台南地區登革熱資料分析–以北區為例

R Computing for Data Analytics

期末報告

組別：YOLO

成員：蔡耀賢

謝承豫

蕭博修

魏曉晨

一、研究動機

在熱帶氣候的台灣每逢夏季都會遇到的氣候性疾病–登革熱，尤其去年(2015 年)暑假台灣的南部地區登革熱病例數更是突破以往的紀錄，且登革熱往往無法即時控疫情。

在 2015 年的台南北區為登革熱的重大疫區，因此我們想針對台南北區的蚊子密度(以陽性住戶當作指標)、會影響蚊子生長的氣候(降雨量和氣溫)和噴藥次數來與病例數做迴歸分析，探討其中的因果的相關性。

二、資料來源

主要使用了台南市政府的開放資料平台上 2015 年的登革熱相關資料集(http://data.tainan.gov.tw/dataset/dengue-dist)，分別為：

(1). 7~12 月台南市本土登革熱病例數，資料維度包含編號、確診日、區別、里別、道路名稱、緯度座標、經度座標。
在此資料集我們使用了**確診日**和**區別**這兩個維度。

(2). 7~10 月登革熱病媒蚊密度調查結果，資料維度包含日期、區別、里別、調查種類、調查戶數、陽性戶數、調查容器數(戶內)、調查容器數(戶外)、陽性容器數(戶內)、陽性容器數(戶外)、布氏指數、布氏級數、容器指數、容器級數。
在此資料集我們使用了**日期**、**區別**、陽性戶數當作**蚊子密度**這三個維度。

(3). 7~12 月臺南市 104 年登革熱噴藥場次，資料維度包含序號、孳清/噴藥、區域、里別、日期、星期、集合時間、集合地點、緯度、經度、鎖匠、警員(安全維護)、刑警(V8)、噴工數、支援人力、可支援人力。
在此資料集我們使用了**孳清/噴藥**、**日期**和**區別**、這三個維度。

(4). 台南氣象資料來自於中央氣象局網站以及中央氣象局的觀測資料查詢系統，分別拿出了台南 7~12 月的**日均溫**和**日雨量**。
每日雨量資料來源：
http://www.cwb.gov.tw/V7/climate/dailyPrecipitation/dP.htm
每日氣溫資料來源：
http://e-service.cwb.gov.tw/HistoryDataQuery/index.jsp

由於蚊子密度在 10/23 後在資料集裡就無資料(初期在處理時在 R code 上設為 N/A)，觀察資料發現在 10 月開始蚊子密度就呈下降趨勢，因想做限定於蚊子存活的旺季區間，因此我們取 **7/1 到 10/23 的時間作為分析**，其中還有少許無蚊子密度的日期，使用移動平均法(取前三期的平均)來補缺值。

三、模型定義與實作

我們整理並彙整各資料來源後，共分成 14 個維度，如下表：

維度代稱	說明
Month	月份
Day	日期
Date	月份/日期/西元年
Rain	降雨量(mm)
Temperature	溫度(℃)
Illness case	病例數(人)
Density	蚊子密度(陽性住戶幾間)
Spraying	噴藥次數(次)
Illness_cum	累積病例數(人)
t	時間
t^2	時間平方
Spray_cum	累計噴藥次數(次)
Rain_cum	累積降雨量(mm)
Density_cum	累積蚊子密度(累加陽性住戶數)

資料分析將以迴歸分析進行，嘗試建立迴歸模型，以找出變數之間的關聯性以及這些變數的解釋力強度；亦即找尋雨量、氣溫、蚊子密度噴藥次數與時間等因素，跟登革熱確診人數的關係。主要使用函數：lm(formula, data)。

首先，先將整理好的資料集 District A2_d.csv 匯入 R 程式，將各欄位指定相對應的變數名稱。(其中 y.illness 代表病例數、y.illness_cum 代表累積病例數、x.rain 代表降雨量、x.temp 代表氣溫、x.spray 代表噴藥次數、x.density 代表蚊子密度、x.t 代表時間、x.t2 代表時間平方。)

```
District.A2_d <- read.csv("E:/District A2_d.csv")
View(District.A2_d)

y.illness = District.A2_d$Illness.Case
y.illness_cum = District.A2_d$Illness_cum

x.rain = District.A2_d$Rain
x.spray = District.A2_d$Spraying
x.temp = District.A2_d$Temparature
x.density = District.A2_d$Density_movaverage
x.t = District.A2_d$t
x.t2 = District.A2_d$t.2
```

