報告題名：

影響美國各洲死力不同的因素

作者：張秀媛、張淑盈

系級：統計與精算研究所

學號：M9700162、M9618737

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 一 學期

摘要

死亡，是存活在這個世界上的所有生物將來都必須面對的議題，影響死亡的因素很多，且各個因素造成死亡的機率也大不相同，因此選取了一月份平均溫度、七月份平均溫度、相對溼度、每年度降雨量、受教育年數中位數、人口密度、非白種人比例、白種人比例、人口數、每個家庭人數、家庭收入中位數、碳氫化合物潛在污染、一氧化二氮潛在污染、二氧化硫潛在污染為其潛在的重要變數，以「向前選取法」、「向後消去法」、「逐步選取法」、「校正後的複判定係數法」、「CP 選取法」、「AIC 準則」及「SBC 準則」這七種方法選出了每種方法中的重要變數，而後綜合結果，選出最終模式，還要檢查是否有影響點跟異常點、檢測是否有多重共線性問題，最後要檢查誤差是否符合四個基本假設：殘差平均是否為零、殘差是否來自常態、殘差間是否獨立、殘差變異數是否為常數。

最終選取變數為「一月份平均溫度」、「每年度降雨量」、「非白種人比例」、「二氧化硫潛在污染」是影響各城市死力不同最重要的變數。

關鍵字

死力、向前選取法、向後消去法、逐步迴歸法、校正後的複判定係數法、CP 選取法、AIC 準則、SBC 準則、散佈圖、殘差分析、複迴歸分析、選擇模式。

目次

研究動機

人類的平均壽命隨著科技的進步，生活品質的改善而增加，以往一些會大幅影響人類壽命的因素也逐一被人們克服，但同時也有一些新的影響因素出現，例如因機車數量增加，而產生了一些環境的汙染，進而成為影響人類死亡的因素之一。所以我們想要研究到底什麼因素影響死亡率有多高，哪些因素對死亡沒有影響。

因此，由不論是軍事、農業、體育、醫學、科技等各方面，都排名世界第一的美國，做為研究對象，從中挑選 58 個城市，探討影響各個城市死力不同的因素。藉由研究第一強國的方式，而推演至其它小國未來可能也會因為相同的原因，而導致死力的上升或下降，進而做到事先防範的目的。

決定變數

我們想要研究影響死亡的因素，但不以「死亡率」這個會受單位影響的變數作為研究對象，而是以某一個 X 歲的人，在 t 時間點瞬間死亡的強度，也就是死力(Y)，這個沒有單位的變數作為我們的應變數，由於此變數沒有單位，所以計算出來的值會是一個純粹的數字，也就會是一個客觀公認的數值，可直接觀察出它的大小。

一般來說，全世界一月份的溫度都會偏低，所以我們想要探討一月份的溫度，對死力造成的嚴重性，因此以一月份平均溫度(華氏溫標)(x1)作為第一個要討論的自變數。又七月份可說是一年中氣溫最高的時候，所以我們也想探討高溫度是否也會對死力造成影響，因此以七月份平均溫度(華氏溫標)(x2)作為第二個要討論的自變數。

由於不同的生物都有適合自己乾燥或潮濕的環境，所以不同的溼度將會對他們的健康造成影響，又不同的地理位置會產生不同的氣侯，因此我們以相對溼度(百分比)(x3)作為第三個要討論的自變數。

美國國土面積超過 962 萬平方公里，由於幅員遼闊和廣泛的地理特徵，美國幾乎有著世界上所有的氣候類型，有些地區瀕臨沿海降雨量因而較多，而有些地區則深處內陸降雨量相對地來說就會大幅地減少，因雨量多寡是影響農業生長的重要因素，它會牽涉到人類食物的供給，此外，雨量多寡也會牽涉到生活環境的品質，影響到人們的健康，因此我們以每年度降雨量(厘米)(x4)作為我們第四個要討論的自變數。

至 2006 年 10 月 16 日，美國人口總計已達到 3 億人，是世界上人口僅次於中國、印度人口的第三大國，由於人口眾多，再加上美國是一個種族差異極大的多民族國家，全國有 31 個種族，所以每個人受教育的程度也就會有所差異，因教育程度的不同，對生命的看法也會有個自的想法，因此我們以受教育年數中位

數(年數)(x5)作為第五個要討論的自變數。

目前全美國有大約 77%的人口居住於城市地區，其中又有半數以上集中於 37 座主要的大城市。這些城市也形塑了美國的文化、傳統、和經濟。在 2004 年，全美有 251 個超過了 100,000 人的都市，以及 9 個超過 1,000,000 人的大都市，包括了好幾個重要的全球城市，例如紐約市、洛杉磯、和芝加哥。此外，若將市中心外的都會區域也算進去的話，美國有 50 個超過了 1,000,000 人的大都會。表一我們列出美國幾個著名城市的人口密集度，我們想要知道人口的稠密與疏散，是否會有不同的死力，因此我們以人口密度(人口/每平方公里)(x6)作為第六個要討論的自變數。

大多數的美國人（在 2004 年有 74.67%）是歐洲白人移民的後代，這些移民當時在首批殖民地安居，許多是在內戰後的「南部重建運動」中來到美國的，由於非白人的移民以及少數族群的高出生率的緣故，非拉美裔人的白人比率正在逐漸下跌。我們想要知道白種人與非白種人會不會因為血統、習性或思想的不同而產生不同的死力，因此我們分別以非白種人比例(百分比)(x7)及白種人比例(百分比)(x8)作為我們第七、第八個要討論的自變數。

表一我們列出美國人口數排行前幾名的城市，因美國人口數眾多，我們想要探討城市中人口數的多寡，是否會產生不同的死力效果，像是人口數愈多是不是會引起愈多的紛爭以致造成不必要的死傷等，因此我們以人口數(人)(x9)作為第九個要討論的自變數。

每個家庭的成員數都不盡相同，一般來說成員數愈多，亦即兄弟姊妹愈多，每個人受到的照顧就會相對地減少，死力也就會提升，因此我們想要探討家庭人口數對死力的影響，所以以每個家庭人數(人口數/家庭數)(x10)作為第十個要討論的自變數。

不同的家庭所得，也就會有不同的生活品質，這關係到全家人的健康，影響到死力的上升或下降，因此我們以家庭收入中位數(美元)(x11)作為第十一個要

討論的自變數。

由於科技的進步，不同型態的工廠紛紛成立，汽、機車的數量也逐年增加，它們所排放出來的廢水、廢氣也常造成環境上的污染，因而影響到人們的健康，因此我們分別以碳氫化合物潛在污染(ppm)(x12)、一氧化二氮潛在污染(ppm)(x13)、二氧化硫潛在污染(ppm)(x14)，作為要討論的第十二、第十三、第十四個自變數。

表一

排行	都市	人口 僅限市內	人口密集度 每平方英哩	大都會區域		地區
				百萬	排名	
1	紐約市	**8,143,197**	26,402.9	18.7	1	東北部
2	洛杉磯	**3,844,829**	7,876.8	12.9	2	西部
3	芝加哥	**2,842,518**	12,750.3	9.4	3	中西部
4	休士頓	**2,016,582**	3,371.7	5.2	7	南部
5	費城	**1,463,281**	11,233.6	5.8	4	東北部
6	鳳凰城	**1,461,575**	2,782.0	3.7	14	西部
7	聖安東尼奧	**1,256,509**	2,808.5	1.8	29	南部
8	聖地牙哥	**1,255,540**	3,771.9	2.9	17	西部
9	達拉斯	**1,213,825**	3,469.9	5.7	5	南部
10	聖荷西	**912,332**	5,117.9	1.7	30	西部

第一章 資料分析

第一節 一般敘述統計

由表二知，因每個城市造成死亡的潛在因素不同，以致不同的城市，會存在著不同的死力大小，在這 58 個不同的城市中，最小的死力值為 790.73，最大值為 1113，而平均死力大小為 940.9443，標準差為 62.9416。

由於美國幅員遼闊，所以由表二可看出一月份平均溫度最低溫為華氏 12 度，最高溫為華氏 67 度，58 個城市的平均值為華氏 33.8966 度，而標準差為華氏 10.2113 度。

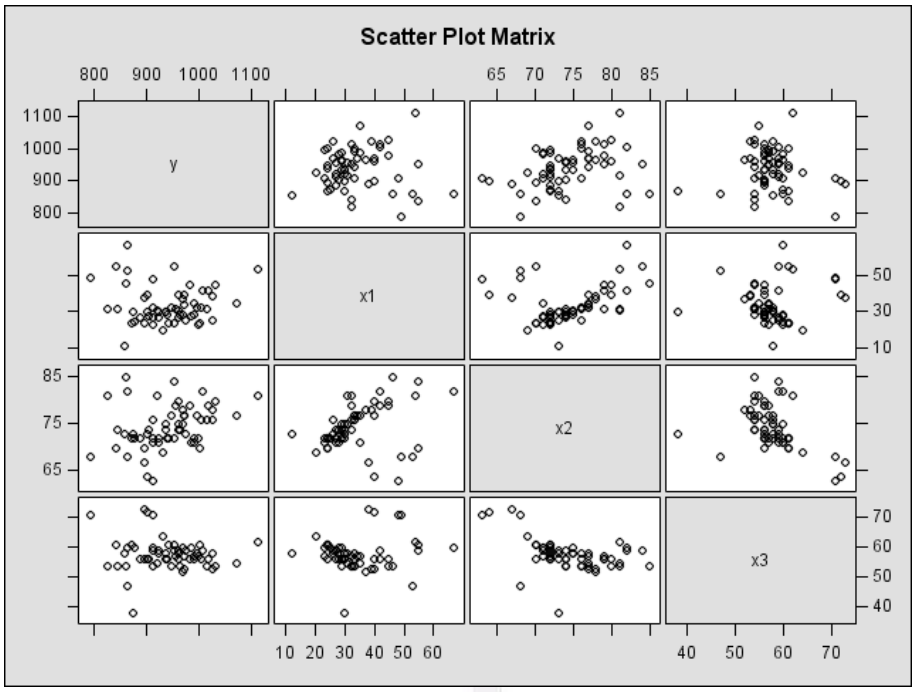
降雨量與城市是否位處沿海地帶有關，所以由表二可看出在這 58 個不同的城市中，最低降雨量為 10 厘米，最高降雨量為 65 厘米，降雨量平均值為 38.5172 厘米，標準差為 11.6743 厘米。

