財經數據分析

R 新手入門第一課: 從線性模式開始

~ 如果是新手，只要會裝 R，剩下的就從這 30 頁開始 ~

R 的線性模式 lm()函數的使用方法，是多數 R 套件的代表。本文就用 **lm()**的介紹，來介紹 R 的函數和物件特徵。這樣，讀者在使用 R 上，比較能夠滿足相關資料分析需求。以免介紹一堆無直接關係的語法和指令。

本文提供程式碼和資料檔，學習者只要一行一行執行，30 頁就輕易入門了。

By

何宗武

世新大學財務金融系教授

R 的線性模式函數 lm()介紹

第 1 節 估計原理--最小平方法

假設我們有兩筆資料，分別為 x 和 y，其散佈圖如下圖：

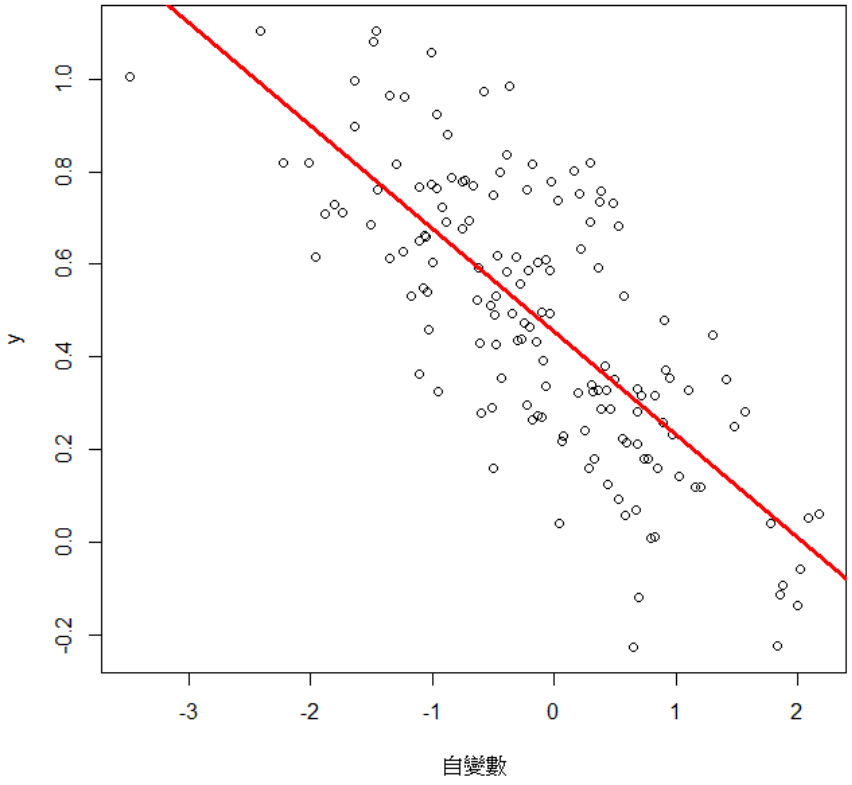

Figure 1.1 資料散佈圖

視覺上可以知道兩者之間為負相關，因此，我們可以目測：用尺在差不多的地方畫一條負斜率的直線，如上圖之斜截式。上圖之直線，稱為配適線(fitted line)，散佈之樣本和這條線的距離稱為殘差或剩餘(residuals)。這條斜截式的直線方程式可以寫成如下

$$\boldsymbol{y}=a+b\boldsymbol{X}+\varepsilon$$

右邊之 $a+bX$ 也稱為 $\boldsymbol{y}$ 的期望值 E$\boldsymbol{y}$，通常寫成 $\hat{y}$ ，ε 則為殘差項。另一個看上式的方式是，隨機變數 $\boldsymbol{y}$ 可以被分割為三份：截距 a，X 的線性組合 bX 和殘差ε。

迴歸分析第一件工作就是「估計」：用數理方法估計出這一條線的截距和斜率(統稱參數或係數)，而不是用目測。對於線性關係的估計，最好的方法就是最小平方法(Least Square)，這個方法求解參數的目標為：使殘差平方和為最小的所有參數。

利用矩陣，一個標準的線性迴歸可以矩陣表示如下：

$$y = X\beta + \varepsilon$$

上式中 y 為樣本數為 n 個資料的被解釋變數(或稱為 n-維向量), X 是解釋變數，假設有 k 個，故 X 為 $n \times k$ 的獨立變數矩陣；循上可知 β 則為 k-維的係數向量(或稱為參數向量)。ε 為殘差項，也就是變數 y 的變異，不被 $X\beta$ 解釋的剩餘部分。

令 b 代表 β 的樣本估計式，則滿足下式的解，即為迴歸係數

$$\min_{\beta}(y - X\beta)'(y - X\beta)$$

上述目標函數，即是「殘差的平方和」，以偏微分解出最適值如下：

$$b = (X'X)^{-1}X'y$$

這個估計式的共變異數矩陣(covariance matrix)的式子為

$$\text{cov}(b) = s^2(X'X)^{-1}$$

上式中 $s^2 = \dfrac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$，且 $\hat{\varepsilon} = y - Xb$。如果是複迴歸，則第 j 個係數的變異數為

$$\text{var}(b_j) = \frac{1}{1-R_j^2}\frac{s^2}{\sum_j (x_{jj} - \bar{x}_j)^2}$$

因為有 k 的係數，所以共變異數矩陣 $cov(b)$為 $k \times k$ 的方陣，長的如下

$$\text{cov}(b) = \begin{bmatrix} \sigma_{11} & \sigma_{21} & \cdots & \sigma_{k1} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{k2} \\ \cdots & \cdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{11} & \cdots & \sigma_{kk} \end{bmatrix}$$

上式中，主對角線 (diagonal) 的 k 個數值，就是係數自己的變異數，取根號後，就是標準差，可以用來檢定個別參數的性質，好比是否顯著異於 0。

離對角線(off-diagonal)就是交叉成分σ_{ij}，為參數間的共變異。可以用來檢定參數間的關係、相關性等結合檢定(joint tests)。後續會繼續說明。

第 2 節　單變數線性迴歸

我們先打開資料 **IS_CA.csv** 這個檔案，這個檔案是全球 140 個國家 1980-2010 年的平均資料，這筆資料是橫斷面資料。變數定義如下：

INVEST=投資率 （投資毛額/GNP）

SAVING=儲蓄率 （儲蓄毛額/GNP）

CA=經常帳餘額（經常帳/GNP）

Countries=國名

Group=經濟發展程度分類

R Code. 載入資料且快速瀏覽

```
1.  temp=read.csv("IS_CA.csv",header=TRUE)
2.  head(temp)
3.  summary(temp)
4.  library(fBasics)
5.  myData=na.omit(temp)
6.  basicStats(myData[,1:3])
7.  colStats(myData[,1:3],mean)
8.  colMeans(myData[,1:3])
9.  skewness(myData[,1:3])
10. kurtosis(myData[,1:3])
11. cov((myData[,1:3])
12. var(myData[,1:3])
13. cor(myData[,1:3])
```

說明

1. 載入外部資料，存入暫存物件 temp。
2. 看看資料變數前 6 筆 （要看最後 6 筆，則用 tail()）
3. 看看資料的統計摘要
4. 載入模組 fBasics （因為要使用較完整的敘述統計函數）
5. 移除缺值（na.omit()是將缺值移除的函數）
6. 對前 3 筆連續資料，看 16 項統計摘要。
7. 計算前 3 欄連續變數的平均數
8. 同前
9. 計算前 3 欄連續變數的偏態
10. 計算前 3 欄連續變數的峰態
11. 計算前 3 欄連續變數的共變異數矩陣

