



統計系在職專班
迴歸分析期中報告

環境變數對登革熱確診病例影響

教授：蘇南誠博士

學生：711378912 蔡宜誼

中 華 民 國 1 1 3 年 1 1 月

目錄

第一章 緒論	2
第一節 研究背景與動機	2
第二節 研究目的	3
第三節 研究步驟	3
第二章 研究方法	4
第一節 研究對象	4
第二節 資料處理	5
第三章 研究結果	7
第一節 敘述性統計	7
第二節 假設檢定	17
第三節 相關係數檢定	22
第四節 簡單線性迴歸	24
第五節 多變量回歸分析	30
第四章 結論與建議	36
第一節 結論	36
第二節 建議	38

第一章 緒論

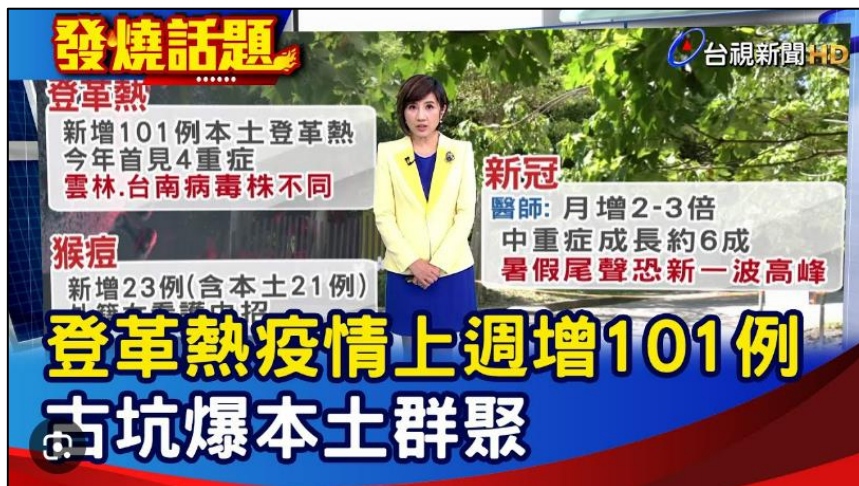
本研究旨在討論環境變數（溫度、降雨量、濕度等）對登革熱確診病例影響。本章主要分為（1）研究背景與動機；（2）研究目的；（3）研究步驟。

第一節 研究背景與動機

世界衛生組織表示，2023 年全球報告了超 500 萬登革熱感染病例，包括 5000 例死亡病例。這一出乎意料的病例激增，對公共衛生構成了潛在的嚴重威脅。尤其是像台灣這樣擁有溫暖濕潤氣候的地區，登革熱的季節性爆發尤為顯著。

聯合新聞網中有提到，近日每天都可以看到登革熱相關的新聞，也出現越來越多登革熱的病例，光是新北在 2024/09/30 日就再增 6 例本土登革熱病例，累計達 49 例(<https://udn.com/news/story/123735/8259545>)。此外，根據台視新聞，疫情的快速上升還伴隨著本土群聚現象，進一步提高了疾病傳播的風險。

登革熱的疫情不僅僅對台灣的衛生造成負擔，還禱顯了氣候變遷、都市化等很多的因素對疾病傳播的交互影響。因此，本研究在探討影響登革熱疫情的變數，包括空氣污染、氣溫與降雨量等變項，期待可以增強對登革熱的預防。



圖片取自台視新聞(<https://www.youtube.com/watch?v=pSDhBLn9RcU>)

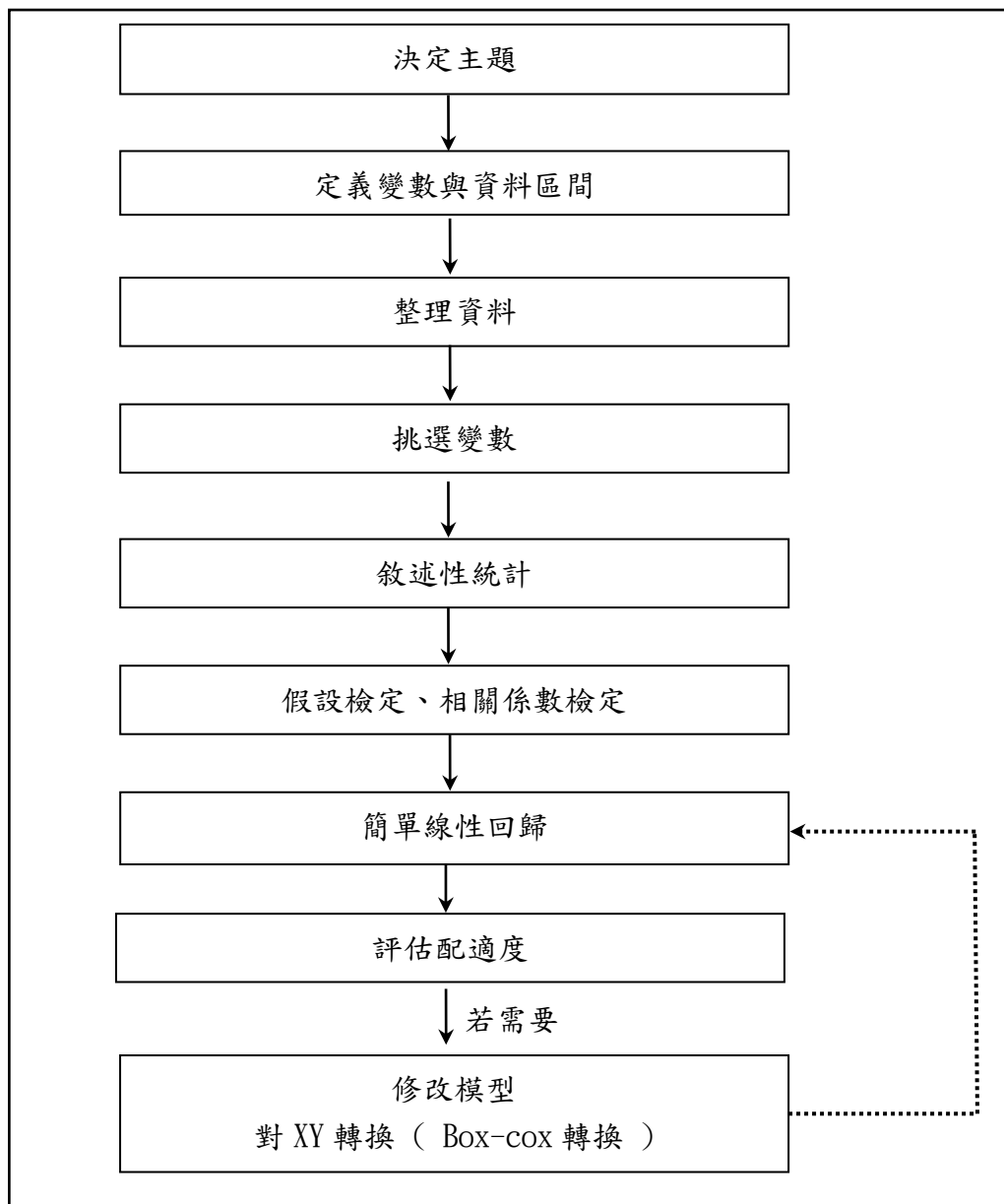
第二節 研究目的

這裡將資料的區間定為 2023 年全年度登革熱確診資料，這裡以高雄地區為例。研究內容區分如下：

- 一、 探討解釋變數（環境）、反應變數（確診百分比）基本統計量
- 二、 分析解釋變數（環境）與反應變數（確診百分比）的相關性
- 三、 建立模型並評估其適配度，並在必要時進行模型修正以提高解釋力。

第三節 研究步驟

本節針對研究主題與研究目的，選用之研究方法與研究步驟如下：

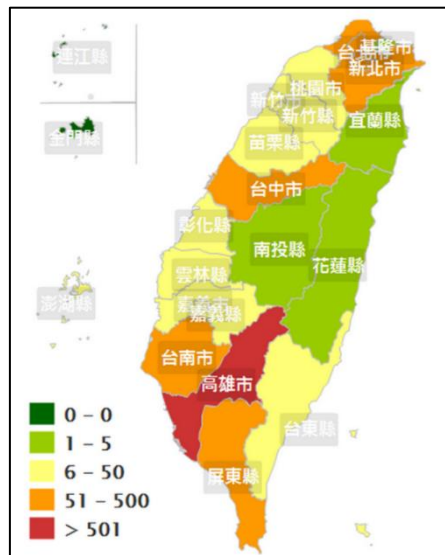


第二章 研究方法

本章共分為兩節，（1）研究對象；（2）資料處理。茲就各節逐一敘述之。

第一節 研究對象

資料的區間定為 2023 年全年登革熱確診資料，這裡以高雄地區為例，從登革熱病例地理分布圖可以看出高雄市為台灣嚴重的縣市。且高雄地區的氣候特徵為高溫濕潤，符合蚊蟲的生存條件。因此這裡我們選擇高雄作為研究對象。



圖片取自衛福部疾管屬登革熱統計資料 記者陳俊廷翻攝

(<https://www.peoplemedia.tw/news/70b63bcd-2977-469b-be59-e5b9338d9a65>)

第二節 資料處理

研究中用到三個資料，分別為登革熱確診數、各空氣品質資料、高雄地區各區人口數。此次研究反應變數為登革熱確診百分比，解釋變數為空氣品質等等的環境因素。

壹、登革熱確診數

資料來源為疾病管制署資料開放平台，登革熱 1998 年起每日確定病例統計。(<https://data.cdc.gov.tw/dataset/dengue-daily-determined-cases-1998>) 原始資料為 1998 年至今登革熱每日確定病例統計，資料粒度為日，其中不列入境外移入資料，原始資料篩選後 2023 年高雄資料共計 3145 筆，研究所用到的資料為 2023 年每日確定病例分組統計後結果，這裡我們分組以高雄地區各區作分組，統計後共計 695 筆資料。

	A	B	C	D	E	F	G	H
1	發病日	通報日	性別	年齡層	居住縣市	居住鄉鎮	是否境外移入	確定病例數
2	1998/1/2	1998/1/7 M		40-44	屏東縣	屏東市	否	1
3	1998/1/3	1998/1/14 M		30-34	屏東縣	東港鎮	是	1
4	1998/1/13	1998/2/18 M		55-59	宜蘭縣	宜蘭市	是	1
5	1998/1/15	1998/1/23 M		35-39	高雄市	苓雅區	否	1
6	1998/1/20	1998/2/4 M		55-59	宜蘭縣	五結鄉	否	1
7	1998/1/22	1998/2/19 M		20-24	桃園市	蘆竹區	是	1
8	1998/1/23	1998/2/2 M		40-44	新北市	新店區	否	1
9	1998/1/26	1998/2/19 F		65-69	台北市	北投區	否	1
10	1998/2/11	1998/2/13 F		25-29	台南市	南區	是	1
11	1998/2/16	1998/2/24 M		20-24	高雄市	楠梓區	是	1
12	1998/2/17	1998/2/23 F		30-34	高雄市	鳳山區	否	1

