LEZIONE (13) 15/06/09

STIMATORE BLUE (segue da pag. 209)

Osservazione

- Se il parametro σ^2 è incognito si dimostra che

$$\hat{\sigma}^2 = \frac{\varepsilon^T \Psi^{-1} \varepsilon}{n - q}$$

è uno stimatore non polarizzato di σ^2.

Ricordiamo che q = numero di parametri (dimensione del vettore θ) e n = numero di dati e che valgono le ipotesi I_1 e l'ipotesi $n > q$ (pag. 206).

- **Nota**

Quanto affermato sullo stimatore BLUE vale anche con l'ipotesi di identificabilità, secondo la quale

$$\text{rank}(\Phi) = q \quad \text{(pag. 199)}$$

L'osservazione fatta dimostra che si può avere una conoscenza incompleta delle caratteristiche statistiche degli errori!

Si ricordi che σ^2 è la varianza dell'errore di misura.

Nonostante ciò si cerca di costruire il modello e con l'analisi dei residui si cerca di individuare le proprietà statistiche dell'errore di misura.

Se Ψ fosse la matrice identità al numeratore si avrebbe la somma dei quadrati dei residui e ci si attenderebbe di dover dividere per n per avere l'aspettativa media del quadrato dei residui.

Al denominatore troviamo $(n - q)$ dato la polarizzazione, con cui nella stima varianza campionaria compare $(n - 1)$ (pag. 155).

Se si dividesse per n si avrebbe una stima polarizzata, una proprietà comportamento meno, stimatore non polarizzato.

- Osservazione

Sempre nel caso di σ^2 incognita, usando $\hat{\sigma}^2$ si può stimare anche la varianza del vettore dei parametri

$$\hat{\Sigma}_\theta = \hat{\sigma}^2 \left(\Phi^T \Psi^{-1} \Phi \right)^{-1}$$

La varianza vera di θ^M è:

$$Var\left[\theta^M\right] = \Sigma_\theta = \sigma^2 \left(\Phi^T \Psi^{-1} \Phi \right)^{-1}$$

Ovviamente non conoscendo σ^2 non si può usare la stima a disposizione.

Rimane aperto il problema degli intervalli di confidenza per il vettore dei parametri θ: da qui si può risalire non si conosce la d.d.p. di θ^M, ossia dello stimatore.

- Nota

$$\theta^M = cY = c\left(\Phi\theta^0 + V\right),$$

essendo $c = \left(\Phi^T \Psi^{-1} \Phi \right)^{-1} \Phi^T \Psi^{-1}$ (pag. 206)

La matrice c non è tutta deterministica, mentre Y è una v.c.

Quindi θ^M risulta il prodotto di una matrice c deterministica, calcolabile, per un vettore di v.c.

Le proprietà statistiche di θ^M dipendono dalle proprietà statistiche di Y.

A meno dello Y è la somma di un termine non noto ma deterministico ($\Phi\theta^0$, non conoscendo il vettore vero dei parametri, e di un termine casuale (V).

Quindi le proprietà statistiche, come la d.d.p. di θ^M dipendono dalla d.d.p. di V, cioè dagli errori di misura.

Le proprietà degli errori di misura determinano le proprietà dello stimatore.

Per poter effettuare una analisi più approfondita occorre introdurre un'ipotesi sulla d.d.p. di V, che indicheremo con I_1.

Ipotesi I_1

$$Y = \underset{I}{\Phi}\,\theta^0 + V,$$

con V distribuito gaussianamente con media nulla e matrice varianza Σ_V.

$$V \sim N(0, \Sigma_V)$$

con

$$\Sigma_V = \sigma^2 \Psi,$$

essendo Ψ nota e > 0 e σ^2 uno scalare eventualmente incognito.

Rispetto all'ipotesi I_2 si sta considerando un vettore di v.c. congiuntamente gaussiane, con matrice di misura di rischio quadratico gaussiana, anche se ciò non è sempre vero.

L'ipotesi I_1 è identica alla I_2 con l'aggiunta della considerazione sulla d.d.p.

Teorema Posta l'ipotesi I_1 e posto $\left(\underset{I}{\Phi} = q\right)$, θ^M risulta distribuito gaussianamente:

$$\theta^M \sim N\left(\theta^0, \sigma^2 \left(\underset{I}{\Phi}^T \Psi^{-1} \underset{I}{\Phi}\right)^{-1}\right)$$

Dimostrazione

$$\theta^M = cY = c\underset{I}{\Phi}\,\theta^0 + cV$$

θ^M risulta una combinazione lineare di v.c. congiuntamente gaussiane e perciò è una v.c. gaussiana. Posto, inoltre, essendo V un vettore di v.c. congiuntamente gaussiane, $c\Phi\theta^0$ è c quantità (una matrice) determinabile.

Intervalli di confidenza per θ^0

1) σ^2 nota

Definiamo la matrice varianza di θ^M:

$$\Sigma_{\theta^M} := \sigma^2 \left(\underset{I}{\Phi}^T \Psi^{-1} \underset{I}{\Phi}\right)$$

e

$$\sigma^2_{\theta^M_i} := \left[\Sigma_{\theta^M}\right]_{ii}$$ (i-esimo elemento sulla diagonale principale di Σ_{θ^M})

Risulta

$$I_{0,95}(\vartheta_i^o) = [\vartheta_i^M - 1{,}96\,\sigma_{\vartheta_i^M} \; ; \; \vartheta_i^M + 1{,}96\,\sigma_{\vartheta_i^M}]$$

essendo $I_{0,95}(\vartheta_i^o)$ ~~essendo~~ l'intervallo di confidenza dell'elemento i-esimo di ϑ_i vero ai valore di affidabilità pari a 0,95.

Nota: non compare la $\sqrt{n}$ poiché non si sta considerando l'intervallo di confidenza della media campionaria.

Negli intervalli di confidenza degli stimatori gaussiani compare sempre la deviazione standard dello stimatore.

2) σ^2 incognita

Definiamo

$$\hat{\Sigma}_{\vartheta^M} := \hat{\sigma}^2 (\Phi^T \Psi^{-1} \Phi) \simeq Var[\vartheta^M]$$

L'"uguale circa" indica che è una stima.

$$\hat{\sigma}^2_{\vartheta_i^M} := [\hat{\Sigma}_{\vartheta_i^M}]_{ii}$$

Si dimostra che vale la seguente proprietà

$$\left(\frac{\vartheta_i^M - \vartheta_i^o}{\hat{\sigma}_{\vartheta_i^M}} \right) \sim t_{n-g}$$

La quantità entro parentesi è distribuita come una t di Student con $n-g$ gradi di libertà e rappresenta una pseudostandardizzazione.

Da ϑ_i^M si è tolto il valore vero e si è diviso per la deviazione standard stimata, non conoscendo quella vera.

