Contents

List of Figures

(4 maggio 2020, versione ancora da riaggiornare, considerate questo materiale un promemoria per alcune formule e dimostrazioni e un supporto alle lezioni, che quest'anno peraltro sono in buona parte registrate e restano in piattaforma)

Consultare il materiale dei notebook, degli esercizi e laboratorio

Per tenere aggiornato il software didattico del mio package MLANP, potete installarlo dal mio repository github. In Rstudio installate devtools e poi

devtools::install_github("marcellochiodi/packages")

1 Regressione Multipla

2 Introduzione

Nelle lezioni precedenti si è vista la teoria generale sull'inferenza nei modelli lineari: ci siamo concentrati sui casi in cui $\mathbf{X}$ è di *rango pieno*, rimandando ai capitoli sull'analisi della varianza l'analisi dei modelli con matrice $\mathbf{X}$ *a rango non pieno*; abbiamo anche parlato dei modelli con vincoli sui parametri e modelli senza vincoli sui parametri: il risultato fondamentale è che se abbiamo due modelli (uno con vincoli e uno senza), si può costruire un test basato sull'incremento di devianza residua dovuto alla presenza dei vincoli.

Precedentemente avevamo visto come in realtà i modelli lineari siano utilizzabili per diversi problemi statistici, in funzione della particolare costruzione e configurazione della matrice $\mathbf{X}$; in questa sezione affrontiamo il caso specifico dei modelli di regressione, e le peculiarità dell'inferenza in questo caso, insieme con una selezione dei problemi inferenziali più comunemente affrontati nelle applicazioni reali. Nella pratica dello statistico le tecniche di regressione lineare multipla costituiscono una tecnica di base che si applica in numerosi problemi con variabili quantitative

almeno come tecnica preliminare di esplorazione dei dati.

3 Scomposizione della devianza empirica col termine noto e k regressori a media nulla:

Se la matrice $\mathbf{X}$ prevede una colonna di costanti uguali ad uno e altre k colonne a media nulla, in modo che $\mathbf{X} = [\mathbf{1_n}|\mathbf{Z}]$, abbiamo un modello di regressione con termine noto e con matrice $\mathbf{X}^\mathsf{T}\mathbf{X}$ partizionata a due blocchi diagonali:

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{pmatrix} n & \mathbf{0}_k^\mathsf{T} \\ \mathbf{0}_k & \mathbf{Z}^\mathsf{T}\mathbf{Z} \end{pmatrix} \quad \text{o:} \quad \mathbf{X}^\mathsf{T}\mathbf{X} = \begin{pmatrix} n & \mathbf{0}_k^\mathsf{T} \\ \mathbf{0}_k & nS_\mathbf{X} \end{pmatrix} \quad \text{dato che:} \quad \mathbf{Z}^\mathsf{T}\mathbf{Z} = nS_\mathbf{X}$$

Quindi tutte le forme quadratiche che hanno come matrice dei coefficienti questa matrice con $(k+1)\times(k+1)$ elementi, saranno scomponibili in una forma quadratica con matrice di $k \times k$ elementi, ed un termine singolo.

Indichiamo ora il termine noto con α, ed il corrispondente stimatore con a, invece che con $\boldsymbol{\beta}_0$ per evitare confusione con i valori $\boldsymbol{\beta}_0$ dell'ipotesi nulla; con $\boldsymbol{\beta}$ indico il vettore dei parametri relativo alle k variabili e con $\mathbf{b}$ il corrispondente stimatore dei minimi quadrati

Chiaramente risulta: $a = M_{\mathbf{y}} = \Sigma_{i=1}^{n} y_i$

Per quanto riguarda la scomposizione della devianza empirica di $\mathbf{y}$ nel modello di regressione multipla, possiamo partire dalla relazione trovata fra $R(\mathbf{b})$ e la somma dei quadrati $\mathbf{y}^\mathsf{T}\mathbf{y}$: [1]

$$\begin{aligned} R(\mathbf{b}) &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \\ &= (\mathbf{y} - \hat{\mathbf{y}})^\mathsf{T}(\mathbf{y} - \hat{\mathbf{y}}) = (\mathbf{y} - M_{\mathbf{y}}\mathbf{1_n} - \mathbf{Zb})^\mathsf{T}(\mathbf{y} - M_{\mathbf{y}}\mathbf{1_n} - \mathbf{Zb}) = \\ &= \mathbf{y}^\mathsf{T}\mathbf{y} - nM_{\mathbf{y}}^2 - \mathbf{b}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Zb}; \end{aligned}$$

dato che $a = M_{\mathbf{y}}$.

Possiamo anche scrivere:

$$R(\mathbf{b}) = (\mathbf{y} - \mathbf{1_n}M_{\mathbf{y}})^\mathsf{T}(\mathbf{y} - \mathbf{1_n}M_{\mathbf{y}}) - \mathbf{b}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Zb}.$$

Quindi nei modelli di regressione multipla, per eliminare l'influenza del termine noto α, che svolge il ruolo di parametro di disturbo, si può direttamente lavorare in termini di scarti, sia per le X che per $\mathbf{y}$.

[1] in effetti adesso dovremmo indicarlo con $R(a, \mathbf{b})$

In ogni caso sarà possibile fare inferenza indipendente su questo termine: nella tavola che segue, $\mathbf{Z}$ è la matrice degli scarti dalle medie delle X:

TOTALE	RESIDUA	SPIEGATA
$(\mathbf{y} - \mathbf{1_n} M_\mathbf{y})^\intercal (\mathbf{y} - \mathbf{1_n} M_\mathbf{y})$	$(\mathbf{y} - \mathbf{1_n} M_\mathbf{y} - \mathbf{Zb})^\intercal (\mathbf{y} - \mathbf{1_n} M_\mathbf{y} - \mathbf{Zb})$	$\mathbf{b}^\intercal \mathbf{Z}^\intercal \mathbf{Zb}$
$\sum_{i=1}^n (\mathbf{y}_i - M_\mathbf{y})^2$	$\sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)^2$	$\sum_{i=1}^n (\hat{\mathbf{y}}_i - M_\mathbf{y})^2$
Devianza totale osservata di $\mathbf{y}$	devianza residua (deviazioni dal valore stimato)	Devianza spiegata dalla regressione lineare sui k regressori (presi globalmente)
gradi di libertà:		
$n-1$	$n-k-1$	k

Table 1: TAVOLA DI SCOMPOSIZIONE DELLA DEVIANZA EMPIRICA NELLA REGRESSIONE MULTIPLA

4 Il coefficiente di determinazione lineare multipla R^2

E' utile anche da un punto di vista descrittivo, utilizzare un indice impiegato anche per le distribuzioni condizionate di vettori aleatori normali.

La bontà della regressione lineare sui k regressori X_j $(j = 1, 2, \ldots, k)$ per spiegare la variabilità della $\mathbf{y}$ può essere misurata dall'indice (compreso fra 0 e 1):

$$\mathbf{R}^2_{\mathbf{y}.12\ldots k} = \frac{\text{DEVIANZA SPIEGATA DALLA REGRESSIONE}}{\text{DEVIANZA TOTALE}} = \\ = 1 - \frac{\text{DEVIANZA RESIDUA}}{\text{DEVIANZA TOTALE}}$$

Se $k = 1$ è ovvio che $R^2_{\mathbf{y}.1} = r^2$

Si può eventualmente calcolare R^2 mediante la formula vista per le distribuzioni condizionate di vettori aleatori normali. link con Notebook R su modelli lineari ; link con Notebook R su significato degli elementi dell'inversa della

$$\mathbf{R}^2_{\mathbf{y}.12\ldots k} = 1 - \frac{1}{\sigma_i^2 cyy}]$$

Dove con *cyy* ho indicato l'elemento diagonale dell'inversa della matrice di correlazione fra tutte le variabili (cioè Y e tutte le X) corrispondente a Y (vedere gli esempi)

Evidentemente possiamo anche utilizzare il complemento ad 1 per misurare l'incidenza del residuo sul totale:

$$1 - \mathbf{R}^2_{\mathbf{y}.12\ldots k} = \frac{\text{DEVIANZA RESIDUA}}{\text{DEVIANZA TOTALE}}$$

Il valore di questa quantità fornisce la porzione di variabilità di $\mathbf{y}$ che *non è spiegata dalla regressione* sulle k variabili.

5 Scomposizione della devianza teorica nella regressione multipla

Scomponiamo ora la devianza teorica:

Si riveda eventualmente la parte relativa alla stima dei parametri con questa particolare matrice $\mathbf{X}$

$$\begin{aligned}\frac{\boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{\varepsilon}}{\sigma^2} &= = R(\mathbf{b})/\sigma^2 + (a-\alpha)n(a-\alpha)/\sigma^2 + (\mathbf{b}-\boldsymbol{\beta})^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}(\mathbf{b}-\boldsymbol{\beta})/\sigma^2 = \\ &= R(\mathbf{b})/\sigma^2 + (M_{\mathbf{y}}-\alpha)^2/(\sigma^2/n) + (\mathbf{b}-\boldsymbol{\beta})^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}(\mathbf{b}-\boldsymbol{\beta})/\sigma^2.\end{aligned}$$

Analogamente a quanto visto in link con cochran , palesemente vale ancora il teorema di Cochran, per la scomposizione in tre parti della devianza complessiva: link con cochran il nuovo termine $\boldsymbol{\alpha})^2/(\sigma^2/n)$ si distribuisce come una χ^2_1, e per il teorema di Cochran risulta indipendente dalle altre due forme quadratiche.

Si ha, considerando quindi il termine noto:

$$R(\alpha,\boldsymbol{\beta}) = R(a,\mathbf{b}) + (\mathbf{b}-\boldsymbol{\beta})^T(\mathbf{Z}^T\mathbf{Z})(\mathbf{b}-\boldsymbol{\beta}) + n(M_{\mathbf{y}}-\alpha)^2$$

oppure

$$\sum_{i=1}^{n}[\mathbf{y}_i - \mathrm{E}[\mathbf{y}_i]]^2 = (\mathbf{y}-\mathbf{1_n}\alpha-\mathbf{Z}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y}-\mathbf{1_n}\alpha-\mathbf{Z}\boldsymbol{\beta})$$

$$
\begin{aligned}
&= (\mathbf{y} - \mathbf{1_n}\alpha - \mathbf{Z}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{1_n}\alpha - \mathbf{Z}\boldsymbol{\beta}) = \\
&= (\mathbf{y} - \mathbf{1_n}M_{\mathbf{y}} - \mathbf{Z}\mathbf{b})^{\mathsf{T}}(\mathbf{y} - \mathbf{1_n}M_{\mathbf{y}} - \mathbf{Z}\mathbf{b}) + \\
&+ (\mathbf{b} - \boldsymbol{\beta})^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}(\mathbf{b} - \boldsymbol{\beta}) + n(M_{\mathbf{y}} - \alpha)^2
\end{aligned}
$$

(rispetto al simbolismo adottato precedentemente si consideri che adesso il valore atteso è: $\mathrm{E}[\mathbf{Y}] = \mathbf{1_n}\alpha.1 + \mathbf{Z}\boldsymbol{\beta}$)

Possiamo rivedere questa relazione in termini di contributi alla devianza teorica di $\boldsymbol{\varepsilon}$:

Forma Quadratica	fonte	gradi di libertà
$(\mathbf{y}-\mathbf{1_n}\alpha-\mathbf{Z}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{1_n}\alpha-\mathbf{Z}\boldsymbol{\beta})$	devianza teorica complessiva di ε. (rispetto al modello vero)	n
$(\mathbf{y}-\mathbf{1_n}M_\mathbf{y}-\mathbf{Zb})^\top(\mathbf{y}-\mathbf{1_n}M_\mathbf{y}-\mathbf{Zb})$	devianza residua	$n-k-1$
$(\mathbf{b}-\boldsymbol{\beta})^\top\mathbf{Z}^\top\mathbf{Z}(\mathbf{b}-\boldsymbol{\beta})$	devianza delle stime dei coefficienti di regressione	k
$n(M_\mathbf{y}-\alpha)^2$	devianza dovuta alla stima del termine noto	1

6 Prova dell'ipotesi di coefficienti di regressione nulli nella regressione multipla.

Dai risultati visti in precedenza e che scaturiscono sostanzialmente dall'ortogonalità fra termine noto e regressori, risulta immediato il test per saggiare l'ipotesi nulla:

$$H_0 : \boldsymbol{\beta} = 0_k,$$

con α qualsiasi contro l'alternativa generica:

$$H_1 : \boldsymbol{\beta} \neq 0_{k.}$$

Si può infatti impiegare la quantità test:

$$F = \frac{\mathbf{b}'\mathbf{Z}^\top\mathbf{Z}\mathbf{b}}{ks^2}$$

che sotto H_0 si distribuisce secondo una variabile aleatoria F di Snedecor con k ed $n-k-1$ gradi di libertà.

Avendo indicato al solito con s^2 la stima corretta della varianza, con $n-k-1$ gradi di libertà, data da:

$$s^2 = \frac{(\mathbf{y} - \mathbf{1_n}M_\mathbf{y} - \mathbf{Z}\mathbf{b})^\top(\mathbf{y} - \mathbf{1_n}M_\mathbf{y} - \mathbf{Z}\mathbf{b})}{n-k-1} = \frac{\Sigma_{i=1}^n(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2}{n-k-1}$$

E' facile vedere che, dal momento che in fondo il test è dato da:

$$F = \frac{\frac{Devianzaspiegata}{k}}{\frac{Devianzaresidua}{n-k-1}}$$

si può esprimere il test in funzione di R^2:

$$F = \frac{\frac{\mathbf{R}^2_{\mathbf{y}.12\ldots k}}{k}}{\frac{1-\mathbf{R}^2_{\mathbf{y}.12\ldots k}}{n-k-1}}$$

Per saggiare ipotesi particolari, del tipo:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$$

, con α qualsiasi si impiegherà ovviamente il test:

$$F = \frac{[\mathbf{b} - \boldsymbol{\beta}_0]^\top \mathbf{Z}^\top \mathbf{Z} [\mathbf{b} - \boldsymbol{\beta}_0]}{k s^2}$$

Tali ipotesi per valori generici $\boldsymbol{\beta}_0$ sono comunque rare nelle applicazioni della regressione multipla, nelle quali usualmente si è interessati al caso $\boldsymbol{\beta}_0 = \mathbf{0}$:

6.1 La regione di rifiuto.

La regione di rifiuto sarà ovviamente costituita dai valori elevati di F, superiori ad $F_{\alpha,k,n-k-1}$ (ossia situati sulla coda destra della corrispondente variabile F di Snedecor). Valori osservati di F elevati danno evidenza contraria ad H_0. Infatti sotto H_1 il valore atteso di $(\mathbf{b})^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}(\mathbf{b})$ nel numeratore del test F per saggiare l'ipotesi $\boldsymbol{\beta} = 0_k$, è dato, dalle formule precedenti, da:

$$\mathrm{E}\left[\mathbf{b}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\mathbf{b}|H_1\right] = k\sigma^2 + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\boldsymbol{\beta}.$$

mentre

$$\mathrm{E}\left[\mathbf{b}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\mathbf{b}|H_0\right] = k\sigma^2$$

Risulta sempre (al solito):

$$\mathrm{E}\left[\mathbf{b}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\mathbf{b}|H_1\right] \geq \mathrm{E}\left[\mathbf{b}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\mathbf{b}|H_0\right]$$

perchè nella forma quadratica $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\boldsymbol{\beta}$, $\mathbf{Z}^{\mathsf{T}}\mathbf{Z}$ è definita positiva; in ogni caso si vede subito che $\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}^{\mathsf{T}}\mathbf{Z}\boldsymbol{\beta}$=$(\mathbf{Z}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{Z}\boldsymbol{\beta})$, che è palesemente una somma di quadrati.

6.2 Verifica di ipotesi particolari nella regressione multipla.

Si può essere interessati ad una particolare ipotesi, quale un vincolo lineare sui coefficienti di regressione, oppure il fatto che, semplicemente, alcuni dei coefficienti di regressione siano nulli e quindi, che i corrispondenti regressori $\mathbf{X}_j$ siano ininfluenti ai fini della spiegazione di $\mathbf{y}$.

Si può seguire la metodologia generale vista precedentemente: si badi però che quella tecnica è soddisfacente solo se applicata:

- per una ipotesi soltanto *a priori*oppure
- per più ipotesi relative a regressori ortogonali a gruppi.

L' ipotesi:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0 \qquad \text{con} \qquad q < k;$$

e

$$\beta_{q+1}, \beta_{q+2}, \ldots, \beta_k \qquad \text{qualsiasi}$$

(che corrisponde a q vincoli definiti da $\mathbf{C} = [\mathbf{I}_q : 0_{k-q}]\,; \mathbf{a} = 0_q$) stabilisce che q coefficienti di regressione siano nulli e quindi stabilisce che i corrispondenti q

regressori siano eliminabili dal modello generale di spiegazione della variabile di risposta. Possiamo effettuare il test generale:

$$F = \frac{\frac{[R(\mathbf{b}_0)-R(\mathbf{b})]}{q}}{\frac{R(\mathbf{b})}{n-k-1}} = \frac{\frac{[\mathbf{b}-\boldsymbol{\beta}_0]_q^T[(\mathbf{Z}^\top\mathbf{Z})^{-1}{}_q]^{-1}[\mathbf{b}-\boldsymbol{\beta}_0]_q}{q}}{\frac{[\mathbf{y}-\mathbf{X}\mathbf{b}]^\top[\mathbf{y}-\mathbf{X}\mathbf{b}]}{n-k-1}}$$

in cui $\mathbf{b}_0$ è lo stimatore di massima verosimiglianza di $\boldsymbol{\beta}$ sotto H_0 (quindi ha q elementi uguali a zero se $H_0 : [\boldsymbol{\beta}_0]_q = 0$).

In effetti si vede facilmente che il test è ora dato da:

$$F = \frac{[\mathbf{b}]_q^T[(\frac{\mathbf{Z}^\top\mathbf{Z}}{n})^{-1}{}_q]^{-1}[\mathbf{b}]_q}{q\frac{s^2}{n}}$$

ove ancora $\mathbf{b}_q^\top$ indica il vettore di q elementi coinvolto dall'ipotesi nulla e $\left\{\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}\right\}_q$ indica il blocco $q \times q$ di $\left(\mathbf{Z}^\top\mathbf{Z}\right)^{-1}$

in cui è esplicito il fatto che la quantità a numeratore misura la distanza da zero di un particolare sottoinsieme di stimatori di coefficienti di regressione.

Ovviamente si distribuisce come una F con q e $n-k-1$ gradi di libertà.

(La dimostrazione completa di quest'ultima eguaglianza è in altra sezione e comunque non essenziale ai fini della preparazione) link con cap.precedente

6.3 Test per un singolo coefficiente (uno solo!)

link con notebook R Nel caso particolare in cui $q = 1$, evidentemente stiamo saggiando l'ipotesi che un singolo coefficiente di regressione sia nullo:

$H_0 : \beta_j = 0$

e gli altri $\boldsymbol{\beta}$ qualsiasi

Il test in questo caso diventa:

$$F = \frac{b_j[(\frac{\mathbf{Z}^\top\mathbf{Z}}{n})^{-1}{}_{jj}]^{-1}b_j}{\frac{s^2}{n}} = \frac{b_j^2}{c_{jj}\frac{s^2}{n}}$$

essendo c_{jj} il j -esimo elemento sulla diagonale di $(\frac{\mathbf{Z}^\top\mathbf{Z}}{n})^{-1}$ (ricordo che essendo $\mathbf{Z}$ la matrice degli scarti dei regressori, allora semplicemente $\frac{\mathbf{Z}^\top\mathbf{Z}}{n} = S_{\mathbf{X}}$ ossia la matrice di varianza e covarianza osservata delle X); essendo $q = 1$ possiamo prendere la radice quadrata di questa quantità, che si distribuisce come una t di Student con $n - k - 1$ gradi di libertà, per ottenere il test:

Test relativo ad un solo regressore

$$t = \frac{b_j}{c_{jj}\sqrt{\frac{s^2}{n}}} \sim t_{n-k-1}$$

Si può eventualmente considerare in questo caso un'alternativa unidirezionale che conduce a regioni di rifiuto sulla coda destra o sulla sinistra. Si noti anche che $\sigma^2 c_{jj}$ è la varianza campionaria di b_j : infatti sappiamo che

$$\mathrm{V}[\mathbf{b}] = \frac{\sigma^2}{n}\left(\frac{\mathbf{Z}^\top\mathbf{Z}}{n}\right)^{-1}$$

, per cui la varianza di b_j, elemento j-esimo di $\mathbf{b}$, è dato da $\frac{\sigma^2}{n}$ per l'elemento j-esimo della diagonale di $S_{\mathbf{X}}^{-1}$, ossia c_{jj}. [2] Con questo test possiamo saggiare una ipotesi su un coefficiente (uno e uno solo!!!); Utilizzare questo test per più di un regressore è una procedura distorta.

[2] Avverto che in alcune parte di questi appunti, ho indicato con c_{ij} gli elementi dell'inversa di $\frac{\mathbf{Z}^\top\mathbf{Z}}{n}$, come in questa sezione, ma qualche volta,in particolare nella versione degli anni precedenti, ho indicato invece gli elementi dell'inversa di $\mathbf{Z}^\top\mathbf{Z}$. Mi scuso dell'inconveniente, segnalatemi eventualmente le sezioni in cuic'è ancora quest'ultima notazione

6.4 Test per l'eliminazione di q regressori in termini di perdita in R^2

Partiamo dalla relazione:

$$Dev(\mathbf{y}) = (Dev(\mathbf{y}) - R(\mathbf{b}_0)) + (R(\mathbf{b}_0) - R(\mathbf{b})) + R(\mathbf{b})$$

Riscriviamo il test per saggiare l'ipotesi che q regressori siano nulli:

$$F = \frac{\frac{[R(\mathbf{b}_0) - R(\mathbf{b})]}{q}}{\frac{R(\mathbf{b})}{n-k-1}}$$

$$F = \frac{\frac{\text{Devianza spiegata da } k \text{ regressori-Devianza spiegata da } k-q \text{ regressori}}{q}}{\frac{\text{Devianza residua [nel modello completo]}}{n-k-1}}$$

Dividendo ora ambo i termini della frazione per $Dev(\mathbf{y})$ si può esprimere questo test in funzione di due diversi indici R^2 :

$$F = \frac{\frac{\mathbf{R}^2_{\mathbf{y}.12\ldots k} - \mathbf{R}^2_{\mathbf{y}.q+1\ldots k}}{q}}{\frac{1-\mathbf{R}^2_{\mathbf{y}.12\ldots k}}{n-k-1}} = \frac{\mathbf{R}^2_{\mathbf{y}.12\ldots k} - \mathbf{R}^2_{\mathbf{y}.q+1\ldots k}}{1 - \mathbf{R}^2_{\mathbf{y}.12\ldots k}} \frac{n-k-1}{q}$$

in cui:

- $\mathbf{R}^2_{\mathbf{y}.q+1\ldots k}$ è la frazione di varianza di $\mathbf{y}$ spiegata dai $k-q$ regressori $\mathbf{X}_{q+1}, \mathbf{X}_{q+2}, \ldots, \mathbf{X}_k$;
- $\mathbf{R}^2_{\mathbf{y}.12\ldots k}$ è la frazione di varianza di $\mathbf{y}$ spiegata da tutti i regressori;

Quindi il test corrisponde a saggiare l'ipotesi che il decremento in $\mathbf{R}^2_{\mathbf{y}.12\ldots k}$ dovuto all'eliminazione dei q regressori $\mathbf{X}_1$, $\mathbf{X}_2$, $\ldots$, $\mathbf{X}_q$ sia dovuto solo al caso, e non dall'influenza dei primi k regressori.[3]

Il numeratore del test F, ossia $\mathbf{R}^2_{\mathbf{y}.12\ldots k} - \mathbf{R}^2_{\mathbf{y}.q+1\ldots k}$, è la proporzione di varianza di $\mathbf{y}$ spiegata dai primi q regressori, al netto degli altri $k-q$. Evidentemente il numeratore del test F è sempre positivo: si tratta sempre, come si era visto prima, di una frazione di varianza.

Il test si riferisce ad una ipotesi relativa ad un insieme *fissato* di q regressori. Successivamente si utilizzeranno queste scomposizioni per arrivare ad un criterio di scelta di $k-q$ particolari regressori.

