

《迴歸分析》

試題評析	本考卷之整體難度偏高，涉及範圍很廣且計算量很大，考生需要對於迴歸分析中一些深入議題非常熟悉才能夠獲得高分。
考點命中	1.《迴歸分析申論題完全制霸》，高點文化出版，趙治勳編著，頁2-8,2-11,2-88,2-108。 2.《迴歸分析申論題完全制霸》，高點文化出版，趙治勳編著，第二篇，第五章。 3.《迴歸分析申論題完全制霸》，高點文化出版，趙治勳編著，第二篇，第二章。 4.《迴歸分析申論題完全制霸》，高點文化出版，趙治勳編著，第二篇，第三章，第六節。

注意：若無特別標示，本試卷採用顯著水準為0.05及95%信心水準為原則。

附表A：t分布 $\alpha=0.025$ 與 $\alpha=0.05$ 右尾臨界值，df為自由度

df	1	2	3	4	5	6	7	8	9	10
$t_{df,0.025}$	12.706	4.303	3.182	2.776	2.571	2.447	2.365	2.306	2.262	2.228
df	11	12	13	14	15	16	17	18	19	20
$t_{df,0.025}$	2.201	2.179	2.160	2.145	2.131	2.120	2.110	2.101	2.093	2.086
df	21	22	23	24	25	26	27	28	29	30
$t_{df,0.025}$	2.080	2.074	2.069	2.064	2.060	2.056	2.052	2.048	2.045	2.042
df	1	2	3	4	5	6	7	8	9	10
$t_{df,0.05}$	6.314	2.920	2.353	2.132	2.015	1.943	1.895	1.860	1.833	1.812
df	11	12	13	14	15	16	17	18	19	20
$t_{df,0.05}$	1.796	1.782	1.771	1.761	1.753	1.746	1.740	1.734	1.729	1.725
df	21	22	23	24	25	26	27	28	29	30
$t_{df,0.05}$	1.721	1.717	1.714	1.711	1.708	1.706	1.703	1.701	1.699	1.697

附表B：F分布 $\alpha=0.05$ 右尾臨界值，df1為分子自由度，df2為分母自由度

$F_{df1,df2,0.05}$

df1/df2	1	2	3	4	5	6
1	161.45	18.51	10.13	7.71	6.61	5.99
2	199.50	19.00	9.55	6.94	5.79	5.14
3	215.71	19.16	9.28	6.59	5.41	4.76
4	224.58	19.25	9.12	6.39	5.19	4.53
5	230.16	19.30	9.01	6.26	5.05	4.39
6	233.99	19.33	8.94	6.16	4.95	4.28

一、考慮一簡單線性迴歸模型 $Y_i = \alpha + \beta X_i + \varepsilon_i$ ， $i=1, \dots, n$ ，其中 Y_i 為因變數， X_i 為自變數， ε_i 為誤差項且與 X_i 獨立。另外，也假設 ε_i ($i=1, \dots, n$) 具有獨立且相同的常態分布 $N(0, \sigma^2)$ ，其中 σ^2 表變異數。（每小題5分，共20分）

(一)請導出參數 α, β 的最小平方估計式 $\hat{\alpha}, \hat{\beta}$ ，並證明其不偏性(unbiasedness)。

(二)如果其他假設不變，但 $\text{Var}(\varepsilon_i) = \sigma^2 X_i^2$ ， $i=1, \dots, n$ 。說明由(一)導出之 $\hat{\beta}$ 是否仍具有不偏性？在此情形下，是否可提供較佳的估計式（以式子說明概念或作法，無需列出詳細結果）？

