- Teorema dovuto a Gauss (segue da pag. 194)

Stiamo usando il metodo dei minimi quadrati, al fine di individuare tra tutti i possibili modelli quello che minimizza la somma dei quadrati dei residui (scarti rispetto al modello)

Esiste una formula esplicita che fornisce il vettore dei parametri ottimali si determina nel modo seguente:

$$f(\theta) := \varepsilon^T \varepsilon = \|\varepsilon\|^2$$

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

Dimostriamo, innanzitutto, che la matrice inversa

$$\left(\Phi^T \Phi\right)^{-1}$$

esiste, considerando che è sufficiente che

$$\operatorname{rank}(\Phi) = p$$

implica che

$$\det\left(\Phi^T \Phi\right) \neq 0$$

Infatti se per assurdo il suddetto determinante fosse nullo, allora esisterebbe un

$$k \neq 0$$

tale che

$$\Phi^T \Phi k = 0$$

da cui seguirebbe che

$$k^T \Phi^T \Phi k = 0$$

$$\left(\Phi k\right)^T \Phi k = 0$$

$$\|\Phi k\|^2 = 0 \quad \text{norma di } (\Phi k) = 0$$

$$\Phi k = 0$$

Se ne deduce che Φ presenterebbe delle colonne linearmente dipendenti, in contraddizione con l'ipotesi fatta

$\text{rank}(\Phi) = q$

Avendo supposto che $\Phi \in \mathbb{R}^{n \times q}$ il rango di Φ indica il numero $\leq$ delle colonne realmente

Ricordiamo che n è il numero dei dati e q il numero di incognite e che

$$\varepsilon_i = Y - \Phi\theta \qquad \text{(pag. 194)}$$

Perciò,

$$J(\theta) = (Y - \Phi\theta)^T (Y - \Phi\theta) =$$

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

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

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

con ε^{LS} uguale ai residui ottenuti con θ^{LS}.

Quindi,

$$J(\theta) = [\varepsilon^{LS} - \Phi(\theta - \theta^{LS})]^T [\varepsilon^{LS} - \Phi(\theta - \theta^{LS})] =$$

$$= (\varepsilon^{LS})^T \varepsilon^{LS} - (\varepsilon^{LS})^T \Phi(\theta - \theta^{LS}) - (\theta - \theta^{LS})^T \Phi^T \varepsilon^{LS} +$$

$$+ (\theta - \theta^{LS})^T \Phi^T \Phi (\theta - \theta^{LS})$$

Osserviamo che

$(\varepsilon^{LS})^T \varepsilon^{LS} \in \mathbb{R}^{1 \times n} \times \mathbb{R}^{n \times 1} = \mathbb{R}$ (scalare)

con ε^{LS} = vettore dei residui

$(\varepsilon^{LS})^T \Phi (\theta - \theta^{LS}) \in \mathbb{R}^{1 \times n} \times \mathbb{R}^{n \times q} \times \mathbb{R}^{q \times 1} = \mathbb{R}$

$(\theta - \theta^{LS})^T \Phi^T \varepsilon^{LS} \in \mathbb{R}^{1 \times q} \times \mathbb{R}^{q \times n} \times \mathbb{R}^{n \times 1} = \mathbb{R}$

$(\theta - \theta^{LS})^T \Phi^T \Phi (\theta - \theta^{LS}) \in \mathbb{R}^{1 \times q} \times \mathbb{R}^{q \times q} \times \mathbb{R}^{q \times 1} = \mathbb{R}$

Tutti i termini sono scalari e uno scalare è uguale al suo trasposto. Ne consegue che il secondo e il terzo termine sono tra loro uguali.

