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 $Y_1,\ldots,Y_N$, come faccio a sapere se il modello

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

li descrive adeguatamente?

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

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

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).

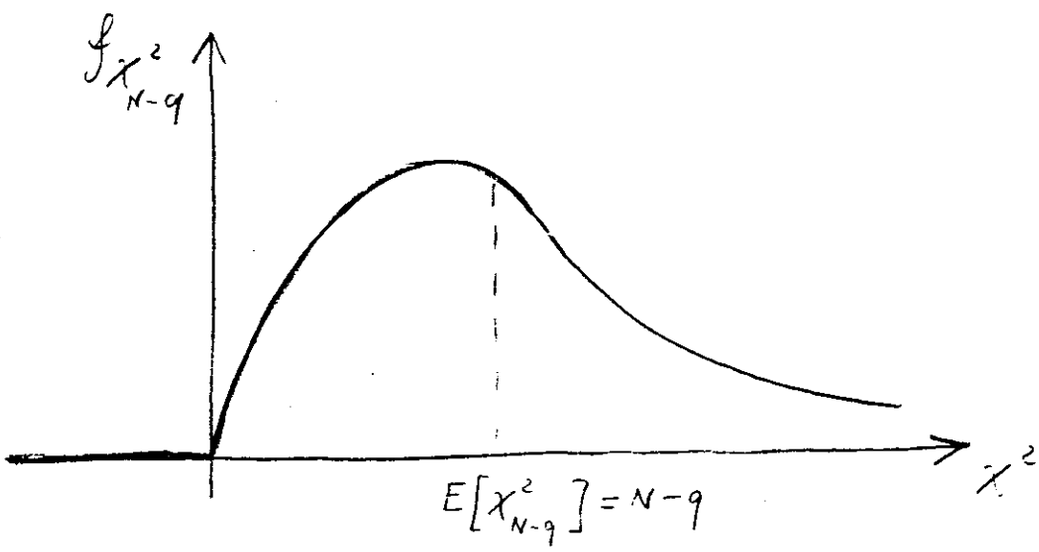
Problema: cosa significa "dello stesso ordine di grandezza"?

Ipotesi I1: $Y = \Phi\theta^\circ + V\ ,\quad V \sim N(0,\sigma^2 I)$

(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 I1, 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).

Appendix Table 3 Quantiles of the d.f. of χ^2

(Reproduced from Table III of Sir Ronald Fisher's *Statistical Methods for Research Workers*, Oliver and Boyd Ltd., Edinburgh, by kind permission of the author and publishers)

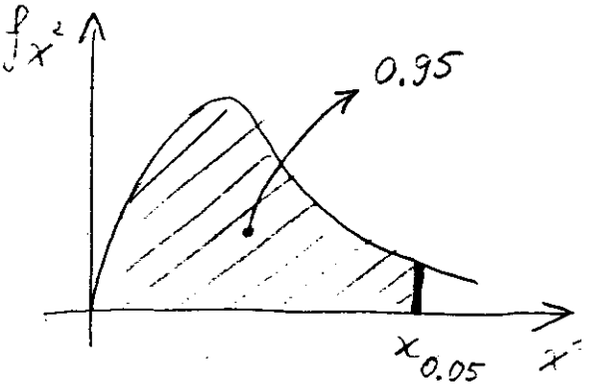
$P = 1 - F$	0·99	0·98	0·95	0·90	0·80	0·70	0·50	0·30	0·20	0·10	0·05	0·02	0·01
$\nu = 1$	$0{\cdot}0^3157$	$0{\cdot}0^3628$	$0{\cdot}0^2393$	0·0158	0·0642	0·148	0·455	1·074	1·642	2·706	3·841	5·412	6·635
2	0·0201	0·0404	0·103	0·211	0·446	0·713	1·386	2·408	3·219	4·605	5·991	7·824	9·210
3	0·115	0·185	0·352	0·584	1·005	1·424	2·366	3·665	4·642	6·251	7·815	9·837	11·345
4	0·297	0·429	0·711	1·064	1·649	2·195	3·357	4·878	5·989	7·779	9·488	11·668	13·277
5	0·554	0·752	1·145	1·160	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086
6	0·872	1·134	1·635	2·204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475
8	1·646	2·032	2·733	3·490	4·594	5·527	7·344	9·524	11·030	13·362	15·507	18·168	20·090
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209
11	3·053	3·609	4·575	5·578	6·989	8·148	10·341	12·899	14·631	17·275	19·675	22·618	24·725
12	3·571	4·178	5·226	6·304	7·807	9·034	11·340	14·011	15·821	18·549	21·026	24·054	26·217
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24·996	28·259	30·578
16	5·812	6·614	7·962	9·312	11·152	12·624	15·338	18·418	20·465	23·542	26·296	29·633	32·000
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22·775	25·038	28·412	31·410	35·020	37·566
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·171	29·615	32·671	36·343	38·932
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·652	41·566	44·314
26	12·198	13·409	15·379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588
30	14·953	16·306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892