Possiamo impostare una tavola di analisi della varianza per la riduzione di variabili:

[3] perchè stiamo saggiando l'ipotesi $H_0 : \beta_1 = \beta_2 = \ldots = \beta_q = 0;\ \forall \beta_j,\ j = q+1, \ldots, k$

TOTALE=	RESIDUA	SPIEGATA da $k-q$ regressori	SPIEGATA da q regressori (al netto degli altri $k-q$)
frazioni di varianza			
1	$1 - \mathbf{R}^2_{\mathbf{y}.12\ldots k}$	$\mathbf{R}^2_{\mathbf{y}.q+1\ldots k}$	$\mathbf{R}^2_{\mathbf{y}.12\ldots k} - \mathbf{R}^2_{\mathbf{y}.q+1\ldots k}$
gradi di libertà:			
$n-1$	$n-k-1$	$k-q$	q

$\mathbf{R}^2_{\mathbf{y}.12\ldots k}$	devianza spiegata da tutti i k regressori
$\mathbf{R}^2_{\mathbf{y}.q+1\ldots k}$	devianza spiegata dagli ultimi $k-q$ regressori
$\mathbf{R}^2_{\mathbf{y}.12\ldots k} - \mathbf{R}^2_{\mathbf{y}.q+1\ldots k}$	devianza in più spiegata dai primi q regressori
$1 - \mathbf{R}^2_{\mathbf{y}.q+1\ldots k}$	devianza non spiegata dagli ultimi $k-q$ regressori

Un indice normalizzato è dato da:

$$\frac{\mathbf{R}^2_{\mathbf{y}.12\ldots k} - \mathbf{R}^2_{\mathbf{y}.q+1\ldots k}}{1 - \mathbf{R}^2_{\mathbf{y}.q+1\ldots k}}$$

coefficiente di determinazione parziale di $\mathbf{Y}$ sui primi q regressori, al netto degli altri k-q regressori

L'indice che è ancora palesemente compreso fra 0 e 1;

misura la frazione ulteriore di varianza spiegata dai q regressori, tenuto conto della regressione sugli altri k-q.

incremento di R^2 in funzione dell'indice di correlazione parziale;

si può dimostrare un collegamento fra il numeratore del test F per la verifica di un ipotesi relativa ad un coefficiente di regresione e la correlazione parziale fra due variabili:

7 La multicollinearità nella regressione multipla.

8 Generalità del problema

In questa sezione affrontiamo un problema cruciale nell'analisi della regressione multipla, molto rilevante in alcuni contesti applicativi, per esempio per l'analisi di aggregati economici o comunque per l'analisi di dati provenienti da indagini osservazionali con molte variabili, che si può riassumere nella domanda:

avere variabili esplicative, ossia regressori, linearmente correlate ha qualche influenza negativa sull'analisi della regressione?

Banalmente si potrebbe pensare che l'unica cosa importante (e su cui finora ci siamo concentrati) è la dipendenza (multipla) della $\mathbf{Y}$ dalle $\mathbf{X}$. Vedremo in questa sezione che è anche importantissimo analizzare la struttura di correlazione interna delle $\mathbf{X}$.

Abbiamo in effetti già visto link con regressori non correlati che avere regressori non correlati porta diversi vantaggi in un problema di regressione multipla, sia dal punto di vista statistico, che dal punto di vista numerico-computazionale:

- con regressori non correlati possiamo dare un significato alle correlazioni della risposta $\mathbf{Y}$ con ciascuno dei regressori, dato che il singolo contributo portato dal regressore $\mathbf{X}_j$ alla spiegazione della varianza di $\mathbf{Y}$ è dato da r^2_{y,X_j}; oltretutto tali contributi risultano indipendenti
- regressori non correlati portano stimatori dei regressori non correlati e quindi, per l'ipotesi di normalità, indipendenti;
- gli stimatori possono essere calcolati separatamente; (sebbene questo aspetto al giorno d'oggi non sia particolarmente rilevante vista la grande potenza di calcolo disponibile).

Nell'altro caso estremo, ossia regressori esattamente collineari, la matrice $\mathbf{X}^\mathsf{T}\mathbf{X}$non sarà invertibile, ma questo non è detto sia un problema insormantabile: potrebbe bastare trovare il regressore o i regressori che portano tale collinearità ed eliminarli dall'analisi (tanto il loro contributo parziale alla spiegazione di $\mathbf{Y}$ è senz'altro nullo); per trovare i regressori che portano la collinearità esatta (ossia tali che esista una loro combinazione lineare che è uguale al vettore nullo) si possono per esempio cercare gli autovalori della matrice di varianza e covarianza delle $\mathbf{X}$ ed esaminare i

coefficienti dell'autovettore corrispondente all'autovalore nullo: [4] i regressori coinvolti nel vincolo lineare esatto sono quelli corrispondenti ai coefficienti non nulli di tale autovettore. `link con esercizio su collinearità esatta`

`Oppure si può ricorrere alla cosiddetta ridge regression di cui non ci occuperemo in questo corso`

Può anche verificarsi il caso di avere una multicollinearità quasi esatta: è il caso in cui, pur non avendosi delle combinazioni lineari esatte di regressori, il determinante della matrice di varianze e covarianze è *quasi nullo* nel senso che è numericamente vicino a zero, ossia risulta vicino al più piccolo numero rappresentabile dalla macchina; in pratica risulterà impossibile calcolare gli elementi dell'inversa o comunque questi risulteranno numericamente instabili; si parla in questo caso di *malcondizionamento*. Tuttavia questo è un problema essenzialmente di calcolo numerico (impossibilità materiale di invertire una matrice non esattamente singolare) e non sono particolarmente interessato a questo aspetto: quello che interessa qui sono le conseguenze statistiche dell'avere regressori correlati (anche molto correlati).

[4] ve ne sarà almeno uno nullo in caso di collinearità esatta

E in tutte le situazioni intermedie? ossia regressori correlati, ma non tanto da portare ad una collinearità esatta? Questi sono i casi più comuni nell'analisi dei dati reali, per cui inizieremo a vedere cosa comporta la correlazione fra i regressori in termini di varianza degli stimatori, partendo dal caso più semplice, con $k = 2$. Inoltre per comodità di notazione supporremo in questa trattazione dela multicollinearità di avere variabili standardizzate (media nulla e varianza unitaria), dal momento che media e dispersione non influenzano le correlazioni.

8.1 Caso di due soli regressori

Supponiamo un caso molto semplice con due soli regressori X_1 e X_2. Consideriamo per semplificare le cose, e focalizzare l'attenzione solo sulle correlazioni, che le variabili X_1 e X_2 siano tutte standardizzate.

Sappiamo che $\mathrm{V}[\mathbf{b}] = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \sigma^2$. Se quindi $k = 2$ si ha:

$$\mathrm{V}[\mathbf{b}] = \sigma^2 \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}^{-1} = \sigma^2 \frac{1}{1 - r_{12}^2} \begin{pmatrix} 1 & -r_{12} \\ -r_{12} & 1 \end{pmatrix}$$

per cui la varianza dei due stimatori è data da:

$$\mathrm{V}[\mathbf{b}_1] = \mathrm{V}[\mathbf{b}_2] = \sigma^2 \frac{1}{1 - r_{12}^2}$$

Collinearità nella regressione a due regressori

La varianza degli stimatori dei coefficienti di regressione è *funzione crescente* del valore assoluto della correlazione fra i regressori r_{12} ed è *funzione crescente* della varianza σ^2 della componente accidentale.

Quindi due regressori molto correlati portano (con $k = 2$) a stimatori con varianza molto elevata, addirittura tendente ad infinito quando r_{12} in valore assoluto tende ad uno!. E' utile ricordare che ovviamente questo ha conseguenze sulla varianza stimata di qualsiasi funzione lineare di tali stimatori $\mathbf{b}$, ad esempio $\hat{\mathbf{y}} = \mathbf{Xb}$, ossia i valori stimati.

Forse quindi potrebbe essere il caso di utilizzare un solo regressore anzichè due se r_{12}^2 fosse troppo vicino ad uno; chiediamoci allora di quanto incrementa la frazione di varianza spiegata di $\mathbf{Y}$ se usiamo due regressori anzichè uno solo; per sem-

plificare l'esempio (ma la sostanza non cambia) supponiamo che la correlazione semplice di $\mathbf{Y}$ con i due regressori sia la stessa, e indichiamola con r_{y1}. La matrice di correlazione fra le tre variabili, necessaria per il calcolo del coefficiente di determinazione multipla, è dunque data da:

$$R(\mathbf{Y}, X_1, X_2) = \begin{pmatrix} 1 & r_{y1} & r_{y1} \\ r_{y1} & 1 & r_{12} \\ r_{y1} & r_{12} & 1 \end{pmatrix}$$

Indichiamo adesso con $\mathbf{C}$ la sua inversa di elemento generico c_ij. Calcoliamo il coefficiente[5] di determinazione multipla $\mathbf{R}^2_{\mathbf{y}.12}$: link con determinazione multipla in

$$\mathbf{R}^2_{\mathbf{y}.12} = 1 - \frac{1}{c_{yy}} =$$

[5]Ricordo che, per essere una matrice di correlazione, $R(\mathbf{Y}, X_1, X_2)$ deve avere determinante positivo, e questo impone il vincolo, verificabile facilmente dallo sviluppo del determinante, $r^2_{y1} < \frac{1+r_{12}}{2}$

$$= 1 - \frac{\begin{vmatrix} 1 & r_{y1} & r_{y1} \\ r_{y1} & 1 & r_{12} \\ r_{y1} & r_{12} & 1 \end{vmatrix}}{\begin{vmatrix} 1 & r_{12} \\ r_{12} & 1 \end{vmatrix}} = 1 - \frac{-r_{12}^2 + 2r_{y1}^2 r_{12} - 2r_{y1}^2 + 1}{1 - r_{12}^2} =$$

$$= \frac{2r_{y1}^2}{1 + r_{12}}$$

La frazione di devianza spiegata da un solo regressore è data ovviamente da: $\mathbf{R}^2_{\mathbf{y}.1} = r^2_{y1}$, per cui l'incremento di spiegazione dovuto all'introduzione del secondo regressore è dato da:

$$\mathbf{R}^2_{\mathbf{y}.12} - \mathbf{R}^2_{\mathbf{y}.1} = \frac{2r_{y1}^2}{1 + r_{12}} - r_{y1}^2 = \frac{r_{y1}^2(1 - r_{12})}{1 + r_{12}}$$

Se adesso ipotizziamo $r_{12} > 0$, vediamo facilmente che tale incremento di devianza spiegata si riduce all'avvicinarsi di r_{12} ad uno! Pertanto, in questa situazione molto

semplice è facile vedere l'effetto di una collinearità positiva fra i due regressori: valori di r_{12} vicini ad uno portano un incremento della varianza campionaria degli stimatori dei coefficienti di regressione (e quindi anche delle varianze campionarie dei valori previsti) ed un incremento sempre minore della spiegazione della varianza di $\mathbf{Y}$!!!

In definitiva, come si vedrà meglio a proposito di scelta di variabili link con sceltavar , quando i regressori sono molto correlati aumentare il numero di variabili con regressori correlati non solo non serve (perchè la devianza spiegata di $\mathbf{Y}$ aumenta poco), ma è pure controproducente (perchè la varianza degli stimatori aumenta).

8.2 Collinearità con k regressori

Adesso passiamo alla situazione generale, con k regressori; ovviamente se i k regressori non sono ortogonali, ossia a correlazione nulla, questo può portare a strutture di interdipendenza (fra i regressori!) di vario tipo.

Si sono già viste alcune delle conseguenze della non ortogonalità dei regressori o fattori sulla distribuzione degli stimatori di massima verosimiglianza e di altre quantità collegate:

- Lo stimatore $\mathbf{b}$ è a componenti correlate (dal momento che ha varianza proporzionale a $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$);
- I contributi alla spiegazione di $\mathbf{Y}$ di ciascuna variabile non sono separabili.
- Non si possono condurre test indipendenti su tutti i singoli coefficienti.
- Le regioni di confidenza dei parametri $\boldsymbol{\beta}$ costruite sulla base del valore critico di F risultano ellissoidali e non sferiche.
- Il luogo dei punti $\mathbf{x}_i$ nello spazio dei regressori che conduce ad intervalli di confidenza di eguale ampiezza per $E(\hat{\psi}_i)$ è il contorno di un ellissoide di equazione:

$$\sigma^2 \mathbf{x}_{(i)^T} (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1} \mathbf{x}_{(i)} = Cost.$$

Il caso estremo è quello in cui il rango di $\mathbf{X}$ (e quindi di $\mathbf{X}^\mathsf{T}\mathbf{X}$) è inferiore a k: supponiamo di non trovarci comunque in questa situazione, perché l'eventuale variabile combinazione lineare esatta delle altre può essere individuata ed eliminata.

Supponiamo di trovarci invece, nell'ambito delle situazioni con dei regressori molto correlati, vicino a questa situazione estrema.

Occorrerà misurare il grado di collinearità fra le variabili indipendenti (indotta

dalle correlazioni fra le $\mathbf{X}_j$) e vedere quanto complessivamente incide sulla distribuzione di $\mathbf{b}$ ed in generale sull'inferenza nella regressione multipla; la situazione con $k = 2$ è stata affrontata nella sezione precedente: ovviamente adesso si possono avere configurazioni differenti della matrice di correlazione fra le X, mentre con $k = 2$ la struttura di correlazione è determinata esclusivamente dal valore di r_{12}. Occorre una strumentazione tecnica più generale ma i concetti di fondo sulle conseguenze della collinearità, delineati nella sezione precedente, non cambieranno.

Consideriamo una matrice delle X a media nulla ed a varianza unitaria (quindi è una matrice di variabili standardizzate); evidentemente ciò corrisponde ad effettuare una traslazione ed un cambiamento di scala sugli assi che non alterano in alcun modo lo studio della dipendenza lineare di $\mathbf{y}$ dalle $\mathbf{X}_j$. Anzi in questo modo si possono fare valutazioni comparative fra i coefficienti di regressione, in quanto non influenzati dalle diverse unità di misura.

Lo studio della multicollinearità riguarda la struttura di correlazione fra le $\mathbf{X}$ e successivamente l'influenza di questa struttura sullo studio della dipendenza di $\mathbf{Y}$ dalle $\mathbf{X}$, sulle proprietà degli stimatori, delle regioni di confidenza, etc.

Come accennato prima, stiamo esaminando essenzialmente le implicazioni di tipo statistico della multicollinearità: lasciamo volutamente da parte le questioni di natura computazionale. È noto, infatti, che dal punto di vista numerico la risoluzione di sistemi di equazioni lineari, in presenza di collinearità, comporta dei problemi di stabilità numerica delle soluzioni.

Con determinante della matrice dei coefficienti prossimo a zero gli errori di troncamento potrebbero svolgere un ruolo determinante sul calcolo delle soluzioni del sistema di equazioni normali.

Se le X sono standardizzate la matrice di varianze e covarianze $\mathbf{S}$ è anche la matrice di correlazione, ed è data da:

$$\mathbf{S} = \frac{\mathbf{X}^\top \mathbf{X}}{n}.$$

Quindi è lo stesso studiare la struttura di $\mathbf{X}^\top \mathbf{X}$ o quella di $\mathbf{S}$.

Dal momento che le x sono a media nulla e a varianza unitaria, si avrà che combinazioni lineari delle x sono a media nulla, e inoltre, dal momento che la somma degli autovalori di $\mathbf{S}$ è uguale alla sua traccia (ossia alla somma delle varianze), tale somma è uguale a k dato che stiamo lavorando con variabili standardizzate.

Infatti:

$$\lambda_i > 0 \qquad i, k = 1, 2, \ldots,$$

(**S** è definita positiva e quindi di rango pieno) Inoltre:

$$\sum_{i=1}^{k} \lambda_i = k \qquad \text{e quindi:} \qquad M(\lambda_i) = \frac{\sum_{i=1}^{k} \lambda_i}{k} = 1$$

Per cui gli autovalori sono limitati fra 0 e k:

$$0 \; < \; \lambda_i \; < \; k \qquad i, k = 1, 2, \ldots,$$

Nella situazione ideale di assenza di correlazioni fra le x si ha:

$$\lambda_1 = \lambda_2 = \ldots = \lambda_k = 1$$

perché $\mathbf{S} = \mathbf{I}$

La situazione è ideale perché le stime dei regressori risultano non correlate e le inferenze sui regressori sono indipendenti.

multicollinearità

Si parla di multicollinearità quando, pur essendo la matrice $\mathbf{S}$ a rango pieno, alcuni dei suoi autovalori sono piccoli, avvicinandosi alla situazione estrema di collinearità esatta e comunque portando ad un incremento della varianza degli stimatori dei coefficienti di regressione.

Questo si verifica quando qualcuna delle variabili x è quasi uguale ad una combinazione lineare di alcune delle altre variabili $\mathbf{X}$.

La situazione limite $\lambda_k = 0$ corrisponde al caso di rango inferiore a k, ossia una variabile è esattamente combinazione lineare delle altre (oppure q variabili sono combinazioni lineari delle altre se $\lambda_{k-q+1} = \lambda_{k-q+2} = \ldots = \lambda_k = 0$)

Nella regressione multipla ci interessa che la $\mathbf{Y}$ sia molto correlata con le $\mathbf{X}$, ma è preferibile che le $\mathbf{X}$ siano poco correlate internamente

Si riveda per analogia la parte relativa all'analisi delle componenti principali per vettori aleatori. Si riveda anche l'interpretazione dell'analisi in componenti principali per variabili statistiche osservate.

Si riveda anche lo schema riportato in un capitolo precedente sull'influenza delle possibili configurazioni di matrice $\mathbf{x}$ sull'inferenza nei modelli lineari.

9 Legami lineari fra regressori

Adesso esamineremo con dettaglio l'influenza delle correlazioni fra i regressori nel caso generale: esistono infatti delle situazioni nelle quali la presenza di correlazioni potrebbe essere importante anche se non si è in una situazione di multicollinearità vera e propria; si vedrà più avanti a questo proposito link con varianza delle previs

la relazione che lega la varianza delle previsioni con la varianza degli stimatori.

Dall'equazione che definisce gli autovettori e gli autovalori della matrice delle varianze e covarianze $\mathbf{S}$ (gli autovalori sono proporzionali a quelli della matrice delle devianze e codevianze $\mathbf{X}^\mathsf{T}\mathbf{X}$, essendo $\mathbf{X}$ una matrice di variabili scartate dalle rispettive medie e possibilmente standardizzate) si ha:

$$\mathbf{S}\boldsymbol{\gamma}_j = \frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{n}\boldsymbol{\gamma}_j = \lambda_j\boldsymbol{\gamma}_j \approx 0 \qquad \text{se} \qquad \lambda_j \approx 0$$

(dato che tutti gli elementi di $\boldsymbol{\gamma}_j$, i-esimo autovettore sono compresi fra 0 e 1, per la condizione di normalizzazione $\boldsymbol{\gamma}_j^\mathsf{T}\boldsymbol{\gamma}_j = 1$)

Allora premoltiplicando per $\boldsymbol{\gamma}_j^\mathsf{T}$ si ha:

$$\frac{(\boldsymbol{\gamma}_j^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\gamma}_j)}{n} = \boldsymbol{\gamma}_j^\mathsf{T}\lambda_j\boldsymbol{\gamma}_j = \lambda_j \approx 0$$

Poniamo $u_j = \frac{\mathbf{X}\boldsymbol{\gamma}_j}{\sqrt{n}}$

così che u_j è una combinazione lineare nelle $\mathbf{X}$, e quindi:

$$(\boldsymbol{\gamma}_j^T\mathbf{X}^\mathsf{T}\mathbf{X}\boldsymbol{\gamma}_j)/n = u_j^T u_j = \lambda_j \approx 0 \qquad \text{(per l'ipotesi fatta)}$$

Allora se λ_j è piccolo il vettore u_j è una combinazione lineare delle $\mathbf{X}$, con media zero e varianza molto piccola, per cui si ha anche:

$$u_j \approx 0 \qquad \text{ossia} \qquad \Rightarrow \mathbf{X}\boldsymbol{\gamma}_j \approx 0$$

- Quindi esiste una combinazione lineare delle variabili quasi nulla;
- le variabili maggiormente coinvolte corrispondono ai più alti coefficienti di $\boldsymbol{\gamma}_j$; ossia le variabili $\mathbf{X}_r$ corrispondenti ai più alti elementi $\gamma_{rj}(r, k = 1, 2, \ldots,)$;

avendo inteso le colonne della matrice $\mathbf{\Gamma}$ di elemento γ_{rj} costituite dagli autovettori di $\mathbf{S}$

Si può giungere a questo tipo di risultato (ossia esistenza di combinazioni lineari quasi esatte fra i regressori), anche considerando che in questo caso una o più variabili risulta combinazione lineare quasi esatta delle altre, ossia avrà una dipendenza lineare elevata dalle altre variabili, in termini di regressione multipla. Risulta naturale dunque, e perfettamente in linea con i concetti della regressione, considerare come misura della collinearità potenziale di ciascuna variabile, il coefficiente di determinazione multipla $R^2_{j.}$, $(j = 1, 2, \ldots, k)$ che misura la parte di varianza di X_j spiegata dalla regressione lineare sugli altri $k-1$ regressori.

In effetti, ricordando le relazioni fra R^2 e gli elementi dell'inversa di $\mathbf{S}$ (si rivedano nella parte relativa alle distribuzioni condizionate di v.a. normali link con distr. cond
e in particolare i notebook con esempi in R usati a lezione), si sa che:

$$R^2_i = 1 - \frac{1}{c_{ii}}$$

R^2_i è il coefficiente di determinazione multipla di $\mathbf{X}_i$ rispetto alle altre $k-1$ variabili,

c_{ii} è l'elemento diagonale di $\mathbf{C}$, l'inversa di $\mathbf{S}$, ed anche:

$$R_i^2 = 1 - \frac{1}{c_{ii}} \Rightarrow \frac{1}{1 - R_i^2} = c_{ii}$$

Ricordando anche che:

$$\lambda_j(\mathbf{C}) = \lambda_j\left(S^{-1}\right) = \frac{1}{\lambda_j\left(S\right)};$$

sommando queste ultime relazioni per tutte le variabili si ha:

$$\sum_{i=1}^{k} 1/(1 - R_i^2) = \sum_{i=1}^{k} c_{ii} = tr[\mathbf{C}] = \sum_{i=1}^{k} 1/\lambda_i$$

Quindi se qualche autovalore è molto piccolo, la traccia di $\mathbf{C}$ è molto grande e questo è direttamente collegato al fatto che qualche correlazione multipla delle x è elevata.

9.1 Multicollinearità e distribuzione campionaria di b

Adesso vediamo le caratteristiche fondamentali dell'influenza della collinearità sulla distribuzione di $\mathbf{b}$ che generalizzano il caso visto per $k = 2$. Rivediamo la matrice

di varianze e covarianze di $\mathbf{b}$:

$$\mathrm{V}[\mathbf{b}] = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \sigma^2(n\mathbf{S})^{-1} = \frac{\sigma^2}{n}\,\mathbf{S}^{-1} = \frac{\sigma^2}{n}\,\mathbf{C}$$

Quindi a parte il fattore $\frac{\sigma^2}{n}$ la struttura delle correlazioni interne fra gli elementi di $\mathbf{b}$ è funzione della struttura delle correlazioni interne fra le $\mathbf{X}$, e non dipende in alcun modo dalla variabile di risposta $\mathbf{y}$: dipende solo dallo schema di valori assunti dai regressori (siano essi osservati o prestabiliti prima di un esperimento).[6] Le varianze invece dipendono al solito dai valori osservati, attraverso il fattore (σ^2/n).

Vediamo come la somma di tali varianze (ossia la traccia della matrice delle varianze e covarianze di $\mathbf{b}$) dipende dagli autovalori di $\mathbf{S}$:

$$\sum_{i=1}^{k} \mathrm{V}[b_i] = tr(\mathrm{V}[\mathbf{b}]) = \frac{\sigma^2}{n}\,tr(\mathbf{S}^{-1}) \;=\; \frac{\sigma^2}{n}\,tr[\mathbf{C}] = \frac{\sigma^2}{n}\sum_{i=1}^{k}\frac{1}{\lambda_i}$$

Ecco finalmente il collegamento fra collinearità e varianza degli stimatori nel modello lineare!

[6] Si tenga presente che la collinarità quasi esatta porta può portare problemi numerici nell'inversione di una matrice, ma non ne porta nella determinazione degli autovalori, poichè per quest'ultimo problema esistono algoritmi robusti anche per matrici di rango non pieno

Conseguenze della multicollinearità

Se vi è multicollinearità (ossia qualche λ_i molto piccolo) la traccia di $\mathbf{C}$ sarà elevata e quindi sarà elevata la somma delle varianze campionarie degli stimatori dei coefficienti di regressione. Sarà conseguentemente elevata anche la varianza di $\hat{\mathbf{y}}_i$

Indici di multicollinearitá:

$$I_p = \frac{\Sigma_{i=1}^{p} \lambda_j}{\Sigma_{i=1}^{k} \lambda_j} = \frac{\text{varianza delle prime } p \text{ componenti}}{\text{somma di tutte le varianze}}$$

$$I_p = \frac{\Sigma_{i=1}^{p} \lambda_j}{k}$$

nel caso di variabili standardizzate.

