

《迴歸分析》

試題評析

今年迴歸分析考題和往年類似皆以簡單計算為主，觀念題為輔，且考題內容皆屬於一般化考題淺顯易解。第一題與講義第一回第二章例1及例4題目類似，而第四題虛擬變數考題與講義第三回中之例8題完全相同，同學應該有較充裕的時間去思考另二題基本的觀念題，因此一般同學要獲取80分以上應該很容易。

一、通常在早餐中，小朋友會喜歡"喜瑞爾" (cercal) 伴隨牛奶食用。令Y表示早餐喜瑞爾保持脆度的壽命（以分鐘計）。而X表示牛奶的溫度（以攝氏表示）。一位調查者想了解喜瑞爾的脆度和牛奶的溫度之間是否有關聯。該調查者事先選擇幾個不同溫度並記錄溫度（x1值）及脆度的壽命（y1值）。用這些樣本資料，我們想知道配適線性迴歸模式 $Y1 = b0 + b1X1 + e1$ 是否合適？所收集的數據如下：

表1

X	Y	X	Y	X	Y	X	Y
40	13.8	45	12.5	50	10.5	55	8.7
40	14.8	45	12.7	50	10.2	55	8.7
40	11.1	45	10.9	50	8.7	55	6.8
40	11.3	45	10.5	50	8.2	55	6.6
40	9.7	45	6.5	50	6.5	55	3.6
40	8.7	45	7.1	50	5.2	55	4.3

我們擬作缺適檢定 (lack-of-fit test)，一些可能會用到的統計量彙整如下：

$$\bar{y} = 9.0667, \quad \bar{x} = 47.5, \quad S_{xx} = \sum(x_i - \bar{x})^2 = 750, \quad S_{yy} = \sum(y_i - \bar{y})^2 = 195.653, \quad S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = -257.5$$

表2

X的水準	$\sum_j (y_j - \bar{y}_j)^2$	自由度 (df)
40		
45		
50		
55		

表3 ANOVA (變異數分析法)

變異來源	平方和 (SS)	自由度 (d.f.)	均方和 (mean square)	F檢定
迴歸				
殘差				
(Lack of fit)				
(Pure error)				
總和	195.653			

依據前述資訊，請將表2及表3繪於試券上，並完成之。在顯著水準 $\alpha = 0.05$ ，作F檢定 (overall effect) 及缺適檢定，需列出虛無假設與對立假設並判斷檢定結果。此迴歸模式是否描述資料適當？（請詳述你的理由）。（三十分）

答：

由表一之資料可知

$$\sum_{i=1}^{n_1} (y_{i1} - \bar{y}_1)^2 = \sum_{i=1}^{n_1} y_{i1}^2 - \frac{\left(\sum_{i=1}^{n_1} y_{i1}\right)^2}{n_1} = 830.16 - \frac{(69.4)^2}{6} = 27.4333$$

$$\sum_{i=1}^{n_2} (y_{i2} - \bar{y}_2)^2 = \sum_{i=1}^{n_2} y_{i2}^2 - \frac{\left(\sum_{i=1}^{n_2} y_{i2}\right)^2}{n_2} = 69.26 - \frac{(60.2)^2}{6} = 35.2533$$

$$\sum_{i=1}^{n_3} (y_{i3} - \bar{y}_3)^2 = \sum_{i=1}^{n_3} y_{i3}^2 - \frac{\left(\sum_{i=1}^{n_3} y_{i3}\right)^2}{n_3} = 426.51 - \frac{(49.3)^2}{6} = 21.4283$$

$$\sum_{i=1}^{n_4} (y_{i4} - \bar{y}_4)^2 = \sum_{i=1}^{n_4} y_{i4}^2 - \frac{\left(\sum_{i=1}^{n_4} y_{i4}\right)^2}{n_4} = 272.63 - \frac{(38.7)^2}{6} = 23.015$$