Note—For values of ν greater than 30 the quantity $\sqrt{(2\chi^2)}$ may be taken to be distributed normally about mean $\sqrt{(2\nu - 1)}$ with unit variance.

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:

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

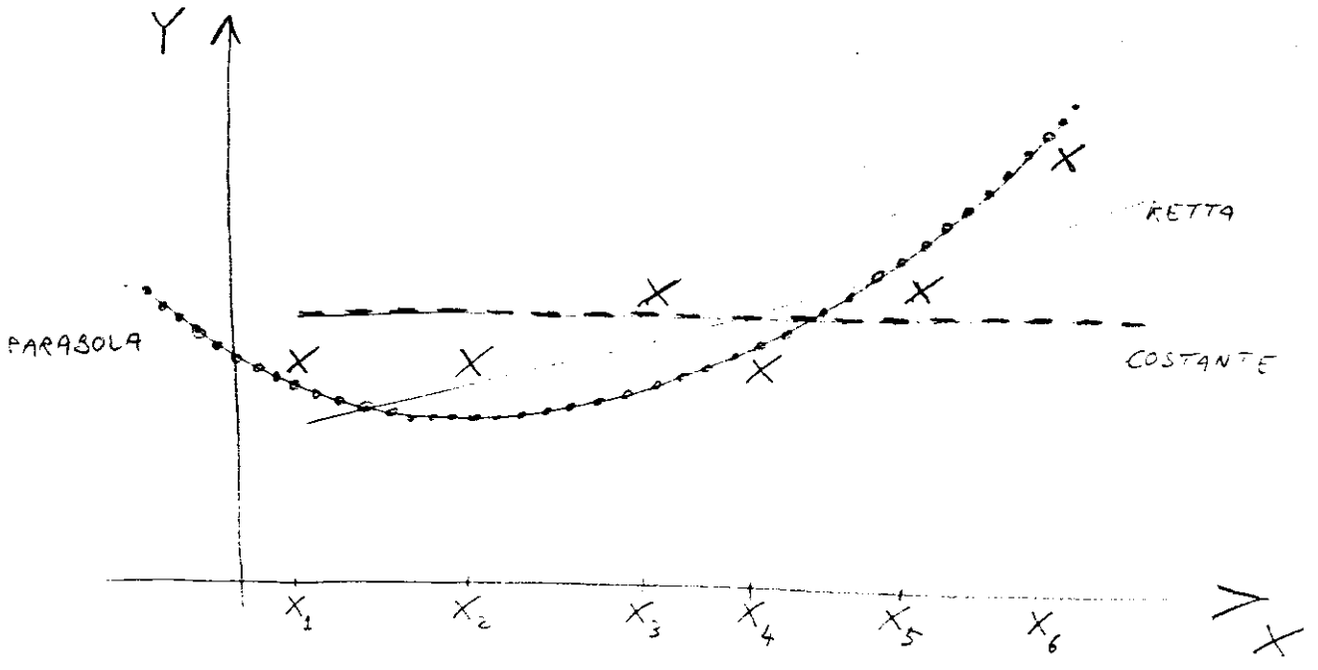
- Il test si basa sull'ipotesi che esista un "modello vero" di tipo lineare che genera i dati e che gli errori siano gaussiani (ipotesi molto semplificativa).