貳、高雄地區各區人口數

資料來源為高雄市政府民政局 (<https://cabu.kcg.gov.tw/Stat/StatRpts/StatRpt1.aspx?yq=112&mq=1&dq=>) 其中以區域作分組，這裡資料粒度為月份，但這裡直接將每月的資料分配到每日的資料。

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	行政區	1	2	3	4	5	6	7	8	9	10	11	12
2	左營區	196003	196192	196479	196620	196769	196824	197023	197009	196953	197026	197078	197276
3	楠梓區	191521	191672	191853	191929	192105	192242	192485	192667	192777	192888	192944	193139
4	前金區	26728	26808	26856	26897	26929	26924	26972	27011	27033	27053	27064	27098
5	小港區	154852	154817	154878	154878	154931	154977	155026	155066	155042	155034	154970	155009
6	鳳山區	356359	356633	356693	356712	356813	356713	356762	356594	356397	356463	356476	356536
7	仁武區	95309	95522	95745	95934	96131	96353	96582	96808	97008	97155	97261	97337
8	大寮區	111575	111578	111574	111663	111770	111735	111845	111906	111910	111923	111916	111986
9	林園區	68459	68374	68340	68371	68303	68292	68325	68354	68328	68299	68254	68216
10	橋頭區	40551	40689	40819	40878	40971	41036	41138	41346	41488	41605	41685	41712

參、各空氣品質資料

資料來源為空氣品質監測網

(https://airtw.moeenv.gov.tw/CHT/Query/His_Data.aspx) 其中將資料作格式的轉換，從中抓取每日環境變數資料的最大值、最小值、平均值作為變數，例如當日最高溫度、當日平均溫度、當日總降雨量等等，這裡資料時間粒度原為為一小時一筆，整理過後為一天一筆資料，這裡原先有 105120 筆資料，經由整理過後剩下 4380 筆資料。



第三章 研究結果

本研究旨在進行台灣高雄地區之研究。本章共分為四節，（1）敘述性統計；（2）假設檢定；（3）相關係數檢定；（4）簡單線性回歸。茲就各節逐一敘述之。

第一節 敘述性統計

此研究的反應變數為登革熱確診百分比，但因分母很大，算出來的數值很小，因此我們的單位為登革熱確診ppm，另外有相當多的變數其中我們挑選看起來最可能影響登革熱確診數的11個連續變數以及2個類別變數作為解釋變數，其中連續變數包含：當日總降雨量、當日最高溫度、當日平均溫度、當日最高懸浮微粒PM2.5、當日平均懸浮微粒PM2.5、當日最高相對濕度、當日平均相對濕度、當日最高臭氧濃度、當日平均臭氧濃度、當日最高一氧化氮濃度、當日平均一氧化氮濃度；類別變數包含地區、季節。

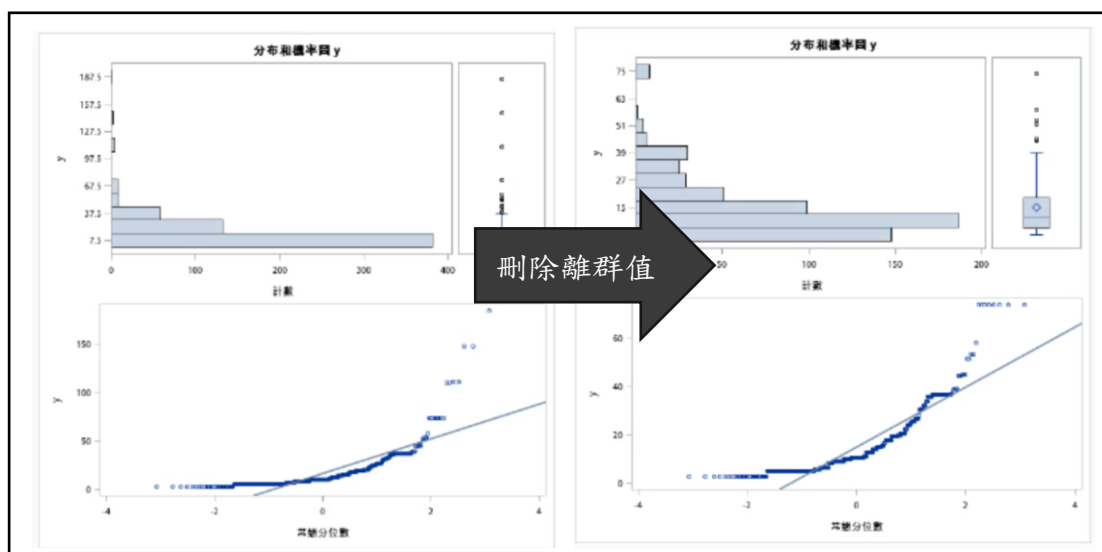
壹、 反應變數 – 登革熱確診 ppm

這裡抓取登革熱確診數除以該地區總人口數即為登革熱確診 ppm，表示每一百萬人中就有幾個人確診登革熱，公式如下：

$$\text{登革熱確診 ppm} = \frac{\text{登革熱確診數}}{\text{人口數}} * 1000000$$

確診ppm 敘述性統計				
UNIVARIATE 程序				
變數: y				
動差				
N	595	總和加權	595	
平均值	16.2874216	總和觀測值	9691.01584	
標準差	17.9439699	變異數	321.986056	
偏態	4.32061172	峰態	27.1517912	
未校正 SS	349101.378	已校正平方和	191259.717	
變異係數	110.170722	標準誤差平均值	0.73563105	
常態性檢定				
檢定	統計值		p 值	
Shapiro-Wilk	W	0.601587	Pr < W	<0.0001
Kolmogorov-Smirnov	D	0.226184	Pr > D	<0.0100
Cramer-von Mises	W-Sq	10.17898	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	55.6781	Pr > A-Sq	<0.0050

最大值為 184.82，最小值為 2.8，平均值為 16.29，峰態=27.15>0，為右偏，絕大多數的值位於平均值的左側。由 Shapiro-Wilk 常態性檢定<0.0001 可以看出他不符合常態分配。



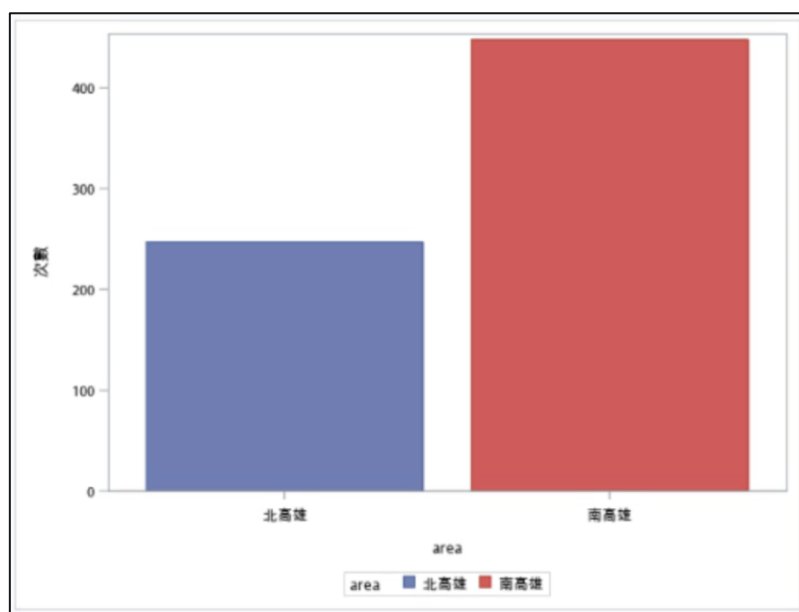
這張圖可以看的出來大於 100 的資料不太多，這裡判定他為離群值，因此後續會把它刪除後得到右圖結果。刪除過後比較能看出整體數字的型態。後續將用這個資料去進行分析。

貳、解釋變數－類別變數

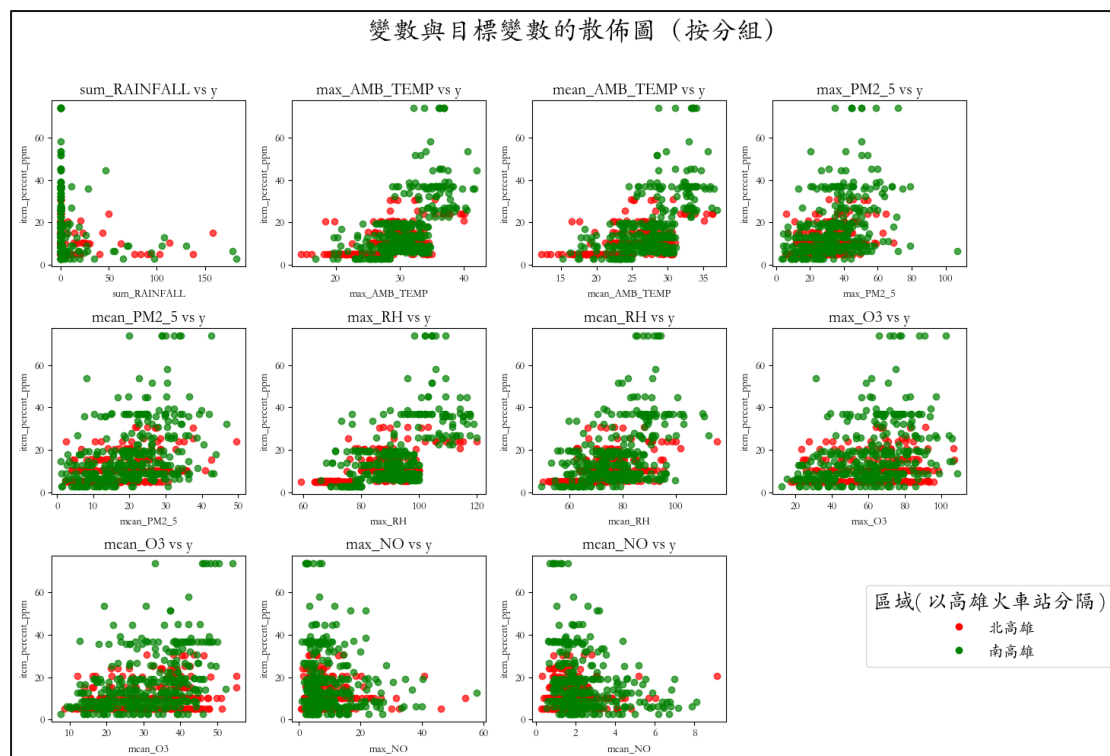
這裡兩個類別變數後續整理加上去的，分別為地區、季節。

甲、地區

這裡的地區以北高雄以及南高雄作為區分，高雄火車站以北為北高雄，以南為南高雄，其中北高雄為「三民區」、「鼓山區」、「左營區」、「楠梓區」、「仁武區」、「橋頭區」、「美濃區」，南高雄為「鹽埕區」、「新興區」、「前金區」、「苓雅區」、「前鎮區」、「小港區」、「大寮區」、「復興區」、「林園區」、「鳳山區」。