In particolare è utile mettere in collegamento la varianza totale delle b_i con il valore minimo, che si ottiene quando i λ_i sono tutti uguali fra loro, ossia nel caso

migliore di regressori non correlati:

$$IV = 1 - \frac{k}{\Sigma_{i=1}^{k} \frac{1}{\lambda_j}} = 1 - \frac{\text{varianze dei } b_i \text{ con regressori non correlati}}{\text{varianze effettive dei } b_i}$$

Più che regole automatiche, l'analisi grafica dell'andamento di $I_p$al variare di p può guidare nell'analisi della multicollinearità in insiemi di dati reali.

---

argomenti avanzati e complementi (non in programma)

9.1.1 Costruzione di un stimatore distorto di β

Per esaminare meglio gli effetti della multicollinearità sulla varianza campionaria dello stimatore $\mathbf{b}$, si può sfruttare la decomposizione spettrale o canonica della matrice $\mathbf{S}^{-1}$, introdotta a proposito delle proprietà degli autovalori e degli autovettori di matrici simmetriche:

$$\mathbf{S}^{-1} = \mathbf{\Gamma \Lambda}^{-1} \mathbf{\Gamma}^{\mathsf{T}} = \frac{\sum_{i=1}^{k} \gamma_i \gamma_i^{\mathsf{T}}}{\lambda_i}$$

mentre per la matrice originaria $\mathbf{S}$ abbiamo la decomposizione di base:

$$\mathbf{S} = \mathbf{\Gamma \Lambda \Gamma}^{\mathsf{T}} = \sum_{i=1}^{k} \lambda_i \gamma_i \gamma_i^{\mathsf{T}}$$

Se invece di prendere tutti i k termini di questa decomposizione, ci limitiamo a prendere i primi q termini, otteniamo un'approssimazione della matrice $\mathbf{S}$ tanto migliore, quanto più sono piccoli gli autovalori corrispondenti ai termini scartati:

$$\mathbf{S} = \mathbf{\Gamma \Lambda \Gamma}^{\mathsf{T}} = \sum_{i=1}^{k} \lambda_i \gamma_i \gamma_i^{\mathsf{T}} \approx \sum_{i=1}^{q} \lambda_i \gamma_i \gamma_i^{\mathsf{T}} S_{(q)}$$

in corrispondenza di questa approssimazione costruiamo una inversa modificata:

$$\mathbf{S}^{-1} = \sum_{i=1}^{k} \gamma_i \gamma_i^{T} / \lambda_i \rightarrow \sum_{i=1}^{q} \gamma_i \gamma_i^{T} / \lambda_i = \mathbf{S}_{(q)}^{-1},$$

in cui stavolta mancano i termini più elevati in valore assoluto.
(evidentemente le stesse scomposizioni, a meno del fattore n, si possono fare sulla matrice $\mathbf{X}^{\mathsf{T}}\mathbf{X}$)
Pertanto, se invece di $\mathbf{b}$ si definisse:

$$b^0 = \mathbf{S}_{(q)}^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}/n$$

si otterrebbe uno stimatore distorto ma con minore varianza!
Infatti:
(controllare bene il seguito)

$$E(b^0) = \mathbf{S}_{(q)}^{-1}\mathbf{X}^\mathsf{T}(\mathbf{X}\boldsymbol{\beta}) =$$

$$\mathbf{S}_{(q)}^{-1}(\mathbf{X}^\mathsf{T}\mathbf{X})\boldsymbol{\beta} = (\mathbf{S}_{(q)}^{-1}/n)(n\mathbf{S}_{(q)} + \mathbf{R}(q))\boldsymbol{\beta} =$$

$$\sum_{i=1}^{q} \frac{\boldsymbol{\gamma}_i^\mathsf{T}\boldsymbol{\gamma}_i}{\lambda_i}(\lambda_i\boldsymbol{\gamma}_i^\mathsf{T}\boldsymbol{\gamma}_i) + \sum_{i=q+1}^{k} \lambda_i\boldsymbol{\gamma}_i^\mathsf{T}\boldsymbol{\gamma}_i$$

A parte l'eventuale impiego effettivo di questo stimatore, l'utilità della sua introduzione sta nell'esplicitazione del legame fra distorsione e varianza campionaria degli stimatori di $\boldsymbol{\beta}$.

9.2 Esempi (sulla collinearità e simili)

10 La scelta delle variabili nella regressione lineare multipla.

Consultare il materiale dei notebook, degli esercizi e laboratorio
(4 maggio 2020, versione ancora da riaggiornare)

10.1 Perchè scegliere variabili?

Abbiamo parlato più volte di sottoinsiemi di variabili; abbiamo anche parlato implicitamente di confronto fra modelli con diversi regressori; in generale ci stiamo cominciando ad occupare adesso di *scelta del modello.*

Si è detto prima di sottoinsiemi di variabili predittive stabiliti a priori e quindi senza riferimento ai particolari dati osservati e si è anche visto come fare inferenza su tali sottoinsiemi stabiliti a priori.

L'analisi della correlazione parziale ci pemette di vedere la correlazione fra la variabile di risposta e un regressore, tenuti costanti gli altri, in termini di incrementi di devianza spiegata e quindi ci permette di vedere qual è il contributo di una variabile al netto delle altre.

Si è pure visto nell'analisi della multicollinearità come la presenza di vincoli

lineari fra i regressori possa abbassare molto la qualità dell'inferenza aumentando la varianza degli stimatori.

Spesso però, date k variabili esplicative, si vuole scegliere un sottoinsieme di q di tali variabili con diverse finalità:

- per effettuare stime o previsioni statistiche a costo inferiore, riducendo il numero di variabili che occorrerà rilevare in futuri studi.
- Per migliorare l'accuratezza delle previsioni eliminando variabili poco informative o comunque poco rilevanti ai fini della previsione di $\mathrm{E}[\mathbf{y}]$
- per descrivere un data-set multivariato, o comunque una relazione multipla in modo parsimonioso e con pochi parametri.
- per stimare coefficienti di regressione con errori standard piccoli, in particolare se alcuni dei regressori sono molto correlati.
- Stime carenti dei coefficienti possono portare buone stime predittive (ossia al solo scopo di stimare valori di $\mathbf{y}$ o di $\mathrm{E}[\mathbf{y}]$).

10.2 Strategie di scelta

La strategia complessiva della scelta di variabili si può articolare in alcune fasi generali:

- decidere quali sono le variabili che costituiscono l'insieme più ampio dei k regressori;
- trovare uno o più sottoinsiemi di variabili che spiegano bene la variabile di risposta;
- applicare una regola di arresto per decidere quante variabili esplicative (regressori) usare;
- stimare i coefficienti di regressione
- saggiare la bontà del modello ottenuto (analisi dei residui, aggiunta di nuove variabili, aggiunta di termini polinomiali, etc.).

Per quanto riguarda il punto b), possiamo esplicitarlo in questo modo:
fissato un numero di regressori ridotto, diciamo p, quale dei ${}^kC_q$ sottoinsiemi dei k regressori originari scegliere?

Sembra logico, e comunque più semplice, almeno in prima istanza, scegliere quello che fornisce la maggior quota di varianza spiegata, ossia il maggiore fra gli R^2;

In aggiunta a questo criterio di massimizzazione globale, avendo fissato q, si può comunque pensare a scopo esplorativo di prendere in esame alcuni sottoinsiemi che forniscono le soluzioni migliori.

Occorrerà possibilmente un qualche algoritmo per ridurre il numero di R^2 da calcolare.

10.2.1 Fonti di distorsioni

Le distorsioni nella stima dei coefficienti sono dovute a diverse fonti:

- una distorsione dovuta all'avere omesso variabili, di ui è possibile fornire una valutazione (in termini di deviazione dal modello completo)
- una distorsione dovuta al procedimento di selezione, che non viene in generale fatto indipendentemente dai dati; in altri termini *i dati mediante i quali si stimano i coefficienti sono gli stessi che hanno portato alla selezione* di un particolare sottoinsieme.

cenno alla cross validation

Quest'ultima distorsione, dovuta alla selezione, può essere distinta in due ulteriori componenti:

- una dovuta alla scelta fra sottoinsiemi delle stesse dimensioni
- l'altra dovuta alla regola di arresto impiegata per scegliere il numero p migliore di regressori. Queste ultime fonti di distorsione in generale non sono valutabili con precisione.

10.3 Criteri di scelta

Che criterio usare per scegliere il numero p più opportuno di variabili da includere nel modello?

Si tenga presente che se A_p è l'insieme ottimo di p variabili e A_{p+1} è l'insieme ottimo con $p+1$ variabili, si ha sempre:

$$R^2_{\mathbf{y}.A_p} \leq R^2_{\mathbf{y}.A_{p+1}}$$

(l'uguaglianza in effetti vale solo in caso di collinearità esatta, che a rigore abbiamo escluso se $\mathbf{S}$ è di rango pieno). check

Inoltre se I_{q+1} è un insieme con $q+1$ variabili e se I_q è un suo sottoinsieme, ossia un insieme di q variabili ottenuto da I_{q+1} eliminando una sola variabile, si ha ancora:

$$R^2_{\mathbf{y}.I_q} \leq R^2_{\mathbf{y}.I_{q+1}} \qquad \text{con:} \qquad I_{q+1} \supset I_q$$

Eventuali test F condotti sugli R^2 saranno comunque distorti, almeno in termini di livelli di significatività. Infatti la devianza che si mette a numeratore non è calcolata su un set di variabili dato a priori, ma in base al fatto che il residuo sia il più basso possibile fra tutti i sottoinsiemi di variabili esplicative.

10.4 Algoritmi di scelta delle variabili.

Si possono comunque avere diversi algoritmi di scelta di variabili, a prescindere dal problema della scelta di p:

- Tutte le regressioni possibili
- Selezione in avanti (forward selection)
- Selezione all'indietro, o eliminazione (backward selection);
- Regressione passo (stepwise regression)

- algoritmi di sostituzione

Il metodo di tutte le regressioni possibili prevede l'esame di tutti i $2^k - 1$ possibili sottoinsiemi di variabili;

$$(2^k - 1 = \sum_{p=1}^{k} {}_kC_p)$$

Computazionalmente oneroso, sebbene esistano ora degli algoritmi di ricerca che consentono di limitare il numero dei confronti, pur trovando l'ottimo assoluto per ciascun numero di regressori q.

Un problema interpretativo si ha quando si ottengono soluzioni non nidificate: alcuni packages di R o altro software possono fornire oltre l'ottimo assoluto per ciascun valore di p, anche un certo numero di soluzioni sub-ottimali, ossia gli r migliori sottoinsiemi.

10.4.1 Metodi che conducono ad ottimi locali

Il metodo della selezione in avanti prevede di partire da un modello senza regressori, e di introdurli uno alla volta secondo che producano il valore più elevato fra i test F.

Evidentemente si trovano soluzioni sub-ottimali, e si rischia di non prendere mai in esame simultaneamente determinati sottoinsiemi di regressori.

Il metodo della selezione all'indietro, consiste nel partire dal modello completo, e ad ogni passo si elimina la variabile cui corrisponde il valore di F più basso.

Anche questo fornisce soluzioni sub-ottimali; tuttavia è molto usato e abbastanza ben interpretabile, in quanto prende comunque in esame una volta tutte le variabili simultaneamente;

inoltre fornisce una graduatoria delle variabili in ordine decrescente di importanza secondo l'ordine di eliminazione;

Il metodo stepwise unisce le due tecniche prima menzionate:

si parte da un modello senza regressori e si segue la tecnica della selezione in avanti; ad ogni passo con una nuova variabile introdotta, si riesamina l'insieme delle variabili introdotte, per vedere se si può eliminarne qualcuna (con procedura backward); successivamente i continua con la selezione in avanti fino a che non si effettuano più modifiche dell'insieme di regressori:

test di ingresso: $F > F_{in}$

test di uscita: $F < F_{out}$

(con $F_{in} > F_{out}$)

Questa tecnica, che risale al 1960, essenzialmente rispondeva all'esigenza pratica di non prendere in esame simultaneamente grossi insiemi di regressori; inoltre nella versione originaria considerava la possibilità di valutare le varie inverse e determinanti di ogni passo a partire da quelli trovati al passo precedente.

La funzione `step` di R consente di effettuare questa scelta basandosi sulla minimizzazione dell' AIC (definito più avanti)

10.4.2 Distorsione degli stimatori con modelli ridotti

Come si è visto:

$$\mathrm{E}\left[\hat{\mu}_i\right] = \mathrm{E}\left[\mathbf{x}_i^{\mathsf{T}}\mathbf{b}\right] = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} = \mathrm{E}\left[\mathbf{y}_i\right]$$

$$\mathrm{V}\left[\hat{\mu}_i\right] = \mathrm{V}\left[\mathbf{x}_i^{\mathsf{T}}\mathbf{b}\right] = \mathbf{x}_i^{\mathsf{T}}\mathrm{V}\left[\mathbf{b}\right]\mathbf{x}_i = \sigma^2\mathbf{x}_i^{\mathsf{T}}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{x}_i$$

Si può facilmente dimostrare che:

$$\sum_{i=1}^{n}\mathrm{V}\left[\hat{\mu}_i\right] = k\sigma^2$$

Infatti

$$\sum_{i=1}^{n}\mathrm{V}\left[\hat{\mu}_i\right] = \sum_{i=1}^{n}h_{ii}\sigma^2 = tr(\mathbf{H})\sigma^2 = k\sigma^2$$

Ovviamente questa relazione presuppone la correttezza del modello; se adesso prendiamo in considerazione la possibilità di lavorare con modelli distorti, dobbiamo vedere cosa succede all'errore quadratico medio della singola previsione:

$E.q.m(\hat{\mu}_i) = E.q.m.\left[\mathbf{x}_i^{\mathsf{T}}\mathbf{b}\right]$

E' opportuno ricordare che:

$$E.q.m.\left[\mathbf{b}\right] = \mathrm{V}\left[\mathbf{b}\right] + \left[\mathrm{E}\left[\mathbf{b}\right] - \boldsymbol{\beta}\right]\left[\mathrm{E}\left[\mathbf{b}\right] - \boldsymbol{\beta}\right]^{\mathsf{T}}$$

e quindi

$$E.q.m(\hat{\mu}_i) = \mathbf{x}_i^{\mathsf{T}}(\mathrm{V}\left[\mathbf{b}\right] + \left[\mathrm{E}\left[\mathbf{b}\right] - \boldsymbol{\beta}\right]\left[\mathrm{E}\left[\mathbf{b}\right] - \boldsymbol{\beta}\right]^{\mathsf{T}})\mathbf{x}_i =$$
$$= \mathbf{x}_i^{\mathsf{T}}\mathrm{V}\left[\mathbf{b}\right]\mathbf{x}_i + \mathbf{x}_i^{\mathsf{T}}\left[\mathrm{E}\left[\mathbf{b}\right] - \boldsymbol{\beta}\right]\left[\mathrm{E}\left[\mathbf{b}\right] - \boldsymbol{\beta}\right]^{\mathsf{T}}\mathbf{x}_i =$$
$$= \mathrm{V}\left[\hat{\mu}_i\right] + bias(\hat{\mu}_i)^2$$

10.5 Confronto fra modelli con un numero differente di regressori

Dobbiamo mediare fra due diverse esigenze (bassa distorione e bassa varianza)

- La prima esigenza ci porterebbe a preferire modelli con più variabili
- La seconda esigenza ci porterebbe a preferire modelli con meno variabili

Possibili criteri:

- C_p di Mallows
- AIC
- Cross validation
- Insieme dati di test

10.5.1 C_p di Mallows

Scelta di un modello di regressione lineare con p regressori su un totale di k regressori:

$$C_p = \frac{RSS_p}{s} - n + 2k$$

Approssima l'errore di previsione (RRS_p è la devianzaa residua per un modello con p regressori)

Si può rappresentare C_p in funzione di p e confrontare con la bisettrice (perchè si dimostra che per un modello non distorto $\mathrm{E}[C_p] = p$)

10.6 AIC

$$AIC = -2\log \mathrm{L}(\text{Mod. con } p \text{ param.}) + 2p$$

Compensa il valore della verosimiglianza (che aumenta col numero dei parametri) per il numero di parametri stimati

E' utilizzabile anche per modelli non lineari (per esempio con i GLM, la regressione logistica, etc.)

Applicato a modelli di regressione lineare multipla, è asintoticamente equivalente al C_p

La dimostrazione del perchè il termine $2p$ sia un buon termine di *penalizzazione* va al di là delle finalità di questo corso.

Faccio solo cenno al fatto che è collegato al fatto che la distribuzione asintotica del rapporto delle verosimiglianze, è una χ^2 con p gradi di libertà (non centrale sotto H_1, con un parametro di non centralità che dipende dalla particolare alternativa)

10.6.1 Cross validation

E' un'approssimazione dell'errore quadratico medio di previsione

$$MSEP = \mathrm{E}\left[\sum_{i=1}^{n}(\mu_i - \hat{\mu}_i)^2\right]$$

Si può confrontare ciascun valore y_i osservato con la stima $\hat{y}_{(-i)}$ ottenuta dalle altre $n-1$ osservazioni (cioè utilizzando la stima $\hat{\boldsymbol{\beta}}_{(-i)}$, ottenuto omettenda la $i-$esima osservazione y_i e la $i-$esima riga di variabili esplicative, $\mathbf{x}_i^{\mathsf{T}}$ e calcolando

poi $\hat{y}_{(-i)} = \mathbf{x}_i^{\mathsf{T}} \hat{\boldsymbol{\beta}}_{(-i)}$).

$$MSE(CV) = \frac{\Sigma_{i=1}^{n} (y_i - \hat{y}_{(-i)})^2}{n}$$

Si ha una stima corretta dell'errore di previsione perchè nel calcolo di $\hat{y}_{(-i)}$ non interviene y_i.

Questa tecnica è molto utilizzata nell'approccio moderno alla costruzione di modelli con buone capacità previsive. Qualche volta viene indicata con la sigla LOOCV (leave one out cross validation), perchè le osservazioni vengono levate dal campione ad una ad una (e poi rimesse...); altre tecniche più complesse prevedono l'esclusione di una frazione fissata di valori, ma non me ne occuperò in questo corso.

L'errore di Cross Validation introdotto può sembrare molto oneroso dal punto di vista computazionale, ma in realtà nel caso del modello lineare (e in molte altre situazioni ad esso riconducibili), non è necessario per calcolare gli scarti stimare n modelli lineari ciascuno con $n - 1$ dati.

Si può vedere, in questo e in altri problemi simili, che la somma dei quadrati di *cross validation* può essere espressa senza bisogno di fare i calcoli espliciti e

ricorrendo a quantità disponibili dopo la stima del modello originario:

$$nMSE(CV) = \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_{ii}} \right)^2$$

essendo h_{ii} l'$i-$esimo elemento diagonale della hat matrix $\mathbf{H}$

Invece di ricorrere alle proprietà delle matrici orlate ragioniamo in questo modo, ricordando le proprietà della *hat matrix* $\mathbf{H} = \mathbf{X} \left(\mathbf{X}^\mathsf{T} \mathbf{X} \right)^{-1} \mathbf{X}^\mathsf{T}$:

ossia $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ e

$\hat{y}_i = \Sigma_{j=1}^{n} h_{ij} y_i = h_{ii} y_i + \Sigma_{i \neq j} h_{ij} y_j$.

Ora consideriamo la stima dei parametri nel modello in cui manca la riga $i-$esima, ossia $\hat{\boldsymbol{\beta}}_{(-i)}$. Costruiamo ora un modello lineare in cui la riga $i-$esima (originariamente costituita da $(y_i, \mathbf{x}_i^\mathsf{T})$), viene rimpiazzata dalla riga $(\hat{y}_{(-i)}, \mathbf{x}_i^\mathsf{T}$. In questo nuovo modello lo stima dei minimi quadrati è sempre $\hat{\boldsymbol{\beta}}_{(-i)}$ (perchè il punto aggiunto $(\hat{y}_{(-i)}, \mathbf{x}_i^\mathsf{T}$ fornisce un residuo empirico nullo, in quanto $\hat{y}_{(-i)} = \mathbf{x}_i^\mathsf{T} \hat{\boldsymbol{\beta}}_{(-i)}$ ed è un punto che giace già sull'iperpiano adattato agli $n - 1$ dati).

Questo nuovo modello ha *la stessa matrice del disegno*$\mathbf{X}$ *del modello originale*, e in particolare per la riga $i-$esima, possiamo riapplicare la proprietà degli

elementi h_{ij}della *hat matrix* $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}$, considerando che in questo modello artificiale l' $i-$esimo valore stimato $(\hat{y}_{(-i)})$ è per costruzione uguale al valore osservato!

Quindi si ha nell'$i-$esimo modello costruito, riapplicando la formula vista per il modello completo:

$\hat{y}_{(-i)} = h_{ii}\hat{y}_{(-i)} + \Sigma_{i\neq j}\, h_{ij}y_j$, da cui:

$\Sigma_{i\neq j}\, h_{ij}y_j = (1 - h_{ii})\hat{y}_{(-i)}$ Ora sostituiamo questa espressione alla stessa sommatoria per il valore stimato nel modello completo:

$\hat{y}_i = h_{ii}y_i + \Sigma_{i\neq j}\, h_{ij}y_j = h_{ii}y_i + (1 - h_{ii})\hat{y}_{(-i)}$

da qui con pochi passaggi ricaviamo una nuova espressione del residuo empirico:

$$y_i - \hat{y}_i = y_i - h_{ii}y_i - (1 - h_{ii})\hat{y}_{(-i)} = (1 - h_{ii})(y_i - \hat{y}_{(-i)})$$

e quindi il risultato da dimostrare:

$$(y_i - \hat{y}_{(-i)}) = \frac{y_i - \hat{y}_i}{(1 - h_{ii})}$$

Questa relazione, utilissima per il calcolo, è molto importante anche dal punto di

vista statistico per far vedere che il residuo di cross-validation è sempre superiore al residuo ordinario, tanto più quanto maggiore h_{ii} *esempi*

10.6.2 Altri criteri di cross-validation

In alcuni contesti si usa anche un errore di cross-validation generalizzato:

$$GCV(\boldsymbol{\alpha}) = \frac{\Sigma_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2}{\left(1 - \frac{tr(\mathbf{H})}{n}\right)^2}$$

10.6.3 Insieme dati di test

Per problemi non lineari o comunque più complessi di quelli base trattati in questo corso, si utilizzano tecniche più complesse dal punto di vista computazionale, ma molto semplici concettualmente.

Suddividiamo i dati in sue insiemi: uno lo utilizziamo per stimare i parametri, l'altro per valutare la bontà dell'adattamento

-

Eventuale esempio in laboratorio

11 Esempio di correlazioni osservate fra molte variabili

Quando si rilevano molte variabili su n soggetti, in particolare in studi osservazionali, è possibile rilevare nella fase esplorativa delle correlazioni, sia semplici che multiple, anche molto consistenti, semplicemente per effetto di fluttuazioni campionarie dovute al cercare correlazioni empiriche alte in una matrice di correlazione con molti elementi.

Infatti si supponga per semplicità che la matrice $n \times p$ delle osservazioni costituisca un campione (multivariato) di ampiezza n proveniente da una distribuzione normale multivariata a p componenti indipendenti, e quindi con correlazioni lineari teoriche $\rho_{ij} = 0$; semplicemente per il fatto che nella matrice di correlazione stimata $p \times p$ si avranno $p(p-1)/2$ indici r_{ij} empirici di correlazione lineare, stime di massima verosimiglianza delle corrispondenti correlazioni lineari ρ_{ij} della popolazione multinormale di provenienza (sebbene tali $p(p-1)/2$ non siano indipendenti perché calcolate su p variabili):

Il più grande di tali indici chiaramente ha una distribuzione campionaria che non ha come valore atteso il valore teorico $\rho_{ij} = 0$.

Per un r_{ij} qualsiasi vale l'usuale trasformazione:

$$r_{ij}\sqrt{\frac{n-2}{1-r_{ij}^2}}$$

che si distribuisce come una t di student, con $n-2$ gradi di libertà, quando $\rho_{ij}=0$, tuttavia in questo caso stiamo *scegliendo* dalla matrice di correlazione *l'elemento* (o gli elementi) *più grande*, per cui non valgono i normali risultati sulla distribuzione di r_{ij}.

