10 Maggio 2021

STIMA LS.

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

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

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 $\Phi^T\Phi x = 0 \Rightarrow$

$x^T\Phi^T\Phi x = 0 \Rightarrow \left((\Phi x)^T\Phi x\right) = \|\Phi x\|^2 \Rightarrow \Phi x = 0 \Rightarrow$ $\underset{N\times q}{\Phi}$ ha colonne linearmente dipendenti

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

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

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

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

Sistema normale

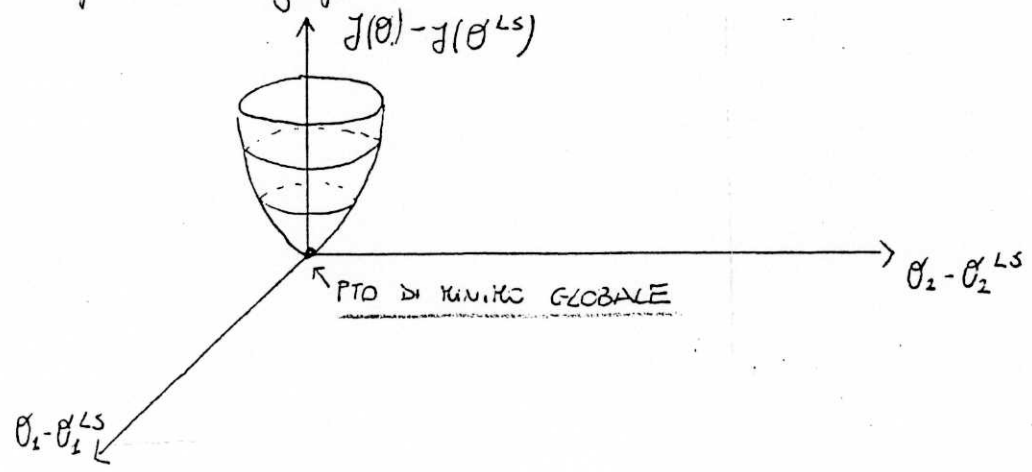
$$\boxed{\theta = (\Phi^T\Phi)^{-1}\Phi^T Y}$$

Hessiana $\dfrac{d^2J(\theta)}{d\theta^2} = 2\Phi^T\Phi \geq 0$ DIADE semidefinita positiva.

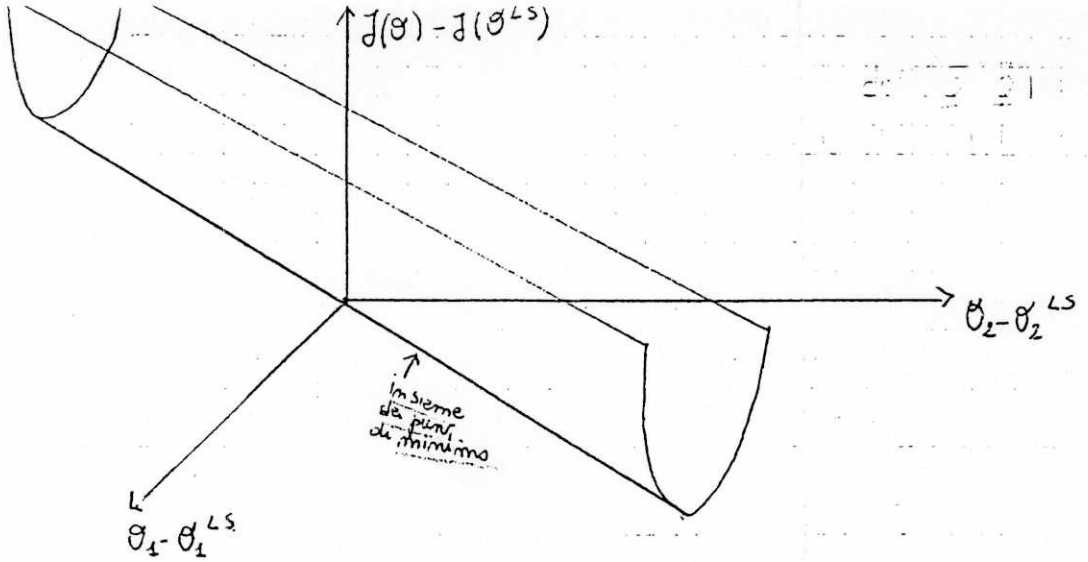
Proposta $(A \geq 0 \text{ e } \det(A)\neq 0) \Rightarrow A > 0$.

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

Interpretazione grafica

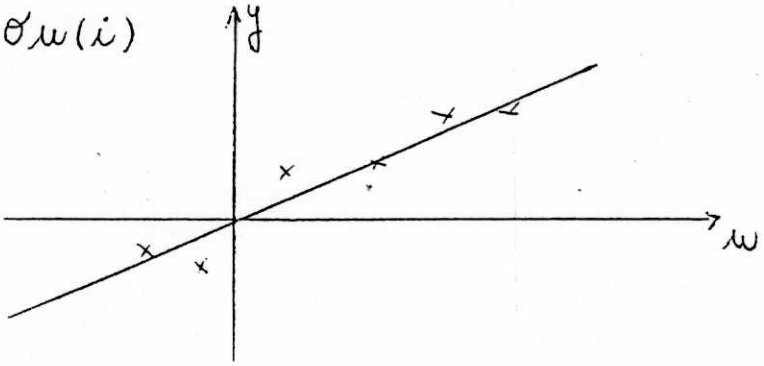


$\det(\Phi^T\Phi)\neq 0$ $\exists$ un'unica soluzione



Esempio: Regressione lineare con $q = 1$.

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



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

$$y(1) = u(1)\theta$$
$$y(2) = u(2)\theta$$

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

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

Il problema dell'identificabilità

Cosa succede se $\text{rank}(\bar{\Phi}) < q$?

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

$$y(t) = \theta_1 u_1(t) + \theta_2 u_2(t) + \theta_3 u_3(t)$$

$$\bar{\Phi} = \begin{bmatrix} u_1(1) & u_2(1) & u_3(1) \\ u_1(2) & u_2(2) & u_3(2) \\ \vdots & \vdots & \vdots \\ u_1(N) & u_2(N) & u_3(N) \end{bmatrix}$$

Supp che $\text{rank}(\bar{\Phi}) = 2 \Rightarrow$ una delle colonne di $\bar{\Phi}$ è combinazione lineare delle altre, ovvero esistono α, β t.c. $u_3(t) = \alpha u_1(t) + \beta u_2(t)$

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

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

$\Rightarrow w_3$ è inutile perché posso ottenere la stessa previsione y usando solo $w_1(t), w$

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 $w_3 \Rightarrow$ da solve

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

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

DI CONDIZIONAMENTO: è un indice del grado di singolarità di una matrice.
Matrice che tende alla singolarità
↳ determinante della matrice che tende a zero

LIMITI DELLA STIMA LS

- affidabilità della stima
- il modello è giusto?
- confronto tra modelli

↓ per dare della risposte

Teoria della stima

STIMA ML

Devo fare ipotesi sulla d.d.p dell'errore di misura $V := Y - \Phi\theta$ $\quad V = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}$

Ipotesi I1 $\quad Y = \Phi\theta + V$, $V \sim N(0, \Sigma_v)$, $\Sigma_v > 0$

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

varianza $\Sigma_v = \sigma^2 I$

Teorema: Se vale l'ipotesi I1 e $\text{rank}(\Phi) = q$, allora → SOMMA DEI QUADRATI DEI RESIDUI (SSR) PESATI

a) $\theta^{ML} = \arg\min_{\theta} J^{ML}(\theta)$, $J^{ML}(\theta) = \varepsilon^T \Sigma_v^{-1} \varepsilon$

b) $\theta^{ML} = (\Phi^T \Sigma_v^{-1} \Phi)^{-1} \Phi^T \Sigma_v^{-1} Y$ cong. lineare

($\Rightarrow \theta^{ML}$ è gaussiano)

c) $E[\theta^{ML}] = \theta$ (stimatore non polarizzato)

d) $\mathrm{Var}[\theta^{HL}] = \left(\bar{\Phi}^T \Sigma_V^{-1} \bar{\Phi}\right)^{-1}$

Dim .....