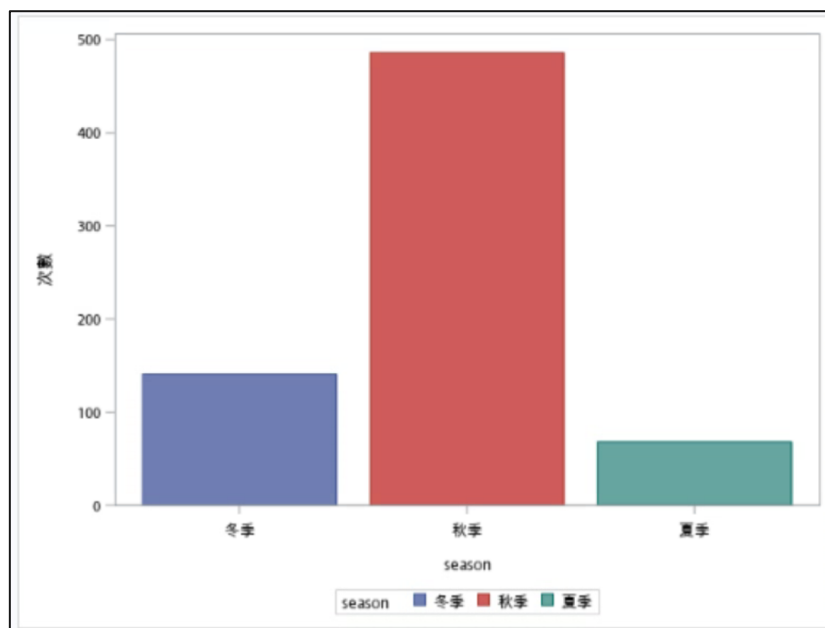
從此圖中可以看出南高雄比北高雄登革熱確診人數更多。



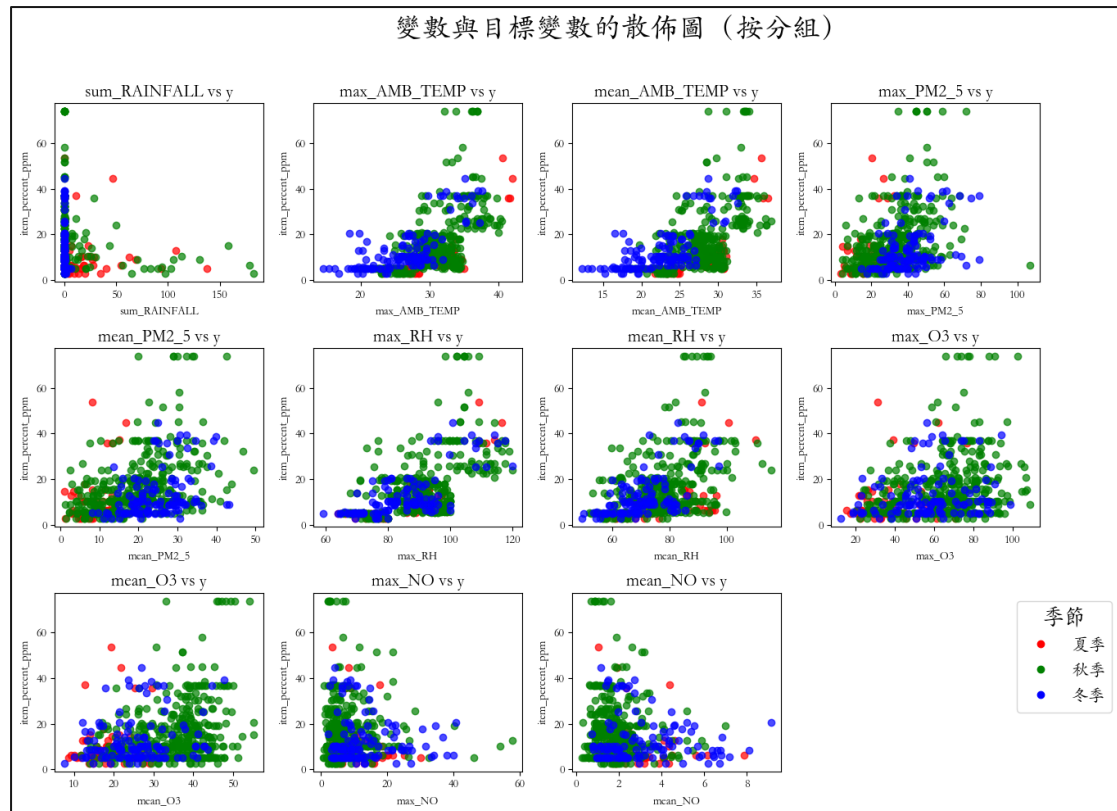
從圖中可以看出，兩組數據的分布沒有明顯的區隔，許多變數的數據點在不同組之間有較大的重疊。因此，目前並不需要進行分組來建立模型，可以直接針對整體數據進行分析或建模。

乙、 季節

這裡的季節分為春季、夏季、秋季、冬季，其中春季為「三月」、「四月」、「五月」，夏季為「六月」、「七月」、「八月」，秋季為「九月」、「十月」、「十一月」，冬季為「十二月」、「一月」、「二月」。

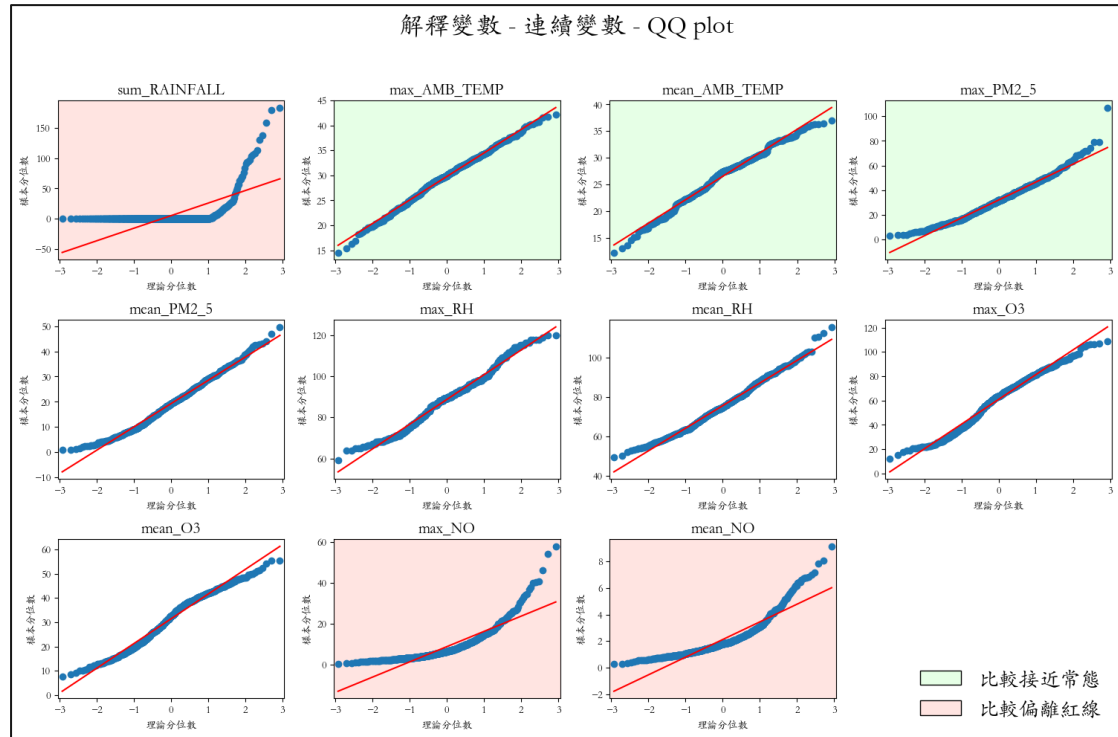


從此圖可以看出大多集中在秋季，猜測可能是因為夏季開始確診登革熱並且經過時間傳染，導致秋季病例數達到高峰，而冬季病例數則逐漸趨緩，可能與氣候條件不利於蚊蟲活動有關。春季沒有登革熱確診案例，也不排斥有沒紀錄到資料。

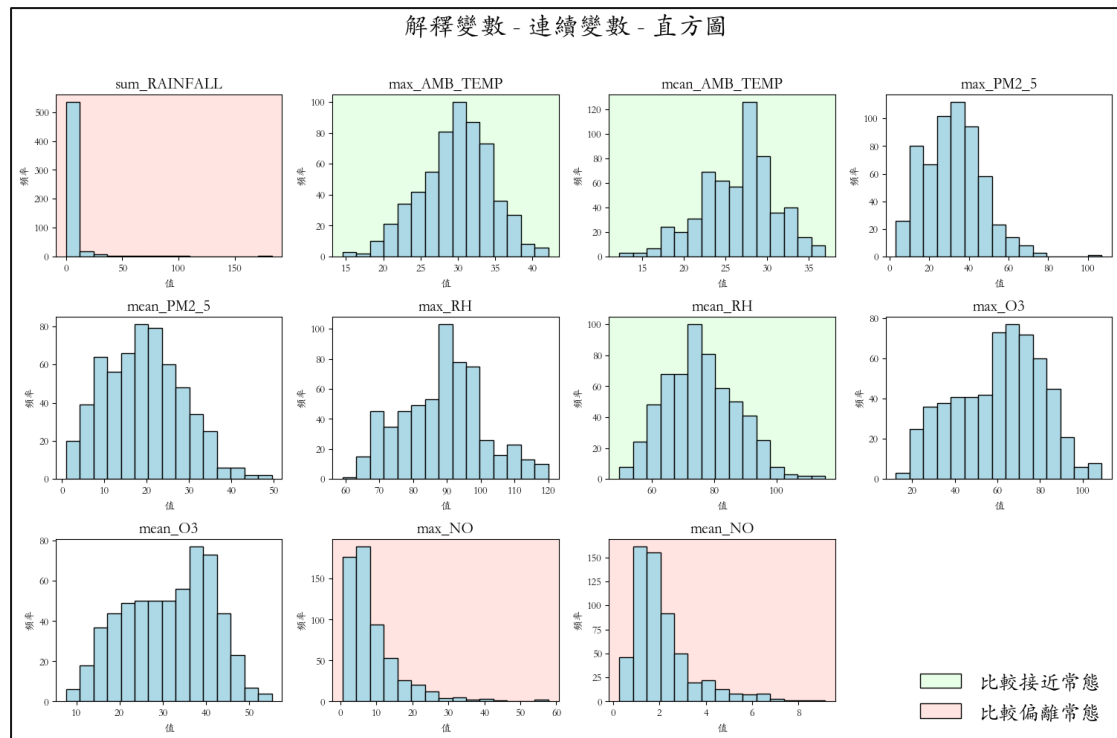


從圖中可以看出，兩組數據的分布沒有明顯的區隔，許多變數的數據點在不同組之間有較大的重疊。因此，目前並不需要進行分組來建立模型，可以直接針對整體數據進行分析或建模。