表二

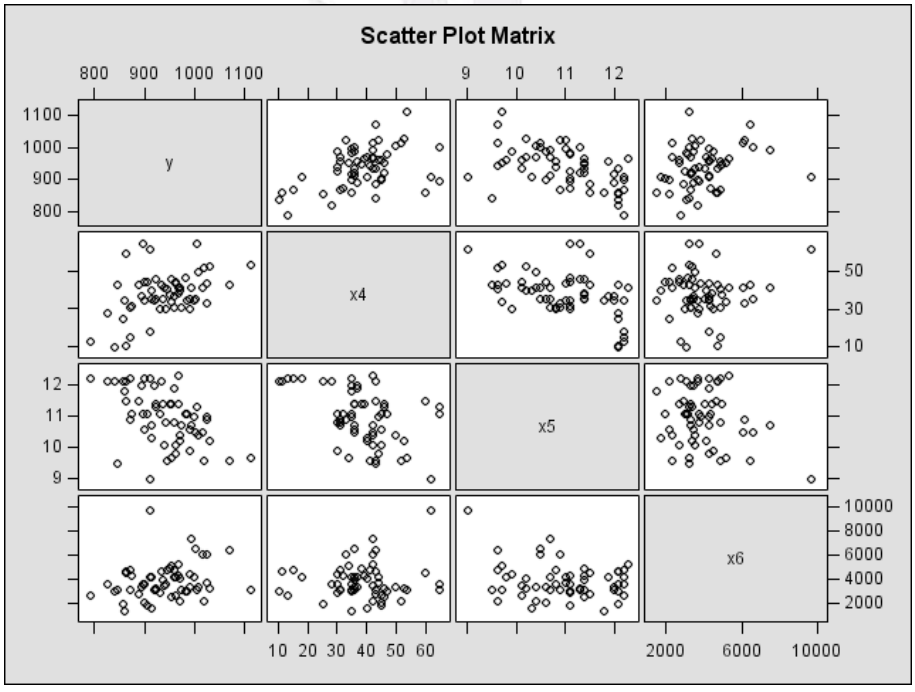
Variable	Label	個數	平均數	標準差	總和	最小值	最大值
y	死力	58	940.9443	62.9416	54575	790.73	1113
x1	一月份平均溫度	58	33.8966	10.2113	1966	12	67
x2	七月份平均溫度	58	74.4483	4.6308	4318	63	85
x3	相對溼度	58	57.7414	5.4276	3349	38	73
x4	每年度降雨量	58	38.5172	11.6743	2234	10	65
x5	受教育年數中位數	58	10.9707	0.8574	636	9	12.3
x6	人口密度	58	3918.0000	1453.0000	227268	1441	9699
x7	非白種人比例	58	11.8793	9.0761	689	0.8	38.5
x8	白種人比例	58	46.5397	4.9743	2699	33.8	62.2
x9	人口數	58	1453670.0000	1550477.0000	84312836	124833	8274961
x10	每個家庭人數	58	3.2426	0.1819	188	2.65	3.53
x11	家庭收入中位數	58	33321.0000	4476.0000	1932593	25782	47966
x12	碳氫化合物潛在污染	58	38.8966	93.3907	2256	1	648
x13	一氧化二氮潛在污染	58	23.1379	47.0545	1342	1	319
x14	二氧化硫潛在污染	58	54.9310	64.0726	3186	1	278

第二節 散佈圖

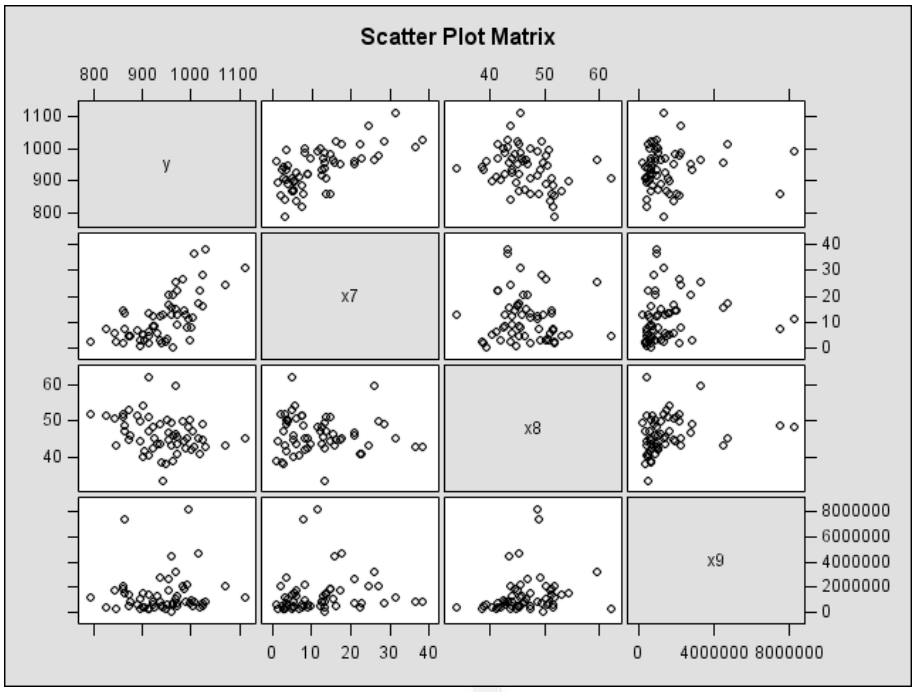
由圖一看出 Y 與 X_1 無線性關係，可知道一月份平均溫度對死力無明顯的關係；Y 與 X_2 有線性關係，但相關性不大，表示死力會隨著七月份平均溫度而呈現正相關；Y 與 X_3 無線性關係，可知道相對溼度對死力無明顯的關係。由圖二看出 Y 與 X_4 有線性關係，但相關性不大，表示死力會隨著每年度降雨量而呈現正相關；Y 與 X_5 有線性關係，但相關性不大，表示死力會隨著受教育年數中位數而呈現負相關；Y 與 X_6 無線性關係，可知道人口密度對死力無明顯的關係。由圖三看出 Y 與 X_7 有線性關係，有顯著的線性相關，表示死力的大小會隨著非白種人比例的大小而呈現正相關；Y 與 X_8 無線性關係，可知道白種人比例對死力無明顯的關係；Y 與 X_9 無線性關係，點約集中在某個範圍，亦即人口數大多集中在少於 3000000 的範圍內，所以死力大小與人口數多寡沒有很明顯的關聯。由圖四看出 Y 與 X_{10} 有線性關係，但相關性不大，表示死力會隨著每個家庭人數而呈現正相關；Y 與 X_{11} 無線性關係，可知道家庭收入中位數對死力無明顯的關係；Y 與 X_{12} 無線性關係，點約集中在某個範圍，亦即碳氫化合物潛在污染的量大多集中在 3ppm~65ppm 的範圍內，所以死力與碳氫化合物潛在污染沒有很明顯的關聯。由圖五看出 Y 與 X_{13} 無線性關係，點約集中在某個範圍，亦即一氧化二氮潛在污染的量大多集中在 1ppm~66ppm 的範圍內，所以死力與一氧化二氮潛在污染沒有很明顯的關聯；Y 與 X_{14} 散佈圖有線性關係，有顯著的線性相關，表示死力會隨著二氧化硫潛在污染而呈現正相關。此時發現似乎有異常點出現，即位於左上方的點(死力為 1113.16，二氧化硫潛在污染為 1ppm)。



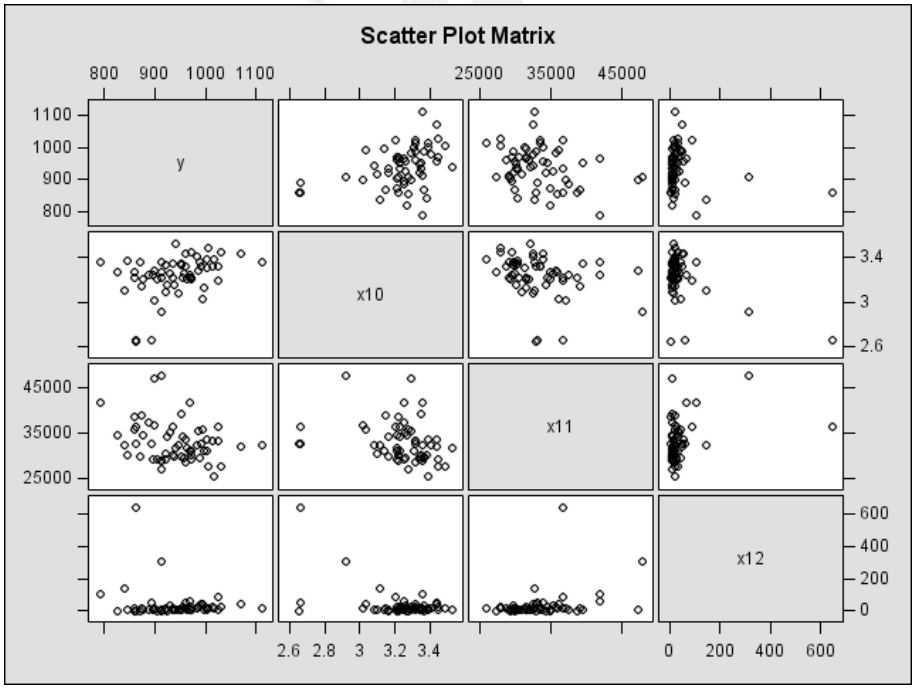
圖一



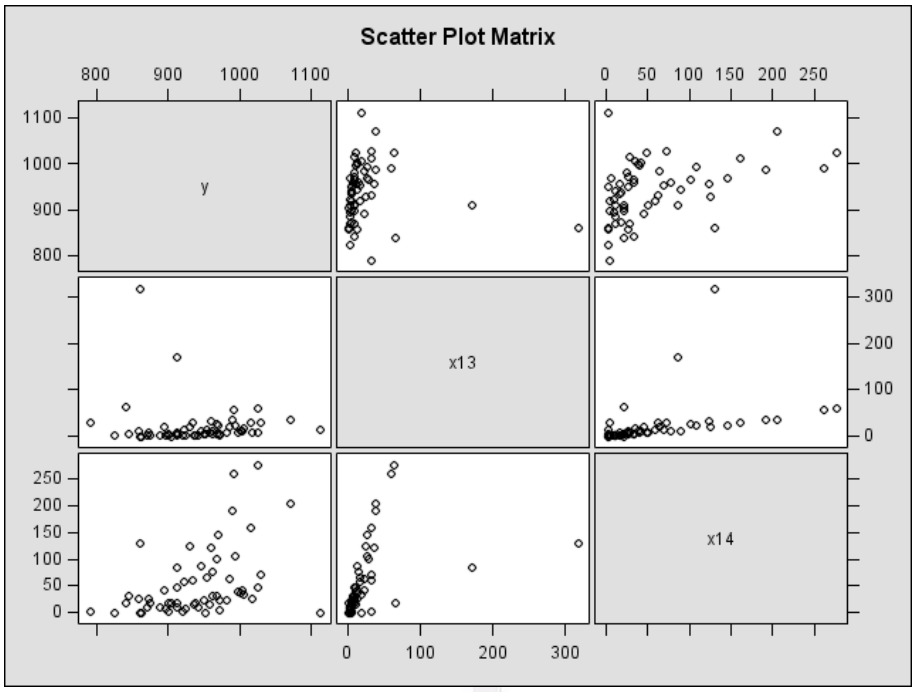
圖二


Scatter Plot Matrix
y
x7
x8
x9

圖三


Scatter Plot Matrix
y
x10
x11
x12

圖四



圖五

第三節 變數間的相關性

由表三的“Pearson 相關係數表”和“Y 對各自變數所作的散佈圖”，可大略的對所搜集到，且欲加以分析解釋的各自變數能有初步的認識及概念：

由圖二看出 Y 與 X_5 有線性關係，且由表三得知它們倆者間的相關係數只有 -50.76%，雖然相關性不大，但約略能看出死力會隨著受教育年數中位數增加而減少。

由圖三看出 Y 與 X_7 有線性關係，有顯著的線性相關，且由表三得知它們倆者間的相關係數達到 64.69%。表示死力的大小會隨著非白種人比例的大小而呈現正相關。

