

《迴歸分析》

試題評析	本年度的試題：除了第一題(三)容易迷失陷阱，比較難拿到15分之外，其他各題都只考公式定義、簡單的運算，因此預估程度中上的同學，分數大概落在70-80分之間，若能考到85分以上，足夠顯示已達上榜水準。
考點命中	第一題：(一)(二)與第二回P54例題2相同，(三)與總複習P67例題9(三)相同 第二題：與第三回P81實例說明相同 第三題：與第三回P63定義、P75例題1相同 第四題：與第一回P92例題2相同 第五題：與第一回P16定理敘述相同

一、在一般迴歸模型及常態誤差假設下，25筆獨立資料所得之配適模型 (fitted model) 為

$$\hat{y} = 2.341 + 1.616x_1 + 0.014x_2,$$

及變異數分析 (ANOVA) 表

Source	Sum of Squares	df	F	PR>F
Model	5550.8	(A)	(D)	0.000
Error	(B)	(C)		
Corrected Total	5784.5			

(一)試填入ANOVA表中 (A)、(B)、(C) 和 (D) 內之數字。(5分)

(二)試問上述迴歸模型是否顯著 ($\alpha = 5\%$) ? (5分)

(三)若X為資料中之設計短陣 (design matrix)，且

$$(X'X)^{-1} = \begin{pmatrix} 0.1132 & -0.00445 & -0.00008 \\ -0.00445 & 0.0027 & -0.00004 \\ -0.00008 & -0.00004 & 0.000001 \end{pmatrix}$$

試檢定 $H_0: \beta_0 = \beta_1 = \beta_2$ vs. $H_1: \beta_i \neq \beta_j$ ，任意 $i, j = 0, 1, 2$ ； i 不等於 j 。請寫出檢定統計量之分布和自由度 ($\alpha = 5\%$)。(臨界值 (critical value) = 3.44。)(15分)

答：

(一)(A)2 (B) $5784.5 - 5550.8 = 233.7$ (C) $n - k = 25 - 3 = 22$

$$(D) \frac{5550.8/2}{233.7/22} = 261.27$$

(二)1. $\begin{cases} H_0: \beta_1 = \beta_2 = 0 & (\text{迴歸模型不顯著}) \\ H_1: \beta_1, \beta_2 \text{不全為} 0 & (\text{迴歸模型顯著}) \end{cases}$

2. $p\text{-value} = 0.000 < \alpha = 0.05$

$\therefore$ reject H_0 ，有充分證據顯示迴歸模型顯著

(三) $H_0: \beta_0 = \beta_1 = \beta_2$

$$\Leftrightarrow H_0: \beta_0 - \beta_1 = 0, \beta_1 - \beta_2 = 0, \beta_0 - \beta_2 = 0$$

$$\Leftrightarrow H_0: \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\Leftrightarrow H_0: \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & -1 \end{bmatrix}_{3 \times 3} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\Leftrightarrow H_0: \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}_{2 \times 3} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Leftrightarrow H_0: [C] \cdot [\beta] = [h]$$

$$Cb_F - h = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} 2.341 \\ 1.616 \\ 0.014 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.725 \\ 1.602 \end{bmatrix}$$

$$C(X^T X)^{-1} C^T = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 0.1132 & -0.00445 & -0.00008 \\ -0.00445 & 0.0027 & -0.00004 \\ -0.00008 & -0.00004 & 0.000001 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1248 & -0.00711 \\ -0.00711 & 0.002781 \end{bmatrix}$$

$$[C(X^T X)^{-1} C^T]^{-1} = \frac{\begin{bmatrix} 0.002781 & 0.00711 \\ 0.00711 & 0.1248 \end{bmatrix}}{0.0002965167}$$

$$1. (1) \begin{cases} H_0 : \beta_0 = \beta_1 = \beta_2 \\ H_1 : \beta_i \neq \beta_j, i, j = 0, 1, 2, i \neq j \end{cases}$$

