

$$\theta^{LS} = \arg \min_{\theta} (Y - \Phi\theta)^T (Y - \Phi\theta)$$

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

Dim:  $\text{rank}(\Phi) = q \Rightarrow \det(\Phi^T \Phi) \neq 0$

Infatti, se per assurdo fosse  $\det(\Phi^T \Phi) = 0 \Rightarrow \exists x \neq 0$  t.c.  $\Phi^T \Phi x = 0 \Rightarrow x^T \Phi^T \Phi x = 0 \Rightarrow ((\Phi x)^T \Phi x) = \|\Phi x\|^2 \Rightarrow \Phi x = 0 \Rightarrow \Phi$  ha colonne linearmente dipendenti.

$\text{rank}(\Phi) < q$  (assurdo) e l'inverso esiste.

Calcolo il gradiente di  $J(\theta) = (Y - \Phi\theta)^T (Y - \Phi\theta)$

$$\frac{dJ(\theta)}{d\theta} = -2(Y - \Phi\theta)^T \Phi$$

$$\frac{dJ(\theta)}{d\theta} = 0 \quad (Y^T - \theta^T \Phi^T) \Phi = 0 \Rightarrow Y^T \Phi = \theta^T \Phi^T \Phi \Rightarrow \boxed{\Phi^T \Phi \theta = \Phi^T Y}$$

questo minimo

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

Hessiano

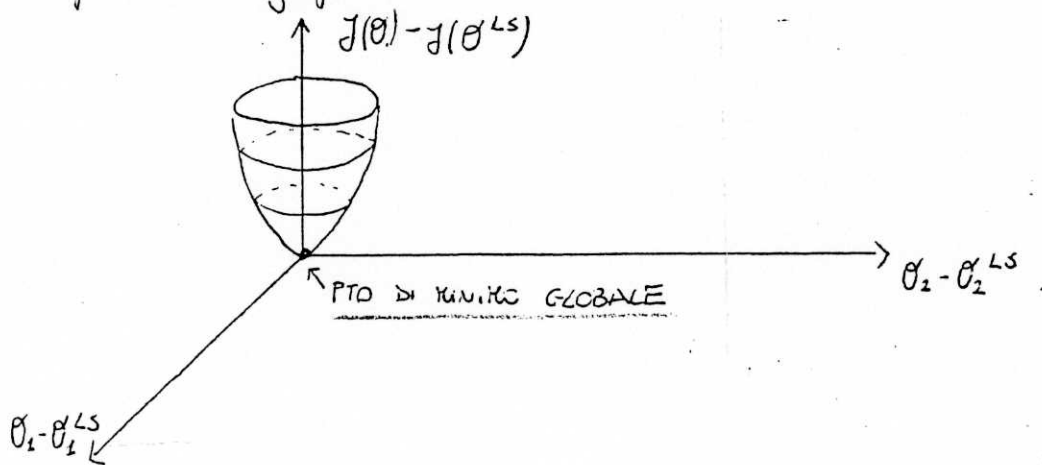
$$\frac{d^2 J(\theta)}{d\theta^2} = 2 \Phi^T \Phi \geq 0$$

DIAGNE semidefinita positiva.

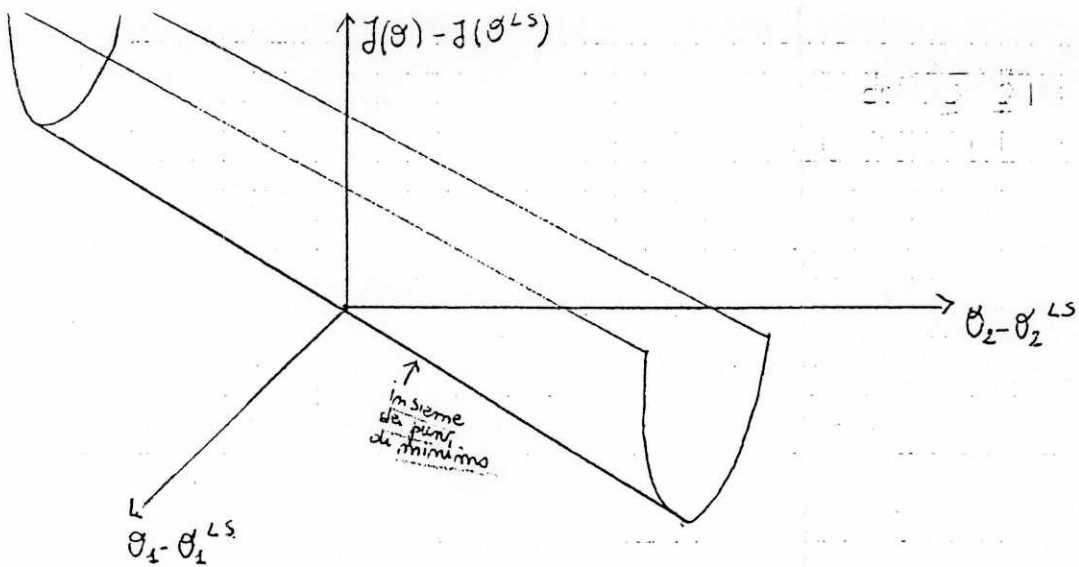
Proposizione ( $A \geq 0$  e  $\det(A) \neq 0$ )  $\Rightarrow A > 0$ .

Dato che  $\det(\Phi^T \Phi) \neq 0 \Rightarrow \Phi^T \Phi > 0$  Ho un punto di minimo

Interpretazione grafica

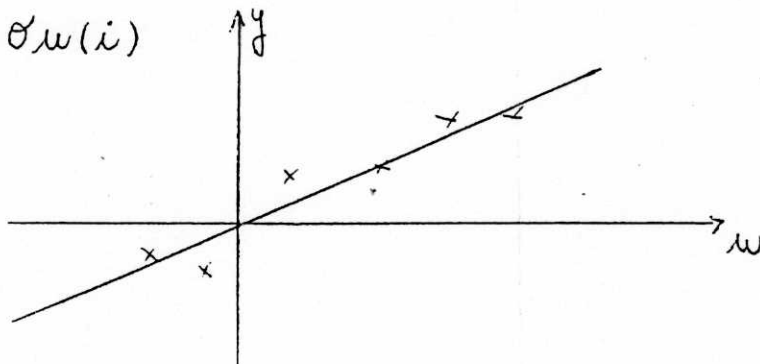


$\det(\Phi^T \Phi) \neq 0 \Rightarrow \exists$  un'unica soluzione



Esempio: Regressione lineare con  $q=1$ .

$$y(i) = \theta w(i)$$



$$Y = \underline{\Phi} \theta$$

$$y(1) = w(1) \theta$$

$$y(2) = w(2) \theta$$

$$Y = \begin{pmatrix} y(1) \\ y(2) \\ \vdots \\ y(N) \end{pmatrix}$$

$$\underline{\Phi} = \begin{pmatrix} w(1) \\ w(2) \\ \vdots \\ w(N) \end{pmatrix}$$

$$\theta^{LS} = (\underline{\Phi}^T \underline{\Phi})^{-1} \underline{\Phi}^T Y = \frac{\sum_{i=1}^N w(i) y(i)}{\sum_{i=1}^N w(i)^2}$$

Il problema dell'identificabilità

Come succede se  $\text{rank}(\underline{\Phi}) < q$ ?

Per semplicità consideriamo  $q=3$  e un pb di regressione lineare.

$$y(t) = \theta_1 w_1(t) + \theta_2 w_2(t) + \theta_3 w_3(t)$$

$$\underline{\Phi} = \begin{bmatrix} w_1(1) & w_2(1) & w_3(1) \\ w_1(2) & w_2(2) & w_3(2) \\ \vdots & \vdots & \vdots \\ w_1(N) & w_2(N) & w_3(N) \end{bmatrix}$$

Supponi che  $\text{rank}(\underline{\Phi}) = 2 \Rightarrow$  una delle colonne di  $\underline{\Phi}$  è combinazione lineare delle altre, ovvero esistono  $\alpha, \beta$  t.c.  $w_3(t) = \alpha w_1(t) + \beta w_2(t)$

caso  $y(t) = \theta_1 w_1(t) + \theta_2 w_2(t) + \theta_3 w_3(t)$   
 $= (\theta_1 + \alpha \theta_3) w_1(t) + (\theta_2 + \beta \theta_3) w_2(t)$

$\Rightarrow w_3$  è inutile perché posso ottenere la stessa previsione  $y$  usando solo  $w_1(t), w_2$   
 Considero un nuovo vettore dei parametri incogniti:

$$\bar{\theta} = \begin{bmatrix} \bar{\theta}_1 \\ \bar{\theta}_2 \end{bmatrix} = \begin{bmatrix} \theta_1 + \alpha \theta_3 \\ \theta_2 + \beta \theta_3 \end{bmatrix}$$

Considero la regressione di  $Y$  su  $w_1$  e  $w_2$ . Se la condizione di identificabilità è soddisfatta trovo un'unica soluzione

$$\bar{\theta}^{LS} = \begin{bmatrix} \bar{\theta}_1^{LS} \\ \bar{\theta}_2^{LS} \end{bmatrix}$$

Se tengo  $w_3 \Rightarrow$  so risolvere

$$\theta^{LS} = [\theta_1^{LS} \quad \theta_2^{LS} \quad \theta_3^{LS}]^T \text{ che soddisfa}$$

DI CONDIZIONAMENTO: è un indice del grado di simpatia di una matrice.