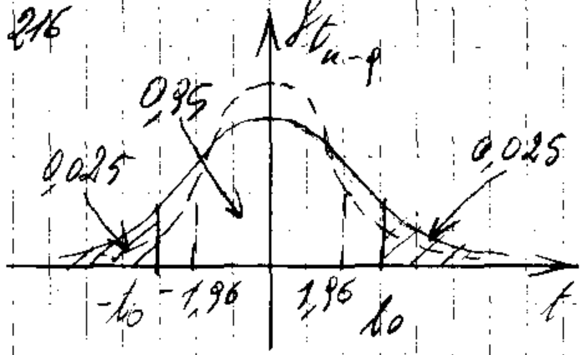
Risulta

$$I_{0,95}(\vartheta_i^o) = [\vartheta_i^M - t_0\,\hat{\sigma}_{\vartheta_i^M} \, , \, \vartheta_i^M + t_0\,\hat{\sigma}_{\vartheta_i^M}]$$

con t_0 che rappresenta

$$P(|t_{n-g}| \le t_0) = 0{,}95$$

Se $n \gg g$ la t_{n-g} diventa una gaussiana e t_0 diventa 1,96.



Nel grafico, a linea continua è rappresentata la t_{n-q}.
A linea tratteggiata appare la v.c. gaussiana standard.

<u>Applicazione: regressione lineare multipla</u>

È un modello del tipo

$$y(t) = \theta_1 u_1(t) + \theta_2 u_2(t) + \dots + \theta_q u_q(t) + v(t)$$

$$t = 1, 2, \dots, n$$

Ipotizziamo che

$$V \sim N(0, \sigma^2 I)$$

con

$$\Psi = I \quad \text{(pag. 208)}$$

e

$$V = [v(1)\ v(2)\ \dots\ v(n)]^T$$

Come esempio, possiamo immaginare di dover stimare il valore del PIL in funzione di alcuni parametri quali investimenti in grandi opere, per la ricerca scientifica, ecc. Ciascun parametro è moltiplicato per un peso diverso.

Si può studiare come cambia il PIL al variare di uno o più parametri.

Occorre scrivere le equazioni per tutti i casi

$$y(n) = \theta_1 u_1(n) + \theta_2 u_2(n) + \dots + \theta_q u_q(n) + v(n)$$

~~esempio~~

In forma matriciale, si ha

$$Y = \begin{bmatrix} u_1(1) & u_2(1) & \dots & u_q(1) \\ u_1(2) & u_2(2) & \dots & u_q(2) \\ \vdots & \vdots & & \vdots \\ u_1(n) & u_2(n) & \dots & u_q(n) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_q \end{bmatrix} + V$$

matrice Φ

Il modello è identificato attraverso i seguenti passi.

- *passo 1*: calcolo di θ^M (vettore delle stime)

$$\theta^M = (\Phi^T \Psi^{-1} \Phi)^{-1} \Phi^T \Psi^{-1} Y$$

- *passo 2*: riconoscendo che σ^2 non è conosciuto, si calcola $\hat{\sigma}^2$ (stima della varianza dell'errore, di misura)

$$\hat{\sigma}^2 = \frac{E^T \Psi^{-1} E}{n - q}$$

Se tutti gli errori hanno la stessa varianza e sono tra loro incorrelati, allora

$$\Psi = I \qquad \text{(pag. 208)}$$

e

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n - q}$$

$$\theta^M = (\Phi^T \Phi)^{-1} \Phi^T Y = \theta^{LS} \qquad \text{(pag. 194)}$$

- *passo 3* calcolo di $\hat{\Sigma}_{\theta^M}$ (matrice varianza-covarianza dei parametri stimati)

$$\hat{\Sigma}_{\theta^M} = \hat{\sigma}^2 (\Phi^T \Psi^{-1} \Phi)^{-1} =$$

$$= \hat{\sigma}^2 (\Phi^T \Phi) \qquad \text{con } \Psi = I$$

- *passo 4*: intervalli di confidenza

Sono i risultati sono presentati nel modo seguente:

$$y(t) = \theta_1^M u_1(t) + \theta_2^M u_2(t) + \dots + \theta_q^M u_q(t) + V(t)$$

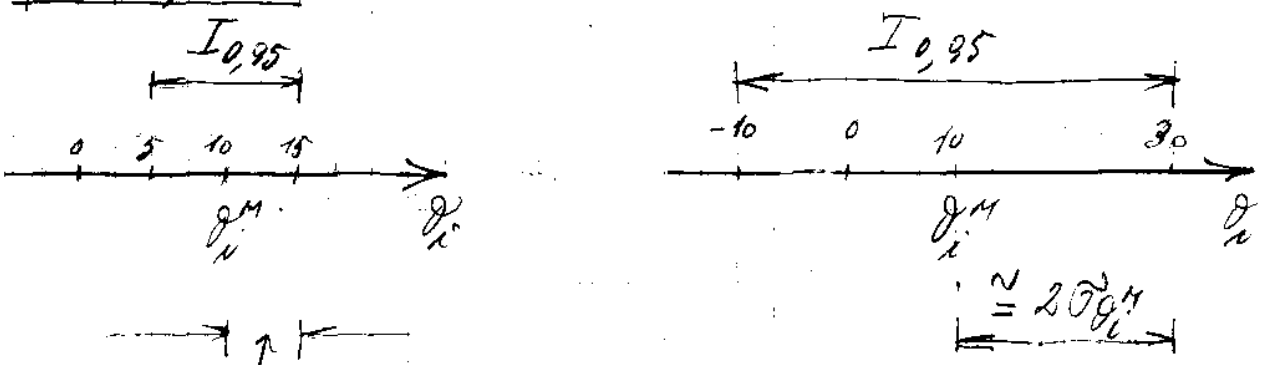
Al posto dei valori di θ si mettono i valori calcolati con la procedura di stima. Dopo di θ^M e posto al vettore si collocano le deviazioni standard

$$y(t) = \theta_1^M u_1(t) + \dots + V(t)$$

$$\left[\hat{\sigma}_{\theta_1^M}\right] \qquad \left[\hat{\sigma}\right]$$

218 Oltre al modello ed ai parametri stimati occorre fornire elementi che permettano di valutare il grado di affidabilità del risultato.

Osservazione



In entrambi i grafici ϑ_i^M è la stima del parametro.

Supponiamo che $\sigma_{\vartheta_i^M}$ sia conosciuto oppure che il numero dei dati ϑ_i^M sia molto grande.

Sotto questa ipotesi accade, a destra ed a sinistra di ϑ_i^M due intervalli pari a $\simeq 2\sigma_{\vartheta_i^M}$ mentre l'intervallo di confidenza del caso a sinistra ~~è~~ viene così non accade per l'intervallo a destra, nel quale l'intervallo di confidenza comprende lo zero.

Di conseguenza non si può dire niente sul parametro potendo anche questo positivo o negativo.

Regola pratica: l'intervallo di confidenza viene se

$$|\vartheta_i^M| > 2\sigma_{\vartheta_i^M}$$

Il valore di ogni parametro deve essere almeno il doppio della propria deviazione standard.