12. 同前
13. 計算前 3 欄連續變數的相關係數矩陣

summary(temp) 是資料的摘要，不是較詳細的敘述統計量。但是，對於資料的狀況、分佈和缺值等等，都有重要資訊。例如，**Countries** 和 **Group** 均是文字，比對的數據，其實就告訴了我們，不同經濟發展程度下有多少國家。例如，Advanced Economies 有 34 國

```
> summary(temp)
     INVEST            SAVING             CA          
 Min.   : 6.992   Min.   :-16.11   Min.   :-24.0701  
 1st Qu.:20.570   1st Qu.: 16.00   1st Qu.: -5.5788  
 Median :23.546   Median : 20.35   Median : -2.9622  
 Mean   :23.919   Mean   : 20.38   Mean   : -1.7747  
 3rd Qu.:26.425   3rd Qu.: 24.67   3rd Qu.:  0.1403  
 Max.   :43.372   Max.   : 43.10   Max.   :122.9456  
 NA's   :14       NA's   :14                         
                Countries
 Albania             :  1
 Algeria             :  1
 Antigua and Barbuda:  1
 Argentina           :  1
 Armenia             :  1
 Australia           :  1
 (Other)             :134
                                  Group
 Advanced Economies                  :34
 Central and Eastern Europe          :14
 Commonwealth of Independent States:13
 Developing Asia                     :27
 Latin America and the Caribbean     :32
 Middle East and North Africa        :20
```

如果需要進一步較詳細的統計資訊，**basicStats()** 可以計算 16 項數據和資料的性質，這個函數只能用於實數資料的統計計算，所以，我們的資料只有前 3 欄是數字，所以，就用 1:3 取前 3 行。

```
> basicStats(myData[,1:3])
                            INVEST     SAVING         CA
      nobs                     140        140        140
```

NAs	14	14	0
Minimum	6.992	-16.107	-24.070
Maximum	43.372	43.102	122.946
1. Quartile	20.570	16.005	-5.579
3. Quartile	26.425	24.668	0.140
Mean	23.919	20.383	-1.775
Median	23.546	20.355	-2.962
Sum	3013.791	2568.217	-248.460
SE Mean	0.488	0.711	1.138
LCL Mean	22.952	18.975	-4.024
UCL Mean	24.886	21.791	0.475
Variance	30.061	63.764	181.243
Stdev	5.483	7.985	13.463
Skewness	0.634	-0.420	6.111
Kurtosis	1.592	2.935	51.919

R 資料格式，基本的有兩種：向量矩陣(vector/matrix)和資料框架(data.frame)。都是用矩陣 **列×行** 的方式，來呼叫部份資料。例如，myData[1:2, 2:3] 就是呼叫資料前 2 列和後 2 行。

下面兩個語法，呼叫同樣的結果。

```
basicStats(myData[,1:3])["Sum", ]
basicStats(myData[,1:3])[9, ]
```

其次，我們執行單變數線性迴歸

$$INVEST=a+b_1 \cdot SAVING$$

這筆資料的投資率，是固定資本形成毛額。這個迴歸可以評估一個國家「儲蓄率」對「投資率」的關係，係數 b_1 也稱為儲蓄保留係數(saving retention coefficient ）衡量了儲蓄對投資的貢獻。依照經濟學原理，這個迴歸係數有兩種可能：

其一、如果是國際資本流動高程度高的封閉體系，則應該是顯著正相關，以反映出整體儲蓄對整體投資的正面誘因。然而，我們還可以看一看相關程度和影響有多大。

其二、如果是國際資本流動高程度高的開放體系，則應該是不顯著正相關，反映出國內投資所需資本，在國際資本市場上融資，所以，國內投資和國內儲蓄沒有統計上的相關性。

這就是著名的 Felstein and Horioka(1980)[1]對國際資本研究的議題。也成為國際資本流動迷思(Puzzle)。

R Code

```
1.  FH_1vlm=lm(INVEST~SAVING, data=myData)
2.  summary(FH_1vlm)
3.  confint(FH_1vlm,  level=0.9)
```

說明

1. 估計線性迴歸，把結果存入暫存物件 FH_1vlm。
2. 迴歸結果摘要
3. 迴歸係數 90%之信任區間。

暫存物件只存在記憶體，軟體關閉就會消失。暫存物件名稱，只要符號是允許的，可以任意定義你喜歡歡的名稱，但是，不可用中文。

lm()還有 2 個重要的技巧：

第 1、如果要將資料所有的變數都放進去，則

```
lm(y ~ ., data=)
```

如果解釋變數很多，上面這個功能就很有用。

第 2、如果要去除截距，則在解釋變數後面增加「**-1**」

```
lm(INVEST~SAVING-1, data=myData)
```

summary()和 **confint**()是兩個處理迴歸後物件 **FH_1vlm** 的原生函數(genetic functions)，詳細的原生函數後面會列表說明。

```
>summary(FH_1vlm)

Call:
lm(formula = INVEST ~ SAVING, data = myData)

Residuals:
    Min      1Q  Median      3Q     Max
-13.265  -3.330  -1.145   2.867  17.766

Coefficients:
```

[1] Feldstein, M. and C. Horioka (1980) Domestic saving and international capital Flows, *Economic Journal*, 90, 314-329

```
             Estimate    Std. Error   t value Pr(>|t|)
(Intercept)  17.3286     1.1902       14.560  < 2e-16 ***
SAVING        0.3233     0.0544        5.944   2.63e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.856 on 124 degrees of freedom
  (14 observations deleted due to missingness)
Multiple R-squared:  0.2217,Adjusted R-squared:  0.2155
F-statistic: 35.33 on 1 and 124 DF,  p-value: 2.627e-08
```

這個估計結果摘要的 output，有幾項可以注意，爾後本書簡單的列出結果。例如，`Call:`是估計函數，殘差的分量用來檢視離群值的狀況。`Signif. codes:` 指稱顯著性。最下方則是基礎 F 檢定。

由以上結果，我們發現兩者是正相關，估計出之係數為 0.323，由極小之 P-value，可知其非常顯著。這個數字的解釋如下，平均而言：

> 其他條件不變，儲蓄率多一單位的國家，比低一單位的國家，投資率要高出 32.3%單位。

因為這筆資料不是時間序列。所以在解釋變動時，要強調**樣本間差異**，而不是成長率。如果是時間數列，則可以用具有時間趨勢意涵的成長率來解釋。

如果覺得小數點顯示非 0 位數太多，要改成至少 4 個有意義的非 0 數字，可以執行 **options**(digits = 4)[2]修改。如果要四捨五入，則可以用函數 round()，以此例，round(**FH_1vlm, 3**)將故計數據，近似到小數第 3 位。

```
> confint(FH_1vlm, level=0.9)
                    5 %       95 %
   (Intercept)      15.3562912  19.3009961
   SAVING       0.2331837        0.4134761
```

令估計係數為 *b* 信任區間的定義：

[2] 這個小數宣告，不是恰有 4 位小數點。**R** 會以 4 位小數自動的判讀與對齊，所以，自動調整對齊和美觀的要求。

$$\hat{b} \pm t_{\frac{1}{2}\alpha,df} \cdot \hat{s}_b$$

$t_{\frac{1}{2}\alpha,df}$ 為特定信任水準之下的臨界值(critical value)

$\hat{s}_b$ 為估計係數的標準差

這個信任區間，將 α=10%分成一半一半在兩端。利用這些資訊，我們可以繪出 intercept 和 SAVING 兩個係數的信任橢圓(confidence ellipse)，下面的程式檔，說明了如何繪製係數的信任橢圓。

R Code. 係數之信任橢圓

```
1. library(ellipse)
2. plot(ellipse(FH_1vlm,c(1,2)),t ype="l")
3. points(coef(FH_1vlm)[1], coef(FH_1vlm)[2], pch=13)
4. abline(v=confint(FH_1vlm)[1,], lty=2)
5. abline(h=confint(FH_1vlm)[2,], lty=2)
```