我們大概可以判斷,在之後所進行的“變數選取法”過程中,這兩個自變數要被選入的機會為最高,且不大輕易的會被移除。

表三

	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14
y	100.00%														
死力															
x1	-1.39%	100.00%													
一月份平均溫度	0.9177	0.0148													
x2	32.47%	31.86%	100.00%												
七月份平均溫度	0.0129	0.0148													
x3	-10.13%	8.62%	-44.20%	100.00%											
相對溼度	0.4493	0.5198	0.0005												
x4	43.35%	5.83%	47.30%	-11.77%	100.00%										
每年度降雨量	0.0007	0.6638	0.0002	0.3787											
x5	**-50.76%**	10.55%	-27.32%	18.61%	-47.36%	100.00%									
受教育年數中位數	<.0001	0.4308	0.038	0.1619	0.0002										
x6	25.36%	-7.95%	-1.18%	-14.93%	8.37%	-23.84%	100.00%								
人口密度	0.0547	0.5531	0.93	0.2634	0.5321	0.0715									
x7	**64.69%**	46.04%	**60.35%**	-11.94%	30.28%	-20.92%	-0.69%	100.00%							
非白種人比例	<.0001	0.0003	<.0001	0.3722	0.0209	0.1151	0.959								
x8	-29.09%	19.61%	-2.97%	1.67%	-11.87%	49.03%	25.05%	-5.95%	100.00%						
白種人比例	0.0268	0.1401	0.8247	0.901	0.375	<.0001	0.0578	0.6575							
x9	6.10%	23.56%	1.62%	-14.32%	-23.57%	19.44%	33.21%	11.59%	20.60%	100.00%					
人口數	0.6491	0.075	0.9039	0.2835	0.0749	0.1436	0.0109	0.3862	0.1208						
x10	36.87%	-31.79%	27.34%	-14.68%	20.30%	-38.80%	-16.21%	35.83%	-32.09%	-30.65%	100.00%				
每個家庭人數	0.0044	0.015	0.0379	0.2714	0.1265	0.0026	0.2242	0.0057	0.0141	0.0193					
x11	-28.21%	19.06%	-20.15%	12.95%	-36.60%	**50.68%**	-0.84%	-10.19%	34.88%	31.21%	-28.03%	100.00%			
家庭收入中位數	0.0319	0.1519	0.1293	0.3325	0.0047	<.0001	0.95	0.4465	0.0073	0.0171	0.0331				
x12	-18.41%	36.11%	-36.04%	-2.62%	-49.51%	29.03%	11.14%	-2.63%	16.40%	**52.88%**	-49.06%	32.59%	100.00%		
碳氫化合物潛在污染	0.1666	0.0054	0.0055	0.8453	<.0001	0.027	0.4052	0.8447	0.2185	<.0001	<.0001	0.0125			
x13	-8.38%	33.32%	-33.74%	-5.28%	-46.00%	22.82%	15.75%	1.91%	12.63%	**54.60%**	-45.14%	31.07%	**98.38%**	100.00%	
一氧化二氮潛在污染	0.5315	0.0106	0.0096	0.6937	0.0003	0.0849	0.2377	0.8871	0.3446	<.0001	0.0004	0.0176	<.0001		
x14	42.04%	-9.66%	-7.39%	-11.63%	-13.12%	-23.07%	42.09%	15.97%	-7.31%	36.49%	-0.48%	6.40%	27.78%	40.57%	100.00%
二氧化硫潛在污染	0.001	0.4708	0.5816	0.3845	0.3262	0.0815	0.001	0.2313	0.5856	0.0049	0.9714	0.6332	0.0348	0.0016	

第四節 建立模式

一、選取重要變數

（一）向前選取法(Forward Selection)

用向前選取法時，第一步為選取與依變數有最大相關的變數，意即對模式的貢獻最大者，進入迴歸方程式中，第二步為對尚未進入迴歸方程式中的預測變數加以考驗，決定納入哪一個變數，依此類推，直到沒有其他的變數符合選取的標準為止。

由表四得知，最後選取的變數有 X_1(一月份平均溫度)、X_2(七月份平均溫度)、X_4(每年度降雨量)、X_5(受教育年數中位數)、X_7(非白種人比例)、X_{14}(二氧化硫潛在污染)。

（二）向後消去法(Backward Selection)

用向後消去法時，第一步為先將全部的預測變數放入迴歸方程式中，第二步為淘汰對模式貢獻最小的變數，依序對留在迴歸方程式中的預測變項加以考驗，決定淘汰哪一個變數，直到所有預測變數均達到標準為止。

由表五得知，消去的變數有 X_3(相對濕度)、X_5(受教育年中位數)、X_9(人口數)、X_{10}(每個家庭人數)、X_{11}(家庭收入中位數)、X_{14}(二氧化硫潛在污染)。

（三）逐步選取法(Stepwise Selection)

用逐步選取法時，第一步為依據向前選取的方式，根據對模式貢獻最大者，挑選預測變項進入迴歸模式中，之後的每一步驟中，已被納入模式的所有預測變數，利用向後消去法的考驗，將不重要的變數剔除，依此類推，直到沒有變數被選取或剔除為止。

由表六得知，最後留下的變數有 X_1（一月份平均溫度）、X_2（七月份平均溫度）、X_4（每年度降雨量）、X_5（受教育年數中位數）、X_7（非白種人比例）、X_{14}（二氧化硫潛在污染）。

（四）校正後的複判定係數法（R^2_{adj} Selection）

校正後的複判定係數法為估算全部可能的迴歸模式之 R^2_{adj} 值，相互比較，以選取最大之 R^2_{adj} 為最佳最有效的迴歸模式。

由表七得知，最後所選取的變數為 X_1（一月份平均溫度）、X_2（七月份平均溫度）、X_4（每年度降雨量）、X_5（受教育年數中位數）、X_6（人口密度）、X_7（非白種人比例）、X_8（白種人比例）、X_{10}（每個家庭人數）、X_{12}（碳氫化合物潛在污染 ）、X_{13}（一氧化二氮潛在污染）。

（五）C_p 選取法

C_p 選取法為估算全部可能的迴歸模式之 C_p 值，選取的迴歸模式必須使得 C_p 值夠小且滿足 C_p 值接近 p 的條件，其中 p 為參數個數。

由圖六，也就是 C_p 圖可知，當 p=7 時，p 值最接近 C_p 值，所以選擇參數個數為 7 的組合。

由表八可知，當 p=7，且 Cp 最接近 p 時選擇的參數為 X_1（一月份平均溫度）、X_2（七月份平均溫度）、X_4（每年度降雨量）、X_7（非白種人比例）、X_8（白種人比例）、X_{14}（二氧化硫潛在污染）。

(六) *AIC and SBC Criteria*

AIC(*Akaike's information criterion*)與 *SBC*(*Schwarz' Bayesian criterion*)主要是用以判斷新增的預測變數是否適當，其定義如下：

$$AIC_p = n\ln SSE_p - n\ln n + 2p$$
$$SBC_p = n\ln SSE_p - n\ln n + [\ln n]p$$

上述定義的涵義，首先是 $n\ln SSE_P$，當 p 越大時，$n\ln SSE_P$ 將會降低，而第二項 $n\ln n$ 在樣本數固定下，為一個固定之常數，最後第三項將遠著 p 越大時而越大，在此準則下，只要 $2p$（對於 AIC_P）或是 $[\ln n]p$（對於 SBC_p）不是太大，SSE_p 越小越佳，因此我們在選取迴歸模式時，要選取使得 AIC、SBC 越小的越好。

由表九可知，以 AIC 準則所選擇的變數為 X_1、X_4、X_5、X_7、X_{14}，以 SBC 準則所選擇的變數為 X_1、X_2、X_4、X_5、X_7、X_{14}。

（七）綜合結果

綜合了以上七種方法，我們選取出現四次以上的變數，有 X_1（一月份平均溫度）、X_2（七月份平均溫度）、X_4（每年度降雨量）、X_5（受教育年數中位數）、X_7（非白種人比例）、 X_{14}（二氧化硫潛在污染）。

表四

Summary of Forward Selection								
Step	Variable Entered	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x7	非白種人比例	1	0.4185	0.4185	51.0884	40.3	<.0001
2	x5	受教育年數中位數	2	0.1449	0.5634	26.9017	18.25	<.0001
3	x1	一月份平均溫度	3	0.0729	0.6363	15.7281	10.82	0.0018
4	x14	二氧化硫潛在污染	4	0.044	0.6802	9.7844	7.29	0.0093
5	x4	每年度降雨量	5	0.0288	0.709	6.5854	5.14	0.0276
6	x2	七月份平均溫度	6	0.0174	0.7264	5.4446	3.24	0.0778

表五

Summary of Backward Elimination								
Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x3	相對溼度	13	0.0001	0.762	13.0122	0.01	0.9127
2	x11	家庭收入中位數	12	0.0006	0.7613	11.1291	0.12	0.7311
3	x14	二氧化硫潛在污染	11	0.0019	0.7594	9.4785	0.36	0.549
4	x9	人口數	10	0.0044	0.755	8.2787	0.85	0.3623
5	x10	每個家庭人數	9	0.006	0.749	7.3624	1.15	0.289
6	x5	受教育年數中位數	8	0.0062	0.7428	6.4871	1.19	0.2807

表六

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr>F
1	x7		非白種人比例	1	0.4185	0.4185	51.0884	40.3	<.0001
2	x5		受教育年數中位數	2	0.1449	0.5634	26.9017	18.25	<.0001
3	x1		一月份平均溫度	3	0.0729	0.6363	15.7281	10.82	0.0018
4	x14		二氧化硫潛在污染	4	0.044	0.6802	9.7844	7.29	0.0093
5	x4		每年度降雨量	5	0.0288	0.709	6.5854	5.14	0.0276
6	x2		七月份平均溫度	6	0.0174	0.7264	5.4446	3.24	0.0778

表七

Adjusted R-Square Selection Method			
Number in Model	Adjusted R-Square	R-Square	Variable in Model
10	0.7028	0.7550	x1 x2 x4 x5 x6 x7 x8 x10 x12 x13
9	0.7019	0.7490	x1 x2 x4 x5 x6 x7 x8 x12 x13
11	0.7019	0.7594	x1 x2 x4 x5 x6 x7 x8 x9 x10 x12 x13
10	0.7014	0.7538	x1 x2 x4 x5 x6 x7 x8 x9 x12 x13
7	0.7013	0.7380	x1 x2 x4 x6 x7 x8 x14
9	0.7013	0.7484	x1 x2 x4 x6 x7 x8 x12 x13 x14
8	0.7008	0.7428	x1 x2 x4 x5 x6 x7 x8 x14
8	0.7008	0.7428	x1 x2 x4 x6 x7 x8 x12 x13
9	0.7005	0.7478	x1 x2 x4 x5 x7 x9 x10 x12 x13
10	0.7005	0.7530	x1 x2 x4 x5 x6 x7 x8 x12 x13 x14

表八

Mallows' Cp Selection Method					
Number in	p	C(p)	C(p)-p	R-Square	Variable in Model
7	8	5.3498	-2.6502	0.7380	x1 x2 x4 x6 x7 x8 x14
6	7	5.4446	-1.5554	0.7264	x1 x2 x4 x5 x7 x14
6	7	5.7923	-1.2077	0.7245	x1 x4 x6 x7 x8 x14
8	9	6.4778	-2.5222	0.7428	x1 x2 x4 x5 x6 x7 x8 x14
8	9	6.4871	-2.5129	0.7428	x1 x2 x4 x6 x7 x8 x12 x13
7	8	6.5183	-1.4817	0.7315	x1 x3 x4 x6 x7 x8 x14
5	6	6.5854	0.5854	0.7090	x1 x4 x5 x7 x14
7	8	6.6248	-1.3752	0.7309	x1 x2 x4 x5 x7 x8 x14
7	8	6.6342	-1.3658	0.7309	x1 x2 x4 x5 x6 x7 x14
7	8	6.6419	-1.3581	0.7308	x1 x2 x4 x5 x7 x10 x14
8	9	6.7315	-2.2685	0.7414	x1 x2 x4 x6 x7 x8 x12 x14
6	7	6.8403	-0.1597	0.7187	x1 x2 x4 x7 x8 x14
7	8	6.8779	-1.1221	0.7295	x1 x2 x4 x5 x7 x9 x14
8	9	7.0028	-1.9972	0.7399	x1 x2 x4 x6 x7 x8 x13 x14

表九

Number in Model	C(p)	R-Square	AIC	SBC	Variables in Model
4	20.1888	0.6342	431.1639	441.4661	x2 x5 x7 x14
4	23.9345	0.6141	434.2654	444.5677	x1 x2 x4 x7
4	24.4067	0.6115	434.6450	444.9472	x2 x4 x5 x7
4	24.9963	0.6084	435.1153	445.4175	x1 x2 x7 x14
4	50.1384	0.4735	452.2816	462.5838	x2 x4 x5 x14
4	53.0688	0.4578	453.9881	464.2903	x1 x4 x5 x14
4	56.0002	0.4420	455.6464	465.9486	x1 x2 x4 x14
4	61.2709	0.4138	458.5137	468.8160	x1 x2 x5 x14
4	78.8354	0.3195	467.1592	477.4614	x1 x2 x4 x5
5	8.2396	0.7090	419.8895	432.2522	x1 x4 x5 x7 x14
5	10.0032	0.6995	421.7454	434.1080	x1 x2 x4 x7 x14
5	12.4516	0.6864	424.2272	436.5898	x1 x2 x5 x7 x14
5	13.5970	0.6803	425.3527	437.7153	x2 x4 x5 x7 x14
5	14.7380	0.6741	426.4527	438.8153	x1 x2 x4 x5 x7
5	52.0904	0.4737	454.2532	466.6159	x1 x2 x4 x5 x14
6	7.0000	0.7264	418.3175	432.7406	x1 x2 x4 x5 x7 x14

