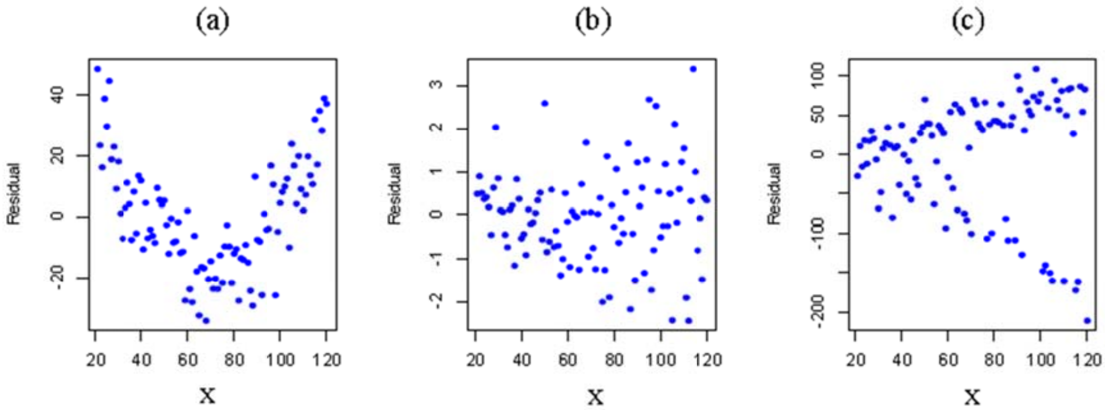
《迴歸分析》

試題評析	今年命題內容包括：觀念問答敘述、矩陣運算、公式的靈活應運。考生必須觀念清楚、公式熟練、計算快速、表達流暢才有可能拿到高分。預估上榜的水準大概80分左右。
考點命中	第一題：(a)《高點迴歸分析講義》第三回，秦大成編撰，頁61。(b)《高點迴歸分析講義》第三回，秦大成編撰，頁11。(c)《高點迴歸分析講義》第三回，秦大成編撰，頁45。 第二題：《高點迴歸分析講義》第三回，秦大成編撰，頁95《例題7》。 第三題：(一)《高點迴歸分析講義》第二回，秦大成編撰，頁64《例題4》。(二)(三)《高點迴歸分析講義》第三回，秦大成編撰，頁93《例題4》。 第四題：(一)(二)《高點迴歸分析講義》第一回，秦大成編撰，頁49《例題1》。(三)《高點迴歸分析講義》第三回，秦大成編撰，頁P71《例題3》。(四)《高點迴歸分析講義》第三回，秦大成編撰，頁P74。(五)《高點迴歸分析講義》第三回，秦大成編撰，頁80《例題1》。

附表：F分佈 $\alpha=0.1$ 臨界值 $F_{df1,df2,0.1}$

df1\df2	15	16	17	18	19	20
1	3.073	3.048	3.026	3.007	2.990	2.975
2	2.695	2.668	2.645	2.624	2.606	2.589
3	2.490	2.462	2.437	2.416	2.397	2.380
4	2.361	2.333	2.308	2.286	2.266	2.249

一、若考慮配適一簡單線性迴歸模型 $y=\alpha+\beta x+\varepsilon$，其中 α、β 為參數，ε 為隨機誤差，且假設其為具均數0，標準差 σ 之常態分配。今於配適模型後，繪出殘差對自變數x的分析圖。請分別針對圖(a)-(c)的結果，說明迴歸模型是否恰當？若模型不恰當時，請指出對於參數估計值是否會有偏差（bias）之影響，對於有關參數的假設檢定是否正確，另外也請提出修正的方法。（18分）



答：

(a)(1)迴歸模型設定不恰當

修正模型：$Y=\beta_0+\beta_1 x+\beta_2 x^2+\varepsilon$

$\quad\quad=\beta_0+\beta_1 x_1+\beta_2 x_2+\varepsilon$

(2)迴歸模型遺漏一個關鍵自變數，參數估計值(OLSE)”可能”偏差

①當 $r_{x_1x_2}=0 \Rightarrow E[\hat{\alpha}]\neq\alpha$ (偏)，但 $E[\hat{\beta}]=\beta$(不偏)

