quante sono le colonne di $\mathbf{A}$. Per la definizione di autovalore e autovettore, si ha $\mathbf{Av} = \lambda\mathbf{v}$, ma anche $\mathbf{A}^2\mathbf{v} = \mathbf{A}(\mathbf{Av}) = \mathbf{A}(\lambda\mathbf{v}) = \lambda^2\mathbf{v}$. Essendo $\mathbf{A}$ idempotente:

$$\mathbf{A}^2\mathbf{v} = \mathbf{Av} \;\Rightarrow\; \lambda^2\mathbf{v} = \lambda\mathbf{v} \;\Rightarrow\; (\lambda^2 - \lambda)\mathbf{v} = 0 \;\Rightarrow\; \lambda(\lambda - 1) = 0 \;\Rightarrow\; \lambda \in \{0, 1\} \qquad \square$$

Proposizione A.22. *Il rango di una matrice idempotente è uguale alla sua traccia.*

Dimostrazione. Per la proposizione precedente, una matrice idempotente è simile ad una matrice diagonale avente solo 1 e/o 0 sulla diagonale principale e il cui rango è quindi uguale alla sua traccia, cioè al numero degli 1 sulla diagonale principale. Ma matrici simili hanno la stessa traccia e lo stesso rango, quindi per qualsiasi matrice idempotente il rango è uguale alla traccia. $\square$

A.3 Immagine di una matrice

È noto che una qualsiasi matrice può essere considerata come associata ad un'applicazione lineare e che, quindi, si usa parlare di *immagine* di una matrice; ad esempio, data un'applicazione lineare $L : \mathbb{R}^n \to \mathbb{R}^m$, ad essa può essere associata una matrice $\underset{m,n}{\mathbf{A}}$ tale che, per ogni $\mathbf{v} \in \mathbb{R}^n$, $L(\mathbf{v}) = \mathbf{Av} \in \mathbb{R}^m$. L'immagine di una matrice è quindi l'insieme di tutti i vettori $\mathbf{Av}$, che coincide con l'immagine dell'applicazione associata.

È noto anche che, essendo il prodotto $\mathbf{Av}$ una combinazione lineare delle colonne di $\mathbf{A}$ (di cui gli elementi di $\mathbf{v}$ sono i coefficienti), la dimensione dell'immagine di una matrice è uguale al suo rango e che questo è uguale non solo al numero delle colonne linearmente indipendenti, ma anche al numero delle righe linearmente indipendenti (quindi il rango di una matrice e della sua trasposta sono uguali).

Proposizione A.23. *Data una matrice* $\underset{m,p}{\mathbf{A}} : \underset{m,n-p}{\mathbf{B}}$*, cioè una matrice di* m *righe le cui prime* p *colonne siano costituite dalla matrice* $\mathbf{A}$ *e le restanti* $n-p$ *dalla matrice* $\mathbf{B}$*, si ha:*

$$\operatorname{Im}(\mathbf{A} : \mathbf{B}) = \operatorname{Im}(\mathbf{A}) + \operatorname{Im}(\mathbf{B}) \qquad \dim \operatorname{Im}(\mathbf{A} : \mathbf{B}) \leq \dim \operatorname{Im}(\mathbf{A}) + \dim \operatorname{Im}(\mathbf{B})$$

Dimostrazione. Segue dalla definizione di immagine di una matrice: l'immagine di $\mathbf{A} : \mathbf{B}$ è lo spazio generato dalle sue colonne ed è quindi lo spazio generato dall'unione delle colonne di $\mathbf{A}$ e di quelle di $\mathbf{B}$, è quindi la somma delle immagini delle due matrici sue componenti.

