4 具影響力觀察值之偵測

4.1 傳統線性迴歸中的偵測方法

最小平方法的缺點是估計的結果容易受到某一些特殊觀察值的影響，在本章中，我們將引進傳統線性迴歸中偵測具影響力觀察值的概念與技巧，在模糊線性迴歸中架構出一套找出具影響力觀測值的方法。

在傳統迴歸中，若某個觀察值的自變數或應變數部份明顯偏離其他自變數與應變數的值，此時我們稱這樣的觀察值爲離群值。若某一觀察值被移除後，會造成最小平估計有巨大的改變，我們稱此觀察值爲具影響力的觀察值。

在自變數離群值的部分，大都以槓桿值$h_{ii} = x_i'(X'X)^{-1}x_i$作爲判斷，其中$0 \le h_{ii} \le 1$，$h_{ii}$越大，代表第$i$筆資料越有可能是離群值。若要偵測應變數部分的離群值，我們可利用殘差值$e_i = y_i - \hat{y}_i$做爲偵測基準，其中$\hat{y}_i$爲第i筆觀察值的預測值。當$|e_i|$越大，代表第i筆資料越有可能是離群值。或者可利用

$$e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1 - h_{ii}}$$

來作判斷，其中$\hat{y}_{(i)}$爲扣除第i筆資料後，再對第i筆資料作出的預測值。若$|e_{(i)}|$值越大，代表第i筆資料越有可能是離群值。

但是造成迴歸模型有巨大改變的，除了有可能是離群值外，還有可能是一些具影響力的觀察值。而這些具影響力的觀察值，在傳統迴歸中，大都使用Cook距離

$$CD_i \equiv \frac{\|\hat{Y} - \hat{Y}_{(i)}\|^2}{ps^2} = \frac{e_i^2}{ps^2}\frac{h_{ii}}{(1 - h_{ii})^2}$$

來判斷，其中$\hat{Y}_{(i)}$爲扣除第i筆資料後的預測值向量，p爲參數個數，$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$爲均方差(mean square error)。若CD_i值越大，代表第i筆資料越有可能是具影響力的觀察值。利用Cook距離的優點之一是，無論自變數或應變數的測量單位爲何，最後都不會影響CD_i的大小。

4.2 模糊線性迴歸中的偵測方法

在模糊環境下，令e_i爲模糊數y_i與$\hat{y}_i$間之距離，$e_{(i)}$爲y_i與$\hat{y}_{(i)}$間之距離。本節中我們僅在模型I的架構下進行討論，推導出有關e_i,$e_{(i)}$與CD_i的重要公式，作爲偵測離群值與有影響力觀察值之用。對模型II而言，不論是利用簡單距離公式或Yang和Ko的距離公式，雖然皆可獲得最小平方估計，但卻無法導出如同模型I所能獲致有關e_i、$e_{(i)}$與CD_i的結果。至於最複雜的模型III，則更無法處理了。

在簡單距離公式下可導出有關e_i，$e_{(i)}$之結果如下(計算過程詳見附錄二)：

$$\begin{aligned} e_i^2 &= (c_i - x_i\hat{a})^2 + (s_i - x_i\hat{r})^2 \\ &= (e_i^c)^2 + (e_i^s)^2 \qquad (4.1) \\ e_{(i)}^2 &= (c_i - x_i\hat{a}_{(i)})^2 + (s_i - x_i\hat{r}_{(i)})^2 \\ &= (\frac{e_i}{1-h_{ii}})^2 \qquad (4.2) \end{aligned}$$

其中$e_i^c = c_i - x_i\hat{a}$爲資料中心的殘差，$e_i^s = s_i - x_i\hat{r}$爲資料分展度的殘差，且$\hat{a}$與$\hat{r}$定義如(2.4)。

同理，在Yang和Ko的距離公式下可導出有關e_i，$e_{(i)}$之結果如下(計算過程詳見附錄二)：

$$\begin{aligned} e_i^2 &= d_{LR}^2(y_i, \hat{y}_i) \\ &= 3(e_i^c)^2 + 2l^2(e_i^s)^2 \qquad (4.3) \\ e_{(i)}^2 &= d_{LR}^2(y_i, \hat{y}_{(i)}) \\ &= 3(\frac{e_i^c}{1-h_{ii}})^2 + 2l^2(\frac{e_i^s}{1-h_{ii}})^2 \\ &= (\frac{e_i}{1-h_{ii}})^2 \qquad (4.4) \end{aligned}$$