②當 $r_{x_1x_2}\neq 0 \Rightarrow E[\hat{\alpha}]\neq\alpha$ (偏)，且 $E[\hat{\beta}]\neq\beta$(偏)

(3)①當 $r_{x_1x_2}=0 \Rightarrow$ 不影響 β 檢定，但 α 檢定不正確

②當 $r_{x_1x_2}\neq 0 \Rightarrow \alpha,\beta$ 檢定皆不正確

(b) (1)根據殘差圖：異質變異性，迴歸模型設定依然恰當，但求出之 OLSE 仍具不偏性，但不具效性，因此影響參數檢定不正確

(2)修正方法：求加權 OLSE

(c) (1) 迴歸模型設定不恰當

修正模型：$Y = \beta_0 + \beta_1 x + \beta_2 D + \beta_3 xD + \varepsilon$ (X, D 有交互作用)

$$= \begin{cases} \beta_0 + \beta_1 x + \varepsilon & \text{, if } D = 0 \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x + \varepsilon, & \text{if } D = 1 \end{cases}$$

(2)迴歸模型未設定啞變數 D 及交互作用，使得參數估計值(OLSE)產生偏差，且參數檢定不正確。

二、根據下列3變數，6個觀察值的資料

Y	1	0	1	1	0	0
X1	1	-2	1	0	0	0
X2	0	1	2	2	1	0

(一)令Y、X1、X2分表各變數觀察值所形成的向量，另定義X0為長度等於6且元素均等於1的向量。在以向量表示法的迴歸模型M：$Y = \beta_0 X0 + \beta_1 X1 + \beta_2 X2 + \varepsilon$ 中，如何將 $\beta_0 X0 + \beta_1 X1 + \beta_2 X2$ 更精簡的以矩陣與參數向量表示？另外，在一般情形下，此時ε之機率分佈為何？（4分）

(二)計算迴歸模型M中之參數向量的最小平方估計量及估計其變異數共變異數矩陣（variance-covariance matrix）。（8分）

(三)令 $\hat{Y}$ 為長度等於6的向量，其元素為迴歸模型M對Y的配適值（fitted values），則存在一矩陣H使得 $\hat{Y} = HY$，計算此矩陣H。（4分）

(四)計算迴歸模型M中的變異數膨脹因子（variance inflation factor, vif） vif(X1)與vif(X2)。（4分）

答：

(一) (1) $$\begin{bmatrix} 1 & 1 & 0 \\ 1 & -2 & 1 \\ 1 & 1 & 2 \\ 1 & 0 & 2 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} = X\beta$$

(2) $$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{bmatrix} \sim N(0_{6\times 1}, \sigma^2 \cdot I_6)$$

(二) $$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & -2 & 1 \\ 1 & 1 & 2 \\ 1 & 0 & 2 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 6 & 0 & 6 \\ 0 & 6 & 0 \\ 6 & 0 & 10 \end{bmatrix},$$

$$(X^T X)^{-1} = \begin{bmatrix} 5/12 & 0 & -1/4 \\ 0 & 1/6 & 0 \\ -1/4 & 0 & 1/4 \end{bmatrix}$$

$$X^TY=\begin{bmatrix}1&1&1&1&1&1\\1&-2&1&0&0&0\\0&1&2&2&1&0\end{bmatrix}\begin{bmatrix}1\\0\\1\\1\\0\\0\end{bmatrix}=\begin{bmatrix}3\\2\\4\end{bmatrix}$$

(1)OLSE $\hat{\beta}=(X^TX)^{-1}X^TY=\begin{bmatrix}5/12&0&-1/4\\0&1/6&0\\-1/4&0&1/4\end{bmatrix}\begin{bmatrix}3\\2\\4\end{bmatrix}=\begin{bmatrix}1/4\\1/3\\1/4\end{bmatrix}$

(2) $SSTO=\sum y^2-\frac{(\sum y)^2}{n}=3-\frac{3^2}{6}=\frac{3}{2}$

$SSR=Y^TY-\hat{\beta}^T X^TY=\frac{7}{12}\ \Rightarrow\ SSE=\frac{11}{12}$, $MSE=\frac{SSE}{n-k}=\frac{11}{36}$

variance-covariance matrix (dispersion matrix)