VALIDAZIONE E SCELTA DEL MODELLO

Contenuti:

- Validazione: Test χ^2
- Test F
- Crossvalidazione
- Final Prediction Error (FPE)
- Akaike Information Criterion (AIC)
- Minimum Description Length (MDL)
- Un esempio semplice
- Conclusioni

Motivazione: Una volta nota la struttura del modello (cioè $\Phi(U,\theta)$, è relativamente facile stimare θ. Problemi aperti:

(i) come capire se un dato modello è "buono";
(ii) come confrontare due o più modelli.

VALIDAZIONE: TEST χ^2

Problema: Avendo a disposizione N dati $Y_1,...,Y_N$, come faccio a sapere se il modello

$$Y = \Phi\theta^\circ + V, \quad Var[V] = \sigma^2 I$$

li descrive adeguatamente?

Nota 1, pag. 218

Idea: $\theta^{LS} \equiv \theta^\circ \Rightarrow \varepsilon = Y - \Phi\theta^{LS} \equiv V$ (il vettore dei residui è una "stima" del vettore degli errori di misura). Perciò mi aspetto che

$$\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i^2 \equiv \sigma^2$$

2/218

Se conosco σ^2 è bene controllare che σ^2 e la varianza campionaria dei residui abbiano lo stesso ordine di grandezza (idea ovvia: se ho errori di misura dell'ordine di 10^{-3} e i residui sono dell'ordine di 10^{-1}, il modello è sbagliato perché non riesce a spiegare i dati).

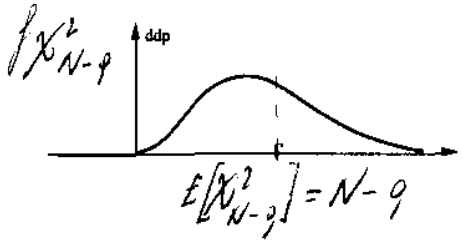
Problema: cosa significa "dello stesso ordine di grandezza"?

Ipotesi II: $Y = \Phi\theta^\circ + V, \quad V \sim N(0,\sigma^2 I)$

(Suppongo di aver azzeccato la struttura del "modello vero" che genera i dati e che gli errori siano gaussiani con varianza nota)

Teorema: Sotto l'Ipotesi II, dividendo per σ^2 la somma dei quadrati dei residui (SSR: Sum of Squared Residuals) della stima LS si ottiene una V.C. di tipo χ^2 "con N-q gradi di libertà" (N = n° di dati, q = n° di parametri):

$$\frac{\varepsilon^T\varepsilon}{\sigma^2} \sim \chi^2(N-q)$$



Nota: Risulta $E[\chi^2(N-q)] = N-q$, $Var[\chi^2(N-q)] = 2(N-q)$. Inoltre, per N-q "grande" la f.d.d. della V.C. $\chi^2(N-q)$ è circa gaussiana.

Idea: Utilizzando la f.d.d. del χ^2 ho un criterio numerico per valutare quando la SSR è "anormale" ($\Rightarrow$ modello sbagliato).

Tabella: Valori di $\chi^2_{\alpha,n}$

n	α=0.975	α=0.95	α=0.05	α=0.025
1	0.00	0.00	3.84	5.02
2	0.05	0.10	5.99	7.38
3	0.22	0.35	7.81	9.35
4	0.48	0.71	9.49	11.14
5	0.83	1.15	11.07	12.38
6	1.24	1.64	12.59	14.45
7	1.69	2.17	14.07	16.01
8	2.18	2.73	15.51	17.53
9	2.70	3.33	16.92	19.02
10	3.25	3.94	18.31	20.48
11	3.82	4.57	19.68	21.92
12	4.40	5.23	21.03	23.34
13	5.01	5.89	22.36	24.74
14	5.63	6.57	23.68	26.12
15	6.27	7.26	25.00	27.49
16	6.91	7.96	26.30	28.85
17	7.56	8.67	27.59	30.19
18	8.23	9.39	28.87	31.53
19	8.91	10.12	30.14	32.85
20	9.59	10.85	31.41	34.17
25	13.12	14.61	37.65	40.65
30	16.79	18.49	43.77	46.98
40	24.43	26.51	55.76	59.34
50	32.36	34.76	67.50	71.42
60	40.48	43.19	79.08	83.30
70	48.76	51.74	90.53	95.02
80	57015	60.39	101.88	106.63
90	65.65	69.13	113.14	118.14
100	74.22	77.93	124.34	129.56

3/218

Esempio: 10 dati, 3 parametri *(N-q = 10-3 = 7)*. Dalle tabelle della distribuzione χ^2 risulta $P(\chi^2(7) < 14.07) = 0.95 \Rightarrow$ nel 95% dei casi $\varepsilon^T\varepsilon/\sigma^2 < 14.07$.
Se accade che $\varepsilon^T\varepsilon/\sigma^2 > 14.07$, sospetto che il modello non sia buono (sotto l'ipotesi che sia buono, accade raramente che la somma dei quadrati dei residui sia così grande).

Test χ^2: Fissato il livello di significatività α (tipicamente, $\alpha = 0.05$) cerco sulle tabelle x_α tale che $P(\chi^2(N-q) < x_\alpha) = 0.95$. Poi adotto la seguente regola:

- $\frac{\varepsilon^T\varepsilon}{\sigma^2} < x_\alpha \Rightarrow$ accetto il modello
- $\frac{\varepsilon^T\varepsilon}{\sigma^2} > x_\alpha \Rightarrow$ respingo il modello

Estensione: $V \sim N(0,\Sigma_V)$, dove Σ_V è una matrice nota.
Basta usare $\varepsilon^T\Sigma_V^{-1}\varepsilon$ al posto di $\varepsilon^T\varepsilon/\sigma^2$.

Punti deboli:

- Può essere difficile capire quali sono i motivi per cui viene scartato il modello. Almeno 4 possibilità:

  a) la relazione $Y = \Phi\theta^\circ + V$ spiega male i dati;

  b) V non è gaussiano;

  c) il valore di σ^2 è sbagliato per difetto;

  d) gli errori di misura non hanno tutti la stessa varianza.

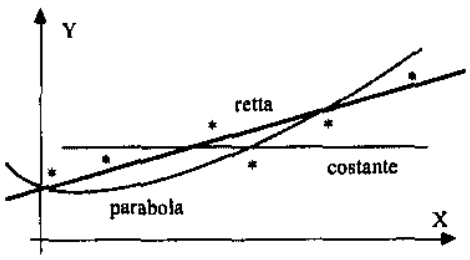
- Il test si basa sull'ipotesi che esista un "modello vero" di tipo lineare che genera i dati e che gli errori siano gaussiani (ipotesi molto semplificativa).

TEST *F*

Confronto tra modelli: Spesso, non è chiaro quale sia il modello "giusto" e i considera un ventaglio di possibilità. Come si fa a scegliere il modello "ottimo"?

Modelli "matrioska": Una sequenza di classi di modelli in cui ciascuna classe comprende al suo interno (come casi particolari) le classi precedenti.
Esempio: M_1 = {*rette*}, M_2 = {*parabole*}, M_3 = {*cubiche*}, ...

Problema tipico: Conosco N coppie (X_i,Y_i) e voglio identificare un modello $Y = f(X)$. Se mi limito ai polinomi di ordine k, cosa è meglio? Una retta, una parabola, una cubica, ...?