Inoltre, alcune delle $\operatorname{rk}(\mathbf{A})$ colonne linearmente indipendenti di $\mathbf{A}$ potrebbero risultare linearmente dipendenti da alcune delle $\operatorname{rk}(\mathbf{B})$ colonne linearmente indipendenti di $\mathbf{B}$, e viceversa, da cui la disuguaglianza delle dimensioni. □

Proposizione A.24. *Date due matrici* $\mathbf{A}$ *e* $\mathbf{B}$*, l'immagine del prodotto* $\mathbf{AB}$ *è un sottoinsieme dell'immagine di* $\mathbf{A}$*:*

$$\operatorname{Im}(\mathbf{AB}) \subseteq \operatorname{Im}(\mathbf{A})$$

Dimostrazione. $\mathbf{ABv} = \mathbf{A}(\mathbf{Bv}) \subseteq \operatorname{Im}(\mathbf{A})$. □

Proposizione A.25. *L'immagine di una matrice* $\mathbf{A}$ *è uguale all'immagine del suo prodotto per la sua trasposta e sono uguali anche i ranghi.*

$$\operatorname{Im}(\mathbf{AA}') = \operatorname{Im}(\mathbf{A}) \qquad \operatorname{rk}(\mathbf{AA}') = \operatorname{rk}(\mathbf{A})$$

Dimostrazione. Per l'uguaglianza delle immagini si tratta di dimostrare che valgono sia $\operatorname{Im}(\mathbf{A}) \subseteq \operatorname{Im}(\mathbf{AA}')$ che $\operatorname{Im}(\mathbf{AA}') \subseteq \operatorname{Im}(\mathbf{A})$. La seconda inclusione segue dalla proposizione precedente.

Se $\mathbf{v}$ è un vettore appartenente al complemento ortogonale di $\operatorname{Im}(\mathbf{AA}')$, $\mathbf{v}$ appartiene anche al complemento ortogonale di $\operatorname{Im}(\mathbf{A})$:

$$\begin{aligned} \mathbf{v} \in \operatorname{Im}(\mathbf{AA}')^{\perp} \quad &\Rightarrow \quad \mathbf{v}'\mathbf{AA}' = \mathbf{0} \quad \Rightarrow \quad \mathbf{v}'\mathbf{AA}'\mathbf{v} = \mathbf{0} \quad \Rightarrow \quad \|\mathbf{Av}\| = \mathbf{0} \\ &\Rightarrow \quad \mathbf{Av} = \mathbf{0} \quad \Rightarrow \quad \mathbf{v} \in \operatorname{Im}(\mathbf{A})^{\perp} \end{aligned}$$

Ne segue $\operatorname{Im}(\mathbf{AA}')^{\perp} \subseteq \operatorname{Im}(\mathbf{A})^{\perp}$, quindi si ha anche $\operatorname{Im}(\mathbf{A}) \subseteq \operatorname{Im}(\mathbf{AA}')$. L'uguaglianza dei ranghi segue da quella delle immagini. □

Proposizione A.26. *Date due matrici* $\mathbf{A}$ *e* $\mathbf{C}$ *con lo stesso numero di righe,* $\operatorname{Im}(\mathbf{C})$ *è un sottospazio di* $\operatorname{Im}(\mathbf{A})$ *solo se* $\mathbf{C} = \mathbf{AB}$*, dove* $\mathbf{B}$ *sia una matrice moltiplicabile per* $\mathbf{A}$ *e con lo stesso numero di colonne di* $\mathbf{C}$*:*

$$\operatorname{Im}(\underset{m,p}{\mathbf{C}}) \subseteq \operatorname{Im}(\underset{m,n}{\mathbf{A}}) \quad \Rightarrow \quad \underset{m,p}{\mathbf{C}} = \underset{m,n}{\mathbf{A}}\ \underset{n,p}{\mathbf{B}}$$

Dimostrazione. $\operatorname{Im}(\mathbf{C})$ è lo spazio generato dalle colonne di $\mathbf{C}$. Perché questo sia incluso nell'immagine di $\mathbf{A}$, per ciascuna colonna $\mathbf{c}_i$ di $\mathbf{C}$ deve esservi un vettore $\mathbf{b}_i$ tale che $\mathbf{Ab}_i = \mathbf{c}_i$. Quindi $\mathbf{C} = \{\mathbf{c}_1 : \cdots : \mathbf{c}_p\}$ deve essere uguale a $\mathbf{AB}$ con $\mathbf{B} = \{\mathbf{b}_1 : \cdots : \mathbf{b}_p\}$. □