說明

1. 載入模組 **ellipse** 以畫製橢圓
2. 利用前述迴歸的第 1、第 2 個係數當橢圓圓心畫橢圓。Figure 1.2
3. 將這 2 個係數標在上面圖文框
4. 用第 1 個係數信任區間的 2 個值，畫 2 條垂直線(***v***ertical line)
5. 用第 2 個係數信任區間的 2 個值，畫 2 條水平線 (***h***orizontal line)

Figure 1.2 的橢圓面積內，沒有座標(0, 0)，且是狹長型負相關。可知兩個係數的相關性都非常地顯著。

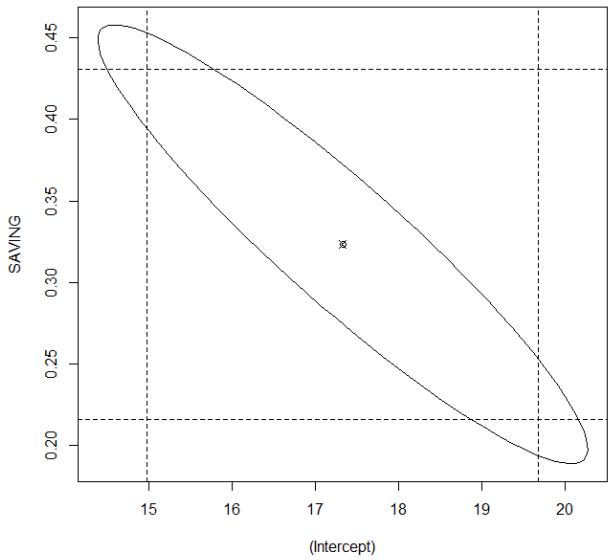

Figure 1.2 係數之信任橢圓

迴歸結果物件 **FH_1vlm** 內，據有很多其他資訊，不是直接用 **summary**()就可以全叫出來。這是因為 R 是用串列 **list**()的原理在處理資訊。要進一步去看，可以用 **names**()函數於兩個物件：

names(FH_1vlm) 和 **names(summary**(FH_1vlm)**)**

names(FH_1vlm) 可以看到 FH_1vlm 內的物件，

```
> names(FH_1vlm)
 [1] "coefficients" "residuals"  "effects"   "rank"
 [5] "fitted.values" "assign"     "qr"  "df.residual"
 [9] "na.action"     "xlevels"    "call"    "terms"
[13] "model"
```

names(summary(FH_1vlm)**)** 可以看到 **summary**(FH_1vlm)內的物件，

```
> names(summary(FH_1vlm))
[1]"call"     "terms"  "residuals"     "coefficients"
[5]"aliased"   "sigma"  "df"    "r.squared"
[9]"adj.r.squared" "fstatistic" "cov.unscaled"
[12]"na.action"
```

用符號 **$** 擷取資訊。這兩個物件的子物件，有一些名稱重複；但是，內容是不同的。例如，係數(coefficients)，兩者皆有，但是，FH_1vlm$coef 只有估計值，summary(FH_1vlm) $coef 較為詳細。如下：

```
> FH_1vlm$coef
(Intercept)      SAVING
 17.3286436   0.3233299
```

```
> summary(FH_1vlm)$coef
             Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 17.328644 1.19014654  14.560092 1.293509e-28
SAVING       0.3233299  0.05439557  5.944049  2.627034e-08
```

用 R 進行實證分析，必須要將結果到文字編輯軟體。因為 LaTex 的說明，對於使用 Tex 的人，應該都可以知道。目前，我們假設讀者使用的是微軟的 WORD，如果是這樣，可以將係數表格，存入 csv 格式，再於 WORD 內插入物件即可。

```
myTable01=summary(FH_1vlm)$coef
write.csv(myTable01, file= "myTable01.csv")
```

當我們在 R 執行完線性迴歸 **lm()**後，且也將迴歸結果存成物件，例如上例存成 FH_1vlm，R 對這個物件的迴歸後統計分析相當多，我們稱之為原生函數(genetic functions)，如表 1 所列。

表 1.　迴歸結果物件的原生函數

函數	說明
print()	將資訊列印於螢幕
summary()	資訊摘要
coef()	係數資訊
residuals()	迴歸殘差
fitted()	迴歸配適值資訊
anova()	變異數分析
predict()	迴歸推測
plot()	繪圖
confint()	迴歸係數之信任區間
deviance()	計算模型的 SSE
vcov()	共變異數矩陣 variance-covariance matrix
logLik()	對數概似值
AIC()	AIC 值
BIC()	BIC 值

接下來，就是對上表原生函數使用的介紹

R Code. 迴歸後診斷(1)

```
1.  with(plot(SAVING, INVEST), data=myData)
2.  abline(FH_1vlm)
3.  AIC(FH_1vlm)
4.  BIC(FH_1vlm)
5.  cbind(AIC(FH_1vlm), BIC(FH_1vlm))
6.  rbind(AIC(FH_1vlm), BIC(FH_1vlm))
7.  anova(FH_1vlm)
    說明
1.  和語法 with(plot(INVEST~ SAVING),data=myData) 相同。Figure
    1.3
2.  用迴歸係數以斜截式畫直線
```

3. 迴歸結果的 AIC 值
4. 迴歸結果的 BIC 值
5. 將 AIC 和 BIC 水平(左右)合併
6. 將 AIC 和 BIC 垂直(上下)合併
7. ANOVA 分析

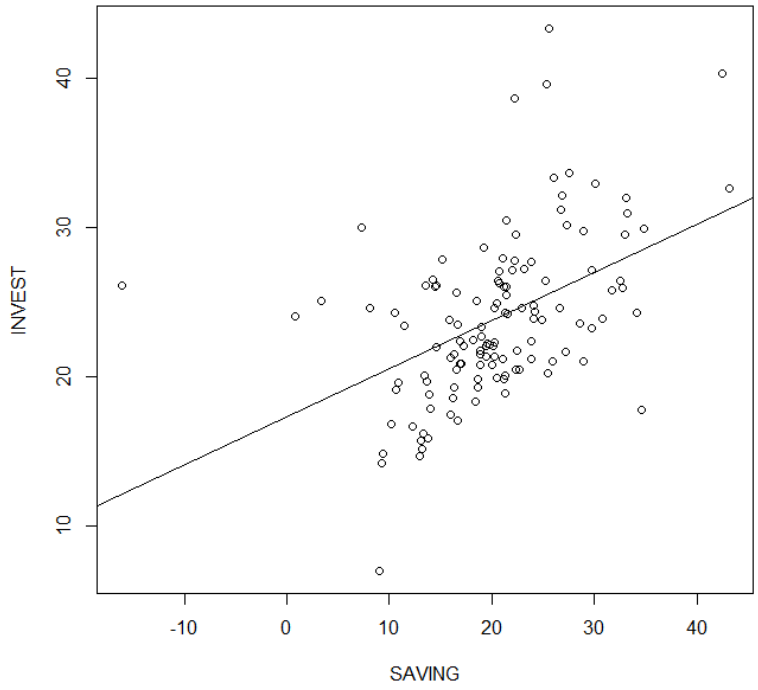

Figure 1.3

這裡我們使用了函數 with()，

```
with(plot(SAVING, INVEST), data=myData)
```

因為 **plot**()不像 **lm**(…, data=)裡面可以宣告資料集，如果單獨使用，就必須這樣

```
plot(myData$SAVING, myData$INVEST)
```

