Journal of the KSQC
Vol. 13, No. 2, 1985

Rejecting Outliers by Maximum Modified Normed Residual

Soon Kwi Kim*

ABSTRACT

One may be particularly interested in identifying which are the genuinely exceptional observations, in order to create a new insight into the phenomena under study. To detect outliers, many statistics have been proposed such as the maximum normed residual (MNR), a statistic equivalent to the maximum normed residual C. Daniel proposed, studentized residual, standardized residual, and so on.

This paper gives a procedure for calculating critical values of the maximum modified normed residual and the distribution of the modified normed residual.

1. Introduction

The maximum normed residual (MNR) has been proposed as a test statistic in connection with the problem of rejecting outliers. Several tests are discussed, among them tests based on the MNR, and tables of critical values are included. T. S. Ferguson (1961) has shown that for designs with the property that all residuals have a common variance, the procedure based on the MNR has the optimum property of being admissible among all invariant procedures.

Designs with the property that the residuals have a common variance include all ordinary factorial designs, where the different levels of each factor are replicated equally often, balanced incomplete blocks and Latin squares. (Note that for these designs the diagonal terms of the hat matrix are all the same)

Stefansky (1972) has obtained bounds for the fractiles of the MNR in the case of two factors. This paper proposes a modified normed residual which is defined as $z_i = e_i / (\sum_{j \neq i} e_j^2(i))$, where $e_j(i) = y_i - \underset{\sim}{x}_j' \cdot b(i)$, $e_i = y_i - \underset{\sim}{x}_i' \cdot b(i) = y_i - \tilde{y}_i$, and $b(i)$ the regression coefficients estimated with the i th row deleted and derives the distribution function of z_i and the joint distribution function of z_i and z_j

* Department of Computer Science & Statistics, Seoul National University.

2. Distribution Theory

For the work in later sections the distribution of z_i and the joint distribution of z_i and z_j are needed. We shall now obtain these distributions, taking for definiteness $i=n+1$, $j=1$

As is well know, $\tilde{e}_{n+1} \sim N(0,(1-h_{n+1})\sigma^2)$

Where h_{n+1} is the diagonal element of the hat matrix $H=X(X'X)^{-1}X$.

Then,

$$e_{n+1}=y_{n+1}-\tilde{y}_{n+1}=(y_{n+1}-\hat{y}_{n+1})+(\hat{y}_{n+1}-\tilde{y}_{n+1}) \text{ (where } \hat{y}_{n+1} \text{ is } \underset{\sim}{x}'_{n+1}b)$$

$$=\tilde{e}_{n+1}+\frac{h_{n+1}}{1-h_{n+1}}\tilde{e}_{n+1}=\frac{\tilde{e}_{n+1}}{1-h_{n+1}}$$

And,

$$e_1^2(n+1)+\cdots+e_n^2(n+1)=\underset{\sim}{y}'_n[I-X(n+1)[X'(n+1)\,X(n+1)]^{-1}\,X'(n+1)]\underset{\sim}{y}_n$$

where $\underset{\sim}{y}'_n=(y_1,\ldots,y_n)$ and $X(n+1)$ is a matrix with the $(n+1)$th row deleted.

With no confusion we denote $e_j^2(n+1)$ by e_j^2 for $j=1,2\cdots\cdots n$.

Then,

$$e_1^2+e_2^2+\cdots+e_n^2 \sim \sigma^2\,X^2(n-k-1), \quad \text{where the rank of the matrix } X \text{ is } K+1.$$

Lemma. $y_{n+1}-\tilde{y}_{n+1}$ and $e_1^2+\cdots+e_n^2$ are independent.

Proof. $\tilde{y}_{n+1}=(\underset{\sim}{x}'_{n+1})\cdot b(n+1)=\underset{\sim}{x}'_{n+1}[X'(n+1)\,X(n+1)]^{-1}\,X'(n+1)\;\underset{\sim}{y}_n$

$$e_1^2+\cdots+e_n^2=\underset{\sim}{y}'_n(I-X(n+1)[X'(n+1)\,X(n+1)]^{-1}\,X'(n+1))\underset{\sim}{y}_n$$

Hence $\tilde{y}_{n+1}$ and $e_1^2+\cdots+e_n^2$ are independent.

And y_{n+1} and $e_1^2+\cdots+e_n^2$ are also independent.

It follows that the density function of z_n is

$$f(y)=\sqrt{1-h_{n+1}}\cdot\frac{\Gamma[\frac{1}{2}(n-k)]}{\Gamma[\frac{1}{2}(n-k-1)]}\cdot\frac{1}{\sqrt{\pi}}\;\frac{1}{[1+(1-h_{n+1})y^2]^{\frac{n-k}{2}}} \qquad \cdots\cdots\cdots\cdots \quad (2.1)$$

Next, we shall derive the joint distribution function of z_1 and z_{n+1}.