Idea (stupida): Considero i diversi modelli (retta, parabola, cubica, ...) e ne stimo i parametri mediante LS. Poi, tra i modelli stimati, scelgo quello che minimizza la SSR.

Fatto: Per i modelli matrioska, la SSR decresce sempre al crescere dell'ordine del modello (dato che la stima LS si basa sulla minimizzazione di SSR, non è possibile che la migliore parabola abbia una SSR maggiore di una retta)

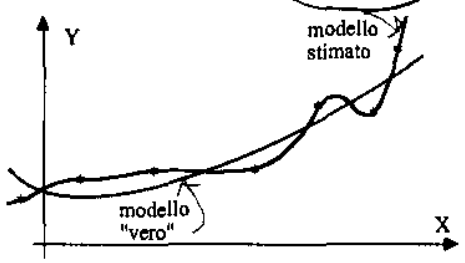
$\Downarrow$

Usare la minimizzazione di SSR per scegliere il modello ottimo conduce a scegliere sempre il modello più complesso (per assurdo, $N = 100 \Rightarrow$ polinomio di ordine 99).

Che male c'è ad eccedere con il n° di parametri?

- Nessun problema se i dati fossero privi di rumore.
  Esempio: stimo con una parabola dei dati che stanno su una retta ⇒ il coefficiente del termine quadratico risulta = *0* ⇒ non commetto errori.

- Se c'è rumore ed ho troppi parametri, il modello stimato tende ad essere influenzato dal rumore (riproduce oscillazioni che non hanno significato fisico ma che sono frutto degli errori di misura)



Principio di parsimonia: Non usare parametri addizionali per descrivere un fenomeno se essi non sono necessari.

Idea: Dati M_{k-1} e M_k, sappiamo che $SSR_k < SSR_{k-1}$. Sceglierò M_k solo se SSR_k è "molto più piccola" di SSR_{k-1}.

Problema: Cosa vuol dire "molto più piccola"?

Teorema: Sotto l'Ipotesi I1,

$$f = (N-k)\ \frac{SSR_{k-1} - SSR_k}{SSR_k}$$

è una V.C. distribuita come una F di Fisher con $(1,N-k)$ gradi di libertà

Osservazioni:

- f è un indice della riduzione % di SSR che ottengo passando dal modello M_{k-1} (meno complesso) al modello M_k (più complesso).

- Per $N-k$ "grande", si ha $F(1,N-k) = \chi^2(1)$

Test F: Fissato un livello di significatività α (tipicamente, $\alpha = 0.05$) cerco sulle tabelle f_α tale che $P(F(1,N-k) < f_\alpha) = 0.95$. Poi adotto la seguente regola:

- $f < f_\alpha \Rightarrow$ scelgo il modello M_{k-1}
- $f > f_\alpha \Rightarrow$ scelgo il modello M_k

Osservazioni:

- Non è necessario conoscere σ^2
- $V \sim N(0,\sigma^2\Psi)$, dove Ψ è una matrice nota ⇒ basta usare $\varepsilon^T\Psi^{-1}\varepsilon$ al posto di $SSR = \varepsilon^T\varepsilon$.
- L'esito del test dipende dalla scelta del livello di significatività α (α "piccolo": aumenta la probabilità di sottostimare l'ordine; α "grande": aumenta la probabilità di sovrastimare l'ordine)

Punti deboli:

- L'Ipotesi I1 è restrittiva. Esiste un "modello vero"? Ammesso che esista, appartiene alla classe di modelli considerati?.
- Il test si applica solo a classi di modelli "matrioska"

Approccio alternativo

Fatto: Supponiamo che valga I1 e che il modello vero $\in M_k$. Allora $\theta_j^\circ = 0 \Rightarrow \theta_j^{LS} \sim G(0,\sigma_\theta^2) \Rightarrow \theta_j^{LS}/\sigma_\theta \sim G(0,1) \Rightarrow P(|\theta_j^{LS}/\sigma_\theta| \le 1.96) = 0.95$.

⇓

Se $\theta_j^\circ = 0$, nel *95%* dei casi risulta $\theta_j^{LS} < 1.96\ \sigma_\theta$

Morale: Se un parametro stimato è maggiore del doppio del valore della sua *SD*, è verosimile che il parametro sia $\neq 0$. In caso contrario, può essere conveniente azzerare quel parametro.

E' bene diffidare dei modelli in cui i parametri stimati non sono almeno 2÷3 volte più grandi delle loro SD.

CROSSVALIDAZIONE

Idea: Se ho abbastanza dati, li divido in due gruppi:

1) dati di identificazione Y^I;
2) dati di validazione Y^V.

Uso il primo gruppo per identificare vari modelli (mediante *LS*, per es.). Poi uso il secondo gruppo per testare i modelli e capire quale è il migliore.

Procedura:

- Considero vari modelli (rette, parabole, cubiche, ...) e ne identifico i parametri mediante *LS*

  $\theta^{LS} = (\Phi^T\Phi)^{-1}\Phi^T Y^I$

- Con i modelli identificati cerco di prevedere i dati di validazione e calcolo i relativi residui

  $SSR^V = \varepsilon^{VT}\varepsilon^V, \quad \varepsilon^V = Y^V - \hat{Y}, \quad \hat{Y} = \Phi^V\theta^{LS}$

- Scelgo il modello che minimizza SSR^V

Osservazioni:

- L'assegnamento di un'osservazione al set di identificazione o a quello di ~~identificazione~~ VALIDAZIONE deve essere casuale.
- Se uso modelli matrioska, SSR^I decresce al crescere dell'ordine. All'inizio anche SSR^V decresce. Ad un certo punto i parametri diventano troppi ed il modello identificato cerca di seguire troppo fedelmente i dati di identificazione $\Rightarrow$ SSR^V comincia a salire.
- Talvolta la crossvalidazione suggerisce l'uso di modelli in cui alcuni parametri hanno *SD* elevata. Vale la pena di azzerare il valore di questi parametri e ricalcolare SSR^V. Se SSR^V aumenta di poco *(1÷2%)* rispetto al minimo può essere conveniente adottare il modello semplificato.
- Non faccio nessuna ipotesi sul meccanismo "vero" di generazione dei dati. Non pretendo di trovare il modello "vero", ma solo il modello "migliore" (in termini di ~~SSR~~) entro una certa rosa di possibilità.
- Limitazione fondamentale: bisogna avere parecchi dati altrimenti sia l'identificazione che la validazione diventano poco affidabili.

E se non ho abbastanza dati per formare due gruppi?

Ordinary Cross Validation (OCV): Metto da parte il dato i-esimo e calcolo $\theta^{LS(i)}$ usando tutti i dati rimanenti. Poi calcolo l'errore che commetto cercando di indovinare Y_i usando $\theta^{LS(i)}$

$\varepsilon^{(i)} = Y_i - \Phi^{(i)}\theta^{LS(i)}$ ($\Phi^{(i)}$: i-esima riga di Φ)

Ripeto la procedura per $i = 1,...,N$ e uso come indice di bontà del modello:

$$OCV = \frac{1}{N}\sum_{i=1}^{N} \varepsilon^{(i)2}$$

Problema: Sembra necessario risolvere N problemi di stima LS (computazionalmente oneroso).

