探討抽菸量與癌症之間關聯性之回歸分析

作者：林橙莉、詹雅竹、林美惠、李玲慧、劉威麟、鐘愉翔

系級：應統所博一、統精所碩二

學號：P9522017、M9416494、M9416481、M9431905、M9416505、M9485005

開課老師：林文欽

課程名稱：回歸分析

開課系所：統精碩一

開課學年：95　學年度 第 一 學期

中文摘要

本文所分析之資料為 1960 年蒐集美國 43 州與哥倫比亞特區已銷售之菸頭數與每十萬人當中不同癌症各自之死亡率，其中癌症包含了膀胱癌、肺癌、腎臟癌與白血症，利用回歸分析探討癌症死亡率與銷售菸頭數之間的關係。主要目的是探討各種癌症之死亡率對菸頭銷售數的影響。開始我們由簡單回歸分析開始探討每一種癌症死亡率對菸頭銷售數之關係，利用偏回歸分析選擇出較佳之複回歸模型。接著針對選擇出之複回歸模型，我們進行完整之殘差分析與影響點分析之探討。並與逐步回歸選模分析，選出之適當模型比較，發現所選出之最佳模型是一致的。在文獻資料顯示抽菸對肺癌的形成有直接的影響，對罹患膀胱癌與腎臟癌抽菸也會造成影響。而一些白血症的形成也是由於抽菸的關係。最後，我們放入地區之虛擬變數探討不同地區之間，癌症致死亡率對地區菸頭銷售量的影響。

而一般文獻均是探討抽菸對癌症之影響，但我們此組資料是探討各種癌症之死亡率對菸頭銷售數的影響。雖然，反因為果，但統計方法之運用與解釋角度卻是正確的，堪可參考。

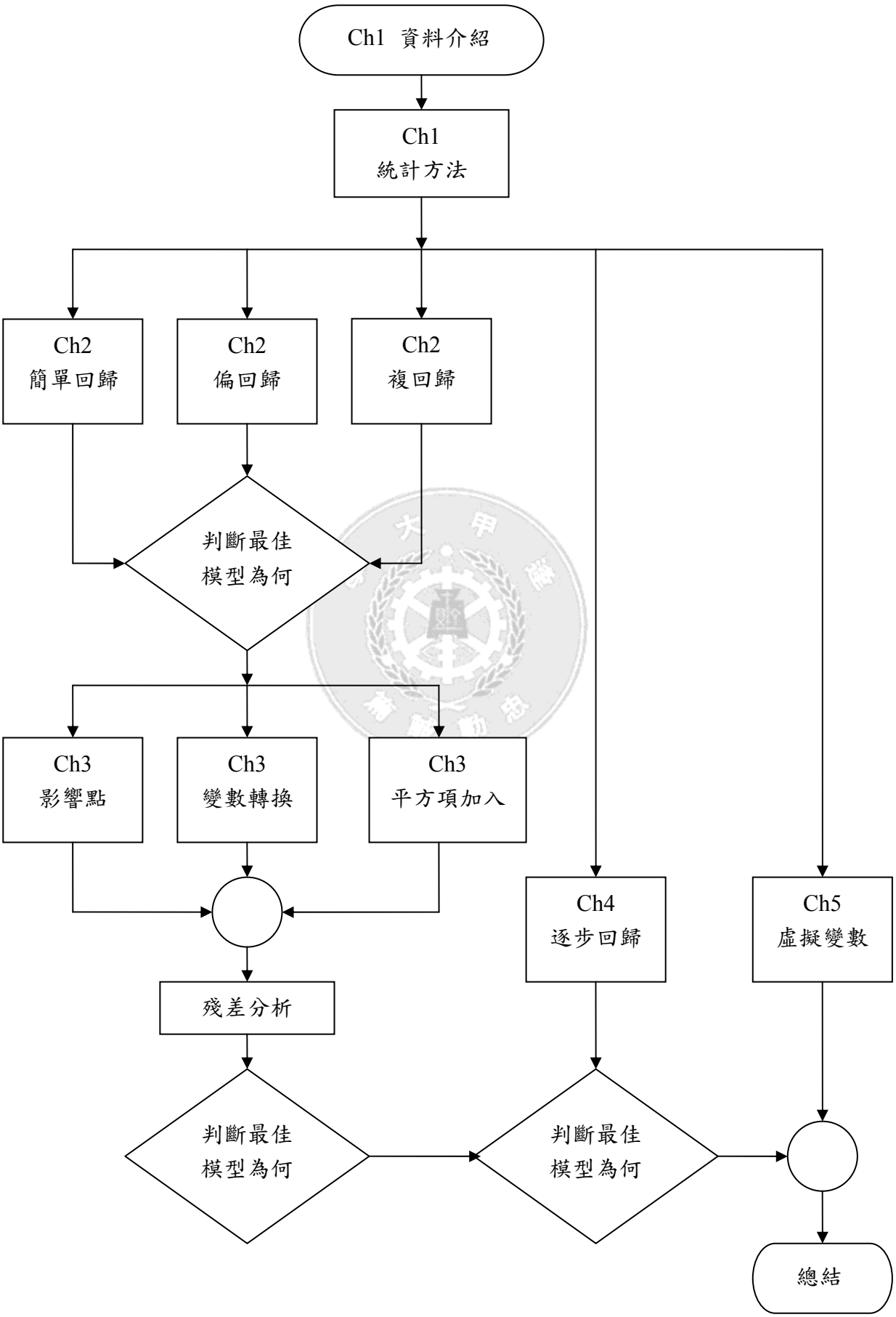
關鍵字：複回歸分析、殘差分析診斷、逐步回歸分析

Contents

Tables

Figures

研究流程圖


Ch1 資料介紹
Ch1
統計方法
Ch2
簡單回歸
Ch2
偏回歸
Ch2
複回歸
判斷最佳
模型為何
Ch3
影響點
Ch3
變數轉換
Ch3
平方項加入
Ch4
逐步回歸
Ch5
虛擬變數
殘差分析
判斷最佳
模型為何
判斷最佳
模型為何
總結

第一章 資料介紹與分析方法陳述

第一節

此組資料為 1960 年蒐集美國 43 州與哥倫比亞特區之已抽香菸頭數(賣出)與每十萬人當中不同癌症各自之死亡率，其中癌症包含了膀胱癌、肺癌、腎臟癌與白血症，分析癌症與抽菸之間的關係。而我們參考 Fraumeni, J.F. 1968 此篇論文，將地區做細分，分析抽菸與癌症是否會因地區的不同受環境影響而有所差異。

變數名稱介紹

y：已抽過香菸頭數(Number of cigarettes smoked)

x_1：各州每十萬人當中罹患膀胱癌死亡率

(Deaths per 100K population from bladder cancer)

x_2：各州每十萬人當中罹患肺癌死亡率(Deathes per 100K population from lung cancer)

x_3：各州每十萬人當中罹患腎臟癌死亡率(Deaths per 100K population from kidney cancer)

x_4：各州每十萬人當中罹患白血症死亡率(Deaths per 100K population from leukemia)

增加 3 個虛擬變數分別為

x_4：菸頭量＞平均數 23.77 設為 1

菸頭數＜平均數 23.77 設為 0

x_5：依地區劃分設 1 為北西部(Northwest)

2 為中西部(Midwest)

3 為南部(South)

4 為西部(West)

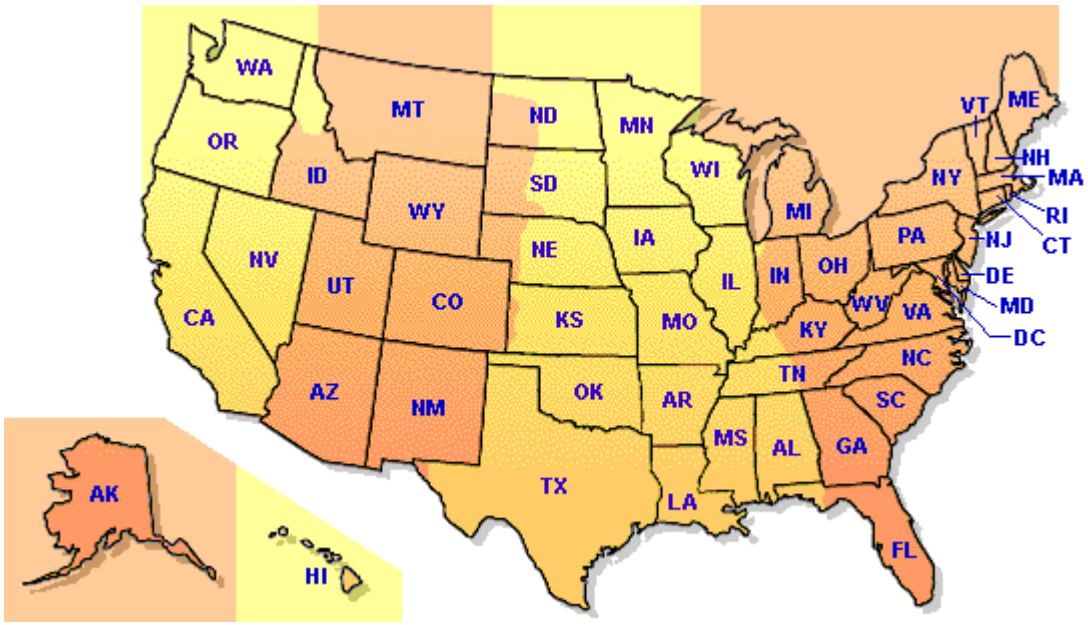
x_6：依地區劃分設 1 為南部與西部

0 為北西部與中西部

資料共 44 筆與 7 個解釋變數。資料來源：http://lib.stat.cmu.edu

Note: 在經過簡單回歸模型配適後發現白血症的死亡率(x_4)對菸頭量並無解釋能力，因此在刪除該解釋變數後加入虛擬變數由 x_4 開始。

44 州分布在美國各地的位置：



虛擬變數中各個值所包含的州

$x_5 = 1$ ⇒北西區： 康乃狄克州(CT)、緬因州(ME)、麻薩諸塞州(MA)、新澤西州 (NJ)、紐約州(NY)、賓夕法尼亞州(PA)、羅德島州(RI)、猶他州(UT)。

2 ⇒中西區： 伊利諾州(IL)、印第安納州(IN)、愛荷華州(IO)、堪薩斯州(KS)、密西根州(MI)、明尼蘇達州(MN)、密蘇里州(MO)、內布拉斯加州(NB)、北達科他州(ND)、俄亥俄州(OH)、南達科他州(SD)、威斯康辛州(WI)。

3 ⇒ 南區： 阿拉斯加州(AK)、阿拉巴馬州(AL)、阿肯色州(AR)、德拉威州(DE)、哥倫比亞特區(DC)、佛羅里達州(FL)、肯塔基州(KY)、路易斯安納州(LA)、馬里蘭州(MD)、密西西比州(MS)、奧克拉荷馬州(OK)、南卡羅來納州(SC)、田納西州(TE)、德州(TX)、西維吉尼亞州(WV)。

4 ⇒ 西區： 亞利桑那州(AZ)、加州(CA)、愛達荷州(ID)、蒙大拿州(MT)、內華達州(NE)、新墨西哥州(NM)、猶他州(UT)、華盛頓州(WA)、懷俄明州(WY)。

$x_6 = 1$ ⇒南區與西區：阿拉斯加州(AK)、阿拉巴馬州(AL)、阿肯色州(AR)、德拉威州(DE)、哥倫比亞特區(DC)、佛羅里達州(FL)、肯塔基州(KY)、路易斯安納州(LA)、馬里蘭州(MD)、密西西比州(MS)、奧克拉荷馬州(OK)、南卡羅來納州(SC)、田納西州(TE)、德州(TX)、西維吉尼亞州(WV)、亞利桑那州(AZ)、加州(CA)、愛達荷州(ID)、蒙大拿州(MT)、內華達州(NE)、新墨西哥州(NM)、猶他州(UT)、華盛頓州(WA)、懷俄明州(WY)。

0 ⇒北西區與中西區：康乃狄克州(CT)、緬因州(ME)、麻薩諸塞州(MA)、新澤西州(NJ)、紐約州(NY)、賓夕法尼亞州(PA)、羅德島州(RI)、猶他州(UT)、伊利諾州(IL)、印第安納州(IN)、愛荷華州(IO)、堪薩斯州(KS)、密西根州(MI)、明尼蘇達州(MN)、密蘇里州(MO)、內布拉斯加州(NB)、北達科他州(ND)、俄亥俄州(OH)、南達科他州(SD)、威斯康辛州(WI)。

第二節

我們將利用回歸分析，針對此組資料做探討。而回歸分析(Regression Analysis)是一種統計分析方法，它利用一組預測變數(或稱獨立變數)的數值，對某一準則變數(或稱應變數)做預測，它也可以做為評估預測變數對準則變數的影響程度。很不幸地，迴歸(Regression)的名字取得不理想，從字面上並不能表現出這種方法的重要性及其應用，取名實際上來自於 1885 年高登(Galton)所寫的論文“Regression Toward Mediocrity in Heredity Stature”。大致來說，其意義為：如果一些未知的獨立變數之影響程度消失，其應變異數應些一迴歸線。迴歸的主要目的是做預測，目標是發展一種能以一個或多個預測變數的數值來做為應變數預測的方法。迴歸分析就是找出變數間的關係式。我們將變數分成兩類，一類變數是做為預測提供者，稱為獨立變數(Independent Variable)或稱為預測變數(Predictor Variable)，以 x 表示，另一類是我們真是關心的被想預測者，稱為反應變數(Dependent Variable)或準則變數(Response Variable)，以 y 表示。

首先，我們針對每個解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與白血症(x_4)與反應變數菸頭數(y)配適簡單線性回歸模型，依變數解釋能力高低，將所有解釋變數放入模型中並選擇出最合適的模型。並對模型進行殘差診斷，包含對(1) 預測變數的殘差圖、(2) 配適值的殘差圖、(3)殘差之常態機率圖。利用 *VIF* 診斷預測變數間是否有存在多元共線性存在。若由殘差圖型發現殘差非固定數或非常態性時，我們可能考慮變數變換等矯正方法。接著使用標準化後的殘差值來判斷觀測值 y 是否存在離群值，和計算帽子矩陣槓桿值判斷觀測值 x 是否存有離群值。辨認出離群點後，緊接著探討這些離群值是否具影響力。且我們亦會介紹逐步迴歸分析，藉由向前選擇法(Forward selection)、向後消去法(Backward elimination)、逐步選擇法

(Stepwise regression)，選擇出對回歸模型具有較佳解釋能力之解釋變數之組合。最後，考慮在模型中加入了虛擬變數$x_4=1$表菸頭量＞平均數 23.77，$x_4=0$表菸頭量＜平均數 23.77；依地區(AREA)劃分設$x_5=1$為北西部、$x_5=2$中西部、$x_5=3$南部與$x_5=4$西部；依地區(WEST)劃分$x_6=1$設南部與西部$x_6=0$為北西部與中西部，探討因虛擬變數設定而產生之組別間的差異。

第二章 簡單線性回歸分析與複回歸

前言

本章首先針對每個解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與白血症(x_4)與反應變數菸頭數(y)配適簡單線性回歸模型。接著考慮放入兩個解釋變數配進入模型中配適複回歸模型，而利用偏回歸圖觀察若已有一個解釋變數在模型中，加入另一解釋變數進入模型是否對模型有幫助，並利用 Contour 圖形觀察解釋變數之間是否有交互作用存在。依序放入三個解釋變數與四個解釋變數分別配適模型，最後依整體現象選擇出最合適的模型。

第一節

此節為針對每個解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與白血症(x_4)與反應變數菸頭數(y)配適簡單線性回歸模型。

由 Table 2.1.1－2.1.8 看出 x_1、x_2、x_3 的 p-value<0.01，參數檢定皆顯著，表示具解釋能力。而 x_4 的 p-value = 0.6587>0.01，參數不顯著。而相對於其他變數而言，x_1 的 $R^2 = 0.4951$ 為最大，$\sqrt{MSE} = 4.0071$ 為最小，因此就簡單回歸而言解釋變數 x_1 對 y 的解釋能力最高。另外，由 Figure 2.1.4 發現其散佈圖點的散佈情況完全沒有呈現直線的樣子，由 x_4 的 $R^2 = 0.0047$ 解釋 y 總變異能力僅有 0.47%，且 $\sqrt{MSE} = 5.6260$ 為最大，加上 x_4 的參數檢定並不顯著，因此 x_4 對 y 而言可能不具解釋能力。

$$\hat{y} = 8.1657 + 4.064x_1 \tag{2.1.1}$$

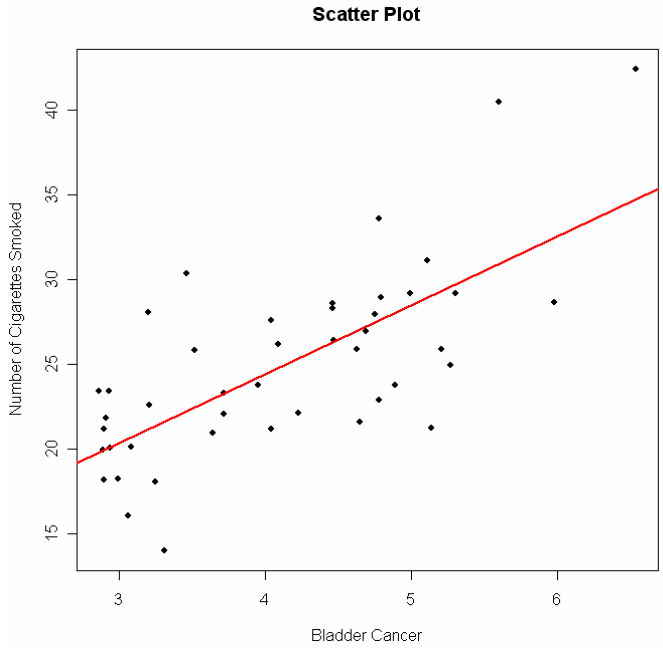


Figure 2.1.1: Scatter plot Number of cigarettes smoked on x_1(Bladder)

Table 2.1.1. Parameter estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	8.1657	2.6789	3.05	0.0040
x_1	1	4.0640	0.6333	6.42	<.0001

Table 2.1.2. Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	1	661.2562	661.2562	41.18	<.0001
Error	42	674.3890	16.0569		
Total	43	1335.6453	1335.6453		
Root MSE		4.0071	R-Square		0.4951
Dependent Mean		24.9141	Adj R-Sq		0.4831
Coeff Var		16.0837			

$$\hat{y} = 6.8473 + 0.9193x_2 \tag{2.1.2}$$

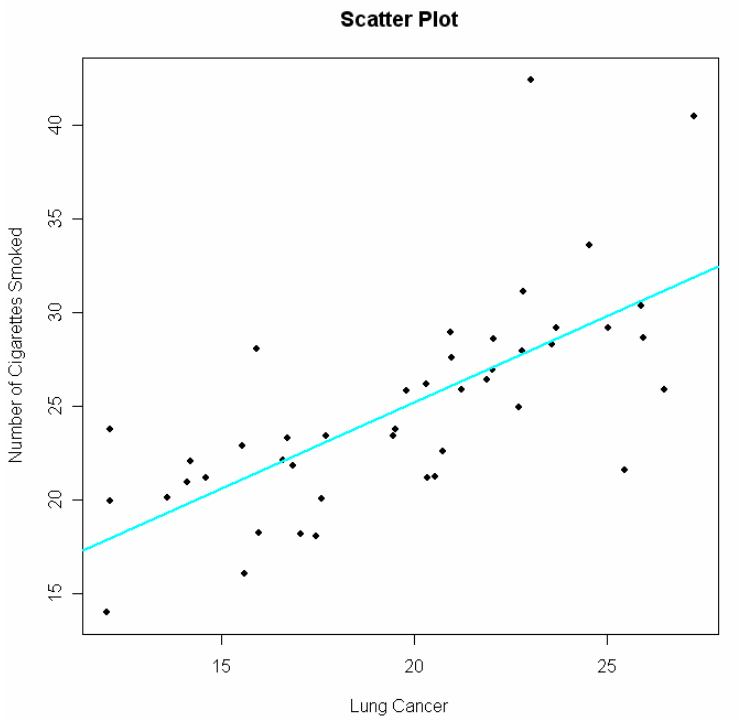


Figure 2.1.2: Scatter plot Number of cigarettes smoked on x_2 (Lung Cancer)

Table 2.1.3. Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	6.8473	2.9289	2.34	0.0242
x_2	1	0.9193	0.1458	6.31	<.0001

Table 2.1.4. Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	1	649.6181	649.6181	39.77	<.0001
Error	42	686.0271	16.3340		
Total	43	1335.6453			
Root MSE		4.0415	R-Square		0.4864
Dependent Mean		24.9141	Adj R-Sq		0.4741
Coeff Var		16.2219			

$$\hat{y} = 10.2902 + 5.233x_3 \tag{2.1.3}$$

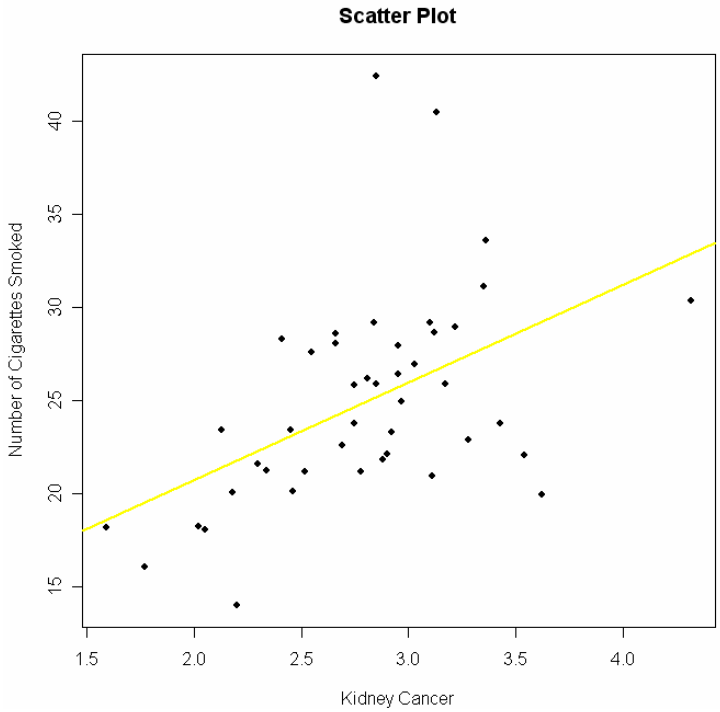


Figure 2.1.3: Scatter plot Number of cigarettes smoked on x_3 (Kidney Cancer)

Table 2.1.5. Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	10.2902	4.1103	2.50	0.0163
x_3	1	5.2330	1.4466	3.62	0.0008

Table 2.1.6. Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	1	317.2807	317.2807	13.09	0.0008
Error	42	1018.3646	24.2468		
Total	43	1335.6453			
Root MSE		4.9241	R-Square		0.2375
Dependent Mean		24.9141	Adj R-Sq		0.2194
Coeff Var		19.7643			

$$\hat{y} = 28.9982 - 0.598x_4 \qquad (2.1.4)$$

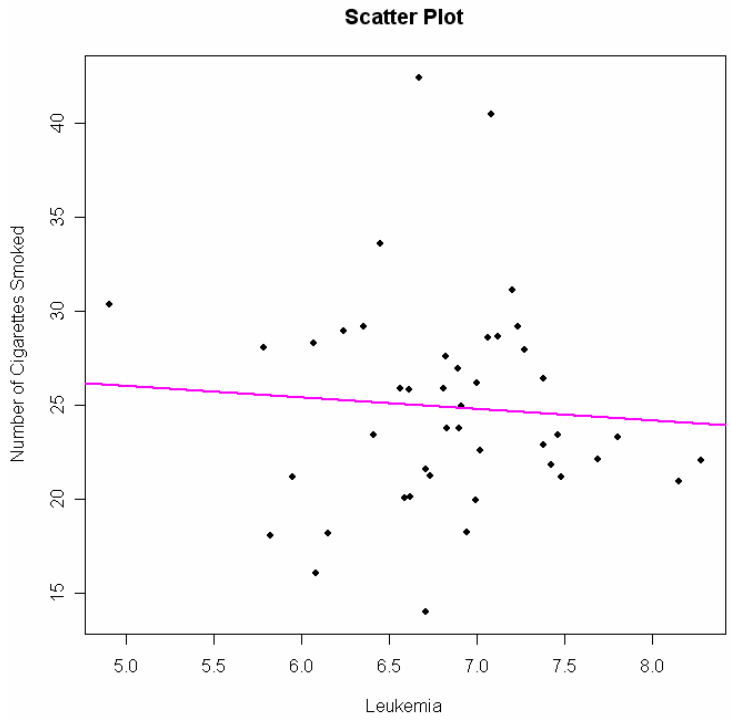


Figure 2.1.4: Scatter plot Number of cigarettes smoked on x_4 (Leukemia)

Table 2.1.7. Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	28.9982	9.2200	3.15	0.0030
x_4	1	-0.5980	1.3442	-0.44	0.6587

Table 2.1.8. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	1	6.2638	6.2638	0.2	0.6587
Error	42	1329.3815	31.6519		
Total	43	1335.6453			
Root MSE		5.6260	R-Square		0.0047
Dependent Mean		24.9141	Adj R-Sq		-0.019
Coeff Var		22.5816			

第二節

此節為針對解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與白血症(x_4)，依序選擇兩個解釋變數與反應變數菸頭數(y)配適複回歸模型。並利用偏回歸圖觀察若已有一個解釋變數在模型中，加入另一解釋變數進入模型是否對模型有幫助，並利用 Contour 圖形觀察解釋變數之間是否有交互作用存在。

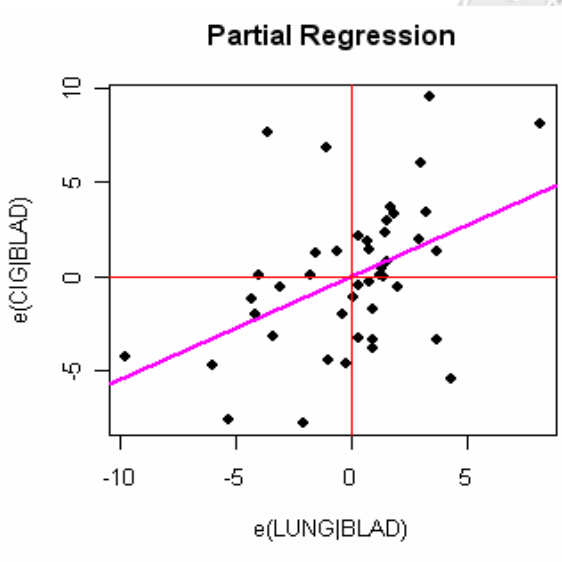


Figure 2.2.1 : Partial regression scatter plot

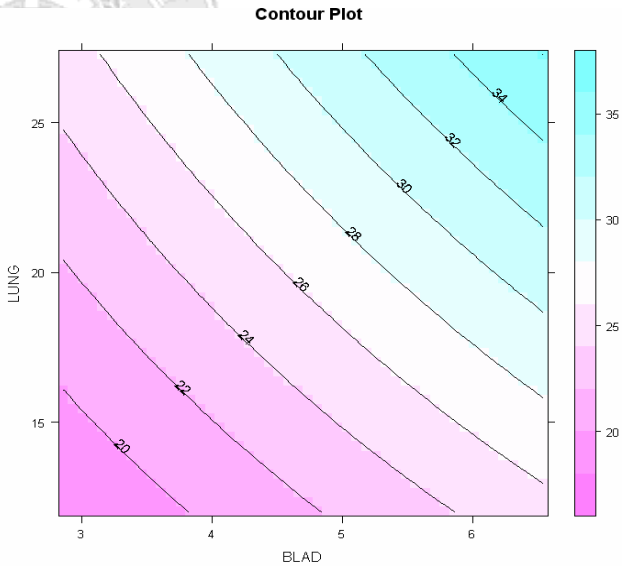


Figure 2.2.2 : ($x_1 * x_2$) contour plot

$$\hat{y} = 3.9373 + 2.4922x_1 + 0.5448x_2 \quad (2.2.1)$$

$$\hat{y} = 9.2118 + 1.1812x_1 + 0.2761x_2 + 0.0647x_1x_2 \quad (2.2.2)$$

我們從 Figure 2.2.1 的偏回歸圖可看到，在x_1已加入模型中的情況下再加入x_2後的散佈狀況大致呈一直線，表示模型在x_1解釋完後，x_2的加入對模型依舊有解釋能力。由 Figure 2.2.2

看出x_1與x_2的 contour plot 並未呈曲線狀，交互作用並不顯著；而由 Table 2.2.1－2.2.4 看出對x_1與x_2此兩變數而言，未加入交互作用項x_1*x_2前，x_1與x_2的 p-value<0.01，兩參數估計值皆為顯著，但加入交互作用項x_1*x_2後，x_1與x_2及x_1*x_2的 p-value>0.01，參數估計值皆變為不顯著；未加入交互作用項x_1*x_2前的$R^2_{adj}=0.5719$較加入後的$R^2_{adj}=0.5630$來的大，解釋力較高，而未加入交互作用項x_1*x_2前的$\sqrt{MSE}=3.6465$較加入後的$\sqrt{MSE}=3.6844$來的小；因此模型並不適合加入交互作用項x_1*x_2。

Table 2.2.1. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	3.9373	2.7898	1.41	0.1657
x_1	1	2.4922	0.7658	3.25	0.0023
x_2	1	0.5448	0.1748	3.12	0.0033

Table 2.2.2. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	790.4571	395.2285	29.72	<.0001
Error	41	545.1882	13.2973		
Total	43	1335.6453			
Root MSE		3.6465	R-Square		0.5918
Dependent Mean		24.9141	Adj R-Sq		0.5719
Coeff Var		14.6365			

Table 2.2.3. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	9.2118	13.3850	0.69	0.4953
x_1	1	1.1812	3.3430	0.35	0.7257
x_2	1	0.2761	0.6895	0.40	0.6910
x_1*x_2	1	0.0647	0.1605	0.40	0.6890

Table 2.2.4. Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	792.6628	264.2209	19.46	<.0001
Error	40	542.9825	13.5746		
Total	43	1335.6453			
Root MSE		3.6844	R-Square		0.5935
Dependent Mean		24.9141	Adj R-Sq		0.5630
Coeff Var		14.7883			

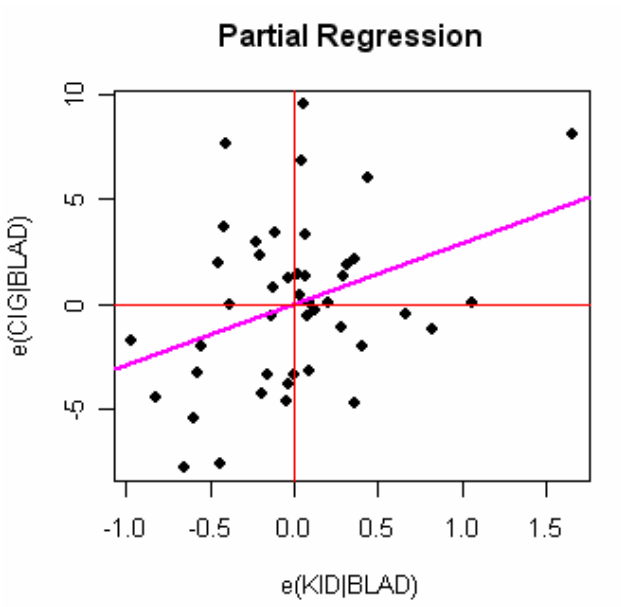


Figure 2.2.3 : Partial regression scatter plot

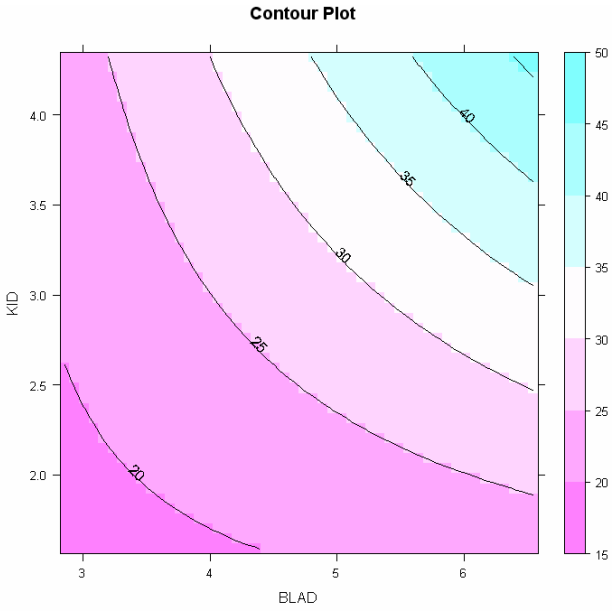


Figure 2.2.4 : ($x_1 * x_3$) contour plot

$$\hat{y} = 2.3783 + 3.5052x_1 + 2.895x_3 \quad (2.2.3)$$

$$\hat{y} = 21.0707 - 1.8944x_1 - 3.7462x_3 + 1.8918x_1x_3 \quad (2.2.4)$$

我們從 Figure 2.2.3 的偏回歸圖可看到，在 x_1 已加入模型中的情況下再加入 x_3 後的散佈狀況大致呈一直線，表示模型在 x_1 解釋完後，x_3 的加入對模型依舊有解釋能力。由 Figure 2.2.4 看出 x_1 與 x_3 的 contour plot 雖稍微呈曲線狀，但交互作用仍未達顯著的標準；由 Table 2.2.5－2.2.8 看出對 x_1 與 x_3 此兩變數而言，加入交互作用項 $x_1 * x_3$ 後的 $R^2_{adj} = 0.5402$ 雖然較加入前 $R^2_{adj} = 0.5369$ 來的大，解釋力較高，而加入交互作用項 $x_1 * x_3$ 後的 $\sqrt{MSE} = 3.7794$ 也較加入前的 $\sqrt{MSE} = 3.7928$ 來的小；但未加入交互作用項 $x_1 * x_3$ 前，x_1 與 x_3 的 p-value<0.01，兩參數估計值皆為顯著，但加入交互作用項 $x_1 * x_3$ 後，x_1 與 x_3 及 $x_1 * x_3$ 的 p-value>0.01，參數估計值皆變為

不顯著；因此模型並不適合加入交互作用項$x_1 * x_3$。

Table 2.2.5. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	2.3783	3.4820	0.68	0.4984
x_1	1	3.5052	0.6422	5.46	<0.0001
x_3	1	2.8950	1.1938	2.43	0.0198

Table 2.2.6. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	745.8597	372.9298	25.92	<.0001
Error	41	589.7856	14.3850		
Total	43	1335.6453			

Root MSE	3.7928	R-Square	0.5584
Dependent Mean	24.9141	Adj R-Sq	0.5369
Coeff Var	15.2234		

Table 2.2.7. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	21.0707	16.8103	1.25	0.2173
x_1	1	-1.8944	4.7943	-0.40	0.6948
x_3	1	-3.7462	5.9678	-0.63	0.5335
$x_1 * x_3$	1	1.8918	1.6647	1.14	0.2625

Table 2.2.8. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	764.3063	254.7688	17.84	<.0001
Error	40	571.3389	14.2835		
Total	43	1335.6453			

Root MSE	3.7794	R-Square	0.5722
Dependent Mean	24.9141	Adj R-Sq	0.5402
Coeff Var	15.1695		

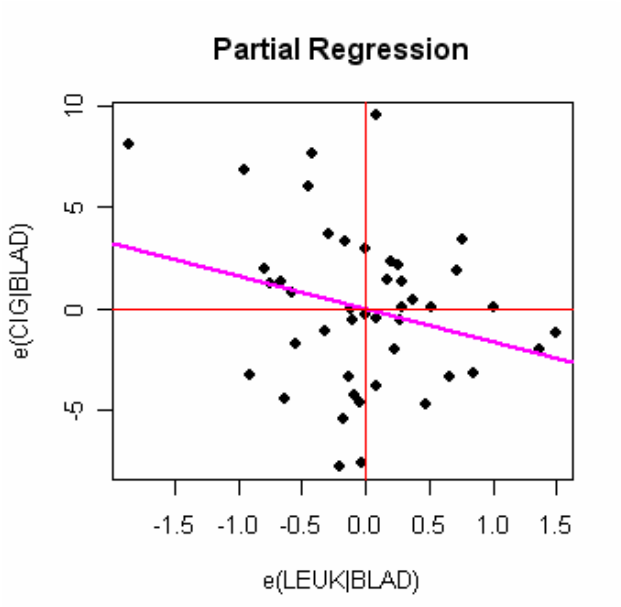


Figure 2.2.5 : Partial regression scatter plot

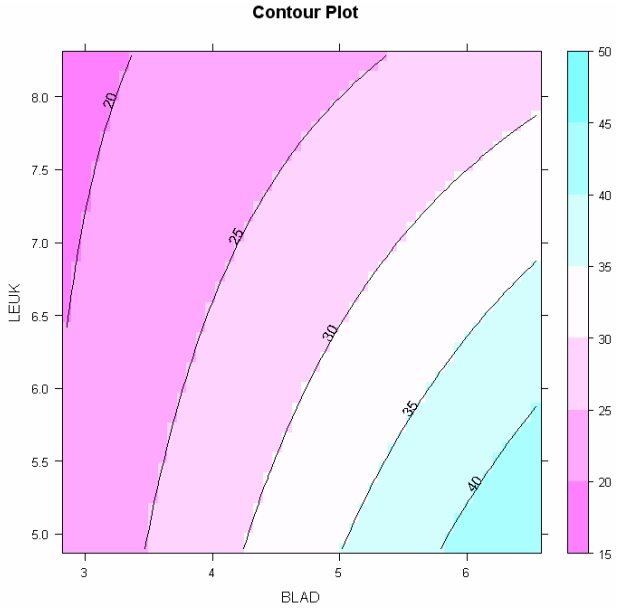


Figure 2.2.6 : ($x_1 * x_4$) contour plot

$$\hat{y} = 18.6245 + 4.2397x_1 - 4.2397x_4 \quad (2.2.5)$$

$$\hat{y} = -10.5373 + 12.2147x_1 + 2.668x_4 - 1.1722x_1x_4 \quad (2.2.6)$$

我們從 Figure 2.2.5 的偏回歸圖可看到，在 x_1 已加入模型中的情況下再加入 x_4 後的散佈狀況較似隨機分佈，並不呈一直線，表示模型在 x_1 解釋完後， x_4 的加入對模型並無解釋能力。由 Figure 2.2.6 看出 x_1 與 x_4 的 contour plot 曲線狀並不明顯，交互作用不顯著；而由 Table 2.2.9－2.2.12 看出對 x_1 與 x_4 此兩變數而言，未加入交互作用項 $x_1 * x_4$ 前， x_1 的 p-value<0.01，而 x_4 的 p-value>0.01， x_1 的參數估計值為顯著， x_4 的參數估計值不顯著，但加入交互作用項 $x_1 * x_4$ 後， x_1 與 x_4 及 $x_1 * x_4$ 的 p-value>0.01，參數估計值皆變為不顯著；未加入交互作用項 $x_1 * x_4$ 前的 $R^2_{adj} = 0.5064$ 較加入後的 $R^2_{adj} = 0.5006$ 來的大，解釋力較高，而未加入交互作用項 $x_1 * x_4$ 前的 $\sqrt{MSE} = 3.9158$ 較加入後的 $\sqrt{MSE} = 3.9387$ 來的小；因此模型並不適合加入變數 x_4 及交互作用項 $x_1 * x_4$ 。

Table 2.2.9. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	18.6245	6.5980	2.82	0.0073
x_1	1	4.2397	0.6272	6.76	<0.0001
x_4	1	-4.2397	0.9481	-1.73	0.0917

Table 2.2.10. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	706.9820	353.4910	23.05	<.0001
Error	41	628.6632	15.3333		
Total	43	1335.6453			

Root MSE	3.9158	R-Square	0.5293
Dependent Mean	24.9141	Adj R-Sq	0.5064
Coeff Var	15.7171		

Table 2.2.11. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	-10.5373	40.7896	-0.26	0.7975
x_1	1	12.2147	11.0244	1.11	0.2745
x_4	1	2.6680	6.0178	0.44	0.6599
$x_1 * x_4$	1	-1.1722	1.6178	-0.72	0.4729

Table 2.2.12. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	715.1268	238.3756	15.37	<.0001
Error	40	620.5185	15.5130		
Total	43	1335.6453			

Root MSE	3.9387	R-Square	0.5354
Dependent Mean	24.9141	Adj R-Sq	0.5006
Coeff Var	15.8089		

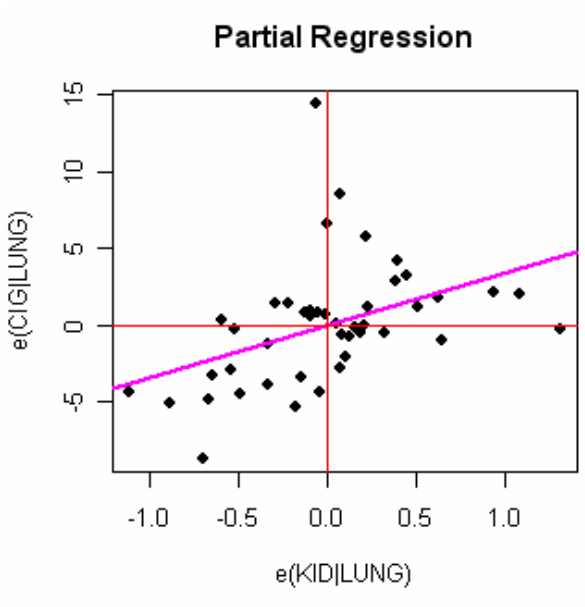


Figure 2.2.7 : Partial regression scatter plot

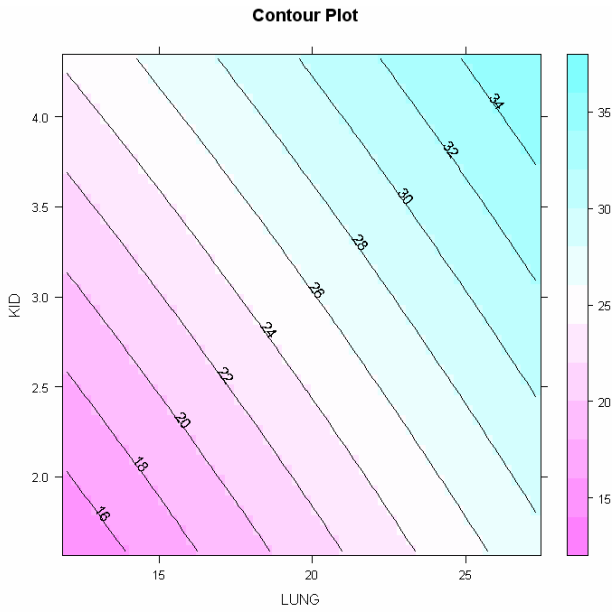


Figure 2.2.8 : ($x_2 * x_3$) contour plot

$$\hat{y} = -0.3064 + 0.8017x_2 + 3.3866x_3 \quad (2.2.7)$$

$$\hat{y} = -2.1452 + 0.8998x_2 + 4.0175x_3 - 0.0333x_2x_3 \quad (2.2.8)$$

我們從 Figure 2.2.7 的偏回歸圖可看到，在 x_2 已加入模型中的情況下再加入 x_3 後的散佈狀況大致呈一直線，表示模型在 x_2 解釋完後，x_3 的加入對模型依舊有解釋能力。由 Figure 2.2.8 看出 x_2 與 x_3 的 contour plot 並未呈曲線狀，交互作用並不顯著；而由 Table 2.2.13－2.2.16 看出對 x_2 與 x_3 此兩變數而言，未加入交互作用項 $x_2 * x_3$ 前，x_2 與 x_3 的 p-value<0.01，兩參數估計值皆為顯著，但加入交互作用項 $x_2 * x_3$ 後，x_2 與 x_3 及 $x_2 * x_3$ 的 p-value>0.01，參數估計值皆變為不顯著；未加入交互作用項 $x_2 * x_3$ 前的 $R^2_{adj} = 0.5573$ 較加入後的 $R^2_{adj} = 0.5465$ 來的大，解釋力較高，而未加入交互作用項 $x_2 * x_3$ 前的 $\sqrt{MSE} = 3.7082$ 較加入後的 $\sqrt{MSE} = 3.7534$ 來的小；因此模型並不適合加入交互作用項 $x_2 * x_3$。

Table 2.2.13. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	-0.3064	3.6024	-0.09	0.9326
x_2	1	0.8017	0.1394	5.75	<0.0001
x_3	1	3.3866	1.1358	2.98	0.0048

Table 2.2.14. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	771.8778	385.9389	28.07	<.0001
Error	41	563.7675	13.7504		
Total	43	1335.6453			

Root MSE	3.7082	R-Square	0.5779
Dependent Mean	24.9141	Adj R-Sq	0.5573
Coeff Var	14.8838		

Table 2.2.15. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	-2.1452	14.2827	-0.15	0.8814
x_2	1	0.8998	0.7499	1.20	0.2372
x_3	1	4.0175	4.8751	0.82	0.4148
$x_2 * x_3$	1	-0.0333	0.2504	-0.13	0.8947

Table 2.2.16. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	772.1276	257.3759	18.27	<.0001
Error	40	563.5177	14.0879		
Total	43	1335.6453			

Root MSE	3.7534	R-Square	0.5781
Dependent Mean	24.9141	Adj R-Sq	0.5465
Coeff Var	15.0653		

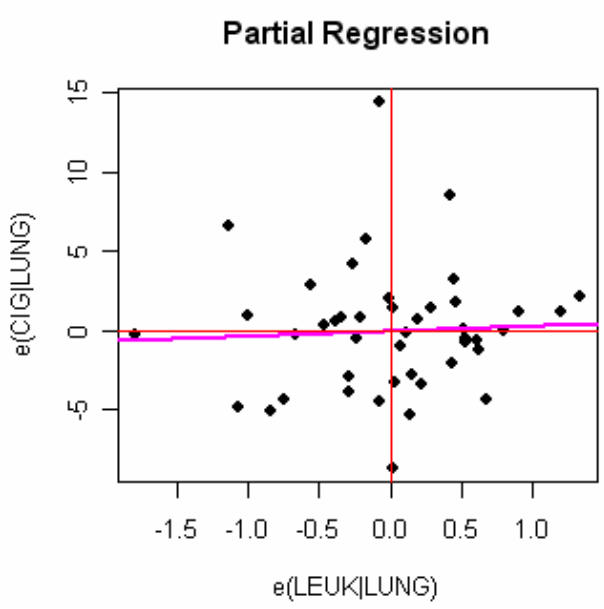


Figure 2.2.9 : Partial regression scatter plot

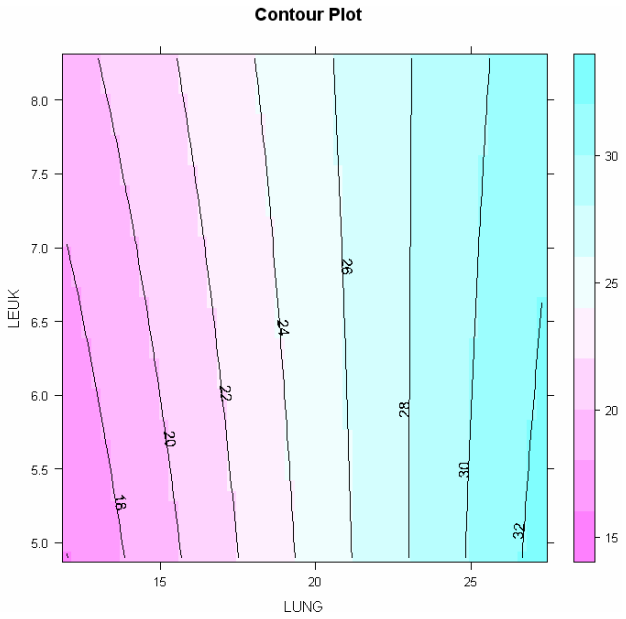


Figure 2.2.10 : ($x_2 * x_4$) contour plot

$$\hat{y} = 4.4249 + 0.9269x_2 + 0.3328x_4 \qquad (2.2.9)$$

$$\hat{y} = -7.0722 + 1.5311x_2 + 2.0189x_4 - 0.0889x_2x_4 \qquad (2.2.10)$$

我們從 Figure 2.2.9 的偏回歸圖可看到，在 x_2 已加入模型中的情況下再加入 x_4 後的散佈狀況較似隨機分佈，並不呈一直線，表示模型在 x_2 解釋完後，x_4 的加入對模型並無解釋能力。由 Figure 2.2.10 看出 x_2 與 x_4 的 contour plot 未呈曲線狀，交互作用不顯著；而由 Table 2.2.17－2.2.20 看出對 x_2 與 x_4 此兩變數而言，未加入交互作用項 $x_2 * x_4$ 前，x_2 的 p-value<0.01，而 x_4 的 p-value>0.01，x_2 的參數估計值為顯著，x_4 的參數估計值不顯著，但加入交互作用項 $x_2 * x_4$ 後，x_2 與 x_4 及 $x_2 * x_4$ 的 p-value>0.01，參數估計值皆變為不顯著；未加入交互作用項 $x_2 * x_4$ 前的 $R^2_{adj} = 0.4628$ 較加入後的 $R^2_{adj} = 0.4517$ 來的大，解釋力較高，而未加入交互作用項 $x_2 * x_4$ 前的 $\sqrt{MSE} = 4.0849$ 較加入後的 $\sqrt{MSE} = 4.127$ 來的小；因此模型並不適合加入變數 x_4 及交互作用項 $x_2 * x_4$。

Table 2.2.17. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	4.4249	7.7735	0.57	0.5723
x_2	1	0.9269	0.1491	6.22	<0.0001
x_4	1	0.3328	0.9874	0.34	0.7378

Table 2.2.18. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	651.5134	325.7567	19.52	<.0001
Error	41	684.1319	16.6861		
Total	43	1335.6453			