Recall that

$$z_1=\frac{e'_1}{\sqrt{e_2'^2+e_3'^2+\cdots+e_{n+1}'^2}} \qquad \text{and} \qquad z_{n+1}=\frac{e_{n+1}}{\sqrt{e_1^2+\cdots+e_n^2}}$$

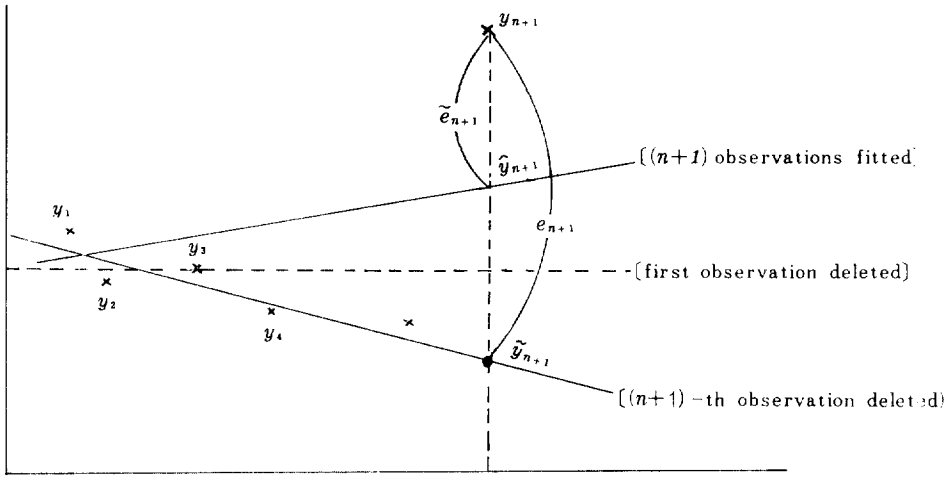

Fig. 1. (Case of K=1)

From the notation of Park S. H. (1984), we know that the following identities holds.

$$(n-k-1)s^2(n+1)=(n-k)\cdot s^2-\frac{\tilde{e}_{n+1}^2}{1-h_{n+1}} \qquad \cdots\cdots\cdots\cdots \quad (2.2)$$

$$(n-k-1)s^2(1)=(n-k)\cdot s^2-\frac{\tilde{e}_1^2}{1-h_1}$$

Hence, we can derive the relation between n_i and z_i (See Formula (4.1)) (Assume the residuals all have the same variance). From the joint distribution function of n_1 and n_{n+1}, the following joint distribution function $h(\cdot)$ is derived.

$$h(z_1, z_{n+1})=(n+1)(n-k-2)\cdot \frac{(1-h_1)(1-h_{n+1})}{[1+(1-h_1)z_1^2]^{\frac{3}{2}}[1+(1-h_{n+1})z_{n+1}^2]^{\frac{3}{2}}}$$

$$\cdot\left\{1-(n+1)\left[\frac{(1-h_1)^2z_1^2}{1+(1-h_1)\cdot z_1^2}-2\rho\cdot\frac{(1-h_1)z_1(1-h_{n+1})z_{n+1}}{\sqrt{1+(1-h_1)z_1^2}\sqrt{1+(1-h_{n+1})z_{n+1}^2}}\right.\right.$$

$$\left.\left.+\frac{(1-h_{n+1})^2z_{n+1}^2}{1+(1-h_{n+1})z_{n+1}^2}\right]\Big/[(n-k)(1-\rho^2)]\right\}^{\frac{n-k-4}{2}}$$

if $\dfrac{(1-h_1)^2\cdot z_1^2}{1+(1-h_1)\cdot z_1^2}-2\rho\cdot\dfrac{(1-h_1)\cdot z_1\cdot(1-h_{n+1})\cdot z_{n+1}}{\sqrt{1+(1-h_1)z_1^2}\cdot\sqrt{1+(1-h_{n+1})z_{n+1}^2}}+\dfrac{(1-h_{n+1})^2\cdot z_{n+1}^2}{1+(1-h_{n+1})\cdot z_{n+1}^2}$

$$\leq\frac{(n-k)\cdot(1-\rho^2)}{n+1}$$

where ρ denotes the correlation coefficient between $\tilde{e}_1$ and $\tilde{e}_{n+1}$.

3. The Computational Procedures

We now consider a procedure for computing significance points of $z^{(1)}$ defined by $\max_i |z_i|$. By Bonferroni's inequality we have for any D.

$$n \cdot P_r(|z_i| > D) - \binom{n}{2} P_r(|z_i| > D, |z_j| > D) \leq P_r(z^{(1)} > D) \leq n \cdot P_r(|z_i| > D)$$

Hence the number D_1 for which $n \cdot P_r(|z_i| > D_1) = \alpha$ is an upper bound for D_α (D_α is defined by the equation $P_r(z^{(1)} > D_\alpha) = \alpha$)

From (2.1) this equation can be written as

$$\frac{\alpha}{2n} = \int_{D_1}^{\infty} f(y)dy$$

The above equation can be solved for D_1 by Newton's method as described in Quesenberry, C. P. (1961).

