Journal of the KSQC
Vol. 15, No. 1, 1987

二元配置法에서의 異常値 發見方法에 대한 硏究

A Study On Detecting Outliers In Two-Way Tables

姜 銀 美*

ABSTRACT

Basic problems in the study of detecting outliers from data of experimental designs are that they are difficult to detect and their presence influences the analysis of variance of the data set.

This article is concerned with mainly detecting outliers in two-way tables with no replications. Various methods are reviewed and their relations to the Andrews-Pregibon's Statistic and Cook's Statistic are derived.

1. 序 論

실험계획법에서의 異常値 發見方法은 여러가지 어려운 점들이 있어서 깊이 연구되지 못하였다. 즉, 異常値의 存在가 유의한 효과들을 유의하지 않게 하거나 유의하지 않은 효과들을 유의하게 나타나게 하는 등 분산분석에 영향을 주고 또한 발견하기가 힘들기 때문이다. 二元配置法에서 충분한 반복이 있는 경우는 각 水準의 組合에서의 범위를 이용하여 비교적 쉽게 異常値를 發見할 수 있으나 반복 측정하는데에 실험자가 너무 많은 시간을 소비하여야 한다. 그러므로 本 硏究에서는 반복없는 경우에 대하여 주로 다루기로 한다.

Barnett와 Lewis(1978)는 실험계획법에서의 異常値 發見方法을 殘差를 기준으로 한 것과 그렇지 않은 것으로 나누고 있다. 本 硏究에서도 이를 따르기로 한다. 그리고 회귀분석에서의 기존 方法과 이들의 관계를 考察하여 보기로 하겠다.

2. 殘差를 기준으로 한 方法

2-1. 最大殘差를 이용한 方法

殘差를 이용한 異常値 發見方法은 Daniel(19

* 성신여대 통계학과 조교수

60)이 제안하였는데 반복없는 $R\times C$ 二元配置法 模型을 사용하였다.

$$y_{ij}=\mu+\alpha_i+\beta_i+\epsilon_{ij}$$

여기에서 $i=1,\ldots R$이며 $j=1,\ldots C$이고 ϵ_{ij}들은 독립적으로 $N(0,\sigma^2)$을 따르고 母數模型이므로 $\Sigma\alpha_i=0$와 $\Sigma\beta_j=0$를 가정하였다.

Daniel(1960)은 편기된 값은 殘差에 특이한 형태를 주는 것을 보였다. 그러므로 가장 큰 殘差를 갖는 측정치를 異常値로 간주한다. 제안된 統計量은 다음과 같다.

$$t_{1,\alpha}=e_{max}/s_1$$

여기서

$$s_1^2=\frac{\mathrm{SSE}-\frac{RC}{(R-1)(C-1)}e_{max}^2}{(R-1)(C-1)-1}$$

이며 e_{max}는 가장 큰 殘差를 말하고 SSE는 殘差自乘合이다. s_1^2은 異常値를 제외한 模型에서의 殘差自乘平均이다. Daniel(1960)은 $\alpha=0.2$에서의 기각치의 근사값을 구하였다.

2-2. MNR을 이용한 方法

Ferguson(1961)은 모든 殘差가 同一한 분산을 갖는다면 MNR(Maximum Normed Residual)을 이용한 方法이 모든 不變的(invariant) 方法 중에서 가장 좋은 성질을 갖는다고 밝혔다. MNR $z^{(1)}$은 다음 z_i들 중 가장 큰 절대값을 갖는 것이다.

$$z_i=e_i/(\sum_{i=1}^{n}e_i^2)^{1/2}$$

물론 여기에서 e_i, $(i=1,\ldots n)$는 殘差이다. A_i를 $[|z_i|>D]$인 事象이라고 한다면

$$[|z|^{(1)}>D]=\bigcup_{i=1}^{n}A_i$$ 이다.

계산을 편리하게 하기 위하여 다음을 정의한다.

$$S_1=\sum_{i=1}^{n}P(A_i)$$

$$S_2=\sum_{i<j\le n}P(A_iA_j),$$

$$\vdots$$

$$S_K=\sum P(A_{i_1}A_{i_2}\ldots A_{i_K})$$

여기서 合은 모든 가능한 $i_1<i_2<\cdots<i_K\le n$의 조합을 말한다.