或

```
attach(myData)
plot(SAVING, INVEST)
```

上述 attach()的做法，要小心一點。如果整個工作空間只有一個檔案，變

數就不會有重複的名稱，那就可以一開始就用 **attach**() 宣告資料，這樣，像 **lm**()的函數，都不用在 argument 內宣告資料集。 我習慣使用 **with**()或宣告函數內資料集，而不使用全域 **attach**()函數。因為，當手上有許多跨國資料時，很多變數的名稱，因為慣用或通用，所以多是一樣的，例如，GDP。這樣的話，**attach**()多個檔案，會造成衝突。讀者還可以選擇自己習慣的作法。

接下來我們再來診斷殘差。

R Code. 迴歸後診斷(2)：迴歸殘差診斷圖

```
1. par(mfrow=c(2,2))
2. plot(FH_1vlm)
3. par(mfrow=c(1,1))
```

說明

1. 將圖形框分割成 2×2 四格的視窗.
2. Regression diagnosis plot。Figure 1.4
3. 將圖形框還原成 1×1 單格的視窗

par()是 parameter 的簡稱，有許多 arguments，如下

表 2. **par()**常用的宣告

Argument	說明
axes	Should axes be drawn?
bg	Background color
cex	Size of a print or symbol
col	Color
las	Orientation of axis labels
lty, lwd	Line type and line width
main, sub	Title and subtitle
mar	Size of margins
mfcol,mfrow	Array defining layout for several graphs on a plot
pch	Plotting symbol
type	Types
xlab, ylab	Axis labels

xlim, ylim	Axis ranges
xlog, ylog, log	Logarithmic scales

R 對 **lm()**函數的物件，繪製的殘差診斷圖有 6 個，內建為如下圖 4 個。

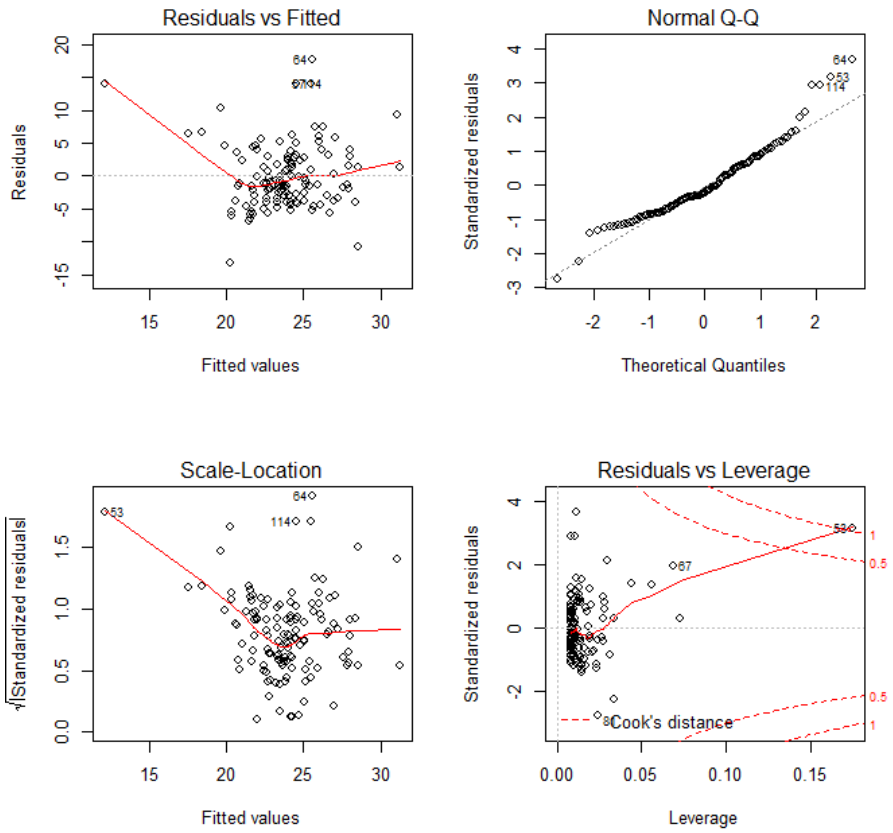

Figure 1.4

殘差診斷圖四個，必須詳細解讀。我們從左上角順時針一個一個來解釋。

第 1 個是左上角的「殘差與配適值」散佈圖，一個好的迴歸，這個圖會沒有任何的類型，例如正相關或負相關。圖形中的一條微微凹折的線就是「類型曲線」。好的模型，會和水平線很相近。如果這條類型曲線有明顯的斜率，則代表了模型的設定不完善，可能是有遺漏變數或者非線性關係。本圖代表模型配適有問題，且是離群值造成的。

第 2 個是右上角，診斷殘差是否是常態分佈。如果是常態分佈，則散佈點會和理論畫出的那一條直線重疊。如果散佈點的形狀像是 S 型或是香蕉型，就沒有常態的性質。線性模型沒有常態其實也沒有很嚴重，如果違反常態，就是顯著性

檢定的標準嚴格一點就可以，例如，用 1%的水準來看，不用 5%。當然，這還要看整體狀況。

第 3 個是第 1 個圖的正值版本：將殘差標準化後，取絕對值再開根號。這是用來判斷變異數是否是固定常數(同質變異)的方法。如果視同質變異，則畫出的補助線，會是水平的，如果有趨勢，就不是同質變異。我們的補助線就顯現出負斜率的趨勢。所以，告訴我們有異質變異。

最後一個是 Cook's Distance 圖，它將標準化殘差和槓桿(leverage)值放一起。圖形右上角的弧形虛線稱為 Cook contour，越接近它的點，對模型有最大的離群影響(outlier influence)。這個圖形告訴我們，第 53 國的行為最與眾不同，對模型估計影響最大，同時，他也有著最大的 leverage 值。

迴歸後的殘差診斷圖，會標出離群值的編號。Figure 1.4 右下角的診斷圖中，有一項 Cook's distance，公式如下：

$$D_i^2 = \frac{(\hat{y} - \hat{y}_i)'(\hat{y} - \hat{y}_i)}{k\hat{\sigma}^2}$$

k=解釋變數個數。Cook's distance 計算了每一個樣本點的資訊，圖內的標號是依照 Cook's distance 判斷出來的離群樣本(outliers)。如果原始資料的列名稱有字串，好比人名或地名，則這裡會直接顯示文字。

R Code. 迴歸預期與圖形分析

```
1. FH_pred=predict(FH_1vlm, interval="confidence")
2. with(plot(INVEST~SAVING),data=myData)
3. abline(FH_1vlm)
4. with(lines(FH_pred[,2] ~ SAVING, lwd=0.1, lty=4, col=2),
   data=myData)
5. with(lines(FH_pred[,3] ~ SAVING, lwd=0.1, lty=4, col=2),
   data=myData)
6. legend("topleft", c("regression line", "low", "upper"),
   lty=c(1,4,4), lwd=0.1, bty="n")
```