Esempio:

Da una distribuzione normale multivariata con 30 componenti indipendenti e standardizzate è stato estratto un campione di 100 osservazioni (la matrice dei dati è stata costruita per simulazione, ossia mediante generazione di numeri pseudo-casuali). Dal campione di osservazioni, con $n=100$ e $p=30$ è stata calcolata la matrice delle stime delle correlazioni lineari:

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
X1	1.00	-.07	-.03	-.13	.08	-.10	-.04	.06	.15	-.08	-.00	-.02	-.03	.00	-.03
X2	-.07	1.00	-.00	.03	.07	.07	.02	.04	-.00	.01	-.25	.05	-.14	.01	-.02
X3	-.03	-.00	1.00	.14	-.09	-.02	-.01	.08	.16	-.04	-.08	-.02	-.08	.04	-.05
X4	-.13	.03	.14	1.00	-.13	.07	.18	-.14	.18	-.12	-.05	-.03	-.09	.10	.15
X5	.08	.07	-.09	-.13	1.00	.02	.02	-.10	.06	.03	-.08	.01	-.02	-.17	.12
X6	-.10	.07	-.02	.07	.02	1.00	-.01	-.17	-.01	.00	.03	.13	.04	.02	.03
X7	-.04	.02	-.01	.18	.02	-.01	1.00	-.05	-.05	.09	-.04	-.04	.02	-.09	.02
X8	.06	.04	.08	-.14	-.10	-.17	-.05	1.00	.05	-.02	-.04	-.01	-.18	-.03	.03
X9	.15	-.00	.16	.18	.06	-.01	-.05	.05	1.00	.08	.09	-.20	-.05	-.01	.00
X10	-.08	.01	-.04	-.12	.03	.00	.09	-.02	.08	1.00	.16	-.18	-.01	.04	.14
X11	-.00	-.25	-.08	-.05	-.08	.03	-.04	-.04	.09	.16	1.00	-.23	.04	-.08	-.20
X12	-.02	.05	-.02	-.03	.01	.13	-.04	-.01	-.20	-.18	-.23	1.00	.25	.09	.05
X13	-.03	-.14	-.08	-.09	-.02	.04	.02	-.18	-.05	-.01	.04	.25	1.00	-.10	-.17
X14	.00	.01	.04	.10	-.17	.02	-.09	-.03	-.01	.04	-.08	.09	-.10	1.00	-.02
X15	-.03	-.02	-.05	.15	.12	.03	.02	.03	.00	.14	-.20	.05	-.17	-.02	1.00
X16	.08	.00	-.19	-.13	.05	.04	-.09	.03	-.18	-.03	-.03	.09	.25	-.01	.12
X17	.20	-.04	.05	.06	-.08	-.13	.14	-.05	.01	-.01	-.03	.00	.02	.02	.00
X18	-.02	-.06	.05	.03	-.03	-.13	-.09	.26	.10	-.07	.10	.08	-.00	.10	.01
X19	.22	-.04	-.08	.02	.01	.19	-.05	.02	-.09	-.13	.04	.21	.23	.00	.04
X20	.19	-.04	.08	-.13	.01	-.06	-.03	.23	.01	-.07	-.10	-.11	-.09	.15	.03
X21	.03	.06	-.11	-.09	.10	.06	.12	-.23	-.27	-.08	.04	.20	.09	-.02	-.17
X22	-.18	-.03	.14	.01	.12	-.05	.02	.12	-.13	.02	-.13	-.06	-.15	.17	.07
X23	.04	-.13	.04	.05	.04	-.18	.14	.10	.05	.08	.17	.19	-.11	-.11	.03
X24	-.05	.10	-.06	-.03	-.05	-.11	.13	.00	-.13	.06	-.01	.07	-.11	.01	.13
X25	.02	-.01	-.08	-.05	-.00	-.14	.08	-.09	-.08	-.14	-.11	.15	.06	.17	.01
X26	-.08	.16	-.12	-.01	.12	.13	-.10	-.16	-.06	.00	.13	-.07	.01	-.07	.02
X27	.00	.07	-.08	-.09	-.12	-.10	-.01	.05	.01	-.01	.00	.08	.11	.09	-.03
X28	-.02	-.03	.03	-.03	.12	-.22	.03	-.05	-.09	-.00	-.20	-.07	-.03	.02	.02
X29	.09	-.02	-.07	-.04	-.13	.06	-.03	-.06	-.14	-.17	.00	-.01	-.07	-.08	.04
X30	.10	-.01	.13	-.17	.08	-.14	-.05	-.06	.16	.03	-.11	.00	.03	.16	-.02

	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
X1	.08	.20	-.02	.22	.19	.03	-.18	.04	-.05	.02	-.08	.00	-.02	.09	.10
X2	.00	-.04	-.06	-.04	-.04	.06	-.03	-.13	.10	-.01	.16	.07	-.03	-.02	-.01
X3	-.19	.05	.05	-.08	.08	-.11	.14	.04	-.06	-.08	-.12	-.08	.03	-.07	.13
X4	-.13	.06	.03	.02	-.13	-.09	.01	.05	-.03	-.05	-.01	-.09	-.03	-.04	-.17
X5	.05	-.08	-.03	.01	.01	.10	.12	.04	-.05	-.00	.12	-.12	.12	-.13	.08
X6	.04	-.13	-.13	.19	-.06	.06	-.05	-.18	-.11	-.14	.13	-.10	-.22	.06	-.14
X7	-.09	.14	-.09	-.05	-.03	.12	.02	.14	.13	.08	-.10	-.01	.03	-.03	-.05
X8	.03	-.05	.26	.02	.23	-.23	.12	.10	.00	-.09	-.16	.05	-.05	-.06	-.06
X9	-.18	.01	.10	-.09	.01	-.27	-.13	.05	-.13	-.08	-.06	.01	-.09	-.14	.16
X10	-.03	-.01	-.07	-.13	-.07	-.08	.02	.08	.06	-.14	.00	-.01	-.00	-.17	.03
X11	-.03	-.03	.10	.04	-.10	.04	-.13	.17	-.01	-.11	.13	.00	-.20	.00	-.11
X12	.09	.00	.08	.21	-.11	.20	-.06	.19	.07	.15	-.07	.08	-.07	-.01	.00
X13	.25	.02	-.00	.23	-.09	.09	-.15	-.11	-.11	.06	.01	.11	-.03	-.07	.03
X14	-.01	.02	.10	.00	.15	-.02	.17	-.11	.01	.17	-.07	.09	.02	-.08	.16
X15	.12	.00	.01	.04	.03	-.17	.07	.03	.13	.01	.02	-.03	.02	.04	-.02
X16	1.00	-.05	.02	.26	-.02	.20	-.12	-.01	.11	-.02	-.14	.06	-.12	.14	.08
X17	-.05	1.00	.01	.10	.02	*.20*	-.20	-.08	.10	.16	-.15	-.05	-.02	-.11	.11
X18	.02	.01	1.00	-.01	-.11	-.16	.02	.01	.01	-.06	-.10	.14	.08	-.18	-.00
X19	.26	.10	-.01	1.00	-.03	.05	-.13	-.06	.10	.13	-.26	-.11	-.02	.00	-.05
X20	-.02	.02	-.11	-.03	1.00	-.13	.07	.02	.03	-.10	.05	-.10	.10	.12	.07
X21	.20	.20	-.16	.05	-.13	1.00	.14	.01	-.00	.23	.11	.11	-.06	-.08	-.06
X22	-.12	-.20	.02	-.13	.07	.14	1.00	.04	-.01	.12	.11	-.06	.22	-.18	.02
X23	-.01	-.08	.01	-.06	.02	.01	.04	1.00	.20	.05	-.20	-.16	.19	-.06	-.08
X24	.11	.10	.01	.10	.03	-.00	-.01	.20	1.00	.08	-.12	.12	.04	-.15	.02
X25	-.02	.16	-.06	.13	-.10	.23	.12	.05	.08	1.00	-.08	.01	.13	-.24	-.04
X26	-.14	-.15	-.10	-.26	.05	.11	.11	-.20	-.12	-.08	1.00	-.05	.04	.05	-.04
X27	.06	-.05	.14	-.11	-.10	.11	-.06	-.16	.12	.01	-.05	1.00	-.21	-.01	.07
X28	-.12	-.02	.08	-.02	.10	-.06	.22	.19	.04	.13	.04	-.21	1.00	.02	-.13
X29	.14	-.11	-.18	.00	.12	-.08	-.18	-.06	-.15	-.24	.05	-.01	.02	1.00	.02
X30	.08	.11	-.00	-.05	.07	-.06	.02	-.08	.02	-.04	-.04	.07	-.13	.02	1.00

verificare

	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15
X1	1.00	-.07	-.03	-.13	.08	-.10	-.04	.06	.15	-.08	-.00	-.02	-.03	.00	-.03
X2	-.07	1.00	-.00	.03	.07	.07	.02	.04	-.00	.01	*-.25*	.05	-.14	.01	-.02
X3	-.03	-.00	1.00	.14	-.09	-.02	-.01	.08	.16	-.04	-.08	-.02	-.08	.04	-.05
X4	-.13	.03	.14	1.00	-.13	.07	.18	-.14	.18	-.12	-.05	-.03	-.09	.10	.15
X5	.08	.07	-.09	-.13	1.00	.02	.02	-.10	.06	.03	-.08	.01	-.02	-.17	.12
X6	-.10	.07	-.02	.07	.02	1.00	-.01	-.17	-.01	.00	.03	.13	.04	.02	.03
X7	-.04	.02	-.01	.18	.02	-.01	1.00	-.05	-.05	.09	-.04	-.04	.02	-.09	.02
X8	.06	.04	.08	-.14	-.10	-.17	-.05	1.00	.05	-.02	-.04	-.01	-.18	-.03	.03
X9	.15	-.00	.16	.18	.06	-.01	-.05	.05	1.00	.08	.09	-.20	-.05	-.01	.00
X10	-.08	.01	-.04	-.12	.03	.00	.09	-.02	.08	1.00	.16	-.18	-.01	.04	.14
X11	-.00	-.25	-.08	-.05	-.08	.03	-.04	-.04	.09	.16	1.00	-.23	.04	-.08	-.20
X12	-.02	.05	-.02	-.03	.01	.13	-.04	-.01	-.20	-.18	-.23	1.00	.25	.09	.05
X13	-.03	-.14	-.08	-.09	-.02	.04	.02	-.18	-.05	-.01	.04	.25	1.00	-.10	-.17
X14	.00	.01	.04	.10	-.17	.02	-.09	-.03	-.01	.04	-.08	.09	-.10	1.00	-.02
X15	-.03	-.02	-.05	.15	.12	.03	.02	.03	.00	.14	-.20	.05	-.17	-.02	1.00
X16	.08	.00	-.19	-.13	.05	.04	-.09	.03	-.18	-.03	-.03	.09	.25	-.01	.12
X17	.20	-.04	.05	.06	-.08	-.13	.14	-.05	.01	-.01	-.03	.00	.02	.02	.00
X18	-.02	-.06	.05	.03	-.03	-.13	-.09	.26	.10	-.07	.10	.08	-.00	.10	.01
X19	.22	-.04	-.08	.02	.01	.19	-.05	.02	-.09	-.13	.04	.21	.23	.00	.04
X20	.19	-.04	.08	-.13	.01	-.06	-.03	.23	.01	-.07	-.10	-.11	-.09	.15	.03
X21	.03	.06	-.11	-.09	.10	.06	.12	-.23	-.27	-.08	.04	.20	.09	-.02	-.17
X22	-.18	-.03	.14	.01	.12	-.05	.02	.12	-.13	.02	-.13	-.06	-.15	.17	.07
X23	.04	-.13	.04	.05	.04	-.18	.14	.10	.05	.08	.17	.19	-.11	-.11	.03
X24	-.05	.10	-.06	-.03	-.05	-.11	.13	.00	-.13	.06	-.01	.07	-.11	.01	.13
X25	.02	-.01	-.08	-.05	-.00	-.14	.08	-.09	-.08	-.14	-.11	.15	.06	.17	.01
X26	-.08	.16	-.12	-.01	.12	.13	-.10	-.16	-.06	.00	.13	-.07	.01	-.07	.02
X27	.00	.07	-.08	-.09	-.12	-.10	-.01	.05	.01	-.01	.00	.08	.11	.09	-.03
X28	-.02	-.03	.03	-.03	.12	-.22	.03	-.05	-.09	-.00	-.20	-.07	-.03	.02	.02
X29	.09	-.02	-.07	-.04	-.13	.06	-.03	-.06	-.14	-.17	.00	-.01	-.07	-.08	.04
X30	.10	-.01	.13	-.17	.08	-.14	-.05	-.06	.16	.03	-.11	.00	.03	.16	-.02

	X16	X17	X18	X19	X20	X21	X22	X23	X24	X25	X26	X27	X28	X29	X30
X1	.08	.20	-.02	*.22*	.19	.03	-.18	.04	-.05	.02	-.08	.00	-.02	.09	.10
X2	.00	-.04	-.06	-.04	-.04	.06	-.03	-.13	.10	-.01	.16	.07	-.03	-.02	-.01
X3	-.19	.05	.05	-.08	.08	-.11	.14	.04	-.06	-.08	-.12	.08	.03	-.07	.13
X4	-.13	.06	.03	.02	-.13	-.09	.01	.05	-.03	-.05	-.01	-.09	-.03	-.04	-.17
X5	.05	-.08	-.03	.01	.01	.10	.12	.04	-.05	-.00	.12	-.12	.12	-.13	.08
X6	.04	-.13	-.13	.19	-.06	.06	-.05	-.18	-.11	-.14	.13	-.10	-.22	.06	-.14
X7	-.09	.14	-.09	-.05	-.03	.12	.02	.14	.13	.08	-.10	-.01	.03	-.03	-.05
X8	.03	-.05	.26	.02	.23	-.23	.12	.10	.00	-.09	-.16	.05	-.05	-.06	-.06
X9	-.18	.01	.10	-.09	.01	-.27	-.13	.05	-.13	-.08	-.06	.01	-.09	-.14	.16
X10	-.03	-.01	-.07	-.13	-.07	-.08	.02	.08	.06	-.14	.00	-.01	-.00	-.17	.03
X11	-.03	-.03	.10	.04	-.10	.04	-.13	.17	-.01	-.11	.13	.00	-.20	.00	-.11
X12	.09	.00	.08	.21	-.11	.20	-.06	.19	.07	.15	-.07	.08	-.07	-.01	.00
X13	.25	.02	-.00	.23	-.09	.09	-.15	-.11	-.11	.06	.01	.11	-.03	-.07	.03
X14	-.01	.02	.10	.00	.15	-.02	.17	-.11	.01	.17	-.07	.09	.02	-.08	.16
X15	.12	.00	.01	.04	.03	-.17	.07	.03	.13	.01	.02	-.03	.02	.04	-.02
X16	1.00	-.05	.02	.26	-.02	.20	-.12	-.01	.11	-.02	-.14	.06	-.12	.14	.08
X17	-.05	1.00	.01	.10	.02	.20	-.20	-.08	.10	.16	-.15	-.05	-.02	-.11	.11
X18	.02	.01	1.00	-.01	-.11	-.16	.02	.01	.01	-.06	-.10	.14	.08	-.18	-.00
X19	.26	.10	-.01	1.00	-.03	.05	-.13	-.06	.10	.13	-.26	-.11	-.02	.00	-.05
X20	-.02	.02	-.11	-.03	1.00	-.13	.07	.02	.03	-.10	.05	-.10	.10	.12	.07
X21	.20	.20	-.16	.05	-.13	1.00	.14	.01	-.00	.23	.11	.11	-.06	-.08	-.06
X22	-.12	-.20	.02	-.13	.07	.14	1.00	.04	-.01	.12	.11	-.06	.22	-.18	.02
X23	-.01	-.08	.01	-.06	.02	.01	.04	1.00	.20	.05	-.20	-.16	.19	-.06	-.08
X24	.11	.10	.01	.10	.03	-.00	-.01	.20	1.00	.08	-.12	.12	.04	-.15	.02
X25	-.02	.16	-.06	.13	-.10	.23	.12	.05	.08	1.00	-.08	.01	.13	-.24	-.04
X26	-.14	-.15	-.10	-.26	.05	.11	.11	-.20	-.12	-.08	1.00	-.05	.04	.05	-.04
X27	.06	-.05	.14	-.11	-.10	.11	-.06	-.16	.12	.01	-.05	1.00	-.21	-.01	.07
X28	-.12	-.02	.08	-.02	.10	-.06	.22	.19	.04	.13	.04	-.21	1.00	.02	-.13
X29	.14	-.11	-.18	.00	.12	-.08	-.18	-.06	-.15	-.24	.05	-.01	.02	1.00	.02
X30	.08	.11	-.00	-.05	.07	-.06	.02	-.08	.02	-.04	-.04	.07	-.13	.02	1.00

Inoltre nella tavola che segue sono riportati nella parte sinistra, per ciascuna variabile, la minima e la massima correlazione lineare r_{ij} ; nella parte destra si

riporta per ciascuna variabile, il coefficiente di determinazione multipla R^2 che esprime la porzione di variabilità spiegata dalla regressione multipla (lineare) su tutte le altre 29 variabili:

	Min. r_{ij}	Max r_{ij}		R^2 (Var. $\mathbf{X}_i$ con tutte le altre)
X1	-.18	.22	X1	.249
X2	-.25	.16	X2	.211
X3	-.19	.16	X3	.164
X4	-.17	.18	X4	.336
X5	-.17	.12	X5	.222
X6	-.22	.19	X6	.288
X7	-.10	.18	X7	.183
X8	-.23	.26	X8	.364
X9	-.27	.18	X9	.306
X10	-.18	.16	X10	.259
X11	-.25	.17	X11	.427
X12	-.23	.25	X12	.434
X13	-.18	.25	X13	.380
X14	-.17	.17	X14	.303
X15	-.20	.15	X15	.280
X16	-.19	.26	X16	.367
X17	-.20	.20	X17	.301
X18	-.18	.26	X18	.271

Ricordo che i valori critici di r ad un livello di significatività α per un test bilaterale sono:

$$r_\alpha = \sqrt{\frac{t_\alpha^2}{t_\alpha^2 + n - 2}}$$

essendo t_α il valore critico ad un livello α per una t con $n-2$ gradi di libertà.

Nel nostro caso, lavorando al 5

$$r_\alpha = \sqrt{\frac{1.9845^2}{1.9845^2 + 98}} = 0.197$$

Per quanto riguarda R^2 analogamente ricaviamo (dalla distribuzione F):

$$R_\alpha^2 = \frac{kF_\alpha}{kF_\alpha + n - k - 1}$$

essendo k il numero dei regressori e F_α il valore critico ad un livello α per una F di Snedecor con k ed $n-k-1$ gradi di libertà.

Nel nostro caso:

$$R^2_\alpha = \frac{29x1.6294}{29x1.6294 + 70} = 0,403$$

Di seguito sono riportati anche gli autovalori ricavati dalle 30 variabili standardizzate:

	C	omponenti principali	
i	Autovalore	varianza	varianza cumulata
1	2.300	7.668	7.67
2	1.999	6.662	14.33
3	1.925	6.417	20.75
4	1.690	5.634	26.38
5	1.621	5.402	31.78
6	1.560	5.200	36.98
7	1.529	5.098	42.08
8	1.429	4.764	46.85
9	1.332	4.440	51.29
10	1.206	4.021	55.31
11	1.135	3.784	59.09
12	1.105	3.682	62.77
13	1.009	3.363	66.14
14	.968	3.227	69.36
15	.899	2.996	72.36
16	.885	2.949	75.31
17	.854	2.845	78.15

12 Allontanamento dalle assunzioni di base nel modello lineare e nell'analisi della varianza: Analisi dei residui

13 Tipi di allontanamenti dalle assunzioni di base

Studio degli effetti dell'allontanamento dalle assunzioni di base del modello lineare sugli stimatori dei parametri, sulle stime di previsione e sui test.

Possibili allontanamenti dalle assunzioni di base:

- Nella componente sistematica:
  - Per esempio:
  - non linearità e/o non additività
  - irrilevante solo nell'AOV a 1 via
  - esclusione di variabili rilevanti (o di componenti polinomiali importanti)
  - esclusione di effetti di interazione particolari
  - esempio: AOV a due vie con $m = 1$
  - errore nella scala di misurazione della $\mathbf{y}$ e/o delle x (trasformazioni non lineari delle variabili)

– (questi tipi di allontanamento hanno degli aspetti in comune)

- nella componente accidentale:
  - non additività fra componente sistematica e accidentale
  - valore atteso non nullo (equivale alla non corretta specificazione del modello)
  - non normalità
  - esempio: errori distribuiti secondo una normale di ordine p.
  - oppure secondo un modello lineare generalizzato (GLM): coinvolge anche gli altri aspetti
  - sui momenti secondi:
  - eteroscedasticità
  - esempio: regressione ponderata; (varianze funzione dei valori attesi).
  - correlazione fra le componenti
  - esempio: dipendenza temporale; dipendenza territoriale.
  - in generale $V(\boldsymbol{\varepsilon}) \neq \sigma^2 \mathbf{I}_n$.

effetti su:

- proprietà stimatori
- non distorsione
- efficienza relativa
- valori previsti
- proprietà test
- livello di significatività effettivo
- potere del test
- Effetti della non normalità sulla distribuzione di F
- Effetti della non indipendenza
- Effetti della eteroscedasticità
- Trasformazioni
- Finalità delle trasformazioni nel modello lineare
- Tecniche alternative: test non parametrici e semiparametrici nella AOV

- Test non parametrici
- Test di permutazione

14 Analisi dei residui:

Adeguatezza del modello (estensione dell'analisi per modelli di regressione più generali).

Validità del legame funzionale ipotizzato.

Identificazione di eventuali fattori o variabili trascurati.

Eliminazione di variabili poco importanti.

identificazione di trasformazioni non lineari delle variabili, rispetto alle quali valgano le assunzioni di additività, normalità, indipendenza e omoscedasticità

Identificazione di allontanamenti dalle ipotesi di base per la distribuzione della componente accidentale

outlier (valore distante dalla maggioranza delle osservazioni) ? ==¿ campione eterogeneo, miscuglio di popolazioni

15 Aspetti peculiari dell'analisi dei residui:

Nella regressione: esame del legame funzionale fra la $\mathbf{y}$ e le $\mathbf{X}$; esame del range (eventualmente multivariato) all'interno del quale è plausibile l'ipotesi di linearità.

Nell'analisi della varianza: validità delle ipotesi riguardanti l'additività di effetti; identificazione delle particolari combinazioni di livelli dei fattori che portano interazioni significative

L'adeguatezza di un modello di regressione Ω_0, può essere saggiata più formalmente considerando un modello più ampio Ω_1, che fa ipotesi più generali sulla distribuzione degli errori e/o sul legame funzionale della componente sistematica; Ω_1 dovrebbe includere Ω_0 come caso particolare, fissando alcuni parametri di Ω_1 e costruendo i test relativi. ESEMPI

Si può costruire una differente famiglia di modelli e saggiarne l'adeguatezza mediante criteri basati sulla log-verosimiglianza

La costruzione di una qualsiasi famiglia di modelli presuppone che si abbiano delle idee precise sul tipo di allontanamento dalle assunzioni di base.

15.1 Definizione generale di residuo.

In generale su in un modello di regressione si ipotizza:

$$\mathbf{Y} = g(\mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\varepsilon})$$

con $g(\cdot)$ qualsiasi (ingloba sia il legame di $\mathbf{y}$ con i parametri che quello fra la componente sistematica e accidentale)

allora se $\boldsymbol{\beta}^*$ è lo stimatore di massima verosimiglianza, i valori stimati di $\mathbf{y}$ sono:

$$\mathbf{y}* = g(\mathbf{X}, \boldsymbol{\beta}^*, \boldsymbol{\varepsilon})$$

e quindi un residuo generale è:

r soluzione delle equazioni: $\mathbf{Y} = g(\mathbf{X}, \boldsymbol{\beta}^*, r)$

In particolare si ha, con errori indipendenti:

r_i soluzione dell'equazione:

$$\mathbf{y}_i = g(\mathbf{X}_{(i)}, \boldsymbol{\beta}^*, r_i)$$

15.2 Caratteristiche dei residui empirici nei modelli lineari

Se si considerano i residui empirici nel modello lineare, si ha che, indicando con e il vettore dei residui empirici:

$$\mathbf{e} = \mathbf{y} - \mathbf{Xb}$$

se il modello è correttamente specificato:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

i residui empirici sono allora esprimibili come:

$$\mathbf{e} = (\mathbf{y} - \mathbf{Xb}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \mathbf{Xb};$$

e quindi:

$$\mathbf{e} = \varepsilon + \mathbf{X}(\boldsymbol{\beta} - \mathbf{b}).$$

Se invece in realtà vi è una generica distorsione δ, dipendente da variabili escluse, da componenti non lineari trascurate, o comunque da una errata specificazione del modello di varia natura, e quindi se: $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} + \delta$, allora ovviamente si ha:

$$\mathbf{e} = \boldsymbol{\varepsilon} + \mathbf{X}(\boldsymbol{\beta} - \mathbf{b}) + \delta$$

Il residuo empirico è dunque composto sostanzialmente da tre componenti (non separabili):

Residuo empirico =	$\mathbf{e}=$
componente accidentale +	$\boldsymbol{\varepsilon}$ +
componente legata all'accuratezza degli stimatori +	$\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})+$
componente legata alla non corretta specificazione del modello	δ

Se il modello è correttamente specificato comunque il residuo è formato da due componenti: $\boldsymbol{\varepsilon}$ e $\mathbf{X}(\boldsymbol{\beta}-\mathbf{b})$. Con queste limitazioni il residuo empirico dà informazioni su $\boldsymbol{\varepsilon}$

Se valgono le assunzioni di base:

$$E(\boldsymbol{\varepsilon}) = 0, \mathbf{V}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I},$$

allora:

$$E(\mathbf{e}) = 0$$

peró, per la matrice di varianza e covarianza si ottiene:

$$V(\mathbf{e}) = E(\mathbf{e}\mathbf{e}^T) = E[(\mathbf{y} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y})(\mathbf{y} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y})^T] =$$
$$= \sigma^2(\mathbf{I}_n - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})$$

o più rapidamente (e ricordando l'idempotenza di $\mathbf{H}$ e di $[\mathbf{I} - \mathbf{H}]$):

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = [\mathbf{I} - \mathbf{H}]\mathbf{y} \Rightarrow \mathrm{V}[\mathbf{e}] = [\mathbf{I} - \mathbf{H}]\mathrm{V}[\mathbf{y}][\mathbf{I} - \mathbf{H}] =$$
$$= [\mathbf{I} - \mathbf{H}]\sigma^2\mathbf{I}[\mathbf{I} - \mathbf{H}] = \sigma^2[\mathbf{I} - \mathbf{H}]^2 = \sigma^2[\mathbf{I} - \mathbf{H}]$$

Pertanto i residui empirici risulteranno in generale correlati e con varianza teorica differente.

