報告題名：

體脂肪與身體 13 個測量值的關係

作者：何佩珊、李淑如、林君亭、賴永偉、林瑋民、林宗仁
系級：統計三乙
學號：D9561201、D9538375、D9538358、D9561232、D9561156、D9452814

開課老師：陳婉淑 教授

課程名稱：迴歸分析

開課系所：統計系

開課學年：97 學年度 第 1 學期

中文摘要

在本報告中，資料來自卡內基美隆大學(Carnegie Mellon University，CMU)的圖書館，探討體脂肪的多寡與年齡、體重、身高、頸圍、胸圍、腹圍、臀圍、大腿圍、膝圍、踝圍、二頭肌圍、前臂圍、手腕圍等十三的變數有統計關係。利用 VIF 檢測變數之間是否有共線性，發現體重與身高這兩個變數具有多重共線性，並以 BMI 來取代體重、身高。

計算 BMI 的公式如下：

$$BMI=體重\ kg/（身高\ m）^2$$

選擇重要變數，利用(一)逐步選取法(Stepwise Selection) (二)向前選取法(Forward selection) (三)倒退消去法(Backward elimination) (四)全部子集迴歸等四個方法選擇重要變數。

最後對殘差做分析，符合誤差項四個基本假設。結果顯示，藉由觀測年齡、頸圍、腹圍、臀圍、大腿圍、前臂圍、手腕圍可以預測體脂肪多寡。

關鍵字：體脂肪、迴歸分析、多重共線性、選擇重要變數

目 次

第一章　緒論

第一節 研究背景、動機與目的

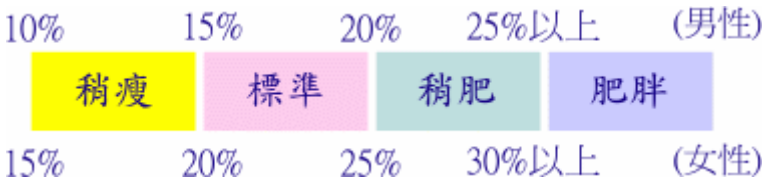
隨著時代的變遷和科技的發達，現代人在生活上的飲食已經有了大幅的改變，讓食物精緻化、美味可口已經是最基本的要求，但也因爲如此，加工過的食物中常含有高熱量、高蛋白等，它們會加重身體的負擔，讓我們體重上升、體脂肪過高，也會導致許多疾病的發生，例如：高血壓、糖尿病和許多心血管的慢性疾病等，這些病例大多是來自自己的體脂肪過高，而導致疾病發生。

體脂肪率是指身體成份中，脂肪組織所佔的比率。體脂肪指數越高者，代表危害身體健康的百分比就越高。判斷肥胖的標準不再只是看體重，體脂肪率是現在醫學上的健康指標。而身型的胖瘦與體脂肪率的高低實際上並沒有必然的關係，體內囤積了過多的脂肪，即表示自身的健康狀況亮起了紅燈，基於生活水平的提升，國人越來越注重身體的健康。

身體組織成份中，有五十%至六十%是水份，佔第二位的即是脂肪。體重合乎標準，但體脂肪量高是典型的現代人體型。所以判斷肥胖的標準不再只是看體重，現在醫學上最新的健康指標是「體脂肪率」。「體脂肪率」即爲身體中的脂肪重量除以體重得到的比率。過高的體脂肪率是造成各種慢性疾病的主要導火線；體脂肪指數越高者，代表危害身體健康的比率越高。體脂肪檢測乃是利用阻抗原理，及阻抗較低者表示導電性較高之肌肉較多、脂肪較少；反之表示肌肉少、脂肪多。

可能的疾病

一般而言，成年男子體脂肪率超過 25%，成年女子體脂肪率超過 30%，即表示有肥胖症。



體脂肪與身體 13 個測量值的關係

醫生建言

根據不同的體型，給予下列的飲食與運動建議：

瘦肥型 ■體脂肪高，體重低於標準 10%以上。 醫生建言　減少高脂肪與甜食攝取，但注意均衡營養攝取，尤其是蛋白質的補充。培養運動習慣，降低體內脂肪。	**微胖型** ■體脂肪高，體重標準，是現代人最典型體型。 醫生建言　減少高熱量、高脂肪食物攝取，並勤做脂肪燃燒運動。	**肥胖型** ■體脂肪高，體重高於標準 10%以上 醫生建言　請由專家指導飲食減肥，並加強脂肪燃燒的運動。
稍瘦型 ■體脂肪標準，但體重低於標準體重 10%以上。 醫生建言　多做強化肌力運動，促進體內新陳代謝。	**標準型** ■體脂肪及體重均標準 醫生建言　維持飲食與運動，才可使體型繼續標準。	**強壯型** ■體脂肪標準，但體重高於標準，大都爲運動型的人。 醫生建言　減少脂肪與醣類並加強脂肪燃燒運動。
過瘦型 ■體脂肪不足，體重低於標準 10%以上。 醫生建言　注意均衡營養攝取，增加肌力訓練。	**活力型** ■體脂肪少，但重標準，大多爲運動型的人。 醫生建言　繼續運動，以維持體型。	**肌肉發達型** ■體脂肪少，但體重高於標準典型的運動員。 醫生建言　繼續運動，避免脂肪堆積。

現代人都十分注重外表，也流行各種不同的「減肥」方法，甚至傳播媒體是一而再、再而三的播放著，「瘦」、「更瘦」已經是時代的潮流。有許多人用了錯誤的減肥方法，讓自己成爲名副其實的「脂肪人」。醫學上指出，少吃不動的減重方法，只能消除身體的水分與肌肉，身體的脂肪卻不動如山，一但恢復飲食，胖的還是脂肪。以「快速減肥」爲例，最容易減掉肌肉後，看似體重下降，因此恢復正常飲食時，身體卻將熱量以脂肪形式儲存，體重機看似恢復原來的體重，但是身體組織已經改變，減肥也越來越困難。理想的減重是以一個月不超過「四」公斤，一天至少攝取 1200 大卡的熱量，並且配合著運動以包持肌肉的活動，避免減掉肌肉，真正達到減肥效果。

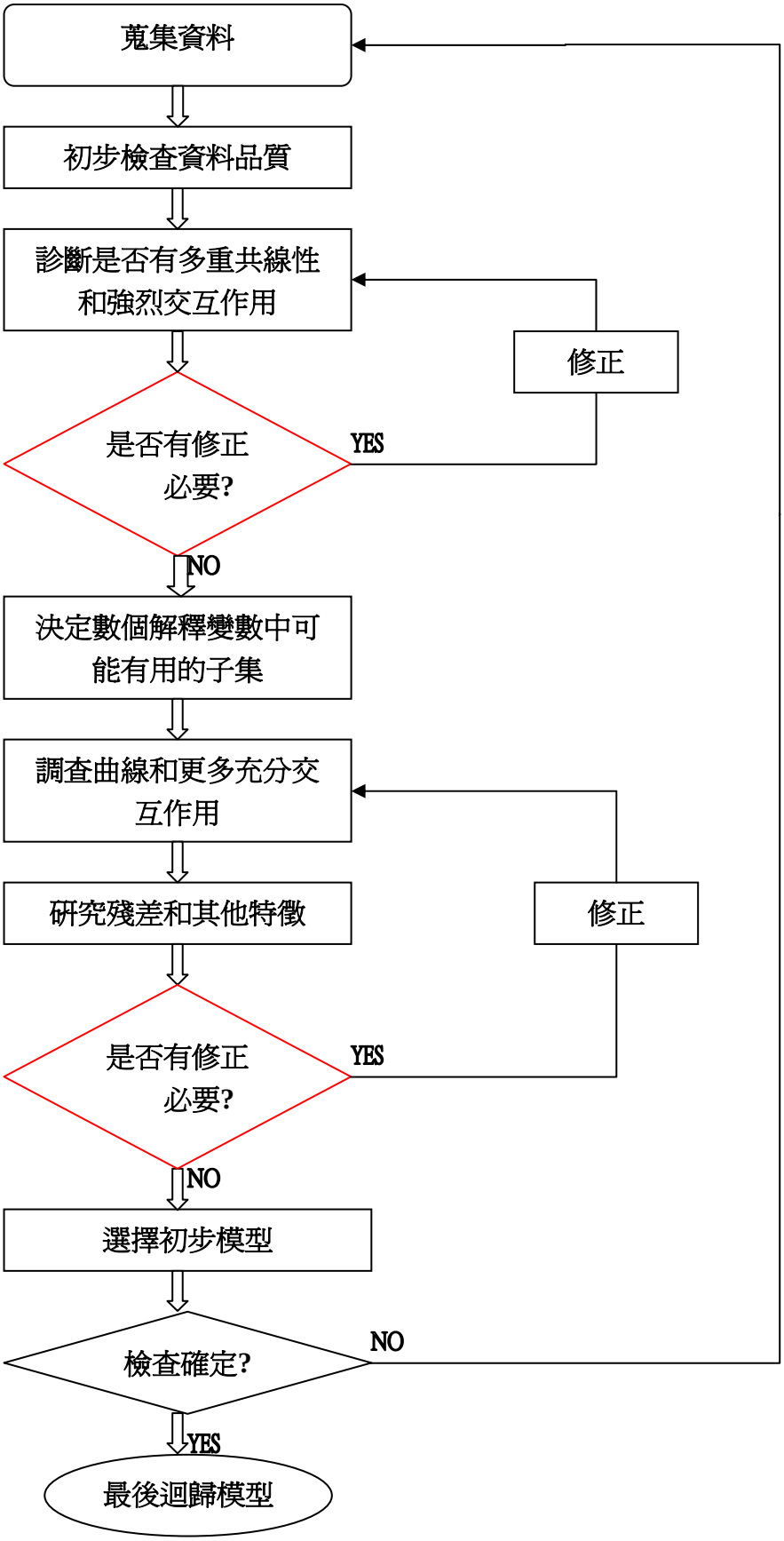
因爲體脂肪率是否被控制在正常範圍內，是無法藉由目測或一般體重計獲得數值，雖然後來有方便簡易的體脂肪計因應而生，但是體脂肪的測量卻有許多限制，如懷孕、裝置心臟節律器者禁止測量，而測量時需赤裸雙腳，且將雙腳腳底請擦拭乾淨，腳底若有塵埃或髒污時，則無法正確測量，過度飲食或有極端脫水狀況時，測量值也會有誤差，而且還有最佳測量時段，起床 3 小時以上和飲食後 3 小時以上，最好每次在同一時間測量。爲了瞭解體脂肪是否維持在正常範圍內以及不受體脂肪計的限制，我們研究年齡、體重、身高、頸圍、胸圍、腹圍、臀圍、大腿圍、膝圍、踝圍、二頭肌圍、前臂圍、手腕圍身體的十三個量測指標和體脂肪的相關性，透過身體各種量測指標和體脂肪率的相關程度，推測人體內所佔的體脂肪多寡。