說明

1. 使用預期函數 predict()，計算預期值之信任區間，把結果存入暫存物件 FH_pred。。
2. 繪製資料散佈圖。Figure 1.5

3. 繪製配適線。這一條就是預期的均線
4. 以虛線畫上界
5. 以虛線畫下界
6. 在圖文框內，加上三條線的文字說明

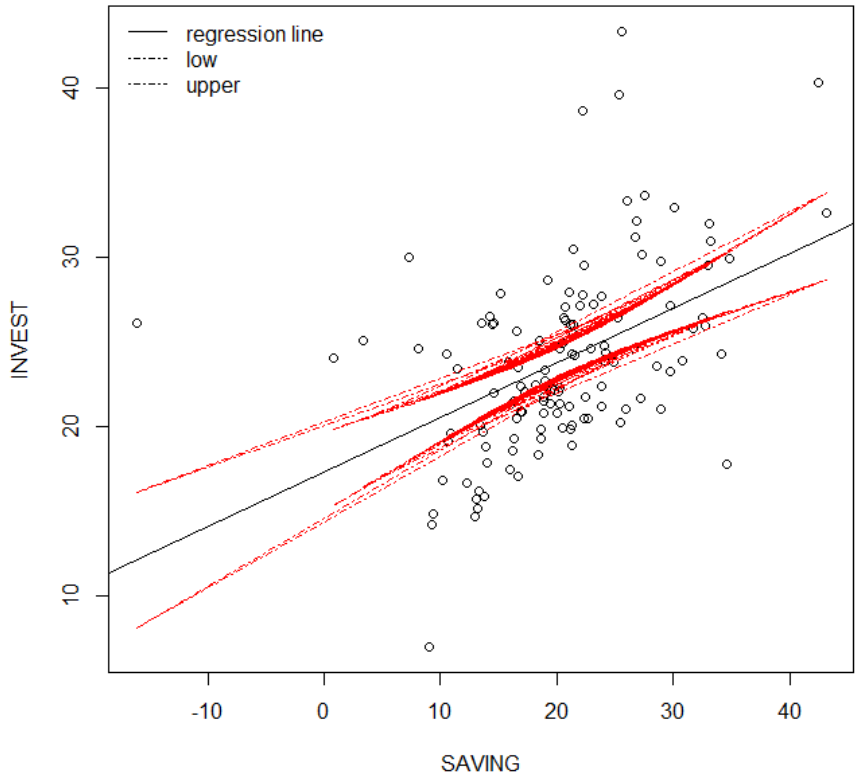

Figure 1.5

估計完畢，接下來是檢定參數的特定關係。好比，如果某理論認為儲蓄對投資的影響是 20%，則我們可以從事如下程式的檢定

R Code. 假設檢定

```
1. library(car)
2. linearHypothesis(FH_1vlm, "SAVING=0.2")
```

說明

1. 載入假設檢定模組 **car**
2. 檢定虛無假設 H_0: b_1=0.1

這個檢定結果如下

```
> linearHypothesis(FH_1vlm, "SAVING=0.2")
```

```
Linear hypothesis test

Hypothesis:
SAVING = 0.2

Model 1: restricted model
Model 2: INVEST ~ SAVING

    Res.Df RSS     Df  Sum of Sq      F     Pr(>F)
1    125    3045.6
2    124    2924.4  1    121.23        5.1406 0.0251 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

上面其實就是一個 ANOVA 的觀念。看最後的 P-值 Pr(>F) = 0.0251，可以知道：在 0.05 的顯著水準，F 檢定拒絕了係數為 0.2 的虛無假設。

第 3 節　連續變數線性複迴歸

3.1 相異兩個解釋變數

$$INVEST=a+b_1 \cdot SAVING+ b_2 \cdot CA$$

R Code. 複迴歸

```
1.  FH_2vlm=lm(INVEST~SAVING +CA, data=myData)
2.  summary(FH_2vlm)
3.  anova(FH_1vlm, FH_2vlm)
4.  linearHypothesis(FH_2vlm, "SAVING+CA=0")
```

說明

1. 執行迴歸，把結果存入暫存物件 FH_2vlm。
2. 結果摘要
3. ANOVA 比較單變數和雙變數
4. 線性假設檢定 H_0: $b_1+ b_2 =0$

3.2 多項式迴歸-- 解釋變數的冪次方

對於偏離線性關係，一個最簡單的情況，就是解釋變數的冪次進入迴歸。如果是平方，就有 U 型的二次曲線。

$$INVEST = a + b_1 \cdot SAVING + b_2 \cdot SAVING^2$$

R Code. 多項式迴歸

```
1. FH_2v2m=lm(INVEST~SAVING+I(SAVING^2), data=myData)
2. summary(FH_2v2m, corr =TRUE)
```

說明

1. 執行二次迴歸，把結果存入暫存物件 FH_2v2m。
2. 結果。corr=TRUE 會計算估計係數的相關性。

在 **lm()**內，要處理同樣一個解釋變數的平方，好比多項式迴歸的作法，直接平方是不行的，例如，

lm(INVEST~SAVING+SAVING^2)

是不行的。必須用隔離函數 **I()**才能發揮作用[3]。上式的結果如下，

```
> summary(FH_2v2m, corr =TRUE)

Coefficients:
             Estimate    Std. Error  t value Pr(>|t|)
(Intercept) 20.237217   1.418935    14.262   < 2e-16 ***
SAVING      -0.060552   0.122925    -0.493   0.623178
I(SAVING^2) 0.010269    0.002978   3.449     0.000772 ***
---

Residual standard error: 4.656 on 123 degrees of freedom
Multiple R-squared:  0.2904, Adjusted R-squared:  0.2788
F-statistic: 25.16 on 2 and 123 DF,  p-value: 6.896e-10

Correlation of Coefficients:
           (Intercept) SAVING
SAVING      -0.86
I(SAVING^2)  0.59       -0.91
```

我們看最後面故係數的相關係數，SAVING 和 SAVING^2 兩個解釋變數係

[3] 隔離的英文 **I**nsulation，取其字首。

數的相關係數高達 –0.9：這意味著一個係數大，另一個就會小。為了處理這樣的問題，我們可以用正交多項式迴歸 *orthogonal polynomial regression*, 函數 **poly()** 可以取代解釋變數的輸入。

讀者可以練習執行

```
summary(lm(INVEST~poly(SAVING,2),data=myData),corr =TRUE)
```

比較看一看，和上述結果差異有多大。

第 4 節 因子和交互效果

4.1 因子迴歸

因子變數是說這筆資料不是連續變數，而是虛擬變數。如果原始資料就是字串，那 R 會自動辨認其為因子。如果是數值，則需要宣告因子轉換 **factor()**。如下例，

$$INVEST=a+b_1 \cdot SAVING+b_2 \cdot CA+b_3 \cdot Group$$

變數 Group 是的文字，我們看程式

R Code. 具因子變數之複迴歸

```
1.  FH_3vlm=lm(INVEST ~ SAVING+Group, data=myData)
2.  summary(FH_3vlm)
3.  linearHypothesis(FH_3vlm, "SAVING=0")
4.  library(lmtest)
5.  waldtest(FH_3vlm, .~. -SAVING)
```

說明

1. 執行迴歸，把結果存入暫存物件 FH_3vlm。
2. 結果摘要
3. 線性假設檢定 H_0: $b_1=0$
4. 載入 Wald test 模組 **lmtest**
5. Wald test H_0: $b_1=0$

這個估計結果如下：

```
> summary(FH_3vlm)
```

```
Call:
lm(formula = INVEST ~ SAVING + Group, data = myData)

Residuals:
     Min       1Q   Median       3Q      Max
-13.1989  -2.9183  -0.5072   2.8354  17.0393

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                             14.8400     1.5438   9.613  < 2e-16 ***
SAVING                                   0.3712     0.0569   6.524 1.75e-09 ***
GroupCentral and Eastern Europe          2.9960     1.6264   1.842  0.06794 .
GroupCommonwealth of Independent States  5.0033     1.6181   3.092  0.00248 **
GroupDeveloping Asia                     1.9885     1.3841   1.437  0.15342
GroupLatin America and the Caribbean     1.7428     1.2219   1.426  0.15640
GroupMiddle East and North Africa        0.1560     1.3604   0.115  0.90892
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.722 on 119 degrees of freedom
Multiple R-squared:  0.2938,    Adjusted R-squared:  0.2581
F-statistic: 8.249 on 6 and 119 DF,  p-value: 1.79e-07
```

4.2 交互效果(Interaction terms)

交互效果是最簡單的線性模式擴充，也是一個描述係數非常數的最簡單方法。如下

$$INVEST=a+b_1 \cdot SAVING+b_2 \cdot Group+b_4 \cdot (SAVING \cdot Group)$$

上式之中，原本是由下式出發

$$INVEST=a+b_1 \cdot SAVING$$

但是，當研究者考慮係數 b_1 是否會因為其他變數的高低，例如 Group，有所不同時，則隱含了這樣的另一條方程式

$$b_1 = c + d \cdot Group$$

這樣的方程式因為 b_1 不是資料，所以是無法估計的，因此必須帶入原式

$$INVEST=a+(c+ d \cdot Group) \cdot SAVING$$

$$= a+c \cdot SAVING + d \cdot (SAVING \cdot Group)$$

其餘都是符號問題。這就是 interaction term 的由來。這樣才能解釋所估計出來的係數。我們的例子中，Group 是文字變數，所以是群組的意義。Interaction 項，

通常是由兩個連續變數所構成，如前例的平方項。有興趣的讀者，可以將前面的平方項，用一樣的方式，簡單推導一番。

R Code.