參、解釋變數 - 連續變數



從這張圖中可以看出當日最高溫度、當日平均溫度、當日最高懸浮微粒 PM2.5 藍點與紅線較為吻合可以判斷它們分佈比較接近常態，後續可以挑選這些變數做為主要的解釋變數；當日總降雨量大概可以看出他可能有很多極端值；當日最高一氧化氮濃度、當日平均一氧化氮濃度可以看出比較偏離紅線，後續選擇主要變數可以避免這些變數。



從這張圖中可以看出當日最高溫度、當日平均溫度、平均相對濕度，是比較明顯的單峰且比較對稱，後續可以選擇他做為主要的解釋變數；而其他變數（如 當日總降雨量、一氧化氮濃度相關變數）分布偏態或存在極端值，需進一步變換或處理以提升模型的穩定性和解釋能力。

變數	平均值	最小值	最大值	單位	符合常態
sum_RAINFALL 當日總降雨量	5.35	0	211	mm	不符合
max_AMB_TEMP 當日最高溫度	29.72	14.56	42.12	°C	不符合
mean_AMB_TEMP 當日平均溫度	26.52	12.25	36.94	°C	不符合
max_PM2_5 當日最高懸浮微粒 PM2.5	32.29	3.2	107	$\mu\text{g}/\text{m}^3$	不符合
mean_PM2_5 當日平均懸浮微粒 PM2.5	19.42	0.95	49.65	$\mu\text{g}/\text{m}^3$	不符合
max_RH 當日最高相對濕度	88.57	59	120	%	不符合
mean_RH 當日平均相對濕度	75.76	49.53	115.6	%	不符合
max_O3 當日最高臭氧濃度	60.9	12.3	108.9	ppb	不符合
mean_O3 當日平均臭氧濃度	31.58	7.61	55.31	ppb	不符合
max_NO 當日最高一氧化氮濃度	9.06	0.6	57.8	ppb	不符合
mean_NO 當日平均一氧化氮濃度	2.23	0.26	9.15	ppb	不符合

第二節 假設檢定

壹、卡方檢定 (地區 * 季節)

這裡我們針對地區以及季節去做卡方檢定看他彼此之間是否獨立。假設檢定如下：

H_0 ：地區、季節兩變數是獨立的

H_1 ：地區、季節兩變數不是獨立的

FREQ 程序						area * season 之表格的統計值			
次數 百分比 列百分比 欄百分比	area * season的表格					統計值	DF	值	機率
	area	season							
		冬季	秋季	夏季	總計				
		北高雄	43	176	28	247			
			6.25	25.58	4.07	35.90			
17.41	71.26		11.34						
30.50	36.74		41.18						
南高雄	98	303	40	441					
	14.24	44.04	5.81	64.10					
	22.22	68.71	9.07						
	69.50	63.26	58.82						
總計	141	479	68	688					
	20.49	69.62	9.88	100.00					

統計值	DF	值	機率
卡方	2	2.7597	0.2516
概度比卡方	2	2.7886	0.2480
Mantel-Haenszel 卡方	1	2.7084	0.0998
Phi 係數		0.0633	
列聯係數		0.0632	
Cramer V		0.0633	

這裡可以看到佔比最高的是秋季的南高雄，佔了比較大的部分。另外可以看到卡方檢定 p value 大於 0.05 不拒絕虛無假設，這裡我們沒有足夠證據證明他們有關聯性。

貳、 Two sample t test (地區)

TTEST 程序				
變數: y				
方法	變異數	DF	t 值	Pr > t
集區	均等	586	-7.53	<.0001
Satterthwaite	不均等	504.16	-8.50	<.0001

變異數相等性				
方法	分子自由度	分母自由度	F 值	Pr > F
Folded F	345	241	5.28	<.0001

1. F 檢定

這裡我們針對 F 檢定的假設檢定如下：

$$H_0: \text{兩組(北高雄、南高雄)變異數相等, } \sigma_1^2 = \sigma_2^2$$

$$H_1: \text{兩組(北高雄、南高雄)變異數不相等, } \sigma_1^2 \neq \sigma_2^2$$

可以看到變異數相等性 F 檢定中 p value < 0.0001 小於 0.05 拒絕虛無假設，說明兩組變異數在統計上不相等。

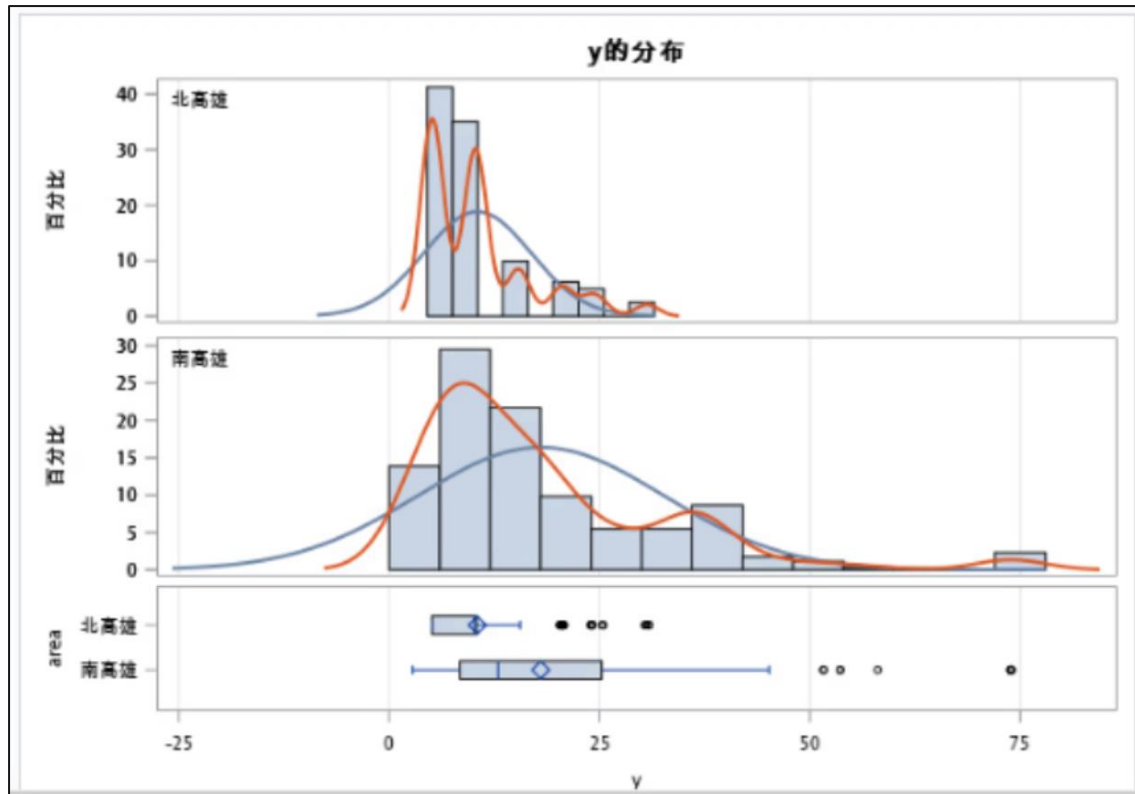
2. T 檢定

這裡我們針對 T 檢定的假設檢定如下：

$$H_0: \text{兩組(北高雄、南高雄)}\mu \text{ 相等, } \mu_1 = \mu_2$$

$$H_1: \text{兩組(北高雄、南高雄)}\mu \text{ 不相等, } \mu_1 \neq \mu_2$$

因此上面變異數不相等，因此 t 檢定結果查看 Satterthwaite t test 的 p value 小於 0.05 拒絕虛無假設，兩組(北高雄、南高雄) μ 在統計上存在顯著差異。



從上圖兩個變異數以及 μ 都有顯著差異，再次驗證了剛剛的結果，南高雄可以看的出來核密度估計曲線尾部較有拉長的趨勢，雖然部分數據都集中在中間，但尾部異常值較多且數值偏大。箱型圖中可以看出來南高雄的箱子比較大，數據範圍也較寬整理分佈是比較分散的。

參、 Anova (季節)

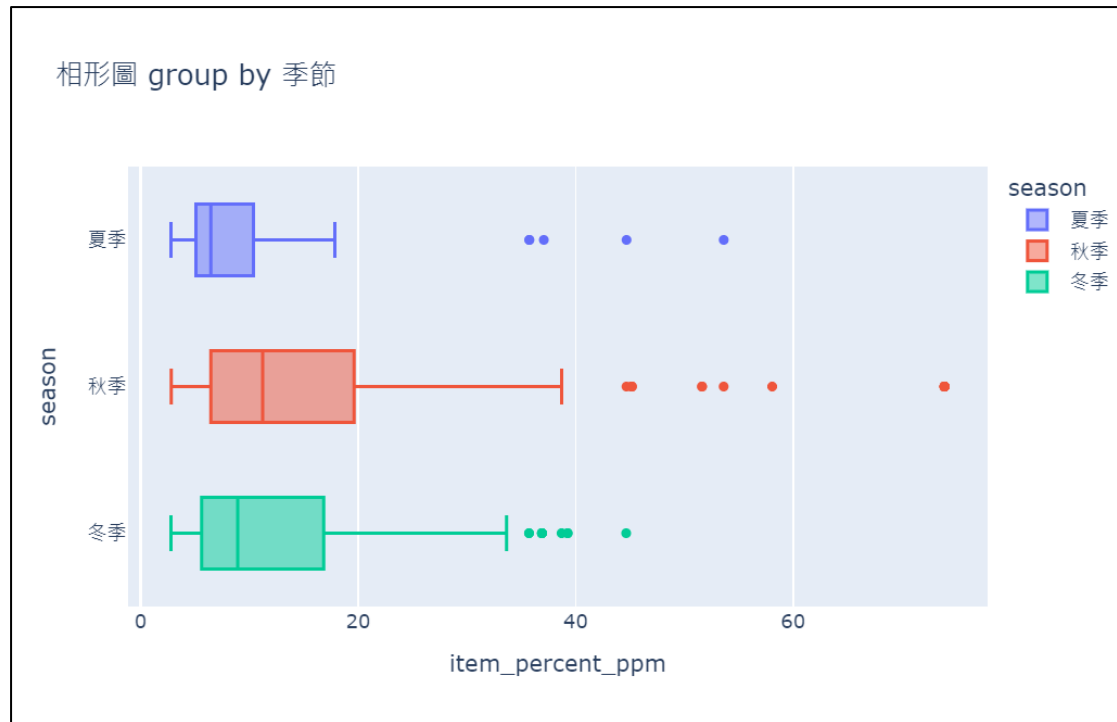
這裡我們針對季節做 anova 變異數檢定。

H_0 ：每組(秋季、冬季、夏季) μ 無顯著差異

H_1 ：每組(秋季、冬季、夏季)至少有一組 μ 存在顯著差異

ANOVA 程序					
應變數: y					
來源	DF	平方和	均方	F 值	Pr > F
模型	2	2447.97592	1223.98796	8.05	0.0004
誤差	585	88918.09291	151.99674		
已校正的總計	587	91366.06884			

結果顯示 P value = 0.0004 小於 0.05，拒絕虛無假設，表示在統計上組別之間至少有一組的 μ 不相等，後續也可以進一步探討每兩組之間的 μ 是否存在顯著差異。



