THE EGYPTAIN STATISTICAL JOURNAL
ISSR, CAIRO UNIV., VOL.,41, NO.2, 1997

Detection of Outliers in Mixed Model Analysis

A.K.A. Ali[1], C. Y. Lin[2] and E. B. Burnside[3]

[1]Department of Animal Production
King Saud University
P.O. Box 2460

[2]Dairy and Swine Research and Development Centre
Agriculture and Agri-Food Canada
P.O. Box 90, 2000 Route 108 East
Lennoxville, QC, Canada J1M 1Z3

[3]Nova Scotia Agricultural College
Truro, N. S., Canada B2N 5E3

INTRODUCTION

Field data may contain some influential observations that have an effect on fitting the model. An outlier may be an influential observation with large residual. The presence of the outliers is an indication of inadequate model, incorrect data or both. Data inadequacies are often caused by gross measurement or recording errors. As Hamper et al (1986) reported, routine data contained about 1-10% gross errors, and even the highest quality data cannot be guaranteed free from gross errors. Detection of influential observation as an outlier is important in developing a robust technique to: 1) estimate parameters insensitive to the influence of this observation, 2) increase the power of test statistics, 3) narrow the confidence intervals, and 4) keep the integrity of data-model combination.

Several studies have described outliers diagnostics as the statistics that reveal observations having a large influence on parameter estimates, which is known to be sensitive to departures from the assumptions under which it is derived "nonrobust". Diagnostics measures have been derived to detect individual or multiple cases that deviate abnormally from the data.

A. Diagnostics of individual outlier

Hoaglin and Welsch (1987) found that, in least square solution of linear model, the hat matrix $H=X(X'X)^{-1}X'$ is a projection of the data vector y into their estimates $\hat{y} = H y$ and estimates residuals such that $e = (I - H) y$. The diagonal elements of H matrix 'leverage point' describes how far away the individual observation is from the centroid of all data in the space of independent variables. The diagonal elements ($0 \leq h_{ii} \leq 1$) of zero indicate a point with no influence on fit. Some researchers (Cook and Weisberg, 1982; Hocking, 1983; Stevans, 1984) used the leverage points (h_{ii}) to determine the cutoff value (CFV) where $h_{ii} = 2p / N$, where p is the number of independent variables and N is the total number of observations. If h_{ii} is greater than 2p/N, the total number observation is declared as outlier.

Rousseuw and Leroy (1987) found a one-to-one relationship between squared Mahalanobis distance and diagonal elements (h_{ii}) of hat matrix given by:

$MD_i^2 = (N-1)(h_{ii} - 1/N)$.

The outliers in the y-direction are completely neglected by the hat matrix (or function of h_{ii}) because H matrix is only based on the independent variables. Several studies (Hoaglin and Welsch, 1978; Draper and John, 1981; Atkinson, 1982; Cook and Weisberg, 1982; Hocking, 1983; Hocking and Pendleton, 1983; Paul, 1983; Stevans, 1984; Atkinson, 1985) used the hat matrix in computing the covariance among residuals such that $Cov(e) = \sigma^2 (I - H)$. Covariance among residuals was taken into consideration for scaling residuals and deriving other statistics such as:

1) Standardized residuals $t_i = e_i / s_i$ where s_i is an estimate of σ and e_i is the residual of the ith observation.

2) Studentized residuals $t_i = e_i / [s(1-h_{ii})^{.5}]$.

Observation with the largest absolute value of t_i is taken as most likely to be a contaminant.

3) Jackknifed residuals $J_i = e_i / [s_{(i)}(1-h_{ii})^{.5}]$ where $s_{(i)}$ is the estimate of s without including the ith observation in the analysis.

4) Cook's statistic (1977, 1979) is defined as the squared standardized distance over b when estimated with the ith observation such that:

$CD_i^2 = (b - b_i)' X'X (b - b_i) / ps^2 = (y_{(i)} - y)' (y_{(i)} - y) / ps^2$

where:

b = the least square estimate (LSE) of the regression coefficient.

b_i = the LSE of the regression coefficient after deleting the ith observation.

p = the number of independent variables.

s^2 = an estimate of σ^2.

$y_{(i)}$ = the data vector after deleting the ith observation.

This statistic is invariant under nonsingular linear transformation. Furthermore, by definition it measures either the change in the estimate of b relative to its variance or the change in the fitted value vector.