```
1. FH_4vlm=lm(INVEST~SAVING+Group+Group*SAVING,
   data=myData)
2. summary(FH_4vl m)
   說明
1. 略
```

```
> summary(FH_4vlm)

Call:
lm(formula = INVEST ~ SAVING + Group + Group * SAVING, data = myData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.8987 -2.7839 -0.8556  2.8868 15.4483

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                    13.9827     3.0406   4.599 1.11e-05 ***
SAVING                                          0.4086     0.1281   3.190  0.00184 **
GroupCentral and Eastern Europe                 6.3508     4.8035   1.322  0.18878
GroupCommonwealth of Independent States         9.6650     3.7991   2.544  0.01230 *
GroupDeveloping Asia                           -8.4189     4.2872  -1.964  0.05200 .
GroupLatin America and the Caribbean            4.0163     3.9238   1.024  0.30820
GroupMiddle East and North Africa               3.5312     4.3658   0.809  0.42030
SAVING:GroupCentral and Eastern Europe         -0.1833     0.2410  -0.761  0.44843
SAVING:GroupCommonwealth of Independent States -0.2503     0.1660  -1.508  0.13437
SAVING:GroupDeveloping Asia                     0.4648     0.1800   2.582  0.01108 *
SAVING:GroupLatin America and the Caribbean    -0.1213     0.1894  -0.640  0.52320
SAVING:GroupMiddle East and North Africa       -0.1443     0.1796  -0.803  0.42339
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.425 on 114 degrees of freedom
Multiple R-squared:  0.4059,    Adjusted R-squared:  0.3486
F-statistic: 7.081 on 11 and 114 DF,  p-value: 4.707e-09
```

第 1 步的其他輸入方式如下：

lm(INVEST ~ SAVING + Group + Group:SAVING, data=myData)

lm(INVEST ~ SAVING + Group + Group/ SAVING, data=myData)

lm(INVEST ~ (Group + SAVING)^2, data=myData)

第 3 種作法，在多個變數的交叉互動下，相當好用。唯一會受影響的是估計結果的參數排序，R 會自動處理重複問題，相當方便。

第 5 節 迴歸診斷

5.1 異質殘差檢定

古典迴歸假設迴歸殘差是同質變異(homoskedasticity)，一個同質變異的迴歸，概念上是這樣：假設一條樣本數為 1,000 的迴歸： Y=a+bX+u，u 是殘差項。變異數是殘差的平方和，同質變異就是 var(u)=σ^2。有時更廣義的寫法，是用分配表示：u~N(0, σ^2)。所謂的同質變異，就是說，1,000 個 e 的樣本殘差：任取若干子樣本，此樣本計算的變異數，在統計上皆相等。反之，異質變異就是說，依此計算的變異數，統計上不相等。

異質變異的問題，如果在時間序列資料，可以用某時間前後區分出兩個不同的變異數，計量學者稱之為結構變動(structural breaks)，著名的 Chow 檢定和其他結構變動檢定，多是基於這樣的思路。

如果資料沒有考慮殘差異質變異的性質，OLS 的估計往往會過度顯著。因此，檢定殘差是否具有異質性就相當重要。文獻上對於異質殘差的檢定有許多，時間序列資料的稱為 ARCH，我們就不在這裡談。我們簡介兩種：Breusch-Pagan test 和 Goldfeld-Quandt。

Goldfeld-Quandt 檢定，先估計一 k 個變數之迴歸式：

$$Y_i = a + bX_i + u_i$$

Goldfeld-Quandt 檢定的虛無假設為資料是同質變異。要檢定此虛無假設，假設變異數σ_i^2和資料 X_i 正相關：

$$\sigma_i^2 = \sigma^2 X_i^2$$

σ^2 為一常數。如果上式為真，則當 X 的值越大，則其變異數越大。Goldfeld-Quandt 則設計如下檢定程序：

第 1 步。將原始資料依照 X 小大排序，X 由小排到大。

第 2 步。將中間 c 個樣本移除，留下雙端兩群資料，這兩群樣本數皆相等。

第 3 步。對這兩群資料配適原 LS 迴歸。

第 4 步。令第 1 群的殘差平方和為 RSS_1，自由度=$\frac{n-c}{2}-k$；第 2 群為 RSS_2，自由度=$\frac{n-c}{2}-k$。依此，Goldfeld-Quandt 計算

$$\lambda = \frac{\frac{RSS_2}{\text{df}}}{\frac{RSS_1}{\text{df}}} = \frac{RSS_2}{RSS_1}$$

df 為自由度。

如果虛無假設是正確的，這兩群資料的迴歸變異數會幾乎相等。在常態分配假設下，λ 會是 F 分配。

Goldfeld-Quandt 檢定需要挑一個解釋變數來排序，因此，挑哪一個來排序的結果就很關鍵。Breusch-Pagan 的檢定，則對所有的解釋變數蒐集資訊。假設某三變數迴歸：

第 1 步。$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{3i} + e_i$

第 2 步。$e_i^2 = c + d_1 X_{1i} + d_2 X_{2i} + d_3 X_{3i} + \varepsilon_i$

第 3 步。檢定同質變異　$H_0 : d_1 = d_2 = d_3 = 0$

第 3 步的檢定量為 $\frac{1}{2}(ESS) \sim \chi^2_{m-1}$　m 是第 2 步解釋變數的個數，ESS 是第 2 步迴歸殘差的平方和。

R Code. 殘差異質性檢定(heteroskedasticity)

```
1.  gqtest(FH_2vlm, order.by=~SAVING, data=myData)
2.  bptest(FH_2vlm, data=myData)
```

說明

1. 執行 Goldfeld-Quandt test，以 SAVING 排序。
2. 執行 Breusch-Pagan test

```
> gqtest(FH_2vlm, order.by=~SAVING, data=myData)
Goldfeld-Quandt test

data:  FH_2vlm
GQ = 0.6445, df1 = 60, df2 = 60, p-value = 0.9542

> bptest(FH_2vlm, data=myData)
studentized Breusch-Pagan test
```

```
data:  FH_2vlm
BP = 21.5234, df = 2, p-value = 2.12e-05
```

5.2 迴歸函數形式判定

迴歸方程式如果函數形式設定錯誤，如何判定？文獻上有許多檢定方法。Ramsey(1969)[4]的 **RESET** 檢定是較早提出來的一種。假設某雙變數迴歸：

第 1 步。執行迴歸：$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + e_i$

令迴歸配適值為 $\hat{Y}_i$ ，且相關係數為 R_A^2

第 2 步。執行迴歸：$Y_i = a + b_1 X_{1i} + b_2 X_{2i} + b_3 \hat{Y}_i^2 + b_4 \hat{Y}_i^3 + v_i$ ，相關係數為 R_B^2

第 3 步。$F = \dfrac{R_B^2 - R_A^2}{1 - R_A^2} \dfrac{\text{第1步之自由度}}{\text{第2步新增之解釋變數個數}}$

如果 F 值大於 0.05 的臨界值，則拒絕了模型函數設定形式是完整的假設。這個檢定一般稱為 **RESET**。使用這個統計量時，必須對被解釋變數和殘差的關係，做一些檢視。如果有曲度性(curvilinear)關係，則可以 **RESET** 檢定配適關係。

Rainbow test[5] 則和 **RESET** 與的類似，且使用了 Goldfeld-Quandt 的排序方法來做非線性區度檢定。Harvey and Collier(1977)[6] 提出的統計量是檢定 Recursive residuals，一般視為 CUSUM 的延伸：漸進上，如果模型的函數設定是正確的，則 Recursive residuals 的平均數為 0。這些檢定的技術層次，我們就不細說，直接看 **R** 如何執行就可以。

R Code. 函數形式檢定(functional form)

