殘差的獨立性與相關性

洪文洋　游添旭　鄭宇欽
德霖技術學院通識中心講師

摘要

在迴歸分析中，採用殘差圖去檢測迴歸模型，其原由都是根據經驗法則，並未探究殘差 e 與預估迴歸式 $\hat{Y}$ 之間的相關性，使得此檢測方法缺乏理論基本架構的支持。因此，本篇論文將證明殘差 e 不是獨立之隨機變數，並進一步推導殘差 e 與預估迴歸式 $\hat{Y}$ 之間為零相關，而這些證明結果可提供，藉由殘差圖去檢測迴歸模型的檢測方法有一個充分的理論基礎。

關鍵字：殘差，獨立性，相關性

The Independence and Correlation of Residual

Wen-Yang Hung　Tian-Shu Yu　Yu-Chin Cheng
Lecturer, General Education Center, De Lin Institute to Technology

Abstract

In the regression analysis, residual plots were adopted to test regression model based on empirical rule. Similarly, this test method lacks a theoretical and fundamental structure since the correlation was not investigated between residual e and least square line $\hat{Y}$. In this study, we prove a group of residuals to be not a random variable. Furthermore, we induce the relation between residual e and least square line $\hat{Y}$ which is uncorrelated. These results will provide an adequate fundamental of theory for the method testing regression model by residual plots.

Key：Residual，Independence，Correlation

壹、緒論

在迴歸分析中，殘差圖可用來檢測簡單迴歸模型之(1)直線假設 (2)誤差ε_i的齊一性 (3)誤差ε_i的獨立性(自我相關性) (4)離群值 (5)誤差ε_i的常態性 (6)忽略重要自變數(預測變數)。

通常殘差圖是採用(e ,$\hat{Y}$)或(e ,X)為橫軸與縱軸的變數，然並未探究殘差e與預估迴歸式$\hat{Y}$的相關性，而殘差e與預估迴歸式$\hat{Y}$之間的相關性，將會影響由殘差圖所判斷出來之結果的客觀性；因此本篇論文將針對殘差e與預估迴歸式$\hat{Y}$之間是否為零相關進行證明，其證明結果將可提供給藉由 (e ,$\hat{Y}$)圖形探討迴歸模型時，不會因其相關性，喪失檢測迴歸模型的能力。[2~10]

當採用殘差圖去檢測時的假設有：(1)若殘差圖不具有規化性時，則假設成立。 (2)若殘差圖具有規化性時，則假設不成立。

貳、 文獻探討

殘差圖在迴歸分析上的應用很廣泛，其研究探討的方向有[1]：

一、檢測迴歸函數是否為線性

當殘差圖落在以 O 為中心線的水平帶內，且正、負都沒有任何規則性的傾向，則表示迴歸函數為線性模型。若殘差圖呈現拋物線圖形，則表示迴歸函數為拋物線型式：

$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ ；同理，若殘差圖呈現一個反曲點，則迴歸函數為曲線型式：

$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$ 。

二、檢測變異數齊一性

若殘差圖顯示出變異數可能系統性的隨預估迴歸式$\hat{Y}$遞增或遞減，則表示變異數並不具有齊一性；但若殘差圖落在以 O 為中心線的水平帶內，且正、負並沒有任何規則的傾向，則表示變異數具有齊一性。

三、檢測離群值

離群值(outlier)即是極端的觀測值。殘差的離群值容易影響$\hat{Y}$的配適不良，使得殘差圖呈現規則性(系統性)變化。

四、檢測誤差項ε_t的獨立性(自我相關性)

當誤差項ε_t無自我相關(獨立性)時，殘差圖將呈現無規則性；當誤差項ε_t存在自我相關(非獨立)時，ε_t將隨ε_{t-1}遞增($\rho > 0$)，或隨ε_{t-1}遞減($\rho < 0$)，此時殘差e將呈現遞增或遞減趨勢。

五、誤差項ε_t的非常態性

一般採用常態機率圖時，圖中每一個殘差都對應到它在常態分佈下的期望值，若圖接近直線，則表示和常態性符合；若明顯偏離直線，則表示誤差分佈不是常態。至於誤差項分析明顯偏離常態性的三個機率圖形，若呈現凹向向上，表示分佈右偏的機率圖；若呈顯凹向向下，表示分佈左偏的機率圖；若呈現對稱，但尾巴有較厚的分佈，則表示在兩尾的部份有比常分布大的機率。