Commento

Ipotesi forti (V gaussiana)

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

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

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

Relazione con L.S.

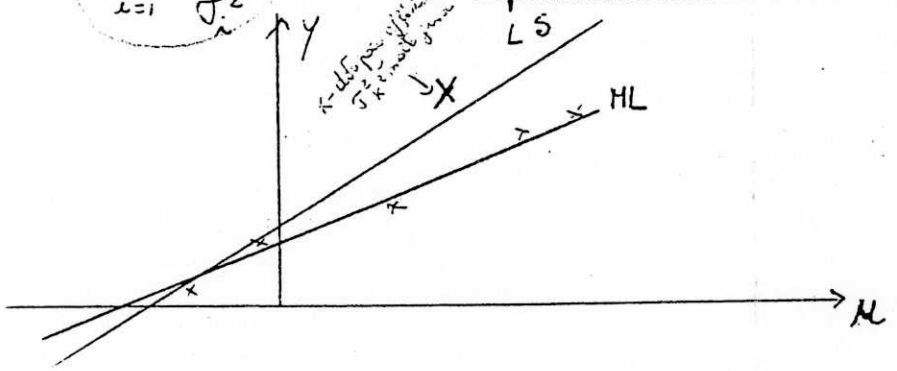
Se $\Sigma_V = \sigma^2 I \Rightarrow \theta^{HL} = \theta^{LS}$

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

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

$$J^{HL}(\theta) = \varepsilon^T \Sigma_V^{-1} \varepsilon = [\varepsilon_1 \varepsilon_2 \dots \varepsilon_N] \begin{bmatrix} \frac{1}{\sigma_1^2} & & 0 \\ & \frac{1}{\sigma_2^2} & \\ 0 & & \ddots \\ & & & \frac{1}{\sigma_N^2} \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{bmatrix} =$$

$= \sum_{i=1}^{N} \frac{\varepsilon_i^2}{\sigma_i^2}$ (somma pesata dei quadrati degli errori residui)



Vogliamo trovare gli intervalli di confidenza

STIMA DELLA VARIANZA DEL DISTURBO

Spesso Σ_V è nota a meno di un coefficiente moltiplicativo ($\Sigma_V = \sigma^2 I$)

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

Allora

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

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

$$J_\psi^{HL} = \mathcal{E}^T(\theta)\,\psi^{-1}\,\mathcal{E}(\theta)$$

NOTA: $(\sigma^2)^{HL}$ è polarizzato

Per avere una stima non polarizzata

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

STIMA ML: INTERVALLI DI CONFIDENZA

(a) Σ_V nota.

$\theta^{HL} \sim N(\theta, \Sigma_{\theta_{HL}})$ e sono gaussiane, $\Sigma_{\theta_{HL}} = (\Phi^T \Sigma_V^{-1} \Phi)^{-1}$

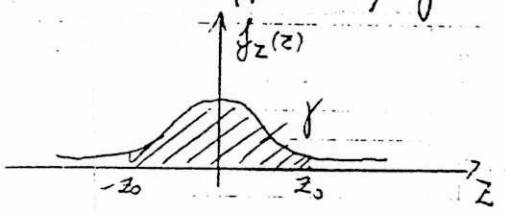
anche le marginali sono gaussiane

⇓

intervalli di confidenza per uno stimatore gaussiano con varianza nota.

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

$$\sigma^2_{\theta_i^{HL}} = \left[\Sigma_{\theta^{HL}}\right]_{ii} \qquad z_0 \text{ è t.c. } P(|Z| \le z_0) = \gamma$$



11 maggio 2001

(b) $\Sigma_V = \sigma^2 \psi$ (ψ matrice nota, σ^2 scalare incognito)

Stimare

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

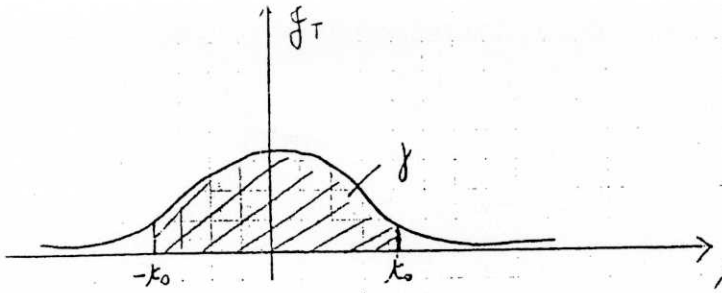
Proprietà: Risulta che $\dfrac{\theta_i^{HL} - \theta_i}{\hat{\sigma}_{\theta_i^{HL}}} \sim T_{N-q}$ (t di Student)

N: n° dei dati

q: n° dei parametri

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

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



APPLICAZIONE: IL PROBLEMA DELLA REGRESSIONE LINEARE

$$y(t) = \theta_1 u_1(t) + \theta_2 u_2(t) + \dots + \theta_q u_q(t) + v(t) \quad t = 1, \dots, N$$

$$\mathrm{Var}[V] = \sigma^2 I \qquad V = \begin{bmatrix} v(1) \\ \vdots \\ v(N) \end{bmatrix}$$

Se molte V è gaussiana

$\Downarrow$

$$\theta^{ML} = \theta^{LS}$$

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

ponendo $\varphi(t) = [u_1(t) \; u_2(t) \; \dots \; u_q(t)]^T$, si vede che $\theta^{LS} = (\Phi^T \Phi)^{-1} \Phi^T Y$

vettore dei regressori

$$= \left[\sum_{t=1}^{N} \varphi(t)\varphi(t)^T\right]^{-1} \left[\sum_{t=1}^{N} \varphi(t) y(t)\right]$$

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

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

$$y(t) = \varphi(t)^T \theta + v(t)$$

Stimo $\mathrm{Var}[\theta^{ML}]$ come $\hat{\Sigma}_{\theta^{ML}} = \hat{\sigma}^2 \left[\sum_{t=1}^{N} \varphi(t)\varphi(t)^T\right]^{-1}$

Spero i risultati sono espressi come segue:

$$y(t) = \theta_1^{ML} u_1(t) + \theta_2^{ML} u_2(t) + \dots + \theta_q^{ML} u_q(t) + \varepsilon(t)$$

(ho espresso ai parametri stimati)

$$\varepsilon(t) = y(t) - \varphi(t)^T \theta^{ML}$$

$$v(t) = y(t) - \varphi(t)^T \theta$$

$$\left[\hat{\sigma}_{\theta_1^{ML}}\right] \quad \left[\hat{\sigma}_{\theta_2^{ML}}\right] \quad \left[\hat{\sigma}_{\theta_q^{ML}}\right] \quad \left[\hat{\sigma}\right]$$

Regola per capire quali regressori vanno tolti del modello [$\exists$ variabili che non influenzano il modello

Problema inferenziale: come faccio ad essere sicuro che $\theta_K \neq 0$? (Se non sono sicuro può essere meglio togliere u_K dal modello)

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

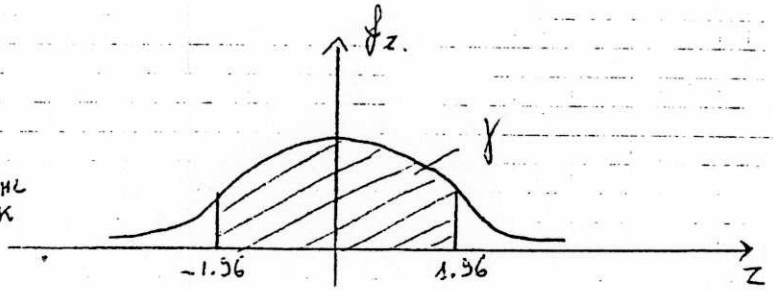
$$\theta_K = 0 \Rightarrow \frac{\theta_K^{HL} - E(\theta_K^{HL})}{\sqrt{Var[\theta_K^{HL}]}} \sim Z \Rightarrow \frac{\theta_K^{HL}}{\sigma_K^{HL}} \sim Z$$

$$\left(\frac{\theta_K^{HL}}{\hat{\sigma}_K^{HL}} \sim T_{N-q} \right)$$

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

$$\left| \frac{\theta_K^{HL}}{\sigma_{\theta_K^{HL}}} \right| \leq 1.96$$

Idea: Se $|\theta_K^{HL}| \leq 1.96\, \sigma_{\theta_K^{HL}}$



non c'è niente di assurdo se non riesco a smentire l'ipotesi $\theta_K = 0$

Invece se $|\theta_K^{HL}| > 1.96\, \sigma_{\theta_K^{HL}}$

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

$\Rightarrow$ "ASSURDO" (poco probabile)