$cov(Y)=cov(X\beta+\varepsilon)=cov(\varepsilon)=\sigma^2\cdot I_6$

$\therefore\ \widehat{cov}(Y)=MSE\cdot I_6=\frac{11}{36}\cdot I_6$

(三) $H=X(X^TX)^{-1}X^T$

$$=\begin{bmatrix}1&1&0\\1&-2&1\\1&1&2\\1&0&2\\1&0&1\\1&0&0\end{bmatrix}\begin{bmatrix}5/12&0&-1/4\\0&1/6&0\\-1/4&0&1/4\end{bmatrix}\begin{bmatrix}1&1&1&1&1&1\\1&-2&1&0&0&0\\0&1&2&2&1&0\end{bmatrix}$$

$$=\begin{bmatrix}\frac{7}{12}&\frac{-1}{6}&\frac{1}{12}&\frac{-1}{12}&\frac{1}{6}&\frac{5}{12}\\ \frac{-1}{6}&\frac{5}{6}&\frac{-1}{6}&\frac{1}{6}&\frac{1}{6}&\frac{1}{6}\\ \frac{1}{12}&\frac{-1}{6}&\frac{7}{12}&\frac{5}{12}&\frac{1}{6}&\frac{-1}{12}\\ \frac{-1}{12}&\frac{1}{6}&\frac{5}{12}&\frac{5}{12}&\frac{1}{6}&\frac{-1}{12}\\ \frac{1}{6}&\frac{1}{6}&\frac{1}{6}&\frac{1}{6}&\frac{1}{6}&\frac{1}{6}\\ \frac{5}{12}&\frac{1}{6}&\frac{-1}{12}&\frac{-1}{12}&\frac{1}{6}&\frac{5}{12}\end{bmatrix}$$

(四) $n=6$, $\sum x_1=0$, $\sum x_2=6$, $\sum x_1x_2=0$

$S_{12}=\sum x_1x_2-\frac{(\sum x_1)(\sum x_2)}{n}=0-\frac{0}{6}=0\ \Rightarrow r_{12}=0$

$R_{XX}=\begin{bmatrix}1&r_{12}\\r_{12}&1\end{bmatrix}=I\ \Rightarrow\ R_{xx}^{-1}=I\ \Rightarrow\ Vif(X_1)=Vif(X_2)=1$

三、三高（高血壓、高血糖、高血脂）與許多重大慢性病皆有重要關係。為了解個人體質、生活習慣等對於三高的影響因子，並對社會大眾提出建議與注意事項。因此，研究人員由臺灣數個醫學中心，採用隨機抽樣法蒐集了10000個就診慢性病者的資料進行調查分析。該資料測量每個人的血壓（以收縮壓為例，單位為mmHg）及其他相關變數如下：

性別（男性為1，女性為0），年齡（25～85歲），身體質量指數BMI（定義為身高/體重^2，單位

為m/kg^2），量血壓習慣（有量血壓習慣者為1，反之為0），量血糖習慣（有量血糖習慣者為1，反之為0），量血脂習慣（有量血脂習慣者為1，反之為0），喝酒習慣（平均每天喝1瓶600c.c.啤酒或相當之酒類以上者為1，反之為0），抽煙習慣（有抽煙習慣者為1，反之為0），外食頻率（每週外食次數），運動習慣（有運動習慣者為1，反之為0），睡眠品質（睡眠品質佳者為1，反之為0）。研究者建立血壓(y)對所有解釋變數的迴歸模型，得到如下表（LM1）之結果，其殘差分析也無明顯瑕疵。

(一)模型LM1之所有變數的解釋力為多少？一般來說，此解釋力算是高、中或低？並解釋表中「F-statistic：4961 on 11 and 9988 DF, p-value：< 2.2e-16」之意義。（4分）

(二)在模型LM1下，以兩人之不同的性別、年齡及BMI解釋參數估計值所代表之意義。（6分）

(三)為了去蕪存菁，研究人員去除兩個非常不顯著的變數並得到下表模型LM2之結果。根據LM1及LM2，請就下面1.或2.擇一回答（兩項均答者不予評分）。（10分）

1. 說明LM1與LM2何者較佳或差不多，並建議大眾那些變數為三高影響因子應儘量避免或注意？
2. 此分析結果不適合用來推薦三高影響因子（說明原因及提出改進方法，此結論是否與題(一)結論矛盾？）。

