Contents

1 Introduzione ai Modelli Lineari: Parte I **8**

2 Il modello lineare di dipendenza per variabili normali. **9**

3 I modelli statistici. **14**

4 Il modello lineare generale. **14**

4.1 Impieghi del modello lineare 21

5 Problemi di inferenza **24**

5.1 Componente sistematica e componente casuale. 26

5.2 Caratteristiche essenziali degli elementi del modello lineare 31

5.3 Caratteristiche più dettagliate degli elementi del modello: 32

5.4 Ipotesi sulle ε 39

6 Introduzione ai Modelli Lineari: Parte II **41**

DRAFT

7 La matrice delle X **41**
7.1 Osservazioni ripetute. 42
7.2 Disegni fattoriali 43
7.2.1 Disegni 2^k 50

8 Regressione multipla. **53**
8.1 Relazione di regressione in termini di scarti 55
8.2 Regressione polinomiale: 59
8.2.1 Polinomi in più variabili e superfici di risposta 60

9 Fattori qualitativi e analisi della varianza **61**
9.1 Regressori del tipo 0/1 (dummy variables) 61
9.2 Analisi della varianza ad effetti fissi ed un criterio di classificazione 65
9.3 Analisi della varianza ad effetti fissi con due criteri di classificazione 72
9.4 Analisi della covarianza 74
9.5 Rette o piani di regressione con pendenze diverse: termini polinomiali moltiplicativi 78

DRAFT

9.6 Modelli autoregressivi 82

10 Generalizzazioni **85**

10.0.1 Regressione logistica 87
10.0.2 Regressione piecewise 88
10.0.3 Approssimazione di modelli non lineari 88
10.0.4 Modelli additivi generalizzati 88

11 [**90**

12 Assunzioni di base nel modello lineare **94**

13 verosimiglianza **99**

13.1 La stima di massima verosimiglianza di σ^2 104
13.2 Verosimiglianza profilo rispetto a β. 108
13.3 Verosimiglianza profilo e test basati sul rapporto delle verosimiglianze 112
13.4 Costruzione del test del rapporto delle verosimiglianze 119

DRAFT

13.5 Costruzione del test LR per confronto di ipotesi generiche nel modello lineare 122

14 Intervalli e regioni di confidenza **126**

15 Minimi quadrati ordinari: stima dei β_j. **127**

15.1 Soluzione mediante derivate 128

15.2 Minimizzazione di $R(\beta)$ senza uso di derivate (modelli a rango non pieno) 131

15.3 Teorema di Gauss-Markov 136

15.4 Regressori a media nulla 139

15.5 Distribuzione di $\mathbf{b}$ 141

16 Test di significatività nei modelli lineari **148**

17 Distribuzione della devianza residua nei modelli lineari **148**

17.1 Devianza residua in funzione dei valori osservati 149

17.2 Devianza residua in funzione della componente accidentale ε: 153

DRAFT

17.2.1 intervalli di confidenza per la varianza 158

18 Scomposizione della devianza **159**

18.0.1 La scomposizione della somma dei quadrati $\mathbf{y}^\top\mathbf{y}$ 159

18.1 Scomposizione di $R(\beta)$ 162

18.2 Test F per la verifica di ipotesi nel modello lineare: distribuzione nulla 166

18.2.1 Statistiche sufficienti nel modello lineare. 171

18.2.2 Matrice di informazione 172

18.3 Distribuzioni sotto H_0 e sotto H_1. 173

18.4 Scomposizione della devianza e test nel caso di gruppi di regressori ortogonali 179

19 Configurazioni della matrice X e di $\mathbf{X}^\top\mathbf{X}$ **186**

20 Modello lineare: Verifica di ipotesi generali **189**

20.0.1 Esempio: Analisi della varianza ad una via. 191

20.0.2 Esempio sulla scelta delle variabili. 194

DRAFT

21 La stima dei parametri del modello lineare con vincoli lineari sui parametri **195**
21.0.1 Minimi quadrati vincolati 196
21.1 Modello lineare: Scomposizione della devianza per il problema soggetto a vincoli: 201
21.2 Prove di ipotesi particolari nel modello lineare 208

22 Test e regioni di confidenza nei modelli lineari **211**
22.0.1 Regioni di confidenza simultanee per i parametri 211
22.1 regioni di confidenza per funzioni lineari dei parametri 212
22.1.1 regioni di confidenza relative a sottoinsiemi di parametri . . 213
22.1.2 Intervalli di confidenza per $E(\mathbf{y}_i)$ 214
22.1.3 errori di previsione 215

DRAFT

List of Figures

1 campione da una normale bivariata 10
2 distribuzioni condizionate normali in corrispondenza di valori fissati 17

3 box-plot con retta di regressione e spezzata di regressione 43
4 Piano fattoriale di 3 regressori $2 \times 4 \times 3$ 50
5 Esempio di disegno fattoriale completo e incompleto 50
6 Una relazione con un termine moltiplicativo $x_1 \times x_2$ ha relazioni parziali lineari 79
7 Pendenza diversa come effetto interazione fra due fattori x_1, x_2 . . 79
8 verosimiglianza rispetto a μ e σ^2 di un campione proveniente da una normale e verosimiglianza profilo rispetto a μ (proiettata sul piano verticale in fondo) 116

DRAFT

1 Introduzione ai Modelli Lineari: Parte I

L'approccio tecnico scelto in questo corso ci consentirà di affrontare in modo simile gli aspetti inferenziali relativi alla regressione multipla, all'analisi della varianza e della covarianza;

inoltre costituirà una buona base per alcuni tipi di GLM (Generalized linear models) sia per l'interpretazione dei parametri che per l'inferenza.

DRAFT

2 Il modello lineare di dipendenza per variabili normali.

Per quanto visto nelle lezioni sulla normale multivariata, la distribuzione di una componente Y_1 (e in generale di più componenti), condizionatamente a valori qualsiasi $\mathbf{Y}_2$ di altre componenti, è normale, con valore atteso che è funzione lineare di $\mathbf{Y}_2$, e matrice di varianze e covarianze indipendente dai particolari valori condizionanti; quindi le regressioni sono tutte lineari e omoscedastiche.

Pertanto se si ha a disposizione un campione casuale semplice da una normale multivariata, non esiste alcun problema di identificazione del modello di regressione, né di scelta della funzione, perché tutte le distribuzioni condizionate sono note.

DRAFT

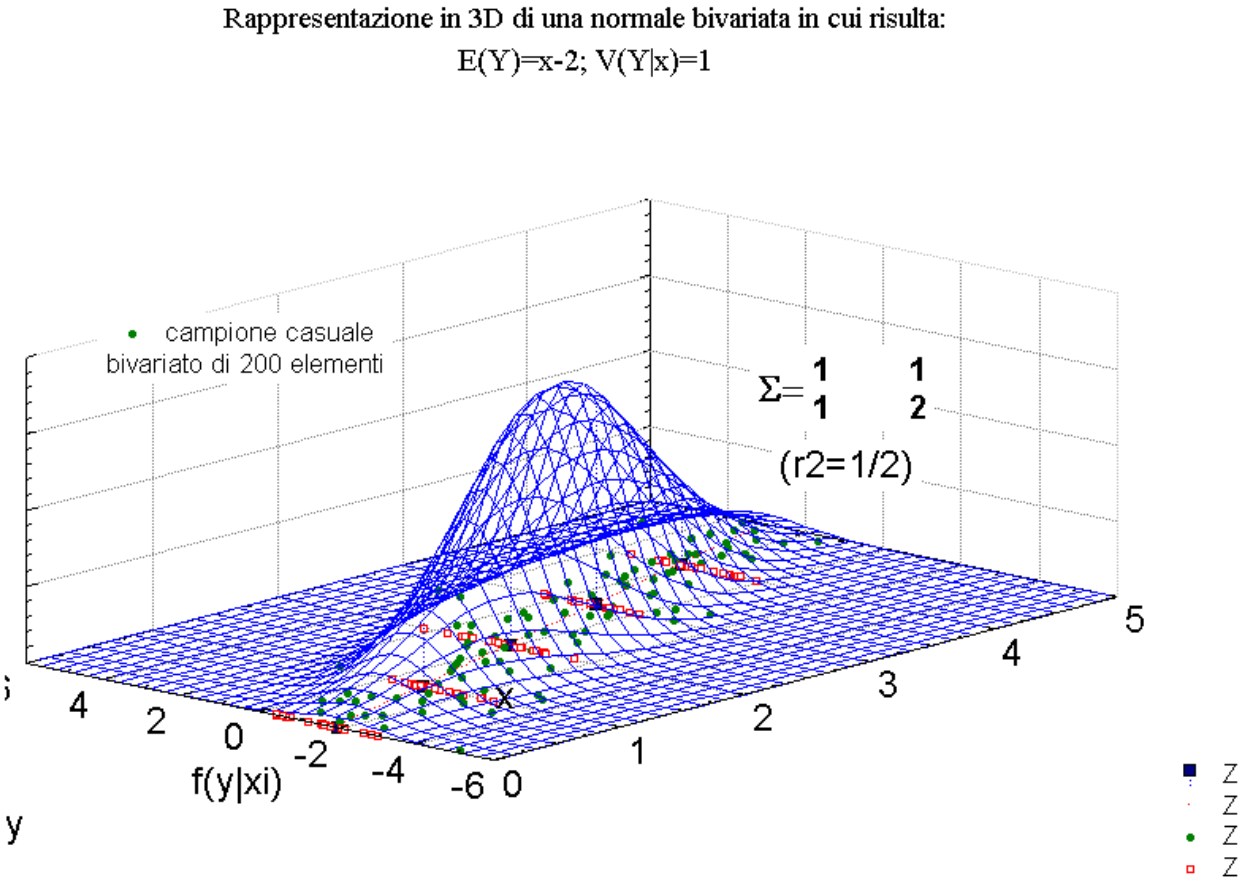

Figure 1: campione da una normale bivariata

Tuttavia sono rari i casi in cui nello studio della dipendenza di uno o più fenomeni, si può ragionevolmente ipotizzare di avere un campione casuale semplice da una distribuzione normale multipla, perché spesso ci si trova in altre situazioni, fra cui essenzialmente si hanno le seguenti:

DRAFT

- I dati costituiscono un campione casuale semplice proveniente da una distribuzione multivariata *non normale*.
- I dati non costituiscono un campione casuale semplice ma, per esempio, i valori delle variabili esplicative sono stati opportunamente selezionati o predisposti
- oppure si ha un campione non probabilistico o comunque un archivio di dati che non costituisce un campione.
- Il modello da cui si possono selezionare i dati è effettivamente una distribuzione normale multivariata (almeno approssimativamente normale), e si può estrarre un campione casuale semplice, tuttavia l'interesse dell'analisi è limitato allo studio della distribuzione di una delle componenti $\mathbf{y}$ condizionatamente a valori particolari o estremi delle altre componenti $\mathbf{X}$: è noto anche nell'analisi della regressione semplice, che l'inferenza è migliore (ossia le bande di confidenza della relazione di regressione sono più strette) se si selezionano unità con valori delle x più vicine a quelle di interesse.

DRAFT

Ovviamente verranno trattati nei capitoli successivi i problemi relativi alla stima dei parametri sulla base di un campione di osservazioni p-variate, che verranno affrontati estendendo opportunamente le tecniche impiegate quando si studia la dipendenza di una variabile $\mathbf{y}$ da una variabile esplicativa $\mathbf{x}$.

In effetti anche nel caso di campioni casuali semplici da distribuzioni non normali multivariate, si possono cercare le migliori (nel senso dei minimi quadrati) relazioni lineari fra le speranze matematiche di $\mathbf{y}$ e particolari valori di $\mathbf{X}$.

In ogni caso, come si apprestiamo a discutere diffusamente, i valori delle x possono anche non essere delle determinazioni di variabili casuali, ma valori anche scelti in modo non casuale.

DRAFT

Nei paragrafi che seguono verranno affrontati diversi aspetti relativi alla *versatilità del modello lineare* ed alle diverse possibilità interpretative del modello e dei suoi parametri: alcuni dei concetti fondamentali relativi a particolari modelli lineari vengono introdotti fra breve, prima che vengano affrontati gli aspetti inferenziali.

versatilità del modello lineare

3 I modelli statistici.

Prima di iniziare lo studio del modello lineare, che ci accompagnerà per tutto (o quasi) il corso, vale la pena di fare una citazione:

Utilità dei modelli statistici

All models are wrong, but some are useful

(G.E.P. Box)

(Tutti i modelli (statistici) sono sbagliati, ma alcuni sono utili)

DRAFT

4 Il modello lineare generale.

Per modello lineare in generale si intende un modello nel quale una variabile di risposta osservabile $\mathbf{Y}$ è spiegata da una combinazione lineare di k variabili esplicative $\mathbf{X}_j$, secondo dei parametri incogniti β_j , più una componente accidentale

ε (non osservabile) , secondo la generica relazione lineare:

Equazione del modello lineare

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \ldots + \mathbf{X}_k\beta_k + \boldsymbol{\varepsilon}$$

In particolare comunque ci occuperemo di modelli lineari di dipendenza nei quali le $\mathbf{X}_j$ non sono variabili casuali, ma costanti note, che assumono n valori in $\Re_k$ (tutti distinti oppure con ripetizioni, questo si vedrà meglio in seguito).

Non ci stiamo occupando della distribuzione simultanea di $k+1$ variabili aleatorie , perché le $\mathbf{X}_j$ sono variabili i cui valori possono addirittura essere prefissati ed assegnati.

La generica osservazione i -esima è quindi caratterizzata da un particolare vettore

DRAFT

di valori delle k variabili $\mathbf{X}_j$, indicato con:

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \dots \\ x_{ij} \\ \dots \\ x_{ik} \end{pmatrix}$$

Eventualmente le x potranno essere dei valori particolari **(fissati!)** di variabili casuali, nel caso in cui studiamo le distribuzioni condizionate della variabile aleatoria $\mathbf{Y}$, condizionatamente agli n valori di k variabili aleatorie $\mathbf{X}_j$, e ipotizzeremmo in quel caso l'esistenza di $k+1$ variabili aleatorie *osservabili*. Anche in questa situazione però non ci occuperemmo della distribuzione congiunta delle $\mathbf{X}_j$, ma solo di $f(\mathbf{Y}|\mathbf{X}_{nxk})$, ossia la distribuzione di $\mathbf{Y}$ condizionatamente a particolari valori delle x.

DRAFT

Rappresentazione in 3D di una regressione lineare normale omoscedastica;

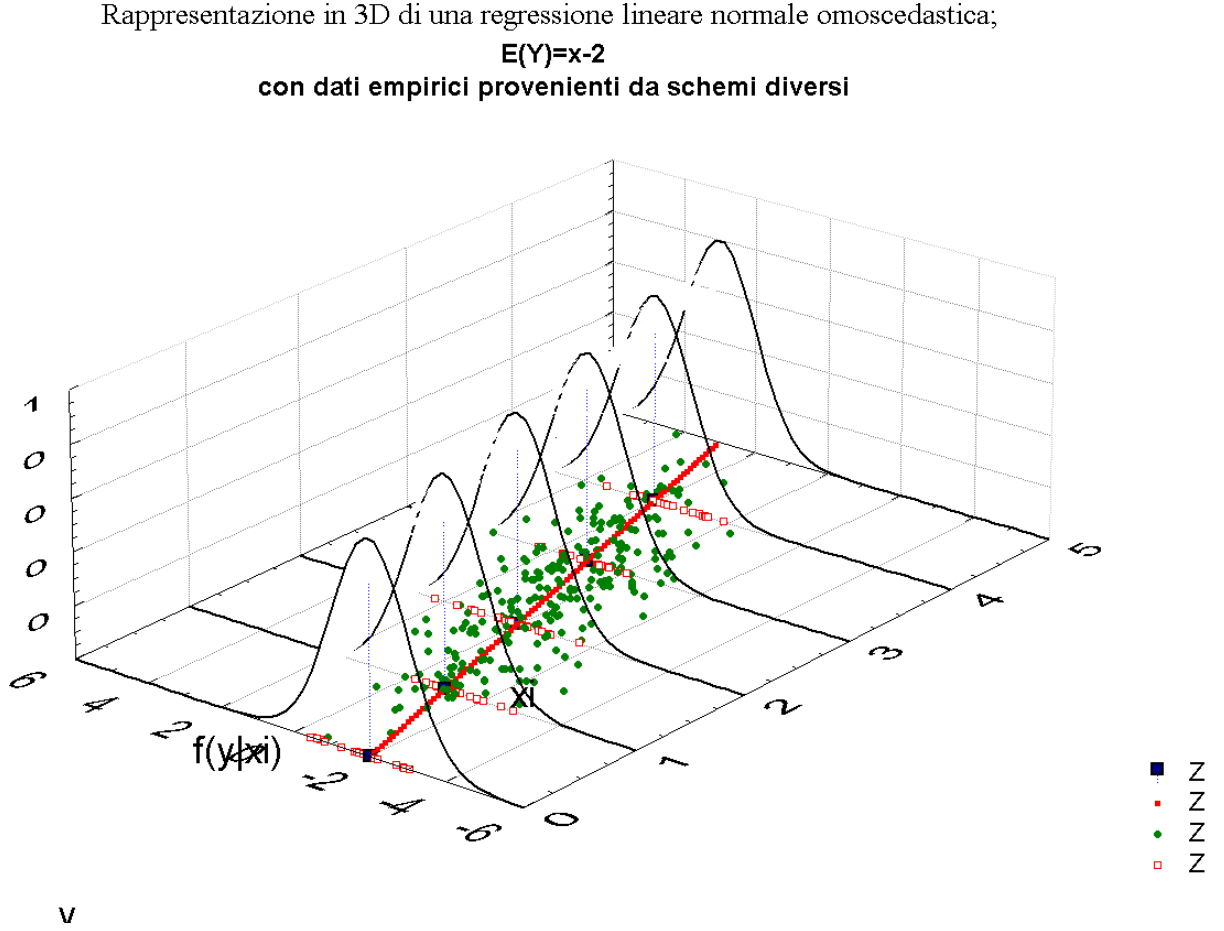

Figure 2: distribuzioni condizionate normali in corrispondenza di valori fissati

E' più opportuno allora fornire l'equazione per la variabile casuale $\mathbf{Y}_i$ corrispondente alla generica i-esima osservazione:

Equazione del modello lineare per l'i-esima osservazione

$$\mathbf{Y}_i = x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ik}\beta_k + \boldsymbol{\varepsilon}_i = \mathbf{x}_i^T\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$$

DRAFT

Il vettore delle n osservazioni può essere quindi così espresso formalmente in termini matriciali:

MODELLO LINEARE GENERALE

$$\mathbf{Y}_{[n\times 1]} = \mathbf{X}_{[n\times k]}\boldsymbol{\beta}_{[k\times 1]} + \boldsymbol{\varepsilon}_{[n\times 1]}$$

L' equazione deve essere *lineare nei parametri* $\boldsymbol{\beta}$.

Esplicitando l'espressione matriciale in n equazioni, si ha:
$\mathbf{Y}_{[n\times 1]} = \mathbf{X}_{[n\times k]}\boldsymbol{\beta}_{[k\times 1]} + \boldsymbol{\varepsilon}_{[n\times 1]}$ equivale a:

$$\begin{bmatrix} Y_1 \\ \dots \\ \dots \\ Y_i \\ \dots \\ \dots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ik} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \dots \\ \boldsymbol{\beta}_j \\ \dots \\ \boldsymbol{\beta}_k \end{bmatrix} + \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \dots \\ \dots \\ \boldsymbol{\varepsilon}_i \\ \dots \\ \dots \\ \boldsymbol{\varepsilon}_n \end{bmatrix}$$

ed anche a:

DRAFT

$$\begin{pmatrix} Y_1 \\ \dots \\ \dots \\ Y_i \\ \dots \\ \dots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11}\beta_1 + & x_{12}\beta_2 + & \dots & +x_{1k}\beta_k \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ x_{i1}\beta_1 + & x_{i2}\beta_2 + & \dots & +x_{ik}\beta_k \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ x_{n1}\beta_1 + & x_{n2}\beta_2 + & \dots & +x_{nk}\beta_k \end{pmatrix} + \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \dots \\ \dots \\ \boldsymbol{\varepsilon}_i \\ \dots \\ \dots \\ \boldsymbol{\varepsilon}_n \end{pmatrix}$$

DRAFT

Versatilità del modello lineare

L'utilità e la versatilità di tale modello per la descrizione di fenomeni reali risiede nella possibilità di dare un significato agli elementi di $\mathbf{X}$ e di $\boldsymbol{\beta}$.
Il nome lineare presuppone in generale che il modello sia lineare nei parametri β_j

4.1 Impieghi del modello lineare

La formulazione di tale modello per la speranza matematica di una v.a., sebbene molto semplice, permette di trattare diversi tipi di situazioni e di risolvere differenti problemi di inferenza.

Apparentemente siamo solo passati da una dipendenza da una variabile, ad una a più variabili.

In effetti possiamo rendere questo modello molto flessibile, come cominciamo a vedere adesso e come vedremo in tutto il resto del corso

DRAFT

In funzione di particolari configurazioni che può assumere la matrice $\mathbf{X}$, si può adattare questa impostazione a situazioni particolari.

Ad esempio:

- per l'analisi della regressione lineare multipla, se le colonne della matrice x sono n osservazioni di k variabili quantitative,
- per l'analisi della regressione polinomiale, se le colonne della matrice $\mathbf{X}$ sono le potenze di una o più variabili quantitative,
- oppure per l'analisi della varianza se le k colonne di x sono delle variabili dicotomiche indicatrici (dummy variables) di appartenenza ad un gruppo;
- per l'analisi della covarianza (una particolare analisi con variabili qualitative e quantitative);
- per particolari analisi di disegni sperimentali a più vie con interazioni fino ad un ordine massimo fissato.
- Analisi di superfici di risposta
- Analisi discriminante

DRAFT

- Analisi dei modelli di crescita
- altri modelli

soltanto alcune di queste problematiche verranno trattate in questi appunti;
si rivedano comunque gli esempi tratti dalla sezione di problemi introduttivi

DRAFT

5 Problemi di inferenza

In generale in un modello lineare possiamo avere diversi problemi di inferenza, in particolare di stima e di prova delle ipotesi, in funzione della natura dei dati e del tipo di problema. Ad esempio:

- stimare il vettore dei parametri $\boldsymbol{\beta}$ nel caso generale;
- stimare il vettore dei parametri $\boldsymbol{\beta}$ nel caso in cui vengono imposti dei vincoli su alcune delle sue componenti (alcune componenti nulle o uguali, per esempio)
- Il valore del vettore dei parametri $\boldsymbol{\beta}$ è uguale ad un certo valore β_0 ?
- Costruzione di una regione di confidenza per il vettore dei parametri $\boldsymbol{\beta}$;
- Costruzione di un intervallo di confidenza per una delle componenti di $\boldsymbol{\beta}$; (o per una combinazione lineare delle componenti di $\boldsymbol{\beta}$, ad esempio $\beta_1 - (\beta_2 + \beta_3)/2$).
- Inferenza su r componenti di $\boldsymbol{\beta}$; le altre $k - r$ componenti di $\boldsymbol{\beta}$ non interessano e svolgono però il ruolo di parametri di disturbo.

DRAFT

- Gli effetti di alcune variabili $\mathbf{X}_j$ sono uguali? Ossia alcuni dei parametri sono uguali?
- Alcuni dei parametri sono uguali subordinatamente al valore di altre variabili $\mathbf{X}_j$?
- Qual è la combinazione di fattori che fornisce la risposta media $\mathbf{Y}$ più elevata?
- Subordinatamente al fatto che alcuni effetti siano significativamente diversi da zero, quali hanno condotto alla significatività?
- Una o più fra le variabili $\mathbf{X}_j$ può essere eliminata, senza che questo riduca in modo sostanziale la spiegazione della variabile di risposta? Eliminare una variabile esplicativa $\mathbf{X}_j$ dal modello corrisponde ad ipotizzare $\beta_j = 0$.
- **Anche se β_j è significativamente diverso da zero, può comunque convenire lavorare con un modello ridotto anche se distorto?**

DRAFT

5.1 Componente sistematica e componente casuale.

Possiamo interpretare le due componenti fondamentali del modello che forniscono la risposta $\mathbf{Y}$ come:

- $\mathbf{X}\boldsymbol{\beta}$ la componente sistematica del modello;
- $\boldsymbol{\varepsilon}$ la componente accidentale, che qui sto supponendo additiva, per semplicità, e per comodità interpretativa.

Se:

$$\mathrm{E}[\boldsymbol{\varepsilon}] = 0_n$$

(come è ovvio assumere se $\boldsymbol{\varepsilon}$ è effettivamente una componente accidentale additiva) allora:

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta},$$

DRAFT

Approssimazione alla funzione di regressione

Dall'espressione

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta},$$

è evidente che $\mathbf{X}\boldsymbol{\beta}$ può essere vista come un'approssimazione alla vera funzione di regressione

e quindi il modello è schematizzabile come:

$$\mathbf{Y} = \mathrm{E}[\mathbf{Y}] + \boldsymbol{\varepsilon}$$

DRAFT

In questo caso quindi possiamo vedere la variabile $\mathbf{Y}$ come una variabile casuale, di cui abbiamo un campione di n osservazioni, la cui speranza matematica è funzione lineare di k variabili $\mathbf{X}_j$ secondo la relazione:

$$\mathrm{E}[\mathbf{y}_i] = \sum_{j=1}^{k} x_{ij}\beta_j \qquad i = 1, 2, \ldots, n$$

questa proprietà è in stretta relazione con l'ipotesi di additività della componente accidentale.

L'assunzione $\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}_n$ presuppone la validità del modello per le speranze matematiche e quindi implicitamente si ipotizza:

- che la componente accidentale (che ha un effetto additivo) sia a media nulla: questo in effetti è quasi scontato quando parliamo di errori accidentali additivi;
- che le k variabili siano le uniche rilevanti ai fini della spiegazione della speranza matematica di $\mathbf{Y}$, o meglio della spiegazione di sue variazioni.
- Il modello per la parte sistematica non è distorto, perchè: $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$.

