Practice#1：基本的迴歸分析計算

目的：

迴歸分析是統計學上重要的應用工具。多數的統計相關軟體都有提供迴歸分析的功能指令。透過指令的呼叫，可以計算出許多相關的統計量，例如SSE，R^2。本單元想進一步深入 MATLAB 的運用，借迴歸分析的一些相關公式，利用 MATLAB 程式計算出相關的統計量，對於 MATLAB 在計算上的了解將有些幫助。至於變異數分析及參數的區間估計與檢定並不在此列。

簡單的線性迴歸模式如下：

$$Y = \beta_0 + \beta_1 X \tag{1}$$

這定義了應變數（或輸出變數）Y與自變數（或輸入變數）X間的關係，而係數 β_0，β_1是決定這個關係的參數。當然，在實際的情況下，這兩個參數值通常是未知的，已知的只有 X與 Y的量測資料(measurements)。迴歸分析的第一步便是透過這些已知的資料，估計未知的參數，這也是這個單元的重點。其中最重要的估計方法為「最小平方法」，其問題與結果如下

$$\min_{\beta_0,\beta_1} \sum_{k=1}^{N} [\ y_k - (\beta_0 + \beta_1 x_k)\ ]^2 \equiv \min_{\underline{\beta}} \left\| X\underline{\beta} - \underline{y} \right\|_2^2 \tag{2}$$

$$\text{其解（以矩陣表示）}\ X^T X \underline{\hat{\beta}} = X^T \underline{y} \tag{3}$$

式(2)中 x_k與 y_k代表已知的資料（共N筆資料），式(3)的矩陣 $X, \underline{y}, \underline{\hat{\beta}}$定義於(4)。這個單元提示如何應用 *Matlab* 來做參數估計。至於從問題(式 2)到結果(式 3)的推導過程將在課堂中仔細的推敲。此外，這個階段的重點還有：如何以矩陣的方式求解？什麼是梯度向量(gradient vector)？*Matlab* 的程式從何處下手？資料將如何安排建構成矩陣，方便在MATLAB 中進行運算。別忘了，MATLAB 是 Matrix Lab.。

練習：

有幾組資料提供下列練習及作業之用：(1) *regress_data1.txt* (2) regress_data2.*txt* 。 這些資料可以從網站上直接下載。資料來源參考[1] 。網址：(http://web.ntpu.edu.tw/~ccw/statmath/stat_comp.htm)。

1. 對相關資料畫散佈圖。對兩組資料間的關係做初步的了解。
2. 針對這些資料，利用線性迴歸模式，在 MATLAB 下建立最基本的資料矩陣 X 及 $\underline{y}$ 。其中

$$\underline{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \tag{4}$$

N 表示資料的數量。其中 X 的結構很特殊，第一行都是 1，善用 Matlab 的 ones 指令，可以很精簡的完成類似的資料矩陣。

3. 從最小平方法導出的 *Normal Equations* (3)計算 $\underline{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}$。
4. 計算下列統計量的值（在 Matlab 中以向量相乘的方式）
   1. $SSTO = \sum (Y_i - \overline{Y})^2$ ；$\overline{Y}$ 為均值
   2. $SSE = \sum (Y_i - \hat{Y}_i)^2$ ：*殘差平方合(Sum of Squares due to Errors)*
   3. $SSR = \sum (\hat{Y}_i - \overline{Y})^2 = SSTO - SSE$：*迴歸平方合(Sum of Squares due to Regression)*
   4. $R^2 = \dfrac{SSR}{SSTO}$
5. 在散佈圖上畫一條迴歸線 $y = \hat{\beta}_0 + \hat{\beta}_1 x$。並在圖上適當的地方印上 R^2 *值*。
6. 畫兩張殘差圖：“*Residuals vs. x*”及“*Residuals vs.* $\hat{y}$“ 。其中，
   1. 擬合值(*fitted value*) $\hat{y}$ ：$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$，利用已知的 x 值與參數估計值計算得到的 Y 值稱為擬合值。
   2. 殘差(*residuals*)：擬合值 $\hat{y}$ 與觀察值 y 的誤差。
   3. 預測值(*predicted value*)：利用已知資料以外的 x 值與估計參數計算得到的 Y 值稱為預測值。

觀察：

1. 迴歸分析有其嚴密的理論推導過程，其中應用最小平方法的觀念來計算 β_0, β_1 尤其重要。這部分是課堂講解的重點。也是初學迴歸時最痛恨的計算，但是用矩陣與電腦來計算就顯得輕鬆多了。
2. 用線性代數的空間觀念(space)來解釋最小平方法，是了解線性代數很棒的途徑。也提供往後的各項練習的基礎。這部分也是課堂講解的重點。
3. 畫殘差圖的目的是什麼呢？從上面的資料你觀察出什麼？
4. Matlab 也提供了一些關於迴歸分析的工具，試著找出來用看看。是不是有

些程式就不必寫了呢？

5. 迴歸分析是一個比較複雜的資料分析工具。它的複雜來自資料本身的多樣性，並非一套方法適用所有的資料。為方便研究工作的進行，有時候需要自行模擬資料。其中最簡單的模式如：$Y = \beta_0 + \beta_1 X + \varepsilon \qquad \varepsilon \in N(0,\sigma^2)$，將參數決定好，便可以隨時產生資料。請試著產生幾組資料並套入之前的程式；觀察誤差項ε的分佈對結果的影響？或是如何增加誤差項ε對結果的影響？
6. 從簡單的線性迴歸模型到最小平方解(1)(2)(3)，不難看出當自變數從一個變成多個時，其參數的估計都是一樣的。只需要在資料矩陣 X 上擴充即可套入原程式。試試看這個模式：$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$。你可以利用前面的方式自行產生需要的資料，再做估計。

作業：

1. 同以上的練習，但是採用資料(2)，分析〔平均溫度〕與〔電費〕的關係。
2. 資料(2)還提供了月份的資料，可以進一步分析〔月份〕與〔平均溫度〕，〔月份〕與〔電費〕的關係。
3. 對以上的練習，請寫下結果的觀察。你可以從殘差圖、R^2 或其他統計量或圖表來觀察。
4. 假設$f(\underline{x})$為一多變量函數，計算其梯度向量$\nabla f(\underline{x})$當

   a. $f(\underline{x}) = \underline{b}^T \underline{x}$

   b. $f(\underline{x}) = \underline{x}^T \underline{a}$

   c. $f(\underline{x}) = \underline{x}^T \underline{A}\ \underline{x}$

5. 仔細且清楚的寫出從式(2)以矩陣梯度向量的方式，推導出式（3）的Normal Equations。
6. 仔細且清楚的寫出從向量空間的正交概念(orthogonal)推導出式（3）的Normal Equations。

參考文獻：

陳順宇，迴歸分析一三版，華泰書局。

1.