模型LM1	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	97.487	0.627	155.365	0.0000
性別	19.564	0.113	173.786	0.0000
年齡	0.452	0.005	86.894	0.0000
身體質量指數BMI	1.249	0.458	2.729	0.0064
量血壓習慣	2.070	0.108	19.084	0.0000
量血糖習慣	0.557	0.100	5.545	0.0000
量血脂習慣	3.012	0.311	9.697	0.0000
喝酒習慣	-0.741	0.294	-2.522	0.0117
抽煙習慣	0.046	0.049	0.936	0.3494
外食頻率	-1.827	0.979	-1.866	0.0621
運動習慣	2.933	0.858	3.418	0.0006
睡眠品質	-0.005	0.019	-0.284	0.7764
Residual standard error: 4.923 on 9988 degrees of freedom Multiple R-squared: 0.8453, Adjusted R-squared: 0.8451 F-statistic: 4961 on 11 and 9988 DF, p-value: < 2.2e-16				

模型LM2	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	97.551	0.624	156.414	0.0000
性別	19.570	0.111	176.780	0.0000
年齡	0.452	0.005	86.912	0.0000
身體質量指數BMI	1.247	0.457	2.726	0.0064
量血壓習慣	2.070	0.108	19.081	0.0000
量血糖習慣	0.556	0.100	5.532	0.0000
量血脂習慣	3.013	0.311	9.702	0.0000
喝酒習慣	-0.746	0.294	-2.539	0.0111
外食頻率	-1.836	0.979	-1.876	0.0607
運動習慣	2.934	0.858	3.420	0.0006

```
Residual standard error: 4,923 on 9990 degrees of freedom
Multiple R-squared: 0.8453, Adjusted R-squared: 0,8451
F-statistic: 6064 on 9 and 9990 DF, p-value: < 2.2e-16
```

答：

(一)(1) $R^2=0.8453$ 表示 LM1 之所有自變數對迴歸模型的解釋力達 84.53 %，此解釋力>>0.49，算是很高。

(2) $F=4961\gg F_\alpha(11\ ,9988)\ \Leftrightarrow\ p-value<2.2e-6\ll\alpha$

表示：LM1 迴歸模型非常顯著

(二)(1)性別參數估計值意義：男性比女性平均血壓高出 19.564mmHg

(2)年齡參數估計值意義：年齡每增加一歲平均血壓增加 0.452mmHg

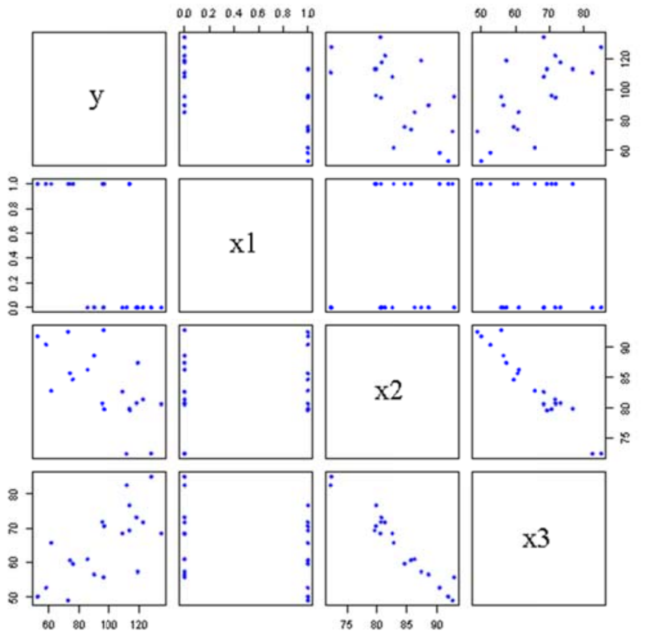
(3)BMI 參數估計值意義：BMI 每增加一單位平均血壓增加 1.249mmHg

(三) 1. (1)兩個模型之調整判定係數與 p-value 均相同，但因模型 LM2 自變數較少可以降低共線性，所以 LM2 較佳。

(2)BMI、量血壓習慣、量血糖習慣、量血脂習慣、喝酒習慣、外食頻率、運動習慣：$p-value<\alpha=0.1$，∴上述變數均為影響三高的重要因子，應盡量避免或注意

【備註】本題只能擇一回答，兩項均答不予評分，因此僅解 1.