Feller(1960)는 다음을 증명하였다.

$$P[|z|^{(1)}>D]=S_1-S_2+S_3-\cdots\cdots\pm S_n$$

그러므로

$$S_1-S_2+\cdots\cdots-S_{2r}\le P[|z|^{(1)}>D]$$

$$\le S_1-S_2+\cdots\cdots+S_{2r-1}$$

의 Bonferroni 부등식이 성립한다.

$M=\max|z|^{(K)}$일때 $D>M_K$이면

$$P[|z|^{(1)}>D]=S_1-S_2+\cdots\cdots\pm S_K$$

이 성립한다.

M_K는 Stefansky(1971)에 의하여 구할 수 있다.

Stefansky(1972)에서 반복없는 二元配置法 경우의 MNR의 기각치를 Newton方法을 이용하여 구하였는데 Daniel의 $t_{1,\alpha}$가 MNR과 同直(equivalent)가 된다는 점을 감안하면 Daniel(1960)의 $\alpha=0.2$에서의 기각치는 檢定力이 떨어짐을 알 수 있다.

Galpin과 Hawkins(1981)는 이 결과를 약간 수정하고 二元配置法의 경우로 확장하였다.

3. 殘差를 기준으로 하지 않은 方法

3-1. Q_K를 이용한 方法

Gentleman과 Wilk(1975 a, b)는 K개의 異常値가 있는 경우에 대하여 異常値를 제외시켰을 때 감소되는 殘差自乘合을 Q_K로 하고 이 Q_K를 최대로 하는 부분집합을 $\binom{n}{K}$개의 모든 가능한 부분집합 중에서 찾아서 시뮬레이션(simulation)을 이용하여 Q_K의 확률분포를 그려 검정한다. 이 경우 유의하지 않으면 $(K-1)$개의 異常値로 가정하고 다시 검정하며 또 유의

하지 않으면 $(K-2)$개로 가정하고 검정하는 方法을 사용한다.

John과 Draper(1978)도 Q_K를 이용하여 異常値를 판단하는 方法을 제안하였는데 一般線形式으로 표현하였다. 즉,

$$E(y)=E\begin{pmatrix}y_1\\y_2\end{pmatrix}=\begin{pmatrix}X_1\\X_2\end{pmatrix}\beta=X\beta \qquad \cdots\cdots\cdots\cdots (3.1.1)$$

관측치들은 K개의 異常値 부분에 해당하는 y_2와 異常値아닌 $n-K$개의 원소를 가진 y_1으로 표현되었다. 이 模型에서 殘差 e는 다음과 같이 나타난다.

$$e=\begin{pmatrix}e_1\\e_2\end{pmatrix}=(I-R)y$$

$$=\begin{pmatrix}I-R_{11} & -R_{12}\\ -R_{21} & I-R_{22}\end{pmatrix}\begin{pmatrix}y_1\\y_2\end{pmatrix} \qquad \cdots\cdots\cdots\cdots (3.1.2)$$

여기에서 $R_{ij}=X_i(X'X)^{-1}X_j'$는 $R=X(X'X)^{-1}X'$의 部分行列이다. 異常値를 제거함으로써 감소되는 殘差自乘合은

$$Q_K=e_2'(I-R_{22})^{-1}e_2$$

로 유도된다. 또한,

$$Q_K=Q_{K-1}+U_K^2/V(U_K)$$

의 式이 성립하는데 U_K는 K개 중에서 $K-1$개의 異常値를 제외한 나머지 값을 말하며 $V(U_K)\sigma^2$은 U_K의 분산이다.

John과 Draper(1980)는 二元配置法에서 3개 以下의 異常値 存在時에 대하여 論했으며 Gentleman(1980)이 더 확장하였다.