Metica che tende alla simpatia  $\rightarrow$  determinante della matrice che tende a zero

$$\begin{cases} \bar{\theta}_1^{LS} = \theta_1^{LS} + \alpha \theta_3^{LS} \\ \bar{\theta}_2^{LS} = \theta_2^{LS} + \beta \theta_3^{LS} \end{cases}$$

LIMITI DELLA STIMA LS

- affidabilità della stima
  - il modello è giusto?
  - confronto tra modelli
- $\Downarrow$  per dare della risposta  
 Geometria della stima

STIMA ML

Devo fare ipotesi sulla d.d.p dell'errore di misura

$$V = Y - \Phi \theta$$

$$V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}$$

Ipotesi I1

$$Y = \Phi \theta + V, \quad V \sim N(0, \Sigma_V), \quad \Sigma_V > 0$$

Se gli errori sono indipendenti

$$\Sigma_V = \begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \sigma_N^2 \end{bmatrix}$$

se inoltre hanno tutti la stessa

risultato

$$\Sigma_V = \sigma^2 I$$

Teorema

Se vale l'ipotesi I1 e  $\text{rank}(\Phi) = q$ , allora

a)  $\theta^{ML} = \arg \min_{\theta} J^{ML}(\theta), \quad J^{ML}(\theta) = \epsilon^T \Sigma_V^{-1} \epsilon$

b)  $\theta^{ML} = (\Phi^T \Sigma_V^{-1} \Phi)^{-1} \Phi^T \Sigma_V^{-1} Y$  comp. opportuno  
 $(\Rightarrow \theta^{ML} \text{ è gaussiano})$

c)  $E[\theta^{ML}] = \theta$  (Stima non偏uso)

SOMMA DEI QUADRATI DEI RESIDUI (SQR)

$$d) \text{Var}[\theta^{HL}] = (\Phi^T \Sigma_V^{-1} \Phi)^{-1}$$

Dim  
 2  
 argomenti

Ipotesi forti (V gaussiane)

Si verifica che  $\text{Var}[\theta^{HL}] = S^{-1}$

(S: matrice di informazione di Fisher)  $\rightarrow$  si è raggiunto il limite di Cramer-Rao

$\Rightarrow \theta^{HL}$  è lo stimatore a minima varianza

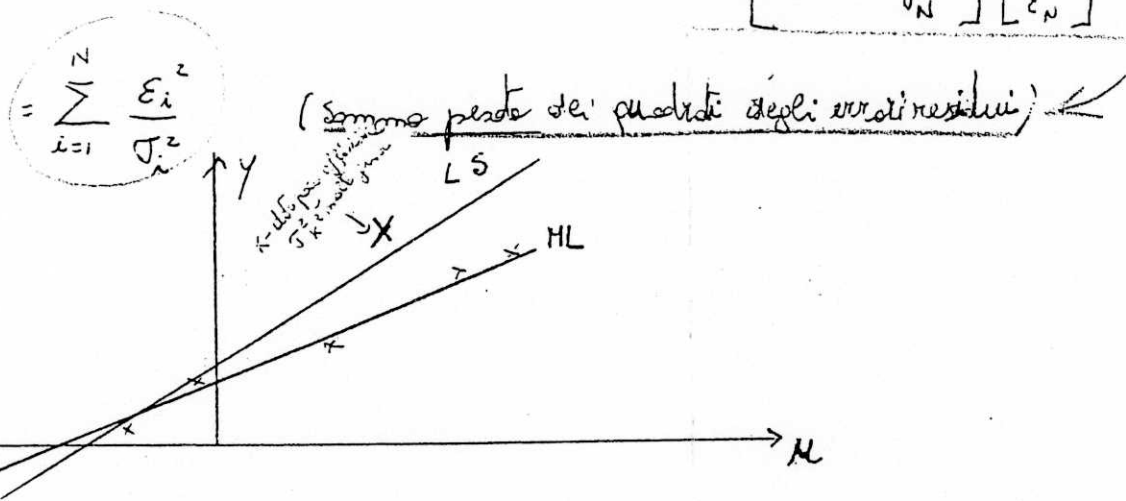
Relazione con L.S.

Se  $\Sigma_V = \sigma^2 I \Rightarrow \theta^{HL} = \theta^{LS}$

$$\left( \theta^{HL} = \frac{(\Phi^T \Phi)^{-1} \Phi^T Y}{\sigma^2} = \theta^{LS} \right)$$

$$\text{Se } \Sigma_V = \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_N^2 \end{bmatrix}$$

$$J^{HL}(\theta) = \varepsilon^T \Sigma_V^{-1} \varepsilon = [\varepsilon_1 \varepsilon_2 \dots \varepsilon_N] \begin{bmatrix} \frac{1}{\sigma_1^2} & & & 0 \\ & \ddots & & \\ & & \frac{1}{\sigma_N^2} & \\ 0 & & & \ddots \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} =$$



Vogliamo trovare gli intervalli di confidenza

STIMA DELLA VARIANZA DEL DISTURBO

Spero  $\Sigma_V$  è moltiplicativo (  $\Sigma_V = \sigma^2 I$  )

Teorema Vale I1 con  $\Sigma_V = \sigma^2 \Psi$  con  $\Psi$  matrice nota e  $\sigma^2$  scalare incognito

Allora

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

(ii) 
$$(\sigma^2)^{HL} = \frac{(Y - \Phi \theta^{HL})^T \Psi^{-1} (Y - \Phi \theta^{HL})}{N} = \frac{J_{\Psi}^{HL}(\theta^{HL})}{N}$$

$$J_{\psi}^{HL} = \varepsilon^T(\theta) \Psi^{-1} \varepsilon(\theta)$$

NOTA:  $(\hat{\sigma}^2)^{HL}$  è polarizzato  
 Per avere una stima non polarizzata

$$\hat{\sigma}^2 = \frac{J_{\psi}^{HL}(\theta^{HL})}{N-q}$$

## STIMA HL: INTERVALLI DI CONFIDENZA

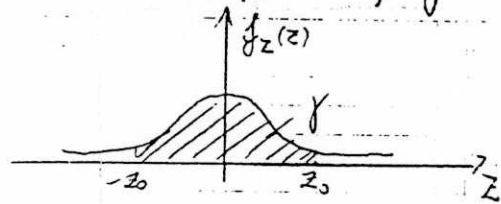
(a)  $\Sigma_V$  nota.

$\theta^{HL} \sim N(\theta, \Sigma_{\theta^{HL}})$   $\Sigma_{\theta^{HL}} = (\Phi^T \Sigma_V^{-1} \Phi)^{-1}$   
 anche le marginali sono gaussiane  $\Downarrow$  intervalli di confidenza per uno stimatore gaussiano con varianza nota.

$$I_{\gamma}(\theta_i) = [\theta_i^{HL} - z_0 \sigma_{\theta_i^{HL}}, \theta_i^{HL} + z_0 \sigma_{\theta_i^{HL}}]$$

$$\sigma_{\theta_i^{HL}}^2 = \left[ \Sigma_{\theta^{HL}} \right]_{ii}$$

$z_0$  è t.c.  $P(|Z| \leq z_0) = \gamma$



11 maggio 2001

(b)  $\Sigma_V = \sigma^2 \Psi$  ( $\Psi$  matrice nota,  $\sigma^2$  scalare incognito)

Stimare

$$\hat{\Sigma}_{\theta^{HL}} = \hat{\sigma}^2 (\Phi^T \Psi^{-1} \Phi)^{-1}, \quad \hat{\sigma}_{\theta_i^{HL}}^2 = \left[ \hat{\Sigma}_{\theta^{HL}} \right]_{ii}$$

Proprietà

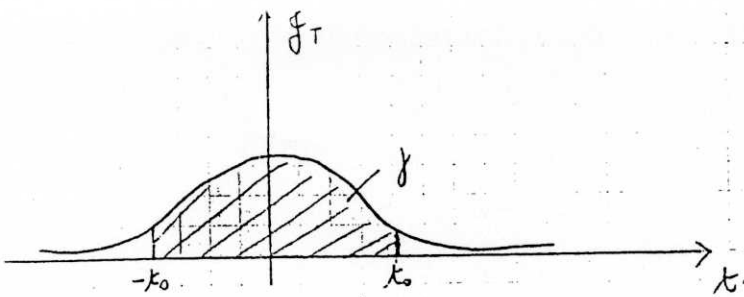
Risulta che  $\frac{\theta_i^{HL} - \theta_i}{\hat{\sigma}_{\theta_i^{HL}}^2} \sim T_{N-q}$  (t di Student.)