$$\Leftrightarrow \begin{cases} H_0 : \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ H_1 : \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} \neq \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{cases}$$

$$\Leftrightarrow \begin{cases} H_0 : [C]_{S \times k} \cdot [\beta]_{k \times 1} = [h]_{S \times 1} \\ H_1 : [C]_{S \times k} \cdot [\beta]_{k \times 1} \neq [h]_{S \times 1} \end{cases}$$

$$(2) F = \frac{(Cb_F - h)^T [C(X^T X)^{-1} C^T]^{-1} (Cb_F - h) / S}{SSE(F) / (n - k)}$$

$$= \frac{\begin{bmatrix} 0.725 & 1.602 \end{bmatrix} \cdot \frac{\begin{bmatrix} 0.002781 & 0.00711 \\ 0.00711 & 0.1248 \end{bmatrix}}{0.0002965167} \cdot \begin{bmatrix} 0.725 \\ 1.602 \end{bmatrix} / 2}{233.7 / (25 - 3)}$$

$$= 53.70 > 3.44 \text{ (critical value)}$$

$\therefore$ reject H_0 , 有充分證據顯示 $\beta_i \neq \beta_j, i, j = 0, 1, 2, i \neq j$

$$2. \text{檢定統計量 } F = \frac{(Cb_F - h)^T [C(X^T X)^{-1} C^T]^{-1} (Cb_F - h) / S}{SSE(F) / (n - k)} \xrightarrow{H_0} F(2, 22)$$

二、迴歸模型中解釋變數間若存在共線性 (multicollinearity) 對估計結果影響甚鉅。變異數膨脹因子 (variance inflation factor, VIF) 是判斷共線性的一個指標。

(一) VIF 之意涵為何？試說明 VIF 和共線性的關係。(10分)

(二) 某組資料有 5 個解釋變數 x_1, x_2, x_3, x_4, x_5 所得迴歸係數估計結果如下：

Variable	b_j	se	VIF
Intercept	0.830	0.318	
x_1	-0.012	0.647	(0.002) ⁻¹
x_2	0.199	0.483	(0.001) ⁻¹
x_3	-0.117	0.178	(0.010) ⁻¹
x_4	-0.367	0.294	(0.008) ⁻¹
x_5	0.186	0.147	(0.009) ⁻¹

其中 b_j 為迴歸係數 B_j 之估計值，se 為其標準誤。試評估解釋變數中是否存在共線性？(5分)

答：

(一) 1. 定義： $VIF_j = \frac{1}{1 - R_j^2}$ ：

用來判斷第 j 個解釋變數與其他解釋變數之間共線性之強弱

2. VIF_j 越大 $\Rightarrow$ 第 j 個解釋變數與其他解釋變數之間共線性越強

用法： $\overline{VIF} = \frac{\sum_{j=1}^{p-1} VIF_j}{p-1}$ 越大 $\Rightarrow$ 迴歸模型共線性越大

【版權所有，重製必究！】

其中： $p-1$ 為迴歸模型中解釋變數的個數

(二) $\because VIF_j > 1, j = 1, 2, 3, 4, 5 \Rightarrow \overline{VIF} > 1$

$\therefore$ 解釋變數之間存在共線性

三、在有三個解釋變數之一組資料配適迴歸模型後，得到所有部分集合之變數選擇 (all possible subsets selection) 結果如下 (b_j 為迴歸係數 B_j 之估計值)：

$p-1$	R^2	C_p	b_0	b_1	b_2	b_3
1	0.032	75.45	0.77	0.325		
1	0.705	11.85	0.27		1.361	
1	0.707	11.67	-0.36			1.103
2	0.759	8.79	-0.29	-0.463		1.255
2	0.808	3.12	-0.19		0.781	0.633
2	0.830	2.03	-0.93	0.655	1.487	
3	0.831	4.00	-0.12	0.737	1.589	-0.094

(一)何謂 C_p 試說明其意義。(10分)