Root MSE	4.0849	R-Square	0.4878
Dependent Mean	24.9141	Adj R-Sq	0.4628
Coeff Var	16.3958		

Table 2.2.19. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	-7.0722	29.1920	-0.24	0.8098
x_2	1	1.5311	1.4853	1.03	0.3088
x_4	1	2.0189	4.2423	0.48	0.6367
$x_2 * x_4$	1	-0.0889	0.2174	-0.41	0.6848

Table 2.2.20. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	654.3615	218.1205	12.81	<.0001
Error	40	681.2838	17.0321		
Total	43	1335.6453			

Root MSE	4.127	R-Square	0.4899
Dependent Mean	24.9141	Adj R-Sq	0.4517
Coeff Var	16.5649		

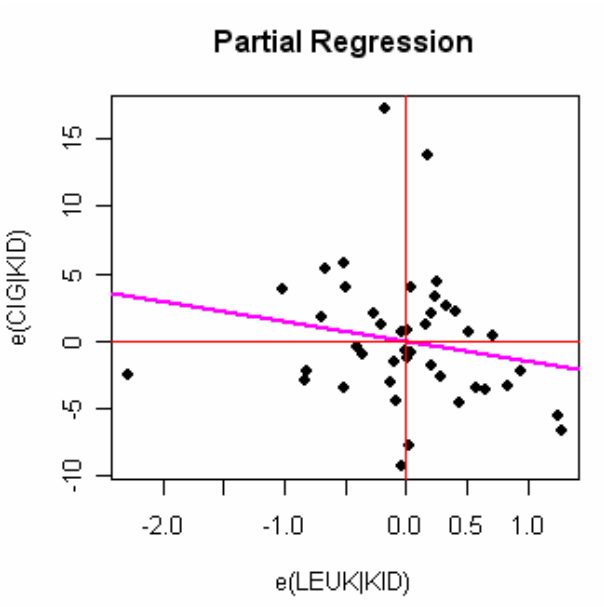


Figure 2.2.11 : Partial regression scatter plot

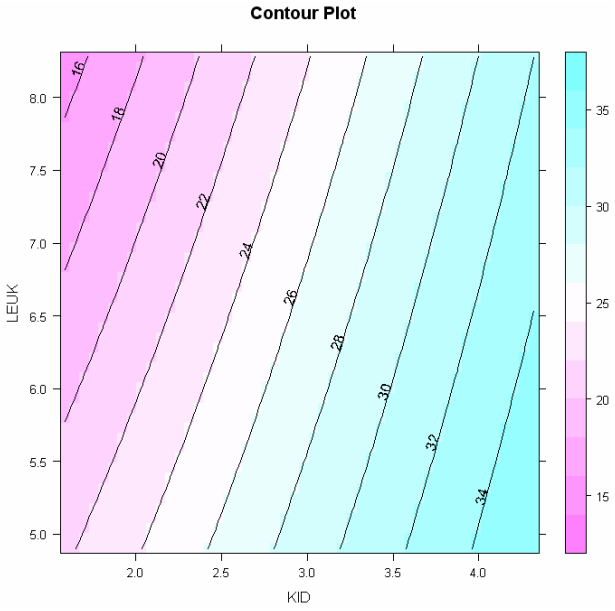


Figure 2.2.12 : ($x_3 * x_4$) contour plot

$$\hat{y} = 19.2708 + 5.5702x_3 - 1.4529x_4 \qquad (2.2.11)$$

$$\hat{y} = 24.9447 + 3.8297x_3 - 3.8297x_4 + 3.8297x_3x_4 \qquad (2.2.12)$$

我們從 Figure 2.2.11 的偏回歸圖可看到，在 x_3 已加入模型中的情況下再加入 x_4 後的散佈狀況較似隨機分佈，並不呈一直線，表示模型在 x_3 解釋完後，x_4 的加入對模型並無解釋能力。由 Figure 2.2.12 看出 x_3 與 x_4 的 contour plot 未呈曲線狀，交互作用不顯著；而由 Table 29－32 看出對 x_3 與 x_4 此兩變數而言，未加入交互作用項 $x_3 * x_4$ 前，x_3 的 p-value<0.01，而 x_4 的 p-value>0.01，x_3 的參數估計值為顯著，x_4 的參數估計值不顯著，但加入交互作用項 $x_3 * x_4$ 後，x_3 與 x_4 及 $x_3 * x_4$ 的 p-value>0.01，參數估計值皆變為不顯著；未加入交互作用項 $x_3 * x_4$ 前的 $R^2_{adj} = 0.2284$ 較加入後的 $R^2_{adj} = 0.2094$ 來的大，解釋力較高，而未加入交互作用項 $x_3 * x_4$ 前的 $\sqrt{MSE} = 4.8958$ 較加入後的 $\sqrt{MSE} = 4.9554$ 來的小；因此模型並不適合加入變數 x_4 及交互作用項 $x_3 * x_4$。

Table 2.2.21. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	19.2708	8.4209	2.29	0.0273
x_3	1	5.5702	1.4646	3.80	0.0005
x_4	1	-1.4529	1.1911	-1.22	0.2295

Table 2.2.22. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	352.9391	176.4696	7.36	0.0019
Error	41	982.7061	23.9684		
Total	43	1335.6453			
Root MSE		4.8958	R-Square		0.2642
Dependent Mean		24.9141	Adj R-Sq		0.2284
Coeff Var		19.6506			

Table 2.2.23. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	24.9447	41.7954	0.60	0.5540
x_3	1	3.8297	12.6385	0.30	0.7634
x_4	1	-3.8297	6.6353	-0.36	0.7242
$x_3 * x_4$	1	3.8297	2.0223	0.14	0.8904

Table 2.2.24. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	353.4113	117.8038	4.8	0.006
Error	40	982.2339	24.5559		
Total	43	1335.6453			
Root MSE		4.9554	R-Square		0.2646
Dependent Mean		24.9141	Adj R-Sq		0.2094
Coeff Var		19.8899			

第三節

由前兩節所配適的簡單線性回歸，發現解釋變數白血症(x_4)放入模型中解釋能力似乎較為低。而在兩兩變數的模型中，我們發現偏回歸圖，在考慮白血症(x_4)加入模型的情形，其偏回歸圖形幾乎呈水平線， 表示白血症(x_4)加入模型中似乎並無幫助。而本節我們先利用相關係數矩陣觀察變數之間的關係，並比較將所有解釋變數放入模型中與只放三個解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與反應變數菸頭數(y)的情況，並選擇較適當的模型。

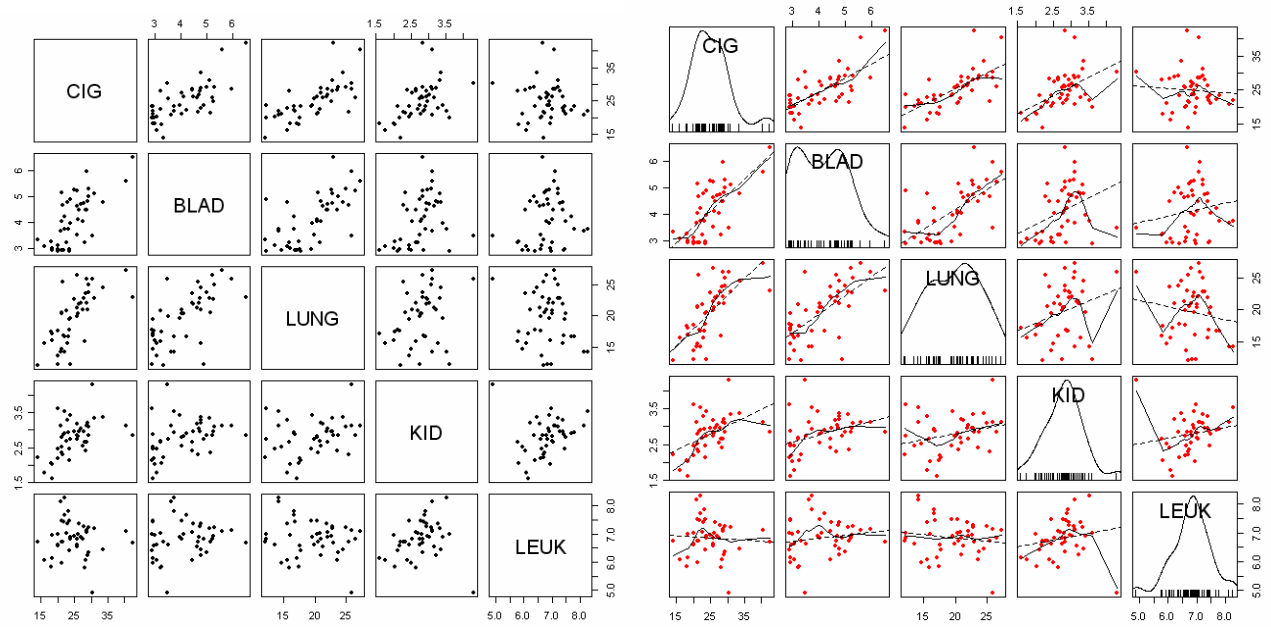


Figure 2.3.1 : Scatterplot matrix for four regressor variables

$$\hat{y} = 6.5897 + 2.378x_1 + 0.433x_2 + 2.9272x_3 - 1.1954x_4 \qquad (2.3.1)$$

我們由 Figure 2.3.2 大概發現 x_4 與 y 的相關係數散佈幾乎接近圓形，可看出 x_4 與 y 的相關性非常低；而 x_1 對 y 以及 x_2 對 y 的相關係數散佈較接近一直線，可看出 x_1、x_2 與 y 的相關性較高；而 x_3 對 y 的相關性則僅次於 x_1 與 x_2。而從 Table 2.3.1 的相關係數矩陣亦可看出此現象，相較於其他變數而言 $r_{y,x_1} = 0.7036$ 呈現高度正相關，可知 x_1 與 y 有線性關係；而 $r_{y,x_4} = -0.0685$ 呈現低度負相關，x_4 與 y 之線性關係最低。

而由 Table 2.3.2 我們可以發現 將四個解釋變數均放入的模型，除了 x_4 的 p-value>0.01 參數值不顯著外，其他三個變數 p-value 皆大於 0.01，參數值均顯著，在模型均具解釋能力。因此，我們判斷 x_4 也許應該從模型中剔除。

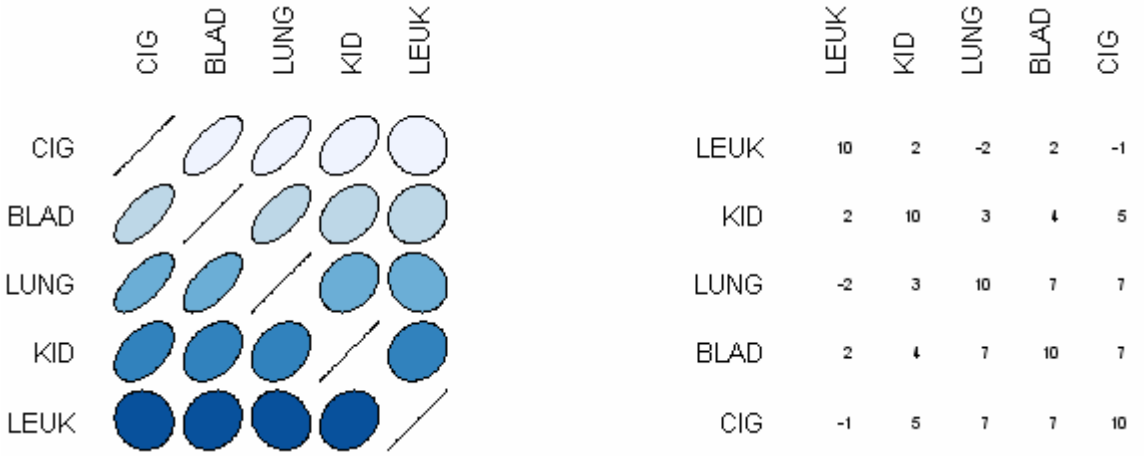


Figure 2.3.2 : correlation coefficient plot

Table 2.3.1. correlation coefficient matrix					
	y	x_1	x_2	x_3	x_4
y	1	0.7036	0.6974	0.4874	-0.0685
x_1	0.7036	1	0.6585	0.3588	0.1622
x_2	0.6974	0.6585	1	0.2827	-0.1516
x_3	0.4874	0.3588	0.2827	1	0.1887
x_4	-0.0685	0.1622	-0.1516	0.1887	1

Table 2.3.2. Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	6.5897	6.7021	0.98	0.3316
x_1	1	2.3780	0.7756	3.07	0.0039
x_2	1	0.4330	0.1755	2.47	0.0181
x_3	1	2.9272	1.0919	2.68	0.0107
x_4	1	-1.1954	0.8940	-1.34	0.1889

Table 2.3.3. Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	882.8688	220.7172	19.01	<.0001
Error	39	452.7765	11.6097		
Total	43	1335.6453			
Root MSE		3.4073	R-Square		0.661
Dependent Mean		24.9141	Adj R-Sq		0.6262
Coeff Var		13.6762			

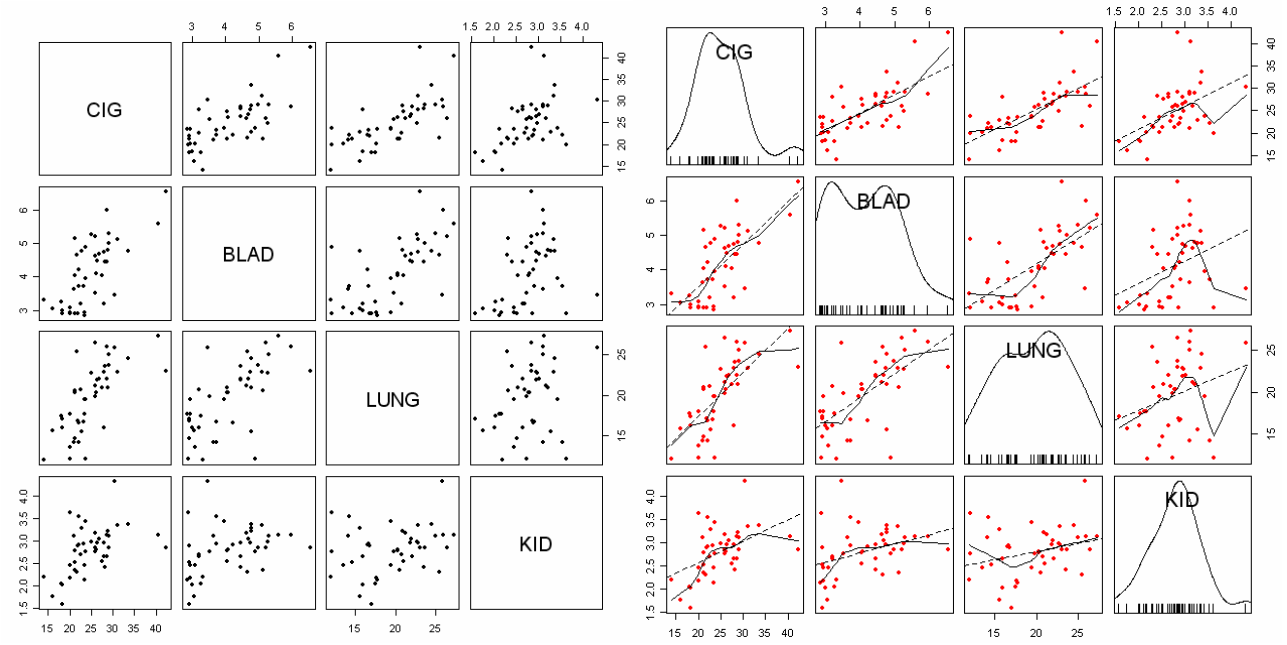


Figure 2.3.3 : Scatterplot matrix for three regressor variables

$$\hat{y} = -1.1916 + 2.0544x_1 + 0.5179x_2 + 2.6701x_3 \qquad (2.3.2)$$

由 Table 2.3.4 我們可以看到，剔除解釋變數 x_4 後的模型，參數值均為顯著，且其 $R^2_{Adj} = 0.6189$。與模型 2.3.1 之 $R^2_{Adj} = 0.6262$ 相較下，其解釋總變異能力只降低 0.73%。近一步觀察解釋變數之間的相關性，由 VIF_{x_1}、VIF_{x_2}、VIF_{x_3} 均接近 1 小於 10，表示這三個變數並無強烈的多元共線性問題存在。

Table 2.3.4. Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value	Variance Inflation
Intercept	1	-1.1916	3.3579	-0.35	0.7246	0
x_1	1	2.0544	0.7441	2.76	0.0087	1.8727
x_2	1	0.5179	0.1653	3.13	0.0032	1.7734
x_3	1	2.6701	1.0853	2.46	0.0183	1.1528

Table 2.3.5. Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	0.0184	0.0184	127.0766	0.0013
Error	39	0.0056	0.0001		
Total	43	0.0240			
Root MSE		3.4407	R-Square		0.6455
Dependent Mean		24.9141	Adj R-Sq		0.6189
Coeff Var		13.8103			

第四節

綜合以上結果，我們可以看到，無論是單一解釋變數的簡單線性回歸模式、兩解釋變數的複模式、討論加入交互作用項的兩解釋變數的複回歸模式及四個解釋變數均加入的回歸模式，解釋變數x_4在模型中的參數檢定均不顯著。而由相關係數矩陣與圖形也可發現，x_4與y的線性關係度很低。因此，我們判斷解釋變數x_4應從模型中踢除。我們認為模型 2.3.2 為最佳模型，下一章我們將對模型 2.3.2 進行殘差分析診斷。

第三章 模型之診斷及矯正策略

前言

本章主要是針對模型進行殘差診斷，包含對(1) 預測變數的殘差圖、(2) 配適值的殘差圖、(3)殘差之常態機率圖。利用*VIF*診斷預測變數間是否有存在多元共線性存在。若由殘差圖型發現殘差非固定數或非常態性時，我們可能考慮變數變換等矯正方法。接著使用標準化後的殘差值來判斷觀測值y是否存在離群值，和計算帽子矩陣槓桿值判斷觀測值x是否存有離群值。辨認出離群點後，緊接著探討這些離群值是否具影響力。使用 Cook's D 判斷其對所有配適值之影響、而$DFFITS_i$為判斷其對單一配適值之影響與$DFBETAS_{ji}$為用於判斷其對迴歸係數之影響。$COVRATIO_i > 1$時，表示觀察值i可以改善估計精確度；$COVRATIO_i < 1$時，表示觀察值i降低估計精確度，$COVRATIO_i > 1+3p/n$ or $COVRATIO_i < 1-3p/n$則第i筆觀察值可能為影響點。最後，因我們知道最小平方法容易受影響點影響，嚴重時可能會扭曲其餘觀測值之配適情形。也可能會導致遺漏重要的變數或選用不正確的函數形式。所以我們可能會比較其與刪去影響點之後配模的差別。

第一節

根據第二章結果所選擇之最佳模型 2.3.2 進行模型之診斷，包含殘差圖形診斷、判斷離群值與影響點等分析。從 Figure 3.1.1 模型 2.3.2 之殘差圖(a)發現殘差變異數不一致且略呈曲線，殘差圖(b)(c)(d)殘差點的散佈亦無在 0 上下均勻散佈，且均呈現有離群值存在的現象。而由 Figure 3.1.2 殘差機率圖發現明顯觀察出離群值且為輕尾分佈(Light – tailed errors)。接下來我們利用數值方法判斷是否有離群值與影響點存在。

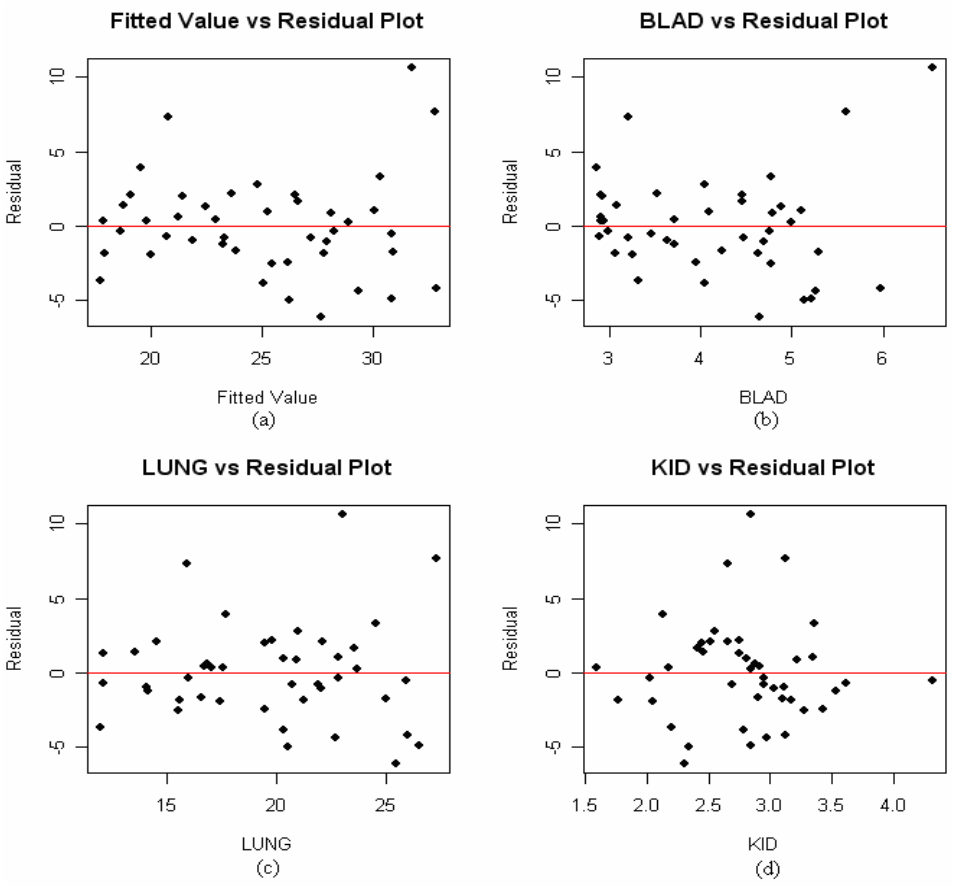


Figure 3.1.1: Residual Plot for Model 2.3.2

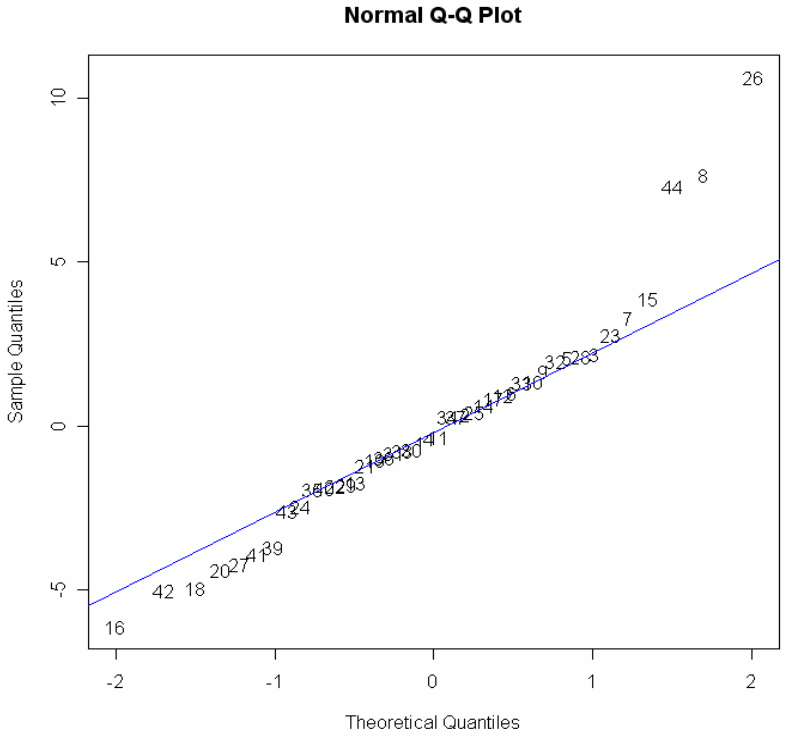


Figure 3.1.2: Normal probability plot of residuals for Model 2.3.2

離群值分析(n = 44, p = 4)

1. 對 x 之影響

利用 hat matrix 之對角線來檢視 x 之離群值，因 h_{ii} 表示每個解釋變數之元素與各解釋變數平均之距離量度，而判別式為：

$$h_{ii} > 2\overline{h}\ ,\ \overline{h} = \frac{p}{n},$$

計算結果臨界值為 0.1818，由 Table3.2.1 可以得知，在此條件下符合的觀察值有第 1、26、30、33 等四筆。

2. 對 y 之影響

利用 d_i 、t_i 來判斷 y 之離群值，其判別式為：

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}} > 3 \Rightarrow e_i > 3\sqrt{MS_{Res}}$$

$$t_i > t_{\alpha/2n,n-p-1}$$

由 Table 3.1.1 顯示，其中 d_i 大於 $3\sqrt{MSE}$ =10.32 的只有第 26 筆資料，而在 $t_{0.0011,39} = 3.5134$ 條件下，沒有任何資料大於臨界值但第 26 筆 $t_i = 3.4785$ 接近此值，因此結果為第 26 筆觀察值可能為 y 之影響點。

3. 對 $\hat{y}$ 之影響

為考慮第 i 筆觀察值對所有配適值之影響，為一比較綜合影響之量數，其意涵在於檢測第 i 筆是否為影響全體配適值結果之影響點，其判別式為：

$$Cook's\ D > \ F_{0.5,p,n-p} \approx 1$$

由 Table 3.1.1 結果顯示沒有任何一筆 *Cook's D* 值大於 1，只有在第 26 筆資料是 0.8172 最接近 1，因此考慮第 26 筆為影響點。

4. 對 $\hat{y}_i$ 之影響

計算由全體配適值減去捨棄第 i 筆所估計配適值之差除以全體之標準差估計值，其涵義為加入第 i 筆觀察值導致配適值增減多少倍的標準差估計值，其判別式為：

$$DFFITS_i > 2\sqrt{p/n}$$

由 Table 3.1.2 結果顯示在臨界值 $2\sqrt{p/n}$ = 0.603 下符合的觀察值有第 8、16、26 等三筆可能為影響點。

5. 對回歸係數($\hat{\beta}$'s)之影響

$DFBETAS_{ji}$ 其涵義本身指出納入一個觀察值將導致估計的回歸係數會增大或減少，此絕對量顯示相對於此回歸係數之估計的標準誤其差異量大小，大的 $DFBETAS_{ji}$ 值直接表示第 i 筆觀察質對第 j 個回歸係數具有較大的衝擊，因此作為辨認影響點的依據，其判別式為：

$$DFBETAS_{j,i} > 2/\sqrt{n},$$

其臨界值為0.3015，由 Table 3.1.2 結果顯示第 26 筆資料對所有回歸係數而言有明顯的效果；而個別對於 $\hat{\beta}_0$ 而言第 8、26 這兩筆符合判斷標準，對於 $\hat{\beta}_1$ 而言第 26、42 這兩筆符合判斷標準，對於 $\hat{\beta}_2$ 而言第 8、26 這兩筆符合判斷標準，對於 $\hat{\beta}_3$ 而言第 26、42 這兩筆符合判斷標準。其中 $\hat{\beta}_0$ 最大影響力都出現在第 8 筆資料，而 $\hat{\beta}_1$、$\hat{\beta}_2$、$\hat{\beta}_3$ 最大影響力都出現在第 26 筆資料。

6. 對精確度之影響

主要顯示出去除某一筆觀測值後與全體之變異數之比例，$COVRATIO_i > 1$ 時，表示加入第 i 筆觀察值可以改善估計精確度；$COVRATIO_i < 1$ 時，表示加入第 i 筆觀察值降低估計精確度，一般來說其臨界值難以估計，因此我們參考 Belsley, kuh, and welsch [1980]所提供的結果，其判別式如下：

$$COVRATIO \begin{cases} > 1+3\dfrac{p}{n} \\ < 1-3\dfrac{p}{n} \end{cases}$$

其臨界值應大於1.2727 或小於0.7273，由 Table 3.1.2 結果顯示第 1、2、30、33 這四筆觀察值可以改善估計精確度，而第 8、26、44 這三筆觀察值則會降低估計的精確度。

綜合以上判斷標準可以確定第 26 筆觀察值為影響點，而第 8 筆觀察值則需要我們多加注意其可能為影響點。

Table 3.1.1: Residual analysis

Obs	y	$\hat{y}$	e_i	R_i	T_i	h_{ii}	*PRESS*	*COOK's D*	Obs	y	$\hat{y}$	e_i	R_i	T_i	h_{ii}	*PRESS*	*COOK's D*
1	30.34	30.8536	-0.5136	-0.1979	-0.1955	0.4312	-0.9029	0.0074	23	27.56	24.7816	2.7784	0.8218	0.8184	0.0344	2.8776	0.0060
2	18.20	17.8411	0.3589	0.1136	0.1122	0.1571	0.4258	0.0006	24	23.75	26.1799	-2.4299	-0.7314	-0.7271	0.0677	-2.6064	0.0097
3	25.82	23.6362	2.1838	0.6478	0.6431	0.0402	2.2752	0.0044	25	23.32	22.8956	0.4244	0.1258	0.1243	0.0390	0.4416	0.0002
4	18.24	18.6200	-0.3800	-0.1155	-0.1141	0.0854	-0.4155	0.0003	26	42.40	31.7802	10.6198	3.4785	4.1127	0.2127	13.4886	0.8172
5	28.60	26.5026	2.0974	0.6205	0.6157	0.0349	2.1733	0.0035	27	28.64	32.8628	-4.2228	-1.3018	-1.3136	0.1112	-4.7512	0.0530
6	31.10	30.0739	1.0261	0.3077	0.3042	0.0604	1.0921	0.0015	28	21.16	19.0503	2.1097	0.6343	0.6295	0.0655	2.2575	0.0071
7	33.60	30.3133	3.2867	0.9909	0.9907	0.0708	3.5369	0.0187	29	29.14	30.9308	-1.7908	-0.5387	-0.5339	0.0667	-1.9187	0.0052
8	40.46	32.7924	7.6676	2.3532	2.5033	0.1032	8.5495	0.1592	30	19.96	20.6877	-0.7277	-0.2392	-0.2364	0.2181	-0.9307	0.0040
9	28.27	26.6118	1.6582	0.4993	0.4945	0.0683	1.7797	0.0046	31	26.38	27.2042	-0.8242	-0.2432	-0.2403	0.0298	-0.8496	0.0005
10	20.10	18.7369	1.3631	0.4114	0.4071	0.0725	1.4697	0.0033	32	23.44	21.4418	1.9982	0.6069	0.6020	0.0843	2.1822	0.0085
11	27.91	28.2507	-0.3407	-0.1009	-0.0996	0.0367	-0.3537	0.0001	33	23.78	22.4684	1.3116	0.4423	0.4378	0.2573	1.7659	0.0169
12	26.18	25.2264	0.9536	0.2806	0.2773	0.0240	0.9771	0.0005	34	29.18	28.9058	0.2742	0.0817	0.0807	0.0488	0.2883	0.0001
13	22.12	23.8330	-1.7130	-0.5110	-0.5062	0.0506	-1.8043	0.0035	35	18.06	19.9955	-1.9355	-0.5850	-0.5801	0.0753	-2.0932	0.0070
14	21.84	21.1973	0.6427	0.1937	0.1914	0.0701	0.6912	0.0007	36	20.94	21.8973	-0.9573	-0.2914	-0.2881	0.0883	-1.0501	0.0021
15	23.44	19.5425	3.8975	1.1863	1.1925	0.0882	4.2743	0.0340	37	20.08	19.7834	0.2966	0.0898	0.0886	0.0777	0.3216	0.0002
16	21.58	27.6821	-6.1021	-1.8824	-1.9469	0.1123	-6.8741	0.1121	38	22.57	23.3259	-0.7559	-0.2285	-0.2258	0.0753	-0.8175	0.0011
17	28.92	28.0906	0.8294	0.2465	0.2436	0.0440	0.8676	0.0007	39	14.00	17.7021	-3.7021	-1.1460	-1.1506	0.1185	-4.1996	0.0441
18	25.91	30.8345	-4.9245	-1.4980	-1.5225	0.0872	-5.3948	0.0536	40	25.89	27.7734	-1.8834	-0.5578	-0.5529	0.0368	-1.9554	0.0030
19	26.92	27.9475	-1.0275	-0.3038	-0.3003	0.0335	-1.0631	0.0008	41	21.17	25.0642	-3.8942	-1.1462	-1.1508	0.0249	-3.9935	0.0084
20	24.96	29.3310	-4.3710	-1.3075	-1.3196	0.0560	-4.6304	0.0254	42	21.25	26.2580	-5.0080	-1.5324	-1.5596	0.0978	-5.5510	0.0637
21	22.06	23.2564	-1.1964	-0.3758	-0.3717	0.1438	-1.3973	0.0059	43	22.86	25.4286	-2.5686	-0.8024	-0.7988	0.1344	-2.9674	0.0250
22	16.08	17.8995	-1.8195	-0.5632	-0.5583	0.1184	-2.0637	0.0106	44	28.04	20.7292	7.3108	2.1764	2.2888	0.0469	7.6703	0.0582

Table 3.1.2: Diagnostics for Leverage and Influence

						$DFBETAS_{ji}$										$DFBETAS_{ji}$			
Obs	e_i	T_i	h_{ii}	$COVRATIO_i$	$DFFITS_i$	β_0	β_1	β_2	β_3	Obs	e_i	T_i	h_{ii}	$COVRATIO_i$	$DFFITS_i$	β_0	β_1	β_2	β_3
1	-0.5136	-0.1955	0.4312	1.9380	-0.1703	0.0942	0.1151	-0.0924	-0.1285	23	2.7784	0.8184	0.0344	1.0706	0.1546	0.0427	-0.0323	0.0664	-0.0643
2	0.3589	0.1122	0.1571	1.3111	0.0484	0.0399	-0.0117	0.0079	-0.0377	24	-2.4299	-0.7271	0.0677	1.1247	-0.1960	0.0716	0.0605	-0.0018	-0.1580
3	2.1838	0.6431	0.0402	1.1052	0.1315	0.0276	-0.0860	0.0583	0.0109	25	0.4244	0.1243	0.0390	1.1497	0.0250	0.0070	-0.0010	-0.0115	0.0088
4	-0.3800	-0.1140	0.0854	1.2083	-0.0348	-0.0308	0.0091	0.0010	0.0208	26	10.6198	4.1127	0.2127	0.3327	2.1375	-0.4290	1.9376	-0.7622	-0.5485
5	2.0974	0.6157	0.0349	1.1031	0.1171	0.0020	0.0070	0.0459	-0.0424	27	-4.2228	-1.3136	0.1112	1.0471	-0.4647	0.2528	-0.2638	-0.0634	0.0188
6	1.0261	0.3042	0.0604	1.1666	0.0771	-0.0495	0.0242	0.0025	0.0358	28	2.1097	0.6295	0.0655	1.1372	0.1666	0.1203	-0.0600	-0.0476	-0.0048
7	3.2867	0.9907	0.0708	1.0781	0.2734	-0.1859	-0.0478	0.1385	0.1323	29	-1.7908	-0.5339	0.0666	1.1516	-0.1427	0.0820	-0.0395	-0.0513	-0.0101
8	7.6676	2.5033	0.1032	0.6799	0.8490	-0.5209	0.1769	0.4225	0.0136	30	-0.7277	-0.2364	0.2181	1.4071	-0.1248	-0.0026	0.0276	0.0570	-0.0930
9	1.6582	0.4945	0.0683	1.1584	0.1339	0.0142	-0.0076	0.0771	-0.0775	31	-0.8242	-0.2403	0.0298	1.1339	-0.0421	0.0124	0.0008	-0.0140	-0.0058
10	1.3631	0.4071	0.0725	1.1730	0.1138	0.0885	-0.0078	-0.0611	-0.0137	32	1.9982	0.6020	0.0843	1.1646	0.1827	0.0747	-0.1419	0.0992	-0.0292
11	-0.3407	-0.0996	0.0367	1.1476	-0.0195	0.0073	-0.0031	-0.0064	-0.0007	33	1.3116	0.4378	0.2573	1.4609	0.2577	0.0706	0.2007	-0.2361	-0.0152
12	0.9536	0.2773	0.0240	1.1249	0.0435	0.0029	-0.0076	0.0098	0.0012	34	0.2742	0.0807	0.0488	1.1625	0.0183	-0.0057	0.0053	0.0063	-0.0036
13	-1.7130	-0.5062	0.0506	1.1354	-0.1169	-0.0261	-0.0557	0.0852	-0.0194	35	-1.9355	-0.5801	0.0753	1.1563	-0.1656	-0.1345	0.0401	-0.0162	0.1103
14	0.6427	0.1914	0.0701	1.1856	0.0526	0.0154	-0.0366	0.0052	0.0196	36	-0.9573	-0.2880	0.0883	1.2035	-0.0897	-0.0186	-0.0114	0.0628	-0.0428
15	3.8975	1.1925	0.0882	1.0516	0.3708	0.2637	-0.2055	0.1127	-0.1726	37	0.2966	0.0886	0.0777	1.1989	0.0257	0.0186	-0.0136	0.0068	-0.0117
16	-6.1021	-1.9469	0.1123	0.8602	-0.6925	-0.0305	0.0445	-0.4496	0.4169	38	-0.7559	-0.2257	0.0753	1.1904	-0.0644	-0.0105	0.0520	-0.0415	-0.0025
17	0.8294	0.2436	0.0440	1.1505	0.0523	-0.0236	0.0185	-0.0092	0.0238	39	-3.7021	-1.1506	0.1185	1.0984	-0.4218	-0.3444	-0.1199	0.3009	0.1487
18	-4.9245	-1.5225	0.0872	0.9623	-0.4706	0.1933	-0.0435	-0.2872	0.0964	40	-1.8834	-0.5529	0.0368	1.1135	-0.1081	0.0486	-0.0197	0.0006	-0.0491
19	-1.0275	-0.3003	0.0335	1.1344	-0.0559	0.0224	-0.0104	-0.0101	-0.0114	41	-3.8942	-1.1508	0.0248	0.9929	-0.1837	-0.0194	0.0431	-0.0515	0.0030
20	-4.3710	-1.3196	0.0560	0.9843	-0.3215	0.1151	-0.1950	0.0149	0.0188	42	-5.0080	-1.5596	0.0978	0.9629	-0.5135	-0.1259	-0.3741	0.1388	0.3266
21	-1.1964	-0.3717	0.1438	1.2742	-0.1523	0.0103	-0.0080	0.0898	-0.1098	43	-2.5686	-0.7988	0.1344	1.1980	-0.3147	0.0156	-0.1967	0.2548	-0.1135
22	-1.8195	-0.5583	0.1183	1.2158	-0.2046	-0.1855	0.0185	0.0181	0.1519	44	7.3108	2.2888	0.0469	0.7013	0.5075	0.3036	-0.1817	-0.1232	0.0400

第二節

因我們知道最小平方法容易受影響點影響，嚴重時可能會扭曲其餘觀測值之配適情形。也可能會導致遺漏重要的變數或選用不正確的函數形式。故我們將由第一節離群值分析所得到的第 8 筆觀察值和第 26 筆觀察值兩個影響點刪去，比較在刪去影響點後之差異，且我們亦可觀察在少了此兩筆影響點下，模型之配適情形。

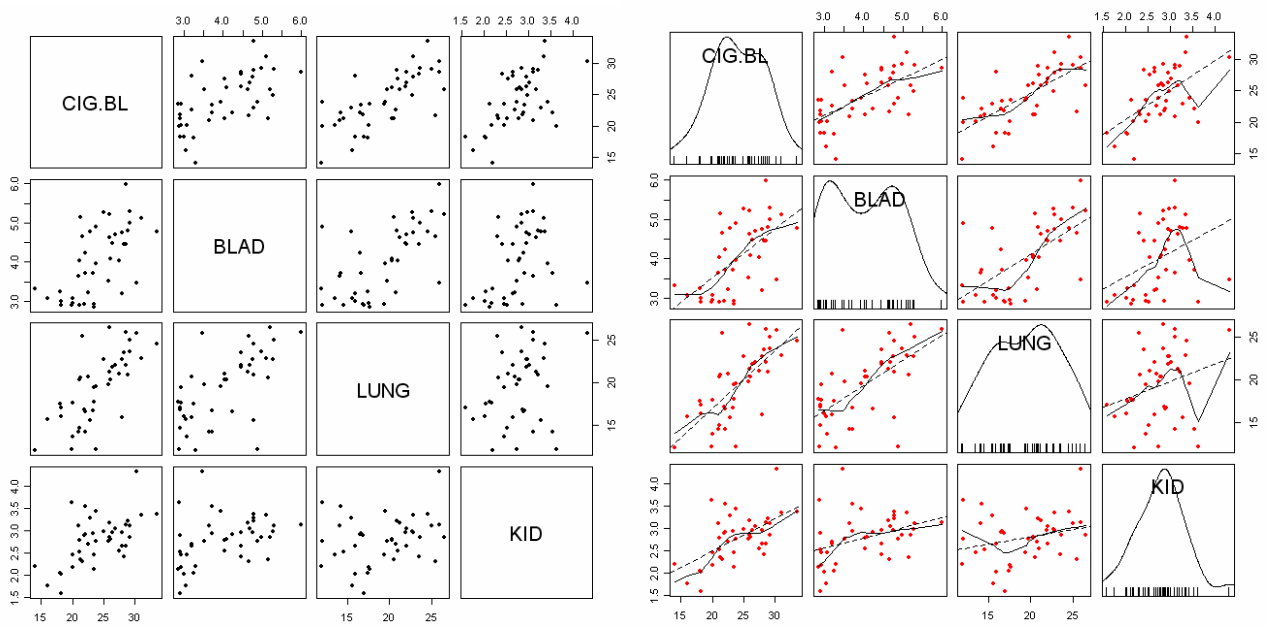


Figure 3.2.1: Scatterplot matrix for three regressor variables

在刪去第 26 筆及第 8 筆觀察值後，令模型為：

$$y^* = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \beta_3 x_3^* + \varepsilon \qquad (3.2.1)$$

Figure 3.2.1 為刪去兩個影響點後所有變數的多重散佈圖，比較轉換前 Figure 2.3.3 多重散佈圖我們可以明顯的發現 y^* 成鐘形散佈，以模型 3.2.1 作模型參數估計及殘差分析，結果從 Table 3.2.1 可以看到此參數估計式為：

$$\hat{y}^* = 2.0274 + 0.5962 x_1^* + 0.5567 x_2^* + 3.1969 x_3^* \qquad (3.2.2)$$

並發現在 $\alpha = 0.05$ 下常數項和解釋變數膀胱癌(x_1^*)的參數檢定卻不顯著，而由 Table 3.2.2 可看出 $R^2 = 68.10\%$，$R_{adj}^2 = 65.58\%$顯示在刪去兩個影響點後模型解釋能力較模型 2.3.2 提高。Figure 3.2.2 為模型 3.2.2 之殘差圖其散佈情況有較均勻，且由 Figure 3.2.3 殘差常態機率圖發現符合常態假設。接下來我們利用數值方法判斷是否有離群值與影響點存在。

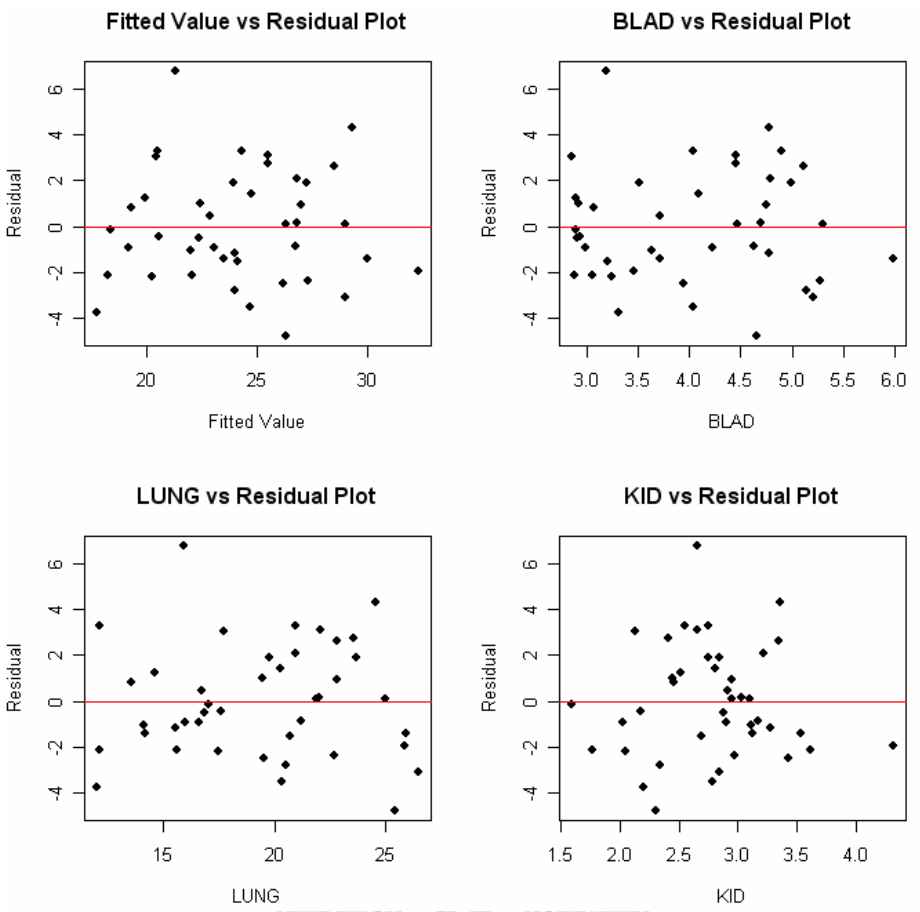


Figure 3.2.2: Residual Plot for Model 3.2.2

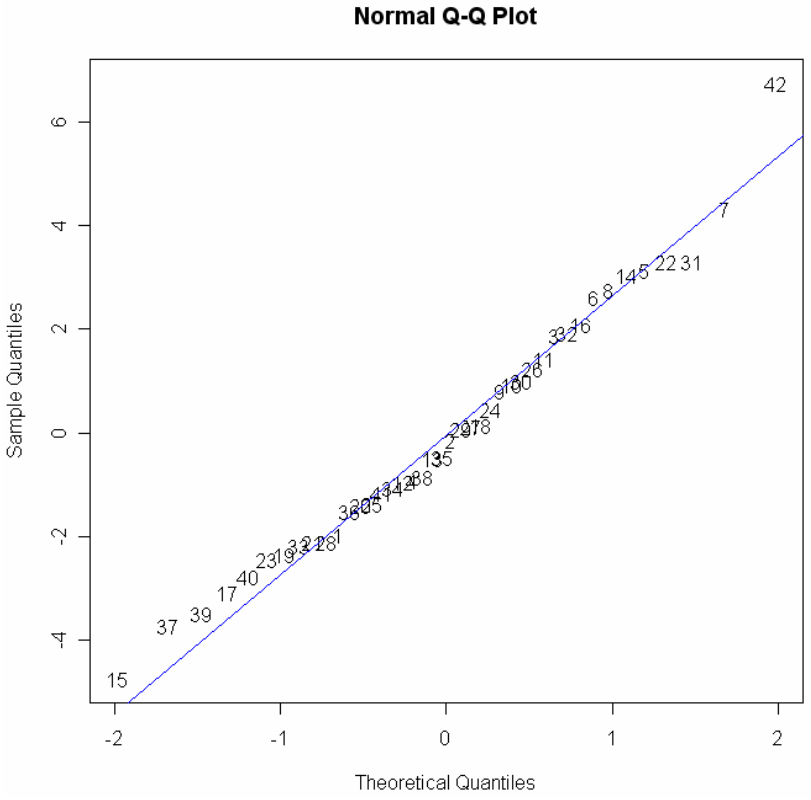


Figure 3.2.3: Normal probability plot of residuals for Model 3.2.2

Table 3.2.1: Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	2.0274	2.5405	0.80	0.4298
x_1^*	1	0.5962	0.6082	0.98	0.3332
x_2^*	1	0.5567	0.1249	4.46	<0.0001
x_3^*	1	3.1969	0.8050	3.97	0.0003

Table 3.2.2: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	519.1032	173.0344	27.04	<.0001
Error	38	243.1312	6.3982		
Total	41	762.2344			
Root MSE		2.5295	R-Square		0.6810
Dependent Mean		24.1276	Adj R-Sq		0.6558
Coeff Var		10.4837			

離群值分析(n=42,p=4)

1. 對x^*之影響

利用 hat matrix 之對角線來檢視x^*之離群值，因h_{ii}表示每個解釋變數之元素與各解釋變數平均之距離量度，而判別式為：

$$h_{ii} > 2\bar{h}\ ,\ \bar{h} = \frac{p}{n},$$

計算結果臨界值為 0.1905，由 Table3.2.3 可以得知，在此條件下符合的觀察值有第 1、 28 等兩筆。

2. 對y^*之影響

利用d_i、t_i來判斷y^*之離群值，其判別式為：

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}} > 3 \Rightarrow e_i > 3\sqrt{MS_{Res}}$$

$$t_i > t_{\alpha/2n,n-p-1}$$

由 Table 3.2.3 顯示，沒有任何觀察值的d_i大於$3\sqrt{MSE}$=7.5885，而在$t_{0.0005,37}$=3.7551 條件下，也沒有任何資料大於臨界值或接近此值，因此可能經由刪去對y^*之影響點而消除了對y^*所產生的影響。