Si ottiene

$$J(\theta) = (\varepsilon^{LS})^T \varepsilon^{LS} - 2(\varepsilon^{LS})^T \Phi(\theta - \theta^{LS}) + (\theta - \theta^{LS})^T \Phi^T \Phi (\theta - \theta^{LS})$$

Osserviamo che il secondo termine è nullo. Infatti

$$2(\theta - \theta^{LS}) \cdot \Phi \cdot (\varepsilon^{LS})^T =$$

$$= 2(\theta - \theta^{LS})^T \cdot \Phi^T \varepsilon^{LS} =$$

$$= 2(\theta - \theta^{LS})^T \cdot \Phi^T (Y - \Phi \theta^{LS}) =$$

$$= 2(\theta - \theta^{LS})^T \cdot \Phi^T (Y - \Phi (\Phi^T \Phi)^{-1} \Phi^T Y) =$$

(Rap. 19c)

$$= 2(\theta - \theta^{LS})^T \Phi^T [I - \Phi (\Phi^T \Phi)^{-1} \Phi^T] Y =$$

$$= 2(\theta - \theta^{LS})^T [\Phi^T - \Phi^T \Phi (\Phi^T \Phi)^{-1} \Phi^T] Y =$$

$$= 2(\theta - \theta^{LS})^T [\Phi^T - I \cdot \Phi^T] \cdot Y = 0,$$

essendo

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

una matrice per la proprietà vista in cui si ottiene la matrice identità.

Risulta:

$$J(\theta) = (\varepsilon^{LS})^T \varepsilon^{LS} + (\theta - \theta^{LS})^T \Phi^T \Phi (\theta - \theta^{LS})$$

Il primo termine non dipende da θ ma da θ^{LS} che ha un valore determinato, non minimizzabile! La minimizzazione $J(\theta)$ occorre intervenire sul 2° termine, ponendo

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

ed osservando che si ha una forma quadratica pari negativa.

- Proprietà delle matrici.

Data una matrice D,

$$D^T D$$

è detta diade ed è sempre semidefinita positiva.

202 Ne consegue che

$$k^T D^T D k \geq 0 \qquad \forall k \in \mathbb{R}^n$$

- Proprietà sulle matrici

Se una matrice S è semidefinita positiva, ($S \geq 0$)

e

$$\det(S) \neq 0,$$

allora S è definita positiva.

$$k^T S k > 0 \qquad \forall k \neq 0$$

Allora d scelta anche simmetrica.

Se una matrice semidefinita positiva, tutti gli autovalori non sono negativi.

Se il determinante è diverso da zero (allora) gli autovalori non sono tutti esattamente maggiori di zero e la matrice è definita positiva.

~~Il terzo termine di $J(\theta)$ richiede questa proprietà~~

Il termine $\Phi^T \Phi$ è una diade, perciò il terzo termine di $J(\theta)$ si annulla solo se

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

e questa assicura la scelta migliore per minimizzare $J(\theta)$.

Interpretazione grafica

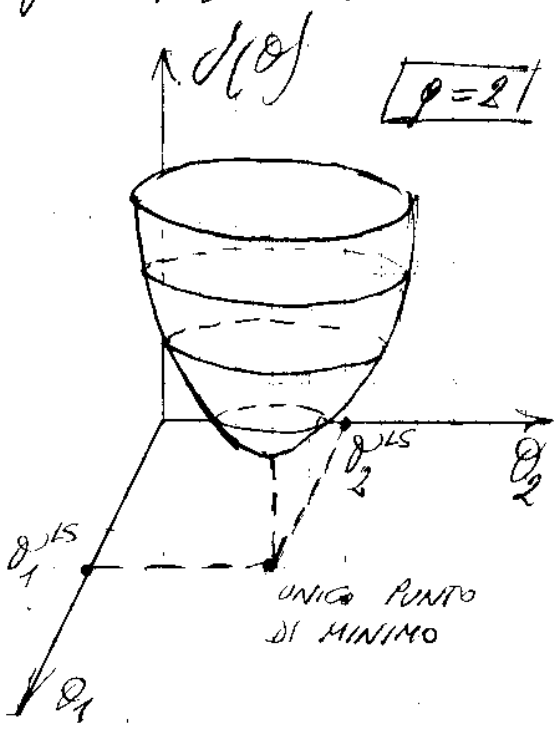
Caso $q = 2$.

Il piano $\theta_1 \theta_2$ è il piano dei parametri.