$N$ : n° dei dati

$q$ : n° dei parametri

$$I_{\gamma}(\theta_i) = [\theta_i^{HL} - t_0 \hat{\sigma}_{\theta_i^{HL}}, \theta_i^{HL} + t_0 \hat{\sigma}_{\theta_i^{HL}}]$$

$t_0$  è t.c.  $P(|T_{N-q}| \leq t_0) = \gamma$



APPLICAZIONE: IL PROBLEMA DELLA REGRESSIONE LINEARE

$$y(t) = \sigma_1 u_1(t) + \sigma_2 u_2(t) + \dots + \sigma_q u_q(t) + n(t) \quad t=1, \dots, N$$

$$\text{Var}[V] = \sigma^2 I$$

$$V = \begin{bmatrix} n(1) \\ \vdots \\ n(N) \end{bmatrix}$$

Se inoltre  $V$  è gaussiano

$$\sigma^{HL} = \sigma^{LS}$$

$$\Phi = \begin{bmatrix} u_1(1) & u_2(1) & \dots & u_q(1) \\ \vdots & \vdots & \ddots & \vdots \\ u_1(N) & u_2(N) & \dots & u_q(N) \end{bmatrix}$$

Definendo  $\varphi(t) = [u_1(t) \ u_2(t) \ \dots \ u_q(t)]^T$ , si vede che  $\sigma^{LS} = \frac{(\Phi^T \Phi)^{-1} \Phi^T Y}{\left[ \sum_{t=1}^N \varphi(t) \varphi(t)^T \right]^{-1} \left[ \sum_{t=1}^N \varphi(t) y(t) \right]}$

(se  $\text{rank}(\Phi) = q$  CONDIZIONE DI IDENTIFICABILITÀ)

$$\hat{\sigma}^2 = \frac{J(\sigma^{HL})}{N-q} = \frac{\sum_{t=1}^N \varepsilon(t)^2}{N-q} = \frac{\sum_{t=1}^N (y(t) - \varphi(t)^T \sigma^{HL})^2}{N-q}$$

$$y(t) = \varphi(t)^T \sigma + n(t)$$

Stima  $\text{Var}[\sigma^{HL}]$  come  $\sum \sigma^{HL} = \hat{\sigma}^2 \left[ \sum_{t=1}^N \varphi(t) \varphi(t)^T \right]^{-1}$

Spesso i risultati sono espressi come segue:

$$y(t) = \sigma_1^{HL} u_1(t) + \sigma_2^{HL} u_2(t) + \dots + \sigma_q^{HL} u_q(t) + \varepsilon(t)$$

Il parametro  $\varepsilon(t)$  è il parametro di disturbo

$$\varepsilon(t) = y(t) - \varphi(t)^T \sigma^{HL}$$

$$n(t) = y(t) - \varphi(t)^T \sigma$$

$$\left[ \hat{\sigma}_{\sigma_1^{HL}} \right]$$

$$\left[ \hat{\sigma}_{\sigma_2^{HL}} \right]$$

$$\left[ \hat{\sigma}_{\sigma_q^{HL}} \right]$$

$$\left[ \hat{\sigma} \right]$$

Regole per capire quali regressioni vanno tolte dal modello ( $\exists$  variabile che non influenza il modello)

Problema incerto: come faccio ad essere sicuro che  $\theta_k \neq 0$ ? (Se non sono sicuro può essere meglio togliere  $\mu_k$  dal modello)

Procedo come "per assurdo", faccio l'ipotesi che  $\theta_k = 0$  e controllo se i risultati sperimentali lo smentiscono

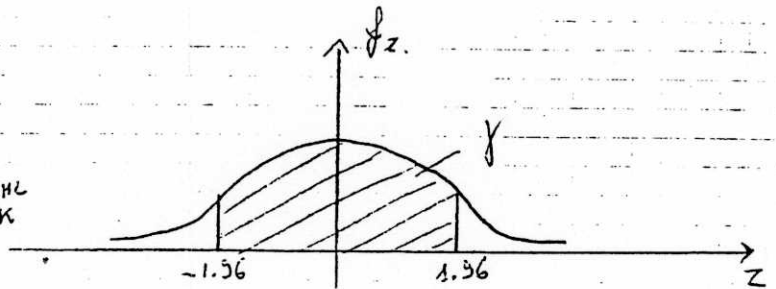
$$\theta_k = 0 \Rightarrow \frac{\theta_k^{HL} - E(\theta_k^{HL})}{\sqrt{\text{Var}[\theta_k^{HL}]}} \sim Z \Rightarrow \frac{\theta_k^{HL}}{\sigma_k^{HL}} \sim Z$$

$$\left( \frac{\theta_k^{HL}}{\hat{\sigma}_k^{HL}} \sim T_{N-9} \right)$$

Nel 95% dei casi  $|Z| \leq 1.96$

$$\left| \frac{\theta_k^{HL}}{\hat{\sigma}_k^{HL}} \right| \leq 1.96$$

Idea: Se  $\left| \theta_k^{HL} \right| \leq 1.96 \hat{\sigma}_k^{HL}$



non ci smentisce di assurdo  $\Rightarrow$  non userei

Smentire l'ipotesi  $\theta_k = 0$

Invece se  $\left| \theta_k^{HL} \right| > 1.96 \hat{\sigma}_k^{HL}$

mi trovo in una situazione che si verifica solo nel 5% dei casi (ipotesi  $\theta_k = 0$ )

$\Rightarrow$  "ASSURDO" (non probabile)

$\Rightarrow$  respingo l'ipotesi  $\theta_k = 0 \Rightarrow$  il parametro  $\theta_k$  è SIGNIFICATIVAMENTE  $\neq 0$

CONCLUSIONE

$$\left| \theta_k^{HL} \right| > 2 \hat{\sigma}_k^{HL}$$

sono "sicuro" che  $\theta_k \neq 0$

$$\left| \theta_k^{HL} \right| < 2 \hat{\sigma}_k^{HL}$$

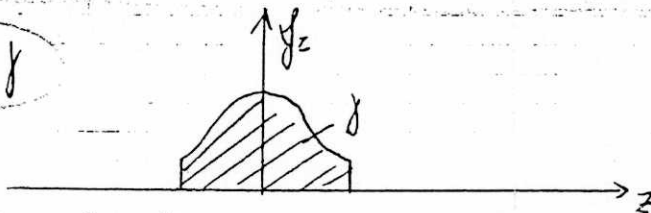
non so garantire che  $\theta_k \neq 0$

Esistono 2 tipi di errore:

$F^+$  (falso positivo) affermo  $\theta_k \neq 0$  mentre in realtà  $\theta_k = 0$

$F^-$  (falso negativo) affermo  $\theta_k = 0$  quando in realtà  $\theta_k \neq 0$

$$P(F^+) = 1 - \gamma$$



Cost di ipotesi (soggettivo: ognuno può scegliere un "nuovo"  $\gamma$ )

# VALIDAZIONE TEST $\chi^2$

Avendo a disposizione  $N$  dati  $y_1, \dots, y_N$  come faccio a sapere se il modello.

$$Y = \Phi \theta^0 + V, \text{ var}[V] = \sigma^2 I$$

li dedurre successivamente?

Faccio  $\theta^{LS} \equiv \theta^0 \Rightarrow \varepsilon = Y - \Phi \theta^{LS} \equiv V$  (il vettore dei residui è una "sintesi" del vettore degli errori di misura). Perciò mi aspetto che

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

Se guardo  $\sigma^2$  e bene controllate che  $\sigma^2$  e la varianza campionaria dei residui abbiano lo stesso ordine di grandezza (vale a dire 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)

Ipotesi  $H_1$   $Y = \Phi \theta^0 + V$   $V \sim N(0, \sigma^2 I)$

Teorema sotto  $H_1$ , dividendo per  $\sigma^2$  la somma dei quadrati residui della stima LS si ottiene una V.C di tipo  $\chi^2$  con  $N-q$  g.d.l.  
( $N$  n° di dati,  $q$  n° di parametri)

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

FTAB per  $\chi^2$

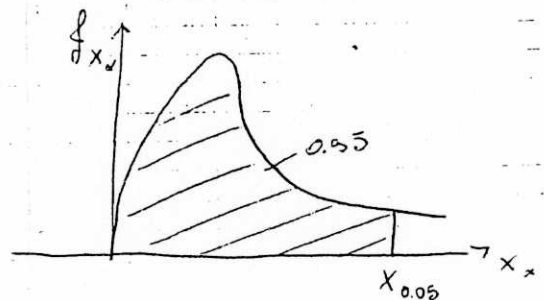
95% di cui  $\frac{\varepsilon^T \varepsilon}{\sigma^2} < 14.07$

Se accade  $\frac{\varepsilon^T \varepsilon}{\sigma^2} > 14.07 \Rightarrow$  modello non buono

- 1) Fissare un livello di signif.  $\alpha$
- 2) Cercare su TAB il valore  $\chi_{\alpha}$

$\frac{\varepsilon^T \varepsilon}{\sigma^2} < \chi_{\alpha} \Rightarrow$  non respingo il modello

$\frac{\varepsilon^T \varepsilon}{\sigma^2} > \chi_{\alpha} \Rightarrow$  respingo il modello



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

Beste  $\varepsilon^T \Sigma_V^{-1} \varepsilon$

Possibile: 1) può essere difficile capire qual è il valore per cui viene scartato il modello.