Lemma del "lasciane-uno-fuori" ("leave-out-one" lemma):

$$OCV = \frac{1}{N}\sum_{i=1}^{N} \frac{\varepsilon_i^2}{(1 - H_{ii})^2}$$

$H = \Phi(\Phi^T\Phi)^{-1}\Phi^T$

$\varepsilon = Y - \Phi\theta^{LS}$

Osservazioni:

- Grazie al "leave-out-one lemma" basta risolvere una sola stima *LS* e calcolare gli elementi sulla diag. principale di H.
- L'uso di *OCV* è in genere più oneroso che crossvalidare dividendo i dati in due gruppi (nel calcolo di θ^{LS} in genere si evita di calcolare esplicitamente $(\Phi^T\Phi)^{-1}$, che è invece richiesta da *OCV*).
- Punto debole: quando gli errori di misura non hanno tutti la stessa varianza.
- Per ridurre i calcoli, si può ricorrere ad una approssimazione di *OCV(Generalized Cross Validation)*:

$$GCV = \frac{1}{N}\frac{\sum_{i=1}^{N}\varepsilon_i^2}{\left[\frac{1}{N}Tr(I-H)\right]^2} = \frac{1}{N}\frac{\sum_{i=1}^{N}\varepsilon_i^2}{\left[\frac{N-q}{N}\right]^2} = \frac{N}{(N-q)^2}SSR$$

FINAL PREDICTION ERROR (FPE)

Idea: Se faccio delle ipotesi sul meccanismo di generazione dei dati posso cercare di minimizzare SSR^V senza doverla calcolare esplicitamente.

Ipotesi I2: $Y = \Phi\theta^o + V$, $E[V] = 0$, $Var[V] = \sigma^2 I$.
(rispetto a I1, non faccio ipotesi sulla gaussianità del rumore)

Supponiamo di considerare un vettore θ e di sottoporlo a validazione. Ipotizzando $\Phi^V = \Phi^I$, si dimostra che, se estraggo a caso un campione Y^V di N dati di validazione (estrazione #1):

$$E[SSR^V] = N\sigma^2 + (\theta-\theta^o)^T\Phi^T\Phi(\theta-\theta^o)$$

Osservazione (ovvia): $E[SSR^V]$ è minimizzata da $\theta = \theta^o$.

Supponiamo ora che $\theta = \theta^{LS}$ sia la stima ottenuta da un campione estratto a caso di N dati Y^I di identificazione (estrazione #2). Si dimostra che

$$E[E[SSR^V]] = \sigma^2(N+q)\ ,\ q = dim(\theta)$$

(ho due medie perché ho due estrazioni casuali)

E' la formulazione matematica del principio di parsimonia: se il modello vero è una retta ($q = 2$) e per identificare uso una parabola o una cubica ($q = 3$, $q = 4$) peggioro inutilmente le prestazioni (medie) in validazione del mio modello.

In molti casi σ^2 non è nota. Uno stimatore non polarizzato di σ^2 è fornito da:

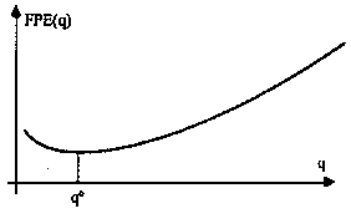
$$\hat{\sigma}^2 = \frac{SSR}{N-q}$$

Criterio FPE: Tra diversi modelli matrioska caratterizzati dal numero q di parametri, scelgo quello che minimizza la media (stimata) della SSR^V, ovvero il cosiddetto *Final Prediction Error*

$$FPE = \frac{N+q}{N-q} SSR$$

Osservazioni:

- Al crescere di q dapprima FPE diminuisce perché diminuisce SSR. Se però q cresce troppo, FPE cresce (c'è $N-q$ al denominatore) $\Rightarrow \exists$ un minimo di FPE al variare di q.
- Computazionalmente efficiente: per calcolare FPE basta solo conoscere θ^{LS}.



Confronto criteri soggettivi/oggettivi

Criteri soggettivi: quelli basati sui test statistici (vedi test F) che richiedono la scelta "soggettiva" del livello di significatività α.

Criteri oggettivi: quelli basati sulla minimizzazione di una cifra di merito (*OCV, GCV, FPE, AIC, MDL*). Non c'è da scegliere nessun livello di significatività.

La differenza è meno profonda di quanto sembri.

Fatto (di facile ma noiosa dimostrazione): Per N "grande", scegliere tra M_k e M_{k+1} basandosi su FPE equivale ad usare il Test F con $\alpha = 0.157$

⇓

Per $N \to \infty$, FPE ha il 15.7% di probabilità di scegliere (erroneamente) il modello più complicato anche se quello giusto è quello più semplice.

⇓

Per $N \to \infty$, FPE sovrastima (in media) l'ordine del modello (non è uno stimatore consistente dell'ordine del modello)

AKAIKE INFORMATION CRITERION (AIC)

Criterio ricavato in base alla "distanza" tra la d.d.p. vera dei dati e quella generata da un dato modello stimato (vale sotto l'Ipotesi I1).

Criterio AIC: Tra diversi modelli matrioska caratterizzati dal numero q di parametri, scelgo il modello che minimizza

$$AIC = \frac{2q}{N} + ln(SSR)$$

Osservazione: Per $N \to \infty$ si dimostra che FPE e AIC sono equivalenti (infatti, si vede che $lim_{N\to\infty} ln\ FPE = AIC$).

⇓

Per $N \to \infty$, anche AIC sovrastima (in media) l'ordine del modello

MINIMUM DESCRIPTION LENGTH (MDL)

Idea: Immagino di dover trasmettere i dati. Invece di trasmettere il vettore Y, posso trasmettere θ^{LS} e il vettore dei residui ε. Il ricevitore potrà ricostruire i dati calcolando $Y = \Phi\theta^{LS}+\varepsilon$.

32/224

Vantaggio: se il modello è buono, l'errore di predizione è piccolo e, per una data precisione, bastano pochi bit per codificarlo.

Se l'ordine q del modello aumenta i residui diventano più piccoli (occorrono meno bit per ε) ma aumenta il n° dei parametri da codificare (occorrono più bit per θ^{LS}).

Criterio Minimum Description Length: scelgo il modello che conduce alla codifica più compatta. Si dimostra che (sotto I1) ciò equivale a minimizzare la cifra di merito

$$MDL = \frac{ln(N)\ q}{N} + ln(SSR)$$

33/224

Osservazione: La penalità su q è più pesante che in AIC. Infatti MDL conduce a modelli più parsimoniosi. Anzi si dimostra che MDL è uno stimatore *consistente* dell'ordine del modello (per $N \to \infty$ l'ordine indicato da MDL converge all'ordine vero).

UN ESEMPIO SEMPLICE

Esempio tratto da (J.V. Beck e K.J. Arnold, "Parameter Estimation in Engineering and Science, Wiley 1977).

