

利用關鍵變數預測公共自行車共享系統的租借總數量

一、研究背景與動機：

各個國家為了改善都市道路交通擁擠、環境污染及能源損耗等問題，讓民眾有更好的生活，都設有公共自行車租賃系統，推廣民眾騎乘自行車作為短程接駁交通工具，響應全球節能減碳風潮。目前世界各地共有超過 500 項的自行車共享計劃，例如台北都會區的 youbike 自行車共享系統。

在自行車共享租賃的過程中從加入會員、租賃和退貨的整個過程都採自動化方式，通過系統，用戶能夠輕鬆地從特定的位置租借一輛自行車並於另一個位置還車。但實際上，在日常生活中可以觀察到，每個不同的租借點租用的數量差異很大，有時租借點車位滿載，有時租借點無車可借，雖然大多時候政府已努力提供相關的資訊供民眾查詢租賃站狀況，但受到假日、天氣、季節等等因素的影響，租賃站使用效率仍有進步的空間，是個可以深入研究的項目。

二、研究目的與資料來源：

不同的天氣、季節或是否假日，自行車租賃系統被乘客租用的數量差異很大，這可能帶來困擾，例如：營運人員無法提供足夠數量的自行車或是不知道何時可進行維修作業，若能預知租用數量則自行車需要維修時，可以選擇在租用數量較少的時間；租用數量大時，可以提供更多的數量，增加營業額。

還有很多行業也適合使用天氣、季節或假日與否來預測銷售量或備貨量，例如：超商可以根據天氣來預測關東煮、雨衣等的銷售數量，餐廳也可以根據季節或假日與否來預測每天須準備的備貨量，減少因準備太多數量導致浪費食材。

本次研究想透過觀察天氣、季節或假日與否等因素進行自行車租賃系統的每小時租用數量，在著名的數據建模和數據分析競賽平台 kaggle 上也曾透過加州大學爾灣分校機器學習資料庫(UCI Machine Learning Repository)的授權取得資料集舉辦相關競賽，本次研究一樣使用 UCI 資料庫的 Bike Sharing Dataset 來進行探討。

該資料集是美國華盛頓特區的自行車資源共享系統 Capital Bikeshare 的真實資料集，Capital Bikeshare 從 2010 年 9 月起開辦，和台北都會區的 youbike 公共自行車租賃系統一樣，在市區與各大



景點均有密集設立的取還車站，採會員制，每次借車的前三十分鐘免費。

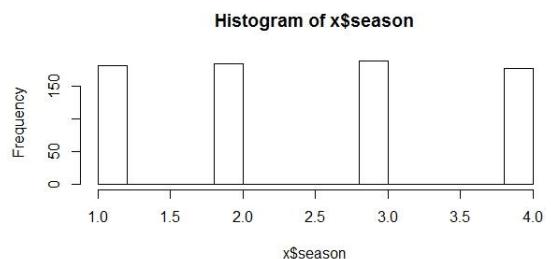
資料集包含 2011 年至 2012 年在 Capital bikeshare 系統中租賃自行車的每時數量，來自公開在官網 <http://capitalbikeshare.com/system-data> 的資料，相應的天氣訊息，摘自 <http://www.freemeteo.com>。另外工作日或假日的定義是來自 <https://dchr.dc.gov/page/holiday-schedules>。

三、變數介紹：

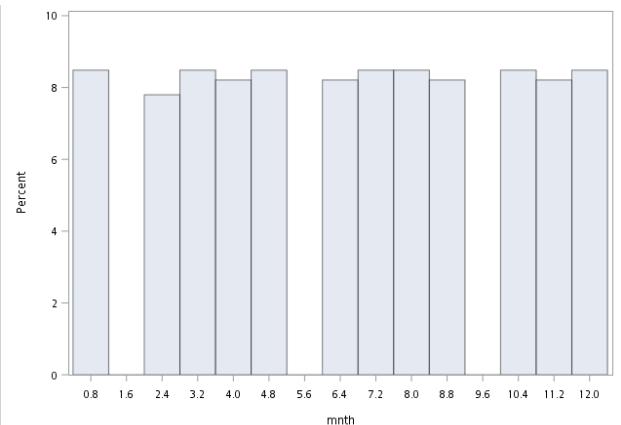
原始資料集共有 16 個變數，731 個觀察值，變數分別介紹如下：

- instant：編號
- dteday：日期
- season：季節 (1:春天, 2:夏天, 3:秋天, 4:冬天)
- yr：年份 (0:2011 年, 1:2012 年)
- mnth：月份 (1-12)
- holiday：是否為國定假日 (0:否, 1:是)
- weekday：星期幾
- workingday：是否為工作日 (0:周末或國定假日, 1:工作日，非周末且非假日)
- weathersit：天氣狀況
 - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
 - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
 - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- temp：標準化攝氏溫度，將溫度除以 41 進行標準化
- atemp: 標準化體感攝氏溫度，將溫度除以 50 進行標準化
- hum: 標準化濕度，將濕度除以 100 進行標準化
- windspeed: 標準化風速，將風速除以 67 進行標準化
- cnt: 此時段租借總數量(已註冊使用者+臨時性使用者)
- casual: 臨時使用者於此時段租借的數目
- registered: 已註冊會員於此時段租借的數目
 - cnt 欄位為本次研究中的反應變數。
 - instant(編號)、dtedat(日期)與欲預測的結果無關，忽略這兩個欄位。另外，由於希望預測未來的年份，所以忽略 yr 年份(0:2011 1:2012)此欄位。causal 與 registered 欄位因為 causal 加 registered 等於 cnt 所以兩個欄位已經隱含了反應變數欄位的資訊，所以忽略這兩個欄位。
 - 因此本次共使用 10 個解釋變數來預測反應變數 cnt。
 - 對所有的解釋變數作敘述統計和繪圖如下：

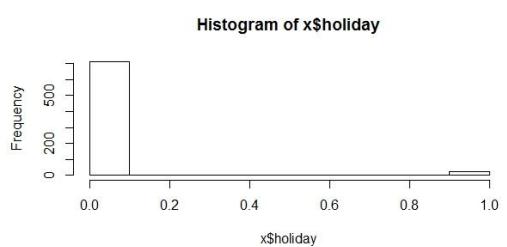
season	次數	百分比	累計 次數	累計 百分比
1	181	24.76	181	24.76
2	184	25.17	365	49.93
3	188	25.72	553	75.65
4	178	24.35	731	100.00



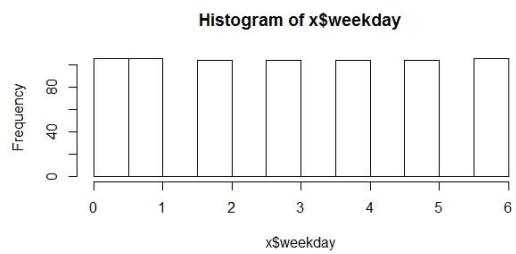
mnth	次數	百分比	累計 次數	累計 百分比
1	62	8.48	62	8.48
2	57	7.80	119	16.28
3	62	8.48	181	24.76
4	60	8.21	241	32.97
5	62	8.48	303	41.45
6	60	8.21	363	49.66
7	62	8.48	425	58.14
8	62	8.48	487	66.62
9	60	8.21	547	74.83
10	62	8.48	609	83.31
11	60	8.21	669	91.52
12	62	8.48	731	100.00



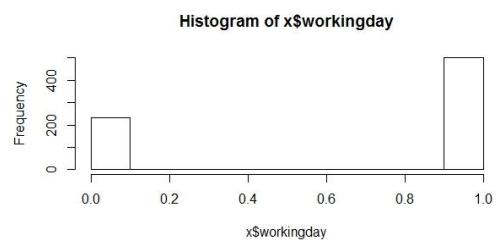
holiday	次數	百分比	累計 次數	累計 百分比
0	710	97.13	710	97.13
1	21	2.87	731	100.00



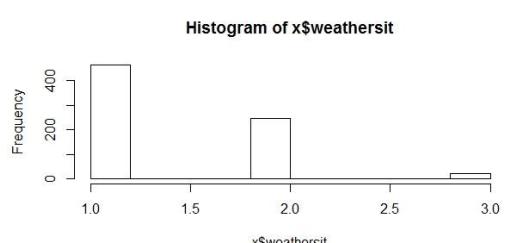
weekday	次數	百分比	累計 次數	累計 百分比
0	105	14.36	105	14.36
1	105	14.36	210	28.73
2	104	14.23	314	42.95
3	104	14.23	418	57.18
4	104	14.23	522	71.41
5	104	14.23	626	85.64
6	105	14.36	731	100.00



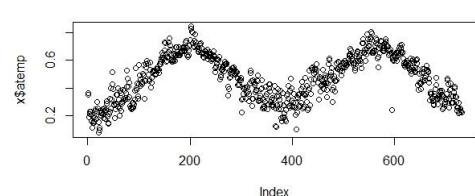
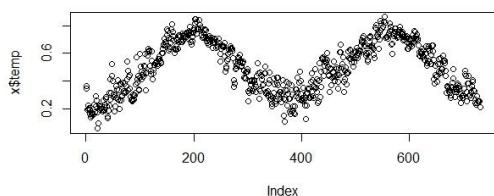
workingday	次數	百分比	累計 次數	累計 百分比
0	231	31.60	231	31.60
1	500	68.40	731	100.00

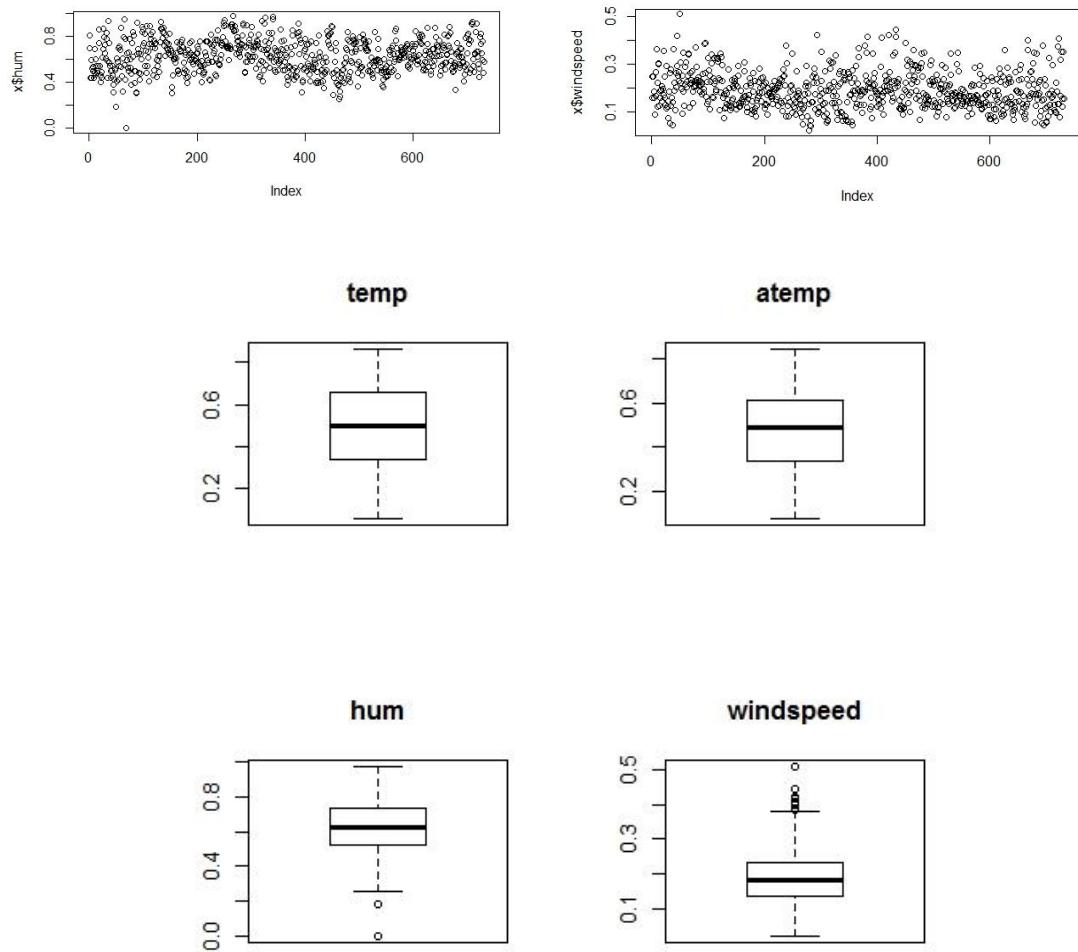


weathersit	次數	百分比	累計 次數	累計 百分比
1	463	63.34	463	63.34
2	247	33.79	710	97.13
3	21	2.87	731	100.00