$\Rightarrow$ respingo l'ipotesi $\theta_K = 0$ $\Rightarrow$ "il parametro θ_K è SIGNIFICATIVAMENTE $\neq 0$"

CONCLUSIONE

$|\theta_K^{HL}| > 2\, \sigma_{\theta_K^{HL}}$ sono "sicuro" che $\theta_K \neq 0$

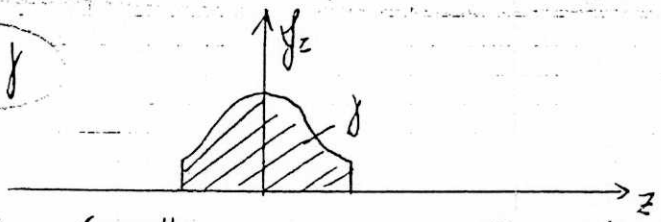
$|\theta_K^{HL}| < 2\, \sigma_{\theta_K^{HL}}$ non so garantire che $\theta_K \neq 0$

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 $\theta_K \neq 0$

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



Test di ipotesi (soggettivo: ognuno può scegliere un "suo" γ)

VALIDAZIONE TEST χ^2

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

$$y = \Phi\theta^0 + V \quad , \quad \mathrm{Var}[V] = \sigma^2 I$$

li descrive adeguatamente?

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

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

Se anche σ^2 è ben stimabile che σ^2 e la varianza campionaria dei residui abbiano lo stesso ordine di grandezza (idea anche 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 I1 $\quad y = \Phi\theta^0 + V \qquad V \sim N(0, \sigma^2 I)$

Teorema sotto I1, dividendo per σ^2 la somma dei quadrati dei residui della stima LS si ottiene una V.C di tipo χ^2 con $N-q$ g.d.l"

(N n° dati $\quad$ q n° parametri)

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

$\exists$ TAB per χ^2

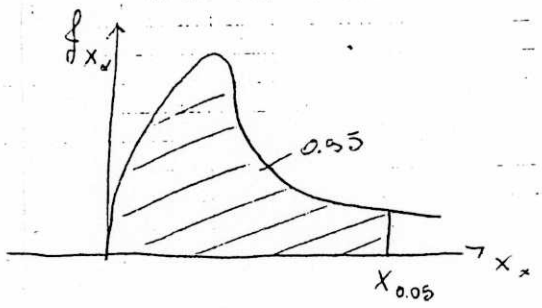
95% dei casi $\quad \frac{\varepsilon^T\varepsilon}{\sigma^2} < 14.07$

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

- Fissare un livello di signif. α
- Cercare su TAB il valore $X_\alpha \Rightarrow$

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

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



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

Basta usare $\varepsilon^T\Sigma_V^{-1}\varepsilon$

Punti deboli 1) può essere difficile capire quali sono i motivi per cui viene scartato il modello:

- $Y = \Phi\theta^* + V$ spiega male i dati
- V non è gaussiana
- Il valore di σ^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)

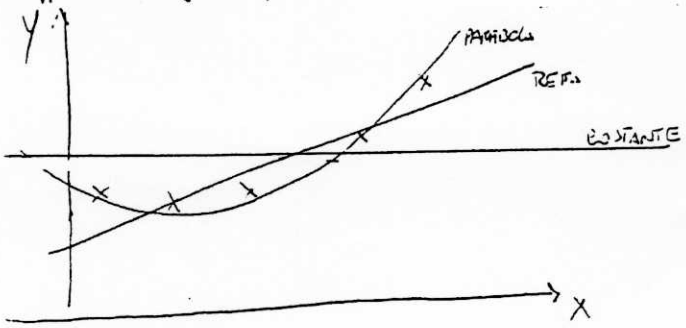
TEST F

Confronto tra modelli

MODELLI "MATRIOSKA" Una sequenza di classi di modelli in cui ciascuna classe comprende una intorno (come casi particolari) le classi precedenti

Es. M_1 = [rette] M_2 = [parabole] M_3 = [cubiche] ....

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 con LS

poi Fra modelli sceglo quello che minimizza SSR (la somma dei quadrati dei residui)

FATTO: SSR decresce sempre al crescere dell'ordine del modello (dato che LS ci garantisce la minimizzazione di SSR, non è possibile che la migliore parabola abbia una SSR più alta della SSR della migliore retta)

Usare il minimo di SSR conduce a scegliere sempre il modello più complesso (per esempio $N=100$ $\Rightarrow$ polinomio di ordine 99) 100 punti e 100 parametri passa per tutti i punti

- Che male c'è nell'eccedere con il numero di parametri?
  - Nessun pb se i dati fossero privi di rumore.
    es. Trovo 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 (riproduce 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 $SSR_K < SSR_{K-1}$

Scelgo M_K se se SSR_K è "molto più piccola" di SSR_{K-1}

Teorema Sotto l'ipotesi I1 $\left(Y = \Phi\theta^* + V \quad V \sim N(0, \sigma^2 I)\right)$

$$f = (N-K)\frac{SSR_{K-1} - SSR_K}{SSR_K} =$$ è una F di Fisher con $(1, N-K)$ g. di l.

- Per $N-K$ grande $F(1, N-K) \to \chi^2(1)$ (tende)

TAB F di Fisher.

Test F: Fissato un livello di significatività α (tipicamente 0.05) cerco f_α t.c. $P(F(1, N-K) \ldots) = 0.95$

$f < f_\alpha$ scelgo M_{K-1}

$f > f_\alpha$ scelgo M_K

- Non è necessario conoscere σ^2
- è soggettivo (va scelto il livello di significatività)

$V \sim N(0, \sigma^2 \Psi)$ Ψ matrice nota $\Rightarrow \varepsilon^T \Psi^{-1} \varepsilon$ al posto di $SSR = \varepsilon^T \varepsilon$

LIMITI

- Presuppone l'esistenza di un modello vero
- Si applica solo a modelli matrioska.

pag 30: Parametrizzazione: servono abbastanza dati

FPE e AIC hanno una prob. > 0 di sovrastimare il modello

⇒ è meglio usare + di un criterio x valutare un modello

APPLICAZIONE: REGRESSIONE LINEARE

$q = 1, 2$

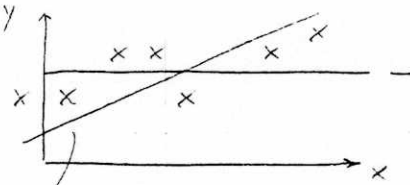
Dati: $y(1), y(2) \dots y(N)$ $\quad$ Hp: $V \sim N(0, \sigma^2 I)$ $\quad$ $V = \begin{bmatrix} v(1) \\ v(2) \\ \vdots \\ v(N) \end{bmatrix}$

$x(1), x(2) \dots x(N)$

2 modelli alternativi:

(a) $y(t) = \bar{\theta}_1 + v(t)$

(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)

Stimiamo (a):

$Y = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix}$ $\quad$ $\Phi = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ (matrice Φ) $\quad$ $\bar{\theta}_1^{ML} = (\Phi^T \Phi)^{-1} \Phi^T Y = N^{-1} \sum_{i=1}^{N} y(i) = M_Y$ (Media campionaria)

(manca una matrice che in questo caso è $= I$)

(b) $\Phi = \begin{bmatrix} 1 & x(1) \\ 1 & x(2) \\ \vdots & \vdots \\ 1 & x(N) \end{bmatrix}$ $\quad$ $\theta^{ML} = \begin{bmatrix} N & \Sigma x(i) \\ \Sigma x(i) & \Sigma x(i)^2 \end{bmatrix}^{-1} \begin{bmatrix} \Sigma y(i) \\ \Sigma x(i) y(i) \end{bmatrix}$

