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).

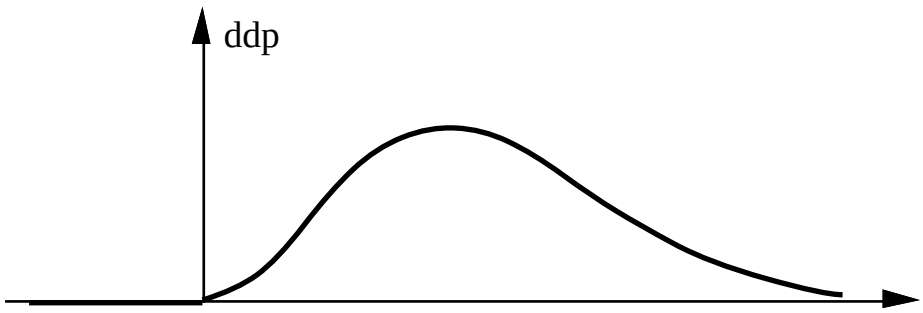
Problema: cosa significa "dello stesso ordine di grandezza"?

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

(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):

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



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

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.

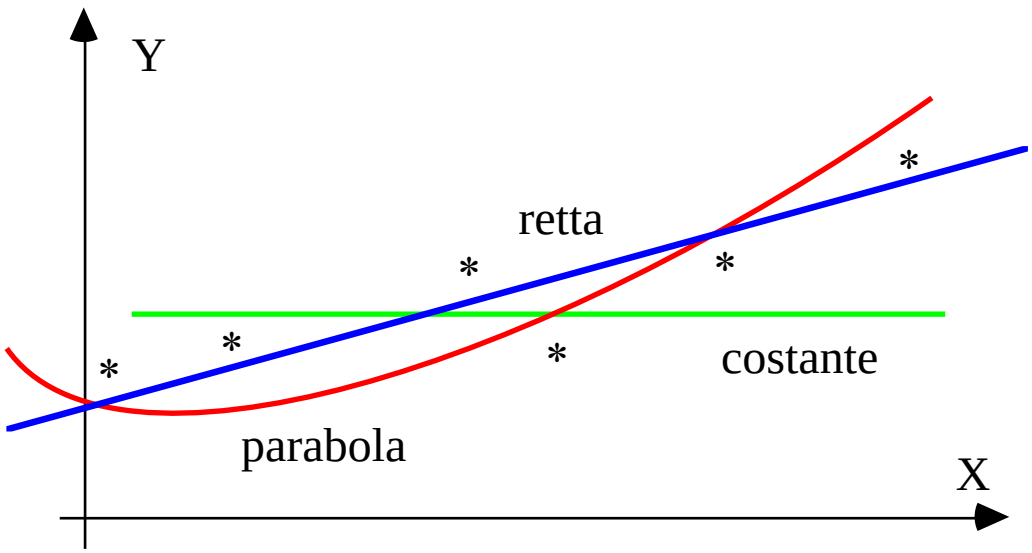
- Il test si basa sull'ipotesi che esista un "modello vero" di tipo lineare che genera i dati e che gli errori siano gaussiani (ipotesi molto semplificativa).

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)