此圖為用以季節作分組的箱型圖，可以看得出來它們彼此之間的變異數、 μ 都有顯著差異。

第三節 相關係數檢定

這裡我們針對所有連續的解釋變數對反應變數做相關係數檢定，假設檢定如下：

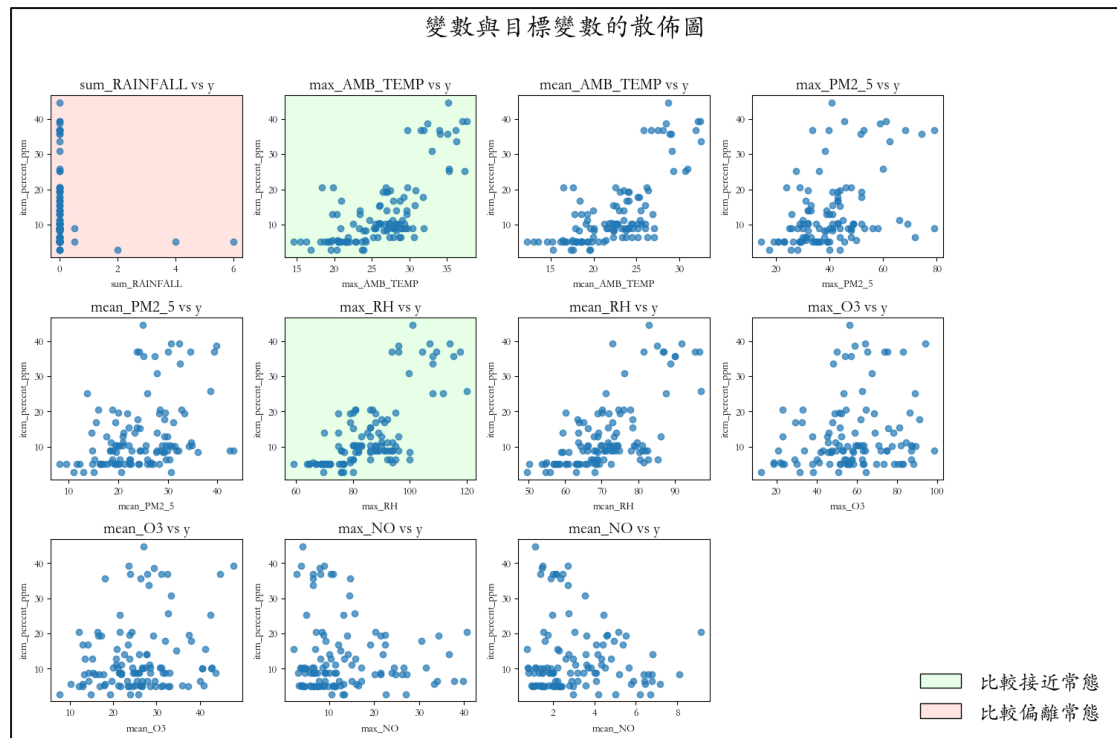
H_0 ：該解釋變數與反應變數 y 之間相關性等於 $Rho0(0.5)$

H_1 ：該解釋變數與反應變數 y 之間偏離 $Rho0(0.5)$

Pearson 相關統計值 (Fisher z 轉換)										
變數	常變數	N	樣本相關	Fisher z	偏差調整	相關估計值	95% 信賴界限		H0: Rho=Rho0	p 值
sum_RAINFALL	y	588	-0.12029	-0.12088	-0.0001025	-0.12019	-0.199114	-0.039720	0.50000	<.0001
max_AMB_TEMP	y	585	0.57707	0.65805	0.0004941	0.57674	0.519983	0.628421	0.50000	0.0090
mean_AMB_TEMP	y	585	0.54287	0.60822	0.0004648	0.54254	0.482710	0.597338	0.50000	0.1583
max_PM2_5	y	588	0.40060	0.42437	0.0003412	0.40032	0.330147	0.466089	0.50000	0.0024
mean_PM2_5	y	588	0.39135	0.41339	0.0003333	0.39106	0.320335	0.457456	0.50000	0.0010
max_RH	y	587	0.58958	0.67702	0.0005031	0.58925	0.533775	0.639671	0.50000	0.0021
mean_RH	y	587	0.48545	0.53009	0.0004142	0.48513	0.420722	0.544674	0.50000	0.6350
max_O3	y	588	0.27048	0.27739	0.0002304	0.27027	0.193645	0.343619	0.50000	<.0001
mean_O3	y	588	0.32026	0.33194	0.0002728	0.32002	0.245513	0.390764	0.50000	<.0001
max_NO	y	588	-0.08569	-0.08590	-0.0000730	-0.08561	-0.165326	-0.004788	0.50000	<.0001
mean_NO	y	588	-0.12297	-0.12359	-0.0001047	-0.12286	-0.201716	-0.042426	0.50000	<.0001

其中可以看到總雨量、當日最高一氧化氮濃度、當日平均一氧化氮濃度 為低度負相關；當日平均懸浮微粒 PM2.5、當日最高臭氧濃度、當日平均臭氧濃度為低度正相關；當日最高溫度、當日平均溫度、當日最高懸浮微粒 PM2.5、當日最高相對濕度、當日平均相對濕度為中度正相關。

從檢定結果可以看出來除了當日最高相對濕度、當日平均溫度外變數 p value 都小於 0.05，拒絕虛無假設，說明在統計上相關性顯著偏離 $Rho0(0.5)$ ，而當日最高相對濕度、當日平均溫度 p value 大於 0.05，不拒絕虛無假設，在統計上這裡沒有足夠證據證明相關性顯著偏離 $Rho0(0.5)$ 。

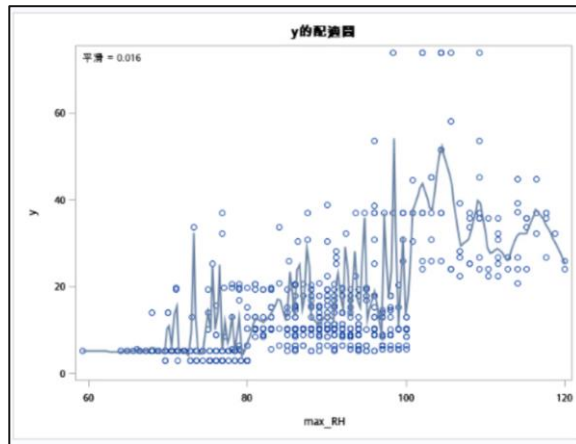


參照前面相關係數以及散佈圖的結果，後續會挑選相關性較高的當日最高溫度、最高相對濕度去建立模型，這兩個變數大約為 0.57~0.58，接近高度相關，做為主要解釋變數。

第四節 簡單線性迴歸

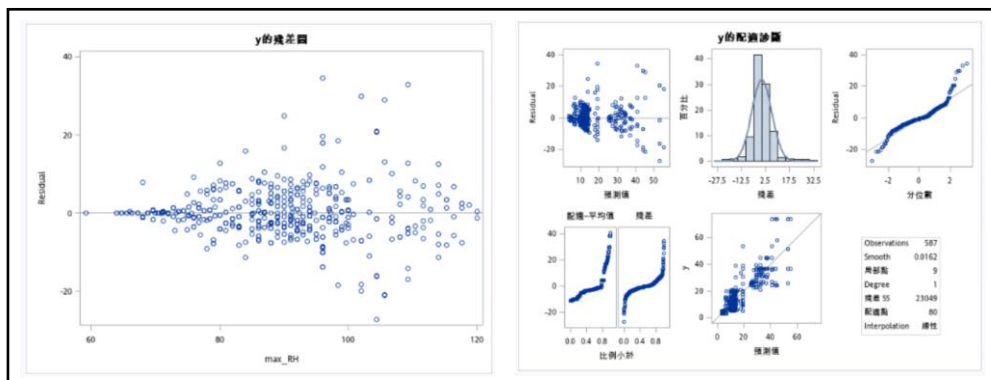
壹、 最高相對濕度*登革熱確診 ppm

甲、 檢查線性關係



從圖中曲線反應主要趨勢，從此圖可以看得出來解釋變數對反應變數非線性關係 Y 且波動非常的大，趨勢不穩定，目前還是針對探索性的分析，因此這邊決定直接查看 LOESS 模型進行非線性關係的分析。

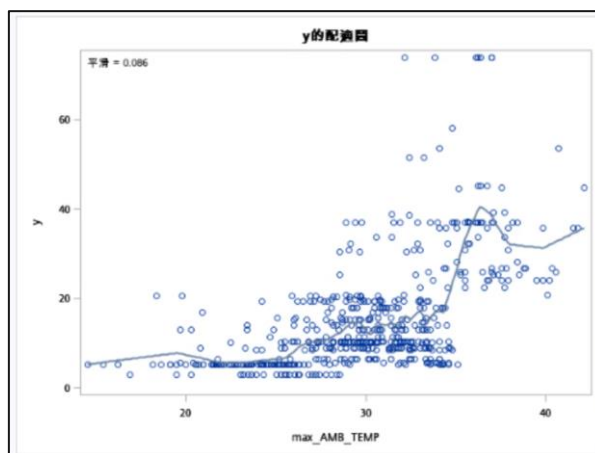
乙、 無母數回歸 LOESS



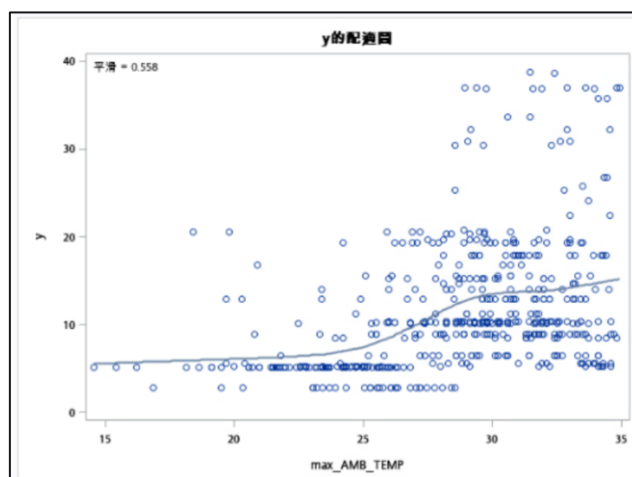
從殘差圖中可以看出殘差大約都分佈在 0 附近，隨著最高相對濕度增加殘差的變異也逐漸變大。從配適診斷圖中可以看出這裡的配適度適存在一定誤差的。