↓ 2 parametri θ_1^{ML} e θ_2^{ML}

$\theta_1^{ML} = M_Y - \theta_2^{ML} M_X$

$\theta_2^{ML} = \dfrac{S_{XY}}{S_{XX}}$ $\quad$ dove $\quad$ $S_{XY} = \dfrac{\sum_1^N (x(i) - M_X)(y(i) - M_Y)}{N}$ $\quad$ $S_{XX} = \dfrac{\sum_1^N (x(i) - M_X)^2}{N}$

(S_{XY}) è la COVARIANZA CAMPIONARIA

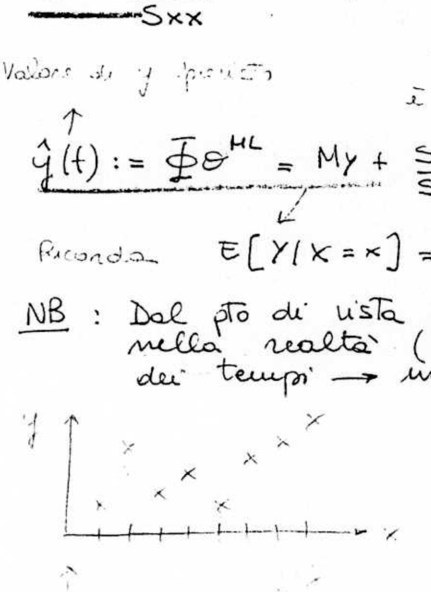
Valore di y predetto

$\hat{y}(t) := \Phi \theta^{ML} = M_Y + \dfrac{S_{XY}}{S_{XX}} (x(t) - M_X)$

è come se avessimo sostituito ai 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 VC e NON.

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

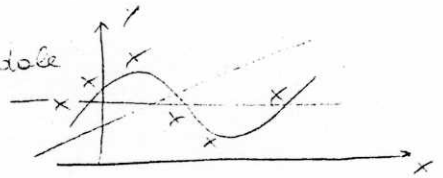
Devo decidere se θ_2 serve o meno:

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

Vd. Tab di FISHER

× fare il Test 2 bisogna Hp la gaussianità $\Rightarrow$ ML

Ricordiamo: la dipendenza potrebbe essere rimuovibile $\Rightarrow$ tra cost e retta scelgo la retta ML non è quella mia



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

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

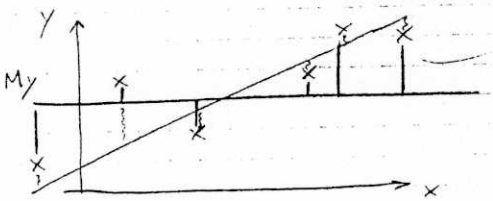
Scarti attorno alla retta di regressione

Scarti attorno a M_Y

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

$$0 \le R^2 \le 1$$

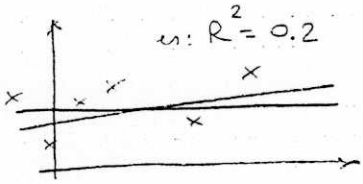
Sempre ≤ 1 (vale sopra il modello è + complicato).



M_Y = scarti attorno a M_Y

mm = scarti attorno alla retta

- $R^2 \cong 0$ (scarti quasi sovrapponibili).

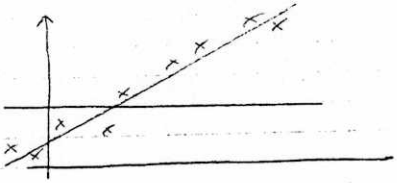


es: $R^2 = 0.2$

migliora poco passando da (a) a (b).

- $R^2 \cong 1$

es. $R^2 = 0,91$



il modello (b) è molto meglio

B est	Se rimuovo l'ipotesi di gaussianità, cosa posso dire?
L inear	IPOTESI I2: $Y = \Phi\theta + V$ $\rightarrow$ $\exists$ un modello vero
U nbiased	$E[V] = 0$ $Var[V] = \sigma^2 \Psi$
E stimator	(σ^2 scalare eventualmente incognito) $\square$

Teorema (Gauss & Markov)

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

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

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

Allora:

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

(*) stimatore lineare nel senso θ^M dipende linearmente da Y

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

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

OSSERVAZIONI

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

- Si dimostra che (se σ^2 è incognito) $\hat{\sigma}^2 = \frac{J^M(\theta^M)}{N-q}$ è uno stimatore NON polarizzato di σ^2
- Usando σ^2 (o $\hat{\sigma}^2$) posso stimare anche $\text{Var}[\theta^M] = \Sigma_{\theta^M} = \hat{\sigma}^2 (\Phi^T \Psi^{-1} \Phi)^{-1}$
- ATTENZIONE! Dato che non conosco la ddp di V:
  - NON posso dire che $J^M(\theta^M) \sim \chi^2_{N-q}$ (Niente + validazione del modello!)
  - NON posso calcolare gli intervalli di confidenza (perché non so se il parabl. (se non con Tchebycheff). → poco preciso ⇒ poco usato (intervalli troppo grandi)

Idea: fornire solo le deviazioni standard (SD) dei parametri

- Cade il test F
- Posso usare FPE (perché basato sulla Hp I2) → a volte in una anche AIC

STIMA DI BAYES DA FARE!

Ho info. a priori su θ → vettore di VC

Ipotesi I3: ∃ modello vero $Y = \Phi(H) + V$, $V \sim N(0, \Sigma_V)$, $\Sigma_V > 0$ (def pos.), $(H) \sim N(m_\theta, \Sigma_\theta)$, $\Sigma_\theta > 0$ e V e θ indipendenti

TEOREMA: Sotto Hp I3:

(a) $\theta^B := \arg\min_\theta \{ (\varepsilon^T \Sigma_V^{-1} \varepsilon) + (\theta - m_\theta)^T \Sigma_\theta^{-1} (\theta - m_\theta) \}$ → somma dei quadrati dei residui ($\Sigma_V = I$ → pesata se $\Sigma_V \neq I$)

(b) $\theta^B = (\Phi^T \Sigma_V^{-1} \Phi + \Sigma_\theta^{-1})^{-1} (\Phi^T \Sigma_V^{-1} Y + \Sigma_\theta^{-1} m_\theta)$ → valore che ho (es: risultati sondaggi telefonici)

(c) $\text{Var}[\theta^B] = [\Phi^T \Sigma_V^{-1} \Phi + \Sigma_\theta^{-1}]^{-1}$ → Matrice varianza che misura l'affidabilità dei dati (pochi dati ⇒ Σ_θ grande ⇒ Σ_θ^{-1} è piccola ⇒ i dati pesano poco).

COMMENTI

- $\Sigma_\theta^{-1} \to 0 \Rightarrow \theta^B \to \theta^{ML}$ (matrice Var della info a priori; se → 0 vuol dire che vale poco ⇒ i dati sperimentali valgono poco)
- NON è più necessario ipotizzare $\text{rank}(\Phi) = q$ ($\Sigma_\theta > 0 \Rightarrow (\Phi^T \Sigma_V^{-1} \Phi + \Sigma_\theta^{-1}) > 0$)

  La somma di una matrice semi def pos e 1 def pos è una matrice def pos ⇒ Hp $\text{rank}(\Phi) = q$ NON serve e che posso fare è inverso. ($\Phi^T \Sigma_V^{-1} \Phi$ è comunque ≥ 0; se $\Sigma_\theta > 0$ anche $\Sigma_\theta^{-1} > 0$)

Sfruttando le info a priori, posso avere $q > N$ (più incognite che dati) → sono comunque dati ("pomati") Φ rappresenta solo i dati misuri

VARIABILI INDIPENDENTI AFFETTE DA ERRORE

Problema: relazione tra statura dei padri e dei figli.