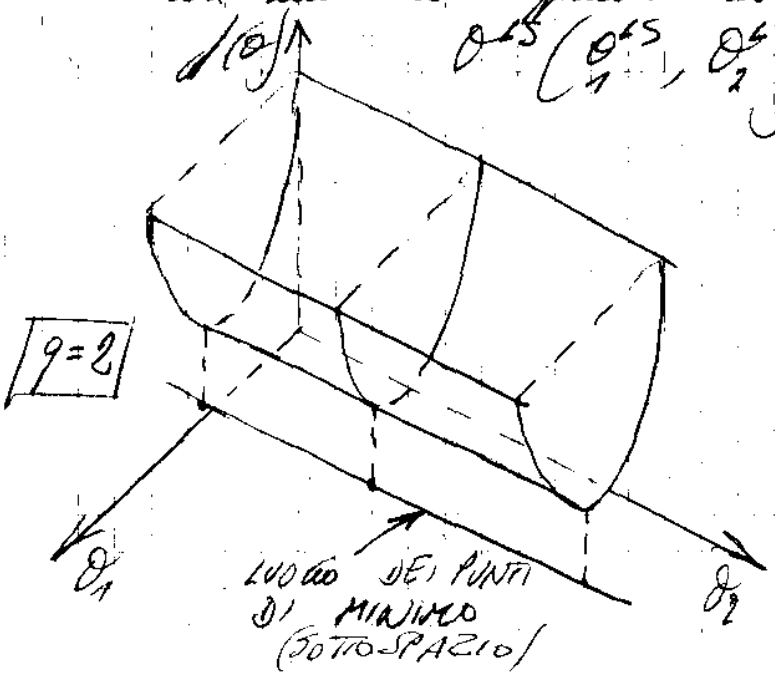
Dell'asse delle ascisse si legge la somma dei quadrati dei residui.

Si decidere che la somma più piccola possibile si ha per un solo valore che, se

$$\text{rank}(\Phi) = q$$



$J(\theta)$ è una forma quadratica definita positiva con un unico punto di minimo globale in $\theta^{LS} (\theta_1^{LS}, \theta_2^{LS})$



Caso $\text{rank}(\Phi) = 1$

e $q = 2$

si ha un luogo di punti di minimo che forma una retta.

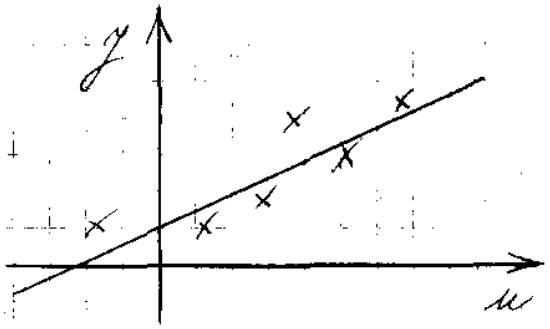
Le rette possibili sono infinite.

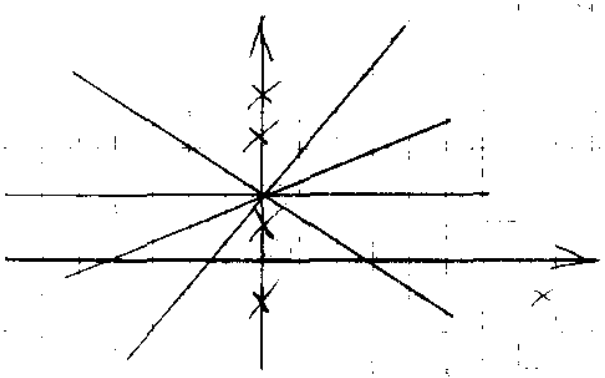
-) **Regressione lineare** ($q > 1$)

è un metodo di stima mediante modello lineare

cambiando la posizione della retta cambia la somma dei quadrati dei residui.

Esiste una posizione ottimale scegliendo la quale la somma dei quadrati dei residui aumenta.





Nel caso rappresentato le rimanenti sono state calcolate solo in corrispondenza ad $\theta = 0$.