Proposizione A.27. *Date due matrici* $\mathbf{A}$ *e* $\mathbf{B}$*, se* $\operatorname{Im}(\mathbf{B}) \subseteq \operatorname{Im}(\mathbf{A})$ *allora* $\mathbf{AA}^{-}\mathbf{B} = \mathbf{B}$*, quale che sia l'inversa generalizzata di* $\mathbf{A}$*. Analogamente, se* $\operatorname{Im}(\mathbf{B}') \subseteq \operatorname{Im}(\mathbf{A}')$ *allora* $\mathbf{BA}^{-}\mathbf{A} = \mathbf{B}$*.*

Dimostrazione. Se $\mathrm{Im}(\mathbf{B}) \subseteq \mathrm{Im}(\mathbf{A})$, per la proposizione precedente esiste una matrice $\mathbf{M}$ tale che $\mathbf{B} = \mathbf{AM}$, quindi:

$$\mathbf{AA}^{-}\mathbf{B} = \mathbf{AA}^{-}\mathbf{AM} = \mathbf{AM} = \mathbf{B}$$

Se invece $\mathrm{Im}(\mathbf{B}') \subseteq \mathrm{Im}(\mathbf{A}')$, esiste una matrice $\mathbf{N}$ tale che $\mathbf{B}' = \mathbf{A}'\mathbf{N}'$ e $\mathbf{B} = (\mathbf{N}')'(\mathbf{A}')' = \mathbf{NA}$, quindi:

$$\mathbf{BA}^{-}\mathbf{A} = \mathbf{NAA}^{-}\mathbf{A} = \mathbf{NA} = \mathbf{B}$$ □

Proposizione A.28. *Date tre matrici* $\mathbf{A}, \mathbf{B}, \mathbf{C}$*, si ha* $\mathrm{Im}(\mathbf{B}') \subseteq \mathrm{Im}(\mathbf{A}')$ *e* $\mathrm{Im}(\mathbf{C}) \subseteq \mathrm{Im}(\mathbf{A})$ *se e solo se* $\mathbf{BA}^{-}\mathbf{C}$ *è invariante rispetto alla scelta dell'inversa generalizzata di* $\mathbf{A}$*.*

Dimostrazione. Se $\mathrm{Im}(\mathbf{B}') \subseteq \mathrm{Im}(\mathbf{A}')$ e $\mathrm{Im}(\mathbf{C}) \subseteq \mathrm{Im}(\mathbf{A})$, allora per la proposizione A.26 esistono due matrici $\mathbf{M}$ e $\mathbf{N}$ tali che $\mathbf{B} = \mathbf{NA}$ e $\mathbf{C} = \mathbf{AM}$. Se $\mathbf{A}_1^{-}$ e $\mathbf{A}_2^{-}$ sono due inverse generalizzate di $\mathbf{A}$, si ha:

$$\begin{aligned}\mathbf{BA}_1^{-}\mathbf{C} - \mathbf{BA}_2^{-}\mathbf{C} &= \mathbf{NAA}_1^{-}\mathbf{AM} - \mathbf{NAA}_2^{-}\mathbf{AM} = \mathbf{N}(\mathbf{AA}_1^{-}\mathbf{A} - \mathbf{AA}_2^{-}\mathbf{A})\mathbf{M} \\ &= \mathbf{N}(\mathbf{A} - \mathbf{A})\mathbf{M} = \mathbf{O}\end{aligned}$$

Si può dimostrare anche l'implicazione inversa. □

Proposizione A.29. *Il prodotto di due matrici* $\mathbf{A}$ *e* $\mathbf{B}$ *è nullo se e solo se l'immagine dell'una è inclusa nel complemento ortogonale dell'immagine dell'altra:*

$$\mathrm{Im}(\mathbf{B}'\mathbf{A}) = \mathbf{O} \quad \Leftrightarrow \quad \mathrm{Im}(\mathbf{B}) \subseteq \mathrm{Im}(\mathbf{A})^{\perp}$$