3. 對 $\hat{y}$ 之影響

為考慮第 i 筆觀察值對所有配適值之影響，為一比較綜合影響之量數，其意涵在於檢測第 i 筆是否為影響全體配適值結果之影響點，其判別式為：

$$Cook's\ D > \ F_{0.5,p,n-p} \approx 1$$

由 Table 3.2.3 結果顯示沒有任何一筆 *Cook's D* 值大於 1。

4. 對 $\hat{y}_i$ 之影響

計算由全體配適值減去捨棄第 i 筆所估計配適值之差除以全體之標準差估計值，其涵義為加入第 i 筆觀察值導致配適值增減多少倍的標準差估計值，其判別式為：

$$DFFITS_i > 2\sqrt{p/n}$$

由 Table 3.2.4 結果顯示在臨界值 $2\sqrt{p/n} = 0.6172$ 下符合的觀察值有第 1、15、31、42 等四筆可能為影響點。

5. 對回歸係數($\hat{\beta}$'s)之影響

$DFBETAS_{ji}$ 其涵義本身指出納入一個觀察值將導致估計的回歸係數會增大或減少，此絕對量顯示相對於此回歸係數之估計的標準誤其差異量大小，大的 $DFBETAS_{ji}$ 值直接表示第 i 筆觀察質對第 j 個回歸係數具有較大的衝擊，因此作為辨認影響點的依據，其判別式為：

$$DFBETAS_{j,i} > 2/\sqrt{n},$$

其臨界值為 0.3086，由 Table 3.2.4 結果顯示第 1 筆資料對所有回歸係數而言有明顯的效果；而個別對於 $\hat{\beta}_0$ 而言第 1、7、37、42 這四筆符合判斷標準，對於 $\hat{\beta}_1$ 而言第 1、31 這二筆符合判斷標準，對於 $\hat{\beta}_2$ 而言第 1、15、31、37 這四筆符合判斷標準，對於 $\hat{\beta}_3$ 而言第 1、15 這五筆符合判斷標準。其中 $\hat{\beta}_0$ 最大影響力出現在第 1 筆資料，$\hat{\beta}_1$ 最大影響力出現在第 31 筆資料，$\hat{\beta}_2$ 最大影響力出現在第 31 筆資料，$\hat{\beta}_3$ 最大影響力出現在第 1 筆資料。

6. 對精確度之影響

主要顯示出去除某一筆觀測值後與全體之變異數之比例，$COVRATIO_i > 1$ 時，表示加入第 i 筆觀察值可以改善估計精確度；$COVRATIO_i < 1$ 時，表示加入第 i 筆觀察值降低估計精確度，一般來說其臨界值難以估計，因此我們參考 Belsley, kuh, and welsch [1980]所提供的結果，其判別式如下：

$$COVRATIO\begin{cases} > 1+3\dfrac{p}{n} \\ < 1-3\dfrac{p}{n} \end{cases}$$

其臨界值應大於 1.2857 或小於 0.7143，由 Table 3.2.4 結果顯示第 1、2、28 這三筆觀察值可以改善估計精確度，而第 42 筆觀察值則會降低估計的精確度。

綜合以上判斷標準，雖然由 Figure 3.2.4 Influence Index Plot 第 42 筆觀察值其標準化殘差高過於 3 但未超過其臨界值 $t_{0.0005,37}=3.7551$ 故在此不視為影響點。所以在刪去先前兩個影響點(第 8、26 筆)後，已無其他明顯的影響點存在。

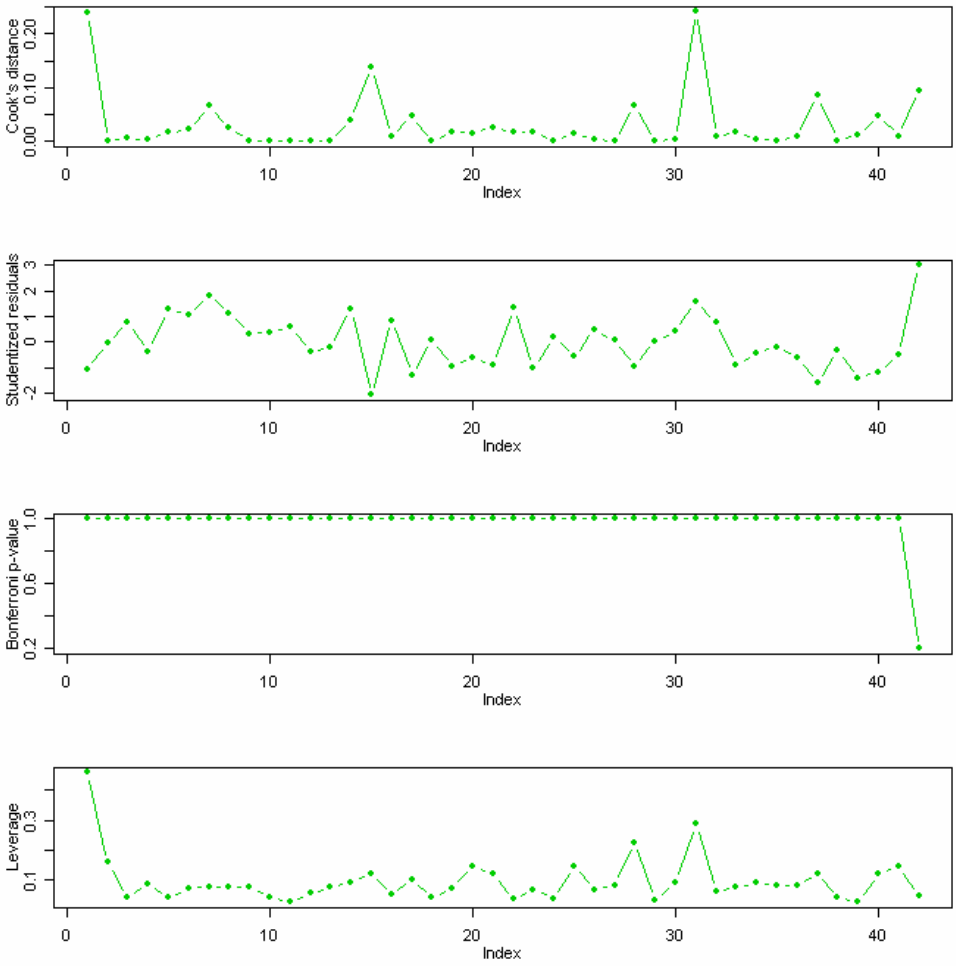


Figure 3.2.4: Influence Index Plot for Model 3.2.2

Table 3.2.3: Residual analysis

Obs	y	$\hat{y}$	e_i	R_i	T_i	h_{ii}	*PRESS*	*COOK's D*	Obs	y	$\hat{y}$	e_i	R_i	T_i	h_{ii}	*PRESS*	*COOK's D*
1	30.34	32.3072	-1.9672	-1.0582	-1.0600	0.4599	-3.6423	0.2384	22	27.56	24.2668	3.2932	1.3259	1.3397	0.0359	3.4158	0.0164
2	18.20	18.3305	-0.1305	-0.0562	-0.0555	0.1582	-0.1550	0.0002	23	23.75	26.2026	-2.4526	-1.0046	-1.0047	0.0684	-2.6327	0.0185
3	25.82	23.9393	1.8807	0.7594	0.7551	0.0415	1.9622	0.0063	24	23.32	22.8764	0.4436	0.1789	0.1766	0.0391	0.4616	0.0003
4	18.24	19.1632	-0.9232	-0.3819	-0.3776	0.0867	-1.0109	0.0035	25	28.64	30.0123	-1.3723	-0.5872	-0.5821	0.1465	-1.6078	0.0148
5	28.60	25.4757	3.1243	1.2605	1.2706	0.0397	3.2536	0.0164	26	21.16	19.9342	1.2258	0.5022	0.4972	0.0689	1.3165	0.0047
6	31.10	28.4921	2.6079	1.0699	1.0720	0.0714	2.8084	0.0220	27	29.14	29.0252	0.1148	0.0474	0.0468	0.0829	0.1251	0.0001
7	33.60	29.2848	4.3152	1.7761	1.8302	0.0774	4.6774	0.0662	28	19.96	22.0699	-2.1099	-0.9484	-0.9471	0.2264	-2.7274	0.0658
8	28.27	25.5114	2.7586	1.1335	1.1378	0.0742	2.9797	0.0258	29	26.38	26.3085	0.0715	0.0288	0.0284	0.0337	0.0740	0.0000
9	20.10	19.2875	0.8125	0.3339	0.3300	0.0745	0.8779	0.0022	30	23.44	22.4337	1.0063	0.4172	0.4126	0.0907	1.1066	0.0043
10	27.91	26.9820	0.9280	0.3752	0.3709	0.0440	0.9707	0.0016	31	23.78	20.4754	3.3046	1.5475	1.5775	0.2872	4.6364	0.2413
11	26.18	24.7493	1.4307	0.5729	0.5677	0.0252	1.4677	0.0021	32	29.18	27.2633	1.9167	0.7819	0.7778	0.0607	2.0405	0.0099
12	22.12	23.0553	-0.9353	-0.3802	-0.3759	0.0542	-0.9889	0.0021	33	18.06	20.2324	-2.1724	-0.8933	-0.8908	0.0756	-2.3500	0.0163
13	21.84	22.3436	-0.5036	-0.2072	-0.2045	0.0766	-0.5453	0.0009	34	20.94	21.9944	-1.0544	-0.4368	-0.4321	0.0892	-1.1576	0.0047
14	23.44	20.4004	3.0396	1.2610	1.2711	0.0918	3.3469	0.0402	35	20.08	20.5467	-0.4667	-0.1924	-0.1900	0.0806	-0.5076	0.0008
15	21.58	26.3196	-4.7396	-1.9993	-2.0856	0.1217	-5.3961	0.1384	36	22.57	24.0860	-1.5160	-0.6251	-0.6200	0.0807	-1.6490	0.0086
16	28.92	26.8336	2.0864	0.8466	0.8434	0.0508	2.1981	0.0096	37	14.00	17.7195	-3.7195	-1.5694	-1.6013	0.1221	-4.2366	0.0856
17	25.91	28.9851	-3.0751	-1.2839	-1.2953	0.1034	-3.4297	0.0475	38	25.89	26.7343	-0.8443	-0.3409	-0.3369	0.0416	-0.8809	0.0013
18	26.92	26.7789	0.1411	0.0569	0.0562	0.0396	0.1469	0.0000	39	21.17	24.6459	-3.4759	-1.3923	-1.4103	0.0259	-3.5682	0.0129
19	24.96	27.3114	-2.3514	-0.9659	-0.9650	0.0738	-2.5387	0.0186	40	21.25	24.0119	-2.7619	-1.1648	-1.1704	0.1212	-3.1429	0.0468
20	22.06	23.4668	-1.4068	-0.6013	-0.5962	0.1444	-1.6443	0.0153	41	22.86	24.0079	-1.1479	-0.4912	-0.4863	0.1464	-1.3449	0.0104
21	16.08	18.1942	-2.1142	-0.8905	-0.8881	0.1191	-2.4000	0.0268	42	28.04	21.3010	6.7390	2.7310	3.0059	0.0483	7.0809	0.0946

Table 3.2.4: Diagnostics for Leverage and Influence

Obs	e_i	T_i	h_{ii}	$COVRATIO_i$	$DFFITS_i$	$DFBETAS_{ji}$			
						β_0	β_1	β_2	β_3
1	-1.9672	-1.0600	0.4599	1.8276	-0.9781	0.5013	0.6670	-0.5525	-0.7387
2	-0.1305	-0.0555	0.1581	1.3211	-0.0240	-0.0196	0.0060	-0.0039	0.0183
3	1.8807	0.7551	0.0415	1.0919	0.1572	0.0322	-0.1014	0.0730	0.0162
4	-0.9232	-0.3776	0.0867	1.1996	-0.1163	-0.1024	0.0324	0.0029	0.0668
5	3.1243	1.2706	0.0397	0.9766	0.2584	-0.0164	0.0448	0.0922	-0.0951
6	2.6079	1.0720	0.0714	1.0601	0.2972	-0.1964	0.1210	0.0056	0.1140
7	4.3152	1.8302	0.0774	0.8527	0.5302	-0.3713	-0.0435	0.2627	0.2356
8	2.7586	1.1378	0.0742	1.0473	0.3222	0.0106	0.0132	0.1756	-0.1844
9	0.8125	0.3300	0.0745	1.1881	0.0936	0.0733	-0.0091	-0.0497	-0.0103
10	0.9280	0.3709	0.0440	1.1466	0.0796	-0.0337	0.0218	0.0232	-0.0005
11	1.4307	0.5677	0.0252	1.1024	0.0913	0.0011	-0.0083	0.0205	0.0012
12	-0.9353	-0.3759	0.0542	1.1585	-0.0900	-0.0160	-0.0473	0.0654	-0.0113
13	-0.5036	-0.2045	0.0766	1.1994	-0.0589	-0.0188	0.0427	-0.0075	-0.0230
14	3.0396	1.2711	0.0918	1.0325	0.4042	0.2866	-0.2326	0.1274	-0.1724
15	-4.7396	-2.0856	0.1217	0.8121	-0.7762	0.0187	-0.0226	-0.4787	0.4608
16	2.0864	0.8434	0.0508	1.0862	0.1952	-0.0950	0.0861	-0.0346	0.0742
17	-3.0751	-1.2953	0.1034	1.0392	-0.4398	0.2018	-0.0913	-0.2425	0.0964
18	0.1411	0.0562	0.0396	1.1580	0.0114	-0.0051	0.0034	0.0018	0.0017
19	-2.3514	-0.9650	0.0738	1.0875	-0.2723	0.1102	-0.1837	0.0187	0.0287
20	-1.4068	-0.5962	0.1444	1.2516	-0.2449	0.0127	-0.0104	0.1427	-0.1745
21	-2.1142	-0.8881	0.1191	1.1608	-0.3265	-0.2926	0.0305	0.0313	0.2389
22	3.2932	1.3397	0.0359	0.9548	0.2585	0.0556	-0.0334	0.1089	-0.1076
23	-2.4526	-1.0047	0.0684	1.0724	-0.2723	0.0988	0.0807	-0.0087	-0.2187
24	0.4436	0.1766	0.0391	1.1539	0.0356	0.0097	-0.0010	-0.0162	0.0123
25	-1.3723	-0.5821	0.1465	1.2569	-0.2412	0.1364	-0.1541	-0.0224	0.0211
26	1.2258	0.4972	0.0689	1.1635	0.1353	0.0992	-0.0541	-0.0358	-0.0007
27	0.1148	0.0468	0.0829	1.2129	0.0141	-0.0084	0.0054	0.0043	0.0003
28	-2.1099	-0.9471	0.2264	1.3068	-0.5124	-0.0318	0.1380	0.2201	-0.3815
29	0.0715	0.0284	0.0337	1.1512	0.0053	-0.0018	0.0005	0.0016	0.0005
30	1.0063	0.4126	0.0907	1.2012	0.1303	0.0538	-0.1026	0.0723	-0.0154
31	3.3046	1.5775	0.2872	1.2031	1.0014	0.2230	0.7993	-0.9053	-0.0979
32	1.9167	0.7778	0.0607	1.1101	0.1977	-0.0735	0.0789	0.0571	-0.0431
33	-2.1724	-0.8908	0.0756	1.1056	-0.2547	-0.2038	0.0611	-0.0250	0.1663
34	-1.0544	-0.4321	0.0892	1.1972	-0.1352	-0.0292	-0.0165	0.0944	-0.0631
35	-0.4667	-0.1900	0.0806	1.2054	-0.0562	-0.0405	0.0309	-0.0154	0.0234
36	-1.5160	-0.6200	0.0807	1.1612	-0.1837	-0.0314	0.1480	-0.1209	-0.0130
37	-3.7195	-1.6013	0.1221	0.9693	-0.5971	-0.4780	-0.1677	0.4313	0.2130
38	-0.8443	-0.3369	0.0416	1.1467	-0.0702	0.0342	-0.0196	0.0008	-0.0274
39	-3.4759	-1.4103	0.0259	0.9263	-0.2298	-0.0123	0.0358	-0.0646	0.0062
40	-2.7619	-1.1704	0.1212	1.0947	-0.4347	-0.0591	-0.3383	0.1232	0.2700
41	-1.1479	-0.4862	0.1464	1.2706	-0.2014	0.0165	-0.1335	0.1613	-0.0613
42	6.7390	3.0058	0.0483	0.4878	0.6771	0.4124	-0.2609	-0.1519	0.0646

第三節

而我們知道影響點(第 8、26 筆)之形成原因，並非資料輸入錯誤，而是起因於哥倫比亞特區(第 8 筆)與內華達州(第 26 筆)均為光觀旅遊勝地，且哥倫比亞特區為首都故每天上班會湧入大批的通勤工作者，造成此兩州菸的銷售量較高。於是我們在本節選擇利用 Box-Cox 轉換，希望藉於轉換 y 可以矯正模型 2.3.2 誤差具非常態性及誤差為不等之變異數之情況。最佳的轉換 λ 值是選取使 SSE 最小的 λ 值或使 $\ln L(\beta,\sigma^2,\lambda)$ 之值最大。

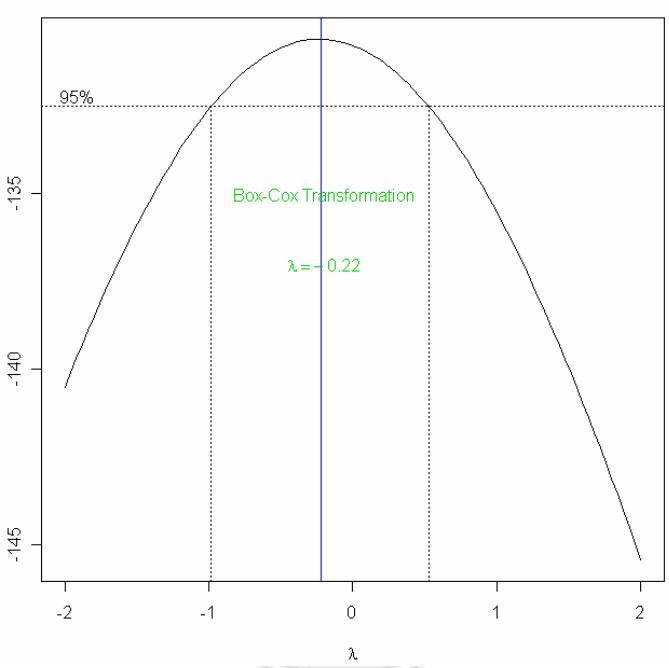


Figure 3.3.1: The plot of max($\ln L(\beta,\sigma^2,\lambda)$)

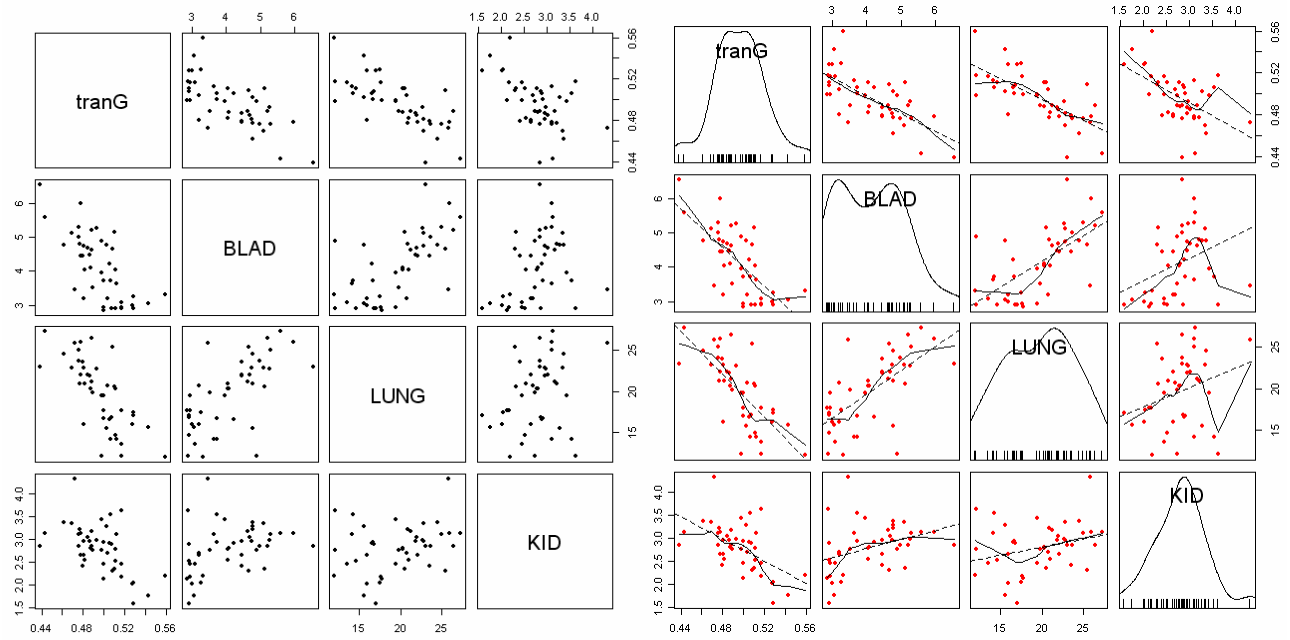


Figure 3.3.2: Scatterplot matrix for three regressor variables

使用 Box-Cox 對 y 做轉換 $y^{(\lambda)}=\begin{cases} y^{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$.從-2 到 2 之間任意選取$\lambda$，使模型為：

$$y^{(\lambda)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \tag{3.3.1}$$

由 Figure 3.3.2 可知，經由 Box-Cox 轉換使 $\max(\ln L(\beta,\sigma^2,\lambda))$之$\lambda$值為 -0.22。我們可以從 Figure 3.3.3 散佈圖矩陣看出 $y^{(-0.22)}$ 呈現鐘型分布，接著以模型 3.3.1 作模型參數估計及殘差分析，結果從 Table 3.3.1 可以清楚發現在$\alpha = 0.05$下參數均顯著，故此參數估計式為：

$$\hat{y}^{(-0.22)} = 0.6134 - 0.0070x_1 - 0.0025x_2 - 0.0143x_3 \tag{3.3.2}$$

在 Figure 3.3.3 模型 3.3.2 之殘差圖，發現殘差變異數不一致的情形有改善與 Figure 3.3.4 轉換後殘差之常態性亦有改善。而由 Table 3.3.2 可看出 $R^2 = 68.43\%$，$R^2_{adj} = 66.06\%$相對提高，但因為應變數不同無法使用判定係數做彼此模型間優劣之比較，因為轉換前後應變數之總變異是不相同，故以參數顯著性及變數間解釋意義作為選擇依據。解釋變數為百分比之單位；而應變數為數量的資料，所以轉換後模型相對於轉換前更為適合。接下來我們對轉換後的模型做殘差檢定判斷是否有離群值。

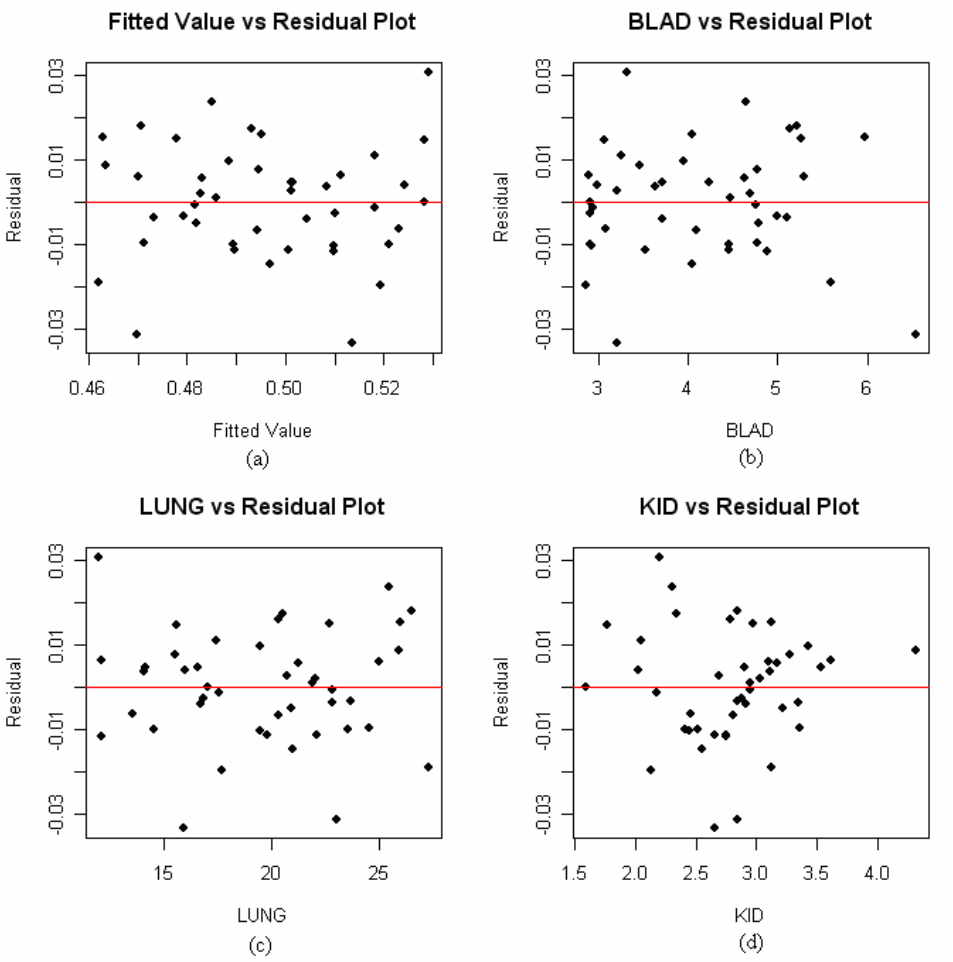


Figure 3.3.3.: Residual Plot for Model 3.3.2

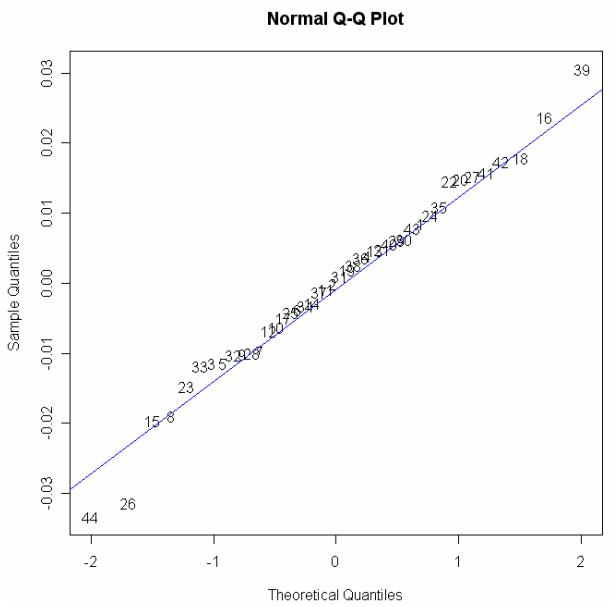


Figure 3.3.4: Normal probability plot of residuals for Model 3.3.2

Table 3.3.1: Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	0.6134	0.0134	45.63	<.0001
x_1	1	-0.0070	0.0030	-2.36	0.0232
x_2	1	-0.0025	0.0007	-3.72	0.0006
x_3	1	-0.0143	0.0043	-3.30	0.0020

Table 3.3.2: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	0.0165	0.0055	28.9	<.0001
Error	40	0.0076	0.0002		
Total	43	0.0240			
Root MSE		0.0138	R-Square		0.6843
Dependent Mean		0.496	Adj R-Sq		0.6606
Coeff Var		2.7770			

離群值分析(n=44,p=4)

1. 對 x 之影響

利用 hat matrix 之對角線來檢視 x 之離群值，因 h_{ii} 表示每個解釋變數之元素與各解釋變數平均之距離量度，而判別式為：

$$h_{ii} > 2\bar{h}\ ,\ \bar{h} = \frac{p}{n},$$

計算結果臨界值為 0.1818，由 Table3.3.3 可以得知，在此條件下符合的觀察值有第 1、26、30、33 等四筆。

2. 對 y 之影響

利用 d_i、t_i 來判斷 y 之離群值，其判別式為：

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}} > 3 \Rightarrow e_i > 3\sqrt{MS_{Res}}$$

$$t_i > t_{\alpha/2n,n-p-1}$$

由 Table 3.3.3 顯示，沒有任何觀察值的 d_i 大於 $3\sqrt{MSE}$ =0.0414，而在 $t_{0.0011,39}$ =3.5134 條件下，也沒有任何資料大於臨界值或接近此值，因此可能經由對 y 做轉換消去對 y 之影響點。

3. 對 $\hat{y}$ 之影響

為考慮第 i 筆觀察值對所有配適值之影響，為一比較綜合影響之量數，其意涵在於檢測第 i 筆是否為影響全體配適值結果之影響點，其判別式為：

$$Cook's\ D > \ F_{0.5,p,n-p} \approx 1$$

由 Table 3.3.3 結果顯示沒有任何一筆 *Cook's D* 值大於 1，只有在第 26 筆資料是 0.4454 較其他資料的 *Cook's D* 值來的大。

4. 對 $\hat{y}_i$ 之影響

計算由全體配適值減去捨棄第 i 筆所估計配適值之差除以全體之標準差估計值，其涵義為加入第 i 筆觀察值導致配適值增減多少倍的標準差估計值，其判別式為：

$$DFFITS_i > 2\sqrt{p/n}$$

由 Table 3.3.4 結果顯示在臨界值 $2\sqrt{p/n}$ = 0.603 下符合的觀察值有第 1、16、26、39 等四筆可能為影響點。

5. 對回歸係數($\hat{\beta}$'s)之影響

$DFBETAS_{ji}$ 其涵義本身指出納入一個觀察值將導致估計的回歸係數會增大或減少，此絕對量顯示相對於此回歸係數之估計的標準誤其差異量大小，大的 $DFBETAS_{ji}$ 值直接表示第 i 筆觀察質對第 j 個回歸係數具有較大的衝擊，因此作為辨認影響點的依據，其判別式為：

$$DFBETAS_{j,i} > 2/\sqrt{n},$$

其臨界值為0.3015，由 Table 3.3.4 結果顯示第 1 筆資料對所有回歸係數而言有明顯的效果；而個別對於 $\hat{\beta}_0$ 而言第 1、8、15、22、39、44 這六筆符合判斷標準，對於 $\hat{\beta}_1$ 而言第 1、26、42 這三筆符合判斷標準，對於 $\hat{\beta}_2$ 而言第 1、16、26、33、39 這五筆符合判斷標準，對於 $\hat{\beta}_3$ 而言第 1、16、22、26、39 這五筆符合判斷標準。其中 $\hat{\beta}_0$ 最大影響力出現在第 39 筆資料，$\hat{\beta}_1$ 最大影響力出現在第 26 筆資料，$\hat{\beta}_2$ 最大影響力出現在第 39 筆資料，$\hat{\beta}_3$ 最大影響力出現在第 1 筆資料。

6. 對精確度之影響

主要顯示出去除某一筆觀測值後與全體之變異數之比例，$COVRATIO_i > 1$ 時，表示加入第 i 筆觀察值可以改善估計精確度；$COVRATIO_i < 1$ 時，表示加入第 i 筆觀察值降低估計精確度，一般來說其臨界值難以估計，因此我們參考 Belsley, kuh, and welsch [1980]所提供的結果，其判別式如下：

$$COVRATIO \begin{cases} > 1+3\dfrac{p}{n} \\ < 1-3\dfrac{p}{n} \end{cases}$$

其臨界值應大於1.2727 或小於0.7273，由 Table 3.3.4 結果顯示第 1、2、21、30、33 這四筆觀察值可以改善估計精確度，而第 26、44 這兩筆觀察值則會降低估計的精確度。

綜合以上判斷標準並由 Figure 3.3.5 Influence Index Plot 第 26 筆觀察值明顯的不同於其他資料，故第 26 筆觀察值需要我們多加注意其在對 y 做轉換後仍可能為影響點。

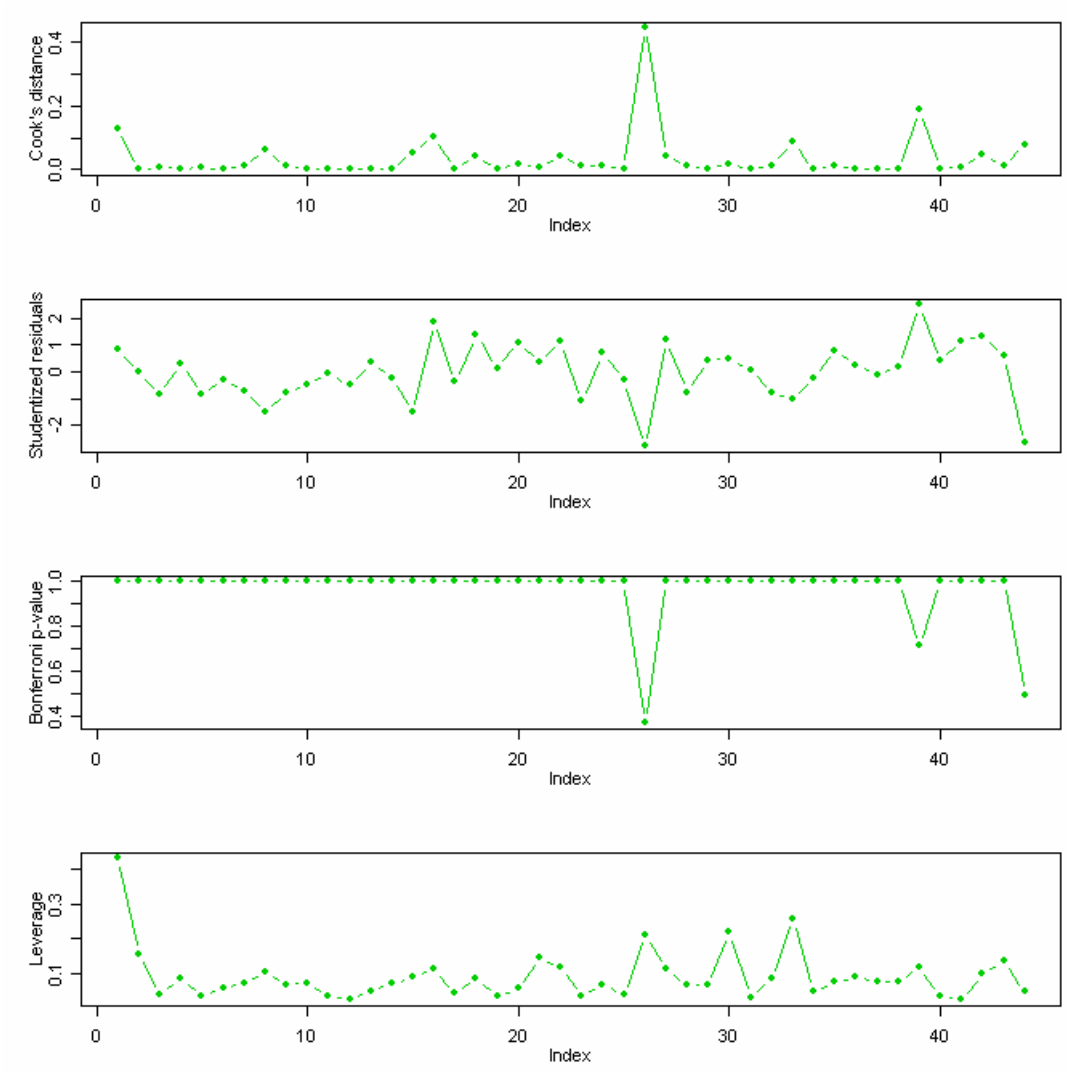


Figure 3.3.5: Influence Index Plot for Model 3.3.2

Table 3.3.3: Residual analysis

Obs	y^*	$\hat{y}^*$	e_i	R_i	T_i	h_{ii}	*PRESS*	*COOK's D*	Obs	y^*	$\hat{y}^*$	e_i	R_i	T_i	h_{ii}	*PRESS*	*COOK's D*
1	0.4720	0.4635	0.0086	0.8239	0.8206	0.4312	0.0150	0.1287	23	0.4821	0.4968	-0.0147	-1.0873	-1.0898	0.0344	-0.0152	0.0105
2	0.5282	0.5283	-0.0001	-0.0067	-0.0067	0.1571	-0.0001	0.0000	24	0.4981	0.4885	0.0097	0.7273	0.7230	0.0677	0.0104	0.0096
3	0.4891	0.5005	-0.0114	-0.8477	-0.8447	0.0402	-0.0119	0.0075	25	0.5002	0.5043	-0.0041	-0.3066	-0.3031	0.0390	-0.0043	0.0010
4	0.5279	0.5241	0.0038	0.2906	0.2872	0.0854	0.0042	0.0020	26	0.4385	0.4699	-0.0314	-2.5680	-2.7747	0.2127	-0.0399	0.4454
5	0.4782	0.4896	-0.0114	-0.8437	-0.8406	0.0349	-0.0118	0.0064	27	0.4780	0.4628	0.0153	1.1755	1.1813	0.1112	0.0172	0.0432
6	0.4695	0.4733	-0.0038	-0.2859	-0.2826	0.0604	-0.0041	0.0013	28	0.5110	0.5210	-0.0100	-0.7527	-0.7485	0.0655	-0.0107	0.0099
7	0.4615	0.4712	-0.0097	-0.7291	-0.7247	0.0708	-0.0104	0.0101	29	0.4762	0.4701	0.0061	0.4578	0.4533	0.0667	0.0065	0.0037
8	0.4431	0.4621	-0.0190	-1.4571	-1.4786	0.1032	-0.0212	0.0611	30	0.5176	0.5114	0.0062	0.5101	0.5053	0.2181	0.0079	0.0181
9	0.4794	0.4895	-0.0101	-0.7589	-0.7548	0.0683	-0.0108	0.0106	31	0.4868	0.4858	0.0009	0.0698	0.0689	0.0298	0.0010	0.0000
10	0.5168	0.5231	-0.0063	-0.4743	-0.4696	0.0725	-0.0068	0.0044	32	0.4996	0.5098	-0.0102	-0.7765	-0.7726	0.0843	-0.0112	0.0139
11	0.4808	0.4816	-0.0008	-0.0626	-0.0619	0.0367	-0.0009	0.0000	33	0.4980	0.5098	-0.0118	-0.9930	-0.9928	0.2573	-0.0159	0.0854
12	0.4876	0.4944	-0.0068	-0.5019	-0.4971	0.0240	-0.0070	0.0016	34	0.4761	0.4793	-0.0033	-0.2424	-0.2395	0.0488	-0.0034	0.0008
13	0.5060	0.5013	0.0047	0.3528	0.3489	0.0506	0.0050	0.0017	35	0.5291	0.5182	0.0109	0.8195	0.8161	0.0753	0.0117	0.0137
14	0.5074	0.5102	-0.0028	-0.2106	-0.2081	0.0701	-0.0030	0.0008	36	0.5121	0.5085	0.0036	0.2766	0.2734	0.0883	0.0040	0.0019
15	0.4996	0.5192	-0.0196	-1.4899	-1.5138	0.0882	-0.0215	0.0537	37	0.5169	0.5182	-0.0013	-0.0975	-0.0963	0.0777	-0.0014	0.0002
16	0.5088	0.4851	0.0236	1.8214	1.8780	0.1123	0.0266	0.1049	38	0.5038	0.5012	0.0025	0.1905	0.1882	0.0753	0.0027	0.0007
17	0.4770	0.4820	-0.0050	-0.3723	-0.3683	0.0440	-0.0052	0.0016	39	0.5596	0.5290	0.0305	2.3609	2.5128	0.1185	0.0346	0.1873
18	0.4887	0.4708	0.0179	1.3628	1.3781	0.0872	0.0196	0.0444	40	0.4888	0.4832	0.0056	0.4135	0.4092	0.0368	0.0058	0.0016
19	0.4846	0.4828	0.0018	0.1362	0.1345	0.0335	0.0019	0.0002	41	0.5109	0.4951	0.0158	1.1626	1.1679	0.0249	0.0162	0.0086
20	0.4927	0.4779	0.0149	1.1105	1.1138	0.0560	0.0157	0.0183	42	0.5105	0.4932	0.0173	1.3246	1.3376	0.0978	0.0192	0.0476
21	0.5063	0.5016	0.0048	0.3728	0.3688	0.1438	0.0056	0.0058	43	0.5024	0.4946	0.0078	0.6083	0.6034	0.1344	0.0090	0.0144
22	0.5428	0.5281	0.0146	1.1322	1.1363	0.1184	0.0166	0.0430	44	0.4803	0.5136	-0.0333	-2.4778	-2.6591	0.0469	-0.0350	0.0755

Table 3.3.4: Diagnostics for Leverage and Influence

Obs	e_i	T_i	h_{ii}	$COVRATIO_i$	$DFFITS_i$	$DFBETAS_{ji}$			
						β_0	β_1	β_2	β_3
1	0.0086	0.8206	0.4312	1.8168	0.7145	-0.3953	-0.4832	0.3876	0.5392
2	-0.0001	-0.0067	0.1571	1.3128	-0.0029	-0.0024	0.0007	-0.0005	0.0022
3	-0.0114	-0.8447	0.0402	1.0722	-0.1728	-0.0362	0.1130	-0.0766	-0.0143
4	0.0038	0.2872	0.0854	1.1997	0.0878	0.0775	-0.0230	-0.0024	-0.0523
5	-0.0114	-0.8406	0.0349	1.0672	-0.1599	-0.0028	-0.0096	-0.0627	0.0579
6	-0.0038	-0.2825	0.0604	1.1681	-0.0716	0.0460	-0.0225	-0.0023	-0.0332
7	-0.0097	-0.7247	0.0708	1.1288	-0.2000	0.1360	0.0350	-0.1013	-0.0968
8	-0.0190	-1.4786	0.1032	0.9920	-0.5014	0.3077	-0.1045	-0.2496	-0.0080
9	-0.0101	-0.7548	0.0683	1.1208	-0.2044	-0.0217	0.0115	-0.1177	0.1183
10	-0.0063	-0.4696	0.0725	1.1664	-0.1313	-0.1020	0.0090	0.0705	0.0158
11	-0.0008	-0.0619	0.0367	1.1483	-0.0121	0.0045	-0.0019	-0.0040	-0.0004
12	-0.0068	-0.4971	0.0240	1.1055	-0.0780	-0.0052	0.0137	-0.0176	-0.0022
13	0.0047	0.3489	0.0506	1.1511	0.0806	0.0180	0.0384	-0.0587	0.0134
14	-0.0028	-0.2081	0.0701	1.1848	-0.0571	-0.0167	0.0398	-0.0056	-0.0213
15	-0.0196	-1.5138	0.0882	0.9658	-0.4707	-0.3347	0.2609	-0.1430	0.2191
16	0.0236	1.8780	0.1123	0.8817	0.6680	0.0294	-0.0429	0.4337	-0.4022
17	-0.0050	-0.3683	0.0440	1.1415	-0.0790	0.0357	-0.0279	0.0140	-0.0360
18	0.0179	1.3781	0.0872	1.0023	0.4259	-0.1750	0.0394	0.2599	-0.0873
19	0.0018	0.1345	0.0335	1.1428	0.0250	-0.0101	0.0046	0.0045	0.0051
20	0.0149	1.1138	0.0560	1.0342	0.2714	-0.0972	0.1646	-0.0126	-0.0158
21	0.0048	0.3688	0.1438	1.2745	0.1511	-0.0102	0.0080	-0.0891	0.1089
22	0.0146	1.1363	0.1183	1.1018	0.4163	0.3775	-0.0376	-0.0369	-0.3091
23	-0.0147	-1.0898	0.0344	1.0165	-0.2058	-0.0568	0.0430	-0.0884	0.0856
24	0.0097	0.7230	0.0677	1.1254	0.1949	-0.0712	-0.0602	0.0018	0.1571
25	-0.0041	-0.3031	0.0390	1.1407	-0.0611	-0.0170	0.0025	0.0280	-0.0216
26	-0.0314	-2.7747	0.2127	0.6837	-1.4421	0.2894	-1.3073	0.5143	0.3701
27	0.0153	1.1813	0.1112	1.0817	0.4179	-0.2274	0.2372	0.0570	-0.0169
28	-0.0100	-0.7485	0.0655	1.1184	-0.1981	-0.1431	0.0713	0.0566	0.0057
29	0.0061	0.4533	0.0666	1.1609	0.1211	-0.0696	0.0336	0.0435	0.0086
30	0.0062	0.5053	0.2181	1.3787	0.2669	0.0056	-0.0591	-0.1219	0.1989
31	0.0009	0.0689	0.0298	1.1400	0.0121	-0.0036	-0.0002	0.0040	0.0017
32	-0.0102	-0.7726	0.0843	1.1373	-0.2345	-0.0958	0.1821	-0.1273	0.0375
33	-0.0118	-0.9928	0.2573	1.3483	-0.5843	-0.1602	-0.4551	0.5355	0.0344
34	-0.0033	-0.2395	0.0488	1.1565	-0.0542	0.0170	-0.0158	-0.0186	0.0107
35	0.0109	0.8161	0.0753	1.1183	0.2329	0.1893	-0.0564	0.0229	-0.1551
36	0.0036	0.2734	0.0883	1.2045	0.0851	0.0177	0.0108	-0.0596	0.0407
37	-0.0013	-0.0963	0.0777	1.1987	-0.0279	-0.0202	0.0148	-0.0073	0.0127
38	0.0025	0.1882	0.0753	1.1923	0.0537	0.0088	-0.0434	0.0346	0.0021
39	0.0305	2.5128	0.1185	0.6887	0.9211	0.7521	0.2618	-0.6572	-0.3247
40	0.0056	0.4092	0.0368	1.1293	0.0800	-0.0360	0.0146	-0.0004	0.0363
41	0.0158	1.1679	0.0248	0.9890	0.1864	0.0197	-0.0437	0.0523	-0.0030
42	0.0173	1.3376	0.0978	1.0251	0.4404	0.1079	0.3209	-0.1191	-0.2801
43	0.0078	0.6034	0.1344	1.2317	0.2378	-0.0118	0.1486	-0.1925	0.0857
44	-0.0333	-2.6591	0.0469	0.5962	-0.5897	-0.3527	0.2111	0.1431	-0.0465

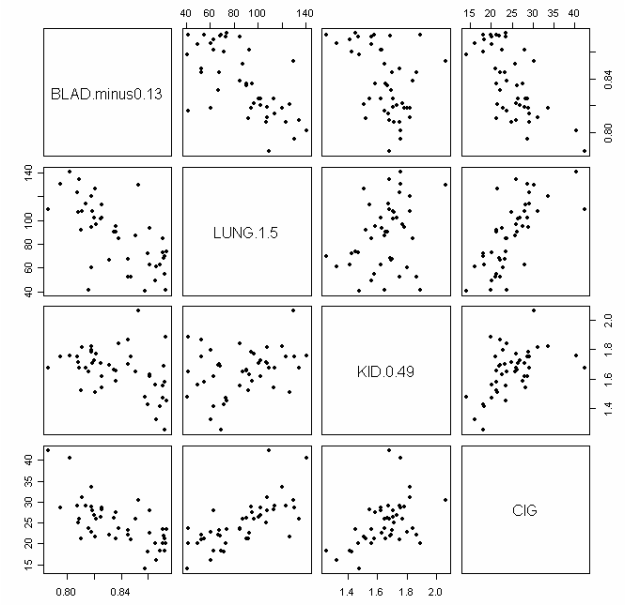


Figure 3.3.6: The plot of max($\ln L(\beta,\sigma^2,\lambda)$) for three regressor

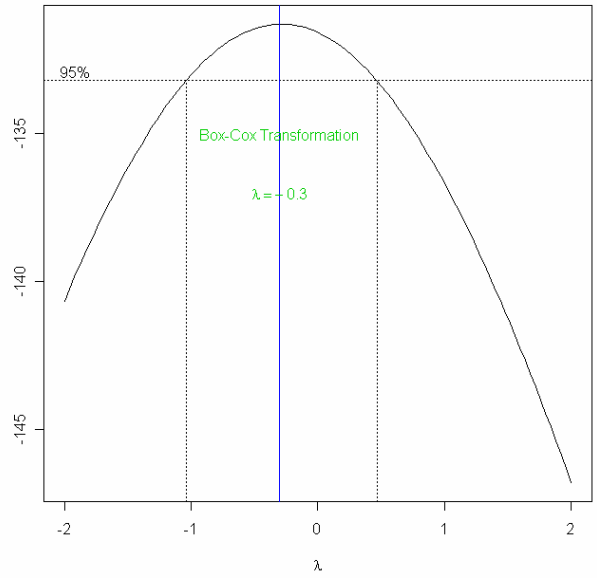


Figure 3.3.7: The plot of max($\ln L(\beta,\sigma^2,\lambda)$) for Response

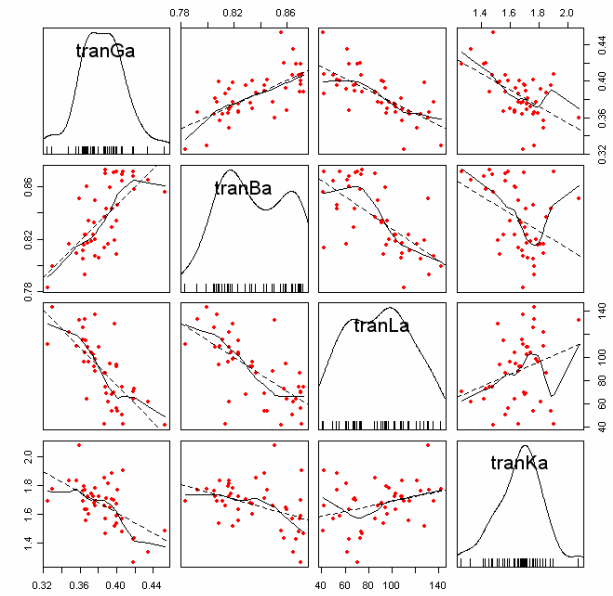


Figure 3.3.8: Scatterplot matrix for three regressor variables

從 Figure 3.1.1 殘差圖變異數不一致，我們知道除了考慮對反應變數做轉換，亦可考慮反應變數與解釋變數均做轉換。我們先對解釋變數做轉換，使模型為：

$$y = b_0 + b_1 x_1^{-0.13} + b_2 x_2^{1.5} + b_3 x_3^{0.49} + e \qquad (3.3.3)$$