In ogni caso non si sta implicitamente assumendo l'esistenza di relazioni di causa effetto fra le X e la Y , ma semplicemente che la conoscenza delle $\mathbf{X}$ può spiegare meglio la variabilità del fenomeno $\mathbf{Y}$ (nel senso che ne diminuisce la variabilità).

Ricordo inoltre che non è necessario ipotizzare modelli distributivi per le $\mathbf{X}_j$, perché, almeno per ora, si sta supponendo che siano dei termini fissati, secondo differenti possibili schemi che vengono esaminati fra breve.

Ad esempio nella regressione lineare semplice si ipotizza:

$$\mathbf{Y}_i = \beta_0 + x_i\beta_1 + \boldsymbol{\varepsilon}_i$$

DRAFT

con

$$\mathrm{E}\left[\mathbf{Y}_i\right] = \beta_0 + x_i\beta_1$$

DRAFT

DRAFT

Elemento e Dimensioni	Caratteristiche
$\mathbf{Y}$ vettore n elementi	Vettore aleatorio osservabile; è la variabile di risposta di interesse, ossia quella di cui si cerca di studiare (e di spiegare) la variabilità;
$\mathbf{X}$ matrice $n \times k$ elementi	Matrice di costanti note. Le k componenti (vettori di n elementi) sono variabili non aleatorie osservate senza errori Sono le k variabili esplicative che si pensa influenzino la risposta $\mathbf{Y}$. Si vedranno dopo alcune delle numerose configurazioni che può assumere la matrice $\mathbf{X}$.
$\boldsymbol{\beta}$ vettore k elementi	Vettore di parametri incogniti; $\boldsymbol{\beta}$ andrà stimato dai dati del campione. In generale i β_j sono dei parametri fissi; in certi modelli, che non tratteremo in questo corso se non per alcuni cenni, alcuni dei coefficienti sono considerati come effetti casuali, e quindi come variabili aleatorie.
$\boldsymbol{\varepsilon}$	Vettore aleatorio non osservabile direttamente;

DRAFT

5.3 Caratteristiche più dettagliate degli elementi del modello:

Elemento	Caratteristiche
$\mathbf{Y}$	Vettore aleatorio *osservabile* di n componenti

- è la variabile di risposta di interesse, ossia quella di cui si cerca di studiare (e di spiegare) la variabilità;
- è una variabile quantitativa;
- solo in casi speciali si considerano $\mathbf{Y}$ qualitative (ad esempio presenza/assenza; oppure successo/insuccesso). In questo corso non affronteremo, almeno non con queste tecniche, casi di risposte $\mathbf{y}$ qualitative non dicotomiche.
- Ci stiamo occupando essenzialmente di modelli nei quali la risposta $\mathbf{y}_i$ è univariata; diversamente, con risposte multiple, abbiamo modelli multivariati.
- Si considera $\mathbf{Y}$ come vettore aleatorio, perché si pensa che la sua distribuzione possa per qualche aspetto (speranza matematica, varianza, etc.) variare in funzione delle X_j .

DRAFT

- Il modello è multiplo se si hanno diverse colonne nella matrice $\mathbf{X}$
- con $\mathbf{y}$ indichiamo il vettore dei valori osservati
- Di solito è utile vedere (preliminarmente) se la variabilità osservata della $\mathbf{Y}$ è dovuta solo alla variabilità naturale o anche a fattori sistematici (ossia la dipendenza dalle $\mathbf{X}$).
- Le n unità dovrebbero essere gli elementi di un campione casuale; tuttavia questo modello viene utilizzato anche per analisi esplorative su dati osservazionali o comunque non provenienti da un campione

DRAFT

Elemento	Caratteristiche e Dimensioni
$\mathbf{X}$	Matrice di *costanti note* di $n \times k$ elementi • Le k componenti (vettori di n elementi) sono variabili non aleatorie osservate senza errori • o comunque con un eventuale errore di ordine di grandezza molto inferiore rispetto a quello di $\mathbf{Y}$. • I valori delle x potrebbero essere n valori particolari assunti da un vettore aleatorio p-dimensionale. In questo caso studiamo la distribuzione condizionata di $\mathbf{y}$ per quei particolari valori di $\mathbf{X}$. • Le $\mathbf{X}_j$ sono le k variabili esplicative che si pensa influenzino la risposta $\mathbf{Y}$.

DRAFT

Le configurazioni di $\mathbf{X}$ possono essere numerose:

• quantitative

- variabili indicatrici (0/1 o -1/1)
- variabili miste

La matrice delle $\mathbf{X}$ (o meglio l'intero insieme dei dati) può provenire da:

- studi osservazionali: in cui si scelgono le k variabili, ma gli n valori di ciascuna variabile sono quelli osservati negli n individui scelti, per cui non è possibile in generale pianificare particolari combinazioni degli $n \times k$ valori.
- esperimenti pianificati: in cui si scelgono non solo le k variabili, ma anche tutto lo schema degli $n \times k$ valori, per cui è possibile stabilire in anticipo quali valori verranno utilizzati per ciascuna delle k variabili ed inoltre quali combinazioni di valori dei fattori (o delle variabili) verranno impiegate insieme.
- dati ricavati da statistiche ufficiali o archivi e/o databases o dati prelevati da archivi remoti in rete: *possibilmente si tratta di dati raccolti non per finalità statistiche* e pertanto potrebbero essere poco affidabili, di qualità non nota e molto probabilmente non costituiscono nè un campione casuale nè una popolazione completa. [1]

DRAFT

[1] Ovviamente questa considerazione riguarda l'intero dataset osservato, compresa la $\mathbf{y}$.

Elemento	Caratteristiche
$\boldsymbol{\beta}$	Vettore di *parametri incogniti* di k elementi: $$\boldsymbol{\beta} = \{\beta_1, \beta_2, \ldots, \beta_j, \ldots, \beta_k\}^{\mathsf{T}}$$ **$\boldsymbol{\beta}$ andrà stimato dai dati del campione**

- Ciascun parametro esprime la dipendenza (lineare) dalla corrispondente variabile esplicativa.
- In generale gli elementi di $\boldsymbol{\beta}$ sono dei parametri fissi, se non precisato diversamente;
- in certi modelli alcuni dei coefficienti sono considerati come effetti casuali, e quindi come variabili aleatorie.
- Ciascun parametro esprime la dipendenza (lineare) dalla corrispondente variabile esplicativa.
- Pertanto β_j misura l'incremento medio della risposta $\mathbf{Y}$ in corrispondenza di un incremento unitario della j-esima variabile $\mathbf{X}_j$.

DRAFT

- Se $\mathbf{X}_j$ è una variabile indicatrice (0/1) della presenza di una certa caratteristica (non quantitativa), allora β_j misura l'effetto medio della presenza di tale caratteristica sulla risposta $\mathbf{Y}$.

In generale:

$$\beta_j = \frac{\partial \mathrm{E}\left[\mathbf{Y}_i\right]}{\partial x_{ij}}$$

se il modello è lineare però vale anche:

$$x_{ij} = \frac{\partial \mathrm{E}\left[\mathbf{Y}_i\right]}{\partial \beta_j}$$

DRAFT

Elemento	Caratteristiche
$\boldsymbol{\varepsilon}$	Vettore aleatorio *non osservabile* direttamente di n elementi *I*n funzione delle diverse ipotesi fatte sulla natura della distribuzione di $\boldsymbol{\varepsilon}$ (che può dipendere in generale da un insieme di parametri $\boldsymbol{\theta}$) si hanno differenti stimatori dei parametri incogniti del modello.

- $\boldsymbol{\varepsilon}$ rappresenta la componente accidentale, che viene supposta additiva, in modo tale che se è anche con speranza matematica nulla (come spesso si può ipotizzare) si ha:

DRAFT

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$$

- In effetti $\boldsymbol{\varepsilon}$ a rigore dovrebbe essere una variabile aleatoria non dipendente da variabili esterne, che esprime semplicemente l'errore sperimentale, o l'errore di misurazione
- nel caso in cui il modello non sia correttamente specificato, $\boldsymbol{\varepsilon}$ finirà per inglobare le variabili ed i fattori non esplicitati nella parte sistematica, e quindi perderà la

sua natura di componente accidentale. Se ne riparlerà molto più in là quando ci porremi il rpoblema di validare il modello assunto, mediante analisi dei residui.

5.4 Ipotesi sulle ε

Per potere dare una risposta, anche approssimativa, ad alcune delle domande relative agli elementi di $\boldsymbol{\beta}$, e quindi per la costruzione di stimatori e test, e per fare in generale inferenza (almeno muovendosi in un contesto parametrico), occorrerà fare ovviamente delle ipotesi, più o meno restrittive, sulla distribuzione di $\boldsymbol{\varepsilon}$. Questa distribuzione dipenderà in generale da un vettore di parametri $\boldsymbol{\theta}$:

$$\boldsymbol{\varepsilon} \sim \phi(\boldsymbol{\theta}).$$

DRAFT

E' ovvio che, anche ammettendo di conoscere la forma funzionale ϕ, occorre stimare il vettore di parametri $\boldsymbol{\theta}$.

ε non è osservabile!

Va tenuto presente che $\boldsymbol{\varepsilon}$ non è direttamente osservabile, come accade invece, ad esempio, quando si osserva un campione proveniente da una normale univariata di parametri incogniti μ (costante) e σ^2.

$\boldsymbol{\theta}$ svolge in generale il ruolo di parametro di disturbo.

Ovviamente il numero dei parametri incogniti θ_s non dovrà essere elevato, diversamente non sarà possibile stimarli.

Esempio: se si suppone $\boldsymbol{\varepsilon} \sim N(0, \boldsymbol{\Sigma})$ non possono essere incogniti **tutti** gli elementi della matrice di varianza e covarianza $\boldsymbol{\Sigma}$ (perchè sarebbero $n(n+1)/2$ parametri)

Quanto interagiscono la stima di θ e quella di $\boldsymbol{\beta}$? E' possibile in qualche modo verificare a posteriori la validità delle ipotesi fatte sulla distribuzione delle ε ?

Le possibili scelte verranno analizzate successivamente alla discussione sul significato della parte sistematica.

DRAFT

6 Introduzione ai Modelli Lineari: Parte II

7 La matrice delle X

La struttura ed il metodo di scelta delle X_j, insieme con la parametrizzazione scelta determina in parte il tipo di analisi.

Sostanzialmente le X_j (tutte o alcune) possono provenire da:

studi osservazionali Questo caso si presenta quando non è possibile in generale stabilire a priori la matrice $\mathbf{X}$: si sceglieranno solo le k particolari variabili da analizzare e le n unità che costituiscono il campione. Eventualmente potremo, entro certi limiti, operare alcune trasformazioni sulle x in modo da ricondurci a schemi particolari.

DRAFT

esperimenti pianificati con: variabili controllabili

in cui alcune variabili (h), e tutto lo schema degli $n \times h$ valori corrispondenti della matrice $\mathbf{X}$, vengono pianificati in anticipo, per cui si stabilisce in partenza il range di valori di ciascuna variabile esplicativa e le combinazioni di valori delle variabili esplicative che si vogliono osservare, in funzione delle risposte che si

vogliono ottenere dall'esperimento. *Con un esperimento mal pianificato, in cui ad esempio non sono previste alcune combinazioni di livelli di variabili, non si potranno per esempio condurre tutti i test che si possono effettuare con dati provenienti da un esperimento ben pianificato.*

variabili note ma il cui valore non è pianificabile Ad esempio vengono selezionati alcuni soggetti in base al sesso ed alla condizione lavorativa, per cui si stabilisce in anticipo quante osservazioni fare per tutte le combinazioni sesso x condizione lavorativa mentre per le altre variabili non è possibile pianificare dei valori particolari.

Particolari tipi di esperimenti con randomizzazione Ad esempio clinical trial

DRAFT

7.1 Osservazioni ripetute.

Alcune delle righe della matrice $\mathbf{X}$ potrebbero essere (volutamente o per caso) replicate. Nel caso di presenza di osservazioni ripetute per ciascuna combinazione di fattori, l'analisi potrà anche dire qualcosa di più:

FIGURA DA FARE

Figure 3: box-plot con retta di regressione e spezzata di regressione

- sulla bontà delle assunzioni fatte sulla distribuzione degli errori
- sulla forma funzionale della relazione (se lineare o meno).
- Sulla variabilità della componente accidentale per ciascuna combinazione di fattori.

DRAFT

7.2 Disegni fattoriali

Un disegno si dice fattoriale se vengono pianificate le osservazioni di tutte le possibili combinazioni dei livelli dei k fattori.

Pertanto se ogni fattore X_j può assumere m_j livelli ($j = 1, 2, \ldots, k$), si avranno: $C = \Pi_{j=1}^{k} m_j$ distinte combinazioni, ciascuna delle quali può anche essere replicata, per ottenere la matrice $\mathbf{X}$.

Esempio: In un esperimento farmacologico si vuole stimare l'effetto di un farmaco (tre dosi: una nulla, una media, una alta) su pazienti con una particolare pa-

tologia. Si vuole verificare anche l'effetto su pazienti sani, e vedere se il sesso del paziente influenza il tipo di risposta. Complessivamente si hanno i seguenti fattori e corrispondenti livelli:

Fattore	livelli (o modalità qualitative)
dosi di un farmaco	3 livelli di dose (bassa, media, alta)
Sesso	2 livelli
Condizione sperimentale	2 livelli: malati e sani
Totale:	12 combinazioni

DRAFT

Ai fini di quest'esempio non ha importanza specificare anche la variabile di risposta $\mathbf{y}$. Le 12 possibili combinazioni sono dunque:

	DOSE	SESSO	CONDIZIONE
1	Alta	F	Sano
2	Alta	F	Malato
3	Alta	M	Sano
4	Alta	M	Malato
5	Media	F	Sano
6	Media	F	Malato
7	Media	M	Sano
8	Media	M	Malato
9	Bassa	F	Sano
10	Bassa	F	Malato
11	Bassa	M	Sano
12	Bassa	M	Malato

DRAFT

Si può convenire di assegnare i seguenti valori numerici: (ovviamente supponendo

che per la dose i tre gradi bassa,media, alta corrispondano ad una scala numerica equispaziata, ossia:

dose(alta)-dose(media)=dose(media)-dose(bassa)

DOSE	Valore
Alta	+1
Media	0
Bassa	-1

SESSO	Valore
M	+1
F	-1

CONDIZIONE	Valore
Sano	+1
Malato	-1

Si ottiene la seguente matrice $\mathbf{X}$ dei regressori:

DRAFT

$\mathbf{X} =$

	DOSE	SESSO	CONDIZIO
1	+1	+1	+1
2	+1	+1	-1
3	+1	-1	+1
4	+1	-1	-1
5	0	+1	+1
6	0	+1	-1
7	0	-1	+1
8	0	-1	-1
9	-1	+1	+1
10	-1	+1	-1
11	-1	-1	+1
12	-1	-1	-1

DRAFT

Se i livelli sono quantitativi ed equispaziati (come in questo esempio), l'analisi risulta *ortogonale* (nel senso che i fattori sono non correlati, e vedremo in seguito che vantaggio porta nell'inferenza e nell'interpretazione dei parametri)

Anche nell'esempio che segue si ha un disegno bilanciato:

X1: 5 LIVELLI; X2 e X3 con 3 livelli

LIVELLI ORIGINALI			SCARTI DALLE MEDIE		
X1	X2	X3	Z1	Z2	Z3
1	0	0	-2	-1	-1
2	0	0	-1	-1	-1
3	0	0	0	-1	-1
4	0	0	1	-1	-1
5	0	0	2	-1	-1
1	1	0	-2	0	-1
2	1	0	-1	0	-1
3	1	0	0	0	-1
4	1	0	1	0	-1
5	1	0	2	0	-1
1	2	0	-2	1	-1
2	2	0	-1	1	-1
3	2	0	0	1	-1
4	2	0	1	1	-1
5	2	0	2	1	-1
1	0	1	-2	-1	0
2	0	1	-1	-1	0
3	0	1	0	-1	0
4	0	1	1	-1	0
5	0	1	2	-1	0
1	1	1	-2	0	0
2	1	1	-1	0	0
3	1	1	0	0	0
4	1	1	1	0	0
5	1	1	2	0	0
1	2	1	-2	1	0
2	2	1	-1	1	0
3	2	1	0	1	0
4	2	1	1	1	0
5	2	1	2	1	0
1	0	2	-2	-1	1
2	0	2	-1	-1	1
3	0	2	0	-1	1
4	0	2	1	-1	1
5	0	2	2	-1	1
1	1	2	-2	0	1
2	1	2	-1	0	1
3	1	2	0	0	1
4	1	2	1	0	1
5	1	2	2	0	1
1	2	2	-2	1	1
2	2	2	-1	1	1
3	2	2	0	1	1
4	2	2	1	1	1
5	2	2	2	1	1

DRAFT

$$\mathbf{Z}^\top\mathbf{Z} = \begin{pmatrix} 90 & 0 & 0 \\ 0 & 30 & 0 \\ 0 & 0 & 30 \end{pmatrix}$$

- Anche se non si riesce a costruire un disegno fattoriale, perché troppo oneroso, sarà opportuno di solito ricorrere a *disegni ortogonali* , ossia schemi di disegni sperimentali con variabili esplicative non correlate.

disegni ortogonali

- L'opportunità di avere l'ortogonalità dei fattori (ossia variabili non correlate) è pienamente giustificata solo nell'ambito della teoria normale completa sui minimi quadrati.
- Comunque è ragionevole fare in modo che i fattori non siano correlati (se possibile).
- In un esperimento a molti fattori sarà opportuno che siano bilanciate in corrispondenza a ciascuna coppia di fattori, le possibili combinazioni di coppie di livelli.

DRAFT

FIGURA DA FARE

Figure 4: Piano fattoriale di 3 regressori $2 \times 4 \times 3$

FIGURA DA FARE

Figure 5: Esempio di disegno fattoriale completo e incompleto

7.2.1 Disegni 2^k

Un caso particolare di disegno fattoriale si ha nel caso di k fattori qualitativi dicotomici, per cui le variabili assumeranno il valore 1 o 0 secondo che la caratteristica è presente o assente; è conveniente anche utilizzare i valori 1 e -1, in modo che in un piano fattoriale completo le variabili risulteranno centrate (ossia con media nulla) e con varianza unitaria.

Per esaminare tutte le combinazioni (senza repliche) occorre prevedere 2^k osservazioni.

DRAFT

Esempio 7.1 *Disegno fattoriale completo 4 fattori a due livelli -1,1.*
farmaco si/no;
sesso M/F;
malato si/no;

ospedalizzato si/no;
Si ottiene una matrice (centrata, ossia con medie nulle) con $16 = 2^4$ *righe:*

DRAFT

	Z1	Z2	Z3	Z4
1	1	1	1	1
2	1	1	1	-1
3	1	1	-1	1
4	1	1	-1	-1
5	1	-1	1	1
6	1	-1	1	-1
7	1	-1	-1	1
8	1	-1	-1	-1
9	-1	1	1	1
10	-1	1	1	-1
11	-1	1	-1	1
12	-1	1	-1	-1
13	-1	-1	1	1
14	-1	-1	1	-1
15	-1	-1	-1	1
16	-1	-1	-1	-1

DRAFT

$$\mathbf{Z}^\top\mathbf{Z} = \begin{pmatrix} 16 & 0 & 0 & 0 \\ 0 & 16 & 0 & 0 \\ 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 16 \end{pmatrix}$$

8 Regressione multipla.

DRAFT

L'informazione campionaria, relativa a n unità, è costituita da:

- Un vettore di n valori osservati $\mathbf{y}$ della variabile di risposta quantitativa $\mathbf{Y}$.
- La matrice $\mathbf{X}$ (n righe e k colonne) è data dai valori di k regressori quantitativi, noti, per ciascuna delle n osservazioni

Figura da inserire blocchi

$$\mathbf{y}_{[n\times 1]}, \mathbf{X}_{[n\times k]}$$

Le n unità osservate sono quindi costituite da k+1 variabili e sono schematizzabili nelle n righe:

$$(\mathbf{y}|\mathbf{X}) = \left(\begin{array}{c|cccc} \mathbf{y}_1 & x_{11} & x_{12} & \dots & x_{1k} \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline \mathbf{y}_i & x_{i1} & x_{i2} & \dots & x_{ik} \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline \dots & \dots & \dots & \dots & \dots \\ \hline \mathbf{y}_n & x_{n1} & x_{n2} & \dots & x_{nk} \end{array}\right)$$

La dipendenza (del valore atteso) di $\mathbf{y}$ dalle X_j è espressa quindi dalla relazione:

$$\mathrm{E}[\mathbf{y}_i] = \alpha + \sum_{j=1}^{k} x_{ij}\beta_j$$

DRAFT

abbiamo quindi k coefficienti di regressione incogniti β_j che esprimono la dipendenza media (parziale) della risposta da ciascun regressore.

In generale nel modello si considera anche un termine noto incognito α , che esprime la risposta media corrispondente a valori nulli dei regressori;

α di solito non è oggetto di particolare interesse ed usualmente svolge il ruolo di parametro di disturbo.

La relazione è analoga, almeno formalmente, alla relazione di regressione lineare che studia la dipendenza della speranza matematica di una variabile aleatoria rispetto ai valori (fissati!) di altre k variabili aleatorie.

Non si confonda la regressione multipla (una variabile di risposta e molti regressori) con la regressione multivariata (molte variabili di risposta e uno o più regressori).

DRAFT

8.1 Relazione di regressione in termini di scarti

Per comodità interpretativa, e per motivi più tecnici che si vedranno al momento di affrontare i problemi di stima, convenzionalmente si può porre:

la prima colonna ($j = 0$) composta tutta da 1 (in modo da prevedere la presenza di un termine noto);

le altre colonne costituite dagli scarti semplici rispetto alla media di ciascuna variabile.

Con la posizione:

$$z_{ij} = x_{ij} - M(\mathbf{X}_j) \qquad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, k$$

la matrice $\mathbf{X}$ può essere messa nella forma più conveniente:

$$\mathbf{X} = \begin{pmatrix} 1 & z_{11} & \dots & z_{1j} & \dots & z_{1k} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & z_{i1} & \dots & z_{ij} & \dots & z_{ik} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & z_{n1} & \dots & z_{nj} & \dots & z_{nk} \end{pmatrix}$$

$$\begin{pmatrix} \text{Media variabile:} & 1 & 0 & \dots & 0 & \dots & 0 \end{pmatrix}$$

Per i parametri si ha:

$$\boldsymbol{\beta}^{\mathsf{T}} = \{\beta_0, \beta_1, \dots, \beta_j, \dots, \beta_k\}$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_j \\ \vdots \\ \beta_k \end{pmatrix}$$

DRAFT

Termine noto

Coefficiente di regressione parziale variabile 1

$\vdots$

variabile j

$\vdots$

variabile k

Quindi la matrice dei regressori e il vettore dei coefficienti risultano partizionati in:

$$\mathbf{X} = [\mathbf{1_n}|\mathbf{Z}]$$
$$\boldsymbol{\beta}^{\mathsf{T}} = [\beta_0|\beta_{1,k}]$$

DRAFT

Il legame lineare è ora dato da:

$$E(\mathbf{y}_i) = \sum_{j=0}^{k} z_{ij}\beta_j$$

Per cui la risposta viene vista come somma di:

- un effetto generale, β_0 , corrispondente a livelli nulli degli scarti z_{ij} , e quindi a livelli medi dei regressori originari x_{ij}

- k singoli effetti proporzionali agli scarti dei singoli regressori dalla propria media.

Dal punto di vista interpretativo, la riscrittura in termini di scarti consente di dare un significato logico, ed utile per i confronti, al termine noto.

Rispetto alla parametrizzazione originaria si ha:

$$E(\mathbf{y}_i) = \sum_{j=0}^{k} z_{ij}\beta_j = \beta_0 + \sum_{j=1}^{k} z_{ij}\beta_j = \beta_0 - \sum_{j=1}^{k} M(\mathbf{X}_j)\beta_j + \sum_{j=1}^{k} x_{ij}\beta_j$$

Quindi:

i coefficienti di regressione sono sempre uguali (si sono solo effettuate delle traslazioni di assi!)

Per il termine noto:

$$\alpha = \beta_0 - \sum_{j=1}^{k} M(\mathbf{X}_j)\beta_j$$

DRAFT

L'utilità teorica e pratica di queste posizioni sarà chiarita nella parte relativa all'inferenza nella regressione lineare. In ogni caso continuerò ad indicare la matrice del disegno o dei regressori con $\mathbf{X}$, precisando eventualmente se si tratta di scarti o di variabili originarie.

L' ipotesi nulla che più spesso si vuole verificare (almeno preliminarmente) è:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0; \qquad \text{con} \qquad \beta_0 \qquad \text{qualsiasi.}$$

Ossia che il valore atteso della variabile dipendente sia costante ed indipendente dai regressori.

Figura da inserire ESEMPIO

DRAFT

8.2 Regressione polinomiale:

Dal momento che la linearità va intesa rispetto ai parametri, e non rispetto alle $\mathbf{X}_j$, il modello lineare comprende anche la regressione polinomiale in una o più variabili:

Regressione polinomiale di grado k in un regressore Z se

$$\mathrm{E}[\mathbf{y}_i] = \sum_{j=0}^{k} \beta_j z_i^j; \qquad i = 1, 2, \ldots, n$$

Ci si riporta al caso generale del modello lineare ponendo:

$$x_{ij} = z_i^j \beta_j; \qquad i = 1, 2, \ldots, n; \quad j = 0, 1, \ldots, k.$$

Anche in questo caso si continua a parlare di modelli lineari, pochè il termine lineare si riferisce sempre ai parametri e non ai regressori.

Si noti come però i regressori risultino in generale correlati, a meno che non si faccia ricorso a particolari trasformazioni del modello polinomiale basate sui polinomi ortogonali.

8.2.1 Polinomi in più variabili e superfici di risposta

DRAFT

E' immediata la generalizzazione alle superfici polinomiali di grado k in p regressori.