第二節 研究方法與流程

迴歸分析(regression analysis)在統計上爲重要工具之一，它是用來表示兩個或兩個以上計量變數間的關係，或藉由一群自變數來預測某一應變數的相關資訊。迴歸分析通常被認爲是在 1885 年由 Francis Galton 所提出的，有許多學者在他之前也曾經提出類似的觀念，但這些觀念通常都是理論上的，並沒有實際的應用，舉例來說，相關係數在之前只是一個數學上的理論，而 Galton 的主要成就，就是將理論的東西實際應用在現實生活中。

Galton 是從他對於遺傳學上的觀察中，發現在遺傳上所觀察到的特徵，可以用迴歸到母群的平均數來解釋，也就是說，親代的特徵在很極端的情況下，他的子代所遺傳到的特徵會比較接近於平均數，不會是相同或更加極端的，而迴歸分析用於預測實際問題方面，目的在減少決策之風險，且在成本極小的情況下，尋求改進預測的準確度及估計在不確定量之決策過程所產生的損失。

流程圖：


蒐集資料
初步檢查資料品質
診斷是否有多重共線性
和強烈交互作用
修正
是否有修正
必要?
YES
NO
決定數個解釋變數中可
能有用的子集
調查曲線和更多充分交
互作用
研究殘差和其他特徵
修正
是否有修正
必要?
YES
NO
選擇初步模型
檢查確定?
NO
YES
最後迴歸模型

第二章 資料分析

第一節 一般敘述統計

資料來源：卡內基美隆大學(Carnegie Mellon University，CMU)的圖書館；
日期：1995.10.02
網址：http://lib.stat.cmu.edu/datasets/bodyfat

解釋變數：

Yi：體脂肪
Xi1：年齡單位爲年(year)
Xi2：體重單位爲磅(lbs)
Xi3：身高單位爲英吋(inch)
Xi4：頸部周長單位爲公分(cm)
Xi5：胸膛周長單位爲公分(cm)
Xi6：腹部周長單位爲公分(cm)
Xi7：臀圍單位爲公分(cm)
Xi8：大腿圍單位爲公分(cm)
Xi9：膝圍單位爲公分(cm)
Xi10：踝圍單位爲公分(cm)
Xi11：二頭肌圍單位爲公分(cm)
Xi12：前臂圍單位爲公分(cm)
Xi13：腕部周長單位爲公分(cm)
Xi14：BMI

表 2.1.1：X1~X14 的基本統計量　樣本數 n=222

	平均數	中位數	標準差	偏態	峰態	最小值	最大值
年齡 X1	42.5090	42	11.2928	0.3800	0.1558	22	81
體重 X2	178.8932	176.50	29.4379	1.3574	6.0727	118.50	363.15
身高 X3	70.3074	70.375	9.7900	-5.6184	60.0620	29.500	77.750
頸部周長 X4	37.9351	37.85	2.4433	0.6805	3.2351	31.10	51.20
胸膛周長 X5	100.3815	99.45	8.2611	0.7593	1.4884	79.30	136.20
腹部周長 X6	92.0347	90.65	10.6593	0.9424	2.9702	69.40	148.10
臀圍 X7	99.9667	99.3	7.1669	1.6653	8.5446	85.0	147.7
腿圍 X8	59.6113	59.05	5.3597	0.8397	2.6416	47.20	87.30
膝圍 X9	38.5644	38.4	2.3029	0.5863	1.4170	33.5	49.1
踝圍 X10	23.1640	22.9	1.7352	2.3543	11.9853	19.1	33.9
二頭肌圍 X11	32.3243	32.05	3.0170	0.3754	0.5944	24.80	45.00
前臂圍 X12	28.6811	28.7	2.0068	-0.0980	0.7583	21.0	34.9
腕部周長 X13	18.1896	18.2	0.9101	0.3904	0.4754	16.1	21.4
BMI X14	0.1797	0.1719	0.0701	12.1538	167.5495	0.1251	1.1501

第二節 散佈圖

散佈圖(Scatter Diagram)是相關與迴歸分析最基本的工具，其 y 軸表示應變數(Dependent variable)，x 軸表示獨立變數(Independent variable)，如果散佈圖上的點愈接近形成一直線，表示這兩個變數關係愈密切。反之，則表示這兩個變數沒有什麼關係，也就是說獨立變數對應變數沒有什麼預測能力。

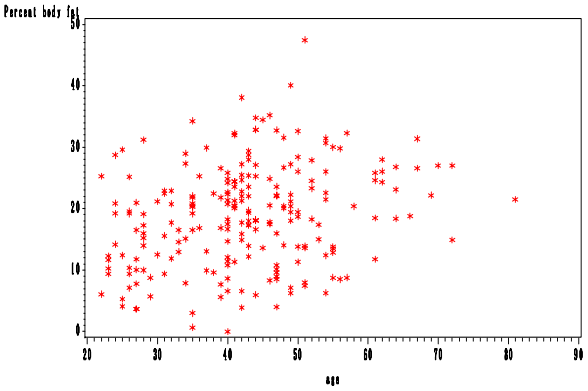
由圖 2.2.1 爲 Y(體脂肪)對 X1(年齡)的散佈圖和圖 2.2.4 爲 Y(體脂肪)對 X4(頸圍)的散佈圖可以看出其相關性並不高。

圖 2.2.2 爲 Y(體脂肪)對 X2(體重)的散佈圖，由散佈圖的分散情形可以看出體脂肪與體重有明顯的正相關，其相關性頗高。

圖 2.2.3 爲 Y(體脂肪)對 X3(身高)的散佈圖，由散佈圖的分散情形可以看出身高頗爲集中，其範圍多數在 65 英吋至 78 英吋之間，且有明顯的離群值，其相關性低。

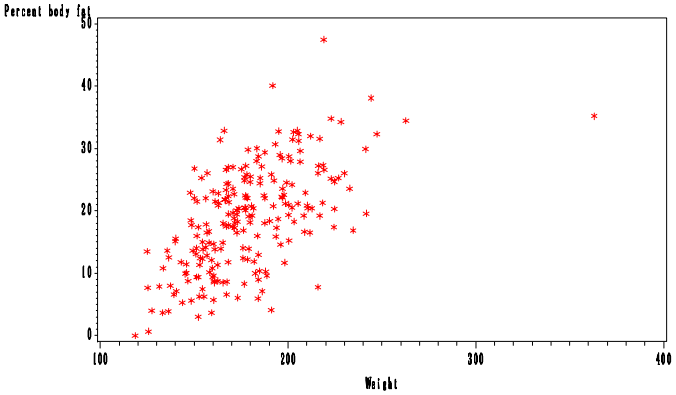
Correlation Coefficient(Y,X1)=0.26119

圖 2.2.1：Y(體脂肪)對 X1(年齡)的散佈圖



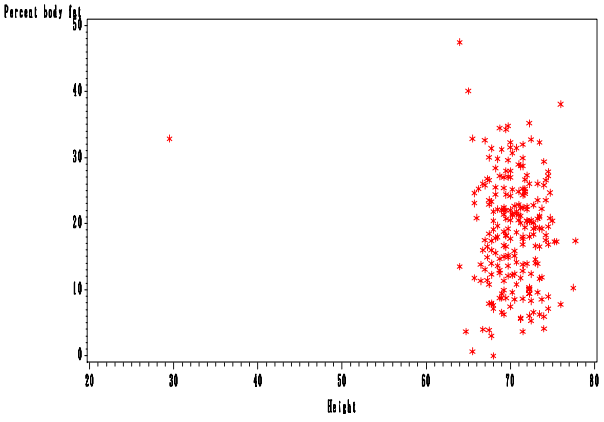
Correlation Coefficient(Y,X2)=0.59018

圖 2.2.2：Y(體脂肪)對 X2(體重)的散佈圖



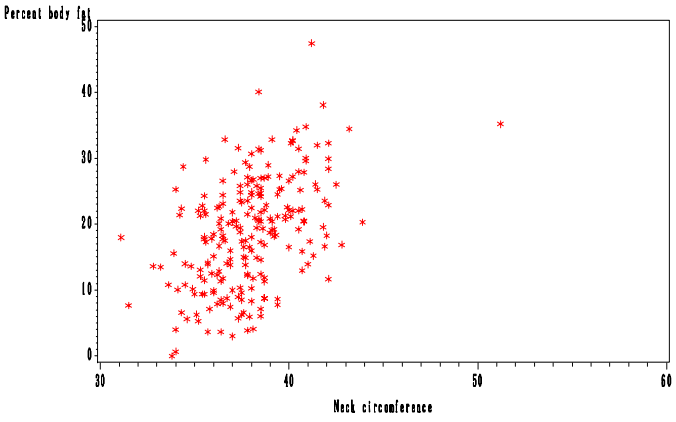
Correlation Coefficient(Y,X3)= - 0.11062