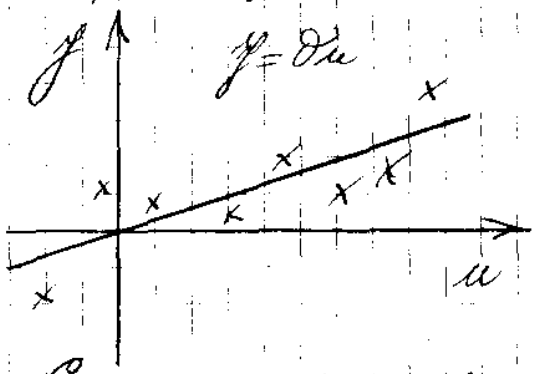
Qualsiasi retta passante per quelle rappresentate non ha sottratto la somma dei quadrati dei residui che non tutti calcolati nell'origine.

I residui dipendono solo dall'intercetta e non dal coefficiente angolare. Non ci sono quindi ripetizioni per determinare quest'ultimo.

Esistono infinite soluzioni ottime.

Il caso migliore è quello rappresentato dalla rotazione

- Esempio: regressione lineare con retta passante per l'origine



La equazione generica che rappresenta il modello è

$$y(i) = x(i)\theta + e(i)$$

cioè

$$\begin{cases} y(1) = x(1)\theta + e(1) \\ y(2) = x(2)\theta + e(2) \\ \dots \\ y(n) = x(n)\theta + e(n) \end{cases}$$

Le equazioni scritte e riscritte accorpando a

$$Y = \Phi^T \theta + e$$

essendo

$$\Phi = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$$

con $Y \in \mathbb{R}^{n\times 1}$ $\quad \Phi \in \mathbb{R}^{n\times 1}$ $\quad e \in \mathbb{R}^{n\times 1}$

$\theta \in \mathbb{R}$ (scalare)

Chiedere che $\text{rank}(\Phi) = 1$ significa chiedere che

$$\text{rank}\begin{bmatrix} x(1) \\ x(2) \\ \dots \\ x(n) \end{bmatrix} = 1$$

cioè verifica che almeno una $x(i)$ sia diversa da zero.

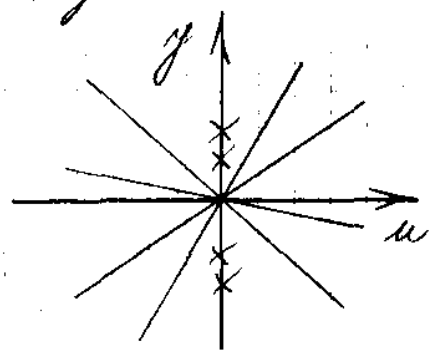
La formula risolutiva è (pag. 199)

$$\theta^{LS} = \left[\Phi^T \Phi\right]^{-1} \Phi^T Y =$$

$$= \left[\begin{bmatrix} x(1) & x(2) & \dots & x(n) \end{bmatrix} \begin{bmatrix} x(1) \\ x(2) \\ \dots \\ x(n) \end{bmatrix}\right]^{-1} \begin{bmatrix} x(1) & x(2) & \dots & x(n) \end{bmatrix} \begin{bmatrix} y(1) \\ y(2) \\ \dots \\ y(n) \end{bmatrix} =$$

$$= \frac{\sum x(i)\, y(i)}{\sum x(i)^2}$$

che rappresenta la famiglia delle regressioni li-
neari passanti per l'origine



Il caso rank $(\Phi) = 0$
corrisponde alla situazione rap-
presentata a destra in cui
tutte le x_i sono nulle.
Tutti i modelli passanti per
l'origine sono equivalenti es-
sendo ridotti ad uno solo in
tutte le circostanze.

Osservazione il metodo dei minimi quadrati può
fare a meno del concetto di probabili-
tà.

La teoria ha dei limiti e pone alcuni quesiti:

- qual è l'affidabilità della stima? di $\hat{\theta}^{LS}$ ad es.
- il modello è giusto? funziona bene in tutti i casi?
- confronto tra modelli: occorre avere degli strumenti per poter valutare se un modello è migliore di un altro.