Regressione polinomiale di grado k in p regressori Z_h

$$E(\mathbf{y}_i) = \sum_{j=0}^{k} \cdots \sum_{j=0}^{k} \beta_{j1,j2,\ldots,jp} \prod_{\sum j_h = j} (z_{ih})^{jh}; i = 1, 2, \ldots, n$$

In particolare se $k = 2$ e se i coefficienti dei termini di secondo grado in ciascun regressore sono nulli, si possono convenientemente quantificare ed inserire nel modello degli effetti di interazione moltiplicativi del tipo $\beta_{hr} z_{ih} z_{ir}$ (interazione del primo ordine fra il regressore r-esimo ed h-esimo; Termini moltiplicativi che coinvolgono k regressori sono relativi ad effetti di interazione di grado $k - 1$

9 Fattori qualitativi e analisi della varianza

9.1 Regressori del tipo 0/1 (dummy variables)

Esiste un modo formale di esplicitare la matrice $\mathbf{X}$ in modo da trattare anche variabili esplicative di tipo qualitativo. Vediamo come prima con un esempio relativo ad una situazione nota.

Si supponga la situazione classica del confronto delle medie μ_1 e μ_2 di due popolazioni normali con uguale varianza sulla base delle informazioni di due campioni casuali semplici indipendenti.

Per la speranza matematica della variabile casuale associata alla generica osservazione abbiamo:

$$\mathrm{E}\left[\mathbf{Y}_i\right] = \mu_j \qquad \text{per} \qquad j = 1, 2,$$

secondo che l'unità i-esima appartenga al primo o al secondo campione.

Possiamo indicare sinteticamente:

$$\mathrm{E}\left[\mathbf{Y}_i\right] = x_{i1}\mu_1 + x_{i2}\mu_2$$

introducendo due regressori con la convenzione che per le unità del primo campione si ha: $x_{i1} = 1$ e $x_{i2} = 0$, per le unità del secondo campione si ha invece:

DRAFT

$x_{i1} = 0$ e $x_{i2} = 1$.

Oppure si può parametrizzare con:

$$\mathrm{E}\left[\mathbf{Y}_i\right] = \mu_1 + x_{i2}(\mu_2 - \mu_1)$$

e l'ipotesi da verificare sarà:

$$H_0 : \delta = (\mu_2 - \mu_1) = 0$$

con μ_1 qualsiasi.

(oppure si vorranno costruire intervalli di confidenza per δ)

L'aspetto essenziale di questo esempio è che anche questa situazione standard è riconducibile ad un modello lineare.

DRAFT

Esempio: Si hanno due campioni indipendenti di 14 osservazioni relative ad una variabile quantitativa, suddivise in due gruppi A e B, rispettivamente di numerosità 6 e 8.

A	2; 3; 3,1; 4; 5; 5,3.
B	3; 4,1; 4,3; 4,8; 6, 6,5; 7; 7,2.

Potremmo pensare di avere rilevato 3 variabili su 14 individui nel modo che segue:

DRAFT

$\mathbf{y}$	$\mathbf{x_A}$	$\mathbf{x_B}$
2	1	0
3	1	0
3,1	1	0
4	1	0
5	1	0
5,3	1	0
3	0	1
4,1	0	1
4,3	0	1
4,8	0	1
6	0	1
6,5	0	1
7	0	1
7,2	0	1

DRAFT

Sarà bene che da ora in poi lo studente si abitui a questa impostazione, in particolare per problemi con più variabili, perché risulta estremamente comoda in particolare per le situazioni complesse; (per la situazione dell'esempio, ossia test t a due campioni, non v'è alcun motivo pratico di ricorrere a tale formulazione, perché l'impostazione standard è quella più utile)

9.2 Analisi della varianza ad effetti fissi ed un criterio di classificazione

DRAFT

La versatilità del modello lineare, almeno da un punto di vista formale, si coglie per situazioni apparentemente lontane da quelle della regressione multipla, ossia per lo studio della dipendenza in media di una variabile quantitativa da una qualitativa (o più variabili qualitative).

Si supponga di avere n osservazioni suddivise in k gruppi indipendenti secondo le k modalità di un criterio di classificazione semplice (o mutabile sconnessa).

Si suppone che i gruppi siano internamente omogenei, ma che le medie dei gruppi possano essere in generale diverse:

$$E(\mathbf{Y}_i) = \mu_j$$

La matrice $\mathbf{X}$ è ora composta da k colonne costituite dagli n indicatori dell'appartenenza delle unità ai gruppi:

MATRICE del disegno:

DRAFT

$$
\mathbf{X} = \begin{array}{c}
\begin{array}{cccccc} \multicolumn{6}{c}{\text{Gruppo}} \\ 1 & 2 & \ldots & j & \ldots & k \end{array} \\
\left(\begin{array}{cccccc}
1 & 0 & & 0 & & 0 \\
\ldots & 0 & \ldots & \ldots & \ldots & 0 \\
1 & 0 & \ldots & \ldots & \ldots & 0 \\
0 & 1 & \ldots & \ldots & \ldots & 0 \\
0 & \ldots & \ldots & \ldots & \ldots & 0 \\
0 & 1 & \ldots & \ldots & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & \ldots & 1 & \ldots & 0 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
0 & 0 & \ldots & 0 & \ldots & 1 \\
0 & 0 & \ldots & 0 & \ldots & \ldots \\
0 & 0 & \ldots & 0 & \ldots & 1
\end{array} \right)
\end{array}
\begin{array}{c}
i \\
\left\{ \begin{array}{c}
1 \\
\ldots \\
n_1 \\
\ldots \\
\ldots \\
n_1 + n_2 \\
\ldots \\
\ldots \\
n_1 + n_2 + \ldots + n_j \\
\ldots \\
\ldots \\
n_1 + n_2 + \ldots + n_k
\end{array} \right.
\end{array}
$$

DRAFT

$$\boldsymbol{\beta} = \begin{pmatrix} \mu_1 \\ \dots \\ \mu_j \\ \dots \\ \mu_k \end{pmatrix}$$

DRAFT

Abbiamo:

n_j osservazioni per ogni trattamento o gruppo:

$$n_j = \sum_{i=1}^{n} x_{ij}; \qquad j = 1, 2, \ldots k.$$

ogni unità $\mathbf{U}_i$ appartiene ad un solo trattamento:

$$\sum_{j=1}^{k} x_{ij} = 1; \qquad i = 1, 2, \ldots, n$$

$x_{ij} = 1$ se e solo se l'unità $\mathbf{U}_i$ appartiene al j -esimo trattamento

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

L' ipotesi nulla di interesse è di solito quella di omogeneità:

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

oppure

$$H_0 : \beta_1 - \beta_k = \beta_2 - \beta_k = \ldots = \beta_{k-1} - \beta_k = 0$$

Con questa parametrizzazione $\mathbf{X}$ ha rango pieno k , ma l'ipotesi nulla di omegenità far le medie impone $k - 1$ vincoli

Altro modo di impostare l' analisi della varianza a una via:

DRAFT

$\beta_j = \mu_j - \mu$	effetto del trattamento (o del gruppo) $j; j = 1, 2, \ldots k$.
$\beta_{k+1} = \mu$	media generale;

e stavolta la matrice del disegno è:

DRAFT

$$
\mathbf{X} = \left(\begin{array}{cccccccc}
 & & & Effetti & & & & \\
\text{gr.1} & \text{gr.2} & \ldots & \text{gr.}j & \ldots & \text{gr.}k & \text{generale} & \\
1 & 0 & & 0 & & 0 & 1 & \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 1 & \\
1 & 0 & \ldots & \ldots & \ldots & 0 & 1 & \\
0 & 1 & \ldots & \ldots & \ldots & 0 & 1 & \\
0 & \ldots & \ldots & \ldots & \ldots & 0 & \ldots & \\
0 & 1 & \ldots & \ldots & \ldots & 0 & 1 & \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \\
0 & 0 & \ldots & 1 & \ldots & 0 & 1 & \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \\
0 & 0 & \ldots & 0 & \ldots & 1 & 1 & \\
0 & 0 & \ldots & 0 & \ldots & \ldots & \ldots & \\
0 & 0 & \ldots & 0 & \ldots & 1 & 1 &
\end{array}\right)
\left\{\begin{array}{c}
i \\
1 \\
\ldots \\
n_1 \\
\ldots \\
\ldots \\
n_1 + n_2 \\
\ldots \\
\ldots \\
n_1 + n_2 + \ldots + n_j \\
\ldots \\
\ldots \\
n_1 + n_2 + \ldots + n_k
\end{array}\right.
$$

DRAFT

$$\boldsymbol{\beta} = \begin{pmatrix} \mu_1 - \mu \\ \dots \\ \mu_j - \mu \\ \dots \\ \mu_k - \mu \\ \mu \end{pmatrix}$$

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$ e μ qualsiasi (k vincoli)

In questo caso però $\mathbf{X}$ ha una colonna linearmente dipendente dalle altre, per cui ha rango k invece di $k + 1$.

DRAFT

9.3 Analisi della varianza ad effetti fissi con due criteri di classificazione

E' possibile estendere il disegno precedente all'analisi della varianza a due vie, per la quale si può impostare un modello lineare con rs colonne, con:

$\mathbf{X}_{ijm} = 1$ se $\mathbf{U}_i$ appartiene al j -esimo trattamento di riga e all' m -esimo trattamento di colonna.

Oppure si può partire da una matrice del disegno sperimentale semplificata con

$r + s + 1$ colonne x e Z , tali che:

$x_{i0} = 1$ effetto generale;

$x_{ij} = 1$ se $\mathbf{U}_i$ appartiene al j -esimo trattamento di riga

$z_{im} = 1$ se $\mathbf{U}_i$ appartiene all' m -esimo trattamento di colonna

e introdurre nel modello di descrizione dei dati dei termini moltiplicativi (che saranno 1 solo se $\mathbf{U}_i$ appartiene ad una riga e ad una colonna) per considerare l'effetto di interazione:

$$\mathbf{y}_{ijk} = \beta_0 + \sum_{j=1}^{r} \alpha_j x_{ij} + \sum_{m=1}^{s} \eta_m z_{imj} + \sum_{j=1}^{r} \sum_{m=1}^{s} \gamma_{jm} x_{ij} z_{im} + \varepsilon_{ijk}$$

DRAFT

In pratica si considerano le due matrici di appartenza ai gruppi per i due criteri di classificazione separatamente; se nel modello occorre tener conto dell'appartenza simultanea (termini di interazione) si farà riferimento ai termini moltiplicativi $x_{ij}z_{im}$, che sono uguali ad 1 solo per le unità che appartengono alla modalità j -esima del primo criterio di classificazione ed alla modalità m -esima del secondo criterio di classificazione.

Le ipotesi da verificare sono quelle usuali (si vedranno in dettaglio nella parte inferenziali relativa all'analisi della varianza a due vie); con questa parametrizzazione

però, peraltro molto comoda e naturale, il modello ha parametri ridondanti (rango $= rs$; parametri $1 + r + s + rs$).

In modo analogo si possono impostare modelli a più vie.

9.4 Analisi della covarianza

(L'utilità dell'analisi della covarianza verrà esaminata più avanti)

Supponendo di avere n osservazioni suddivise in k gruppi secondo un criterio di classificazione semplice e relative ad una variabile di risposta $\mathbf{y}$ e ad una singola variabile concomitante x ci si può ricondurre al modello lineare generale ponendo:

$$z_{ij} = x_{ij} - M_j(x) \qquad j = 1, 2, \ldots, k$$

ove $M_j(x)$ è la media di x per le sole osservazioni del gruppo j .

DRAFT

La matrice $\mathbf{X}$ sarà composta da $2k$ colonne, di cui le prime k sono date da:

$$\mathbf{X}_1 = \begin{pmatrix} z_{1,1} & \dots & 0 \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ z_{n1,1} & \dots & 0 \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & z_{ij} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \dots & z_{n1,k} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \dots & z_{nk,k} \end{pmatrix}$$

DRAFT

mentre le altre k colonne sono costituite dalla matrice di appartenenza ai gruppi:

$$\mathbf{X}_2 = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & 0 \\ \dots & 0 & \dots & \dots & \dots & 0 \\ 1 & 0 & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 \\ 0 & 0 & \dots & 0 & \dots & \dots \\ 0 & 0 & \dots & 0 & \dots & 1 \end{pmatrix}$$

DRAFT

per cui la matrice $\mathbf{X}$ è costituita dalle colonne di $\mathbf{X}_1$ e $\mathbf{X}_2$ affiancate ossia:

$$\mathbf{X} = \mathbf{X}_1 | \mathbf{X}_2,$$

e i $2k$ parametri sono:

$$\boldsymbol{\beta}^{\mathsf{T}} = \beta_1, \dots, \beta_j, \dots, \beta_k, \alpha_1, \dots, \alpha_j, \dots, \alpha_k$$

Ipotesi di interesse:

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

con β_1 , α_1 qualsiasi ($2k - 2$ vincoli)

rette di regressione uguali nei k gruppi.

In generale si possono costruire disegni più complessi, con più variabili concomitanti e con più regressori, considerando un modello lineare del tipo:

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\varepsilon}$$

in cui:

- $\mathbf{X}_1$ è una matrice a più regressori,
- $\mathbf{X}_2$ è una matrice di indicatori per più criteri di classificazione,
- β_1 è il vettore dei parametri che esprimono la dipendenza della variabile di risposta dalle variabili concomitanti
- β_2 è il vettore dei parametri che esprimono la dipendenza della variabile di risposta dai fattori di classificazione.

DRAFT

9.5 Rette o piani di regressione con pendenze diverse: termini polinomiali moltiplicativi

Una relazione polinomiale con termini lineari e termini misti di 2° grado può esprimere la presenza di effetti di interazione in un modello lineare:

Esempio 1:

Si supponga una dipendenza in media della risposta $\mathbf{y}$ da due fattori quantitativi secondo la relazione:

$$E(\mathbf{y}_i) = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i1}x_{i2}\beta_{12}$$

DRAFT

Se il parametro β_{12} fosse uguale a zero avremmo un classico piano di regressione: $E(\mathbf{y}_i) = x_{i1}\beta_1 + x_{i2}\beta_2$, in cui parametri sarebbero interpretabili nel modo già visto (modello additivo).

Se invece tale parametro è diverso da zero, è presente un effetto di interazione fra i regressori $\mathbf{X}_1$ e $\mathbf{X}_2$: infatti per esempio la dipendenza di $\mathbf{y}$ dal regressore $\mathbf{X}_1$, per ciascuno dei possibili livelli di $\mathbf{X}_2$, è sempre lineare, ma l'inclinazione, e quindi la forza della dipendenza di $\mathbf{y}$ da $\mathbf{X}_1$, dipendono dal particolare livello assunto da $\mathbf{X}_2$. Il parametro β_1 non misura più la dipendenza parziale di $\mathbf{y}$ da $\mathbf{X}_1$, per qualsiasi livello di $\mathbf{X}_2$, ma solo la dipendenza media rispetto a tutti i livelli di $\mathbf{X}_2$.

FIGURA DA FARE

Figure 6: Una relazione con un termine moltiplicativo $x_1 \times x_2$ ha relazioni parziali lineari

FIGURA DA FARE

Figure 7: Pendenza diversa come effetto interazione fra due fattori x_1, x_2

Esempio di polinomio di secondo grado per effetto interazione: Supponiamo per esempio

$$\beta_1 = 1, \qquad \beta_2 = 3, \qquad \beta_{12} = 2,$$

per cui:

$$\mathrm{E}[\mathbf{y}_i] = x_{i1}1 + x_{i2}3 + x_{i1}x_{i2}2$$

L'effetto interazione fra $\mathbf{X}_1$ e $\mathbf{X}_2$ è tale da modificare anche il tipo di dipendenza di $\mathbf{y}$ da $\mathbf{X}_1$ (da negativa a positiva)

Si vedano nel grafico seguente le tre rette di regressione ottenute per tre diversi valori di $\mathbf{X}_2$ (-1;0;+1)

DRAFT

Esempio 2: (confronto fra due rette) Pendenza diversa come effetto interazione fra un fattore (o regressore) quantitativo e un fattore qualitativo:

Si supponga che la relazione di una risposta $\mathbf{y}$ da un regressore $\mathbf{X}_1$ dipenda anche da una variabile dicotomica: In questo caso la differenza di pendenza può essere inserita nel modello lineare mediante l'introduzione di un termine moltiplicativo, che non altera la linearità delle relazioni parziali, ma consente l'interpretazione dell'interazione fra i due fattori. ($\mathbf{X}_1$ può essere formato da un gruppo di regressori: l'esempio resta sostanzialmente inalterato) Per semplicità possiamo considerare la variabile dicotomica $\mathbf{X}_2$ con due livelli:-1 e +1, per cui ci riportiamo formalmente al caso precedente:

DRAFT

$$E(\mathbf{y}_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i1}x_{i2}\beta_{12} =$$

$$= (\beta_0 + x_{i2}\beta_2) + x_{i1}(\beta_1 + x_{i2}\beta_{12})$$

e quindi:

$$\mathrm{E}\left[\mathbf{y}_i\right] = \begin{cases} (\beta_0 - \beta_2) + x_{i1}(\beta_1 - \beta_{12}) & \text{se} \quad x_{i2} = -1 \\ (\beta_0 + \beta_2) + x_{i1}(\beta_1 + \beta_{12}) & \text{se} \quad x_{i2} = +1 \end{cases}$$

Da cui risulta evidente, ed utile da un punto di vista interpretativo, che β_2 rappresenta un effetto (medio) del fattore $\mathbf{X}_2$ sul livello medio di $\mathbf{y}_i$, mentre β_{12} rappresenta l'effetto (medio) del fattore $\mathbf{X}_2$ sulla relazione fra $\mathbf{y}$ e $\mathbf{X}_1$, per cui rappresenta un *effetto di interazione* (di primo ordine).

Risulta quindi irrilevante o comunque poco interessante dal punto di vista pratico, con questa interpretazione dei parametri, un test costruito per la verifica dell'ipotesi: $H_0 : \beta_1 = 0$, perché questo misurerebbe l'effetto marginale del primo regressore, senza tenere conto del livello dell'altro regressore (o meglio per un livello nullo, o medio, del secondo fattore). Se per esempio il fattore dicotomico $\mathbf{X}_2$ fosse il sesso (M=-1;F=+1), tale effetto marginale sarebbe di nessun interesse, perché ogni soggetto sarà o M o F, e quindi anche se risultasse $\beta_1 = 0$, in effetti la dipendenza della risposta dal regressore $\mathbf{X}_1$ sarebbe $-\beta_{12}$ per i maschi e $+\beta_{12}$ per le femmine. Eventualmente occorrerebbe prima saggiare l'ipotesi: $H_0 : \beta_{12} = 0$

Termini moltiplicativi con più termini possono servire per quantificare effetti di interazione di ordine superiore al primo.

Abbiamo già fatto cenno a questo argomento quando abbiamo parlato di distribuzioni condizionate nella normale multivariata; ricordo infatti che in una dis-

DRAFT

tribuzione normale multivariata la correlazione fra due variabili condizionata ai valori singoli di un'altra variabile o di più variabili è sempre la stessa, indipendentemente dai livelli assunti dalla III variabile. In altri termini nella distribuzione normale multivariata si è già visto che la dipendenza di $\mathbf{y}$ da x non varia in funzione dei livelli di una terza variabile z : questo è analogo al concetto di assenza di interazione, con l'avvertenza che in effetti il concetto di interazione può essere introdotto senza la necessità di riferirsi ad un modello probabilistico multivariato.

DRAFT

9.6 Modelli autoregressivi

Un caso speciale è costituito dall'osservazione di una serie temporale, cioè si dispone di n osservazioni eseguite ad intervalli di tempo uguali.

Si può pensare, in assenza di informazioni esterne o comunque di altre variabili, di volere studiare la dipendenza della serie dalla stessa serie spostata di uno o più unità temporali; in pratica si ipotizza che Y_t, osservazione al tempo t, o meglio, la sua speranza matematica $\mathrm{E}[Y_t]$, dipenda linearmente dall'osservazione precedente y_{t-1}.

Supponiamo quindi di volere spiegare la variabilità di una serie mediante i soli

valori della serie stessa in tempi precedenti; sarà in realtà opportuno fare delle ipotesi sul processo stocastico che ha generato la serie (ossia che sia stazionario), per cui la serie non ha certamente componenti di trend.

Possiamo, prima di ipotizzare particolari processi stocastici che possono avere generato la serie, adottare un approccio analogo alla regressione lineare, cercando la relazione di regressione che fa dipendere Y_t da Y_{t-1}. In pratica impostiamo un modello di regressione (detto *modello autoregressivo*) nel quale la serie originaria svolge il ruolo della variabile di risposta, mentre la Y_{t-1} svolge il ruolo di regressore o variabile esplicativa.

modello autoregressivo

DRAFT

serie originaria serie arretrata di una unità temporale

$$\left\{\begin{array}{c} y_2 \\ y_3 \\ \vdots \\ y_t \\ y_{t+1} \\ \vdots \\ y_n \end{array}\right\} \qquad \left\{\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_{t-1} \\ y_t \\ \vdots \\ y_{n-1} \end{array}\right\}$$

DRAFT

Evidentemente questo approccio presuppone serie equiintervallate

Ovviamente la dipendenza da valori precedenti può essere estesa anche a valori distanziati di più di un intervallo temporale:

Si può proseguire il ragionamento pensando che y_t sia influenzato non solo dalla precedente determinazione y_{t-1} ma anche da y_{t-2} e dalle precedenti osservazioni fino a y_{t-k}.

$$\begin{array}{cccc} \text{serie originaria} & \text{serie } y_{t-1} & \text{serie } y_{t-2} & \dots & \text{serie } y_{t-k} \\ \left\{ \begin{array}{c} y_{k+1} \\ y_{k+2} \\ \vdots \\ y_t \\ y_{t+1} \\ \vdots \\ y_n \end{array} \right. & \left\{ \begin{array}{c} y_k \\ y_{k+1} \\ \vdots \\ y_{t-1} \\ y_t \\ \vdots \\ y_{n-1} \end{array} \right. & \left\{ \begin{array}{c} y_{k-1} \\ y_k \\ \vdots \\ y_{t-2} \\ y_{t-1} \\ \vdots \\ y_{n-2} \end{array} \right. & \dots & \left\{ \begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_{t-k} \\ y_{t-k+1} \\ \vdots \\ y_{n-k} \end{array} \right. \end{array}$$

DRAFT

10 Generalizzazioni e modelli non lineari (cenni)

Possiamo pensare che la speranza matematica della risposta sia una funzione qualsiasi dei parametri e delle variabili indipendenti X_j:

Modello non lineare con errori additivi.

$$\mathbf{Y} = f(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon} \qquad \text{con:} \qquad \mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$$

$\mathbf{f}(\cdot)$ vettore di funzioni non lineari.

Modello non lineare con legame qualsiasi fra componente accidentale e sistematica.

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

DRAFT

Modello non lineare con errori moltiplicativi.

$$\mathbf{Y}_i = f_i(\mathbf{X}_; \boldsymbol{\beta}) \times \boldsymbol{\varepsilon}_i$$

GLM: Generalized Linear Models Modelli Lineari Generalizzati:

$$\mathrm{E}[\mathbf{Y}] = h(\mathbf{X}\boldsymbol{\beta}) \qquad \eta(\mathrm{E}[\mathbf{Y}]) = \mathbf{X}\boldsymbol{\beta}$$

La speranza matematica della variabile di risposta è funzione ($h(.)$ non lineare) del *predittore lineare* $\mathbf{X}\ \boldsymbol{\beta}$.

Si tratta ancora di modelli non lineari, ma con la particolarità che la dipendenza dalle X_j è scomposta in due parti:

- la funzione di legame (unica)
- un predittore lineare $\mathbf{x}_i^\top \boldsymbol{\beta}$

Questa impostazione consente di attribuire alla matrice $\mathbf{X}$ e al vettore di parametri $\boldsymbol{\beta}$ significati simili a quelli assunti nei modelli lineari.

Una sottoclasse di GLM molto impiegata nelle applicazioni è quella in cui la distribuzione della componente accidentale appartiene alla famiglia di distribuzioni esponenziale.

DRAFT

10.0.1 Regressione logistica

La probabilità del verificarsi di un evento (variabile di risposta) dipende dalle variabili $\mathbf{X}_j$.

10.0.2 Regressione piecewise

Una relazione di regressione può essere individuata da una spezzata, ossia da una retta che cambia inclinazione in corrispondenza dei livelli delle variabili esplicative. Nel caso in cui i punti di cambio dell'inclinazione non siano noti, il problema è configurabile nell'ambito dei modelli non lineari (*non lineari rispetto ai parametri*!)

10.0.3 Approssimazione di modelli non lineari

Eventualmente un modello lineare può essere visto come approssimazione del primo ordine di un modello non lineare

DRAFT

10.0.4 Modelli additivi generalizzati

I modelli additivi generalizzati, noti in letteratura come GAM (Generalized additive models), sono modelli che generalizzano i modelli lineari e le tecniche di regressione non parametrica.

Si suppone che la dipendenza della risposta quantitativa osservabile $\mathbf{Y}$ da k

regressori $vecx_j$ sia modellizzabile attraverso la relazione:

$$\mathrm{E}[\mathbf{Y}] = g\left(\sum_{j=1}^{k} f(x_j)\right). \tag{1}$$

Le singole $f(x_j)$ andranno stimate in modo non parametrico.

I modelli GAM rapprentano una estensione comoda della regressione non parametrica al caso multivariato, ipotizzando la separabilità dei vari contributi. E' anche possibile confrontare modelli con un numero di regressori differenti e scegliere quindi un sottoinsieme di variabili esplicative che spiegano bene le variazioni (non lineari) di $\mathbf{Y}$.