- $y = \Phi \theta + V$  spiega male i dati
- $V$  non è gaussiana
- Il valore di  $\sigma^2$  è sbagliato per difetto
- Gli errori di misura non hanno tutti la stessa varianza

2) Il test si basa sull'ipotesi che  $\exists$  un "modello vero" di tipo lineare che genera i dati e che gli errori sono gaussiani (ipotesi semplificative)

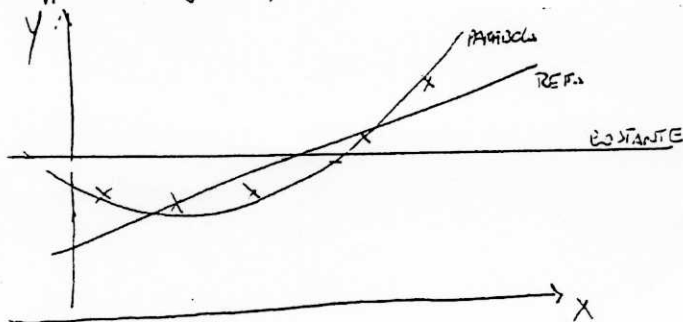
**TEST F**

Confronto tra modelli

MODELLI "MARIOSKA" Una sequenza di classi di modelli in cui ciascuna classe comprende al suo interno (come casi particolari) le classi precedenti

Es.  $M_1 = [retta]$     $M_2 = [parabola]$     $M_3 = [cubiche] \dots$

Pb.  $N$  coppie  $(x_i, y_i)$ , trovare una curva che approssimi al meglio i dati sperimentali.



Idea stupida: considero i diversi modelli (retta, parabola, ...) e ne stimo i parametri in LS

poi fra i modelli scelgo quello che minimizza SSR (la somma dei quadrati dei residui)

FATTO: SSR decresce sempre al crescere dell'ordine del modello (dato che LS minimizza la minimizzazione di SSR, non è possibile che la miglior parabola abbia una SSR che una retta)

Usare la minimizzazione di SSR conduce a scegliere sempre il modello più complesso (per esempio  $N=100 \Rightarrow$  polinomio di ordine 99) 100 punti  $\Rightarrow$  con 100 parametri posso passare per tutti i punti

• Che male c'è nell'eccedere con il molti parametri?

• Nessun pb se i dati fossero privi di rumore.

Es. Fitto con una parabola dei dati su una retta  $\Rightarrow$  il coeff del termine quadratico risulta  $= 0 =$  non commetto errori

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

PRINCIPIO DI PARSIMONIA: non usare più parametri di quelli che servono

Idea:  $M_{k-1}$  e  $M_k$ , sappiamo che  $SSR_k < SSR_{k-1}$

Scegliamo  $M_k$  solo se  $SSR_k$  è "molto più piccola" di  $SSR_{k-1}$

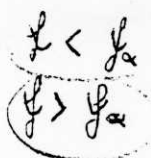
Generalmente  $SSR \sim \frac{1}{N}$  per  $I \downarrow$  ( $y = \Phi \theta + V$     $V \sim N(0, \sigma^2 I)$ )

$$F = (N-k) \frac{SSR_{k-1} - SSR_k}{SSR_k} = \text{è una } F \text{ di Fisher con } (1, N-k) \text{ g. di l.}$$

• Per  $N-k$  grande  $F(1, N-k) \approx \chi^2(1)$   
↑  
rende.

TAB F di Fisher.

Test F Finché un livello di significatività ipotesi 203 è scelto  $\alpha$   $\rightarrow P(F > F_{\alpha}) = \alpha$



scelgo  $H_{k-1}$   
 scelgo  $H_k$

$\alpha = 0.05$

- Non è necessario conoscere  $\sigma^2$
- è sufficiente (o scelto il livello di significatività)
- $V \sim N(0, \sigma^2 \Psi)$   $\Psi$  matrice nota  $\Rightarrow \varepsilon^T \Psi^{-1} \varepsilon$  al posto di  $SSR = \varepsilon^T \varepsilon$
- Si presuppone l'esistenza di un modello vero
- Si applica solo a modelli matriciali.

pag 30: Considerazione: sempre abbastanza dati

FPE e AIC hanno una proba > 0 di convergenza al modello

→ il miglior valore + di un criterio + valutare un modello

APPLICAZIONE: REGRESSIONE LINEARE

$q = 1, 2$

Dati:  $y(1), y(2) \dots y(N)$   
 $x(1), x(2) \dots x(N)$

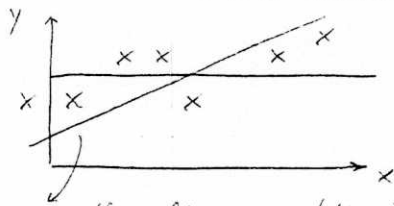
$H_p: V \sim N(0, \sigma^2 I)$

$V = \begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(N) \end{bmatrix}$

2 modelli alternativi:

(a)  $y(t) = \theta_1 + v(t)$

(b)  $y(t) = \theta_1 + \theta_2 x(t) + v(t)$



modello + semplice = cost (y non dipende da x)

modello lineare (y è fz lineare di x)

Media campionaria

Stimiamo (a):

$Y = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix}$

$\Phi = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$   
matrice  $\Phi$

$\theta_1^{ML} = (\Phi^T \Phi)^{-1} \Phi^T Y = N^{-1} \sum_{i=1}^N y(i) = M_y$

matrice una matrice che in questo caso è = I

(b)  $\Phi = \begin{bmatrix} 1 & x(1) \\ 1 & x(2) \\ \vdots & \vdots \\ 1 & x(N) \end{bmatrix}$

$\theta^{ML} = \begin{bmatrix} N & \sum x(i) \\ \sum x(i) & \sum x(i)^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y(i) \\ \sum x(i)y(i) \end{bmatrix}$

2 parametri  $\theta_1^{ML}$  e  $\theta_2^{ML}$

$\theta_1^{ML} = M_y - \theta_2^{ML} M_x$

$\theta_2^{ML} = \frac{S_{xy}}{S_{xx}}$

dove

$S_{xy} = \frac{\sum (x(i) - M_x)(y(i) - M_y)}{N}$        $S_{xx} = \frac{\sum (x(i) - M_x)^2}{N}$

Valore di y previsto

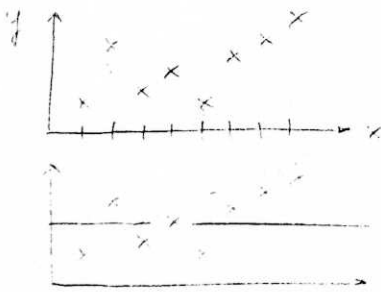
è la COVARIANZA CAMPIONARIA

$\hat{y}(t) := \Phi \theta^{ML} = M_y + \frac{S_{xy}}{S_{xx}} (x(t) - M_x)$

è come se avessimo sostituito su valori teorici quelli campionari

Ricorda  $E[Y|X=x] = E[Y] + Cov[X,Y] Var[X]^{-1} (x - E[X])$

NB: Dal pto di vista concettuale ↑ immagino che x sia una V.C. Ma nella realtà ( $\hat{y}(t)$ ) potrebbe NON esserlo (es: x potrebbero essere dei tempi → uniforme).



La formula di  $\hat{y}(t)$  vale sempre, per V.C. e NON.

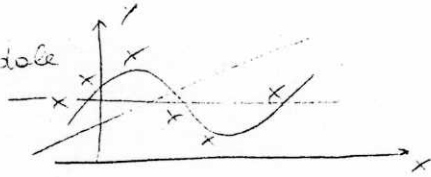
Dobbiamo decidere se il modello è (a) o (b)

Devo decidere se  $\theta_2$  serve o meno: Vd. Tab di FISHER

$$f := \frac{J(\bar{\theta}_1^{ML}) - J(\theta^{ML})}{J(\theta^{ML}) / (N-2)} \sim F(1, N-2)$$

x fare il Test di omogeneità Hp la gaussianità  $\Rightarrow$  ML

Rischioso: la dipendenza potrebbe essere nonlineare  $\Rightarrow$  tra costi e rete selgo la rete ML non è quella vera



Introduciamo un indice normalizzato che misura in che grado il modello (b) è superiore ad (a)

Coeff. di Determinazione (Multipla):  $R^2 := \frac{\bar{J}(\bar{\theta}_1^{LS}) - J(\theta^{LS})}{\bar{J}(\bar{\theta}_1^{LS})} = 1 - \frac{J(\theta^{LS})}{\bar{J}(\bar{\theta}_1^{LS})}$

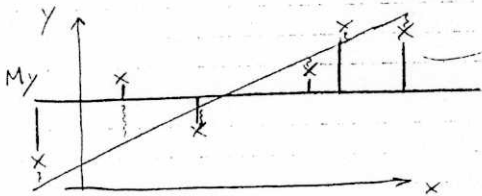
Scarti attorno alla retta di regressione

Scarti attorno a  $M_y$

in questo caso LS = ML, Testi non devo fare Hp iniziali.

$$0 \leq R^2 \leq 1$$

Sempre  $\leq 1$  (che sopra il modello è + complicato).

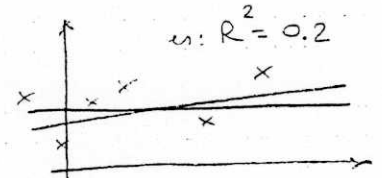


scarti attorno a  $M_y$

scarti attorno alla retta

$R^2 \approx 0$  (scarti quasi sovrapposti).

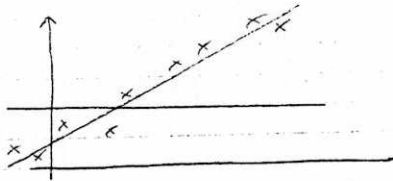
migliore poco pancia da (a) a (b).



il modello (b) è molto meglio

$R^2 \approx 1$

es.  $R^2 = 0,91$



B est

L linear

U unbiased

E estimator

Se rimuovo l'ipotesi di gaussianità, cosa fanno due?