變數	N	平均值	標準差	最小值	最大值
temp	731	0.4953848	0.1830510	0.0591304	0.8616670
atemp	731	0.4743540	0.1629612	0.0790696	0.8408960
hum	731	0.6278941	0.1424291	0	0.9725000
windspeed	731	0.1904862	0.0774979	0.0223917	0.5074630





- 各個解釋變數對反應變數的相關係數：

解釋變數	相關係數
season	0.40610037
mnth	0.27997711
holiday	-0.06834772
weekday	0.06744341
workingday	0.06115606
weathersit	-0.29739124
temp	0.62749401
atemp	0.63106570
hum	-0.10065856
Windspeed	-0.23454500

四、用簡單迴歸分析探討溫度和用戶租借自行車數量之間的關係：

欲利用迴歸分析探討溫度和用戶租借自行車數量之間的關係，得到迴歸係數之最小平方估計值分別為 $b_0=1214.6$ ， $b_1=6640.7$ ，因此所估計出的迴歸函數為：
 $\hat{y}=1214.6+6640.7x$

解釋變數 x 表示標準化後的攝氏溫度，反應變數 y 表示自行車租借總數量， \hat{y} 表示預測變數 x 在該水準下之估計迴歸函數值， x 和 y 兩者間存在有正比例性質的迴歸關係。

此迴歸直線的斜率=6640.7，又因為此處是經標準化的解釋變數 x 攝氏溫度，表示每增加一度攝氏溫度，所增加的自行車租借總數量其機率分配的期望值為 $6640.7/41$ =約 162 輛，而截距=1214.6，因模型範圍不包括 $x=0$ ，截距本身不具備任何特殊意義。結果之適用範圍約為標準化攝氏溫度介於 0.05913 至 0.86167 之間(即原始攝氏溫度介於 2.42 至 35.33 度之間)。

如果要估計解釋變數 x 在該水準下之平均反應，將 x 值代入所估計出的迴歸函數，例如我們關心當攝氏溫度 20.5 度下之用戶租借自行車數量，則將攝氏溫度 20.5 度轉換為標準化攝氏溫度 $x=0.5$ 其點估計為: $\hat{y}=1214.6+6640.7x=4534.997$ ，換句話說，若固定攝氏溫度 20.5 度去看多次的用戶租借自行車數量，平均用戶租借自行車數量為 4534.997，由於多少會有一些變異性，所以模型中會透過誤差項來詮釋。

誤差平方和 SSE 其自由度 $n-2$ ，誤差均方 $MSE=SSE/(n-2)$ ，標準差的估計量為 MSE 之平方根，此迴歸函數的 $SSE=1660846807$ ，因為 SSE 之自由度為 729，所以 $MSE=s^2=1660846807/729=2278254$ ，則對任意 x 計算 y 之機率分配其標準差之點估計值 $s=1509.38845$ 。

b_1 的抽樣分配是指固定解釋變數 x 的水準，重複抽樣得到不同樣本造成不同 b_1 之值的現象， b_1 的期望值 $E\{b_1\}$ =斜率參數 β_1 ，由於 b_1 是常態分配，其標準化統計量 $b_1-\beta_1/s\{b_1\}$ 服從自由度 $n-2$ 的 t 分配，可用來算出斜率參數 β_1 ，假設想以 95% 的信賴係數來估計 β_1 ，首先先得到 $s\{b_1\}=305.19$ ，又 $t(.975; 729)=1.96$ ，則 β_1 的 95% 的信賴區間 6043.54 至 7238.88 之間，故在 95% 的信賴係數下，我們估計攝氏溫度每增加一度，用戶租借自行車的增加量介於 147.40 到 176.56 輛之間。

藉由迴歸模型來檢定是否溫度和用戶租借自行車數量之間具有線性關聯，即是否 $\beta_1=0$ ，於是假設： $H_0: \beta_1=0$ vs $H_a: \beta_1 \neq 0$

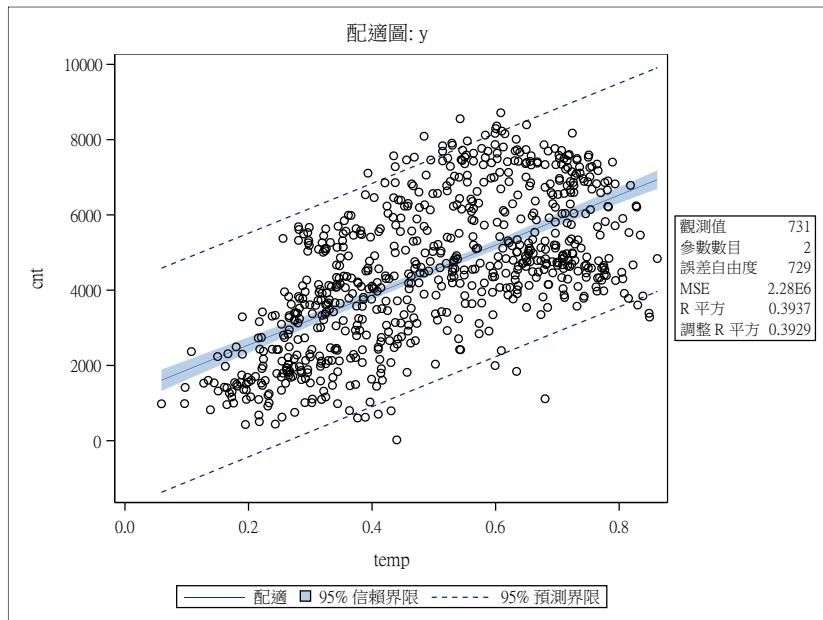
型一錯誤控制在 $\alpha=0.05$ ，由於之前 β_1 的 95% 信賴區間沒有包含 0，可以得知拒絕 H_0 ，或利用統計量 $b_1-\beta_1/s\{b_1\}$ ，這個檢定統計量在顯著水準控制在 α

下的決策法則是：若絕對值的 $t \leq t(.975; 729)$ ，則不拒絕 H_0 ；若絕對值的 $t > t(.975; 729)$ ，則拒絕 H_0

此例 $t=21.76$ ， $p\text{-value} < .0001$ ，所以拒絕 H_0 ，也就是 $\beta_1 \neq 0$ ，氣溫和用戶租借自行車數量之間有線性關聯。

若改利用 F 檢定對 β_1 檢定，待檢定的假說也是 $H_0: \beta_1 = 0$ vs $H_a: \beta_1 \neq 0$ ， $\alpha=0.05$ ， $F(0.95, 1, 729)=3.84$ ，決策規則為若 $F^*=\text{MSR}/\text{MSE} \leq 3.84$ ，則不拒絕 H_0 ，若 $F^* > 3.84$ ，則拒絕 H_0 ，計算得出 $F^*=\text{MSR}/\text{MSE}=473.47$ ， $p\text{-value} < .0001$ ，拒絕 H_0 ，即 $\beta_1 \neq 0$ ，和上述結論相同。

由於 $\beta_0 = 0$ 不涵蓋 $x=0$ ， β_0 沒意義，所以不對 β_0 做推論。



迴歸模型，可看到 $E\{y\}$ 的 95% 信賴區間和 95% 的信賴帶

接下來探討當 $x=x_h$ 時的平均反應 y 記為 $E\{y_h\}$ ，假設要找 $x=0.5$ 時 \hat{y}_h 的 95% 信賴區間，先利用之前的點估計值 $\hat{y}=1214.6+6640.7x=4534.997$ ， $s=55.84$ ， $t(.975; 729)=1.96$ ，95%的信賴區間在 4425.36 至 4644.63 之間，則在信賴係數 0.95 下，我們獲得的結論是，當攝氏溫度為 20.5 度時，用戶租借自行車數量在 4425.36 至 4644.63 輛之間的某處。由上圖中也可以看到整個迴歸線 $E\{y\}$ 的信賴區間。

	temp	cnt	下列項目的預測值 y	平均值預測值的標準誤差	平均值的 95% C.I. 下限	平均值的 95% C.I. 上限
367	0.5	5169	4534.9971183	55.844514126	4425.3618588	4644.6323778

新 y_h 的預測區間(同樣在攝氏溫度為 20.5 度時)，由 $\hat{y}_h=4534.997$ ， $s^2\{\hat{y}_h\}=23393.087$ ， $MSE=2278254$ ， $s\{\text{pred}\}=1510.42$ ，故得 95% 預測區間 $4534.997 \pm (1510.42)(1.96)=(1569.70, 7500.29)$

在信賴係數 0.95 下，我們當溫度為攝氏 20.5 度時，新的預測用戶租借自行車數量會落在 1569.70 至 7500.29 之間的某處，可以看到個別結果會偏離平均反

應且當 x_h 離 x 的平均數更遠時，預測區間會更寬，使得估計的 \hat{y}_h 會變得較不精確。

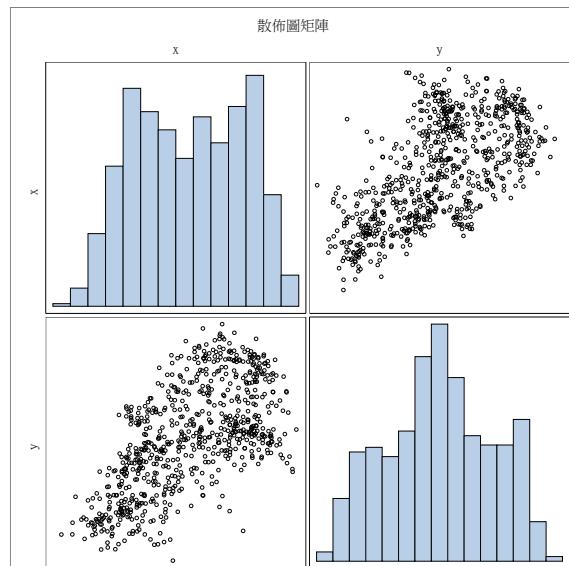
	temp	cnt	下列項目的預測值 y	95% C.I. 下限 (個別預測值)	95% C.I. 上限 (個別預測值)	個別預測的標準誤差
367	0.5	5169	4534.9971183	1569.7028591	7500.2913775	1510.4211721