Dimostrazione. Se $\mathbf{v}$ è un elemento dell'immagine di $\mathbf{B}$, esiste un vettore $\mathbf{u}$ tale che $\mathbf{Bu} = \mathbf{v}$; se $\mathbf{w}$ è un elemento dell'immagine di $\mathbf{A}$, esiste un vettore $\mathbf{x}$ tale che $\mathbf{Ax} = \mathbf{w}$ e si ha:

$$\mathbf{v}'\mathbf{w} = \mathbf{u}'\mathbf{B}'\mathbf{Ax} = 0$$

ovvero $\mathbf{v} \in \mathrm{Im}(\mathbf{A})^{\perp}$. □

Proposizione A.30. *Se una matrice* $\mathbf{A}$ *ha* m *righe, allora la dimensione dell'immagine del suo complemento ortogonale è* $m - \mathrm{rk}(\mathbf{A})$*.*

Dimostrazione. Si può vedere $\mathbf{A}$ come associata all'applicazione $L : \mathbb{R}^n \to \mathbb{R}^m$. L'immagine di $\mathbf{A}$ è un sottospazio di $\mathbb{R}^m$ di dimensione pari al rango di $\mathbf{A}$; essendo $\mathbb{R}^m = \mathrm{Im}(\mathbf{A}) \oplus \mathrm{Im}(\mathbf{A})^{\perp}$, la dimensione di $\mathrm{Im}(\mathbf{A})^{\perp}$ è $m - \mathrm{rk}(\mathbf{A})$. □

Proposizione A.31. *Date due matrici* $\mathbf{A}$ *e* $\mathbf{B}$*, se* $\mathrm{Im}(\mathbf{A}) \subseteq \mathrm{Im}(\mathbf{B})$ *e* $\mathrm{rk}(\mathbf{A}) = \mathrm{rk}(\mathbf{B})$ *allora* $\mathrm{Im}(\mathbf{A}) = \mathrm{Im}(\mathbf{B})$*.*

Dimostrazione. Se ciascun elemento di $\mathbf{A}$ è anche elemento di $\mathbf{B}$, ciò vale anche per gli elementi delle basi; poiché l'uguaglianza dei ranghi implica l'uguaglianza delle dimensioni, quindi delle numerosità delle basi, le due immagini hanno le stesse basi, quindi sono uguali. □

Proposizione A.32. *Date due matrici* $\mathbf{A}$ *e* $\mathbf{B}$*,* $\mathrm{rk}(\mathbf{AB}) \leq \min\{\mathrm{rk}(\mathbf{A}), \mathrm{rk}(\mathbf{B})\}$*.*

Dimostrazione. Per la proposizione A.24, $\mathrm{Im}(\mathbf{AB}) \leq \mathrm{Im}(\mathbf{A})$, quindi $\mathrm{rk}(\mathbf{AB}) \leq \mathrm{rk}(\mathbf{A})$ e, analogamente, $\mathrm{rk}(\mathbf{AB}) = \mathrm{rk}(\mathbf{B}'\mathbf{A}') \leq \mathrm{rk}(\mathbf{B}') = \mathrm{rk}(\mathbf{B})$. □

Proposizione A.33. *Date due matrici* $\mathbf{A}$ *e* $\mathbf{B}$, $\text{rk}(\mathbf{A}+\mathbf{B}) \leq \text{rk}(\mathbf{A}) + \text{rk}(\mathbf{B})$.

Dimostrazione. Si ha:

$$\text{rk}(\mathbf{A}+\mathbf{B}) \leq \text{rk}(\mathbf{A}:\mathbf{B}) \leq \text{rk}(\mathbf{A}) + \text{rk}(\mathbf{B})$$

La prima disuguaglianza vale in quanto $\mathbf{A}+\mathbf{B}$ ha un numero di colonne pari alla metà di quello di $\mathbf{A}:\mathbf{B}$, la seconda per la proposizione A.23. □

Segue un risultato di particolare interesse per i modelli lineari.