並再對 y 做轉換，使用 Box-Cox 對 y 做轉換 $y^{(\lambda)} = \begin{cases} y^{\lambda}, & \lambda \neq 0 \\ \ln y, & \lambda = 0 \end{cases}$. 從-2 到 2 之間任意選取 λ，使模型為：

$$\hat{y}^{(\lambda)} = b_0 + b_1 x_1^{-0.13} + b_2 x_2^{1.5} + b_3 x_3^{0.49} + e \qquad (3.3.4)$$

由，Figure 3.3.6 為解釋變數轉換次方圖形之呈現，Figure 3.3.7 可知，經由 Box-Cox 轉換使 $\max(\ln L(\beta,\sigma^2,\lambda))$之$\lambda$值為 -0.3。接著以模型 3.3.3 作模型參數估計及殘差分析，結果從 Table 3.3.5 可以清楚發現在$\alpha = 0.05$下參數除了$x_1^{-0.13}$不為顯著但仍可接受外其餘參數均顯著，故此參數估計式為：

$$\hat{y}^{(-0.3)} = 0.3207 + 0.2268x_1^{-0.13} - 0.0004x_2^{1.5} - 0.0538x_3^{0.49}$$

(3.3.5)

而由 Table 3.3.6 可看出$R^2 = 67.51\%$，$R^2_{adj} = 65.17\%$較僅對y作轉換的R^2的R^2_{adj}相對變小，所以僅對y轉換的模型相對於對反應變數與解釋變數均轉換更為適合。

Table 3.3.5: Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	0.3207	0.1205	2.66	0.0112
$x_1^{-0.13}$	1	0.2268	0.1246	1.82	0.0761
$x_2^{1.5}$	1	-0.0004	0.0001	-3.84	0.0004
$x_3^{0.49}$	1	-0.0538	0.0162	-3.31	0.0020

Table 3.3.6: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	0.0181	0.0061	27.7	<.0001
Error	40	0.0087	0.0002		
Total	43	0.0269			

Root MSE	0.0148	R-Square	0.6751
Dependent Mean	0.3846	Adj R-Sq	0.6507
Coeff Var	3.8417		

第四節

由於 Figuare 3.1.1 殘差圖變異數不一致且略呈曲線狀，我們除了可以考慮變數轉換做矯正。我們知道當殘差有曲線時，亦有可能是模型中需加入平方項。故此節我們考慮在模型 2.3.2 中放入平方項，我們藉由檢定模型中是否須放平方項與逐步迴歸分析幫我們選取出適合加入的平方項之解釋變數配適模型。

因為由 Table 3.4.1 檢定結果可以發現 Linear 顯著(P-Value < 0.0001 < α = 0.05)，則表示解釋變數與反應變數間有線性關係；且 Quadratic 顯著(P-Value = 0.0156 < α = 0.05)，所以模型考慮加入二次項的解釋變數；然而 Crossproduct 不顯著(P-Value = 0.4610 > α = 0.05)，則此模型不考慮加入交互作用項的解釋變數。而後利用逐步回歸與所有回歸式的比較選取法來選擇變數，其中發現需要加入x_1^2項。

Table 3.4.1: Test if need the second order term (The RSREG Procedure)

Regression	DF	Type I Sum of Squares	R-Square	F Value	P-value
Linear	3	862.1085	0.6455	29.48	<.0001
Quadratic	3	116.3397	0.0871	3.98	0.0156
Crossproduct	3	25.7441	0.0193	0.88	0.4610
Total Model	9	1004.1924	0.7518	11.45	<.0001

在加入x_1^2後使得模型為:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \varepsilon \qquad (3.4.1)$$

對此模型做參數估計、ANOVA、殘差檢定並畫出散佈圖。由 Figuare 3.4.1 散佈圖矩陣可觀察資料之散佈情形，由 Table 3.4.2 結果顯示其參數估計式為：

$$\hat{y} = 23.5358 - 11.1275 x_1 + 0.5443 x_2 + 3.3897 x_3 + 1.5126 x_1^2 \qquad (3.4.2)$$

從 Table 3.4.2－3.4.3 參數估計看到在$\alpha = 0.05$下所有參數估計皆顯著表示此模型是合適的，且從 Table 3.4.3 得知R_{adj}^2=68.28%相較於模形 2.3.2 其R_{adj}^2=61.89%較大。Figuare 3.4.2 殘差圖 (a) 發現配適值與殘差分佈隱約看出其曲線現象未完全消除且明顯有離群值存在，在殘差圖(b)(c)(d)(e)解釋變數與殘差分布圖上面也明顯可以發現離群值存在。Figure 3.4.3 其殘差常態圖仍略呈輕尾分佈現象。

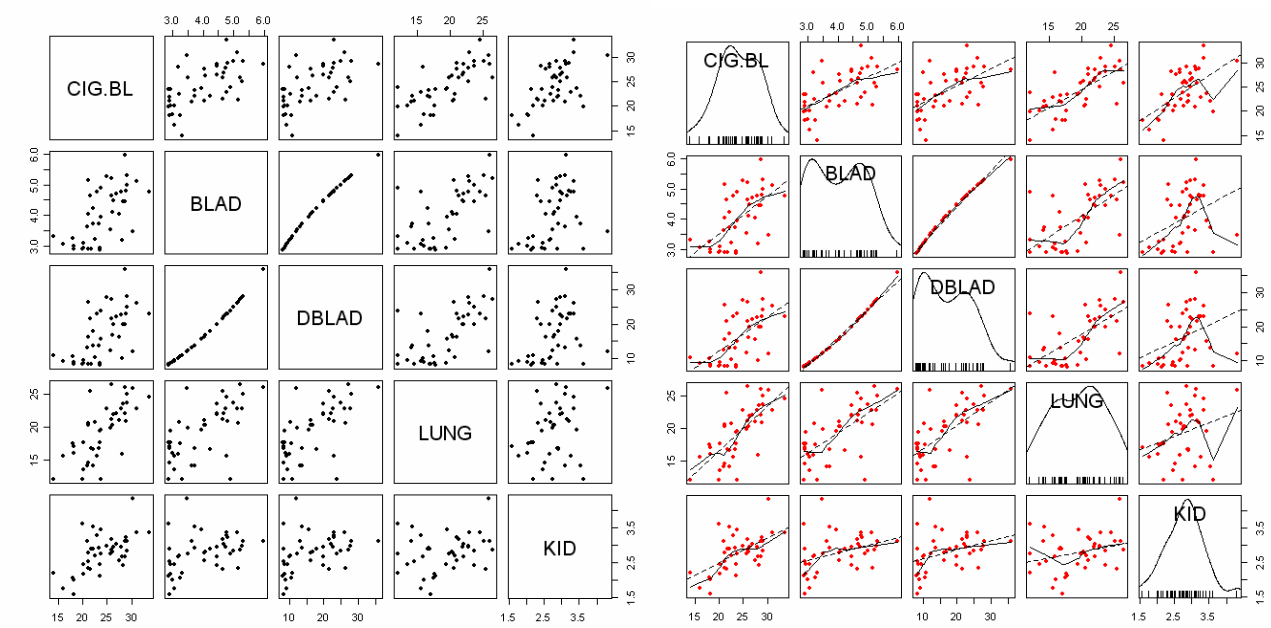


Figure 3.4.1: Scatterplot matrix for four regressor variables

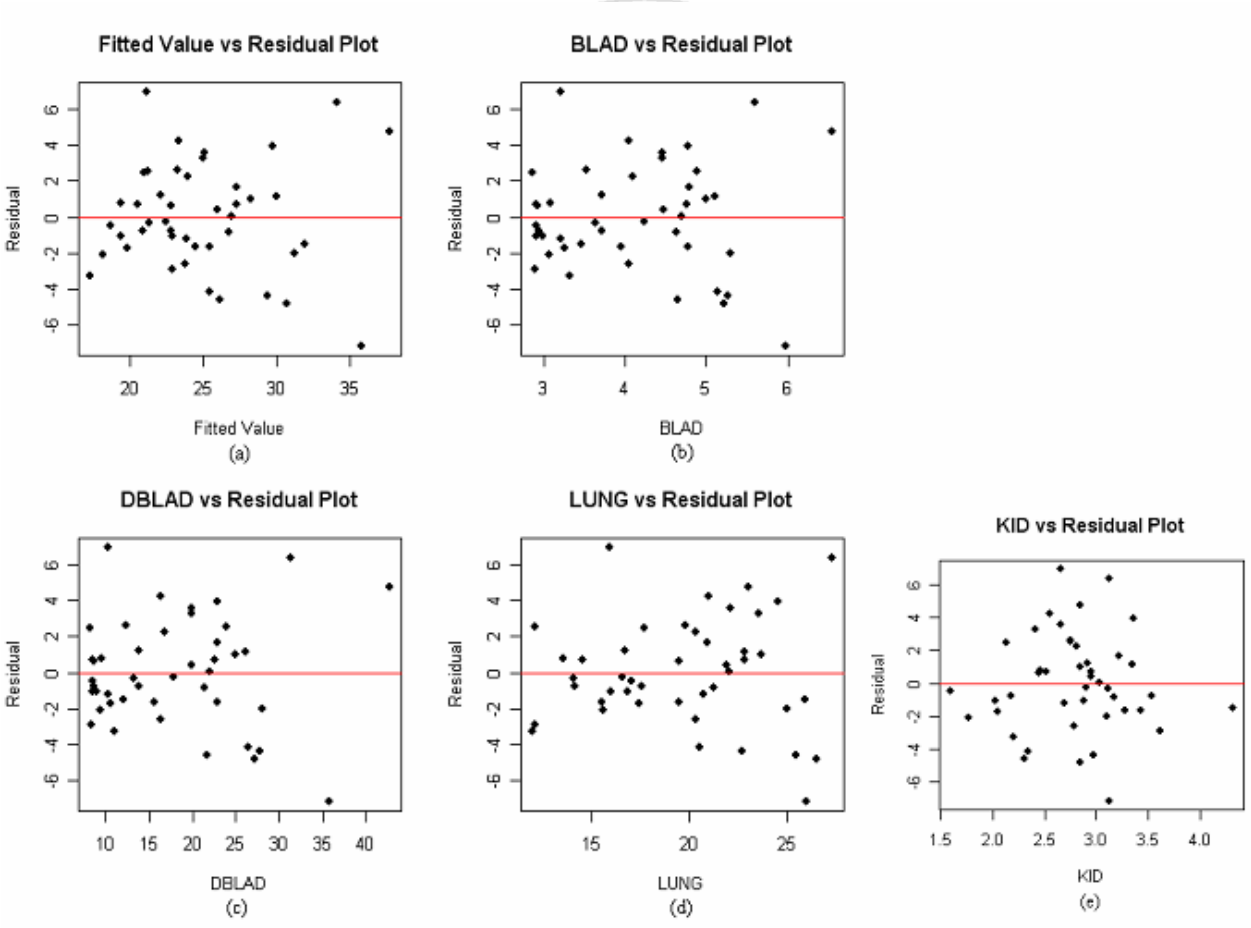


Figure 3.4.2: Residual Plot for Model 3.4.2

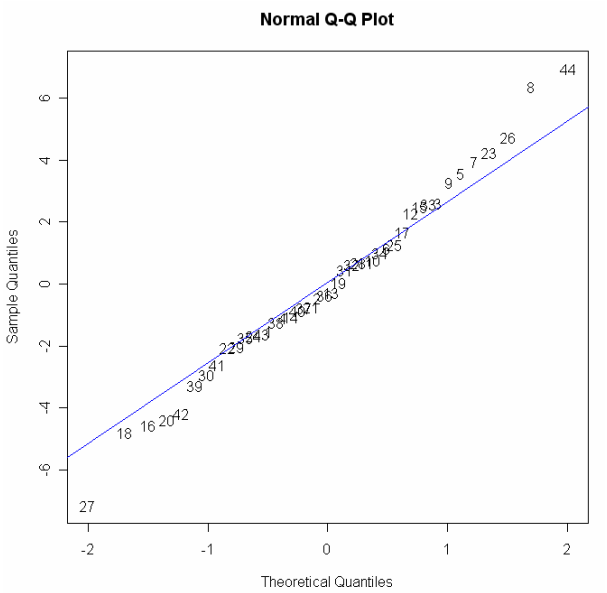


Figure 3.4.3: Normal probability plot of residuals for Model 3.4.2

Table 3.4.2: Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	23.5358	8.7641	2.69	0.0106
x_1	1	-11.1275	4.4297	-2.51	0.0163
x_2	1	0.5443	0.1510	3.60	0.0009
x_3	1	3.3897	1.0185	3.33	0.0019
x_1^2	1	1.5126	0.5023	3.01	0.0045

Table 3.4.3: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	951.4438	237.8610	24.15	<.0001
Error	39	384.2015	9.8513		
Total	43	1335.6453			
Root MSE		3.1387	R-Square		0.7123
Dependent Mean		24.9141	Adj R-Sq		0.6828
Coeff Var		12.5980			

離群值分析(n= 44, p= 5)

1. 對 x 之影響

利用 hat matrix 之對角線來檢視 x 之離群值，因 h_{ii} 表示每個解釋變數之元素與各解釋變數平均之距離量度，而判別式為：

$$h_{ii} > 2\bar{h}\ ,\ \bar{h} = \frac{p}{n},$$

計算結果臨界值為 0.2273，由 Table3.4.4 可以得知，在此條件下符合的觀察值有第 1、 26、30、33 等四筆。

2. 對 y 之影響

利用 d_i 、 t_i 來判斷 y 之離群值，其判別式為：

$$d_i = \frac{e_i}{\sqrt{MS_{Res}}} > 3 \Rightarrow e_i > 3\sqrt{MS_{Res}}$$

$$t_i > t_{\alpha/2n,n-p-1}$$

由 Table 3.4.4 顯示，沒有任何觀察值的 d_i 大於 $3\sqrt{MSE}$ =9.4161，而在 $t_{0.0005,38}$ = 3.76 條件下，也沒有任何資料大於臨界值或接近此值，因此可能經由加入 x^2 項而消除了對 y 所產生的影響。

3. 對 $\hat{y}$ 之影響

為考慮第 i 筆觀察值對所有配適值之影響，為一比較綜合影響之量數，其意涵在於檢測第 i 筆是否為影響全體配適值結果之影響點，其判別式為：

$$Cook's\ D > \ F_{0.5,p,n-p} \approx 1$$

由 Table 3.4.4 結果顯示只有第 26 筆 *Cook's D* 值大於 1。

4. 對 $\hat{y}_i$ 之影響

計算由全體配適值減去捨棄第 i 筆所估計配適值之差除以全體之標準差估計值，其涵義為加入第 i 筆觀察值導致配適值增減多少倍的標準差估計值，其判別式為：

$$DFFITS_i > 2\sqrt{p/n}$$

由 Table 3.4.5 結果顯示在臨界值 $2\sqrt{p/n}$ = 0.6742 下符合的觀察值有第 8、26、27 等三筆可能為影響點。

5. 對回歸係數($\hat{\beta}$'s)之影響

$DFBETAS_{ji}$ 其涵義本身指出納入一個觀察值將導致估計的回歸係數會增大或減少，此絕對量顯示相對於此回歸係數之估計的標準誤其差異量大小，大的 $DFBETAS_{ji}$ 值直接表示第 i 筆觀察質對第 j 個回歸係數具有較大的衝擊，因此作為辨認影響點的依據，其判別式為：

$$DFBETAS_{j,i} > 2/\sqrt{n},$$

其臨界值為 0.3015，由 Table 3.4.5 結果顯示沒有任何資料對所有回歸係數有明顯的效果；但個別對於 $\hat{\beta}_0$ 而言第 26、27 這兩筆符合判斷標準，對於 $\hat{\beta}_1$ 而言第 8、26、27、30 這四筆符合判斷標準，對於 $\hat{\beta}_2$ 而言第 1、8、16、26 這四筆符合判斷標準，對於 $\hat{\beta}_3$ 而言第 1、16、30、41 這四筆符合判斷標準，對於 $\hat{\beta}_4$ 而言第 8、26、27 這四筆符合判斷標準。其中 $\hat{\beta}_0$、$\hat{\beta}_1$、$\hat{\beta}_2$、$\hat{\beta}_4$ 最大影響力出現在第 26 筆資料，$\hat{\beta}_3$ 最大影響力則是出現在第 30 筆資料。

6. 對精確度之影響

主要顯示出去除某一筆觀測值後與全體之變異數之比例，$COVRATIO_i > 1$ 時，表示加入第 i 筆觀察值可以改善估計精確度；$COVRATIO_i < 1$ 時，表示加入第 i 筆觀察值降低估計精確度，一般來說其臨界值難以估計，因此我們參考 Belsley, kuh, and welsch [1980]所提供的結果，其判別式如下：

$$COVRATIO \begin{cases} > 1+3\dfrac{p}{n} \\ < 1-3\dfrac{p}{n} \end{cases}$$

其臨界值應大於 1.3409 或小於 0.6591，由 Table 3.4.4 結果顯示第 1、2、33 這三筆觀察值可以改善估計精確度，而第 27、44 筆觀察值則會降低估計的精確度。

綜合以上判斷標準，第 26 筆資料對 x 和 y 而言皆為離群值且會影響對回歸係數的估計。並由 Figure 3.4.4 Influence Index Plot 觀察到第 26 筆觀察值明顯與其他資料有所不同，第 26、8 筆資料則會影響配適值及回歸估計參數，且第 27 筆觀察值則由 $COVRATIO$ 判斷出其會降低估計的精確度，因此我們判定第 8，26，27 筆觀察值為影響點。

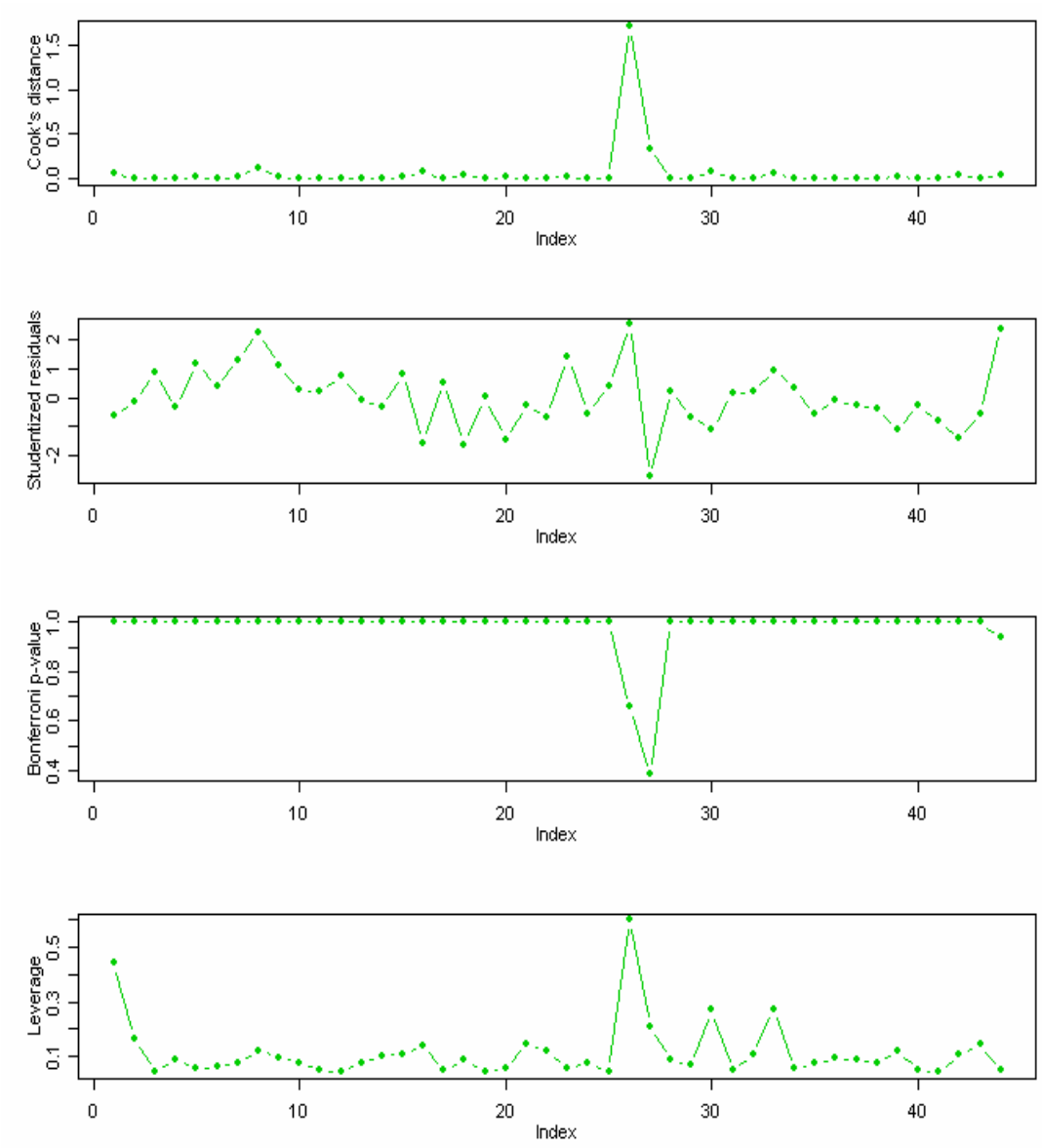


Figure 3.4.4: Influence Index Plot for Model 3.4.2

Table 3.4.4: Residual analysis

Obs	y	$\hat{y}$	e_i	R_i	T_i	h_{ii}	*PRESS*	*COOK's D*	Obs	y	$\hat{y}$	e_i	R_i	T_i	h_{ii}	*PRESS*	*COOK's D*
1	30.34	31.8729	-1.5329	-0.6543	-0.6495	0.4429	-2.7515	0.0681	23	27.56	23.3322	4.2278	1.3878	1.4050	0.0580	4.4879	0.0237
2	18.20	18.6570	-0.4570	-0.1593	-0.1573	0.1646	-0.5471	0.0010	24	23.75	25.4233	-1.6733	-0.5541	-0.5491	0.0741	-1.8073	0.0049
3	25.82	23.2077	2.6123	0.8504	0.8474	0.0422	2.7274	0.0064	25	23.32	22.0614	1.2586	0.4107	0.4063	0.0468	1.3204	0.0017
4	18.24	19.3326	-1.0926	-0.3651	-0.3610	0.0911	-1.2020	0.0027	26	42.40	37.6554	4.7446	2.3874	2.5503	0.5991	11.8340	1.7033
5	28.60	25.0249	3.5751	1.1744	1.1804	0.0594	3.8007	0.0174	27	28.64	35.7860	-7.1460	-2.5565	-2.7658	0.2069	-9.0097	0.3409
6	31.10	29.9539	1.1461	0.3767	0.3726	0.0605	1.2199	0.0018	28	21.16	20.4704	0.6896	0.2301	0.2272	0.0881	0.7561	0.0010
7	33.60	29.6593	3.9407	1.3058	1.3181	0.0755	4.2627	0.0279	29	29.14	31.1761	-2.0361	-0.6717	-0.6669	0.0673	-2.1831	0.0065
8	40.46	34.1106	6.3494	2.1597	2.2720	0.1226	7.2367	0.1304	30	19.96	22.8783	-2.9183	-1.0896	-1.0923	0.2718	-4.0076	0.0886
9	28.27	24.9939	3.2761	1.0988	1.1018	0.0976	3.6305	0.0261	31	26.38	25.9337	0.4463	0.1457	0.1439	0.0479	0.4687	0.0002
10	20.10	19.3426	0.7574	0.2511	0.2481	0.0766	0.8203	0.0011	32	23.44	22.8092	0.6308	0.2125	0.2099	0.1053	0.7050	0.0011
11	27.91	27.2184	0.6916	0.2259	0.2232	0.0487	0.7270	0.0005	33	23.78	21.2055	2.5745	0.9634	0.9625	0.2751	3.5515	0.0705
12	26.18	23.9019	2.2781	0.7422	0.7379	0.0437	2.3821	0.0050	34	29.18	28.1898	0.9902	0.3244	0.3207	0.0545	1.0473	0.0012
13	22.12	22.3917	-0.2717	-0.0900	-0.0888	0.0739	-0.2934	0.0001	35	18.06	19.7953	-1.7353	-0.5751	-0.5701	0.0758	-1.8776	0.0054
14	21.84	22.8920	-1.0520	-0.3538	-0.3498	0.1023	-1.1719	0.0029	36	20.94	21.2953	-0.3553	-0.1188	-0.1173	0.0924	-0.3914	0.0003
15	23.44	20.9433	2.4967	0.8433	0.8401	0.1101	2.8057	0.0176	37	20.08	20.8645	-0.7845	-0.2621	-0.2590	0.0908	-0.8629	0.0014
16	21.58	26.1483	-4.5684	-1.5683	-1.5993	0.1387	-5.3037	0.0792	38	22.57	23.8097	-1.2397	-0.4113	-0.4069	0.0779	-1.3444	0.0029
17	28.92	27.2533	1.6667	0.5453	0.5404	0.0518	1.7578	0.0033	39	14.00	17.2705	-3.2705	-1.1111	-1.1146	0.1206	-3.7188	0.0338
18	25.91	30.6941	-4.7841	-1.5956	-1.6290	0.0874	-5.2423	0.0488	40	25.89	26.7368	-0.8468	-0.2766	-0.2733	0.0488	-0.8903	0.0008
19	26.92	26.8868	0.0332	0.0108	0.0107	0.0461	0.0348	0.0000	41	21.17	23.7635	-2.5935	-0.8450	-0.8418	0.0438	-2.7122	0.0065
20	24.96	29.3378	-4.3778	-1.4356	-1.4560	0.0560	-4.6376	0.0245	42	21.25	25.4207	-4.1707	-1.4051	-1.4235	0.1057	-4.6634	0.0467
21	22.06	22.8022	-0.7423	-0.2559	-0.2528	0.1461	-0.8692	0.0022	43	22.86	24.4786	-1.6186	-0.5575	-0.5525	0.1445	-1.8919	0.0105
22	16.08	18.1400	-2.0600	-0.6992	-0.6946	0.1190	-2.3382	0.0132	44	28.04	21.0988	6.9412	2.2671	2.4016	0.0484	7.2942	0.0523

Table 3.4.5: Diagnostics for Leverage and Influence

Obs	T_i	h_{ii}	$COVRATIO_i$	$DFFITS_i$	$DFBETAS_{ji}$					Obs	T_i	h_{ii}	$COVRATIO_i$	$DFFITS_i$	$DFBETAS_{ji}$				
					β_0	β_1	β_2	β_3	β_4						β_0	β_1	β_2	β_3	β_4
1	-0.6495	0.4429	1.9341	-0.5790	0.0226	0.1520	-0.3149	-0.4412	-0.0938	23	1.4050	0.0580	0.9383	0.3485	-0.1821	0.2108	0.1023	-0.1607	-0.2220
2	-0.1573	0.1646	1.3586	-0.0698	-0.0336	0.0172	-0.0119	0.0481	-0.0149	24	-0.5491	0.0741	1.1822	-0.1554	0.0618	-0.0381	0.0013	-0.1056	0.0457
3	0.8474	0.0422	1.0827	0.1779	-0.0241	0.0214	0.0745	0.0047	-0.0393	25	0.4063	0.0468	1.1690	0.0900	-0.0264	0.0358	-0.0397	0.0196	-0.0367
4	-0.3610	0.0911	1.2315	-0.1143	-0.0609	0.0326	0.0014	0.0574	-0.0285	26	2.5503	0.5991	1.2890	3.1174	2.2154	-2.2160	-0.5157	0.1241	2.5036
5	1.1803	0.0594	1.0111	0.2965	-0.1769	0.1901	0.0780	-0.1247	-0.1903	27	-2.7658	0.2069	0.5738	-1.4125	-0.7030	0.8591	-0.1970	-0.1846	-0.9605
6	0.3725	0.0605	1.1902	0.0946	-0.0258	0.0094	0.0028	0.0414	-0.0049	28	0.2272	0.0880	1.2402	0.0706	0.0489	-0.0387	-0.0153	0.0067	0.0358
7	1.3181	0.0755	0.9850	0.3768	-0.1755	0.0840	0.1789	0.1493	-0.0949	29	-0.6669	0.0673	1.1519	-0.1792	0.0190	0.0101	-0.0650	-0.0165	-0.0179
8	2.2720	0.1226	0.6865	0.8493	0.1499	-0.3094	0.4067	0.0915	0.3383	30	-1.0923	0.2718	1.3397	-0.6673	-0.2823	0.3134	0.2554	-0.5026	-0.2967
9	1.1018	0.0976	1.0783	0.3624	-0.1748	0.1936	0.1628	-0.2171	-0.1985	31	0.1439	0.0479	1.1927	0.0323	-0.0212	0.0195	0.0073	-0.0012	-0.0198
10	0.2481	0.0766	1.2232	0.0714	0.0344	-0.0171	-0.0363	-0.0042	0.0165	32	0.2099	0.1053	1.2653	0.0720	0.0393	-0.0394	0.0368	-0.0025	0.0321
11	0.2232	0.0487	1.1891	0.0505	-0.0291	0.0258	0.0130	-0.0044	-0.0250	33	0.9625	0.2751	1.3926	0.5929	-0.0866	0.2177	-0.5334	-0.0682	-0.1510
12	0.7378	0.0437	1.1089	0.1576	-0.0963	0.1013	0.0202	-0.0216	-0.1057	34	0.3207	0.0545	1.1882	0.0770	-0.0314	0.0279	0.0234	-0.0198	-0.0250
13	-0.0888	0.0739	1.2282	-0.0251	0.0116	-0.0154	0.0159	-0.0001	0.0141	35	-0.5701	0.0758	1.1807	-0.1632	-0.0345	-0.0064	-0.0152	0.1083	0.0126
14	-0.3497	0.1023	1.2482	-0.1181	-0.0720	0.0758	-0.0135	-0.0509	-0.0662	36	-0.1173	0.0924	1.2523	-0.0374	0.0047	-0.0085	0.0260	-0.0152	0.0078
15	0.8401	0.1101	1.1671	0.2955	0.1894	-0.1529	0.0879	-0.0887	0.1320	37	-0.2590	0.0908	1.2414	-0.0819	-0.0482	0.0369	-0.0216	0.0260	-0.0311
16	-1.5993	0.1386	0.9545	-0.6416	0.2531	-0.2706	-0.3580	0.4036	0.2796	38	-0.4069	0.0779	1.2083	-0.1183	-0.0270	0.0358	-0.0761	-0.0095	-0.0217
17	0.5404	0.0518	1.1558	0.1263	-0.0644	0.0549	-0.0234	0.0400	-0.0492	39	-1.1146	0.1205	1.1024	-0.4126	-0.0659	-0.0714	0.2945	0.1529	0.0543
18	-1.6290	0.0874	0.8903	-0.5042	0.0960	-0.0322	-0.3053	0.1063	0.0253	40	-0.2733	0.0488	1.1855	-0.0619	0.0372	-0.0319	0.0021	-0.0165	0.0307
19	0.0107	0.0461	1.1937	0.0024	-0.0014	0.0013	0.0003	0.0001	-0.0012	41	-0.8418	0.0438	1.0857	-0.1801	0.1060	-0.1122	-0.0311	0.0300	0.1185
20	-1.4560	0.0560	0.9195	-0.3547	0.0434	-0.0319	0.0163	0.0199	-0.0011	42	-1.4235	0.1057	0.9820	-0.4893	0.0846	-0.1843	0.1348	0.3224	0.1333
21	-0.2528	0.1461	1.3223	-0.1046	0.0148	-0.0138	0.0618	-0.0696	0.0131	43	-0.5525	0.1445	1.2788	-0.2271	0.0601	-0.0803	0.1805	-0.0627	0.0600
22	-0.6946	0.1190	1.2135	-0.2553	-0.0983	0.0221	0.0214	0.1793	-0.0188	44	2.4016	0.0484	0.5903	0.5416	0.2016	-0.1244	-0.1236	0.0634	0.0963

我們在此亦考慮模型 3.3.1 加入平方項，得到模型如下：

$$y^{(-0.22)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \varepsilon \tag{3.4.3}$$

由 Table 3.4.6 結果顯示其參數估計式為：

$$\hat{y} = 0.5485 + 0.0276 x_1 - 0.0025 x_2 - 0.0162 x_3 - 0.004 x_1^2 \tag{3.4.4}$$

Table 3.4.6-3.4.7 為 y 做轉換後對 x_1, x_2, x_3, x_1^2 的參數估計表和 ANOVA 表，而由 Table3.4.7 可看出 $R^2 = 70.99\%$，$R_{adj}^2 = 68.02\%$並沒有較未轉換 y 前相對提高，但因轉換前後反應變數尺度不同，故無法使用判定係數做彼此模型間優劣之比較，而以參數顯著性及變數間解釋意義作為選擇依據。我們可以看到參數估計表內的 x_1 和 x_1^2 的 P-Value 都不為顯著因此以不對 y 轉換的模型 3.4.1 較佳。

Table 3.4.6: Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	0.5485	0.0373	14.69	<0.0001
x_1	1	0.0276	0.0189	1.46	0.1521
x_2	1	-0.0025	0.0006	-3.93	0.0003
x_3	1	-0.0162	0.0043	-3.74	0.0006
x_1^2	1	-0.0040	0.0021	-1.86	0.0711

Table 3.4.7: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	0.0171	0.0043	23.86	<0.0001
Error	39	0.0070	0.0002		
Total	43				
Root MSE		0.0134	R-Square		0.7099
Dependent Mean		0.4960	Adj R-Sq		0.6802
Coeff Var		2.6959			

接著我們想要比較影響點在配適模型上所造成之差異。故在刪去影響點後，模型中考慮加入平方項，檢定模型中是否須放平方項與逐步迴歸分析選取出最合適的模型。

原始資料(y)去除影響點之後是否需要考慮平方項之解釋變數，檢驗結果由 Table 3.4.8 看到 Linear 顯著(P-Value < 0.0001 < α = 0.05)，則解釋變數與反應變數有線性關係；且 Quadratic

不顯著(P-Value = 0.2700 > α = 0.05)，所以模型不考慮加入二次項的解釋變數；然而 Crossproduct 不顯著(P-Value = 0.2551 > α = 0.05)，則此模型不考慮加入交互作用項的解釋變數。而後利用逐步迴歸與所有回歸式的比較選取法來選擇變數並驗證是否需要平方項。

其中發現不需要加入$(x_1^*)^2$項，使得模型為: $y^* = b_0 + b_1x_1^* + b_2x_2^* + b_3x_3^*$；此模型參數估計式為$\hat{y}^{*1}$= 0.5962+ 0.5567$x_2^*$+ 3.1969$x_3^*$OVA，從 Table 3.4.8 參數估計看到$x_1^*$不顯著，從 Table 3.4.10 得知 MSE=6.3982，R_{adj}^2=0.6558，去除參數不顯著之變數x_1^*，故模型為: $y = b_0 + b_2x_2^* + b_3x_3^*$；此模型之參數式為$\hat{y}^{*1}$= 2.3987+ 0.6308$x_1^*$+ 3.41$x_3^*$。從 Table 3.4.9 參數估計看到所有參數估計皆顯著表示此模型是合適的，從 Table 3.4.12 得知 MSE=6.3918，R_{adj}^2=65.62%皆有明顯的改善，故$\hat{y}^{*1}$= 2.3987+ 0.6308x_1^*+ 3.41x_3^*為刪去影響點之最佳回歸模型。

Table 3.4.8: Test if need the second order term (The RSREG Procedure)

Regression	DF	Type I Sum of Squares	R-Square	F Value	P-value
Linear	3	519.1032	0.6810	28.73	<.0001
Quadratic	3	24.7297	0.0324	1.37	0.2700
Crossproduct	3	25.6528	0.0337	1.42	0.2551
Total Model	9	569.4857	0.7471	10.51	<.0001

Table 3.4.9: Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	2.0274	2.5405	0.80	0.4298
x_1	1	0.5962	0.6082	0.98	0.3332
x_2	1	0.5567	0.1249	4.46	<.0001
x_3	1	3.1969	0.8050	3.97	0.0003

Table 3.4.10: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	519.1032	173.0344	27.04	<.0001
Error	38	243.1312	6.3982		
Total	41	762.2344			

Root MSE	2.5295	R-Square	0.6810
Dependent Mean	24.1276	Adj R-Sq	0.6558
Coeff Var	10.4837		

Table 3.4.11: Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	2.3987	2.5108	0.96	0.3453
x_2	1	0.6308	0.0994	6.35	<.0001
x_3	1	3.4100	0.7747	4.4	<.0001

Table 3.4.12: Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	512.95523	256.4776	40.13	<.0001
Error	39	249.27913	6.3918		
Total	41	762.23436			
Root MSE		2.5282	R-Square		0.6730
Dependent Mean		24.1276	Adj R-Sq		0.6562
Coeff Var		10.4784			

第五節 結論

首先由模型 2.3.2 我們發現殘差與預測值之散佈圖隱約呈現非線性曲線關係，且發現第 8、26 兩筆觀測值為影響點。首先我們考慮比較刪除第 8、26 這兩筆影響點是否能消除非線性殘差的情形，結果顯示在此條件下未能消除非線性的曲線關係，因此接下來我們使用對變數轉換和加入平方項兩種方法檢測是否能消除殘差的非線性情況。接著我們嘗試藉經由對 y 做轉換討論是否消除非線性的散佈情況並探討在轉換後是否能減緩其離群值對模型所產生的影響，結果發現經過轉換後的模型較能解釋反應變數與應變數的關係，但第 26 點仍為影響點。於是我們更進一步的對所有變數做轉換，結果顯示此模型相對於僅對 y 做轉換的模型並沒有較佳的解釋能力。最後我們加入平方項的模型，而此模型結果和模型 2.3.2 比較下有較佳的解釋能力，但其殘差顯示能有非線性的情況存在且有離群值存在，故分別考慮並比較模型 3.3.1 加入平方項的模型與模型 3.4.1 刪除影響點的模型之影響，但此兩模型皆沒有比模型 2.3.2 加入平方項的模型來的好。因此，本章得到兩個較佳的模型，分別為模型 3.3.2 和模型 3.4.2。

但比較轉換後模型 3.3.2 其殘差之矯正情況與加入平方項模型 3.4.2 之殘差矯正情形，我們認為轉換後之模型對殘差之矯正情形較佳。且考慮在模型精簡原則，我們選擇模型 3.3.2 作為最佳模型。

第四章 逐步回歸建立回歸模型

前言

一般在建立回歸模式時，希望盡可能包含主要之解釋變數以減少誤差，並求得較精確之結果。因解釋變數過多常導致模式冗大而分析困難。而本章介紹逐步迴歸分析，藉由向前選擇法(Forward selection)、向後消去法(Backward elimination)、逐步選擇法(Stepwise regression)，選擇出對回歸模型具有較佳解釋能力之解釋變數之組合，而得到一個精簡之回歸模式。故逐步回歸程序乃利用選擇一組新集合的解釋變數以產生一較佳的回歸模式，將一連串的變數簡化，藉由逐步回歸將與反應變數有顯著關係之解釋變數選出，刪除對此回歸模式不顯著之相關解釋變數。解釋變數之選取或消去通常使用 F 檢定之顯著水準來決定其去留，而解釋變數之偏 R^2 值為其對模式之影響效果，可作為模式選入之參考標準。此外，我們也可以利用所有回歸式的比較選取法(All possible regressions procedure method)中的指標來考慮模型，選擇最佳變數之根據為以下四點：(P：參數個數)

a. $R_p^2 = \frac{SS_R(p)}{SS_T} = 1 - \frac{SS_{\mathrm{Re}s}(p)}{SS_T}$ 及 $R_{Adj,p}^2 = 1 - \left(\frac{n-1}{n-p}\right)(1-R_p^2)$ 為最大。

b. $C_p = \frac{SS_{\mathrm{Re}s}(p)}{\hat{\sigma}^2} - n + 2p$ 值最小及 $C_p \to P$。

c. $AIC(p) = n\ln\hat{\sigma}^2 + 2p$ 值最小。

d. $MS_{\mathrm{Re}s}(p) = \frac{SS_{\mathrm{Re}s}(p)}{n-p}$ 值最小。

第一節

本節為考慮反應變數(y)與解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與其解釋變數之平方項，利用逐步回歸分析(Stepwise regression methods)來選擇較佳模型之組合。而逐步回歸包含以下三種方法：1.向前選擇法(Forward selection) 2. 向後消去法(Backward elimination) 3. 逐步選擇法(Stepwise regression)。

假設考慮之解釋變數為 $X_k = \{x_1, x_2, x_3, x_1^2, x_2^2, x_3^2\}$，k = 1, …, 6；以下為說明逐步回歸選取法之結果：($\alpha = 0.05, F_{in} = F_{out} \approx 4$)

1. 向前選擇法(Forward selection)

此法一開始假設模式中沒有任何解釋變數，而後第一步驟選入之變數為 x_1^2(k = 4)，由 Table 4.1.1 得知此解釋變數之 $F_4^* = 45.01 > F_{in}$，統計上顯著(P-value < 0.05)，由 ANOVA Table 4.1.2 得知，R^2 = 0.5173，C_p = 26.7792；第二步驟選入之變數為 x_2(k = 2)，由 Table 4.1.3 得知

此解釋變數之 $F^*_{2|4}=9.71>F_{in}$，統計上顯著(P-value = 0.0033 < 0.05)，由 ANOVA Table 4.1.4 得知，$R^2=0.6098$，$C_p=15.9912$；第三步驟選入之變數為 x_3 (k = 3)，由 Table 4.1.5 得知此解釋變數之 $F^*_{3|4,2}=6.71>F_{in}$，統計上顯著(P-value = 0.0133 < 0.05)，由 ANOVA Table 4.1.6 得知，$R^2=0.6658$，$C_p=10.2365$; 第四步驟選入之變數為 x_1 (k = 1)，由 Table 4.1.7 得知此解釋變數之 $F^*_{1|4,2,3}=6.31>F_{in}$，統計上顯著(P-value = 0.0163 < 0.05)，由 ANOVA Table 4.1.8 得知，$R^2=0.7123$，$C_p=5.7972$；由 Table 4.1.9 得知依自變數的重要性逐漸選入的優先順序為 x_1^2, x_2, x_3, x_1。

2. 向後消去法(Backward elimination)

此法一開始假設模式中包括所有解釋變數，第一步驟去除之變數為 x_2^2 (k = 5)由 Table 4.1.10 得知解釋變數之 $F^*_5=0.61<F_{out}$，統計上不顯著(P-value = 0.44 > 0.05)，去除 x_2^2 後由 ANOVA Table 4.1.13 得知，$R^2=0.7282$，$C_p=5.6094$；第二步驟去除之變數為 x_3^2 (k = 6)由 Table 4.1.12 得知解釋變數之 $F^*_{6|5}=2.21<F_{out}$，統計上不顯著(P-value = 0.1453 > 0.05)，去除 x_3^2 後由 Table 4.1.14 得知剩下之解釋變數 x_1, x_2, x_3, x_1^2 統計上皆顯著，由 ANOVA Table 4.1.15 得知，$R^2=0.7123$，$C_p=5.7972$；由 Table 4.1.16 得知依自變數的重要性依序去除 x_2^2, x_3^2。

3. 逐步選擇法(Stepwise regression)

此方式為 Forward 和 Backward 合併使用，發現選變數之過程與 Forward selection 相同，選入的變數亦相同，在此不做贅述，其結果呈現於 Table 4.1.17－4.1.25。

由逐步回歸選取法得知，建議模式為

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_3+\hat{\beta}_4x_1^2 \qquad (4.1.1)$$
$$=23.5358-11.1275x_1+0.5443x_2+3.3897x_3+1.5126x_1^2$$

而後利用所有回歸式的比較選取法中之選模指標(Table 4.1.26)：R^2_p, $R^2_{Adj,p}$, C_p, AIC, $MS_{Res}(p)$，得知 R^2_p=0.7123, $R^2_{Adj,p}$=0.6828, C_p=5.7972, AIC=105.3470, $MS_{Res}(p)$=9.8513 與 Figure 4.1.1－4.1.3 清楚看出，依參數精簡原則，所有回歸式的比較選取法所建議的模型與逐步回歸選取法所建議的模型一致，而且與模型 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_3$ (R^2_p=0.6455, $R^2_{Adj,p}$=0.6189, C_p=13.0510, AIC=112.5458, $MS_{Res}(p)$=11.8384)比較。此兩種方法均建議之模型 4.1.1 為較佳之模式。故在放入反應變數(y)與解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與其解釋變數之平方項下，其模型 4.1.1 為我們所選取之最佳回歸模型。

Table 4.1.1 Parameter estimates (Forward Selection: Step 1)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	16.3030	1.4128	2043.8377	133.15	<.0001
x_1^2	1	0.4812	0.0717	690.9604	45.01	<.0001

Table 4.1.2 Analysis of Variance (Forward Selection: Step 1)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	1	690.9604	690.9604	45.01	<.0001
Error	42	644.6849	15.3496		
Total	43	1335.6453			
R-Square		0.5173	C(p)		26.7792

Table 4.1.3 Parameter estimates (Forward Selection: Step 2)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	9.0606	2.6560	147.9408	11.64	0.0015
x_2	1	0.5259	0.1688	123.4553	9.71	0.0033
x_1^2	1	0.3084	0.0857	164.7976	12.96	0.0008

Table 4.1.4 Analysis of Variance (Forward Selection: Step 2)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	814.4157	407.2079	32.03	<.0001
Error	41	521.2296	12.7129		
Total	43	1335.6453			
R-Square		0.6098	C(p)		15.9912

Table 4.1.5 Parameter estimates (Forward Selection: Step 3)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	3.0397	3.4053	8.8920	0.80	0.3774
x_2	1	0.4864	0.1588	104.6439	9.38	0.0039
x_3	1	2.7053	1.0445	74.8636	6.71	0.0133
x_1^2	1	0.2657	0.0819	117.4015	10.52	0.0024

Table 4.1.6 Analysis of Variance (Forward Selection: Step 3)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	889.2793	296.4264	26.56	<.0001
Error	40	446.366	11.1592		
Total	43	1335.6453			
R-Square		0.6658	C(p)		10.2365

Table 4.1.7 Parameter estimates (Forward Selection: Step 4)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	23.5358	8.7641	71.0460	7.21	0.0106
x_1	1	-11.1275	4.4297	62.1645	6.31	0.0163
x_2	1	0.5443	0.1510	127.9864	12.99	0.0009
x_3	1	3.3897	1.0185	109.1184	11.08	0.0019
x_1^2	1	1.5126	0.5023	89.3353	9.07	0.0045

Table 4.1.8 Analysis of Variance (Forward Selection: Step 4)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	951.4438	237.8610	24.15	<.0001
Error	39	384.2015	9.8513		
Total	43	1335.6453			
R-Square		0.7123	C(p)		5.7972

Table 4.1.9 Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	P-value
1	x_1^2	1	0.5173	0.5173	26.7792	45.01	<.0001
2	x_2	2	0.0924	0.6098	15.9912	9.71	0.0033
3	x_3	3	0.0561	0.6658	10.2365	6.71	0.0133
4	x_1	4	0.0465	0.7123	5.7972	6.31	0.0163

Table 4.1.10 Parameter estimates (Backward Elimination: Step 0)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	6.0356	13.9553	1.8058	0.19	0.6679
x_1	1	-13.4205	4.5977	82.2551	8.52	0.0059
x_2	1	1.4838	1.1349	16.5024	1.71	0.1991
x_3	1	13.0829	7.6285	28.3945	2.94	0.0947
x_1^2	1	1.7350	0.5153	109.4280	11.34	0.0018
x_2^2	1	-0.0234	0.0299	5.8829	0.61	0.4400
x_3^2	1	-1.6284	1.3125	14.8592	1.54	0.2225

Table 4.1.11 Analysis of Variance (Backward Elimination: Step 0)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	6	978.4482	163.0747	16.89	<.0001
Error	37	357.1970	9.6540		
Total	43	1335.6453			
R-Square		0.7326	C(p)		7.0000

Table 4.1.12 Parameter estimates (Backward Elimination: Step 1)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	12.0304	11.5921	10.2909	1.08	0.3059
x_1	1	-13.0405	4.5483	78.5437	8.22	0.0067
x_2	1	0.6062	0.1544	147.2173	15.41	0.0004
x_3	1	14.3242	7.4225	35.5844	3.72	0.0611
x_1^2	1	1.6693	0.5058	104.0735	10.89	0.0021
x_3^2	1	-1.8813	1.2653	21.1215	2.21	0.1453

Table 4.1.13 Analysis of Variance (Backward Elimination: Step 1)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	5	972.5653	194.5131	20.36	<.0001
Error	38	363.0799	9.5547		
Total	43	1335.6453			
R-Square		0.7282	C(p)		5.6094

Table 4.1.14 Parameter estimates (Backward Elimination: Step 2)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	23.5358	8.7641	71.0460	7.21	0.0106
x_1	1	-11.1275	4.4297	62.1645	6.31	0.0163
x_2	1	0.5443	0.1510	127.9864	12.99	0.0009
x_3	1	3.3897	1.0185	109.1184	11.08	0.0019
x_1^2	1	1.5126	0.5023	89.3353	9.07	0.0045

Table 4.1.15 Analysis of Variance (Backward Elimination: Step 2)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	951.4438	237.8610	24.15	<.0001
Error	39	384.2015	9.8513		
Total	43	1335.6453			
R-Square		0.7123	C(p)		5.7972

Table 4.1.16 Summary of Backward Elimination

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	P-value
1	x_2^2	5	0.0044	0.7282	5.6094	0.6100	0.4400
2	x_3^2	4	0.0158	0.7123	5.7972	2.2100	0.1453

Table 4.1.17 Parameter estimates (Stepwise Selection: Step 1)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	16.3030	1.4128	2043.8377	133.15	<.0001
x_1^2	1	0.4812	0.0717	690.9604	45.01	<.0001

Table 4.1.18 Analysis of Variance (Stepwise Selection: Step 1)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	1	690.9604	690.9604	45.01	<.0001
Error	42	644.6849	15.3496		
Total	43	1335.6453			
R-Square		0.5173	C(p)		26.7792

Table 4.1.19 Parameter estimates (Stepwise Selection: Step 2)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	9.0606	2.6560	147.9408	11.64	0.0015
x_2	1	0.5259	0.1688	123.4553	9.71	0.0033
x_1^2	1	0.3084	0.0857	164.7976	12.96	0.0008