判定係數 $R^2=SSR/SSTO=0.3937$ 表示使用解釋變數 x 削減總變異的比值，表示考量了氣溫後，用戶租借自行車數量的變異削減了 39.37%。

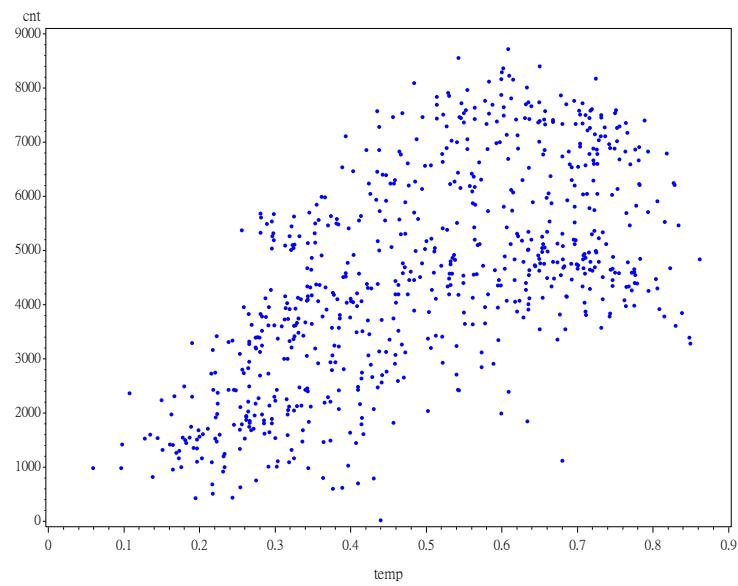
相關係數 $r=0.627$ ，兩變量屬中度相關。

Pearson 相關係數, N = 731		
Prob > r (位於 H0 底下): Rho=0		
	x	y
x	1.00000	0.62749
temp		<.0001
y	0.62749	1.00000
cnt	<.0001	

簡單統計值							
變數	N	平均值	標準差	中位數	最小值	最大值	標籤
x	731	0.49538	0.18305	0.49833	0.05913	0.86167	temp
y	731	4504	1937	4548	22.00000	8714	cnt



解釋變數 x 與反應變數 y 之散佈圖矩陣



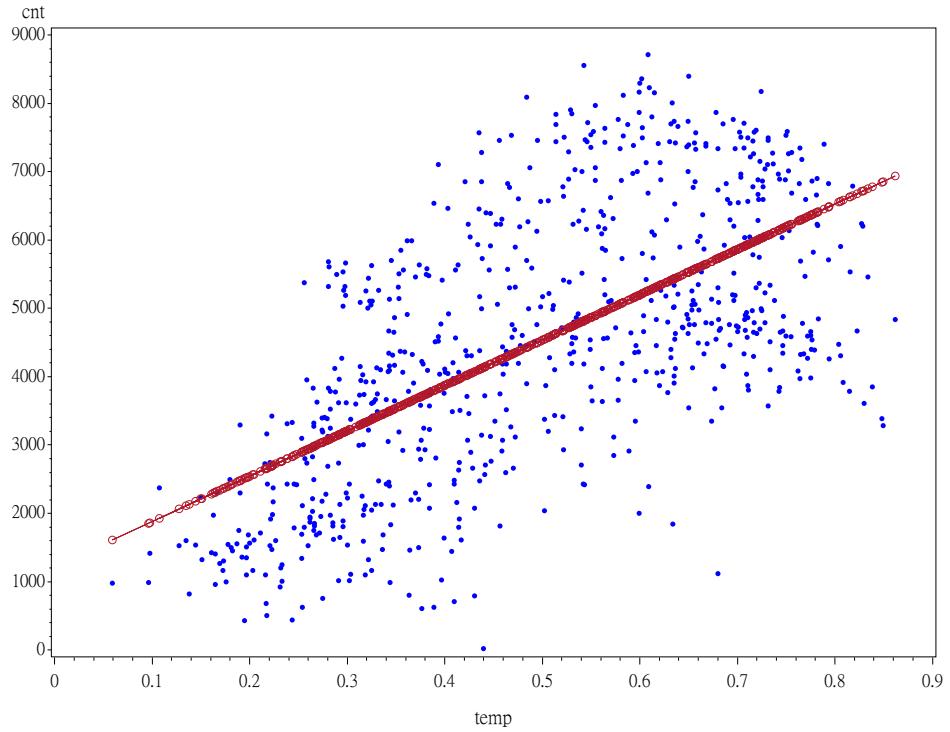
解釋變數 x 與反應變數 y 之散佈圖

變異數分析					
來源	自由度	平方和	平均值 平方	F 值	Pr > F
模型	1	1078688585	1078688585	473.47	<.0001
誤差	729	1660846807	2278254		
已校正的總計	730	2739535392			

根 MSE	1509.38845	R 平方	0.3937
應變平均值	4504.34884	調整 R 平方	0.3929
變異係數	33.50958		

參數估計值						
變數	標籤	自由度	參數 估計值	標準 誤差	t 值	Pr > t
Intercept	Intercept	1	1214.64212	161.16353	7.54	<.0001
x	temp	1	6640.71000	305.18803	21.76	<.0001

Predicted & Observed Count vs Temperature

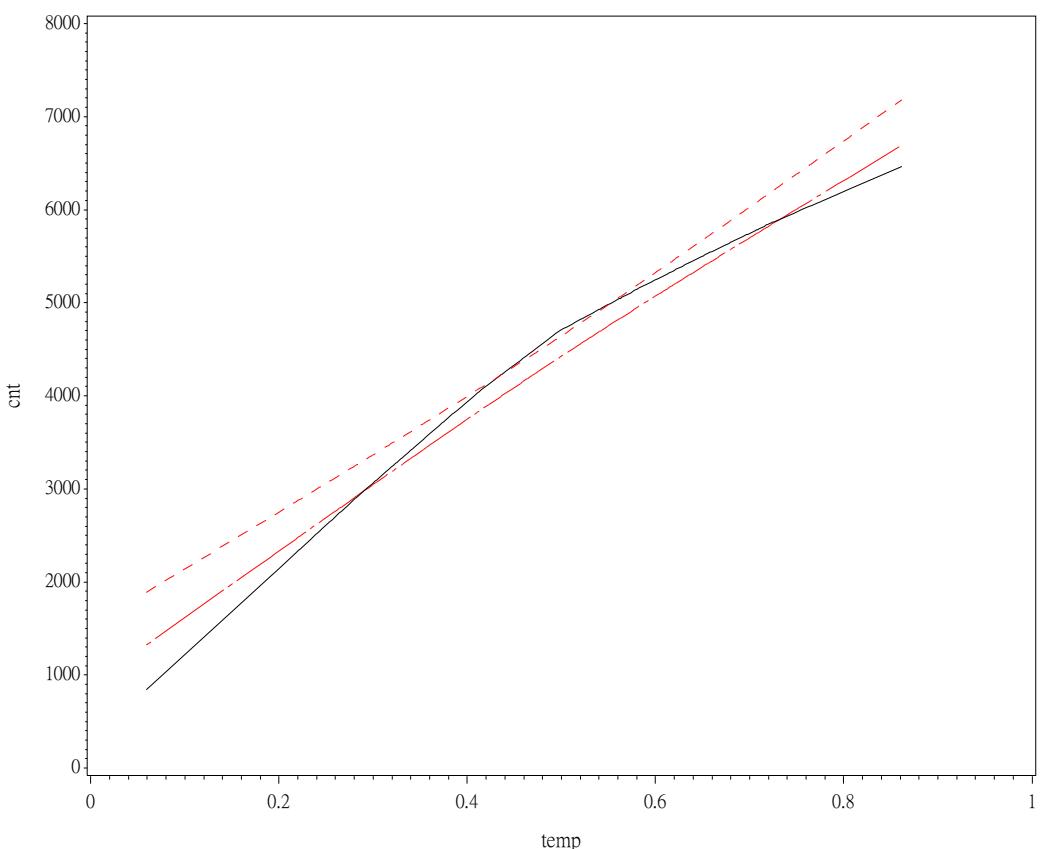
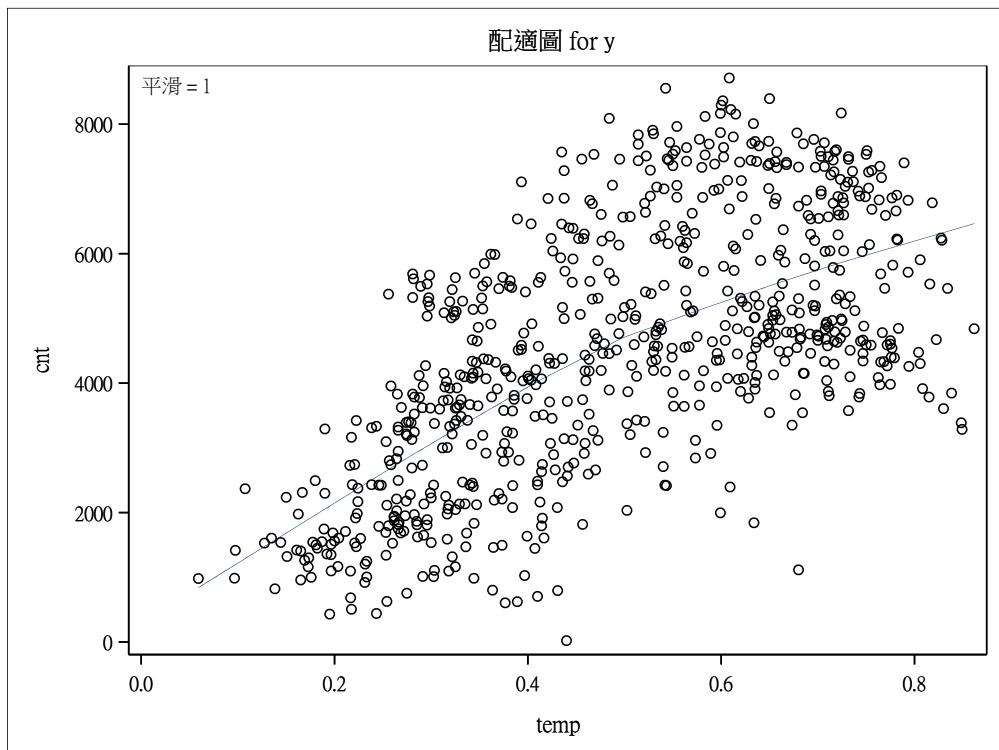


迴歸模型，可看到預測出的迴歸線和觀察值之間的關係

五、模型適當性檢查與矯正：

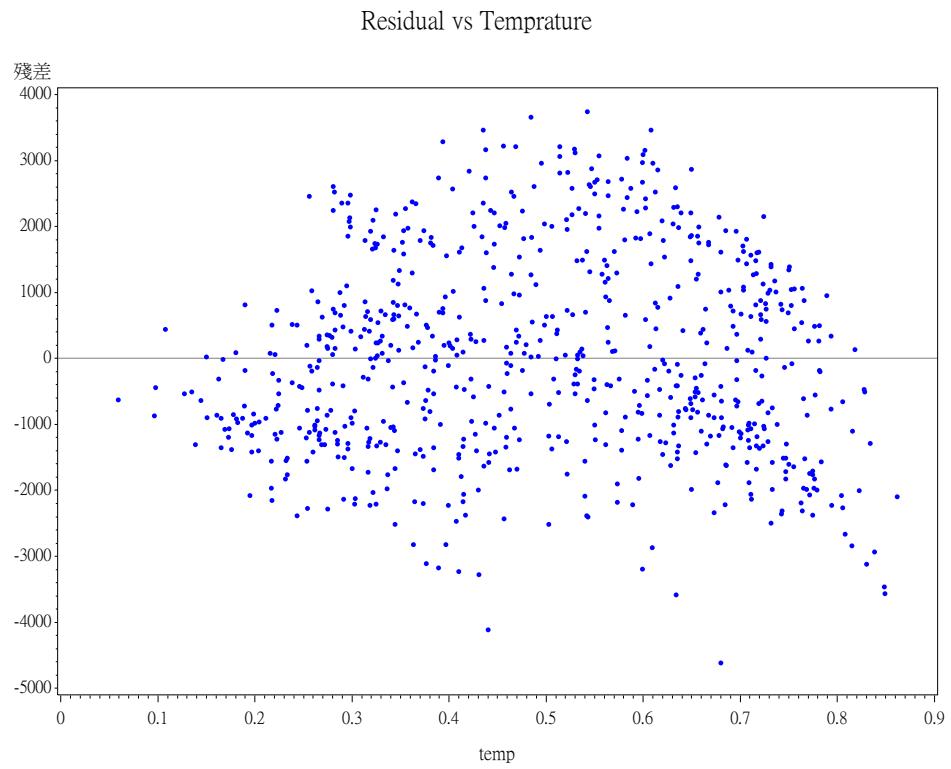
→透過 Lowess 平滑曲線確認所配適的迴歸函數：

配適摘要	
配適方法	kd 樹狀結構
調合	線性
觀測值數目	731
配適點數目	9
kd 樹狀結構儲存區大小	146
本機多項式的次數	1
平滑參數	0.99990
本機鄰域中的點	730
殘差平方和	1568885322

