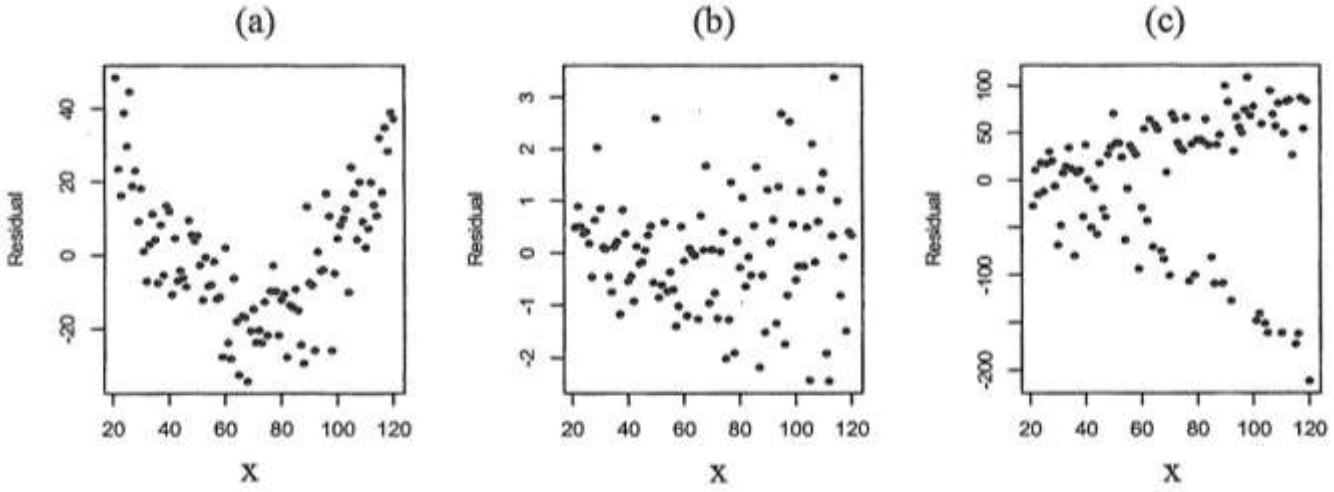
104 年公務人員高等考試三級考試試題

類　科：統計
科　目：迴歸分析

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	2.026	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	2.38
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 $y=\alpha\quad\beta x\quad$，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。



【擬答】：

(a)原始資料可能有二次或更高次的相關，違背線性的前提假設。
OLSE之點估計是否有偏誤不一定，但因以線性估計法，會產生較大的變異，所以不為有效的估計值，對於假設檢定的結果會不正確。
可利用二次式之迴歸模式來配適，並重新檢驗迴歸模型是否符合假設。

(b)殘差散布隨著X變大而變大，此時為變異數不同質，違反變異數同質的假設。OLSE之點估計仍不偏，但因為變異數不同質，所以不為有效的估計值，對於假設檢定的結果會不正確。可利用加權最小平方法校正。

(c)從殘差的散布情況，可以發現除了隨著X變大而變大以外，同時伴隨著影響點的出現(下方的殘差離中心較遠)。所以OLSE之點估計可能會產生偏誤，且因為變異數不同質，所以亦不為有效的估計值，對於假設檢定的結果會不正確。可利用加權最小平方法校正變異數異質，並且透過檢查原始資料、刪除離群值、或藉由穩健迴歸等方式來防護影響點的過度影響。

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

㈠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＝β_0X0　β_1X1　β_2X2　ε 中，如何將β_0X0　β_1X1　β_2X2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？

㈡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）。

㈢今$\hat{Y}$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。

㈣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)。

【擬答】：

㈠此為兩自變數之線性迴歸模型

$$\text{假設 } Y=\begin{bmatrix}1\\0\\1\\1\\0\\0\end{bmatrix}\quad X=\begin{bmatrix}1&1&0\\1&-2&1\\1&1&2\\1&0&2\\1&0&1\\1&0&0\end{bmatrix}\quad \beta=\begin{bmatrix}\beta_{0.}\\\beta_1\\\beta_2\end{bmatrix}$$