Table 4.1.20 Analysis of Variance (Stepwise Selection: Step 2)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	814.4157	407.2079	32.03	<.0001
Error	41	521.2296	12.7129		
Total	43	1335.6453			
R-Square		0.6098	C(p)		15.9912

Table 4.1.21 Parameter estimates (Stepwise Selection: Step 3)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	3.0397	3.4053	8.8920	0.80	0.3774
x_2	1	0.4864	0.1588	104.6439	9.38	0.0039
x_3	1	2.7054	1.0445	74.8636	6.71	0.0133
x_1^2	1	0.2657	0.0819	117.4015	10.52	0.0024

Table 4.1.22 Analysis of Variance (Stepwise Selection: Step 3)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	889.2793	296.4264	26.56	<.0001
Error	40	446.366	11.1592		
Total	43	1335.6453			
R-Square		0.6658	C(p)		10.2365

Table 4.1.23 Parameter estimates (Stepwise Selection: Step 4)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	23.5358	8.7641	71.0460	7.21	0.0106
x_1	1	-11.1275	4.4297	62.1645	6.31	0.0163
x_2	1	0.5443	0.1510	127.9864	12.99	0.0009
x_3	1	3.3897	1.0185	109.1184	11.08	0.0019
x_1^2	1	1.5126	0.5023	89.3353	9.07	0.0045

Table 4.1.24 Analysis of Variance (Stepwise Selection: Step 4)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	951.4438	237.8610	24.15	<.0001
Error	39	384.2015	9.8513		
Total	43	1335.6453			
R-Square		0.7123	C(p)		5.7972

Table 4.1.25 Summary of Stepwise Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	P-value
1	x_1^2	1	0.5173	0.5173	26.7792	45.01	<.0001
2	x_2	2	0.0924	0.6098	15.9912	9.71	0.0033
3	x_3	3	0.0561	0.6658	10.2365	6.71	0.0133
4	x_1	4	0.0465	0.7123	5.7972	6.31	0.0163

Table 4.1.26.1 Summary of All Possible Regressions

Number of Regressors in Model	Regressors in Model	R_p^2	$R_{Adj,p}^2$	C_p	AIC	$MS_{Res}(p)$
1	x_1^2	0.5173	0.5058	26.7792	122.1212	15.3496
1	x_2^2	0.4957	0.4837	29.7749	124.0520	16.0382
1	x_1	0.4951	0.4831	29.8561	124.1032	16.0569
1	x_2	0.4864	0.4741	31.0616	124.8560	16.3340
1	x_3	0.2375	0.2194	65.4866	142.2376	24.2468
1	x_3^2	0.2009	0.1819	70.5564	144.3031	25.4121
2	$x_2\ x_1^2$	0.6098	0.5907	15.9912	114.7680	12.7129
2	$x_1^2\ x_2^2$	0.6081	0.5889	16.2259	114.9589	12.7682
2	$x_1\ x_2$	0.5918	0.5719	18.4729	116.7454	13.2973
2	$x_1\ x_2^2$	0.5912	0.5713	18.5583	116.8119	13.3174
2	$x_3\ x_1^2$	0.5875	0.5673	19.0760	117.2128	13.4393
2	$x_1^2\ x_3^2$	0.5864	0.5662	19.2230	117.3260	13.4739
2	$x_2\ x_3$	0.5779	0.5573	20.3975	118.2199	13.7504
2	$x_3\ x_2^2$	0.5760	0.5553	20.6581	118.4158	13.8118
2	$x_2\ x_3^2$	0.5599	0.5384	22.8900	120.0589	14.3373
2	$x_2^2\ x_3^2$	0.5584	0.5369	23.0907	120.2037	14.3846
2	$x_1\ x_3$	0.5584	0.5369	23.0925	120.2051	14.3850
2	$x_1\ x_3^2$	0.5578	0.5362	23.1856	120.2720	14.4069
2	$x_1\ x_1^2$	0.5254	0.5023	27.6612	123.3783	15.4608

Table 4.1.26.2 Summary of All Possible Regressions

Number of Regressors in Model	Regressors in Model	R_p^2	$R_{Adj,p}^2$	C_p	AIC	$MS_{Res}(p)$
2	$x_2\ x_2^2$	0.4962	0.4717	31.6965	126.0025	16.4109
2	$x_3\ x_3^2$	0.2912	0.2566	60.0676	141.0288	23.0913
3	$x_2\ x_3\ x_1^2$	0.6658	0.6407	10.2365	109.9458	11.1592

3	$x_2\ x_1^2\ x_3^2$	0.6610	0.6356	10.9037	110.5762	11.3202
3	$x_3\ x_1^2\ x_2^2$	0.6602	0.6348	11.0070	110.6729	11.3451
3	$x_1^2\ x_2^2\ x_3^2$	0.6552	0.6293	11.7092	111.3254	11.5146
3	$x_1\ x_2\ x_3$	0.6455	0.6189	13.0510	112.5458	11.8384
3	$x_1\ x_3\ x_2^2$	0.6407	0.6138	13.7039	113.1276	11.9960
3	$x_1\ x_2\ x_3^2$	0.6403	0.6133	13.7682	113.1845	12.0115
3	$x_1\ x_2^2\ x_3^2$	0.6353	0.6080	14.4537	113.7864	12.1770
3	$x_1\ x_2\ x_1^2$	0.6307	0.6029	15.1002	114.3466	12.3330
3	$x_1\ x_1^2\ x_2^2$	0.6273	0.5994	15.5592	114.7401	12.4438
3	$x_2\ x_3\ x_1^2$	0.6165	0.5878	17.0546	115.9981	12.8047
3	$x_3\ x_2^2\ x_3^2$	0.6134	0.5844	17.4851	116.3537	12.9086
3	$x_2\ x_3\ x_3^2$	0.6118	0.5827	17.7015	116.5314	12.9608
3	$x_1\ x_1^2\ x_3^2$	0.6116	0.5825	17.7321	116.5564	12.9682
3	$x_2\ x_1^2\ x_2^2$	0.6098	0.5805	17.9911	116.7679	13.0307
3	$x_1\ x_2\ x_2^2$	0.5920	0.5614	20.4421	118.7214	13.6223
3	$x_3\ x_1^2\ x_3^2$	0.5875	0.5565	21.0736	119.2109	13.7747
3	$x_2\ x_3\ x_2^2$	0.5782	0.5466	22.3578	120.1900	14.0846
3	$x_2\ x_2^2\ x_3^2$	0.5603	0.5274	24.8285	122.0145	14.6809
3	$x_1\ x_3\ x_3^2$	0.5585	0.5254	25.0838	122.1987	14.7425
4	$x_1\ x_2\ x_3\ x_1^2$	0.7123	0.6828	5.7972	105.3470	9.8513
4	$x_1\ x_3\ x_1^2\ x_2^2$	0.7029	0.6724	7.1050	106.7696	10.1750
4	$x_1\ x_2\ x_1^2\ x_3^2$	0.7015	0.6709	7.2954	106.9729	10.2222

4	$x_1\ x_1^2\ x_2^2\ x_3^2$	0.6922	0.6606	8.5848	108.3258	10.5413
4	$x_2\ x_3\ x_1^2\ x_3^2$	0.6694	0.6354	11.7453	111.4758	11.3237
4	$x_2\ x_3\ x_1^2\ x_2^2$	0.6689	0.6349	11.8096	111.5376	11.3396
4	$x_2\ x_1^2\ x_2^2\ x_3^2$	0.6648	0.6304	12.3730	112.0755	11.4791
4	$x_3\ x_1^2\ x_2^2\ x_3^2$	0.6648	0.6304	12.3731	112.0756	11.4791
4	$x_1\ x_2\ x_3\ x_3^2$	0.6502	0.6144	14.3898	113.9486	11.9783
4	$x_1\ x_2\ x_3\ x_2^2$	0.6469	0.6107	14.8464	114.3618	12.0913
4	$x_1\ x_3\ x_2^2\ x_3^2$	0.6469	0.6106	14.8589	114.3731	12.0944
4	$x_1\ x_2\ x_2^2\ x_3^2$	0.6422	0.6055	15.4997	114.9465	12.2531
4	$x_1\ x_2\ x_1^2\ x_2^2$	0.6309	0.5930	17.0721	116.3224	12.6423

Table 4.1.26.3 Summary of All Possible Regressions

Number of Regressors in Model	Regressors in Model	R_p^2	$R_{Adj,p}^2$	C_p	AIC	$MS_{Res}(p)$
4	$x_1\ x_3\ x_1^2\ x_3^2$	0.6179	0.5788	18.8588	117.8354	13.0846
4	$x_2\ x_3\ x_2^2\ x_3^2$	0.6137	0.5741	19.4408	118.3172	13.2286
5	$x_1\ x_2\ x_3\ x_1^2\ x_3^2$	0.7282	0.6924	5.6094	104.8591	9.5547
5	$x_1\ x_2\ x_3\ x_1^2\ x_2^2$	0.7214	0.6848	6.5392	105.9336	9.7910
5	$x_1\ x_3\ x_1^2\ x_2^2\ x_3^2$	0.7202	0.6834	6.7094	106.1275	9.8342
5	$x_1\ x_2\ x_1^2\ x_2^2\ x_3^2$	0.7113	0.6733	7.9412	107.5059	10.1471
5	$x_2\ x_3\ x_1^2\ x_2^2\ x_3^2$	0.6710	0.6277	13.5203	113.2589	11.5645
5	$x_1\ x_2\ x_3\ x_2^2\ x_3^2$	0.6506	0.6047	16.3350	115.8988	12.2796

6	$x_1\ x_2\ x_3\ x_1^2\ x_2^2\ x_3^2$	0.7326	0.6892	7.0000	106.1403	9.6540

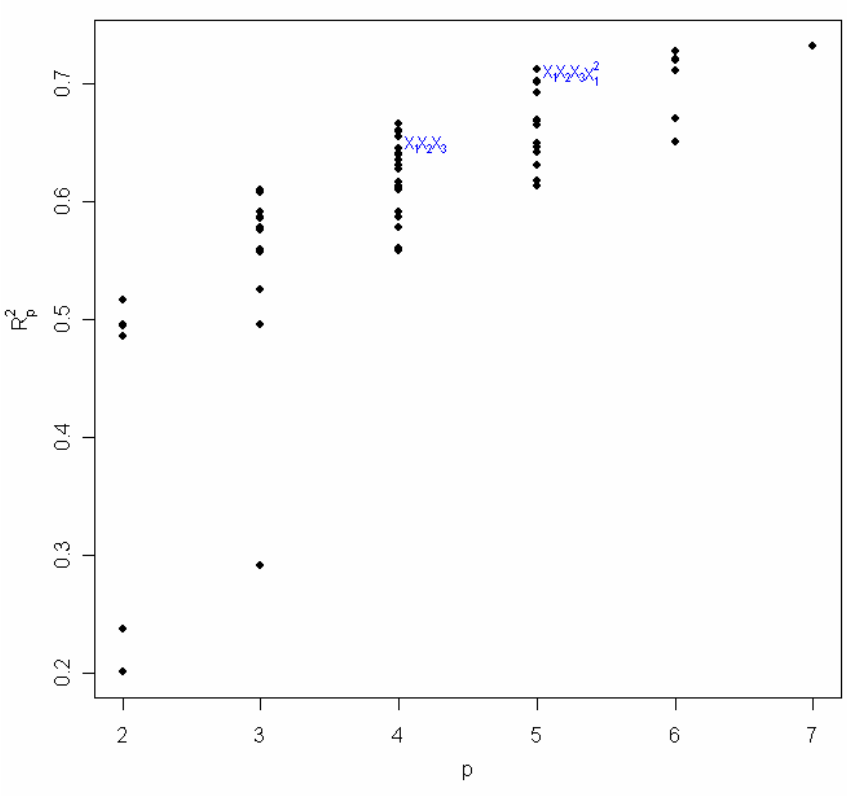


Figure 4.1.1 Plot R_p^2 versus p

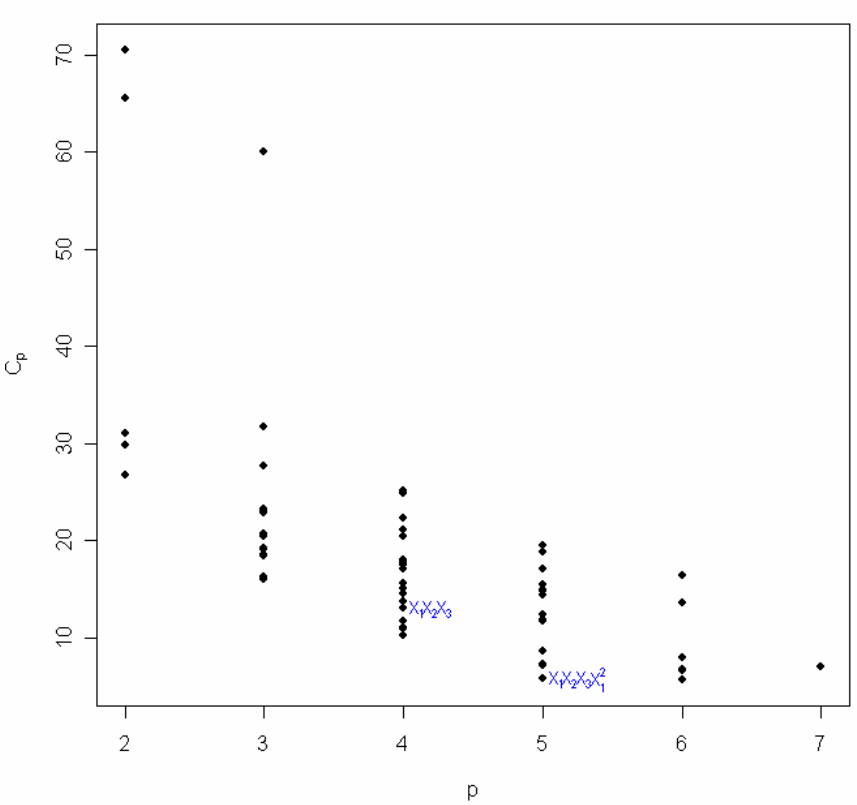


Figure 4.1.2 The C_p plot

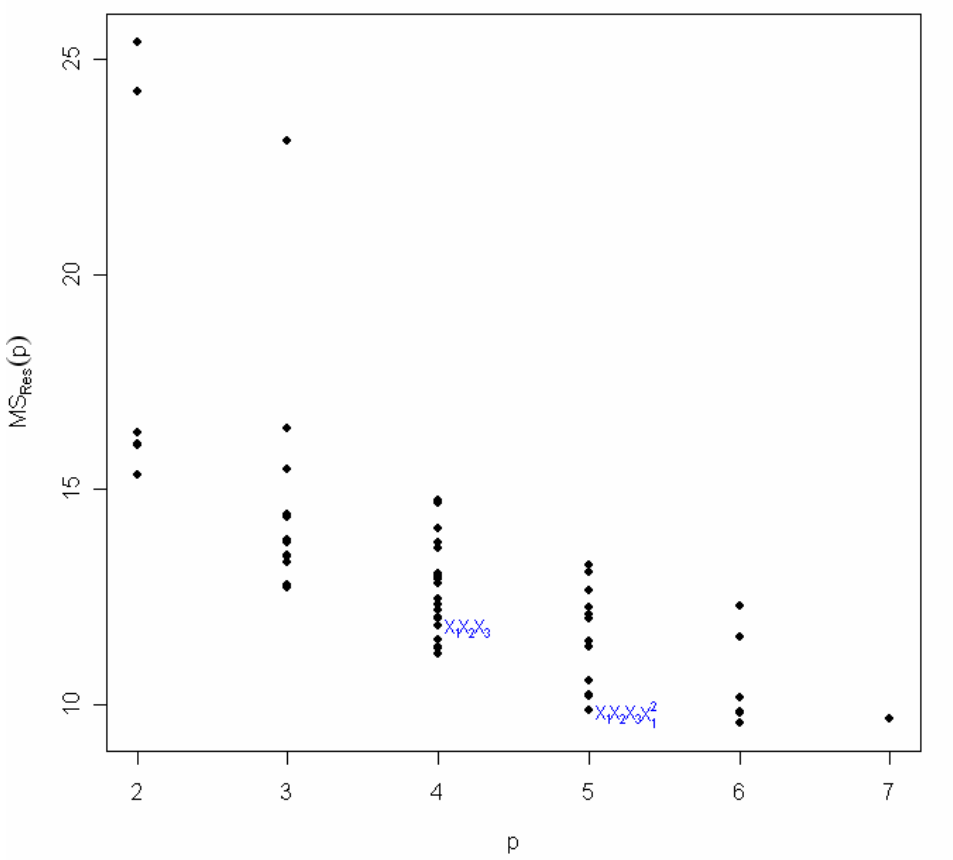


Figure 4.1.3 Plot $MS_{Res}(p)$ versus p

第二節

本節為考慮反應變數經轉換後($y^{(-0.22)}$)與解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與其解釋變數之平方項，利用逐步回歸分析(Stepwise regression methods)來選擇較佳模型之組合。而逐步回歸包含以下三種方法：1.向前選擇法(Forward selection) 2. 向後消去法(Backward elimination) 3. 逐步選擇法(Stepwise regression)。

假設考慮之解釋變數為 $X_k = \{x_1, x_2, x_3, x_1^2, x_2^2, x_3^2\}$，k = 1, …, 6；以下為說明逐步回歸選取法之結果：($\alpha = 0.05, F_{in} = F_{out} \approx 4$)

1. 向前選擇法(Forward selection)

此法一開始假設模式中沒有任何解釋變數，而後第一步驟選入之變數為 x_2(k = 2)，由 Table 4.2.1 得知此解釋變數之 $F_2^* = 44.65 > F_{in}$，統計上顯著(P-value < 0.05)，由 ANOVA Table4.2.2 得知，$R^2 = 0.5153$，$C_p = 34.0582$；第二步驟選入之變數為 x_3(k = 3)，由 Table 4.2.3 得知此解釋變數之 $F_{3|2}^* = 14.26 > F_{in}$，統計上顯著(P-value = 0.0005 < 0.05)，由 ANOVA Table 4.2.4 得知，$R^2 = 0.6404$，$C_p = 16.9496$；第三步驟選入之變數為 x_1^2(k = 4)，由 Table 4.2.5 得知

此解釋變數之 $F^*_{4|2,3} = 7.02 > F_{in}$，統計上顯著(P-value = 0.0115 < 0.05)，由 Table 4.2.6 得知，R^2 = 0.6940，C_p = 10.7462;由 Table 4.2.7 得知依解釋變數的重要性逐漸選入的優先順序為 x_2, x_3, x_1^2。

2. 向後消去法(Backward elimination)

此法一開始假設模式中包括所有解釋變數，第一步驟去除之變數為 x_2^2 (k = 5)由 Table 4.2.8 得知解釋變數之 $F_5^* = 2.01 < F_{out}$，統計上不顯著(P-value = 0.1649 > 0.05)，由 Table 4.2.10 得知剩下之解釋變數 $x_1, x_2, x_3, x_1^2, x_3^2$ 統計上皆顯著，由 ANOVA Table 4.2.11 得知，R^2 = 0.7447，C_p = 7.0072；由 Table 4.2.12 得知依解釋變數的重要性去除 x_2^2。

3. 逐步選擇法(Stepwise regression)

此方式為 Forward 和 Backward 合併使用，發現選變數之過程與 Forward selection 相同，選入的變數亦相同，在此不做贅述，其結果呈現於 Table 4.2.13－4.2.19 中。

由逐步回歸選取法得知，建議模式為

$$\hat{y}^{(-0.22)} = \hat{\beta}_0 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_1^2 \tag{4.2.1}$$

然而轉換後的資料配適模式會傾向次方較高的模式，所以我們保留低次方項修改模型為

$$\hat{y}^{(-0.22)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_1^2 \tag{4.2.2}$$

結果由先前參數估計表(Table 3.4.7)得知，轉換後的資料 y 加入解釋變數 x_1 的一次項與平方項不顯著(P-value > α = 0.05)，故資料經轉換後不適合加入二次項，而後利用所有回歸式的比較選取法中之選模指標(Table 4.2.20)：$R_p^2, R_{Adj,p}^2, C_p, AIC, MS_{Res}(p)$，與 Figure 4.2.1－4.2.3 清楚看出，依參數精簡原則，所有回歸式的比較選取法所建議的模型與逐步回歸選取法所建議的模型一致，在此經轉換後的資料 y 之回歸模式，我們以

$$\begin{aligned}\hat{y}^{(-0.22)} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 \\ &= 2.3987 + 0.6308 x_2 + 3.41 x_3 \\ (R_p^2 &= 0.6843, R_{Adj,p}^2 = 0.6606, C_p = \ \ 2.2328)\end{aligned} \tag{4.2.3}$$

為較佳模式，然而與前模型(4.1.1)比較，綜合選模指標得知，對 y 轉換後的模型 4.2.3 為我們所選取之最佳模型；若考慮包含平方項則以不對 y 轉換的模型 4.1.1 較佳。故在放入反應變數經轉換後($y^{(-0.22)}$)與解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與其解釋變數之平方項下，其模

型 4.2.3 為我們所選取之最佳回歸模型。

Table 4.2.1 Parameter estimates (Forward Selection: Step 1)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	0.5749	0.0121	0.6293	2268.34	<0.0001
x_2	1	-0.0040	0.0006	0.0124	44.65	<0.0001

Table 4.2.2 Analysis of Variance (Forward Selection: Step 1)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	1	0.0124	0.0124	44.65	<.0001
Error	42	0.0117	0.0003		
Total	43	0.0240			
R-Square		0.5153	C(p)		34.0582

Table 4.2.3 Parameter estimates (Forward Selection: Step 2)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	0.6104	0.0141	0.3948	1872.02	<0.0001
x_2	1	-0.0034	0.0005	0.0083	39.49	<0.0001
x_3	1	-0.0168	0.0045	0.0030	14.26	0.0005

Table 4.2.4 Analysis of Variance (Forward Selection: Step 2)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	0.0154	0.0077	36.5	<.0001
Error	41	0.0087	0.0002		
Total	43	0.0240			
R-Square		0.6404	C(p)		16.9496

Table 4.2.5 Parameter estimates (Forward Selection: Step 3)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	0.5993	0.0138	0.3456	1879.59	<0.0001
x_2	1	-0.0024	0.0006	0.0025	13.69	0.0006
x_3	1	-0.0145	0.0042	0.0022	11.75	0.0014
x_1^2	1	-0.0009	0.0003	0.0013	7.02	0.0115

Table 4.2.6 Analysis of Variance (Forward Selection: Step 3)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	0.0167	0.0056	30.25	<.0001
Error	40	0.0074	0.0002		
Total	43	0.0240			
R-Square		0.6940	C(p)		10.7462

Table 4.2.7 Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	P-value
1	x_2	1	0.5153	0.5153	34.0582	44.65	<.0001
2	x_3	2	0.1251	0.6404	16.9496	14.26	0.0005
3	x_1^2	3	0.0537	0.6940	10.7462	7.020	0.0115

Table 4.2.8 Parameter estimates (Backward Elimination: Step 0)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	0.6648	0.0563	0.0219	139.25	<0.0001
x_1	1	0.0424	0.0186	0.0008	5.22	0.0282
x_2	1	-0.0094	0.0046	0.0007	4.16	0.0485
x_3	1	-0.0759	0.0308	0.0010	6.08	0.0184
x_1^2	1	-0.0054	0.0021	0.0011	6.83	0.0129
x_2^2	1	0.0002	0.0001	0.0003	2.01	0.1649
x_3^2	1	0.0100	0.0053	0.0006	3.55	0.0674

Table 4.2.9 Analysis of Variance (Backward Elimination: Step 0)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	6	0.0182	0.0030	19.3	<.0001
Error	37	0.0058	0.0002		
Total	43	0.0240			
R-Square		0.7578	C(p)		7.0000

Table 4.2.10 Parameter estimates (Backward Elimination: Step 1)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	0.6209	0.0477	0.0274	169.71	<0.0001
x_1	1	0.0396	0.0187	0.0007	4.49	0.0408
x_2	1	-0.0029	0.0006	0.0034	21.13	<0.0001
x_3	1	-0.0850	0.0305	0.0013	7.76	0.0083
x_1^2	1	-0.0050	0.0021	0.0009	5.68	0.0223
x_3^2	1	0.0118	0.0052	0.0008	5.18	0.0286

Table 4.2.11 Analysis of Variance (Backward Elimination: Step 1)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	5	0.0179	0.0036	22.17	<.0001
Error	38	0.0061	0.0002		
Total	43	0.0240			
R-Square		0.7447	C(p)		7.0072

Table 4.2.12 Summary of Backward Elimination

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	P-value
1	x_2^2	5	0.0131	0.7447	7.00724	2.01	0.1649

Table 4.2.13 Parameter estimates (Stepwise Selection: Step 1)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	0.5749	0.0121	0.6293	2268.34	<0.0001
x_2	1	-0.0040	0.0006	0.0124	44.65	<0.0001

Table 4.2.14 Analysis of Variance (Stepwise Selection: Step 1)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	1	0.0124	0.0124	44.65	<.0001
Error	42	0.0117	0.0003		
Total	43	0.0240			
R-Square		0.5153	C(p)		34.0582

Table 4.2.15 Parameter estimates (Stepwise Selection: Step 2)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	0.6104	0.0141	0.3948	1872.02	<0.0001
x_2	1	-0.0034	0.0005	0.0083	39.49	<0.0001
x_3	1	-0.0168	0.0045	0.0030	14.26	0.0005

Table 4.2.16 Analysis of Variance (Stepwise Selection: Step 2)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	0.0154	0.0077	36.5	<.0001
Error	41	0.0087	0.0002		
Total	43	0.0240			
R-Square		0.6404	C(p)		16.9496

Table 4.2.17 Parameter estimates (Stepwise Selection: Step 3)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	0.5992	0.0138	0.3456	1879.59	<0.0001
x_2	1	-0.0024	0.0006	0.0025	13.69	0.0006
x_3	1	-0.0145	0.0042	0.0022	11.75	0.0014
x_1^2	1	-0.0009	0.0003	0.0013	7.02	0.0115

Table 4.2.18 Analysis of Variance (Stepwise Selection: Step 3)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	0.0167	0.0056	30.25	<.0001
Error	40	0.0074	0.0002		
Total	43	0.0240			
R-Square		0.6940	C(p)		10.7462

Table 4.2.19 Summary of Stepwise Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	P-value
1	x_2	1	0.5153	0.5153	34.0582	44.65	<.0001
2	x_3	2	0.1251	0.6404	16.9496	14.26	0.0005
3	x_1^2	3	0.0537	0.6940	10.7462	7.020	0.0115

Table 4.2.20.1 Summary of All Possible Regressions

Number of Regressors in Model	Regressors in Model	R_p^2	$R_{Adj,p}^2$	C_p	AIC	$MS_{Res}(p)$
1	x_2	0.5153	0.5038	34.0582	-358.4039	0.0002774
1	x_2^2	0.5135	0.5019	34.3368	-358.2386	0.0002785
1	x_1^2	0.4788	0.4664	39.6311	-355.2115	0.0002983
1	x_1	0.4755	0.4630	40.1397	-354.9314	0.0003002
1	x_3	0.2940	0.2771	67.8764	-341.8539	0.0004041
1	x_3^2	0.2456	0.2276	75.2664	-338.9385	0.0004318
2	$x_2\ x_3$	0.6404	0.6228	16.9496	-369.5350	0.0002109
2	$x_3\ x_2^2$	0.6270	0.6088	18.9843	-367.9352	0.0002187
2	$x_2\ x_3^2$	0.6143	0.5955	20.9356	-366.4538	0.0002262
2	$x_2\ x_1^2$	0.6042	0.5848	22.4813	-365.3147	0.0002321
2	$x_2^2\ x_3^2$	0.6012	0.5818	22.9296	-364.9897	0.0002338
2	$x_1\ x_2$	0.5983	0.5787	23.3714	-364.6718	0.0002355
2	$x_1^2\ x_2^2$	0.5961	0.5764	23.7141	-364.4268	0.0002368
2	$x_1\ x_2^2$	0.5910	0.5711	24.4854	-363.8803	0.0002398
2	$x_3\ x_1^2$	0.5893	0.5693	24.7482	-363.6956	0.0002408
2	$x_1^2\ x_3^2$	0.5810	0.5606	26.0140	-362.8169	0.0002457
2	$x_1\ x_3$	0.5752	0.5545	26.9036	-362.2096	0.0002491
2	$x_1\ x_3^2$	0.5681	0.5470	27.9906	-361.4788	0.0002532
2	$x_2\ x_2^2$	0.5160	0.4924	35.9480	-356.4694	0.0002838

Table 4.2.20.2 Summary of All Possible Regressions

Number of Regressors in Model	Regressors in Model	R_p^2	$R_{Adj,p}^2$	C_p	AIC	$MS_{Res}(p)$
2	$x_1\ x_1^2$	0.4792	0.4538	41.5666	-353.2472	0.0003053
2	$x_3\ x_3^2$	0.3714	0.3408	58.0388	-344.9682	0.0003686
3	$x_2\ x_3\ x_1^2$	0.6940	0.6711	10.7462	-374.6491	0.000184
3	$x_2\ x_3\ x_3^2$	0.6936	0.6706	10.8146	-374.5847	0.000184
3	$x_3\ x_2^2\ x_3^2$	0.6849	0.6613	12.1428	-373.3538	0.000189
3	$x_1\ x_2\ x_3$	0.6843	0.6606	12.2328	-373.2716	0.000190
3	$x_2\ x_1^2\ x_3^2$	0.6811	0.6572	12.7291	-372.8212	0.000192
3	$x_3\ x_1^2\ x_2^2$	0.6809	0.6570	12.7507	-372.8016	0.000192
3	$x_1\ x_3\ x_2^2$	0.6717	0.6471	14.1558	-371.5514	0.000197

3	$x_1\ x_2\ x_3^2$	0.6714	0.6468	14.2008	-371.5120	0.000197
3	$x_1^2\ x_2^2\ x_3^2$	0.6679	0.6430	14.7454	-371.0372	0.000200
3	$x_1\ x_2^2\ x_3^2$	0.6589	0.6333	16.1231	-369.8586	0.000205
3	$x_2\ x_3\ x_2^2$	0.6457	0.6191	18.1345	-368.1926	0.000213
3	$x_2\ x_2^2\ x_3^2$	0.6189	0.5903	22.2299	-364.9838	0.000229
3	$x_2\ x_1^2\ x_2^2$	0.6076	0.5782	23.9480	-363.7043	0.000236
3	$x_1\ x_2\ x_1^2$	0.6059	0.5763	24.2219	-363.5037	0.000237
3	$x_1\ x_2\ x_2^2$	0.6009	0.5709	24.9857	-362.9492	0.000240
3	$x_1\ x_1^2\ x_2^2$	0.5972	0.5670	25.5421	-362.5496	0.000242
3	$x_1\ x_3\ x_1^2$	0.5950	0.5646	25.8862	-362.3043	0.000243
3	$x_3\ x_1^2\ x_3^2$	0.5939	0.5634	26.0483	-362.1892	0.000244
3	$x_1\ x_1^2\ x_3^2$	0.5845	0.5534	27.4769	-361.1876	0.000250
3	$x_1\ x_3\ x_3^2$	0.5789	0.5473	28.3458	-360.5893	0.000253
4	$x_2\ x_3\ x_1^2\ x_3^2$	0.7146	0.6853	9.6115	-375.7032	0.000176
4	$x_2\ x_3\ x_1^2\ x_2^2$	0.7115	0.6820	10.0739	-375.2392	0.000178
4	$x_1\ x_2\ x_3\ x_1^2$	0.7099	0.6802	10.3211	-374.9930	0.000179
4	$x_1\ x_2\ x_3\ x_3^2$	0.7065	0.6764	10.8371	-374.4837	0.000181
4	$x_3\ x_1^2\ x_2^2\ x_3^2$	0.7037	0.6733	11.2757	-374.0555	0.000183
4	$x_2\ x_1^2\ x_2^2\ x_3^2$	0.7002	0.6695	11.8048	-373.5442	0.000185
4	$x_1\ x_2\ x_3\ x_2^2$	0.6986	0.6677	12.0442	-373.3149	0.000186
4	$x_1\ x_3\ x_2^2\ x_3^2$	0.6963	0.6652	12.4025	-372.9737	0.000187
4	$x_2\ x_3\ x_2^2\ x_3^2$	0.6953	0.6640	12.5601	-372.8246	0.000188
4	$x_1\ x_3\ x_1^2\ x_2^2$	0.6941	0.6627	12.7438	-372.6513	0.000189

Table 4.2.20.3 Summary of All Possible Regressions

Number of Regressors in Model	Regressors in Model	R_p^2	$R_{Adj,p}^2$	C_p	AIC	$MS_{Res}(p)$
4	$x_1\ x_2\ x_1^2\ x_3^2$	0.6925	0.6610	12.9760	-372.4333	0.000190
4	$x_1\ x_2\ x_2^2\ x_3^2$	0.6871	0.6550	13.8072	-371.6616	0.000193
4	$x_1\ x_1^2\ x_2^2\ x_3^2$	0.6771	0.6440	15.3329	-370.2793	0.000199
4	$x_1\ x_2\ x_1^2\ x_2^2$	0.6099	0.5699	25.6015	-361.9594	0.000240
4	$x_1\ x_3\ x_1^2\ x_3^2$	0.6027	0.5620	26.6988	-361.1567	0.000245
5	$x_1\ x_2\ x_3\ x_1^2\ x_3^2$	0.7447	0.7111	7.0072	-378.6125	0.000162
5	$x_1\ x_2\ x_3\ x_1^2\ x_2^2$	0.7346	0.6997	8.5497	-376.9060	0.000168
5	$x_1\ x_3\ x_1^2\ x_2^2\ x_3^2$	0.7306	0.6951	9.1636	-376.2449	0.000170
5	$x_2\ x_3\ x_1^2\ x_2^2\ x_3^2$	0.7237	0.6873	10.2153	-375.1349	0.000175

5	$x_1\ x_2\ x_1^2\ x_2^2\ x_3^2$	0.7181	0.6810	11.0791	-374.2437	0.000178
5	$x_1\ x_2\ x_3\ x_2^2\ x_3^2$	0.7131	0.6754	11.8302	-373.4831	0.000181
6	$x_1\ x_2\ x_3\ x_1^2\ x_2^2\ x_3^2$	0.7578	0.7186	7.0000	-378.9369	0.000157

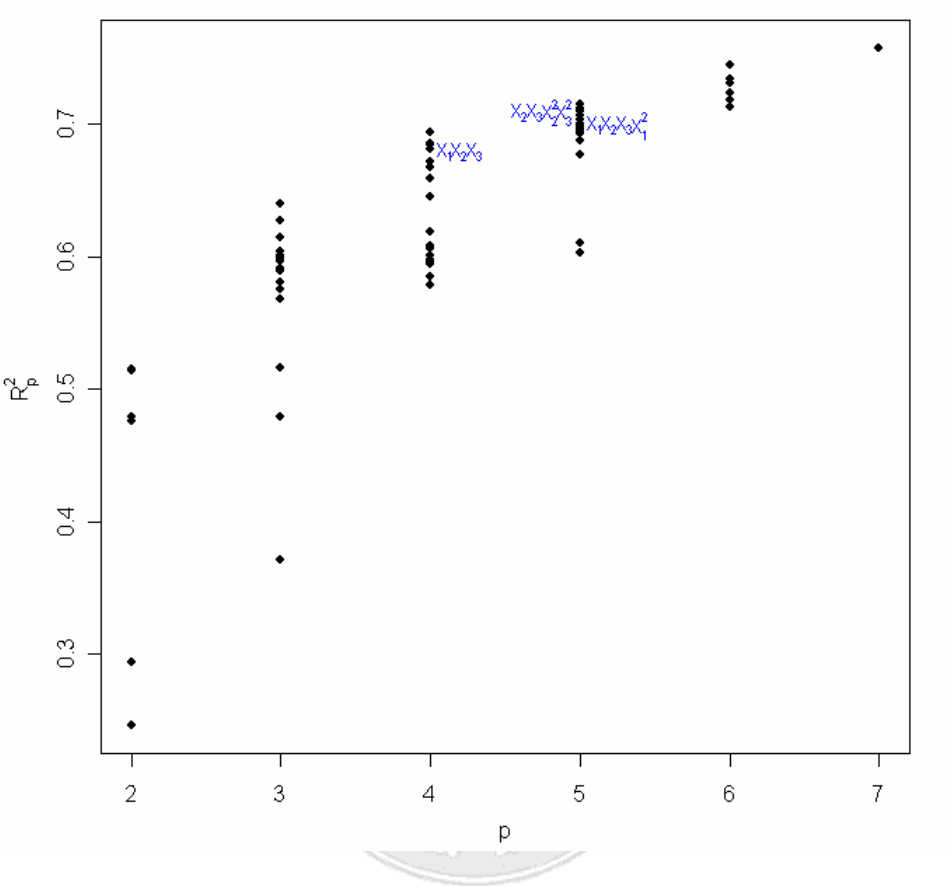


Figure 4.2.1 Plot R_p^2 versus p

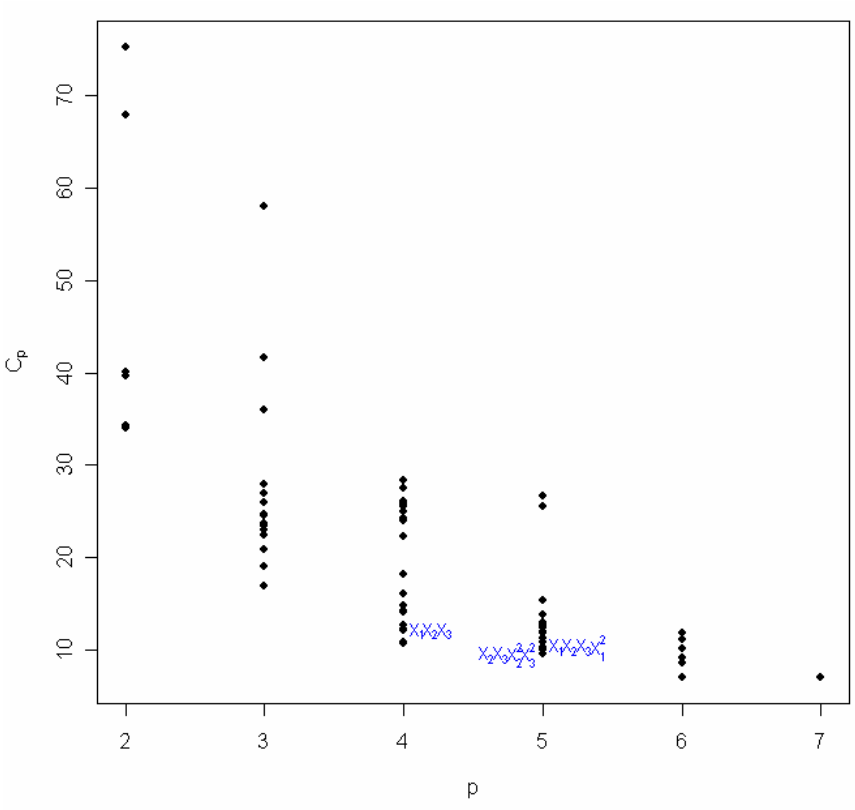


Figure 4.2.2 The C_p plot

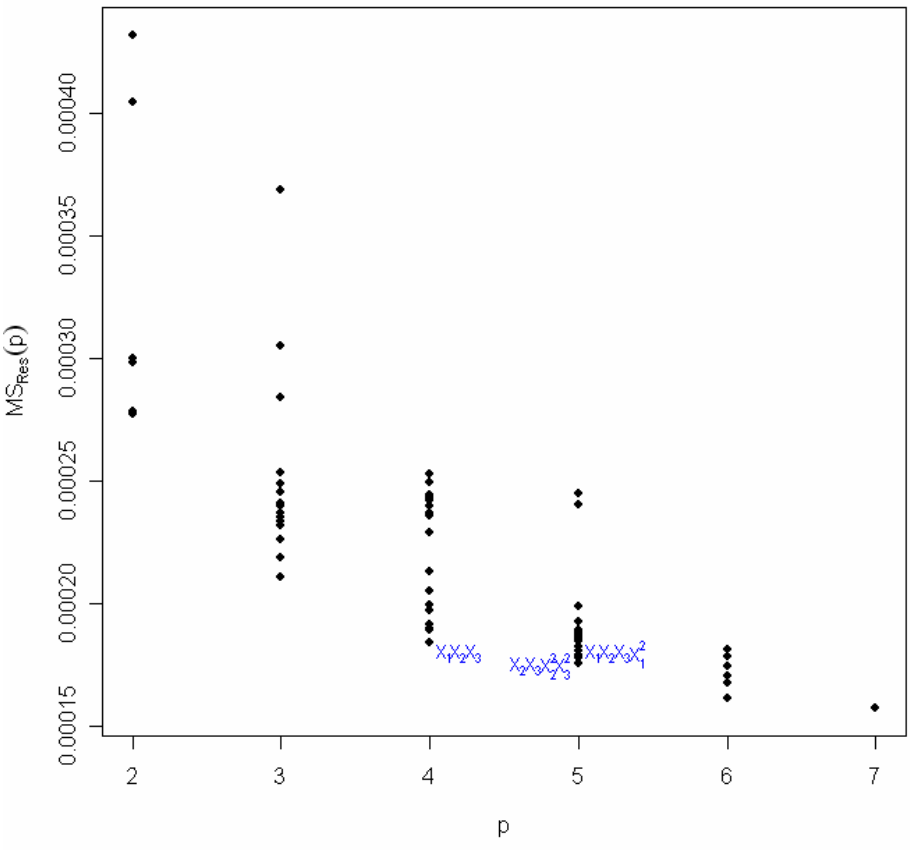


Figure 4.2.3 Plot $MS_{Res}(p)$ versus p

第三節

本節為考慮去除影響點之資料，反應變數後y^*與解釋變數膀胱癌(x_1^*)、肺癌(x_2^*)、腎癌(x_3^*)與解釋變數膀胱癌(x_1^*)之平方項，利用逐步回歸分析(Stepwise regression methods)來選擇較佳模型之組合。而逐步回歸包含以下三種方法：1.向前選擇法(Forward selection) 2. 向後消去法(Backward elimination) 3. 逐步選擇法(Stepwise regression)。

假設考慮之解釋變數為$X_k = \{x_1^*, x_2^*, x_3^*, (x_1^*)^2\}$， k = 1, …, 4；以下先說明逐步回歸選取法之結果： $(\alpha = 0.05, F_{in} = F_{out} \approx 4)$

1. 向前選擇法(Forward selection)

此法一開始假設模式中沒有任何解釋變數，而後第一步驟選入之變數為x_2^* (k = 2)，由 Table 4.3.1 得知此解釋變數之$F_2^* = 41.71 > F_{in}$，統計上顯著(P-value < 0.05)，由 ANOVA Table4.3.2 得知，$R^2 = 0.5105$，$C_p = 19.2627$；第二步驟選入之變數為x_3^* (k = 3)，由 Table 4.3.3 得知此解釋變數之$F_{3|2}^* = 19.38 > F_{in}$，統計上顯著(P-value < 0.05)，由 ANOVA Table4.3.4 得知，$R^2 = 0.6730$，$C_p = 2.2561$；由 Table 4.3.5 得知依自變數的重要性逐漸選入的優先順序為x_2^*, x_3^*。

2. 向後消去法(Backward elimination)

此法一開始假設模式中包括所有解釋變數，第一步驟去除之變數為$(x_1^*)^2$(k = 4)由 Table 4.3.6 得知解釋變數之$F_4^* = 0.31 < F_{out}$，統計上不顯著(P-value = 0.5795 > 0.05)，去除$(x_1^*)^2$後由 ANOVA Table 4.3.9 得知，$R^2 = 0.6810$，$C_p = 3.3126$；第二步驟去除之變數為x_1^* (k = 1)由 Table 4.3.8 得知解釋變數之$F_{1|4}^* = 0.96 < F_{out}$，統計上不顯著(P-value = 0.3332 > 0.05)，由 Table 4.3.10 得知剩下之解釋變數x_2^*, x_3^*統計上皆顯著，由 ANOVA Table 4.3.11 得知，R^2 = 0.6730，C_p = 2.2561；由 Table 4.3.12 得知依自變數的重要性依序去除$(x_1^*)^2, x_1^*$。

3. 逐步選擇法(Stepwise regression)

此方式為 Forward 和 Backward 合併使用，發現選變數之過程與 Forward selection 相同，選入的變數亦相同，在此不做贅述，其結果呈現於 Table 4.3.13－4.3.17。

由逐步回歸選取法得知，建議模式為

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_2 x_2^* + \hat{\beta}_3 x_3^* \qquad (4.3.1)$$

$$= 2.3987 + 0.6308 x_2^* + 3.41 x_3^*$$

而後利用所有回歸式的比較選取法中之選模指標(Table 4.3.18)：R_p^2，$R_{Adj,p}^2$，C_p，AIC, MSE，得知R_p^2=0.6730，$R_{Adj,p}^2$=0.6562，C_p=2.2561, AIC=80.7980，$MS_{Res}(p)$=6.39177 與 Figure 4.2.1－4.2.3 清楚看出，依參數精簡原則，所有回歸式的比較選取法所建議的模型與逐步回歸選取法所建議的模型一致，且與未去除影響點之模型 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_3$ (R_p^2=0.6455, $R_{Adj,p}^2$=0.6189，C_p=13.0510, AIC=112.5458，$MS_{Res}(p)$=11.8384)比較，此兩種方法建議之模型 4.3.1 為較佳之模式。故在放入反應變數去除影響點後 y^* 與解釋變數膀胱癌(x_1^*)、肺癌(x_2^*)、腎癌(x_3^*)與其解釋變數之平方項下，其模型 4.3.1 為我們所選取之最佳回歸模型。

Table 4.3.1 Parameter estimates (Forward Selection: Step 1)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	9.6329	2.2932	164.5969	17.65	0.0001
x_2^*	1	0.7475	0.1157	389.1066	41.71	<0.0001

Table 4.3.2 Analysis of Variance (Forward Selection: Step 1)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	1	389.1066	389.1066	41.71	<.0001
Error	40	373.1277	9.3282		
Total	41	762.2344			
R-Square		0.5105	C(p)		19.2627

Table 4.3.3 Parameter estimates (Forward Selection: Step 2)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	2.3987	2.5108	5.8338	0.91	0.3453
x_2^*	1	0.6308	0.0994	257.3525	40.26	<0.0001
x_3^*	1	3.4100	0.7746	123.8486	19.38	<0.0001

Table 4.3.4 Analysis of Variance (Forward Selection: Step 2)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	512.9552	256.4776	40.13	<.0001
Error	39	249.2791	6.3918		
Total	41	762.2344			
R-Square		0.6730	C(p)		2.2561

Table 4.3.5 Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	P-value
1	x_2^*	1	0.5105	0.5105	19.2627	41.71	<.0001
2	x_3^*	2	0.1625	0.6730	2.2561	19.38	<.0001

Table 4.3.6 Parameter estimates (Backward Elimination: Step 0)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	-3.2198	9.7293	0.7136	0.11	0.7426
x_1^*	1	3.3640	4.9885	2.9631	0.45	0.5043
x_2^*	1	0.5606	0.1263	128.4520	19.71	<0.0001
x_3^*	1	3.1000	0.8306	90.7625	13.93	0.0006
$(x_1^*)^2$	1	-0.3362	0.6013	2.0367	0.31	0.5795

Table 4.3.7 Analysis of Variance (Backward Elimination: Step 0)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	521.1399	130.2850	19.99	<.0001
Error	37	241.0945	6.5161		
Total	41	762.2344			
R-Square		0.6837	C(p)		5.0000

Table 4.3.8 Parameter estimates (Backward Elimination: Step 1)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	2.0274	2.5405	4.0748	0.64	0.4298
x_1^*	1	0.5962	0.6082	6.1479	0.96	0.3332
x_2^*	1	0.5567	0.1249	127.0589	19.86	<0.0001
x_3^*	1	3.1969	0.8050	100.9135	15.77	0.0003

Table 4.3.9 Analysis of Variance (Backward Elimination: Step 1)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	519.1032	173.0344	27.04	<.0001
Error	38	243.1312	6.3982		
Total	41	762.2344			
R-Square		0.6810	C(p)		3.3126

Table 4.3.10 Parameter estimates (Backward Elimination: Step 2)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	2.3987	2.5108	5.8338	0.91	0.3453
x_2^*	1	0.6307	0.0994	257.3525	40.26	<0.0001
x_3^*	1	3.4100	0.7747	123.8486	19.38	<0.0001

Table 4.3.11 Analysis of Variance (Backward Elimination: Step2)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	512.9552	256.4776	40.13	<.0001
Error	39	249.2791	6.3918		
Total	41	762.2344			

R-Square	0.6730	C(p)	2.2561

Table 4.3.12 Summary of Backward Elimination

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	P-value
1	$(x_1^*)^2$	3	0.0027	0.6810	3.3126	0.31	0.5795
2	x_1^*	2	0.0081	0.6730	2.2561	0.96	0.3332

Table 4.3.13 Parameter estimates (Stepwise Selection: Step 1)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	9.6329	2.2932	164.5969	17.65	0.0001
x_2^*	1	0.7475	0.1157	389.1066	41.71	<0.0001

Table 4.3.14 Analysis of Variance (Stepwise Selection: Step 1)

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	1	389.1066	389.1066	41.71	<.0001
Error	40	373.1277	9.3282		
Total	41	762.2344			
R-Square		0.5105	C(p)		19.2627

Table 4.3.15 Parameter estimates (Stepwise Selection: Step 2)

Variable	DF	Parameter Estimate	Standard Error	Type II SS	F Value	P-value
Intercept	1	2.3987	2.5108	5.8338	0.91	0.3453
x_2^*	1	0.6308	0.0994	257.3525	40.26	<0.0001
x_3^*	1	3.4100	0.7747	123.8486	19.38	<0.0001