圖 2.2.3：Y(體脂肪)對 X3(身高)的散佈圖



Correlation Coefficient(Y,X4)=0.46432

圖 2.2.4：Y(體脂肪)對 X4(頸圍)的散佈

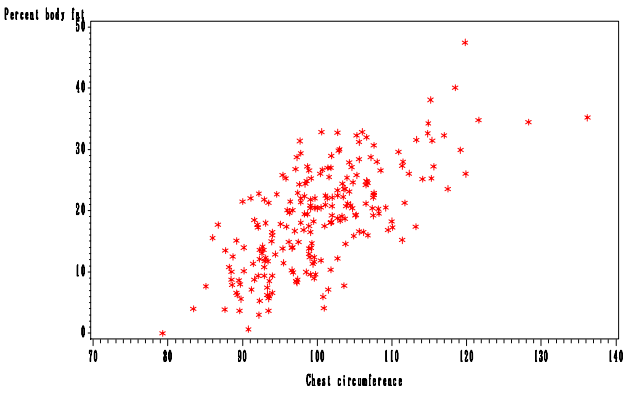


體脂肪與身體 13 個測量值的關係

由圖 2.2.5、圖 2.2.6、圖 2.2.7 與圖 2.2.8 的散佈圖，可以看出體脂肪與胸圍、腹圍、臀圍和大腿爲有高度正相關。而圖 2.2.9 爲 Y(體脂肪)對 X9(膝圍)的散佈圖，其相關性並不高。圖 2.2.10 爲 Y(體脂肪)對 X10(踝圍)的散佈圖，踝圍的值多數介於 19 至 28 之間，且有 2 個明顯的異常點，其相關性低。

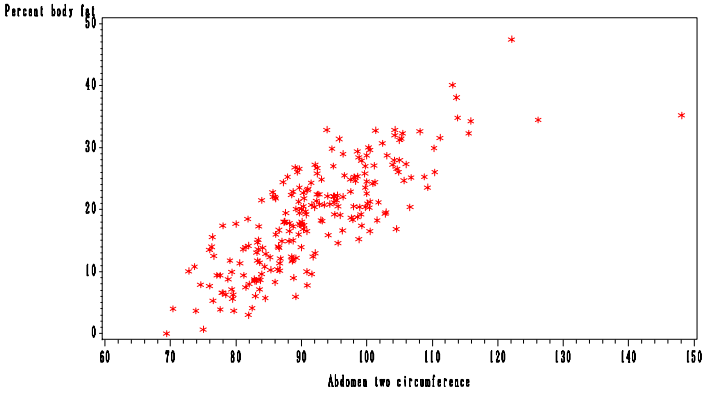
Correlation Coefficient(Y,X5)=0.68903

圖 2.2.5：Y(體脂肪)對 X5(胸圍)的散佈圖



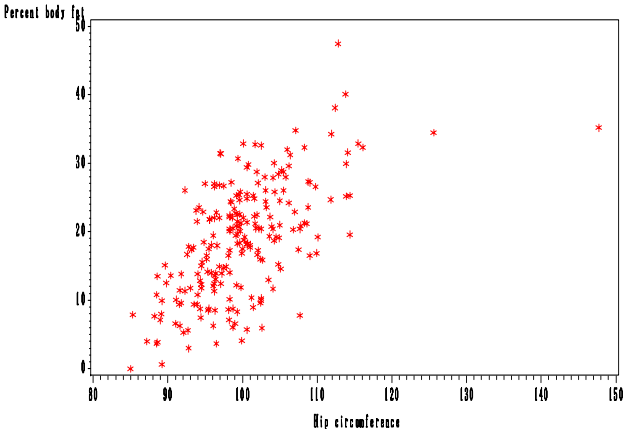
Correlation Coefficient(Y,X6)=0.80318

圖 2.2.6：Y(體脂肪)對 X6(腹圍)的散佈圖



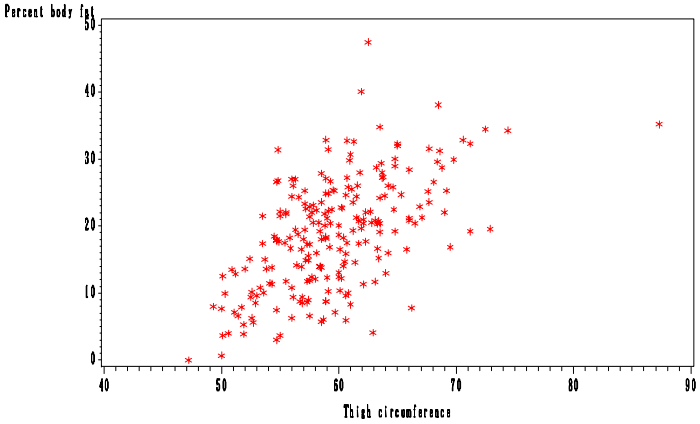
Correlation Coefficient(Y,X7)=0.60835

圖 2.2.7：Y(體脂肪)對 X7(臀圍)的散佈圖



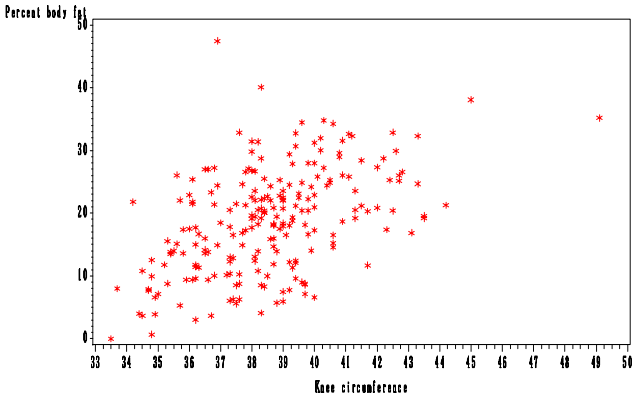
Correlation Coefficient(Y,X8)=0.57495

圖 2.2.8：Y(體脂肪)對 X8(大腿圍)的散佈圖



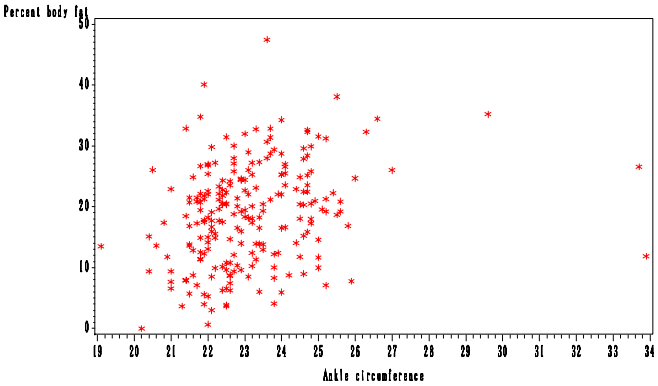
Correlation Coefficient(Y,X9)=0.48631

圖 2.2.9：Y(體脂肪)對 X9(膝圍)的散佈圖



Correlation Coefficient(Y,X10)=0.26119

圖 2.2.10：Y(體脂肪)對 X10(踝圍)的散佈圖

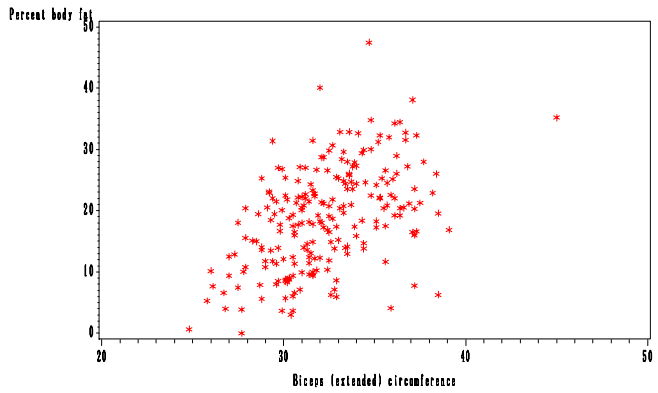


體脂肪與身體 13 個測量值的關係

由圖 2.2.11、圖 2.2.12、圖 2.2.13 與圖 2.2.14 的散佈圖可以看出體脂肪與二頭肌圍、前臂圍、手腕爲和 BMI 的相關性並不高。且由圖 2.2.14：Y(體脂肪)對 X14(BMI)的散佈圖，可以看出 BMI 的值多數介於 0.125 至 0.3 之間，且有 1 個明顯的異常點。