所以迴歸模型M：$Y=\beta_0X_0+\beta_1X_1+\beta_2X_2+\varepsilon$ 可表示為

$$Y=X\beta+\varepsilon=\begin{bmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{bmatrix}_{n\times 1}\sim MN\left[\mu=\begin{bmatrix}0\\0\\\vdots\\0\end{bmatrix}_{n\times 1},\sum=I_n\sigma^2=\begin{bmatrix}\sigma^2&&\\&\sigma^2&\\&&\sigma^2\end{bmatrix}_{n\times n}\right]$$

誤差項ε服從多元常態分配

㈡ $X'X=\begin{bmatrix}6&0&6\\0&6&0\\6&0&10\end{bmatrix}\quad (X'X)^{-1}=\frac{1}{144}\begin{bmatrix}60&0&-36\\0&24&0\\-36&0&36\end{bmatrix}\quad (X'Y)=\begin{bmatrix}3\\2\\4\end{bmatrix}$

$$\text{由 }\hat{\beta}=(X'X)^{-1}(X'Y)=\frac{1}{144}\begin{bmatrix}60&0&-36\\0&24&0\\-36&0&36\end{bmatrix}\begin{bmatrix}3\\2\\4\end{bmatrix}=\begin{bmatrix}0.1\\0.3333\\0.4\end{bmatrix}$$

$$SSE=Y'Y-\hat{\beta}'(X'Y)$$

$$=3-\begin{bmatrix}0.1&0.3333&0.4\end{bmatrix}\begin{bmatrix}3\\2\\4\end{bmatrix}=0.4334$$

$$\hat{\sigma}^2=MSE=\frac{SSE}{n-3}=\frac{0.4334}{3}=0.1445$$

$$Cov(\hat{\beta})=\hat{\sigma}^2(X'X)^{-1}=0.1445\times\frac{1}{144}\begin{bmatrix}60&0&-36\\0&24&0\\-36&0&36\end{bmatrix}$$

$$=\begin{bmatrix}0.0602&0&-0.0361\\0&0.0241&0\\-0.0361&0&0.0361\end{bmatrix}$$

㈢$\hat{Y}=X\hat{\beta}=X(X'X)^{-1}(X'Y)=HY$

所以$H=X(X'X)^{-1}X'$

$$=\frac{1}{144}\begin{bmatrix}1&1&0\\1&-2&1\\1&1&2\\1&0&2\\1&0&1\\1&0&0\end{bmatrix}\begin{bmatrix}60&0&-36\\0&24&0\\-36&0&36\end{bmatrix}\begin{bmatrix}1&1&1&1&1&1\\1&-2&1&0&0&0\\0&1&2&2&1&0\end{bmatrix}$$

$$=\frac{1}{144}\begin{bmatrix}60&24&-36\\24&-48&0\\-12&24&36\\-12&0&36\\24&0&0\\60&0&-36\end{bmatrix}\begin{bmatrix}1&1&1&1&1&1\\1&-2&1&0&0&0\\0&1&2&2&1&0\end{bmatrix}$$

$$=\frac{1}{144}\begin{bmatrix}84&-24&12&-12&24&60\\-24&120&-24&24&24&24\\12&-24&84&60&24&-12\\-12&24&60&60&24&-12\\24&24&24&24&24&24\\60&24&-12&-12&24&60\end{bmatrix}$$

㈣計算$VIF(X_1)$，可考慮為X_1應變數，為X_2自變數作迴歸

即$X_1=\alpha_0+\alpha_1X_2$，可得$R^2=\frac{SS^2_{X_1X_2}}{SS_{X_1}SS_{X_2}}=0$，$VIF(X_1)=\frac{1}{1-R^2}=1$

同理計算$VIF(X_2)$，可考慮為X_2應變數，為X_1自變數作迴歸

即$X_2=\alpha_0+\alpha_1X_1$，可得$R^2=\frac{SS^2_{X_1X_2}}{SS_{X_1}SS_{X_2}}=0$，$VIF(X_2)=\frac{1}{1-R^2}=1$