Se si vuole eliminare la diversa variabilità dei residui empirici si possono standardizzare gli stessi mediante gli elementi $\sigma^2(1 - h_{ii})$ sulla diagonale principale della matrice definita sopra.

Si ottengono così residui standardizzati (o residui studentizzati).

$h_{ii} \to 0$ al crescere di n (sotto condizioni piuttosto generali di regolarità).

casi particolari di matrice H

riprendere esempi su carta

esempio dell'AOV a una via

esempio della regressione semplice

15.3 Residui particolari

Poniamo:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$$

(detta *hat matrix* poichè è la matrice che trasforma $\mathbf{y}$ in ; $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$) indichiamo con h_i l'elemento diagonale $i-$esimo della matrice $\mathbf{H}$

residui studentizzati:

$$r_i = \frac{e_i}{s\sqrt{1-h_i}}$$

dal momento che $var(e_i) = \sigma^2(1-h_i)$

Sappiamo anche che

$$\sum_{i=1}^{n} h_i = k \qquad \text{e quindi} \qquad M(h) = \frac{k}{n}$$

Valori alti di h_i *(leverage values)* implicano una varianza bassa di e_i e quindi forzeranno $\hat{y}_i$ ad essere vicino a y_i

I valori di $\mathbf{H}$ *dipendono solo dai valori di* $\mathbf{X}$

15.3.1 qualche esempio di matrice H

Supponiamo di avere una matrice di regressori $\mathbf{Z}$ a media nulla (ed anche $\mathbf{y}$ a media nulla). Allora abbiamo:

$$\mathbf{H} = \mathbf{Z}\frac{S_{\mathbf{Z}}^{-1}}{n}\mathbf{Z}^{\mathsf{T}}$$

e quindi

$$h_i = \mathbf{z}_i^{\mathsf{T}}\frac{S_{\mathbf{Z}}^{-1}}{n}\mathbf{z}_i$$

se $k = 1$:

$$h_i = \frac{z_i^2}{n\sigma_x}$$

Nell'AOV a una via (n_i osservazioni per ogni gruppo) si ha

$$h_i = \frac{1}{n_i}$$

$$\hat{y}_{(i)} = x_i^{\mathsf{T}} \hat{\boldsymbol{\beta}}_{(i)}$$

(stima dell' $i-$esima osservazione basata sulle altre $n-1$)

Residuo di cross-validation:

$$y_i - \hat{y}_{(i)}$$

si può standardizzare con:

$$t_i = \frac{y_i - \hat{y}_{(i)}}{s.e(y_i - \hat{y}_{(i)})} = r_i \left(\frac{n-k-1}{n-p-r_i^2}\right)^{1/2}$$

Una misura globale (standardizzata) dell'influenza della simgola osservazione è data dalla *distanza di Cook*:

$$D_i = \frac{(y_i - \hat{y}_{(i)})^{\mathsf{T}} (y_i - \hat{y}_{(i)})}{k s^2}$$

Si può dimostrare che si ha:

$$D_i = r_i^2 \frac{h_i}{1 - h_i}$$

15.5 grafici dei residui empirici

Il modo migliore per avere informazioni sulla plausibilità delle assunzioni fatte sulla distribuzione di $\boldsymbol{\epsilon}$ è quello di esaminare la distribuzione dei residui empirici $e_i = \hat{y}_i - y_i$, pur con le avvertenze fatte nei paragrafi precedenti: va ancora ricordato che la difficoltà fondamentale nel fare ipotesi sulle $\boldsymbol{\epsilon}$ è che si tratta di *variabili aleatorie non osservabili*.

In ogni caso se il modello è non distorto si ha per i residui empirici:

$$\mathbf{e} = \boldsymbol{\varepsilon} + \underbrace{\mathbf{X}(\boldsymbol{\beta} - \mathbf{b})}_{\text{effetto accuratezza stimatore}}$$

e quindi le informazioni su $\boldsymbol{\varepsilon}$ sono sintetizzate in $\mathbf{e}$ a meno dell'effetto dovuto agli stimatori $\mathbf{X}$ ($\boldsymbol{\beta}$ -$\mathbf{b}$), in generale dell' ordine $O(n^{-1/2})$. Per valori grandi di n tale effetto sarà trascurabile, come visto a proposito della valutazione di $\mathbf{Ve}$

Possiamo ottenere differenti tipi di grafici riassunti schematicamente nella tabella:

ascissa	ordinata	impiego
valore predetto: $\hat{y}_i$	residuo: $e_i = \hat{y}_i - y_i$	allontanamenti di vario tipo dalle ipotesi di base
valore predetto: $\hat{y}_i$	(residuo)^2 $e_i^2 = (\hat{y}_i - y_i)^2$	evidenzia un'eventuale presenza di eteroschedasticità
(valore predetto)^2: $\hat{y}_i{}^2$	residuo: $e_i = \hat{y}_i - y_i$	serve per vedere eventualmente se è adeguata l'ipotesi di linearità
regressore j: x_{ij}	residuo: $e_i = \hat{y}_i - y_i$	serve per vedere eventualmente se vi sono indicazioni di non linearità rispetto al regressore j-esimo
regressore j: x_{ij}	deleted residual: d_i	serve per vedere eventualmente se vi sono indicazioni di non linearità rispetto al regressore j-esimo
residuo al tempo $i-1$: e_{i-1}	residuo: e_i	evidenzia l'eventuale presenza di autocorrelazione seriale (di intervallo 1) anche non lineare
normal probability		evidenzia l'allontanamento dalla

FIGURA DA FARE

Figure 1:

15.5.1 Variabile di risposta con poche modalità distinte: residui allineati su poche righe.

Il grafico ottenuto rappresentando i residui in corrispondenza dei valori predetti è costituito dai punti: $P_i : (\hat{y}_i, \hat{y}_i - y_i)$.

Se le osservazioni y_i assumono solo pochi valori distinti, diciamo s valori distinti, con $s < n$ nel grafico *residui contro valori predetti* i punti si disporranno lungo s rette; infatti se abbiamo ad esempio r_j osservazioni per ciascuna modalità y_i, i corrispondenti r_j punti $P_i : (\hat{y}_i, \hat{y}_i - y_i)$ si disporranno lungo la retta di equazione:

$$f(z) = z - y_i$$

15.5.2 Esempio

Nel grafico sono riportati i residui in funzione dei valori predetti della regressione lineare multipla fra la variabile "CRANIOCM"(Circonferenza cranica) e altre tre variabili antropometriche.

FIGURA DA FARE

Figure 2:

I punti si dipongono lungo 12 segmenti di rette parallele, perchè i valori distinti della variabile dipendente CRANIOCM sono solo 12, come riportato nell'istogramma.

16 Analisi della varianza: variabili esplicative qualitative

17 Analisi della varianza: modelli a rango non pieno

Per degli esempi di modelli lineari a rango non pieno è opportuno rivedere la parte introduttiva sui modelli lineari, in particolare per quanto riguarda le particolari configurazioni della matrice $\mathbf{X}$.

`Rivedere i problemi su dati reali già introdotti`

Se nel modello lineare la matrice $\mathbf{X}$ risulta a rango non pieno (**perché è stata volutamente strutturata in questo modo**), per ottenere le stime dei parametri non si può procedere nel modo ordinario.

Ad esempio si ottiene una matrice a rango non pieno se, per osservazioni suddivise in k gruppi, ipotizziamo i k valori attesi μ_j dati da $\mu_j = \mu + \eta_j$, ossia da una media generale più k effetti di gruppo. La matrice $\mathbf{X}$ corrispondente a questi $k + 1$ parametri risulterà singolare (si rivedano le configurazioni della matrice $\mathbf{X}$ nell'introduzione sui modelli lineari)

Possibili soluzioni:

- Riparametrizzazione

Per esempio nell'analisi della varianza ad una via esprimendo tutti i parametri in termini di contrasto di ciascuna media rispetto alla media di un particolare gruppo (di solito il primo o l'ultimo).

Prendendo come riferimento il primo gruppo:

$$\beta_j = \mu_j - \mu_1 \qquad j = 2, 3, k$$

Questo approccio è molto comodo quando la variabile esplicativa qualitativa è inserita in un modello di dipendenza lineare in cui è presente un termine noto (è il metodo di default di R)

- Aggiunta di vincoli sui parametri: ad esempio in un modello di analisi della varianza ad una via in cui si è usata, per comodità interpretativa, la parametrizzazione:

$$\mu_j = \mu + \eta_j, \qquad j = 1, 2, \ldots, k$$

si potrebbe aggiungere il vincolo:

$$\mu = \sum_{j=1}^{k} \mu_j n_j / n \qquad \text{che equivale a:} \qquad \sum_{j=1}^{k} \eta_j n_j = 0$$

- Modifica della matrice $\mathbf{X}$ in modo da eliminare la singolarità
- Uso dell'inversa generalizzata. Questo approccio è utile essenzialmente da un punto di vista teorico perché consente di esaminare alcune proprietà degli stimatori.

  Per far questo, indichiamo una delle soluzione dei minimi quadrati in modo generale, facendo ricorso all'inversa generalizzata (indicando con $\mathbf{A}^-$ l'inversa generalizzata [7]di $\mathbf{A}$):

  $$\mathbf{b} = (\mathbf{X}^\mathsf{T}\mathbf{X})^-\mathbf{X}^\mathsf{T}\mathbf{y}$$

  tenendo presente che stavolta questa soluzione non è unica e ricordando che comunque $\mathbf{b}$ è una soluzione di: $\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{b} = \mathbf{X}^\mathsf{T}\mathbf{b}$

  In effetti dal punto di vista del calcolo non conviene ricorrere all'inversa generalizzata, piuttosto conviene aggiungere delle altre equazioni, o lasciare non

[7] si veda eventualmente il concetto di inversa generalizzata nella parte dedicata al calcolo matriciale, ma comunque non è un approccio che userò ulteriormente in questa pagine

specificati alcuni parametri incogniti. In questo corso di norma non verrà impiegato questo approccio,

17.0.1 Funzioni stimabili

In effetti si può dimostrare che nei modelli a rango non pieno non tutte le funzioni lineari dei parametri sono *stimabili*; non è un argomento che tratto ulteriormente, ma basterà un esempio:

Se impostiamo un modello di analisi della varianza ad una via con $\mathrm{E}[Y_i j] = \mu + \eta_j$, se gli effetti η_j non sono vincolati, i parametri non saranno stimabili individualmente, ma soltanto i loro contrasti, p.e. $\eta_j - \eta_h$.

18 Analisi della varianza ad una via

Il modello più semplice di analisi della varianza si ha per il modello ad una via ad effetti fissi.

Effetti fissi e casuali

Per modello ad effetti fissi si intende un modello nel quale i parametri incogniti, come fatto fino ad ora, rappresentano delle costanti, sebbene non note.
In un modello ad effetti casuali invece i parametri, o almeno alcuni di essi, sono delle realizzazioni di variabili casuali, per le quali occorre ipotizzare un modello particolare.

18.0.1 Modelli ad effetti fissi:assunzioni di base

Si ha un modello ad una via quando le n osservazioni sono suddivise in k gruppi secondo le k modalità di un criterio di classificazione in generale qualitativo. Se il

criterio di classificazione è quantitativo esiste la possibilità di effettuare analisi più forti di quelle che vengono esposte in queste pagine, tenendo presenti le tecniche di regressione lineare, come si vedrà più avanti.

Il modello per $\mathbf{y}$:

$$\mathbf{y}_i = \mu_j + \varepsilon_i \qquad i = 1, 2, \ldots, n$$

Il modello per le medie:

$$\mu_j = \mu + \eta_j \qquad j = 1, 2, \ldots, k$$

media gruppo j = media generale + effetto gruppo j.

Il modello è detto a effetti fissi perché si suppone che le $\mu_j$siano dei parametri fissi, sebbene incogniti, relativi a k particolari gruppi.

In definitiva il modello per le osservazioni diventa:

$$\mathbf{y}_i = \mu + \eta_j + \varepsilon_i \qquad i = 1, 2, \ldots, n \quad j = 1, 2, \ldots, k$$

Osservazione = media generale + effetto gruppo j+errore accidentale

In effetti, per i motivi anticipati prima, occorre fissare un vincolo sui parametri, dato che questa parametrizzazione (1 media generale + k effetti dei gruppi) ha introdotto un nuovo parametro; infatti anche se i parametri adesso sono $k + 1$,

in realtà la parte sistematica del modello è chiaramente dipendente soltanto da k parametri degli effetti medi.

Il vincolo che si impiega è:

$$\mu = \frac{\Sigma_{j=1}^{k} \mu_j n_j}{n} \qquad \text{e quindi} \qquad \sum_{j=1}^{k} \eta_j n_j = 0$$

Non è una scelta univoca, ma presenta il vantaggio che, qualora si ipotizzi assenza di effetti, la stima di μ sotto H_0 coincide con quella sotto H_1. Diversamente si potrebbe pensare ad un sistema di vincoli più naturale, indipendente dalle ampiezze campionarie, con pesi uguali:

$$\mu = \frac{\Sigma_{j=1}^{k} \mu_j}{k} \quad \text{e quindi} \qquad \sum_{j=1}^{k} \eta_j = 0.$$

Scopo dell'analisi è quello di fare inferenza sulle relazioni che intercorrono fra le k medie delle popolazioni. In particolare si può essere interessati a saggiare l'ipotesi che le k medie siano tutte uguali, contro un'alternativa generica che almeno una sia diversa dalle altre: ipotesi di omogeneità:

$$H_0 : \mu_j = \mu \qquad j = 1, 2, \ldots, k$$

(ossia **Le medie sono uguali ad un valore comune non specificato**)
equivalente a:

$$H_0 : \eta_j = 0 \qquad j = 1, 2, \ldots, k,$$

(ossia **li effetti di gruppo sono tutti nulli**)

questo tipo di impostazione, in termini di effetti, è utile in particolare con disegni di analisi della varianza a più vie.

In questo caso si ipotizza sempre un vincolo sui parametri:
$\Sigma_{j=1}^{k} \eta_j n_j = 0$, in modo che questi risultino stimabili.

In questo modo si vede che in entrambi i casi l'ipotesi nulla fissa $k - 1$ vincoli sui parametri. L'ipotesi alternativa H_1 consiste temporaneamente nella semplice negazione di H_0. Impostato in questo modo il problema, si tratta banalmente di *un'estensione a k medie del test t per il confronto di due medie mediante due campioni indipendenti.*

Assunzioni fatte sulla componente accidentale:

In corrispondenza delle ipotesi di base sulla distribuzione degli errori già fatte nel contesto generale dei modelli lineari, si hanno le seguenti assunzioni nell'analisi della varianza ad una via:

$$\varepsilon_i (N(0, \sigma^2);$$

e quindi

$$\mathbf{y}_i \sim N(\mu_j, \sigma^2) \qquad \text{con} \qquad \mathbf{U}_i \in G_j j = 1, 2, \ldots, k$$

Ipotesi:

- $\varepsilon_i, \varepsilon_r$ indipendenti (sia se $\mathbf{U}_i$ e $\mathbf{U}_r$ appartengono allo stesso gruppo sia nel caso i gruppi siano diversi); è utile vedere adesso questa ipotesi generale di indipendenza scomposta in due assunzioni particolari, con riferimento ai k campioni:
  - gli errori sono indipendenti all'interno dei k campioni (ciascun gruppo è un campione casuale semplice);
  - i k campioni sono indipendenti.
- Omoscedasticità;
- Ipotesi di normalità.

(in effetti si potrebbe semplicemente assumere la non correlazione, e l'indipendenza scaturirebbe una volta fatta anche l'ipotesi di normalità)

Si noti che le ultime due assunzioni (normalità e omoscedasticità) riguardano esclusivamente le k popolazioni (o universi) teoriche e non hanno relazione con il campionamento; le prime due assunzioni riguardano invece le relazioni fra le unità e fra i campioni e sono quindi collegati essenzialmente al meccanismo di acquisizione dei dati. (in effetti, però, un qualche collegamento fra i due gruppi di assunzioni esiste se per esempio si assume soltanto la non correlazione fra le osservazioni entro i campioni: se si assume anche la normalità, questa implica l'indipendenza).

In pratica si sta ipotizzando che il modello da cui provengono i dati è del tipo rappresentato nella figura che segue (se è vera l'ipotesi nulla):

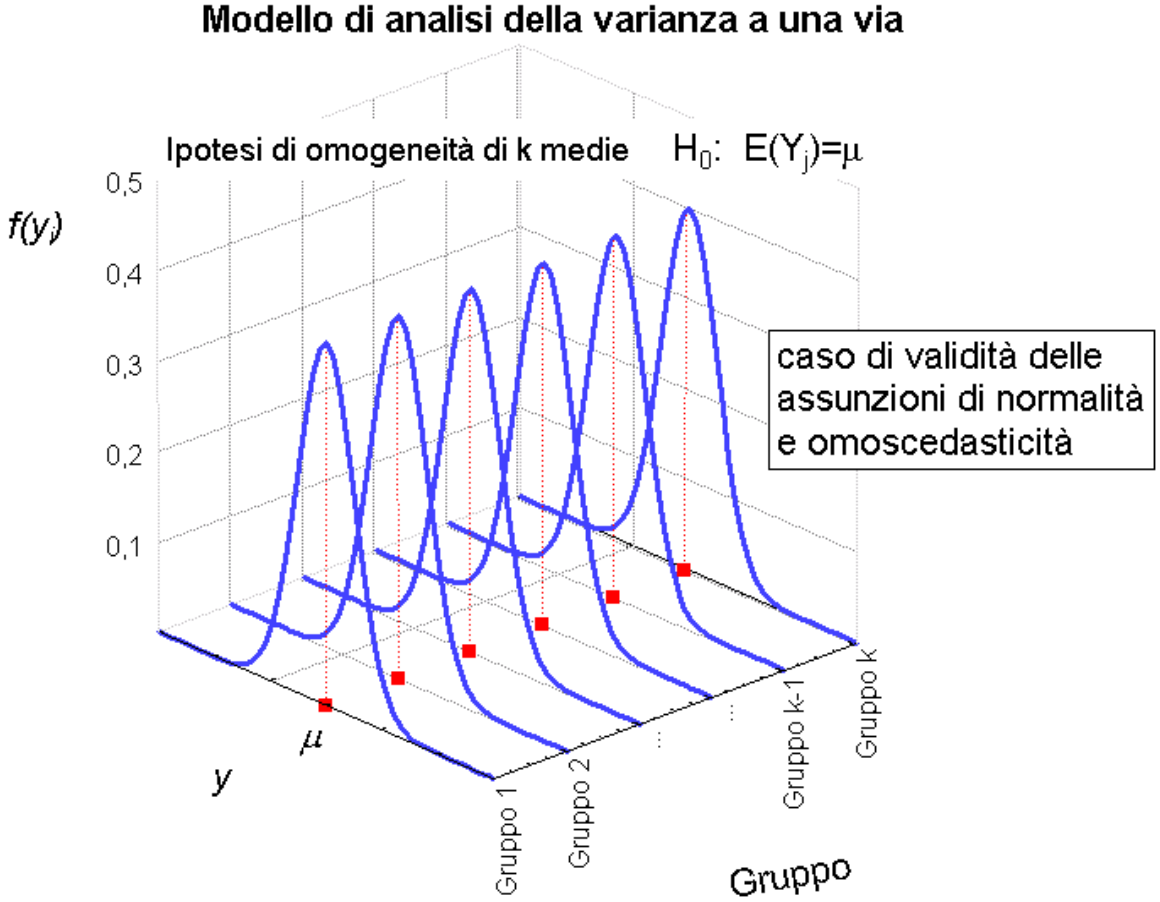

Figure 3: Omeogeneità delle medie

Se invece non vale l'ipotesi nulla di omogeneità delle medie, si potrà avere una situazione come quella della figura che segue:

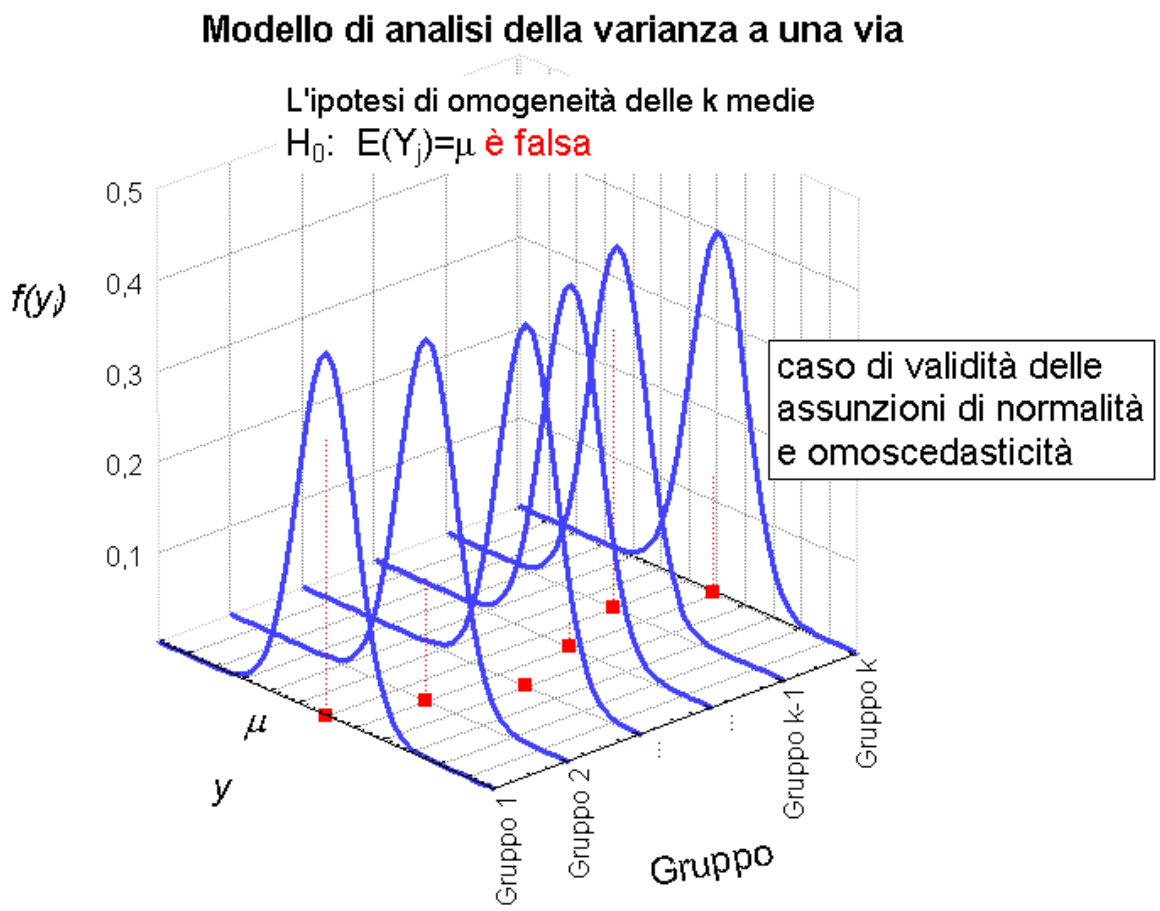

Figure 4: Eterogeneità delle medie

In ogni caso si tratta di k popolazioni, normali, con la stessa varianza, per le quali l'eventuale effetto sperimentale (ossia l'appartenenza ad un particolare gruppo) comporta in sostanza semplicemente uno slittamento dell'intera distribuzione (o di più distribuzioni).

18.1 Modelli ad effetti fissi: allontanamenti dalle assunzioni di base

In definitiva, come si è visto, se valgono le assunzioni di base, H_0 consiste nell'ipotesi che le k popolazioni siano in tutto uguali, ossia che *i k campioni in realtà si possono considerare provenienti da un'unica popolazione*.