四、一個學習效果評量相關分析的報告裏，資料內容由20人（男女各半）的4個變數(y, x1, x2, x3)所構成。其中y為學習效果（其平均值96.2且標準差為24.47），x1=1或0表男性及女性，x2（其平均值83.6且標準差為5.9）與x3（其平均值65且標準差為10.3）分別表某性向測驗的兩種分數。下圖為資料之4個變數間的散佈圖；此外，下表也列出配適學習效果y與不同解釋變數之迴歸模型的 R^2。



Model	Variables in model	R^2
M1	x1	0.397
M2	x2	0.413
M3	x3	0.487
M4	x2, x3	0.504
M5	x1, x2	0.676
M6	x1, x3	0.697
M7	x1,x2, x3	0.697

(一)考慮模型M1，完成下面的分析表，說明填入之F value及t value的值所代表意義。（12分）

Analysis of Variance Table: Response: y

	Df	Sum Sq	Mean Sq	F value
x1				
Residuals				
Total				

Coefficients:

	Estimate	Std. Error	t value
Intercept			
x1			

(二)考慮模型M1，計算y在x1=1之信心水準為90%的預測區間。(5分)

(三)在M1-M7模式中，給定進入模式水準(entry level) α=0.1，採用F檢定法，列出前進選取(forward selection)程序與其最終選定之模式。(10分)

(四)根據準則Akaike Information Criterion(AIC)，依序列出M1-M7模式中的最佳3個模型。(10分)

(五)針對M7模式，在顯著水準α=0.1下，檢定x2與x3之係數是否同時等於0。(5分)

答：

(一)$SSTO = S_{YY} = (n-1)S_Y^2 = (20-1)\times 24.47^2 = 11376.8371$

$$S_{X_1X_1} = \sum {X_1}^2 - \frac{(\sum X_1)^2}{n} = 10 - \frac{10^2}{20} = 5$$

$$R^2 = \frac{SSR}{SSTO} \Rightarrow 0.397 = \frac{SSR}{11376.8371} \Rightarrow SSR = 4516.6043$$

$$SSR = \hat{\beta}^2 \times S_{X_1X_1} \Rightarrow \hat{\beta} = \pm\sqrt{\frac{4516.6043}{5}}(\text{取}-) = 30.0553$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{X}_1 = 96.2 - (-30.0553)\times\frac{10}{20} = 111.2277$$

$$S_{\hat{\beta}} = \sqrt{\frac{MSE}{S_{X_1X_1}}} = \sqrt{\frac{381.1240}{5}} = 8.7307$$

$$S_{\hat{\alpha}} = \sqrt{\left(\frac{1}{n} + \frac{\bar{X}_1^2}{S_{X_1X_1}}\right)MSE} = \sqrt{\left(\frac{1}{20} + \frac{0.5^2}{5}\right)\cdot 381.1240} = 6.1735$$

(1)

	Df	Sum Sq	Mean Sq	F value
X_1	1	4516.6043	4516.6043	11.85
Residuals	18	6860.2328	381.1240	
Total	19	11376.8371		

	Estimate	Std.Error	t value
Intercept	111.2277	6.1735	18.02
X_1	−30.0553	8.7307	−3.4425

(2) $F = 11.85 > F_{0.1}(1,18) = 3.007$：表示迴歸模型顯著

兩個參數 t 值皆顯著：表示兩個參數皆顯著異於 0 。

(二) $\hat{Y}_{X_1=1} = 111.2277 - 30.0553 \times 1 = 81.1724$

$$Y_{X_1=1}\text{之 } 90\%\ P.I. = \hat{Y}_{X_1=1} \pm t_{\frac{0.1}{2}}(20-2)\sqrt{[1+\frac{1}{n}+\frac{(x_i-\bar{X}_1)^2}{S_{X_1X_1}}]MSE}$$

$$= 81.1724 \pm \sqrt{3.007}\sqrt{[1+\frac{1}{20}+\frac{(1-0.5)^2}{5}]\times 381.1240}$$