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（有量血糖習慣為 1，反之為 0），量血脂習慣（有量血脂習慣為 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。

㈠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。

㈡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。

㈢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。

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㈠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				

【擬答】：

㈠在LM1中，調整後判定係數為0.8451，代表以此11個自變數來解釋血壓，有84.51%的解釋力，屬高解釋力。

F-statistic：4961 on 11 and 9988 DF, p-value<2.2e-16

代表F統計量為4961，在自由度為11與9988時

相對的p值小於2.2×10^{-16}，代表p值非常小，模型顯著成立。

㈡在LM1中，男性相對於女性的收縮壓，高19.564 mmHg

年齡每增加一歲，收縮壓會高0.452 mmHg

每增加一單位BMI，收縮壓會高1.249 mmHg

㈢兩個模型調整後判定係數相同，可知兩模型沒有差異

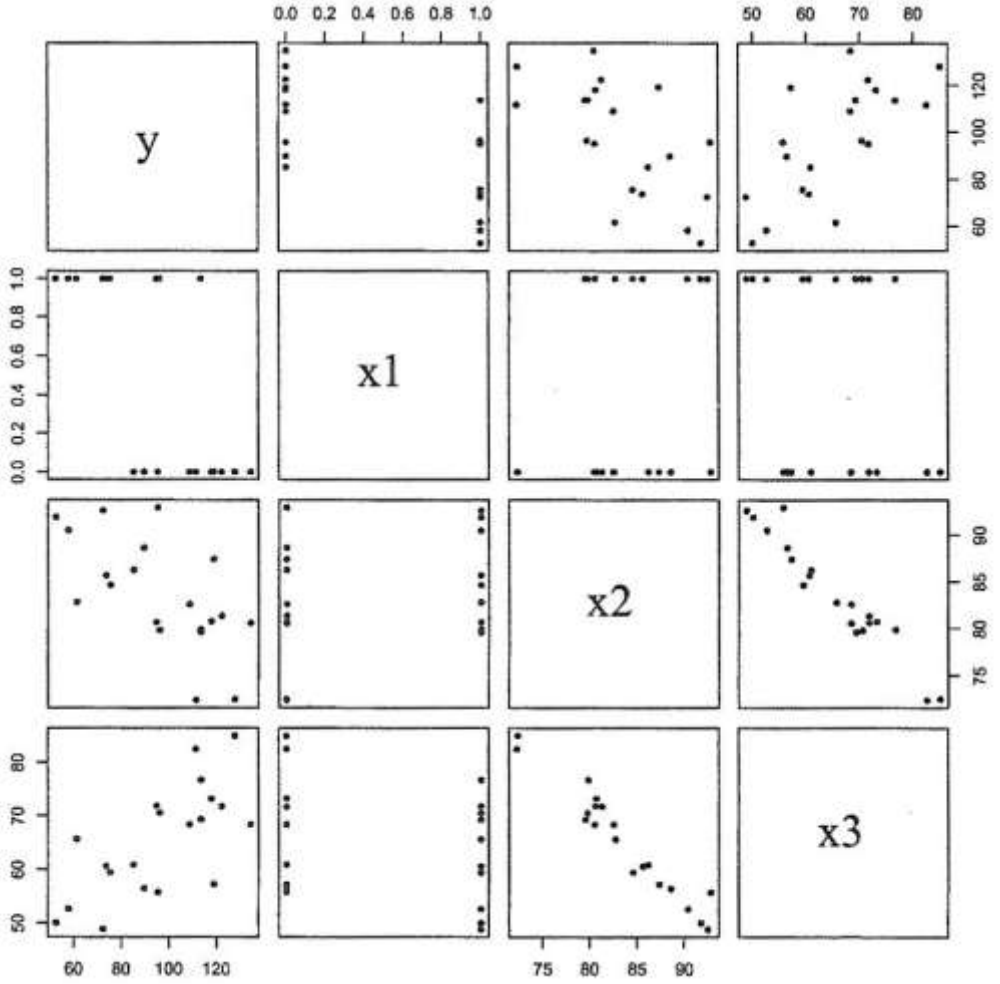
但選擇LM2可少估計2個自變數，所以應考慮LM2即可

再者，以LM2中可以發現，除了外食頻率不顯著外，其餘變項皆顯著，但這些變項的單位不同，所以僅考慮迴歸係數大小，並不足以找出對收縮壓影響最大的因素，需有標準化迴歸係數方可得知。