An equivalent formula that makes use of h_{ii} is:

$CD_i^2 = (1/p) * t_i^2 * h_{ii} / (1 - h_{ii})$

The ratio $h_{ii} / (1 - h_{ii})$ measure the influence of the ith observation on estimating the parameters, since t_i is also an outlier measure, combining t_i and the ratio $h_{ii} / (1 - h_{ii})$ results in a measure of the overall impact of any single observation on the solution.

5) Belsley et al (1980) used a similar function to CD_i^2 with using $s_{(i)}$ as:

$DFFITS_i = (e_i / s_{(i)}) * h_{ii}^{.5} / (1 - h_{ii})$

$DFFITS_i$ measures the influence on the prediction when the ith observation is deleted. Observation with $DFFITS_i > 2(p/n)^{.5}$ should be scrutinized. Belsley et al (1980) defined a diagnostic based on the change in the ith regression coefficient, namely

$DFBETAS_j(i) = b_j - b_j(i) = (X'X)^{-1} X_i e_i / (1 - h_{ii})$

where X_i is the ith column of X matrix, $b_j(i)$ is the least square regression coefficient after deleting the ith observation, and the cut-off value for DFBETAS is $2/(n)^{.5}$.

Hocking (1983) pointed out that $DFFITS_i$ and $DFBETAS_j(i)$ are not invariant under nonsingular transformation, and Cook's distance is preferable.

6) Graphing is an effective way of detecting the outlier. Probability plot or Rankit Plot of Daniel and Wood(1980) can be used to check if the residuals are approximately normally distributed. The ordered values of residuals e_i^* (the vertical coordinate) could be plotted against $\Phi^{-1}(e_i - 3/8) / (n + 1/4)$ (horizontal coordinate)..

Φ^{-1} is the inverse of the standard normal distribution function , and n is the number of non-missing data.

Other diagnostic functions like Studentized residuals, Jackknifed residuals can be plotted against the predicted value of y. Cook and Weisberg (1982) suggested to plot ordinary, absolute and Studentized residuals against $(1 - h_{ii})s_i$.

Single-case diagnostics, although they are simple from a computational point of view, often fails to reveal the impact of small group of cases because the influence of one point could be masked or obscured by another. In other words, two or several outliers can act togther in a complicated way to enforce or to offset each other's influence. Also, swamping problem which statisticaly defined as a problem arise from including a "non-outlier" in a group of observations judged to be outliers. Therefore, multiple diagnostics is an urgent solution to the routine detection of outliers.

B. Diagnostics of multiple outliers

An influential subset of outliers is a natural generalization of an influential outlier. Cook and Weisberg (1980) recommended the use of CD_I^2 where I is an index set corresponding to the subset cases. Note that CD_I^2 reduces to CD_i^2 in the single case. Andrews and Pregibon (1978) proposed the determinantal ratio:

$R_I = |(Z'_I Z_I)| / |(Z'Z)|$

where:

R_I is a measure of the remoteness of the subset of cases with index I , If $R_I = 1$ then the subset of outliers is not influential.

Z is the X matrix with the vector of observation y appended, and

Z_I is the Z matrix with the rows of Z in the index set I deleted.

Gray and Ling (1984) considered the modified hat matrix $H^* = Z(Z'Z)^{-1}Z'$

where H^* has the same properties as H, such as symmetry and idempotency. They have shown that $H^* = H + (e'e /$ sums of squares due to residuals). So H^* combines the leverage (h_{ii}) and the residual information (i.e., information contained in H and e). Subsets of cases jointly influential are often associated with submatrices of H^* containning several elements with large absolute values. The jointly influential cases exhibit large values of determinant of h_{ii} and a nearly block diagonal structure can be observed in H^*.

The objective of this study is to detect single or multiple outliers under the analysis of mixed linear model and to investigate the effects of the outliers on the single or linear combination of the parameters estimated by mixed model analysis.

METHODS

The mixed linear model considered is,

$$y = Xb + Zu + e. \qquad (1)$$

where:

y is (n*1) vector of observation.
X is (n*p) incidence matrix.
b is (p*1) vector of a fixed effect.
Z is (n*q) incidence matrix associated with random vector.
u is (q*1) vector of random effects. E(u) = 0, V(u) = G.
e is (n*1) vector of random error. E(e) = 0, V(e) = R = Is^2.