IPOTESI I2:  $Y = \Phi \theta + V \rightarrow \exists$  un modello vero

$$E[V] = 0 \quad \text{Var}[V] = \sigma^2 \Psi$$

( $\sigma^2$  scalare eventualmente incognito)

□

Teorema (Gauss & Markov)

MARKOV

Si consideri la cifra di merito  $J^{(M)}(\theta) = \varepsilon^T \Psi^{-1} \varepsilon = (Y - \Phi \theta)^T \Psi^{-1} (Y - \Phi \theta)$

che è minimizzata da  $\theta^M := (\Phi^T \Psi^{-1} \Phi)^{-1} \Phi^T \Psi^{-1} Y$  (stimatore di MARKOV)

Allora:

NON fanno due che è lo stimatore ML perché ho Hp I2 (NON ho presupposto la gaussianità).

(i)  $\theta^M$  è, tra tutti gli stimatori lineari e non polarizzati, lo stimatore che minimizza  $\text{Var}[\hat{\theta} - \theta^0]$  ( $\rightarrow$  BLUE)

(\*) Stimatore lineare non  $\theta^M$  dipende linearmente da Y

$\rightarrow$  NON è il miglior stimatore possibile (potrebbe essere NON lineare), ma è comodo e tra i lineari è il migliore.

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

OSSERVAZIONI

•  $\text{Var}[V] = \sigma^2 I \Rightarrow \theta^M = \theta^{LS}$

- Si dimostra che (se  $\hat{\sigma}^2$  è incognito)  $\hat{\sigma}^2 = \frac{J^H(\theta^H)}{N-9}$  è uno stimatore NON polarizzato di  $\sigma^2$
  - Usando  $\hat{\sigma}^2$  (o  $\hat{\sigma}^2$ ) posso stimare anche  $\text{Var}[\theta^H] = \sum_{\theta^H} = \hat{\sigma}^2 (\Phi^T \Psi^{-1} \Phi)^{-1}$
  - **ATTENZIONE!** Dato che non conosco la ddp di  $V$ :
    - NON posso dire che  $J^H(\theta^H) \sim \chi^2_{N-9}$  (Niente - validazione del modello!)
    - NON posso calcolare gli intervalli di confidenza (come 10%) o se il param (se non con Tchebycheff). poco preciso  $\Rightarrow$  poco sicuro per i suoi limiti (intervalli troppo grandi).
- Idea: fornire solo le deviazioni standard (SD) dei parametri
- Cade il test F
  - Posso usare FPE (numero basati sulla  $H_0: \mathbb{Z}$ )  $\rightarrow$  a volte  $\rightarrow$  una anche AIC

**STIMA DI BAYES DA FARE!**

Ho info. a priori su  $\theta$

Ipotesi I3:  $\exists$  modello vero  $Y = \Phi(\theta) + V$ ,  $V \sim N(0, \Sigma_V)$ ,  $\Sigma_V > 0$ ,  
 $\theta \sim N(m_\theta, \Sigma_\theta)$ ,  $\Sigma_\theta > 0$  e  $V$  e  $\theta$  indipendenti

TEOREMA: sotto  $H_0$  I3:

- (a)  $\theta^B := \arg \min_{\theta} \{ \epsilon^T \Sigma_V^{-1} \epsilon + (\theta - m_\theta)^T \Sigma_\theta^{-1} (\theta - m_\theta) \}$
- (b)  $\theta^B = (\Phi^T \Sigma_V^{-1} \Phi + \Sigma_\theta^{-1})^{-1} (\Phi^T \Sigma_V^{-1} Y + \Sigma_\theta^{-1} m_\theta)$
- (c)  $\text{Var}[\theta^B] = [\Phi^T \Sigma_V^{-1} \Phi + \Sigma_\theta^{-1}]^{-1}$

$\rightarrow$  somma dei quadrati del residuo ( $\Sigma_V =$  pesata su  $\Sigma_V \neq I$ )  
 valore che ho (es. sondaggi, telefonate)  
 Matrice covarianza che misura l'affidabilità dei dati (pochi dati  $\Rightarrow \Sigma_\theta$  grande  $\Rightarrow \Sigma_\theta^{-1}$  è piccola  $\Rightarrow$  i dati pesano poco).

COMMENTI

- $\Sigma_\theta^{-1} \rightarrow 0 \Rightarrow \theta^B \rightarrow \theta^{ML}$   
 $\rightarrow$  matrice var delle info a priori  
 $\rightarrow$  se  $\rightarrow$  mol. div. che vale poco  $\Rightarrow$  i dati sperimentali valgono poco
- NON è più necessario ipotizzare  $\text{rank}(\Phi) = 9$ . ( $\Sigma_\theta > 0 \Rightarrow (\Phi^T \Sigma_V^{-1} \Phi + \Sigma_\theta^{-1}) > 0$ )  
 La somma di una matrice semi-def pos. e 1 def. pos. è una matrice def. pos.  $\Rightarrow H_0$   $\text{rank}(\Phi) = 9$  NON serve  
 + anche posso fare il inverso  $\rightarrow$  comunque se  $\Sigma_\theta > 0$  anche  $\geq 0$   $\Sigma_\theta^{-1} > 0$

Sfruttando le info a priori, posso avere  $9 > N$  (più incognite che dati)  
 $\rightarrow$  non comunque dati ("pomati")  $\Phi$  rappresentato solo i dati nuovi

VARIABILI INDIPENDENTI AFFETTE DA ERRORE

Problema: relazione tra stature dei padri e dei figli.  $\epsilon = 0$

$x(i) =$  statura  $i$ -esimo padre - statura media popolazione  $(\rightarrow$  non devianza)

$y(i) =$  statura  $i$ -esimo figlio - statura media pop.

Idea:  $\exists$  una relazione del tipo  $y(t) = \theta x(t)$   $\rightarrow$  statura del padre  $\rightarrow$  statura del figlio

$\downarrow$  GALTON 1822-1911

Calcolo la regressione di  $y$  su  $x \rightarrow y(t) = \epsilon_x x(t) + \epsilon_y(t)$   $\rightarrow$  residuo

Soluzione:  $\hat{\theta}_x^{LS} := \frac{\sum_{t=1}^N x(t)y(t)}{\sum_{t=1}^N x(t)^2}$

(NB)  $\rightarrow$   $y$  non 12  $\rightarrow$  regressione  $\rightarrow$   $\epsilon^H = 0$

$$E_x^{LS} = \frac{\sum_{t=1}^N \frac{x(t)y(t)}{N}}{\sum_{t=1}^N \frac{x(t)^2}{N}} \xrightarrow{N \rightarrow \infty} \frac{\text{Cov}[X, Y]}{\text{Var}[X]}$$

Cov. campionaria      Var. campionaria

Hp che da una generazione all'altra NON cambia

$$\approx \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]\text{Var}[Y]}} = \rho_{xy} \text{ coeff. di correlazione}$$

$|\rho_{xy}| \leq 1$  dovrebbe essere  $\geq 0$

$\text{Var}[X] \approx \text{Var}[Y]$

$\Rightarrow 0 < \theta_x^{LS} < 1 \rightarrow$  la popolazione tende verso la media

**regression to mediocrity** ( $\theta_x^{LS} < 1!$ )

il termine regressione deriva da qui e c'è un errore di fondo!

17-5-2001

$$y(t) = \theta x(t)$$

Idea: Regressione di  $y$  su  $x$        $y(t) = \theta_x x(t) + \epsilon_y(t)$

$$\theta_x^{LS} = \frac{\sum x(t)y(t)}{\sum x(t)^2}$$

pedice      residuo

Si trova che  $0 < \theta_x < 1$

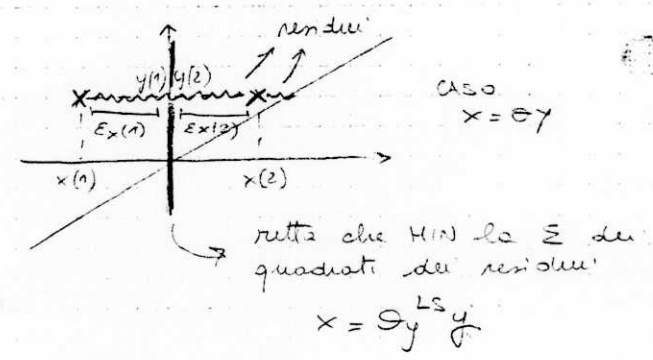
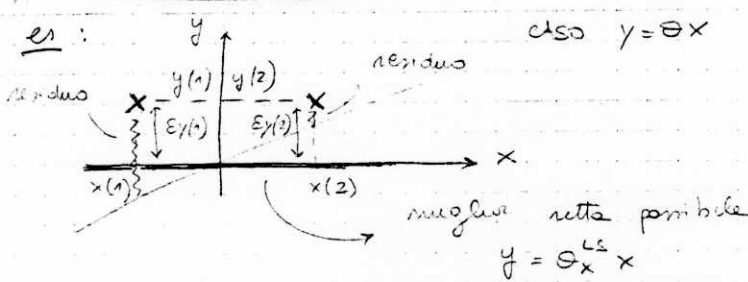
Tuttavia ...