A.4 Proiezione ortogonale sull'immagine di una matrice

Proposizione A.34. *Data una matrice* $\mathbf{A}$, *la matrice* $\mathbf{A}\mathbf{A}^-$ *è una matrice di proiezione su* $\text{Im}(\mathbf{A})$. *Inoltre, la matrice di proiezione ortogonale su* $\text{Im}(\mathbf{A})$ *è* $\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$.

Dimostrazione. Sia $\mathbf{A}$ una matrice $n \times p$. $\underset{n,p}{\mathbf{A}}\ \underset{p,n}{\mathbf{A}^-}$ è una matrice di proiezione su $\text{Im}(\mathbf{A}) \subseteq \mathbb{R}^n$ in quanto:

a) dato un vettore $\mathbf{v}$, per la proposizione A.24 $\text{Im}(\mathbf{A}\mathbf{A}^-) \subseteq \text{Im}(\mathbf{A})$, quindi:

$$(\mathbf{A}\mathbf{A}^-)\mathbf{v} \in \text{Im}(\mathbf{A})$$

b) se $\mathbf{v}$ stesso appartiene a $\text{Im}(\mathbf{A})$, esiste un $\mathbf{x}$ tale che $\mathbf{v} = \mathbf{A}\mathbf{x}$, quindi:

$$(\mathbf{A}\mathbf{A}^-)\mathbf{v} = \mathbf{A}\mathbf{A}^-\mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{x} = \mathbf{v}$$

Quanto a $\underset{n,p}{\mathbf{A}}(\underset{p,n}{\mathbf{A}'}\ \underset{n,p}{\mathbf{A}})^-\underset{p,n}{\mathbf{A}'}$, per la proposizione A.25 e per la simmetria di $\mathbf{A}'\mathbf{A}$:

$$\text{Im}(\mathbf{A}') = \text{Im}(\mathbf{A}'\mathbf{A}) = \text{Im}[(\mathbf{A}'\mathbf{A})']$$

e, per la proposizione A.27:

$$\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{A} = \mathbf{A}$$

Quindi $(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ è un'inversa generalizzata di $\mathbf{A}$ e $\mathbf{P} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'$ è una matrice di proiezione. Per un qualsiasi vettore $\mathbf{v} \in \text{Im}(\mathbf{A}) \subseteq \mathbb{R}^n$ esiste un $\mathbf{x} \in \mathbb{R}^p$ tale che $\mathbf{A}\mathbf{x} = \mathbf{v}$; se $\mathbf{y} \in \text{Im}(\mathbf{A})^\perp$, $\mathbf{v}'\mathbf{y} = (\mathbf{A}\mathbf{x})'\mathbf{y} = \mathbf{x}'\mathbf{A}'\mathbf{y} = 0$, ovvero $\mathbf{A}'\mathbf{y} = \mathbf{0}$, quindi:

$$\mathbf{P}\mathbf{y} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}'\mathbf{y} = \mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{0} = \mathbf{0} \qquad (\mathbf{I}_n - \mathbf{P})\mathbf{y} = \mathbf{y}$$

Inoltre, per qualsiasi vettore $\mathbf{v} \in \mathbb{R}^n$ si ha, ancora per la proposizione A.27:

$$\mathbf{A}'(\mathbf{I}_n - \mathbf{P})\mathbf{v} = [\mathbf{A}' - \mathbf{A}'\mathbf{A}(\mathbf{A}'\mathbf{A})^-\mathbf{A}']\mathbf{v} = [\mathbf{A}' - \mathbf{A}']\mathbf{v} = \mathbf{0} \quad \Rightarrow \quad (\mathbf{I}_n - \mathbf{P})\mathbf{v} \in \text{Im}(\mathbf{A})^\perp$$

Quindi $\mathbf{P}$ è la matrice di proiezione ortogonale su $\in (\mathbf{A})$. □

Se $\mathbf{A}$ è una matrice di riparametrizzazione a rango pieno, $\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'$ è la matrice hat $\mathbf{H}$, che è appunto la matrice di proiezione ortogonale su $\text{Im}(\mathbf{A})$.