APPROXIMATE SIGNIFICANCE LEVELS FOR DETECTING OUTLIERS IN LINEAR MODELS

Technical Report No. 367

by

R. Dennis Cook

Department of Applied Statistics
University of Minnesota
St. Paul, Minnesota 55108

and

P. Prescott

Department of Mathematics
University of Southampton
Highfield, Southampton
England, SO9 5NH.

January 1980

ABSTRACT

Present methods for assessing the accuracy of 1st order Bonferroni p-values or critical values associated with the maximum absolute studentized residual as a criterion for detecting a single outlier in a linear model require numerical integration. We present a relatively simple alternative method which is suitable for routine use. The application to analyses of designed experiments and regression models is discussed.

KEY WORDS: Bonferroni bounds, Factorial designs, Maximum absolute studentized residual, Outliers, P-values, Regression.

1. INTRODUCTION

The maximum absolute studentized residual is an accepted criterion for detecting the presence of a single outlier in a least squares analysis based on a general linear model. However, because of the complexity of the associated distribution, exact p-values are difficult to obtain and the few published critical values are based on the 1st order Bonferroni upper bound or large scale simulations.

Tietjen, Moore and Beckman (1973) used simulation to determine critical values for simple linear regression. Prescott (1975) demonstrated that these critical values are close to those obtained using the 1st order Bonferroni upper bound and suggested that this bound would produce adequate critical values for linear models in general. Following Prescott's suggestion, Lund (1975) constructed tables of critical values. According to Barnett and Lewis (1978), Lund (1975) provides the most useful tabulation to date. More recently, Moses (1978) provides charts for finding the upper percentage points of Student's t in the range 0.01 to 0.00001.

Stefansky (1971,1972) and Srikantan (1961) show that for sufficiently small samples the usual critical values based on the 1st order Bonferroni upper bound will be exact. Stefansky (1971) and Prescott (1977) provide sufficient conditions for determining when these critical values are exact in models with constant residual variance.

At present, the 1st order Bonferroni upper bound appears to be the only practically useful tool for determining critical values or p-values. Despite the fact that there are many results (see Barnett and Lewis (1978) for further discussion) to suggest that these values should usually be adequate, there is no convenient method for an assessment of their accuracy in a given problem.

Here, we provide a relatively simple method for assessing the goodness of p-values and critical values based on the 1st order Bonferroni upper bound. The method applies to general linear models, incorporates the configuration of the carrier values and depends on sufficient conditions for the Bonferroni values to be exact. The general method is presented in Section 2 and is applied to several examples in Section 3.

2. BOUNDS IN LINEAR MODELS

Consider the usual linear model

$$\underset{\sim}{Y} = \underset{\sim}{X}\,\underset{\sim}{\beta} + \underset{\sim}{e} \qquad (1)$$

where $\underset{\sim}{Y}$ is an $n\times 1$ vector of observations, $\underset{\sim}{X}$ is an $n\times p$ matrix of known constants and $\underset{\sim}{\beta}$ is a $p\times 1$ vector of unknown parameters. Under the assumption of no outliers, the elements of the $n\times 1$ vector $\underset{\sim}{e}$ are assumed to be independently distributed normal random variables with mean zero and constant variance σ^2. Let $\underset{\sim}{r} = (r_i)$ denote the $n\times 1$ vector of residuals from a least-square fit, $s^2 = \underset{\sim}{r}'\underset{\sim}{r}/(n-p)$, $\mathrm{Var}(r_i) = v_i\sigma^2$ and ρ_{ij} = correlation between r_i and r_j. The studentized residuals t_i are given by $t_i = r_i/(v_i s^2)^{\frac{1}{2}}$, $i = 1,2,\ldots,n$.

The basic test statistic for a single outlier in the above linear model is taken to be the maximum absolute studentized residual, $\max|t_i|$. This statistic is usually associated with a mean-slippage alternative, i.e. the model (1) is correct except that the expectation of one unknown element of $\underset{\sim}{e}$ is nonzero.

In the following, we find it convenient to use

$$w_i = t_i/(n-p) = r_i/(\underset{\sim}{r}'\underset{\sim}{r}v_i)^{\frac{1}{2}} \qquad (2)$$

rather than t_i .