$x(i)$ = statura i-esimo padre − statura media popolazione (→ mean centered) ($E = 0$)

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

Idea: ∃ una relazione del tipo $y(t) = \theta x(t)$ → scostamento dell'altezza del padre dalla media

↓ GALTON 1822-1911 → scostamento del figlio

Calcolo la regressione di y su x → $y(t) = \theta_x x(t) + \varepsilon_y(t)$ → residuo

Soluzione: $\theta_x^{LS} := \frac{\sum_{t=1}^N x(t) y(t)}{\sum_{t=1}^N x(t)^2}$ (Hp garantisce che θ^{ML})

NB

Cov. campionaria =

$$\theta_x^{LS} = \frac{\sum_{t=1}^{N} \frac{x(t)\,y(t)}{N}}{\sum_{t=1}^{N} \frac{x(t)^2}{N}} \xrightarrow{N\to\infty} \frac{Cov[x,y]}{Var[x]}$$

(Var. campionaria)

Hp che da una persona all'altra non cambi

$$\simeq \frac{Cov[x,y]}{\sqrt{Var[x]\,Var[y]}} = \rho_{xy} \text{ coeff. di correlazione}$$

$Var[x] \cong Var[y]$

$|\rho_{xy}| \le 1$ (dovrebbe essere ≥ 0)

$\Rightarrow 0 < \theta_x^{LS} < 1$ → 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

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

predice

Idea: Regressione di y su x $\quad y(t) = \theta_x\, x(t) + \varepsilon_y(t)$ (residuo)

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

Si trova che $0 < \theta_x < 1$

Tuttavia ...

predice

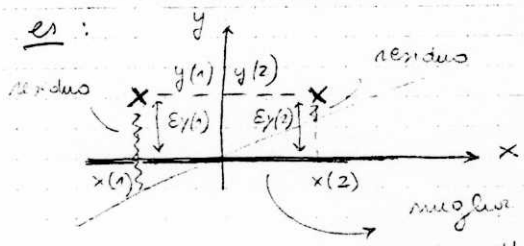
Regressione di x su y: $x(t) = \theta_y\, y(t) + \varepsilon_x(t)$

Soluzione: $\theta_y^{LS} = \frac{\sum x(t)\,y(t)}{\sum y(t)^2}$ avendo scambiato x e y si pensa che $\theta_y = \frac{1}{\theta_x}$ ($y = \theta x \to x = \frac{1}{\theta} y$) → NB non è vero!

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

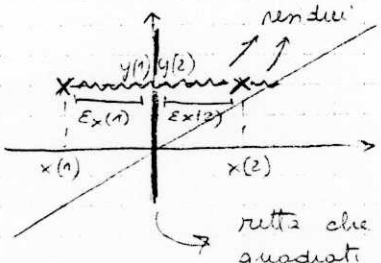
Minimizzo cose ≠

es:



caso $y = \theta x$

unica retta possibile $y = \theta_x^{LS} x$



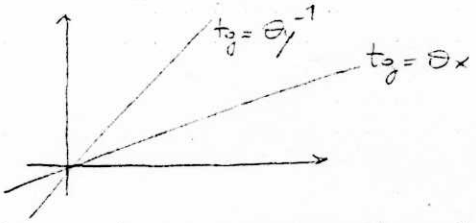
caso $x = \theta y$

retta che MIN la Σ dei quadrati dei residui

$x = \theta_y^{LS} y$

la diff tra le 2 rette di regressione è molto grande (sono ⊥)

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



anche $\theta_y < 1$ cioè anche i padri tendono verso la mediocrità ⇒ impossibile

Differenza fondamentale tra le 2 regressioni:

Nella 1° le x sono considerate prive di rumore mentre le y sono rumorose

$$y(t) = \theta_x\, x(t) + \varepsilon_y(t)$$

Invece nella 2ª regressione le parti si scambiano:

$$x(t) = \theta_y\, y(t) + \varepsilon_x(t)$$ Presupposto sbagliato: [illegible] esatta di x o di y

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

però dispongo solo di misure rumorose

$$y(t) = \tilde{y}(t) + \varepsilon_y(t)$$
$$x(t) = \tilde{x}(t) + \varepsilon_x(t)$$

} altezze reali ma affette da rumore…

altezza VERA (teorica) che si potrebbe raggiungere

NB (tutto questo fatto fuori, senza rumore, serve x fare previsioni, non x conoscere leggi di natura)

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

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

↳ ho molte + incognite

Tuttavia $\tilde{y}(t) = \theta\, \tilde{x}(t)$ ⇒ le incognite sono solo θ e $\tilde{x}$ $\qquad t = 1, \dots N$

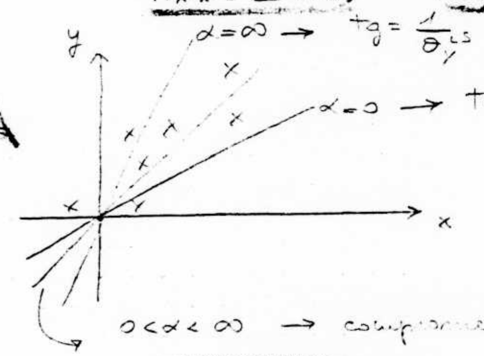
Massimizzo la verosimiglianza …. (verosim → log → derivate parziali …)

Risultato:

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

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

$M_{xx} = \sum x(t)^2 \qquad M_{xy} = \sum x(t) y(t) \qquad \alpha := \dfrac{\sigma_x^2}{\sigma_y^2}$ → varianza dell'errore su x ÷ var. errore su y



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

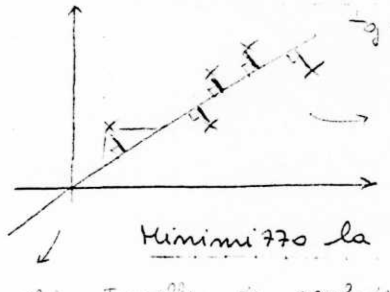
$\alpha = \infty \Rightarrow \sigma_y = 0 \Rightarrow$ regressione di x su y

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

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

Interpretazione per $\alpha = 1$

Regr. di y su x → residui verticali
" x su y → " orizzontali



→ residui ⊥ alla retta → MIN $\sum(\Delta x^2 + \Delta y^2)$ = distanza euclidea della retta dal pto sperimentale

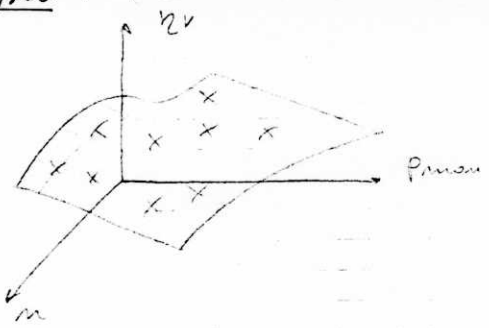
Minimizzo la somma dei quadrati delle distanze dei pti dalla retta

∃ gli intervalli di confidenza

NB Questa categoria di modelli è molto + difficile da gestire perché il modello NON è + lineare nei parametri (… θ·x …). Di solito si fanno entrambe le regressioni lineari (su x e su y) e se vengono simili si sa che con un modello + complicato andrei in mezzo 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 $\in$ a una certa classe

(monomi, sinusoidi, wavelet, ...)

Gergo: h_i sono dette regressori → Nel caso della conducibilità termica c'era solo $1, u_1, u_1^2, u_1^3, u_1^4$.

Se considero monomi di grado ≤ 3 i possibili regressori sono:

$1, u_1, u_2, u_1^2, u_1u_2, u_2^2, u_1^3, u_1^2u_2, u_1u_2^2, u_2^3$

↳ mettendo insieme questi monomi posso ottenere tante superfici ≠

Modello "ricco":

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

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

Forte rischio di sovraparametrizzazione 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 ≤ 1

$$g(u_1, u_2) = \begin{cases} \theta_1 \\ \theta_1 u_1 \\ \theta_1 u_2 \\ \theta_1 + \theta_2 u_1 \\ \theta_1 + \theta_2 u_2 \\ \theta_1 u_1 + \theta_2 u_2 \\ \theta_1 + \theta_2 u_1 + \theta_3 u_2 \end{cases}$$

(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: **esplosione del n° dei modelli**!