Per risolvere questi quesiti occorre avere delle nozioni relative alla teoria della stima.

STIMATORE BLUE

Consideriamo sotto l'ipotesi I_2 secondo la quale vale la relazione

$$Y = \Phi\theta^\circ + V \qquad E[V] = 0$$
$$Var[V] = \sigma^2\Psi$$

Si sta formulando un modello probabilistico dei dati secondo il quale i dati osservati sono spiegabili da un determinato modello matematico ($\Phi\theta^\circ$) al quale si aggiunge un termine di errore descritto da un vettore di variabili casuali aventi valore atteso nullo e con determinata matrice varianza

$$Y \in R^{n\times 1} \qquad \Phi \in R^{n\times q} \qquad \theta^\circ \in R^{q\times 1}$$
$$V \in R^{n\times 1}$$

La matrice varianza è composta da due termini,

Ψ, matrice nota $n\times n$

σ^2, valore eventualmente incognito

Esempio per giustificare l'uso della matrice Ψ

Si considerino delle misure di concentrazione del tipo

$$Y_i = c_i + V_i$$

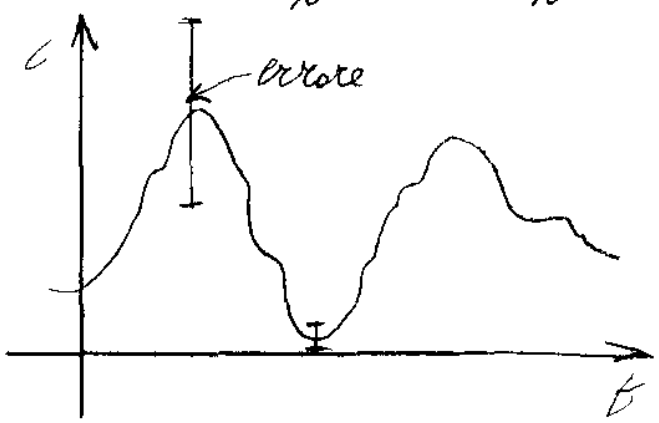
c_i = misura della concentrazione

V_i = errore di misura

Può capitare che la varianza di V_i sia

$$Var[V_i] = k^2 c_i^2$$

nel senso che l'entità di cui sbaglio dipende dalla concentrazione.



Laddove la concentrazione è piccola, ho un errore piccolo ed un errore grande quando la concentrazione è grande. Come mi muovo nello schema da presentato una concentrazione di modulo variabile nel tempo.

k è detto coefficiente di variazione (CV)

La matrice varianza è una matrice diagonale

$$\mathrm{Var}[V] = \begin{bmatrix} \mathrm{Var}[V_1] & & & 0 \\ & \mathrm{Var}[V_2] & & \\ & & \ddots & \\ 0 & & & \mathrm{Var}[V_n] \end{bmatrix}$$

Abbiamo ipotizzato che le misure siano incorrelate tra loro.

Otteniamo

$$\mathrm{Var}[V] = k^2 \begin{bmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{bmatrix} = k^2 \begin{bmatrix} \gamma_1^2 & & & 0 \\ & \gamma_2^2 & & \\ & & \ddots & \\ 0 & & & \gamma_n^2 \end{bmatrix} =$$

$$= \sigma^2 \Psi$$

dove si ignora la concentrazione vera (σ_i), dove è la grandezza che si sta misurando.

Al suo posto si considera la misura (γ_i) e si pone

$$\Psi = \begin{bmatrix} \gamma_1^2 & & & 0 \\ & \gamma_2^2 & & \\ & & \ddots & \\ 0 & & & \gamma_n^2 \end{bmatrix}$$ (matrice nota delle grandezze misurate)

Si è considerato come spesso succede un modello a meno di una costante (non si conosce il valore di k, assegnato con 0).

Il modello dell'errore di misura è parzialmente noto: si sa che le misure più grandi hanno maggiore varianza ma con coefficiente di proporzionalità ignoto.