Regressione di  $x$  su  $y$ :  $x(t) = \theta_y y(t) + \epsilon_x(t)$

Soluzione:  $\theta_y^{LS} = \frac{\sum x(t)y(t)}{\sum y(t)^2}$  avendo scambiato  $x$  e  $y$  si pensa che  $\theta_y = \frac{1}{\theta_x}$  ( $y = \theta_x x \rightarrow x = \frac{1}{\theta_x} y$ )  $\rightarrow$  Ma NON è vero!

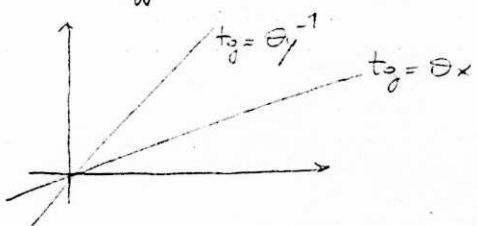
**OSSERVAZIONE** In generale  $\theta_x^{LS} \neq \frac{1}{\theta_y^{LS}} \rightarrow$  i residui sono  $\neq$

Minimizzo cose  $\neq$



la diff. tra le 2 rette di regressione è molto grande (sono  $\perp$ )

Spesso le differenze tra le 2 rette sono meno "drammatiche"



anche  $\theta_y < 1$  cioè anche i pedici tendono verso la mediocrità  $\Rightarrow$  impossibile

Differenza fondamentale tra le 2 regressioni:

Nella 1<sup>a</sup> le  $x$  sono considerate prive di rumore mentre le  $y$  sono rumorose

$$y(t) = \theta_x x(t) + \epsilon_y(t)$$

Invece nella 2<sup>a</sup> regressione le parti si scambiano:

$x(t) = \theta_y y(t) + \epsilon_x(t)$  presupposto sbagliato: avere errore sulla  $x$  o di

Idea: esiste una relazione (legge di natura)  $\tilde{y}(t) = \theta \tilde{x}(t)$

però dispongo solo di misure rumorose

$y(t) = \tilde{y}(t) + \epsilon_y(t)$   
 $x(t) = \tilde{x}(t) + \epsilon_x(t)$

alternativa VERA (teoria) che si potrebbe raggiungere

ME (tutto quanto fatto finora, senza rumori, come se fosse perfettamente noto e conoscere legge di natura)

$\epsilon_y \sim N(0, \sigma_y^2 I_N)$      $\epsilon_x \sim N(0, \sigma_x^2 I_N)$      $\epsilon_y = \begin{bmatrix} \epsilon_y(1) \\ \epsilon_y(2) \\ \vdots \\ \epsilon_y(N) \end{bmatrix}$

Incognite:  $\theta, \tilde{y}, \tilde{x}$      $\tilde{y} = \begin{bmatrix} \tilde{y}(1) \\ \vdots \\ \tilde{y}(N) \end{bmatrix}$

ho molte incognite  
 Tuttavia  $\tilde{y}(t) = \theta \tilde{x}(t) \Rightarrow$  le incognite sono solo  $\theta$  e  $\tilde{x}$      $t=1, \dots, N$

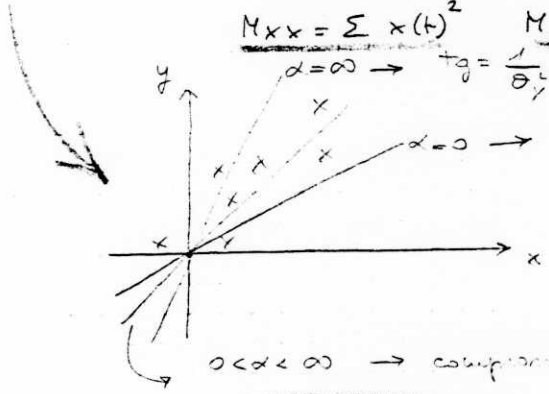
Minimizzo la rovinosità (versione  $\rightarrow$  log  $\rightarrow$  derivato parziali ...)

Risultato:  $\theta^{HL} = \frac{\alpha M_{yy} - M_{xx} + \sqrt{(\alpha M_{yy} - M_{xx})^2 + 4\alpha M_{xy}}}{2\alpha M_{xy}}$

dove  $M_{yy} = \sum y(t)^2$

$M_{xx} = \sum x(t)^2$      $M_{xy} = \sum x(t)y(t)$

$\alpha := \frac{\sigma_x^2}{\sigma_y^2}$      $\rightarrow$  varianza dell'errore su  $x$  / var. errore su  $y$



$\alpha=0 \rightarrow \theta_y = \frac{M_{xy}}{M_{xx}}$      $\rightarrow$  regressione di  $y$  su  $x$

$\alpha=0 \Rightarrow \sigma_x=0 \Rightarrow$  regressione di  $y$  su  $x$

$\alpha=\infty \Rightarrow \sigma_y=0 \Rightarrow$  regressione di  $x$  su  $y$

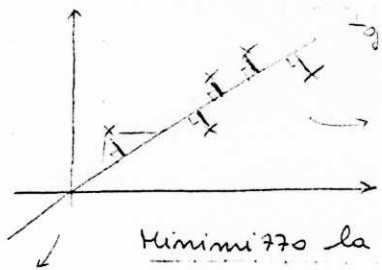
$0 < \alpha < \infty \rightarrow$  compromesso tra le varianze

$\tilde{x}(t)^{HL} = \frac{x(t) + \theta^{HL} \alpha y(t)}{1 + (\theta^{HL})^2 \alpha}$

$\tilde{y}(t)^{HL} = \theta^{HL} \tilde{x}(t)^{HL}$

Interpretazione per  $\alpha=1$

Regr. di  $y$  su  $x \rightarrow$  residui verticali  
 $x$  su  $y \rightarrow$  " orizzontali



residui  $\perp$  alla retta  $\rightarrow$   $\min \sum (\Delta x^2 + \Delta y^2) =$  distanza euclidea della retta dal pto sperimentale

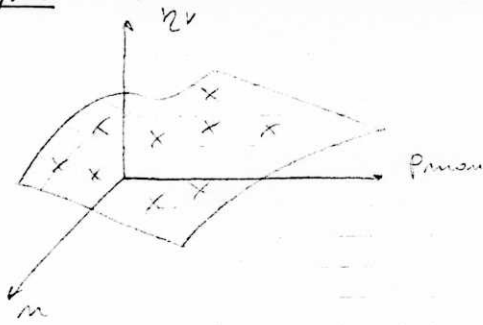
Minimizzo la somma dei quadrati delle distanze dei pti dalla retta

$\ni$  gli intervalli di confidenza

ME Questa categoria di modelli è molto difficile da gestire, anche il modello NON è + lineare nei parametri (...  $\theta \cdot x$  ...). È solito si fanno entrambe le regressioni su  $x$  e su  $y$  e se vengono simili si sa che con un modello + complicato, anche se meglio tra le 2.

# STEPWISE REGRESSION

Esempio: (rendimento volumetrico di un motore a scoppio)



Voglio identificare

$$\eta_v(n, P_{man}) = g(u_1, u_2)$$

Idea: uso un modello scelto all'interno di una classe abbastanza ampia

$g(u_1, u_2) = \sum \theta_i h_i(u_1, u_2)$  dove  $h_i$  sono funzioni e a una certa classe (monomi, sinusoidi, wavelet, ...)

Scelgo:  $h_i$  sono dette regressori

Nel caso della conducibilità termica c'è solo  $1, u_1, u_1^2, u_1^3, u_2$

Se considero monomi di grado  $\leq 3$  i possibili regressori sono:

$$1, u_1, u_2, u_1^2, u_1 u_2, u_2^2, u_1^3, u_1^2 u_2, u_1 u_2^2, u_2^3$$

mettendo insieme questi monomi posso ottenere tante superfici  $\neq$

Modello "naïf":

$$g(u_1, u_2) = \theta_1 \cdot 1 + \theta_2 u_1 + \theta_3 u_2 + \theta_4 u_1^2 + \dots + \theta_{10} u_2^3$$

NON è un buon modello al crescere del n° dei parametri il modello insegue gli errori

Forte rischio di sovrapparametrizzazione se uso tutti i regressori

IDEA: Stimo tutti i modelli possibili e scelgo il "migliore" (F-test, FPE, ...)

Esempio: se mi limito a monomi di grado  $\leq 1$

$$g(u_1, u_2) = \begin{cases} \theta_1 \\ \theta_1 u_1 \\ \theta_1 u_2 \\ \theta_1 + \theta_2 u_1 \\ \theta_1 + \theta_2 u_2 \\ \theta_1 u_1 + \theta_2 u_2 \\ \theta_1 + \theta_2 u_1 + \theta_3 u_2 \end{cases} \quad (\text{regressori: } 1, u_1, u_2)$$

→ tutte le combinazioni possibili dei regressori

PB Quando i regressori sono qualche decina tutti i possibili modelli diventano Troppi

PROBLEMA: uscita esplosiva del n° dei modelli!