OCCHIO: col Test F se A > B e B > C NON ho la garanzia che (A > C) ⇒ (A migliore di C) devo fare tutti i confronti

Vediamo una tecnica per ridurre il n° di confronti:

STEPWISE REGRESSION

1) Dato un modello di partenza (anche modello vuoto ⇒ y = rumore → assenza di modello) classifico secondo un criterio di importanza i regressori che non fanno parte del modello

(Possibile criterio: riduzione di SSR (somma dei quadrati dei residui) quando inserisco il nuovo regressore)

NB Comunque calcolo il modello SSR di minimizza, ma lo fa in modi ≠ a seconda del modello

2) Considero il 1° in classifica e lo sottopongo ad un Test di ingresso (FPE, AIC ...) → confronto FPE nuovo (con + nuovo regressore) con FPE precedente

Se lo supera inserisco il nuovo regressore. Altrimenti ripeto per il secondo in classifica. → non è detto che il secondo in classifica è il 2° più preciso

3) Calcolo un Test di espulsione (FPE, AIC ...) per ognuno dei regressori del modello. (Tolgo il regressore associato a FPE + alto e lo rimetto con FPE + basso)
Se il Test scatta elimino il regressore.
Poi procedo a ulteriori Test di espulsione sul modello semplificato

4) Torno al pto 1)

Con la cross-validazione non serve fare la classifica → si prende il regressore che migliora la cross-validazione e si toglie quello che la fa peggiorare

FORWARD REGRESSION: parto da un modello semplice e continuo ad aggiungere funzioni non lineari.

BACKWARD REGRESSION: modello complesso → continuo a togliere funzioni non mi servono.

Con questa tecnica non è detto che trovo il modello migliore (OTTIMO) perché aggiungo 1 pezzo alla volta (xò è il migliore locale)

VANTAGGIO: Posso lavorare con un n° appropriato di regressori fino ad ottenere un modello + semplice
Ottengo un modello SUBOTTIMALE

22-5-2001

Modello lineare

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

(θ: parametri liberi usati a stimare il modello)

PB: Cosa succede quando la dipendenza da θ non è lineare?

STIMA DI MODELLI NON LINEARI NEI PARAMETRI

$Y = \Phi(\theta) + V$ → errore di misura

a scatola nera è meglio un modello lineare nei parametri

Caso tipico: identif a scatola grigia. Leggi fisiche ⇒ $\Phi(\theta)$

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

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

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

→ $\theta^{ML} = \arg\min_{\theta} \varepsilon^T \Sigma_V^{-1} \varepsilon$ somma dei quadrati dei residui pesati

pesare = moltiplicare per una matrice diagonale

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

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

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 = $\varepsilon = Y - \Phi\theta$ → $\frac{d(\varepsilon^T\varepsilon)}{d\theta}$ = lineare in $\theta = 0$ (infatti $\varepsilon^T\varepsilon$ è come essere un prodotto di θ).

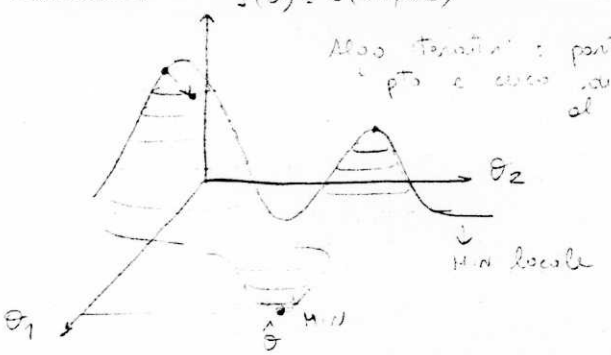
⇒ eq. normali

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

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

attraverso procedure iterative.

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



Algo iterativi: parto da un θ^0 e cerco di avvicinarmi al MIN di successione

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

Supponiamo di aver trovato la stima $\hat{\theta}$, allora posso linearizzare 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

$\left.\frac{d\Phi(\theta)}{d\theta}\right|_{\theta=\hat{\theta}}$ è 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

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

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

(*) MATRICE DI SENSITIVITÀ

Posso usare il modello linearizzato per:

- Stimare la var σ^2 del disturbo e la var di $\hat{\theta}$ ($\Rightarrow$ intervalli di confidenza) ⊙
  ↳ sostituisco al modello lineare il mod. linearizzato e faccio come al solito
  ⊙ NON devono essere troppo grandi VS potrei finire al di fuori della zona in cui vale la linearità
- Validare il modello (test χ^2)
- Effettuare confronti tra modelli (TEST F, FPE, AIC, MDL ...)

Calcolo della stima: Algoritmo di GAUSS-NEWTON

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

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

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

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

Per lo sviluppo di Taylor: $\Phi(\theta) \cong \Phi(\theta^k) + \Phi_\theta^k (\theta - \theta^k) \Rightarrow$

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

$\varepsilon(\theta)$ vettore dei residui

$$J(\theta) \cong J^k(\tilde{\theta}^k) := (\tilde{Y}^k - \Phi_\theta^k \tilde{\theta}^k)^T \Psi^{-1} (\tilde{Y}^k - \Phi_\theta^k \tilde{\theta}^k)$$ → varietà: dipende in modo quadratico da $\tilde{\theta}$ $\Rightarrow$ posso MINIMIZZARE $J^k(\tilde{\theta}^k)$

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

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

Iterazione base dell'algo G-N

Ricordando che $\theta = \theta^k + \tilde{\theta}^k$ (non $\hat{\theta}$ che avevo)

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

L'inizializzazione è molto importante (il meglio per la convergenza a MIN globale)

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

lancio + volte l'algo, se ottengo valori ≠ x individuazione il MIN globale devo ... fare I e ... corrispondenza dei MIN trovati

• Può accadere che $\det\left((\Phi_\theta^k)^T \Psi^{-1} \Phi_\theta^k\right) \cong 0$ → NON è detto che il pb NON sia identificabile (potrebbe accadere solo localmente)

Si modifica l'algoritmo:

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

↳ sto spostando tutti gli autovalori verso dx ⇒ i λ = 0 diventano > 0

↓
Variante di

LEVENBERG - MARQUARDT

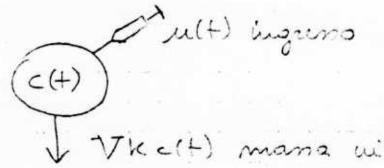
(PB) Se α_k troppo grande diventa dominante

• Convergenza quadratica nell'intorno del MINIMO:

$$\|\theta^{k+1} - \theta^k\| \le \gamma \|\theta^k - \theta^{k-1}\|^2$$

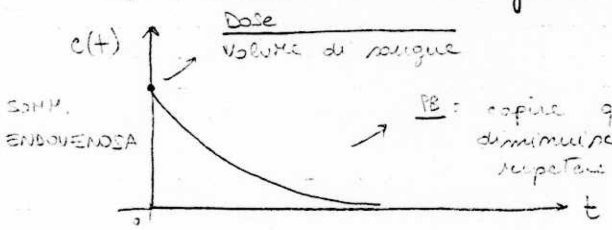
→ NON garantisce la convergenza globale ⇒ spesso si usano algo ibridi

Esempio: Cinetica di un farmaco



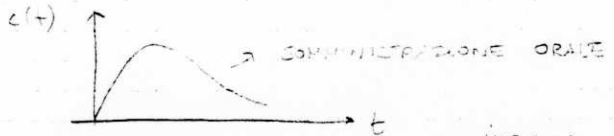
c(t): concentrazione del farmaco

V = volume del compartimento



PB: capire quanto velocemente diminuisce x sapere quando ripetere la somministrazione

Idea: sangue come 1 solo serbatoio → compartimenti omogenei

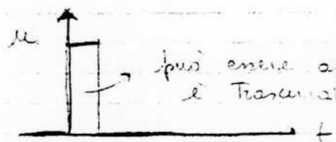


$$c(t) = \frac{m(t)}{V}$$

$\dot{m}(t) = -k\,m(t) + u(t)$ (variazione di massa; $k\,m(t)$: massa in uscita; $u(t)$: ingresso)

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

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



può essere approssimato a δ(t) perché la durata è trascurabile rispetto al tempo tot

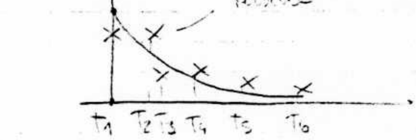
$c(0^-) = 0$

$y_i = c(t_i) + v_i$ (y_i: misure; 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 raramente cost → dipende dalla concentrazione (dir. prop.)