Table 4.3.16 Analysis of Variance (Stepwise Selection: Step 2)					
Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	2	512.9552	256.4776	40.13	<.0001
Error	39	249.2791	6.3918		
Total	41	762.2344			
R-Square		0.6730	C(p)		2.2561

Table 4.3.17 Summary of Stepwise Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	P-value
1	x_2^*	1	0.5105	0.5105	19.2627	41.71	<.0001
2	x_3^*	2	0.1625	0.6730	2.2561	19.38	<.0001

Table 4.3.18 Summary of All Possible Regressions						
Number of Regressors in Model	Regressors in Model	R_p^2	$R_{Adj,p}^2$	C_p	AIC	$MS_{Res}(p)$
1	x_2^*	0.5105	0.4982	19.2627	95.7385	9.32819
1	x_1^*	0.3692	0.3534	35.7884	106.3877	12.02025
1	$(x_1^*)^2$	0.3539	0.3378	37.5765	107.3933	12.31153
1	x_3^*	0.3353	0.3187	39.7511	108.5848	12.66579
2	$x_2^* x_3^*$	0.6730	0.6562	2.2561	80.7980	6.39177
2	$x_1^* x_2^*$	0.5486	0.5255	16.7994	94.3303	8.82166
2	$x_2^* (x_1^*)^2$	0.5423	0.5189	17.5379	94.9136	8.94503
2	$x_1^* x_3^*$	0.5143	0.4894	20.8119	97.4066	9.49205
2	$x_3^* (x_1^*)^2$	0.5102	0.4851	21.2924	97.7603	9.57234
2	$x_1^* (x_1^*)^2$	0.3807	0.3489	36.4482	107.6179	12.10455
3	$x_1^* x_2^* x_3^*$	0.6810	0.6558	3.3126	81.7491	6.39819
3	$x_2^* x_3^* (x_1^*)^2$	0.6798	0.6545	3.4547	81.9088	6.42257
3	$x_1^* x_2^* (x_1^*)^2$	0.5646	0.5303	16.9290	94.8154	8.73308
3	$x_1^* x_3^* (x_1^*)^2$	0.5152	0.4769	22.7131	99.3335	9.72491
4	$x_1^* x_2^* x_3^* (x_1^*)^2$	0.6837	0.6495	5.0000	83.3958	6.51607

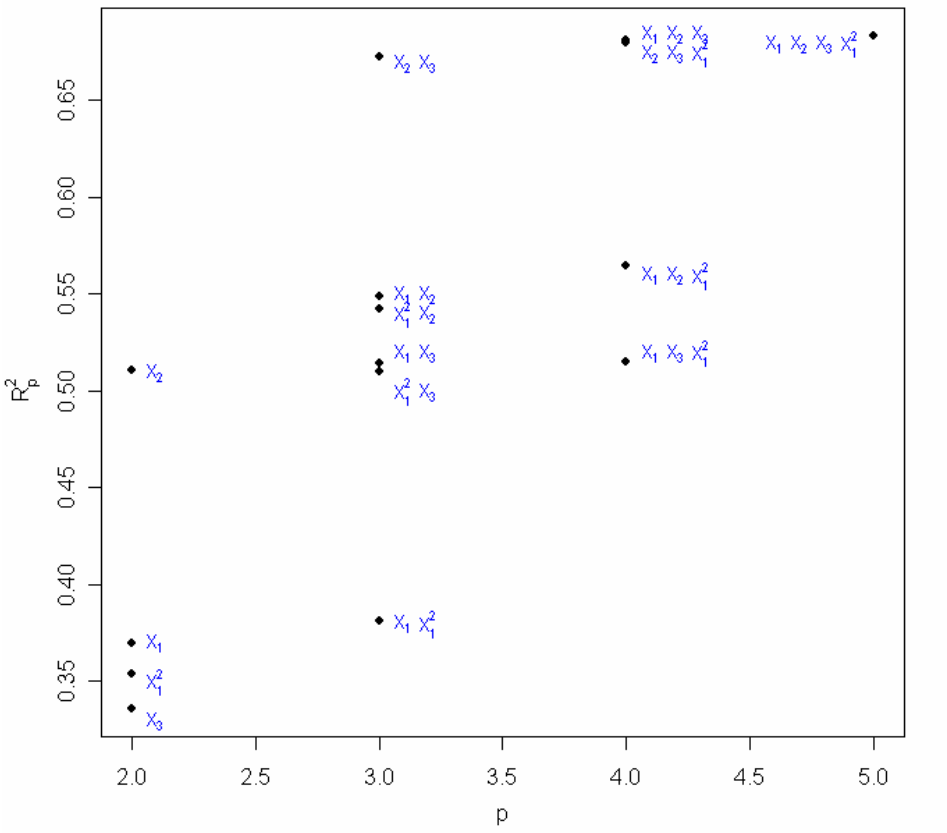


Figure 4.3.1 Plot R_p^2 versus p

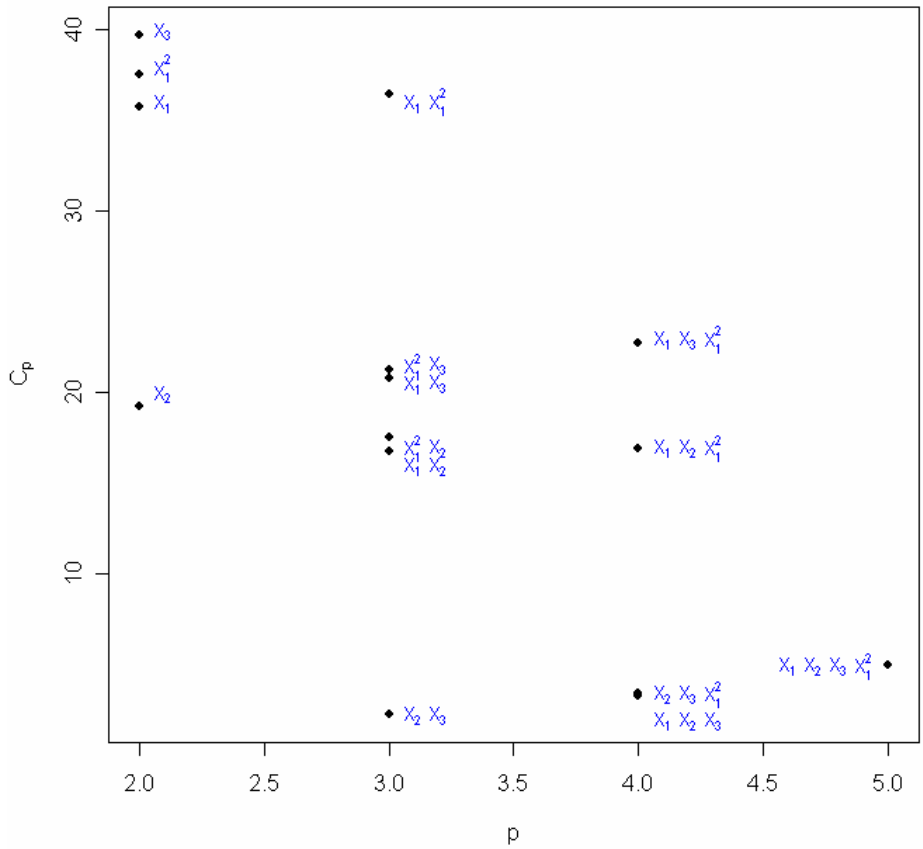


Figure 4.3.2 The C_p plot

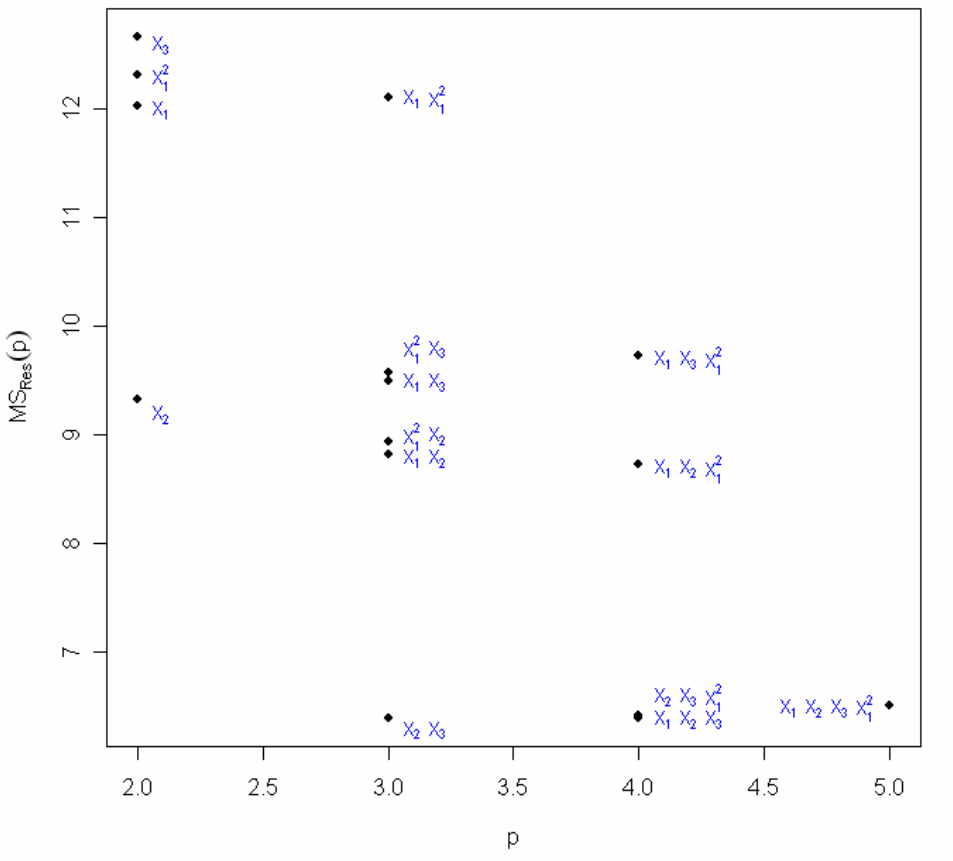


Figure 4.3.3 Plot $MS_{Res}(p)$ versus p

第四節

綜合以上三種情形之逐步回歸分析與所有回歸式比較選模結果得知以下三點：

1. 反應變數 y 適合加入平方項，則最佳模式為

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_1^2$$

$$= 23.5358\text{-}11.1275 x_1 + 0.5443 x_2 + 3.3897 x_3 + 1.5126 x_1^2$$

2. 反應變數 y 經轉換後為 $y^{(-0.22)}$，不適合加入二次項，故最佳模式為

$$\hat{y}^{(-0.22)} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

$$= 0.6134 - 0.0070 x_1 - 0.0025 x_2 - 0.0143 x_3$$

3. 反應變數 y 去除影響點(第 8 與 26 筆資料)之後為 $\hat{y}^*$，不適合加入二次項，則最佳模式為

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_2 x_2^* + \hat{\beta}_3 x_3^*$$

$$= 2.3987 + 0.6308 x_2^* + 3.41 x_3^*$$

第五章 屬質的解釋變數

前言

本章考慮在模型中加入了虛擬變數，探討虛擬變數的組別之間的差異對模型是否有影響。而分別加入模型中考慮的 3 個虛擬變數分別為：$x_4=1$ 表菸頭量＞平均數 23.77，$x_4=0$ 表菸頭量＜平均數 23.77；依地區(AREA)劃分設 $x_5=1$ 為北西部、$x_5=2$ 中西部、$x_5=3$ 南部與 $x_5=4$ 西部；依地區(WEST)劃分 $x_6=1$ 設南部與西部 $x_6=0$ 為北西部與中西部。

第一節

我們先利用簡單的散佈圖來觀察不同的組別之間是否由圖型可明顯看出差異。由 Figure 5.1.1 可看出，若依菸頭量＞平均數 23.77 與＜平均數 23.77 可將資料明顯的分成兩群。

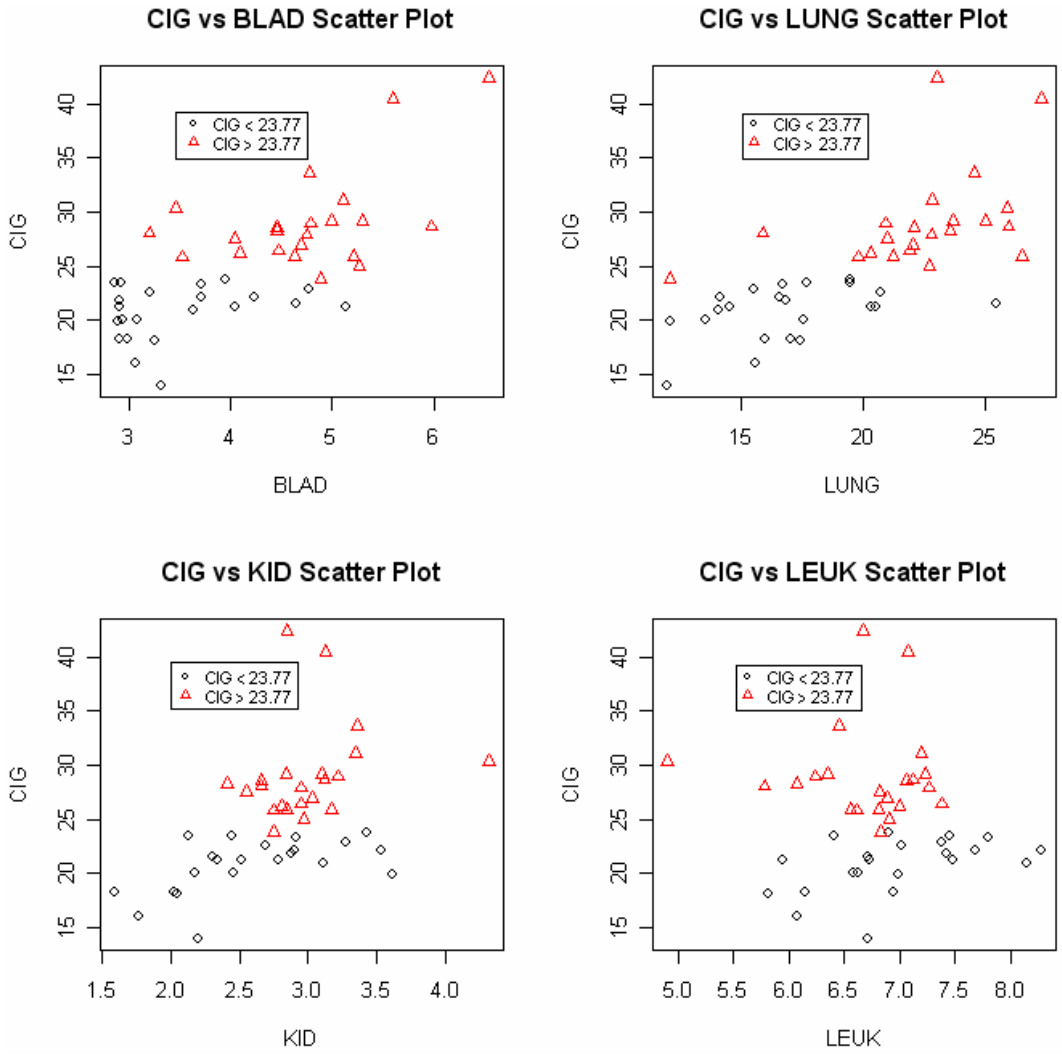


Figure 5.1.1 Scatter Plot to Separate the Dummy Variable x_4

Figure 5.1.2 是依地區(AREA)： 1 為北西部 2.中西部 3.南部 4.西部，利用散佈圖標示出來，但我們很難由圖型判斷出菸頭數是否受地區的不同有所影響。

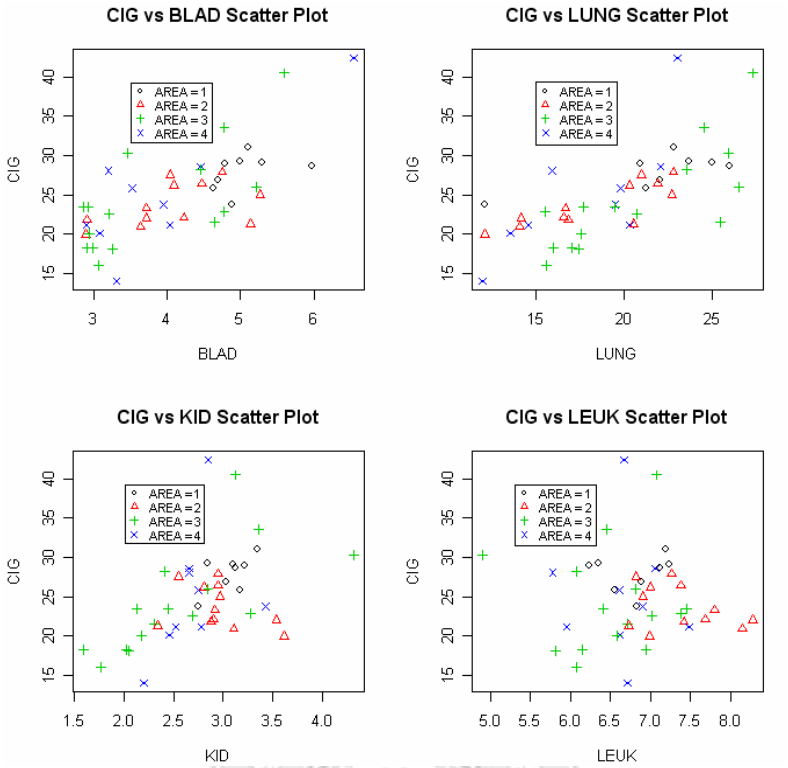


Figure 5.1.2 Scatter Plot to Separate the Dummy Variable x_5

Figure 5.1.3 依地區劃分設南部與西部為 1 而北西部與中西部為 0 利用散佈圖分別標示出。

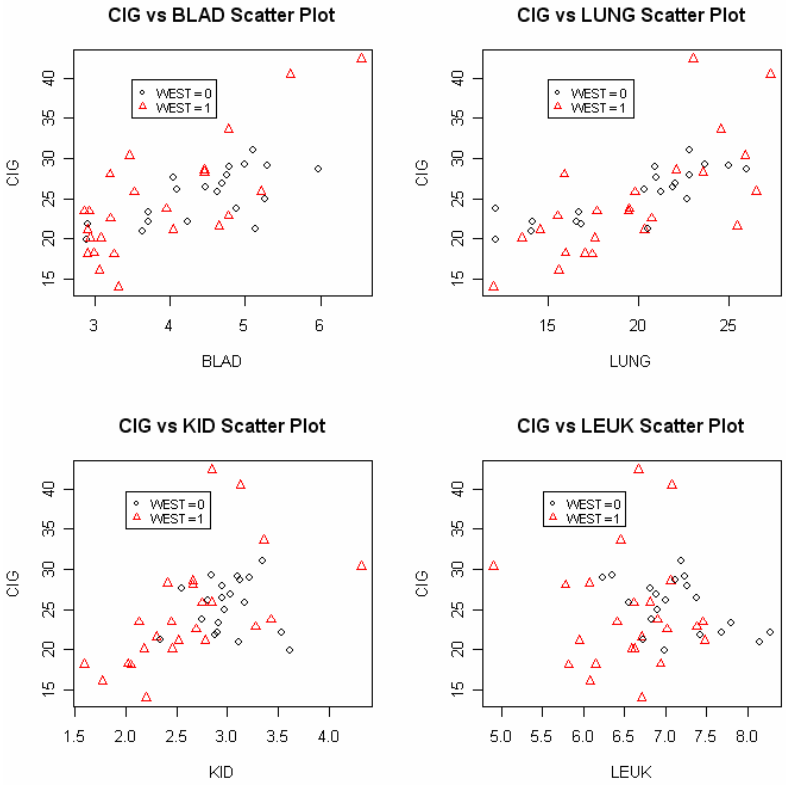


Figure 5.1.3 Scatter Plot to Separate the Dummy Variable x_6

Figure 5.1.4 與 Figure 5.1.5 是取菸頭量＞平均數 23.77 的資料，在分別看是否受地區影響。由圖形我們並沒有看出明顯的差異。

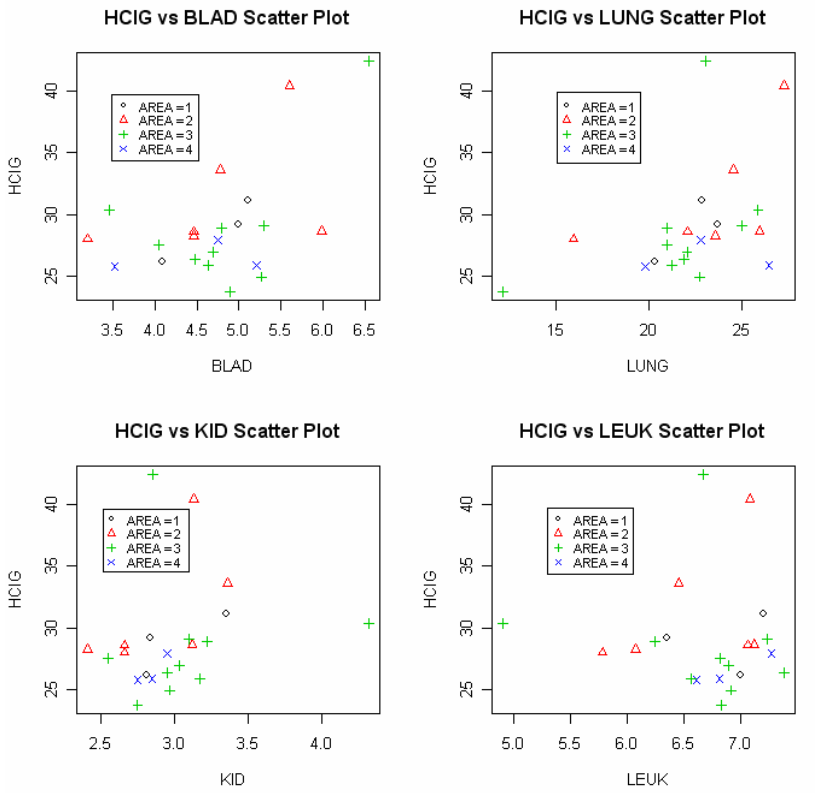


Figure 5.1.4 Scatter Plot to Separate the Dummy Variable x_5

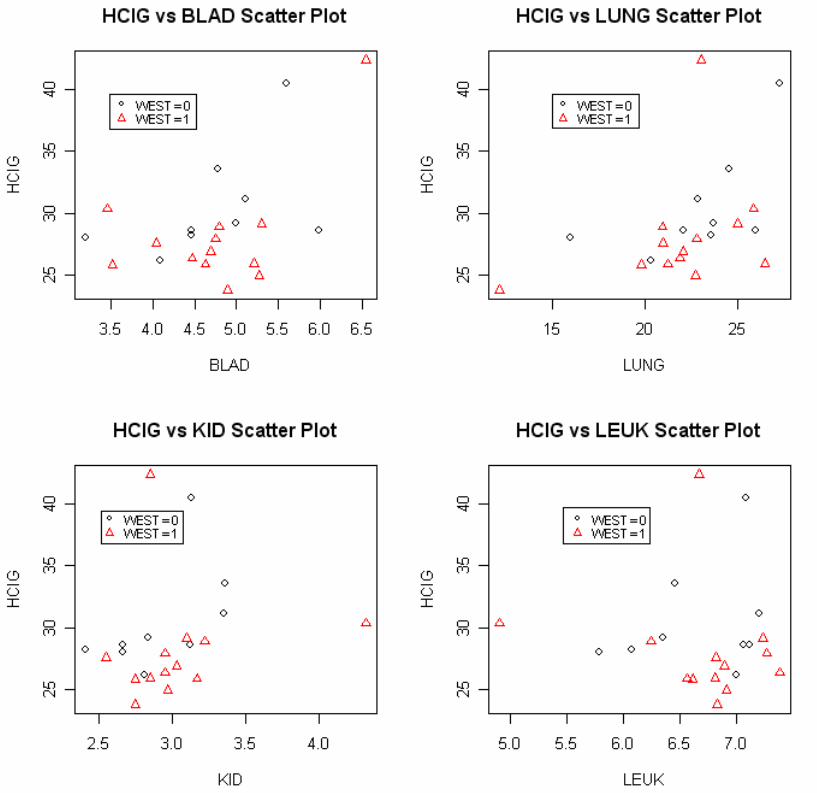


Figure 5.1.5 Scatter Plot to Separate the Dummy Variable x_6

Figure 5.1.6 與 Figure 5.1.7 是取菸頭量＜平均數 23.77 的資料，看是否受地區影響。Figure 5.1.7 是依地區劃分設南部與西部為 1 而北西部與中西部為 0，發現北西部與中西部菸頭數的點散佈似乎都高於南部與西部。

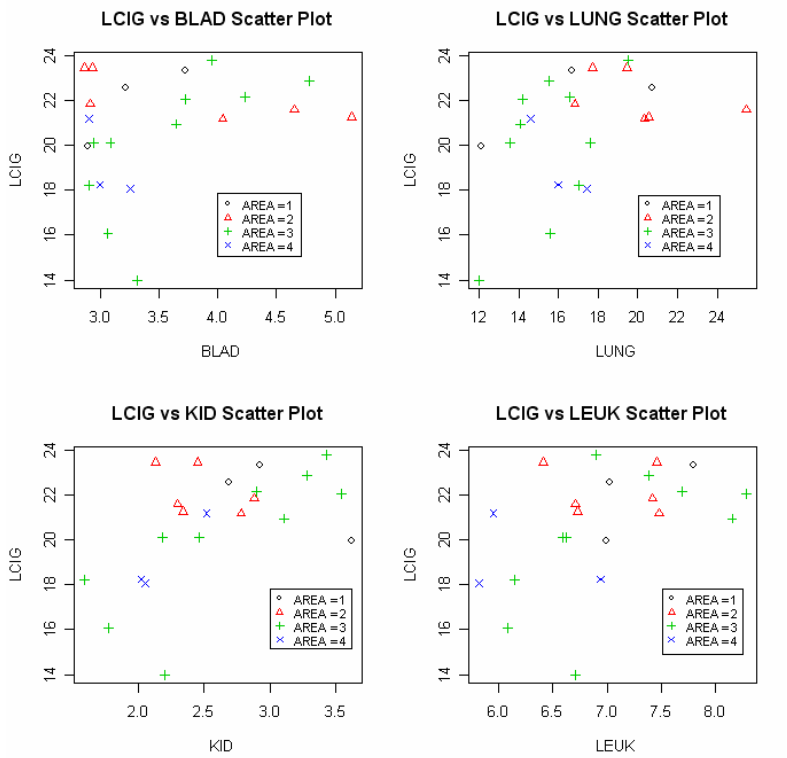


Figure 5.1.6 Scatter Plot to Separate the Dummy Variable x_5

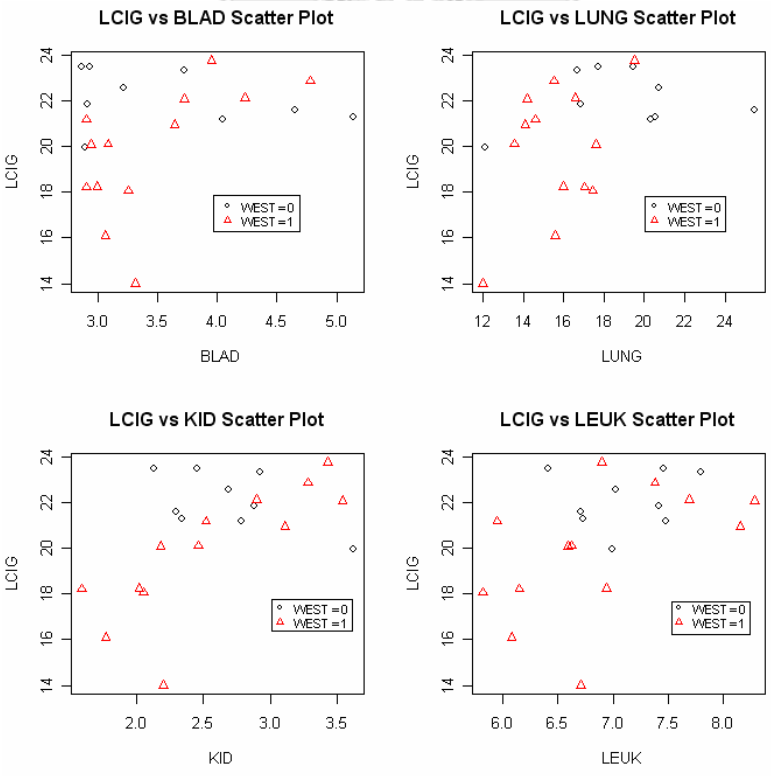


Figure 5.1.7 Scatter Plot to Separate the Dummy Variable x_6

第二節

此節將延續前面章節所選出之解釋變數膀胱癌(x_1)、肺癌(x_2)與腎癌(x_3)與分別加入的虛擬變數對反應變數菸頭數(y)配適複迴歸模型。

模型 5.2.1 為反應變數菸頭數(y)與解釋變數膀胱癌(x_1)、肺癌(x_2)與腎癌(x_3)配適模型。可用來比較當我們加入新的虛擬變數進入模型中與未加入有何不同。由 Table 5.2.1 可知參數估計均為顯著，解釋變數均具解釋能力。Table 5.2.2 可看出模型檢定顯著，而整體解釋變數的解釋能力，可由 $R^2_{Adj}=61.89\%$ 知解釋了總變異的 61.89%。

$$\begin{aligned}\hat{y} &= \hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_3\\ &= -1.1916+2.0544x_1+0.5179x_2+2.6701x_3\end{aligned} \tag{5.2.1}$$

Table 5.2.1 Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	-1.1916	3.3579	-0.355	0.7246
x_1	1	2.0544	0.7441	2.761	0.0087
x_2	1	0.5179	0.1653	3.134	0.0032
x_3	1	2.6701	1.0853	2.460	0.0183

Table 5.2.2 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	862.11	287.37	24.27	<.0001
Error	40	473.54	11.84		
Total	43	1335.65			
Root MSE		3.441	R-Square		0.6455
			Adj R-Sq		0.6189

模型 5.2.2 為反應變數菸頭數(y)與解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與菸頭數之虛擬變數(x_4)配適模型。由 Table 5.2.3 可知參數估計除了膀胱癌(x_1)此變數不顯著外，其他均為顯著，解釋變數均具解釋能力。Table 5.2.4 可看出模型檢定顯著，而整體解釋變數的解釋能力，可由 $R^2_{Adj}=68.51\%$ 知解釋了總變異的 68.51%。由 Figure 5.2.1 之配適線可看出截距有差異。這很符合直覺，因為配適線就是依菸頭量＞平均數 23.77 與菸頭量＜平均數 23.77 畫出

兩條不同的配適線。Figure 5.2.2 為此模型之殘差常態圖，如同前面所分析之結果，可明顯看到影響點觀測值 8 與觀測值 26 對整體的影響。

$$\hat{y} = 5.1954 + 1.2842x_1 + 0.3371x_2 + 2.0459x_3 + 4.1683x_4 \quad (5.2.2)$$

當 $x_4 = 1$ 時

$$E(y) = (\beta_0 + \beta_4) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$
$$\hat{y} = 9.3637 + 1.2842x_1 + 0.3371x_2 + 2.0459x_3 \quad (5.2.3)$$

當 $x_4 = 0$ 時

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$
$$\hat{y} = 5.1954 + 1.2842x_1 + 0.3371x_2 + 2.0459x_3 \quad (5.2.4)$$

Table 5.2.3 Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	5.1954	3.6941	1.406	0.1675
x_1	1	1.2842	0.7214	1.780	0.0828
x_2	1	0.3371	0.1613	2.089	0.0433
x_3	1	2.0459	1.0072	2.031	0.0491
$(x_4 = 1)$	1	4.1683	1.3582	3.069	0.0039

Table 5.2.4 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	954.33	238.5825	24.39	<.0001
Error	39	381.42	9.78		
Total	43	1335.75			
Root MSE		3.127	R-Square		0.7144
			Adj R-Sq		0.6851

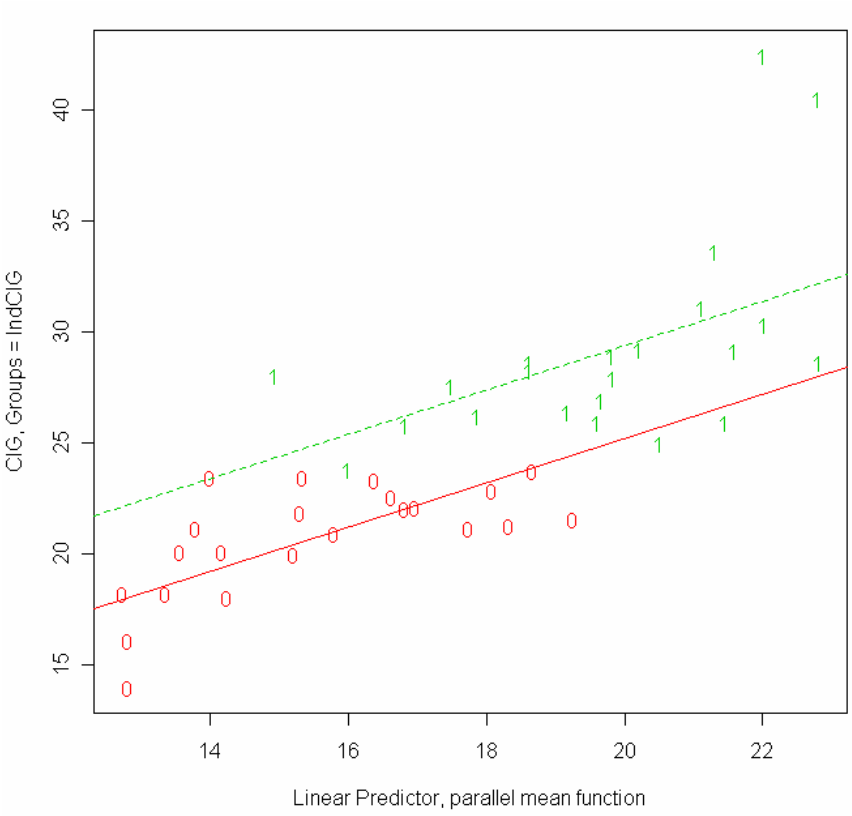


Figure 5.2.1 Response function for Model 5.2.2

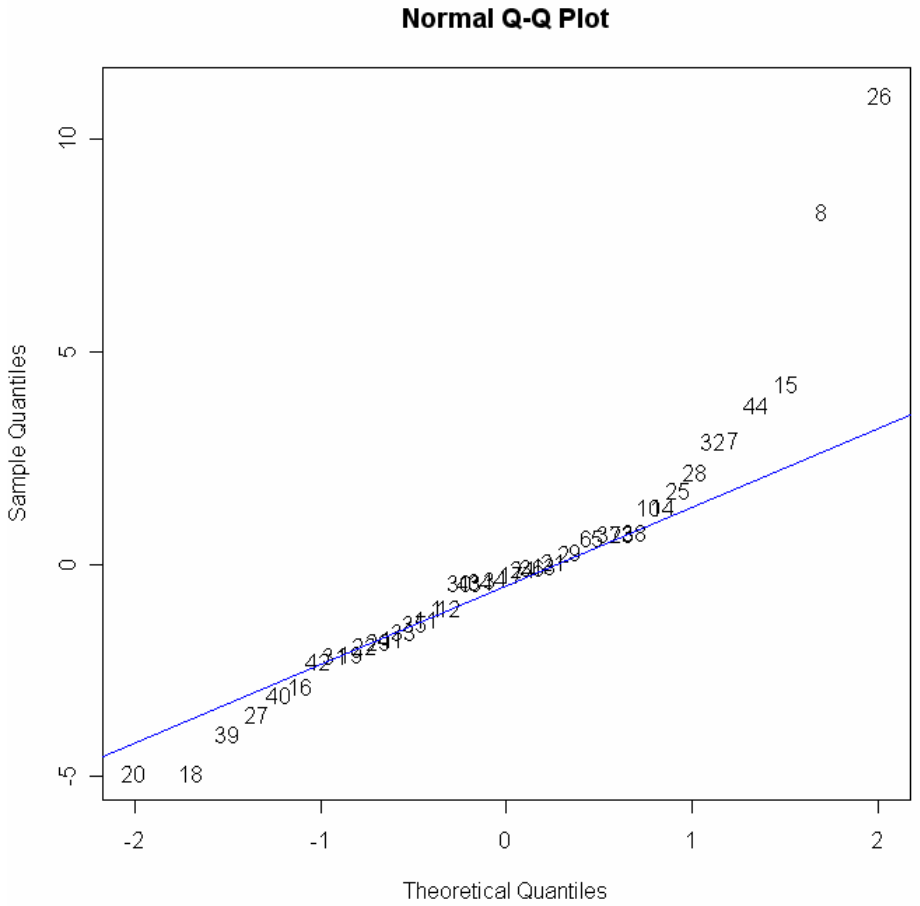


Figure 5.2.2 Normal probability plot of residuals for Model 5.2.2

模型 5.2.5 為反應變數菸頭數(y)與解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與地區之虛擬變數(x_5)配適模型。。由 Table 5.2.5 可知參數估計除了虛擬變數(x_5)此變數不顯著外，其他均為顯著，解釋變數均具解釋能力。Table 5.2.6 可看出模型檢定顯著，而整體解釋變數的解釋能力，可由 $R^2_{Adj} = 62.59\%$ 知解釋了總變異的 62.59%。由 Figure 5.2.3 可看出依地區(AREA)：1 為北西部 2.中西部 3.南部 4.西部此四區之配適線其截距不同，而北西部與中西部之配適線是幾乎重合的。由此模型配適下，地區此參數並不顯著，而圖形的呈現讓我們考慮或許可將地區依北西部與中西部為一區，而南部與西部為一區。再從新配適模型比較地區性的差異，我們將在下一個模型呈現此結果。Figure 5.2.4 為此模型之殘差常態圖，有點像輕尾分佈(Light – tailed errors)，而我們由前面分析已知是有少數的極端的觀察值存在所影響。

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_5 x_5 \tag{5.2.5}$$

當 $x_5 = 1$ 時

$$\begin{aligned} E(y) &= (\beta_0 + \beta_5) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ \hat{y} &= -4.085 + 2.3206 x_1 + 0.4490 x_2 + 3.0847 x_3 \end{aligned} \tag{5.2.6}$$

當 $x_5 = 2$ 時

$$\begin{aligned} E(y) &= (\beta_0 + \beta_5) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ \hat{y} &= -4.02 + 2.3206 x_1 + 0.4490 x_2 + 3.0847 x_3 \end{aligned} \tag{5.2.7}$$

當 $x_5 = 3$ 時

$$\begin{aligned} E(y) &= (\beta_0 + \beta_5) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ \hat{y} &= -2.8687 + 2.3206 x_1 + 0.449 x_2 + 3.0847 x_3 \end{aligned} \tag{5.2.8}$$

當 $x_5 = 4$ 時

$$\begin{aligned} E(y) &= (\beta_0 + \beta_5) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ \hat{y} &= -1.27 + 2.3206 x_1 + 0.499 x_2 + 3.0847 x_3 \end{aligned} \tag{5.2.9}$$

Table 5.2.5 Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	-4.0850	4.2571	-0.960	0.3435
x_1	1	2.3206	0.8521	2.723	0.0098
x_2	1	0.4990	0.1884	2.649	0.0118
x_3	1	3.0847	1.1555	2.670	0.0112
$(x_5 = 2)$	1	0.0650	1.6672	0.039	0.9691
$(x_5 = 3)$	1	1.2163	1.8299	0.665	0.5104
$(x_5 = 4)$	1	2.8150	1.8160	1.550	0.1296

Table 5.2.6 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	6	905.71	150.9517	12.99	<.0001
Error	37	429.93	11.62		
Total	43	1335.64			
Root MSE		3.409	R-Square		0.6781
			Adj R-Sq		0.6259

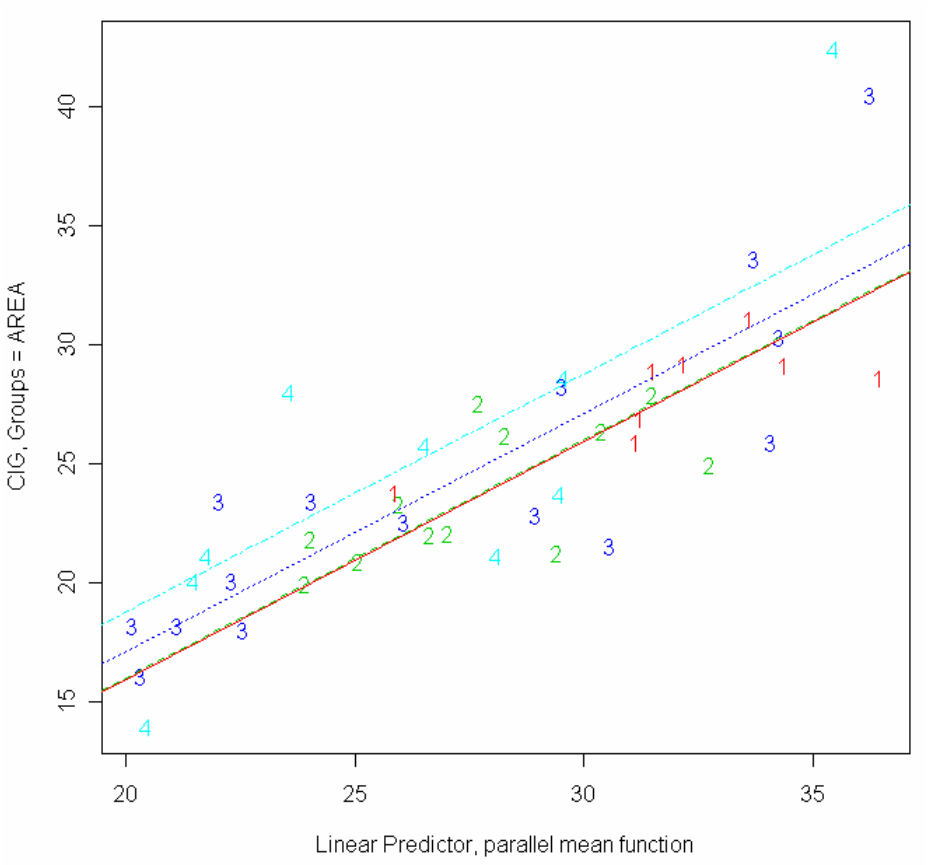


Figure 5.2.3 Response function for Model 5.2.5

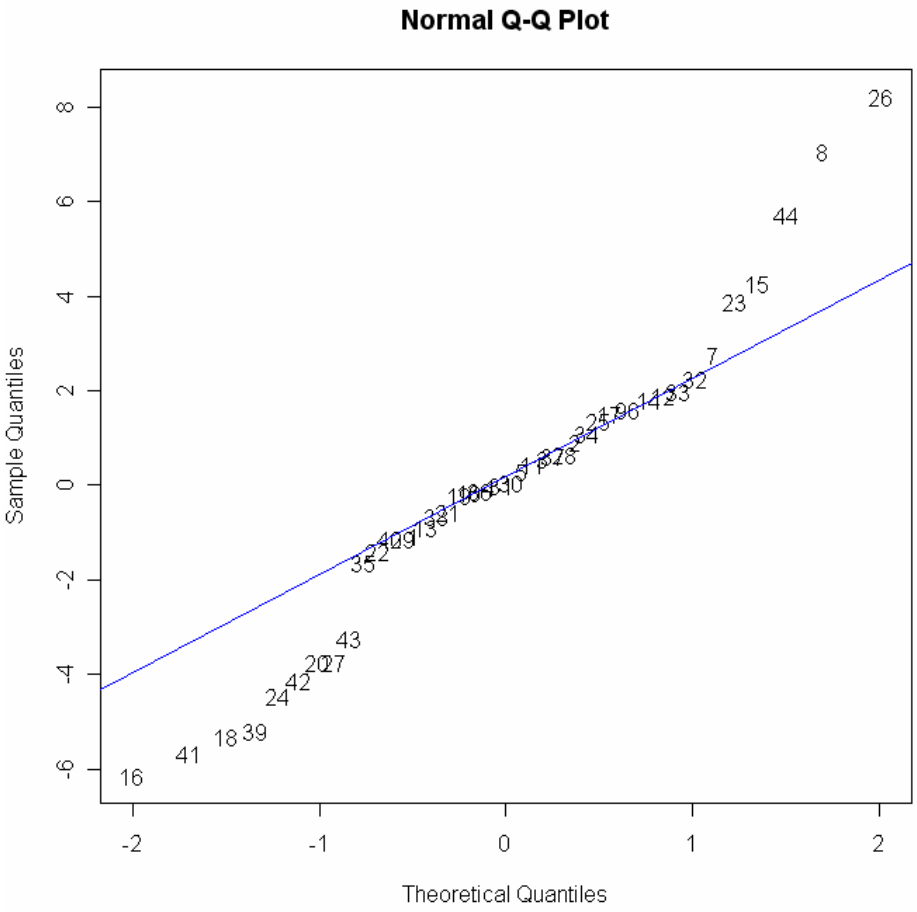


Figure 5.2.4 Normal probability plot of residuals for Model 5.2.5

模型 5.2.10 為反應變數菸頭數(y)與解釋變數膀胱癌(x_1)、肺癌(x_2)、腎癌(x_3)與地區之虛擬變數(x_6)配適模型。。由 Table 5.2.7 可知參數估計除了虛擬變數(x_6)此變數較不顯著外，其他均為顯著，解釋變數均具解釋能力。Table 5.2.8 可看出模型檢定顯著，而整體解釋變數的解釋能力，可由 $R^2_{Adj}=63.51\%$ 知解釋了總變異的 63.51%。依地區(WEST)劃分 1 為南部與西部 0 為北西部與中西部此兩區，由 Figure 5.2.5 可看出此兩區之配適線其截距不同。如同 Figure 5.2.6，Figure 5.2.7 此模型之殘差常態圖，為輕尾分佈(Light – tailed errors)。

$$\hat{y}=-4.1796+2.5464x_1+0.4239x_2+3.2855x_3+1.9923x_4 \tag{5.2.10}$$

當 $x_6=1$ 時

$$\begin{aligned} E(y)&=(\beta_0+\beta_6)+\beta_1x_1+\beta_2x_2+\beta_3x_3\\ \hat{y}&=-2.1873+2.5464x_1+0.4239x_2+3.2855x_3 \end{aligned} \tag{5.2.11}$$

當 $x_6=0$ 時

$$\begin{aligned} E(y)&=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3\\ \hat{y}&=-4.1796+2.5464x_1+0.4239x_2+3.2855x_3 \end{aligned} \tag{5.2.12}$$

Table 5.2.7 Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	-4.1796	3.7423	-1.117	0.2709
x_1	1	2.5464	0.7856	3.241	0.0024
x_2	1	0.4239	0.1712	2.476	0.0177
x_3	1	3.2855	1.1242	2.922	0.0058
($x_6=1$)	1	1.9923	1.1946	1.668	0.1034

Table 5.2.8 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	893.63	223.4075	19.71	<.0001
Error	39	442.01	11.33		
Total	43	1335.64			
Root MSE		3.367	R-Square		0.6691
			Adj R-Sq		0.6351

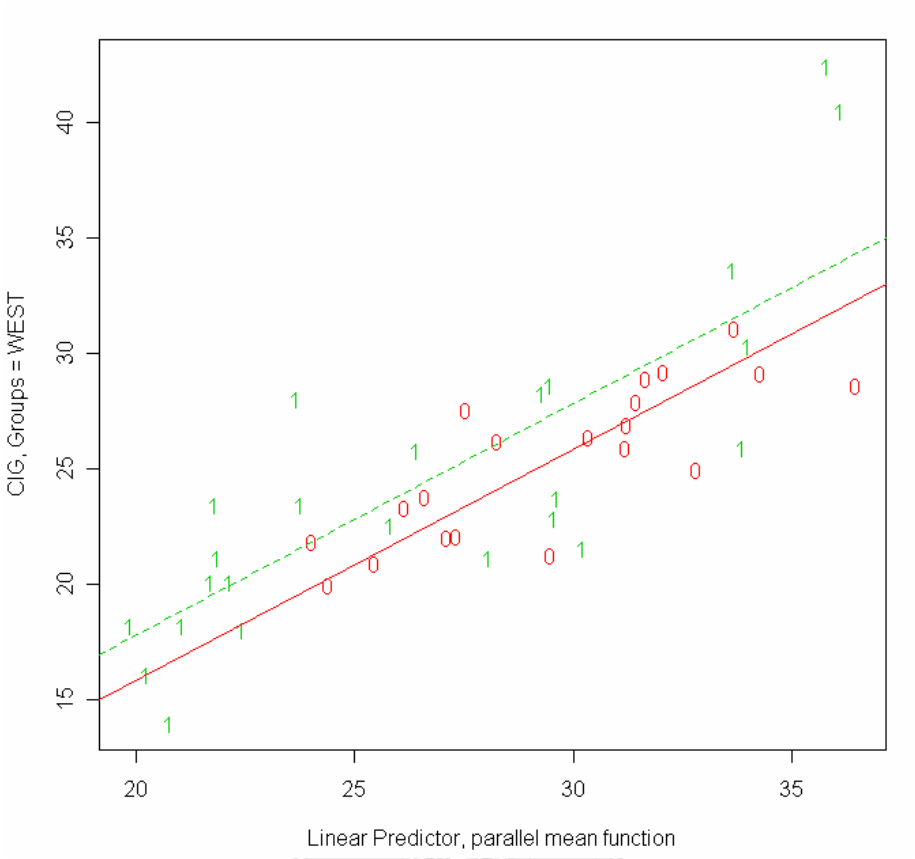


Figure 5.2.5 Response function for Model 5.2.10

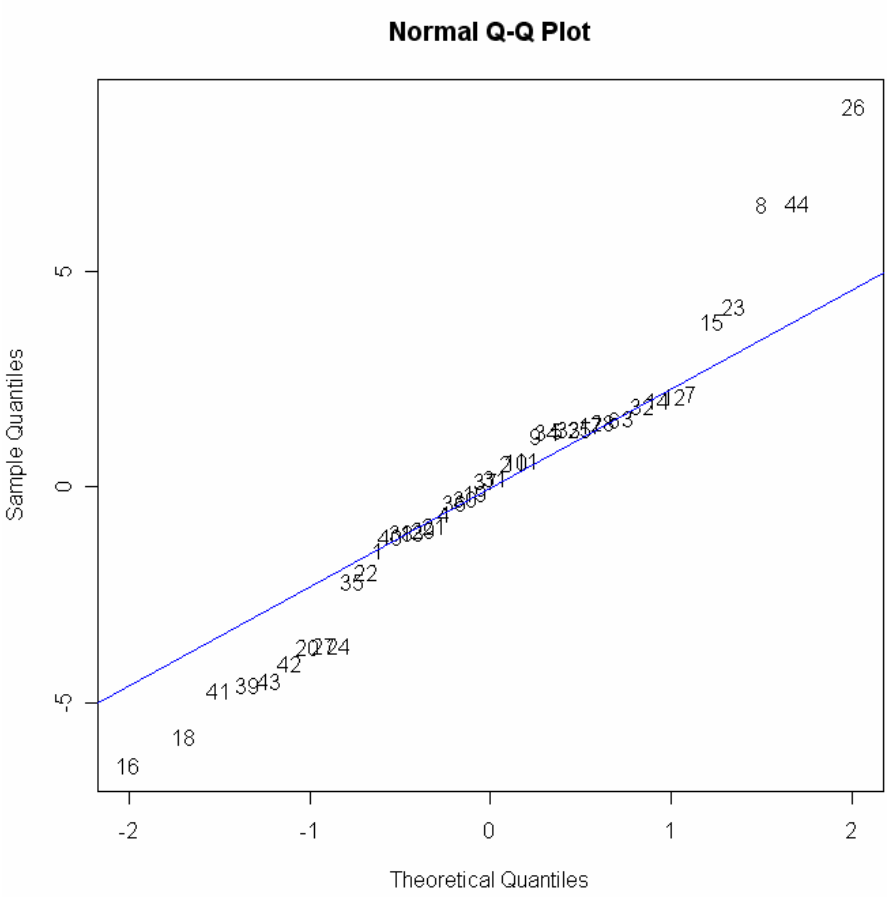


Figure 5.2.6 Normal probability plot of residuals for Model 5.2.10

第三節

此節將如同第二節針對解釋變數膀胱癌(x_1^*)、肺癌(x_2^*)與腎癌(x_3^*)與分別加入的虛擬變數對反應變數菸頭數(y^*)配適複迴歸模型。但我們將影響點觀測值 8 與觀測值 26 刪去，想比較影響點之差異，且我們亦可觀察少了此兩影響點下，模型之配適情形。因為我們知道最小平方法容易受影響點影響，嚴重時可能會扭曲其餘觀測值之配適情形。也可能會導致遺漏重要的變數或選用不正確的函數形式。

