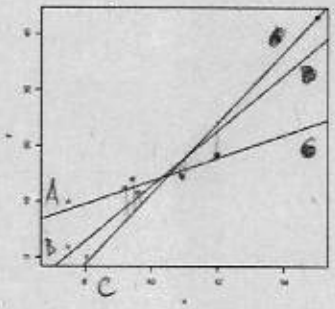
Insegnamento di "Data mining"
'Compitino' del 27 aprile 2015

1. Si spieghi cosa si intende per valore-p.

2. La figura rappresenta un diagramma di dispersione relativo a due variabili X e Y insieme a tre rette tra cui quella dei minimi quadrati.



   a. Si dica, giustificando la risposta, quale delle tre è la retta ai minimi quadrati;

   b. si dica, giustificando la risposta, quale delle seguenti tre possibilità è il coefficiente angolare della retta dei minimi quadrati: $(2, 5, 6)$;

   c. si dica, giustificando la risposta, quale delle seguenti tre possibilità è l'intercetta della retta ai minimi quadrati: $(-7, -52, -37)$;

   d. dimostrare che la somma degli scarti dalla retta dei minimi quadrati è nulla.

3. Si hanno le seguenti osservazioni sulla temperatura
   a Spokane: (18,37; 18,04; 19,57; 21,56; 20,13; 18,08; 22,70; 17,74)
   e a Lewiston: (20,48; 19,41; 22,46; 23,79; 21,18; 19,22; 23,74; 19,18).

   a. Proporre un confronto tra i due gruppi di osservazioni utilizzando un opportuno metodo grafico e commentare il risultato.

   b. Scomporre opportunamente la varianza. Sono confermate le conclusioni di cui al punto precedente?

4. Data la seguente matrice di varianze-covarianze di tre variabili X, Y, Z,

$$\begin{bmatrix} 4 & 4{,}8 & 2{,}4 \\ 4{,}8 & 16 & 0 \\ 2{,}4 & 0 & 9 \end{bmatrix}$$

   a. calcolare la matrice di correlazione (la cui diagonale è formata di elementi tutti pari a 1 e gli elementi fuori dalla diagonale sono le correlazioni tra coppie di variabili);

   b. sapendo che i valori delle medie delle tre variabili sono pari rispettivamente a $3, 7, 2$, calcolare i coefficienti di regressione multipla dei minimi quadrati per l'equazione

$$x_i = \alpha + \beta y_i + \gamma z_i$$

   c. calcolare i parametri e la percentuale di varianza spiegata della regressione

$$y_i = a + b x_i$$

5. Il dataset `spot.dat` contiene i dati relativi ad un esperimento condotto in 22 supermercati comparabili e volto a valutare l'efficacia, in termini di volume di vendite, di due diverse campagne promozionali, radiofonica e postale. Per ciascun supermercato vengono forniti i dati (in migliaia di Euro) relativi alle vendite effettuate (vendite) e alle spese sostenute dal supermercato per la campagna radiofonica (radio) e postale (posta). Per esplorare la relazione tra volume di vendite e spese sostenute, si è adattato un modello di regressione lineare utilizzando un software statistico [in questo caso si è utilizzato il comando `lm()` dell'ambiente R, ma analoghi risultati si sarebbero ottenuti usando un diverso software] ed ottenendo l'output di seguito riportato. In esso, alcune quantità sono state omesse ed altre sostituite con un punto interrogativo (?).

```
Call:
lm(formula = vendite ~ radio + posta)

Residuals:
    Min      1Q  Median      3Q     Max
-278.33 -79.66  -17.11  117.47  290.76

Coefficients:
            Estimate  Std. Error  t value     Pr(>|t|)
(Intercept) 156.430     126.758    1.234        0.232
radio        13.081       1.759    7.435     4.89e-07 ***
posta        16.795       2.963    5.668     1.83e-05 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: ?  on ? degrees of freedom
Multiple R-Squared: 0.8087, Adjusted R-squared: 0.7886
F-statistic: 40.16 on ? and ? DF, p-value: 1.501e-007
```

(a) Si scriva il modello statistico corrispondente.

(b) Indicata con X la matrice di regressione e sapendo che il primo elemento di $\underset{3\times3}{V} = (X^TX)^{-1}$ vale $V_{11} = 0.6363$, si completino i valori mancanti dell'output.

(c) Si fornisca una interpretazione dei coefficienti relativi alle variabili esplicative. Le campagne pubblicitarie hanno un effetto statisticamente significativo sulle vendite?

(d) A parità delle altre condizioni, è più conveniente investire nella campagna radiofonica o postale? Si motivi la risposta.

(e) Si fornisca una previsione delle vendite per un supermercato che ha investito 20000 Euro sia per la campagna pubblicitaria che per quella postale.

(f) Per valutare la bontà dell'adattamento del modello sono state effettuale alcune analisi grafiche. I risultati sono riportati nella figura seguente. Si commentino i risultati.