- (三) 如果其他假設不變，但 $\text{Cov}(\varepsilon_i, \varepsilon_{i+1}) = \rho\sigma^2$ ， $i=1, \dots, n-1$ 。說明由(一)導出之 $\hat{\beta}$ 是否仍具有不偏性？試舉例說明何種類型的數據會較容易發現 $\rho \neq 0$ 的情形。如何檢定 $\rho=0$ （以式子說明概念或作法，無需列出詳細結果）？
- (四) 假設自變數 X_i 無法直接被觀察到，而是觀察到一個替代變數 W_i ， $i=1, \dots, n$ ， $W_i = X_i + \delta_i$ ， δ_i 為白噪音（white noise）與其他變數均獨立，且 δ_i （ $i=1, \dots, n$ ）具有獨立且相同的常態分布 $N(0, 1)$ 。此時若將 W_i 取代最小平方估計式 $\hat{\beta}$ 中的 X_i ，並令所得之新估計式為 $\hat{\beta}_w$ 。說明此 $\hat{\beta}_w$ 是否仍具有不偏性？當 n 很大時， $\hat{\beta}_w$ 的漸近偏差為何？在此情形下是否可提供較佳的估計式（以式子說明概念或作法，無需列出詳細結果）？

答：

$$(一) \min \sum_{i=1}^n e_i^2 = \min \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \min_{\hat{\alpha}, \hat{\beta}} \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} X_i)^2$$

$$\text{令 } \frac{\partial \sum_{i=1}^n e_i^2}{\partial \hat{\alpha}} = 2 \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} X_i) (-1) = 0$$

$$\frac{\partial \sum_{i=1}^n e_i^2}{\partial \hat{\beta}} = 2 \sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta} X_i) (-X_i) = 0$$

$$\begin{cases} n\hat{\alpha} + \sum X_i \hat{\beta} = \sum Y_i \\ \sum X_i \hat{\alpha} + \sum X_i^2 \hat{\beta} = \sum X_i Y_i \end{cases}$$

$$\text{可得 } \hat{\beta} = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}} = \frac{SS_{XY}}{SS_X} \text{ 與 } \hat{\alpha} = \bar{Y} - \hat{\beta} \bar{X}$$

$$E(\hat{\beta}) = \frac{\sum (X_i - \bar{X}) E(Y_i)}{\sum (X_i - \bar{X})^2} = \frac{\sum (X_i - \bar{X}) (\alpha + \beta X_i)}{\sum (X_i - \bar{X})^2} = \beta \frac{\sum (X_i - \bar{X})^2}{\sum (X_i - \bar{X})^2} = \beta$$

$$E(\hat{\alpha}) = E(\bar{Y}) - E(\hat{\beta}) \bar{X} = (\alpha + \beta \bar{X}) - \beta \bar{X} = \alpha$$

$$(二) Y_i = \alpha + \beta X_i + \varepsilon_i, \varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2 X_i^2), i=1, 2, \dots, n$$

只要 $E(\varepsilon_i)=0$ ，由(一)所得之 $\hat{\beta}$ 仍然具有不偏性

在異質性變異數情況下，可以使用加權最小平方法以得出較佳之估計式

$$\frac{Y_i}{X_i} = \alpha \frac{1}{X_i} + \beta + \frac{\varepsilon_i}{X_i} \Rightarrow Y_i^* = \beta + \alpha X_i^* + \varepsilon_i^*$$

$$\text{其中 } V(\varepsilon_i^*) = V\left(\frac{\varepsilon_i}{X_i}\right) = \frac{V(\varepsilon_i)}{X_i^2} = \frac{\sigma^2 X_i^2}{X_i^2} = \sigma^2 \text{ (齊質性變異數)}$$

$$(三) Y_i = \alpha + \beta X_i + \varepsilon_i, \varepsilon_i \stackrel{id}{\sim} N(0, \sigma^2), i=1, 2, \dots, n$$

其中 $\text{Cov}(\varepsilon_i, \varepsilon_{i+1}) = \rho\sigma^2$

只要 $E(\varepsilon_i)=0$ ，由(一)所得之 $\hat{\beta}$ 仍然具有不偏性

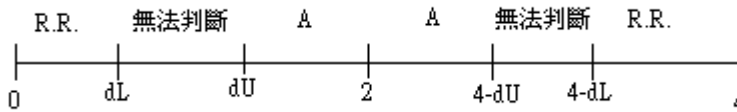
於時間序列之數據下較容易出現 $\rho \neq 0$ 之情形

可以利用Durbin-Watson Test以檢定 ρ 是否為0，檢定如下：

$$H_0: \rho = 0 \quad \text{vs} \quad H_1: \rho \neq 0$$

$$\text{T.S.: } d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

R.R. :