Let $\alpha_i = \Pr(|w_i|>d)$ and $\alpha_{ij} = \Pr(|w_i|>d\ ,\ |w_j|>d)$, $i \neq j$. The 1st order Bonferroni bounds applied to the events $\{|w_i|>d\}$ yields

$$\sum_i \alpha_i - \sum_{i<j} \alpha_{ij} \leq \Pr(\max|w_i|>d) \leq \sum_i \alpha_i \qquad (3)$$

Since, under the assumption of no outliers, the $|w_i|$'s are identically distributed, the upper bound can be conveniently expressed in terms of a single random variable F which follows an F-distribution with 1 and $(n-p-1)$ degrees of freedom

$$\sum_i \alpha_i = n \Pr(w_i^2 > d^2) = n \Pr\left[F > d^2(n-p-1)/(1-d^2)\right] \qquad (4)$$

Evaluation of the lower bound in (3) is more difficult since the α_{ij}'s would have to be determined by numerical integration (cf. Stefansky, 1972). Instead, we shall use an approximation for the α_{ij}'s which is obtained as follows:

Since the distribution of (w_i,w_j) is symmetric,

$$\alpha_{ij} = 2 \Pr(w_i > d,\ w_j > d) + 2 \Pr(w_i > d,\ -w_j > d) .$$

Clearly,

$$\Pr(w_i > d,\ \pm w_j > d) \leq \Pr(w_i \pm w_j > 2d)$$

and therefore

$$\alpha_{ij} \leq 2 \Pr(w_i + w_j > 2d) + 2 \Pr(w_i - w_j > 2d)$$

$$= \Pr\left[(w_i + w_j)^2 > 4d^2\right] + \Pr\left[(w_i - w_j)^2 > 4d^2\right]$$

$$= \beta_{ij}^{+} + \beta_{ij}^{-} , \text{ say.}$$

This in combination with (3) implies that

$$\alpha - \sum_{i<j} (\beta_{ij}^{+} + \beta_{ij}^{-}) \leq \Pr(\max|w_i| > d) \leq \alpha \tag{5}$$

where $\alpha = \sum \alpha_i$.

The evaluation of the lower bound in (5) is straightforward once it is recognized that

$$(w_i \pm w_j)^2 (n-p-1) / [2(1 \pm \rho_{ij}) - (w_i \pm w_j)^2] \tag{6}$$

has an F-distribution with 1 and $n-p-1$ degrees of freedom. Evidently, from (6)

$$\beta_{ij}^{\pm} = 0 \quad \text{when} \quad 2d^2 > (1 \pm \rho_{ij}) \ . \tag{7}$$

Let, $c(\pm) = \{(i,j) \mid i < j,\ 2d^2 < (1 \pm \rho_{ij})\}$.

Using (5), (6) and (7) we obtain the final form

$$\alpha - \beta^{+} - \beta^{-} \leq \Pr(\max_i |w_i| > d) \leq \alpha \tag{8}$$

where

$$\alpha = n \Pr\left[F > d^2(n-p-1)/(1-d^2)\right]$$

$$\beta^{+} = \sum_{c(+)} \Pr\left[F > d^2(n-p-1)/(\tfrac{1}{2}(1 + \rho_{ij}) - d^2)\right]$$

and

$$\beta^{-} = \sum_{c(-)} \Pr\left[F > d^2(n-p-1)/(\tfrac{1}{2}(1 - \rho_{ij}) - d^2)\right] \ .$$

Several comments on this final form are in order. First, the lower bound depends on the joint distribution of (w_i, w_j) through the residual correlations ρ_{ij} which depend only on the configuration of the carrier values and are thus known.

Second, it follows immediately from (8) that the upper bound is exact when $c(+)$ and $c(-)$ are empty, i.e.

$$1 + \max_{i<j} |\rho_{ij}| < 2d^2 \tag{9}$$

This is equivalent to the sufficient conditions given by Prescott (1977), Stefansky (1971,1972) and Srikantan (1961). Note also that since $d^2 < 1$, the upper bound can never be exact if $\rho_{ij} = \pm 1$ for some $i \neq j$.