接著，一一建立各迴歸模型，如下表：

模型代號	迴歸模型
Model 1	y.illness~x.rain+x.temp+x.spray+x.density
Model 2	y.illness~x.rain+x.temp+x.spray+x.density+x.t
Model 3	y.illness~x.rain+x.temp+x.spray+x.density+x.t+x.t2
Model 4	y.illness_cum~x.rain+x.temp+x.spray+x.density
Model 5	y.illness_cum~x.rain+x.temp+x.spray+x.density+x.t
Model 6	y.illness_cum~x.rain+x.temp+x.spray+x.density+x.t+x.t2
Model 1_log	log(y.illness+1)~x.rain+x.temp+x.spray+x.density
Model 2_log	log(y.illness+1)~x.rain+x.temp+x.spray+x.density+x.t
Model 3_log	log(y.illness+1)~x.rain+x.temp+x.spray+x.density+x.t+x.t2
Model 4_log	log(y.illness_cum +1)~x.rain+x.temp+x.spray+x.density
Model 5_log	log(y.illness_cum +1)~x.rain+x.temp+x.spray+x.density+x.t
Model 6_log	log(y.illness_cum +1)~x.rain+x.temp+x.spray+x.density+x.t+x.t2
Model 1_1og_s	log(y.illness+1)~x.rain+x.temp+x.density
Model 2_log_s	log(y.illness+1)~x.rain+x.temp+x.density+x.t
Model 3_log_s	log(y.illness+1)~x.rain+x.temp+x.density+x.t+x.t2
Model 4_1og_s	log(y.illness_cum +1)~x.rain+x.temp+x.density
Model 5_log_s	log(y.illness_cum +1)~x.rain+x.temp+x.density+x.t
Model 6_log_s	log(y.illness_cum +1)~x.rain+x.temp+x.density+x.t+x.t2

第一組模型 model1~model6 跑的是線性迴歸，分別看時間 t 與時間平方 t2 的影響力；其中前三個模型是以病例數作為 y 變數，後三個模型則是累積病例數。第二組模型 model 1_log~model6_log 的 x 與 y 變數邏輯都與前組相同，不同的是此六個模型跑的是指數迴歸。最後，第三組模型則是將模型 model 1_log~model6_log 中的 x.spray(噴藥次數)去除，希望看有噴藥與否對病例數的影響。

各模型將以 lm 函數建立迴歸，再以 summary(model)查看迴歸結果。以 model1 為例，程式碼及迴歸結果如下：

```
#model 1
#y = illness, x:{rain,spray,temp,density}
xy1 = data.frame(x.rain,x.spray,x.temp,x.density,y.illness)
model1 = lm(y.illness~x.rain+x.spray+x.temp+x.density,xy1)
summary(model1)
```

```
> summary(model1)

Call:
lm(formula = y.illness ~ x.rain + x.spray + x.temp + x.density,
    data = xy1)

Residuals:
   Min     1Q Median     3Q    Max
-64.89 -24.12 -10.36  23.27 110.93

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -114.42473   86.61432  -1.321 0.189215
x.rain         0.07232    0.17836   0.405 0.685913
x.spray       13.18288    2.99282   4.405 2.47e-05 ***
x.temp         5.64704    2.97462   1.898 0.060263 .
x.density     -2.42312    0.65243  -3.714 0.000322 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.18 on 110 degrees of freedom
Multiple R-squared:  0.3085,    Adjusted R-squared:  0.2833
F-statistic: 12.27 on 4 and 110 DF,  p-value: 2.782e-08
```