$$(四) Y_i = \alpha + \beta X_i + \varepsilon_i, \varepsilon_i \stackrel{id}{\sim} N(0, \sigma^2), i = 1, 2, \dots, n$$

$$W_i = X_i + \delta_i, \delta_i \stackrel{id}{\sim} N(0, 1), i = 1, 2, \dots, n$$

$$Y_i = \alpha + \beta(W_i - \delta_i) + \varepsilon_i = \alpha + \beta W_i + (\varepsilon_i - \beta \delta_i) = \alpha + \beta W_i + \varepsilon_i^*$$

$$\hat{\beta}_w = \frac{\sum W_i Y_i - \frac{\sum W_i \sum Y_i}{n}}{\sum W_i^2 - \frac{(\sum W_i)^2}{n}} = \frac{SS_{WY}}{SS_W}$$

由於 $E(\varepsilon_i^*) = E(\varepsilon_i - \beta \delta_i) = E(\varepsilon_i) - \beta E(\delta_i) = 0 - \beta(0) = 0$ ， $\hat{\beta}_w$ 仍然具有不偏性

$$\text{漸近偏差為 } \lim_{n \rightarrow \infty} E(\hat{\beta}_w - \beta) = \lim_{n \rightarrow \infty} [E(\hat{\beta}_w) - \beta] = \lim_{n \rightarrow \infty} [\beta - \beta] = \lim_{n \rightarrow \infty} 0 = 0$$

二、在一調查薪資結構的研究中，吾人欲了解薪資（Y）與以下兩變數（X1, X2）的關係，其中 X1 表性別（女性為F，男性為M），X2 表區域別（分為A, B, C三個區域），收集資料如下表：

Y	6	4	3	4	4	2
X1	F	F	F	M	M	M
X2	A	A	B	B	C	C

一般來說，統計軟體的語法建立Y與兩變數的迴歸模型分析時，模式部分可寫為 $Y \sim X1 + X2$ （R軟體）或 $Y = X1 X2$ （SAS軟體），或是直接點選 X1, X2 為自變數進行迴歸分析。請依據此精神與上述之資料，

(一) 定義一個設計矩陣（design matrix），並說明此設計矩陣各個欄（column）的意義。

寫下線性迴歸模型，以矩陣形式列出正規方程式（normal equation），解正規方程式求出參數估計值，列出三區域之兩兩比較薪資差異的估計值。（14分）

(二) 完成下面之ANOVA表。（8分）

Analysis of Variance Table: Response: Y

變異來源	自由度 (d. f.)	平方和 (SS)	均方和 (MS)	F值 F value
迴歸				
殘差				
總和		8.833		

(三)計算性別薪資差異(男性對女性)的95%信賴區間,估計一個男性在區域A的平均薪資及其95%信賴區間。最後,根據ANOVA表格中F值說明其代表之意義。(10分)

答:

$$(一) \text{令 } X_1 = \begin{cases} 1, \text{男性} \\ 0, \text{女性} \end{cases}, X_{21} = \begin{cases} 1, \text{區域A} \\ 0, \text{其他} \end{cases}, X_{22} = \begin{cases} 1, \text{區域B} \\ 0, \text{其他} \end{cases}$$

$$\text{本題之設計矩陣為 } X = [1, X_1, X_{21}, X_{22}] = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

$$\text{模型: } Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{21i} + \beta_3 X_{22i} + \varepsilon_i, i = 1, 2, \dots, 6$$

$$\text{假設: } \varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

$$\text{模型以矩陣表示為 } Y = X\beta + \varepsilon, \varepsilon \sim N(0, I_6\sigma^2)$$

$$\text{正規方程式: } X^T X \hat{\beta} = X^T Y$$

$$\text{可得 } \hat{\beta} = (X^T X)^{-1} X^T Y = \begin{bmatrix} 2 \\ 1 \\ 3 \\ 1 \end{bmatrix} \quad \text{其中 } (X^T X)^{-1} = \begin{bmatrix} 2.5 & -2 & -2.5 & -1.5 \\ -2 & 2 & 2 & 1 \\ -2.5 & 2 & 3 & 1.5 \\ -1.5 & 1 & 1.5 & 1.5 \end{bmatrix}$$

$$\text{區域 A 與區域 C 之薪資差異為 } \hat{\beta}_2 = 3$$

$$\text{區域 B 與區域 C 之薪資差異為 } \hat{\beta}_3 = 1$$

$$\text{區域 A 與區域 B 之薪資差異為 } \hat{\beta}_2 - \hat{\beta}_3 = 3 - 1 = 2$$

【版權所有，重製必究！】

(二)