A migliore di C

OCCHIO: col Test F se  $A > B$  e  $B > C$  NON ho la garanzia che  $A > C \Rightarrow$  devo fare tutti i confronti

Vediamo una tecnica per ridurre il n° di confronti:

## STEPWISE REGRESSION

anche modello vuoto ( $\Rightarrow y = \text{numero} \rightarrow$  uscita di modelli)

1) Dato un modello di partenza classifica secondo un criterio di importanza i regressori che non fanno parte del modello

(Possibile criterio: riduzione di SSR  $\rightarrow$  quando inserisco il nuovo regressore)

NB Complicazione: il modello SSR diminuisce, ma lo fa in modi  $\neq$  a seconda del modello

2) Considero il 1° di classifica e lo sottopongo ad un Test di ingenuità (FPE, AIC, ...)  $\rightarrow$  se supera inserisco il nuovo regressore. Altrimenti ripeto per il secondo di classifica.



3) Calcolo un Test di espulsione (FPE, AIC, ...) per ciascuno dei regressori del modello. (Togli il regressore, calcola FPE e lo rimpiazzi con FPE vecchi). Se il Test scatta elimino il regressore. Poi procedo a ulteriori test di espulsione nel modello semplificato.

4) Torno al pto. 1)

Con la considerazione non serve fare la classificazione  $\rightarrow$  si prende il regressore che migliora la considerazione e si toglie quello che ha la peggio.

FORWARD REGRESSION: parto da un modello semplice e continuo ad aggiungere finché non migliora.

BACKWARD REGRESSION: modello completo  $\rightarrow$  continuo a togliere finché non mi sento.

OTTIMO

Con questa tecnica non è detto che Trov il modello migliore (che aggiorni il passo alla volta (x0 e il migliore locale))

VANTAGGIO: Posso lavorare con un n' spropositato di regressori fino ad ottenere un modello + semplice. Otengo un modello SUBOTTIMALE.

22-5-2001

Modello lineare

$$Y = \Phi(\theta, \mu) = \Phi(\mu)\theta$$

parametri liberi  
scelto e stimare  
il modello

PB: Cosa succede quando la dipendenza da  $\theta$  NON è lineare?

### STIMA DI MODELLI NON LINEARI NEI PARAMETRI

$$Y = \Phi(\theta) + V \rightarrow \text{errore di misura}$$

è scatole nera è meglio un modello lineare nei parametri.

Caso tipico: identif. a scatola grigia. Leggi finché  $\Rightarrow \Phi(\theta)$

$$\Rightarrow \varepsilon(\theta) := Y - \Phi(\theta)$$

LS:  $\theta^{LS} = \arg \min_{\theta} \varepsilon^T \varepsilon$   $\rightarrow$  MIN la somma dei quadrati dei residui

ML: Ipotesi:  $Y = \Phi(\theta) + V, V \sim N(0, \Sigma_V)$

$\rightarrow \theta^{ML} = \arg \min_{\theta} \varepsilon^T \Sigma_V^{-1} \varepsilon$  somma dei quadrati dei residui pesati  
pesare p una matrice diagonale

Gauss-Markov: Ipotesi:  $E[V] = 0, \text{Var}[V] = \sigma^2 \Psi$  nota

$$\theta^H = \arg \min_{\theta} \varepsilon^T \Psi^{-1} \varepsilon$$

Bayes: Ipotesi:  $V \sim N(0, \Sigma_V), \theta \sim N(m_{\theta}, \Sigma_{\theta})$   $Y = \Phi(\theta) + V$

$$\theta^B = \arg \min_{\theta} \left\{ \varepsilon^T \Sigma_V^{-1} \varepsilon + (\theta - m_{\theta})^T \Sigma_{\theta}^{-1} (\theta - m_{\theta}) \right\}$$

Cosa è cambiato rispetto al caso lineare nei parametri?

Caso lineare =  $\varepsilon = Y - \Phi\theta \rightarrow \frac{d(\varepsilon^T \varepsilon)}{d\theta} = \text{lineare in } \theta = 0$  (che  $\varepsilon^T \varepsilon$  è come una un generato da  $\theta$ ).

$\Rightarrow$  eq. normali

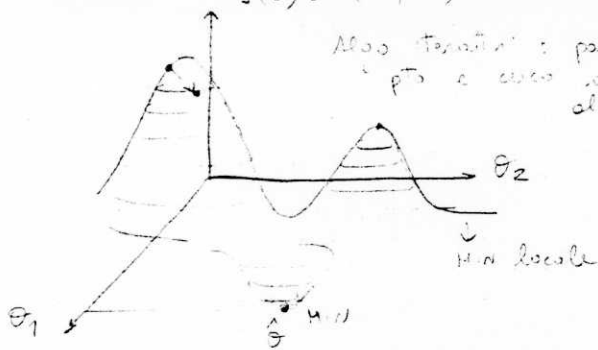
Adesso:  $\varepsilon = Y - \Phi(\theta) \rightarrow \frac{d(\varepsilon^T \varepsilon)}{d\theta}$  non lineare di  $\theta = 0 \rightarrow$  NON trovo più una soluzione esplicita!

Cosa fare?  $J(\theta) = \varepsilon(\theta)^T \varepsilon(\theta)$  Cerco di trovare  $\theta$  che minimizza  $J(\theta)$

attraverso procedure iterative.

Problema

$$J(\theta) = J(\theta_1, \theta_2)$$



Algo iterativo: parto da un pto e cerco di avvicinarmi al MIN in successive

Nel caso lineare nei parametri  $J$  è una "modellata"  $\Rightarrow$  caso + semplice

Supponiamo di aver trovato la stima  $\hat{\theta}$ , allora pono linearizzazione il modello

$$Y = \Phi(\hat{\theta}) + \left. \frac{d\Phi(\theta)}{d\theta} \right|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \dots + V$$

sviluppo di Taylor

il vettore è un gradiente  $\Rightarrow$  vettore riga

Se definisco  $\tilde{Y} := Y - \Phi(\hat{\theta})$ ,  $\hat{\Phi}_\theta := \left. \frac{d\Phi(\theta)}{d\theta} \right|_{\theta=\hat{\theta}}$   $\tilde{\theta} := \theta - \hat{\theta}$  ottengo un modello lineare approssimato

$$\tilde{Y} = \hat{\Phi}_\theta \tilde{\theta} + V \rightarrow \text{ho trascurato i termini di ordine superiore al 1°}$$

$\rightarrow$  vale per  $\|\theta - \hat{\theta}\|$  piccolo

⊛ MATRICE DI SENSITIVITÀ

Posso usare il modello linearizzato per:

- Stimare la var  $\sigma^2$  del disturbo e la var di  $\hat{\theta}$  ( $\Rightarrow$  intervalli di confidenza)

$\rightarrow$  Sostituisco al modello lineare il mod. linearizzato e faccio come al solito

⊙ NON devono essere troppo grandi vs poter finire al di fuori della zona in cui vale la linearità

- Validare il modello (test  $\chi^2$ )

- Effettuare confronti tra modelli (test F, FPE, AIC, HDL...)

**Calcolo della stima**: Algoritmo di GAUSS-Newton

Dobbiamo calcolare  $\hat{\theta} = \arg \min_{\theta} (Y - \Phi(\theta))^T \Psi^{-1} (Y - \Phi(\theta))$

Sia  $\theta^k$  l'approssimazione di  $\hat{\theta}$  al passo  $k$ :

Matrice di sensitività:  $\hat{\Phi}_\theta^k := \left. \frac{d\Phi(\theta)}{d\theta} \right|_{\theta=\theta^k}$   $\tilde{\theta}^k := \theta - \theta^k$

$$\tilde{Y}^k := Y - \Phi(\theta^k)$$

Per lo sviluppo di Taylor:  $\Phi(\theta) \cong \Phi(\theta^k) + \hat{\Phi}_\theta^k (\theta - \theta^k) \Rightarrow$

$$Y - \Phi(\theta) \cong Y - \Phi(\theta^k) - \hat{\Phi}_\theta^k (\theta - \theta^k) = \tilde{Y}^k - \hat{\Phi}_\theta^k \tilde{\theta}^k$$

$\varepsilon(\theta)$  vettore dei residui

$$J(\theta) \cong J^k(\tilde{\theta}^k) := (\tilde{Y}^k - \hat{\Phi}_\theta^k \tilde{\theta}^k)^T \Psi^{-1} (\tilde{Y}^k - \hat{\Phi}_\theta^k \tilde{\theta}^k) \rightarrow \text{vettore } \tilde{\theta}^k \text{ dipende in modo quadratico da } \tilde{\theta}^k \Rightarrow \text{pono MINIMIZZARE } J^k(\tilde{\theta}^k)$$

La minimizzazione di  $J^k(\tilde{\theta}^k)$  è un pb

Standard:

$$\hat{\tilde{\theta}}^k = \left( (\hat{\Phi}_\theta^k)^T \Psi^{-1} \hat{\Phi}_\theta^k \right)^{-1} (\hat{\Phi}_\theta^k)^T \Psi^{-1} \tilde{Y}^k$$

Iterazione base dell'algo G-N

Ricordando che  $\theta = \theta^k + \tilde{\theta}^k$

è logico che  $\theta^{k+1} = \theta^k + \hat{\tilde{\theta}}^k$ , ovvero  $\theta^{k+1} = \theta^k + \left( (\hat{\Phi}_\theta^k)^T \Psi^{-1} \hat{\Phi}_\theta^k \right)^{-1} (\hat{\Phi}_\theta^k)^T \Psi^{-1} (Y - \Phi(\theta^k))$

L'iterazione è molto importante (l'angolo parte da un pto globale)