六、檢測隨機性

將殘差時間順序排列，採用連檢定(run test)的方法檢定殘差是否具有隨機性。

經由上述六點的闡釋，殘差圖在迴歸分析上的應用如右列：(1)殘差圖雖爲非正式的分析法，但在許多情況的應用上，具有檢驗模型的適切性。(2)殘差圖分析法，不僅適用於簡單線性迴歸模型，也適用於更複雜的迴歸或統計模型。(3)大部份的殘差圖分析，都是藉由電腦完成，幾乎所有的迴歸程式都提供配適值及殘差，並且一般都可查出各種殘差圖。

參、研究理論

迴歸模型假設：

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad , \quad i = 1, 2, \cdots n$$

$$= \mu_i + \varepsilon_i$$

<符號說明>

(1) 誤差 $\varepsilon_i = Y_i - \mu_i$

殘差 $e_i = Y_i - \hat{Y}_i$

(2) $S_{xx} = \sum_{i=1}^{n} (X_i - \bar{X})^2$

<定理一>

設迴歸線模型：$Y_i = \alpha + \beta X_i + \varepsilon_i \quad , \quad i = 1, 2, \cdots n$

設迴歸方程式：$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i \quad , \quad i = 1, 2, \cdots n$

殘差：$e_i = Y_i - \hat{Y}_i \quad , \quad i = 1, 2, \cdots n$

則殘差 e_1 , e_2, $\cdots$, e_n 不是獨立之隨機變數

<證明>

當 $i \neq j$

則(1) $COV(Y_i, Y_j) = 0$

(2) $COV(Y_i, \hat{Y}_j) = COV(Y_i, \hat{\alpha} + \hat{\beta} X_j) = COV(Y_i, \bar{Y} - \hat{\beta}\bar{X} + \hat{\beta} X_j)$

$$= COV(Y_i, \bar{Y} + \hat{\beta}(X_j - \bar{X}))$$

$$= COV(Y_i, \bar{Y}) + (X_j - \bar{X}) COV(Y_i, \frac{\sum_{k=1}^{n} (X_k - \bar{X}) Y_k}{S_{xx}})$$

$$= \frac{\sigma^2}{n} + \frac{(X_j - \bar{X})(X_i - \bar{X})\sigma^2}{S_{xx}}$$

(3)同理

$$COV(\hat{Y}_i, Y_j) = \frac{\sigma^2}{n} + \frac{(X_j - \bar{X})(X_i - \bar{X})\sigma^2}{S_{xx}}$$

$$= COV(Y_j, \hat{Y}_j)$$

$$(4)COV(\hat{Y}_i,\hat{Y}_j)=COV(\hat{\alpha}+\hat{\beta}X_i,\hat{\alpha}+\hat{\beta}X_j)$$

$$=COV(\hat{\alpha},\hat{\alpha})+COV(\hat{\alpha},\hat{\beta}X_j)+COV(\hat{\beta}X_i,\hat{\alpha})+COV(\hat{\beta}X_i,\hat{\beta}X_j)$$

$$=(\frac{1}{n}+\frac{\bar{X}^2}{S_{xx}})\sigma^2-\frac{\bar{X}\cdot X_j}{S_{xx}}\sigma^2-\frac{\bar{X}\cdot X_i}{S_{xx}}\sigma^2+\frac{X_i\cdot X_j}{S_{xx}}\sigma^2$$

$$=\frac{\sigma^2}{n}+\frac{(X_i-\bar{X})(X_j-\bar{X})\sigma^2}{S_{xx}}$$

$$=COV(Y_i,\hat{Y}_j)$$

$$\therefore COV(e_i,e_j)$$

$$=COV(Y_i-\hat{Y}_i,Y_j-\hat{Y}_j)$$

$$=COV(Y_i,Y_j)-2COV(Y_i,\hat{Y}_j)+COV(\hat{Y}_i,\hat{Y}_j)$$

$$=0-2COV(Y_i,\hat{Y}_j)+COV(\hat{Y}_i,\hat{Y}_j)$$

$$=-COV(Y_i,\hat{Y}_j)$$

$$=-\left[\frac{\sigma^2}{n}+\frac{(X_i-\bar{X})(X_j-\bar{X})}{S_{xx}}\sigma^2\right]$$

