工程統計期末報告

土木系大二　109612054　吳巽言

報告題目	Old Faithful Geyser Data
報告學生 (照片與姓名)	 吳巽言
Abstract **摘要(10%)**	老忠實間歇泉（Old Faithful Geyser）位於美國黃石國家公園（Yellowstone National Park, Wyoming, USA）西南區的上間歇泉盆地 Upper Geyser Basin，會穩定間隔一段時間噴發，是著名的觀光景點。但噴發與噴發之間間隔時長從 43 到 96 分鐘，且呈雙峰分佈。因此尋找噴發時長與噴發間隔時長間的關係用以預測噴發間隔時長，得知需等待多久才看得到下次噴發。噴發時長及噴發間隔時長皆為雙峰型分佈，兩者之間也大約成正相關，對其進行線性回歸後可得決定係數達 0.81 的線性關係式 waiting= 33.47 + 10.73 x eruptions。而對資料曲自然指數後進行計算可由指數關係得決定係數達 0.82 的曲線關係式 waiting= 38.59 x eruptions^ 0.5。由這兩個關係是可以有效地以上一次噴發時長預測到噴發的間隔時長。 **Keywords:** Old Faithful Geyser, bimodal distribution
Motivation **動機(10%)**	老忠實間歇泉是黃石公園著名的景點之一，許多遊客皆會前往觀賞它噴發，因此就有需要預測噴發與噴發之間的間隔長來讓遊客判斷等待時間還有多久。雖然現有的資料可以大略得知多項噴發間隔時長的紀錄，但其範圍從 43.0 到 96.0 分鐘，且呈雙峰型分佈，因此有必要有更值得信任的判斷/預測方式。資料中的另一項為噴發時長，如果兩者之間有相關性的話即可用以預測噴發間隔時長。

Data
資料(15%)

老忠實間歇泉（Old Faithful Geyser）

位於美國黃石國家公園（Yellowstone National Park, Wyoming, USA）西南區的上間歇泉盆地 Upper Geyser Basin，會穩定間隔一段時間噴發。

於 1870 年被 Washburn Expedition 發現，因為他規律的噴發頻率而命名為老忠實間歇泉。一天約可噴發 20 次，大約有 90%信心預測對噴發時間。

噴發高度大約為 100-180 feet，較集中於 130-140 feet。科學家估計總水量大約為 3700 加侖到 8400 加侖，由噴發時長決定。噴發水溫則非常高溫：95.6°C，蒸氣溫度甚至 176.6°C。

這份老忠實間歇泉的資料框架包含 2 種變數：噴發時長（分鐘）、噴發間隔時長（分鐘）的 272 份觀察資料。

[,1]	eruptions	numeric	Eruption time in mins
[,2]	waiting	numeric	Waiting time to next eruption (in mins)

基本資料：

	噴發時長（分鐘） Eruptions (min)	噴發間隔時長（分鐘） Waiting (min)
最小值 Min.	1.600	43.0
第一四分位數 1st Qu.	2.163	58.0
中位數 Median	4.000	76.0
平均數 Mean	3.488	70.9
第三四分位數 3rd Qu.	4.454	82.0
最大值 Max.	5.100	96.0
眾數 Mode	1.867	78
標準差 Standard Deviation	1.141371	13.59497

Methodology **方法(10%)**	（1） 繪製直方圖，如圖 1。即可明顯看出兩項資料規律皆為雙峰分佈。 （2） 繪製分位圖，如圖 2，可看出在兩個數值範圍內接觀測到資料點，並皆偏離直線且漸趨水平，因此可判斷噴發時長與噴發間隔時長確實都是雙峰型分佈。 （3） 繪製時間對噴發間隔時間的散佈圖，如圖 3，可看出兩筆資料有正相關，但非線性相關。因此進行線性回歸。 （4） 線性回歸：使用 lm()函式處理資料，再計算得預測直線。 （5） 冪變換 (power transformation)：將資料取自然對數，再計算得預測曲線。 （6） 線性回歸即冪變換後的資料可看出兩筆資料的關係。最後繪製處理過後的資料的散佈圖，如圖 4 。
Results & **Discussion** **結果與討論(30%)**	由直方圖及密度曲線看出噴發時間（eruptions）與噴發間隔時間（waiting）後。 由分位圖看出在兩個數值範圍內接觀測到資料點，並皆偏離直線且漸趨水平，判斷出噴發時長與噴發間隔時長確實都是雙峰型分佈。 發現兩筆資料皆為雙峰關係後，分別找他們的偏值及峰值： eruptions: skewness = -0.415841, kurtosis = 1.4994 waiting: skewness = -0.4163188, kurtosis = 1.857369 由偏值為負且峰值為正可知道是偏左偏的高峽峰雙峰分佈。 圖 3 時間對噴發間隔時間的散佈圖，可看出兩筆資料有正相關，但非線性相關。 進行線性回歸後， 得到線性關係為：waiting= 33.47 + 10.73 x eruptions 決定係數 r^2: 0.81 標準誤差 se: 5.91 對資料曲自然對數後可以得到指數型分佈的預測曲線， 得到指數關係為： waiting= 38.59 x eruptions^ 0.5 決定係數 r^2: 0.82 標準誤差 se 5.83 最後畫圖可以得到經過冪回歸的資料散佈圖以及線性回歸的直線及曲線。 由決定係數=0.81 及 0.82 可以得知老忠實間歇泉的噴發時長及噴發間隔時長確實有一定程度的相關性， 直線 waiting= 33.47 + 10.73 x eruptions 及曲線 waiting= 38.59 x eruptions^ 0.5 兩關係式可以用以預測資料的噴發時長與噴發間隔時長。