```
1.  resettest(FH_2vlm, data=myData)
2.  raintest(FH_2vlm, order.by=~SAVING, data=myData)
3.  harvtest(FH_2vlm, data=myData)

    說明
1.  執行 RESET test
2.  執行 Rainbow test
```

[4] Ramsey J.B.(1969) Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society*, series B, vol. 31, pp.350-371

[5] Utts J.M. (1982) *The Rainbow Test for Lack of Fit in Regression*. *Communications in Statistics - Theory and Methods* 11, 1801–1815.

[6] Harvey, A. and Collier G. (1977). Testing for Functional Misspecification in Regression Analysis. *Journal of Econometrics*, 6, 103-119.

```
3.  執行 Harvey-Collier test
```

套件 **lmtest** 有一些基本的檢定量，如果需要更多，讀者可以參考套件 **fRegression** 內所附帶的迴歸後檢定。

5.3 穩健共變異數的異質變異修正

Robust Covariance for heteroskedasticity

我們介紹當殘差出現異質變異，則需要修正這種異質性，再重新計算的共變異數，稱為穩健共變異數(Robust Covariance)。這裡有兩個問題須要釐清處：

(1) 此處所稱的穩健共變異數，和文獻上處理離群值(outiers)的 Robust Regression 不同。例如，分量迴歸，L-和 M-估計法。這些估計方法，會重新估計 LS 模型的方程式，全體參數都會重新計算。

(2) 處理異質變異的 Robust Covariance 只會修正標準差，不會改變原先 LS 估計的參數。如果原來 OLS 的係數檢定很顯著，一般而言，修正異質變異的穩健標準差，會讓原先的檢定結果更穩健，更保守，也就是降低顯著性；例如，p-value 從 0.002 變成 0.04。

LS 假設殘差為同質，共變異數的估計式為 $\hat{\sigma}^2(X'X)^{-1}$。當殘差有異質變異時，LS 估計的變異數就不再正確，一般的情況會過度膨脹 t 統計量，而使的所估計的係數都很顯著。 如果殘差異值變異的特徵，有某種類型或規律，例如，分群分組就可以辨認出來，則我們可以由群組的方向去修正。有時後分組變數是產業，則控制產業的集群(cluster)，就可以降低影響；有時後是時間點，如前述之結構變動，分期之後，就可以解決。

像這樣的問題，在於如果殘差是異質的，但是我們將所有殘差視為同質，則計算期望值時，每個樣本點的權重都一樣，所以會導致變異數低估，顯著性會高估。當然，某些特殊情況會例外。

考慮異質變異和序列相關的一般形式則為

$$(X'X)^{-1}X'\Omega X(X'X)^{-1}$$

式中 $\Omega=\text{diag}(\omega_1, \omega_2, \omega_3, \ldots \omega_N)$

ω矩陣是由殘差所估計的共變異數矩陣，它是一個方陣，和係數對應的

變異數Ω則是它的主對角線成分。R 模組程式分成五種

Constant: $\omega_i = \hat{\sigma}^2$

HC0: $\omega_i = \hat{\varepsilon}_i^{\,2}$

HC1: $\omega_i = \frac{N}{N-k}\hat{\varepsilon}_i^{\,2}$

HC2: $\omega_i = \frac{\hat{\varepsilon}_i^{\,2}}{1-h_{ii}}$

HC3: $\omega_i = \frac{\hat{\varepsilon}_i^{\,2}}{(1-h_{ii})^2}$

HC4: $\omega_i = \frac{\hat{\varepsilon}_i^{\,2}}{(1-h_{ii})^{\delta_i}}$

k 是解釋變數的個數，h_{ii} 為ω矩陣主對角線成分，$\bar{h}$ 為其平均值，$\delta_i = \min\left\{4, \frac{h_{ii}}{\bar{h}}\right\}$。

對於考慮殘差異質性和序列相關的穩健共變異數，R 主要是 **sandwich**。穩健共變異數的考慮，不會改變所估係數的值，只會修正共變異數，也就是標準差。R 做穩健共變數估計選項，非常簡易，只需要將 **lm()**的迴歸物件用 vcovHC 處理即可重新計算共變異數矩陣。請看以下程式碼：

R Code.　Robust Covariance

```
1.  coeftest(FH_2vlm)
2.  library(sandwich)
3.  coeftest(FH_2vlm, vcovHC)
4.  coeftest(FH_2vlm, vcov = vcovHC(FH_2vlm,
    type = "HC1"))
```

說明

1. 無修正之估計結果。
2. 載入檢定套件 **sandwich**
3. **vcovHC** 計算穩健共變異數，修正方法內定 HC0。
4. **vcovHC** 計算穩健共變異數，修正方法選用 HC1。

vcovHC()的一般形式，有很多選項。我們將有修正與無修正的結果並列比較，如下：

```
> coeftest(FH_2vlm)

t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.90313    1.28496   5.372 3.74e-07 ***
SAVING       0.73175    0.05420  13.500  < 2e-16 ***
CA          -0.68305    0.06291 -10.858  < 2e-16 ***
---

> coeftest(FH_2vlm, vcovHC)

t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.9031     4.7774   1.445 0.151010
SAVING       0.7317     0.2035   3.597 0.000465 ***
CA          -0.6831     0.1554  -4.397 2.35e-05 ***
---
```

比較發現，考慮穩健共變異數後的標準差會比較大，使的顯著性檢定的結果趨於保守，不會太輕易就拒絕虛無假設，也使的迴歸結果較為可靠。

第 6 節 時間序列迴歸：dynlm()

6.1 時間序列線性迴歸

如果我們要執行時間序列資料的線性迴歸，例如 $Y_t = a + b_0 X_t$，如果殘差沒有序列相關修正問題時，這個迴歸基本上可以用 **lm()**處理。但是，如果我們的迴歸方程式有落後期，如下

$$Y_t = a + b_0 X_t + b_1 X_{t-1} + b_2 X_{t-2}$$

則 **lm()**就會有一些麻煩，因為落後的資料會出現缺值，所以，必須另外處理。我們看下面程式例子：台灣經濟成長率和自己的落後期迴歸

我們發現，這個殘差圖沒有時間刻度，雖然第 5 步就給予了時間框架，但是 **lm()**迴歸後，這些附上去的時間性質都會消失。

使用套件 **dynlm** 可以將上面程式的第 6、7 步省略為 1 步，而且會保留時間序列的刻度。看下面例子

R Code.　Dynamic Linear Model

```
1.  myTS=read.csv("taiwan3v.csv")
2.  head(myTS)
3.  myTS_Dates=ts(myTS, start=c(1962,1), freq=4)
4.  head(myTS_Dates)
5.  library(dynlm)
6.  dynlm_twn=dynlm(Growth ~ L(Saving) + L(Saving,
    4),data=myTS_Dates)
7.  summary(dynlm_twn)
8.  plot(dynlm_twn$resid)
```

說明

1. 讀取資料
2. 看資料前 6 筆
3. 賦予時間框架
4. 看新資料前 6
5. 載入套件 **dynlm**
6. 對資料執行 **dynlm**() 迴歸
7. 並將結果 **summary**() 出來
8. 畫殘差圖。Figure 1.6