又此模型的迴歸係數並不合理，舉例來說，有固定量血壓、量血糖、量血脂習慣的民眾，理應較重視自身健康，收縮壓應較低，不過迴歸係數反倒為正；並且有喝酒習慣的迴歸係數為負，代表有喝酒相對沒喝酒反倒有較低的收縮壓，與常理不合。可見此模型可能出現共線性問題，因為有量血壓習慣者，很容易有量血糖、量血脂的習慣，又或者有喝酒習慣者，外食頻率可能相對較高，這些自變數間易產生高度相關，產生共線性問題，導致迴歸分析的結果不可信。

雖然模型估計結果不可信，但解釋力高卻是事實，代表這些變項確實可以有效解釋收縮壓，兩者並不矛盾。

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

㈠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。

Analysis of Variance Table：Response：y

	Df	Sum Sq	Mean Sq	F value
x1				
Residuals				
Total				

Coefficients：

	Estimate	Std. Error	t value
Intercept			
x1			

㈡考慮模型 M1，計算 y 在 x1＝1 之信心水準為 90%的預測區間。

㈢在 M1-M7 模式中，給定進入模式水準（entry level）$\alpha=0.1$，採用 F 採定法，列出前進選取（forward selection）程序與其最終選定之模式。

㈣根據準則 Akaike Information Criterion（AIC），依序列出 M1 M7 模式中的最佳3個模型。

㈤針對 M7 模式，在顯著水準 $\alpha=0.1$下，檢定 x2與 x3之係數是否同時等於0。

【擬答】：

㈠ $SSTO=SS_Y=(20-1)\times 24.47^2=11376.8371$

$$R^2=\frac{SSR(X_1)}{SSTO}\Rightarrow 0.397=\frac{SSR(X_1)}{11376.8371}\Rightarrow SSR(X_1)=4516.6043$$

因為男女各半，所以 $SS_{X_1}=20\times 0.5^2=5$

ANOVA表

	Df	Sum Sq	Mean Sq	F value
X1	1	4516.6043	4516.6043	11.851
Residuals	18	6860.2328	381.124	
Total	19	11376.8371		

利用 $\hat{\beta}_1=r\cdot\frac{S_Y}{S_X}=-\sqrt{0.397}\cdot\frac{24.47}{0.5}=-30.8361$

（因為散布圖可以看出x1越大y越小，呈現負相關）

$T_1^*=-\sqrt{F^*}=-\sqrt{11.851}=-3.443$

又 $T_1^*=\frac{\hat{\beta}_1}{S(\hat{\beta}_1)}\Rightarrow S(\hat{\beta}_1)=\frac{-30.8361}{-3.443}=8.956$

$\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X}=96.2-(-30.8361)\times 0.5=111.6181$

$$S(\hat{\beta}_0)=\sqrt{MSE\left(\frac{1}{n}+\frac{\overline{X}_1^2}{SS_{X_1}}\right)}=\sqrt{381.124\left(\frac{1}{20}+\frac{0.5^2}{5}\right)}=6.174$$

$$T_0^*=\frac{\hat{\beta}_0}{S(\hat{\beta}_0)}=\frac{111.6181}{6.1735}=18.08$$

係數估計表

	Estimate	Std. Error	T value
Intercept	111.6181	6.174	18.08
X1	-30.8361	8.956	-3.443