Alternatively, D_1 can be found from tables of the t-distribution as follows

$$1 - \frac{\alpha}{2n} = \int_{-\infty}^{D_1} f(y)dy = P_r(Y \leq D_1) = P_r\left(t(n-k-1) \leq D_1 \sqrt{(n-k-1)(1-h_{n+1})}\right)$$

where $t(n-k-1)$ denotes the random variables of the t-distribution with $n-p-1$ df.

If $D_1 > M_2$, $D_1 = D_\alpha$, so that no further computations are necessary. Otherwise it is necessary to solve for D_2 in the equation $n \cdot P_r(|z_1| > D_2) - \binom{n}{2} P_r(|z_i| > D_2, |z_j| > D) = \alpha$

An iterative procedure for solving this equation is described in Quesenberry, C. P. (1961). If $D_2 > M_3$, $D_2 = D_\alpha$ and no further computations are necessary. If $D_2 \leq M_3$ but D_1 and D_2 agree to sufficiently many decimal places, no further computations are necessary either. Otherwise it is necessary to find the second upper bound D_3 for D_α by solving for D_3 in the equation.

$$\alpha = n \cdot P_r(|z_i| > D_3) - \binom{n}{2} P_r(|z_i| > D_3, |z_j| > D_3) + \binom{n}{3} P_r(|z_i| > D_3, |z_j| > D_3, |z_k| > D_3)$$

If the solution D_3 is greater than M_4 or if D_3 and D_2 agree to sufficiently many decimal places, no further computations are necessary. Otherwise it is necessary to calculate the second lower bound D_4 for D_α. It should be clear now how the scheme can be continued until D_α is found with sufficient accuracy.

4 Relation between z_i and the Quantity $n_i = e_i / (\sum_1^{n+1} e_i^2)$

Under the assumption that the errors ε are independently and identically distributed $N(0, \sigma^2)$, the

joint distribution of a set of P n_i's has the form of an inverted t-distribution with $\nu=n-k-p$ df. From (2.2) we can derive the relation between z_i and n_i.

$$n_i = \frac{(1-h_i)z_i}{\sqrt{1+(1-h_i)\cdot z_i^2}} \quad \text{for} \quad i=1, 2, \cdots, n+1.$$

From the fact that the matrix H is idempotent, if the variances of the residuals are same, the value of the element h_i is the mean of the rank of the matrix H.

5. Conclusions

The critical values for $z^{(1)}$ are found by solving $P(z^{(1)}>D)=\alpha$ for D.

Feller (1960) proved that $P(z^{(1)}>D)=s_1-s_2+s_3, \cdots \pm s_n$

where $s_k=\sum P(|z_{j_1}|>D, |z_{j_2}|>D, \cdots, |z_{j_k}|>D)$, summation being over all K-tuple $j_1<j_2<\cdots<j_k\leq n$, so that there are $\binom{n}{k}$ terms in the summation.

To obtain the more precise value of D such that $P(z^{(1)}>D)=\alpha$, D_4 and D_5 etc. must be computed. Extension of the method to higher-way cross-classification design is conceptually easy, however the computations needed for s_2 and s_3, s_4 are proportional to the number of distinct correlations and correlation triples. respectively.

For $K\geq 4$, s_k remains computationally intractable. And, for most higher-way cross-classifications these numbers increase dramatically as the order of the layout increases, so that the exact calculations may become excessively laborious.

REFERENCES

1. Feller, W. (1960), *An introduction to prob. Theory and its applications,* Vol. 1, New York, John Wiley & Sons.
2. Jacqueline S. G. and Douglas M. H. (1981), "*Rejection of a single outlier in Two-or Three way Layouts,*" Technometrics, 23, 65-70.
3. John J. A. and Prescott P. (1975), "*Critical values of a test to detect outliers in Factorial experiments,*" Appl. Statistics, 24, 56-59.
4. Mcmillan R. G. (1971), "*Tests for one or two outliers in normal samples with unknown variance,*" Technometrics, 13, 87-100.
5. Park S. H. (1984), 회귀분석, 대명사, 서울
6. Stefansky W. (1971), "*Rejecting outliers by maximum normed residual,*" Ann. Math. Statist. 42, 35-45.
7. Stefansky (1972), "*Rejecting outliers in factorial designs,*" Technometrics, 14, 469-479.
8. Welsh R.E., Kuh E. and Bersley D.A *Regression diagonistics,* New York, John Wiley.