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.

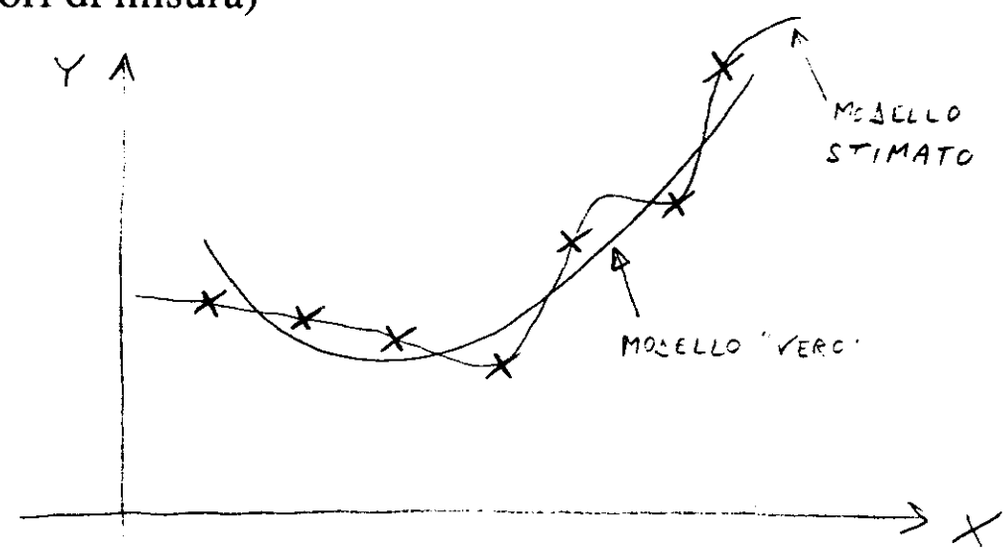
Fatto: Per i modelli matrioska, la SSR decresce sempre al crescere dell'ordine del modello (dato che la stima LS si basa sulla minimizzazione di SSR, non è possibile che la migliore parabola abbia una SSR maggiore di una retta)

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 $\Rightarrow$ il coefficiente del termine quadratico risulta = *0* $\Rightarrow$ 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 $F(1,N-k) = \chi^2(1)$

Appendix Table 7 5 per cent. points of the variance ratio F
(values at which the d.f. = 0·95)

(Reproduced from Sir Ronald Fisher and Dr F. Yates: *Statistical Tables for Biological, Medical and Agricultural Research*, Oliver and Boyd Ltd., Edinburgh, by kind permission of the authors and publishers)