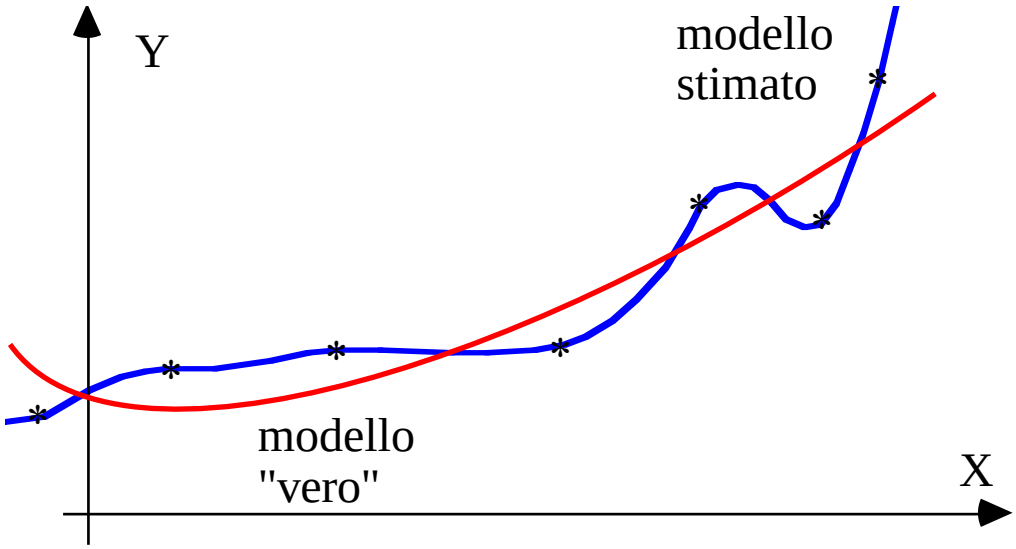
$\Downarrow$

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 $\Rightarrow$ 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

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

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

(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:

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

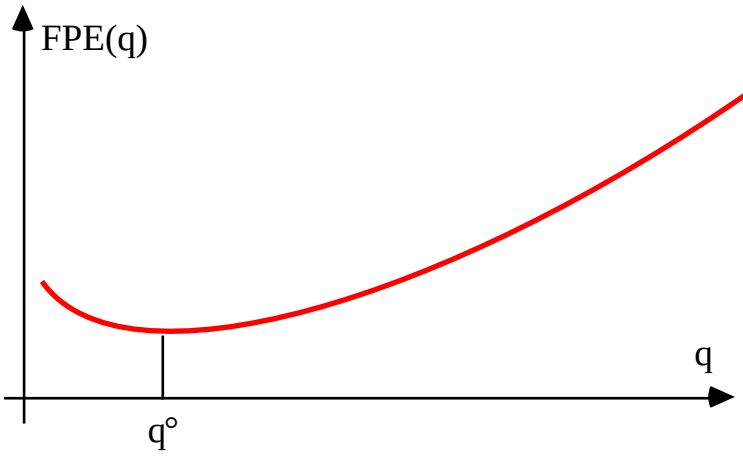
Criterio FPE: Tra diversi modelli matrioska caratterizzati dal numero q di parametri, scelgo quello che minimizza la media (stimata) della SSR^V, ovvero il cosiddetto *Final Prediction Error*

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

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	35.7853	4.53	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)

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

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:

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

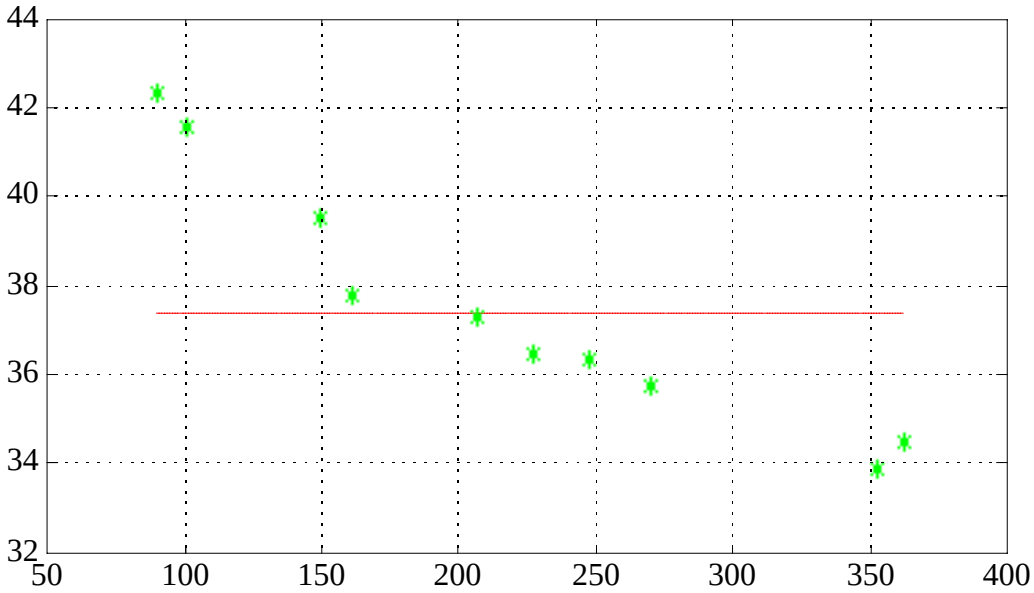
4. $$\Phi = \begin{bmatrix} 1 & T(1) & T(1)^2 & T(1)^3 \\ 1 & T(2) & T(2)^2 & T(2)^3 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & T(10) & T(10)^2 & T(10)^3 \end{bmatrix}, \qquad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix}$$