ANOVA TABLE				
source	d. f.	SS	MS	F
Reg	3	4.833	1.611	$F^* = 0.8055$
Error	2	4	2	
Total	5	8.833		

其中 $SSE = Y^T Y - \hat{\beta}^T X^T X \hat{\beta} = 4$

(三) 男性對女性之薪資差異估計值為 $\hat{\beta}_1 = 1$

$\hat{\beta}_1$ 之 95% 信賴區間為 $(\hat{\beta}_1 \mp t_{0.025(2)} S(\hat{\beta}_1)) = (1 \mp 4.303 \times 2) = (-7.606, 9.606)$

令 $X_h = [1 \ 1 \ 1 \ 0]$

男性在區域A之平均薪資為 $\hat{y}|_{X_1=1, X_2=1, X_3=0} = 2 + 1 \times 1 + 3 \times 1 + 1 \times 0 = 6$ (單位)

男性在區域A平均薪資之95%信賴區間為

$(\hat{y}|_{X_1=1, X_2=1, X_3=0} \mp t_{0.025(2)} \sqrt{s^2(\hat{y}|_{X_1=1, X_2=1, X_3=0})}) = (6 \mp 4.303\sqrt{5}) = (-3.6218, 15.6218)$

其中 $s^2(\hat{y}|_{X_1=1, X_2=1, X_3=0}) = X_h^T (X^T X)^{-1} M S E X_h = 5$

(二) 中 ANOVA 代表

H_0 : 模型不適當 vs H_1 : 模型適當

R.R.: Reject H_0 at $\alpha = 0.05$ if $F^* > F_{(3,2)0.05} = 19.16$

$Q F^* = 0.8055 \quad \therefore$ don't reject H_0

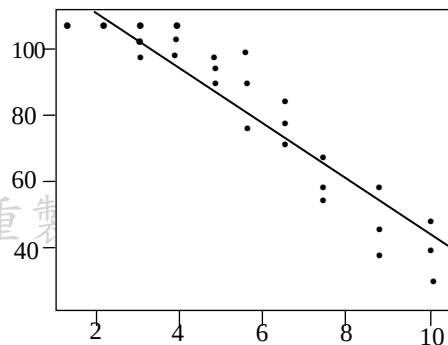
我們沒有足夠證據去推論模型是適當的

三、在一個關於放射線對腫瘤及壽命的影响研究中，研究人員利用老鼠設計了一項為期兩年的實驗。此實驗設計30隻老鼠每週照射不同劑量的放射線（劑量範圍為1~10），並記錄其壽命（單位：週）。數據形式如下表：

X (劑量)	1	1	1	2	2	2	3	3	...	...	8	8	9	9	9	10	10	10
Y (壽命)	104	104	104	104	104	98	104	94	...	...	53	56	44	36	56	37	26	46

根據資料，研究人員完成一迴歸分析及配適圖如下：

```
Call:
lm(formula = Y~X)
Residuals:
    Min       IQ   Median       3Q      Max
-16.745  -5.830  -1.500    5.113   22.028
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  123.3111   3.6872   33.44  <2e-16 ***
X             -8.0566   0.5942  -13.56  7.95e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.
Residual standard error: 9.349 on 28 degrees of freedom
Multiple R-squared:  0.8678,
```



(一) 根據分析結果，求X與Y之相關係數，完成下面之變異數分析表 (ANOVA table) 並說明此模型是否恰當？另，預測當X=15時之壽命，說明是否認同此預測值？(15分)

變異來源	自由度 (d. f.)	平方和 (SS)	均方和 (MS)	F值 F value
迴歸				
殘差				...
總和			...	...

