Logistic curve fitting to dengue fever data

參賽學生 ： 數學系 周一 王庭萱
指導老師 ： 數學系 林惠婷 老師

一、簡介

通過對登革熱疫情的偶然瞭解，我們對其提出了很多數學問題。在觀察確診病例的數據時，我們發現其與 Logistc 模型十分相似，通過正式的數學計算和分析，我們一步步將數據擬合在一條 Logistic 曲線上，通過歸納和整理，最終組成了此次報告的形態。此次研究的方法基於一篇非常實用且有趣的文章，Fitting a Logistic Curve to Data, by David Arnold, February 24, 2002. 在此文章中，作者細緻地闡述了擬合數據之方法，為我們提供了方法和思考方向。

二、數據整理

2015 年於台灣台南市所爆發的登革熱疫情十分嚴重，高頻的傳染及死亡使得此次疫情在短時間內得到了政府的高度重視。雖然此次疫情的傳播很快進入了根絕期，但由於登革熱病毒自身種類之間存在較大差異，且至今並未研究出有效之治療方法，所以我們認為，對其作出進一步的瞭解還是很有必要的。首先我們將台南市的確診數據之累積人數進行了整理，得出表 1。並將數據之散點圖繪出，命為圖 1。

時間	0	1	2	3	4	5	6	7
累積病例	141	477	989	2088	3367	6192	9071	12487
時間	8	9	10	11	12	13	14	15
累積病例	15516	17753	19367	20405	21182	21698	22022	22226

表 1

表 1 中之數據來自衛生福利部疾病管制署（網址：www.cdc.gov.tw）之南區台南市登革熱本土病例及境外移入病例趨勢圖（2015 年 01 週-2016 年 09 週）。其中時間以週為單位，累積病例以個為單位。

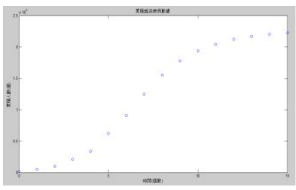

圖 1

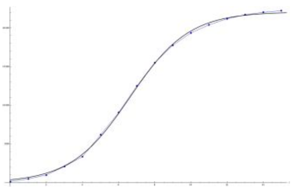

圖 2

三、Logistic 方程

首先介紹 Logistic 方程的基本理論。

我們使用 $\frac{dm}{dt}=rm(1-\frac{m}{K})$ 的形式定義 Logistic 方程，其中 t 為時間，$m=m(t)$ 是累積病例數，且 r 和 K 均為正參數。通過分離參數，解得此方程，$m(t)=\frac{K}{1+Ce^{-rt}}$，其中 C 為任意常數。經過計算，其二階導數為 $m''(t)=\frac{CKr^2e^{rt}(C-e^{rt})}{(C+e^{rt})^3}$，令 $m''(t)=0$，得到 $C=e^{rt}$。我們令 $C=e^{rt_0}$，即曲線之反曲點於 $t=t_0$ 取得，則方程的解可化為 $m(t)=\frac{K}{1+e^{-r(t-t_0)}}$，當 $t\to\infty$ 時，累積病例的值達到飽和狀態 K。

四、擬合方法

我們將運用最小平方法（Least square method）將數據 $(t_i,m_i), i=1,2,...,n$ 擬合在 Logistic 曲線上。此問題的重點是，在不同的 r,K 和 t_0 的情況下，將誤差 e 最小化，其中 $e=\sum_{i=1}^{n}(m(t_i)-m_i)^2$。此處對於誤差的描述，有以下幾點標注及分析。首先，我們繪製出預想結果之圖像，即所有數據點分佈在一條 Logistic 曲線上，記為圖 2。此外，簡單地說，最小平方法即是對數據點 (t_i,m_i) 以及曲線點 $(t_i,m(t_i))$ 之間的距離進行優化，並取得最優解。最後，需要強調的是，最小化誤差 e 是本次研究的最終目的。

1.估計參數

本次研究的難點在於如何用最小平方法解決非線性方程之問題。在 Logistic 函數 $m(t)=\frac{K}{1+e^{-r(t-t_0)}}$ 中，存在三個無法確定的參數。但如果我們估計出 K 的值，那麼問題就轉化為僅含有兩個參數的問題，即含有 r 和 t_0 兩個參數。這是此次研究運用的主要方法。