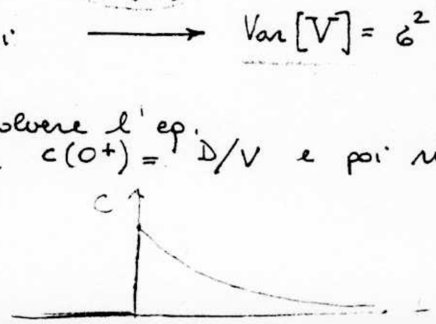
Errori di misura: Hp $\frac{\sqrt{Var[v_i]}}{c(t_i)} = cost = \sigma$ ⇒ l'errore percentuale è cost

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

($c(t_i)$ NON nota)

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

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



$$\Phi(\theta) = \begin{bmatrix} c(t_1) \\ \vdots \\ c(t_N) \end{bmatrix} = \begin{bmatrix} \frac{D}{V} e^{-kt_1} \\ \vdots \\ \frac{D}{V} e^{-kt_N} \end{bmatrix} = D \begin{bmatrix} \frac{1}{\theta_1} e^{-\theta_2 t_1} \\ \vdots \\ \frac{1}{\theta_1} e^{-\theta_2 t_N} \end{bmatrix}$$

Per usare G-N serve $\frac{\partial}{\partial \theta} \Phi(\theta)$

$$\frac{d}{d\theta}\left(\frac{D}{\theta_1} e^{-\theta_2 t}\right) = \left[-\frac{1}{\theta_1^2} e^{-\theta_2 t} \quad -\frac{t}{\theta_1} e^{-\theta_2 t} \right]$$

($\partial/\partial\theta_1$ sulla prima componente, $\partial/\partial\theta_2$ sulla seconda)

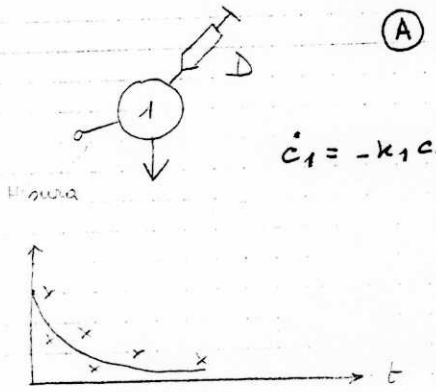
$$\Phi_\theta^k = D \begin{bmatrix} -\frac{1}{(\theta_1^k)^2} e^{-\theta_2^k t_1} & -\frac{t_1}{\theta_1^k} e^{-\theta_2^k t_1} \\ \cdots & \\ -\frac{1}{(\theta_1^k)^2} e^{-\theta_2^k t_N} & \cdots \end{bmatrix}$$

($N \times 2$)

23-5-2001

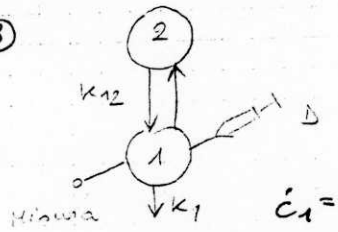
Esempio : Scelta tra 2 modelli.

Modello con 1 compartimento



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

Modello con 2 compartimenti



Si usa la conservazione della massa

$$\dot{c}_1 = -k_1 c_1 + k_{12}(c_2 - c_1) + \frac{1}{V}$$

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

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):

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

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

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

Il problema si riduce alla scelta tra i modelli:

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

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

(*) è una delle hp x applicare Tecniche x la scelta

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

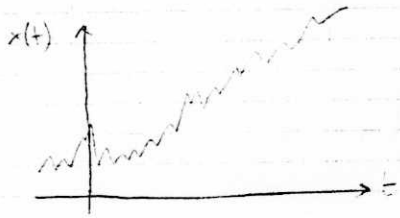
$\Rightarrow$ F-Test, AIC, FPE, MDL .... $\Rightarrow$ decido quale è il migliore.

INIZIA LA 3° PARTE $\rightarrow$ 2° COMPITINO

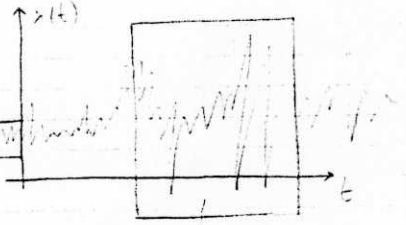
LUCIDI

PROCESSI CASUALI STAZIONARI

che faccia ha un P.C. stazionario ergodico? È più facile dire quando non lo è

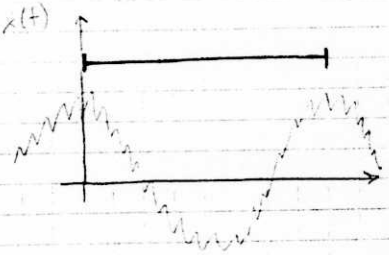


NO xché $m_x(t) \neq cost$



→ la varianza esplode

NO perché $Var[x(t)] \neq cost$



No perché $E[x(t)] \neq cost$

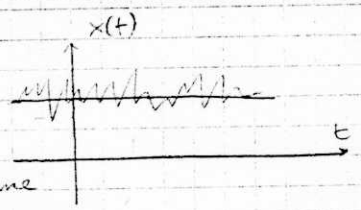
→ potrebbe essere CICLOSTAZIONARIO

ovvero stazionario x una certa lunghezza della finestra

Per verificare se un PC NON è stazionario dovrei identificare una di queste 3 condizioni

Valor medio periodico → x rendere il proc stazionario si può sottrarre il valor medio periodico (→ normalizzo).

CARATTERIZZAZIONE DEI P.C. STAZIONARI ERGODICI



IDEA: Usare i momenti

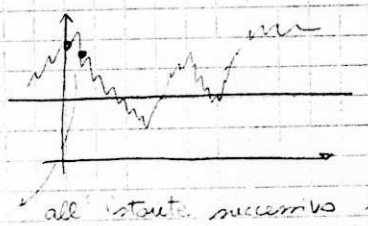
- m_x: valor medio
  o valore atteso di insieme
- $E[x(t)^2]$: valor quadratico medio

Per l'ergodicità: $E[x(t)^2] = \lim_{T\to\infty} \frac{1}{T} \sum_{i=1}^{T} x(i)^2$ → POTENZA MEDIA del processo

fz di autocovarianza

- $\gamma_{xx}(\tau) := E[(x(t)-m_x)(x(t+\tau)-m_x)]$ Mi dà info sulla struttura di correlazione

$x(t) > m_x \overset{?}{\Rightarrow} x(t+\tau) > m_x$



Ci sono processi che hanno molta memoria del passato (si discostano lentamente da certi valori).

all'istante successivo potrei ancora essere sopra la media

NOTA: se conosco m_x e $\gamma_{xx}(\tau)$ non mi serve $E[x(t)^2]$ perché $E[x(t)^2] = Var[x(t)] + m_x^2 =$

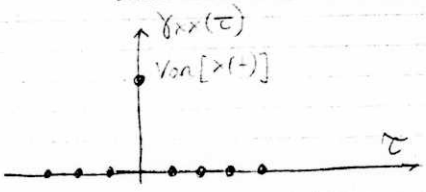
$= \gamma_{xx}(0) + m_x^2$ Autocov $(\tau=0)$ = Var.

PARENTESI: RUMORE BIANCO (WHITE NOISE)

NB $x(t_1)$ non può essere incorrelato da se stesso.