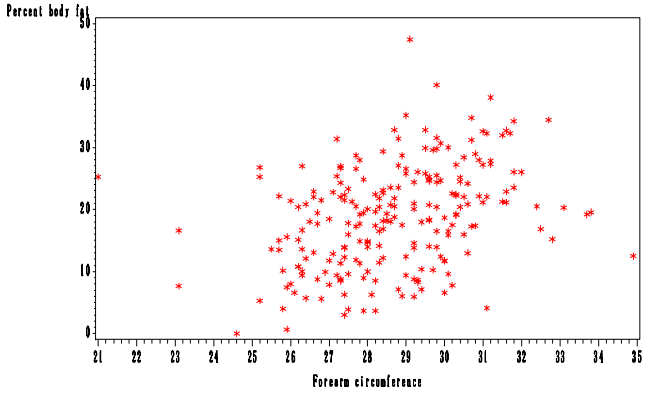
Correlation Coefficient(Y,X11)=0.48651

圖 2.2.11：Y(體脂肪)對 X11(二頭肌圍)的散佈圖



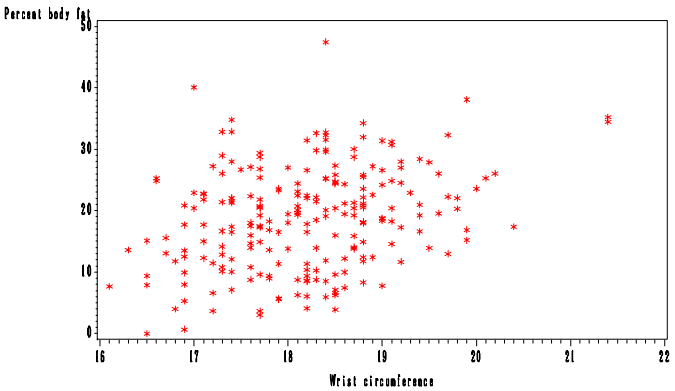
Correlation Coefficient(Y,X12)=0.35435

圖 2.2.12：Y(體脂肪)對 X12(前臂圍)的散佈圖



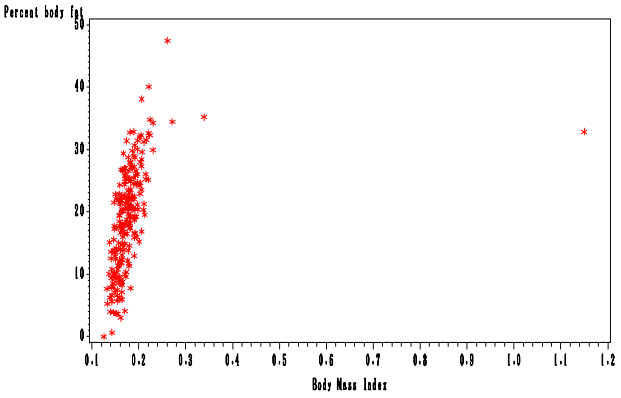
Correlation Coefficient(Y,X13)=0.30149

圖 2.2.13：Y(體脂肪)對 X13(手腕圍)的散佈圖



Correlation Coefficient(Y,X14)=0.35995

圖 2.2.14：Y(體脂肪)對 X14(BMI)的散佈圖



由上述可知，解釋變數 X6(腹圍)、X5(胸圍)、X7(臀圍)、X2(體重)、X8(大腿圍)對 Y 的解釋能力較高，但因爲身高與體重有高度相關，預期在選擇重要變數的時候，X6(腹圍)、X5(胸圍)、X7(臀圍)、X8(大腿圍)會被選上。

第三節 檢測 Full model 多重共線性

多重共線性是指在一個迴歸模型裡，獨立變數之間有高度相關。而多重共線性會使個別誤差變很大，造成 t-stats 小，導致各別參數相關性不顯著。可利用 Variance Inflation Factor(簡稱 VIF)做診斷，若 Variance Inflation Factor＞10，則變數之間有多重共線性。

由表 2.3.1，可以得知 X2(體重)的 VIF=31.53657，X6(腹圍)的 VIF=11.57146，X7(臀圍)的 VIF=14.53428，皆大於 10，其中以 X2(體重)最爲嚴重。

由表 2.3.1 可知配適迴歸線

$$\hat{Yi}=-11.98607+0.08345x_1-0.08695x_2-0.07752x_3-0.52669x_4+0.04488x_5+0.94488x_6-0.26612x_7+0.31417x_8$$
$$-0.11646x_9+0.18832x_{10}+0.14500x_{11}+0.48590x_{12}-1.85875x_{13}$$

表 2.3.1　Full model 的參數估計

Variable	Parameter Estimate	Standard Error	Variance Inflation
Intercept	-11.98607	17.88313	0
X1	0.08345*	0.03485	1.95069
X2	-0.08695	0.05375	31.53657
X3	-0.07752	0.09542	1.64729
X4	-0.52669*	0.23623	4.19578
X5	0.04488	0.10222	8.98201
X6	0.94488**	0.08992	11.57146
X7	-0.26612	0.14989	14.53428
X8	0.31417*	0.15213	8.37328
X9	-0.11646	0.26141	4.56473
X10	0.18832	0.22114	1.85457
X11	0.14500	0.17596	3.54973
X12	0.48590*	0.20041	2.03730
X13	-1.85875**	0.56568	3.33863

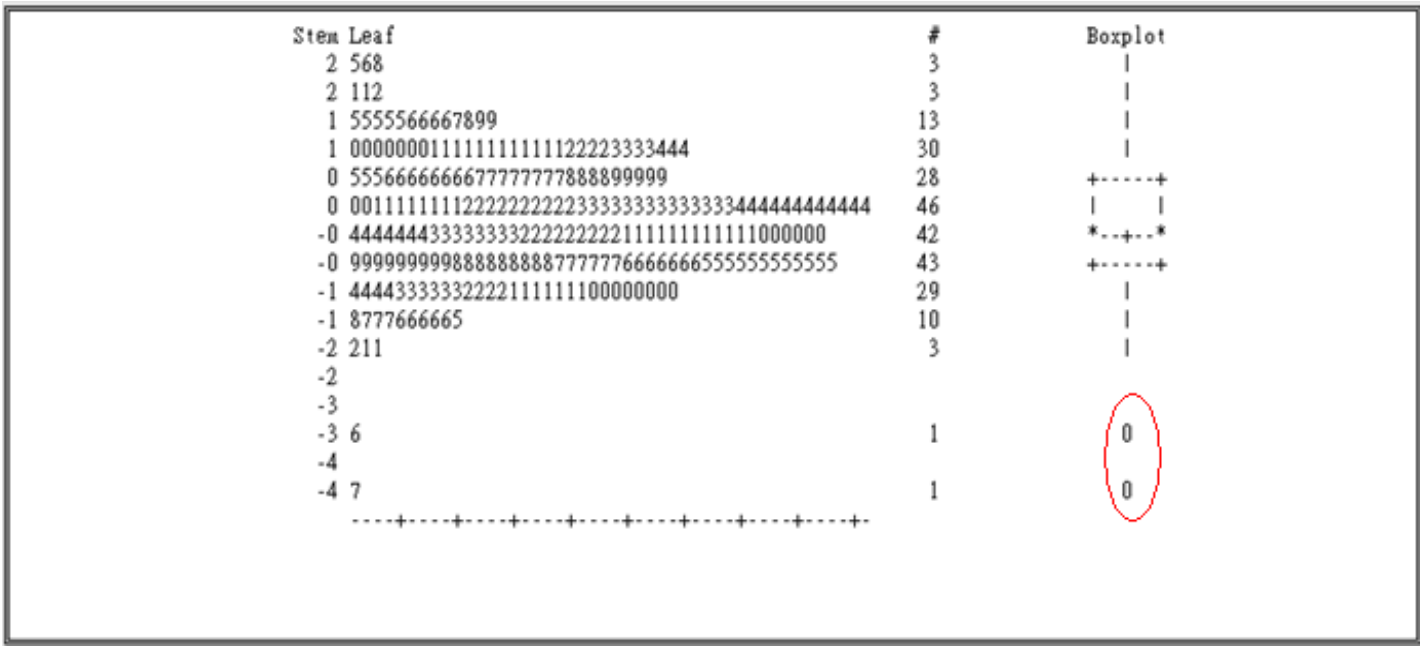
Parameter Estimate：
*：p-value＜0.05
**：p-value＜0.01
標註紅色表示不顯著

體脂肪與身體 13 個測量值的關係

第四節 離群值

圖 2.4.1 爲資料標準化所做的莖葉圖，且因爲資料過度集中，所以將莖分割，以便看出資料的分布情形。由圖十四可以判斷此資料爲常態分布，但有二個值偏離分布，經由分析後，此二筆資料與其他資料不符，因此建議移除，而其他筆資料的標準化殘差，其絕對值皆在三以內。

圖 2.4.1：莖葉圖



體脂肪與身體 13 個測量值的關係

第五節 變數轉換與刪除離群值之模型的 VIF

爲了改善多重共線性的問題，做變數轉換，設 X14=(0.45359237 × X2) / (0.3048 ×X3)^2，X14 爲 Body Mass Index，以 X14(Body Mass Index)取代 X2(Weight)與 X3(Height)，重新配適迴歸模型，其所有變數的 VIF 皆小於 10。