(二)由於實驗時間的限制，事實上有8隻老鼠壽命記錄在104週時還是活著的狀態。試問若預算足夠而得以完整觀察所有老鼠的壽命時（如實驗時間3年），則迴歸分析的參數估計會如何變動（可配合圖形說明），亦即實驗數據因經費限制而對於真實之「壽命與輻射劑量關係」的分析結果可能產生怎樣的影響？（5分）

答：

$$(一)r_{XY} = (\text{sign}\hat{\beta}_1)\sqrt{R^2} = -\sqrt{0.8678} = -0.9316$$

ANOVA TABLE				
source	d. f.	SS	MS	F
Reg	1	16064.8449	16064.8449	$F^* = 183.8$
Error	28	2447.3064	87.4038	
Total	29	18512.1513		

$$\text{其中 } SST = \frac{SSE}{1-R^2} = 18512.1513$$

H_0 : 模型不適當 vs H_1 : 模型適當

R.R.: Reject H_0 at $\alpha = 0.05$ if $F^* > F_{(1,28)0.05} = t_{0.025(28)}^2 = 4.194$

$QF^* = 183.8 \quad \therefore \text{reject } H_0$

我們有足夠證據去推論模型是適當的

$$\text{當 } X=15 \text{ 時, } \hat{y} = 123.3111 - 8.0566 \times 15 = 2.4621$$

我並不認同此估計值 $\hat{y} = 2.4621$ ，因為本例中樣本資料X值介於1至10間，在利用樣本迴歸線進行預測時，當有興趣討論之自變數X=15值越遠離樣本時，這個預測值之變異程度會越大的。

(二)題意是當樣本中有8筆Y的數值變大，對於迴歸分析中參數估計之影響。由於當Y值變大會使得

$$\hat{\beta}_1 = \frac{SS_{XY}}{SS_X} = \frac{\sum (X_i - \bar{X})Y_i}{SS_X} \text{ 之分子變大, 故 } \hat{\beta}_1 \text{ 會變大。}$$

四、一組資料內含Y及 $X_1 \sim X_5$ 等變數，資料有31筆觀察值。為了進行變數選取，考慮Y對 $X_1 \sim X_5$ 之一階（first order）所有可能迴歸模式。經由分析整理得到下表：

no. of variables	X1	X2	X3	X4	X5	adjr2	Cp	no. of variables	X1	X2	X3	X4	X5	adjr2	Cp
1	0	0	0	0	1	0.142	14.5	3	1	0	1	0	1	0.371	5.4
1	0	0	1	0	0	0.142	14.5	3	1	0	0	1	1	0.361	5.8
1	0	0	0	1	0	0.14	14.6	3	0	1	1	0	1	0.294	8.8
1	1	0	0	0	0	0.014	20.8	3	0	0	1	1	1	0.277	9.6
1	0	1	0	0	0	0.008	21	3	0	1	0	1	1	0.263	10.2
2	0	0	1	0	1	0.288	8.3	3	1	1	0	0	1	0.21	12.6
2	0	0	0	1	1	0.286	8.4	3	1	1	1	0	0	0.178	14

2	1	0	0	0	1	0.189	12.9	3	1	0	1	1	0	0.169	14.5
2	1	0	1	0	0	0.185	13.1	3	1	1	0	1	0	0.156	15.1
2	1/	0	0	1	0	0.176	13.5	3	0	1	1	1	0	0.128	16.3
2	0	1	0	0	1	0.163	14.1	4	1	1	1	0	1	0.377	6.1
2	0	1	1	0	0	0.137	15.3	4	1	0	1	1	1	0.361	6.7
2	0	0	1	1	0	0.126	15.8	4	1	1	0	1	1	0.343	7.5
2	0	1	0	1	0	0.115	16.4	4	0	1	1	1	1	0.322	8.4
2	1	1	0	0	0	0.021	20.7	4	1	1	1	1	0	0.164	15.3
3	1	0	1	0	1	0.371	5.4	5	1	1	1	1	1	0.401	6