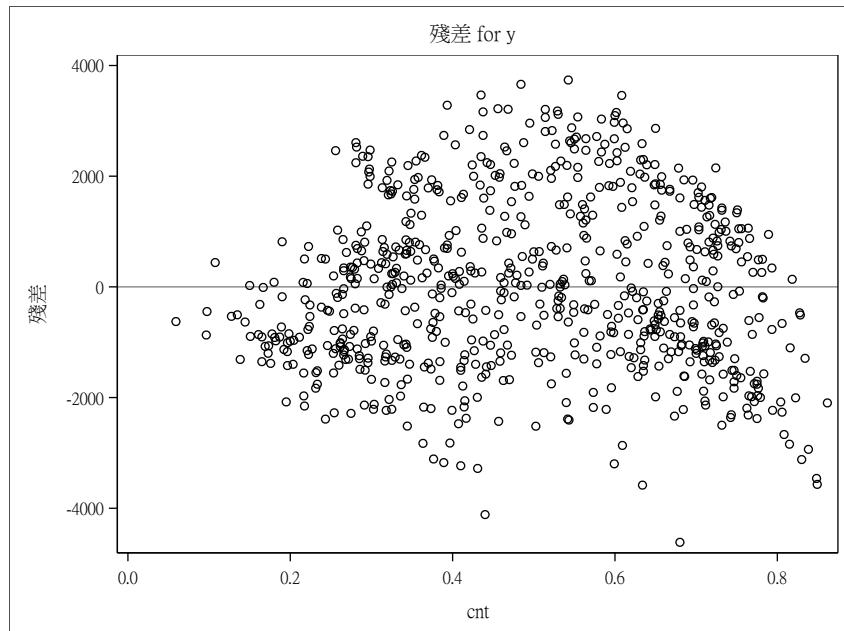


由於平滑曲線沒有完全落在信賴區間帶內，
僅能在 $x=0.3 \sim 0.7$ 區間才能當作支持此迴歸函數的證據。

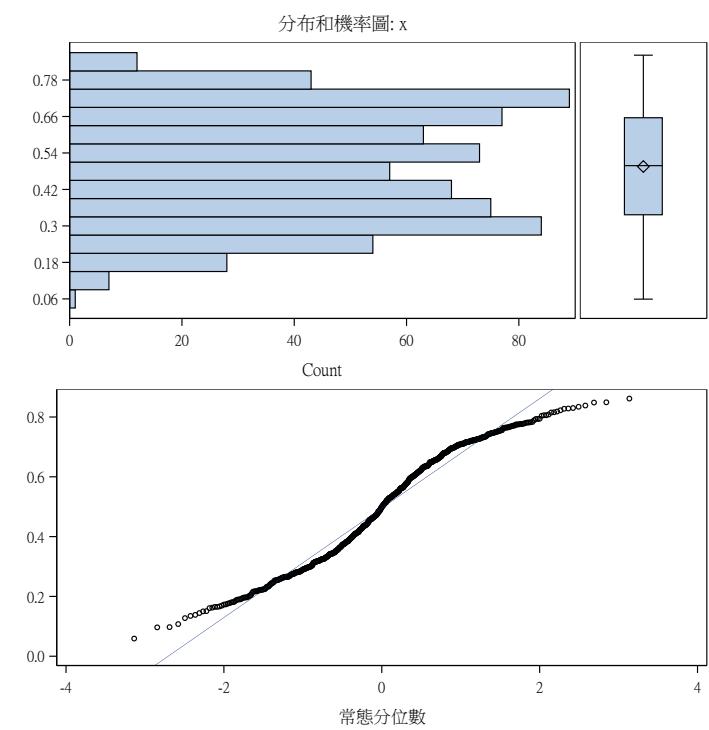
→利用殘差分析評估模型適當性：



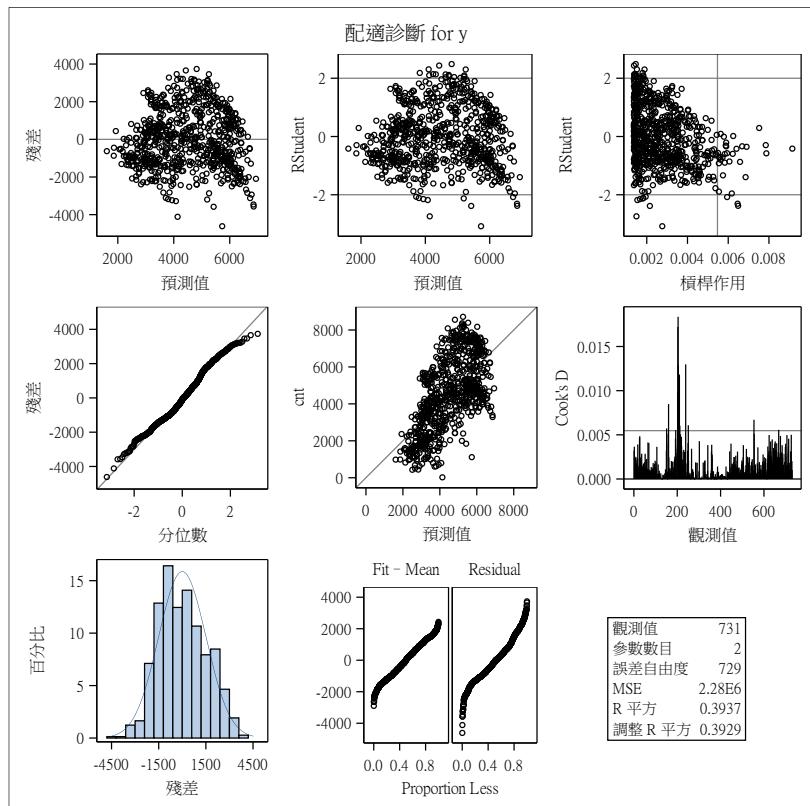
對解釋變數氣溫的殘差圖，看出有些許 outlier，大致分布散亂均勻



對反應變數的殘差圖，看出有些許 outlier，大致分布散亂均勻

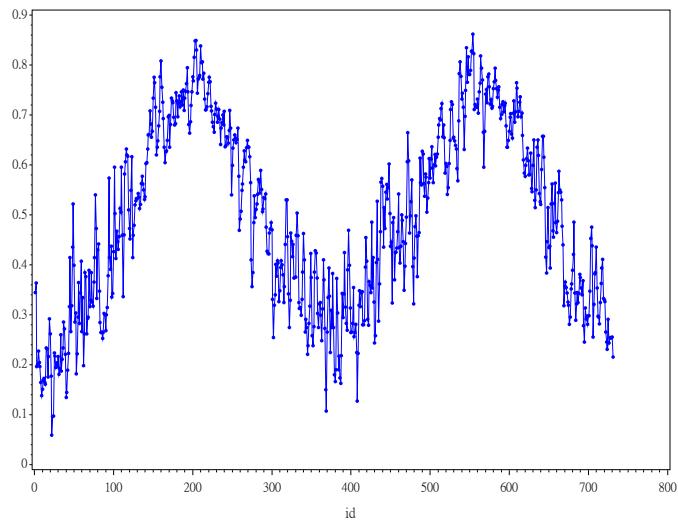


直方圖、盒形圖、QQ 圖

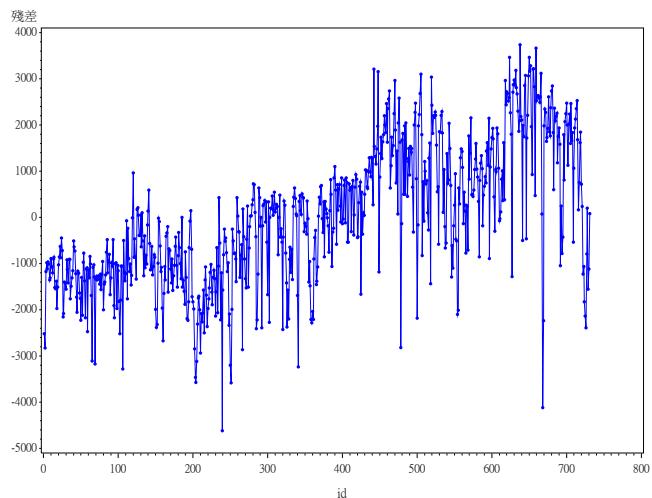


從殘差圖中可以看到在 $+/-4$ 倍的殘差之外有些 outlier，
但未有明顯證據 outlier 是因紀錄、計算錯誤而產生因此暫不考慮忽略 outlier

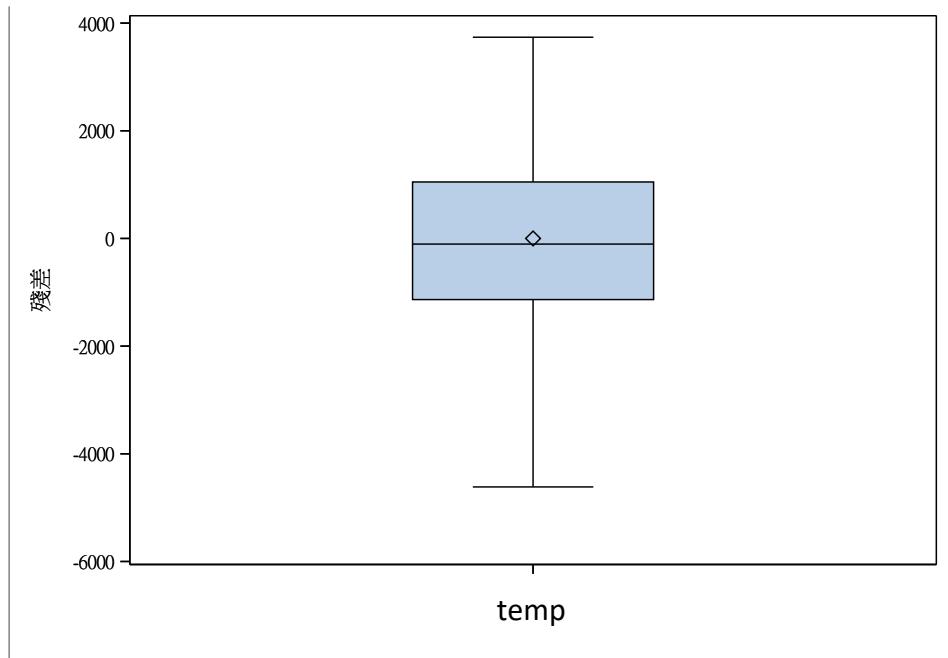
極端觀測值			
最低值		最高值	
值	觀測值	值	觀測值
0.0591304	22	0.834167	546
0.0965217	23	0.838333	210
0.0973913	24	0.848333	203
0.1075000	369	0.849167	204
0.1275000	408	0.861667	554