3-2. Tetrad를 이용한 方法

Bradu와 Hawkins(1982)는 여러개의 異常値를 한번에 發見할 수 있는 方法을 제안했다. 그들은 模型을 다음과 같이 가정하였다.

$$Y_{ij}\sim N(\mu_{ij},\ \sigma^2)$$

여기서 $\mu_{ij}=\mu+\alpha_i+\beta_i+\delta_{ij}$

그리고 T라는 부분집합을 정의하였다. 즉, $(i,j)\notin T$이면 $\delta_{ij}=0$이고 $(i,j)\in T$이면 $\delta_{ij}\neq 0$이다. 여기서 T에 들어가는 관측치를 異常値로 한다. 그리고 Tetrad란 다음과 같이 정의 하였다. 즉,

$$T_{ij:eg}=Y_{ij}-Y_{ej}-Y_{ig}+Y_{eg}$$

여기서 殘差는 (i,j)를 포함하는 모든 Tetrad의 平均으로 표현됨을 알 수 있다. 즉,

$$e_{ij}=\sum_e\sum_g T_{ij:eg}/RC$$

Bradu와 Hawkins(1982)는 δ_{ij}의 추정치로 (i,j)를 포함하는 $e\neq i$이고 $g\neq j$인 모든 Tetrad의 중앙값(median)을 제안하였다. Huber(1977)에 의하면 이 統計量은 상당히 robust하다고 할 수 있다. Tetrad의 중앙값을 半정규확률지(half normal plot)를 이용하여 그리면 쉽게 異常値를 發見할 수 있다.

4. 회귀분석의 異常値 推定 統計量들과의 관계

Draper와 John(1981)은 Q_K와 Andrews-Pregibon(1978)統計量, Cook(1977)의 統計量과의 관계를 유도하였다. 模型(3.1.1)에서 $X_1^*=(X\vdots y)$이고 y_2를 異常値 部分이라면 다음 模型을 쓸 수 있다.

$$E\begin{pmatrix}y_1\\y_2\end{pmatrix}=\begin{pmatrix}X_1 & 0\\X_2 & I\end{pmatrix}\begin{pmatrix}\beta\\\gamma\end{pmatrix}=$$

$$=\left(X\vdots D\right)\begin{pmatrix}\beta\\\gamma\end{pmatrix} \qquad \cdots\cdots (4.1.1)$$

여기에서 $X_2^*=(X\vdots D\vdots y)$라면 AP統計量은 다음과 같다.

$$R_{ij\cdots}^{(K)} = |X_2^{*\prime} X_2^*| / |X_1^{*\prime} X_1^*|$$

Q_K와는 다음 관계식이 성립한다.

$$R_{ij\cdots}^{(K)} = (1 - Q_K / \mathrm{RSS}) |I - \mathrm{RSS}|$$

여기서 RSS는 殘差自乘合이고 Q_K는 (3.1.3)을 말하며 R_{22}는 (3.1.2)의 定意를 따른다.

Cook의 統計量은 다음과 같다.

$$C_{ij\cdots} = (b - b^*)' X' X (b - b^*) / ps^2$$

여기서 $\underline{b}$는 (3.1.1)의 β의 最小自乘推定値이고 $\underline{b}^*$는 (4.1.1)模型의 β의 最小自乘推定値이다. $s^2 = \mathrm{RSS}/(n-p)$이고 $ij\cdots$는 K관측치가 $\underline{y}_2$로 뽑힌 것을 의미한다.

Q_K와의 관계식은 다음과 같다.

$$C_{ij\cdots} = \frac{Q_K}{Ps^2}\left(\frac{\underline{C}'\underline{C}}{Q_K} - 1\right)$$

여기서 $\underline{C}$는 (4.1.1)模型의 $\underline{\gamma}$의 最小自乘推定値이다.