二、個別係數檢定及最終模式

由表十得知，**配適的迴歸線**如下：

$$y_i = 1214.092 - 1.55877x_1 - 2.47975x_2 + 1.3872x_4 - 14.7357x_5 + 4.9419x_7 + 0.2517x_{14}$$

且知只有 X_2 係數的 p-value 大於 0.05，所以這個係數是不顯著的，無法拒絕虛無假設，因此先將 x2 這個變數刪除，再做一次個別係數檢定。

由表十一得知，**配適的迴歸線**如下：

$$y_i = 1364.325 - 1.79717x_1 - 3.17277x_4 + 0.97566x_5 - 20.9967x_7 + 5.59901x_{14}$$

刪除 x2 這個變數後，又發現 x5 係數的 p-value 大於 0.05，所以這個係數是不顯著的，無法拒絕虛無假設，因此再將 x5 這個變數刪除，又再做一次個別係數檢定。

刪除 x2、x5 這些不顯著的變數後，再做一次檢定，檢定結果可由表十二得知剩餘變數中所有係數的 p-value 皆小於 0.05，可拒絕虛無假設，所以它們都是顯著的，得**最終模式的配適迴歸線**如下：

$$y_i = 869.4615 - 1.8086x_1 + 1.62043x_4 + 4.42679x_7 + 0.3238x_{14}$$

表十

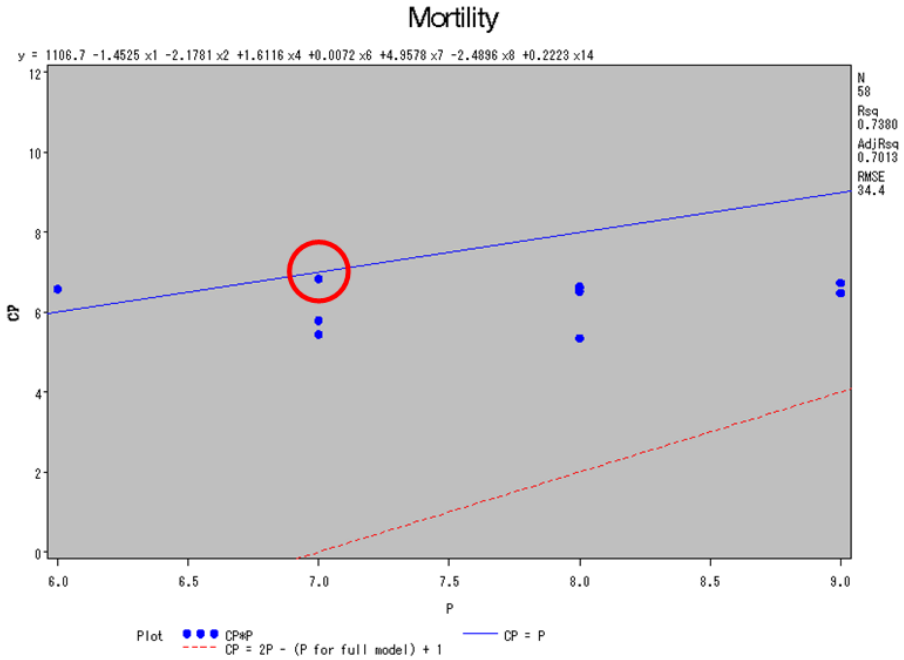
Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > | t |	Variance Inflation
Intercept	Intercept	1	1214.092	129.6183	9.37	<.0001	0.0000
x1	一月份平均溫度	1	-1.55877	0.53163	-2.93	0.005	1.3865
x2	七月份平均溫度	1	-2.47975	1.37771	-1.8	0.0778	1.9151
x4	每年度降雨量	1	1.3872	0.50818	2.73	0.0087	1.6559
x5	受教育年數中位數	1	-14.7357	6.58789	-2.24	0.0297	1.5010
x7	非白種人比例	1	4.9419	0.72016	6.86	<.0001	2.0101
x14	二氧化硫潛在污染	1	0.2517	0.08066	3.12	0.003	1.2566

表十一

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > | t |	Variance Inflation
Intercept	Intercept	1	1364.325	130.065	10.49	<.0001	0.0000
x1	一月份平均溫度	1	-1.79717	0.5686	-3.16	0.0026	1.3579
x4	每年度降雨量	1	-3.17277	1.4695	-2.16	0.0355	1.8653
x5	受教育年數中位數	1	0.97566	0.5304	1.84	0.0716	1.5444
x7	非白種人比例	1	-20.9967	6.7817	-3.1	0.0032	1.3618
x14	二氧化硫潛在污染	1	5.59901	0.7443	7.52	<.0001	1.8382

表十二

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > | t |	Variance Inflation
Intercept	Intercept	1	869.4615	25.40552	34.22	<.0001	0.0000
x1	一月份平均溫度	1	-1.8086	0.55018	-3.29	0.0018	1.3442
x4	每年度降雨量	1	1.62043	0.44796	3.62	0.0007	1.1647
x7	非白種人比例	1	4.42679	0.66062	6.7	<.0001	1.5310
x14	二氧化硫潛在污染	1	0.3238	0.07996	4.05	0.0002	1.1177



圖六 C_P 圖

第二章 診斷模式

第一節　多重共線性的診斷

多重共線性(Multicollinearity)是指在迴歸模式中，某些自變數或所有自變數之間具有高度線性相關的現象。而變異數膨脹因子(Variance Inflation Factor，簡稱 VIF)是一種經常使用於發現多重共線性的正式診斷方法，它主要在量測迴歸係數之變異數，相對於預測變數間的無線性關係之膨脹量。判斷準則為檢定結果中若發現 VIF 大於 10 的話，就表示自變數間有高度的相關性存在。

由表十二得知，因為各變數的 VIF 值皆**小於 10**，故自變數間無高度的相關性存在，亦即**無多重共線性發生**。

另外，VIF 之平均值 $(\overline{\text{VIF}}) = \frac{\sum_{K=1}^{4}(VIF)_K}{4} = 1.2894$ 也小於 10，故無嚴重的多重共線性存在。

第二節　殘差四個基本假設的檢定

（一） 檢查殘差平均是否為 0

要檢查殘差平均數是否為 0，首先我們必須設立虛無假設（$H_0:\mu=0$）與對立假設（$H_1:\mu\neq 0$），利用 Student's test、Sign test signed 與 Rank test 這三種檢定方法來判斷殘查平均數等於 0 是否顯著。

由表十三可看出三種檢定的p-value值皆明顯大於0.05，所以不能拒絕虛無假設，即殘差平均數為0，符合基本假設$E(\varepsilon_i)=0$。

（二） 檢查殘差是否來自常態

要檢查殘差是否來自常態分配，首先我們必須設立虛無假設（$H_0:e_i \sim Normal$）與對立假設（$H_1:$反對H_0），利用 Shapiro-Wilk test、Kolmogorov-Smimov test、Cramer-von Mises 與 Anderson-Darling test 這四種檢定方法來判斷殘查來自常態分配是否顯著。

由表十三可看出 p-value 的值皆明顯大於 0.05，所以不能拒絕虛無假設，即殘差來自常態分配，符合基本假設$\varepsilon_i \sim Normal$。或由圖七可看出，常態分配機率圖（P-P 圖）為一 45^0 之直線，所以殘差來自常態分配。

（三） 檢查殘差間是否獨立

要檢查殘差間是否獨立，首先我們必須設立虛無假設（$H_0:\rho=0$）與對立假設（$H_1:\rho>0$），利用 Durbin-Watson D test 這種檢定方法來判斷殘差間獨立是否顯著。

由表十三可知 d=1.666，接近 2，所以我們可以說$\hat{\rho}$=0，即殘差間沒有相關，符合基本假設$Cov(\varepsilon_i,\varepsilon_j)=0 \ \ for\ all\ i\neq j$。

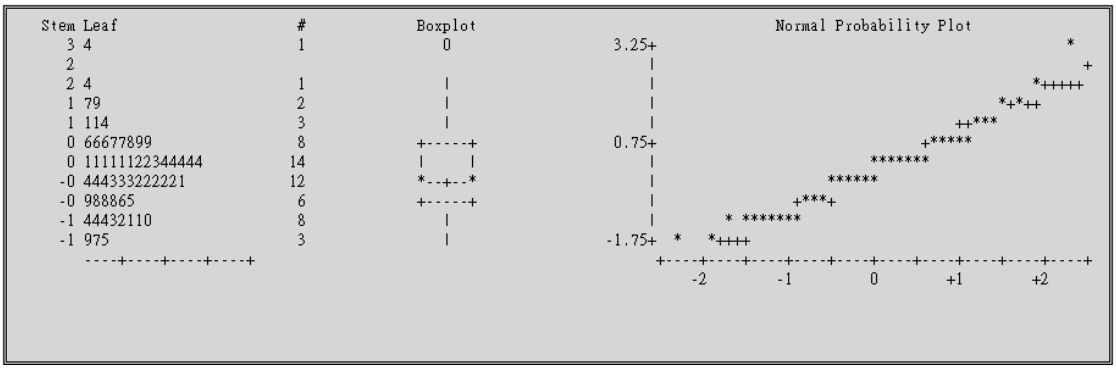
（四） 檢查殘差變異數是否為常數

由圖八至圖十二的殘差圖中可以看出，殘差落在以 0 為中心線的區塊內，呈現均勻散佈，沒有逐漸變大或逐漸變小的趨勢，所以我們可以說誤差的變異為一常數，也就是符合基本假設$Var(\varepsilon_i)=\sigma^2$。

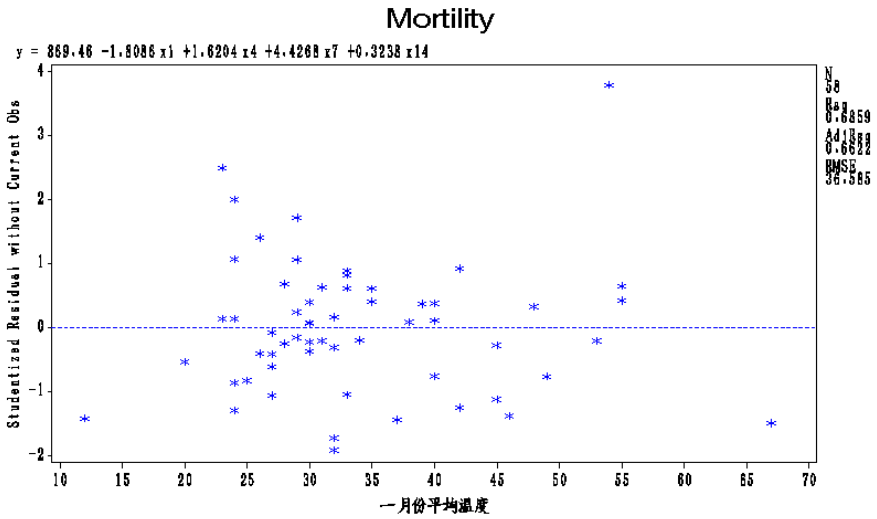
綜合上述，我們得到$\varepsilon_i \overset{iid}{\sim} Normal(0,\sigma^2)\quad i=1,2,.....,n$，亦即滿足誤差的四個基本假設。