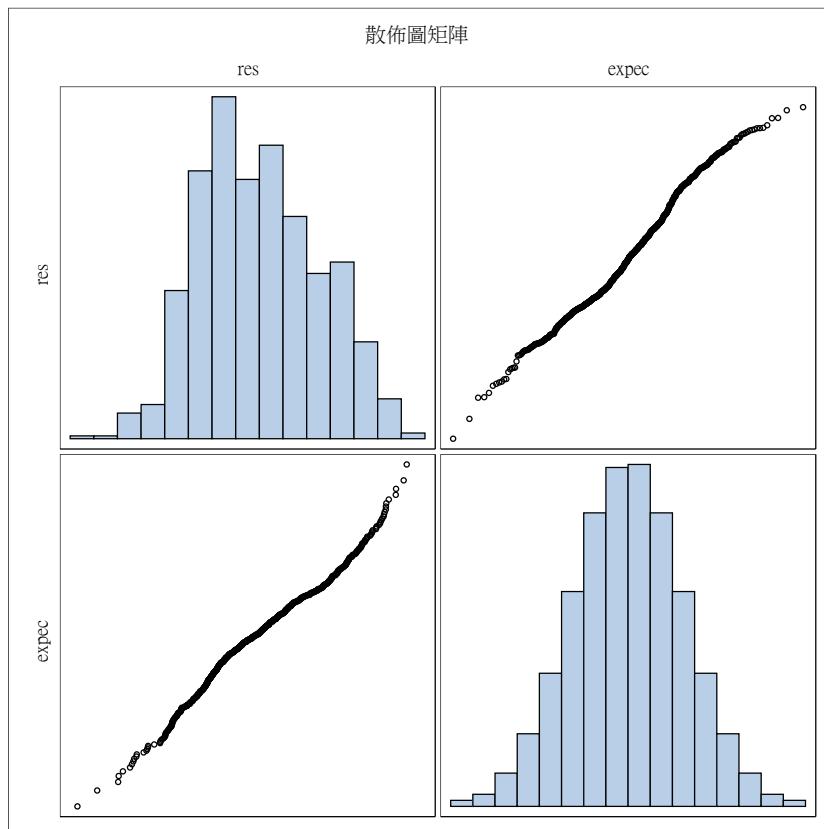
溫度與時間順序圖，從微觀看每一筆相鄰資料呈上下跳動似無規律，但從整體看可發現由於該資料是兩個年度的資料，溫度變化有其規律性



對殘差作時間順序圖檢查是否非常數誤差變異數



殘差盒形圖，大部分資料在兩倍標準差內



觀察殘差與其所對應常態期望值兩者間的相關係數，越高表示越符合常態分配，算出相關係數為0.99364， $p\text{-value}<0.0001$ ，可以認為誤差項未嚴重偏離常態分配。

簡單統計值							
變數	N	平均值	標準差	中位數	最小值	最大值	標籤
res	731	0	1508	-104.39441	-4615	3738	殘差
expec	731	0	1506	0	-4734	4734	

Pearson 相關係數, N = 731 Prob > rl (位於 H0 底下): Rho=0		
	res	expec
res	1.00000	0.99364
殘差		<.0001
expec	0.99364 <.0001	1.00000

總結用殘差分析觀察是否模型偏離共 6 種情形：

- 迴歸函數是不是直線型式：未看出非直線型式的可能
- 誤差項滿不滿足常數變異數：未看出非常數變異數的可能
- 誤差項滿不滿足常態：大部分資料在盒形圖對稱且在常態機率圖對角線上，應符合常態
- 誤差項獨不獨立：對反應變數的殘差圖，大致分布散亂均勻，應屬獨立
- 是否除了少數離群值外，模型配適適當：初步探索分析結果該簡單迴歸模型除了少數離群值外，配適尚可
- 模型中有沒有其他未考慮到的一個或多個重要解釋變數：可加入其他解釋變數使模型更配適

→檢定誤差項變異數是否為常數：

- Brown-Forsythe 檢定： H_0 : 誤差項變異數是常數 vs H_a : 誤差項變異數不是常數，將資料 731 個依氣溫大小排序後分兩群，算出統計量絕對值 = $2.55 > 1.96$ ，推翻虛無假設，由於若統計量絕對值小於等於 1.96，則結論是誤差項變異數是常數；反之，若統計量絕對值大於 1.96，則結論是誤差項變異數不是常數，由此檢定發現該誤差項變異數不是常數，亦即誤差項變異數會隨 X 之水準不同而改變，此一檢定之 p-value 為 0.0111。

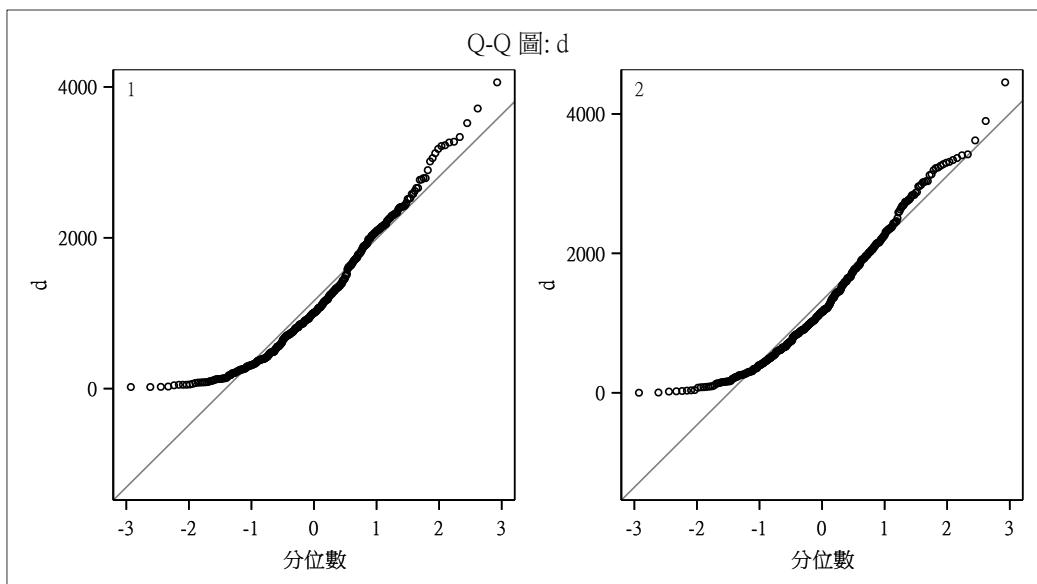
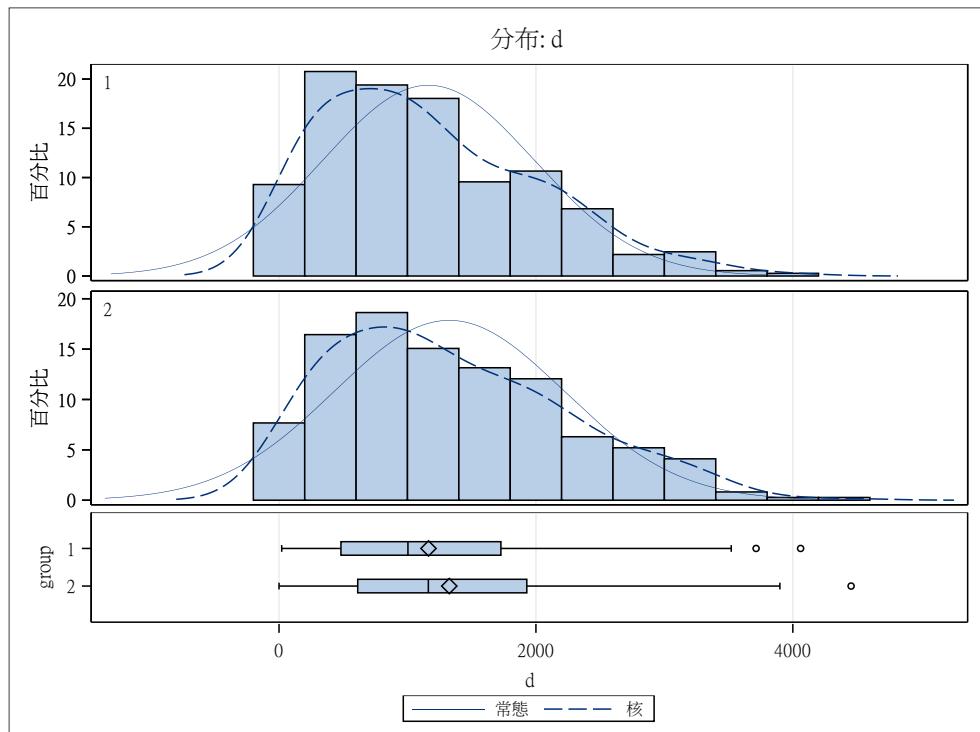
Obs	group	_TYPE_	_FREQ_	殘差
1	1	0	366	-53.899
2	2	0	365	-161.545

group	N	平均值	標準差	標準誤差	最小值	最大值
1	366	1163.6	823.8	43.0583	20.9375	4060.7
2	365	1325.5	893.6	46.7751	0	4453.8
Diff (1-2)		-161.9	859.4	63.5691		

group	方法	平均值	95% CL 平均值		標準差	95% CL 標準差	
1		1163.6	1078.9	1248.2	823.8	768.1	888.2
2		1325.5	1233.5	1417.4	893.6	833.2	963.6
Diff (1-2)	集區	-161.9	-286.7	-37.0885	859.4	817.4	905.9
Diff (1-2)	Satterthwaite	-161.9	-286.7	-37.0731			

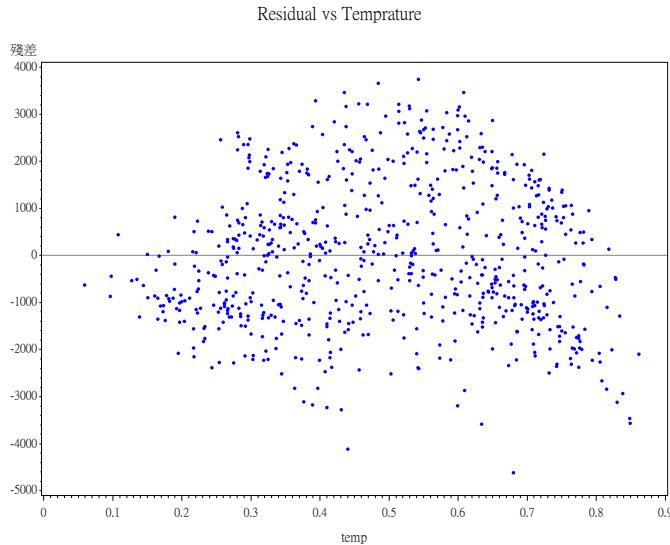
方法	變異數	自由度	t 值	Pr > t
集區	均等	729	-2.55	0.0111
Satterthwaite	不均等	723.89	-2.55	0.0111