(1 磅=0.45359237 公斤；1 英吋=0.3048 公尺)

由表 2.5.1 可知配適迴歸線

$$\hat{Y}i=3.69382+0.11081X_1+3.74185X_{14}-0.55347X_4-0.04494X_5+0.91000X_6-0.28632X_7+0.31168X_8-0.38080X_9$$
$$+0.17815X_{10}+0.13120X_{11}+0.32219X_{12}-2.26906X_{13}$$

表 2.5.1：變數轉換後的參數估計

Variable	Parameter Estimates	Standard Error	Variance Inflation
Intercept	3.69382	7.53466	0
X1	0.11081**	0.03357	1.81947
X14	3.74185	4.67252	1.32799
X4	-0.55347*	0.23481	3.61403
X5	-0.04494	0.09522	7.08769
X6	0.91000**	0.08816	9.62280
X7	-0.28632*	0.13642	9.45804
X8	0.31168*	0.14948	7.06588
X9	-0.38080	0.24626	3.67123
X10	0.17815	0.21843	1.69838
X11	0.13120	0.17336	3.15720
X12	0.32219	0.20811	2.20871
X13	-2.26906**	0.55165	3.01543

Parameter Estimate：
*：p-value＜0.05
**：p-value＜0.01
標註紅色表示不顯著

第六節 選擇重要變數

變數轉換：以X14取代X2、X3
1 逐步選取法(Stepwise Selection)：最後選取重要變數 X6、X13、X1、X4、X12、X7、X8
2 向前選取法(Forward selection)：最後選擇重要變數 X6、X13、X1、X4、X12、X7、X8
3 倒退消去法(Backward elimination)：最後選擇重要變數 X1、X4、X6、X7、X8、X12、X13
4 全部子集迴歸：最後選擇重要變數 X1、X4、X6、X7、X8、X12、X13

(一) 逐步選取法(Stepwise Selection)

逐步選取法(Stepwise Selection)是結合“向前”、“向後”選取法而成。開始時以向前選取法選入一個變數，而後每當選入一個新預測變數後，就利用向後選取法式看看在模式中已存在的預測有無 p-value 大於 0.15 的變數，若有，則該預測變數就會被排除在模式之外，接著再進行向前選取法；若無，則繼續向前選取；這樣向前與向後選取輪流使用，直到沒有預測變數可在選進來，也沒有預測變數會被去除。

在顯著水準 0.1500 下，最後選擇的變數是 X6(腹圍)、X13(手腕圍)、X1(年齡)、X4(脖子周長)、X12(前臂圍)、X7(臀圍)、X8(腿圍)

表 2.6.1：逐步選取法

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x6	1	0.6657	0.6675	71.2301	436.17	<0.0001
2	x13	2	0.0536	0.7193	27.0097	41.63	<0.0001
3	x1	3	0.0188	0.7382	12.7855	15.59	0.0001
4	x4	4	0.0034	0.7415	11.8789	2.82	0.0947
5	x12	5	0.0043	0.7459	10.1346	3.67	0.0566
6	x7	6	0.0026	0.7485	9.8626	2.24	0.1358
7	x8	7	0.0060	0.7545	6.6786	5.22	0.0234

(二) 向前選取法(Forward selection)

向前選取法(Forward selection)是選取進入模式的預測變數越選越多，每一步驟都是選取“剩餘”解釋能力最強的一個預測變數進入模式，但其解釋能力也必須通過事先訂好的門檻(顯著水準 0.1500)。

在顯著水準 0.1500 下，最後選擇的變數是 X6(腹圍)、X13(手腕圍)、X1(年齡)、X4(脖子周長)、X12(前臂圍)、X7(臀圍)、X8(腿圍)

表 2.6.2：向前選取法

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x6	1	0.6657	0.6675	71.2301	436.17	<0.0001
2	x13	2	0.0536	0.7193	27.0097	41.63	<0.0001
3	x1	3	0.0188	0.7382	12.7855	15.59	0.0001
4	x4	4	0.0034	0.7415	11.8789	2.82	0.0947
5	x12	5	0.0043	0.7459	10.1346	3.67	0.0566
6	x7	6	0.0026	0.7485	9.8626	2.24	0.1358
7	x8	7	0.0060	0.7545	6.6786	5.22	0.0234

體脂肪與身體 13 個測量值的關係

(三)倒退消去法(Backward elimination)

倒退消去法(Backward elimination)的選擇預測區間過程，剛好跟向前選取法相反，再一開始時是所有預測變數都放在模式內，然後再將解釋能力差的變數一一去掉，直到所有放在模式中的預測變數其顯著水準皆小於設定的門檻(顯著水準 0.1000)才停止。

在顯著水準 0.1000 下，最後選擇的變數是 X1(年齡)、X4(脖子周長)、X6(腹圍)、X7(臀圍)、X8(腿圍)、X12(前臂圍)、X13(手腕圍)

表 2.6.3：倒退消去法

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x5	11	0.0003	0.7585	11.2425	0.24	0.6229
2	x11	10	0.0004	0.7581	9.6254	0.38	0.5360
3	x10	9	0.0006	0.7574	8.1849	0.56	0.4538
4	x14	8	0.0008	0.7566	6.8756	0.70	0.4048
5	x9	7	0.0021	0.7545	6.6786	1.82	0.1786

體脂肪與身體 13 個測量值的關係

(四)全部子集迴歸(All-subsets Regression)

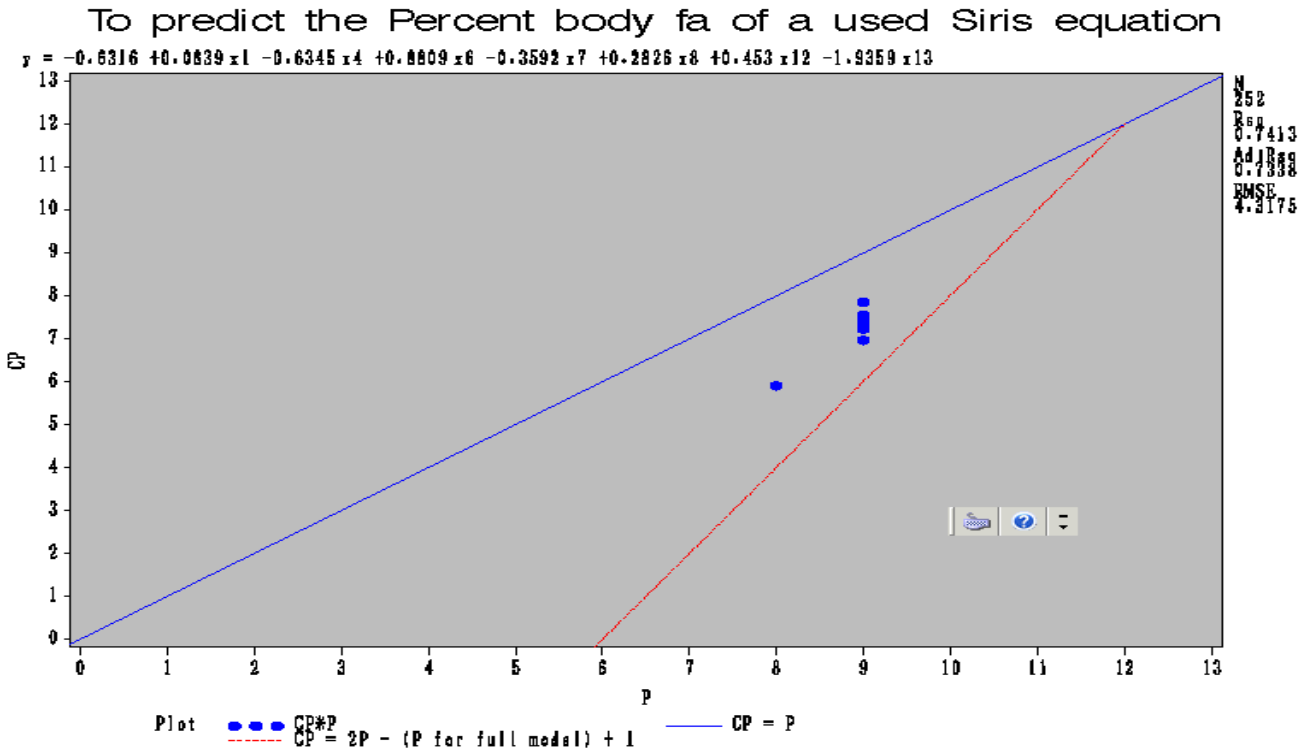
全部子集迴歸的挑選方法有 RMSE、R^2、adj R^2、Mallow’ s Cp criterion 和 AIC 等等。其 root MSE 是愈小愈好，而 R^2 則是愈大愈好，但是因爲變數愈多，則 R^2 就會愈大，所以選用 adj R^2 比較適當；而 Mallow’ s Cp criterion 和 AIC 也是愈小愈好，且要接近參數個數。

表八爲 Cp 選取法前 6 個最佳模式，則最後選擇的變數是 X1(年齡)、X4(脖子周長)、X6(腹圍)、X7(臀圍)、X8(腿圍)、X13(手腕圍)