DRAFT

11 [

Stima dei parametri del modello lineare] Stima dei parametri del modello lineare (modelli a rango pieno)

Adesso, dopo avere visto alcuni dei più importanti impieghi del modello lineare per la descrizione di relazioni statistiche di dipendenza di natura varia, e le diverse interpretazioni dei parametri e delle variabili del modello, passiamo ad affrontare i problemi di inferenza, ossia di stima dei parametri, costruzione di test, di intervalli e regioni di confidenza ed altro.

L'approccio che seguiremo, di tipo parametrico, **è fondato interamente sulla verosimiglianza** e viene esposto prima con riferimento ad un modello lineare generico a rango pieno; una volta esposte le caratteristiche fondamentali dell'inferenza per il caso generico, si passerà ad esaminare problemi relativi a modelli particolari, principalmente per l'*analisi della regressione multipla* e per l'*analisi della varianza*.

DRAFT

Si supponga che:

Modello lineare generale

$$\mathbf{Y}_{[n\times 1]} = \mathbf{X}_{[n\times k]}\beta_{[k\times 1]} + \boldsymbol{\varepsilon}_{[n\times 1]}$$

DRAFT

con:

$\mathbf{Y}_{[n\times 1]}$ il vettore degli n valori osservati della variabile di risposta

$\mathbf{X}_{[n\times k]}$ una matrice nota, costituita dagli n valori osservati delle k variabili esplicative (regressori)

$\beta_{[k\times 1]}$ il vettore di k parametri da stimare in generale completamente incognito.

$\boldsymbol{\varepsilon}_{[n\times 1]}$ un vettore di n variabili casuali non osservabili, la cui distribuzione dipende in genere da un vettore $\boldsymbol{\theta}$ incognito di parametri di disturbo.

DRAFT

Ovviamente per potere stimare i parametri $\boldsymbol{\beta}$ e $\boldsymbol{\theta}$ mediante il metodo della massima verosimiglianza occorre fare delle ipotesi sulla distribuzione congiunta delle componenti di $\boldsymbol{\varepsilon}$.

In ogni caso sarà necessario fare qualche ipotesi su $f(\boldsymbol{\varepsilon}; \boldsymbol{\theta})$ se si vuole fare inferenza mediante la verosimiglianza rispetto ai parametri sia per i problemi di stima puntuale, che per la verifica di ipotesi e la costruzione di intervalli di confidenza di vario tipo.

In questa prima parte considereremo esclusivamente approcci di tipo parametrico.

DRAFT

12 Assunzioni di base nel modello lineare

Le ipotesi più semplici che possiamo fare nell'approccio parametrico sono:

DRAFT

a)	momento primo $\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}_n,$	per cui $\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}$ e quindi $\mathbf{X}\ \boldsymbol{\beta}$ è la componente sistematica ed $\boldsymbol{\varepsilon}$ è la componente accidentale additiva.
b)	momento secondo $\mathrm{V}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}_n$	La matrice di varianza e covarianza della componente accidentale è diagonale con elementi uguali, ossia b1) gli errori sono non correlati; b2) gli errori hanno la stessa varianza (ipotesi di omoscedasticità);
c)	distribuzione $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}; \sigma^2\mathbf{I}_n)$	Nel caso di normalità degli errori, le assunzioni a) e b) che specificano i primi due momenti multivariati, identificano in modo univoco la distribuzione della componente accidentale $\boldsymbol{\varepsilon}$.

DRAFT

Con queste ipotesi si vedrà che il metodo di stima della massima verosimiglianza conduce al metodo dei minimi quadrati.

Ancora una volta va notato che tali ipotesi sono equivalenti a quelle della distribuzione di una componente di una distribuzione normale multivariata condizionatamente alle altre componenti: pertanto se le osservazioni $\{\mathbf{y}, \mathbf{X}\}$ sono un campione di ampiezza n proveniente da una normale multivariata a $k+1$ componenti, possiamo direttamente utilizzare le relazioni viste per la normale multivariata. In questo caso la relazione stimata sarà la stima di una funzione di regressione esatta, ossia del valore atteso di una distribuzione condizionata (e non un'approssimazione alla vera funzione di regressione). In effetti, in generale, secondo l'impostazione fin qui seguita, non occorre neanche che le $\mathbf{X}$ siano delle determinazioni di variabili aleatorie, per potere impostare un modello di dipendenza lineare.

DRAFT

Altre implicazioni delle ipotesi di base:

- **Data l'assunzione di normalità, la non correlazione fra le componenti di $\boldsymbol{\varepsilon}$ implica anche l'indipendenza fra tali componenti.**

- In caso di validità della b1) e della b2) non solo si ha l'indipendenza, ma la forma della distribuzione di ciascuna $\mathbf{y}_i$ dipende solo dalla corrispondente componente accidentale ε_i .
- Sono quindi esclusi, con questa restrizione, i modelli con componente accidentale autoregressiva o comunque con una componente di dipendenza temporale o spaziali e/o territoriale. I modelli lineari possono essere adattati al trattamento di modelli di tipo autoregressivo.
- Le assunzioni (a),(b) e (c) implicano che le ε_i abbiano la stessa distribuzione, che quindi non dipende in alcun modo né dai particolari valori x_{ij} né dai valori dei parametri $\beta_{j.}$ e quindi nemmeno dalle y_i (o dalle $\mathrm{E}[\mathbf{Y}_i]$). Sono quindi escluse tutte le situazioni, che pure si presentano nelle applicazioni, in cui la variabilità dell'errore dipende in qualche misura da $\mathrm{E}[\mathbf{Y}_i]$.

DRAFT

- L' additività fra componente accidentale e sistematica implica che non vi sia collegamento fra l'assegnazione delle varie unità e gli errori accidentali.

DRAFT

13 La funzione di verosimiglianza nel modello lineare.

In un primo momento costruiamo la funzione di verosimiglianza del modello lineare rispetto ai parametri $\boldsymbol{\beta}$ ed alla varianza (o i parametri $\boldsymbol{\theta}$ da cui dipende la matrice di varianze e covarianze $\boldsymbol{\Sigma}(\boldsymbol{\theta})$).

È inutile per ora precisare se questa verosimiglianza ci servirà per costruire degli stimatori puntuali, o degli stimatori per intervallo, o per costruire dei test. In ogni caso per fare inferenza su $\boldsymbol{\beta}$, su σ^2 o su loro funzioni, l'analisi della verosimiglianza è essenziale, perché ci permette di costruire un criterio per la plausibilità di determinati valori dei parametri alla luce dell'evidenza campionaria.

DRAFT

Con le assunzioni a), b) e c) fatte prima siamo in grado di costruire la verosimiglianza campionaria, dal momento che abbiamo una osservazione $\mathbf{y}$ da una normale multivariata a n componenti di parametri $\mathbf{X}\boldsymbol{\beta}$ e $\sigma^2\mathbf{I}_n$, in modo che:

$$\mathrm{E}[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta} \qquad \text{e} \qquad \mathrm{V}[\mathbf{Y}] = \sigma^2\mathbf{I}_n;$$

quindi in definitiva:

$$\mathbf{Y} \sim \mathcal{N}_n(\mathbf{X}\boldsymbol{\beta}; \sigma^2\mathbf{I}_n)$$

per cui la verosimiglianza campionaria è data da:

Verosimiglianza del modello lineare

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = (2\pi)^{-\frac{n}{2}} |\mathrm{V}[\mathbf{Y}]|^{-1/2} \times$$
$$\times \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}} [\mathrm{V}[\mathbf{Y}]]^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right]$$
$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right]$$

Funzione di verosimiglianza campionaria per il modello lineare con le assunzioni di base (12)

DRAFT

Rispetto alla notazione precedente il vettore $\boldsymbol{\theta}$ di parametri della componente accidentale è composto dal solo σ^2, in quanto chiaramente la distribuzione di $\boldsymbol{\varepsilon}$ dipende solo da σ^2.

Otteniamo comunque lo stesso risultato senza bisogno di ricorrere alla normale multivariata, almeno se valgono le ipotesi semplificatrici, perchè in questo caso il campione $\mathbf{y}$ è costituito da n osservazioni indipendenti estratte da una distribuzione normale, ciascuna di valore atteso $\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}$.

Quindi otteniamo la verosimiglianza dal prodotto delle n densità:

$$L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = \prod_{i=1}^{n} f(y_i) \propto \frac{1}{\sigma^n} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta})^2\right\}$$

DRAFT

Il logaritmo della verosimiglianza campionaria per i $k+1$ parametri del modello, ossia le k componenti di $\boldsymbol{\beta}$ e σ^2 è quindi dato, trascurando la costante $-(\frac{n}{2})\log(2\pi)$, da:

Log Verosimiglianza di $\{\boldsymbol{\beta}, \sigma^2\}$ per un modello lineare con l'assunzione di normalità, indipendenza e omoscedasticità

$$\begin{aligned} \log L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) &= \\ &= -\frac{n}{2}\log(\sigma^2) - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \end{aligned} \tag{2}$$

Abbiamo definitivamente eliminato il termine costante $-(\frac{n}{2})\log(2\pi)$; d'altra parte la verosimiglianza rispetto a parametri $\boldsymbol{\theta}$ è una qualsiasi quantità proporzionale alla densità congiunta rispetto al campione.

DRAFT

oppure, esplicitando la forma quadratica:

$$\log L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2}\log(\sigma^2) - \frac{\Sigma_{i=1}^{n}\left(\mathbf{y}_i - \Sigma_{j=1}^{k} x_{ij}\beta_j\right)^2)}{2\sigma^2}$$

Con altre ipotesi su $\mathrm{E}[\mathbf{Y}]$ e su $\mathrm{V}[\mathbf{Y}]$ si giunge ovviamente a differenti funzioni di verosimiglianza (e quindi a differenti stimatori!).

DRAFT

13.1 La stima di massima verosimiglianza di σ^2

Ricaviamo prima la stima di massima verosimiglianza di σ^2, considerando noto $\boldsymbol{\beta}$.

Derivando la (2) rispetto a σ^2 otteniamo:

$$\frac{\partial \log L[\boldsymbol{\beta}, \sigma^2|\mathbf{y}])}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2(\sigma^2)^2}$$

Uguagliando a zero e risolvendo rispetto a $\hat{\sigma}^2(\boldsymbol{\beta})$, si ottiene facilmente il valore $\hat{\sigma}^2(\boldsymbol{\beta})$ che massimizza la verosimiglianza; Ho usato adesso la notazione $\hat{\sigma}^2(\boldsymbol{\beta})$ per esplicitare il fatto che tale stimatore è funzione del particolare valore assunto dal vettore di parametri $\boldsymbol{\beta}$.

DRAFT

Stimatore di Max. ver. $\hat{\sigma}^2(\boldsymbol{\beta})$ di σ^2, per un modello con errori indipendenti e omoscedastici, in funzione di $\boldsymbol{\beta}$

Per ciascun $\boldsymbol{\beta}$ si ha:

$$\hat{\sigma}^2(\boldsymbol{\beta}) = \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{n}$$

che si può anche scrivere:

$$\hat{\sigma}^2(\boldsymbol{\beta}) = \frac{\Sigma_{i=1}^{n}[\mathbf{y}_i - \Sigma_{j=1}^{k} x_{ij}\beta_j]^2}{n}$$

DRAFT

Dati osservati

Si vede dunque che con queste ipotesi la verosimiglianza campionaria dipende dalle osservazioni campionarie solo attraverso la somma dei quadrati degli scarti fra valori osservati e valori attesi:

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

E' quindi evidente che l'inferenza poggerà essenzialmente sullo studio della distribuzione di tali quantità sotto differenti ipotesi relative alle componenti di $\boldsymbol{\beta}$.

DRAFT

Si potrà anche avere il caso di osservazioni ancora distribuite normalmente ma con matrice di varianze e covarianze qualsiasi:

sotto queste ipotesi più generali la verosimiglianza sarà funzione dei dati ancora attraverso una forma quadratica, ma solo in alcuni casi particolari sarà possibile ottenere delle soluzioni esplicite per gli stimatori di massima verosimiglianza.

Non affronterò queste problematiche in queste pagine

DRAFT

Tornando al nostro caso semplificato, con errori non correlati e con varianze uguali, è stato immediato trovare $\hat{\sigma}^2(\boldsymbol{\beta})$, stimatore di massima verosimiglianza della varianza.

Si vedranno poi le caratteristiche di questo stimatore, distorsione, efficienza, etc., anche in funzione del fatto che $\boldsymbol{\beta}$ sia noto o sia da stimare.

13.2 Verosimiglianza profilo rispetto a β.

DRAFT

Sostituendo ora nella verosimiglianza campionaria tale valore di $\hat{\sigma}^2(\boldsymbol{\beta})$ al valore incognito del parametro di disturbo σ^2, otteniamo una quantità che è funzione solo del vettore $\boldsymbol{\beta}$ dei parametri di interesse (ossia la verosimiglianza profilo rispetto a $\boldsymbol{\beta}$)

$$L(\boldsymbol{\beta}, \hat{\sigma}^2(\boldsymbol{\beta}); \mathbf{y}) = \left(\frac{1}{\hat{\sigma}^2(\boldsymbol{\beta})}\right)^{\frac{n}{2}} \times \exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\hat{\sigma}^2(\boldsymbol{\beta})}\right\}$$

$$= \left[\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{n}\right]^{-\frac{n}{2}} \times \exp\left\{-\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{n}}\right\}$$

In definitiva si ha:

verosimiglianza profilo rispetto a $\boldsymbol{\beta}$

$$\begin{aligned} L(\boldsymbol{\beta}, \hat{\sigma}^2(\boldsymbol{\beta}); \mathbf{y}) &= \left[\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{n}\right]^{-\frac{n}{2}} \times \exp\left(-\frac{n}{2}\right) \\ &= \hat{\sigma}^2(\boldsymbol{\beta})^{-\frac{n}{2}} \times \exp\left(-\frac{n}{2}\right) \end{aligned} \tag{3}$$

E' evidente che questa espressione è massima rispetto a $\boldsymbol{\beta}$ quando:

$$[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}]^\mathsf{T}[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}] \text{ è un minimo rispetto a } \boldsymbol{\beta} \quad . \tag{4}$$

DRAFT

Analogamente per il logaritmo di tale verosimiglianza profilo si ha:

Log-verosimiglianza profilo rispetto a $\boldsymbol{\beta}$

$$\log L(\boldsymbol{\beta}, \hat{\sigma}^2(\boldsymbol{\beta}); \mathbf{y}) = \left(-\frac{n}{2}\right) - \left(\frac{n}{2}\right) \log \hat{\sigma}^2(\boldsymbol{\beta}) =$$

$$= \left(-\frac{n}{2}\right) - \left[\frac{n}{2}\right] \log \left(\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{n}\right)$$

DRAFT

La verosimiglianza profilo è uno strumento tecnico utile per fare inferenza nel caso generale di presenza di parametri di disturbo; nel nostro caso l'interesse preminente dell'inferenza è per i parametri $\boldsymbol{\beta}$: il parametro σ^2 è soltanto, di solito, un parametro di disturbo, nel senso che non è oggetto diretto dell'inferenza ma è incognito ed è comunque necessario stimarlo dai dati per fare inferenza sul parametro di interesse (multiplo) $\boldsymbol{\beta}$.

Ancora vediamo che la verosimiglianza profilo è funzione dei dati solo attraverso la forma quadratica già vista $\mathbf{R}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$: è evidente che la possibilità di ricavare la verosimiglianza profilo in modo così semplice rispetto a $\boldsymbol{\beta}$, è stata determinata dal fatto che, con le assunzioni semplificatrici fatte in questo caso, lo stimatore di massima verosimiglianza della varianza è in forma esplicita.

DRAFT

13.3 Verosimiglianza profilo e test basati sul rapporto delle verosimiglianze

Si può facilmente mettere in collegamento la verosimiglianza profilo e i test basati sul rapporto delle verosimiglianze, come si vedrà fra poco; se si richiama il metodo di costruzione del rapporto delle verosimiglianze si noterà come sia a numeratore sia a denominatore i parametri di disturbo vengono sostituiti dai valori che massimizzano la verosimiglianza ossia dai valori più plausibili alla luce dei dati osservati.

Questo ci fa vedere la corrispondenza con la verosimiglianza profilo normalizzata, per un generico parametro $\boldsymbol{\theta}$, con parametri di disturbo $\boldsymbol{\psi}$.

DRAFT

Indichiamo con $\hat{\boldsymbol{\psi}}(\boldsymbol{\theta})$ lo stimatore di massima verosimiglianza del parametro di disturbo $\boldsymbol{\psi}$ in funzione dei valori di $\boldsymbol{\theta}$, ossia quello che massimizza la verosimiglianza $L(\boldsymbol{\theta}, \boldsymbol{\psi})$ rispetto a $\boldsymbol{\psi}$. La verosimiglianza profilo per $\boldsymbol{\theta}$ è data da:

$$l_p(\boldsymbol{\theta}) = L\left(\boldsymbol{\theta}, \hat{\boldsymbol{\psi}}(\boldsymbol{\theta})\right) = \max_{\boldsymbol{\psi}} L(\boldsymbol{\theta}, \boldsymbol{\psi}) \tag{5}$$

La verosimiglianza profilo normalizzata è data da:

$$\bar{l}_p(\boldsymbol{\theta}) = \frac{l_p(\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta}} l_p(\boldsymbol{\theta})}$$

DRAFT

e quindi, applicando la 5, si ottiene:

$$\bar{l}_p(\boldsymbol{\theta}) = \frac{l_p(\boldsymbol{\theta})}{\max\limits_{\boldsymbol{\theta}} l_p(\boldsymbol{\theta})} = \frac{\max\limits_{\boldsymbol{\psi}} L(\boldsymbol{\theta}, \boldsymbol{\psi})}{\max\limits_{\boldsymbol{\psi}, \boldsymbol{\theta}} L(\boldsymbol{\theta}, \boldsymbol{\psi})}$$

che è il test basato sul rapporto delle verosimiglianze, per saggiare un'ipotesi su $\boldsymbol{\theta}$, essendovi dei parametri $\boldsymbol{\theta}$ incogniti, ma non coinvolti da H_0.

verosimiglianza profilo normalizzata

verosimiglianza profilo rispetto a $\boldsymbol{\theta}$ normalizzata = rapporto delle verosimiglianze per un test su $\boldsymbol{\theta}$

DRAFT

La figura 8 qui riportata chiarisce il significato e l'utilità dei vari tipi di verosimiglianza.

DRAFT

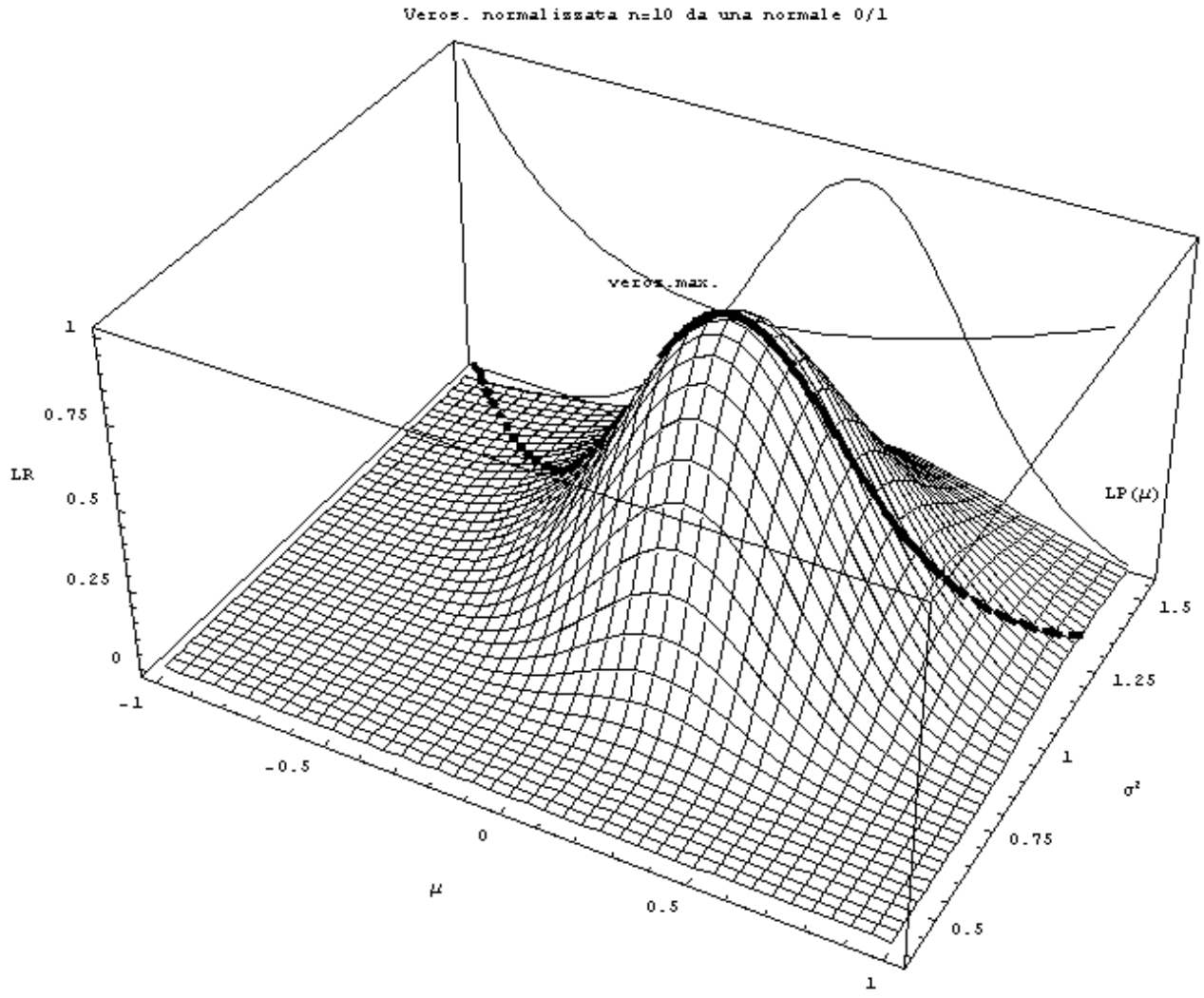

Figure 8: verosimiglianza rispetto a μ e σ^2 di un campione proveniente da una normale e verosimiglianza profilo rispetto a μ (proiettata sul piano verticale in fondo)

La superficie rappresenta la verosimiglianza normalizzata rispetto ai due parametri $\{\mu, \sigma^2\}$ per un campione estratto da una distribuzione normale standardizzata; tale verosimiglianza è rappresentata sull' asse z mentre sugli assi x e y sono rappresentati i due parametri di posizione μ e di varianza σ^2 di una distribuzione normale. Il punto di massimo è raggiunto ovviamente in corrispondenza della media campionaria e della varianza campionaria $P_{max} = \{M, s^2\}$.

La curva rappresentata nel piano xy, per comodità rappresentata sopra la superficie, rappresenta la relazione fra lo stimatore di massima verosimiglianza di σ^2 e il parametro di posizione, ossia $s^2(\mu) = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$.

La curva in grassetto rappresentata sulla superficie è data dai valori della verosimiglianza standardizzata in corrispondenza dello stimatore ottimale della varianza. Questa è la verosimiglianza profilo rispetto al parametro medio; la curva rappresentata sul piano xz (piano verticale in fondo) è la proiezione della verosimiglianza profilo che è funzione soltanto del parametro medio μ, ossia è $L(\mu, s^2(\mu))$.

E' da considerare che nel caso di un modello lineare generale non sarà possibile una tale rappresentazione grafica poiché abbiamo k parametri da stimare, ossia le componenti di $\boldsymbol{\beta}$; tuttavia la relazione che lega la varianza stimata ai parametri

DRAFT

della parte sistematica è sempre la stessa, ossia di tipo quadratico.

DRAFT

E' facile già da queste espressioni della verosimiglianza e in particolare della verosimiglianza profilo, costruire i rapporti di verosimiglianza per la verifica di particolari ipotesi sugli elementi di $\boldsymbol{\beta}$, in quanto la verosimiglianza profilo è funzione soltanto di $\hat{\sigma}^2(\boldsymbol{\beta})$ e quindi solo di $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Infatti vogliamo verificare ad esempio l'ipotesi

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \qquad \forall \sigma^2$$

contro l'alternativa generica:

$$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0 \qquad \forall \sigma^2$$

DRAFT

Costruiamo il test del rapporto delle verosimiglianze [2] rapportando la verosimiglianza massima sotto H_0 e quella massima sotto H_1:

- Sotto H_0 non vi sono parametri di disturbo da stimare (tranne σ^2 la cui influenza è stata eliminata in quanto stiamo lavorando con la verosimiglianza profilo su $\boldsymbol{\beta}$).
- Sotto H_1 oltre σ^2 va stimato il vettore $\boldsymbol{\beta}$

DRAFT

[2] Indicherò spesso questo test con la sigla inglese LR (Likelihood Ratio), più raramente con la sigla italiana TRV (test del rapporto delle verosimiglianze)

Indichiamo con $\hat{\boldsymbol{\beta}}$ la stima di massima verosimiglianza di $\boldsymbol{\beta}$ sotto H_1, ossia quello che porta al massimo non vincolato della verosimiglianza, dato che l'ipotesi alternativa H_1 ora esplicitata non impone alcun vincolo sui parametri.

Pertanto otteniamo la relazione:

$$LR(H_0, H_1) = \frac{\max L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}|H_0)}{\max L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}|H_1)} = \frac{L(\boldsymbol{\beta}_0, \hat{\sigma}^2(\boldsymbol{\beta}_0); \mathbf{y})}{L(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2(\hat{\boldsymbol{\beta}}); \mathbf{y})} =$$

$$= \left(\frac{\hat{\sigma}^2(\boldsymbol{\beta}_0)}{\hat{\sigma}^2(\hat{\boldsymbol{\beta}})}\right)^{-\frac{n}{2}} = \left(\frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^\mathsf{T}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0)}\right)^{\frac{n}{2}} = \left(\frac{R(\hat{\boldsymbol{\beta}})}{R(\boldsymbol{\beta}_0)}\right)^{\frac{n}{2}}$$

Come è noto valori alti di LR (vicini ad uno) indicheranno la plausibilità dell'ipotesi nulla; valori di LR bassi, vicini a zero, indicano invece che i dati non sono conformi con l'ipotesi nulla , e quindi che il valore di $\boldsymbol{\beta}_0$, sulla base dell'evidenza campionaria, è poco plausibile.