因為此次分析建立的模型眾多，此處將資料集以各迴歸模型跑出的結果(包含 R 平方值、各 x 變數的顯著程度與正負關聯性)整理如下表：

迴歸模型	y	x.rain	x.spray	x.temp	x.density	x.t	x.t2	R^2
Model 1	y.illness	+	+	+***	-***			*0.3085*
Model 2	y.illness	+*	+***	+**	-	+***		*0.4346*
Model 3	y.illness	+	+**	+**	+	+***	-***	*0.6253*
Model 4	y.illness_cum	-***	+	-***	-***			*0.5302*
Model 5	y.illness_cum	-	-**	-	+***	+***		*0.9281*
Model 6	y.illness_cum	+	-*	+***	+*	+	+***	*0.9781*
Model 1_log	log(y.illness+1)	-	+***	-	-***			*0.4388*
Model 2_log	log(y.illness+1)	+	+**	+***	-*	+***		*0.6177*
Model 3_log	log(y.illness+1)	-*	+*	+	+	+***	-***	*0.892*
Model 4_log	log(y.illness_cum +1)	-**	-***	-***	-***			*0.6069*
Model 5_log	log(y.illness_cum +1)	+**	+	+**	-**	+***		*0.9197*
Model 6_log	log(y.illness_cum +1)	-	-	-*	+	+***	-***	*0.9966*
Model 1_1og_s	log(y.illness+1)	-*		-*	-***			*0.356*
Model 2_log_s	log(y.illness+1)	+		+***	-*	+***		*0.5927*
Model 3_log_s	log(y.illness+1)	-**		+	+	+***	-***	*0.887*
Model 4_1og_s	log(y.illness_cum +1)	-***		-***	-***			*0.5631*
Model 5_log_s	log(y.illness_cum +1)	+**		+**	-**	+***		*0.9178*
Model 6_log_s	log(y.illness_cum +1)	-		-*	+	+***	-***	*0.9966*

四、討論與意涵

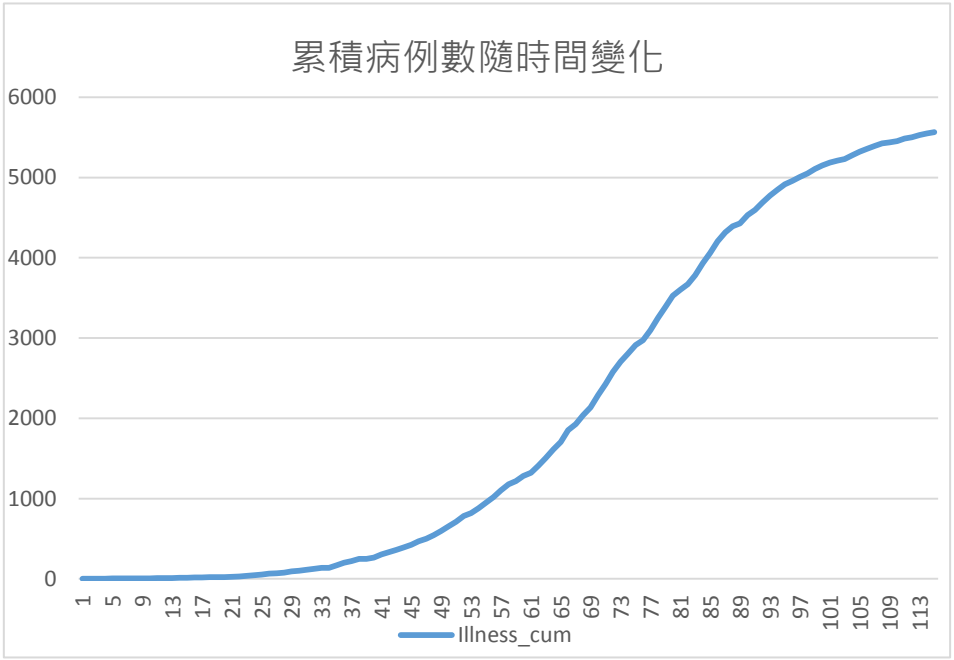
以 2015 年台南北區登革熱的相關開放資料，針對 x 變數降雨量、氣溫、噴藥次數、蚊子密度、時間與 y 變數病例數做迴歸分析，我們可以從結果得到以下發現：

(1). 登革熱病例數與時間有非常顯著的關聯性

比較以上各模型，可以發現每組迴歸中，若加上時間 x.t 或是時間平方 x.t2，都會使迴歸式的 R 平方值大幅提升 0.2~0.4 不等；表示時間對疾病的趨勢是一個很重要的解釋因素。