La conducibilità termica k di alcuni campioni di ferro è stata misurata a diverse temperature T (°F). I risultati sperimentali sono riportati nella seguente tabella.

T (°F)	100	161	227	270	382	90	148	208	247	352
k:	41.5	37.7875	35.4975	35.7854.63		42.545	39.6375	37.552866.36		33.815

Le prime cinque misure sono state prese in condizioni sperimentali diverse rispetto alle ultime cinque. In particolare, si sa che la varianza dell'errore di misura per i secondi cinque dati è quattro volte maggiore della varianza dell'errore per i primi cinque.
Ci si pone l'obiettivo di identificare un modello che descriva la dipendenza di k nei confronti della temperatura. Si considerano i seguenti modelli:

1. $k = \theta_1$
2. $k = \theta_1 + \theta_2 T$
3. $k = \theta_1 + \theta_2 T + \theta_3 T^2$
4. $k = \theta_1 + \theta_2 T + \theta_3 T^2 + \theta_4 T^3$
5. $k = \theta_1 + \theta_2 T + \theta_3 T^2 + \theta_4 T^3 + \theta_5 T^4$

Problema #1: Stima dei parametri

Soluzione: Minimi quadrati pondorati (WLS) (alcuni dati sono più affidabili di altri)

$$\theta = (\Phi^T Q \Phi)^{-1} \Phi^T Q Y$$

34/224

Vettore dei dati e delle variabili indipendenti:

$$Y = \begin{bmatrix} k(1) \\ k(2) \\ \dots \\ k(10) \end{bmatrix}, \quad U = \begin{bmatrix} T(1) \\ T(2) \\ \dots \\ T(10) \end{bmatrix}$$

Matrice $\Phi(U)$ e vettore θ nei 5 modelli:

35/224

1. $\Phi = \begin{bmatrix} 1 \\ 1 \\ \dots \\ 1 \end{bmatrix}$, $\theta = [\ \theta_1\ ]$

2. $\Phi = \begin{bmatrix} 1 & T(1) \\ 1 & T(2) \\ \dots & \dots \\ 1 & T(10) \end{bmatrix}$, $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$

3. $\Phi = \begin{bmatrix} 1 & T(1) & T(1)^2 \\ 1 & T(2) & T(2)^2 \\ \dots & \dots & \dots \\ 1 & T(10) & T(10)^2 \end{bmatrix}$, $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$

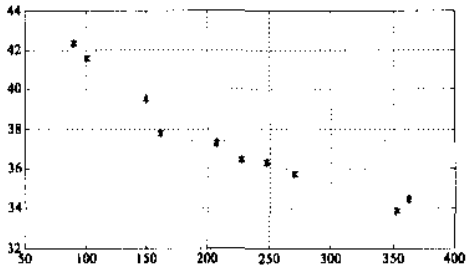
4. $\Phi = \begin{bmatrix} 1 & T(1) & T(1)^2 & T(1)^3 \\ 1 & T(2) & T(2)^2 & T(2)^3 \\ \dots & \dots & \dots & \dots \\ 1 & T(10) & T(10)^2 & T(10)^3 \end{bmatrix}$, $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}$

5. $\Phi = \begin{bmatrix} 1 & T(1) & T(1)^2 & T(1)^3 & T(1)^4 \\ 1 & T(2) & T(2)^2 & T(2)^3 & T(2)^4 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & T(10) & T(10)^2 & T(10)^3 & T(10)^4 \end{bmatrix}$, $\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \end{bmatrix}$

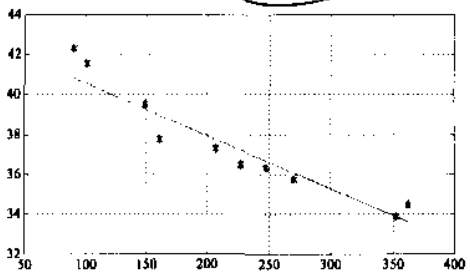
Matrice Q di WLS

$$Q = diag(\ 1\ \ 1\ \ 1\ \ 1\ \ 1\ \ 1/4\ \ 1/4\ \ 1/4\ \ 1/4\ \ 1/4\ )$$

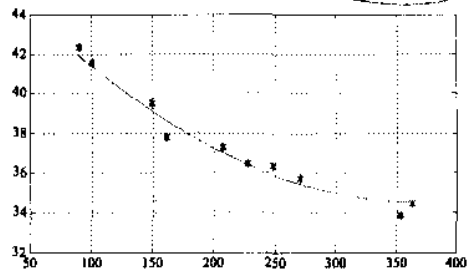
```
theta1 =
   3.7372e+01
sigmatheta1 =
   8.4336e-01
```



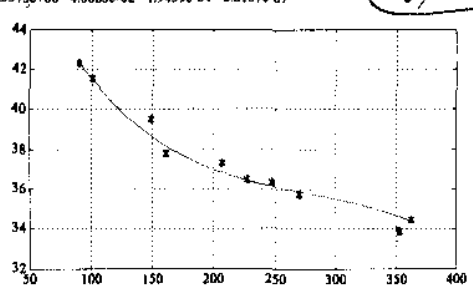
```
theta2 =
   4.3225e+01  -2.6486e-02
sigmatheta2 =
   7.9222e-01   3.3203e-03
```



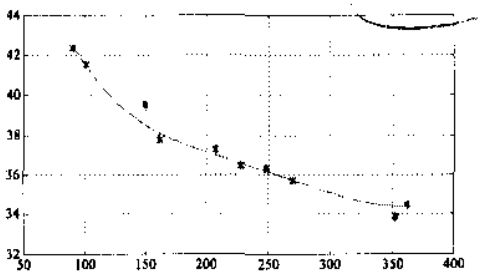

```
theta3 =
   4.7519e+01  -7.0826e-02   9.6665e-05
sigmatheta3 =
   1.0335e+00   9.8890e-03   2.1206e-05
```



```
theta4 =
   5.3391e+01  -1.6786e-01   5.6508e-04  -6.8092e-07
sigmatheta4 =
   2.5573e+00   4.0883e-02   1.9459e-04   2.8187e-07
```




```
theta5 =
   6.2542e+01  -3.7495e-01   2.1788e-03  -5.8582e-06   5.8511e-09
sigmatheta5 =
   1.0209e+01   2.2728e-01   1.7526e-03   5.6051e-06   6.3143e-09
```



Problema #2: Scelta del modello ottimo

Osservazioni:

- Come previsto, all'aumentare dell'ordine del modello si ottiene una maggior aderenza ai dati sperimentali. Caso limite: con dieci parametri (polinomio di ordine 9) posso interpolare i dati.
- Nei modelli 1-3 le deviazioni standard delle stime dei parametri sono accettabilmente inferiori ai valori dei parametri stimati.
- Nel modello 4 la SD di θ_4^{ML} è di poco superiore al 40% del valore stimato del parametro. Nel modello 5 la SD di θ_4^{ML} è circa uguale a θ_4^{ML} e lo stesso accade per la SD di θ_5^{ML}. Ciò potrebbe significare che il coefficiente del termine cubico (e a maggior ragione quello del termine di quarto grado) non è significativamente diverso da zero.

