統計應用數學與計算

汪群超

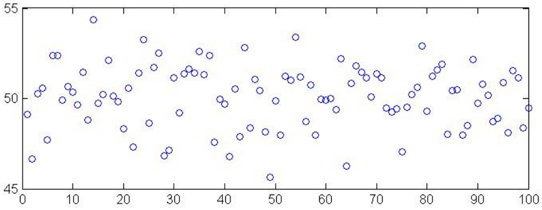
國立台北大學 統計系

May 28, 2013

1. 從平均數說起
   - 算數平均
   - 最小平方法
   - 最大概似函數
   - 貝氏估計
2. 迴歸分析
   - 更多的資訊
3. 時間序列
4. 最大概似函數
5. 線性代數的角色

問題1

使用電子體重計來量測A女的體重，由於體重計本身的誤差，為求較精準，共量測了100次。希望從這100個量測值$y_1, y_2, \cdots, y_{100}$估計 A 女的體重。



巴比倫人的方法

紀元前三百年(戰國時期)

$$\hat{x} = \frac{1}{100}\sum_{i=1}^{100} y_i$$

Legendre 及 Gauss

西元1806,1809年

假設x為 A 女真正的體重,$y_i - x$代表第i次的測量誤差

A女體重的最小平方估計為

$$\hat{x}_{LS} \triangleq \min_{x} \sum_{i=1}^{100} (y_i - x)^2$$

Bernoulli

西元1777年

假設$p(y_1, y_2, \cdots, y_{100}|x)$為量測值的條件式聯合機率密度函數, A女體重的最大 (對數) 概似估計為

$$\hat{x}_{ML} \triangleq \max_x \log p(y_1, y_2, \cdots, y_{100}|x)$$

Bayes

西元1763年

提出貝氏定理: $p(x|y) = \frac{p(y|x)p(x)}{p(y)}$

A女體重的貝氏估計 (或稱 Maximum A Posteriori, MAP 估計) 為

$$\hat{x}_{MAP} \triangleq \max_x \log p(x|y_1, y_2, \cdots, y_{100})$$

Gauss

西元1809年

Gauss建立了量測值的模式 $\mathbf{y} = x + \epsilon$

並做了以下的假設:

- 量測誤差值服從常態分配 $\epsilon_i \sim N(0, \sigma^2)$
- 量測誤差值$\epsilon_1, \epsilon_2, \cdots$為獨立變數
- 未知數x為均等分配 (意即, 對x一無所知)

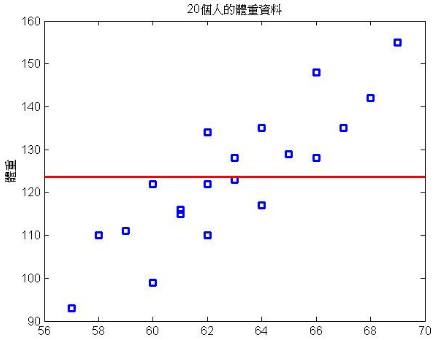
Gauss得到以下結論

$$\hat{x}_{LS} = \hat{x}_{ML} = \hat{x}_{MAP} = \frac{1}{100}\sum_{i=1}^{100} y_i$$

奠定最小平方法的立論基礎，更說明了平均數的內涵。

問題2

已知20個人的體重資料，要預測第21人的體重？



加入身高資料，對預測有幫助嗎？可以從一個人的身高預測其體重？

迴歸分析:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

體重 (Y)	身高 (X)
93	57
110	58
99	60
112	60
⋮	⋮

計算聯立方程式(System of Linear Equations) 的解:β_0, β_1

$$
\begin{aligned}
93 &= \beta_0 + 57\beta_1 \\
110 &= \beta_0 + 58\beta_1 \\
99 &= \beta_0 + 60\beta_1 \\
112 &= \beta_0 + 60\beta_1 \\
\vdots &= \vdots
\end{aligned}
$$

有解? 無解? 無限多組解?

假設 $\hat{\beta}_0, \hat{\beta}_1$為上述聯立方程式的解, 在已知身高$(x)$的情況下, 體重的預測

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

問題3

已知 $y_1, y_2, \cdots, y_N$ 為 N 筆依時間排序的資料, 欲預測尚未發生的第 $N+1$ 筆 y_{N+1}
假設: 每筆時間資料都與其前 p 筆資料有關, 其關係假設為

$$y_n = a_1 y_{n-1} + a_2 y_{n-2} + \cdots + a_p y_{n-p}$$

計算聯立方程式(System of Linear Equations) 的解:$a_1, a_2, \cdots, a_p$

$$
\begin{array}{lclclclcl}
y_{p+1} & = & a_1 y_p & + & a_2 y_{p-1} & + & \cdots & + & a_p y_1 \\
y_{p+2} & = & a_1 y_{p+1} & + & a_2 y_p & + & \cdots & + & a_p y_2 \\
\vdots & & & & & \vdots & & & \\
y_N & = & a_1 y_{N-1} & + & a_2 y_{N-2} & + & \cdots & + & a_p y_{N-p}
\end{array}
$$

有解? 無解? 無限多組解?

假設 $\hat{a}_1, \hat{a}_2, \cdots, \hat{a}_p$為上述聯立方程式的解, 則未知資料 y_{N+1} 的預測為

$$\hat{y}_{N+1} = \hat{a}_1 y_N + \hat{a}_2 y_{N-1} + \cdots + \hat{a}_p y_{N-p+1}$$

問題4

假設 X 為一個服從多項分配的變數,

$X \sim M(N, \theta_1, \theta_2, \theta_3)$, $\theta_1 + \theta_2 + \theta_3 = 1$

已知 $N = 10$ 次的試驗中, 這三項的次數分別為

$2, 3, 5$, 如何對母體的參數 $\theta_1, \theta_2, \theta_3$ 做最好的估計?

假設 $N = Z_1 + Z_2 + Z_3$, 概似函數寫成

$$L(\theta) = \begin{pmatrix} N \\ Z_1 Z_2 Z_3 \end{pmatrix} \theta_{\mathbf{1}}^{\mathbf{Z_1}} \theta_{\mathbf{2}}^{\mathbf{Z_2}} \theta_{\mathbf{3}}^{\mathbf{Z_3}}$$

最大概似函數的參數估計:$\max_\theta L(\theta)$

$$\begin{aligned} \frac{\partial L(\theta)}{\partial \theta_1} &= 0 \\ \frac{\partial L(\theta)}{\partial \theta_2} &= 0 \end{aligned}$$

Matrix Representation

$$A\mathbf{x} = \mathbf{b}$$

where

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$