ν_2 \ ν_1	1	2	3	4	5	6	8	12	24	∞
1	161·40	199·50	215·70	224·60	230·20	234·00	238·90	243·90	249·00	254·30
2	18·51	19·00	19·16	19·25	19·30	19·33	19·37	19·41	19·45	19·50
3	10·13	9·55	9·28	9·12	9·01	8·94	8·84	8·74	8·64	8·53
4	7·71	6·94	6·59	6·39	6·26	6·16	6·04	5·91	5·77	5·63
5	6·61	5·79	5·41	5·19	5·05	4·95	4·82	4·68	4·53	4·36
6	5·99	5·14	4·76	4·53	4·39	4·28	4·15	4·00	3·84	3·67
7	5·59	4·74	4·35	4·12	3·97	3·87	3·73	3·57	3·41	3·23
8	5·32	4·46	4·07	3·84	3·69	3·58	3·44	3·28	3·12	2·93
9	5·12	4·26	3·86	3·63	3·48	3·37	3·23	3·07	2·90	2·71
10	4·96	4·10	3·71	3·48	3·33	3·22	3·07	2·91	2·74	2·54
11	4·84	3·98	3·59	3·36	3·20	3·09	2·95	2·79	2·61	2·40
12	4·75	3·88	3·49	3·26	3·11	3·00	2·85	2·69	2·50	2·30
13	4·67	3·80	3·41	3·18	3·02	2·92	2·77	2·60	2·42	2·21
14	4·60	3·74	3·34	3·11	2·96	2·85	2·70	2·53	2·35	2·13
15	4·54	3·68	3·29	3·06	2·90	2·79	2·64	2·48	2·29	2·07
16	4·49	3·63	3·24	3·01	2·85	2·74	2·59	2·42	2·24	2·01
17	4·45	3·59	3·20	2·96	2·81	2·70	2·55	2·38	2·19	1·96
18	4·41	3·55	3·16	2·93	2·77	2·66	2·51	2·34	2·15	1·92
19	4·38	3·52	3·13	2·90	2·74	2·63	2·48	2·31	2·11	1·88
20	4·35	3·49	3·10	2·87	2·71	2·60	2·45	2·28	2·08	1·84
21	4·32	3·47	3·07	2·84	2·68	2·57	2·42	2·25	2·05	1·81
22	4·30	3·44	3·05	2·82	2·66	2·55	2·40	2·23	2·03	1·78
23	4·28	3·42	3·03	2·80	2·64	2·53	2·38	2·20	2·00	1·76
24	4·26	3·40	3·01	2·78	2·62	2·51	2·36	2·18	1·98	1·73
25	4·24	3·38	2·99	2·76	2·60	2·49	2·34	2·16	1·96	1·71
26	4·22	3·37	2·98	2·74	2·59	2·47	2·32	2·15	1·95	1·69
27	4·21	3·35	2·96	2·73	2·57	2·46	2·30	2·13	1·93	1·67
28	4·20	3·34	2·95	2·71	2·56	2·44	2·29	2·12	1·91	1·65
29	4·18	3·33	2·93	2·70	2·54	2·43	2·28	2·10	1·90	1·64
30	4·17	3·32	2·92	2·69	2·53	2·42	2·27	2·09	1·89	1·62
40	4·08	3·23	2·84	2·61	2·45	2·34	2·18	2·00	1·79	1·51
60	4·00	3·15	2·76	2·52	2·37	2·25	2·10	1·92	1·70	1·39
120	3·92	3·07	2·68	2·45	2·29	2·17	2·02	1·83	1·61	1·25
∞	3·84	2·99	2·60	2·37	2·21	2·09	1·94	1·75	1·52	1·00

Lower 5 per cent. points are found by interchange of ν_1 and ν_2, i.e. ν_1 must always correspond to the greater mean square.

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 $\Rightarrow$ 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 (o complementare) al test F

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

$$\Downarrow$$

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

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

- Scelgo il modello che minimizza SSR^V

Osservazioni:

- L'assegnamento di un'osservazione al set di identificazione o a quello di identificazione 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 $(1 \div 2\%)$ 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^V) 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 $\theta^{LS(i)}$

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

Ripeto la procedura per $i = 1,\ldots,N$ 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: $Y = \Phi\theta^\circ + V$, $E[V] = 0$, $Var[V] = \sigma^2 I$.
(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):

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

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

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:

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

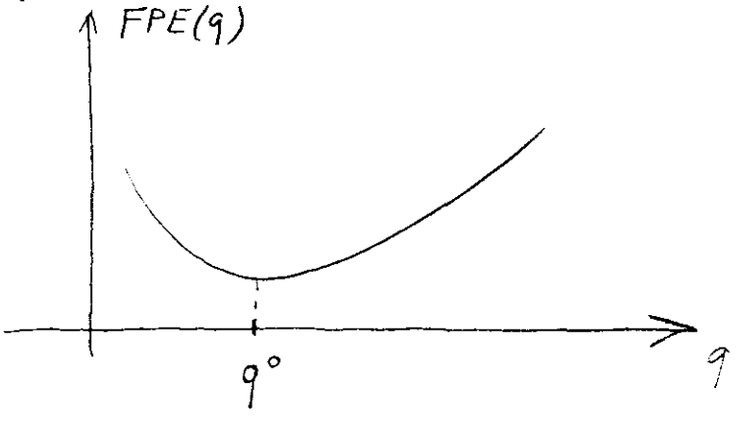
Criterio FPE: Tra diversi modelli matrioska caratterizzati dal numero q di parametri, scelgo quello che minimizza la media (stimata) della SSR^V, ovvero il cosiddetto *Final Prediction Error*