→ Teorema di Gauss-Markov

Se vale l'ipotesi I_2, definiamo

$$\varepsilon := Y - \Phi\theta$$

Allora, la cifra di merito

$$J^M(\theta) := \varepsilon^T \Psi^{-1} \varepsilon$$

Andrej Andreevič Markov: Rjazan 1856 – San Pietroburgo 1922
matematico e statistico russo

è minimizzata da

$$\theta^{M} = \left[\Phi^T \Psi^{-1} \Phi\right]^{-1} \Phi^T \Psi^{-1} Y$$

θ^M è detto stimatore di Markov o stimatore BLUE.

Inoltre:

1) θ^M è lo stimatore lineare e non polarizzato, quello che minimizza la varianza di θ

$$\mathrm{Var}[\theta] = \min$$

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

Nota: BLUE = Best Linear Unbiased Estimator

- Best è riferito alla varianza che risulta minima
- Lineare indica che dipende linearmente da Y
- Unbiased indica che lo stimatore di classe è non polarizzato e, per questo tipo di stimatori, l'obiettivo è minimizzare la varianza.

Osservazioni:

- se gli errori V_i hanno tutti la stessa varianza allora la varianza si scrive

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

essendo σ^2 la varianza comune degli errori da cui segue che

$$\Psi = I$$
$$\theta^M = \theta^{LS} \quad \text{(vedi pag. 204)}$$

- se gli errori V_i sono incorrelati e tali che

$$\mathrm{Var}[V_i] = \sigma_i^2,$$

se, cioè, hanno tutte varianze diverse, allora

$$Var[V] = \Psi,$$

con

$$\Psi = \begin{bmatrix} \sigma_1^2 & & \mathbf{0} \\ & \sigma_2^2 & \\ \mathbf{0} & & \ddots \\ & & & \sigma_n^2 \end{bmatrix}$$

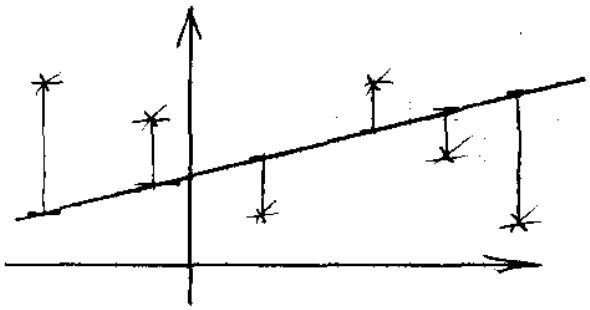
e risulta

$$J^M = \varepsilon^T \Psi^{-1} \varepsilon =$$

$$= \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{bmatrix} \begin{bmatrix} 1/\sigma_1^2 & & \mathbf{0} \\ & 1/\sigma_2^2 & \\ \mathbf{0} & & \ddots \\ & & & 1/\sigma_n^2 \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdots \\ \varepsilon_n \end{bmatrix} =$$

$$= \frac{\varepsilon_1^2}{\sigma_1^2} + \frac{\varepsilon_2^2}{\sigma_2^2} + \cdots + \frac{\varepsilon_n^2}{\sigma_n^2}$$

Di conseguenza J^M non minimizza più la somma dei quadrati dei residui (la somma dei quadrati dei residui pesati).



Nello scarto rappresentato a fianco si nota che i dati hanno diversa varianza ed il modello rappresentato è quello dei minimi quadrati.

Si dovrebbe cercare di seguire più fedelmente i dati meno precisi (a varianza maggiore) tenendo in minor considerazione quelli a varianza minore.

Il metodo è detto anche <u>dei minimi quadrati pesati</u> o

Weighted least squares.