貳、 當日最高溫度*登革熱確診 ppm

甲、 觀察線性關係

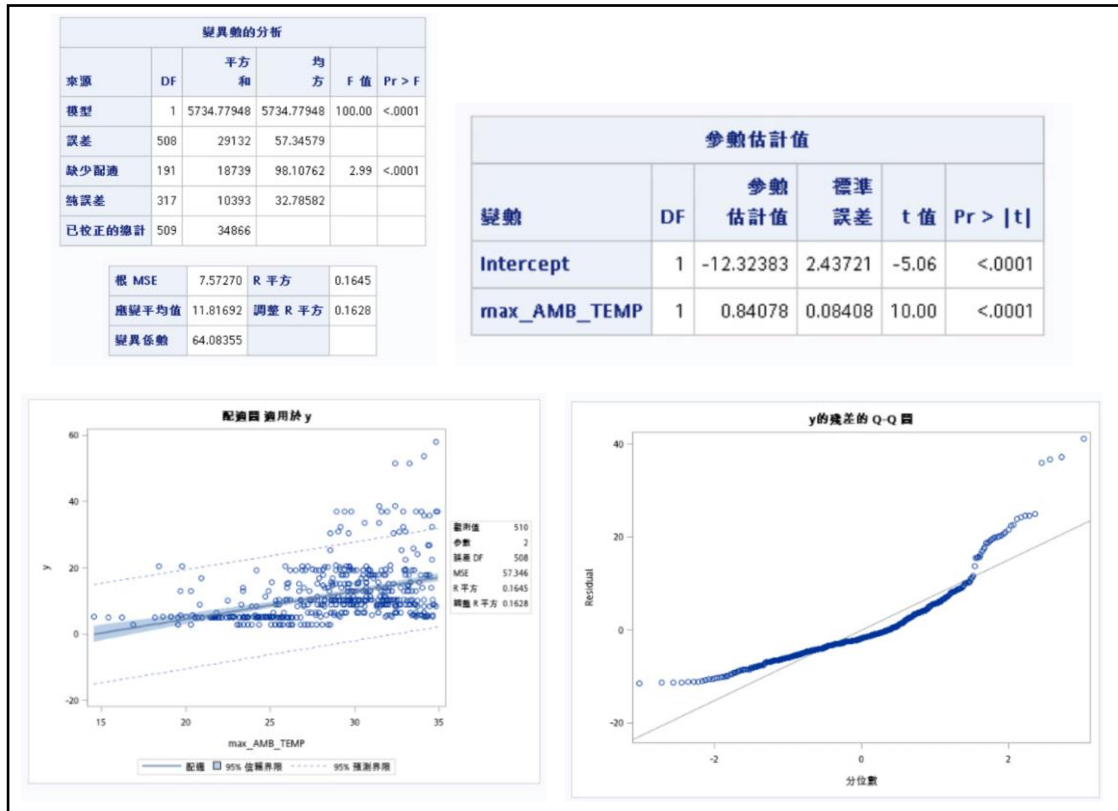


從圖中曲線反應主要趨勢，從此圖可以看得出來解釋變數對反應變數非線性關係 Y 仍然還是存在一些過高的數值，後續考慮把大於 60 的數值拿掉或是直接進入到 LOESS 模型，這裡的趨勢較剛剛平滑，因此這裡我決定刪除登革熱確診 ppm 大於 60 的數值以及當日最高溫度大於 35 的數值。



經過刪除之後有比較平滑一點，但仍然還是不符合線性，後續先把他帶入線性回歸裡面看看狀況。

乙、 回歸結果



- 針對模型整體顯著性虛無假設如下：

H_0 ：模型中解釋變數對反應變數沒有影響

H_1 ：模型中解釋變數對反應變數有影響

此處因為只有放入一個解釋變數，因此可以一起看 F 檢定以及 T 檢定的結果，就結果可以看出來 p value < 0.0001 小於 0.05，拒絕 H_0 ，說明模型具有統計上的顯著性。

- 針對配適度建立虛無假設：

H_0 ：模型沒有缺少配適

H_1 ：模型存在缺少配飾

從結果可看出 p value < 0.0001 小於 0.05，拒絕 H_0 ，說明這裡配適不足。

- 得到線性回歸方程式：

$$\hat{Y} = -12.3238 + 0.8408X$$

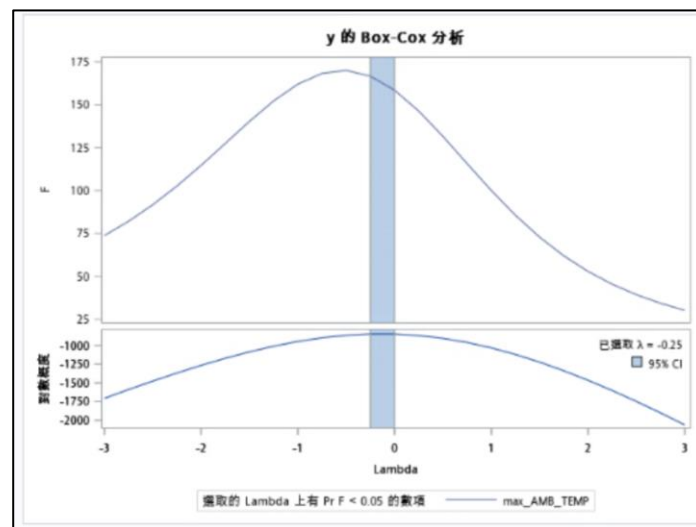
- R 平方也很小，因此需要針對解釋變數或是反應變數進行轉換。
- 從 y 的殘差 qq 圖也可以看的出來大部分沿著對角線分佈，但在兩端明顯有偏離，希望後續可以處理這個偏離的問題。

丙、常態性檢定和常數變異數檢定

UNIVARIATE 程序 變數: res (殘差)					Brown Forsythe Test using F test, in GLM						The modified Breusch-Pagan test					
常態性檢定					GLM 程序						MODEL 程序					
檢定	統計值	Pr < W	Pr > D	p 值	res 變異數均齊性的 Brown 和 Forsythe 檢定 來自群組中位數之絕對差的 ANOVA						不等變異性檢定					
Shapiro-Wilk	W	0.848242	Pr > D	<0.0001	來源	DF	平方和	均方	F 值	Pr > F	方程式	檢定	統計值	DF	Pr > ChiSq	變數
Kolmogorov-Smirnov	D	0.147936	Pr > D	<0.0100	組	1	679.2	679.2	19.39	<.0001	y	Breusch-Pagan	25.94	1	<.0001	max_AMB_TEMP, 1
Cramer-von Mises	W-Sq	3.321426	Pr > W-Sq	<0.0050	殘差	508	17797.0	35.0334								
Anderson-Darling	A-Sq	19.14829	Pr > A-Sq	<0.0050												

從此圖可以看出他拒絕他是常態，兩群的變異數是不一樣的，且他是有線性關係的。

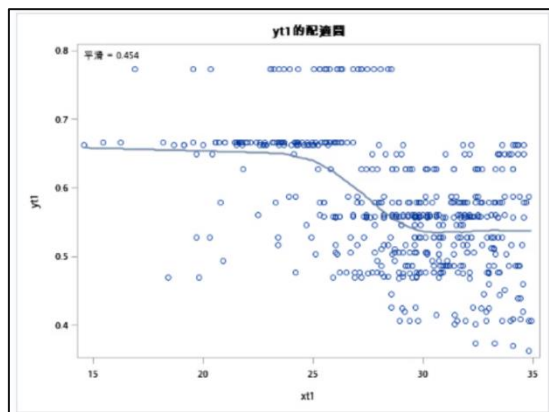
丁、轉換準備



後續為了改善模型，針對 Box-cox 的結果對 Y 取 $\lambda = -0.25$ 做運算，希望可以改善模型的配適度。

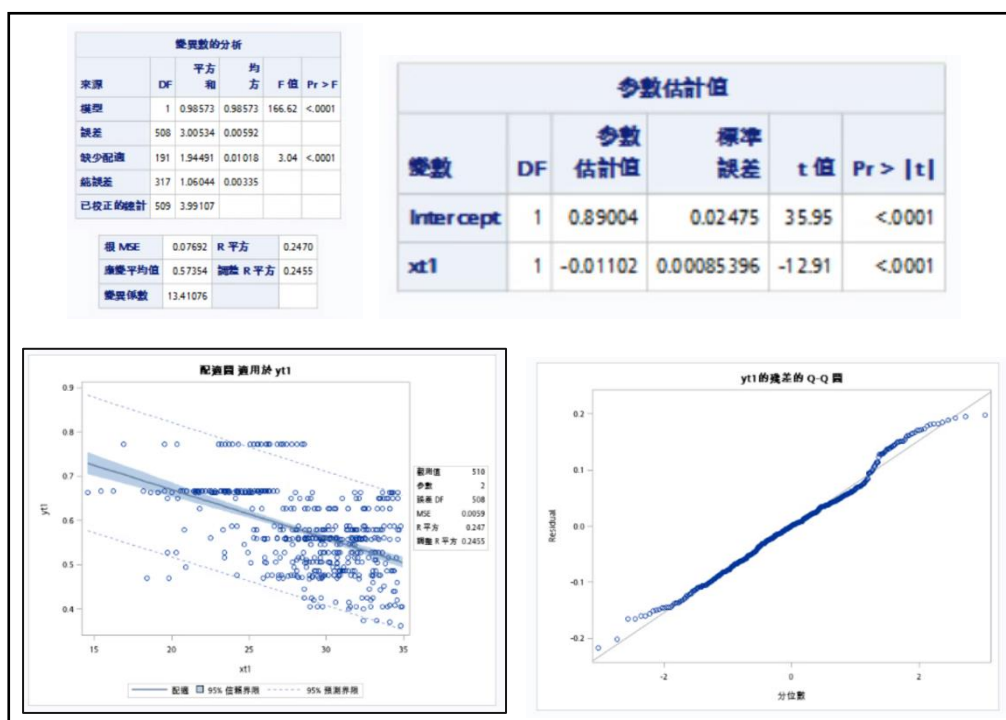
參、(當日最高溫度)*(登革熱確診 ppm** -0.25)

甲、觀察線性關係



看起來分佈不均勻。

乙、回歸結果



- 針對模型整體顯著性虛無假設如下：

H_0 ：模型中解釋變數對反應變數沒有影響

H_1 ：模型中解釋變數對反應變數有影響

此處因為只有放入一個解釋變數，因此可以一起看 F 檢定以及 T 檢定的結果，就結果可以看出來 $p \text{ value} < 0.0001$ 小於 0.05，拒絕 H_0 ，說明模型具有統計上的顯著性。

- 針對配適度建立虛無假設：

H_0 ：沒有缺少配適

H_1 ：存在缺少配飾

從結果可看出 $p \text{ value} < 0.0001$ 小於 0.05，拒絕 H_0 ，說明這裡仍然配適不足。

- R 平方雖然還是很小，但是相較剛剛有提升一些。
- 根 MSE 預測誤差有再減少。
- 得到線性回歸方程式：

$$\hat{Y} = 0.8900 - 0.01102X$$

- 從 y 的殘差 qq plot 也可以看的出來大部分沿著對角線分佈，但在兩端偏離有縮小，更接近常態。

丙、常態性檢定和常數變異數檢定

UNIVARIATE 敘序 變數: res (殘差)					Brown Forsythe Test using F test, in GLM					The modified Breusch-Pagan test						
常態性檢定					GLM 敘序					MODEL 敘序						
檢定	統計值	p 值			res 變異數均齊性的 Brown 和 Forsythe 檢定 來自群組中位數之絕對差的 ANOVA					不等變異性檢定						
Shapiro-Wilk	W	0.989568	Pr < W	0.0011	來源	DF	平方和	均方	F 值	Pr > F	方程式	檢定	統計值	DF	Pr > ChiSq	變數
Kolmogorov-Smirnov	D	0.042569	Pr > D	0.0236	gp	1	0.00684	0.00684	2.99	0.0845	y	Breusch-Pagan	25.94	1	< 0.001	xt1, 1
Cramer-von Mises	W-Sq	0.160918	Pr > W-Sq	0.0185	誤差	508	1.1637	0.00229								
Anderson-Darling	A-Sq	1.355619	Pr > A-Sq	< 0.0050												