表 2.6.4：C(p) Selection Method

Number in Model	C(p)	R-Square	AIC	Variables in Model
7	6.6786	0.7545	637.7611	x1 x4 x6 x7 x8 x12 x13
8	6.8756	0.7566	637.8706	x1 x4 x6 x7 x8 x9 x12 x13
6	7.8253	0.7509	639.0222	x1 x4 x6 x7 x8 x13
8	7.9748	0.7553	639.0250	x1 x14 x4 x6 x7 x8 x12 x13
9	8.1849	0.7574	639.1421	x1 x14x4 x6 x7 x8 x9 x12 x13
7	8.2213	0.7527	639.3659	x1 x4 x6 x7 x8 x9 x13

圖 2.6.1：Cp 最適模型



體脂肪與身體 13 個測量值的關係

第七節 偵測影響點

對於包含多個自變數的迴歸模型及資料，是很容易辨別其對 X 或 Y 值所產生的極端觀察值，以及了解是否對適合迴歸構成顯著性的影響。

(一) COVRATIO

決策規則：

若 $\text{COVRATIO}_i > 1+3\frac{p}{n}$ 或 $\text{COVRATIO}_i < 1-3\frac{p}{n}$ ，則表示第 i 筆資料有可能爲影響點。

P=8，n=220，Cov Ratio $_i > 1+3\frac{8}{220} = 1.1090909$ 或 Cov Ratio $_i < 1-3\frac{8}{220} = 0.8909091$

用 COVRATIO 檢測影響點，發現資料表 2.7.1 的觀察值可能爲影響點

表 2.7.1：COVRATIO

Observation	COVRATIO
5	1.1293
15	1.1305
36	1.1837
40	1.1783
41	1.1714
57	1.1152
77	1.1096
104	1.2402
138	0.8695
173	1.4318
204	1.2407

體脂肪與身體 13 個測量值的關係

(二) The hat matrix elements h_i

決策規則：

若 $H_{ii}>2(\frac{p}{n})$，則第 i 筆資料可能爲影響點。

p=8，n=220，$H_{ii}>2(\frac{8}{200})=0.07272727$

用 Hat Diag 檢測影響點，發現資料表 2.7.2 的觀察值可能爲影響點。

表 2.7.2：Hat Diag

obs	Hat Diag
5	0.0854
15	0.0888
36	0.1365
40	0.1253
41	0.1208
77	0.0743
104	0.1648
167	0.0778
173	0.2955
203	0.0774
204	0.1631
214	0.1049

(三) Cook's distance statistic D_i

決策規則：

若 $D_i>F_{0.5;\,p,\,n-p}$，則第 i 筆資料可能爲影響點。

因爲 $F_{0.5;\,p,\,n-p}$ 近似於 1，所以當 $D_i>1$，則第 i 筆資料可能爲影響點。

用 Cook’s Distance 檢測影響點，發現無資料的 $D_i>1$，所以無觀察值可能爲影響點。

(四)DFBETAS

決策規則：

若 $|\text{DFbetas}_i| > \frac{2}{\sqrt{n}}$，則第 i 筆資料可能爲影響點。

n=220，$|\text{DFbetas}| > \frac{2}{\sqrt{200}} = 0.134839972$

用 DFBETAS 檢測影響點，其顯示有兩個值大於 0.1348 列入表 2.7.3，表示其觀察值可能爲影響點。

表 2.7.3：DFBETAS

z	Intercept	x1	x4	x6	x7	x8	x12	x13
3	0.2460	-0.1909	-0.1322	0.1006	0.0248	0.0190	-0.1697	-0.0529
12	0.1683	0.0054	-0.0214	0.1908	-0.1335	-0.0419	0.0134	-0.0321
20	0.0793	0.0119	-0.0859	0.0785	-0.1430	0.0245	-0.0058	0.1417
28	0.1600	-0.1333	0.2102	0.0803	-0.1122	0.0101	-0.1258	-0.1272
36	-0.0008	0.0299	-0.0084	0.0210	0.1619	-0.1378	0.0503	-0.1480
38	-0.1590	0.1264	0.1650	-0.1184	-0.0571	0.1606	-0.0351	-0.0010
60	0.0008	0.1442	-0.2182	0.0076	-0.0062	0.0832	0.1277	0.0359
73	-0.0676	-0.2396	-0.0874	0.1248	-0.0143	-0.1005	-0.0205	0.2012
74	-0.0743	0.1965	-0.0494	-0.1505	0.0873	0.0135	-0.0219	0.0616
79	-0.0527	0.1236	-0.0055	0.1228	-0.0395	-0.1257	-0.1394	0.1991
80	-0.0276	0.2564	0.2342	-0.1576	0.1611	-0.0603	-0.2552	-0.1374
106	0.0474	0.0760	-0.1559	-0.0980	-0.0031	0.1908	0.0246	-0.0149
110	-0.1472	0.0265	0.0734	-0.1635	0.0977	-0.0743	0.0334	0.0609
126	-0.0219	0.0793	0.2097	-0.2138	0.1236	-0.0963	0.0597	-0.1638
133	-0.0415	-0.0169	-0.1250	-0.0304	0.1377	-0.1837	0.1945	0.0238
173	-0.0867	-0.1178	0.1109	0.0033	0.0619	0.0278	-0.5315	0.1992
198	0.0729	-0.0767	-0.0869	0.1402	-0.2384	0.1084	-0.0019	0.1835
202	0.1760	-0.1501	0.0096	0.2459	-0.2205	0.0136	-0.0453	0.0265

(四) DFFITS

決策規則：

若 $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$，則第 i 筆資料可能爲影響點。

n=200，$|DFFITS_i| > 2\sqrt{\frac{8}{200}} = 0.381385035$

用 DFFITS 檢測影響點，發現資料表 2.7.5 的觀察值可能爲影響點。

表 2.7.5：DFFITS

Obs	DFFITS
3	0.4285
79	0.4810
80	0.4932
173	0.5687
205	0.4494
214	0.5547

當資料有 220 筆時，R-Square=0.7524。由表 2.7.6 可知觀察值第 173 筆可能爲影響點，刪除後 R-Square=0.7572，知觀察值第 36 筆可能爲影響點，刪除後 R-Square=0.7497，因爲第 173 筆資料與 36 筆資料對迴歸模型影響不大，故建議保留所有資料。

表 2.7.6：總結

Influence Analysis	Observation
COVRATIO	5,15,36,40,41,57,77,104,138,173,204
Hat Value	5,15,36,40,41,77,104,167,173,203,204,214
Cook’s Distance	無
DFbetas(有 2 筆以上)	3,12,20,28,36,38,60,73,74,79,80,106,110,126,133,173,198,202
DFFITS	3,79,80,173,205,214

第八節 最終模型

在第六節用四種方法選取重要變數且經由是否爲影響點等篩選，決定移除變數 X2(體重)、X3(身高)、X5(胸圍)、X9(膝圍)、X10(踝圍)、X11(二頭肌圍)；選取 X1(年齡)、X4(頸圍)、X6(腹圍)、X7(臀圍)、X8(大腿圍)、X12(前臂圍)、X13(手腕圍)七個變數留在模型中。

最終模型之配適迴歸線：

$$\hat{Yi} = -0.26802 + 0.11122X_1 - 0.55994X_4 + 0.88358X_6 - 0.31137X_7 + 0.30498X_8 + 0.35391X_{12} - 2.35413X_{13}$$

Std Error (7.12605) (0.03237) (0.22463) (0.06936) (0.12887) (0.13412) (0.19670) (0.50157)

表 2.8.1：參數估計表

Variable	Parameter Estimate	Standard Error	Variance Inflation
Intercept	-0.26802	7.12605	0
X1	0.11122	0.03237	1.70168
X4	-0.55994	0.22463	3.32587
X6	0.88358	0.06936	5.98973
X7	-0.31137	0.12887	8.48687
X8	0.30498	0.13412	5.72012
X12	0.35391	0.19670	1.98414
X13	-2.35413	0.50157	2.50662

體脂肪與身體 13 個測量值的關係

檢定 Model $Yi=\beta_0+\beta_1X_1+\beta_4X_4+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_{12}X_{12}+\beta_{13}X_{13}$
其解釋變數與被解釋變數是否有線性關係。

虛無假設 Ho： $\beta_i=0$　i=1、4、6、7、8、12、13

對立假設 H1： β_i 至少有一不爲零

由表2.8.2可看出p-value的值皆明顯＜0.05，所以拒絕Ho。

由表2.8.2可看出R-Square=0.7524，也就是X1(年齡)、X4(頸圍)、X6(腹圍)、X7(臀圍)、X8(大腿圍)、X12(前臂圍)、X13(手腕圍)能解釋 75.24% 的Y(體脂肪)。

表2.8.2：變異數分析表

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	11165	1595.01284	92.04	<.0001
Error	212	3673.76215	17.32907		
Corrected Total	219	14839			
R-Square	0.7524				

第三章 殘差分析

對誤差項四個基本假設：

1.E(εi）＝0

2. Var(ε_i)=σ^2

3. Cov(ε_i,ε_j)=0 ，$\forall$ i≠j

4. ε_i~Normal

第一節 檢測殘差平均是否爲零

虛無假設 Ho： $\mu = 0$

對立假設 H1： $\mu \neq 0$

由表3.1.1可看出p-value的值皆明顯>0.05，所以接受Ho，即殘差平均數爲零，

符合基本假設 E(ε_i)＝0。

表 3.1.1

Tests for Location: Mu0=0		
Test	Statistic	p Value
Student's t	0.018306	0.9854
Sign	-5	0.5441
Signed Rank	-202	0.8313