故知 (表二) 為

X的水準	$\sum_j (y_j - \bar{y}_j)^2$	自由度 (df)
40	27.4333	5
45	35.2533	5
50	21.4283	5
55	23.015	5

$$\text{又 } \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{-257.5}{750} = -0.3433$$

$$\therefore \text{SSE} = S_{yy} - \hat{\beta}_1^2 \cdot S_{xx} = 195.653 - (-0.3433)^2 \cdot (750) = 107.2618$$

$$\text{又 SSPE} = \sum_{i=1}^n \sum_{j=1}^4 (y_j - \bar{y}_j)^2 = 27.4333 + 35.2533 + 21.4283 + 23.015 = 107.1299$$

$$\therefore \text{SSL} = \text{SSE} - \text{SSPE} = 107.2618 - 107.1299 = 0.1319$$

故 (表三) 為

變異來源	平方和 (SS)	自由度 (d.f.)	均方和 (mean square)	F檢定
迴歸	88.3912	1	88.3912	18.1297
殘差	107.2618	22	4.8755	
(Lack of fit)	0.1319	2	0.06595	0.0123
(Pure error)	107.1299	20	5.3565	
總和	195.653	23		

$$\text{又 } \textcircled{1} H_0: \beta_1 = 0$$

$$\textcircled{2} H_1: \beta_1 \neq 0$$

$$\textcircled{3} \alpha = 0.05$$

$$\textcircled{4} C = \{F \mid F > F_{0.05}(1, 22) = 4.3\}$$

$$\textcircled{5} \text{計算: } \because F = 18.1297 \in C$$

$$\textcircled{6} \text{結論: 拒絕 } H_0$$

而缺適檢定為

$$\textcircled{1} H_0: \mu_{y|x} = \beta_0 + \beta_1 X$$

$$\textcircled{2} H_1: \mu_{y|x} \neq \beta_0 + \beta_1 X$$

$$\textcircled{3} \alpha = 0.05$$

$$\textcircled{4} C = \{F - F > F_{0.05}(2, 20) = 3.49\}$$

$$\textcircled{5} \text{計算: } \because F = 0.0123 \notin C$$

$$\textcircled{6} \text{結論: 不拒絕 } H_0, \text{ 即 } X \text{ 與 } Y \text{ 適合直線迴歸}$$

二、從流行病學的角度，我們有興趣分析台灣地區每日SARS（嚴重急性呼吸道症候群）可能病例的累積人數。我們選擇解釋變數是時間 t ，樣本收集從四月二十一日至五月二十日。 Y 表示每日可能病例的累積人數。觀察散佈圖（因篇幅限制，無法提供）發現 Y 與 t 的關係成曲線狀而非線性關係。擬初步配適模式如下：

$\ln Y_t = \beta_0 + \beta_1 t + \varepsilon_t$, $t = 1, \dots, n$ ，此處 $\ln Y_t$ 表示對 Y 作自然對數轉換。

所作的迴歸分析，請參見下列報表。

(一)列出上述迴歸線之估計式。（十分）

(二)評論此迴歸模式是否適當。為什麼適當或不適當？請詳述你的理由。若你認為模式不合適，請列出建議模式。（十五分）

Dependent Variable: lnY					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr < F
Model	1	18.25579	18.25579	2391.95	< .0001
Error	29	0.22133	0.00769		
Corrected Total	30	18.47713			
Root MSE		0.08736			
Dependent Mean		4.79724	R-Square	0.9880	
Coeff Var		1.82110	Adj R-Sq	0.9876	
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept (β_0)	1	3.42448	0.03216	106.49	< .0001
Intercept (β_1)	1	0.08580	0.00175	48.91	< .0001
Durbin-Watson D				0.527	
Number of Observations				31	
1st Order Autocorrelation				0.685	

答：

(一)迴歸方程式之估計式為

$$\ln Y_t = 3.42448 + 0.0858t$$

(二)由電腦報表得知，自我相關之檢定其檢定統計量為

$$\text{Durbin-Watson } D = 0.527$$

因 D 太小落入危險域，故此模式存在正自我相關，因此可說此迴歸模式並不適當。若要消除自身相關，可增加自變數個數，亦即尋找失落之關鍵變數，使模式改為多元迴歸模式去作分析，ex: $\ln Y_t = \beta_0 + \beta_1 t + \beta_2 X + \varepsilon_t$