表十三

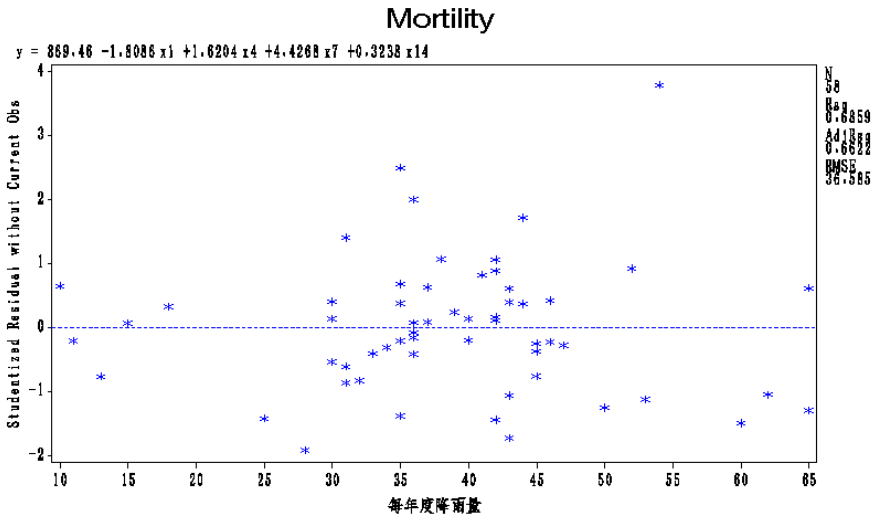
Tests for Location:Mu0=0				
Test	Statistic		p Value	
Student's t	t	-0.05489	Pr > | t |	0.95640
Sign	M	0	Pr >= | M	1.00000
Signed Rank	S	-48.5	Pr >= | S |	0.71000
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.964248	Pr < W	0.0852
Kolmogorov-Smirnov	D	0.076418	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.053347	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.414776	Pr > A-Sq	>0.2500
Tests for Independence				
Durbin-Watson D			1.666	
Number of Observations			58	
1st Order Autocorrelation			0.156	


Stem Leaf # Boxplot
3 4 1 0
2 1
2 4 1
1 79 2
1 114 3
0 66677899 8
0 11111122344444 14
-0 444333222221 12
-0 988865 6
-1 44432110 8
-1 975 3
Normal Probability Plot
3.25+
0.75+
-1.75+
-2 -1 0 +1 +2

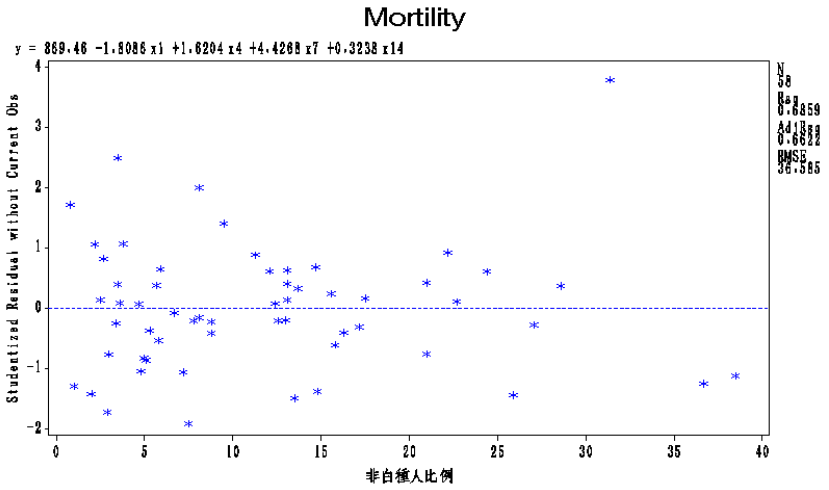
圖七


Mortility
y = 869.46 -1.8086 x1 +1.6204 x4 +4.4268 x7 +0.3238 x14
Studentized Residual without Current Obs
一月份平均溫度

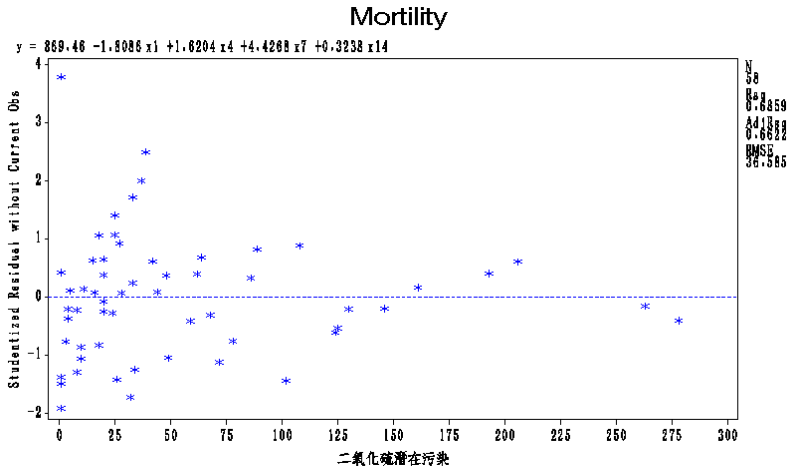
圖八


Mortility
y = 869.46 -1.8086 x1 +1.6204 x4 +4.4268 x7 +0.3238 x14
Studentized Residual without Current Obs
每年度降雨量

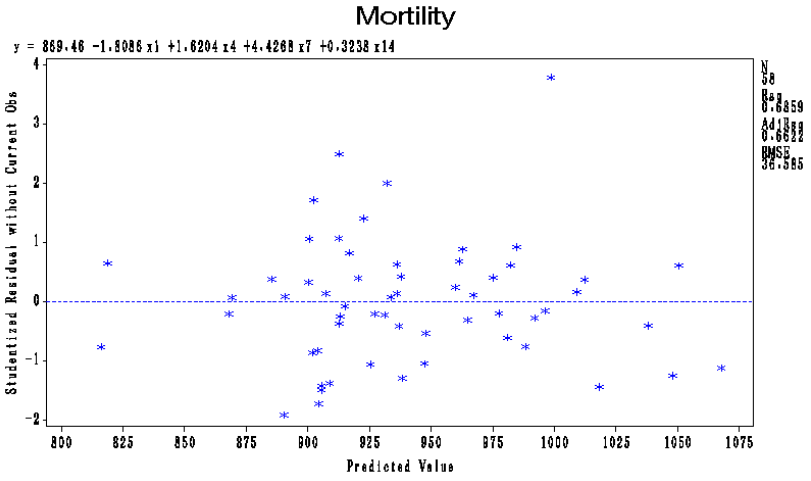
圖九


Mortility
y = 869.46 -1.8086 x1 +1.6204 x4 +4.4268 x7 +0.3238 x14
Studentized Residual without Current Obs
非白種人比例

圖十


Mortility
y = 869.46 -1.8086 x1 +1.6204 x4 +4.4268 x7 +0.3238 x14
Studentized Residual without Current Obs
二氧化硫潛在污染

圖十一


Mortility
y = 869.46 -1.8086 x1 +1.6204 x4 +4.4268 x7 +0.3238 x14
Studentized Residual without Current Obs
Predicted Value

圖十二

第三節 檢查影響點與異常點

一、 異常點偵測

偵測異常點的方法有 Residual、Studentized Residual(STUDENT)以及 Studentized Deleted Residual(RSTUDENT)這三種方法。表十四中的 0 與 1，是由 Excel 的 IF 函數判斷出來的結果，其中 1 代表滿足異常點的條件；反之，0 代表一定不會是異常點。

(一) Residual

判斷準則為

若 $|e_i| > 3\hat{\sigma}$，表示可能為異常點，其中 $\hat{\sigma} = \sqrt{MSE} = 36.58459$ 。

由表十四發現用 Excel 的 IF 函數做判斷，在觀察值 36 的地方出現 1 的結果，因此由 Residual 的偵測方法，得知觀察值 36 可能為異常點。

(二) Studentized Residual(STUDENT)

判斷準則為

若 $|r_i| > 3$ ，表示可能為異常點。

由表十四發現用 Excel 的 IF 函數做判斷，在觀察值 36 的地方出現 1 的結果，因此由 STUDENT 的偵測方法，再次得到觀察值 36 可能為異常點。

（三）Studentized Deleted Residual(RSTUDENT)

判斷準則為

若 $|t_i| > 3$ ，表示可能為異常點。

由表十四發現用 Excel 的 IF 函數做判斷，在觀察值 36 的地方出現 1 的結果，因此由 RSTUDENT 的偵測方法，又再次得到觀察值 36 可能為異常點。

表十四

異常點						
Obs	Residual (ei)	root mse= 36.58459	Student Residual (ri)	3	Rstudent (ti)	3
1	-15.1545	0	-0.42	0	-0.4164	0
2	85.1695	0	2.38	0	2.494	0
3	59.8124	0	1.684	0	1.7147	0
4	-9.6817	0	-0.276	0	-0.2732	0
5	20.7352	0	0.61	0	0.606	0
6	-37.3219	0	-1.121	0	-1.1242	0
32	-18.9024	0	-0.536	0	-0.5325	0
33	-47.9213	0	-1.406	0	-1.4191	0
34	-27.2455	0	-0.759	0	-0.7564	0
35	-8.0592	0	-0.224	0	-0.2224	0
36	114.5349	1	3.384	1	3.7858	1
37	31.8215	0	0.885	0	0.8836	0
38	5.7756	0	0.164	0	0.1623	0
39	-5.0732	0	-0.158	0	-0.1562	0
40	3.116	0	0.0873	0	0.0865	0
41	37.8627	0	1.06	0	1.0609	0
42	29.2045	0	0.823	0	0.8209	0
51	38.1219	0	1.065	0	1.066	0
52	49.6394	0	1.389	0	1.4013	0
53	4.8905	0	0.137	0	0.1361	0

二、 影響點偵測

偵測影響點的方法有 DFFITS、The hat matrix elements h_{ii}、*Cook's* distance statistic D_i、DFBETAS、COVRATIO 這五種方法。

表十五中的 0 與 1，是由 Excel 的 IF 函數判斷出來的結果，其中 1 代表滿足影響點的條件；反之，0 代表一定不會是影響點。

(一) DFFITS

判斷準則為：

若 $|DFFITS_i| > 2\sqrt{\frac{p}{n}}$ ，表示可能為影響點。

其中 $2\sqrt{\frac{p}{n}} = 2 \times \sqrt{\frac{5}{58}} = 0.58722$ 。

利用 Excel 的 IF 函數做判斷，在觀察值 31、36、57 的地方出現 1 的結果，如表十五，亦即，這些觀察值的 $|DFFITS_i|$ 值皆大於 0.58722，故用 DFFITS 的偵測方法，得到這些點可能為影響點。

(二) The hat matrix elements h_{ii}

判斷準則為：

若 $h_{ii} > 2\frac{p}{n}$ ，表示可能為影響點。

其中 $2 \times \frac{p}{n} = 2 \times \frac{5}{58} = 0.17241$ 。

利用 Excel 的 IF 函數做判斷，在觀察值 6、12、28、31、39、46、48、57 的地方出現 1 的結果，如表十五，亦即，這些觀察值的 h_{ii} 值皆大於 0.17241，故用 hat value 的偵測方法，得到這些點可能為影響點。

（三）*Cook's* distance statistic D_i

判斷準則為：

若 $D_i > 1$，表示可能為影響點。

利用 Excel 的 IF 函數做判斷，如表十五所示，結果沒有任何觀察值有出現 1，故用 Cook's D 的偵測方法，並無發現影響點存在。

（四）DFBETAS

判斷準則為：

若 $|DFBETA_i| > \frac{2}{\sqrt{n}}$，表示可能為影響點。

其中 $\frac{2}{\sqrt{n}} = \frac{2}{\sqrt{58}} = 0.26261$ 。

利用 Excel 的 IF 函數做判斷，如表十五所示，在許多觀察值的地方都有出現一次以上 1 的結果，因此我們選取出現兩次以上 1 的結果的觀察值，它們分別為觀察值 31、33、36、57，亦即，這些觀察值的 $|DFBETA_i|$ 值皆大於 0.26261，故用 DFBETAS 的偵測方法，得到這些點可能為影響點。