```
> summary(dynlm_twn)

Time series regression with "ts" data:
Start = 1963(1), End = 2010(4)

Coefficients:
             Estimate        Std. Error   t value     Pr(>|t|)
(Intercept) 0.020824     0.004312   4.829     2.82e-06 ***
L(Growth)    0.842932     0.048173  17.498     < 2e-16 ***
L(Growth,4) -0.129412    0.048167  -2.687     0.00786 **
---
```

```
Residual standard error: 0.02546 on 189 degrees of freedom
Multiple R-squared:  0.6338,Adjusted R-squared:  0.6299
F-statistic: 163.5 on 2 and 189 DF,  p-value: < 2.2e-16
```

我們再看殘差如下圖，時間刻度保持的很好，直接以時間序列的繪圖方式將散佈點接起來。

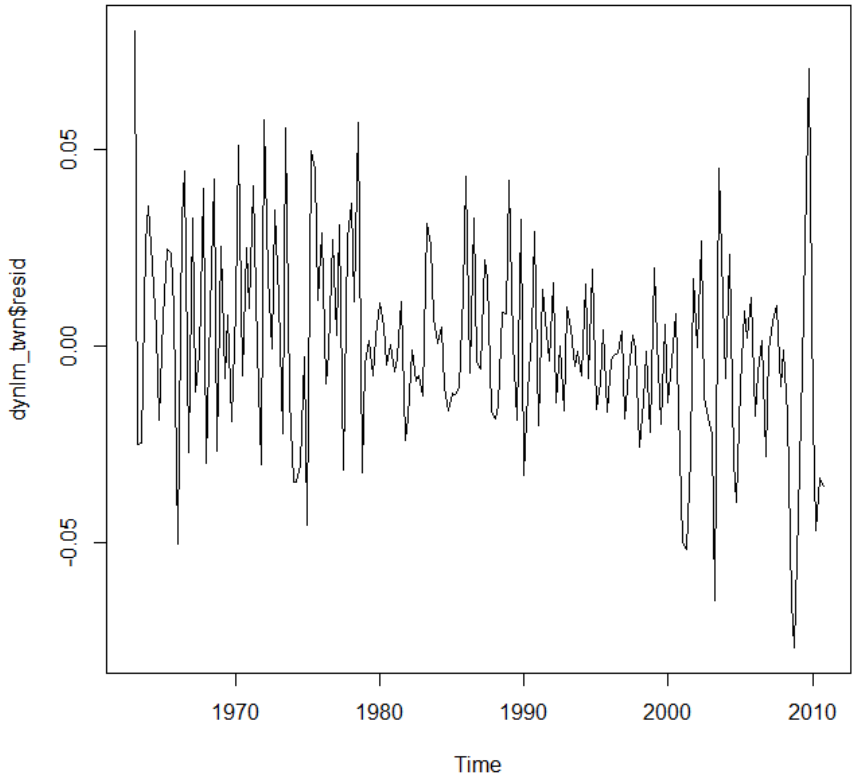

Figure 1.6

dynlm()是一個很好用的時間序列迴歸函數。讀者可以用 **lm()**做做看，在 **lm()**內，很多時間序列功能都無法使用。

第 7 節　線性重合檢定

在複迴歸模型，因為解釋變數有多個，所以解釋變數之間如果彼此有強烈線性相關時，我們稱為共線性 (collinear) 問題。當線性重合存在時，解釋變數之間的正交性質就會受到影響，最極端的行情況就是產生奇異問題(singularity)，此時，估計出來的係數不穩定，也非唯一。檢查共線性的方法有以下多種：

1. 計算解釋變數之間成對的相關係數矩陣。
2. 將解釋變數之間互相迴歸。如果有很高的 R^2，則線性重和就存在。

3. 令 X 代表解釋變數矩陣，計算矩陣 X^TX 的特徵值(eigenvalues)，將特徵值從大到小排序後，特徵值相對偏小的變數會有線性重合問題。假設最大特徵值為λ_1，最小特徵值為λ_p，統計學者建構了一個指數 condition number 如下:

$$\sqrt{\frac{\lambda_1}{\lambda_p}}$$

這個數字若大於 30 被視為是夠大，也就是矩陣出現 singularity。

4. 由參數因為線性重合存在時，會導致所估計參數的變異數龐大，可由下式看出

$$\text{var}(b_j) = \frac{1}{1-R_j^2}\frac{s^2}{\sum_j (x_{jj}-\bar{x}_j)^2}$$

第 j 個自變數與其他自變數迴歸的偏相關係數接近 1 時，第 1 項分母會趨向 0，所以此變異數會膨脹。因此，統計學者定義 variance inflation factor(VIF)如下：

$$VIF_j = \frac{1}{1-R_j^2}$$

VIF 越大，代表線性重合的問題越嚴重。**R** 計算 VIF 有套件 **DAAG** 和套件 **faraway**，都稱為 **vif**()。**faraway** 必須先宣告 model.matrix 再剔除截距項，較為囉唆。本書建議使用 **DAAG** 內的較為簡便。

R Code. VIF 檢定

```
1. data("CPS1985", package="AER")
2. head(CPS1985)
3. g=lm(wage ~ education + experience + age, data=CPS1985)
4. summary(g)
5. DAAG:::vif(g)
6. library(DAAG)
7. vif(g)
```

說明

1. 自套件 AER 載入其資料 CPS1985
2. 看一看 CPS1985 的前 6 筆資料，確認前四筆是數值，其餘是字串
3. 執行三變數線性迴歸，並將迴歸後結果存入物件 g。

4. 迴歸結果摘要
5. Call 模組 DAAG 內的 **vif**()函數，計算三個解釋變數的 **vif** 值
   → 這個方法，不需要載入模組，可以避免只是小小使用一個函數，就載入整個模組進來佔據記憶體。下面就是載入模組再檢定的例子。
6. 載入套件 **DAAG**
7. 計算三個解釋變數的 **vif**()值。

估計結果摘要如下

```
> summary(g)
Coefficients:
             Estimate    Std. Error  t value   Pr(>|t|)
(Intercept) -4.76987    7.04271     -0.677      0.499
education    0.94833    1.15524     0.821       0.412
experience   0.12756    1.15571     0.110       0.912
age          -0.02241   1.15475     -0.019      0.985
Residual standard error: 4.604 on 530 degrees of freedom
Multiple R-squared: 0.202,     Adjusted R-squared: 0.1975
F-statistic: 44.73 on 3 and 530 DF,  p-value: < 2.2e-16
```

根據上面的估計結果，Adjusted R-squared＝0.1975，已經接近 20%，但是，解釋變數的顯著性極低。沒有一個在 10%是顯著異於 0 的。這一般就隱含了線性重和。

vif()的結果如下

```
> vif(g)
  education    experience     age
  229.5738     5147.9190  4611.4008
```

若 VIF<10 的解釋變數可以接受，所以在實務上，我們會逐次剔除 VIF 大於 10 的解釋變數。這三個解釋變數的 VIF 值都相當龐大，以 education 為例，解釋如下： $\sqrt{229.5738}=15.152$，我們這樣解釋：教育目前的標準差(1.155)，比無線性重合時高出 15 倍多。檢視其餘兩個，我們發現所有解釋變數之間都有極大的 VIF，所以，這個線性迴歸的估計和檢定均不可靠。克服線性重合的方法是依照理論剔除多餘的解釋變數，或利用 ridge regression，**R** 有 **lm.ridge()**可以處理，或工具變數方法(Two-stage LS)，此處我們就不詳述。