Occorre ora capire come definire i pesi da adottare, poiché è invece delle varianze

Esercizio 1

Avete a disposizione i dati relativi ad una popolazione di studenti universitari iscritti a Ingegneria. Per ognuno di essi si conoscono il peso, l'altezza, il sesso ed i voti degli esami di Analisi A, Analisi B, Algebra e Geometria, Fisica A. Vi viene richiesto di analizzare i dati. Con riferimento alle variabili *peso* e *altezza* degli studenti maschi dovete:

1. Stimare media, varianza e deviazione standard.

2. Descrivere in modo approssimato la ddp mediante istogramma.

3. Stimare e disegnare le funzioni di distribuzione.

4. Dovete inoltre valutare se ciascuna VC è distribuita approssimativamente in modo gaussiano. A tale scopo vi viene richiesto di confrontare i grafici al punto (3) con la funzione di distribuzione di una gaussiana avente come media e varianza la media e la varianza campionaria calcolate al punto (1).

Vi viene inoltre richiesta un'analisi bivariata delle VC peso e altezza degli studenti maschi. Si richiede di

5. Stimare la matrice di covarianza e la matrice di correlazione.

6. Disegnare lo scatter plot dei dati *normalizzati*.

Indicazioni operative. Le analisi dell'Esercizio 1 sono implementate nel file statistiche.m. Seguendo le istruzioni che vi verranno date in aula, copiate i files statistiche.m, creadat.m, vccont.m, vcdiscr.m, ed eventuali file *.mat nella vostra directory. Analizzate quindi il file statistiche.m per familiarizzarvi con i principali comandi MatLab.

Esercizio 2

a. Considerando le VC *voti* (di Analisi A, Analisi 2, Geometria e Algebra e Fisica A) svolgere le analisi dei punti (1)-(4) (si consideri l'intera popolazione studentesca).

b. Calcolare la matrice di covarianza e di correlazione del vettore di VC formato dai 4 voti, dal peso e dall'altezza.

c. Con riferimento alle coppie di variabili casuali (voti Analisi A, voti Analisi B), (voti Fisica A, voti Geometria), (altezza, voti Analisi 1) disegnare gli scatter plot dei dati normalizzati.

Esercizio 3

a. Progettare lo stimatore lineare ottimo (in termini di errore quadratico medio) che in base al peso dello studente ne predice l'altezza (si suppone di conoscere il sesso dello studente).
Suggerimento:si usi il valore atteso condizionato come predittore e si ricordi che per V.C. X e Y congiuntamente gaussiane,

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

b. Progettare uno stimatore che in base ai voti di Analisi 1, Algebra e Geometria, Fisica 1 predice quello di Analisi 2.

210 LEZIONE (12) 08/06/09 - parte 2ª
MATLAB (3)

- Esercizio 1

- load MNPV2003.f copia il file MNPV2003.mat che contiene dati relativi a studenti

- who mostra le variabili caricate
  - s = sesso (0 = maschio - 1 = femmina)
  - h = altezza in cm
  - w = peso in kg
  - a1 = voto di Analisi 1
  - a2 = " " " 2
  - f1 = voto di Fisica A
  - g = voto di Algebra e Geometria
  - cd1 = ?
  - sede

- D = [s h w a1 a2 f1 g cd1 sede]
  crea e mostra i dati in una matrice D

- mean(D)
  calcola la media dei dati caricati in D e relativi alle 9 colonne.
  0,1882 è la media della 1ª colonna ed indica che le ragazze sono circa il 19%

- M = mean(D) memorizza le medie nella matrice M

- mh = M(2) memorizza l'altezza media nella variabile mh

- std(D) calcola la deviazione standard delle 9

colonne di dati

-) var = std(D).^2 calcola la Varianza e la riscrive nella variabile var.
-) vh = var(2)

riscrive nella variabile vh la varianza relativa ~~ai pesi~~ alle altezze

-) dp = std(D) memorizza le deviazioni standard nel vettore dp
-) dsh = dp(2) memorizza in dsh la deviazione standard delle altezze.
-) hist(h, 16) visualizza l'istogramma delle altezze con 16 bins
-) corrcoef(D) costruisce la matrice dei coefficienti di correlazione

si noti che la correlazione tra peso e altezza è negativa ($0 = M - 1 = F$).
Peso e altezza si correlano poco con i voti.
I voti si correlano abbastanza bene tra loro.

-) cov(D) copia la matrice delle covarianze, poco significativa.

N.B. la soluzione completa dell'esercizio si trova in rete.