The mixed model in equivelent form is:

$y = X b + e$

where $E(e) = 0$ and $V(e) = V = ZGZ' + R$

Henderson's (1963) generalization of the Gauss-Markov theorem states that the best linear estimates of b and u are obtained by solving the overdetermined system of equations:

$$\begin{bmatrix} R^{-1/2}X & R^{-1/2}Z \\ 0 & G^{-1/2} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} R^{-1/2}y \\ G^{-1/2}0 \end{bmatrix} \qquad (2)$$

If C denotes the coefficient matrix of the LHS, then (2) could be rewritten as

$$C \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} R^{-1/2}y \\ G^{-1/2}0 \end{bmatrix}$$

where the last q rows are augmented observations that express the prior information about u.

$$C'C \begin{bmatrix} b \\ u \end{bmatrix} = C' \begin{bmatrix} R^{-1/2}y \\ G^{-1/2}0 \end{bmatrix}$$

which yields the normal equations of the mixed model

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y + G^{-1}0 \end{bmatrix} \qquad (3)$$

Solution to (3) yields b and u which can be used to find the predicted value.
An outlier is an observation with a large residual .

Procedure:

1) Set up the mixed model equations as in (3).
2) Compute $e = y - \hat{y}$.
3) Compute H matrix such that $H = C'(C'C)^{-} C$. The leverage points are the diagonal elements of $C'(C'C)^{-} C$.
4) Compute Studentized, Jackknifed, Cook's distance,...etc
5) Compute H^* and R_I of multiple case diagnostic. If G is not diagonal, it is cumbersome to get $G^{-1/2}$ due to the possibility o f negative covariances. However, H matrix can be computed because the concern, in single case diagnostic, is the diagonal elements h_{ii}. Moreover, an estimate of H matrix can be found via the covariance of the predicted value,

$Cov(y) = \sigma^2 H$ where σ^2 could be replaced by the estimated variance [$s^2 = e'e / (N-p)$].

6) The influence of the outliers or influential observations on the estimated parameters can be found by the following steps:

a) Estimate the standardized difference of the estimated parameters such as:

$(b - b_{(i)}) / s_{(i)} * (f_{ii})^{.5} * 100$

where:

b and $b_{(i)}$ are the solution to the mixed model equation (3) before and after deleting the outliers, $s_{(i)}$ is the standard deviation after deleting the ith observation, and f_{ii} is the diagonal elements of the inverse coefficient matrix $(C'C)^{-1}$.

b) To find the effect of outliers on a linear function of selected subset of the parameters of fixed or random effects, one has to specify the linear combination by using the matrix L which is q*q matrix with rank q. The extension to Cook (1977) can be found such that

$CD_{(s)} = (s - s_{(i)})' M^{-1} (s - s_{(i)}) / qs^{-1}$

where

$M = L (C'C)^{-} L$

q = the number of unknown parameters to be selected for linear combination.

s^2 = an estimate of σ^2.

NUMERICAL EXAMPLE

Given the data in Table 1 (adapted from Schaeffer,1975), the sire model used is,

$y = m + HYS + Sire + e$

Number of HYS = 6, number of sire = 6 and total number of observations = 40

$V(e) = R = I s^2_e$

Assume $s^2_e / s^2_s = 15$ and

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1/2 & 3/4 & 3/4 & 3/4 \\ 0 & 1/2 & 1 & 3/4 & 3/4 & 3/4 \\ 0 & 3/4 & 3/4 & 5/4 & 3/4 & 1 \\ 0 & 3/4 & 3/4 & 3/4 & 5/4 & 1 \\ 0 & 3/4 & 3/4 & 1 & 1 & 9/8 \end{bmatrix}$$

$$\begin{bmatrix} X'X & X'Z \\ X'Z & Z'Z + A^{-1} s^2_e / s^2_s \end{bmatrix} \begin{bmatrix} b \\ u \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

The following postulated scinarios were implemented to study the effect of single and multiple outliers on the solution of mixed model equations:

Single outlier: observation "20" was deliberately set to be 10.

Multiple outliers: observations "4", "20" and "35" were set to be 20, 10 and 30, respectively..

a) Sol (1) where there is no outliers as given in Table 1.

b) Sol (2) has single outlier (observation "20" =10).

c) Sol (3) where single outlier is deleted (i.e, observation "20" deleted).

d) Sol (4) has multiple outliers (obs "4" =20, obs "20" =10, obs "35" =30).

e) Sol (5) where multiple outliers are deleted (obs "4", "20", and "35" deleted).

f) Sol (6) is DFBETAS$_i$ when observation "20" is an outlier.
g) Sol (7) is DFBETAS$_j$ (i) when obs "4", "20", and "35" are outliers.