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	364	365	1.18	0.1205



2. Breusch-Pagan 檢定： $H_0: \gamma_1$ 等於 0 vs $H_a: \gamma_1$ 不等於 0，算出統計量=6.91595，由於該統計值大於當風險控制在 $\alpha=0.05$ 時的臨界值，所以拒絕虛無假設，誤差項變異數不為常數，此一檢定之機率值為 0.00854。

Obs	ssrs	sse	nobs	tests	pv
1	7.1401E13	1660846806.9	731	6.91595	0.00854



3. 由於 Brown-Forsythe 檢定和 Breusch-Pagan 檢定結果顯示殘差變異數不為常數，推測可能因為原始資料本身在前後端變異數比較大的關係造成，故再取氣溫介於 0.2985~0.7 之區段重新再做一次 Brown-Forsythe 檢定：令 H_0 : 誤差項變異數是常數 vs H_a : 誤差項變異數不是常數，將資料 471 個依氣溫大小排序後分兩群，算出統計量絕對值 = $1.52 < 1.96$ ，不拒絕虛無假設，結論是誤差項變異數是常數，此一檢定之 p-value 為 0.1304。

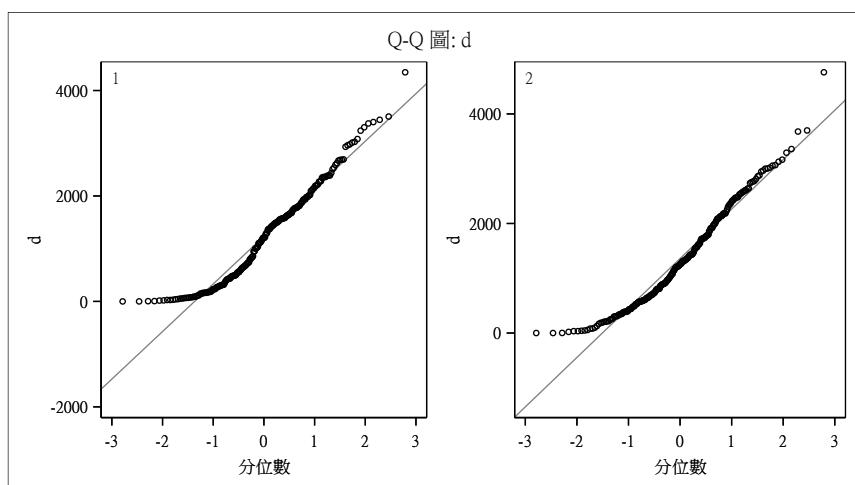
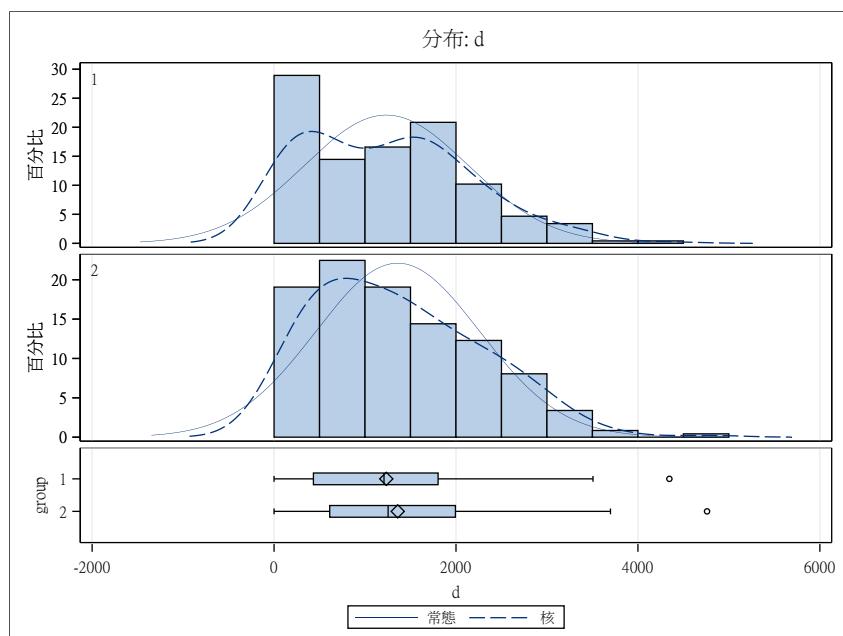
Obs	group	_TYPE_	_FREQ_	殘差
1	1	0	235	10.493
2	2	0	236	-223.849

group	N	平均值	標準差	標準誤差	最小值	最大值
1	235	1233.7	902.4	58.8667	0	4345.7
2	236	1359.7	902.4	58.7411	0.1786	4759.8
Diff (1-2)		-126.0	902.4	83.1613		

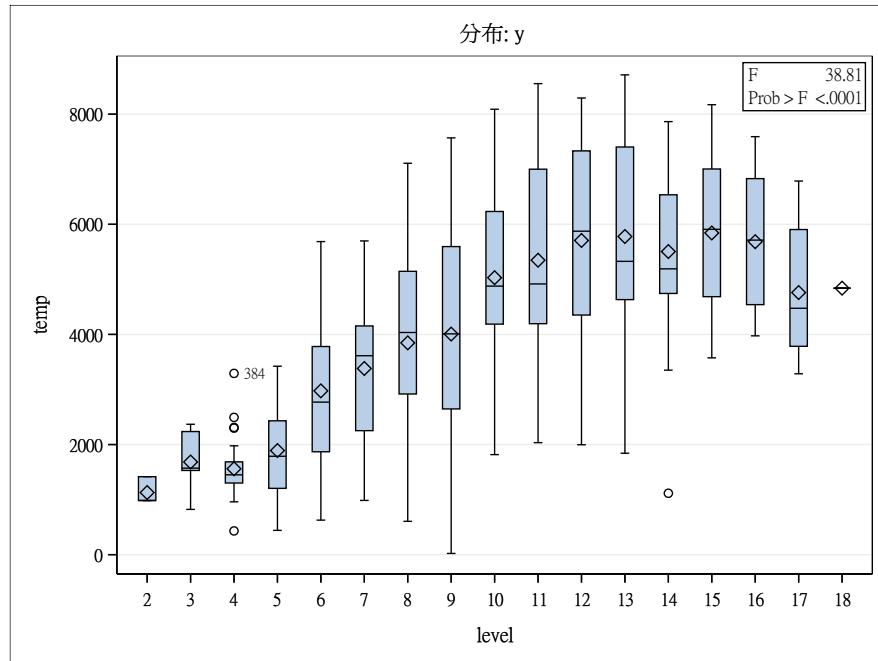
group	方法	平均值	95% CL 平均值		標準差	95% CL 標準差	
			1117.7	1349.7		827.5	992.3
1		1233.7			902.4		
2		1359.7	1244.0	1475.4	902.4	827.7	992.1
Diff (1-2)	集區	-126.0	-289.4	37.3995	902.4	848.2	964.1
Diff (1-2)	Satterthwaite	-126.0	-289.4	37.3996			

方法	變異數	自由度	t 值	Pr > t
集區	均等	469	-1.52	0.1304
Satterthwaite	不均等	468.99	-1.52	0.1304

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	234	235	1.00	0.9998



→利用 lack of fit 的 F 檢定來看配適程度，統計量 F^* =配適不良均方(MSLF)/純誤差均方(MSPE)，結果顯示為 $F^*>$ 臨界值，顯示配適不佳， $p\text{-value}=0.0056$ 。



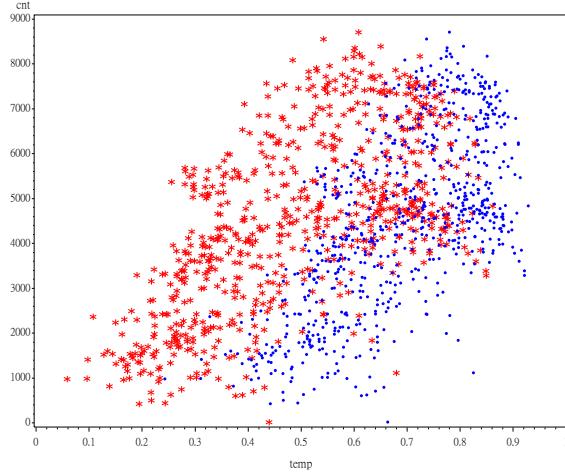
類別層級資訊		
類別	層級	值
level	17	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

來源	自由度	平方和	平均值	F 值	Pr > F
			平方		
模型	1	1078688585	1078688585	473.47	<.0001
誤差	729	1660846807	2278254		
配適不足	497	1231665008	2478199	1.34	0.0056
純誤差	232	429181799	1849922		
已校正的總計	730	2739535392			

→矯正測量：

- 觀察 lowess 無母數迴歸的圖，可能是非線性迴歸型態，凹口向下，考慮對 x 進行轉換 $x'=x^{1/2}$ ，對轉換後的資料畫散佈圖，可合理假設其線性關係，由於資料散佈的範圍並未改變，所以不需對 y 進行轉換，配適函數： $y=-1813.7+9150.9x^{1/2}$ ，無明顯證據指出不良配適，殘差的常態機率圖也並

無偏離常態分配之明顯態勢，故簡單線性迴歸模型對此轉換後的資料應該是合適的，但與原始簡單線性迴歸模型差異不大。

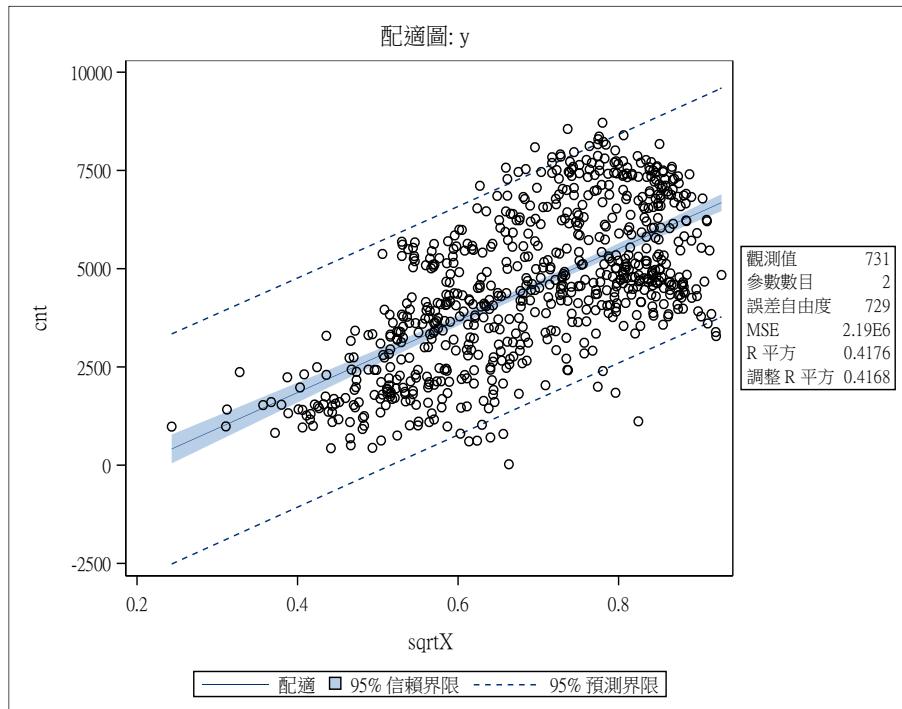
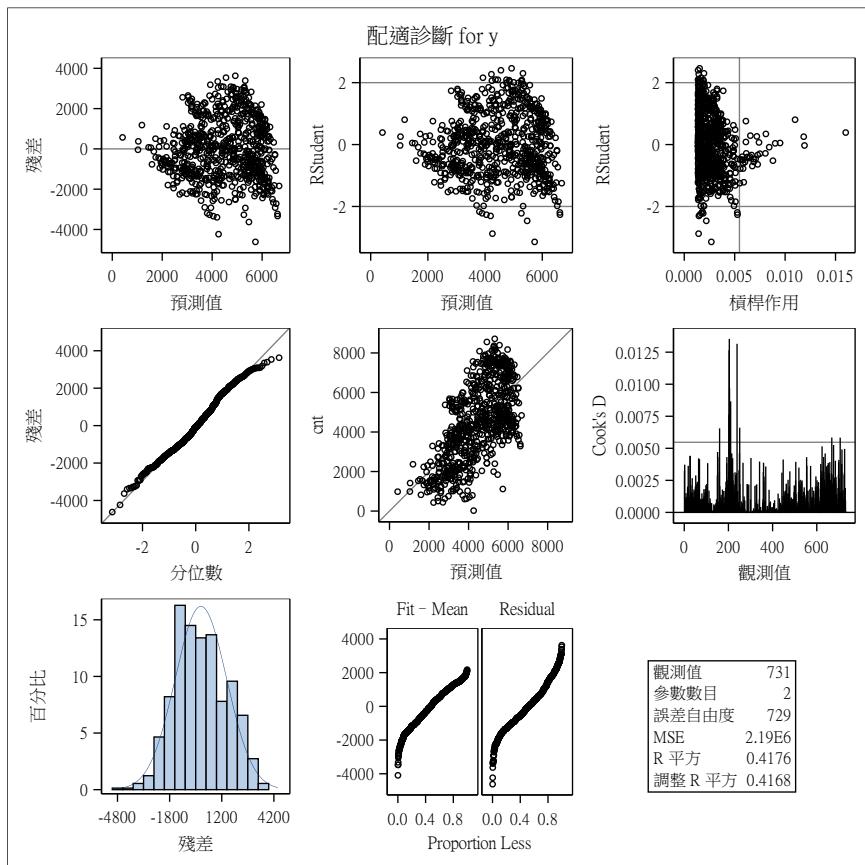