$$\neq 0$$

則e_i與e_j不是獨立的（independent），即 e_1，e_2，‧‧‧‧，e_n 不是獨立之隨機變數。

<定理二>

設$\hat{Y}_i=\hat{\alpha}+\hat{\beta}X_i+\varepsilon_i, i=1,2,\ldots n$爲迴歸估計式

且$e_i=Y_i-\hat{Y}, i=1,2,\ldots n$爲殘差　，　則$\hat{Y}_i$與$e_i$零相關。

<證明>

$$\sum_{i=1}^{n} e_i=0 \Leftrightarrow \sum_{i=1}^{n} Y_i-\sum_{i=1}^{n} \hat{Y}_i=0 \Leftrightarrow \sum_{i=1}^{n} Y_i=\sum_{i=1}^{n} \hat{Y}_i$$

$$\Leftrightarrow \bar{Y}=\bar{\hat{Y}}$$

$$\sum_{i=1}^{n} e_i=0 \Leftrightarrow \bar{e}=0$$

$$\gamma_{e\hat{Y}}=\frac{\sum_{i=1}^{n}(e_i-\bar{e})(\hat{Y}_i-\bar{\hat{Y}}_i)}{\sqrt{\sum_{i=1}^{n}(e_i-\bar{e})^2}\sqrt{\sum_{i=1}^{n}(\hat{Y}_i-\bar{\hat{Y}}_i)^2}}=\frac{\sum_{i=1}^{n}e_i(\hat{Y}_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}e_i^2}\sqrt{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}}$$

$$=\frac{\sum_{i=1}^{n}e_i\hat{Y}_i-\bar{Y}\sum_{i=1}^{n}e_i}{\sqrt{\sum_{i=1}^{n}e_i^2}\sqrt{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}}=\frac{0-0}{\sqrt{\sum_{i=1}^{n}e_i^2}\sqrt{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}}=0$$

$\therefore e_i$與$\hat{Y}_i$零相關

肆、結論

在迴歸分析中，都會討論殘差圖如何用來判斷迴歸模型及誤差 ε_i 的假設，且認爲殘差圖在理論基礎上之嚴密性不足，因此以經驗法則而言，僅供參考用。但根據本篇論文之定理一，證明殘差 e 不是獨立之隨機變數；定理二，證明殘差 e 與迴歸預測式 $\hat{Y}$ 零相關；因此，證明的結果表示殘差 e 與迴歸預測式 $\hat{Y}$ 之間不具有直線關係，使得殘差圖在各方面的應用上，不致因其直線關係，降低客觀的判斷能力。

伍、參考文獻

1、 John Neter, William Wasserman, Michael H. Kutner, “Applied Linear Regression Models”, P111～122, Richard. Irwin, Inc. (1983)

2、 Franklin A. Graybill , Second Printing, “Theory and Application of The Linear Model” ,P214～216, Wadsworth Publishing Co.(1976)

3、 Ronalde Walpole, Ray Monde Myers, “Probability and Statistics for Engineers and Scientists” (3e), P163～167, Macmillan Publishing Co , Inc. (1985) .

4、 Neil A. Weiss, ”Introductory Statistics ”(4e), P793～804, Addison Wesley Longman , Inc. (1995)

5、 James T. Mcclave ,Terry Sincich ,”Statistics ” (8e),P670～684 ,Prentice Hall, Inc. (2000) .

6、 William R. Dillon, Matthew Goldstein,”Multivariate Analysis / Methods and Applications” ,P260～262 , 華泰書局, (1984) .

7、 Mario F. Triola, ”Elementary Satistics ”(7e), P503～504, Addison Wesley Longman, Inc. (1998) .

8、 9、Prem S. Mann, ”Introductory Statistics ”(3e), P627～646 , John Wiley ＆ Sons , Inc. (1998) .

9、 Ronald E. Walpole, ”Introduction to Statistics ”(3e), P345～356, Macmillan Publishing Co , Inc. (1982) .

10、Thomas H. Wonnacott, Ronald J. Wonnacott，”Introductory Statistics ”(5e), P461～464 ,John Wiley ＆ Sons, Inc. (1990) .

.