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

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 $\alpha = 0.157$

$$\Downarrow$$

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.

$$\Downarrow$$

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 $lim_{N\to\infty} ln\ FPE = AIC$).

$$\Downarrow$$

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 $Y = \Phi\theta^{LS}+\varepsilon$.

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

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

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:	100	161	227	270	362	90	149	206	247	352
k:	41.6	37.7875	36.4975	35.785	34.53	42.345	39.5375	37.3525	36.36	33.915

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 ponderati (WLS) (alcuni dati sono più affidabili di altri)

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

Vettore dei dati e delle variabili indipendenti:

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

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

1. $$\Phi = \begin{bmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{bmatrix}, \qquad \theta = \begin{bmatrix} \theta_1 \end{bmatrix}$$

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

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

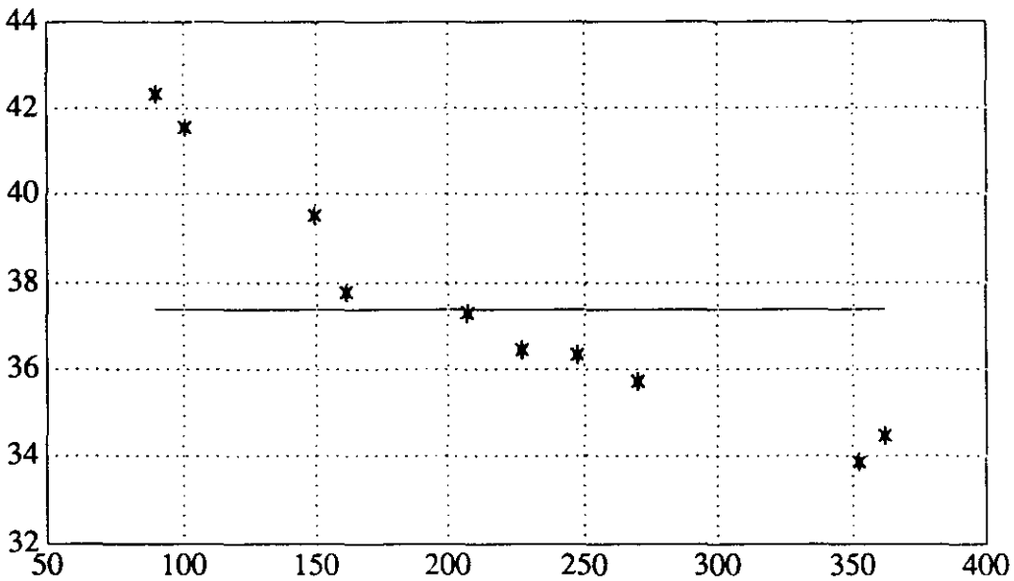
4. $$\Phi = \begin{bmatrix} 1 & T(1) & T(1)^2 & T(1)^3 \\ 1 & T(2) & T(2)^2 & T(2)^3 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & T(10) & T(10)^2 & T(10)^3 \end{bmatrix}, \qquad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}$$