Ci preoccuperemo dopo della costruzione effettiva dei test e della loro distribuzione campionaria. **Non vi è dubbio comunque che qualsiasi test dovrà essere funzione di tale quantità** $\frac{R(\hat{\boldsymbol{\beta}})}{R(\boldsymbol{\beta}_0)}$

DRAFT

In generale comunque se vogliamo saggiare una generica ipotesi nulla H_0 contro una più generale H_1, essendo H_0 un caso particolare di H_1, possiamo pensare ciascuna ipotesi $H_i (i = 0, 1)$ come un sistema di vincoli $g_i(\beta)$ imposti sugli elementi di $\boldsymbol{\beta}$.

Ad esempio $g_0(\boldsymbol{\beta})$ potrebbe consistere del fatto che una superficie sia di primo grado, mentre $g_1(\boldsymbol{\beta})$ potrebbe essere l'alternativa che la superficie sia di secondo grado (ma non un polinomio di grado superiore).

Indicando ora con $\hat{\boldsymbol{\beta}}_i$ la stima di massima verosimiglianza di $\boldsymbol{\beta}$ sotto H_i, possiamo nel caso generale costruire il test:

$$
\begin{aligned}
LR &= \frac{\max L(\boldsymbol{\beta}, \sigma^2; \mathbf{y} | g_0(\boldsymbol{\beta}))}{\max L(\boldsymbol{\beta}, \sigma^2; \mathbf{y} | g_1(\boldsymbol{\beta}))} = \\
&= \frac{L(\hat{\boldsymbol{\beta}}_0, \hat{\sigma}^2(\hat{\boldsymbol{\beta}}_0); \mathbf{y})}{L(\hat{\boldsymbol{\beta}}_1, \hat{\sigma}^2(\hat{\boldsymbol{\beta}}_1); \mathbf{y})} = \left(\frac{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}_0)}{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}_1)} \right)^{-\frac{n}{2}} = \\
&= \left\{ \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_1)^\mathsf{T} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_1)}{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0)^\mathsf{T} (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_0)} \right\}^{\frac{n}{2}}
\end{aligned}
$$

DRAFT

Il criterio del rapporto delle verosimiglianze conduce ad un test sensibile e ad uno strumento generalmente molto utile per l'inferenza statistica sebbene non possegga almeno per piccoli campioni le proprietà ottimali che un test dovrebbe avere secondo la teoria di Neyman-Pearson.

Il problema della verifica di ipotesi, ossia della costruzione di un test di significatività, si può riassumere come segue: sulla base dei dati osservati la famiglia di distribuzioni dell'ipotesi alternativa H_1 si adatta significativamente meglio ai dati della famiglia parametrica rappresentata dall'ipotesi nulla H_0 ? Rifiutiamo H_0 a favore di H_1 se questo miglioramento è significativo.

Sebbene questo test non possegga tutte le proprietà ottimali richieste, risponde comunque ai requisiti fissati da Fisher per la verifica di ipotesi nell'indagine scientifica: lo scopo dei test è di attestare l'evidenza che i dati forniscono in merito a certe ipotesi più o meno definite; criteri di ottimalità quali potenza, ampiezza, non distorsione, sono importanti ma non sono necessariamente la cosa più importante nelle applicazioni.

DRAFT

DRAFT

rapporto delle verosimiglianze per i parametri di un modello lineare

Dalla costruzione del test del rapporto delle verosimiglianze per i parametri di un modello lineare con l'ipotesi di normalità, omoscedasticità, non correlazione, si vede che tale rapporto dipende esclusivamente dai rapporti fra le varianze stimate sotto le diverse ipotesi.

$$\mathrm{LR} = \left[\frac{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}_1)}{\hat{\sigma}^2(\hat{\boldsymbol{\beta}}_0)}\right]^{\frac{n}{2}} \qquad \hat{\boldsymbol{\beta}}_i = \hat{\boldsymbol{\beta}} | H_i \quad i = 0, 1$$

DRAFT

- la varianza a numeratore è quella relativa all'ipotesi più generale cioè quella che impone meno vincoli sui parametri (che sarà più piccola nell'ambito della famiglia parametrica considerata);
- la varianza a denominatore è quella calcolata sotto l'ipotesi di esistenza di qualche vincolo sui parametri.

È quindi evidente che l'ipotesi di normalità implica che le quantità sufficienti per fare inferenza sono le varianze stimate sotto i vincoli imposti dalle due differenti ipotesi.

14 Intervalli e regioni di confidenza

Potremo usare il risultato precedente per la costruzione di intervalli e regioni di confidenza ad un livello di probabilità fiduciaria $1-\alpha$ basati sull'inversione del test LR, ossia prendendo valori di $\boldsymbol{\beta}$ che non risultano esterni alla regione di rifiuto, ossia tali che $LR > k_\alpha$, secondo le tecniche che si vedranno più avanti.

DRAFT

Per trovare dunque il massimo incondizionato della verosimiglianza occorre trovare $\hat{\boldsymbol{\beta}}$ che da ora in poi indico per comodità di notazione con $\mathbf{b}$: abbiamo visto precedentemente (4) che il massimo della verosimiglianza si ottiene minimizzando rispetto a $\boldsymbol{\beta}$ la quantità $R(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$. Risolvere questo problema consiste dunque nel risolvere il problema dei Minimi Quadrati

Minimi Quadrati

DRAFT

Minimi quadrati Ordinari

Va trovato il minimo (rispetto a $\mathbf{b}$) di $(\mathbf{y}-\mathbf{Xb})^{\mathsf{T}}(\mathbf{y}-\mathbf{Xb})$ ossia il minimo della somma dei quadrati degli *scarti fra i valori osservati* y_i *e valori calcolati* $\mathbf{Xb}$. (indicati con $\hat{y}_i$)

$$\min_{\mathbf{b}} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Questo metodo di stima va sotto il nome di metodo dei **Minimi Quadrati**

Ordinari; l'aggettivo *Ordinari* è relativo al fatto che si prende una somma di scarti non ponderati, poichè abbiamo fatto le ipotesi semplificatrici; il termine inglese è **Ordinary Least Squares: OLS**. [3]

In forma matriciale:

$$\min_{\mathbf{b}} R(\mathbf{b}),$$

con:

$$\begin{aligned} R(\mathbf{b}) &= \sum_{i=1}^{n}(y_i - \sum_{j=1}^{k} x_{ij}\beta_j)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= (\mathbf{y}_{[n\times 1]} - \mathbf{X}_{[n\times k]}\mathbf{b}^{\mathsf{T}}_{[k\times 1]})(\mathbf{y}_{[n\times 1]} - \mathbf{X}_{[n\times k]}\mathbf{b}_{[k\times 1]}) = \\ &= \mathbf{y}^{\mathsf{T}}\mathbf{y} - 2\mathbf{b}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y} + \mathbf{b}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{b} \end{aligned}$$

essendo $\hat{y}_i$ l' i -esimo valore stimato.

DRAFT

15.1 Soluzione mediante derivate

Derivando $R(\mathbf{b})$ ($= \mathbf{y}^{\mathsf{T}}\mathbf{y} - 2\mathbf{b}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{y} + \mathbf{b}^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\mathbf{X})\mathbf{b}$) rispetto al vettore $\mathbf{b}$ si ottiene:

$$\frac{\partial R(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}^{\mathsf{T}}\mathbf{y} + 2(\mathbf{X}^{\mathsf{T}}\mathbf{X})\mathbf{b}$$

[3] L'acronimo OLS è abbastanza standard nella notazione internazionale

Uguagliandole a $\mathbf{0}$ (vettore nullo):

$$-2\mathbf{X}^\mathsf{T}\mathbf{y} + 2(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{b} = \mathbf{0};$$

Occorre risolvere, in $\mathbf{b}$, il **sistema di equazioni normali**:

Sistema di equazioni normali

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

DRAFT

Temporaneamente imponiamo la restrizione che $\mathbf{X}$ sia di rango k, e quindi esiste, e unica, l'inversa di $\mathbf{X}^\mathsf{T}\mathbf{X}$. Diversamente potremmo ricorrere ad una riparametrizzazione oppure all'uso dell'inversa generalizzata.

Soluzione generale dei minimi quadrati nei modelli lineari a rango pieno

(Sono stimatori di massima verosimiglianza con le ipotesi semplificatrici)

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

la soluzione esiste unica avendo supposto $\mathbf{X}$ di rango k, e fornisce certamente il minimo di $(\mathbf{y}-\mathbf{X}\mathbf{b})^\mathsf{T}(\mathbf{y}-\mathbf{X}\mathbf{b})$, ossia il minimo di $\Sigma_{i=1}^{n}(y_i-\hat{y}_i)^2$

DRAFT

Si tratta certamente di un minimo, in quanto le condizioni del secondo ordine, riguardanti l'Hessiano, sono sempre soddisfatte, infatti:

$$\frac{\partial R(\mathbf{b})}{\partial \mathbf{b}} = -2\mathbf{X}^\mathsf{T}\mathbf{y} + 2(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{b} \qquad \text{e}$$

$$\frac{\partial^2 R(\mathbf{b})}{\partial \mathbf{b}\partial \mathbf{b}^\mathsf{T}} = 2(\mathbf{X}^\mathsf{T}\mathbf{X})$$

che è sempre definita positiva (avendo assunto $\mathbf{X}$ di rango pieno k) e quindi il punto di stazionarietà (ossia quello che annulla le derivate prime) fornisce il

minimo assoluto della funzione.

15.2 Minimizzazione di $R(\beta)$ senza uso di derivate (modelli a rango non pieno)

Possiamo ricavare il valore $\mathbf{b}$ che minimizza $R(\boldsymbol{\beta})$ anche senza fare uso di derivate (*e addirittura senza neppure ipotizzare che* $\mathbf{X}$ *sia di rango pieno*).

Utilizziamo un procedimento simile a quello che si impiega per dimostrare la seconda proprietà della media aritmetica (ossia che la media aritmetica minimizza la somma dei quadrati degli scarti da un'origine qualsiasi, proprietà che ora generalizziamo); per fissare le idee ricordo che

$$\sum_{i=1}^{n}(y_i - M_y)^2 \leq \sum_{i=1}^{n}(y_i - a)^2 \quad \forall y_i,\ a \qquad \text{essendo:} \qquad M_y = \frac{\Sigma_{i=1}^{n}\, y_i}{n};$$

infatti:

$$\sum_{i=1}^{n}(y_i - a)^2 = \sum_{i=1}^{n}[(y_i - M_y) + (M_y - a)]^2 =$$

$$= \sum_{i=1}^{n}(y_i - M_y)^2 + n(M_y - a)^2 \geq \sum_{i=1}^{n}(y_i - M_y)^2$$

(il doppio prodotto $2(M_y - a)\,\Sigma_{i=1}^{n}(y_i - M_y)$ è ovviamente nullo per la prima proprietà della media aritmetica)

DRAFT

Passiamo ora al caso generale dei modelli lineari.

Sia $\mathbf{b}$ tale che: $\boxed{\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{b} = \mathbf{X}^\mathsf{T}\mathbf{y}}$

E' importante sottolineare che non è affatto necessario adesso assumere che $\mathbf{X}$ sia a rango pieno.

Trasformiamo ora la quantità da minimizzare (devianza) $R(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{\varepsilon}$.

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

(anche uguale a: $(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])^\mathsf{T}(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])$)=

(Aggiungendo e sottraendo $\mathbf{Xb}$)

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

sviluppando il prodotto in cui compare il binomio formato dai due termini: $(\mathbf{y} - \mathbf{Xb})$ e $(\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta})$

$$\begin{aligned} R(\boldsymbol{\beta}) \;=\; & [\mathbf{y} - \mathbf{X}\mathbf{b}]^\mathsf{T}[\mathbf{y} - \mathbf{X}\mathbf{b}] + \\ & [\mathbf{X}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}]^\mathsf{T}[\mathbf{X}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}] + \\ & 2[\mathbf{X}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}]^\mathsf{T}[\mathbf{y} - \mathbf{X}\mathbf{b}] \end{aligned}$$

DRAFT

Nell'ultimo termine in $[\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}]^\mathsf{T}$ si mette in evidenza $\mathbf{X}^\mathsf{T}$ ottenendo:

$$2[\mathbf{b} - \boldsymbol{\beta}]^\mathsf{T}\mathbf{X}^\mathsf{T}[\mathbf{y} - \mathbf{Xb}] = 2[\mathbf{b} - \boldsymbol{\beta}]^\mathsf{T}[\mathbf{X}^\mathsf{T}\mathbf{y} - \mathbf{X}^\mathsf{T}\mathbf{Xb}] = \mathbf{0}$$

l'ultimo termine è nullo per l'ipotesi fatta su $\mathbf{b}$.

In definitiva si ha (indicando $[\mathbf{y} - \mathbf{Xb}]^\mathsf{T}[\mathbf{y} - \mathbf{Xb}]$ con $R(\mathbf{b})$):

$$\begin{aligned} R(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \\ &= [\mathbf{y} - \mathbf{Xb}]^\mathsf{T}[\mathbf{y} - \mathbf{Xb}] + [\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}]^\mathsf{T}[\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}] = \\ &= R(\mathbf{b}) + [\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}]^\mathsf{T}[\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}] \geq R(\mathbf{b}) \end{aligned}$$

DRAFT

dal momento che $[\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}]^\mathsf{T}[\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}] \geq 0$.

Pertanto è dimostrato che $\mathbf{b}$ minimizza $R(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, dal momento che $[\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}]^\mathsf{T}[\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}]$ è sempre non negativa. [4]

Abbiamo ipotizzato all'inizio: $\mathbf{X}^\mathsf{T}\mathbf{Xb} = \mathbf{X}^\mathsf{T}\mathbf{y}$

[4] Si vedrà più avanti, a proposito della scomposizione della devianza, un' interpretazione molto importante della quantità $[\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}]^\mathsf{T}[\mathbf{Xb} - \mathbf{X}\boldsymbol{\beta}]$

SOLUZIONE GENERALE DEI MINIMI QUADRATI NEI MODELLI LINEARI (ANCHE A RANGO NON PIENO)

$\mathbf{b}$ deve essere tale che:

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

(la soluzione non è unica se $\mathbf{X}$ non è di rango pieno)
e dà il minimo di $(\mathbf{y} - \mathbf{X}\mathbf{b})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{b})$

DRAFT

Faccio notare un'altra proprietà interessante: da $\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{b} = \mathbf{X}^\mathsf{T}\mathbf{y}$ viene fuori che il minimo si ottiene quando:

SOLUZIONE GENERALE in funzione dei valori stimati

$$\mathbf{X}^\mathsf{T}\hat{\mathbf{y}} = \mathbf{X}^\mathsf{T}\mathbf{y}$$

A questo punto *se supponiamo* $\mathbf{X}$ *a rango pieno* possiamo esplicitare la soluzione (perchè esiste allora l'inversa di $\mathbf{X}^\mathsf{T}\mathbf{X}$):

SOLUZIONE GENERALE DEI MINIMI QUADRATI NEI MODELLI LINEARI A RANGO PIENO

(Sono stimatori di massima verosimiglianza con le ipotesi semplificatrici)

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

la soluzione esiste unica avendo supposto $\mathbf{X}$ di rango k
e dà il minimo di $(\mathbf{y} - \mathbf{X}\mathbf{b})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{b})$

DRAFT

Il metodo dei minimi quadrati ordinari (OLS: Ordinary Least Squares) COINCIDE con il metodo della massima verosimiglianza se e solo se:

- la distribuzione di ε è una normale a n variabili
- a componenti indipendenti
- e con uguale varianza σ^2

(altrimenti occorre impiegare metodi diversi da quello dei minimi quadrati)

Pertanto gli stimatori dei minimi quadrati godranno delle proprietà asintotiche ottimali degli stimatori M.V. soltanto nel caso gaussiano, diversamente saranno soltanto *i migliori stimatori lineari non distorti*.

DRAFT

15.3 Teorema di Gauss-Markov

Date le assunzioni a) e b), ossia errori a media nulla, non correlati ed a varianze uguali,[5] gli stimatori dei minimi quadrati hanno comunque una proprietà ottimale:

In un modello lineare, con le assunzioni ricordate sopra, omoschedasticità e non correlazione, gli stimatori dei minimi quadrati di un qualsiasi insieme di funzioni

[5] Ma non includendo la normalità!

lineari dei parametri β_j sono a varianza minima nella classe degli stimatori non distorti e lineari nelle $\mathbf{y}_i$. In effetti si può anche dimostrare che sono gli stimatori con la minima varianza generalizzata. chiarire

La dimostrazione non è complicatissima: consideriamo un generico stimatore $\mathbf{t}$ non distorto di $\boldsymbol{\beta}$, lineare nelle osservazioni, per il quale quindi:

$$\mathbf{t} = \mathbf{A}_{k\times n}\mathbf{y} \quad \mathrm{E}[\mathbf{t}] = \boldsymbol{\beta}$$

Sostituendo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, abbiamo:

$\mathbf{t} = \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}$ e data la correttezza di $\mathbf{t}$, deve essere:

$\mathrm{E}[\mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}] = \boldsymbol{\beta}$ e quindi svolgendo dentro il valore atteso e dato che si è assunto $\mathrm{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, deve aversi:

$\mathbf{A}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$ per cui: $\mathbf{A}\mathbf{X} = \mathbf{I}$ e $\mathbf{t} = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon}$

Se ora andiamo a calcolare la matrice di varianze e covarianze di $\mathbf{t}$ otteniamo:

$\mathrm{V}[\mathbf{t}] = \mathbf{A}\mathrm{V}[\boldsymbol{\varepsilon}]\mathbf{A}^{\mathsf{T}} = \sigma^2\mathbf{A}\mathbf{A}^{\mathsf{T}}$

Confrontando con la matrice di varianze e covarianze di $\mathbf{b}$ si ha:

$$\mathrm{V}[\mathbf{t}] - \mathrm{V}[\mathbf{b}] = \sigma^2\mathbf{A}\mathbf{A}^{\mathsf{T}} - \sigma^2\mathbf{X}^{\mathsf{T}}\mathbf{X}\sigma^2(\mathbf{A}\mathbf{A}^{\mathsf{T}} - \mathbf{X}^{\mathsf{T}}\mathbf{X})$$

DRAFT

e dato che $\mathbf{AX} = \mathbf{I}$, lo pre–moltiplichiamo e post-moltiplichiamo nel termine $\mathbf{X}^\mathsf{T}\mathbf{X}$:

$$\mathrm{V}[\mathbf{t}] - \mathrm{V}[\mathbf{b}] = \sigma^2(\mathbf{A}\mathbf{A}^\mathsf{T} - \mathbf{X}^\mathsf{T}\mathbf{X}) = \sigma^2(\mathbf{A}\mathbf{A}^\mathsf{T} - \mathbf{A}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{A}^\mathsf{T}) =$$

$$= \sigma^2\mathbf{A}(\mathbf{I} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})\mathbf{A}^\mathsf{T} =$$

(data l'idempotenza di $\mathbf{I} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$)

$$= \sigma^2\mathbf{A}(\mathbf{I} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})(\mathbf{I} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})\mathbf{A}^\mathsf{T} =$$

e infine, ponendo $\mathbf{B} = \mathbf{A}(\mathbf{I} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T})$, si ottiene:

$$\mathrm{V}[\mathbf{t}] - \mathrm{V}[\mathbf{b}] = \sigma^2\mathbf{B}\mathbf{B}^\mathsf{T}$$

e quindi la differenza fra le matrici di varianze e covarianze di $\mathbf{t}$ e di $\mathbf{b}$ (stimatore dei minimi quadrati) è sempre semi-definita positiva.

In effetti questo teorema non dimostra affatto la superiorità assoluta degli stimatori dei minimi quadrati, e potrebbe considerarsi una proprietà sufficiente per rendere inutile l'assunzione di normalità: infatti il teorema asserisce solo che sono i migliori fra gli stimatori non distorti *lineari nelle osservazioni*.

DRAFT

Intanto non è detto che la non distorsione sia una proprietà in assoluto necessaria, ma fondamentalmente nulla obbliga a restringersi agli stimatori lineari.

```
Assumere la linearità nelle osservazioni equivale
 ad assumere \chiarire la normalità.
```

Ad esempio nella derivazione della normale:

imponendo la condizione che dato un campione di n osservazioni indipendenti z_i, con scostamenti accidentali da una vera quantità $\mu = \mathrm{E}[Z]$, il miglior stimatore di $\mathrm{E}(\mathbf{X})$ sia la media aritmetica delle osservazioni, Gauss dimostrò che la distribuzione degli errori è normale.

DRAFT

15.4 Variabili a media zero (regressione in termini di scarti)

Se $\mathbf{X}$ è posta nella forma conveniente vista prima, 8.1 ossia prima colonna tutta uguale ad 1, e k colonne di scarti dei regressori dalle rispettive medie, $\mathbf{X}$ avrà un totale di $k+1$ colonne, supposte linearmente indipendenti (dal momento che il rango di $\mathbf{X}$ è in questo caso $k+1$).

Questa forma della matrice dei regressori viene utilizzata quando si vuole es-

plicitamente inserire un'ordinata all'origine β_0 fra i parametri del modello e per semplificare alcune scomposizioni successive.

Si vede facilmente che in questo caso:

$$\begin{aligned}
\mathbf{X}^\mathsf{T}\mathbf{X} &= \begin{pmatrix} n & \mathbf{0}_k^\mathsf{T} \\ \mathbf{0}_k & n\mathbf{S}_\mathbf{X} \end{pmatrix} \\
\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} &= \begin{pmatrix} 1/n & \mathbf{0}_k^\mathsf{T} \\ 0_k & S_\mathbf{X}^{-1}/n \end{pmatrix} \\
\mathbf{X}^\mathsf{T}\mathbf{y} &= \begin{pmatrix} n\,M_\mathbf{y} \\ n\,\mathrm{Cov}(\mathbf{X},\mathbf{y}) \end{pmatrix}
\end{aligned}$$

DRAFT

avendo indicato:

con $\mathbf{S}_\mathbf{X}$ la matrice delle varianze e covarianze dei k regressori e

con $\mathrm{cov}(\mathbf{X},\mathbf{y})$ il vettore delle covarianze fra la $\mathbf{y}$ e le X_j .

In questo modo è possibile separare la stima del termine noto da quella dei coefficienti di regressione:

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

$$= \begin{pmatrix} 1/n & \mathbf{0}_k^\mathsf{T} \\ 0_k & (S_\mathbf{X})^{-1}/n \end{pmatrix} \begin{pmatrix} nM_\mathbf{y} \\ n \quad \mathrm{Cov}(\mathbf{X}, \mathbf{y}) \end{pmatrix} =$$
$$= \begin{pmatrix} b_0 \\ \mathbf{b}_k \end{pmatrix} = \begin{pmatrix} M_\mathbf{y} \\ (S_\mathbf{X})^{-1} cov(\mathbf{X}, \mathbf{y}) \end{pmatrix}$$

Si noti l'analogia fra questa espressione $\mathbf{b}_k = (S_\mathbf{X})^{-1} cov(\mathbf{X}, \mathbf{y})$ e l'espressione del vettore dei coefficienti di regressione nelle distribuzioni condizionate della normale multivariata.

DRAFT

15.5 Distribuzione campionaria di b (minimi quadrati ordinari)

In ogni caso, qualunque sia la scelta della $\mathbf{X}$, comunque di rango k (e k colonne), lo stimatore $\mathbf{b}$ è dato in generale da:

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

Per ipotesi $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ e quindi $\mathbf{b}$ è una combinazione lineare delle $\mathbf{Y}$ per cui potremmo direttamente applicare le regole per il calcolo dei momenti di combinazioni lineari di variabili casuali.

Per la speranza matematica di $\mathbf{b}$ si ha:

$$\mathrm{E}[\mathbf{b}] = \mathrm{E}\left[\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Y}\right] =$$

$$= \mathrm{E}\left[\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}(\mathbf{X}\beta + \boldsymbol{\varepsilon})\right] =$$

$$= \mathrm{E}\left[\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\boldsymbol{\beta}\right] + \mathrm{E}\left[\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{\varepsilon}\right] =$$

$$= \mathrm{E}[\boldsymbol{\beta}] + \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathrm{E}[\boldsymbol{\varepsilon}] = \boldsymbol{\beta}$$

DRAFT

Momento primo di b

$$\mathrm{E}[\mathbf{b}] = \boldsymbol{\beta}$$

$\mathbf{b}$ è uno stimatore corretto di $\boldsymbol{\beta}$.

Per ottenere il risultato è stato sufficiente assumere soltanto: $\mathrm{E}[\varepsilon] = \mathbf{0}_n$.

Quindi perché b sia corretto per β è sufficiente che il modello lineare sia non distorto.

Per la matrice di varianze e covarianze campionarie di $\mathbf{b}$ si ha

$$\begin{aligned}
\mathrm{V}[\mathbf{b}] &= \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathrm{V}[\mathbf{Y}]\,\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} = \\
&= \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\sigma^2\mathbf{I}_n\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} = \\
&= \sigma^2\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} = \sigma^2\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}
\end{aligned}$$

DRAFT

Momento secondo di b

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

avendo assunto oltre a $\mathrm{E}[\boldsymbol{\varepsilon}] = 0_n$, anche:

$$\mathrm{V}[\boldsymbol{\varepsilon}] = \sigma^2 \mathbf{I}_n$$

(omoscedasticità e non correlazione)
qualunque sia la forma della distribuzione delle singole componenti ε_i
Quindi la struttura della matrice di varianze e covarianze di $\mathbf{b}$ *dipende dalla struttura della matrice* $\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$ e quindi dalla struttura delle matrici $(\mathbf{X}^\mathsf{T}\mathbf{X})$ e $\mathbf{X}$.