OSS • Non c'è garanzia di convergenza. Se anche converge può finire in un minimo locale.

lavoro = volta e luogo, se stengo valori  $\theta$   $\rightarrow$  individuare il MN globale caso a caso  $\rightarrow$  si convergono da MN trovati

• Può accadere che  $\det((\Phi_{\theta}^k)^T \Psi^{-1} \Phi_{\theta}^k) \cong 0$   $\rightarrow$  NON è detto che il p.o. MN sia soluto finale (potrebbe accadere solo localmente)

Si modifica l'algoritmo:

$$\theta^{k+1} = \theta^k + ((\Phi_{\theta}^k)^T \Psi^{-1} \Phi_{\theta}^k + \alpha_k I)^{-1} (\Phi_{\theta}^k)^T \Psi^{-1} (Y - \Phi(\theta^k))$$

↓  
Variante di

LEVENBERG - MARQUARDT

$\rightarrow$  no spettrale tra gli autovalori verso  $\alpha \rightarrow 0$   
 $\rightarrow$   $\alpha = 0$  diventano  $> 0$

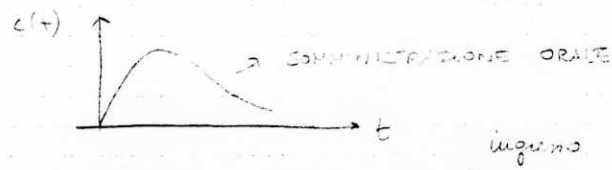
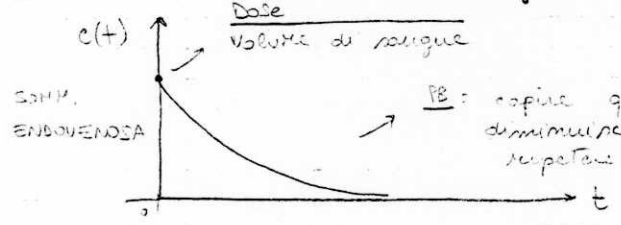
(FE) La  $\alpha$  troppo grande diventa dominante

• Convergenza quadratica nell'intorno del MINIMO:

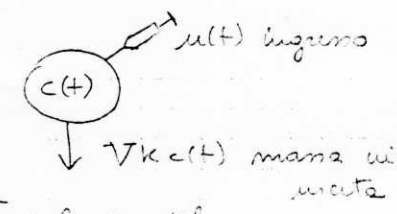
$\|\theta^{k+1} - \theta^k\| \leq \gamma \|\theta^k - \theta^{k-1}\|^2$   $\rightarrow$  NON garantisce la convergenza globale  $\Rightarrow$  spesso in un solo colpo i bridi.

Esempio: Cinetica di un farmaco

$c(t)$ : concentrazione del farmaco



$$c(t) = \frac{m(t)}{V}$$



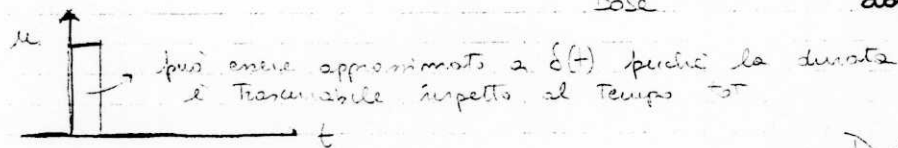
$V =$  volume del compartimento

Idea: sangue come 1 solo compartimento  $\rightarrow$  omogeneo

$\dot{m}(t) = -km(t) + u(t)$   
Variazione di massa  $\rightarrow$  massa in uscita

$$\dot{c}(t) = -kc(t) + \frac{1}{V} u(t)$$

$u(t) = D \delta(t)$  (iniezione endovenosa di una dose di massa  $D$ ).

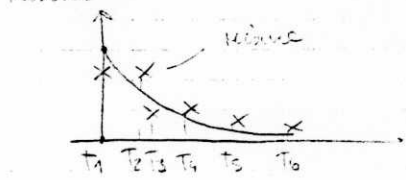


$$c(0^-) = 0$$

$y_i = c(t_i) + (v_i)$  errore di misura

$$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} k \\ V \end{bmatrix}$$

$D$  è nota



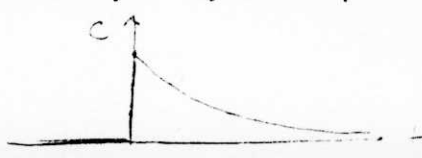
La varianza degli errori di misura è costante  $\rightarrow$  dipende dalla concentrazione (di prop.)

Errori di misura:  $H_p \frac{\sqrt{\text{Var}[v_i]}}{c(t_i)} = \text{cost} = \sigma$   $\Rightarrow$  l'errore percentuale è cost

$\Rightarrow \text{Var}[v_i] = \sigma^2 c(t_i) \cong \sigma^2 y_i$   $\rightarrow \text{Var}[V] = \sigma^2 \begin{bmatrix} y_1^2 & y_2^2 & & 0 \\ & y_2^2 & & \\ & & \dots & \\ 0 & & & y_N^2 \end{bmatrix} = \sigma^2 \Psi$

Per trovare  $\Phi(\theta)$  devo risolvere l'eq. differenziale. (basta porre  $c(0^+) = D/V$  e poi risolvere  $\dot{c} = kc$ )

$$c(t) = \frac{D}{V} e^{-kt} \quad t \geq 0$$



$$\Phi(\theta) = \begin{bmatrix} c(t_1) \\ \vdots \\ c(t_N) \end{bmatrix} = \begin{bmatrix} \frac{D}{V} e^{-k t_1} \\ \vdots \\ \frac{D}{V} e^{-k t_N} \end{bmatrix} = D \begin{bmatrix} \frac{1}{\theta_1} e^{-\theta_2 t_1} \\ \vdots \\ \frac{1}{\theta_1} e^{-\theta_2 t_N} \end{bmatrix}$$

Per usare  $\theta_j - N$  suve  
 $\frac{\partial}{\partial \theta} \Phi(\theta)$

$$\frac{d}{d\theta} \left( \frac{D}{\theta_1} e^{-\theta_2 t} \right) = \begin{bmatrix} -\frac{1}{\theta_1^2} e^{-\theta_2 t} & -\frac{t}{\theta_1} e^{-\theta_2 t} \end{bmatrix}$$

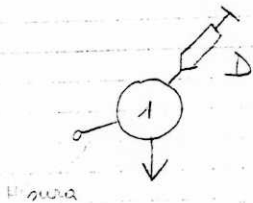
$$\Phi_g^k = D \begin{bmatrix} -\frac{1}{(\theta_1^k)^2} e^{-\theta_2^k t_1} & -\frac{t_1}{\theta_1^k} e^{-\theta_2^k t_1} \\ \dots & \dots \\ -\frac{1}{(\theta_1^k)^2} e^{-\theta_2^k t_N} & \dots \end{bmatrix}$$

(N x 2)

23-5-2001

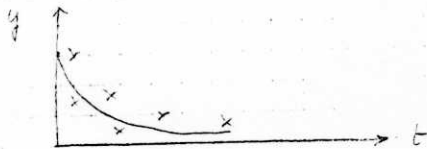
Esempio: Scelta tra 2 modelli.

Modello con 1 compartimento

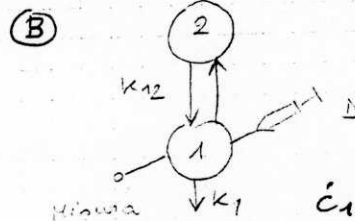


(A)

$$\dot{c}_1 = -k_1 c_1 + \frac{\mu}{V_1}$$



Modello con 2 compartimenti



(B)

Si usa la compensazione della massa

$$\dot{c}_1 = -k_1 c_1 + k_{12} (c_2 - c_1) + \frac{\mu}{V_1}$$

$$\dot{c}_2 = -k_{12} (c_2 - c_1)$$

Scambio tra i compartimenti  
 2 è isolato dal resto del mondo a parte 1

Facendo i conti si vede che, per il secondo modello (B):

$$c_1(t) = D(a e^{-\alpha t} + b e^{-\beta t}) \quad \text{dove } a, b, \alpha, \beta \text{ sono opportune funzioni di } k_1, k_{12} \text{ e } V_1$$

Per il modello A:  $c_1(t) = D a e^{-\alpha t}$       $a = \frac{1}{V}$ ,  $\alpha = k_1$

Sono 2 modelli gerarchici perché A è un sottocaso di B (si ottiene con  $b=0$ ).

Il problema si riduce alla scelta tra i modelli:

(A)  $c_1(t) = D a e^{-\alpha t}$       $\theta_A = \begin{bmatrix} a \\ \alpha \end{bmatrix}$

(B)  $c_1(t) = D(a e^{-\alpha t} + b e^{-\beta t})$       $\theta_B = \begin{bmatrix} a \\ \alpha \\ b \\ \beta \end{bmatrix}$

è una delle hp x applicazione tecniche x la scelta

Come fare? Stimo  $\hat{\theta}_A$  e  $\hat{\theta}_B$ , (linearizzo x gli interv. di confidenza)

⇒ F-Test, AIC, FPE, MDL ... ⇒ decido quale è il migliore.

INIZIA LA 3<sup>a</sup> PARTE → 2<sup>o</sup> COMPITINO

LOCIDI

PROCESSI CASUALI STAZIONARI