（五）COVRATIO

判斷準則為

若 $COVRATIO_i > 1 + 3 \times \frac{p}{n}$ 或 $COVRATIO_i < 1 - 3 \times \frac{p}{n}$ ，

表示可能為影響點。

其中 $1 + 3 \times \frac{p}{n} = 1 + 3 \times \frac{5}{58} = 1.25862$

$1 - 3 \times \frac{p}{n} = 1 - 3 \times \frac{5}{58} = 0.74138$

利用 Excel 的 IF 函數做判斷，如表十五所示，在觀察值 2、12、28、31、36、39、46、48 的地方出現 1 的結果，亦即，這些觀察值的 $COVRATIO_I$ 值不是大於 1.25862，就是小於 0.74138，故用 COVRATIO 的偵測方法，得到這些點可能為影響點。

表十五

影響點																		
Obs	DFFITS	p=5 n=58	Hat Diag H	>2p/n	Cook's D	>1	DFBETAS Intercept		x1		x4		x7		x14		Cov Ratio	
2	0.5276	0	0.0428	0	0.051	0	0.3397	1	-0.2603	0	-0.0504	0	-0.1137	0	-0.0925	0	0.6532	1
3	0.4239	0	0.0576	0	0.035	0	-0.0367	0	0.0437	0	0.2072	0	-0.3062	1	0.0139	0	0.8865	0
6	-0.5132	0	0.1725	1	0.052	0	0.0842	0	0.0489	0	-0.0479	0	-0.4009	1	0.0324	0	1.1788	0
9	0.3951	0	0.0377	0	0.03	0	0.2935	1	-0.2596	0	-0.0726	0	0.0568	0	-0.1191	0	0.7905	0
12	-0.2258	0	0.2338	1	0.01	0	0.0024	0	0.0289	0	0.0035	0	-0.0107	0	-0.1982	0	1.4127	1
27	-0.3776	0	0.0459	0	0.027	0	0.0533	0	-0.086	0	-0.1717	0	0.2696	1	-0.0035	0	0.8738	0
28	-0.1131	0	0.2262	1	0.003	0	0.0072	0	-0.0744	0	0.0561	0	0.0344	0	-0.0417	0	1.4155	1
30	-0.5603	0	0.1673	0	0.062	0	-0.0152	0	0.1255	0	0.0144	0	-0.4809	1	0.1611	0	1.1393	0
31	-1.0417	1	0.3284	1	0.212	0	0.8358	1	-0.8979	1	-0.5383	1	0.5042	1	-0.0697	0	1.329	1
33	-0.5527	0	0.1317	0	0.06	0	-0.517	1	0.4221	1	0.2461	0	-0.0943	0	0.1814	0	1.0476	0
36	1.5535	1	0.1441	0	0.386	0	-0.6424	1	0.5219	1	0.3341	1	0.6917	1	-0.4608	1	0.3805	1
39	-0.0845	0	0.2264	1	0.001	0	0.018	0	-0.0097	0	-0.0143	0	0.0268	0	-0.0802	0	1.4186	1
46	0.3548	0	0.2305	1	0.025	0	0.0287	0	0.2264	0	-0.2097	0	-0.0892	0	-0.0379	0	1.3729	1
47	0.1169	0	0.1129	0	0.003	0	0.0128	0	0.0612	0	-0.0762	0	-0.0012	0	0.0176	0	1.2272	0
48	-0.3581	0	0.1799	1	0.026	0	-0.0619	0	-0.2019	0	0.2101	0	0.1102	0	0.0719	0	1.2683	1
52	0.3057	0	0.0454	0	0.018	0	0.2656	1	-0.1783	0	-0.1539	0	0.0966	0	-0.1415	0	0.9573	0
54	-0.3838	0	0.0661	0	0.029	0	-0.0131	0	0.0762	0	0.0271	0	-0.2767	1	-0.074	0	0.968	0
55	-0.4351	0	0.049	0	0.036	0	-0.3046	1	0.064	0	0.2557	0	-0.0282	0	0.2536	0	0.8221	0
57	-0.6005	1	0.177	1	0.071	0	0.1811	0	0.0409	0	-0.501	1	0.296	1	0.0129	0	1.1404	0
58	-0.4176	0	0.1373	0	0.035	0	0.2349	0	-0.098	0	-0.3697	1	0.2467	0	-0.0929	0	1.1487	0

三、 異常點與影響點綜合結果

綜合上述異常點及影響點的偵測方法，得到表十六，得知觀察值 36(也就是“紐奧良”，位於路易斯安那州)為異常點兼影響點，我們回至原始資料檢查，發現紐奧良的死力，為所有城市中最大的，但觀察我們所考慮到的所有預測變數(X_1-X_{14})，未發現有與其它城市特別不同之處，因此，我們認為也許是一些較特殊的因素，我們沒有納入考量，像是紐奧良位於沿海一帶，就曾經因為發生海嘯，而導致重大傷亡與損失的事故。

此外，由表十六得知移除第一個觀察值 36 後，其判定係數R^2及修正後判定係數R_a^2皆大幅下降，也就是說各變數對死力的解釋能力反而降低，因此我們回到應變數對自變數的散佈圖作觀察，其中圖一裡的第一行的第二個圖是死力對一月份平均溫度的散佈圖，一般來說一月份平均溫度愈高，死力應該會愈低，所以我們假設它們之間有一個負相關存在，且從此圖中可看出左下方與右上方的兩個點會影響到它們負相關的關係，回到原始資料得知這兩個點分別為觀察值 33 和觀察值 36，再加上由表十六得知觀察值 33 在 DFBETAS 的偵測方法中，偵測為可能的影響點，所以我們決定同時將這兩個觀察值移除再做一次判定係數R^2及修正後判定係數R_a^2的檢查。從表十六得知同時移除觀察值 33 和觀察值 36 後，其判定係數R^2及修正後判定係數R_a^2皆比未移除任何觀察值前提高，因此，我們決定將這兩個觀察值同時移除，然後再做一次檢測。

在同時移除觀察值 33 和觀察值 36 後，由表十六得知無觀察值為異常點，而觀察值 31(也就是“邁阿密”，位於佛羅里達州)為影響點，我們再次回至原始資料檢查，在所有考慮到的預測變數(X_1-X_{14})及應變數(Y)中，它的一月份平均溫度為華氏 67 度，除了居 58 個城市之冠外，相較於其它城市，它有偏高的趨勢，因此，我們認為也許是這個因素使得整個模型改變。但由表十六得知移除觀察值 31 後，其判定係數R^2及修正後判定係數R_a^2皆比未移除觀察值 31 前降低，也就

是說各變數對死力的解釋能力降低了，所以我們決定不再移除任何觀察值(包含觀察值 31)。

表十六

Methods	Obs								
Residual(ei)	36								
Student Residual (ri)	36								
Rstudent (ti)	36								
DFFITS	31	36	57						
Hat Diag H	6	12	28	31	39	46	48	57	
Cook's D	-								
DFBEATS	31	33	36	57					
Cov Ratio	2	12	28	31	36	39	46	48	
Methods	Obs(同時移除第33、36筆)								
Residual(ei)	-								
Student Residual (ri)	-								
Rstudent (ti)	-								
DFFITS	55								
Hat Diag H	6	12	28	30	31	37	44	46	
Cook's D	-								
DFBEATS	2	9	31	44	53	55			
Cov Ratio	2	6	12	28	30	31	37	44	53
						R^2		R^2_a	
步驟一	未移除觀察值					69%		66%	
步驟二	移除觀察值36					30%		24%	
步驟三	同時移除觀察值36、3					73%		70%	
步驟四	移除觀察值31					72%		70%	

第三章 刪除觀察值後之迴歸分析

第一節　選取重要變數

刪除完異常點及影響點後，我們再次利用上述的七種選取變數的方法：向前選取法、向後選取法、逐步選取法、Adjusted R-Square 選取法、C_p 選取法、AIC 準則以及 SBC 準則，來選取重要變數。結果七種方法選取到的變數皆為 X_1（一月份平均溫度）、X_4（每年度降雨量）、X_7（非白種人比例）、X_{14}（二氧化硫潛在污染），所以我們決定不再移除任何變數，仍然將這四個變數留在模式中。其中使用 C_p 選取法所得到的 C_p 圖如圖十三所示。

第二節　最終模式

由表十七得知，上述選取到的四個變數的係數，它們的 p-value 皆小於 0.05，可拒絕虛無假設，所以它們都是顯著的，並得出刪除觀察值 33 和觀察值 36 後的**最終模式配適迴歸線**如下：

$$y_i = 899.8715 - 2.34495x_1 + 1.35348x_4 + 4.0768x_7 + 0.34096x_{14}$$

表十七

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > | t |	Variance Inflation
Intercept	Intercept	1	899.8715	23.98839	37.51	<.0001	0.0000
x1	一月份平均溫度	1	-2.34495	0.50663	-4.63	<.0001	1.2746
x4	每年度降雨量	1	1.35348	0.39863	3.4	0.0013	1.1457
x7	非白種人比例	1	4.0768	0.58675	6.95	<.0001	1.4243
x14	二氧化硫潛在污染	1	0.34096	0.07077	4.82	<.0001	1.1313

第三節　殘差四個基本假設的檢定

（一）檢查殘差平均是否為 0

要檢查殘差平均數是否為 0，首先我們必須設立虛無假設（$H_0:\mu=0$）與對立假設（$H_1:\mu\neq 0$），利用 Student's test、Sign test signed 與 Rank test 這三種檢定方法來判斷殘查平均數等於 0 是否顯著。

由表十八可看出三種檢定的p-value值皆明顯大於0.05，所以不能拒絕虛無假設，即殘差平均數為0，符合基本假設$E(\varepsilon_i)=0$。

（二）檢查殘差是否來自常態

要檢查殘差是否來自常態分配，首先我們必須設立虛無假設（$H_0:e_i \sim Normal$）與對立假設（$H_1:$反對H_0），利用 Shapiro-Wilk test、Kolmogorov-Smimov test、Cramer-von Mises 與 Anderson-Darling test 這四種檢定方法來判斷殘查來自常態分配是否顯著。

由表十八可看出 p-value 的值皆明顯大於 0.05，所以不能拒絕虛無假設，即殘差來自常態分配，符合基本假設$\varepsilon_i \sim Normal$。或由圖十四可看出，常態分配機率圖（P-P 圖）為一 45^0 之直線，所以殘差來自常態分配。

（三）檢查殘差間是否獨立

要檢查殘差間是否獨立，首先我們必須設立虛無假設（$H_0:\rho=0$）與對立假設（$H_1:\rho>0$），利用 Durbin-Watson D test 這種檢定方法來判斷殘差間獨立是否顯著。

由表十八可知 d=1.906，接近 2，所以我們可以說$\hat{\rho}$=0，即殘差間沒有相關，符合基本假設$Cov(\varepsilon_i,\varepsilon_j)=0 \ \ for\ all\ i\neq j$。

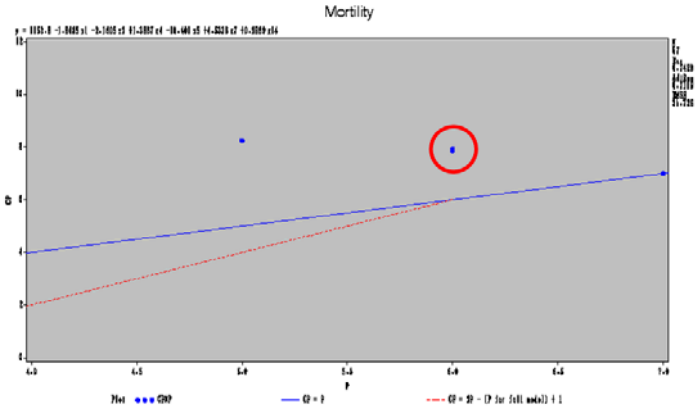
（四）檢查殘差變異數是否為常數

藉由觀察殘差散佈圖，以一月份平均溫度之殘差圖為例，如圖十五所示，可發現殘差落在以 0 為中心線的區塊內，呈現均勻散佈，沒有逐漸變大或逐漸變小的趨勢，所以我們可以說誤差的變異為一常數，也就是符合基本假設$Var(\varepsilon_i)=\sigma^2$。