第二節 檢測殘差變異數爲常態

由圖 3.2.1 至 3.2.7 可看出殘差散佈圖沒有特定的形狀，即重要變數(年齡、頸圍、腹圍、臀圍、大腿圍、前臂圍、手腕圍)的變異數均爲常數，符合基本假設 $Var(\varepsilon_i) = \sigma^2$ 。

圖 3.2.1： X1(年齡)的殘差圖

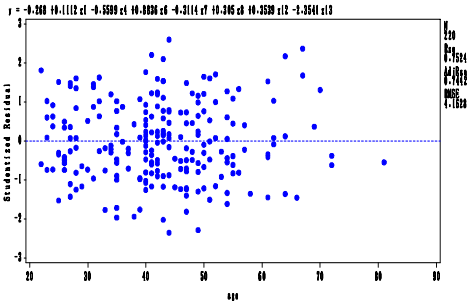


圖 3.2.2： X4(頸圍)的殘差圖

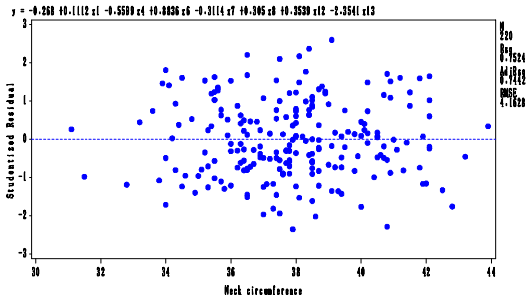


圖 3.2.3： X6(腹圍)的殘差圖

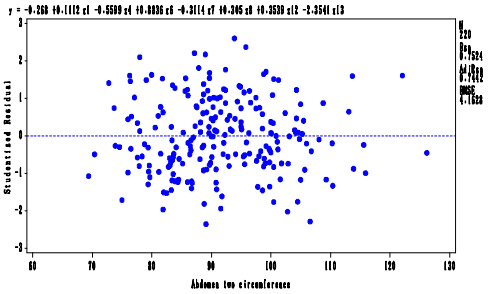


圖 3.2.4： X8(大腿圍)的殘差圖

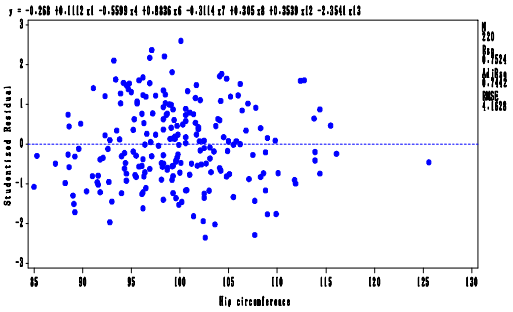


圖 3.2.5： X12(前臂圍)的殘差圖

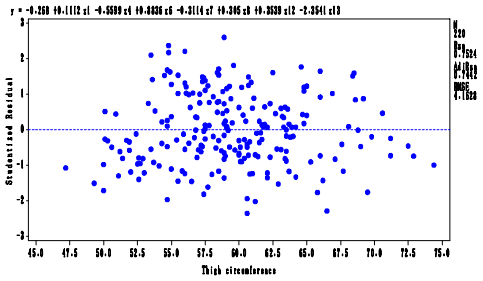


圖 3.2.6： X13(手腕圍)的殘差圖

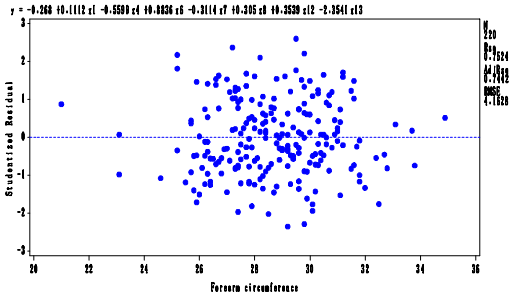
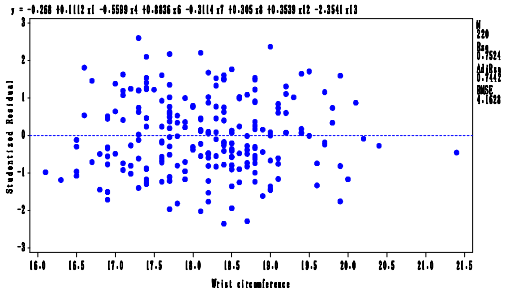


圖 3.2.7：預測值(體脂肪)的殘差圖



第三節 檢測殘差相關係數爲零

虛無假設 Ho： $\rho = 0$

對立假設 H_1： $\rho > 0$

由表 3.3.1 可看出 Pr < DW＝0.0761，大於 0.05，接受虛無假設。

虛無假設 Ho： $\rho = 0$

對立假設 H_1： $\rho < 0$

由表 3.3.1 可看出 Pr > DW＝0.9239，大於 0.05，接受虛無假設。

檢定殘差之間沒有正、負相關，所以可以說殘差沒有自我相關。

符合基本假設 $Cov(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$

表 3.3.1

Ordinary Least Squares Estimates

Durbin-Watson	1.8196
Pr < DW	0.0761
Pr > DW	0.9239

Note: Pr<DW is the p-value for testing positive autocorrelation, and Pr>DW is the p-value for testing negative autocorrelation.

體脂肪與身體 13 個測量值的關係

第四節 檢測誤差是否爲常態

虛無假設 Ho：誤差項來自常態

對立假設 H_1：誤差項非常態

由表 3.4.1 得知 p-Value 皆大於 0.05，接受虛無假設。

符合基本假設 ε_i~Normal 。

表 3.4.1 常態性檢定

Tests for Normality		
Test	—Statistic——	——p Value———
Shapiro-Wilk	0.990302	0.1477
Kolmogorov-Smirnov	0.050689	>0.1500
Cramer-von Mises	0.120552	0.0618
Anderson-Darling	0.698425	0.0711

由圖 3.4.1、圖 3.4.2 與圖 3.4.3 可看出誤差項來自常態。

圖 3.4.1：Q-Q PLOT

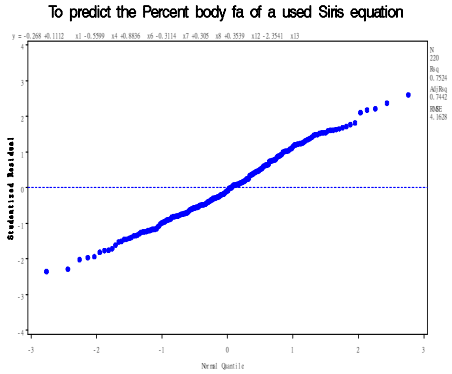


圖 3.4.2：莖葉圖、盒型圖與常態機率圖

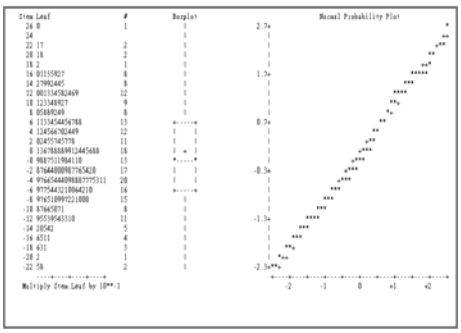
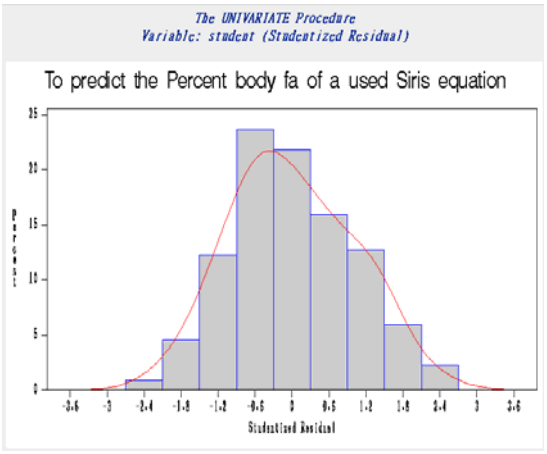


圖 3.4.3：直方圖



第四章　分析結果總結

觀察 222 筆成年男性十三個量測值(年齡、體重、身高、頸圍、胸圍、腹圍、臀圍、大腿圍、膝圍、踝圍、二頭肌圍、前臂圍、手腕圍)與體脂肪的關係，且應用散佈圖與相關係數的檢定方法，發現體脂肪與體重、胸圍、腹圍、臀圍、大腿圍確實有高度相關。

以十三個測量值爲解釋變數，體脂肪爲被解釋變數，利用 Variance Inflation Factor 檢測變數之間的多重共線性，發現體重、腹圍與臀圍具有多重共線性關係，設 X14 爲 BMI 值，做變數轉換，在以莖葉圖、盒型圖與標準化殘差檢測異常點，分析後建議將異常點刪除。

以(一)逐步選取法(Stepwise Selection)、(二)向前選取法(Forward selection)、(三)倒退消去法(Backward elimination)及(四)全部子集迴歸的 Mallow’s Cp criterion 選取重要變數，最後採用逐步選取法選擇到的重要變數：頸圍、腹圍、臀圍、大腿圍、前臂圍、手腕圍、年齡，隨著上述變數的數值增加，體脂肪增加的機率也會較高。

