Practice#1：多項式迴歸分析計算

目的：

簡單線性迴歸分析建立在自變數與因變數的線性假設上。當兩者間的關係不再是線性，而是比較近似多項式的關係時，譬如二次多項式，其模型可以寫成

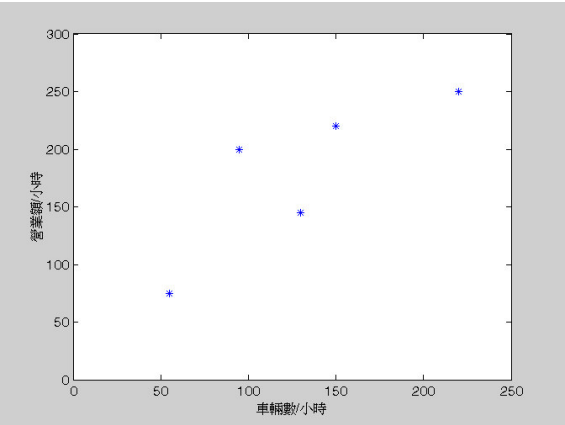
$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

或 p 次多項式，寫為

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p$$

練習：

1. 在前一個練習的作業，有關〔月份〕與〔平均溫度〕的關係中，簡單線性迴歸分析似乎得到一個很差的結果。從散佈圖看來，兩者間的關係比較像一條曲線。在這個練習裡面，請試著利用p 次多項式的來滿足兩變數間的關係。至於 p 等於多少，自行判斷。此時的 R^2 是否進步一些了呢？
2. 逐漸提高 p 值，觀察R^2 值的變化。(調整你的程式，讓這件工作做起來方便)
3. 右圖是紀錄關於加油站每小時的車輛數與營業額的關係。使用適當的迴歸模型，預測當每小時的車輛數為 183 時的營業額？



車輛數/小時	營業額/小時
150	220
55	75
220	250
130	145
95	200

註 1：從線性的迴歸模式到多項式迴歸模式，哪一個比較恰當？判斷的依據是什麼？

註 2：MATLAB 也提供相關的功能。在畫出上面的散佈圖後，直接在圖形視窗上點選『Tools』→『Basic Fitting』。

觀察：

1. 逐漸提高 p 值的同時，R^2 值是否也跟著提高呢？R^2 值得提高是否意味著較好的模式呢？

2. 如何決定適當的項次？
3. 在數學的範疇裡，類似的迴歸的觀念被稱為 curve-fitting，其中的 extrapolation 與預測的意思接近。在迴歸分析或時間序列的模式中，都可以看到 curve-fitting 的影子。一般而言，在數學上curve-fitting 講究的是找到一條線可以貫穿所有的點。不過迴歸分析或時間序列更講究預測的能力，是一種 model-fitting。對過去資料的配適能力不求精準，反而注重對未來資料的預測能力，因此常常在模型配適的過程中，加入一些其他的判斷因素。譬如對資料之新舊給于不同的權重。或是刻意降低配適精準度，以提高預測能力、、、等。練習 3 是很好的例子。

作業：

1. 資料 regress_data3.txt 是有關〔車速〕與〔耗油量〕的關係。先試著以線性迴歸模式來處理，再利用多項式迴歸來證明其適當性。至於p=?請自行判斷，並說明你的理由。最好將程式設計的比較彈性些，可以在執行階段再來決定多項式的項次。
2. 除了利用殘差圖或一些必要的統計量來幫忙決定 p 值外，也可以將多項式的圖與散佈圖畫在一起，看看合適與否？
3. 迴歸分析的目的在於應用最後的決定的最佳模式進行預測。你也可以試試看，不同模式的預測能力。

參考文獻：

陳順宇，迴歸分析一三版，華泰書局。