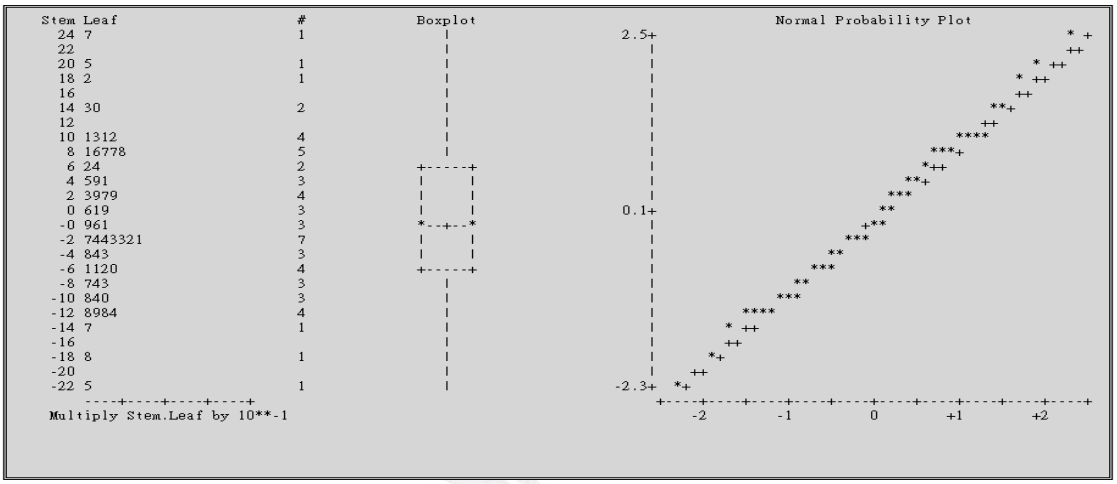
綜合上述，我們得到$\varepsilon_i \overset{iid}{\sim} Normal(0,\sigma^2)\quad i=1,2,.....,n$，亦即滿足誤差的四個基本假設。

表十八

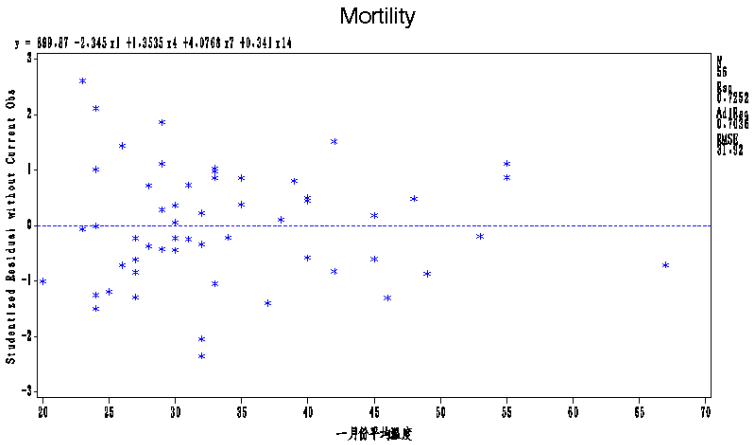
Tests for Location:Mu0=0				
Test	Statistic		p Value	
Student's t	t	-0.04889	Pr > |t|	0.96120
Sign	M	-2	Pr >= |M|	0.68890
Signed Rank	S	-13	Pr >= |S|	0.91670
Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.992862	Pr < W	0.9843
Kolmogorov-Smirnov	D	0.072811	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.024503	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.157073	Pr > A-Sq	>0.2500
Tests for Independence				
Durbin-Watson D			1.906	
Number of Observations			56	
1st Order Autocorrelation			0.034	


Mortility

圖十三


Stem Leaf #
24 7 1
22
20 5 1
18 2 1
16
14 30 2
12
10 1312 4
8 16778 5
6 24 2
4 591 3
2 3979 4
0 619 3
-0 961 3
-2 7443321 7
-4 843 3
-6 1120 4
-8 743 3
-10 840 3
-12 8984 4
-14 7 1
-16
-18 8 1
-20
-22 5 1
Multiply Stem.Leaf by 10**-1
Boxplot
Normal Probability Plot
2.5+
0.1+
-2.3+
-2
-1
0
+1
+2

圖十四


Mortility
Studentized Residual without Current Obs
一月份平均溫度

圖十五

第四章 結論

在資料分析中，我們使用了七種方法來選取變數，分別是「向前選取法」、「向後消去法」、「逐步選取法」、「校正後的複判定係數法」、「CP 選取法」、「AIC 準則」以及「SBC 準則」，經過這些選取變數的方法我們選取了 X_1(一月份平均溫度)、X_2(七月份平均溫度)、X_4(每年度降雨量)、X_5(受教育年數中位數)、X_7(非白種人比例)、 X_{14}(二氧化硫潛在污染)為最佳模型的變數。因而我們可以大概推估此六項變數與死力有相關性，而死力與 X_3(相對溼度)、X_6(人口密度)、X_8(白種人比例)、X_9(人口數)、X_{10}(每個家庭人數)、X_{11}(家庭收入中位數)、X_{12}(碳氫化合物潛在污染)、X_{13}(一氧化二氮潛在污染) 此八項變數與死力較無相關性，其配適迴歸線為

$$y_i = 1214.092 - 1.55877x_1 - 2.47975x_2 + 1.3872x_4 - 14.7357x_5 + 4.9419x_7 + 0.2517x_{14}$$

一月份平均溫度(X_1)每增加華氏一度，死力則會減少 1.55877。

七月份平均溫度(X_2)每增加華氏一度，死力則會減少 2.47975。

每年度降雨量(X_4)每增加一厘米，死力則會增加 1.3872。

受教育年數中位數(X_5)每增加一年，死力則會每減少 14.7357。

非白種人比例(X_7)每增加一個百分比，死力則會增加 4.9419。

二氧化硫潛在污染(X_{14})每增加一個 ppm，死力則會增加 0.2517。

但在第一次個別係數的檢定中發現，x2 的係數不顯著，因此先將 x2 這個變數刪除，再做一次個別係數檢定。在第二次個別係數的檢定中發現，x5 的係數不顯著，因此再將 x5 這個變數刪除，又再做一次個別係數檢定。

刪除 x2、x5 這些個別係數不顯著的變數後，再做一次檢定，發現剩下變數的係數皆顯著，**最終模式的配適迴歸線**如下：

$$y_i = 869.4615 - 1.8086x_1 + 1.62043x_4 + 4.42679x_7 + 0.3238x_{14}$$

一月份平均溫度(X_1)每增加華氏一度，死力則會減少 1.8086。

每年度降雨量(X_4)每增加一厘米，死力則會增加 1.62043。

非白種人比例(X_7)每增加一個百分比，死力則會增加 4.42679。

二氧化硫潛在污染(X_{14})每增加一個 ppm，死力則會增加 0.3238。

此外我們偵測異常點和影響點以免有異常點或影響點讓整個分析有誤，第一次檢查的結果發現觀察值 36 為一個異常點兼影響點，且由表十八得知移除第一個觀察值 36 後，其判定係數R^2及修正後判定係數R_a^2皆大幅下降，也就是說各變數對死力的解釋能力反而降低，因此我們回到應變數對自變數的散佈圖作觀察，其中圖一裡的第一行的第二個圖是死力對一月份平均溫度的散佈圖，一般來說一月份平均溫度愈高，死力應該會愈低，所以我們假設它們之間有一個負相關存在，且從此圖中可看出左下方與右上方的兩個點會影響到它們負相關的關係，回到原始資料得知這兩個點分別為觀察值 33 和觀察值 36，再加上由表十八得知觀察值 33 在 DFBETAS 的偵測方法中，偵測為可能的影響點，所以我們決定同時將這兩個觀察值移除再做一次判定係數R^2及修正後判定係數R_a^2的檢查。從表十八得知同時移除觀察值 33 和觀察值 36 後，其判定係數R^2及修正後判定係數R_a^2皆有比未移除任何觀察值前提高，因此，我們決定將這兩個觀察值同時移除，然後再做一次檢測。第二次檢查的結果，發現無觀察值為異常點，而發現觀察值 31 為一個影響點，移除觀察值 31 後，其判定係數R^2及修正後判定係數R_a^2皆比未移除觀察值 31 前降低，也就是說各變數對死力的解釋能力降低了，所以我們決定不再移除任何觀察值(包含觀察值 31)。

同時將觀測值 33、36 移除後，再作一次迴歸分析，選取的重要變數為X_1(一月份平均溫度)、X_4(每年度降雨量)、X_7(非白種人比例)、 X_{14}(二氧化硫潛在污染)這四個變數，並將它們留在模式中。其配適迴歸線為

$$y_i = 899.8715 - 2.34495x_1 + 1.35348x_4 + 4.0768x_7 + 0.34096x_{14}$$

由此發現每年度降雨量、非白種人比例、二氧化硫潛在污染與死力呈現正相關，只有一月份平均溫度與死力呈現負相關。

一月份平均溫度(X_1)每增加華氏一度，死力則會減少 2.34495。

每年度降雨量(X_4)每增加一厘米，死力則會增加 1.35348。

美國的氣候大部分地區屬溫帶和亞熱帶氣候，僅佛羅里達半島南端屬熱帶。阿拉斯加州位於北緯 60 至 70 度之間，屬北極圈內的寒冷氣候區；夏威夷州位於北回歸線以南，屬熱帶氣候區。但由於美國幅員遼闊，地形複雜，各地氣候差異較大，大體可分為五個氣候區。

氣候類型	特徵	1 月份平均溫度	7 月份平均溫度	年平均降雨量
東北部沿海的溫帶氣候區	1. 受拉布拉多寒流和北方冷空氣的影響，冬季寒冷。 2. 夏季溫和多雨	1 月份平均溫度為 -6℃左右	7 月份平均溫度為 16℃左右	年平均降雨量為 1000 公釐左右
東南部亞熱帶氣候區	受墨西哥灣暖流的影響，氣候溫暖濕潤	1 月份平均溫度為 16℃	7 月份平均溫度為 24～27℃	年平均降雨量為 1500 公釐
中央平原的大陸性氣候區呈大陸性氣候特徵	1.冬季寒冷 2.夏季炎熱	1 月份平均溫度為 -14℃左右，	7 月份平均氣溫高達 27～32℃	年平均降雨量為 1000～1500 公釐
西部高原乾燥氣候區為內陸性氣候	高原上年溫差較大	科羅拉多高原的年溫差高達 25℃		1. 年平均降雨量在 500 公釐以下 2. 高原荒漠地帶降雨量不到 250 公釐
太平洋沿岸的海洋性氣候區	1.冬暖夏涼 2.雨量充沛	1 月份平均氣溫在 4℃以上	7 月份平均氣溫在 20～22℃左右	年平均降雨量為 1500 公釐左右

美國的五個氣候區，它們的 1 月份平均溫度，大部份都過低，這會使得一些生活較貧困的美國人民凍死，因此，若增加華氏一度，死力相對地就會減少。

美國的氣候因所處為熱帶沙漠地區(西南部)或北極大陸地區而異。大都會地區受西部低氣壓的影響，多雲雨天，天氣變化無常。西北太平洋沿岸降雨量最大，西南部則較小，但過多的雨量常會提高死力。因此可能會造成每年度降雨量增加一厘米，死力反而會增加的趨勢。

