Politecnico di Milano
Facoltà di Ingegneria dei Sistemi
I APPELLO DI STATISTICA APPLICATA
7 Febbraio 2007

Nome e cognome:
Numero di matricola:

Problema 1

Si consideri una popolazione di n unità rappresentate dai vettori $\mathbf{x}_i = (x_i, y_i)' \in R^2, i = 1, ..., n$. Indicato con $\bar{\mathbf{x}}$ il baricentro della popolazione, si assuma che la covarianza della popolazione

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})'$$

sia diversa da un multiplo della matrice identità. Si definisca retta di regressione ortogonale la retta che minimizza la somma dei quadrati delle distanze euclidee dei punti (x_i, y_i) dalla retta stessa. Ovvero la retta

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x$$

tale che

$$(\tilde{\beta}_0, \tilde{\beta}_1) = argmin_{(\beta_0,\beta_1)}\left(\sum_{i=1}^{n} d_i^2(\beta_0, \beta_1)\right)$$

ove

$$d_i^2(\beta_0, \beta_1) = min_{(x,y):y=\beta_0+\beta_1 x}\left((x - x_i)^2 + (y - y_i)^2\right).$$

(a) Si dimostri che la retta di regressione ortogonale passa per il baricentro.

(b) Si dimostri che la direzione individuata dalla retta di regressione ortogonale coincide con la direzione individuata dalla prima componente principale.

(c) Si calcolino coefficiente angolare ed intercetta della retta di regressione ortogonale relativa alle osservazioni contenute nel file `Pb1.txt`.

(d) Si calcoli la somma dei quadrati delle distanze euclidee delle osservazioni contenute nel file `Pb1`.txt dalla retta di regressione ortogonale calcolata al punto precedente.

Problema 2

In due pozzi petroliferi della Georgia vengono misurate quotidianamente temperatura (^oC), densità (g/cm^3) e concentrazione di CO_2 ($\mu g/l$) del petrolio estratto.
Nel file `Pb2A.txt` sono riportate per il pozzo A le 31 osservazioni relative al Gennaio 2007 che si assumono essere realizzazioni di vettori aleatori indipendenti e identicamente distribuiti.
Nel file `Pb2B.txt` sono riportate invece quelle relative al pozzo B; anche esse sono assunte essere realizzazioni di vettori aleatori indipendenti ed identicamente distribuiti.
Si effettui un test per il confronto delle medie nel caso in cui:

(a) le misure relative al pozzo A sono indipendenti da quelle ottenute nello stesso giorno dal pozzo B;

(b) non si assume l'ipotesi enunciata in (a).

Inoltre

(c) si riportino e si verifichino con opportuni strumenti statistici le ipotesi necessarie all'implementazione del test effettuato nel caso (a).

(d) si riportino e si verifichino con opportuni strumenti statistici le ipotesi necessarie all'implementazione del test effettuato nel caso (b).

Problema 3

Nell'Antico Egitto era frequente costruire obelischi a sezione circolare su piattaforme di base quadrata di lato 2 *dam* con lati in direzione dei quattro punti cardinali. Tutti gli obelischi di cui oggi abbiamo conoscenza risultano crollati a causa della loro altezza. Oggi possiamo infatti assumere che i blocchi di roccia provenienti dalla piattaforma basale siano distribuiti in modo uniforme all'interno del perimetro originario della piattaforma mentre quelli provenienti dall'obelisco siano distribuiti secondo una normale bivariata con media nel centro della piattaforma e componenti indipendenti con deviazione standard di 1 *dam*. Supponendo che i blocchi di roccia utilizzati per costruire l'obelisco siano il doppio di quelli utilizzati per la piattaforma:

(a) si individuino analiticamente e si riportino su di un grafico le due regioni del piano R_O e R_P nelle quali si stima sia più probabile trovare, rispettivamente, i blocchi provenienti dall'obelisco e quelli provenienti dalla piattaforma;

(b) sulla base delle due regioni precedentemente individuate, si individui un criterio per la classificazione dei blocchi, si stimi l'AER ad esso associato e se ne discuta il valore.

Problema 4

Nel file `Pb4.txt` sono riportati il numero Y (espresso in migliaia di unità) di autoveicoli immatricolati annualmente in tre paesi dell'Unione Europea (Francia, Germania e Italia) durante un periodo di riferimento di 10 anni. Recenti modelli economici descrivono l'andamento di questa variabile secondo un modello:

$$Y|(X = x, G = g) = \beta_{0g} + \beta_{1g}x^2 + \epsilon$$

con $\epsilon \sim N(0, \sigma^2)$, $x = 1, 2, \ldots, 10$ (anno di riferimento) e $g =$ Francia,Germania,Italia (paese della UE).

(a) Con il metodo dei minimi quadrati, si stimino i 7 parametri del modello.

(b) Utilizzando opportuni test statistici, si affermi se si ritiene necessario inserire nel modello:

1. la variabile X^2;
2. la variabile G;
3. l'effetto della variabile G sul coefficiente che moltiplica il regressore X^2;
4. l'effetto della variabile G sull'intercetta.

(c) Una volta identificato il modello "migliore", si costruiscano tre intervalli per la previsione del numero di autoveicoli immatricolati nei tre stati all'undicesimo anno, di modo che le tre nuove osservazioni caschino contemporaneamente all'interno dei rispettivi intervalli col 95% di probabilità.