模型 5.3.1 為反應變數菸頭數(y^*)與解釋變數膀胱癌(x_1^*)、肺癌(x_2^*)與腎癌(x_3^*)配適模型。由 Table 5.3.1 可發現除了膀胱癌(x_1^*)參數估計不顯著，其餘均為顯著，解釋變數均具解釋能力。Table 5.3.2 可看出模型檢定顯著，而整體解釋變數的解釋能力，可由 $R^2_{Adj}=65.58\%$ 知解釋了總變異的 65.58%。

$$\begin{aligned}\hat{y}^* &= \hat{\beta}_0+\hat{\beta}_1x_1^*+\hat{\beta}_2x_2^*+\hat{\beta}_3x_3^* \\ &= 2.0274+0.5962x_1^*+0.5567x_2^*+3.1969x_3^*\end{aligned} \tag{5.3.1}$$

Table 5.3.1 Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	2.0274	2.5405	0.798	0.4298
x_1^*	1	0.5962	0.6082	0.980	0.3332
x_2^*	1	0.5567	0.1249	4.456	<0.0001
x_3^*	1	3.1969	0.8050	3.971	0.0003

Table 5.3.2 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	3	519.103	173.0343	27.04	<.0001
Error	38	243.131	6.398		
Total	41	762.234			
Root MSE		2.529	R-Square		0.681
			Adj R-Sq		0.6558

模型 5.3.2 為反應變數菸頭數(y^*)與解釋變數膀胱癌(x_1^*)、肺癌(x_2^*)與腎癌(x_3^*)與菸頭數之虛擬變數(x_4^*)配適模型。由 Table 5.3.3 可知參數估計除了膀胱癌(x_1^*)此變數不顯著外，其他均為顯著，解釋變數均具解釋能力。Table 5.3.4 可看出模型檢定顯著，而整體解釋變數的解釋能力，可由 $R^2_{Adj}=82.08\%$ 知解釋了總變異的 82.08%。由 Figure 5.3.1 之配適線可看出截距有顯著的差異。可由圖形直觀看出菸頭量＞平均數 23.77 與菸頭量＜平均數 23.77 的資料基準點上的差異。Figure 5.3.2 為此模型之殘差常態圖，我們認為殘差符合常態性假設。

$$\hat{y}^* = 9.577 - 0.363x_1^* + 0.3484x_2^* + 2.5072x_3^* + 4.7740x_4^* \tag{5.3.2}$$

當 $x_4^*=1$ 時

$$\begin{aligned} E(y^*) &= (\beta_0+\beta_4)+\beta_1x_1^*+\beta_2x_2^*+\beta_3x_3^* \\ \hat{y}^* &= 14.351 - 0.363x_1^* + 0.3484x_2^* + 2.5072x_3^* \end{aligned} \tag{5.3.3}$$

當 $x_4^*=0$ 時

$$\begin{aligned} E(y^*) &= \beta_0+\beta_1x_1^*+\beta_2x_2^*+\beta_3x_3^* \\ \hat{y}^* &= 9.577 - 0.363x_1^* + 0.3484x_2^* + 2.5072x_3^* \end{aligned} \tag{5.3.4}$$

Table 5.3.3 Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	9.5770	2.2240	4.306	0.0001
x_1^*	1	-0.3630	0.4671	-0.777	0.4420
x_2^*	1	0.3484	0.0966	3.607	0.0009
x_3^*	1	2.5072	0.5922	4.234	0.0001
x_4^*	1	4.7740	0.7961	5.997	<0.0001

Table 5.3.4 Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	638.942	159.7355	47.94	<.0001
Error	37	123.292	3.332		
Total	41	762.234			
Root MSE		1.825	R-Square		0.8382
			Adj R-Sq		0.8208

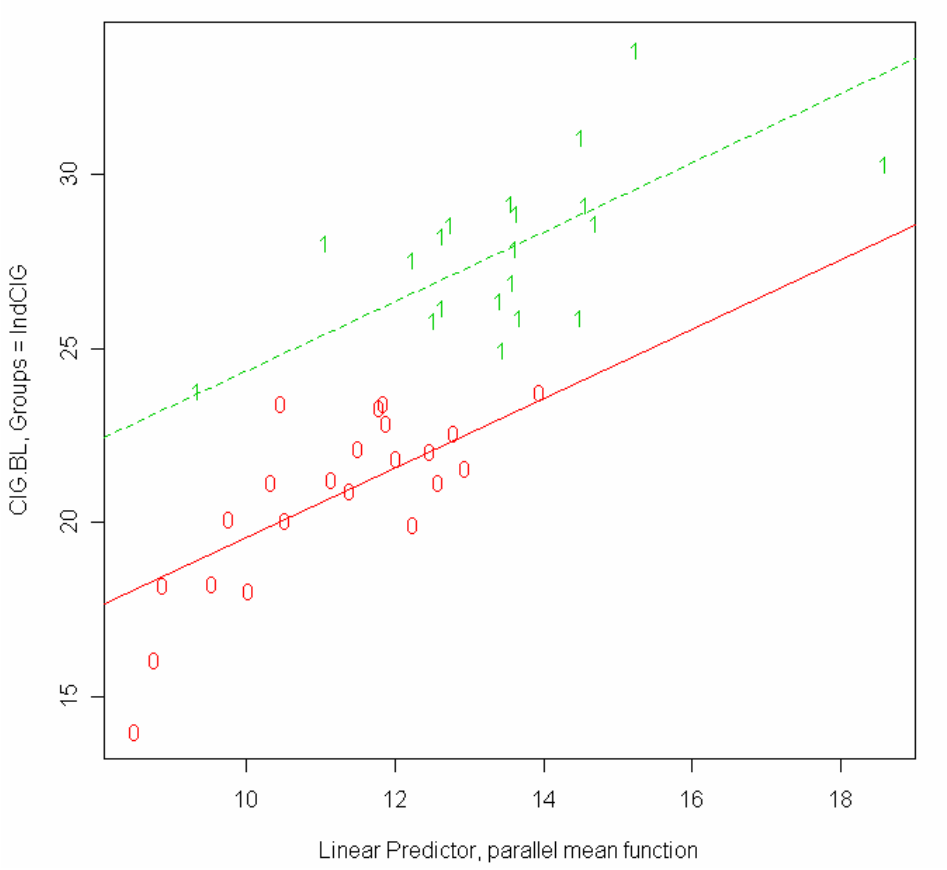


Figure 5.3.1 Response function for Model 5.3.2

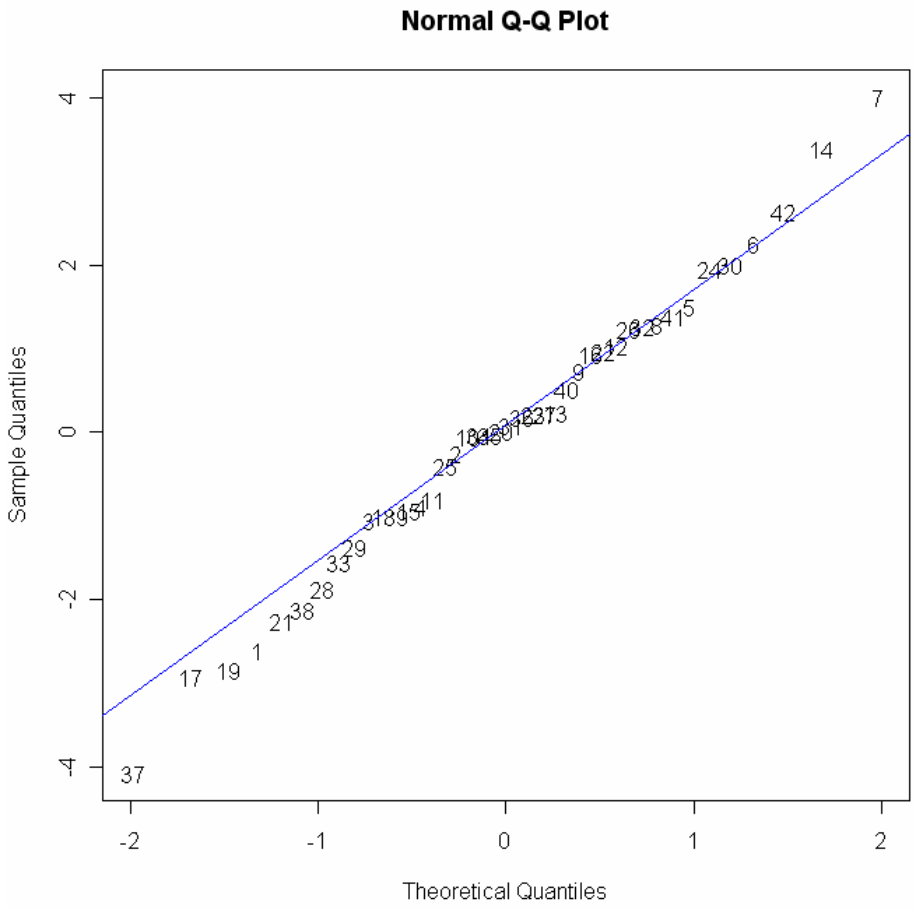


Figure 5.3.2 Normal probability plot of residuals for Model 5.3.2

模型 5.3.5 為反應變數菸頭數(y^*)與解釋變數膀胱癌(x_1^*)、肺癌(x_2^*)與腎癌(x_3^*)與地區之虛擬變數(x_5^*)配適模型。由 Table 5.3.5 可知參數估計膀胱癌(x_1^*)與虛擬變數(x_5^*)不顯著外，其他均為顯著，解釋變數均具解釋能力。Table 5.3.6 可看出模型檢定顯著，而整體解釋變數的解釋能力，可由 $R^2_{Adj}=65.92\%$ 知解釋了總變異的 65.52%。由 Figure 5.3.3 可看出依地區(AREA)：1 為北西部 2.中西部 3.南部 4.西部此四區之配適線呈現其截距上不同，並無之前未刪影響點時，北西部與中西部兩條線幾乎重合的現象。Figure 5.3.4 為此模型之殘差常態圖，可看出有離群值的存在。

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \hat{\beta}_2 x_2^* + \hat{\beta}_3 x_3^* + \hat{\beta}_5 x_5^* \tag{5.3.5}$$

當 $x_5^*=1$ 時

$$\begin{aligned} E(y^*) &= (\beta_0+\beta_5)+\beta_1 x_1^*+\beta_2 x_2^*+\beta_3 x_3^* \\ \hat{y}^* &= 5.5435-0.0691x_1^*+0.6328x_2^*+2.9311x_3^* \end{aligned} \tag{5.3.6}$$

當 $x_5^* = 2$ 時

$$E(y^*) = (\beta_0 + \beta_5) + \beta_1 x_1^* + \beta_2 x_2^* + \beta_3 x_3^*$$
$$\hat{y}^* = 3.7173 - 0.0691 x_1^* + 0.6328 x_2^* + 2.9311 x_3^* \quad (5.3.7)$$

當 $x_5^* = 3$ 時

$$E(y^*) = (\beta_0 + \beta_5) + \beta_1 x_1^* + \beta_2 x_2^* + \beta_3 x_3^*$$
$$\hat{y}^* = 3.0981 - 0.0691 x_1^* + 0.6328 x_2^* + 2.9311 x_3^* \quad (5.3.8)$$

當 $x_5^* = 4$ 時

$$E(y^*) = (\beta_0 + \beta_5) + \beta_1 x_1^* + \beta_2 x_2^* + \beta_3 x_3^*$$
$$\hat{y}^* = 4.3133 - 0.0691 x_1^* + 0.6328 x_2^* + 2.9311 x_3^* \quad (5.3.9)$$

Table 5.3.5 Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	5.5435	3.5668	1.554	0.1291
x_1^*	1	-0.0691	0.7621	-0.091	0.9283
x_2^*	1	0.6328	0.1423	4.446	<0.0001
x_3^*	1	2.9311	0.8580	3.416	0.0016
$(x_5^* = 2)$	1	-1.8262	1.2748	-1.433	0.1609
$(x_5^* = 3)$	1	-2.4454	1.4960	-1.635	0.1111
$(x_5^* = 4)$	1	-1.2302	1.5300	-0.804	0.4268

Table 5.3.6 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	6	540.311	90.0518	14.22	<.0001
Error	35	221.723	6.335		
Total	41	762.234			
Root MSE		2.517	R-Square		0.7091
			Adj R-Sq		0.6592

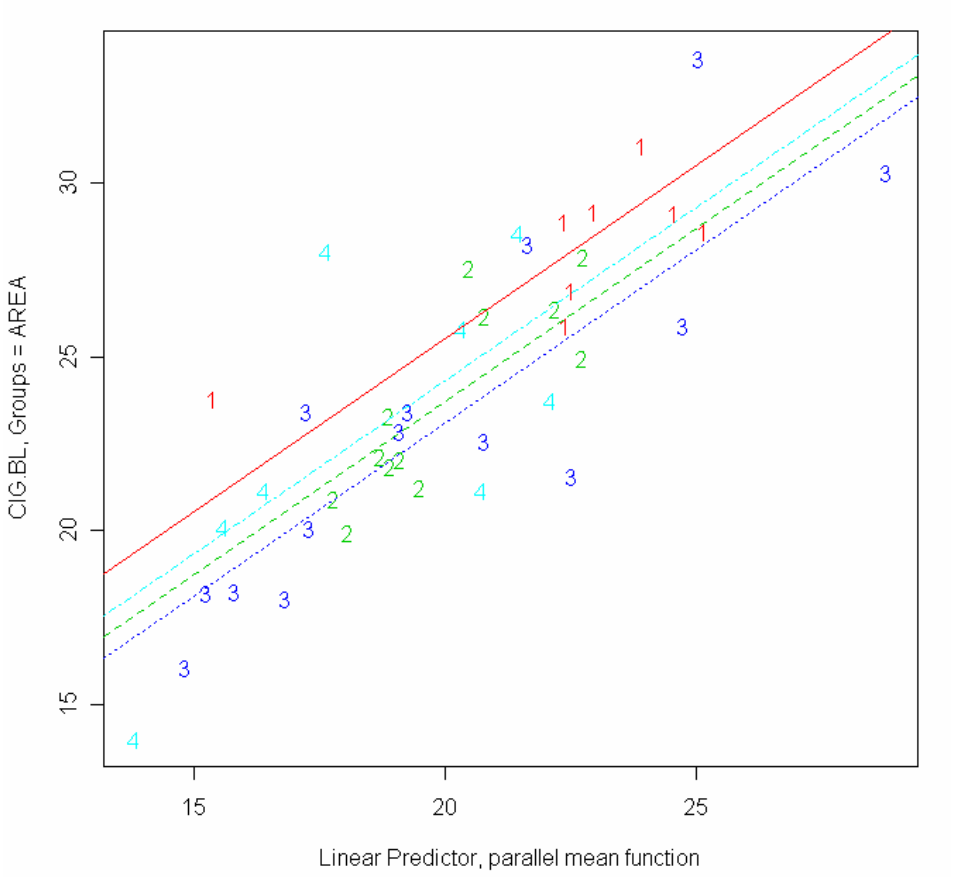


Figure 5.3.3 Response function for Model 5.3.5

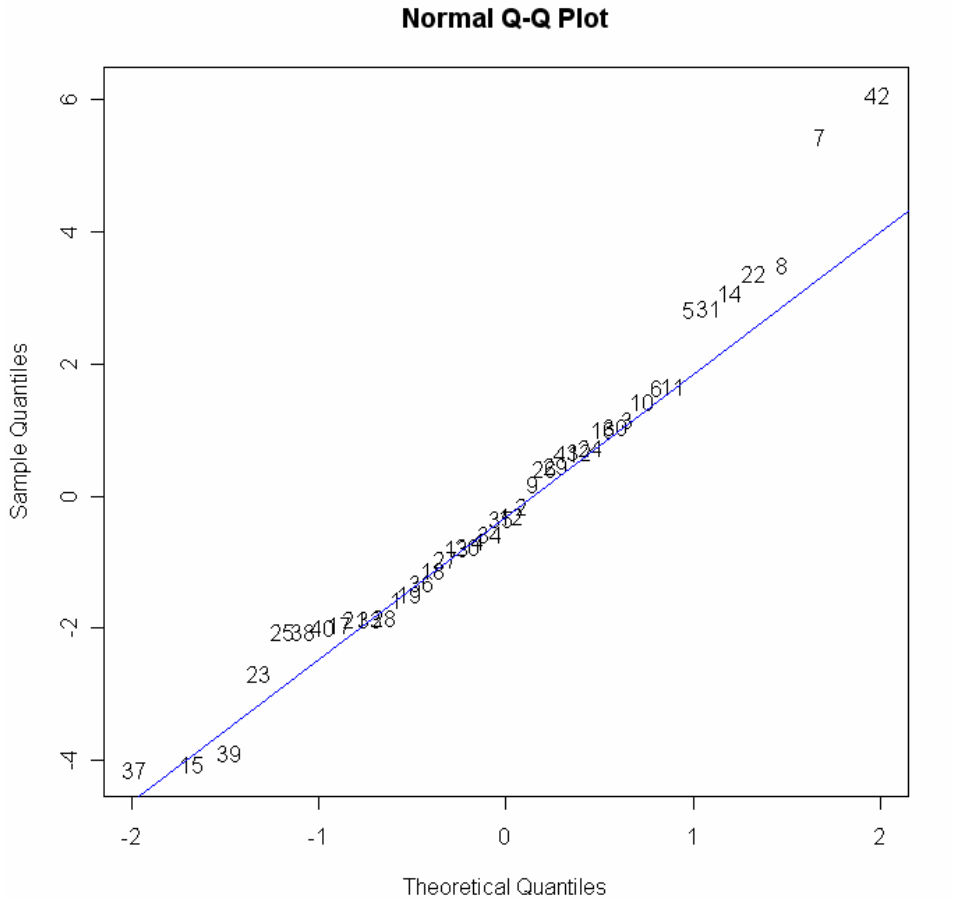


Figure 5.3.4 Normal probability plot of residuals for Model 5.3.5

模型 5.3.10 為反應變數菸頭數(y^*)與解釋變數膀胱癌(x_1^*)、肺癌(x_2^*)與腎癌(x_3^*)與地區之虛擬變數(x_6^*)配適模型。由 Table 5.3.7 可知參數估計膀胱癌(x_1^*)與地區虛擬變數不顯著，其他均為顯著，解釋變數均具解釋能力。Table 5.3.8 可看出模型檢定顯著，而整體解釋變數的解釋能力，可由 $R^2_{Adj}=64.86\%$ 知解釋了總變異的 64.86%。依地區(WEST)劃分 1 為南部與西部 0 為北西部與中西部此兩區，由 Figure 5.3.5 可看出此兩區之配適線其截距上只有些微差異。Figure 5.3.6 為此模型之殘差常態圖，可看出有離群值。

$$\hat{y}^* = 2.8497 + 0.4231x_1^* + 0.5807x_2^* + 3.0732x_3^* - 0.4687x_4^* \tag{5.3.10}$$

當 $x_6^*=1$ 時

$$\begin{aligned} E(y^*) &= (\beta_0+\beta_6)+\beta_1x_1^*+\beta_2x_2^*+\beta_3x_3^* \\ \hat{y}^* &= 2.381+0.4231x_1^*+0.5807x_2^*+3.0732x_3^* \end{aligned} \tag{5.3.11}$$

當 $x_6^*=0$ 時

$$\begin{aligned} E(y^*) &= \beta_0+\beta_1x_1^*+\beta_2x_2^*+\beta_3x_3^* \\ \hat{y}^* &= 2.8497+0.4231x_1^*+0.5807x_2^*+3.0732x_3^* \end{aligned} \tag{5.3.12}$$

Table 5.3.7 Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	P-value
Intercept	1	2.8497	3.1198	0.913	0.3669
x_1^*	1	0.4231	0.7190	0.588	0.5598
x_2^*	1	0.5807	0.1364	4.257	0.0001
x_3^*	1	3.0732	0.8560	3.590	0.0009
$(x_6^*=1)$	1	-0.4687	1.0105	-0.464	0.6455

Table 5.3.8 Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	P-value
Model	4	520.508	130.127	19.92	<.0001
Error	37	241.726	6.533		
Total	41	762.234			
Root MSE		2.556	R-Square		0.6829
			Adj R-Sq		0.6486

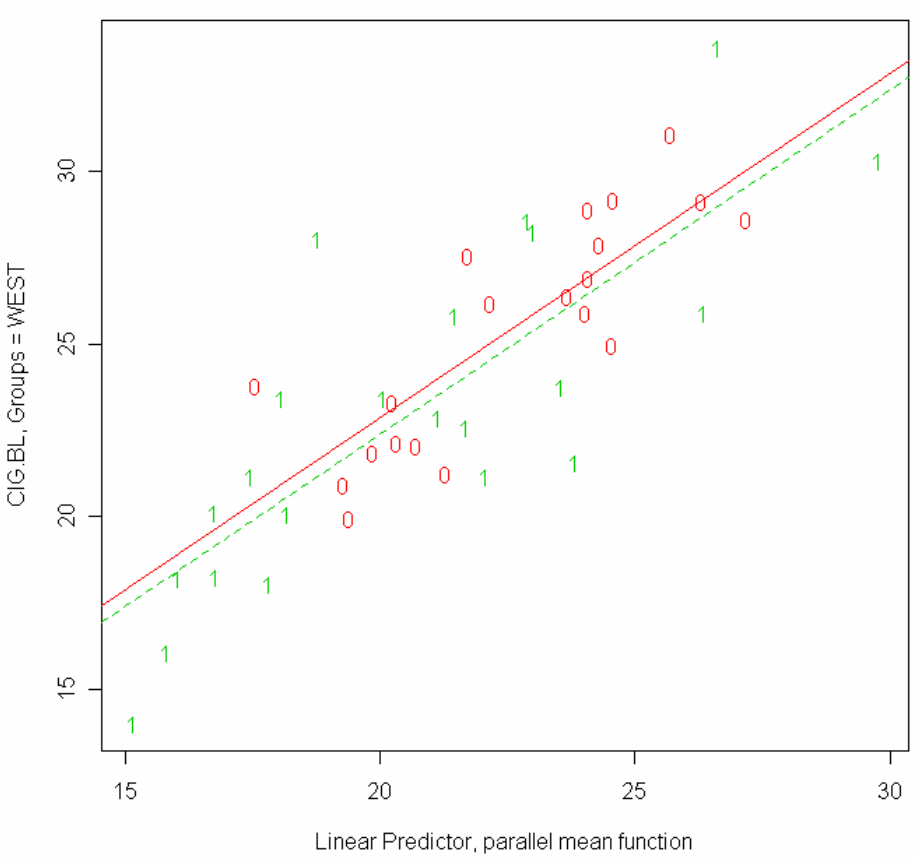


Figure 5.3.5 Response function for Model 5.3.10

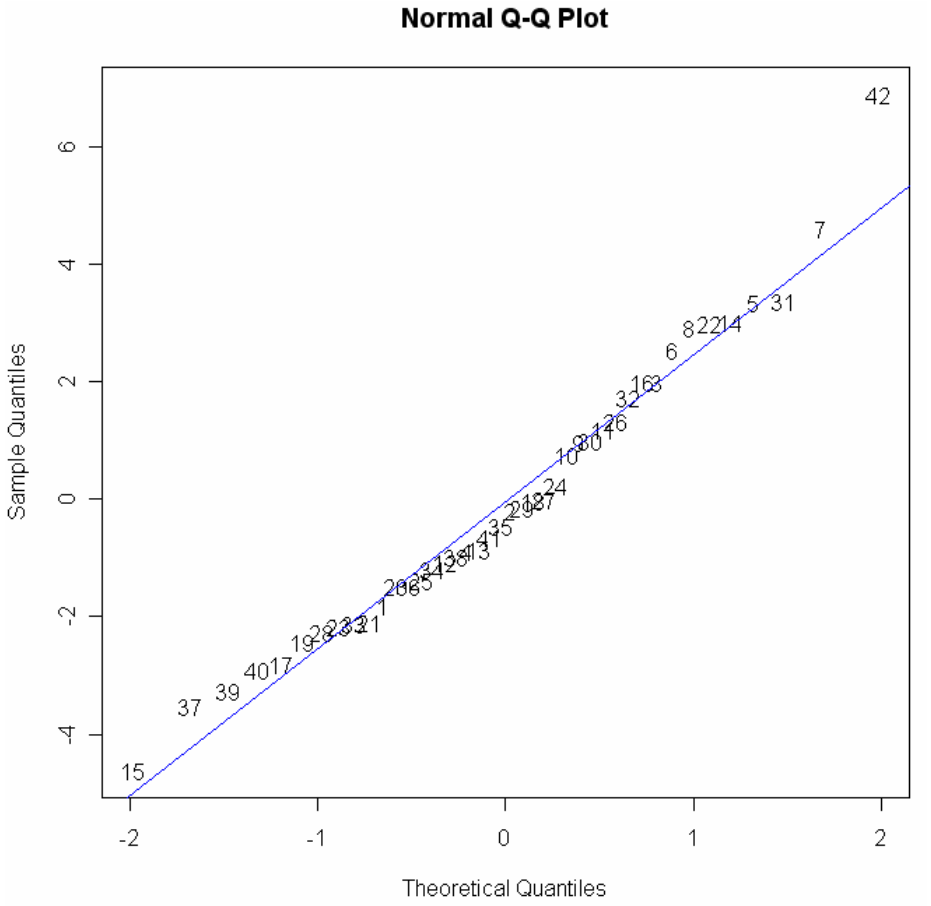


Figure 5.3.6 Normal probability plot of residuals for Model 5.3.10

第四節

由前面我們已知第 8 點觀察值為哥倫比亞特區(District of Columbia)與第 26 點觀察值為內華(Nevada)為影響點。起因於哥倫比亞特區與內華達均為光觀旅遊勝地，且哥倫比亞特區為首都故每天上班會湧入大批的通勤工作者，造成此兩州菸的銷售量較高。而我們發現當刪去此兩個影響點配適模型，解釋變數膀胱癌會變的不顯著。而地區之間的差異雖然檢定結果依然不顯著，但由 Figure 5.2.3 與 Figure 5.3.3 和 Figure 5.2.5 與 Figure 5.3.5 圖上分區的配適線，可看出刪去影響點後地區的差異凸顯出來了。且比較 Figure5.2.1 與 Figure5.3.1 可發現刪去影響點後可明顯看出模型在截距上的差異，而比較 Figure 5.2.3 與 Figure 5.3.3 可發現原本北西部與中西部在影響點之影響下是幾乎重合的兩條線。比較 Figure 5.2.5 與 Figure 5.3.5 依地區(WEST)劃分 1 為南部與西部 0 為北西部與中西部此兩區，刪去影響點後截距差異明顯更小。由 Table5.4.1 明顯的顯現出影響點對整個模型的影響不只殘差與參數檢定的部份，整個模型的解釋能力與模型預測能力都受影響。不考慮哥倫比亞特區與內華達州，我們可以看見更一般化情況。以上考慮的只是加法模型，只能比較截距上的差異，其實我們可以考慮加入交互

作用項的部份。

Table 5.4.1 為此 8 個模型之 R^2_{Adj} 、PRESS 統計量與 R^2_{pred}

	模型 5.2.1	模型 5.2.2	模型 5.2.5	模型 5.2.10	模型 5.3.1	模型 5.3.2	模型 5.3.5	模型 5.3.10
R^2_{Adj}	61.89%	68.51%	62.59%	63.51%	65.58%	82.08%	65.92%	64.86%
PRESS	612.3977	525.625	687.6004	635.5456	304.9741	169.2521	327.2979	318.6864
R^2_{pred}	54.15%	60.65%	48.82%	52.41%	59.99%	76.69%	57.06%	58.19%

第六章 總結

單一解釋變數的簡單線性回歸模式、兩解釋變數討論加入交互作用項的複回歸模式及四個解釋變數均加入的回歸模式，解釋變數白血症在模型中參數檢定均不顯著，且考慮白血症之偏回歸圖型均幾乎呈現水平直線狀態，表示當模型中已有其他解釋變數，考慮放入此變數並無提高模型解釋能力。而由 Table 2.3.1 相關係數矩陣與 Figure 2.3.2 圖形也可發現，白血症與菸頭量的線性關係度很低。因此，我們判斷解釋變數白血症應從模型中踢除。將解釋變數依其解釋模型能力高低將其依序放入模型，得到模型 2.3.2(如下式所列)為最佳模型。

$$\hat{y} = -1.1916 + 2.0544x_1 + 0.5179x_2 + 2.6701x_3$$

接著我們對模型 2.3.2 進行殘差圖型分析診斷其是否符合模型基本假設，由殘差與配適值之散佈圖，觀察出殘差變異數不一致且似乎呈曲線的狀態，而 Figure 3.3.1 殘差常態機率圖明顯觀察出離群值且為輕尾分佈(Light – tailed errors)。而由影響點之診斷，發現觀測點 8 哥倫比亞特區(District of Columbia)與觀察點 26 內華達州(Nevada)為影響點。首先我們考慮比較刪除影響點 8 與 26 這兩筆資料配適模型，試著觀察殘差非線性的情形是否為影響點造成，而我們發現殘差圖散佈情況有較均勻且殘差常態機率圖呈線性。但因為我們知道影響點之形成原因，並非資料輸入錯誤，而是起因於哥倫比亞特區與內華達均為光觀旅遊勝地，且哥倫比亞特區為首都故每天上班會湧入大批的通勤工作者，造成此兩州菸的銷售量較高。因此我們考慮使用變數轉換和加入平方項此兩種方法，希望能矯正殘差的變異數不一致的情況。故我們先嘗試藉由轉換反應變數(菸頭量)看是否能矯正殘差不一致的情況，並探討轉換後是否能減緩其影響點對殘差常態性的影響，結果發現使用轉換後的反應變數配適模型，不但解釋能力提昇，且比較未轉換模型 2.3.2 之殘差圖 Figure 3.3.1 與轉換後模型 3.3.2 之殘差圖 Figure 3.3.5，我們可看出殘差變異數不一致的情形有改善，接著比較未轉換模型 2.3.2 殘差常態機率圖 Figure 3.3.1 與轉換後模型 3.3.2 殘差常態機率圖 Figure 3.3.4，我們發現轉換後殘差常態性亦獲得改善，但內華達州(Nevada)仍為模型之影響點。於是我們更進一步的對反應變數與解釋變數做轉換，結果顯示此模型相對於僅對應變數做轉換的模型並沒有較佳的解釋能力。最後我們由逐步回歸分析之建議加入膀胱癌之平方項配適模型，為模型 3.4.2。而此模型結果和模型 2.3.2 比較下有較佳的解釋能力，但將加入平方項模型之殘差矯正情形與轉換後其殘差之矯正情況比較，我們認為轉換後之模型對殘差之矯正情形較佳。且考慮模型精簡原則，故在經模型診斷與矯正後，我們選擇之最佳模型為模型 3.3.2(如下式所列)

$$\hat{y}^{(-0.22)} = 0.6134 - 0.0070x_1 - 0.0025x_2 - 0.0143x_3$$

在第四章，我們討論利用逐步回歸分析與所有回歸式比較選模此兩種選模方法，選擇最佳模型。

當反應變數 y 未轉換下，選擇之最佳模型為模型 4.1.1(如下式所列)

$$\hat{y} = 23.5358\text{-}11.1275x_1 + 0.5443x_2 + 3.3897x_3 + 1.5126x_1^2$$

而反應變數 y 經轉換後，選擇之最佳模型為模型 4.2.3(如下式所列)

$$\hat{y}^{(-0.22)} = 0.6134 - 0.0070x_1 - 0.0025x_2 - 0.0143x_3$$

當反應變數 y 去除影響點後，選擇之最佳模型為模型 4.3.1(如下式所列)

$$\hat{y}^* = 2.3987 + 0.6308x_2^* + 3.41x_3^*$$

此結果與我們從簡單線性回歸開始配適，直至確定模型 2.3.2，檢驗其是否符合殘差之假設。由殘差不符合假設，選擇變數轉換和加入平方項矯正，所選擇之模型結果是相呼應的。但我們經由簡單線性回歸開始配適，一步步直至選擇出最佳模型 3.3.2，在這過程所呈現之統計現象，能幫助我們更適當與詳細分析此組資料之統計現象。

而第五章是考慮在模型中加入了虛擬變數 $x_4 = 1$ 表菸頭量＞平均數 23.77，$x_4 = 0$ 表菸頭量＜平均數 23.77；依地區(AREA)劃分設 $x_5 = 1$ 為北西部、$x_5 = 2$ 中西部、$x_5 = 3$ 南部與 $x_5 = 4$ 西部；依地區(WEST)劃分 $x_6 = 1$ 設南部與西部 $x_6 = 0$ 為北西部與中西部。探討由虛擬變數設定之組別間的差異，對前面章節所討論之模型是否有影響。但因為我們均只考慮加法模型，故只能比較截距上的差異。未來如果有機會再分析此組資料，我們建議可以考慮加入交互作用項，討論其在斜率上所造成之差異，這是此份報告探討較不足之地方。

參考文獻

1. Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining. Introduction to Linear Regression Analysis, 4rd Edition, 2006.
2. Fraumeni, J. F. 1968. Cigarette Smoking and Cancers of the Urinary Tract: Geographic Variations in the United States, Journal of the National Cancer Institute, 41(5): 1205-1211.
3. John Neter, Michael H Kutner, William Wasserman, Christopher J. Nachtsheim. Applied Linear Regression Models, 4rd Edition, 1996.
4. The Data and Story Library http://lib.stat.cmu.edu/

Appendix 1 SAS與R程式

【SAS Program】

```
dm'log;clear;output;clear;program;recall;';
options ps=55 ls=100 nodate pageno=1 center;
data smoking;/* y = cigar *x1 = bladder * x2 = lung    * x3 = kidney * x4 = leukemia * x5 = area
*/;
INPUT state $ y x1 x2 x3 x4 x5;
IF y >23.77 THEN smoke=1; ELSE smoke=0;
IF x5 = 3 or x5 =4 THEN west=1; ELSE west=0;
/*transformations: ystar1=y^-0.22,ystar2=y^-0.3,x1star=x1^0.13, x2star=x2^1.5, x3star=x3^0.49*/
ystar1=y**-0.22;
ystar2=y**-0.3;
x1x2=x1*x2; x1x3=x1*x3; x1x4=x1*x4; x2x3=x2*x3; x2x4=x2*x4; x3x4=x3*x4;
xx1=x1**2; xx2=x2**2; xx3=x3**2;
x1star=x1**0.13; x2star=x2**1.5;    x3star=x3**0.49;
lable y = 'cigar'      x1 = 'bladder'      x2 = 'lung'    x3 = 'kidney'    x4 = 'leukemia'
ystar1 = 'transformed cigar1'    ystar2 = 'transformed cigar2'    x1star='transformed bladder'
x2star='transformed lung'      x3star='transformed kidney'
x1x2 = 'bladder lung'          x1x3 = 'bladder kidney'          x1x4 = 'bladder leukemia'
x2x3 = 'lung kidney'           x2x4 = 'lung leukemia'           x3x4 = 'kidney leukemia'
xx1 = '2nd order in bladder'      xx2 = '2nd order in lung'      xx3 = '2nd order in kidney';
cards;
AK      30.34      3.46      25.88      4.32      4.90      3
AL      18.20      2.90      17.05      1.59      6.15      3
AZ      25.82      3.52      19.80      2.75      6.61      4
.
.
.
WI      21.25      5.14      20.55      2.34      6.73      2
WV      22.86      4.78      15.53      3.28      7.38      3
WY      28.04      3.20      15.92      2.66      5.78      4
;
proc print data=smoking;
var state y x1 x2 x3 x4 x5 smoke west;
run;
/* Part1. Simple Regression */
proc reg data= smoking;title 'Model 1 : y on x1';
model y = x1/all;
```

```
output out=r1;
run;
proc reg data= smoking;title 'Model 2 : y on x2';
model y = x2/all;
output out=r2;
run;
proc reg data= smoking;title 'Model 3 : y on x3';
model y = x3/all;
output out=r3;
run;
proc reg data= smoking;title 'Model 4 : y on x4';
model y = x4/all;
output out=r4;
run;
/* Part2. Multiple Linear Regression – two variables */
proc reg corr data= smoking;title 'Model P1a : y on x1 x2';
model y = x1 x2/all;
output out=rP1a;
run;
proc reg corr data= smoking;title 'Model P1b : y on x1 x2 x1x2';
model y = x1 x2 x1x2/all;
output out=rP1b;
run;
proc reg corr data= smoking;title 'Model P2a : y on x1 x3';
model y = x1 x3/all;
output out=rP2a;
run;
proc reg corr data= smoking;title 'Model P2b : y on x1 x3 x1x3';
model y = x1 x3 x1x3/all;
output out=rP2b;
run;
proc reg corr data= smoking;title 'Model P3a : y on x1 x4';
model y = x1 x4/all;
output out=rP3a;
run;
proc reg corr data= smoking;title 'Model P3a : y on x1 x4 x1x4';
model y = x1 x4 x1x4/all;
output out=rP3b;
run;
proc reg corr data= smoking;title 'Model P4a : y on x2 x3';
```

```
model y = x2 x3/all;
output out=rP4a;
run;
proc reg corr data= smoking;title 'Model P4b : y on x2 x3 x2x3';
model y = x2 x3 x2x3/all;
output out=rP4b;
run;
proc reg corr data= smoking;title 'Model P5a : y on x2 x4';
model y = x2 x4/all;
output out=rP5a;
run;
proc reg corr data= smoking;title 'Model P5b : y on x2 x4 x2x4';
model y = x2 x4 x2x4/all;
output out=rP5b;
run;
proc reg corr data= smoking;title 'Model P6a : y on x3 x4';
model y = x3 x4/all;
output out=rP6a;
run;
proc reg corr data= smoking;title 'Model P6a : y on x3 x4 x3x4';
model y = x3 x4 x3x4/all;
output out=rP6b;
run;
/* Part3. Multiple Linear Regression – three variables */
proc reg corr data= smoking;title 'Model 5 : y on x1 x2 x3 x4';
model y = x1 x2 x3 x4/all;
output out=r5 p=p5 r=r5 student=Ri5 Rstudent=Ti5 H=H5 PRESS=PRESS5;
run;
/*Diagnostics for Leverage and Influence. Test multicollinearity */
proc reg corr data= smoking;title 'Model 6 : y on x1 x2 x3';
model y = x1 x2 x3/influence collin vif dw;
output out=resid1 p=predy1 r=resid1 student=Ri1 Rstudent=Ti1 H=H1 PRESS=PRESS1
COOKD=COOKD1;
run;
/* Residual analysis */
proc print data=resid1;
var y predy1 resid1 Ri1 Ti1 H1 PRESS1 COOKD1;
run;
/*test if need the second order term */
proc rsreg data= smoking;title 'test if need the second order term' ;
```

```
model y = x1 x2 x3;
run;
/* stepwise1    y on x1 x2 x3 x1^2 x2^2 x3^2 check variable and model */
proc reg data= smoking;title 'Forward Selection';
model y = x1 x2 x3 xx1 xx2 xx3/selection=forward slentry=0.05;
run;
proc reg data= smoking;title 'Backward elimination';
model y = x1 x2 x3 xx1 xx2 xx3/selection=backward slstay=0.05;
run;
proc reg data= smoking;title 'Stepwise regression';
model y = x1 x2 x3 xx1 xx2 xx3/selection=stepwise slentry=0.05 slstay=0.05;
run;
proc reg data= smoking outest=est;title 'All possible regressions';
model y = x1 x2 x3 xx1 xx2 xx3/selection=rsquare adjrsq cp mse press aic;
run;
proc gplot data=est;title 'cp plot';
plot _cp_*_p_/vaxis=0 to 15 by 1.5 haxis=2 to 8 by 1;
run;
proc reg corr data= smoking;title 'Model 7 : y on x1 x2 x3 xx1';
model y = x1 x2 x3 xx1/influence collin vif dw; /*Diagnostics for Leverage and Influence.*/
output out=resid2 p=predy2 r=resid2 student=Ri2 Rstudent=Ti2 H=H2 PRESS=PRESS2
COOKD=COOKD2;
run;
/* Residual analysis */
proc print data=resid2;
var y predy2 resid2 Ri2 Ti2 H2 PRESS2 COOKD2;
run;

/* Part4. Transformation on the Response (ystar1=y^-0.22) */
proc reg corr data= smoking;title 'Model 8 : ystar1 on x1 x2 x3';
model ystar1 = x1 x2 x3/influence dw; /*Diagnostics for Leverage and Influence.*/
output out=resid3 p=predy3 r=resid3 student=Ri3 Rstudent=Ti3 H=H3 PRESS=PRESS3
COOKD=COOKD3;
run;
/* Residual analysis */
proc print data=resid3;
var ystar predy3 resid3 Ri3 Ti3 H3 PRESS3 COOKD3;
run;
/* stepwise2    ystar on x1 x2 x3 x1^2 x2^2 x3^2 check variable and model */
proc reg data= smoking;title 'Forward Selection';
```

```
model ystar1 = x1 x2 x3 xx1 xx2 xx3/selection=forward slentry=0.05;
run;
proc reg data= smoking;title 'Backward elimination';
model ystar1 = x1 x2 x3 xx1 xx2 xx3/selection=backward slstay=0.05;
run;
proc reg data= smoking;title 'Stepwise regression';
model ystar1 = x1 x2 x3 xx1 xx2 xx3/selection=stepwise slentry=0.05 slstay=0.05;
run;
proc reg data= smoking outest=est2;title 'All possible regressions';
model ystar1 = x1 x2 x3 xx1 xx2 xx3/selection=rsquare adjrsq cp mse press aic;
run;
proc gplot data=est2;title 'cp plot';
plot _cp_*_p_/vaxis=0 to 15 by 1.5 haxis=2 to 8 by 1;
run;
/* Part5a.    Transformation on the Regressor (x1star=x1^-0.13, x2star=x2^1.5, x3star=x3^0.49) */
proc reg corr data= smoking;title 'Model 8 : y on x1star x2star x3star';
model y = x1star x2star x3star/influence dw vif; /*Diagnostics for Leverage and Influence.*/
output out=resid p=predy r=resid student=Ri Rstudent=Ti H=H PRESS=PRESS
COOKD=COOKD;
run;
/* Residual analysis */
proc print data=resid;
var y predy resid Ri Ti H PRESS COOKD;
run;
/* Part5b.    Transformation on the Response and Regressor (ystar2=y^-0.3,x1star=x1^0.13,
x2star=x2^1.5, x3star=x3^0.49) */
/*Diagnostics for Leverage and Influence.*/
proc reg corr data= smoking;title 'Model 9 : ystar2 on x1star x2star x3star';
model ystar2 = x1star x2star x3star;
proc reg corr data= smoking;title 'Model 10 : ystar2 on x2star x3star';
model ystar2 = x2star x3star/influence dw;
output out=resid4 p=predy4 r=resid4 student=Ri4 Rstudent=Ti4 H=H4 PRESS=PRESS4
COOKD=COOKD4;
run;
/* Residual analysis */
proc print data=resid4;
var ystar predy4 resid4 Ri4 Ti4 H4 PRESS4 COOKD4;
run;
/* Part6. Indicator Variables Multiple Linear Regression    */
data a; set smoking;
```

```
y1=y;
IF y1 >23.77 then output;
data b; set smoking;
y2=y;
IF y2 <23.77 then output;
/*proc print data=a;
proc print data=b;
run;*/
proc reg data=a;title 'Model 11a : y1star on x1 x2 x3';
model ystar1=x1 x2 x3;
output out=resid11a;
run;
proc reg data=b;title 'Model 11b : y2star on x1 x2 x3';
model ystar1=x1 x2 x3;
output out=resid11b;
run;

/* Part7. Drop extreme points, obs 8, 26 (DC, NV) */
dm'log;clear;output;clear;program;recall;';
options ps=55 ls=100 nodate pageno=1 center;
data smoking2;/* y = cigar *x1 = bladder * x2 = lung    * x3 = kidney * x4 = leukemia * x5 = area
*/;
INPUT state $ y x1 x2 x3 x4 x5;
IF y >23.77 THEN smoke=1; ELSE smoke=0;
IF x5 = 3 or x5 =4 THEN west=1; ELSE west=0;
xx1=x1**2; xx2=x2**2; xx3=x3**2;
lable y = 'cigar'     x1 = 'bladder'     x2 = 'lung'    x3 = 'kidney'    x4 = 'leukemia'
xx1 = '2nd order in bladder'     xx2 = '2nd order in lung'     xx3 = '2nd order in kidney';
cards;
AK      30.34      3.46      25.88      4.32      4.90      3
AL      18.20      2.90      17.05      1.59      6.15      3
AZ      25.82      3.52      19.80      2.75      6.61      4
.
.
.
WI      21.25      5.14      20.55      2.34      6.73      2
WV      22.86      4.78      15.53      3.28      7.38      3
WY      28.04      3.20      15.92      2.66      5.78      4
;
proc reg corr data= smoking2;title 'Model 12 : y on x1 x2 x3 drop obs8, 26';
```

```
model y = x1 x2 x3;
proc reg corr data= smoking2;title 'Model 13 : y on x2 x3 drop obs8, 26';
model y = x2 x3/influence collin vif dw; /*Diagnostics for Leverage and Influence. */
output out=resid5 p=predy5 r=resid5 student=Ri5 Rstudent=Ti5 H=H5 PRESS=PRESS5
COOKD=COOKD5;
run;
/* Residual analysis */
proc print data=resid5;
var y predy5 resid5 Ri5 Ti5 H5 PRESS5 COOKD5;
run;
/*test if need the second order term */
proc rsreg data= smoking2;title 'test if need the second order term' ;
model y = x1 x2 x3;
run;
/* stepwise3    y on x1 x2 x3 x1^2 check variable and model drop obs8, 26 */
proc reg data= smoking2;title 'Forward Selection drop obs8, 26 ';
model y = x1 x2 x3 xx1/selection=forward slentry=0.05;
run;
proc reg data= smoking2;title 'Backward elimination drop obs8, 26';
model y = x1 x2 x3 xx1/selection=backward slstay=0.05;
run;
proc reg data= smoking2;title 'Stepwise regression drop obs8, 26';
model y = x1 x2 x3 xx1/selection=stepwise slentry=0.05 slstay=0.05;
run;
proc reg data= smoking2 outest=est3;title 'All possible regressions drop obs8, 26';
model y = x1 x2 x3 xx1/selection=rsquare adjrsq cp mse press aic;
run;
proc gplot data=est3;title 'cp plot drop obs8, 26';
plot _cp_*_p_/vaxis=0 to 15 by 1.5 haxis=2 to 8 by 1;
run;
/*Add Indicator variables*/
dm'log;clear;output;clear;program;recall;';
options ps=55 ls=100 nodate pageno=1 center;
data smoking;/* y = cigar *x1 = bladder * x2 = lung    * x3 = kidney * x4 = leukemia * x5 = area
*/;
INPUT state $ y x1-x5 IndAREA1-IndAREA4 IndCIG1 IndCIG2 IndWEST1
IndWEST2;/*IndAREA1=0;IndCIG2=0;IndWEST2=0*/;
cards;
AK  30.34    3.46 25.88    4.32 4.9  3    0    0    1    0    0    0    1    0
AL  18.2 2.9  17.05    1.59 6.15 3    0    0    1    0    1    0    1    0
```

```
AZ  25.82  3.52 19.8 2.75 6.61 4  0  0  0  1  0  0  1  0
.
.
.
WI  21.25  5.14 20.55  2.34 6.73 2  0  1  0  0  1  0  0  0
WV 22.86  4.78 15.53  3.28 7.38 3  0  0  1  0  1  0  1  0
WY 28.04  3.2  15.92  2.66 5.78 4  0  0  0  1  0  0  1  0
;
proc reg;title 'Model 14 : y on x1 x2 x3 AREA';
model y = x1-x3 IndAREA2-IndAREA4/influence collin vif dw;
run;
proc reg;title 'Model 15a : y on x1 x2 x3 smoke';
model y = x1-x3 IndCIG1/influence collin vif dw;
run;
proc reg;title 'Model 15b : y on x2 x3 smoke drop obs 8, 26';
model y = x2 x3 IndCIG1/influence collin vif dw;
output out=resid1 p=predy1 r=resid1 student=Ri1 Rstudent=Ti1 H=H1 PRESS=PRESS1
COOKD=COOKD1;
run;
/* Residual analysis */
proc print data=resid1;
var y predy1 resid1 Ri1 Ti1 H1 PRESS1 COOKD1;
run;
proc reg;title 'Model 16 : y on x1 x2 x3 WEST';
model y = x1-x3 IndWEST1/influence collin vif dw;
run;
proc reg;title 'Model 17 : y on x1 x2 x3 WEST';
model y = x1 x2 x3 IndWEST1/influence collin vif dw;
run;
proc reg;title 'Model 18 : y on x1 x2 x3 smoke Area';
model y = x1 x2 x3 IndCIG1 IndAREA2-IndAREA4/influence collin vif dw;
run;
proc reg;title 'Model 19a : y on x1 x2 x3 smoke west';
model y = x1 x2 x3 IndCIG1 IndWEST1/influence collin vif dw;
run;
proc reg;title 'Model 19b : y on x1 x3 smoke west';
model y = x1 x3 IndCIG1 IndWEST1/influence collin vif dw;
output out=resid2 p=predy2 r=resid2 student=Ri2 Rstudent=Ti2 H=H2 PRESS=PRESS2
COOKD=COOKD2;
run;
```

```
/* Residual analysis */
proc print data=resid1;
var y predy2 resid2 Ri2 Ti2 H2 PRESS2 COOKD2;
run;
/* stepwise4    y on x1 x2 x3 AREA Smoke WEST check variable */
proc reg data= smoking;title 'Forward Selection with dummy variables';
model y = x1 x2 x3 IndAREA2-IndAREA4 IndCIG1 IndWEST1/selection=forward slentry=0.05;
run;
proc reg data= smoking;title 'Backward elimination with dummy variables';
model y = x1 x2 x3 IndAREA2-IndAREA4 IndCIG1 IndWEST1/selection=backward slstay=0.05;
run;
proc reg data= smoking;title 'Stepwise regression with dummy variables';
model y = x1 x2 x3 IndAREA2-IndAREA4 IndCIG1 IndWEST1/selection=stepwise slentry=0.05
slstay=0.05;
run;
proc reg data= smoking outest=est3;title 'All possible regressions with dummy';
model y = x1 x2 x3 IndAREA2-IndAREA4 IndCIG1 IndWEST1/selection=rsquare adjrsq cp mse
press aic;
run;
/* drop obs 8, 26*/
dm'log;clear;output;clear;program;recall;';
options ps=55 ls=100 nodate pageno=1 center;
data smoking2;/* y = cigar *x1 = bladder * x2 = lung    * x3 = kidney * x4 = leukemia * x5 = area
*/;
INPUT state $ y x1-x5 IndAREA1-IndAREA4 IndCIG1 IndCIG2 IndWEST1
IndWEST2;/*IndAREA1=0;IndCIG2=0;IndWEST2=0*/;
cards;
AK  30.34    3.46 25.88    4.32 4.9  3   0   0   1   0   0   0   1   0
AL  18.2 2.9 17.05    1.59 6.15 3   0   0   1   0   1   0   1   0
AZ  25.82    3.52 19.8 2.75 6.61 4   0   0   0   1   0   0   1   0
.
.
.
WI  21.25    5.14 20.55    2.34 6.73 2   0   1   0   0   1   0   0   0
WV  22.86    4.78 15.53    3.28 7.38 3   0   0   1   0   1   0   1   0
WY  28.04    3.2  15.92    2.66 5.78 4   0   0   0   1   0   0   1   0
;
proc reg; title 'Model 20 : y on x1 x2 x3 x5 AREA drop obs 8, 26';
model y = x1 x2 x3 IndAREA2-IndAREA4/influence collin vif dw;
run;
```

```
proc reg;title 'Model 21a : y on x1 x2 x3 smoke drop obs 8, 26';
model y = x1 x2 x3 IndCIG1/influence collin vif dw;
run;
proc reg;title 'Model 21b : y on x2 x3 smoke drop obs 8, 26';
model y = x2 x3 IndCIG1/influence collin vif dw;
output out=resid1 p=predy1 r=resid1 student=Ri1 Rstudent=Ti1 H=H1 PRESS=PRESS1
COOKD=COOKD1;
run;
/* Residual analysis */
proc print data=resid1;
var y predy1 resid1 Ri1 Ti1 H1 PRESS1 COOKD1;
run;
proc reg;title 'Model 22 : y on x1 x2 x3 WEST drop obs 8, 26';
model y = x1 x2 x3 IndWEST1/influence collin vif dw;
run;
proc reg;title 'Model 23 : y on x1 x2 x3 smoke Area drop obs 8, 26';
model y = x1 x2 x3 IndCIG1 IndAREA2-IndAREA4/influence collin vif dw;
run;
proc reg;title 'Model 24 : y on x1 x2 x3 smoke west drop obs 8, 26';
model y = x1 x2 x3 IndCIG1 IndWEST1/influence collin vif dw;
run;
/* stepwise5    y on x1 x2 x3 AREA Smoke WEST check variable and model drop obs8, 26 */
proc reg data= smoking2;title 'Forward Selection with dummy variables drop obs8, 26 ';
model y = x1 x2 x3 IndAREA2-IndAREA4 IndCIG1 IndWEST1/selection=forward slentry=0.05;
run;
proc reg data= smoking2;title 'Backward elimination with dummy variables drop obs8, 26';
model y = x1 x2 x3 IndAREA2-IndAREA4 IndCIG1 IndWEST1/selection=backward slstay=0.05;
run;
proc reg data= smoking2;title 'Stepwise regression with dummy variables drop obs8, 26';
model y = x1 x2 x3 IndAREA2-IndAREA4 IndCIG1 IndWEST1/selection=stepwise slentry=0.05
slstay=0.05;
run;
proc reg data= smoking2 outest=est3;title 'All possible regressions with dummy variables drop
obs8, 26';
model y = x1 x2 x3 IndAREA2-IndAREA4 IndCIG1 IndWEST1/selection=rsquare adjrsq cp mse
press aic;
run;
```