$= 81.1724 \pm 35.5055 = (45.6669,\ 116.6779)$

(三) **Step1**： X_3 的 $R^2 = 0.487$ (max)， 且 $F = \frac{(n-2)R^2}{1-R^2} = 17.09 > 3.007$

∴ 選進 X_3 (目前 model 含有自變數：X_3)

Step2：

(1) $SSR(X_3) = 11376.8371 \times 0.487 = 5540.5197$

$SSR(X_2X_3) = 11376.8371 \times 0.504 = 5733.9259$

$SSE(X_2X_3) = 11376.8371 - 5733.9259 = 5642.9112$

$MSR(X_2|X_3) = 5733.9259 - 5540.5197 = 193.4062$

$$F = \frac{MSR(X_2|X_3)}{MSE(X_2X_3)} = \frac{193.4062}{5642.9112/(20-3)} = 0.5826$$

$$< F_{0.1}(1,17) = 3.026$$

(2) $SSR(X_1X_3) = 11376.8371 \times 0.697 = 7929.6554$

$SSE(X_1X_3) = 11376.8371 - 7929.6554 = 3447.1816$

$MSR(X_1|X_3) = 7929.6554 - 5540.5197 = 2389.1357$

$$F = \frac{MSR(X_1|X_3)}{MSE(X_1X_3)} = \frac{2389.1357}{3447.1816/(20-3)} = 11.78$$

$$> F_{0.1}(1,17) = 3.026$$

∴ 再選進 X_1 (目前 model 含有自變數：X_1X_3)

step 3：

$$SSR(X_1X_2X_3) = 11376.8371 \times 0.697 = 7929.6554$$

$MSR(X_1|X_3) = 0$

$$F = \frac{MSR(X_2|X_1X_3)}{MSE(X_1X_2X_3)} = 0 < F_{0.1}(1,18) = 3.007$$

∴ 不再選進 X_2

∴ $X_1\ X_3$ 為最佳自變數組合， 模型 $Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \varepsilon$

(四) $SSE(X_2) = 11376.8371(1-0.413) = 6678.2034$

$SSE(X_3) = 11376.8371(1-0.487) = 5836.3147$

$SSE(X_1X_2) = 11376.8371(1-0.676) = 3686.0952$

$SSE(X_1X_2X_3) = 11376.8371 \times (1-0.697) = 3447.1816$

$AIC_P = n \cdot \ln SSE_P - n \cdot \ln n + 2 \cdot P$

M1：$AIC_1 = 20 \cdot \ln 6860.2328 - 20 \cdot \ln 20 + 2 \cdot 2 = 120.7552$

M2：$AIC_2 = 20 \cdot \ln 6678.2034 - 20 \cdot \ln 20 + 2 \cdot 2 = 120.2174$

M3：$AIC_2 = 20 \cdot \ln 5836.3147 - 20 \cdot \ln 20 + 2 \cdot 2 = 117.5225$

M4：$AIC_3 = 20 \cdot \ln 5642.9112 - 20 \cdot \ln 20 + 2 \cdot 3 = 118.8485$

M5：$AIC_3 = 20 \cdot \ln 3686.0952 - 20 \cdot \ln 20 + 2 \cdot 3 = 110.3318$

M6：$AIC_3 = 20 \cdot \ln 3447.1816 - 20 \cdot \ln 20 + 2 \cdot 3 = 108.9916$

M7：$AIC_4 = 20 \cdot \ln 3447.1816 - 20 \cdot \ln 20 + 2 \cdot 4 = 110.9916$

∴ M6 最佳，M5 次之，M7 再次之

(五) ① $\begin{cases} H_0：\beta_2 = \beta_3 = 0 \\ H_1：\beta_2, \beta_3 \text{不同時為 } 0 \end{cases}$

② $C = \{ F \mid F > F_{0.1}(2, 20-4) = 2.668 \}$

③ $F = \frac{[SSE(X_1)-SSE(X_1X_2X_3)]/2}{SSE(X_1X_2X_3)/(n-3)} = \frac{[6860.2328-3447.1816]/2}{3447.1816/(20-4)} = 7.92 \in C$

$\therefore$ t reject H_0 ，有充分證據顯示：β_2, β_3不同時為 0