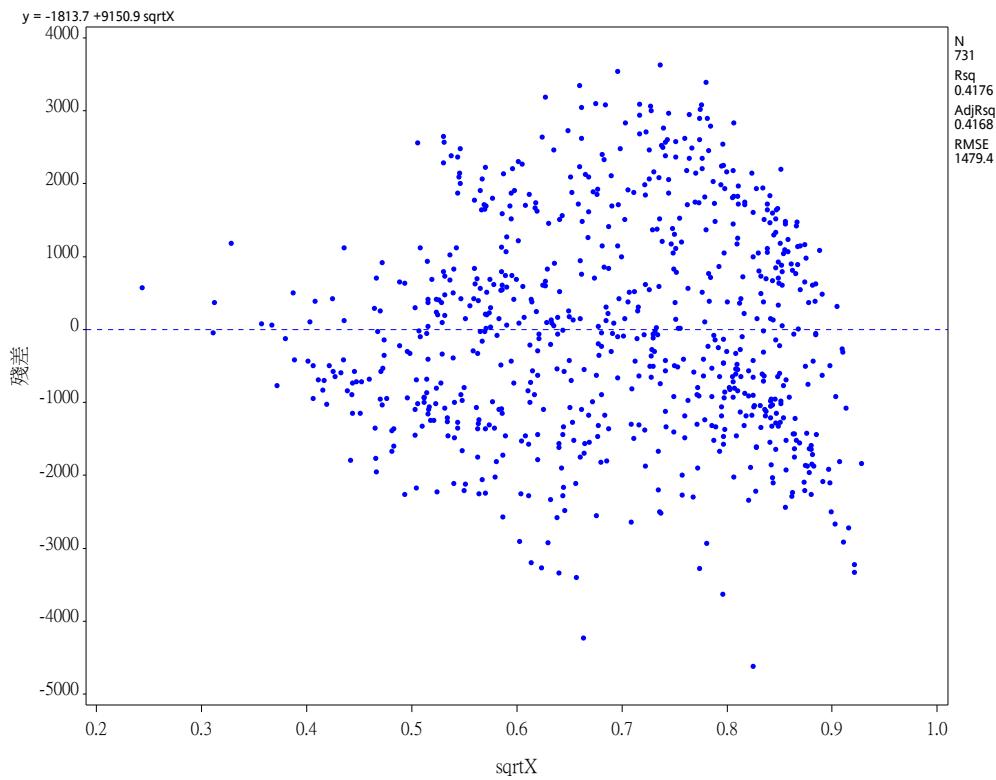
紅色點為原始的 x ，藍色為轉換後的 x'

變異數分析					
來源	自由度	平方和	平均值 平方	F 值	Pr > F
模型	1	1144135806	1144135806	522.80	<.0001
誤差	729	1595399586	2188477		
已校正的總計	730	2739535392			

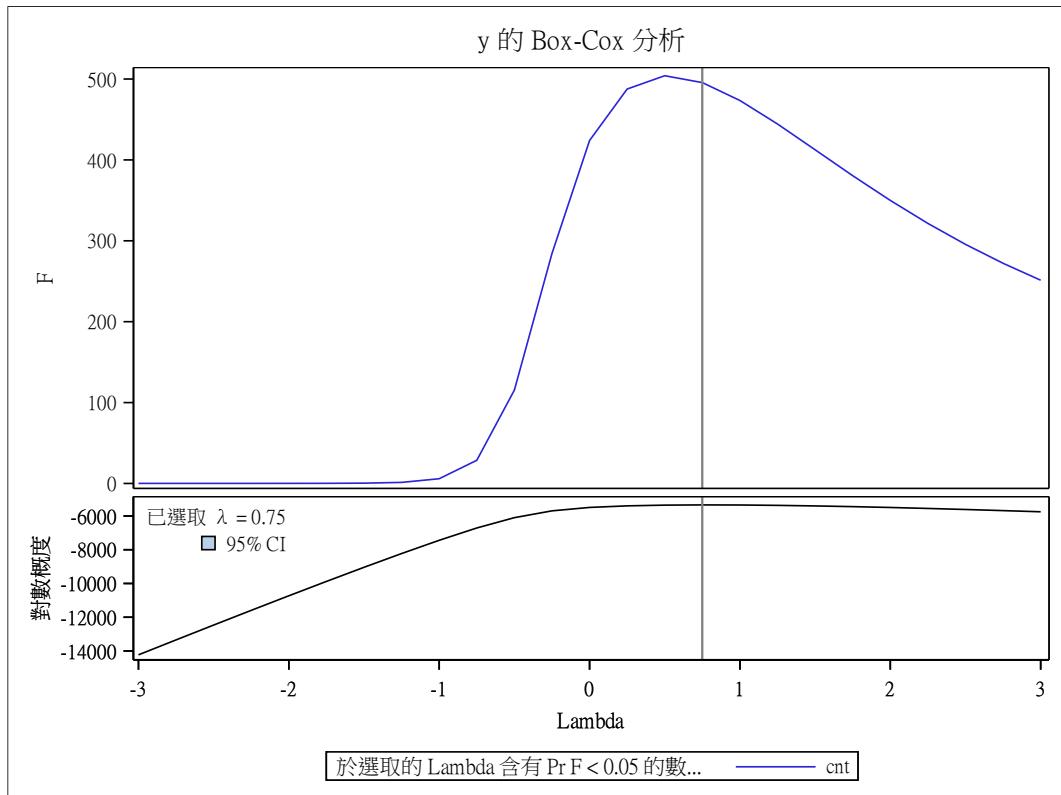
根 MSE	1479.35013	R 平方	0.4176
應變平均值	4504.34884	調整 R 平方	0.4168
變異係數	32.84271		

參數估計值						
變數	標籤	自由度	參數 估計值	標準 誤差	t 值	Pr > t
Intercept	Intercept	1	-1813.72017	281.68784	-6.44	<.0001
sqrtX		1	9150.91642	400.21814	22.86	<.0001

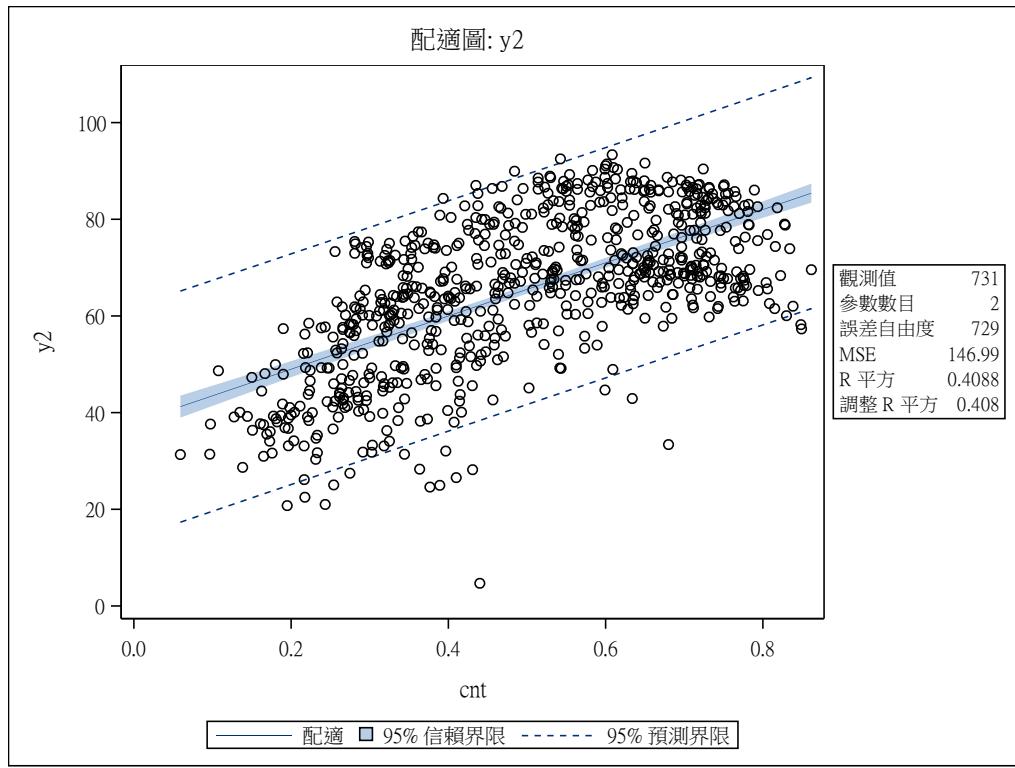
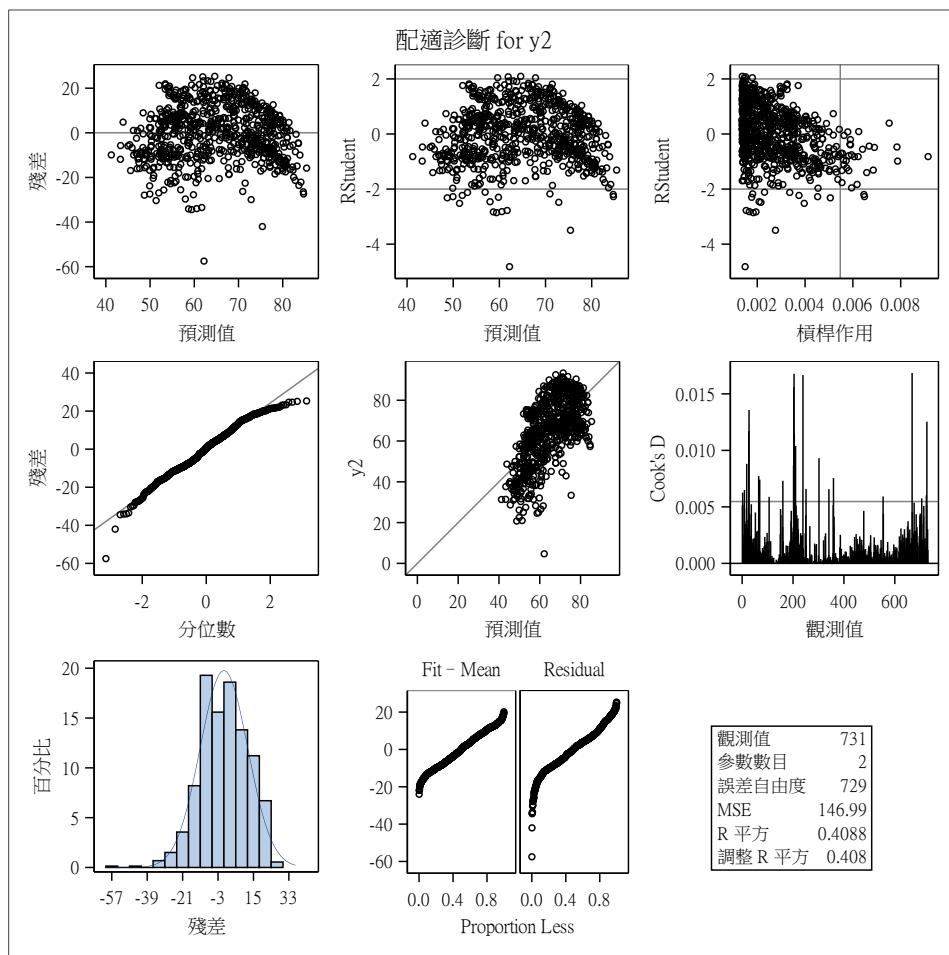