Test F:

Per applicare il test F è utile costruire la seguente tabella, in cui $f = (N-q)\Delta SSR/SSR$ mentre $F_{95}(1,N-q)$ indica il valore (ricavato dalla tabella della F di Fisher) in corrispondenza del quale la funzione di distribuzione della F a $(1,N-q)$ gradi di libertà vale 0.95:

Mod.	*q*	*N-q*	*SSR*	*ΔSSR*	*f*	$F_{95}(1,N-q)$
1	1	9	40.0078			
2	2	8	4.4680	35.5398	63.6341	5.32
3	3	7	1.1259	3.3421	20.7786	5.59
4	4	6	0.5708	0.5551	5.8356	5.99
5	5	5	0.4871	0.0837	0.8587	6.61

Confrontando la penultima e l'ultima colonna si vede che il modello *2* è significativamente migliore (in termini di riduzione dello scarto quadratico) del modello *1*. Lo stesso vale per il modello *3* nei confronti del modello *2*. Per il modello *4* si ha $f < F_{95}(1,N-q)$ e pertanto la riduzione di scarto quadratico non è statisticamente significativa. Vista la vicinanza dei valori di f e $F_{95}(1,N-q)$ è tuttavia opportuno non scartare a priori il modello *4*. Infine, passando dal modello *4* al modello *5* la riduzione dello scarto quadratico è palesemente non significativa, cosicché il modello *5* è da scartare.

Criteri FPE, AIC, MDL:

Mod.	*q*	*FPE*	*AIC*	*MDL*
1	1	48.8984	3.8891	3.9193
2	2	6.7020	1.8969	1.9575
3	3	2.0910	0.7186	0.8094
4	4	1.3318	0.2392	0.3603
5	5	1.4613	0.2807	0.4320

E' interessante notare che, in contrasto con il test F, tutti e tre i criteri "oggettivi" suggeriscono la scelta del modello *4*. In conclusione, anche a causa dei pochi dati disponibili, è difficile scegliere con sicurezza tra il modello *3* e *4*. Tuttavia, tenendo in considerazione le deviazioni standard dei parametri stimati, potrebbe essere più prudente orientarsi verso il modello *3*.

CONCLUSIONI

- L'unico metodo che non richiede ipotesi pesanti è la crossvalidazione.
- Nella pratica, i criteri *FPE, AIC, MDL* vengono usati senza preoccuparsi troppo delle ipotesi.
- *MDL* è meglio di *FPE* e *AIC*, ma solo asintoticamente.
- Suggerimento: usare più di un criterio. Anche l'esame delle *SD* dei parametri è importante e può aiutare a dirimere eventuali discordanze tra i criteri.
- Nella pratica non è essenziale trovare il miglior modello ma spesso basta un buon modello

VALIDAZIONE E SCELTA DEL MODELLO

(cose e commenti degli appunti disponibili in rete)

(1) Si possono le seguenti ipotesi:
- I_1 (pag 216)
- $\Psi = I \Rightarrow \theta^M = \theta^{LS}$ (la stima di Markov coincide con la stima LS)

(2) Immaginiamo di aver stimato la stima $(\theta^{LS} \simeq \theta^0)$.
In questa ipotesi il vettore dei residui $(Y - \hat{Y}\theta^{LS})$ coincide praticamente con il vettore degli errori di misura.
Il vettore dei residui è considerato una stima del vettore degli errori di misura.
Conoscendo la varianza σ^2 degli errori di misura, questa è uguale alla varianza dei residui da cui stimo dividendo per n la somma dei quadrati.

(3) Le righe indicano i gradi di libertà.
Se $P = 0{,}95$ si cerca nella colonna $\alpha = 0{,}05$ $(1-P)$

(4) Esiste la possibilità di scegliere un modello grezzo, e in questo caso, uno vale il 5%.
Questo criterio di validazione molto semplice, richiede di calcolare σ^2, cioè le caratteristiche degli strumenti di misura.

(5) Ψ non è la matrice identità I.

(6) Dato un polinomio di grado 2, di rappresenta una parabola annullando il coefficiente di 2° grado la parabola degenera in una retta.

(7) Scegliendo un modello la cui funzione passa vicino ai dati sperimentali, si cercano i valori dei parametri che minimizzano la somma dei quadrati dei residui.
Si può pensare di usare la stessa criterio per valutare quale è il modello migliore, scegliendo di questo non quello che passa vicino ai dati sperimentali.

(8) Se si dovesse usare come criterio quello dei minimi quadrati, si finirebbe sempre per scegliere il modello più complicato e si sceglierebbe un modello che passa esattamente sopra i dati.

220 (9) Si suol dire che il modello migliora il precedente.

(10) k è la complessità del modello.
Si può adottare il criterio di analizzare quanto scende la somma dei quadrati dei residui e scegliere il modello quando la discesa scende molto. Ciò indica che il modello è migliore del precedente.
Se la discesa è molto piccola essa può essere dovuta al permanere dei trend migliorato.

(11) $$\frac{SSR_{k-1} - SSR_k}{SSR_k}$$ rappresenta la differenza relativa della somma dei quadrati dei residui.

La F di Fisher è una V.C. che ha due gradi di libertà (parametri).
La cifra 1 indica che si sta confrontando un modello con quello immediatamente più complicato.

(12) Quando si hanno tanti dati la F di Fisher tende ad essere una χ^2 ad un grado di libertà.

(13) Dopo aver individuato sulla tabella il valore f_α, lo si confronta con il valore di f
$$f = (N-k)\,\frac{SSR_{k-1} - SSR_k}{SSR_k}$$
Se $f < f_\alpha$ si sceglie il modello più semplice, se invece c'è un miglioramento netto sulla somma dei quadrati dei residui si adotta il modello più complicato.

(14) È il caso in cui $\Psi \neq I$

(15) È un approccio alternativo equivalente al Test F.
Nel caso si stia decidendo tra una retta ed una parabola, anziché analizzare la F di Fisher si considera il coefficiente del termine di secondo grado.
Nel caso prevalga la retta il coefficiente deve essere nullo, altrimenti esso è diverso da zero.
Si analizza l'intervallo di confidenza del coefficiente e si guarda se in esso è compreso lo zero (pag. 218). In questa ipotesi il modello più corretto è la retta. Se l'intervallo non comprende lo zero, il coefficiente non è nullo e si sceglie il

Teoricamente, se il valore vero di un parametro è pari a zero, allora la stima è distribuita gaussianamente con una determinata varianza σ_ϑ^2.

Inoltre il valore del parametro diviso per la sua deviazione standard

$$\vartheta^{45} / \sigma_\vartheta$$

diventa una gaussiana standard.

$\vartheta^{45}/\sigma_\vartheta$ rappresenta la standardizzazione avendo sul lo il valor medio di ϑ^{45}.

Avendo standardizzato abbiamo ottenuto una gaussiana standard e, nel 95% dei casi, che

$$\left| \frac{\vartheta^{45}}{\sigma_\vartheta} \right| < 1{,}96$$

(16) si conclude che, se un parametro vale zero, nel 95% dei casi la sua stima è minore di circa due volte la sua deviazione standard. Si veda anche quanto detto a pag. 218.