Third, calculation of the lower bound in (8) requires knowledge of the ρ_{ij}'s. In many designed experiments, these will have a simple structure so that the lower bound can be calculated without difficulty. For example, in a two-way table with one observation per cell there are only three distinct residual correlations. The correlations for a two-way table along with their respective frequencies of occurrence are given by Stefansky (1972). In other cases, the lower bound may be approximated further by replacing ρ_{ij} in β^+ (β^-) with $\max_{c(+)} \rho_{ij}$ ($\min_{c(-)} \rho_{ij}$). Our experience suggests that this will often be adequate.

Finally for the one-tailed version using $\max w_i$, equation (8) is valid provided the upper bound is replaced by $\alpha/2$ and the lower bound is replaced by $(\alpha-\beta^+)/2$.

3. EXAMPLES

Example 1. Consider a simple random sample of size n which, under the assumption of no outliers, is assumed to be from a single normal distribution. In this case, $\rho_{ij} = -1/(n-1)$, $i \neq j$, and the upper bound will be exact whenever $2d^2 > 1 + 1/(n-1)$. (Note that the corresponding condition given by Barnett and Lewis (1978, page 96) is equivalent to $2d^2 > 1 - 1/(n-1)$.)

Table 1 presents values of d^2 and $\beta^+ + \beta^-$ for various sample sizes. For each value of n, d^2 was determined so that $\alpha = 0.05$. Thus, the d^2 values are the nominal 5% points based on the

1st Bonferroni upper bound. From Table 1 it is clear that, depending on purpose, the upper bound may provide a reasonable approximation for samples as large as 250.

Example 2. In a 2^m factorial design the residual correlations ρ_{ij} will depend on the form of the model. Generally, there will be at most m distinct off diagonal $(i \neq j)$ correlations. For a model with main effects only, the residual correlations are

$$\rho_{ij} = - [p-2k]/(2^m-p) \tag{10}$$

where $k = 1,2,\ldots,m$ and $p = m+1$. In general, p is the number of free parameters in the model including the constant. Notice that $2^m - p$ is just the residual degrees of freedom. The ρ_{ij}'s in (10) occur with frequency $\binom{m}{k} 2^m$ for $k = 1,2,\ldots,m$, respectively.

The residual correlations in a model with all main effects and first order interactions are

$$\rho_{ij} = - [p - 2k(m-k+1)]/(2^m-p), \quad k=1,2,\ldots,m, \tag{11}$$

with respective frequencies $\binom{m}{k} 2^m$.

For main effects and first and second order interactions,

$$\rho_{ij} = - \left[p - 2\left\{k + k(m-k) + \frac{k(m-k)(m-k-1)}{2} + \frac{k(k-1)(k-2)}{3!}\right\}\right] \Big/ (2^m-p) \tag{12}$$

for $k = 1,2,\ldots,m$.

Again, the respective frequencies are $\binom{m}{k} 2^m$.

As an illustration, consider a 2^4 with main effects and first order interactions. Using (11) for $k = 1,2,3,4$, we find that $\rho_{ij} = - 3/5, 1/5, 1/5$ and $-3/5$ with frequencies of occurrence 64, 96, 64 and 16, respectively. Thus, there are only two distinct correlations $-3/5$ and $1/5$ with frequencies 80 and 160, respectively, and the calculation of β^+ in (8) will require the evaluation of two probability statements,

$p_1 = \Pr[F > 5d^2/(0.2-d^2)]$ and $p_2 = \Pr[F > 5d^2/(0.6-d^2)]$.

Recall that the summation in β^+ is over $i < j$ and thus these probabilities are multiplied by half of the respective frequencies of occurrence, $\beta^+ = 40p_1 + 80p_2$. Similar statements hold for β^- . Finally, using (9), we find that α will be exact, i.e. $\beta^+ + \beta^- = 0$, when $d^2 > 0.8$ or, equivalently, $\alpha \lesssim 0.258$.

Table 2 gives $\beta^+ + \beta^-$ and d^2 for various selected α's in three 2^m factorial designs with main effects and first order interactions. From Table 2, it is seen that, at the usual levels, the upper bound will be adequate for most purposes, particularly when judging weight of evidence from p-values. In other models for factorials of the same size, the upper bound may appear more or less precise depending on the residual correlations.

Example 3. To illustrate the use of (8) in regression, we consider the data from Mickey, Dunn and Clark (1967). There are $n=21$ observations and, following Mickey, Dunn and Clark, we use a simple linear regression model, $p=2$.