Se adesso supponiamo delle assunzioni meno rigide, possiamo ovviamente pensare di non far valere (o di rendere più flessibili) una o più delle quattro ipotesi.

Qui accenno soltanto alcune delle implicazioni poste dall'allontanamento dall'ipotesi di normalità o dall'ipotesi di omoscedasticità, lasciando ad altre sezioni la discussione sull'ipotesi di indipendenza entro e fra i campioni.

18.1.1 Eteroscedasticità (varianze non omogenee)

Evidentemente adesso, ammettendo comunque la normalità, anche se H_0 è vera, le k popolazioni non saranno uguali; perché possono avere comunque delle varianze differenti. Il problema di prova delle ipotesi più semplice, ossia quello specificato da H_0, non implica più quindi l'uguaglianza di k popolazioni, ma soltanto l'uguaglianza di k effetti medi.

Se il criterio di classificazione corrisponde a k trattamenti sperimentali, questa situazione corrisponde ad ammettere che i trattamenti possano provocare un alterazione anche nella variabilità fra le unità sperimentali.

In pratica il modello assunto è:

$$\mathbf{y}_i(N(\mu_j, \sigma_j^2) \qquad \text{con} \qquad \mathbf{U}_i \in G_j j = 1, 2, \ldots, k$$

Ovviamente le varianze σ_j^2 andranno stimate dai singoli campioni, e costituiscono quindi degli ulteriori parametri di disturbo nell'inferenza sugli effetti medi; è noto inoltre che nella costruzione di un test già nel caso di due campioni si ricade nel problema di Beherens -Fisher.

Tuttavia è sempre possibile verificare preliminarmente l'adeguatezza dell'assunzione

FIGURA DA FARE

Figure 5: Eteroscedasticità e medie uguali

di omoscedasticità, per campioni costituiti da osservazioni ripetute provenienti da k popolazioni normali. Il test più noto è il test di Bartlett, basato sul rapporto fra le verosimiglianze, corretto per migliore l'approssimazione ad una chi-quadro.

18.1.2 Non normalità

Occorrerebbe qua fare numerosissime distinzioni, perché si hanno ovviamente infinite forme di allontanamento dalla normalità. Qui elenco alcune delle situazioni più plausibili:

- k popolazioni non normali ma dello stesso tipo e tutte note
- k popolazioni non normali appartenenti alla stessa famiglia parametrica, e dipendenti da uno o più parametri incogniti.
  - Ad esempio le popolazioni potrebbero essere delle normali di ordine p o delle uniformi

  - Oppure potrebbero essere k distribuzioni gamma (con un parametro da stimare)
  - Oppure potrebbero essere k distribuzioni esponenziali
- k popolazioni non normali appartenenti ad un'unica famiglia parametrica non nota.
- k popolazioni non normali appartenenti a diverse famiglie parametriche

18.2 Ipotesi di omogeneità delle medie: stimatori e test corrispondenti.

Le stime di massima verosimiglianza dei parametri, in assenza di vincoli sui parametri stessi ossia sotto H_1, si ottengono molto semplicemente considerando che i k campioni sono indipendenti e sono costituiti da osservazioni indipendenti provenienti da universi normali. Quindi è ovvio che tali stimatori sono le medie aritmetiche M_j delle n_j osservazioni relative a ciascun campione $(j = 1, 2, \ldots, k)$.

Tuttavia, se si vuole lavorare con la tecnica dei modelli lineari generali, occorre considerare:

la matrice $\mathbf{X}$, che è ora costituita dalle k colonne di appartenenza delle n unità

ai k gruppi:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

la matrice $\mathbf{X}^\top\mathbf{X}$ è chiaramente costituita da una matrice diagonale, con elemento generico sulla diagonale dato da n_j (si riveda lo schema del la matrice $\mathbf{X}$ riportato nell'introduzione dei modelli lineari, e si verifichi tale relazione: si consideri che le colonne di $\mathbf{X}$ sono ortogonali, per cui gli unici elementi non nulli nel prodotto $\mathbf{X}^\top\mathbf{X}$, sono quelli corrispondenti agli elementi diagonali), per cui:

$$\mathbf{X}^\mathsf{T}\mathbf{X} = Diag(n_1, n_2, \ldots, n_k) = \begin{pmatrix} n_1 & & & \\ & n_2 & & \\ & & \ldots & \\ & & & n_k \end{pmatrix}$$

Il vettore $\mathbf{X}^\mathsf{T}\mathbf{y}$ è costituito dalle k somme delle osservazioni per ciascun gruppo, ossia

$$\mathbf{X}^\mathsf{T}\mathbf{y} = (n_1 M_1, n_2 M_2, \ldots, n_k M_k)^\mathsf{T} = \begin{pmatrix} n_1 M_1 \\ \ldots \\ n_j M_j \\ \ldots \\ n_k M_k \end{pmatrix}$$

per cui in definitiva si ha:
stime di massima verosimiglianza nell'analisi della varianza ad una via:

$$\mathbf{b} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} =$$

$$= Diag(n_1, n_2, \ldots, n_k)^{-1}(n_1 M_1, n_2 M_2, \ldots, n_k M_k)^{\mathsf{T}} =$$

$$\begin{pmatrix} n_1 & & & \\ & n_2 & & \\ & & \ddots & \\ & & & n_k \end{pmatrix}^{-1} \begin{pmatrix} n_1 M_1 \\ \ldots \\ n_j M_j \\ \ldots \\ n_k M_k \end{pmatrix} = \begin{pmatrix} M_1 \\ \ldots \\ M_j \\ \ldots \\ M_k \end{pmatrix}$$

La stima delle μ_j è dunque data dalle stime di massima verosimiglianza di ciascun campione M_j

La stima di massima verosimiglianza di σ^2 si ottiene nel modo usuale con la tecnica generale adottata nei modelli lineari, ossia dalla devianza residua (relativa al modello completo) diviso i gradi di libertà corrispondenti.

18.2.1 M.Q . vincolati: Analisi della varianza ad una via.

La matrice $\mathbf{X}$ è composta da k colonne indicatrici dell'appartenenza delle n unità a k gruppi disgiunti. La parametrizzazione più naturale è quella in cui ogni parametro corrisponde al valor medio di $\mathbf{Y}$ in ciascun gruppo (link con intro mod. li

):

$$\boldsymbol{\beta}^{\mathsf{T}} = \mu_1, \ldots, \mu_j, \ldots, \mu_k$$

L'ipotesi di interesse è:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k$$

Queste $k-1$ uguaglianze corrispondono ad una scelta di $\mathbf{C}$ di $k-1$ righe e k colonne:

$$\mathbf{C}_{[k-1\times k]} = \left(\begin{array}{ccccccc} \underline{Gr.1} & \underline{Gr.2} & \ldots & \underline{Gr.J} & \ldots & \underline{Gr.K} \\ 1 & 0 & \ldots & 0 & \ldots & -1 \\ \ldots & 1 & \ldots & \ldots & \ldots & -1 \\ 0 & 0 & \ldots & 0 & 0 & -1 \\ 0 & 0 & \ldots & 1 & 0 & -1 \\ \ldots & \ldots & \ldots & \ldots & \ldots & -1 \\ 0 & 0 & 0 & 0 & \ldots & -1 \end{array}\right) \left[\begin{array}{c} \underline{vincolo} \\ 1 \\ 2 \\ \ldots \\ j \\ \ldots \\ k-1 \end{array}\right]$$

con $\mathbf{a} = 0_{k-1}$

Scrivere ora $\mathbf{C}\boldsymbol{\beta} = \mathbf{a}$ è come scrivere:

$$\mu_1 - \mu_k = \mu_2 - \mu_k = \ldots = \mu_j - \mu_k = \ldots = \mu_{k-1} - \mu_k = 0.$$

Che costituiscono $(k-1)$ vincoli sui parametri;

Occorre trovare la soluzione di massima verosimiglianza condizionatamente a tali vincoli (lineari).

In effetti per la stima dei parametri sotto H_0 non conviene ricorrere alla tecnica generale di stima con vincoli lineari qualsiasi, ma piuttosto ad un approccio diretto.

Infatti la matrice $\mathbf{X}$ nel modello specificato da H_0 è composta da una colonna di n valori uguali ad 1;

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \dots \\ \dots \\ 1 \end{pmatrix}$$

per cui è facile verificare che la soluzione è data da:

$$\mathbf{b}_0 = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} = \sum_{i=1}^{n} \mathbf{y}_i/n = M(\text{Media generale})$$

D'altra parte è ovvio che sotto l'ipotesi nulla la stima di μ è data dalla media generale, in quanto in questo caso l'ipotesi specifica che la popolazione di provenienza

è un'unica distribuzione normale, e quindi deriva l'usuale risultato sullo stimatore di massima verosimiglianza. La stima di σ^2 sarà adesso costituita dalla devianza residua sotto H_0, divisa per i gradi di libertà corrispondenti.

18.2.2 Scomposizione della varianza.

Secondo l'approccio generale scomponiamo la devianza della componente accidentale, $R(\boldsymbol{\beta})$. Si riveda eventualmente tale parte nei modelli lineari.

Impiego qui la notazione $\mathbf{y}_{ij}$ per indicare la i-esima osservazione del j-esimo gruppo (diversa rispetto alla precedente convenzione, tuttavia questa notazione risulta più utile quando, come adesso, un problema che è formalmente inquadrabile nell'ambito dei modelli lineari, è naturalmente interpretabile anche come confronto fra k popolazioni diverse attraverso k campioni, da cui l'esigenza del doppio indice, uno per le unità e l'altro per i gruppi)

$$R(\boldsymbol{\beta}) = R(\mathbf{b}) + (\mathbf{b} - \boldsymbol{\beta})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})$$

Oppure :

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{b} - \boldsymbol{\beta})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})$$

che nel nostro caso diventa:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mathbf{y}_{ij}-\mu_j)^2=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mathbf{y}_{ij}-M_j)^2+\sum_{j=1}^{k}n_j(\mu-M_j)^2$$

Con riferimento alle devianze residue si ha:

devianza residua sotto H_1 :

$$R(\mathbf{b})=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mathbf{y}_{ij}-M_j)^2$$

devianza residua sotto H_0 :

$$R(\mathbf{b}_0)=\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mathbf{y}_{ij}-M)^2$$

e ricordando che $\mathbf{b}$ è costituito dal vettore delle k medie M_j e che $\mathbf{b}_0$ è invece costituito dalla media generale M si ha:

Scomposizione ricavata dal caso di ipotesi di vincoli lineari qualsiasi della devianza sotto H_0 (rivedere eventualmente)	$R(\mathbf{b}_0) =$ $= R(\mathbf{b}) + (\mathbf{b} - \mathbf{b}_0)^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b} - \mathbf{b}_0) =$
devianza residua sotto H_1	$= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - M_j)^2 +$
devianza residua aggiuntiva dovuta ad H_0	$+ \sum_{j=1}^{k} n_j (M_j - M)^2$

Si ha quindi la scomposizione fondamentale:

Scomposizione della devianza

devianza totale = devianza entro i gruppi+devianza fra i gruppi

Per la stima della varianza ovviamente si ha:

Stima della varianza nell'Analisi della varianza ad una via

$$s^2(n-k)/\sigma^2 \sim \chi^2_{n-k}$$

qualunque sia l'ipotesi vera

$$s^2 = \frac{\text{devianza entro i gruppi}}{n-k} = \frac{\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}(\mathbf{y}_{ij} - M_j)^2}{n-k}$$

s^2 è sempre una stima corretta della varianza

TAVOLA DI SCOMPOSIZIONE DELLA DEVIANZA EMPIRICA

FONTE	DEVIANZA	g.d.l.	Valore atteso	Stima della varianza
TOTALE	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mathbf{y}_{ij}-M)^2$	$n-1$		$\frac{\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}(\mathbf{y}_{ij}-M)^2}{n-1}$
ENTRO I GRUPPI	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mathbf{y}_{ij}-M_j)^2$	$n-k$	$(n-k)\sigma^2$	$\frac{\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}(\mathbf{y}_{ij}-M_j)^2}{n-k}$
FRA I GRUPPI	$\sum_{j=1}^{k} n_j(M_j-M)^2$	$k-1$	$(k-1)\sigma^2+\sum_{j=1}^{k} n_j\eta_j^2$	$\frac{\Sigma_{j=1}^{k} n_j(M_j-M)^2}{k-1}$

Test F per la verifica dell'ipotesi di omogeneità delle medie

$$F = \frac{\frac{\sum_{j=1}^{k} n_j (M_j - M)^2}{k-1}}{\frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - M_j)^2}{n-k}}$$

$(k-1)$ e $(n-k)$ gradi di libertà

Secondo quanto già visto nell'analisi del modello lineare generale, sotto l'ipotesi nulla di omogeneità fra le medie, questo rapporto F si distribuisce come una v.c. F di Snedecor con $(k-1)$ e $(n-k)$ gradi di libertà. La distribuzione di F sotto H_1 é quella di una F non centrale con $(k-1)$ e $(n-k)$ gradi di libertà, e parametro di non centralità $\Sigma_{j=1}^{k} n_j \eta_j^2$.

E' facile vedere che esiste una relazione fra questo rapporto F ed il rapporto di correlazione, $\eta_{yx} = \frac{\sigma_{my}}{\sigma_{\mathbf{y}}}$ (o meglio il suo quadrato η_{yx}^2) impiegato come indice per misurare, in una tavola di contingenza con un carattere quantitativo $\mathbf{y}$ ed

un carattere qualsiasi $\mathbf{X}$, la dipendenza in media della variabile quantitativa $\mathbf{y}$ dalla variabile (qualitativa o quantitativa) $\mathbf{X}$. L'indice η^2_{yx}, che varia fra 0 ed 1, infatti é costruito come rapporto fra σ^2_{my} varianza tra le medie parziali di $\mathbf{y}$ e $\sigma^2_{\mathbf{y}}$, varianza totale di $\mathbf{y}$, ossia sempre quantità che compaiono nella tavola di analisi della varianza.

Ricordando che:

$$\sigma^2_{my} = \frac{\sum_{j=1}^{k} n_j (M_j - M)^2}{n}$$

$$\sigma^2_{\mathbf{y}} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - M)^2}{n}$$

$$\text{e} \qquad \eta^2_{yx} = \frac{\sigma^2_{my}}{\sigma^2_{\mathbf{y}}}$$

si ha:

$$F = \frac{\frac{\sum_{j=1}^{k} n_j (M_j - M)^2}{k-1}}{\frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - M_j)^2}{n-k}} = \frac{\frac{\sum_{j=1}^{k} n_j (M_j - M)^2}{k-1}}{\frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - M)^2 - \sum_{j=1}^{k} n_j (M_j - M)^2]}{n-k}}.$$

Dividendo ora numeratore e denominatore per $n\sigma^2_{\mathbf{y}}$, si ha subito:

$$\mathrm{F} = \frac{\frac{\sigma^2_{my}}{(k-1)\sigma^2_{\mathbf{y}}}}{\left(1 - \frac{\sigma^2_{my;}}{\sigma^2_{\mathbf{y}}}\right)/(k-1)} = \frac{[n-k]\eta^2_{yx}}{[k-1](1-\eta^2_{yx})}$$

18.2.3 Formule per il calcolo

Per il calcolo dei momenti con origine la media aritmetica, è possibile calcolare le tre devianze necessarie per la costruzione della tavola esprimendo le somme dei quadrati degli scarti (momenti con origine la media) in funzione di somme di quadrati (momenti con origine zero); riassumo di seguito tutte le formule utili per il calcolo delle quantità necessarie nell'analisi della varianza a una via.

Formule elementari per l'analisi della varianza ad una via

Numerosità totale	$N = \sum_{j=1}^{k} n_j$	
Media del gruppo j	$M_j = \frac{\Sigma_{i=1}^{n_j} \mathbf{y}_{ij}}{n_j}$	$j = 1, 2, \ldots, k$
Media generale =	$M \frac{\Sigma_{j=1}^{k} \Sigma_{i=1}^{n_j} \mathbf{y}_{ij}}{N} =$	$\frac{\Sigma_{j=1}^{k} n_j M_j}{N}$

Devianza TOTALE =	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mathbf{y}_{ij}-M)^2=$	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\mathbf{y}_{ij}^2-NM^2$
ENTRO I GRUPPI:	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(\mathbf{y}_{ij}-M_j)^2=$	$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\mathbf{y}_{ij}^2-\sum_{j=1}^{k}n_jM_j^2$
FRA I GRUPPI	$\sum_{j=1}^{k}n_j(M_j-M)^2$	$\sum_{j=1}^{k}n_jM_j^2-NM^2$

Le tre quantità essenziali per il calcolo delle devianze interne sono dunque:

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\mathbf{y}_{ij}^2; \qquad \sum_{j=1}^{k}n_jM_j^2=\sum_{j=1}^{k}\frac{\left(\Sigma_{i=1}^{n_j}y_{ij}\right)^2}{n_j}$$

$$NM^2 = \frac{\left(\Sigma_{j=1}^{k} \Sigma_{i=1}^{n_j} y_{ij}\right)^2}{N}$$

Chiaramente queste formule sono utili per il calcolo manuale delle devianze, mentre è ovvio che nella pratica si fa uso di software statistico, programmato da sé, presente in software commerciale o meglio OpenSource, con il quale presumibilmente (e auspicabilmente) lo studente è già familiare, e che senz'altro è indispensabile nella pratica quotidiana di soluzione di problemi reali.

In effetti, per esperienza personale, suggerisco allo studente di effettuare qualche esercizio di analisi della varianza (magari con insiemi di dati poco numerosi, soltanto con qualche decina di dati) esclusivamente con una calcolatrice da tavolo, impiegando le formule sopra riportate.

Sebbene io abbia volutamente trascurato di affrontare problemi di tipo numerico e/o computazionale, va detto che le formule sopra riportate presentano il vantaggio di leggere i dati una sola volta, il che risulta utile per insiemi di dati molto numerosi o registrati su supporti a bassa velocità di accesso; questa maggiore velocità di calcolo viene compensata da una possibile perdita in precisione numerica, dal momento che le somme di quadrati conterranno numeri più grandi che non le

somme di quadrati di scarti.

Considerazioni simili valevano anche per il calcolo dei momenti bivariati (covarianze) necessari per il calcolo della matrice di correlazione.

Un modo conveniente di memorizzare i dati è quello di scrivere in una colonna tutte le osservazioni campionarie $\mathbf{y}_{ij}$, ed in una colonna affiancata un numero, una lettera o anche una sigla alfanumerica identificativa del gruppo di appartenenza.

8

In effetti il più delle volte li si troverà già in questa forma, all'interno di databases con un numero di colonne anche molto maggiore di due, perchè la regola è quella di avere a che fare con osservazioni multivariate!

INSERIRE ESEMPIO

18.2.4 L'analisi della varianza come confronto fra stime di varianze

Presento in questo paragrafo un modo leggermente diverso di impostare l'analisi della varianza ad una via, direttamente come problema di confronto fra varianze campionarie, che mette in luce il ruolo fondamentale dello studio della variabilità fra i gruppi per analizzare l'eterogeneità di un gruppo di medie campionarie. Questa impostazione ci permetterà fra breve di fare anche qualche considerazione

[8] Diffidate di files di k righe) in cui i dati di ciascun gruppo sono memorizzati in una diversa riga: è un formato molto poco versatile e conviene cambiarlo subito nel formato colonnare in n righe e due o più colonne.

sull'adeguatezza della stima della varianza della componente accidentale.

Se l'ipotesi nulla di omogeneità delle medie è vera posso stimare la varianza in due modi diversi:

mediante le singole osservazioni, attraverso la solita quantità:

$$s^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - M)^2/(n-k)$$

(che è una stima corretta della varianza anche sotto H_1); mediante il campione di k medie (ma solo sotto H_0); infatti supponiamo per semplicità che i k campioni siano tutti di numerosità $n_j = m$. Allora ciascuna media M_j è una variabile casuale con distribuzione normale di parametri μ_j e σ^2/m . Se però H_0 è vera le k medie provengono tutte dalla stessa popolazione normale di parametri μ e σ^2/m , e quindi costituiscono un campione casuale semplice (di ampiezza k) da una normale, per cui si può stimare il parametro σ^2/m attraverso la varianza campionaria:

$$s_M^2 = \sum_{j=1}^{k} (M_j - M)^2/(k-1)$$

E' facile vedere che ms_M^2 è uno stimatore di σ^2 e che quindi il rapporto

$$ms_M^2/s^2$$

(che è proprio il rapporto F visto nelle pagine precedenti) si distribuisce sotto l'ipotesi nulla di omogeneità fra le medie, come una F di Snedecor, in quanto rapporto di due stimatori corretti (e indipendenti) di σ^2.

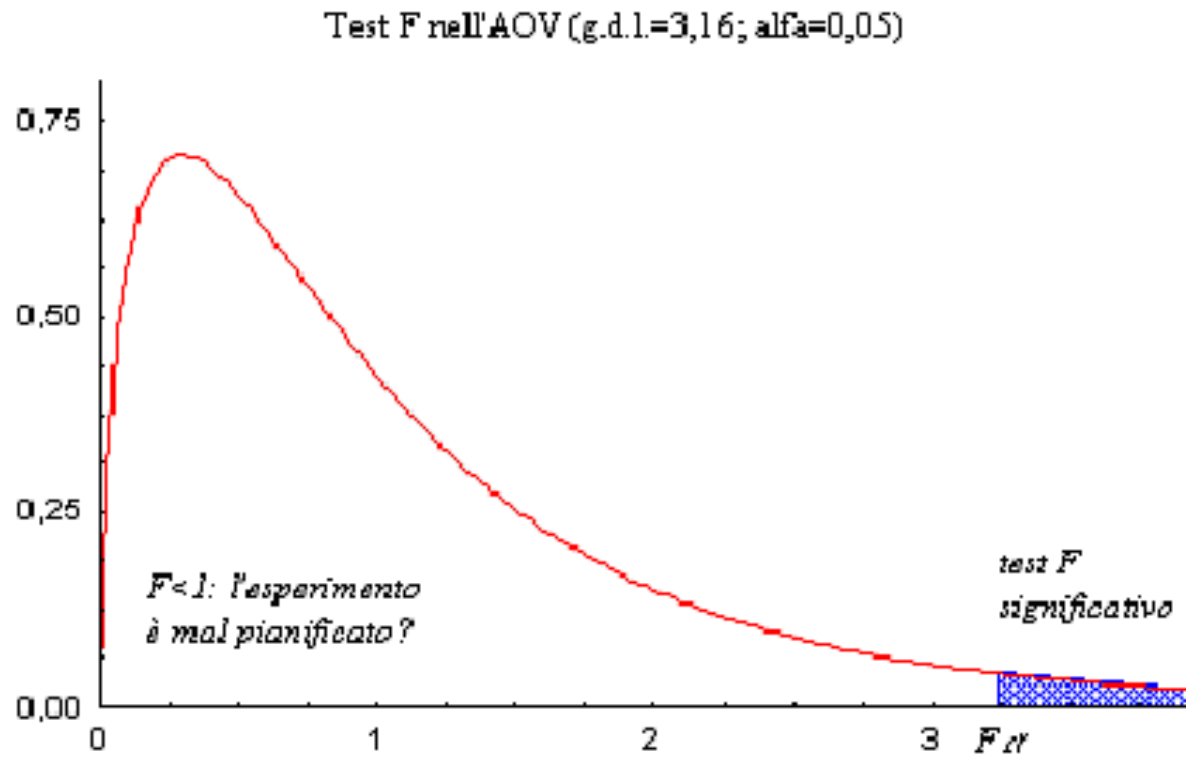

Figure 6: Test dell'analisi della varianza ad una via: valori di F troppo bassi devono inospettirci sull'esperimento

18.2.5 Valore di F inferiore ad uno.

E se il test F risulta inferiore ad 1? E' un caso? Occorre fare qualche considerazione particolare o ci limitiamo a dire che non c'è evidenza contro H_0?

A rigore non dovrebbe importare, perché la zona di rifiuto del test è nella coda destra della distribuzione; tuttavia, per il ragionamento fatto nella sezione precedente, ciò significherebbe che la varianza stimata attraverso le medie è molto più bassa di quella sperimentale, mentre dovrebbero essere sotto H_0 al più dello stesso ordine di grandezza; in effetti nell'AOV la varianza sperimentale rappresenta il metro secondo cui giudichiamo la variabilità fra le medie. In sostanza se F è minore di 1, o addirittura molto più piccolo, ciò vuol dire che la variabilità misurata attraverso le medie degli effetti è inferiore a quella sperimentale. Questo potrebbe essere un indizio di cattiva pianificazione dell'esperimento. Oppure qualche fattore di variazione è stato erroneamente trascurato, e la varianza sperimentale é sovrastimata, e quindi s^2 non é un metro adatto per misurare la variabilità fra le medie degli effetti.