5. 結　論

本 硏究에서는 Barnett와 Lewis의 분류에따라 殘差를 기준으로한 것과 殘差를 기준으로 하지 않은 方法으로 나누었으나 다른 관점으로 보면 殘差를 기준으로 한 것은 한 개의 異常値를 檢定하는데 쓰이는 方法이고 殘差를 기준으로 하지 않은 것은 여러개의 異常値를 檢定하는데 쓰인다고할 수 있다. 그러나 여러개의 異常値를 檢定하는 데에는 Masking과 Swamping의 문제가 따른다. Masking이란 실제보다 적게 異常値가 推定되는 것이고 Swamping은 실제보다 많이 推定되는 것을 말한다. Fieller(1976)는 Q_K가 Masking이나 Swamping의 문제가 있는데 Masking은 K를 충분히 크게 가정하면 해결이 된다고 하였다. Bradu와 Hawkins(1982)는 Tetrad의 중앙값을 이용하면 Masking이나 Swamping의 문제가 해결되며 단 한번에 여러개의 異常値를 찾아낼 수 있다고 하였는데 이론적인 기반이 약하다. (4.1.2)와 (4.1.3)의 式을 살펴보면 회귀분석의 영향을 크게 주는 측정치를 推定하는 統計量에서 異常値를 檢出하는 Q_K部分을 분리시킬 수 있음을 알 수 있다. 즉 (4.1.2)式의 우측변을 보면 AP統計量은 異常値에 의한 部分과 전체 측정치들로부터 얼마만큼 떨어져 있는가 하는 부분으로 나뉘어짐을 알수 있다.

Little(1985)은 Q_K를 $Q_{K_1} = \underline{e}_2' \underline{e}_2$와 $Q_{K_2} = \underline{e}_1' X_1 (X_1' X_1)^{-1} X_1' \underline{e}_1$으로 분리하였는데 Q_{K_2}는 Cook의 統計量을 잘 반영함을 보여주고 있다. 이와 같이 Q_K는 상당히 흥미있는 統計量이므로 더욱 硏究가 필요하다 하겠다.

參 考 文 獻

1. Andrews, D.F., and Pregibon, D. (1978), "Finding the Outliers That Matter," Journal of the Royal Statistical Society, Ser. B, 40, 87-93.
2. Barnett, V., and Lewis, T. (1978), Outliers in Statistical Data, New York: John Wiley.
3. Bradu, D., and Hawkins, D.M. (1982), "Location of Multiple Outliers in Two-Way Tables, Using Tetrads, "Technometrics, 24, 103-108.
4. Cook, R.D. (1977), "Detection of Influential Observations in Linear Regression," Technometrics, 19, 15-18.
5. Draper N.R. and John, J.A. (1980), "Testing for Three or Fewer Outliers in Two-Way Tables," Technometrics, 22, 9-15.
6. ____ (1981), "Influential Observations and Outliers in Regression," Technometrics, 23, 21-26.
7. Feller, W. (1960), An Introduction to Probability Theory and its Applications, Vol. 1, John Wiley and Sons.
8. Ferguson, T.S. (1961), "On the Rejection of Outliers," Preceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1), Berkeley and Los Angeles: University of California Press, 253-287.
9. Fieller, N.R.J. (1976), "Some Problems Related to the Rejection of Outlying Observations," Unpublished Ph.D. Thesis. University of Sheffield.
10. Gentleman, J.F., and Wilk, M.B. (1975a), "Detecting Outliers in a Two-way Table: I Statistical Behavior of Residuals," Technometrics, 17, 1-14.
11. ____ (1975b), "Detecting Outliers II, Supplementing the Direct Analysis of Residuals," Biometrics, 31, 387-410.
12. Gentleman, J.F. (1980), "Finding the K Most Likely Outliers in Two-Way Tables," Technometrics, 22, 591-600.
13. Huber (1977), Robust Statistical Procedures, Society for Industrial and Applied Mathematics, Bristol, England: J.W. Arrowsmith Ltd.
14. John, J.A., and Draper (1978), "On Testing for Two Outliers or One Outliers in Two-Way Tables," Technometrics, 20, 69-78.
15. Little, J.K. (1985), "Influence and Quadratic Form in the Andrews-Pregibon Statistic," Technometrics, Vol. 27, 13-15.
16. Stefansky (1971), "Rejecting Outliers by Maximum Normed Residual," The Annals of Mathematical Statistics, 42, 35-45.
17. ____ (1972), "Rejecting Outliers in Factorial Designs," Technometrics, 14, 469-478.