由(4.2)與(4.4)式可以發現，$e_{(i)}$與e_i間的關係和傳統迴歸分析的結果相同。並且在功能上與傳統迴歸一樣，即$e_{(i)}$的值越大，代表扣除掉第i筆資料後，在當筆資料上的殘差改變越大，則該筆資料越有可能是造成整個模型改變的關鍵。

但是要在模糊環境下導出類似Cook距離的公式，若能先定義出模糊向量間的距離公式，可使整個處理過程較爲方便。令$F_{LR}(\Re)$是所有對稱LR型模糊數的集合，而$\tilde{F}_{LR}(\Re) = \{(X_1, ..., X_p)' | X_i \in F_{LR}(\Re)\}$是所有對稱$LR$型模糊數所構成p維向量的集合。藉由$F_{LR}(\Re)$上的距離測度，我們可以定義一個$\tilde{F}_{LR}(\Re)$上的距離測度，其方法如下：

引理**4.1** *設$d : F_{LR}(\Re) \times F_{LR}(\Re) \to \Re$是距離測度，對任意兩個模糊向量$\mathbb{X} = (X_1, X_2, ..., X_p)'$，$\mathbb{Y} = (Y_1, Y_2, ..., Y_p)' \in \tilde{F}_{LR}(\Re)$，定義*

$$\tilde{d}_{LR}(\mathbb{X}, \mathbb{Y}) = \sqrt{\sum_{i=1}^{p} d^2(X_i, Y_i)} \tag{4.5}$$

則$\tilde{d}_{LR}$亦是$\tilde{F}_{LR}$上的距離測度；若d是完備距離測度時，$\tilde{d}_{LR}$亦為完備距離測度。(詳細證明見附錄三)

當d是簡單距離測度時，定義Cook距離CD_i如下：

$$CD_i \equiv \frac{\tilde{d}^2_{LR}(\hat{\mathbb{Y}}, \hat{\mathbb{Y}}_{(i)})}{ps^2} = \frac{\|X\hat{a} - X\hat{a}_{(i)}\|^2 + \|X\hat{r} - X\hat{r}_{(i)}\|^2}{ps^2}$$

我們可進一步的證明(詳見附錄四)

$$CD_i = \frac{1}{ps^2} \frac{e_i^2 h_{ii}}{(1 - h_{ii})^2} \tag{4.6}$$

此時

$$s^2 = \frac{1}{n-p} \sum_{i=1}^{n} e_i^2 \quad 且 \quad e_i^2 = (e_i^c)^2 + (e_i^s)^2$$

當d是Yang和Ko的距離測度時，定義Cook距離CD_i如下：

$$\begin{aligned} CD_i &\equiv \frac{\tilde{d}^2_{LR}(\hat{\mathbb{Y}}, \hat{\mathbb{Y}}_{(i)})}{ps^2} \\ &= \frac{1}{ps^2}\{\|X\hat{a} - X\hat{a}_{(i)}\|^2 + \|(X\hat{a} - lX\hat{r}) - (X\hat{a}_{(i)} - lX\hat{r}_{(i)})\|^2 \\ &\quad + \|(X\hat{a} + lX\hat{r}) - (X\hat{a}_{(i)} + lX\hat{r}_{(i)})\|^2\} \end{aligned}$$

我們可進一步的證明(詳見附錄四)

$$CD_i = \frac{1}{ps^2} \frac{e_i^2 h_{ii}}{(1 - h_{ii})^2} \tag{4.7}$$

此時

$$s^2 = \frac{1}{n-p}\sum_{i=1}^{n} e_i^2 \quad \text{且} \quad e_i^2 = 3(e_i^c)^2 + 2l^2(e_i^s)^2$$

我們必須注意的是：雖然(4.6)與(4.7)的CD_i公式在符號外表上是一樣的，但實質上是不相同的，因爲兩式中的e_i^2與s^2所代表的內容根本不一樣。一般來說，(4.7)中的s^2值大於(4.6)中的s^2值，因此由(4.6)所計算出的Cook距離會大於(4.7)所計算出的結果。

無論是在簡單距離測度下，或是Yang和Ko的距離測度下，由(4.6)與(4.7)的結果，可以看出CD_i的值皆受槓桿值h_{ii}以及殘差值e_i的影響，而這樣的結論和傳統迴歸分析上偵測有影響力觀測值的結論是一致。

在模型I的架構下我們可以利用完整資料的模糊迴歸結果，計算$e_{(i)}, CD_i$的值。而在模型II的架構下，我們無法導出如(4.1)∼(4.4)的結果，因此僅能藉著逐筆刪除資料，重新做迴歸的方式計算出$e_{(i)}, CD_i$等值。