The largest studentized residual corresponds to observation 19, $r_{19} = 30.28$, $v_{19} = 0.947$ and $\underset{\sim}{r}'\underset{\sim}{r} = 2309$. Thus, $\max|w_i| = |w_{19}| = 0.6475$ and $w_{19}^2 = 0.4193$. The p-value associated with $d = 0.6475$ and based on the 1st order Bonferroni upper bound in (8) is $\alpha = 0.0425$.

Application of the lower bound in (8) at $d = 0.6475$ would require the evaluation of about 420 probability statements. While it would be straightforward to write a code to perform the required calculations, it will usually be sufficient to employ a further approximation as suggested in Section 2. The advantage of this is that the number of probability statements which need to be evaluated is greatly reduced. A small number of evaluations can be handled easily on many hand-held calculators.

A quick inspection of the residual correlation matrix shows that all correlations lie in the interval $[-0.556, 0.202]$. A first lower bound on $\alpha - \beta^+ - \beta^-$ can be obtained by replacing each ρ_{ij} in β^+ (β^-) with $\max \rho_{ij} = 0.202$ ($\min \rho_{ij} = -0.556$). However, this results in negative values for the lower bound on $\alpha - \beta^+ - \beta^-$ at $d = 0.6475$ so that a closer approximation is required.

A second inspection of the residual correlations reveals that one pair has a correlation of -0.556, two other pairs have correlations of -0.30 and of the remaining pairs 17 correlations lie in the interval $[0.002, 0.202]$ and 190 lie in $[-0.221, -0.016]$. A second lower bound on $\alpha - \beta^+ - \beta^-$ can be obtained by using the four values $\{-0.556, -0.30, -0.016, 0.202\}$ in combination with their respective frequencies $\{1, 2, 190, 17\}$ to evaluate β^+ and the four values $\{-0.556, -0.30, -0.221, 0.002\}$ in combination with the same respective frequencies to evaluate β^-. This procedure, which requires the evaluation of only 8 probability statements, produces $\beta^+ + \beta^- < 0.0016$. In short, the true p-value corresponding to $d = 0.6475$ is between 0.0409 and 0.0425.

Table 3 presents upper bounds on $\beta^+ + \beta^-$ for various preselected values of α obtained using the above procedure. Clearly, this procedure produces useful bounds in each case.

4. COMMENT

Recall that the lower bound in (8) can never be exact if $\rho_{ij} = \pm 1$ for some $i \neq j$. While the bounds provided by (8) are valid even if some $\rho_{ij} = \pm 1$ $(i \neq j)$, we can generally expect them to be further apart in this case than in the examples of the previous

section. Discrepant bounds might imply that the structure of the experiment has not been fully exploited.

Consider, for example, a 2×3^2 design with main effects and first order interactions. For this design and model, the simulation study of John and Prescott (1975) gave 0.028 and 0.0610 as the probabilities of exceeding the nominal 5 and 10 percent Bonferroni critical values. Barnett and Lewis (1978, p.244) suggest that this rather large discrepancy should cast some doubt on the accuracy of Bonferroni critical values in general.

The apparent discrepancy in this case is due to the residual correlations. Of the 306 off diagonal correlations, $\pm \frac{1}{2}$ and $\pm \frac{1}{4}$ each occur with frequency 72 and -1 occurs with frequency 18. Straightforward application of (8) shows that the true probabilities of exceeding the nominal 5 and 10 percent critical values must lie in the intervals $[0.025, 0.05]$ and $[0.05, 0.1]$, respectively. These intervals are much wider than those in Section 3. However, because of the -1 correlations, which occur between the + and - levels of the first factor at the same combination of levels of the second and third factors, this approach, while correct, is not the most appropriate.

Since every residual has a -1 correlation with a second residual resulting in 9 pairs, a single outlying observation cannot be identified. However, the pair of observations most likely to contain a single outlier can be identified by considering the residuals at only one level of the first factor. For this approach, the Bonferroni critical values are exact for $\alpha \lesssim 0.519$.

Finally, because of the structure of the residual correlations, the nominal 5 and 10 percent critical values used by John and Prescott (1975) are the actual 2.5 and 5 percent points, respectively, a conclusion which is suggested by their simulation results. For further discussion of difficulties associated with high correlations, see Anscombe (1960).