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

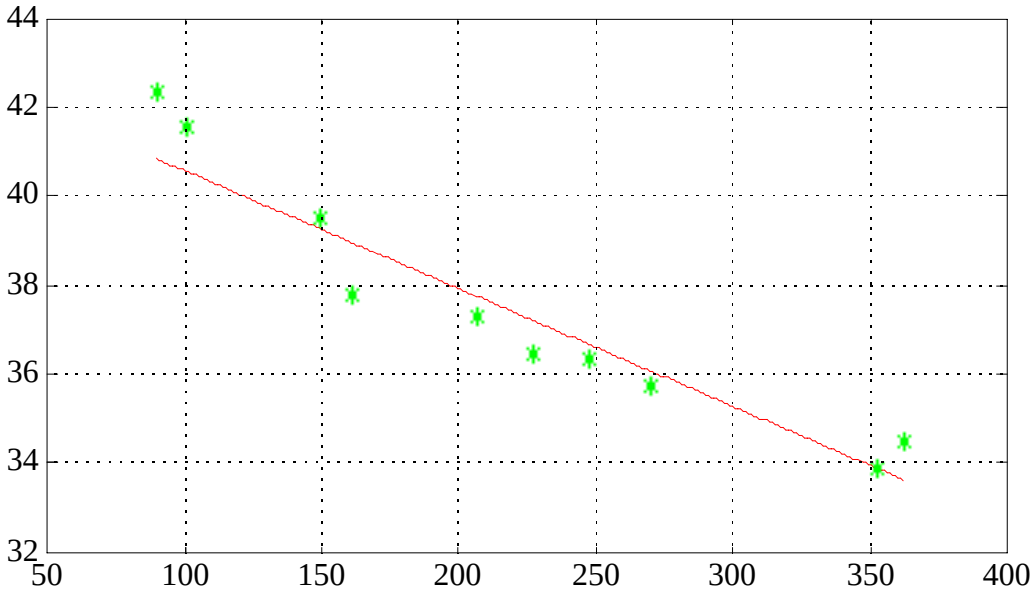
Matrice Q di WLS

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

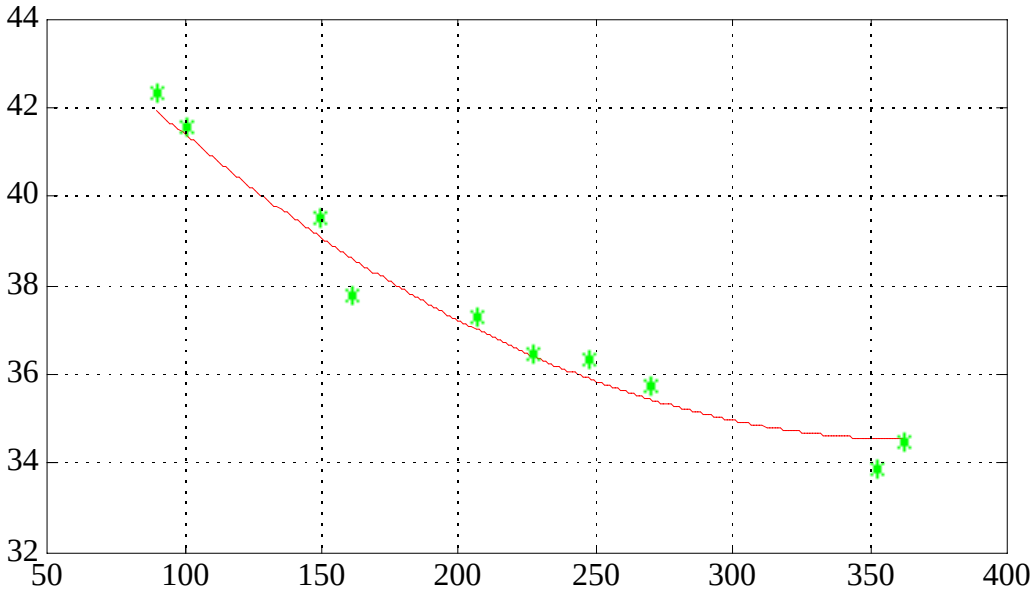
theta1 =
   3.7372e+01
sigmatheta1 =
   8.4336e-01