(二)由表中結果來看，最佳模型為何？為什麼？(10分)

答：

(一) $C_p = \frac{SSE_p}{MSE(Full)} - (n - 2p)$ ：迴歸模型用來搜尋最佳自變數的組合方法，

取 ① C_p 值最小者 ② C_p 值最接近 P 值者：即為最佳自變數組合

(二) $\because C_p = 2.03$ 最小，且最接近 $P = 3$

$\therefore$ 最佳模型： $Y_i = -0.93 + 0.655X_{i1} + 1.487X_{i2}$

四、下述資料為某幼稚園自80年後至90年之幼童入學學費。

年	81	82	83	84	85	86	87	88	89	90
80年後之年分， x	1	2	3	4	5	6	7	8	9	10
學費(千元) y	6.1	6.8	7.5	8.5	9.3	10.5	11.5	12.625	13.975	14.975

(一)由 (y, x) 之散布圖 (scatter plot) 發現 y 和 x 的關係較接近 $y = \alpha e^{\beta x}$ 。欲得一線性模型，須將 y 作何轉換 (transformation)？試寫出轉換後的模型。(10分)

(二)令 z_i 為反應變數 y_i 轉換後的值，且 $\sum_{i=1}^{10} z_i = 22.783$ ， $\sum_{i=1}^{10} x_i z_i = 133.686$ 。試求 α 和 β 的最小平方估計量 (least square estimate)。(10分)

(三)根據(二)之結果，預測100年時該幼稚園之幼童入學學費。(10分)

答：

(一) $Y = \alpha e^{\beta x}$

取 $\ln$ ： $\ln Y = \ln \alpha + \beta x + \varepsilon$

令 $Z = \alpha^* + \beta x + \varepsilon$ ，其中 $Z = \ln Y$ ， $\alpha^* = \ln \alpha$

(二) $S_{XZ} = \sum xz - \frac{(\sum x)(\sum z)}{n} = 133.686 - \frac{55 \times 22.783}{10} = 8.3795$

$S_{XX} = \sum x^2 - \frac{(\sum x)^2}{n} = 385 - \frac{55^2}{10} = 82.5$

$\hat{\beta} = \frac{S_{XZ}}{S_{XX}} = \frac{8.3795}{82.5} = 0.1016$

$\hat{\alpha}^* = \bar{z} - \hat{\beta} \bar{x} = \frac{22.783}{10} - 0.1016 \times \frac{55}{10} = 1.7195$

$\Rightarrow \hat{\alpha} = e^{\hat{\alpha}^*} = 5.5817$

$\therefore (\hat{\alpha}, \hat{\beta}) = (5.5817, 0.1016)$

【版權所有，重製必究！】

$$(三) \hat{Y} = \hat{\alpha}e^{\hat{\beta}x} = 5.5817e^{0.1016x}$$

$$\therefore \hat{Y}_{x=20} = 5.5817e^{0.1016 \cdot 20} = 42.5846$$

五、假設 (y_i, x_i) 滿足 $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$ ， $i=1, \dots, n$ 。令 b_j 為迴歸係數 β_j 之最小平方估計量， $j=0, 1, \dots, p$ 。著名的高斯－馬可夫（Gauss-Markov）定理是對 $\sum_{j=0}^p l_j \beta_j$ 之估計有興趣，其中 $l_0, \dots, l_p$ 是已知的實數。試敘述 Gauss-Markov 定理及其假設條件。（10分）

答：

Gauss-Markov 定理：迴歸模型： $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$

當誤差項 ε_i 滿足：傳統線性迴歸假設

$$\textcircled{1} E\varepsilon_i = 0, \textcircled{2} V(\varepsilon_i) = \sigma^2, \textcircled{3} \text{cov}(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$$

則 OLSE $\sum_{j=0}^p l_j b_j$ 為 $\sum_{j=0}^p l_j \beta_j$ 之 BLUE (Best Linear Unbiased Estimator)

【版權所有，重製必究！】