19 [

Analisi della varianza: altri problemi]Analisi della varianza parte II

20 Divergenza dalla linearità per criteri di classificazione quantitativi.

Riprendiamo in esame la tavola di analisi della varianza per un criterio di classificazione qualitativo:

DEVIANZA	TIPO	g.d.l.	Val. atteso
$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-M)^2$	TOTALE	$n-1$	
$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-M_j)^2$	ENTRO I GRUPPI	$n-k$	$(n-k)\sigma^2$
$\sum_{j=1}^{k} n_j(M_j-M)^2$	FRA I GRUPPI	$k-1$	$(k-1)\sigma^2+\sum_{j=1}^{k} n_j\eta_j^2$

Table 2: TAVOLA DI SCOMPOSIZIONE DELLA DEVIANZA EMPIRICA

Esistono casi in cui è possibile scomporre ulteriormente la devianza fra i gruppi?

DEVIANZA	TIPO	g.d.l	
$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-M)^2$	TOTALE	$n-1$	variabilità complessiva osservata della variabile di risposta $\mathbf{y}$, non considerando l'esistenza di fattori di classificazione.
$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-M_j)^2$	ENTRO I GRUPPI	$n-k$	variabilità residua: non riducibile ulteriormente (a meno che non vi siano variabile concomitanti, si veda l'analisi della covarianza)
$\sum_{j=1}^{k} n_j(M_j-M)^2$	FRA I GRUPPI	$k-1$	variabilità spiegabile dalla classificazione in k gruppi; è ancora scomponibile se: il criterio di classificazione è a più entrate (AV a più vie) oppure se il criterio di classificazione è numerico;

Nella figura 7 è rappresentata la distribuzione dei pesi di un camione di 2498 neonati, suddivisi secondo la durata della gestazione, misurata in settimane.

Appare evidente che il peso medio dipende dal numero di settimane della gestazione, ed un test di analisi della varianza ad una via risulta altamente significativo. Tuttavia in questo caso, dato che il criterio di classificazione (ossia il numero di settimane della gestazione) è ovviamente quantitativo, ovviamente possiamo fare di più:

Ci possiamo chiedere se una retta descrive adeguatamente la dipendenza in media dei pesi dal numero di settimane di gestazione, o se invece non vi sia una significa-

tiva **divergenza dalla linearità** . In effetti in questo caso la rappresentazione sembra suggerire che una relazione polinomiale descriva meglio l'andamento delle medie.

Ovviamente possiamo condurre questa analisi in termini di test di significatività perchè siamo in presenza di **osservazioni ripetute** per ogni modalità della variabile esplicativa: Se avessimo osservazioni singole per ciascuna modalità potremmo giudicare della linearità di una relazione con altri strumenti (grafici o mediante analisi dei residui a posteriori o mediante confronto con regressioni non parametriche).

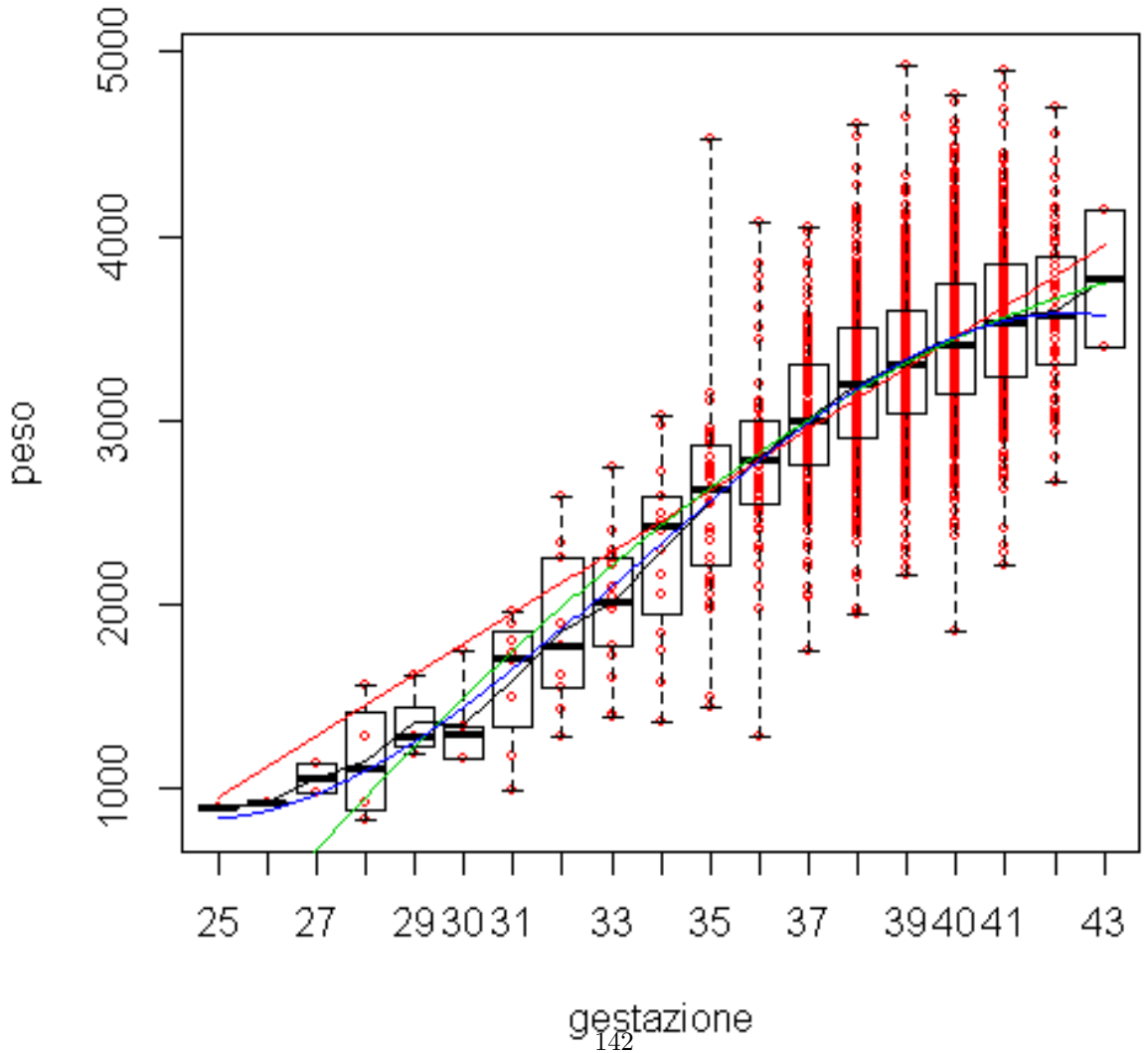

Figure 7: Criterio di classificazione quantitativo: confronto con relazione lineare e con due relazioni polinomiali

20.1 Scomposizione della devianza empirica in 3 componenti

Se il criterio di classificazione è numerico, con k livelli z_j, si può scomporre ulteriormente la devianza fra i gruppi,

$$\sum_{j=1}^{k} n_j (M_j - M)^2,$$

per vedere quanta parte di essa è spiegata dalla regressione lineare delle medie di **y** M_j sui valori z_j.

Se $\hat{M}_j$ è il valore stimato dalla regressione lineare:

$\hat{M}_j = b_{yz} z_j$

si può dimostrare che si può operare algebricamente la scomposizione:

$\Sigma_{j=1}^{k} n_j (M_j - M)^2 =$	$\Sigma_{j=1}^{k} n_j (M_j - \hat{M}_j)^2 +$	$\Sigma_{j=1}^{k} n_j (\hat{M}_j - M)^2$
Fra i gruppi	Divergenza dalla linearita'	Regressione lineare di **Y** su z

la somma dei doppi prodotti è nulla per le equazioni normali.

I due termini si distribuiscono (sotto H_0) secondo due v.c. χ^2 indipendenti, con $k-2$ e 1 grado di libertà [9].

In effetti possiamo considerare lo scarto di ogni osservazione dalla media aritmetica generale $(y_{ij} - M)$; aggiungendo e sottraendo M_j e $\hat{M}_j$ e riordinando i termini otteniamo tre componenti ciascuna con un suo preciso significato:

$$\underbrace{(y_{ij} - M)}_{\text{scarto totale}} = \underbrace{(y_{ij} - M_j)}_{\substack{\text{scarto} \\ \text{dalla} \\ \text{media del} \\ \text{gruppo}}} + \underbrace{(M_j - \hat{M}_j)}_{\substack{\text{Divergenza} \\ \text{dalla} \\ \text{linearità}}} + \underbrace{(\hat{M}_j - M)}_{\substack{\text{Regressione} \\ \text{lineare}}}$$

In definitiva nel caso di un fattore quantitativo Z possiamo scomporre la devianza

[9] Si può applicare il teorema di Cochran (perché la somma dei gradi di libertà coincide col totale)

totale in tre parti (avendo indicato con e_{yz} il rapporto di correlazione che misura la dipendenza in media di y da z):

Tavola di scomposizione della devianza empirica di $\mathbf{y}$ per un criterio di classificazione semplice con k livelli quantitativi z_j

Dev.	TIPO	g.d.l.	Prop. di dev. tot.	Elem.
$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-M)^2$	Totale	$n-1$	1	$\mathbf{y}_{ij}-M$
$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-M_j)^2$	Entro i gruppi	$n-k$	$1-e_{yz}^2$	$\mathbf{y}_{ij}-M_j$
$\sum_{j=1}^{k} n_j(M_j-\hat{M}_j)^2$	Divergenza dalla linearita'	$k-2$	$e_{yz}^2-r_{yz}^2$	$M_j-\hat{M}_j$
$\sum_{j=1}^{k} n_j(\hat{M}_j-M)^2$	Regressione lineare di $\mathbf{y}$ su z	1	r_{yz}^2	$\hat{M}_j-M$

Si può costruire il test per la verifica dell'ipotesi di linearità della regressione:

$$F=\frac{\frac{\sum_{j=1}^{k} n_j(M_j-\hat{M}_j)^2}{k-2}}{\frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij}-M_j)^2}{n-k}}$$

E' facile vedere che questo test si può esprimere anche in questo mdo, evidenziando il ruolo delle frazioni di devianza $e_{yz}^2-r_{yz}^2$

$$F=\frac{\frac{e_{yz}^2-r_{yz}^2}{k-2}}{\frac{1-e_{yz}^2}{n-k}}$$

20.1.1 Differenza fra i test di omogeneità

Il test su r^2_{yz} saggia l'ipotesi che le k medie non varino, contro l'alternativa che varino in modo lineare rispetto a z .

Saggia quindi l'ipotesi che $\beta_{yz} = 0$, contro l'alternativa che $\beta_{yz} \neq 0$, essendo comunque i valori attesi di $\mathbf{y}$ funzioni lineari di Z.

Il test con e^2_{yz} saggia l'ipotesi che le k medie non varino, contro l'alternativa che varino in modo qualsiasi (anche non linearmente rispetto a z).

saggia l'ipotesi che $\mu_j = \mu_r$, contro l'alternativa che per almeno due gruppi si abbia: $\mu_j \neq \mu_r$, essendo i k valori attesi di $\mathbf{y}$ funzioni qualsiasi di Z)

In linea generale questi due test dovrebbero differire per quanto riguarda il potere, dal momento che si riferiscono ad alternative differenti.

Il test su $e^2_{yz} - r^2_{yz}$ saggia l'ipotesi che le k medie varino solo per effetto di una relazione lineare rispetto a z , contro l'alternativa che varino in modo non lineare.

Per esempio, supponendo l'esistenza di una relazione polinomiale di grado $k-1$ dei valori attesi di $\mathbf{y}$ rispetto a Z .

$$\mathrm{E}\left[Y_i\right] = \sum_{j=0}^{k-1} \beta_j z_i^j :$$

saggia l'ipotesi che i $k-2$ coefficienti dei termini di grado 2^o e superiore siano nulli,

$$H_0 : \beta_2 = \beta_3 = \ldots = \beta_{k-1} = 0; \boldsymbol{\beta}_0, \beta_1 \text{qualsiasi}$$

ossia l'ipotesi che la relazione sia lineare, contro l'alternativa che almeno un coefficiente sia diverso da zero, ossia che la relazione sia curvilinea.

Codice utilizzato per l'esempio riportato nella figura 7

21 Analisi della varianza a due vie

Analisi della varianza per due criteri di classificazione qualitativi:

Elementi del modello:

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk};$$

le assunzioni su $\boldsymbol{\varepsilon}$ sono quelle usuali (omoscedasticità, non correlazione entro i gruppi e fra i gruppi, e normalità)

Oltre agli $r \times c$ parametri μ_{ij} possiamo definire tre tipi di medie marginali

Media Riga i (fattore A) $i = 1, \ldots, r$	$\mu_{i.}$	media ponderata delle c medie μ_{ij}
Media Colonna j (fattore B) $j = 1, \ldots, c$	$\mu_{.j}$	media ponderata delle r medie μ_{ij}
Media generale	μ	media ponderata delle rc medie μ_{ij}

...

definizione degli effetti di riga e di colonna:

effetto generale	μ	(un parametro)
effetto Riga (fattore A)	$\alpha_i = \mu_{i.} - \mu$	$i = 1, \ldots, r$
effetto Colonna (fattore B)	$\beta_j = \mu_{.j} - \mu$	$j = 1, \ldots, c$

Modello additivo per μ_{ij}:

$$\mu_{ij} = \mu + \alpha_i + \beta_j;$$

(**separabilità degli effetti**)

e quindi il modello delle osservazioni è:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk};$$

Discussione sull' additività:
esempio con una tavola 2x2

15	20
22	???

Se vale il modello additivo ovviamente il valore nella casella è 27

21.0.1 interazioni.

effetto interazione AxB

Viene definito come **scostamento delle medie dal modello additivo**

$$\gamma_{ij} = \mu_{ij} - \alpha_{i.} - \beta_{.j} - \mu = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu$$
$$i = 1, 2, \ldots, r \quad j = 1, 2, \ldots, c$$

21.0.2 significato delle interazioni

Le interazioni possono essere viste come deviazioni dal modello additivo.

oppure differenze fra gli effetti di riga in corrispondenza dei diversi livelli di colonna

Effetti moltiplicativi per variabili quantitative

esempio con le variabili indicatrici

interpretazione degli effetti dei fattori in presenza di interazione: effetti medi e non parziali

```
additività non è sinonimo qui di linearità: implica che i due fattori ha
effetti che si sommano

 anche le interazioni, se stanno nel modello,
 ci stanno in modo additivo
```

ESEMPI:

	B1	B2	Marg.
A1	12	8	10
A2	6	14	10

Discussione sul numero di parametri indipendenti Impostazione del modello lineare generale

Figura da inserire ALTRO MATERIALE

ESEMPI E GRAFICI (mathematica)

(SU INTERAZIONI COME EFFETTO MOLTIPLICATIVO:

$$\mathrm{E}[\mathbf{y}] = A + bx_1 + cx_2 + dx_1x_2)$$

21.0.3 Influenza della ripartizione delle n osservazioni nelle $r \times c$ celle sull'analisi

Possibili configurazioni delle ampiezze campionarie: (corrispondono a configurazioni differenti della matrice del disegno $\mathbf{X}_{n\times(rc)}$)

n_{ij} proporzionali:

$$n_{ij} = \frac{n_{i.}n_{.j}}{n}$$

Caso bilanciato; in particolare:

n_{ij} uguali:

$$n_{ij} = m$$

Si possono usare pesi uguali per la definizione degli effetti;

- con una sola osservazione per casella ($m = 1$) non sarà possibile stimare nel modo ordinario la varianza σ^2 senza fare opportune ipotesi sulle interazioni $\boldsymbol{\gamma}_{ij}$.
- con n_{ij} qualsiasi (frequenze non proporzionali)
- In questo caso Gli stimatori degli effetti di riga e di colonna risultano non ortogonali.
- problema della scelta dei pesi per la definizione degli effetti riga e di colonna;

- Problemi per la stima degli effetti;
- Problemi per la scomposizione della devianza e per i test.

21.0.4 scomposizione della devianza empirica

Analisi della varianza per due criteri di classificazione qualitativi:

Si scompone facilmente la devianza totale:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{m}(\mathbf{y}_{ijk}-M)^2=$$

Aggiungendo e sottraendo M_{ij} , $M_{i.}$, $M_{.j}$ ed arrangiando opportunamente i termini:

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{m}[(\mathbf{y}_{ijk}-M_{ij})+(M_{ij}-M_{i.}-M_{.j}+M)+(M_{i.}-M)+(M_{.j}-M)]^2=$$

$$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{m}(\mathbf{y}_{ijk}-M_{ij})^2+m\sum_{i=1}^{r}\sum_{j=1}^{c}(M_{ij}-M_{i.}-M_{.j}+M)^2+$$

$$+mc\sum_{i=1}^{r}(M_{i.}-M)^2+mr\sum_{j=1}^{c}(M_{.j}-M)^2$$

I doppi prodotti (nel caso bilanciato) sono tutti nulli.

TAVOLA DI SCOMPOSIZIONE DELLA DEVIANZA EMPIRICA per due criteri di classificazione qualitativi con: $n_{ij} = m; n = rcm$

DEVIANZA	fonte di variab.	g.d.l.	Val. atteso
$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{m}(\mathbf{y}_{ijk} - M)^2$	totale	$rcm-1$	
$\sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{k=1}^{m}(\mathbf{y}_{ijk} - M_{ij})^2$	entro i gruppi	$rc \times (m-1)$	$rc \times (m-1)\sigma^2$
$m\sum_{i=1}^{r}\sum_{j=1}^{c}(M_{ij} - M_{i.} - M_{.j} + M)^2$	interazioni	$(r-1) \times (c-1)$	$(r-1) \times (c-1)\sigma^2 + m\sum_{i=1}^{r}\sum_{j=1}^{c}\boldsymbol{\gamma}_{ij}^2$
$mc\sum_{i=1}^{r}(M_{i.} - M)^2$	righe	$r-1$	$(r-1)\sigma^2 + mc\sum_{i=1}^{r}\alpha_i^2$
$mr\sum_{j=1}^{c}(M_{.j} - M)^2$	colonne	$c-1$	$(c-1)\sigma^2 + mr\sum_{j=1}^{c}\beta_j^2$

Cenno alle formule semplificate per il calcolo delle devianze

La devianza totale è stata scomposta con:

DT=DE+DI+DR+DC

Posto: $s^2 = DE/gdl(DE)$

è immediato ottenere i tre test (nel caso bilanciato, e quindi ortogonale (rivedere, se è il caso)) per la verifica delle ipotesi riguardanti:

Le interazioni:

$$\frac{DI}{gdl(DI)s^2}$$

Gli effetti di riga:

$$\frac{DR}{gdl(DR)s^2}$$

Gli effetti di colonna:

$$\frac{DC}{gdl(DC)s^2}$$

Si distribuiscono sotto la corrispondente ipotesi nulla come delle F con gli opportuni gradi di libertà

Significato dei test se le interazioni sono significativamente diverse da zero

21.0.5 Analisi della varianza a due vie

livello avanzato

costruzione della matrice $\mathbf{X}$ per il caso con frequenze uguali

dimostrazione della ortogonalità

Figura da inserire Tavola della matrice $\mathbf{X}$ NELL' AOV a 2 VIE

Orizzontale

effettuando il prodotto $\mathbf{X}^\mathsf{T}\mathbf{X}$ si ottiene una matrice (simmetrica) in 16 blocchi (4 gruppi di parametri x 4 gruppi di parametri);

riferendoci ancora solo al caso di numerosità uguale m nelle rc caselle: $n = mrc$

	$\boldsymbol{\mu}$	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	$\boldsymbol{\gamma}$		
	(1)	$(r-1)$	$(c-1)$	$(r-1)(c-1)$		
	n	0	0	0	$\boldsymbol{\mu}$	(1)
$\mathbf{X}^\top\mathbf{X} =$	0	$\mathbf{A}$	0	0	$\boldsymbol{\alpha}$	$(r-1)$
	0	0	$\mathbf{B}$	0	$\boldsymbol{\beta}$	$(c-1)$
	0	0	0	$\mathbf{C}$	$\boldsymbol{\gamma}$	$(r-1)(c-1)$

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} \boldsymbol{\mu} & \boldsymbol{\alpha} & \boldsymbol{\beta} & \boldsymbol{\gamma} \\ (1) & (r-1) & (c-1) & (r-1)(c-1) \\ n & 0 & 0 & 0 \\ 0 & \mathbf{A} & 0 & 0 \\ 0 & 0 & \mathbf{B} & 0 \\ 0 & 0 & 0 & \mathbf{C} \end{pmatrix}$$

I quattro gruppi di parametri sono ortogonali
Risultano ortogonali anche se le frequenze sono proporzionali $n_{ij} = n_{i.}n_{.j}/n$
Se le frequenze non sono proporzionali non sarà possibile stimare in modo ortog-

onale gli effetti di riga e di colonna (cambierebbero anche le stime di un gruppo di parametri di effetti in funzione dei valori dell'altro gruppo di parametri)

$$\mathbf{A}_{r-1,r-1} = \begin{pmatrix} 2mc & mc & mc \\ & 2mc & mc \\ & & \dots \\ & & 2mc \end{pmatrix}$$

$$\mathbf{B}_{c-1,c-1} = \begin{pmatrix} 2mr & mr & mr \\ & 2mr & mr \\ & & \dots \\ & & 2mr \end{pmatrix}$$

$$\mathbf{C} = (\text{Omettere});$$

Per stimare i parametri conviene partire dalle definizioni dei parametri e dei vincoli. Infatti è ovvio che, nel modello completo con $r \times c$ parametri, si ha:

$$\hat{\mu}_{ij} = M_{ij}$$

Per cui è immediato vedere che (nel caso di frequenze uguali $n_{ij} = m$) si hanno i seguenti stimatori:

effetto generale	$\hat{\mu} = M$	
effetto Riga (fattore A)	$\hat{\alpha}_i = M_{i.} - M$	$i = 1, \ldots, r$
effetto Colonna (fattore B)	$\hat{\beta}_j = M_{.j} - M$	$j = 1, \ldots, c$
effetto interazione A $\times$ B	$\hat{\gamma}_{ij} = M_{ij} - M_{i.} - M_{.j} + M$	$i = 1, \ldots, r; j = 1, \ldots, c$

21.1 Analisi della varianza a due vie: altre problematiche

21.1.1 Disegni non bilanciati

caso in cui n_{ij} non sono proporzionali:

- la situazione è tipica di studi osservazionali, o indagini esplorative;
- qualche n_{ij} può anche essere nullo;
- per esempio anche in situazioni sperimentali se l'esperimento associato a determinate combinazioni di fattori non può essere condotto a termine.

- Le stime degli effetti interazioni e del residuo sono ortogonali e sono anche ortogonali rispetto alle stime degli altri effetti (scelto un opportuno sistema di pesi).
- Le stime degli effetti di riga non sono comunque ortogonali rispetto alle stime degli effetti di colonna;
- (la matrice $\mathbf{X}^\top\mathbf{X}$che si ottiene impostando l'opportuna matrice $\mathbf{X}(n \times rc)$ di variabili indicatrici, non è diagonale a blocchi.

Non è possibile scomporre la devianza nel modo visto per i piani bilanciati Per stimare i parametri e per effettuare test si può ricorrere alla teoria generale sui modelli lineari:

Si stimano i parametri e la devianza residua nel modo ordinario dal modello con tutti i parametri;

Si stimano le porzioni di devianza attribuibili ai vari gruppi di parametri (ed eventualmente si verificano delle ipotesi nidificate), in sequenza: basta calcolare le stime dei parametri del modello imponendo gli opportuni vincoli, partendo dal gruppo delle interazioni, e dopo gli effetti riga o colonna (procedura stepwise)

L'ordine di esecuzione della procedura per gli effetti riga e colonna determina due tavole di scomposizione della devianza differenti e stime differenti dei parametri (data la non ortogonalità)

Figura da inserire

ESEMPIO

21.1.2 Disegni bilanciati: una sola osservazione per casella ($m = 1$)

Se $m = 1$ non sarà possibile stimare nel modo ordinario la varianza σ^2 senza fare opportune ipotesi sulle interazioni $\boldsymbol{\gamma}_{ij}$.

Infatti non vi sono gradi di libertà per la stima della varianza σ^2 della componente accidentale:

$$\text{devianza entro i gruppi} = \sum_{i=1}^{r} \sum_{j=1}^{c} \sum_{k=1}^{m} (\mathbf{y}_{ijk} - M_{ij})^2 = 0$$

$$(\text{perché} \mathbf{y}_{ijk} = M_{ij}) \qquad \text{gradi di libertà} = rc \times (m-1) = 0$$

Vi sono essenzialmente due possibilità (valide anche per modelli a più vie):

1. Si assume $\boldsymbol{\gamma}_{ij} = 0$, ossia che il modello sia additivo, e si stima la varianza attraverso:

$$s_I^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (M_{ij} - M_{i.} - M_{.j} + M)^2/((r-1)(c-1));$$

$$E(s_I^2) = \sigma^2 + \sum_{i=1}^{r} \sum_{j=1}^{c} \boldsymbol{\gamma}_{ij}^2/((r-1) \times (c-1)) = \sigma^2 \text{se si assume} : \boldsymbol{\gamma}_{ij} = 0$$

(si riveda la scomposizione della devianza per il caso bilanciato)

2. Oppure si possono fare delle assunzioni sulle interazioni, in modo che le interazioni non abbiano (r-1)(c-1) gradi di libertà, ma dipendano da un numero inferiore di parametri:

$$\boldsymbol{\gamma}_{ij} = g_{ij}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}),$$

essendo θ un vettore di $s < (r-1)(c-1)$.