非白種人比例(X_7)每增加一個百分比，死力則會增加 4.0768，這是因為非白種人裡包含許多的黑人，而美國黑人文盲占大多數，過去的黑人許多都沒受過教育，法律的規範難以控制他們的犯罪及暴力行為，加上有許多黑人因為民族意識抬頭卻扭曲其意義者，會向美國白人報復，因此，黑人的死亡率也就較高。

二氧化硫潛在污染(X_{14})每增加一個 ppm，死力則會增加 0.34096，這是因為二氧化硫和二氧化氮的射出物會引起呼吸方面的問題，例如哮喘、乾咳、頭痛、和眼睛、鼻子、喉嚨的過敏，這也是為什麼二氧化硫每增加一個 ppm，死力則會增加的原因。而二氧化硫的主要來源是汽車排放廢氣、燃燒礦物燃料的發電廠及工業渦爐。

附錄

附錄一、原始資料

觀察值	美國城市	變數名稱														
		死力(Y)	一月份平均溫度(X1)	七月份平均溫度(X2)	相對溼度(X3)	每年度降雨量(X4)	受教育年數中位數(X5)	人口密度(X6)	非白種人比例(X7)	白種人比例(X8)	人口數(X9)	每個家庭人數	家庭收入中位數	碳氫化合物潛在污染	一氧化二氮潛在污染	二氧化硫潛在污
			華氏溫標	華氏溫標	百分比	厘米	年數	人口/每平方公里	百分比	百分比	人	人口數/家庭數	美元	ppm	ppm	ppm
	city	Mortality	JanTemp	JulyTemp	RelHum	Rain	Education	PopDensity	%NonWhite	%WC	pop	pop/house	income	HCPot	NOxPot	SO2Pot
1	Akron, OH (亞克朗市，俄亥俄州)	921.87	27	71	59	36	11.4	3243	8.8	42.6	660328	3.34	29560	21	15	59
2	Albany-Schenectady-Troy, NY (紐約州)	997.87	23	72	57	35	11	4281	3.5	50.7	835880	3.14	31458	8	10	39
3	Allentown, Bethlehem, PA-NJ 艾倫鎮，伯利恆(美國賓州)	962.35	29	74	54	44	9.8	4260	0.8	39.4	635481	3.21	31856	6	6	33
4	Atlanta, GA (亞特蘭大，喬治亞州)	982.29	45	79	56	47	11.1	3125	27.1	50.2	2138231	3.41	32452	18	8	24
5	Baltimore, MD (巴爾地摩，馬里蘭州)	1071.29	35	77	55	43	9.6	6441	24.4	43.7	2199531	3.44	32368	43	38	206
6	Birmingham, AL (伯明罕市，美國阿拉巴馬州)	1030.38	45	80	54	53	10.2	3325	38.5	43.1	883946	3.45	27835	30	32	72
7	Boston, MA (波士頓，麻薩諸塞州)	934.7	30	74	56	43	12.1	4679	3.5	49.2	2805911	3.23	36644	21	32	62
8	Bridgeport-Milford, CT (康乃狄克州)	899.53	30	73	56	45	10.6	2140	5.3	40.4	438557	3.29	47258	6	4	4
9	Buffalo, NY (水牛城(又稱布法羅)，位於紐約州)	1001.9	24	70	61	36	10.5	6582	8.1	42.5	1015472	3.31	31248	18	12	37
10	Canton, OH (俄亥俄州)	912.35	27	72	59	36	10.7	4213	6.7	41	404421	3.36	29089	12	7	20
11	Chattanooga, TN-GA (查特怒加市，田納西州)	1017.61	42	79	56	52	9.6	2302	22.2	41.3	426540	3.39	25782	18	8	27
12	Chicago, IL (芝加哥，伊利諾州)	1024.89	26	76	58	33	10.9	6122	16.3	44.9	606387	3.2	36593	88	63	278
13	Cincinnati, OH-KY-IN (辛辛那提市，美國俄亥俄州)	970.47	34	77	57	40	10.2	4101	13	45.7	1401491	3.21	31427	26	26	146
14	Cleveland, OH (克里夫蘭，俄亥俄州第一大城)	985.95	28	71	60	35	11.1	3042	14.7	44.6	1898825	3.29	35720	31	21	64
15	Columbus, OH (哥倫布市，俄亥俄州首府)	958.84	31	75	58	37	11.9	4259	13.1	49.6	124833	3.26	29761	23	9	15
16	Dallas, TX (達拉斯，德克薩斯州)	860.1	46	85	54	35	11.8	1441	14.8	51.2	1957378	3.22	38769	1	1	1
17	Dayton-Springfield, OH (德通市，俄亥俄州)	936.23	30	75	58	36	11.4	4029	12.4	44	942083	3.35	30232	6	4	16
18	Denver, CO (丹佛，科羅拉多州)	871.77	30	73	38	15	12.2	4824	4.7	53.1	1428836	3.15	39099	17	8	28
19	Detroit, MI (底特律，密西根州)	959.22	27	74	59	31	10.8	4834	15.8	43.5	4488072	3.44	33858	52	35	124
20	Flint, MI (福林特，密西根州)	941.18	24	72	61	30	10.8	3694	13.1	33.8	450449	3.53	32000	11	4	11
21	Grand Rapids, MI (激流市，密西根州)	871.34	24	72	61	31	10.9	3226	5.1	45.2	601680	3.37	29915	5	3	10
22	Greensboro-Winston-Salem-High Point, NC (北卡羅萊納	971.12	40	77	53	42	10.4	2269	22.7	41.4	851851	3.45	29450	8	3	5
23	Hartford, CT (哈特福特，康乃狄克州之首府)	887.47	27	72	56	43	11.5	2909	7.2	51.6	715923	3.25	37565	7	3	10
24	Houston, TX (休士頓，德克薩斯州)	952.53	55	84	59	46	11.4	2647	21	46.9	2735766	3.35	39558	6	5	1
25	Indianapolis, IN (印第安納波里，印第安納州首府)	968.67	29	75	60	39	11.4	4412	15.6	46.6	1166575	3.23	31461	13	7	33
26	Kansas City, MO (堪薩斯城，密蘇里州西部)	919.73	31	81	55	35	12	3262	12.6	48.6	914427	3.1	30783	7	4	4
27	Lancaster, PA (賓汐法尼亞)	844.05	32	74	54	43	9.5	3214	2.9	43.7	362346	3.38	30248	11	7	32
28	Los Angeles, Long Beach, CA (洛杉磯，加尼福尼亞州)	861.26	53	68	47	11	12.1	4700	7.8	48.9	7477503	2.66	36624	648	319	130
29	Louisville, KY-IN (路易維耳市，肯塔基-印地安那州)	989.26	35	71	57	30	9.9	4474	13.1	42.6	956756	3.37	29621	38	37	193
30	Memphis, TN-AR-MS (孟菲斯市，田納西州)	1006.49	42	82	59	50	10.4	3497	36.7	43.3	913472	3.49	27910	15	18	34
31	Miami-Hialeah, FL (邁阿密，佛羅里達州)	861.44	67	82	60	60	11.5	4657	13.5	47.3	1625781	2.65	32808	3	1	1
32	Milwaukee, WI (密爾瓦基市，威斯康辛州)	929.15	20	69	64	30	11.1	2934	5.8	44	1397143	3.26	35272	33	23	125
33	Minneapolis-St. Paul, MN-WI (明尼蘇達-威斯康辛)	857.62	12	73	58	25	12.1	2095	2	51.9	2137133	3.28	35871	20	11	26
34	Nashville, TN (那什維爾，田納西州首府)	961.01	40	80	56	45	10.1	2682	21	46.1	850505	3.32	28641	17	14	78
35	New Haven-Meriden, CT (康乃狄克州)	923.23	30	72	58	46	11.3	3327	8.8	45.3	500474	3.16	34364	4	3	8
36	New Orleans, LA (紐奧良，路易斯安那州)	1113.16	54	81	62	54	9.7	3172	31.4	45.5	1256256	3.36	32704	20	17	1
37	New York, NY (紐約)	994.65	33	77	58	42	10.7	7462	11.3	48.7	8274961	3.03	36047	41	26	108
38	Philadelphia, PA-NJ (費城，賓汐法尼亞-紐澤西)	1015.02	32	76	54	42	10.5	6092	17.5	45.3	4716818	3.32	33449	29	32	161
39	Pittsburgh, PA (匹茲堡，賓汐法尼亞州)	991.29	29	72	56	36	10.6	3437	8.1	45.5	2218870	3.32	32934	45	59	263
40	Portland, OR (波特蘭，奧勤岡州)	893.99	38	67	73	37	12	3387	3.6	50.3	1105699	2.66	33020	56	21	44
41	Providence, RI (羅德島)	938.5	29	72	56	42	10.1	3508	2.2	38.8	618514	3.16	30094	6	4	18
42	Reading, PA (賓汐法尼亞州)	946.19	33	77	54	41	9.6	4843	2.7	38.6	312509	3.08	32449	11	11	89
43	Richmond-Petersburg, VA (維吉尼亞州)	1025.5	39	78	53	44	11	3768	28.6	49.5	761311	3.32	33510	12	9	48
44	Rochester, NY (羅契斯特市，紐約州)	874.28	25	72	60	32	11.1	4355	5	46.4	971230	3.21	34896	7	4	18
45	St. Louis, MO-IL (聖路易，密蘇里州東部城市)	953.56	32	79	57	34	9.7	5160	17.2	45.1	1808621	3.23	34546	31	15	68
46	San Diego, CA (聖地牙哥，加尼福尼亞州)	839.71	55	70	61	10	12.1	3033	5.9	51	1861846	3.11	32586	144	66	20
47	San Francisco, CA (舊金山，加尼福尼亞州)	911.7	48	63	71	18	12.2	4253	13.7	51.2	1488871	2.92	47966	311	171	86
48	San Jose, CA (聖荷西(加州西部一城市)，加尼福尼亞州)	790.73	49	68	71	13	12.2	2702	3	51.9	1295071	3.36	41994	105	32	3
49	Seattle, WA (西雅圖，華盛頓州)	899.26	40	64	72	35	12.2	3626	5.7	54.3	1607469	3.02	37069	20	7	20
50	Springfield, MA (春田(伊利諾州首府)，麻薩諸塞州)	904.16	28	74	56	45	11.1	1883	3.4	41.9	515259	3.21	29327	5	1	20
51	Syracuse, NY (西拉鳩斯市，紐約州)	950.67	24	72	61	38	11.4	4923	3.8	50.5	642971	3.34	30114	8	5	25
52	Toledo, OH (托利多，俄亥俄州)	972.46	26	73	59	31	10.7	3249	9.5	43.9	616864	3.22	30497	11	7	25
53	Utica-Rome, NY (紐約)	912.2	23	71	60	40	10.3	1671	2.5	47.4	320180	3.28	27305	5	2	11
54	Washington, DC-MD-VA (華盛頓-馬里蘭-維吉尼亞)	967.8	37	78	52	42	12.3	5308	25.9	59.7	3250822	3.25	41888	65	28	102
55	Wichita, KS (堪薩斯州)	823.76	32	81	54	28	12.1	3665	7.5	51.6	411313	3.27	34812	4	2	1
56	Wilmington, DE-NJ-MD (德拉瓦-紐澤西-馬里蘭州)	1003.5	33	76	56	65	11.3	3152	12.1	47.3	523221	3.39	33927	14	11	42
57	Worcester, MA (麻薩諸塞州)	895.7	24	70	56	65	11.1	3678	1	44.8	402918	3.25	29374	7	3	8
58	York, PA (約克郡，賓汐法尼亞州)	911.82	33	76	54	62	9	9699	4.8	62.2	381255	3.22	28985	8	8	49

參考文獻

1. Kutner, M.H., Nachtsheim, C.J., and Neter, J. (2004)"*Applied Linear Regression Models*" , 4th ed., McGraw Hill.

2. 彭昭英、唐麗英(2003)，"SAS 1-2-3"，第四版，儒林圖書公司出版。

3. 資料來源：*U.S. Department of Labor-Bureau of Labor Statistics*

http://www.bls.gov/bls/proghome.htm