

等 別：三等考試

類 科：統計

科 目：迴歸分析

考試時間：2 小時

座號：\_\_\_\_\_

※注意：(一)可以使用電子計算器。

(二)不必抄題，作答時請將試題題號及答案依照順序寫在試卷上，於本試題上作答者，不予計分。

一、對一組樣本資料  $(x_i, y_i)$ ,  $i = 1, \dots, 200$ ，配適一簡單線性迴歸模型  $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ ，其中  $\epsilon_i$  為 i.i.d.  $N(0, \sigma^2)$ 。已知  $x_1 = 0$  且  $x_i = 1$  當  $2 \leq i \leq 100$ ，令  $\beta_0$  與  $\beta_1$  之最小平方估計值為  $\hat{\beta}_0$  與  $\hat{\beta}_1$ 。試說明下列敘述為何是正確的：

(一)  $(0.99, \bar{y})$  及  $(x_1, y_1)$  必定位於樣本迴歸線上。(10 分)(二)  $(0, y_1)$  必定位於樣本迴歸線上且殘差中至少有一取值為 0。(10 分)

二、美國黃石國家公園為了協助旅客規劃其旅遊路線及行程，對其最受歡迎的景點老忠實間歇泉，提供其噴泉時間的預測值。首先於 1978 年至 1979 年間蒐集了 222 組數據，將第  $k$  次噴泉的持續時間記為  $x_k$ ，用  $x_k$  來預測  $y_k$ 。而  $y_k$  為第  $k+1$  次噴泉爆發時間及第  $k$  次噴泉爆發時間的間距。修正誤差項之相關性後，僅使用 221 組數據，其迴歸分析的結果如下：

ANOVA (Analysis of Variance)

變異來源	自由度	SS	MS	顯著值
迴歸	(a)	18975	18975	0.000
殘差	219	(b)	35.19	
總和	220	(c)		

Predictor	係數	標準誤	t 統計	顯著值
截距	(d)	1.428	23.75	0.000
$X_1$	(e)	0.3822	23.22	0.000

試根據上述資料及電腦分析結果，回答下列問題：

(一) 試問分析結果的(a)至(e)及  $\sum_{k=1}^{221} (y_k - \bar{y})^2$  和  $\sum_{k=1}^{221} (y_k - \hat{y}_k)^2$  之值為何？(9 分)(二) 試寫出樣本迴歸方程式、 $R^2$  統計值及調整  $R^2$  統計值。(9 分)(三) 在  $\alpha = 0.05$  之下，請利用  $F$  檢定及  $t$  檢定判斷此樣本迴歸模式之顯著性？

( $F_{0.05,1,219} = 3.8843$ ,  $F_{0.05,1,220} = 3.8840$ ,  $F_{0.025,1,219} = 5.0937$ ,  $F_{0.025,1,220} = 5.0934$ ) 請寫出各檢定之虛無假設及對立假設並詳述理由。若您認為此樣本迴歸模式不合適，請寫出您的建議模式。(7 分)

(請接背面)

等 別：三等考試

類 科：統計

科 目：迴歸分析

三、已知  $Y_1 = \alpha_1 + \epsilon_1$ ,  $Y_2 = 2\alpha_1 - \alpha_2 + \epsilon_2$ ,  $Y_3 = \alpha_1 + 2\alpha_2 + \epsilon_3$  其中  $\epsilon_i, 1 \leq i \leq 3$  為 i.i.d.  $N(0, \sigma^2)$ 。令  $\hat{\alpha}_1$  及  $\hat{\alpha}_2$  為  $\alpha_1$  及  $\alpha_2$  的最小平方估計量。

(一)當使用矩陣表達時，我們可得出下式：

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \mathbf{X} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

試寫出矩陣  $\mathbf{X}$  及變數間是否有共線性。(8分)

(二)試求  $Cov(\hat{\alpha}_1, \hat{\alpha}_2)$  及  $Var(\hat{\alpha}_1)$ 。(7分)

(三)當使用  $F$  檢定，檢定  $H_0: \alpha_1 = \alpha_2$  時，你需要的  $F$  分配表的自由度為何？(5分)

(四)當  $y_1 = 1.8722$ ,  $y_2 = -0.4853$ ,  $y_3 = 4.5169$  時， $\hat{\alpha}_1 = 0.9031$ ,  $\hat{\alpha}_2 = 1.0004$ ，殘差平方和  $RSS = 14.2495$ 。在  $\alpha = 0.05$  時，請利用  $F$  檢定，檢定  $H_0: \alpha_1 = \alpha_2$  vs  $H_1: \alpha_1 \neq \alpha_2$ 。

( $F_{0.05,1,1} = 161.45$ ,  $F_{0.05,1,2} = 18.51$ ,  $F_{0.05,2,1} = 199.5$ ) (10分)

四、考慮以下數據  $(Y_i, X_i)$ ,  $1 \leq i \leq n$ ，且  $Y_i = X_i + \epsilon_i$ ，其中  $X_i = i/50$  且  $\epsilon_i$  為 i.i.d.  $N(0, 1)$  之隨機變數。當試著以下列 3 個多項式迴歸模式來配適：

$$M_1: Y_i = \beta_{10} + \epsilon_i$$

$$M_2: Y_i = \beta_{20} + \beta_{21}X_i + \epsilon_i$$

$$M_3: Y_i = \beta_{30} + \beta_{31}X_i + \beta_{32}X_i^2 + \epsilon_i$$

且以這 3 個不同的線性迴歸模式中達到 Mallows'  $C_p$  統計量最小者，做為最佳配適模式。當用迴歸模式  $M$  來配適時，其  $C_p$  統計量如下：

$$C_p(M) = \frac{RSS(M)}{\sigma^2} + 2 \times \dim(M) - n$$

於上式中， $RSS(M)$  代表配適迴歸模式  $M$  的殘差平方和， $\dim(M)$  代表迴歸模式  $M$  的未知參數個數。

(一)以  $C_p$  為準則，請使用  $RSS(M_1)$ ,  $RSS(M_2)$ ,  $RSS(M_3)$  此 3 個統計量，寫出迴歸模式  $M_2$  被選取的條件。(6分)

(二)請描述相關於隨機變數  $RSS(M_2)$  及隨機變數  $RSS(M_3)$  的分配分別為何？(6分)

(三)請問迴歸模式  $M_3$  被選取的機率為何？可用標準常態分配函數來表示。(10分)

(四)當  $RSS(M_1) = 138.53$ ,  $RSS(M_2) = 102.93$ ,  $RSS(M_3) = 102.89$  時，那個模型會被選取？試說明理由。(3分)