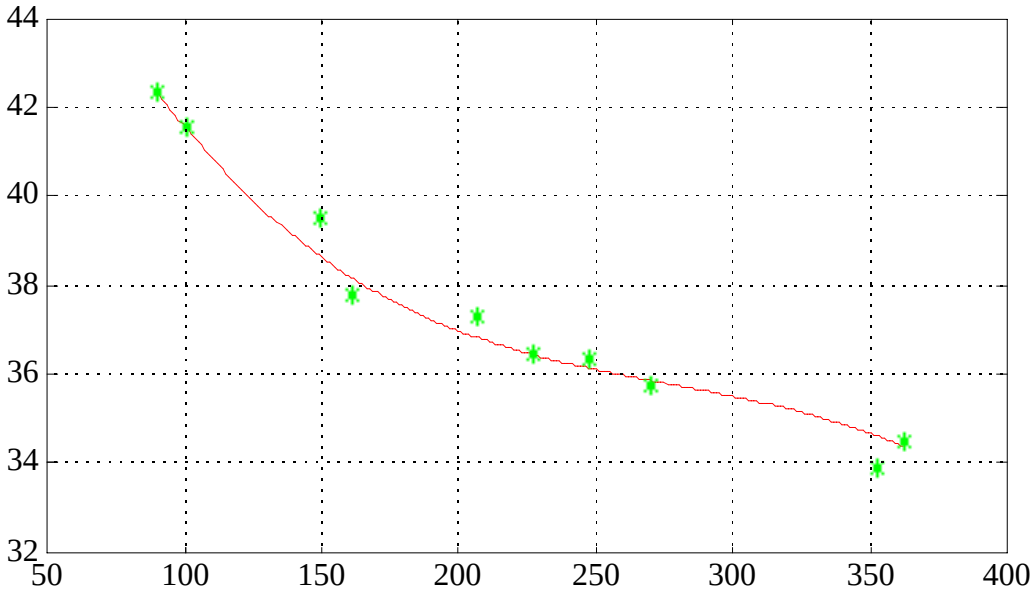
theta2 =
   4.3225e+01 -2.6486e-02
sigmatheta2 =
   7.9222e-01   3.3203e-03



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

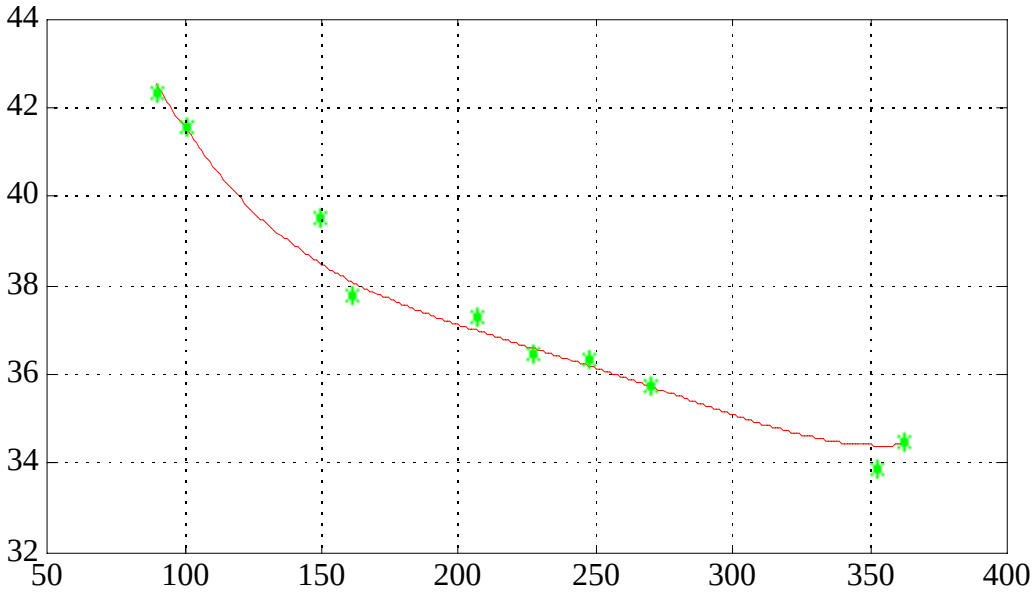


theta5 =

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^{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