Politecnico di Milano
Facoltà di Ingegneria dei Sistemi

Appello di Statistica Applicata

Milano, 5 Luglio 2006

Nome e cognome:
Numero di matricola:

Problema 1

Siano T_1 un campione di 9 osservazioni trivariate $iid \sim N_3(\mu_1, \Sigma)$ e T_2 un campione di 5 osservazioni trivariate $iid \sim N_3(\mu_2, \Sigma)$ stocasticamente indipendenti.
Siano M_1 e M_2 le medie campionarie relative rispettivamente ai due campioni e S_1 e S_2 le rispettive matrici di covarianza campionarie:

$$M_1 = \begin{pmatrix} 4.00 \\ 4.50 \\ 4.00 \end{pmatrix} \qquad M_2 = \begin{pmatrix} 0.50 \\ 0.50 \\ 1.00 \end{pmatrix}$$

$$S_1 = \begin{pmatrix} 0.50 & 0.25 & 0.00 \\ 0.25 & 0.50 & 0.00 \\ 0.00 & 0.00 & 0.25 \end{pmatrix} \quad S_2 = \begin{pmatrix} 1.00 & 0.50 & 0.00 \\ 0.50 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.50 \end{pmatrix}$$

(a) Utilizzando le sole informazioni provenienti dal campione T_2 si implementi un test di livello 10%:

$$H_0\colon\ \mu_2 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \ vs\ H_1\colon\ \mu_2 \neq \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

(b) Ricorrendo al teorema di Hotelling ed ad una stima *pooled* di Σ (utilizzando quindi oltre alle informazioni fornite dal campione T_2 anche quelle fornite dal campione T_1) si implementi un nuovo test di livello 10% per H_0 *vs* H_1.

Problema 2

Un supermercato distribuisce due tipi di volantini pubblicitari, uno studiato appositamente per i clienti con figli (volantino C per clienti C) ed uno studiato appositamente per i clienti senza figli (volantino S per clienti S).
Il supermercato non ha a disposizione nessuna informazione relativa al numero di figli dei propri clienti. D'altra parte, da studi effettuati dalla CONFsupermercati, è noto sia che i clienti C sono il doppio dei clienti S sia che la spesa X (misurata in centinaia di euro) dei clienti C segue approssimativamente una legge uniforme mentre quella dei clienti S segue una legge esponenziale:

$$\begin{aligned} f(x|C) &= \frac{1}{4} \cdot I_{[0,4]}(x) \\ f(x|S) &= e^{-x} \cdot I_{[0,+\infty)}(x) \end{aligned}$$

Il supermercato utilizza la variabile X, osservabile alla cassa al momento del pagamento, per scegliere quale volantino consegnare al cliente.

(a) Si costruisca il criterio su X che minimizza il numero di volantini consegnati ai clienti "sbagliati".

(b) Si calcoli la probabilità che un volantino venga consegnato ad un cliente "sbagliato".

(c) Si calcoli la probabilità che un volantino C venga consegnato ad un cliente S.

(d) Si calcoli la probabilità che un volantino S venga consegnato ad un cliente C.

(e) Che volantino verrà consegnato ad un cliente che spende 100 euro? Con che probabilità sarà stato consegnato il volantino sbagliato?

Problema 3

Lungo le tangenziali milanesi sono presenti quattro centraline che misurano la concentrazione di NO nell'aria.
Le misure relative all'ultimo mese sono riassunte dalla media campionaria $\bar{X}$ e dalla matrice di covarianza campionaria S:

$$\bar{X} = \begin{pmatrix} 12 \\ 10 \\ 10 \\ 12 \end{pmatrix}$$

$$S = \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix}$$

Si esegua l'analisi delle componenti principali riscalata rispetto alla media campionaria. In dettaglio:

(a) Si calcolino i *loadings* relativi alle componenti principali.

(b) Si calcolino le varianze relative alle componenti principali.

(c) Si dia una possibile interpretazione dei risultati ottenuti nei punti (a) e (b).

(d) Il 3 Luglio le centraline hanno registrato i valori $(13, 10, 11, 11)$. Si calcolino gli *scores* relativi.

PS: Per effettuare l'analisi si tenga conto che $\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \\ -1 \end{pmatrix}, \begin{pmatrix} 0 \\ 2 \\ -1 \\ -1 \end{pmatrix}, \begin{pmatrix} 3 \\ -1 \\ -1 \\ -1 \end{pmatrix}$ sono autovettori mutuamente ortogonali della matrice di covarianza campionaria S.

Problema 4

Il dataframe `salary` contiene il salario annuale attuale (in migliaia di dollari), gli anni di servizio e il livello del lavoro (alto 1, basso 0) di 25 dipendenti di una società informatica americana.
È stata implementata in R una regressione lineare, il cui output (parzialmente censurato) è riportato in seguito:

```
> salary
   Salary Years_of_Service Job_Level
1    21.7                3         0
2    24.0                9         0
3    23.8               10         0
...   ...              ...       ...
24   39.9               16         1
25   42.3               18         1

> regression <- lm(Salary ~ Years_of_Service * Job_Level, data = salary)

> summary(regression)

Residuals:
    Min      1Q  Median      3Q     Max
-2.9951 -1.5739 -0.3017  1.5971  3.4983

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 22.5647     1.7368  12.992 1.66e-11
Years_of_Service             0.4230     0.1892   2.236   0.0363
Job_Level                    4.1501     2.2324   1.859   0.0771
Years_of_Service:Job_Level   0.3343     0.2257   1.481   0.1534

Residual standard error: 1.953 on 21 degrees of freedom
Multiple R-Squared: 0.8837,     Adjusted R-squared: 0.8671
F-statistic:  ... on 3 and 21 DF,  p-value: ...
```

(a) Si stimi lo stipendio annuale all'anno 0 e l'incremento annuale dello stipendio annuale per un generico lavoratore di livello basso e per un generico lavoratore di livello alto, avendo cura di mettere in evidenza anche le unità di misura ed il legame coi coefficienti β_0, β_1, β_2 e β_3.

(b) Si effettui il test al 10%:

$$H_0 : \beta_1 = 0 \wedge \beta_2 = 0 \wedge \beta_3 = 0 \quad vs \quad H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0$$

(c) Si effettui il test al 10%:

$$H_0 : \beta_0 = 20 \quad vs \quad H_1 : \beta_0 \neq 20$$

(d) Si calcolino gli intervalli di confidenza di Bonferroni di livello 95% per i coefficienti di regressione β_0, β_1, β_2 e β_3.

Problema 5

Siano β_0, β_1 e σ^2 i parametri che definiscono il modello lineare:

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \epsilon \sim N_1(0, \sigma^2)$$

Siano $\mathbf{Y}$ il vettore delle risposte osservate e $\mathbf{Z}$ la matrice disegno:

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \qquad \mathbf{Z} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \dots & \dots \\ 1 & x_n \end{pmatrix}$$

Ricorrendo al *Lemma del Massimo* ed alle stime ai minimi quadrati $\hat{\beta}_0$, $\hat{\beta}_1$ e s^2 si costruisca una regione di confidenza simultanea ($\forall x \in \Re$) di livello globale γ per $E[Y|x]$.