(17) Si calcola $\hat{Y} = \Phi^V \vartheta^{45}$ che rappresenta la predizione dei dati di validazione.

Si calcolano poi i residui di validazione $E^V = Y^V - \hat{Y}$

successivamente si calcola la somma dei quadrati dei residui dei dati di validazione (SSR^V), cercando di minimizzarla.

(18) Occorre evitare che i due gruppi di dati abbiano caratteristiche diverse come ad esempio che appartengano a due intervalli temporali non distinti.

(19) Se si cerca di prevedere i dati di validazione, cercando il numero ottimo tra i due gruppi di dati di cui disponiamo in validazione è quello in cui si ha il minimo di SSR^V all'interno dei modelli incrementali.

Quando il modello comincia ad imparare il rumore, SSR^V comincia a salire.

(20) non è stata seguita alcuna ipotesi statistica e si è scelto il modello che funziona meglio nel senso che è in grado di prevedere i dati.

(21) Mettendo da parte un dato si identifica il modello

222 e si cerca di prevedere il dato successivo, calcolando il residuo.

Si ripete l'operazione su tutti i dati.

(22) Il risultato che si otterrebbe con la procedura cubica si può ottenere con una unica procedura.

Si risolve il problema dell'identificazione con tutti i dati calcolando la somma dei quadrati dei residui "pesata". Si ottiene lo stesso risultato che si otterrebbe con la procedura cubica.

Si calcola matrice H e si prendono gli elementi della diagonale principale.

(23) Nel calcolo di GCV si ipotizza che tutte le misure abbiano la stessa varianza.

(24)
$$GCV = \frac{N}{(N-p)^2} SSR,$$

con
$$SSR = E^T E$$ (somma dei quadrati dei residui)

Per ogni modello possibile candidato alla scelta si calcola il GCV e si sceglie il modello con il GCV più basso.

(25) Permette di effettuare la CROSSVALIDAZIONE, avente i dati

Con pochi dati, formulando alcune ipotesi statistiche si può prevedere il risultato che si otterrebbe ottenuto con la convalidazione.

Si può usare la somma dei quadrati dei residui in validazione senza calcolarla

(26) L'ipotesi che la matrice I^V per i dati di validazione sia uguale alla matrice I^I calcolata sui dati di identificazione

$$I^V = I^I$$

Ciò significa che i dati di validazione e quelli di identificazione sono sulla stessa scala (vedi le pagg. 191 e 192).

Si può ipotizzare che per ogni k si abbia un dato di validazione ed uno di identificazione.

Operando la validazione per il vettore θ il risultato è una v.c., avendo i dati di validazione un errore di misura.

(27) Il vettore θ citato nella nota precedente è il risultato di una procedura di identificazione basata su dati Y^I aventi errori casuali perciò differenti dagli errori di misura.

Ci sono due diverse sorgenti casuali: una genera i dati di validazione ed una quelli di identificazione

Prima si estraggono dei dati di identificazione (casuali) da cui si calcola θ_2 che viene poi applicato ai dati di validazione; pure casuali.

Si calcola il doppio valore atteso poiché ci sono due eventi casuali.

Ordinariamente la somma dei quadrati dei residui in validazione vale

$$E\left[E\left[SSR^V\right]\right] = \sigma^2 (N + q)$$

con $q = \dim(\theta)$.

(28) Non conviene aumentare q poiché aspettiamo gli errori in validazione. Occorre prendere il valore di q più piccolo.

Vale l'ipotesi I2, secondo la quale i dati sono

(24) generato da un determinato θ^* che ha un certo valore di ρ (ad esempio, un modello parabolico medio di $\rho = 3$).

Vuole ampliare il valore di ρ più piccolo che non renda nullo il modello test.

(29) Al posto di θ^2 si considera la sua stima.

(30) È il risultato della convalidazione incrociata che la sua stima.

(31) Si rappresenta la probabilità di errore per il modello più complicato, quando il modello più semplice è quello corretto.

(32) Si basa sulla teoria della codifica ottimale di Huffman.

Se i dati non sono del tutto casuali ed hanno una certa struttura, si può ridurre la lunghezza del codice, quindi l'idea è assegnare ai risultati più frequenti codici più brevi e considerare tutti i possibili dati.

(33) Rispetto al criterio AIC aggiunge la $\log(n)$ al posto della costante 2. Perciò, vengono generalmente scelti modelli più semplici.

(34) Lo stimatore considerato coincide con quello di Herbert della formula presentata Q coincide con γ^{-1}.

γ è il vettore delle covariabili osservate.

β è il vettore delle temperature

(35) È un esempio di modello "matrioska", ogni volta che si procede si aggiunge una colonna.

(36) Il modello "costante" è palesemente errato, anche se il valore di θ (theta) è almeno il doppio della sua deviazione standard (in matematica).

(37) Il modello "retta" appare graficamente accordo e la condizione di pag. 209 è soddisfatta.

(38) Il modello "parabola" appare più aderente ai dati e la condizione è ancora soddisfatta.

(39) Anche il modello "cubica" è accettato con la condizione relativa ai parametri ancora rispettata anche se si sta avvicinando al limite.

(40) Semplicemente il modello migliore poiché sono quasi

sempre sopra i dati; anche se sembra aver impara- 225
to il tenore, i dati, a parte il primo, non
rispettano la condizione richiesta.

Il modello non è buono.

(41) Per usare il test F conviene costruire una tabella.

Nella prima colonna è indicato il modello, nella se-
conda il numero di parametri, nella terza i dati
meno la complessità del modello, nella quarta la
somma dei quadrati dei residui.

Nella quinta colonna compare l'incremento della som-
ma dei quadrati dei residui. In particolare il
primo dato (35.5388) indica la differenza tra
40.0078 e 4.4688 e rappresenta il miglioramento
di un modello rispetto al precedente.

Quando la differenza è grande conviene adottare
il modello più complesso.

Se il dato si appiattisce non conviene passare al mo-
dello successivo.

f rappresenta l'indicatore per cui si prende la deci-
sione (nota 11, pag. 220).

Nell'ultima colonna compare la F di Fisher a
(1, 8 gradi di libertà).

Occorre confrontare i valori delle ultime due colonne.

Se il valore di f è maggiore di quello dell'ultima co-
lonna si passa al modello più complesso.

La prima discriminatoria riguarda i modelli 2 e 1,
confrontando 63,6341 e 5,32: si sceglie il model-
lo 2.

Nel confronto tra modello 3 e 2 vince il primo (20,7...
contro 5,59).

Tra il modello 3 e 4 c'è ~~una so~~ un sostanziale e-
quilibrio.

In definitiva il modello vincente è il modello 3,
che riguarda la parabola.

(42) Occorre scegliere il valore minimo che, per tutti e
tre i test, riguarda il modello 4.

Vi è una discrepanza: la F di Fisher indica il mo-
dello 3 come vincente, i criteri oggettivi fanno, in-
vece, preferire il modello 4.