DRAFT

Se la matrice $\mathbf{X}$ è una matrice di scarti dalle medie aritmetiche (e le variabili indipendenti sono numeriche in senso stretto), allora $\mathbf{X}^\mathsf{T}\mathbf{X}$ è la matrice di devianze e codevianze dei k regressori; pertanto la struttura dei primi due momenti multi-

variati della distribuzione di $\mathbf{b}$ non dipende solo dalle assunzioni su $\boldsymbol{\varepsilon}$ ma anche dalla struttura della matrice $\mathbf{X}$.

Questo è uno degli aspetti di cui occorre tenere maggiormente conto tutte le volte che è possibile scegliere, in tutto o in parte, come costruire la matrice delle $\mathbf{x}$.

Se (e solo se) le $\mathbf{X}_j$ sono tutte non correlate i b_j saranno tutti non correlati ; se la matrice ($\mathbf{X}^\mathsf{T}\mathbf{X}$) risulta a blocchi (ossia gruppi di variabili internamente correlate ma non fra gruppi diversi), allora è a blocchi anche $V(\mathbf{b})$, ossia i corrispondenti gruppi di stimatori dei coefficienti saranno internamente correlati ma fra gruppi diversi vi sarà assenza di correlazione.

Si rivedranno in contesti particolari alcuni di questi aspetti.

DRAFT

Distribuzione di b

Se, inoltre, vale l'assunzione di normalità, allora:
b segue una distribuzione normale multivariata (in quanto combinazione lineare delle **y**);
b è lo stimatore di massima verosimiglianza (come peraltro abbiamo già ottenuto)

$$\mathbf{b} \sim \mathcal{N}_k(\boldsymbol{\beta}; \sigma^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1})$$

DRAFT

Si possono quindi costruire eventualmente delle regioni di confidenza per i parametri (se σ^2 è noto) che risulteranno in questo caso ellissoidali. Occorrerà distinguere il caso in cui σ^2 sia noto (poco plausibile) dal caso in cui venga stimato. In effetti anche senza assumere la normalità della componente accidentale, sotto condizioni non troppo restrittive sulla matrice $\mathbf{X}$ la distribuzione dello stimatore dei minimi quadrati tende alla normale al divergere di n.

Si rivedrà questa proprietà quando si parlerà dell'allontanamento dalle ipotesi di

normalità.

La plausibilità dell'assunzione di normalità (e delle alltre assunzioni) può essere discussa attraverso l'analisi dei residui

DRAFT

16 Test di significatività nei modelli lineari

Si è visto che i test fondati sui rapporti delle verosimiglianze sono funzioni delle devianze residue. In quseto capitolo ne studieremo la distribuzione, sfruttando le proprietà della distribuzione normale multivariata e alcuni teoremi relativi alla distribuzione di forme quadratiche in variabili normali multiple; il lettore può eventualmente tralasciare le dimostrazioni e andare direttamente ai risultati: ritengo però che alcune delle dimostrazioni di questo capitolo siano di rara essenzialità ed eleganza formale.

DRAFT

17 Distribuzione della devianza residua nei modelli lineari

Abbiamo visto, nella parte relativa alla costruzione della verosimiglianza del modello lineare, che con le ipotesi semplificatrici la verosimiglianza profilo rispetto a $\boldsymbol{\beta}$ e i test basati sui rapporti di verosimiglianza dipendono da $R(\mathbf{b})$ e da $R(\boldsymbol{\beta})$.

Evidentemente occorrerà studiare la distribuzione campionaria di queste quantità (sotto un'ipotesi nulla e nel caso generale).

17.1 Devianza residua in funzione dei valori osservati

Ricordiamo ancora che $\mathbf{b} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$ è lo stimatore di massima verosimiglianza di $\boldsymbol{\beta}$ in un modello lineare (di rango pieno), supponendo la validità delle ipotesi semplificatrici sulla componente accidentale:

$$\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0; \sigma^2\mathbf{I})$$

Trasformiamo la devianza residua $R(\mathbf{b})$, ossia la somma dei quadrati degli scarti fra valori della variabile di risposta osservati e stimati (che è la quantità minimizzata mediante il metodo dei minimi quadrati); l'importanza di tale quantità (e della sua distribuzione campionaria!) è evidente alla luce di quanto abbiamo visto sui test basati sui rapporti di verosimiglianze.

Il vettore $\mathbf{y} - \mathbf{X}\mathbf{b}$ è detto vettore dei residui (empirici). $R(\mathbf{b})$ è quindi la devianza dei residui empirici

Esprimiamo la devianza residua in funzione delle osservazioni:

DRAFT

$$
\begin{aligned}
R(\mathbf{b}) &= \Sigma_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2 = \\
&= \Sigma_{i=1}^{n}(\mathbf{y}_i - \Sigma_{j=1}^{k} x_{ij} b_j)^2 = \\
&= (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}) = \\
&\qquad \text{(sostituendo a } \mathbf{b} \text{ il valore } \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}\text{)} \\
&= (\mathbf{y} - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}) = \\
&= ([\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\mathbf{y})^{\mathsf{T}}([\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\mathbf{y}) = \\
&\qquad \text{mettendo in evidenza } \mathbf{y} \\
&= \mathbf{y}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\mathbf{y} = \\
&= \mathbf{y}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\mathbf{y}
\end{aligned}
$$

DRAFT

Questo risultato finale sarà di importanza fondamentale nell'inferenza sui modelli lineari

Devianza residua nel modello lineare

$$R(\mathbf{b}) = \mathbf{y}^\mathsf{T}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}]\mathbf{y}$$

$R(\mathbf{b})$ è una forma quadratica nelle $\mathbf{y}$

DRAFT

$[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}]$ è simmetrica ed idempotente di rango $n-k$: infatti una qualsiasi matrice $\mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}$ è idempotente di rango k, e quindi lo è anche $[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}]$. [6]

[6] si riveda eventualmente la parte relativa al calcolo matriciale link con matrici

Inoltre:

- $$[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}]\mathbf{X} = \mathbf{0}_{n\times k}$$

  e quindi i residui empirici risultano non correlati con le $\mathbf{X}$ ossia:

  $$Cov(\mathbf{y} - \mathbf{X}\mathbf{b}, \mathbf{X}) = [\mathbf{y} - \mathbf{X}\mathbf{b}]^\mathsf{T}\mathbf{X} = \mathbf{0}$$

  che si ricava direttamente dalle equazioni normali.

- Si ha anche che $\mathbf{y} - \mathbf{X}\mathbf{b}$ ha media nulla.

DRAFT

La matrice $[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}]$

La matrice $[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}]$ fornisce la riduzione nella devianza di $\mathbf{y}$ dovuta alla regressione su $\mathbf{X}$

17.2 Devianza residua in funzione della componente accidentale ε:

Esprimiamo ora $R(\mathbf{b})$ in funzione della componente accidentale $\boldsymbol{\varepsilon}$:

Dall'espressione precedente:

$$R(\mathbf{b}) = \mathbf{y}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\mathbf{y} =$$

(operando sul terzo fattore, esprimendo $\mathbf{y}$ come $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, secondo quanto ipotizzato)

$$= \mathbf{y}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}](\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) =$$

aprendo la parentesi a destra

$$= \mathbf{y}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\mathbf{X}\boldsymbol{\beta} + \mathbf{y}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\boldsymbol{\varepsilon} =$$

e dato che $[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\mathbf{X} = \mathbf{0}_{n\times k}$, ed effettuando le stesse operazioni sul termine $\mathbf{y}^{\mathsf{T}}$, si ha:

$$= \mathbf{y}^{\mathsf{T}}\mathbf{0}_{n\times k}\boldsymbol{\beta} + (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\boldsymbol{\varepsilon} =$$

aprendo la parentesi a sinistra

$$= 0 + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\boldsymbol{\varepsilon} =$$

$$= 0 + 0 + \boldsymbol{\varepsilon}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\boldsymbol{\varepsilon}.$$

DRAFT

In definitiva si ha l'ulteriore espressione per la devianza residua:

Devianza residua in funzione della componente accidentale

$$R(\mathbf{b}) = \boldsymbol{\varepsilon}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\boldsymbol{\varepsilon}$$

La devianza residua $R(\mathbf{b})$ è quindi una forma quadratica nelle $\boldsymbol{\varepsilon}$ ancora con la stessa matrice di coefficienti $[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]$

DRAFT

Calcoliamo il valore atteso di $R(\mathbf{b})$.

Posto $\mathbf{A} = [\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]$, e indicato con a_{ij} l'elemento generico della matrice $\mathbf{A}$ (simmetrica!), si ha per la forma quadratica $\boldsymbol{\varepsilon}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\boldsymbol{\varepsilon}$ l'ovvia espressione:

$$R(\mathbf{b}) = \boldsymbol{\varepsilon}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}^{\mathsf{T}}\mathbf{A}\boldsymbol{\varepsilon} = \sum_{i=1}^{n} a_{ii}\varepsilon_i^2 + \sum_{i \neq j} a_{ij}\varepsilon_i\varepsilon_j$$

Quindi si può vedere facilmente che, essendo $\mathrm{E}\left[\varepsilon_i \varepsilon_j\right] = 0 \quad (i \neq j)$, sviluppando i termini della forma quadratica $R(\mathbf{b})$ si ha:

$$\begin{aligned} \mathrm{E}\left[R(\mathbf{b})\right] &= \sum_{i=1}^{n} a_{ii} \mathrm{E}\left[\varepsilon_i^2\right] + \sum_{i \neq j} a_{ij} \mathrm{E}\left[\varepsilon_i \varepsilon_j\right] = \sum_{i=1}^{n} a_{ii} \sigma^2 + 0 \\ &= \sigma^2 \, tr[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}], \end{aligned}$$

perchè contribuiscono alla speranza matematica solo i termini in ε_i^2,per i quali $\mathrm{E}\left[\varepsilon_i^2\right]$=$\sigma^2$, mentre gli altri termini, in $\varepsilon_i, \varepsilon_j \, (i \neq j)$ hanno valore atteso nullo.

Ricordando ora che la traccia di una matrice idempotente è uguale al suo rango, $n - k$ nel caso di $[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}]$, si ha infine:

DRAFT

Valore atteso della devianza residua

$$\mathrm{E}\left[R(\mathbf{b})\right] = (n-k)\sigma^2$$

avendo ipotizzato soltanto:

$$\mathrm{E}\left[\boldsymbol{\varepsilon}\right] = \mathbf{0} \qquad \text{e} \qquad \mathrm{V}\left[\boldsymbol{\varepsilon}\right] = \sigma^2\mathbf{I}_n$$

(anche senza l'assunzione di normalità); quindi:

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

è *sempre* una stima corretta della varianza.

DRAFT

Inoltre, se vale l' assunzione di normalità, vediamo che $R(\mathbf{b}) = \boldsymbol{\varepsilon}^\mathsf{T}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}]\boldsymbol{\varepsilon}$, $R(\mathbf{b})$ è una forma quadratica in variabili normali ($\boldsymbol{\varepsilon}$) indipendenti, a valore atteso nullo e varianze uguali: possiamo quindi applicare la proprietà link con forme quadrat

dal momento che la matrice dei coefficienti $[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}]$ è idempotente di rango $n - k$, per concludere che $R(\mathbf{b})$ ha una distribuzione proporzionale a quella di una chi-quadro.

Distribuzione della devianza residua, sotto ipotesi di normalità, non correlazione e omoscedasticità

Sotto l'ipotesi che $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(0; \sigma^2\mathbf{I})$:

$$R(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{b}) = (n - k)\ s^2 \sim \sigma^2\chi^2_{n-k}$$

DRAFT

17.2.1 intervalli di confidenza per la varianza

Ovviamente si può usare il risultato precedente per costruire intervalli di confidenza attorno a s^2 per σ^2 basati sui percentili della distribuzione χ^2_{n-k}; abbiamo visto che:

$$(n-k)\ s^2 \sim \sigma^2 \chi^2_{n-k} \quad \text{quindi:} \quad \frac{s^2}{\sigma^2} \sim \frac{\chi^2_{n-k}}{(n-k)}$$

Allora invertendo le relazioni precedenti possiamo ottenere l'intervallo di confidenza per σ^2 con livello di probabilità fiduciaria $1-\alpha$:

$$I_{1-\alpha}(\sigma^2) = \left\{ \frac{(n-k)\ s^2}{\chi^2_{n-k,\frac{\alpha}{2}}};\ \frac{(n-k)\ s^2}{\chi^2_{n-k,1-\frac{\alpha}{2}}} \right\}$$

avendo al solito indicato con $\chi^2_{\nu,\gamma}$ opportuni percentili della distribuzione χ^2 con ν gradi di libertà, ossia tali che: $P\left\{\chi^2_\nu > \chi^2_{n-k,\gamma}\right\} = \gamma$ [7]

DRAFT

[7] Sto seguendo qui la notazione statistica classica, che indicizza i punti critici col valore γ di probabilità della coda *destra*! γ è qundi il complemento ad uno della funzione di ripartizione.

18 Scomposizione della devianza nel modello lineare e verifica di ipotesi.

Effettuiamo alcune scomposizioni delle diverse somme di quadrati (e forme quadratiche in generale) che abbiamo incontrato (fra cui ad esempio: $R(\mathbf{b})$, $R(\boldsymbol{\beta})$, $\mathbf{y}^\mathsf{T}\mathbf{y}$).

18.0.1 La scomposizione della somma dei quadrati $\mathbf{y}^\mathsf{T}\mathbf{y}$

Operiamo sulla devianza di $\mathbf{y}$, (o più precisamente sulla somma dei quadrati $\mathbf{y}^\mathsf{T}\mathbf{y}$) partendo ancora da una delle relazioni trovate per $R(\mathbf{b})$:

$$R(\mathbf{b}) = \sum_{i=1}^{n}(\mathbf{y}_i - \hat{\mathbf{y}}_i)^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{b}) =$$

$$= \mathbf{y}^\mathsf{T}(\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T})\mathbf{y} =$$

aprendo la parentesi

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

sostituendo $\mathbf{b}$ all' espressione $\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$

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

DRAFT

ricordando che, trasponendo il sistema di equazioni normali si ha $\mathbf{y}^\mathsf{T}\mathbf{X} = \mathbf{b}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}$

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

Infine, esplicitando rispetto a $\mathbf{y}^\mathsf{T}\mathbf{y}$), otteniamo l'importante risultato:

Scomposizione Della Devianza Empirica

$$\mathbf{y}^\mathsf{T}\mathbf{y} = R(\mathbf{b}) + \mathbf{b}^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{b}.$$

DRAFT

Tavola Di Scomposizione Della Devianza Empirica	
Forma quadratica	**Fonte di variabilità**
$\mathbf{y}^\intercal\mathbf{y} =$	Somma dei quadrati di $\mathbf{y}$ (devianze se $\mathbf{y}$ è a media nulla)
$(\mathbf{y}-\mathbf{Xb})^\intercal(\mathbf{y}-\mathbf{Xb})+$	devianza residua
$\mathbf{b}^\intercal\mathbf{X}^\intercal\mathbf{Xb}$ ossia $\hat{\mathbf{y}}^\intercal\hat{\mathbf{y}}$	Somma dei quadrati spiegata dal modello lineare

Table 3: Tavola Di Scomposizione Della Devianza Empirica (Somme Dei Quadrati)

DRAFT

18.1 Scomposizione di $R(\beta)$

Per potere ottenere le distribuzioni dei test basati sul rapporto delle verosimiglianze, occorre operare alcune trasformazioni. Operiamo sulla devianza teorica $R(\boldsymbol{\beta}) = \boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{\varepsilon}$

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

(anche uguale a: $(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])^{\mathsf{T}}(\mathbf{Y} - \mathrm{E}[\mathbf{Y}])$)=

(Aggiungendo e sottraendo $\mathbf{Xb}$)

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

$$(= \sum_{i=1}^{n} [(y_i - \hat{y}_i) + (\hat{y}_i - \mathrm{E}[y_i])]^2 =)$$

DRAFT

sviluppiamo quest'ultimo prodotto (che è sempre una somma di quadrati) in cui compare il binomio formato dai due termini: $(\mathbf{y} - Xb)$ e $(Xb - \mathbf{X}\boldsymbol{\beta})$

$$
\begin{aligned}
&= (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{b}) && =_{R(\mathbf{b})} \\
&+ \\
&(\mathbf{X}\mathbf{b} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{X}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}) \\
& && \text{si mette in evidenza } \mathbf{X} \text{ sia a sinistra che a destra e si ottiene } (\mathbf{b} - \boldsymbol{\beta})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta}) \\
&+ \\
&2(\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathsf{T}}(\mathbf{X}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}) \\
& && = 0 \text{ perché } (\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathsf{T}}\mathbf{X} = \mathbf{0} \text{ dalle equazioni dei minimi quadrati} \\
&= \\
&R(\mathbf{b}) + (\mathbf{b} - \boldsymbol{\beta})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})
\end{aligned}
$$

DRAFT

Si può interpretare tale scomposizione in modo leggermente diverso, ponendo l'enfasi non su $\mathbf{b}$, stimatore di $\boldsymbol{\beta}$, bensì su $\mathbf{X}\mathbf{b}$, *stimatore lineare ottimale del valore atteso* $\mathrm{E}[\mathbf{Y}]$. Pertanto:

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

$$
= [\mathbf{X}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}]^{\mathsf{T}}[\mathbf{X}\mathbf{b} - \mathbf{X}\boldsymbol{\beta}] = [\hat{\mathbf{y}} - \mathrm{E}[\mathbf{Y}]]^{\mathsf{T}}[\hat{\mathbf{y}} - \mathrm{E}[\mathbf{Y}]]
$$

In definitiva quindi si ha:

Relazione fra devianza in funzione di $\boldsymbol{\beta}$ e devianza residua

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

Oppure :

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

DRAFT

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

$\boldsymbol{\varepsilon}^\mathsf{T}\boldsymbol{\varepsilon} =$ $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\mathsf{T}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) =$	$(\mathbf{y} - \mathbf{X}\mathbf{b})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{b}) +$	$(\mathbf{b} - \boldsymbol{\beta})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta})$
devianza teorica complessiva di $\boldsymbol{\varepsilon}$ (rispetto al modello vero)	devianza residua	devianza delle stime

DRAFT

Si può anche scrivere:

$$\underbrace{\sum_{i=1}^{n}(y_i - \mathrm{E}[y_i])^2}_{\textit{devianza teorica complessiva}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\textit{devianza residua}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \mathrm{E}[y_i])^2}_{\textit{devianza delle stime rispetto ai valori att}}$$

Questa scomposizione è basilare anche perché possiamo vedere che il rapporto delle verosimiglianze costruito in precedenza link con sectionLR per saggiare l'ipotesi nulla $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, contro l'alternativa generica $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, è funzione

di queste quantità.

Infatti:

$$
\begin{aligned}
LR &= \frac{\max[L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})|H_0]}{\max[L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})|H_1]} = \\
&= \left\{ \frac{[\mathbf{y} - \mathbf{X}\mathbf{b}]^\mathsf{T}[\mathbf{y} - \mathbf{X}\mathbf{b}]}{[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0]^\mathsf{T}[\mathbf{y} - \mathbf{X}\boldsymbol{\beta}_0]} \right\}^{\frac{n}{2}} = \left\{ \frac{R(\mathbf{b})}{R(\boldsymbol{\beta}_0)} \right\}^{\frac{n}{2}} = \qquad (6) \\
&= \left\{ \frac{R(\mathbf{b})}{R(\mathbf{b}) + (\mathbf{b} - \boldsymbol{\beta}_0)^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b} - \boldsymbol{\beta}_0)} \right\}^{\frac{n}{2}} \qquad (7)
\end{aligned}
$$

avendo ora indicato con $\mathbf{b}$ lo stimatore di massima verosimiglianza prima indicato con $\hat{\boldsymbol{\beta}}$.

DRAFT

18.2 Test F per la verifica di ipotesi nel modello lineare: distribuzione nulla

Per esaminare la distribuzione nulla del rapporto delle verosimiglianze[8], o di una sua trasformazione monotona, riprendiamo in esame la scomposizione di base di

[8] Per il lettore disinteressato alla teoria completa direi di saltare la dimostrazione e andare al risultato finale 9

$R(\boldsymbol{\beta})$, e dividiamo tutti i termini per σ^2 :

$$\frac{R(\boldsymbol{\beta})}{\sigma^2} = \frac{R(\mathbf{b})}{\sigma^2} + \frac{(\mathbf{b}-\boldsymbol{\beta})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})}{\sigma^2} \tag{8}$$

con le ipotesi che abbiamo fatto, compresa ovviamente quella di normalità, possiamo vedere che i tre termini si distribuiscono come delle χ^2 , per cui si può applicare il teorema di Cochran; infatti:

- $\frac{R(\boldsymbol{\beta})}{\sigma^2}$ si distribuisce come una χ^2 con n gradi di libertà in quanto somma dei quadrati di n v.c. normali standardizzate:

$$\frac{R(\boldsymbol{\beta})}{\sigma^2} = \frac{\boldsymbol{\varepsilon}^{\mathsf{T}}\boldsymbol{\varepsilon}}{\sigma^2} = \frac{\Sigma_{i=1}^{n}\varepsilon_i^2}{\sigma^2} \sim \chi^2_n$$

- $\frac{R(\mathbf{b})}{\sigma^2}$ si distribuisce come una χ^2 con $n-k$ gradi di libertà in quanto si è già visto link con ref che:

$$R(\mathbf{b}) = (\mathbf{y}-\mathbf{X}\mathbf{b})^{\mathsf{T}}(\mathbf{y}-\mathbf{X}\mathbf{b}) = \boldsymbol{\varepsilon}^{\mathsf{T}}[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]\boldsymbol{\varepsilon}$$

si distribuisce come $\sigma^2\chi^2_{n-k}$, essendo $[\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathsf{T}}]$ idempotente di rango $n-k$.

DRAFT

- $\frac{[\mathbf{b}-\boldsymbol{\beta}]^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}[\mathbf{b}-\boldsymbol{\beta}]}{\sigma^2}$ si distribuisce come una χ^2 con k gradi di libertà in quanto è il numeratore dell'esponente della densità di una normale multivariata:

$$\mathbf{b} \sim N(\boldsymbol{\beta}; \sigma^2 \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1})$$

(infatti poichè $\mathrm{E}[\mathbf{b}] = \boldsymbol{\beta}$ e $\mathrm{V}[\mathbf{b}] = \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\sigma^2$, allora $(\mathrm{V}[\mathbf{b}])^{-1} = \frac{\mathbf{X}^{\mathsf{T}}\mathbf{X}}{\sigma^2}$, e quindi si ricava per la forma quadratica ad esponente nella densità della distribuzione di $\mathbf{b}$:

$$[\mathbf{b} - \mathrm{E}[\mathbf{b}]]^{\mathsf{T}}\mathrm{V}[\mathbf{b}][\mathbf{b} - \mathrm{E}[\mathbf{b}]] = \frac{[\mathbf{b} - \boldsymbol{\beta}]^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}[\mathbf{b} - \boldsymbol{\beta}]}{\sigma^2}$$

Quindi si può applicare il teorema di Cochran ed i termini (B) e (C) risultano indipendenti!

In definitiva la quantità:

$$F = \frac{\frac{[\mathbf{b}-\boldsymbol{\beta}]^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}[\mathbf{b}-\boldsymbol{\beta}]}{k}}{\frac{[\mathbf{y}-\mathbf{X}\mathbf{b}]^{\mathsf{T}}[\mathbf{y}-\mathbf{X}\mathbf{b}]}{n-k}} = \frac{[\mathbf{b} - \boldsymbol{\beta}]^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}[\mathbf{b} - \boldsymbol{\beta}]}{ks^2},$$

la cui distribuzione non dipende da σ^2, è il rapporto fra due variabili casuali χ^2 indipendenti divise per i rispettivi gradi di libertà, e quindi si distribuisce secondo

DRAFT

una F di Snedecor con k ed $n-k$ gradi di libertà, essendo $\boldsymbol{\beta}$ il vero valore del vettore dei parametri.

Test per la verifica di un ipotesi nel modello lineare

Pertanto, per saggiare l'ipotesi nulla:
$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, contro l'alternativa generica

$$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0,$$

possiamo impiegare la quantità:

$$F = \frac{\frac{[\mathbf{b}-\boldsymbol{\beta}_0]^\top \mathbf{X}^\top \mathbf{X}[\mathbf{b}-\boldsymbol{\beta}_0]}{k}}{\frac{[\mathbf{y}-\mathbf{X}\mathbf{b}]^\top[\mathbf{y}-\mathbf{X}\mathbf{b}]}{n-k}} \tag{9}$$

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

DRAFT

La regione di rifiuto sarà costituita dai valori elevati di F, superiori ad $F_{\alpha,k,n-k}$. (ossia situati sulla coda destra della corrispondente variabile F di Snedecor)

Infatti valori osservati di F elevati danno evidenza contraria ad H_0.

La statistica F è funzione monotona decrescente del rapporto delle verosimiglianze LR costruito in precedenza 7. Infatti:

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

$$F = \left(\frac{R(\boldsymbol{\beta}_0)}{R(\mathbf{b})} - 1\right)\frac{n-k}{k} = \left(\frac{1}{LR^{\frac{2}{n}}} - 1\right)\frac{n-k}{k}$$

DRAFT

$\mathbf{b}$ e s^2 costituiscono un set di stimatori congiuntamente sufficienti per $\boldsymbol{\beta}$ e σ^2.

Infatti partendo dalla verosimiglianza del modello lineare, introdotta prima, con le assunzioni fatte, e con le scomposizioni ora viste si può giungere ad una fattorizzazione:

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp[-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} = \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp[-\frac{R(\boldsymbol{\beta})}{2\sigma^2} = \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp[-\frac{R(\mathbf{b})}{2\sigma^2} - \frac{[\mathbf{b} - \boldsymbol{\beta}]^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}[\mathbf{b} - \boldsymbol{\beta}]}{2\sigma^2}] = \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp[-\frac{(n-k)s^2}{2\sigma^2}] \exp[-\frac{[\mathbf{b} - \boldsymbol{\beta}]^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}[\mathbf{b} - \boldsymbol{\beta}]}{2\sigma^2}]. \end{aligned}$$

Quindi la verosimiglianza campionaria rispetto a $\boldsymbol{\beta}$ e σ^2 dipende dalle osservazioni solo attraverso le statistiche $\mathbf{b}$ e s^2 .