從此圖可以看出他拒絕他是常態，沒有足夠證據證明兩群的變異數有差異，且他是有線性關係的。

第五節 多變量回歸分析

在這個章節中，我會以理論性來初步挑選解釋性強的模型、驗證性以評估指標（如 MSE、PRESS）來篩選出更具穩定性模型最後再以自動化的 Stepwise 挑選模型，全面性的篩選模型。

壹、挑選變數

甲、初步篩選

基於擬合性，根據 Cp、AIC、 R^2 、MSE 挑選變數。

各指標前 5 名模型 (ModelIndex)				
rank	Cp 越小排名越前	AIC 越小排名越前	R-Square 越大排名越前	MSE 越小排名越前
1	67	67	78	78
2	78	78	67	67
3	79	79	89	89
4	80	80	90	90
5	81	81	91	91

就此圖來看我可以從中挑選出 model Index 為 67、78、79、89、80、90 等來做下一步的模型。

後續挑選出在特定指標（AIC、SBC、PRESS）中表現最好的前 5 名模型條列出來，作為候選模型進一步分析的基礎，這個表中呈現了適配度（AIC 和 SBC 越小越好）和預測能力（PRESS 越小越好）。

前 5 名模型挑選結果														
類別	modelindex	p	sse	mse	rsquare	adjrsq	cp	aic	sbc	press	varianmodel	NumInModel	aicrank	sbcrank
1	78	9	101.70874	0.20342	0.4831	0.4748	7.2262	-801.6804	-763.56837	105.587	max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3 max_NO	8	2	10
2	67	8	101.96257	0.20356	0.4817	0.4745	6.4703	-802.2819	-758.43229	105.375	max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3 max_NO	7	1	2
6	79	9	101.92195	0.20384	0.4820	0.4737	8.1734	-800.5945	-762.50248	105.992	max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3 max_NO	8	3	12
7	80	9	101.93079	0.20386	0.4820	0.4737	8.2167	-800.5504	-762.45834	105.726	max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3 max_NO mean_NO	8	4	13
8	81	9	101.93583	0.20387	0.4820	0.4737	8.2419	-800.5247	-762.43265	105.693	sum_RAINFALL max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3 max_NO	8	5	14
18	88	8	102.42639	0.20444	0.4795	0.4722	8.6440	-800.0815	-766.22194	105.793	max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3 max_NO	7	8	4
23	56	7	102.79534	0.20477	0.4776	0.4713	8.4509	-800.2514	-770.62425	105.706	max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3	6	6	1
45	45	6	104.66013	0.20907	0.46581	0.4628	15.5840	-793.1904	-767.70575	107.296	max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3	5	36	3

根據結果，每個模型的變數數量較多，可能導致模型過於複雜，增加解釋和實施的難度。建議進一步簡化模型，篩選出對模型影響最大的關鍵變數，以平衡模型的解釋力和簡約性。此外，可以透過檢查變數的共線性或應用變數選擇方法來優化模型結構，從而提升模型的實用性和穩健性。

乙、逐步篩選變數

逐步選擇 的摘要									
步驟	已輸入變數	已移除變數	標籤	變數數目	偏 R 平方	模型 R 平方	C(p)	F 值	Pr > F
1	max_RH		當日最高相對濕度	1	0.2663	0.2663	202.063	184.03	<.0001
2	mean_PM2_5		當日平均懸浮微粒PM2.5	2	0.1358	0.4021	73.1738	114.95	<.0001
3	max_AMB_TEMP		當日最高溫度	3	0.0419	0.4441	34.7469	38.11	<.0001
4	mean_O3		當日平均臭氧濃度	4	0.0072	0.4512	29.8357	6.59	0.0106
5	mean_AMB_TEMP		當日平均溫度	5	0.0078	0.4590	24.3181	7.25	0.0073
6	mean_RH		當日平均相對濕度	6	0.0114	0.4705	15.3041	10.83	0.0011
7	max_O3		當日最高臭氧濃度	7	0.0086	0.4790	9.0519	8.23	0.0043
8		max_RH	當日最高相對濕度	6	0.0015	0.4776	8.4509	1.40	0.2379
9	max_NO		當日最高一氧化氮濃度	7	0.0041	0.4817	6.4703	3.99	0.0462

由此可以看出來模型的 R 平方從 0.2663 提升到 0.4817，顯示這些變數在解釋模型中的變異時非常重要。因此藉由 STEPWISE 可以看出來我們留下了 max_AMB_TEMP、mean_AMB_TEMP、mean_PM2_5、mean_RH、max_O3、mean_O3、max_NO 這些變數。由此處會刪除 max_RH。

丙、前進選擇結果篩選變數

前進選擇 的摘要								
步驟	已輸入變數	標籤	變數數目	偏 R 平方	模型 R 平方	C(p)	F 值	Pr > F
1	max_RH	當日最高相對濕度	1	0.2663	0.2663	202.063	184.03	<.0001
2	mean_PM2_5	當日平均懸浮微粒PM2.5	2	0.1358	0.4021	73.1738	114.95	<.0001
3	max_AMB_TEMP	當日最高溫度	3	0.0419	0.4441	34.7469	38.11	<.0001
4	mean_O3	當日平均臭氧濃度	4	0.0072	0.4512	29.8357	6.59	0.0106
5	mean_AMB_TEMP	當日平均溫度	5	0.0078	0.4590	24.3181	7.25	0.0073
6	mean_RH	當日平均相對濕度	6	0.0114	0.4705	15.3041	10.83	0.0011
7	max_O3	當日最高臭氧濃度	7	0.0086	0.4790	9.0519	8.23	0.0043
8	max_NO	當日最高一氧化氮濃度	8	0.0041	0.4831	7.1292	3.94	0.0478

前向選擇結果表明，8 個變數均對模型具有統計意義，特別是 max_RH 和 mean_PM2_5 是解釋目標變數變異的重要因子。模型的 R 平方達到 0.4831，Cp 值接近變數數量，表明模型的適配性和簡潔性均較好。下一步需要進行模型的驗證和穩定性檢查，以確認其泛化能力。

丁、向後消去結果篩選變數

已移除變數 max_RH: R 平方 = 0.4817、C(p) = 6.4703					
變異數的分析					
來源	DF	平方和	均方	F 值	Pr > F
模型	7	94.78803	13.54115	66.52	<.0001
誤差	501	101.98257	0.20356		
已校正的總計	508	196.77061			

變數	參數估計值	標準誤差	類型 II SS	F 值	Pr > F
Intercept	-1.46604	0.20068	10.86414	53.37	<.0001
max_AMB_TEMP	0.15757	0.01979	12.90604	63.40	<.0001
mean_AMB_TEMP	-0.12431	0.02258	6.16979	30.31	<.0001
mean_PM2_5	0.02486	0.00295	14.47481	71.11	<.0001
mean_RH	0.02218	0.00298	11.27471	55.39	<.0001
max_O3	-0.00728	0.00225	2.12836	10.46	0.0013
mean_O3	0.02226	0.00442	5.17406	25.42	<.0001
max_NO	0.00578	0.00289	0.81276	3.99	0.0462

條件編碼的界限: 18.533, 347.46

留在模型中的所有變數在 0.1500 層級上都是顯著的。

向後消去的摘要								
步驟	已移除變數	標籤	變數數目	偏 R 平方	模型 R 平方	C(p)	F 值	Pr > F
1	sum_RAINFALL	當日總降雨量	10	0.0003	0.4840	10.2949	0.29	0.5874
2	max_PM2_5	當日最高懸浮微粒PM2.5	9	0.0004	0.4836	8.6726	0.38	0.5388
3	mean_NO	當日平均一氧化氮濃度 run	8	0.0005	0.4831	7.1292	0.46	0.4990
4	max_RH	當日最高相對濕度	7	0.0014	0.4817	6.4703	1.35	0.2465

由此圖可以看出我這裡要保留的變數 Pvalue < 0.05 表示對模型有顯著的影響為 max_AMB_TEMP、mean_AMB_TEMP、mean_PM2_5、mean_RH、max_O3、mean_O3、max_NO，其餘變數對目標變數幾乎沒有解釋力。

戊、小結

value	逐步選擇	前進選擇	向後消去
當日總降雨量 sum_RAINFALL			X
當日最高溫度 max_AMB_TEMP	V	V	
當日平均溫度 mean_AMB_TEMP	V	V	
當日最高懸浮微粒 PM2.5 max_PM2_5			X
當日平均懸浮微粒 PM2.5 mean_PM2_5	V	V	
當日最高相對濕度 max_RH	X	V	X
當日平均相對濕度 mean_RH	V	V	
當日最高臭氧濃度 max_O3		V	
當日平均臭氧濃度 mean_O3	V	V	
當日最高一氧化氮濃度 max_NO	V	V	
當日平均一氧化氮濃度 mean_NO			X

後續仍針對前面的結果，因此將會刪除含有 sum_RAINFALL、max_PM2_5、mean_NO、max_RH 的模型進行後續的訓練以及分析。

前 5 名模型挑選結果

varsinmodel
max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 max_RH mean_RH max_O3 mean_O3 max_NO
max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3 max_NO
max_AMB_TEMP mean_AMB_TEMP max_PM2_5 mean_PM2_5 mean_RH max_O3 mean_O3 max_NO
max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3 max_NO mean_NO
sum_RAINFALL max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3 max_NO
max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3 mean_NO
max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH max_O3 mean_O3
max_AMB_TEMP mean_AMB_TEMP mean_PM2_5 mean_RH mean_O3

貳、 訓練模型結果選擇與比較

MODEL	_TYPE_	_DEPVAR_	_RMSE_	_PRESS_	Intercept	max_AMB_TEMP	mean_AMB_TEMP	mean_PM2_5	mean_RH	mean_O3	Iny	max_O3	max_NO
MODEL1	PARMS	Iny	0.46017	51.5343	-1.14503	0.14645	-0.12605	0.022328	0.023492	0.010535	-1	.	.
MODEL1	SEB	Iny	0.46017	.	0.29668	0.02872	0.03366	0.004103	0.004578	0.004005	-1	.	.
MODEL2	PARMS	Iny	0.45546	50.8208	-1.11693	0.17309	-0.15206	0.025895	0.022546	0.022512	-1	-.008321871	.
MODEL2	SEB	Iny	0.45546	.	0.29387	0.03050	0.03502	0.004322	0.004548	0.006357	-1	0.003453150	.
MODEL3	PARMS	Iny	0.45611	51.0754	-1.07692	0.17419	-0.15326	0.026391	0.022435	0.021390	-1	-.008054322	-.002591888
MODEL3	SEB	Iny	0.45611	.	0.30209	0.03060	0.03513	0.004410	0.004559	0.006647	-1	0.003488041	0.004418966