$x(t)$ è un WN se $x(t_1)$ e $x(t_2)$ sono incorrelate $\forall t_1 \neq t_2$, ovvero

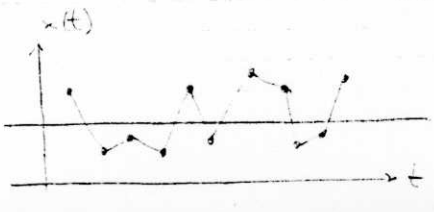
$\gamma_{xx}(t_1, t_2) = 0 \quad t_1 \neq t_2 \Rightarrow$ (PC staz.) $\gamma_{xx}(\tau) = 0 \quad \tau \neq 0$



Notazione: $x(\cdot) \sim WN(m_x, \sigma_x^2)$ → se la media è = 0 specifico solo σ_x^2

- Se $x(t_1)$ e $x(t_2)$ sono indipendenti si dice che $x(\cdot)$ è un WN in senso stretto.

WN = PC in cui un singolo campione è incorrelato da qualunque altro → sapere che un valore è sopra la media NON ci dice assolutamente nulla!



Imprevedibile: PC più casuale che ci possa essere

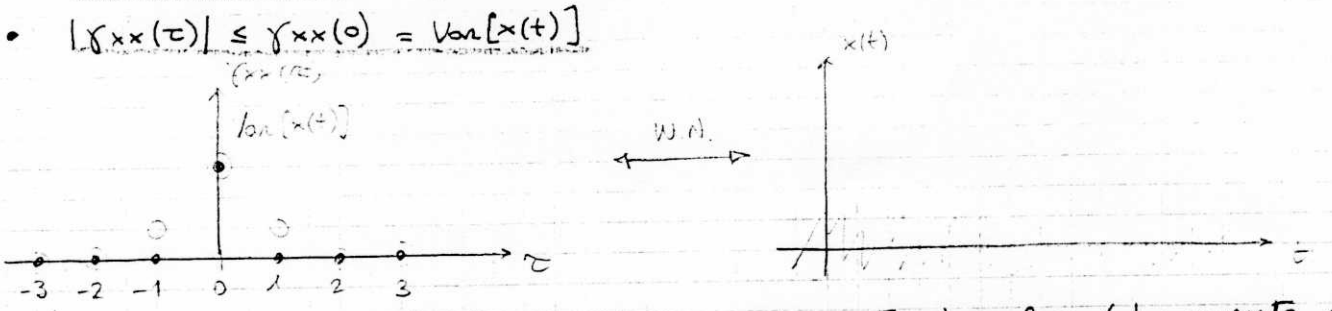
A partire [illegible]

FINE PARENTESI

IMPARIAMO A LEGGERE L'AUTOCOVARIANZA

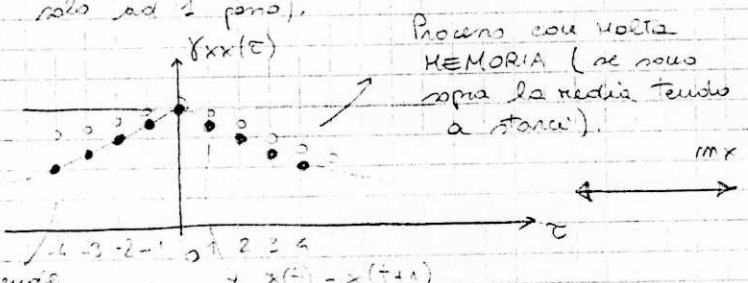
PROPRIETÀ di $\gamma_{xx}(\cdot)$:

- $\gamma_{xx}(\tau) = \gamma_{xx}(-\tau)$ (funzione pari)
- $|\gamma_{xx}(\tau)| \leq \gamma_{xx}(0) = \mathrm{Var}[x(t)]$



$\gamma(\tau) = 0$ ⟷ **andamento irregolare** (puramente casuale)

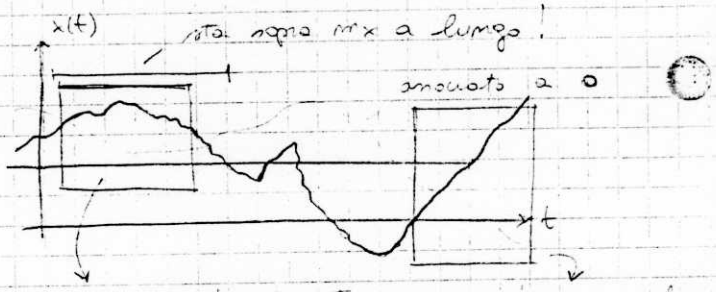
(il processo con γ detta da 0 NON è del tutto casuale (ma ricco e due qualcosa solo ad 1 passo).



Processo con molta MEMORIA (se sono sopra la media tendo a stare).

scende lentamente.

↳ se $x(t)$ è sopra la media è molto probabile che $x(t+1)$ sia sopra la media



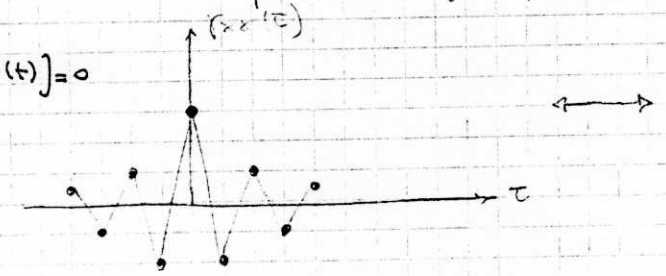
in questa finestra NON sembra neanche un proc stazionario

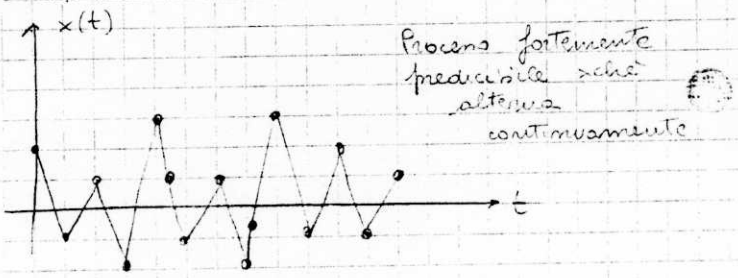
$\gamma(\tau) \cong \gamma(0)$ ⟷ **andamento regolare** (valori adiacenti sono molto correlati).

CASO LIMITE: $\gamma(t) = \gamma(0) \longleftrightarrow x(t) = \text{cost}$ → (NB: NON è + ergodico)

(PE) Se guardo un proc su un periodo troppo breve potrebbe non sembrare stazionario vs se su un periodo lungo potrebbe esserlo.

$E[x(t)] = 0$





Processo fortemente prevedibile xché alterna continuamente

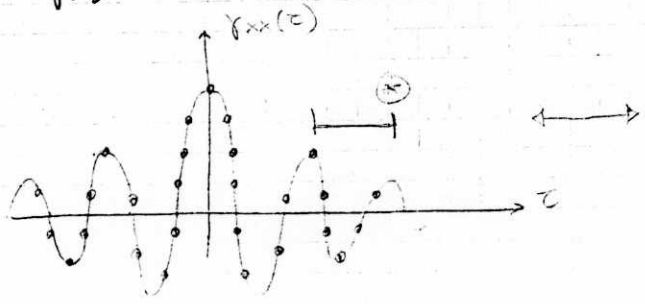
$\gamma(1) < 0 \quad E[x(t)x(t+1)] < 0$

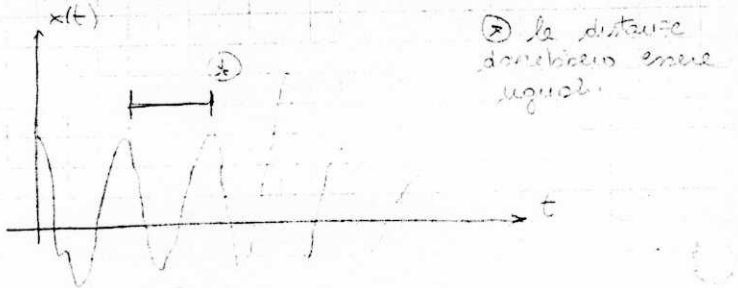
$\gamma(2) > 0 \quad E[x(t)x(t+2)] > 0$

$\gamma(3) < 0$ ...

Se $x(t) > 0$ in media $x(t+1) < 0$

Andamento irregolare dovuto ai cambi di segno.





(*) la distanza dovrebbero essere uguali