DRAFT

18.2.2 Matrice di informazione

Dalla verosimiglianza è anche immediato vedere che l'informazione di Fisher su $\boldsymbol{\beta}$ è ancora funzione della matrice $\mathbf{X}$.

Infatti:

$$I(\boldsymbol{\beta}) = \mathrm{E}\left[\left\{\frac{\partial^2 \log L(\boldsymbol{\beta}, \sigma^2; \mathbf{y})}{\partial\beta\partial\beta^\mathsf{T}}\right\}\right] = -\frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{\sigma^2}$$

$$\mathrm{E}\left[\left\{\frac{\partial(\log L(\boldsymbol{\beta}, \sigma^2; \mathbf{y}))^2}{\partial\beta}\right\}\right] = \frac{\mathbf{X}^\mathsf{T}\mathbf{X}}{\sigma^2}$$

(La matrice delle derivate seconde comunque è costante)

$$\mathbf{V}_{\mathrm{inf}}(\mathbf{b}) = -\left(\mathbf{I}(\boldsymbol{\beta})\right)^{-1} = \sigma^2\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$$

Per cui il valore asintotico della matrice di varianze e covarianze di $\mathbf{b}$ coincide con il valore già trovato per via diretta per n qualsiasi.

DRAFT

Va sottolineato che nella scomposizione vista prima (8) la quantità(B) ossia:

$$\frac{R(\mathbf{b})}{\sigma^2} = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathsf{T}}(\mathbf{y} - \mathbf{X}\mathbf{b})}{\sigma^2}$$

si distribuisce sempre come una v.a. χ^2 con $n - k$ gradi di libertà, sia sotto H_0 che sotto H_1, avendo fatta ovviamente l'assunzione di normalità.

quindi la stima della varianza:

$$s^2 = \frac{R(\mathbf{b})}{n-k} = \frac{\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k}$$

ha sempre una distribuzione proporzionale a quella di una χ^2 con $n - k$ gradi di libertà.

DRAFT

Distribuzione dello stimatore corretto di σ^2

$$s^2 \sim \frac{\sigma^2 \chi^2_{n-k}}{n-k}$$

qualunque sia l'ipotesi vera

Infatti $R(\mathbf{b})$ dipende solo dai valori osservati e non dipende dai particolari valori delle componenti del vettore dei parametri $\boldsymbol{\beta}$.

Si noti inoltre che la distribuzione di s^2 non dipende dalla particolare configurazione (scelta a priori o osservata) della matrice $\mathbf{X}$, se non attraverso le sue dimensioni, n e k .

Diversamente la forma quadratica definita dalla quantità (C) ossia: controllare (C)

$$\frac{(\mathbf{b}-\boldsymbol{\beta})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})}{\sigma^2}$$

si distribuisce come una χ^2 con k gradi di libertà *solo se* $\boldsymbol{\beta}$ *è il vero valore del parametro.*

DRAFT

Pertanto la forma quadratica a numeratore del test F divisa per i gradi di libertà k

$$s_1^2 = \frac{(\mathbf{b}-\boldsymbol{\beta}_0)^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}_0)}{k}$$

è uno stimatore corretto di σ^2 solo sotto H_0 perché:

$$(\mathbf{b}-\boldsymbol{\beta}_0)^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}_0)$$

si distribuisce come $\sigma^2\chi^2_k$ soltanto se è vera H_0

Infatti la distribuzione di s_1^2 dipende dal vero valore assunto dai parametri componenti del vettore $\boldsymbol{\beta}$.

Inoltre, come si vede nelle pagine successive e come si intuisce dalle formule di queste pagine, la distribuzione di s_1^2 nel caso generale (ossia sotto H_1) dipende anche dalla configurazione della matrice $\mathbf{X}$ (scelta a priori o osservata) attraverso il prodotto $\mathbf{X}^\mathsf{T}\mathbf{X}$.

Pertanto è intuibile, sebbene non tratteremo tale argomento in dettaglio, che la scelta del particolare disegno della matrice $\mathbf{X}$, quando possibile, potrebbe influenzare la distribuzione di s_1^2 sotto H_1, e quindi in definitiva *il potere del test*.

In altre parole se per la costruzione di test in particolari contesti sperimentali è necessario operare con certi valori del potere del test, questo obiettivo può essere raggiunto agendo anche sugli elementi della matrice $\mathbf{X}$, ossia sulla configurazione del disegno sperimentale.

In generale se $\boldsymbol{\beta}_0$ è il valore specificato dall'ipotesi nulla e se $\boldsymbol{\beta}$ è il vero valore, allora possiamo calcolare il valore atteso della quantità $(\mathbf{b}-\boldsymbol{\beta}_0)^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}_0)$, effettuando alcune manipolazioni della forma quadratica:

DRAFT

$$\mathrm{E}\left[(\mathbf{b}-\boldsymbol{\beta}_0)^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}_0)\right]=$$

Aggiungendo e sottraendo $\boldsymbol{\beta}$

$$=\mathrm{E}\left[[(\mathbf{b}-\boldsymbol{\beta})-(\boldsymbol{\beta}_0-\boldsymbol{\beta})]^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}[(\mathbf{b}-\boldsymbol{\beta})-(\boldsymbol{\beta}_0-\boldsymbol{\beta})]\right]=$$

sviluppiamo il prodotto in cui compare il binomio formato dai d

termini: $(\mathbf{b}-\boldsymbol{\beta})$ e $(\boldsymbol{\beta}_0-\boldsymbol{\beta})$

$$=\mathrm{E}\left[(\mathbf{b}-\boldsymbol{\beta})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})\right]$$

$= k\sigma^2$ perché la forma quadratica si distribuisce come $\sigma^2\chi^2_k$ essen

$\boldsymbol{\beta}$ il vero valore

$$+$$

$$\mathrm{E}\left[(\boldsymbol{\beta}_0-\boldsymbol{\beta})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\boldsymbol{\beta}_0-\boldsymbol{\beta})\right]$$

é la speranza matematica di una costante

$$-$$

$$\mathrm{E}\left[2(\boldsymbol{\beta}_0-\boldsymbol{\beta})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})\right]=$$

$=\mathbf{0}$ perché é una combinazione lineare del vettore aleatorio $\mathbf{b}-$

che é a media nulla perché: $\mathrm{E}[\mathbf{b}]=\boldsymbol{\beta}$

$$=$$

$$k\sigma^2+(\boldsymbol{\beta}_0-\boldsymbol{\beta})^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\boldsymbol{\beta}_0-\boldsymbol{\beta})$$

DRAFT

argomenti avanzati e complementi (non in programma)

Riassumendo in una tavola questi ultimi risultati:

DRAFT

Quantità		$R(\mathbf{b})$	$R(\boldsymbol{\beta}_0) - R(\mathbf{b})$
Espressioni esplicite		$(\mathbf{y} - \mathbf{X}\mathbf{b})^\top (\mathbf{y} - \mathbf{X}\mathbf{b})$	$(\mathbf{b} - \boldsymbol{\beta}_0)^\top \mathbf{X}^\top \mathbf{X} (\mathbf{b} - \boldsymbol{\beta}_0)$
Interpretazione		Devianza residua	Scostamento dall'ipotesi nulla
Speranza matematica	$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$	$(n-k)\sigma^2$	$k\sigma^2$
	$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$	$(n-k)\sigma^2$	$k\sigma^2 + (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}^\top \mathbf{X} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})$
Distribuzione	$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$	$\sigma^2 \chi^2_{n-k}$	$\sigma^2 \chi^2_k$
	$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0$	$\sigma^2 \chi^2_{n-k}$	$\sigma^2 \chi^2(k, \lambda)$ non centrale; λ: parametro di non centralità $\lambda = (\boldsymbol{\beta}_0 - \boldsymbol{\beta})^\top \mathbf{X}^\top \mathbf{X} (\boldsymbol{\beta}_0 - \boldsymbol{\beta})$

DRAFT

Risulta evidente che $\mathrm{E}[F(H_1] > \mathrm{E}[F(H_0]$ e la regione di rifiuto del test va fissata sulla coda destra della distribuzione di F.

18.4 Scomposizione della devianza e test nel caso di gruppi di regressori ortogonali

Se r gruppi di variabili esplicative sono *ortogonali* (ossia risultano non correlati linearmente se si tratta di regressori scartati dalla media) la matrice $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ risulta composta da r blocchi disposti lungo la diagonale ($r \geq 2$) :

ciascun blocco è composto da un numero qualsiasi k_j di variabili, in modo tale che: $\Sigma_{j=1}^{r} k_j = k$;

Per esempio, un modello con termine noto e regressori quantitativi, $r = 2, k_1 = 1$;

In particolare se tutti i k_j sono uguali ad uno, vuol dire che tutte le variabili risultano ortogonali

eventualmente gli indici delle variabili sono permutati in modo che le variabili di

DRAFT

uno stesso gruppo siano consecutive

$$\mathbf{X}^\mathsf{T}\mathbf{X} = \begin{pmatrix} \mathbf{X}_1^\mathsf{T}\mathbf{X}_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathbf{X}_2^\mathsf{T}\mathbf{X}_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{X}_j^\mathsf{T}\mathbf{X}_j & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{X}_r^\mathsf{T}\mathbf{X}_r \end{pmatrix}$$

Ad esempio tutte le variabili del 1° blocco sono ortogonali a tutte quelle del j -esimo gruppo; all'interno di ciascun gruppo le variabili non sono ortogonali (o comunque non tutte). In corrispondenza a questi r blocchi suddividiamo il vettore dei parametri $\boldsymbol{\beta}$ e quello delle stime $\mathbf{b}$.

$$\boldsymbol{\beta}^\mathsf{T} = (\boldsymbol{\beta}_1^\mathsf{T}, \boldsymbol{\beta}_2^\mathsf{T}, \dots, \boldsymbol{\beta}_j^\mathsf{T}, \dots, \boldsymbol{\beta}_r^\mathsf{T})$$

$$\mathbf{b}^\mathsf{T} = (\mathbf{b}_1^\mathsf{T}, \mathbf{b}_2^\mathsf{T}, \dots, \mathbf{b}_j^\mathsf{T}, \dots, \mathbf{b}_r^\mathsf{T})$$

Il vantaggio per l'inferenza è che i gruppi di stimatori dei corrispondenti parametri saranno a blocchi non correlati (indipendenti data l'assunzione di normalità):

$$Cov(\mathbf{b}_j, \mathbf{b}_s) = 0 (j \neq s)$$

DRAFT

Dal punto di vista numerico, ciascun gruppo di stime è ricavabile da un sottoinsieme di equazioni normali:

$$\mathbf{X}_j^\mathsf{T}\mathbf{X}_j\mathbf{b}_j = \mathbf{X}_j^\mathsf{T}\mathbf{y}$$

e quindi:

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

è la matrice $\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$ risulta ora diagonale a blocchi:

$$\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} = \begin{pmatrix} \left(\mathbf{X}_1^\mathsf{T}\mathbf{X}_1\right)^{-1} & 0 & 0 & 0 & 0 & 0 \\ 0 & \left(\mathbf{X}_2^\mathsf{T}\mathbf{X}_2\right)^{-1} & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \left(\mathbf{X}_j^\mathsf{T}\mathbf{X}_j\right)^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \left(\mathbf{X}_r^\mathsf{T}\mathbf{X}_r\right)^{-1} \end{pmatrix}$$

La matrice di varianze e covarianze di $\mathbf{b}$ è data da:

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

DRAFT

Per cui possiamo scrivere, moltiplicando $\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}$ per lo scalare σ^2 :

$$\mathrm{V}\left[\mathbf{b}\right] = \begin{pmatrix} \mathrm{V}\left[\mathbf{b}_1\right] & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathrm{V}\left[\mathbf{b}_2\right] & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathrm{V}\left[\mathbf{b}_j\right] & 0 & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 0 & \mathrm{V}\left[\mathbf{b}_r\right] \end{pmatrix}$$

In generale è possibile scomporre semplicemente la forma quadratica $(\mathbf{b}-\boldsymbol{\beta})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})$ in r forme quadratiche (due o più) mutuamente indipendenti, se e solo se la matrice $\mathbf{X}$ può essere partizionata in r gruppi di regressori non correlati nel modo visto.

Possiamo in questo caso esprimere la forma quadratica:

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

$$= \sum_{j=1}^{r}(\mathbf{b}_j-\boldsymbol{\beta}_j)^{\mathsf{T}}\mathbf{X}_j^{\mathsf{T}}\mathbf{X}_j(\mathbf{b}_j-\boldsymbol{\beta}_j) = \sum_{j=1}^{r}\mathbf{Q}(\mathbf{b}_j-\beta_j);$$

Evidentemente le singole forme quadratiche si distribuiscono come delle variabili

DRAFT

aleatorie χ^2 con k_j gradi di libertà moltiplicate per σ^2 e sono indipendenti;

Ovviamente sono anche indipendenti rispetto a $R(\mathbf{b})$

per cui le scomposizioni viste prima in questo caso si estendono ulteriormente, scomponendo ciascun termine in r termini.

Si possono quindi costruire dei test F_j con numeratori indipendenti, mettendo a denominatore sempre s^2 (stima corretta della varianza) ed a numeratore l'opportuna forma quadratica $\mathbf{Q}(\mathbf{b}_j - \beta_j)$ divisa per i rispettivi gradi di libertà k_j :

$$F_j = \frac{\frac{[\mathbf{b}_j - \beta_j]^\mathsf{T} \mathbf{X}_j^\mathsf{T} \mathbf{X}_j [\mathbf{b}_j - \beta_j]}{k_j}}{\frac{[\mathbf{y} - \mathbf{X}\mathbf{b}]^\mathsf{T} [\mathbf{y} - \mathbf{X}\mathbf{b}]}{n-k}} = \frac{\frac{\mathbf{Q}(\mathbf{b}_j - \beta_j)}{k_j}}{s^2}$$

DRAFT

I rapporti F_j si distribuiscono secondo una F di Snedecor con k_j ed $n - k$ gradi di libertà

Pertanto, per saggiare un'ipotesi nulla:

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

contro l'alternativa generica

$$H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0,$$

nel caso di r blocchi ortogonali, si può considerare anche per il vettore $\boldsymbol{\beta}_0$ la stessa suddivisione in blocchi:

$$\boldsymbol{\beta}_0^{\mathsf{T}} = \left\{\beta_{10}^{\mathsf{T}}, \beta_{20}^{\mathsf{T}}, \ldots, \beta_{j0}^{\mathsf{T}}, \ldots, \beta_{r0}^{\mathsf{T}}\right\}$$

Per cui l'ipotesi nulla può essere suddivisa in r ipotesi,

$$H_{j0} : \beta_j = \beta_{j0} j = 1, 2, \ldots, r$$

per saggiare ciascuna delle quali possiamo impiegare i test:

$$F_j = \frac{\frac{[\mathbf{b}_j - \beta_{j0}]^{\mathsf{T}} \mathbf{X}^{\mathsf{T}} \mathbf{X} [\mathbf{b}_j - \beta_{j0}]}{k_j}}{\frac{(\mathbf{y} - \mathbf{X}\mathbf{b})^{\mathsf{T}} (\mathbf{y} - \mathbf{X}\mathbf{b})}{n-k}}, \qquad j = 1, 2, \ldots, r.$$

DRAFT

ognuno dei quali sotto H_0 si distribuisce secondo una variabile aleatoria F di Snedecor con k_j ed $n - k$ gradi di libertà (se n è grande questi test sono approssimativamente indipendenti).

E' possibile che l'ipotesi nulla specifichi solo alcuni gruppi di parametri, e non tutti.

Es. $H_0 : \beta_s = \beta_{s0}; \beta_j$ qualsiasi per $j \neq s$

In particolare può interessare:

$$H_0 : \beta_s = 0$$

Rispetto al test che si condurrebbe in presenza di un solo gruppo di regressori, cambia solo a denominatore la stima della varianza, che ha $n - k$ gradi di libertà invece che $n - k_s$. In ogni caso è meglio procedere con la stima con $n - k$ gradi di libertà che è certamente corretta

Se a ciascun gruppo di parametri e di regressori si può fare corrispondere una diversa fonte di variabilità, questo implica che per fare inferenza riguardo a ciascuna componente, *indipendentemente dalle altre*, è necessario che il gruppo di regressori corrispondente a ciascuna sorgente di variazione risulti ortogonale rispetto ai regressori corrispondenti alle altre sorgenti di variabilità. Questi aspetti sottolineano l'importanza di operare, quando possibile, con regressori ortogonali, almeno a gruppi, perché questo implicherà essenzialmente:

L'indipendenza fra i corrispondenti gruppi di stimatori;

L'indipendenza approssimata fra i test relativi ai vari gruppi di parametri, ossia alle differenti sorgenti di variabilità

DRAFT

19 Configurazioni della matrice $\mathbf{X}$ e di $\mathbf{X}^\top\mathbf{X}$

DRAFT

$\mathbf{X}$	$\mathbf{X}^\top\mathbf{X}$	Significato e conseguenze per l'interpretazione del modello e per l'inferenza
Tutte le $\mathbf{X}_j$ sono ortogonali	Diagonale	È il caso migliore: si possono saggiare ipotesi e fare inferenza in generale sui singoli parametri in modo indipendente (anche i valori degli stimatori si trovano in modo indipendente)
Tutte le combinazioni di valori dei fattori	*Fattoriale*	Meglio ancora! Fra l'altro migliorano le proprietà delle regioni di confidenza costruite su $\mathrm{E}[\mathbf{y}_i]$
Gruppi di $\mathbf{X}_j$ sono ortogonali	Diagonale a blocchi	È un caso importante: si possono saggiare ipotesi (e fare inferenza in generale) su gruppi di parametri separatamente
Correlazioni lineari generiche fra le $\mathbf{X}$	A rango pieno ma non diagonale	È il caso generale della regressione multipla, in particolare per studi osservazionali.
Qualcuna delle $\mathbf{X}_j$ è fortemente dipendente linearmente	A rango pieno ma con qualche autovalore vicino a zero	*Multicollinearità*

DRAFT

DRAFT

Comunque sia configurata la matrice $\mathbf{X}$ e quindi $\mathbf{X}^\top\mathbf{X}$, non sempre l'ipotesi d'interesse riguarda tutti i parametri. Infatti l'ipotesi nulla $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ è troppo generica e sono molto poche le situazioni in cui tale ipotesi può essere realistica: si rileggano gli esempi della parte introduttiva link con intro ; è improbabile che si voglia fare inferenza relativamente a valori fissati *per tutti i parametri:* di solito si è interessati ad inferenza sull'uguaglianza fra alcuni parametri, sulla possibilità di eliminarne alcuni (*non tutti!*), sul confronto fra alcuni di essi. In questa sezione verranno formalizzate e generalizzate queste ipotesi.

DRAFT

In generale si vogliono verificare ipotesi relative a sottoinsiemi di valori dei parametri, come ad esempio:

- $H_0 : \beta_1 = \beta_2 = 0;$ $\quad\quad$ β_j qualsiasi per $j > 2$
  che si legge semplicemente così: secondo l'ipotesi nulla le prime due variabili non influenzano la risposta $\mathbf{y}$.
- più in generale potremmo avere:
  $H_0 : \boldsymbol{\beta}_{\{s\}} = \boldsymbol{\beta}_{\{s\}0};$ $\quad\quad$ β_j qualsiasi per $j \notin s$

relativa ad un gruppo di parametri β_s; in pratica l'ipotesi nulla fissa i valori di s dei k parametri.

Può però interessarci un'ipotesi che implichi un confronto fra i valori di alcuni parametri; ad esempio:

$$H_0 : \ \beta_1 = \beta_2 = \beta_3 = \mu;$$

con μ non specificato) e β_j qualsiasi per $j > 3$ ossia *gli effetti dei primi tre fattori sono uguali.*

Quest'ultima ipotesi equivale ad imporre i due vincoli:

DRAFT

$$\begin{aligned} \beta_1 - \beta_3 &= 0 \\ \beta_2 - \beta_3 &= 0 \end{aligned}$$

In effetti queste ipotesi nulle possono essere considerate come delle ipotesi che impongono dei vincoli lineari (anche molto generali) sui valori dei k parametri, secondo la relazione generale:

$$\mathbf{C}\boldsymbol{\beta} = \boldsymbol{\theta}_0.$$

Le 2 uguaglianze dell'ultimo esempio corrispondono ad una scelta di $\mathbf{C}$ di 2 righe e k colonne:

In dettaglio, dato il modello:

$$\mathbf{y}_{[n\times 1]} = \mathbf{X}_{[n\times k]}\beta_{[k\times 1]} + \boldsymbol{\varepsilon}_{[n\times 1]}$$

(supponiamo sempre $\mathbf{X}$ di rango k) in generale siamo interessati a verificare l'ipotesi:

$$H_0 : \ \mathbf{C}_{[q\times k]}\beta_{[k\times 1]} \ = \ \mathbf{a}_{[q\times 1]}.$$

con $q < k$ e q rango di $\mathbf{C}$

DRAFT

20.0.1 Esempio: Analisi della varianza ad una via.

Si riveda l'impostazione della matrice $\mathbf{X}$ nella parte introduttiva sui modelli lineari; 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:

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

L'ipotesi che può interessare non è però che tutti i coefficienti siano nulli, ma che siano uguali fra loro:

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

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

$$\mathbf{C}_{[k-1\times k]} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \text{Gruppo} \\ 1 \ \ 2 \ \ \ldots \ \ j \ \ \ldots \ \ \ldots \ \ k \\ \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \\ \\ \\ \\ \\ \\ \text{vincolo} \end{array}$$

$$\begin{pmatrix} 1 & 0 & \ldots & 0 & \ldots & \ldots & -1 \\ 0 & 1 & \ldots & 0 & \ldots & \ldots & -1 \\ 0 & 0 & \ldots & 0 & \ldots & 0 & -1 \\ 0 & 0 & \ldots & 1 & \ldots & 0 & -1 \\ 0 & 0 & \ldots & 0 & \ldots & \ldots & -1 \\ 0 & 0 & \ldots & 0 & \ldots & 1 & -1 \end{pmatrix} \begin{array}{c} 1 \\ 2 \\ \ldots \\ j \\ \ldots \\ k-1 \end{array} \qquad \mathbf{a} = \begin{pmatrix} 0 \\ 0 \\ \ldots \\ 0 \\ \ldots \\ 0 \end{pmatrix}$$

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

DRAFT

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

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

Riprendiamo l'esempio sull'ipotesi nulla:

$$H_0 : \ \beta_1 = \beta_2 = \beta_3 = \mu;$$

con μ non specificato) e β_j qualsiasi per $j > 3$.
La matrice dei vincoli è costituita da due sole righe:

DRAFT

$$\mathbf{C}_{[2\times k]} = \begin{array}{c} \text{Parametro} \\ 1 \ \ 2 \ \ 3 \ \ \ldots \ \ \ldots \ \ k \\ \begin{pmatrix} 1 & 0 & -1 & \ldots & \ldots & 0 \\ 0 & 1 & -1 & \ldots & \ldots & 0 \end{pmatrix} \end{array} \begin{array}{c} \text{vincolo} \\ 1 \\ 2 \end{array}$$

con:

$$\mathbf{a} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Altro esempio:
se l'ipotesi di interesse è:

$$H_0 : \ \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

questo corrisponde a scegliere:

$$\mathbf{C} = \mathbf{I}_k; \quad \mathbf{a} = \mathbf{0}_k.$$

20.0.2 Esempio sulla scelta delle variabili.

In un modello di regressione multipla si può avere un problema di scelta di variabili link con scelta variabili regressione .

L' ipotesi:

$$H_0 : \ \beta_1 = \beta_2 = \ldots = \beta_q = 0, q < k;$$

e

$\beta_{q+1}, \beta_{q+2}, \ldots, \beta_k$ qualsiasi
corrisponde a q vincoli definiti da:

$$\mathbf{C} = [\mathbf{I}_q \vdots \mathbf{0}_{k-q}]; \quad \mathbf{a} = \mathbf{0}_q$$

ossia i vincoli non coinvolgono i $k - q$ regressori oltre β_q .

Ovviamente $q = 1$ nel caso di ipotesi concernenti un singolo parametro.

DRAFT

In questo caso per costruire il rapporto di verosimiglianza per la verifica dell'ipotesi generale:

$$H_0 : \ \mathbf{C}_{[q\times k]}\beta_{[k\times 1]} = \mathbf{a}_{[q\times 1]} \qquad \mathbf{C} \text{ di rango } \ q$$

(con H_1 : ipotesi alternativa che non fissa alcun vincolo sui parametri) si ha:

$$\begin{aligned} LR &= \frac{\max L[\boldsymbol{\beta}, \sigma^2, \mathbf{y}|H_0]}{\max L[\boldsymbol{\beta}, \sigma^2, \mathbf{y}|H_1]} = \\ &= \frac{\max L[\boldsymbol{\beta}, \sigma^2, \mathbf{y}|\mathbf{C}\boldsymbol{\beta} = \mathbf{a}]}{\max L[\boldsymbol{\beta}, \sigma^2, \mathbf{y}|\boldsymbol{\beta}\epsilon\Re_k]} = \\ &= \left\{\frac{R(\mathbf{b}_0)}{R(\mathbf{b})}\right\}^{-\frac{n}{2}} \end{aligned}$$

essendo $\mathbf{b}$ lo stimatore di massima verosimiglianza non vincolato, e $\mathbf{b}_0$ lo stimatore di massima verosimiglianza sotto i vincoli lineari imposti da H_0.