利用(一)DFFITS(二)Hat value(三)Cooks Distance(四)DFBETAS(五)COVRATIO 等五種方法，來檢定資料，發現並無嚴重的影響點，建議不要刪除資料。
最終模型：

$$\hat{Yi} = -0.26802 + 0.11122X_1 - 0.55994X_4 + 0.88358X_6 - 0.31137X_7 + 0.30498X_8 + 0.35391X_{12} - 2.35413X_{13}$$

Std Error (7.12605)　(0.03237) (0.22463)　(0.06936)　(0.12887)　(0.13412)　(0.19670)　(0.50157)

最後利用「莖葉圖」、「盒型圖」、「常態機率圖」與「殘差圖」等，檢定殘差平均和爲零且誤差項爲常態。

第五章　預測值

$\Sigma_{(Yi-\hat{Yi})^2}$	845.4805
ＭＳＥ	28.18268

三十筆預測值：

資料	Yi	$\hat{Yi}$	Yi-$\hat{Yi}$	(Yi-$\hat{Yi}$)^2
1	11.5	17.1758	-5.6758	32.21471
2	5.2	17.2284	-12.0284	144.6824
3	10.9	20.1132	-9.2132	84.88305
4	12.5	15.2731	-2.7731	7.690084
5	14.8	18.5642	-3.7642	14.1692
6	25.2	22.1572	3.0428	9.258632
7	24.9	17.5268	7.3732	54.36408
8	17.0	19.6121	-2.6121	6.823066
9	10.6	21.0969	-10.4969	110.1849
10	16.1	21.65	-5.5500	30.8025
11	15.4	17.3188	-1.9188	3.681793
12	26.7	24.6481	2.0519	4.210294
13	25.8	21.1367	4.6633	21.74637
14	18.6	23.3031	-4.7031	22.11915
15	24.8	22.0867	2.7133	7.361997

資料	Yi	$\hat{Yi}$	Yi-$\hat{Yi}$	(Yi-$\hat{Yi}$)^2
16	27.3	35.7644	-8.4644	71.64607
17	12.4	14.2381	-1.8381	3.378612
18	29.9	26.4323	3.4677	12.02494
19	17.0	16.4508	0.5492	0.301621
20	35.0	37.0822	-2.0822	4.335557
21	30.4	29.8992	0.5008	0.250801
22	32.6	33.0393	-0.4393	0.192984
23	29.0	30.8463	-1.8463	3.408824
24	15.2	14.358	0.8420	0.708964
25	30.2	30.6277	-0.4277	0.182927
26	11.0	15.0227	-4.0227	16.18212
27	33.6	25.9418	7.6582	58.64803
28	29.3	38.7275	-9.4275	88.87776
29	26.0	24.3615	1.6385	2.684682
30	31.9	26.5648	5.3352	28.46436

第六章　結論與建議

根據《肥胖與肥胖症-指導手冊》一書指出，肥胖的定義爲「脂肪組織的過量囤積」，因此必須正確測量體內脂肪組織的量，也就是體脂肪量。可是，由於精密測量體脂肪量需要大規模的設備，或過於花費時間及經費，故不太利用被利用。於是在判定肥胖時，一般都是採用身高・體重比的方法。近年來在判定肥胖時，除體脂肪量外，脂肪的囤積部位與疾病的關係亦受到重視。

目前國際間廣泛使用的體格指數爲 BMI（Body Mass Index），此指數不只簡便，更與體脂肪量有密切的關係。

本報告研究結果指出 BMI 確實比身高、體重兩個個別變數更適合，且去掉了共線性的問題，決定以 BMI 取代身高、體重。

由迴歸模型指出，體脂肪與年齡、頸圍、腹圍、臀圍、大腿圍、前臂圍、手腕圍有相關。

根據台灣肥胖醫學會祝年豐醫師指出，年過四十歲之後，人的身體新陳代謝率也就自然降低，由於基礎代謝率佔了每天攝取熱量的七成，也就是說，一天吃兩千大卡、有一千四百大卡是提供作爲基礎代謝。

美國加州大學發表一份以三十三名肥胖成年人作爲對象的研究也發現，相較於其他糖類，食用果糖的受試者，腹部肥胖較爲嚴重，膽固醇也比較高。中央肥胖正是代謝症的危險因子之一。高果糖攝取也可能成爲肝臟健康隱憂。台大醫院兼任主治醫師蕭敦仁指出，前年在美國肝病醫學會的年會上發表的一項動物試驗結果顯示，飲食中若含有大量精製糖，尤其是果糖，不但容易有脂肪肝，也可能使肝臟的脂肪產生過氧化反應，引發細胞衰亡、肝纖維化等病變。

建議：

(一)近年來的人喜歡喝飲料，導致攝取過多的果糖，這也是引起體脂肪過高的原因之一。

(二)年齡層越高，　應小心體脂肪的累積。腹圍、大腿圍以及前臂圍較爲寬大的人，即使體重或者 BMI 未超過標準，但並不代表身體是健康的。

參考文獻

參考書籍：

1、Kutner,M.H.,Nachtsheim,C.J.,Neter,J.,and Li.W (2005) Applied Linear Statistical Models,5th Edition,McGraw.Hill
2、黃俊英、林震岩，SAS 精研與實例，民 83 年，台北：華泰
3、吳宗正、鄭淑娥，迴歸分析，民 91 年，台北：華泰
4、陳順宇，迴歸分析，第三版，民 89 年，台北：華泰
5、祝年豐博士，肥胖與肥胖症指導手冊，民 92 年，台北：九州
6、鍾伯光，FIT or FAT 減肥手冊，民 83 年，香港：博益

參考網頁：

1、聯合報新聞網
網址：http://www.wretch.cc/blog/joycehan/2746617
2、國立屏東商業技術學院
網址：http://www.npic.edu.tw/~health/b2_3.htm
3、自由電子報
網址：http://www.libertytimes.com.tw/2009/new/jan/5/today-health3.htm
4、資料來源：卡內基美隆大學(Carnegie Mellon University，CMU)的圖書館
日期：1995.10.02
網址：http://lib.stat.cmu.edu/datasets/bodyfat

附錄

```
dm "output;clear;log;clear;program;recall;graph;cler;";
options ps=55;
title 'To predict the Percent body fa of a used Siris equation ';
data body mass index;
infile    'c：\data.txt';
input y x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13;
x14 = ( 0.45359237 * x2 ) / ( 0.3048 * x3 )**2 ;
label y='Percent body fat'
x1='age' /* in years */
x2='Weight ' /* lbs */
x3='Height '/*inches*/
x4='Neck circumference' /*cm*/
x5='Chest circumference'/*cm*/
x6='Abdomen two circumference'/* cm*/
x7=' Hip circumference'/* cm*/
x8='Thigh circumference'/* cm*/
x9 = 'Knee circumference'/* cm*/
x10 ='Ankle circumference' /* cm*/
x11='Biceps (extended) circumference '/* cm*/
x12='Forearm circumference' /* cm*/
x13='Wrist circumference '/* cm*/
x14='Body Mass Index' ;
ods html;
ods graphics on;
proc corr plots=matrix;
var y x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 ;
run;
/* The experimental PLOTS=SCATTER(NMAXVAR=2) option requests a scatter plot for
the first two variables in the VAR list.
The ALPHA= suboption requests 95% and 99% prediction ellipses.
The ELLIPSE=MEAN and ALPHA= suboptions request 95% and 99% confidence ellipses
for the mean.
*/
proc gplot;
plot y*( x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14);
symbol1 v=star c=red;
```

體脂肪與身體 13 個測量值的關係

```
proc reg;
model y= x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 / vif;
model y= x1 x14 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 / vif;
model y= x1 x4 x6 x7 x8 x12 x13 / vif;
/*p: predicted value
r: residuals
clm: a confidence interval for a mean value of y
cli    : a predictive interval for a individual value of y
dw: Durbin-Watson statistics
vif: variance inflaction factor */
symbol2 v=dot    c=blue;
run;
proc reg;
model y= x1 x14 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13/selection=stepwise;
model y= x1 x14 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13/selection=adjsquare;
model y= x1 x14 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13/selection=forward ;
model y= x1 x14 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13/selection=backward ;
run;
proc reg;
model y= x1 x14 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13/selection=cp aic best=6;
symbol1 v=dot    c=blue;
plot cp.*np.
        /chocking=red cmallows=blue
          vaxis=0 to 8 by 0.5 haxis=0 to 8 by 0.5 crame=ligr;
run;
proc reg;
model y= x1 x4 x6 x7 x8 x12 x13/r influence dw;
proc reg;
model y= x1 x4 x6 x7 x8 x12 x13/p r influence clm cli dw;    /*    influence provides
STUDENT residuals and RStudent */
output out=all student=student rstudent=rstudent;
plot y* (    x1 x4 x6 x7 x8 x12 x13 );
plot y* (    x1 x4 x6 x7 x8 x12 x13 ) / pred;
plot (student. rstudent. )*( x1 x4 x6 x7 x8 x12 x13 predicted.);
plot student. * nqq./vaxis=-4 to 4 by 1 haxis=-3 to 3 by 1.0;
run;
proc univariate;
var y x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12 x13 x14 ;
```

體脂肪與身體 13 個測量值的關係

```
run;
proc univariate normal plot ;
var student;
histogram student / cfill=ltgray    kernel(color=red)    name='MyHist';
run;
ods graphics off;
ods html close;
run;
quit ;
```