Figure & Table

圖與表(20%)

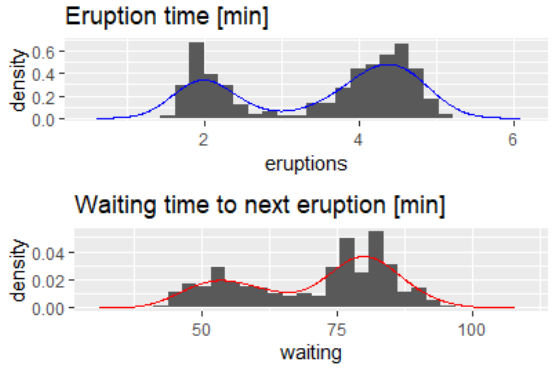

圖 1：噴發時間與噴發間隔時間的直方圖與密度曲線

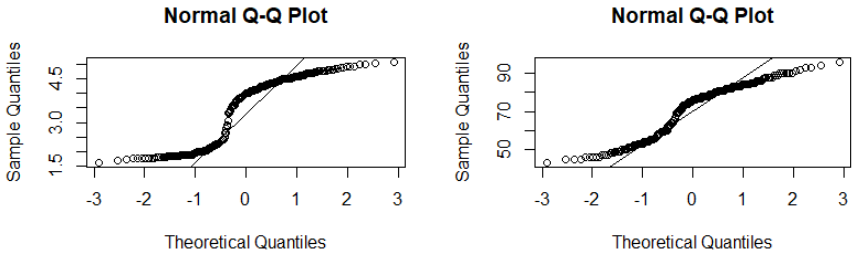

圖 2：噴發時間（左）與噴發間隔時間（右）的分位圖

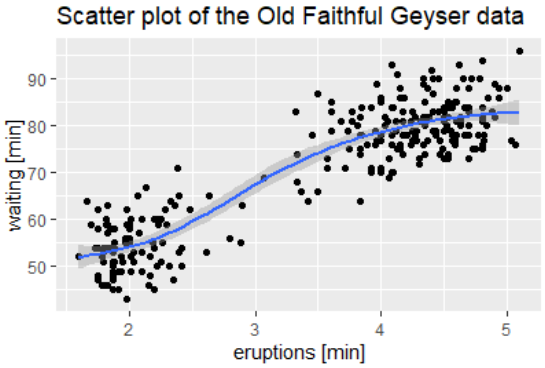

圖 3：噴發時間-噴發間隔時間的散佈圖

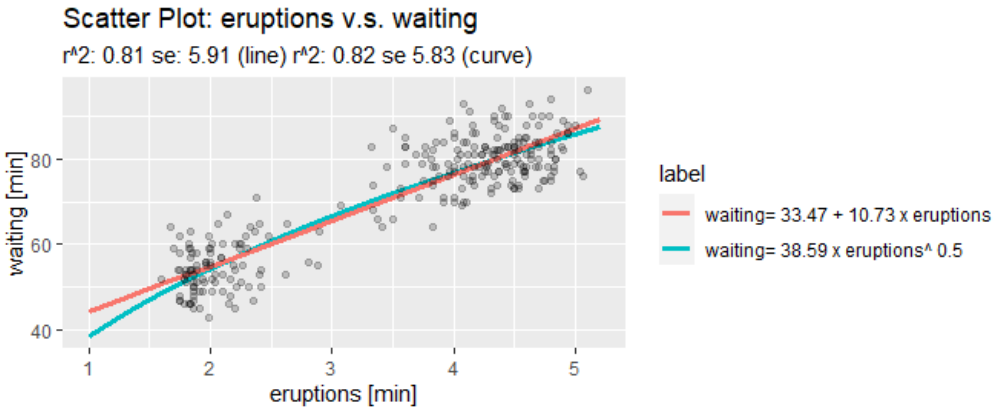

圖 4：線性回歸後的噴發時間-噴發間隔時間散佈圖

References & Resources **參考文獻與資料來源(5%)**	W. Härdle. Härdle, W. (1991). Smoothing Techniques with Implementation in S. New York: Springer. Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. Applied Statistics, 39, 357–365. doi: 10.2307/2347385. https://www.yellowstonepark.com/things-to-do/geysers-hot-springs/about-old-faithful/
Suggestion **課程建議(5%)**	我覺得這學期線上實體並行還不錯，也給我們很多方便。 但希望 PPT 可以詳細一點，有時候自己想要回去看 PPT 找一下解答的時候有點難看懂，就得再回去翻上課錄影找很久，也幸好有錄影！ 謝謝教授！