- (一)以adjusted R^2 為準則，排序選取最佳三個模式。(6分)
 (二)以Mallow's C_p 為準則，排序選取最佳三個模式。(6分)
 (三)採用F檢定法，說明向後消去法(Backward elimination, stay level = 0.05)準則的選模過程，並列出所選取之模式。(10分)
 (四)除變數選擇外，針對模型 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon$ 分析得到另一表。請以第一列的值解釋 dfb.X2(-0.154) 及 dffit(-0.371) 的用途及其大概的原理。(6分)

Obs.	dfb.X1	dfb.X2	dfb.X3	dfb.X4	dfb.X5	dfit	cov.r	cook.d	hat
1	-0.101	-0.154	-0.23	0.201	-0.132	-0.371	2.008	0.024	0.396
2	0.1	0.083	0.072	-0.081	-0.044	0.177	1.608	0.005	0.226
3	-1.145	2.676	2.773	-2.481	1.735	4.332	0.001	0.902	0.23
...									
30	0.019	0.053	0.049	-0.046	-0.016	-0.079	1.636	0.001	0.223
31	0.063	0.07	0.037	-0.048	-0.048	-0.184	1.498	0.006	0.179

答：

(一) R_{adj}^2 越大越好

$$\text{第一：} Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$\text{第二：} Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon_i$$

$$\text{第三：} Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon_i$$

(二) C_p 越接近p越好

$$\text{第一：} Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + \varepsilon_i$$

$$\text{第二：} Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon_i$$

$$\text{第三：} Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + \beta_5 X_{5i} + \varepsilon_i$$

(三)說明：先利用 R_{adj}^2 去計算每一個模型之 R^2 (公式： $R^2 = 1 - (1 - R_{adj}^2) \frac{n-k-1}{n-1}$)，再計算每一個模型

之 F 值 (公式： $F = \frac{n-k-1}{k} \frac{R^2}{1-R^2}$)，則取 F 值小於臨界值 $F_{\alpha(k, n-k-1)}$ 且 F 值最小之模型。

no. of variables(k)	R_{adj}^2	R^2	F	臨界值
1	0.142	0.1706	5.965	$F_{0.05(1,29)} = 4.183$

1	0.142	0.1706	5.965	$F_{0.05(1,29)} = 4.183$
1	0.14	0.1687	5.8851	$F_{0.05(1,29)} = 4.183$
1	0.014	0.0776	2.4397	$F_{0.05(1,29)} = 4.183$
1	0.008	0.041	1.2398	$F_{0.05(1,29)} = 4.183$

模型(0,1,0,0,0)為 F 值小於 $F_{0.05(1,29)}$ 且最小，故不考慮 X_2

由不考慮 X_2 之兩個自變數的模型中整理如下：

no. of variables(k)	R_{adj}^2	R^2	F	臨界值
2	0.288	0.3355	7.0685	$F_{0.05(2,28)} = 3.340$
2	0.286	0.3336	7.0084	$F_{0.05(2,28)} = 3.340$
2	0.189	0.2431	4.4965	$F_{0.05(2,28)} = 3.340$
2	0.185	0.2393	4.4041	$F_{0.05(2,28)} = 3.340$
2	0.176	0.2309	4.2031	$F_{0.05(2,28)} = 3.340$
2	0.126	0.1843	3.1632	$F_{0.05(2,28)} = 3.340$

僅模型(0,0,1,1,0)之 F 值小於 $F_{0.05(2,28)}$ ，故不考慮 X_3, X_4

剩下的自變數 X_1, X_5 建立模型(1,0,0,0,1)之 $R_{adj}^2 = 0.189$ 且 F 值大於 $F_{0.05(2,28)}$ ，故選取之模型為

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_5 X_{5i} + \varepsilon_i$$

(四)

$dfb.Xi$ 與 $dfit$ 分別為影響值檢測方法DFBETAS與DFFITS之統計量，其原理如下：

DFFITS：

在單一應變數配適值下考慮某觀察值影響程度之指標

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

判斷準則：

以 $|(DFFITS)_i| > 2\sqrt{\frac{p}{n}}$ 視為影響程度顯著

DFBETAS：

在迴歸係數下考慮某觀察值影響程度之指標

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)} c_{kk}}}$$

判斷準則：

以 $|(DFBETAS)_i| > \frac{2}{\sqrt{n}}$ 視為影響程度顯著