【R Program】

```
###Regression analyisis Final Report
### need package
library(faraway)
library(acepack)
library(alr3)
library(MASS)
library(MPV)
library(scatterplot3d)
library(gplots)
library(lmtest)
library(car)
library(lattice)
library(ellipse)
###read data
smoke <- read.table("E:/reg/smoking.txt", header=TRUE)
attach(smoke)
str(smoke)
###correlaction matrix
smokecor <- cbind(CIG, BLAD, LUNG, KID, LEUK)
cor(smokecor) -> corr.mtsmokes
ord <- order(corr.mtsmokes [1,])
xc <- corr.mtsmokes [ord, ord]
colors <- c("#EFF3FF","#BDD7E7","#6BAED6","#3182BD","#08519C")
par(mfrow=c(2,2))
plotcorr(corr.mtsmokes, col=colors)
plotcorr(xc, col=colors[5*xc],numbers =TRUE)
###full model
pairs(CIG~ BLAD+LUNG+KID+LEUK+STATE)
scatterplot.matrix(~CIG+BLAD+LUNG+KID+LEUK+STATE)
ylm71 <- lm(CIG~BLAD+LUNG+KID+LEUK)
summary(ylm71)
anova(ylm71)
##simple regression model and plot
y1 <- lm(CIG~BLAD)
summary(y1)
anova(y1)
y2 <- lm(CIG~LUNG)
summary(y2)
```

```
anova(y2)
y3 <- lm(CIG~KID)
summary(y3)
anova(y3)
y4 <- lm(CIG~LEUK)
summary(y4)
anova(y4)
par(mfrow=c(2,2))
plot(CIG~BLAD, pch=16, main="Scatter Plot", xlab="Bladder Cancer", ylab="Number of
Cigarettes Smoked")
abline(y1, col=10 , lwd=2)
plot(CIG~LUNG, pch=16,main="Scatter Plot", xlab="Lung Cancer", ylab="Number of Cigarettes
Smoked")
abline(y2, col=5 , lwd=2)
plot(CIG~KID, pch=16,main="Scatter Plot", xlab="Kidney Cancer", ylab="Number of Cigarettes
Smoked")
abline(y3, col=7, lwd=2)
plot(CIG~LEUK, pch=16,main="Scatter Plot", xlab="Leukemia", ylab="Number of Cigarettes
Smoked")
abline(y4, col=6 , lwd=2)
##two regressor into model
##partial reg BLAD
y11 <- lm(CIG~BLAD+LUNG)
summary(y11)
anova(y11)
y111 <- lm(CIG~BLAD*LUNG)
summary(y111)
anova(y111)
par(mfrow=c(2,2))
chatr <- lm(CIG~BLAD)$res
lhatr <- lm(LUNG~BLAD)$res
plot(lhatr, chatr, xlab="e(LUNG|BLAD)",ylab="e(CIG|BLAD)" ,
main="Partial Regression", pch=16)
lm(chatr~lhatr)$coef
y11$coef
abline(0,y11$coef['LUNG'], col=6,lwd=2)
abline(h=0, v=0, col=2)
##partial reg BLAD
y21 <- lm(CIG~BLAD+KID)
summary(y21)
```

```
anova(y21)
y211 <- lm(CIG~BLAD*KID)
summary(y211)
anova(y211)
par(mfrow=c(2,2))
ckhatr <- lm(CIG~BLAD)$res
lkhatr <- lm(KID~BLAD)$res
plot(lkhatr, ckhatr, xlab="e(KID|BLAD)",ylab="e(CIG|BLAD)" ,
main="Partial Regression", pch=16)
lm(ckhatr~lkhatr)$coef
y21$coef
abline(0,y21$coef['KID'], col=6,lwd=2)
abline(h=0, v=0, col=2)
##partial reg BLAD
y31 <- lm(CIG~BLAD+LEUK)
summary(y31)
anova(y31)
y311 <- lm(CIG~BLAD*LEUK)
summary(y311)
anova(y311)
par(mfrow=c(2,2))
clhatr <- lm(CIG~BLAD)$res
llhatr <- lm(LEUK~BLAD)$res
plot(llhatr, clhatr, xlab="e(LEUK|BLAD)",ylab="e(CIG|BLAD)" ,
main="Partial Regression", pch=16)
lm(clhatr~llhatr)$coef
y21$coef
abline(0,y31$coef['LEUK'], col=6,lwd=2)
abline(h=0, v=0, col=2)
##partial reg LUNG
y41 <- lm(CIG~LUNG+KID)
summary(y41)
anova(y41)
y411 <- lm(CIG~LUNG*KID)
summary(y411)
anova(y411)
par(mfrow=c(2,2))
clkhatr <- lm(CIG~LUNG)$res
llkhatr <- lm(KID~LUNG)$res
plot(llkhatr, clkhatr, xlab="e(KID|LUNG)",ylab="e(CIG|LUNG)" ,
```

```
main="Partial Regression", pch=16)
lm(clkhatr~llkhatr)$coef
y21$coef
abline(0,y41$coef['KID'], col=6,lwd=2)
abline(h=0, v=0, col=2)
## partial reg LEUK
y51 <- lm(CIG~LUNG+LEUK)
summary(y51)
anova(y51)
y511 <- lm(CIG~LUNG*LEUK)
summary(y511)
anova(y511)
par(mfrow=c(2,2))
cllhatr <- lm(CIG~LUNG)$res
lllhatr <- lm(LEUK~LUNG)$res
plot(lllhatr, cllhatr, xlab="e(LEUK|LUNG)",ylab="e(CIG|LUNG)" ,
main="Partial Regression", pch=16)
lm(cllhatr~lllhatr)$coef
y21$coef
abline(0,y51$coef['LEUK'], col=6,lwd=2)
abline(h=0, v=0, col=2)
##partial reg KID
y61 <- lm(CIG~KID+LEUK)
summary(y61)
anova(y61)
y611 <- lm(CIG~KID*LEUK)
summary(y611)
anova(y611)
par(mfrow=c(2,2))
cklhatr <- lm(CIG~KID)$res
kllhatr <- lm(LEUK~KID)$res
plot(kllhatr, cklhatr, xlab="e(LEUK|KID)",ylab="e(CIG|KID)" ,
main="Partial Regression", pch=16)
lm(cklhatr~kllhatr)$coef
y21$coef
abline(0,y61$coef['LEUK'], col=6,lwd=2)
abline(h=0, v=0, col=2)
#contour plot for BLAD*LUNG
ylm1 <- lm(CIG~BLAD+LUNG)
ylm11 <- lm(CIG~BLAD*LUNG)
```

```
right.marginal <- seq(min(BLAD), max(BLAD), length = 50)
left.marginal <- seq(min(LUNG), max(LUNG), length = 50)
yes.marginal    <- list(BLAD=right.marginal, LUNG=left.marginal)
grid <- expand.grid(yes.marginal)
grid[, "fit1"] <- c(predict(ylm11, grid))
contourplot(fit1~ BLAD*LUNG, cuts=10, region = TRUE,
data=grid, main="Contour Plot")
##contour plot for BLAD*KID
ylm2 <- lm(CIG~BLAD+KID)
ylm21 <- lm(CIG~BLAD*KID)
right.marginal <- seq(min(BLAD), max(BLAD), length = 50)
left.marginal <- seq(min(KID), max(KID), length = 50)
yes.marginal    <- list(BLAD=right.marginal, KID=left.marginal)
grid <- expand.grid(yes.marginal)
grid[, "fit1"] <- c(predict(ylm21, grid))
contourplot(fit1~ BLAD*KID, cuts=10, region = TRUE,
data=grid, main="Contour Plot")
#contour plot for BLAD*LEUK
ylm3 <- lm(CIG~BLAD+LEUK)
ylm31 <- lm(CIG~BLAD*LEUK)
right.marginal <- seq(min(BLAD), max(BLAD), length = 50)
left.marginal <- seq(min(LEUK), max(LEUK), length = 50)
yes.marginal    <- list(BLAD=right.marginal, LEUK=left.marginal)
grid <- expand.grid(yes.marginal)
grid[, "fit1"] <- c(predict(ylm31, grid))
contourplot(fit1~ BLAD*LEUK, cuts=10, region = TRUE,
data=grid, main="Contour Plot")
#contour plot for LUNG*KID
ylm4 <- lm(CIG~LUNG+KID)
ylm41 <- lm(CIG~LUNG*KID)
right.marginal <- seq(min(LUNG), max(LUNG), length = 50)
left.marginal <- seq(min(KID), max(KID), length = 50)
yes.marginal    <- list(LUNG=right.marginal, KID=left.marginal)
grid <- expand.grid(yes.marginal)
grid[, "fit1"] <- c(predict(ylm41, grid))
contourplot(fit1~ LUNG*KID, cuts=10, region = TRUE,
data=grid, main="Contour Plot")
#contour plot for LUNG*LEUK
ylm5 <- lm(CIG~LUNG+LEUK)
ylm51 <- lm(CIG~LUNG*LEUK)
```

```
right.marginal <- seq(min(LUNG), max(LUNG), length = 50)
left.marginal <- seq(min(LEUK), max(LEUK), length = 50)
yes.marginal    <- list(LUNG=right.marginal, LEUK=left.marginal)
grid <- expand.grid(yes.marginal)
grid[, "fit1"] <- c(predict(ylm51, grid))
contourplot(fit1~ LUNG*LEUK, cuts=10, region = TRUE,
data=grid, main="Contour Plot")
## contour plot for KID*LEUK
ylm6 <- lm(CIG~KID+LEUK)
ylm61 <- lm(CIG~KID*LEUK)
right.marginal <- seq(min(KID), max(KID), length = 50)
left.marginal <- seq(min(LEUK), max(LEUK), length = 50)
yes.marginal    <- list(KID=right.marginal, LEUK=left.marginal)
grid <- expand.grid(yes.marginal)
grid[, "fit1"] <- c(predict(ylm61, grid))
contourplot(fit1~ KID*LEUK, cuts=10, region = TRUE,
data=grid, main="Contour Plot")
##three regressor into model/Scatterplot matrix, model fitted and Residual plot
par(mfrow=c(2,2))
pairs(CIG~ BLAD+LUNG+KID, pch=16)
scatterplot.matrix(~CIG+BLAD+LUNG+KID, pch=16)
ylm7 <- lm(CIG~BLAD+LUNG+KID)
PRESS(ylm7)
summary(ylm7)
anova(ylm7)
par(mfrow=c(2,2))
plot(ylm7$residuals~ylm7$fitted, xlab="Fitted Value", ylab="Residual", pch=16,
main="Fitted Value vs Residual Plot")
abline(h=0, col=2)
plot(ylm7$residuals~BLAD, xlab="BLAD", ylab="Residual",pch=16,
main="BLAD vs Residual Plot")
abline(h=0, col=2)
plot(ylm7$residuals~LUNG, xlab="LUNG", ylab="Residual",pch=16,
main="LUNG vs Residual Plot")
abline(h=0, col=2)
plot(ylm7$residuals~KID, xlab="KID", ylab="Residual",pch=16,
main="KID vs Residual Plot")
abline(h=0, col=2)
par(mfrow=c(1,1))
qqnorml(ylm7$res, main = "Normal Q-Q Plot",
```

```
        xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
        plot.it = TRUE)
qqline(ylm7$res, col=4)
plot(ylm7,2)
plot(ylm7,4)
## transformtion response
inverse.response.plot(ylm7, key=TRUE, ylab=expression(hat(y)))
boxcox(ylm7)
abline(v=-0.22, col=4)
text(-0.20, -135, c("Box-Cox Transformation"), col=3)
text(-0.20, -137, expression(lambda== -0.22), col=3)
## Scatterplot matrix, model fitted and Residual plot for transformtion response
tranG <- (CIG)^-0.22
pairs(tranG~ BLAD+LUNG+KID, pch=16)
scatterplot.matrix(~tranG+BLAD+LUNG+KID, pch=16)
ylm8 <- lm(tranG~BLAD+LUNG+KID)
PRESS(ylm8)
summary(ylm8)
anova(ylm8)
par(mfrow=c(2,2))
plot(ylm8$residuals~ylm8$fitted, xlab="Fitted Value", ylab="Residual", pch=16,
main="Fitted Value vs Residual Plot")
abline(h=0, col=2)
plot(ylm8$residuals~BLAD, xlab="BLAD", ylab="Residual",pch=16,
main="BLAD vs Residual Plot")
abline(h=0, col=2)
plot(ylm8$residuals~LUNG, xlab="LUNG", ylab="Residual",pch=16,
main="LUNG vs Residual Plot")
abline(h=0, col=2)
plot(ylm8$residuals~KID, xlab="KID", ylab="Residual",pch=16,
main="KID vs Residual Plot")
abline(h=0, col=2)
par(mfrow=c(1,1))
qqnorml(ylm8$res, main = "Normal Q-Q Plot",
        xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
        plot.it = TRUE)
qqline(ylm8$res, col=4)
inf.index(ylm8,pch=16, col=3)
inf.index.lm(ylm8)
vif(ylm8)
```

```
## model fitted and Residual plot for add second-order polynomial
DBLAD <- (BLAD)^2
ylm9 <- lm(CIG~BLAD+DBLAD+LUNG+KID)
PRESS(ylm9)
summary(ylm9)
anova(ylm9)
par(mfrow=c(2,2))
plot(ylm9$residuals~ylm9$fitted, xlab="Fitted Value", ylab="Residual", pch=16,
main="Fitted Value vs Residual Plot")
abline(h=0, col=2)
plot(ylm9$residuals~BLAD, xlab="BLAD", ylab="Residual",pch=16,
main="BLAD vs Residual Plot")
abline(h=0, col=2)
plot(ylm9$residuals~DBLAD, xlab="DBLAD", ylab="Residual",pch=16,
main="DBLAD vs Residual Plot")
abline(h=0, col=2)
plot(ylm9$residuals~LUNG, xlab="LUNG", ylab="Residual",pch=16,
main="LUNG vs Residual Plot")
abline(h=0, col=2)
plot(ylm9$residuals~KID, xlab="KID", ylab="Residual",pch=16,
main="KID vs Residual Plot")
abline(h=0, col=2)
par(mfrow=c(1,1))
qqnorml(ylm9$res, main = "Normal Q-Q Plot",
        xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
        plot.it = TRUE)
qqline(ylm9$res, col=4)
inf.index(ylm9,pch=16, col=3)
avp(ylm9,one.page=TRUE,ask=FALSE,identify.points=FALSE)
plot(ylm9,2)
plot(ylm9,4)
## model fitted for transformtion response and add second-order polynomial
ylm10 <- lm(tranG~BLAD+DBLAD+LUNG+KID)
PRESS(ylm10)
summary(ylm10)
anova(ylm10)
## model fitted for transformtion response and transformation regressor
summary(regressor <- bctrans(CIG~ BLAD+LUNG+KID))
lrt.bctrans(regressor,)
plot(regressor,family="power", pch=16)
```

```
tranGa <- (CIG)^-0.3
tranBa <- (BLAD)^-0.13
tranLa <- (LUNG)^1.5
tranKa <- (KID)^0.5
## Scatterplot matrix, model fitted for transformtion response and transformation regressor
pairs(tranGa~tranBa+tranLa+tranKa, pch=16)
scatterplot.matrix(~tranGa+tranBa+tranLa+tranKa, pch=16)
ylm11 <- lm(tranGa~tranBa+tranLa+tranKa)
PRESS(ylm11)
summary(ylm11)
anova(ylm11)
par(mfrow=c(2,2))
plot(ylm11$residuals~ylm9$fitted, xlab="Fitted Value", ylab="Residual", pch=16,
main="Fitted Value vs Residual Plot")
abline(h=0, col=2)
plot(ylm11$residuals~tranB, xlab="BLAD^-0.13", ylab="Residual",pch=16,
main="BLAD^-0.13 vs Residual Plot")
abline(h=0, col=2)
plot(ylm11$residuals~tranL, xlab="LUNG^1.5", ylab="Residual",pch=16,
main="LUNG^1.5 vs Residual Plot)
abline(h=0, col=2)
plot(ylm11$residuals~tranK, xlab="KID^0.5", ylab="Residual",pch=16,
main="KID^0.5 vs Residual Plot)
abline(h=0, col=2)
##Plot of Dummy Variable
factor(IndCIG) -> indexx
par(mfrow=c(2,2))
plot(tranG~BLAD, pch=unclass(indexx),col=unclass(indexx))
legend(5,0.54, c(expression(CIG<23.77), expression(CIG>23.77)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(tranG~LUNG, pch=unclass(indexx),col=unclass(indexx))
legend(22,0.54, c(expression(CIG<23.77), expression(CIG>23.77)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(tranG~KID, pch=unclass(indexx),col=unclass(indexx))
legend(3.5,0.54, c(expression(CIG<23.77), expression(CIG>23.77)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(tranG~ LEUK, pch=unclass(indexx),col=unclass(indexx))
legend(7.3,0.54, c(expression(CIG<23.77), expression(CIG>23.77)),
                col=1:2, pch=1:2, bty="n", cex=0.8)
par(mfrow=c(2,2))
```

```
plot(CIG~BLAD, pch=unclass(indexx),col=unclass(indexx),main="CIG vs BLAD Scatter Plot")
legend(3.5,40, c(expression(CIG<23.77), expression(CIG>23.77)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(CIG~LUNG, pch=unclass(indexx),col=unclass(indexx), main="CIG vs LUNG Scatter Plot")
legend(15,40, c(expression(CIG<23.77), expression(CIG>23.77)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(CIG~KID, pch=unclass(indexx),col=unclass(indexx),, main="CIG vs KID Scatter Plot")
legend(2.0, 40, c(expression(CIG<23.77), expression(CIG>23.77)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(CIG~ LEUK, pch=unclass(indexx),col=unclass(indexx),, main="CIG vs LEUK Scatter Plot")
legend(5.5,40, c(expression(CIG<23.77), expression(CIG>23.77)),
                col=1:2, pch=1:2, bty="n", cex=0.8)
factor(AREA) -> index2
par(mfrow=c(2,2))
plot(CIG~BLAD, pch=unclass(index2),col=unclass(index2), main="CIG vs BLAD Scatter Plot")
legend(3.5,40, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),
                col=1:4, pch=1:4, bty="n",cex = 0.8)
plot(CIG~LUNG, pch=unclass(index2),col=unclass(index2), main="CIG vs LUNG Scatter Plot")
legend(15,40, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),
                col=1:4, pch=1:4, bty="n",cex = 0.8)
plot(CIG~KID, pch=unclass(index2),col=unclass(index2),, main="CIG vs KID Scatter Plot")
legend(2.0, 40, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),
                col=1:4, pch=1:4, bty="n",cex = 0.8)
plot(CIG~ LEUK, pch=unclass(index2),col=unclass(index2),, main="CIG vs LEUK Scatter Plot")
legend(5.3,40, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),
                col=1:4, pch=1:4, bty="n", cex=0.8)
factor(WEST) -> index3
par(mfrow=c(2,2))
plot(CIG~BLAD, pch=unclass(index3),col=unclass(index3), main="CIG vs BLAD Scatter Plot")
legend(3.5,40, c(expression(WEST==0), expression(WEST==1)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(CIG~LUNG, pch=unclass(index3),col=unclass(index3), main="CIG vs LUNG Scatter Plot")
legend(15,40, c(expression(WEST==0), expression(WEST==1)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(CIG~KID, pch=unclass(index3),col=unclass(index3),, main="CIG vs KID Scatter Plot")
legend(2.0, 40, c(expression(WEST==0), expression(WEST==1)),
```

```
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(CIG~ LEUK, pch=unclass(index3),col=unclass(index3),, main="CIG vs LEUK Scatter Plot")
legend(5.5,40, c(expression(WEST==0), expression(WEST==1)),
                col=1:2, pch=1:2, bty="n", cex=0.8)
##CIG>23.77 Plot of Dummy Variable
hsmoke <- read.table("E:/reg/hsmoking.txt", header=TRUE)
attach(hsmoke)
str(hsmoke)
par(mfrow=c(2,2))
plot(HCIG~HBLAD, pch=unclass(index2),col=unclass(index2), main="HCIG vs BLAD Scatter
Plot")
legend(3.5,40, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),
                col=1:4, pch=1:4, bty="n",cex = 0.8)
plot(HCIG~HLUNG, pch=unclass(index2),col=unclass(index2), main="HCIG vs LUNG Scatter
Plot")
legend(15,40, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),
                col=1:4, pch=1:4, bty="n",cex = 0.8)
plot(HCIG~HKID, pch=unclass(index2),col=unclass(index2),, main="HCIG vs KID Scatter Plot")
legend(2.5, 40, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),
                col=1:4, pch=1:4, bty="n",cex = 0.8)
plot(HCIG~ HLEUK, pch=unclass(index2),col=unclass(index2),, main="HCIG vs LEUK Scatter
Plot")
legend(5.3,40, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),
                col=1:4, pch=1:4, bty="n", cex=0.8)
par(mfrow=c(2,2))
plot(HCIG~HBLAD, pch=unclass(index3),col=unclass(index3), main="HCIG vs BLAD Scatter
Plot")
legend(3.5,40, c(expression(WEST==0), expression(WEST==1)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(HCIG~HLUNG, pch=unclass(index3),col=unclass(index3), main="HCIG vs LUNG Scatter
Plot")
legend(15,40, c(expression(WEST==0), expression(WEST==1)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
plot(HCIG~HKID, pch=unclass(index3),col=unclass(index3),, main="HCIG vs KID Scatter Plot")
legend(2.5, 40, c(expression(WEST==0), expression(WEST==1)),
                col=1:2, pch=1:2, bty="n",cex = 0.8)
```

```
plot(HCIG~ HLEUK, pch=unclass(index3),col=unclass(index3),, main="HCIG vs LEUK Scatter
Plot")
legend(5.5,40, c(expression(WEST==0), expression(WEST==1)),
                col=1:2, pch=1:2, bty="n", cex=0.8)
##CIG<23.77 Plot of Dummy Variable
lsmoke <- read.table("E:/reg/lsmoking.txt", header=TRUE)
attach(lsmoke)
str(lsmoke)
par(mfrow=c(2,2))
plot(LCIG~LBLAD, pch=unclass(index2),col=unclass(index2), main="LCIG vs BLAD Scatter
Plot")
legend(4.0,18, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),col=1:4, pch=1:4, bty="n",cex = 0.8)
plot(LCIG~LLUNG, pch=unclass(index2),col=unclass(index2), main="LCIG vs LUNG Scatter
Plot")
legend(20,18, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),col=1:4, pch=1:4, bty="n",cex = 0.8)
plot(LCIG~LKID, pch=unclass(index2),col=unclass(index2),, main="LCIG vs KID Scatter Plot")
legend(3.0, 18, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)),col=1:4, pch=1:4, bty="n",cex = 0.8)
plot(LCIG~ LLEUK, pch=unclass(index2),col=unclass(index2),, main="LCIG vs LEUK Scatter
Plot")
legend(7.5,18, c(expression(AREA==1), expression(AREA==2), expression(AREA==3),
expression(AREA==4)), col=1:4, pch=1:4, bty="n", cex=0.8)
par(mfrow=c(2,2))
plot(LCIG~LBLAD, pch=unclass(index3),col=unclass(index3), main="LCIG vs BLAD Scatter
Plot")
legend(4.0,18, c(expression(WEST==0), expression(WEST==1)), col=1:2, pch=1:2, bty="n",cex =
0.8)
plot(LCIG~LLUNG, pch=unclass(index3),col=unclass(index3), main="LCIG vs LUNG Scatter
Plot")
legend(20,18, c(expression(WEST==0), expression(WEST==1)), col=1:2, pch=1:2, bty="n",cex =
0.8)
plot(LCIG~LKID, pch=unclass(index3),col=unclass(index3),, main="LCIG vs KID Scatter Plot")
legend(3.0, 18, c(expression(WEST==0), expression(WEST==1)), col=1:2, pch=1:2, bty="n",cex =
0.8)
plot(LCIG~ LLEUK, pch=unclass(index3),col=unclass(index3),, main="LCIG vs LEUK Scatter
Plot")
legend(7.5,18, c(expression(WEST==0), expression(WEST==1)), col=1:2, pch=1:2, bty="n",
cex=0.8)
```

```
## model for Dummy Variable CIG>23.77 and CIG<23.77
mt <- pod(CIG~BLAD+LUNG+KID, data=smoke, group=IndCIG, mean.function="parallel")
lm(CIG~BLAD+LUNG+KID) -> fitt
summary(fitt)
anova(fitt)
pod.lm(fitt, group=IndCIG, mean.function="parallel") -> fitt3
factor(IndCIG) -> indexx
plot(fitt3,pch=c("0", "1"),colors=c(2:3))
plot.pod(fitt3, colors=rainbow(nlevels(fitt3$group)),
   pch=1:nlevels(fitt3$group),key=TRUE,identify=TRUE,
   xlab="Linear Predictor", ylab=as.character(c(formula(fitt3)[[2]])))
lm(CIG~BLAD+LUNG+KID+factor(IndCIG)) -> fitt22
PRESS(fitt22)
summary(fitt22)
anova(fitt22)
lm(CIG~BLAD+LUNG+KID+I(IndCIG)) -> fitt23
PRESS(fitt23)
summary(fitt23)
anova(fitt23)
plot(fitt22$residuals~fitt22$fitted, xlab="Fitted Value", ylab="Residual", pch=16,
main="Fitted Value vs Residual Plot")
abline(h=0, col=2)
qqnorml(fitt22$res, main = "Normal Q-Q Plot",
          xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
          plot.it = TRUE)
qqline(fitt22$res, col=4)
## model for Dummy Variable AREA 1 2 3 4
mt1 <- pod(CIG~BLAD+LUNG+KID, data=smoke, group=AREA, mean.function="parallel")
pod.lm(fitt, group=AREA, mean.function="parallel") -> fitt4
factor(AREA) -> index2
plot(fitt4,pch=c("1","2","3","4"),colors=c(2:5))
plot.parallel(fitt4, colors=rainbow(nlevels(fitt4$group)),
   pch=1:nlevels(fitt4$group),key=TRUE,identify=TRUE,
   xlab="Linear Predictor", ylab=as.character(c(formula(fitt4)[[2]])))
lm(CIG~BLAD+LUNG+KID+factor(AREA)) -> fitt33
PRESS(fitt33)
summary(fitt33)
anova(fitt33)
qqnorml(fitt33$res, main = "Normal Q-Q Plot",
          xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
```

```
          plot.it = TRUE)
qqline(fitt33$res, col=4)
## model for Dummy Variable WEST 1 0
mt2 <- pod(CIG~BLAD+LUNG+KID, data=smoke, group=WEST, mean.function="parallel")
pod.lm(fitt, group=WEST, mean.function="parallel") -> fitt5
plot(fitt5,pch=c("0", "1"),colors=c(2:3))
plot.pod(fitt5, colors=rainbow(nlevels(fitt5$group)),
   pch=1:nlevels(fitt5$group),key=TRUE,identify=FALSE,
   xlab="Linear Predictor", ylab=as.character(c(formula(fitt5)[[2]])))
lm(CIG~BLAD+LUNG+KID+factor(WEST)) -> fitt44
PRESS(fitt44)
summary(fitt44)
anova(fitt44)
qqnorml(fitt44$res, main = "Normal Q-Q Plot",
          xlab = "Theoretical Quantiles", ylab = "Sample Quantiles",
          plot.it = TRUE)
qqline(fitt44$res, col=4)
##drop obs8 and obs 26    the similarly the above
##Plot about the step regression
sel <- read.table("E:/reg/selection1.txt", header=TRUE)
attach(sel)
str(sel)
plot(C.p.~P, xlab=expression(p), ylab=expression(C[p]), pch=16)
text(4.1, 13, expression(X[1]),col=4,cex=0.8)
text(4.2, 13, expression(X[2]),col=4,cex=0.8)
text(4.3, 13, expression(X[3]),col=4,cex=0.8)
text(5.1, 5.8, expression(X[1]),col=4,cex=0.8)
text(5.2, 5.8, expression(X[2]),col=4,cex=0.8)
text(5.3, 5.8, expression(X[3]),col=4,cex=0.8)
text(5.4, 5.8, expression(X[1]^2),col=4,cex=0.8)

plot(MSE~P, xlab=expression(p), ylab=expression(MS[Res](p)), pch=16)
text(4.1,11.8,    expression(X[1]),col=4,cex=0.8)
text(4.2,11.8, expression(X[2]),col=4,cex=0.8)
text(4.3,11.8,    expression(X[3]),col=4,cex=0.8)
text(5.1, 9.8, expression(X[1]),col=4,cex=0.8)
text(5.2, 9.8, expression(X[2]),col=4,cex=0.8)
text(5.3, 9.8, expression(X[3]),col=4,cex=0.8)
text(5.4, 9.8, expression(X[1]^2),col=4,cex=0.8)
plot(R.Square~P, ann=FALSE, pch=16)
```

```
mtext(expression(R[p]^2), side=2, at=.5, line=2.5)
mtext(expression(p), side=1, at=4.5, line=2.5)
text(4.1, 0.65, expression(X[1]),col=4,cex=0.8)
text(4.2, 0.65, expression(X[2]),col=4,cex=0.8)
text(4.3, 0.65, expression(X[3]),col=4,cex=0.8)
text(5.1, 0.71, expression(X[1]),col=4,cex=0.8)
text(5.2, 0.71, expression(X[2]),col=4,cex=0.8)
text(5.3, 0.71, expression(X[3]),col=4,cex=0.8)
text(5.4, 0.71, expression(X[1]^2),col=4,cex=0.8)
se2 <- read.table("E:/reg/selection2.txt", header=TRUE)
attach(se2)
str(se2)
plot(C.p.~P, xlab=expression(p), ylab=expression(C[p]), pch=16)
text(4.1, 12.2, expression(X[1]),col=4,cex=0.8)
text(4.2, 12.2, expression(X[2]),col=4,cex=0.8)
text(4.3, 12.2, expression(X[3]),col=4,cex=0.8)
text(5.1, 10.4, expression(X[1]),col=4,cex=0.8)
text(5.2, 10.4, expression(X[2]),col=4,cex=0.8)
text(5.3, 10.4, expression(X[3]),col=4,cex=0.8)
text(5.4, 10.4, expression(X[1]^2),col=4,cex=0.8)
text(4.6, 9.6, expression(X[2]),col=4,cex=0.8)
text(4.7, 9.6, expression(X[3]),col=4,cex=0.8)
text(4.8, 9.6, expression(X[2]^2),col=4,cex=0.8)
text(4.9, 9.6, expression(X[3]^2),col=4,cex=0.8)
plot(MSE~P, xlab=expression(p), ylab=expression(MS[Res](p)), pch=16)
text(4.1, 0.00018, expression(X[1]),col=4,cex=0.8)
text(4.2, 0.00018, expression(X[2]),col=4,cex=0.8)
text(4.3, 0.00018, expression(X[3]),col=4,cex=0.8)
text(5.1, 0.00018, expression(X[1]),col=4,cex=0.8)
text(5.2, 0.00018, expression(X[2]),col=4,cex=0.8)
text(5.3, 0.00018, expression(X[3]),col=4,cex=0.8)
text(5.4, 0.00018, expression(X[1]^2),col=4,cex=0.8)
text(4.6, 0.000175, expression(X[2]),col=4,cex=0.8)
text(4.7, 0.000175, expression(X[3]),col=4,cex=0.8)
text(4.8, 0.000175, expression(X[2]^2),col=4,cex=0.8)
text(4.9, 0.000175, expression(X[3]^2),col=4,cex=0.8)
plot(R.Square~P, ann=FALSE, pch=16)
mtext(expression(R[p]^2), side=2, at=.5, line=2.5)
mtext(expression(p), side=1, at=4.5, line=2.5)
text(4.1, 0.68, expression(X[1]),col=4,cex=0.8)
```

```
text(4.2, 0.68, expression(X[2]),col=4,cex=0.8)
text(4.3, 0.68, expression(X[3]),col=4,cex=0.8)
text(5.1, 0.7, expression(X[1]),col=4,cex=0.8)
text(5.2, 0.7, expression(X[2]),col=4,cex=0.8)
text(5.3, 0.7, expression(X[3]),col=4,cex=0.8)
text(5.4, 0.7, expression(X[1]^2),col=4,cex=0.8)
text(4.6, 0.71, expression(X[2]),col=4,cex=0.8)
text(4.7, 0.71, expression(X[3]),col=4,cex=0.8)
text(4.8, 0.71, expression(X[2]^2),col=4,cex=0.8)
text(4.9, 0.71, expression(X[3]^2),col=4,cex=0.8)
se3 <- read.table("E:/reg/selection3.txt", header=TRUE)
attach(se3)
str(se3)
par(mfrow=c(1,1))
plot(C.p.~P, xlab=expression(p), ylab=expression(C[p]), pch=16)
text(2.1, 20, expression(X[2]),col=4,cex=0.8)
text(2.1, 36, expression(X[1]),col=4,cex=0.8)
text(2.1, 38, expression(X[1]^2),col=4,cex=0.8)
text(2.1, 40, expression(X[3]),col=4,cex=0.8)
text(3.1, 2.3, expression(X[2]),col=4,cex=0.8)
text(3.2, 2.3, expression(X[3]),col=4,cex=0.8)
text(3.1, 16, expression(X[1]),col=4,cex=0.8)
text(3.2, 16, expression(X[2]),col=4,cex=0.8)
text(3.1, 17, expression(X[1]^2),col=4,cex=0.8)
text(3.2, 17, expression(X[2]),col=4,cex=0.8)
text(3.1, 20.5, expression(X[1]),col=4,cex=0.8)
text(3.2, 20.6, expression(X[3]),col=4,cex=0.8)
text(3.1, 21.5, expression(X[1]^2),col=4,cex=0.8)
text(3.2, 21.6, expression(X[3]),col=4,cex=0.8)
text(3.1, 36, expression(X[1]),col=4,cex=0.8)
text(3.2, 36.1, expression(X[1]^2),col=4,cex=0.8)
text(4.1, 2, expression(X[1]),col=4,cex=0.8)
text(4.2, 2, expression(X[2]),col=4,cex=0.8)
text(4.3, 2, expression(X[3]),col=4,cex=0.8)
text(4.1, 3.45, expression(X[2]),col=4,cex=0.8)
text(4.2, 3.45, expression(X[3]),col=4,cex=0.8)
text(4.3, 3.45, expression(X[1]^2),col=4,cex=0.8)
text(4.1, 17, expression(X[1]),col=4,cex=0.8)
text(4.2, 17, expression(X[2]),col=4,cex=0.8)
text(4.3, 17, expression(X[1]^2),col=4,cex=0.8)
```

```
text(4.1, 23, expression(X[1]),col=4,cex=0.8)
text(4.2, 23, expression(X[3]),col=4,cex=0.8)
text(4.3, 23, expression(X[1]^2),col=4,cex=0.8)
text(4.6, 5, expression(X[1]),col=4,cex=0.8)
text(4.7, 5, expression(X[2]),col=4,cex=0.8)
text(4.8, 5, expression(X[3]),col=4,cex=0.8)
text(4.9, 5, expression(X[1]^2),col=4,cex=0.8)
plot(MSE~P, xlab=expression(p), ylab=expression(MS[Res](p)), pch=16)
text(2.1, 9.2, expression(X[2]),col=4,cex=0.8)
text(2.1, 12, expression(X[1]),col=4,cex=0.8)
text(2.1, 12.3, expression(X[1]^2),col=4,cex=0.8)
text(2.1, 12.6, expression(X[3]),col=4,cex=0.8)
text(3.1, 6.3, expression(X[2]),col=4,cex=0.8)
text(3.2, 6.3, expression(X[3]),col=4,cex=0.8)
text(3.1, 8.8, expression(X[1]),col=4,cex=0.8)
text(3.2, 8.8, expression(X[2]),col=4,cex=0.8)
text(3.1, 9.0, expression(X[1]^2),col=4,cex=0.8)
text(3.2, 9.0, expression(X[2]),col=4,cex=0.8)
text(3.1, 9.5, expression(X[1]),col=4,cex=0.8)
text(3.2, 9.5, expression(X[3]),col=4,cex=0.8)
text(3.1, 9.8, expression(X[1]^2),col=4,cex=0.8)
text(3.2, 9.8, expression(X[3]),col=4,cex=0.8)
text(3.1, 12.1, expression(X[1]),col=4,cex=0.8)
text(3.2, 12.1, expression(X[1]^2),col=4,cex=0.8)
text(4.1, 6.4, expression(X[1]),col=4,cex=0.8)
text(4.2, 6.4, expression(X[2]),col=4,cex=0.8)
text(4.3, 6.4, expression(X[3]),col=4,cex=0.8)
text(4.1, 6.6, expression(X[2]),col=4,cex=0.8)
text(4.2, 6.6, expression(X[3]),col=4,cex=0.8)
text(4.3, 6.6, expression(X[1]^2),col=4,cex=0.8)
text(4.1, 8.7, expression(X[1]),col=4,cex=0.8)
text(4.2, 8.7, expression(X[2]),col=4,cex=0.8)
text(4.3, 8.7, expression(X[1]^2),col=4,cex=0.8)
text(4.1, 9.7, expression(X[1]),col=4,cex=0.8)
text(4.2, 9.7, expression(X[3]),col=4,cex=0.8)
text(4.3, 9.7, expression(X[1]^2),col=4,cex=0.8)
text(4.6, 6.5, expression(X[1]),col=4,cex=0.8)
text(4.7, 6.5, expression(X[2]),col=4,cex=0.8)
text(4.8, 6.5, expression(X[3]),col=4,cex=0.8)
text(4.9, 6.5, expression(X[1]^2),col=4,cex=0.8)
```

```
plot(R.Square~P,ann=FALSE, pch=16)
mtext(expression(R[p]^2), side=2, at=.5, line=2.5)
mtext(expression(p), side=1, at=3.5, line=2.5)
text(2.1, 0.51, expression(X[2]),col=4,cex=0.8)
text(2.1, 0.37, expression(X[1]),col=4,cex=0.8)
text(2.1, 0.35, expression(X[1]^2),col=4,cex=0.8)
text(2.1, 0.33, expression(X[3]),col=4,cex=0.8)
text(3.1, 0.67, expression(X[2]),col=4,cex=0.8)
text(3.2, 0.67, expression(X[3]),col=4,cex=0.8)
text(3.1, 0.55, expression(X[1]),col=4,cex=0.8)
text(3.2, 0.55, expression(X[2]),col=4,cex=0.8)
text(3.1, 0.54, expression(X[1]^2),col=4,cex=0.8)
text(3.2, 0.54, expression(X[2]),col=4,cex=0.8)
text(3.1, 0.52, expression(X[1]),col=4,cex=0.8)
text(3.2, 0.52, expression(X[3]),col=4,cex=0.8)
text(3.1, 0.5, expression(X[1]^2),col=4,cex=0.8)
text(3.2, 0.5, expression(X[3]),col=4,cex=0.8)
text(3.1, 0.38, expression(X[1]),col=4,cex=0.8)
text(3.2, 0.38, expression(X[1]^2),col=4,cex=0.8)
text(4.1, 0.685, expression(X[1]),col=4,cex=0.8)
text(4.2, 0.685, expression(X[2]),col=4,cex=0.8)
text(4.3, 0.685, expression(X[3]),col=4,cex=0.8)
text(4.1, 0.675, expression(X[2]),col=4,cex=0.8)
text(4.2, 0.675, expression(X[3]),col=4,cex=0.8)
text(4.3, 0.675, expression(X[1]^2),col=4,cex=0.8)
text(4.1, 0.56, expression(X[1]),col=4,cex=0.8)
text(4.2, 0.56, expression(X[2]),col=4,cex=0.8)
text(4.3, 0.56, expression(X[1]^2),col=4,cex=0.8)
text(4.1, 0.52, expression(X[1]),col=4,cex=0.8)
text(4.2, 0.52, expression(X[3]),col=4,cex=0.8)
text(4.3, 0.52, expression(X[1]^2),col=4,cex=0.8)
text(4.6, 0.68, expression(X[1]),col=4,cex=0.8)
text(4.7, 0.68, expression(X[2]),col=4,cex=0.8)
text(4.8, 0.68, expression(X[3]),col=4,cex=0.8)
text(4.9, 0.68, expression(X[1]^2),col=4,cex=0.8)
```

Appendix 2 報告花絮

組員介紹

橙莉　應統所博一　P9522017

威麟　統精所碩二　M9416505

美惠　統精所碩二　M9416481

雅竹　統精所碩二　M9416494

玲慧　統精所碩二　M9431905

愉翔　統精所碩二　M9485005

工作分配

橙莉：督導，主要負責Chapter 5、Chapter 6、R程式及圖形部分。

威麟、雅竹：主要負責Chapter 3、報表整理與流程圖。

美惠：主要負責Chapter 4、SAS程式及Appendix。

玲慧：主要負責Chapter 2與報表整理。

愉翔：主要負責Chapter 1。

報告的目標

不奢求寫得完美到送至圖書館留存，只希望老師看到我們這組的報告，想不給100分都難。

報告的感想

橙莉： 最值得稱讚的地方是，我們按著老師給的時間表陸續完成。經歷多次的分工、溝通、協調、整合與互助，盡力的完成此份報告。報告內容或許有不夠完整的地方，但這是「我們的報告」。

威麟：經過這次回歸報告，讓我學習很多，謝謝大家！

雅竹：回歸報告寫了快一個月ㄟ，我要120分！

美惠：這份報告真是讓我「受益良多」耶！總算完工了，老師可以不考試嗎?我想要放假！

玲慧：翻書翻到手沒力，改表改到手斷掉，老師可以不考試嗎？我想看醫生！

愉翔：大家做的真辛苦，好累好累好累……..老師大大可以不考試嗎?

報告討論的進度紀錄表

12 月 28 日 (四)	1. 資料的選取，報告形式與書寫內容討論。 2. 使用軟體 SAS 報表與 R 圖形，及獲得初步結果。 3. 下次討論時間為 12/30 下午 2 點	19:00 ｜ 23:00	美惠－SAS 初步完成 橙莉－R 初步完成，並完成兩個連結 全體回家唸書	橙莉 美惠 雅竹 威麟	愉翔 玲慧
12 月 30 日 (日)	1. 整體報告內容討論與解讀。 2. 程式的討論。 3. 下次討論時間為 1/4 晚上 7 點。	14:00 ｜ 15:00	美惠－附錄與 SAS 程式修改 橙莉－Data Story 雅竹、威麟、玲慧－整理好報表及彙整 全體回家唸書	橙莉 美惠 雅竹 威麟 玲慧	愉翔
1 月 4 日 (四)	1. 統計分析區塊的分配 2. 各章節下星期三交初稿彙整 3. 下次討論時間 1/7 下午 2 點	19:00 ｜ 21:00	各章節分析開始著手，全體回家唸書。 雅竹、威麟－繪製流程圖	全到	
1 月 7 日 (日)	初步討論負責章節內容，問題討論。	14:00 ｜ 15:00	回去各自內容補充與修改	全到	
1 月 13 日 (六)	1. 呈現各章節初步完成結果，章節內容的連貫與統計上的解釋等問題討論。 2. 章節編排與書寫方式的統一。	14:00 ｜ 20:00	分析部分的修改與排版	橙莉 美惠 雅竹 威麟 玲慧	愉翔
1 月 15 日 (一)	1. 總結的討論	10:00 ｜ 12:30	報告完成	全到	