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

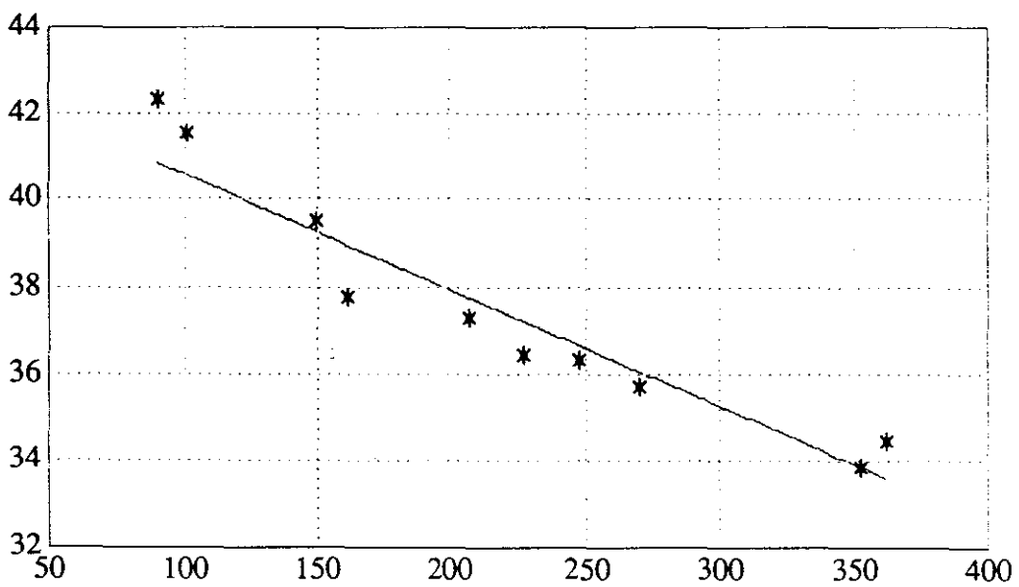
Matrice Q di WLS

$$Q = \text{diag}\{\ 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1/4 \quad 1/4 \quad 1/4 \quad 1/4 \quad 1/4\ \}$$

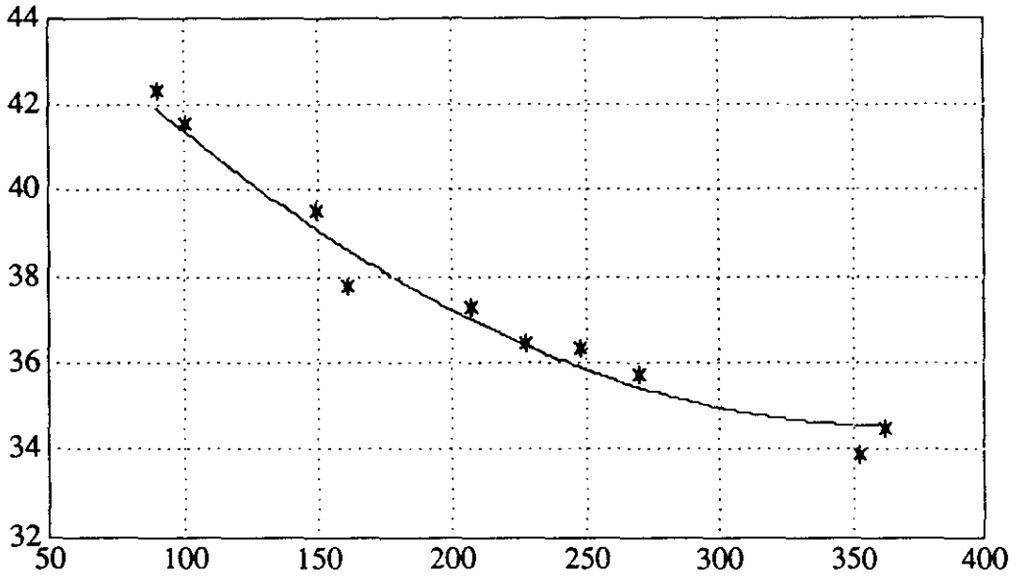
theta1 =
  3.7372e+01
sigmatheta1 =
  8.4336e-01