㈡ $Y_{h|x_1=1}$ 之90%預測區間為

$$\hat{Y}_h \pm t_{\frac{\alpha}{2}}(n-2)\sqrt{(1+\frac{1}{n}+\frac{(x_h-\overline{X}_1)^2}{SS_{X_1}})MSE}$$

$$\Rightarrow (111.6181-30.8361\times 1)\pm\sqrt{3.007}\sqrt{(1+\frac{1}{20}+\frac{(1-0.5)^2}{5})\cdot 381.124}$$

$$\Rightarrow [44.0875, 117.4765]$$

㈢ 1. 先選擇最顯著的變數加入模型（R^2 最大者）

$$F_1^* = (n-2)\frac{R_1^2}{1-R_1^2} = (20-2)\frac{0.397}{1-0.397} = 11.85 > F_{0.1}(1,18)$$

$$F_2^* = (n-2)\frac{R_2^2}{1-R_2^2} = (20-2)\frac{0.413}{1-0.413} = 12.66 > F_{0.1}(1,18)$$

$$F_3^* = (n-2)\frac{R_3^2}{1-R_3^2} = (20-2)\frac{0.487}{1-0.487} = 17.09 > F_{0.1}(1,18)$$

故選擇 x_3 加入模型中

2. 將另一變數 x_1 加入模型，並作偏F檢定

$$F^* = \frac{SSR(x_1|x_3)/1}{SSE(x_1,x_3)/n-3}$$

$$= \frac{SSR(x_1,x_3)-SSR(x_3)}{SSE(x_1,x_3)/20-3}$$

$$= 17\times\frac{R^2(x_1,x_3)-R^2(x_3)}{1-R^2(x_1,x_3)}$$ （分子分母同除SST0）

$$= 17\times\frac{0.697-0.487}{1-0.697} = 11.78 > F_{0.1}(1,17)$$

故在有 x_3 變數下，x_1 需引入模式中

再將另一變數 x_2 加入模型，並作偏F檢定

$$F^* = \frac{SSR(x_2|x_3)/1}{SSE(x_2,x_3)/n-3}$$

$$= \frac{SSR(x_2,x_3)-SSR(x_3)}{SSE(x_2,x_3)/20-3}$$

$$= 17\times\frac{R^2(x_2,x_3)-R^2(x_3)}{1-R^2(x_2,x_3)}$$ （分子分母同除SST0）

$$= 17\times\frac{0.504-0.487}{1-0.504} = 0.58 < F_{0.1}(1,17)$$

故在有 x_3 變數下，x_2 不需引入模式中

所以利用向前選取法，模型為 $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i3}$

㈣ $AIC = n\ln SSE + n\ln n + 2p$

將表格整理如下

Model	*SSR*	*SSE*	p	AIC
M1	4516.6043	6860.2328	2	240.5846

M2	4698.6337	6678.2034	2	240.0467
M3	5540.5197	5836.3174	2	237.3518
M4	5733.9259	5642.9112	3	238.6778
M5	7690.7419	3686.0952	3	230.1611
M6	7929.6555	3447.1816	3	228.8209
M7	7929.6555	3447.1816	4	230.8209

由AIC準則可知，AIC較小即較佳，所以最佳三個模型分別為M6，M5，M7

㈤ $H_0:\beta_2=\beta_3=0$　　$H_1:\beta_2,\beta_3$不全為0

$\alpha=0.01$

$$F^*=\frac{SSR(X_2,X_3\mid X_1)/2}{SSE(X_1,X_2,X_3)/20-4}$$

$$=\frac{16}{2}\frac{SSR(X_1,X_2,X_3)-SSR(X_1)}{SSE(X_1,X_2,X_3)}$$

$$=8\times\frac{\frac{SSR(X_1,X_2,X_3)}{SSTO}-\frac{SSR(X_1)}{SSTO}}{1-\frac{SSR(X_1,X_2,X_3)}{SSTO}}$$

$$=8\times\frac{0.697-0.397}{1-0.697}=7.92\in C$$

$C:\{F^*>F_{0.1}(2,16)=2.668\}$

拒絕H_0，有顯著的證據說，x_2,x_3之係數β_2,β_3不全為0