三、我們考慮配適簡單線性迴歸模式 $Y_i = \beta_0 + \beta_1 f_i + \varepsilon_i$, $i = 1, 2, \dots, n$ 。請寫出下列每一個陳述中所必要用到之假設。請詳述在何處需要用到你所提的假設。

(一)最小平方估計 β_0 和 β_1 。（八分）

(二)作 t 和 F 檢定。（八分）

(三)證明參數的最小平方估計式是不偏的 (unbiased) 及一致的 (consisten) 估計量。（九分）

答：

(一)推求 β_1 , β_2 之最小平方估計式，我們要假設變異數具有同質性，否則判求出之OLS之估計式會不具有有效性，即變異數並非最小變異。

(二)因要發展 t 及 F 檢定統計量去作參數 β_0 , β_1 之統計推論問題時，必要假設母體為常態及誤差項隨機性，否則便不可利用 $\hat{\beta}_1$ 作 β_1 之統計推論。

(三)在驗證不偏性及一致性時，我們須要假設自變數 X 為非隨機變數，且 ε_i 之期望值為0，以及具有隨機獨立性。

四、我們有興趣了解房價與一些變數之間的關係。令Y表示房屋仲介公司對託售房屋之出價金額（萬元）。我們收集有關託售房屋之坪數（X）及房屋坐落之區域，共有三區。考慮配適之線性迴歸模式如下，

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_{1,i} + \beta_3 D_{2,i} + \beta_4 X_i D_{1,i} + \beta_5 X_i D_{2,i} + \varepsilon_i, \quad i = 1, \dots, 30,$$

此處

Y：房屋之出價金額（萬元）

X：房屋之坪數

$$D_1 = \begin{cases} 1 & \text{房屋坐落在區域1} \\ 0 & \text{其他} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{房屋坐落在區域2} \\ 0 & \text{其他} \end{cases}$$

且 ε_i 是獨立具平均數為0、變異數為 σ^2 常態分配之隨機變數。

(一)解釋如何檢定房屋坐落在三個區域之三條迴歸線相互平行。設定義你在解釋中所需用到的任何符號，並請列出虛無假設、對立假設、檢定統計量及決策法則，不需查表。（十分）

(二)請解釋如何檢定房屋坐落在三個區域之三條迴歸線其截距項是相同的。請定義你在解釋中所需用到的任何符號，並請列出虛無假設、對立假設、檢定統計量及決策法則，不需查表。（十分）

答：

(一) $\because E(Y) = (\beta_0 + \beta_1 X + \beta_2 D_1 + \beta_3 D_2 + \beta_4 X D_1 + \beta_5 X D_2 + \beta_5 X D_2)$

$\therefore E(Y | D_1 = 1, D_2 = 0) = \beta_0 + \beta_1 X + \beta_2 + \beta_4 X = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)X \dots\dots\dots (\quad \text{區})$

$E(Y - D_1 = 0, D_2 = 1) = \beta_0 + \beta_1 X + \beta_3 + \beta_5 X = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)X \dots\dots\dots (\quad \text{區})$

$E(Y - D_1 = 0, D_2 = 0) = \beta_0 + \beta_1 X \dots\dots\dots (\quad \text{區})$

故欲檢定三條迴歸線是否相互平行，可檢定

① $H_0: \beta_4 = \beta_5 = 0$

② $H_1: \beta_i$ 不至為0, $i = 4, 5$

③ 檢定統計量為

$$F = \frac{[SSE(X, D_1, D_2) - SSE(X, D_1, D_2 \times D_1, X D_2)] \div [(n-3) - (n-5)]}{SSE(X, D_1, D_2 \times D_1, X D_2) \div (n-5)}$$

④ 決策法則為

若 $F > F_{\alpha}(2, 25)$ ，則拒絕 H_0

(二) ① $H_0: \beta_2 = \beta_3 = 0$

② $H_1: \beta_i$ 不全為0, $i = 2, 3$

③ 檢定統計量為

$$F = \frac{[SSE(X, X D_1, X D_2) - SSE(X, D_1, D_2 \times D_1, X D_2)] \div [(n-3) - (n-5)]}{SSE(X, D_1, D_2 \times D_1, X D_2) \div (n-5)}$$

④ 決策法則為

若 $F > F_{\alpha}(2, 25)$ ，則拒絕 H_0