由此處我們可以得到預測公式，後續我們將再把他帶入後續的模型中。

觀測值	模型編號	統計值類型	截距	均方誤差	Press 統計值	截距	當日最高溫度	當日平均溫度	當日平均懸浮微粒PM2.5	當日平均相對濕度	當日平均氣溫	Iny	當日最高氣溫	當日最高一氧化氮濃度	模型中的線性因子數	模型中的參數數	誤差自由項	平方和誤差	均方誤差	R 平方	調整 R 平方
1	MODEL1	PARMS	Iny	0.46017	51.5343	-1.14503	0.14645	-0.12605	0.022328	0.023492	0.010535	-1	.	.	5	6	231	48.9168	0.21176	0.44246	0.43040
2	MODEL1	SEB	Iny	0.46017	.	0.29668	0.02872	0.03366	0.004103	0.004578	0.004005	-1
3	MODEL2	PARMS	Iny	0.45546	50.8208	-1.11693	0.17309	-0.15206	0.025895	0.022546	0.022512	-1	-.008321871	.	6	7	230	47.7120	0.20744	0.45620	0.44201
4	MODEL2	SEB	Iny	0.45546	.	0.29387	0.03050	0.03502	0.004322	0.004548	0.006357	-1	0.003453150
5	MODEL3	PARMS	Iny	0.45611	51.0754	-1.07692	0.17419	-0.15326	0.026391	0.022435	0.021390	-1	-.008054322	-.002591888	7	8	229	47.6404	0.20804	0.45701	0.44041
6	MODEL3	SEB	Iny	0.45611	.	0.30209	0.03060	0.03513	0.004410	0.004559	0.006647	-1	0.003488041	0.004418966

由這裡面可以看出 R 平方大多都落在 0.45 多左右，顯示這些模型的擬合效果相對穩定，並且能夠解釋一定比例的目標變數變異。

參、 帶入測試資料

觀測值	模型編號	統計值類型	截距	當日最高溫度	當日平均溫度	當日平均懸浮微粒PM2.5	當日平均相對濕度	當日最高氣溫	當日平均氣溫	當日最高一氧化氮濃度	平方和誤差	Press 統計值	均方誤差	C(p)	mspr	調整 R 平方	R 平方
1	MODEL1	PARMS	-1.14503	0.14645	-0.12605	0.022328	0.023492	.	0.010535	.	48.9168	51.5343	0.21176	8.0000	0.20921	0.43040	0.44246
2	MODEL1	SEB	0.29668	0.02872	0.03366	0.004103	0.004578	.	0.004005	8.0000	0.20921	.	.
3	MODEL1c	PARMS	-1.58476	0.14280	-0.10165	0.023188	0.022215	.	0.008798	.	54.8065	57.3731	0.20604	7.0924	.	0.48393	0.49345
4	MODEL1c	SEB	0.26221	0.02555	0.02927	0.003599	0.003968	.	0.003520	7.0924	.	.	.
5	MODEL2	PARMS	-1.11693	0.17309	-0.15206	0.025895	0.022546	-.008321871	0.022512	.	47.7120	50.8208	0.20744	.	0.20677	0.44201	0.45620
6	MODEL2	SEB	0.29387	0.03050	0.03502	0.004322	0.004548	0.003453150	0.006357	0.20677	.	.
7	MODEL2c	PARMS	-1.55161	0.15494	-0.11296	0.025884	0.021318	-0.05332078	0.016772	.	54.1563	57.0668	0.20436	16.6827	.	0.48813	0.49946
8	MODEL2c	SEB	0.26180	0.02634	0.02984	0.003890	0.003984	0.002989411	0.005682	16.6827	.	.	.
9	MODEL3	PARMS	-1.07692	0.17419	-0.15326	0.026391	0.022435	-.008054322	0.021390	-0.002592	47.6404	51.0754	0.20804	.	0.21015	0.44041	0.45701
10	MODEL3	SEB	0.30209	0.03060	0.03513	0.004410	0.004559	0.003488041	0.006647	0.004419	0.21015	.	.
11	MODEL3c	PARMS	-1.81339	0.14675	-0.10360	0.022128	0.022163	-.006317276	0.023259	0.013034	51.8712	55.2858	0.19648	.	.	0.50787	0.52058
12	MODEL3c	SEB	0.26793	0.02594	0.02938	0.003970	0.003914	0.002945399	0.005887	0.003822

三個模型的效果大多差不多，但相比之下模型 3 的預測效果較好，但仍需要進行進一步的優化和調整。可以考慮減少某些變數或者使用不同的參數設定來提高預測準確性。

後續主要可以得到公式：

模型 1

$$\text{yhat1} = -1.14503 + 0.14645 * \text{max_AMB_TEMP} + (-0.12605) * \text{mean_AMB_TEMP} + 0.022328 * \text{mean_PM2_5} + 0.023492 * \text{mean_RH} + 0.010535 * \text{mean_O3}$$

模型 2:

$$\text{yhat2} = -1.11693 + 0.17309 * \text{max_AMB_TEMP} + (-0.15206) * \text{mean_AMB_TEMP} + 0.025895 * \text{mean_PM2_5} + 0.022546 * \text{mean_RH} + 0.022512 * \text{mean_O3} + (-0.008321871) * \text{max_O3}$$

模型 3

$$\text{yhat3} = -1.07692 + 0.17419 * \text{max_AMB_TEMP} + (-0.15326) * \text{mean_AMB_TEMP} + 0.026391 * \text{mean_PM2_5} + 0.022435 * \text{mean_RH} + 0.021390 * \text{mean_O3} + (-0.008054322) * \text{max_O3} + (-0.002591888) * \text{max_NO}$$

後續會選擇 R 平方最高的第三個模型作為主要的模型，R 平方為 0.52 左右。

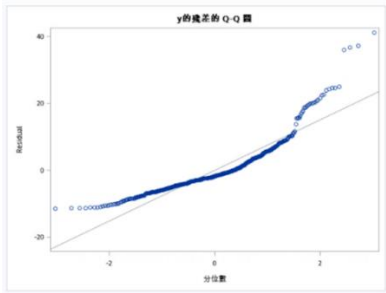
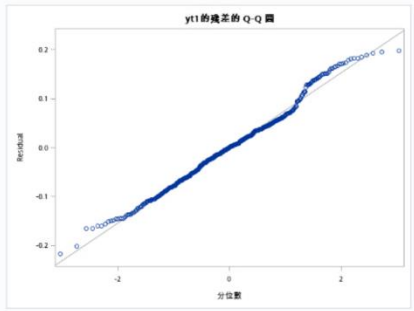
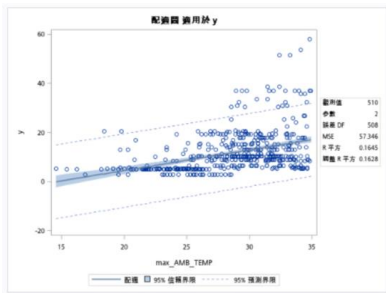
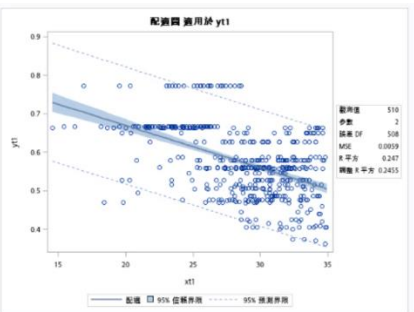
第四章 結論與建議

本章依據研究目的及研究結果，針對分析所得之發現，歸納成以下結論，並提出相關具體建議，以供相關人員參考。

第一節 結論

壹、簡單線性回歸結論

針對線性回歸我們可以得到下表：

	當日最高溫度*登革熱確診 ppm	(當日最高溫度)*(登革熱確診 ppm**-0.25)
F 檢定	< 0.0001	< 0.0001
缺少配適	配適不足	配適不足
R 平方	0.1645	0.2470
方程式	$\hat{Y} = -12.3238 + 0.8408X$	$\hat{Y} = 0.8900 - 0.01102X$
BF/BP	<ol style="list-style-type: none"> 1. 拒絕他是常態 2. 兩群的變異數是不一樣 3. 有線性關係 	<ol style="list-style-type: none"> 1. 拒絕他是常態 2. 沒有足夠證據證明兩群的變異數有差異 3. 有線性關係
殘差圖		 <p>兩端偏離有縮小，更接近常態。</p>
配適圖		

貳、多變量回歸分析結論

value	模型 1	模型 2	模型 3
當日總降雨量 sum_RAINFALL			
當日最高溫度 max_AMB_TEMP	V	V	V
當日平均溫度 mean_AMB_TEMP	V	V	V
當日最高懸浮微粒 PM2.5 max_PM2_5			
當日平均懸浮微粒 PM2.5 mean_PM2_5	V	V	V
當日最高相對濕度 max_RH			
當日平均相對濕度 mean_RH	V	V	V
當日最高臭氧濃度 max_O3		V	V
當日平均臭氧濃度 mean_O3	V	V	V
當日最高一氧化氮濃度 max_NO			V
當日平均一氧化氮濃度 mean_NO			

指標	模型 1	模型 2	模型 3
R-Square (訓練/測試)	0.4304 / 0.4839	0.4420 / 0.4881	0.4404 / 0.5079
預測 R-Square	- / 0.4935	- / 0.4995	- / 0.5206

相較於其他模型，模型 3 更能反映資料特性，具有較好的穩定性與應用價值，因此建議採用模型 3 作為主要分析模型。

模型 3 的預測方程式 $\hat{y}_3 = -1.07692 + 0.17419 \cdot \text{max_AMB_TEMP} + (-0.15326) \cdot \text{mean_AMB_TEMP} + 0.026391 \cdot \text{mean_PM2_5} + 0.022435 \cdot \text{mean_RH} + 0.021390 \cdot \text{mean_O3} + (-0.008054322) \cdot \text{max_O3} + (-0.002591888) \cdot \text{max_NO}$

第二節 建議

從上述可以得到在取對數以及轉換後 R 平方有提升，但仍然不太理想，因此對於後續期末報告有以下的意見。

壹、 修改分組層級

目前模型是以高雄市的行政區作為分組，這能更微觀區域內的細節。然而，未來可以嘗試以縣市為分組基礎，可以獲得比較多變數，變數彼此之間的差異也會比較大，例如教育水平、都市化程度等。雖然可能會犧牲一些微觀層面的細節，但擴大分組層級後，或許更能表示整體趨勢。後續也可以再比較一下兩個模型。

貳、 加入不同變數

加入更多不同的變數，舉例如下：

甲、日照時數：日照時數影響蚊蟲的活動頻率以及活力。

乙、都市化指數：都市化地區可能存在較多的蚊蟲繁殖場所，例如積水的建築工地。

丙、家庭收入：低收入戶家庭可能缺乏足夠的資源防止蚊蟲的孳生以及疾病的傳染。

丁、教育水平：可能影響防疫的意識。