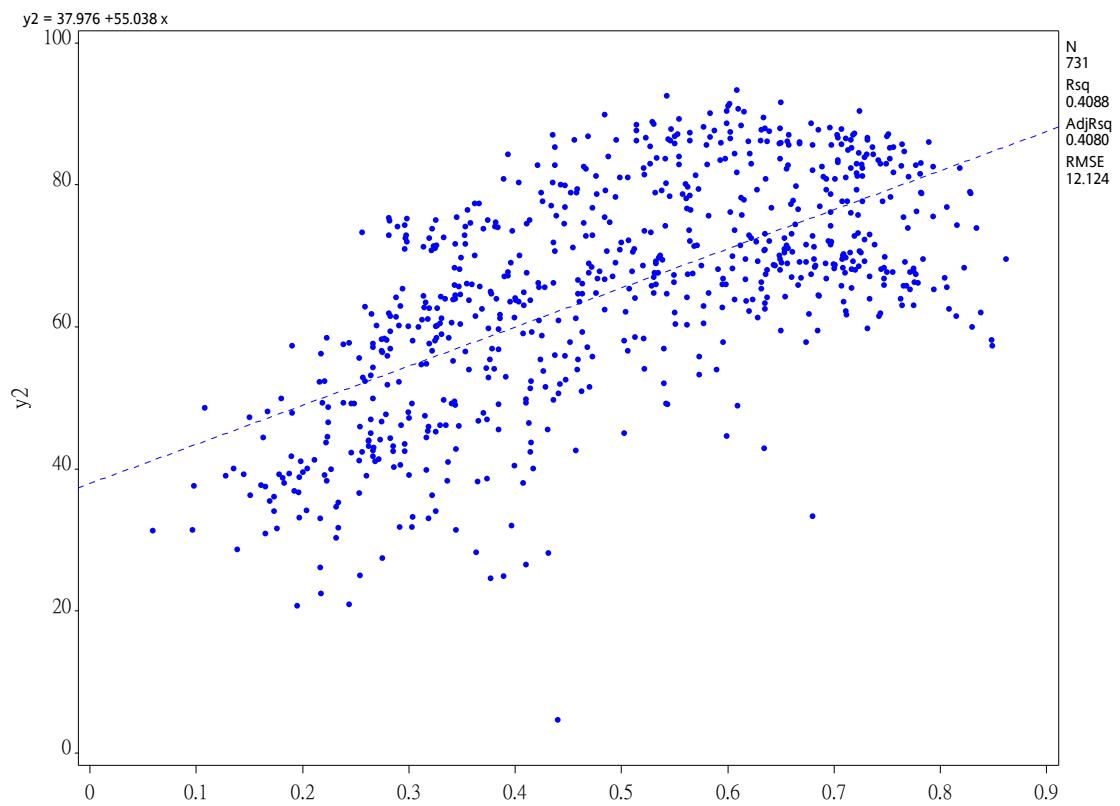


2. 利用 box-cox 找到適合的 λ 值對反應變數進行轉換，經 box-cox 程序得到
 λ 的最大概似估計值是 $\lambda = 0.75$

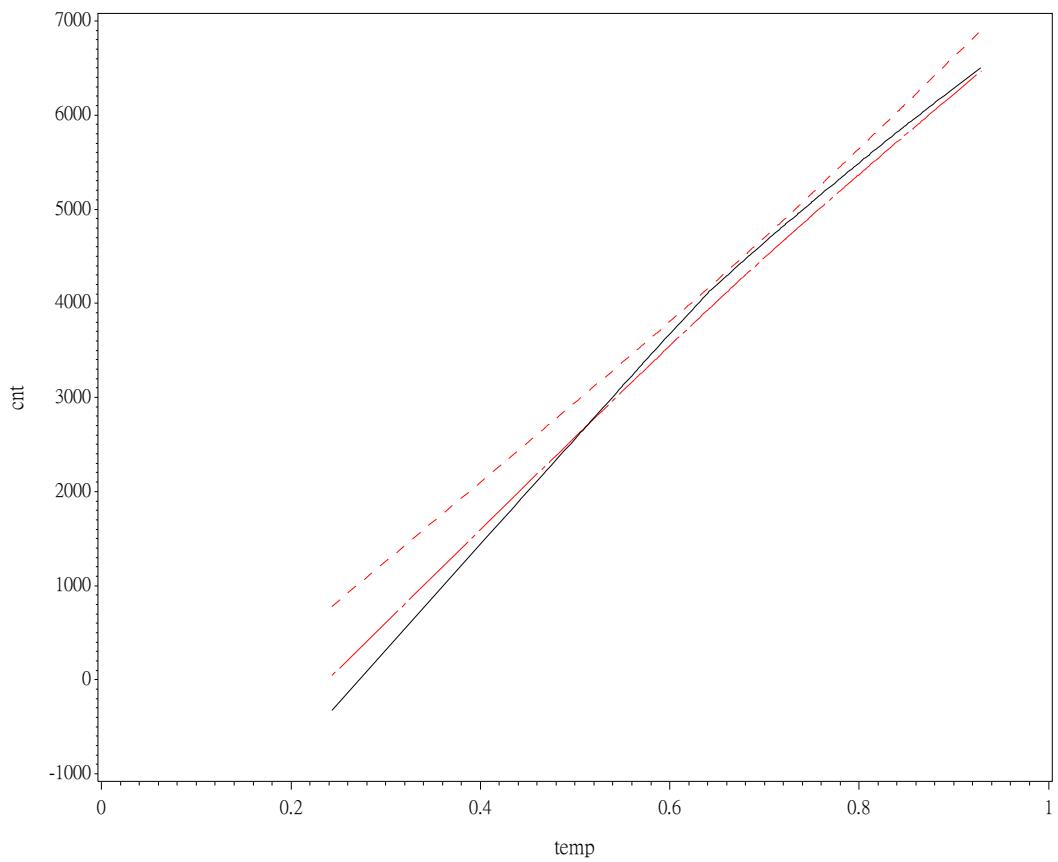


$\lambda = 0.75$ 適用 $Y' = y^{1/2}$ 轉換，結果如圖：

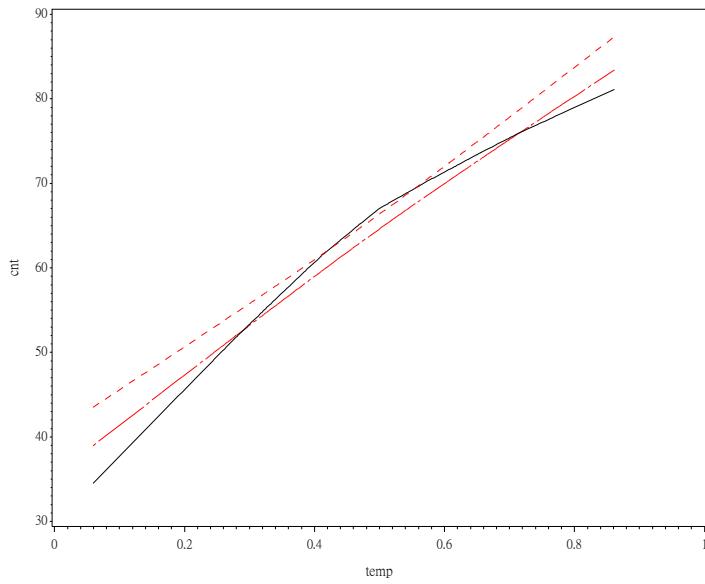




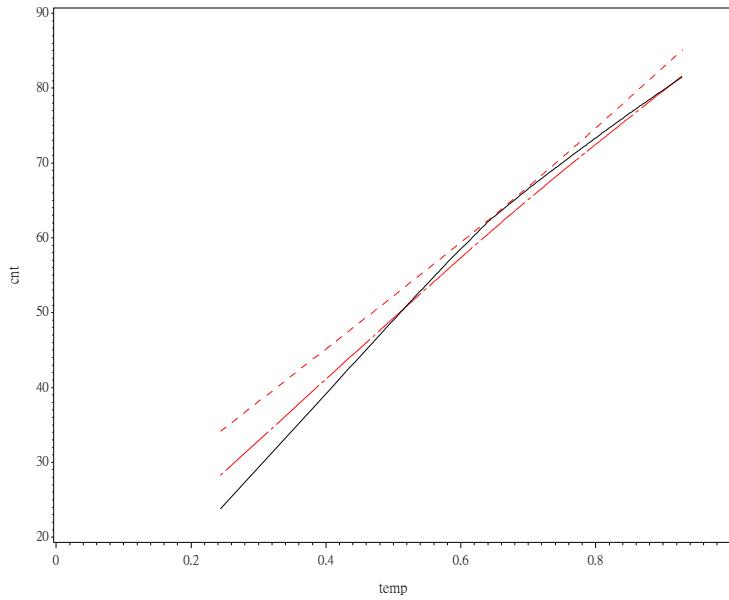
→→→x 轉換成 x' 後再用 lowess 無母數迴歸觀察配適的模型，如下：



→→y 轉換成 y' 後再用 lowess 無母數迴歸觀察配適的模型，如下：



→→x 和 y 皆轉換後用 lowess 無母數迴歸觀察配適的模型，如下：



→→轉換後和原始模型變化不大，畫出的 lowess 平滑曲線仍一半左右的觀察值若在信賴區間帶中。

六、結論：

資料符合常態，lack of fit 結果及 lowess 無母數方法皆顯示可能有非線性關係，之後可加入其他解釋變數讓模型更配適。

七、參考資料：

- [1] Lichman, M. (2013). UCI Machine Learning Repository [<http://archive.ics.uci.edu/ml>]. Irvine, CA: University of California, School of Information and Computer Science.

[2] Fanaee-T, Hadi, and Gama, Joao, "Event labeling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi: 10.1007/s13748-013-0040-3.