Il modello in generale sarà non lineare nei parametri

Modelli moltiplicativi per le interazioni In effetti è opportuno che s sia piccolo in modo da lasciare un numero di gradi di libertà sufficiente per il residuo

In particolare (Tukey):

$$\boldsymbol{\gamma}_{ij} = \theta\alpha_i\beta_j,$$

L'ipotesi moltiplicativa è la più semplice, e si dimostra che, sebbene il modello sia non lineare, si giunge ad una scomposizione additiva conveniente che attribuisce a θ un grado di libertà e $(r-1)(c-1)-1$ gradi di libertà per la stima della varianza.

La validità delle ipotesi restrittive sulle interazioni può essere giudicata anche attraverso un'opportuna analisi dei residui.

21.2 Analisi della varianza a più vie

Analisi della varianza a più vie

interazioni di ordine superiore al primo

Esempio con l'analisi della varianza per una classificazione completa a 3 vie (fattori A,B e C):

$$\mathbf{y}_{ijhk} = \mu + \alpha_i + \beta_j + \chi_h + \boldsymbol{\gamma}_{ij}^{AB} + \boldsymbol{\gamma}_{ih}^{AC} + \boldsymbol{\gamma}_{jh}^{BC} + \boldsymbol{\gamma}_{ijh}^{ABC} + \varepsilon_{ijhk};$$

$$\mathbf{y}_{ijhk} =$$

$\mu+$	effetto generale
$+\alpha_i + \beta_j + \chi_h+$	effetti dei diversi livelli dei tre fattori
$+\boldsymbol{\gamma}_{ij}^{AB} + \boldsymbol{\gamma}_{ih}^{AC} + \boldsymbol{\gamma}_{jh}^{BC}+$	interazioni fra le coppie di fattori (del primo ordine)
$+\boldsymbol{\gamma}_{ijh}^{ABC}+$	interazioni del secondo ordine
$+\varepsilon_{ijhk}$	componente accidentale

Nel caso bilanciato l'analisi è simile a quella a due vie.

21.2.1 Piani 2^k: Piani fattoriali completi e incompleti

si riveda l'introduzione ai modelli lineari, con gli esempi relativi)

21.3 Blocchi randomizzati; Quadrati latini

argomenti non trattati in questi appunti

22 Analisi della varianza con variabili concomitanti: L'analisi della covarianza

Situazioni con variabili quantitative e qualitative (rivedere introduzione ai mod.lineari)

22.0.1 variabili concomitanti

Riduzione della devianza residua.

22.0.2 confronto fra k relazioni di regressione (lineare)

Introduzione del problema in generale

Interessi particolari:

Verificare la significatività degli effetti di uno o più fattori, eliminando l'influenza di variabili concomitanti (con riduzione della devianza residua)

Verificare se l'influenza delle variabili quantitative è la stessa in tutti i gruppi: confronto fra coeffiecienti di regressione lineare.

22.0.3 Assunzioni per l'analisi della covarianza semplificata

l'ipotesi di parallelismo

l'ipotesi di linearità della relazione fra le medie delle variabili.

Matrice del disegno sperimentale per l'analisi della covarianza con:
un fattore qualitativo a k livelli;
n_j osservazioni per ogni trattamento o gruppo
una variabile concomitante quantitativa Z,
misurata come scarto dalla media del gruppo j:

$$\sum_{i=1}^{n_j} z_{ij} = 0; j = 1, 2, \ldots, k$$

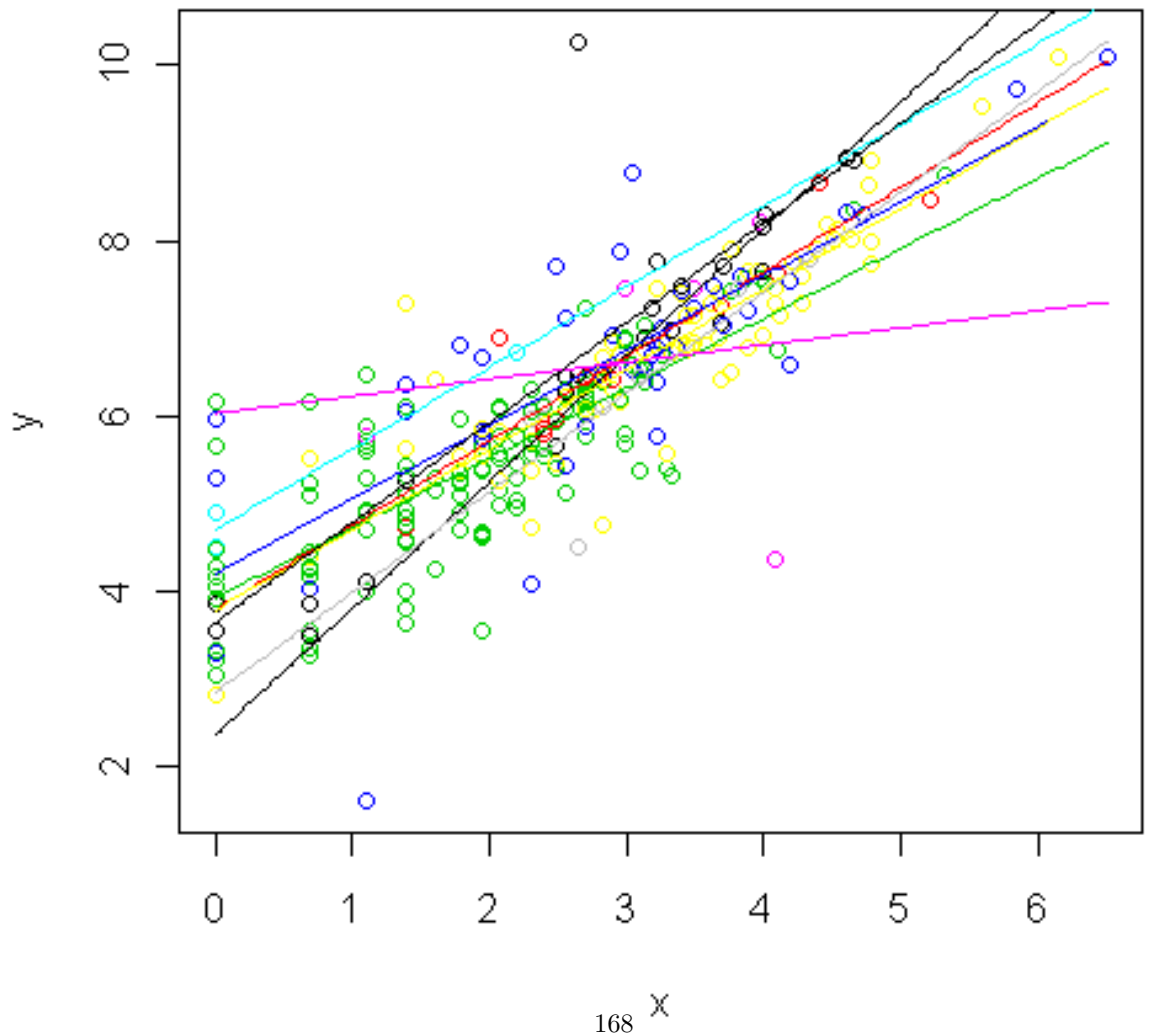

Figure 8: Modello di analisi della covarianza con interazioni: 9 rette distinte

22.0.5 caso semplificato (rette parallele in tutti i gruppi)

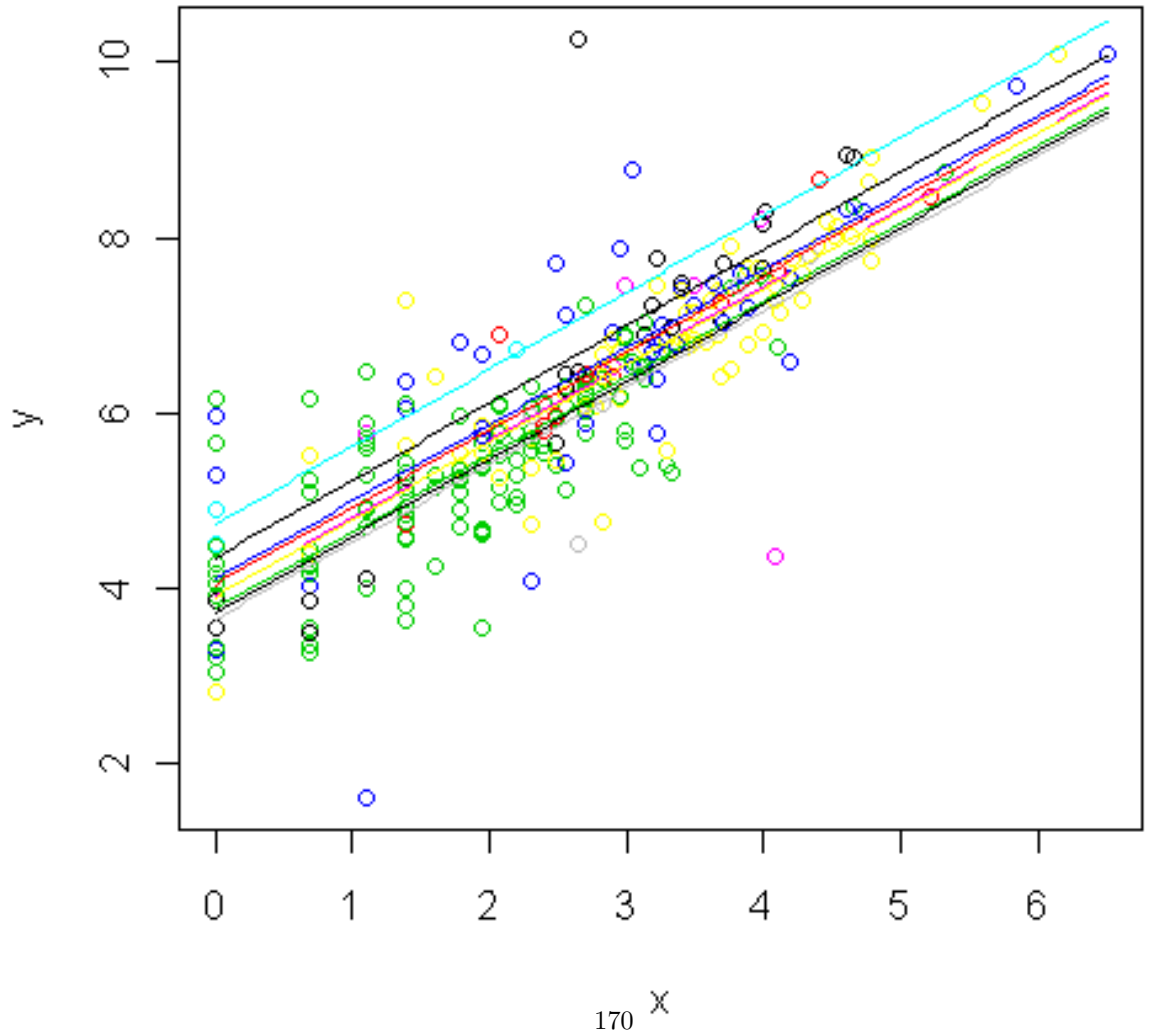

Figure 9: Modello di analisi della covarianza senza interazione: 9 rette parallele

in pratica si ipotizza assenza di interazione fra i due fattori:

parametri da 1 a k	α_j	termine costante della regressione nel gruppo j; $j = 1, 2, \ldots k$.
parametro $(k+1)^o$	$\boldsymbol{\beta}$	pendenza comune alle k rette

Il vettore dei parametri è dato da:

$$\boldsymbol{\theta}^{\mathsf{T}} = \{\alpha_1, \alpha_2, \ldots, \alpha_k, \boldsymbol{\beta}\}$$

La matrice $\mathbf{X}$ (nell'impostazione a rango pieno) ha n righe e $k+1$ colonne:

i	$Rispostaosservata\mathbf{Y}$		$\alpha_1 cost.gruppo1$	$\alpha_2 cost.gruppo2$	$\ldots$	$\alpha_k cost.g$
1	$\mathbf{y}_{1,1}$		1	0		0
$\ldots$	$\ldots$		$\ldots$	$\ldots$	$\ldots$	.
n_1	$\mathbf{y}_{n1,1}$		1	0	$\ldots$	.
	$\mathbf{y}_{1,2}$		0	1	$\ldots$	.
	$\ldots$		0	$\ldots$	$\ldots$	.
n_1+n_2	$\mathbf{y}_{n2,2}$		0	1	$\ldots$	.
	$\ldots$	$\mathbf{X}=$	$\ldots$	$\ldots$	$\ldots$	.
	$\mathbf{y}_{i,j}$		0	0	$\ldots$	1
$n_1+\ldots+n_j$	$\ldots$		$\ldots$	$\ldots$	$\ldots$	.
	$\mathbf{y}_{1,k}$		0	0	$\ldots$	0
	$\ldots$		0	0	$\ldots$	0
$n_1+\ldots+n_k$	$\mathbf{y}_{nk,k}$		0	0	$\ldots$	0

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & & 0 & & 0 & z_{1,1} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & \dots & \dots & \dots & 0 & z_{n1,1} \\ 0 & 1 & \dots & \dots & \dots & 0 & z_{1,2} \\ 0 & \dots & \dots & \dots & \dots & 0 & \dots \\ 0 & 1 & \dots & \dots & \dots & 0 & z_{n2,2} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & & & \\ 0 & 0 & \dots & 1 & \dots & 0 & z_{i,j} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 & z_{1,k} \\ 0 & 0 & \dots & 0 & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 & z_{nk,k} \end{pmatrix}$$

Risposta osservata

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_{1,1} \\ \dots \\ \mathbf{y}_{n1,1} \\ \mathbf{y}_{1,2} \\ \dots \\ \mathbf{y}_{n2,2} \\ \dots \\ \mathbf{y}_{i,j} \\ \dots \\ \mathbf{y}_{1,k} \\ \dots \\ \mathbf{y}_{nk,k} \end{pmatrix}$$

Si vede che:

$$\mathbf{X}^\top\mathbf{X} = \begin{pmatrix} n_1 & 0 & & 0 \\ & \dots & & 0 \\ & & n_k & 0 \\ 0 & 0 & 0 & \Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j} z_{ij}^2 \end{pmatrix}$$

$$\mathbf{X}^{\mathsf{T}}\mathbf{y} = \begin{pmatrix} \Sigma_{i=1}^{n_1} \mathbf{y}_{i1} & 1 \\ \dots & \dots \\ \Sigma_{i=1}^{n_k} \mathbf{y}_{ik} & k \\ \Sigma_{j=1}^{k} \Sigma_{i=1}^{n_j} \mathbf{y}_{ij} z_{ij} & k+1 \end{pmatrix}$$

Per cui le stime di massima verosimiglianza sono:

$$\hat{\alpha}_j = M_j, j = 1, 2, \dots, k;$$

$$\hat{\beta} = \mathbf{b} = \frac{\Sigma_{j=1}^{k} \Sigma_{i=1}^{n_j} \mathbf{y}_{ij} z_{ij}}{\Sigma_{j=1}^{k} \Sigma_{i=1}^{n_j} z_{ij}^2}$$

stima della pendenza comune:
$\hat{\beta} = \mathbf{b}$ è una media ponderata dei $\hat{\beta}_j$ dei singoli gruppi.
La devianza residua è:

$$R(\hat{\theta}) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - M_j - b^* z_{ij})^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - M_j)^2 - b^2 \sum_{j=1}^{k} \sum_{i=1}^{n_j} z_{ij}^2$$

$$= \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - M_j)^2 - [\sum_{j=1}^{k} \sum_{i=1}^{n_j} (\mathbf{y}_{ij} - M_j) z_{ij}]^2 / \sum_{j=1}^{k} \sum_{i=1}^{n_j} z_{ij}^2 =$$

$= DevInt(\mathbf{y}) - [Codev.Int(ZY)]^2/DevInt(Z)$

(Riduzione di devianza residua, rispetto alla AOV, dovuta al fattore concomitante)

Se l'ipotesi nulla impone k-1 vincoli:

$$H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_k (= \alpha)$$

Il modo più semplice di procedere è quello generale di stimare i parametri sotto H_0, e quindi sottrarre dalla devianza residua sotto H_0 quella residua non vincolata:

$$R(\hat{\theta}_0) = DevTot(\mathbf{y}) - [Codev.Tot(ZY)]^2/DevTot(Z)$$

$R(\hat{\theta}_0)$ è ricavata semplicemente dalla regressione di $\mathbf{y}$ su Z considerando i k gruppi come un unico campione.

Si tenga presente che in questo caso vanno considerati gli scarti di Z dalla propria media generale, e non dalle medie dei singoli gruppi.

Per saggiare l'ipotesi si applica la tecnica generale dei modelli lineari:

$$F = \frac{\frac{R(\hat{\theta}_0) - R(\hat{\theta})}{k-1}}{\frac{R(\hat{\theta})}{n-k-1}}$$

Spiegare l'ipotesi in termini di differenza fra medie corrette

Tavola di scomposizione della di $\mathbf{y}$ nel caso di un fattore concomitante Z con pendenza uguale nei k gruppi:

DEVIANZA DI $\mathbf{Y}$	SPIEGATA DALLA REGRESSIONE SU $\mathbf{Z}$	tipo	g.d.l.
DevTot($\mathbf{y}$)	$\frac{Codev.Tot(ZY)]^2}{DevTot(Z)}$	totale	$n-2$
DevInt($\mathbf{y}$)	$\frac{Codev.Int(ZY)]^2}{DevInt(Z)}$	entro i gruppi	$n-k-1$
DevFraGruppi($\mathbf{y}$)	(2)-(4)	fra i gruppi	$k-1$

La tecnica, in particolare nel caso di regressioni con pendenze uguali entro le caselle, è generalizzabile al caso di un modello:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

ove le colonne di Z sono costituite da s variabili concomitanti e la matrice $\mathbf{X}$ (anche di rango non pieno) è la matrice di indicatori associata ad una classificazione anche a più vie.

Caso generale: k coefficienti di regressione distinti La matrice $\mathbf{X}$ sarà composta da $2k$ colonne, di cui le prime k sono come prima costituite dagli indicatori di appartenenza ai gruppi:

$$\mathbf{X}_1 = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \dots & 0 & \dots & \dots & \dots & 0 \\ 1 & 0 & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 \end{pmatrix}$$

Occorre ora affiancare altre k colonne:

$$\mathbf{X}_2 = \begin{pmatrix} z_{1,1} & \dots & 0 \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ z_{n1,1} & \dots & 0 \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & z_{ij} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \dots & z_{n1,k} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \dots & z_{nk,k} \end{pmatrix}$$

dato che ora il vettore di $2k$ parametri è:

$$\boldsymbol{\theta}^{\mathsf{T}} = \{\alpha_1, \dots, \alpha_j, \dots, \alpha_{k,}\beta_1, \dots, \beta_j, \dots, \beta_k\}$$

la matrice $\mathbf{X}$ è costituita dalle colonne di $\mathbf{X}_1$ e $\mathbf{X}_2$ affiancate
Ipotesi di interesse:

$$H_0 : \beta_1 = \dots = \beta_j = \dots = \beta_k;$$

rette di regressione parallele nei k gruppi.

$$H_0 : \beta_1 = \ldots = \beta_j = \ldots = \beta_k; \alpha_1 = \ldots = \alpha = \ldots = \alpha_k$$

rette di regressione uguali nei k gruppi.

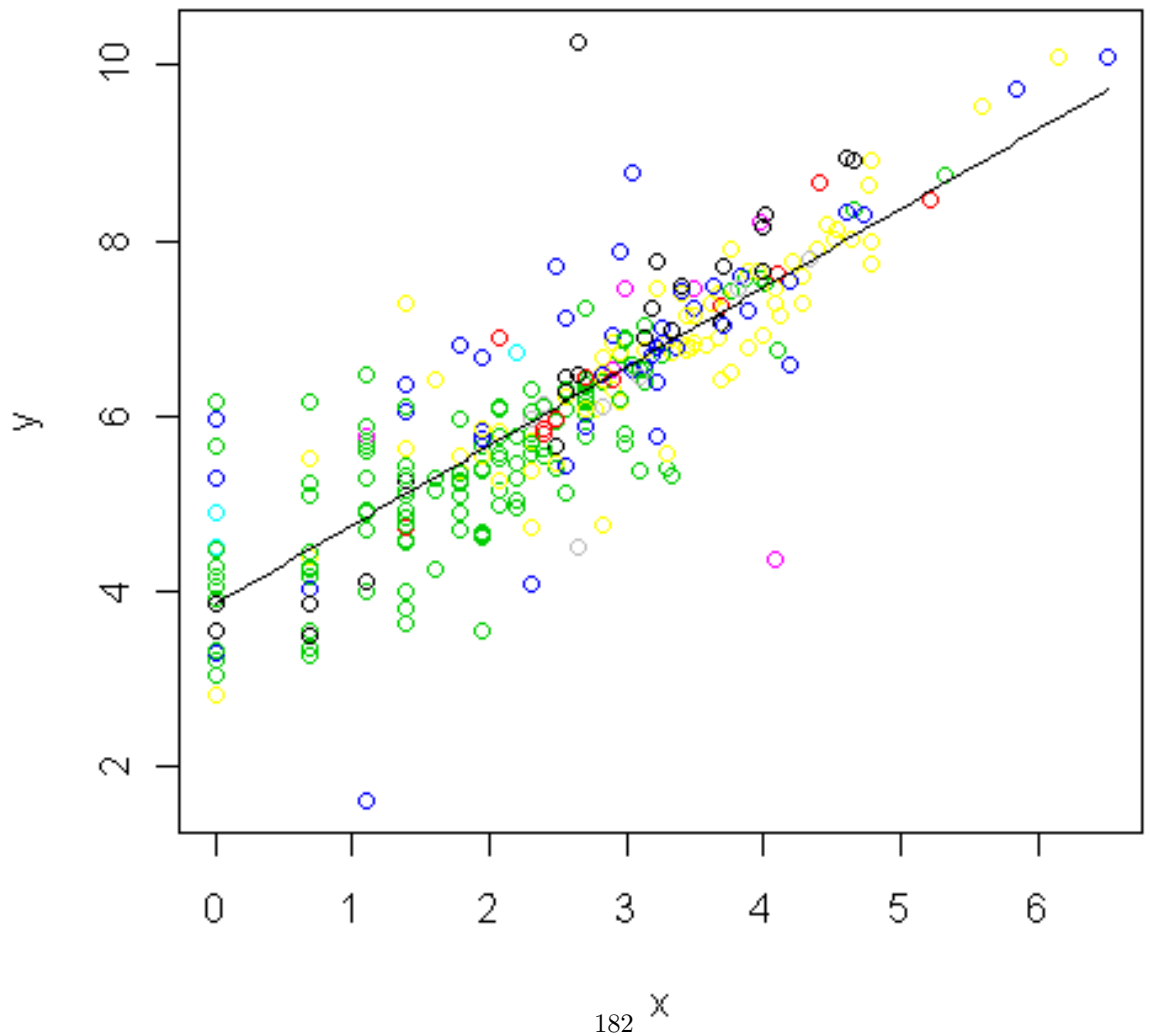

Figure 10: Modello di analisi della covarianza in assenza di effetto di gruppo: una retta unica per tutti i gruppi

E’ possibile (ed è più semplice) procedere per passaggi successivi attraverso modelli nidificati (l’analisi non è ortogonale):

analisidellacovarianza			
fonte	$Fonte$		$g.d.l.$
totale	$DevTot(\mathbf{y})$		$n-1$
residuo delle k regressioni entro i gruppi (pendenze diverse)	$\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}(\mathbf{y}_{ij}-a_j-b_jz_{ij})^2$	$Res1$	$n-2k$
residuo analisi semplificata (pendenze uguali)	$\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}(\mathbf{y}_{ij}-a_j^*-b^*z_{ij})^2$	$Res2$	$n-k-1$
Differenze fra le k pendenze	$Res2-Res1$		$k-1$
residuo regressione unica (trascurando il fattore qualitativo)	$\Sigma_{j=1}^{k}\Sigma_{i=1}^{n_j}(\mathbf{y}_{ij}-a^*-b^{**}z_{ij})^2$	$Res3$	$n-2$
Fra i k gruppi (k medie aggiust.)	$Res3-Res2$		$k-1$

Codice utilizzato per l'esempio riportato nelle figure 8, 9, 10.