通過以上分析，我們令 $m(t)=Kh(t)$，其中 $h(t)=\frac{1}{1+e^{-r(t-t_0)}}$。

記 $H=\langle h(t_0),h(t_1),...,h(t_{15})\rangle$ 和 $M=\langle m_0,m_1,...,m_{15}\rangle$ 是兩個向量，我們對誤差 e 進行化簡，

$$\begin{aligned}e&=\sum_{i=0}^{15}(m(t_i)-m_i)^2=(m(t_0)-m_0)^2+\cdots+(m(t_{15})-m_{15})^2=(Kh(t_0)-m_0)^2+\cdots+(Kh(t_{15})-m_{15})^2\\&=\left\|\langle Kh(t_0)-m_0,...,Kh(t_{15})-m_{15}\rangle\right\|^2=\left\|K\langle h(t_0),...,h(t_{15})\rangle-\langle m_0,...,m_{15}\rangle\right\|^2=\|KH-M\|^2\\&=\langle KH-M,KH-M\rangle=K^2\langle H,H\rangle-K\langle H,M\rangle-K\langle M,H\rangle+\langle M,M\rangle=K^2\langle H,H\rangle-2K\langle H,M\rangle+\langle M,M\rangle\end{aligned}$$

需要強調的是，此表達式中誤差 e 還是含有三個不確定的參數。

2.最小化

接下來的步驟是將誤差 e 進行最小化。在單變量微積分中，可以通過令一階微分為零求出臨界點，從其增減性判斷最小值產生的位置。在多變量微積分中，唯一一點不同是我們必須對相應的變量進行偏微分，從而獲得關於最值的信息。對於誤差函數 e，我們對其取 K 的偏導數，即 $\frac{\partial e}{\partial K}$，並令其值為零，得 $2K\langle H,H\rangle-2\langle H,M\rangle=0$,則 $K=\frac{\langle H,M\rangle}{\langle H,H\rangle}$，將其帶入到誤差函數中，得 $e=\langle M,M\rangle-\frac{\langle H,M\rangle^2}{\langle H,H\rangle}$，記為 ◊。此結果僅含有兩個參數，即 r 和 t_0，而參數 K 已經被代換掉了。

3.構建誤差函數

首先我們估計 t_0 的取值範圍，從表 1 中即可得出，$0\le t_0\le 15$，對於 r，我們將使用一個較為粗略的方式（但結果並非如此）來估計其範圍。

對 $m(t)$ 取導數，得 $m'(t)=\frac{Kre^{-r(t-t_0)}}{(1+e^{-r(t-t_0)})^2}$，則 $m'(t_0)=\frac{Kr}{4}$，或 $r=\frac{4m'(t_0)}{K}$。如果我們對圖 1 中的數據進行估計，可以明顯地看出其反曲點位於 $t=7,8$ 之間，故我們選取 $t_0=7.5$。並且可以更進一步得出靠近反曲點處的切線斜率值，使用 $t=7,8$ 兩組數據進行計算，有 $m'(t_0)\approx\frac{15516-12487}{8-7}\approx 3029$，將此結果帶入 $r=\frac{4m'(t_0)}{K}$ 中，並取累積病例的飽和值 $K\approx 22226$（表 1 數據中累積病例數之最大值），得 $r\approx\frac{4(3029)}{22226}\approx 0.5451$。因此，通過對 r 的初步估計，我們將其範圍定在 $0.2\le r\le 0.9$ 之中。

4.繪製誤差曲面

通過以上參數之估計，取 r，t_0 和誤差 e 分別為 x，y 和 z 軸，用 Matlab 繪製出 ◊ 式之圖像，即所謂誤差曲面，記為圖 3。則最小平方解位於所繪誤差曲面之最低點。為了更清楚地瞭解誤差曲面上最小值的情況，我們將誤差曲面之平面等高線圖也繪製出來，記為圖 4。

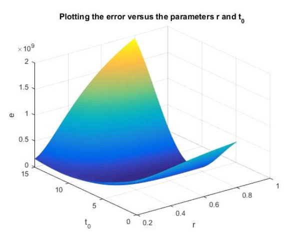

圖 3

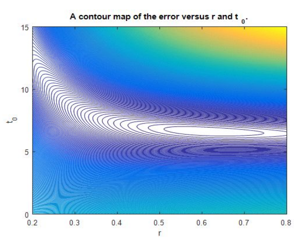

圖 4

5.最優解

通過兩圖，我們可以很清晰地看到，誤差函數的最小值所處之位置，與我們之前估計的數值十分吻合，即我們更加確定，最優解之參數取值為 $(r,t_0)=(0.5451,7.5)$。並計算 $K=\frac{\langle H,M\rangle}{\langle H,H\rangle}\approx 22226.3$，與所估相符。則最優化之 Logistic 表達式為 $m(t)=\frac{22226}{1+e^{-0.5451(t-7.5)}}$。

6.結論曲線

通過 Matlab 繪製出最優之擬合曲線，即右圖。

它與我們先前所估之結果十分相似，故我們得出結論，此次疾病數據之累積人數符合 Logistic 模型。