DRAFT

21.0.1 Minimi quadrati vincolati

Per trovare $\mathbf{b}_0$ occorre risolvere un problema di minimi quadrati vincolati [9]:

$$\begin{aligned} \min_{b0} R(\mathbf{b}_0) &= (\mathbf{y} - \mathbf{X}\mathbf{b}_0)^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{b}_0) = \\ &= \mathbf{y}^\mathsf{T}\mathbf{y} - 2b_0^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{y} + \mathbf{b}_0^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}\mathbf{b}_0 \end{aligned}$$

soggetto a q vincoli lineari:

$$\mathbf{C}\mathbf{b}_0 = \mathbf{a}; \quad \mathbf{C} \text{ è di rango } q$$

DRAFT

Occorre costruire il Lagrangiano $\mathcal{L}(\mathbf{b}_0, \mathbf{d})$ essendo $\mathbf{d}$ un vettore di q moltiplicatori di Lagrange ed uguagliare a $\mathbf{0}$ le derivate di $\mathcal{L}(\mathbf{b}_0, \mathbf{d})$ rispetto al vettore $\mathbf{b}_0$ ed al vettore $\mathbf{d}_{[q\times 1]}$:

$$\begin{aligned} \mathcal{L}(\mathbf{b}_0, \mathbf{d}) &= R(\mathbf{b}_0) + 2(\mathbf{C}\mathbf{b}_0 - \mathbf{a})^\mathsf{T}\mathbf{d} \\ &= (\mathbf{y} - \mathbf{X}\mathbf{b}_0)^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{b}_0) + 2(\mathbf{C}\mathbf{b}_0 - \mathbf{a})^\mathsf{T}\mathbf{d} \end{aligned}$$

[9] Il lettore poco interessato può passare direttamente alla sezione della scomposizione della devianza 21.1

$$\begin{cases} \frac{\partial \mathcal{L}(\mathbf{b}_0,\mathbf{d})}{\partial \mathbf{b}_0} &= 2\mathbf{X}^\mathsf{T}\mathbf{y} + 2(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{b}_0 + 2\mathbf{C}^\mathsf{T}\mathbf{d} \\ \\ \frac{\partial \mathcal{L}(\mathbf{b}_0,\mathbf{d})}{\partial \mathbf{d}} &= 2(\mathbf{C}\mathbf{b}_0 - \mathbf{a}) \end{cases}$$

Occorre uguagliare a $\mathbf{0}$ (vettore nullo) tali derivate parziali e risolvere rispetto a $\mathbf{b}_0$ e $\mathbf{d}$:

$$\begin{cases} -2\mathbf{X}^\mathsf{T}\mathbf{y} + 2(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{b}_0 + 2\mathbf{C}^\mathsf{T}\mathbf{d} &= \mathbf{0} \\ 2(\mathbf{C}\mathbf{b}_0 - \mathbf{a}) &= \mathbf{0} \end{cases} \tag{10}$$

Dal primo gruppo di equazioni 10 ricaviamo successivamente:

$$(\mathbf{X}^\mathsf{T}\mathbf{X})\mathbf{b}_0 = \mathbf{X}^\mathsf{T}\mathbf{y} - \mathbf{C}^\mathsf{T}\mathbf{d}$$

e quindi risolvendo rispetto a $\mathbf{b}_0$:

$$\mathbf{b}_0 = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y} - \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}\mathbf{d};$$

se adesso sostituiamo $\mathbf{b} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{X}^\mathsf{T}\mathbf{y}$, stimatore dei minimi quadrati non vincolato, otteniamo un'espressione più compatta (e statisticamente più interessante

$$\mathbf{b}_0 = \mathbf{b} - \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}\mathbf{d}$$

DRAFT

Dal secondo gruppo di equazioni del sistema (10), ossia il gruppo relativo al soddisfacimento dei vincoli:

$$\mathbf{a} = \mathbf{C}\mathbf{b}_0 \Longrightarrow \mathbf{a} = \mathbf{C}\mathbf{b} - \mathbf{C}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}\mathbf{d};$$

Sono q equazioni indipendenti in q incognite $\mathbf{d}$,

$$\mathbf{C}\mathbf{b} = -\mathbf{C}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}\mathbf{d};$$

con soluzione data da:

$$-\mathbf{d} = [\mathbf{C}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}]^{-1}(\mathbf{a} - \mathbf{C}\mathbf{b})$$

risostituendo nel sistema che fornisce $\mathbf{b}_0$ si ha:

$$\begin{aligned}\mathbf{b}_0 &= \mathbf{b} - \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}\mathbf{d} = \\ &= \mathbf{b} + \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}[\mathbf{C}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}]^{-1}(\mathbf{a} - \mathbf{C}\mathbf{b}) \qquad (11)\end{aligned}$$

Si può facilmente vedere che questa soluzione fornisce il minimo e rispetta i vincoli (premoltiplicando per $\mathbf{C}$). Tutte le inverse citate esistono, per le ipotesi fatte sui ranghi di $\mathbf{X}$ e $\mathbf{C}$.

DRAFT

In realtà di solito conviene risolvere il sistema dei minimi quadrati secondo la parametrizzazione fornita da H_0, se questa è esplicitabile rispetto ai parametri.

La tecnica ora esposta per trovare $\mathbf{b}_0$ *è utile prevalentemente a scopo teorico per vedere la relazione fra* $\mathbf{b}_0$ *e* $\mathbf{b}$*; inoltre è utile per i casi nei quali* $\mathbf{C}\boldsymbol{\beta} = \mathbf{a}$ *non sia esplicitabile in modo semplice.*

Riscriviamo l' espressione di $\mathbf{b}_0$, in modo che sia evidente la relazione lineare fra $\mathbf{b}_0$ e $\mathbf{b}$. Poniamo, per semplicità:

$$\mathbf{G} = \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}[\mathbf{C}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}]^{-1}$$

.

Si ottiene quindi:

$$\begin{aligned}\mathbf{b}_0 &= \mathbf{b} + \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}[\mathbf{C}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}]^{-1}(\mathbf{a} - \mathbf{C}\mathbf{b}) = \\ &= \mathbf{b} + \mathbf{G}(\mathbf{a} - \mathbf{C}\mathbf{b}) = \mathbf{G}\,\mathbf{a} + (\mathbf{I}_k - \mathbf{G}\,\mathbf{C})\mathbf{b}\end{aligned}$$

$\mathbf{b}_0$risulta corretto solo sotto H_0.

Infatti in generale si ha:

$$\mathrm{E}\left[\mathbf{b}_0\right] = \boldsymbol{\beta} + \left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}[\mathbf{C}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}]^{-1}(\mathbf{a} - \mathbf{C}\boldsymbol{\beta})$$

DRAFT

e quindi sotto H_0:

$$\mathrm{E}\left[\mathbf{b}_0\right] = \boldsymbol{\beta}$$

perchè solo sotto H_0 : $\mathbf{a} - \mathbf{C}\boldsymbol{\beta} = 0$

Inoltre per la matrice di varianze e covarianze si ha in generale sotto H_1:

$$\mathrm{V}\left[\mathbf{b}_0\right] = \left(\mathbf{I}_k - \mathbf{GC}\right)^{\mathsf{T}} \mathrm{V}\left[\mathbf{b}\right] \left(\mathbf{I}_k - \mathbf{GC}\right) =$$

$$= \sigma^2 \left\{ \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} - \mathbf{GC}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} - \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}\mathbf{G}^{\mathsf{T}} + \mathbf{GC}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}\mathbf{G}^{\mathsf{T}} \right\}$$

---

verificare

DRAFT

`Questi tre termini risultano uguali in valore assoluto.`

Infine, dopo qualche semplificazione:

$$\begin{aligned} \mathrm{V}\left[\mathbf{b}_0\right] &= \sigma^2 \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} - \sigma^2 [\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}} [\mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}]^{-1} \mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}] = \\ &= \mathrm{V}\left[\mathbf{b}\right] - \sigma^2 [\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}} [\mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}]^{-1} \mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}]. \end{aligned}$$

- Le varianze di ciascun elemento di $\mathbf{b}_0$risultano inferiori a quelle dei corrispondenti elementi di $\mathbf{b}$;
- Si ricordi però che in generale $\mathbf{b}_0$ è distorto.

21.1 Modello lineare: Scomposizione della devianza per il problema soggetto a vincoli:

Procediamo adesso in modo del tutto simile a quanto fatto per il problema non vincolato; anche in questo caso la devianza residua può essere scomposta in una forma conveniente:

DRAFT

$$R(\mathbf{b}_0) = (\mathbf{y} - \mathbf{X}\mathbf{b}_0)^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{b}_0) =$$

Sommando e sottraendo Xb e poi aprendo il quadrato del bi

$$= [(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_0)]^\mathsf{T}[(\mathbf{y} - \mathbf{X}\mathbf{b}) + (\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_0)] =$$

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

=R(**b**)

$$+$$

$$(\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_0)^\mathsf{T}(\mathbf{X}\mathbf{b} - \mathbf{X}\mathbf{b}_0)$$

si mette in evidenza **X** sia a sinistra che a destra e si ottien

$(\mathbf{b} - \mathbf{b}_0)^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b} - \mathbf{b}_0)$

$$+$$

$$2(\mathbf{y} - \mathbf{X}\mathbf{b})^\mathsf{T}\mathbf{X}(\mathbf{b} - \mathbf{b}_0)$$

=0 perché:

$(\mathbf{y} - \mathbf{X}\mathbf{b})^T\mathbf{X} = 0$ dalle equazioni dei minimi quadrati

$$=$$

$$R(\mathbf{b}) + (\mathbf{b} - \mathbf{b}_0)^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b} - \mathbf{b}_0)$$

DRAFT

In definitiva:

$$R(\mathbf{b}_0) = R(\mathbf{b}) + (\mathbf{b} - \mathbf{b}_0)^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b} - \mathbf{b}_0) \qquad (12)$$

Devianza residua supplementare dovuta ad H_0

$$(\mathbf{b}-\mathbf{b}_0)^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b}-\mathbf{b}_0)$$

Misura anche la distanza fra i due stimatori.

In quest'ultima espressione della devianza supplementare, proviamo a evidenziare il ruole della differenza fra gli stimatori [10] e sostituiamo l' espressione di $(\mathbf{b}-\mathbf{b}_0)$:

$$\begin{aligned} R(\mathbf{b}_0)-R(\mathbf{b}) &= (\mathbf{b}-\mathbf{b}_0)^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b}-\mathbf{b}_0) = \\ &= (\mathbf{a}-\mathbf{C}\mathbf{b})^{\mathsf{T}}[\mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}]^{-1}(\mathbf{a}-\mathbf{C}\mathbf{b}) \end{aligned}$$

ottenuta ricordando dalla (11) che:

$$\mathbf{b}_0 = \mathbf{b} + \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}[\mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}]^{-1}(\mathbf{a}-\mathbf{C}\mathbf{b})$$

ed effettuando una serie di banali semplificazioni fra le matrici della forma quadratica. Conviene adesso nella forma quadratica cambiare di segno i due termini es-

[10] ricordo sempre che tale differenza è dovuta all'imposizione di un sistema di vincoli lineari

DRAFT

terni, per cui si ha:

$$R(\mathbf{b}_0)-R(\mathbf{b}) = (\mathbf{a}-\mathbf{C}\mathbf{b})^{\mathsf{T}}[\mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}]^{-1}(\mathbf{a}-\mathbf{C}\mathbf{b}) = (\mathbf{C}\mathbf{b}-\mathbf{a})^{\mathsf{T}}[\mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}]^{-1}(\mathbf{C}\mathbf{b}-$$

Ora si rifletta sul fatto che sotto H_0 si ha: $\mathrm{E}[\mathbf{C}\mathbf{b}] = \mathbf{a}$ e inoltre che

$$\mathrm{V}[\mathbf{C}\mathbf{b}] = \sigma^2\,[\mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}]^{-1}$$

Allora, dividendo per σ^2, possiamo scrivere meglio questa relazione:

$$\frac{R(\mathbf{b}_0)-R(\mathbf{b})}{\sigma^2} = (\mathbf{C}\mathbf{b}-\mathrm{E}[\mathbf{C}\mathbf{b}|H_0])^{\mathsf{T}}\mathrm{V}[\mathbf{C}\mathbf{b}]\,(\mathbf{C}\mathbf{b}-\mathrm{E}[\mathbf{C}\mathbf{b}|H_0]) \qquad (13)$$

DRAFT

Ancora una volta ritroviamo un risultato simile a quello visto in precedenza: *si tratta della forma quadratica ad esponente della densità normale multivariata della v.a.* $\mathbf{C}\mathbf{b}$! Pertanto si distribuisce (sotto H_0) come una χ^2 con q gradi di libertà. Questo è perfettamente ragionevole e coerente con quanto visto finora: in sostanza è come se invece di fare inferenza su $\boldsymbol{\beta}$ fossimo interessati all'inferenza su $\mathbf{C}\boldsymbol{\beta}$: essendo quantità lineari nei parametri, invece di raginare sullo stimatore $\mathbf{b}$ di $\boldsymbol{\beta}$, ragioniamo sullo stimatore $\mathbf{C}\mathbf{b}$ di $\mathbf{C}\boldsymbol{\beta}$!

Possiamo adesso riapplicare il teorema di Cochran alla scomposizione (12); div-

idendo tutto per σ^2 otteniamo infatti:

$$\underbrace{\frac{R(\mathbf{b}_0)}{\sigma^2}}_{\text{si distribuisce come una } \chi^2 \text{ sotto } H_0} = \underbrace{\frac{R(\mathbf{b})}{\sigma^2}}_{\text{si distribuisce come una } \chi^2 \text{ (anche sotto } H_1)} + \underbrace{\frac{(\mathbf{b}-\mathbf{b}_0)^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b}-\mathbf{b}_0)}{\sigma^2}}_{\text{essendo ad esponente di una densità di una normale multivariata (cfr. 13) si distribuisce come una } \chi^2 \text{ (sotto } H_0)}$$

DRAFT

Possiamo finalmente applicare il teorema di Cochran, e quindi concludere che [11] le forme quadratiche sono indipendenti e in particolare che $(\mathbf{b}-\mathbf{b}_0)^\mathsf{T}\mathbf{X}^\mathsf{T}\mathbf{X}(\mathbf{b}-\mathbf{b}_0)$ si distribuisce proporzionalmente ad una χ^2 con q gradi di libertà *indipendentemente da* $R(\mathbf{b})$!

Pertanto è possibile costruire il test per la verifica di una ipotesi qualsiasi semplicemente mettendo a numeratore del test F l'incremento di devianza dovuto ad H_0 (e modificando i gradi di libertà):

[11] sotto H_0

Test per la verifica di un'ipotesi generale

In ogni caso il rapporto:

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

si distribuisce (sotto H_0) come una F con q ed $n-k$ gradi di libertá , se è valida l' ipotesi nulla: $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{a}$(con q numero di gradi di libertá del numeratore, ossia rango di $\mathbf{C}$)

DRAFT

Ovviamente si vede facilmente che questo rapporto è funzione del rapporto delle verosimiglianze (si confronti con la (link con verosimiglianza profilo)).

Un ulteriore utile interpretazione è data dalla seguente scomposizione della devianza totale di $\mathbf{y}$ (banale ma importante per l'interpretazione della differenza fra le devianze residue dei modelli con e senza vincoli):

$$Dev(\mathbf{y}) =$$

$$= \underbrace{(Dev(\mathbf{y}) - R(\mathbf{b}_0))}_{\text{Devianza spiegata sotto } H_0} + \underbrace{R(\mathbf{b})}_{\text{Devianza residua sotto } H_1} + \underbrace{(R(\mathbf{b}_0) - R(\mathbf{b}))}_{\text{Incremento di devianza residua dovuto ai vincoli di } H_0}$$

Occorre ancora una volta ricordare un fatto importantissimo, per la teoria sui modelli lineari: in questa sezione abbiamo trovato la distribuzione esatta di uno strumento generale per verificare ipotesi (e successivamente lo impiegheremo anche per costruire intervalli di confidenza) relative ai parametri di un modello lineare; il test è stato costruito con i ragionamenti sulla verosimglianza visti in merito alla (link con verosimiglianza profilo); ora ci stiamo limitando a vedere qual è la distribuzione esatta di tali quantità nel caso di vincoli lineari sui parametri! Il fatto che riusciamo a ricavare la distribuzione esatta e che questa sia riconducibile ad una F di Snedecor è come una ciliegina sulla torta: **la cosa importante è sapere che il test migliore dipende dalle devianze residue**.

DRAFT

Avendo esposto la teoria generale, ricaviamo alcuni casi particolari molto utili per le applicazioni, in corrispondenza di alcune tipiche configurazioni della matrice $\mathbf{C}$ dei vincoli.

Supponiamo che la matrice $\mathbf{C}$ sia costituita, come nell'esempio della sezione (20.0.2), da:

$$\mathbf{C} = [\mathbf{I}_q : \mathbf{0}_{q \times k}]$$

e quindi l'ipotesi nulla:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{a}$$

specifica solo i valori dei primi q parametri. [12]

La matrice $\left(\mathbf{C}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}\right)^{-1}$ ora risulta costituita dall'inversa del blocco $q \times q$ della matrice $\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$ corrispondente ai q parametri specificati da H_0,ossia $\left\{\left[\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\right]_q\right\}^{-1}$. Infatti, se indichiamo la matrice inversa $\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$ con $\mathbf{A}$, possiamo partizionarla in quattro blocchi (partizionando sia le righe che le colonne in

[12] ovviamente potrebbe trattarsi di un qualsiasi sottoinsieme di q parametri, ma per comodita' ordiniamo i parametri in modo che quelli coinvolti dall'ipotesi nulla siano i primi q.

DRAFT

due gruppi di q e $k-q$ elementi rispettivamente):

$$\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1} = \mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^\mathsf{T} & \mathbf{A}_{22} \end{pmatrix}$$

e quindi si ha per il prodotto $\mathbf{C}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T}$:

$$\begin{aligned} \mathbf{C}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{C}^\mathsf{T} &= \mathbf{C}\mathbf{A}\mathbf{C}^\mathsf{T} = \\ &= [\mathbf{I}_q \vdots \mathbf{0}_{q\times k}] \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{12}^\mathsf{T} & \mathbf{A}_{22} \end{pmatrix} \begin{bmatrix} \mathbf{I}_q \\ \cdots \\ \mathbf{0}_{q\times k} \end{bmatrix} = \\ &= \mathbf{A}_{11} \end{aligned}$$

DRAFT

il vettore di q elementi $(\mathbf{a}-\mathbf{C}\mathbf{b})$ e' semplicemente costituito dalla differenza fra valori ipotizzati e valori stimati sotto H_0.

Se con $[\mathbf{b}_0]_q^T$ indichiamo il vettore di q elementi coinvolto dall'ipotesi nulla e con $[\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}{}_q]$ indichiamo il blocco $q \times q$ nella matrice $\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}$ (ossia quello indicato con $\mathbf{A}_{11}$), il test F visto prima diventa ora:

Test per la verifica di un'ipotesi su q elementi di $\boldsymbol{\beta}$

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

si distribuisce (sotto H_0) come una F con q ed $n-k$ gradi di libertÃ¡ , se e' valida l' ipotesi nulla: $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.

verificare

DRAFT

Va precisato che questo approccio a rigore e' valido per saggiare ipotesi singole, anche concernenti q parametri, ma non gruppi di ipotesi, perchÃ© le forme quadratiche relative a sottoinsiemi differenti di parametri (o di loro combinazioni lineari) non sono indipendenti, se non nel caso visto prima di matrice $\mathbf{X}^\intercal\mathbf{X}$ a blocchi diagonali.

rilevare l'analogia con il modello normale condizionato nell'espressione dellinversa del blocco dell'inversa

precisare

Condurre invece in parallelo test separati sugli elementi di $\boldsymbol{\beta}$ in assenza dei

necessari requisiti di ortogonalita' e' in generale una procedura errata, nel senso che non vengono certamente rispettati i livelli di significativita' nominali. Puo' essere utile, in analisi esplorative, a titolo comparativo, per confrontare verosimiglianze relative a modelli concorrenti, ma non per effettuare test nel vero senso del termine mantenendo un prefissato livello di significativita'.

22 Test e regioni di confidenza nei modelli lineari

DRAFT

L'approccio visto prima, sui test LR per ipotesi che impongono q vincoli lineari sui parametri, a rigore va impiegato solo per saggiare un'ipotesi concernente un unico set di parametri;

oppure occorre avere set di ipotesi ortogonali. In generale se $k > 1$ non esiste un test UMPU.

22.0.1 Regioni di confidenza simultanee per i parametri

La regione di confidenza migliore, ad un livello $1 - \alpha$, e' determinata dai valori $\boldsymbol{\beta}$ per i quali i valori osservati del test F non risultano superiori al valore teorico

$F_{\alpha,k,n-k}$.

Pertanto, dato un campione nel quale $\mathbf{b}$ e' la stima di massima verosimiglianza, tale regione e' delimitata dai valori $\boldsymbol{\beta}$ per i quali:

$$(\mathbf{b}-\boldsymbol{\beta})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}) \leq ks^2F_{\alpha,k,n-k}$$

Nel caso di regressori non ortogonali, tali regioni risulteranno date da ellissoidi con assi obliqui, per cui l'interpretazione delle regioni stesse potra' essere ardua.

Anche la relazione con i singoli intervalli sara' di difficile interpretazione, infatti per ciascun valore di uno dei parametri, l'intervallo ottimo dell'altro varia, sia per posizione che per estensione.

DRAFT

22.1 regioni di confidenza per funzioni lineari dei parametri

In effetti se siamo interessati a particolari combinazioni di parametri $\mathbf{a} = \mathbf{C}\ \boldsymbol{\beta}$, possiamo direttamente costruire regioni di confidenza per tali funzioni lineari dei parametri a partire dalla quantita':

$$R(\mathbf{b}_0) - R(\mathbf{b}) = (\mathbf{a}-\mathbf{C}\mathbf{b})^{\mathsf{T}}[\mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}]^{-1}(\mathbf{a}-\mathbf{C}\mathbf{b});$$

Prendendo in considerazione il corrispondente test F si puo' direttamente costruire la regione (q-dimensionale) costituita da tutti i valori $\mathbf{a}$ per i quali:

$$(\mathbf{a} - \mathbf{Cb})^{\mathsf{T}}[\mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}}]^{-1}(\mathbf{a} - \mathbf{Cb}) \leq qs^2F_{\alpha,q,n-k}$$

22.1.1 regioni di confidenza relative a sottoinsiemi di parametri

Se la matrice $\mathbf{C}$ e' definita da:

$$\mathbf{C} = [\mathbf{I}_q|\mathbf{0}_{q\times k}]$$

(ossia specifica solo i valori di q parametri), allora:

la matrice$(\mathbf{C}\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}\mathbf{C}^{\mathsf{T}})^{-1}$ risulta costituita dall'inversa del blocco $q \times q$ della matrice $\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}$ corrispondente ai q parametri specificati da H_0,ossia$[\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}{}_q]^{-1}$

il vettore di q elementi $(\mathbf{a} - \mathbf{Cb})$ e' semplicemente costruito dalla differenza fra valori dei parametri e valori degli stimatori per soli q dei k parametri.

La regione (q-dimensionale) e' quindi costituita dai valori di β_q per i quali:

$$(\mathbf{b} - \boldsymbol{\beta})_q^{\mathsf{T}}[\left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1}{}_q]^{-1}(\mathbf{b} - \boldsymbol{\beta})_q \leq qs^2F_{\alpha,q,n-k}$$

DRAFT

$[(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}{}_q]$ indica il blocco $q \times q$nella matrice $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$. $[\mathbf{b}-\beta]_q$ indica l'opportuno sottovettore di q elementi

analogia con la normale multivariata condizionata

22.1.2 Intervalli di confidenza per $E(\mathbf{y}_i)$

Per quanto visto prima, e' evidente che lo stimatore migliore di $E(\mathbf{y}_i)$ e' $\mathbf{y}_i* = \mathbf{x}_i^\mathsf{T}\mathbf{b}$, essendo $\mathbf{x}_{(i)}$ il vettore di osservazioni dei regressori corrispondente all'unita' i-esima, e quindi rientriamo nel caso i combinazioni lineari degli stimatori $\mathbf{b}$.

Pertanto, e comunque se il modello e' completo e corretto:

$$E(\mathbf{y}_i*) = E(\mathbf{x}_i^\mathsf{T}\mathbf{b}) = \mathbf{x}_i^\mathsf{T}\boldsymbol{\beta} = E(\mathbf{y}_i)$$

$$V(\mathbf{y}_i*) = V(\mathbf{x}_i^\mathsf{T}\mathbf{b}) = \mathbf{x}_i^\mathsf{T}V(\mathbf{b})\mathbf{x}_{(i)} = \sigma^2\mathbf{x}_i^\mathsf{T}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{x}_{(i)}$$

essendo al solito $\mathbf{x}_i^\mathsf{T}$ l'i-esima riga della matrice $\mathbf{X}$.

Applicando quindi le formule dei paragrafi precedenti, otteniamo l'intervallo di confidenza per $E(\mathbf{y}_i)$ ad un livello di probabilita' fiduciaria $1-\alpha$, dato da:

$$\mathbf{x}_i^\mathsf{T}\mathbf{b} \mp t_{\alpha,n-k}\sqrt{\mathbf{x}_i^\mathsf{T}\left(\mathbf{X}^\mathsf{T}\mathbf{X}\right)^{-1}\mathbf{x}_{(i)}}.$$

DRAFT

Risulta dunque evidente che il luogo dei punti $\mathbf{x}_{(i)}$ per i quali tali intervalli risultano di uguale ampiezza, a parita' di altre condizioni, e' costituito dai punti per i quali

$$\mathbf{x}_i^{\mathsf{T}} \left(\mathbf{X}^{\mathsf{T}}\mathbf{X}\right)^{-1} \mathbf{x}_{(i)} = \text{Costante},$$

ossia dai punti che hanno uguale distanze di MahalaNobis dal centroide dei regressori.

Nelle figure viste a lezione sono mostrati gli effetti dovuti configurazioni diverse delle $\mathbf{X}$.

DRAFT

22.1.3 errori di previsione

Varianza degli errori di previsione e distorsione degli stimatori variano in senso opposto