5. ACKNOWLEDGEMENT

This work was undertaken while the first author was a Hartley Visiting Fellow at the University of Southampton and was supported in part by grant 1-RO1-GM25587 from the National Institute of General Medical Science, U.S. Department of Health, Education and Welfare.

REFERENCES

Anscombe, F.J. (1960). Rejection of outliers. Technometrics, 2, 123-147.

Barnett, V. and Lewis, T. (1978). Outliers in Statistical Data. John Wiley and Sons Ltd.,

John, J.A. and Prescott, P. (1975). Critical values of a test to detect outliers in factorial experiments. Applied Statistics, 24, 56-59.

Lund, R.E. (1975). Tables for an approximate test for outliers in linear models. Technometrics, 17, 473-476.

Mickey, M.R., Dunn, O.J., and Clark, V. (1967). Note on use of stepwise regression in detecting outliers. Computers & Biomed. Res., 1, 105-111.

Moses, L.E. (1978). Charts for finding upper percentage points of Student's t in the range .01 to .00001. Communications in Statistics, B7(5), 479-490.

Prescott, P. (1975). An approximate test for outliers in linear models. Technometrics, 17, 129-132.

Prescott, P. (1977), An upper bound for any linear function of normed residuals. Communications in Statistics, B, 6, 83-88.

Srikantan, K.S. (1961). Testing for a single outlier in a regression model. Sankhya, A, 23, 251-260.

Stefansky, W. (1971). Rejecting outliers by maximum normed residual. Ann. Math. Statist., 42, 35-45.

Stefansky, W. (1972). Rejecting outliers in factorial designs. Technometrics, 14, 469-479.

Tietjen, G.L. Moore, R.H., and Beckman, R.J. (1973). Testing for a single outlier in simple linear regression. Technometrics, 15, 717-721.

TABLE 1 - Bounds for a Single Normal Sample of Size n with $\alpha = 0.05$.

n	d^2	$\beta^+ + \beta^-$
10	0.646	0
15	0.498	8.0×10^{-7}
20	0.406	7.0×10^{-5}
30	0.302	5.0×10^{-4}
50	0.204	1.4×10^{-3}
100	0.117	3.6×10^{-3}
250	0.054	6.0×10^{-3}

TABLE 2 - Bounds for 2^m Factorials with Main Effects and First Order Interactions*.

Upper Bound, α	2^5		2^6		2^7	
	d^2	$\beta^+ + \beta^-$	d^2	$\beta^+ + \beta^-$	d^2	$\beta^+ + \beta^-$
0.0005	0.732	0	0.389	5.4×10^{-8}	0.196	4.3×10^{-7}
0.001	0.696	0	0.369	5.1×10^{-7}	0.184	1.2×10^{-6}
0.005	0.626	2.9×10^{-7}	0.320	2.4×10^{-5}	0.159	2.8×10^{-5}
0.01	0.591	8.5×10^{-6}	0.297	1.1×10^{-4}	0.150	1.3×10^{-4}
0.05	0.498	1.4×10^{-3}	0.243	3.9×10^{-3}	0.121	4.2×10^{-3}
0.10	0.452	7.8×10^{-3}	0.219	1.78×10^{-2}	0.109	1.84×10^{-2}
0.15	0.423	2.0×10^{-2}	0.204	4.28×10^{-2}	0.102	4.34×10^{-2}
0.20	0.402	3.7×10^{-2}	0.194	7.97×10^{-2}	0.098	7.93×10^{-2}

* For a 2^4, $\beta^+ + \beta^- = 0$ for $\alpha \leq 0.258$.

TABLE 3 - Bounds for Various Situations in Example 3.

α	d^2	Upper Bound for $\beta^+ + \beta^-$
0.0005	0.638	3.9×10^{-8}
0.001	0.610	2.1×10^{-7}
0.005	0.537	6.0×10^{-6}
0.01	0.502	2.8×10^{-5}
0.0425	0.419	1.6×10^{-3}
0.05	0.409	2.4×10^{-3}
0.10	0.365	1.4×10^{-2}
0.15	0.338	3.8×10^{-2}
0.20	0.319	7.3×10^{-2}