其中，病例數(y.Illness)、累積病例數(y.Illness_cum)與時間(x.t)呈正顯著相關，顯示整體而言，疾病的傳播會隨著時間持續增加。

另外，累積病例數(y.Illness_cum)與時間平方(x.t2)呈負顯著相關，也就是累積病例數與時間呈現負指數成長關係，顯示登革熱病例累積速度(疫情)會在後期逐漸趨緩。第一個原因，這符合一般流行病發展過程，一開始增長較慢，流行形成一定規模後，速度加快，之後又開始速度減慢，曲線逐漸平穩(如下圖)；同時，也可能是因為登革熱具有季節性，蚊子的活躍時間集中在暑期(大約是下圖前 80 筆資料)，在過了這個旺季後，得病案例就逐漸減少。



所以我們將蚊子密度抽出來與其他變數跑迴歸(結果如下圖)，可以發現單看天氣因素、灑藥與時間對蚊子密度，R 平方值就有 0.41 的解釋力；其中時間的關聯性最顯著，且蚊蟲的活躍仍然被其他未發現因素所影響。

```
lm(formula = x.density ~ x.temp + x.rain + x.spray + x.t + x.t2,
    data = pre.den2)

Residuals:
     Min       1Q   Median       3Q      Max
-13.5524  -2.5158  -0.0283   1.8947  24.6759

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.1744494 13.5363220   2.672  0.00869 **
x.temp      -0.7169499  0.4604220  -1.557  0.12233
x.rain      -0.0214980  0.0249459  -0.862  0.39070
x.spray     -0.1710383  0.3788954  -0.451  0.65259
x.t         -0.2965248  0.0556354  -5.330 5.36e-07 ***
x.t2         0.0015991  0.0004803   3.329  0.00119 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.584 on 109 degrees of freedom
Multiple R-squared:  0.413,     Adjusted R-squared:  0.3861
F-statistic: 15.34 on 5 and 109 DF,  p-value: 2.117e-11
```

(2). 整體而言，指數迴歸的解釋力優於線性迴歸

比較 model1~model6 與 model 1_log~model6_log 兩組迴歸模型，可以發現整體而言，指數迴歸的 R 平方值遠大於線性迴歸，也就是指數迴歸的解釋力比起線性迴歸要好得多。這也與前面的發現一致，即疾病的成長趨勢與時間呈負指數關係。

(3). 病例數和噴藥次數具有顯著正相關性

單看 model1、model12、model13、model1_log、model12_log、model13_log，可以發現 x 變數每日噴藥次數(x.spray)對 y 變數每日病例數(y.illness)皆具有正向顯著相關。若解釋為政府噴灑抑制蚊蟲孳生的藥物越頻繁，登革熱的確診人數就越多，相當不符合直覺常理。因此我們認為這兩個變數之間雖然有統計上的相關，應該沒有直接的因果關係，可能是基於其他干擾因素(氣候因素、蚊子密度或我們未發現的其他因素等)，讓兩個變數都同時被顯著影響；也可說當某些理由使得登革熱正肆虐時，政府也正好意識到嚴重性並著手噴灑藥物。

(4). 只要有考慮到時間，噴藥與否對 R 平方值的影響有限

如前所述，指數迴歸對登革熱病例數的解釋力較強，為了看噴藥對病例數的影響力，所以我們將指數迴歸的 6 個模型各自拿掉噴灑藥物(x.spray)這個變數，建立 model1_log_s~model_6log_s 六個迴歸模型。比較兩組 model 可以發現扣掉 x.spray 的 R 平方值平均僅下降 3%，其中 Model 5_log 和 Model 6_log 甚至下降不到 1%，也就是說整體而言拿掉 x.spray 後 x 變數對 y 變數的解釋力差異性不大。

直覺雖然可以說是因為灑藥與否對登革熱的控制沒有幫助，但我們認為應該是因為噴灑藥物後需要發酵期，並非一灑完藥物就能控制疫情，而是要在一段時間後阻止蚊蟲的生長，而後才能間接抑制疾病發展。

下圖可以看出在沒有噴藥的前期，蚊密度處於整個階段裡的高峰時期，通過噴藥很好控制了蚊子的繁殖狀況，使得在最為炎熱潮濕的時期，蚊子也沒有大量繁殖，表示噴藥與否或許對蚊蟲的抑制仍有一定的影響力，只是尚未發現兩者是何種關聯。