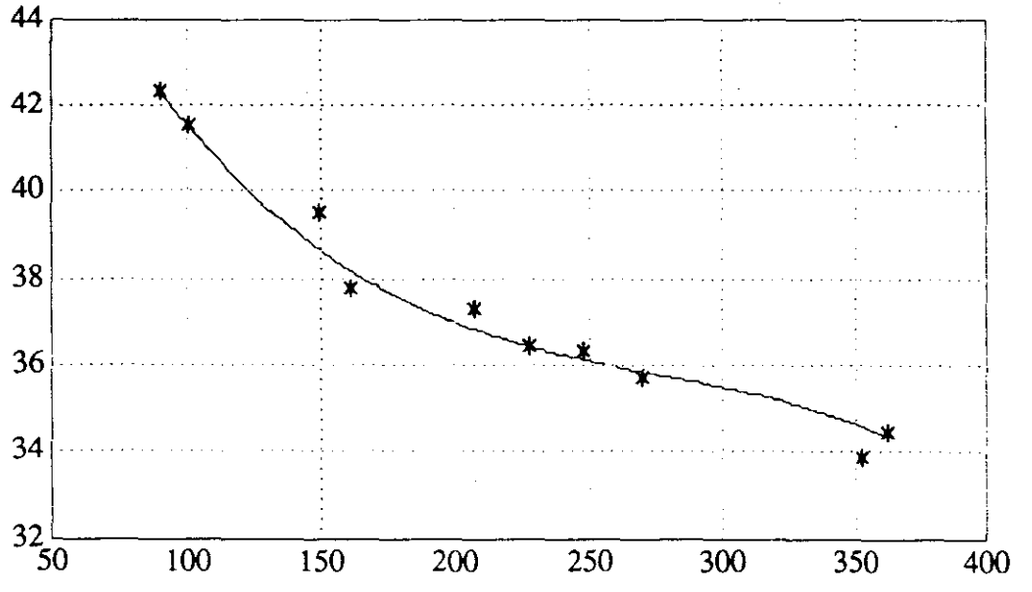
theta2 =
  4.3225e+01 -2.6486e-02
sigmatheta2 =
  7.9222e-01 3.3203e-03



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

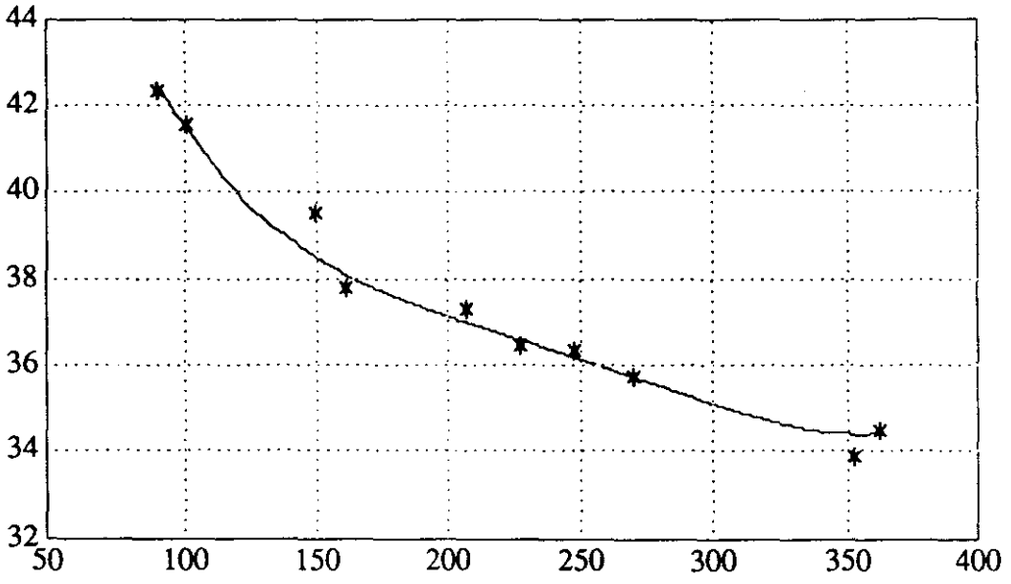


theta5 =

6.2542e+01	-3.7495e-01	2.1788e-03	-5.8682e-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^{WLS} è di poco superiore al 40% del valore stimato del parametro. Nel modello 5 la SD di θ_4^{WLS} è circa uguale a θ_4^{WLS} e lo stesso accade per la SD di θ_5^{WLS}. 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