RESULTS AND DISCUSSION

Table 2 shows different statistics for detecting single outlier (observation "20"). The diagonal elements (h_{ii}) of the Hat matrix have not shown abnormalities because the outliers which ccurred in y vector are completely ignored by the Hat matrix (H). Functions of residuals and h_{ii} elements like standardized residuals, Student, Jackknife, CD_I^2 and DFFITS show that outlier observation "20" has a large distinctive value as compared to the rest of the data. Similar results are observed in Figures 1 and 2.

Solution to mixed model equations by including or excluding the single or multiple outlier(s) are shown in Table 3. In the case of single outlier, observation "20" clearly influenced solutions HYS_3, S_2, S_3, S_5 and S_6, i.e., influenced both fixed and random effects. Solutions HYS_3 and S_5 are affected the most because the outlier of Observation 20 belongs to the subclsses of HYS_3 and S_5. Solutions S_2, S_3 and S_6 are affected to a lesser extent through their relationship to S_5. These results were also supported by the values of DFBETAS (Sol 6). Sol(4) and Sol(5) and FBETAS (Sol 7) show the joint effect of multiple outliers (observation "4," "20" and "35") on the solution vector. Since observations "4", "20" and "35" were set to be outliers in obtaining Sol(4) and these outliers belong to the subclasses of HYS 1, 3 and 6 and Sires 1, 4 and 5, solutions for the corresponding levels of factors are highly affected. Comparison of Sol(1) and Sol(5) in Table 3 revealed that deletion of these outliers did not exhibit much impact on the solution vector.

Cook and Weisberg (1982) ratio (CD_i^2) was estimated after deleting the outlier (observation "20") as an extremely influential observation. The estimate of (CD_i^2) was .355. However, the estimate for multiple outliers, (CD_I^2) for linear combination of random effects was .003. The last estimate is important to detect the impact of influential observations as outliers. The value of D_I^2 is the squared standardized distance over which the parameters move when estimated without the influential observations . A large value (> 0) means a large influence on the parameters. Andrews and Pregibon (1978) ratio after deleting observations "4", "20" and "35" was 0.086. A small ratio (<1) is an indication of influential outliers.

It is important to note that the effects of deleting a single outlier or multiple outliers become less as the sample size increases. DFBETAS$_i$ decreases in proportion to $(n)^{.5}$. The outliers produce large residuals when the chosen model is fitted to the data. Outlier does not necessarily mean that the observation is influential with respect to the fitted equation i.e., does not ecessarily influence the parameters. Deleting observation with small residual may greatly influence the parameters as shown by Andrews and Pregibon (1987). So, if an influential outlier is shown to be due to error, it should be deleted subsequently.

Modifying the hat matrix as proposed by Gray and Ling (1984) has shown a block diagonal structure for the example data (Table 4). The influential observations were isolated in a block on the diagonal H*. The advantages of H* is to avoid the treatment of all possible outliers as a stepwise procedures which was described by Swallow and Kianifard (1996). It should be noted that the diagonal submatrices with large | h_{ij} | elements do not necessarily correspond to the nfluential subset of outliers. They are merely promising candidates for being influential outliers and other diagnostic measures should be used.

Mixed linear model has a vast application in animal breeding. The problem of outliers and the diagnostics of influential observations play an important role in several ways, for example:1) The outlier can be the most important observation in the data set and their identification represents the highest priority for animal breeder especially in identifying superior bull for AI use.
2) Outlier will affect the solution to the mixed-model equations (3) and consequently the

rediction of the linear function K'b + M'u provided K'b is an estimable function, so culling (i.e., delecting observation) has an impact on the linear predictor.

) Robust estimates of variance components could be obtained as a simple modification of REML. Fellner (1986) estimated variance components for mixed model by using the random part of the inverse of the coefficient matrix and an odd bounded monotonic function found by Huber (1964).

-) Fellner (1986) found that outliers could occur among the elements of random effects u. Situations exist in animal breeding in which data have been subjected to prior selection. Henderson (1975) described this as L'u selection i.e., selection on u.

REFERENCES

1) Andrews, D.F. and Pregibon, D. (1987). Finding the outliers that matter. J. Royal Stat. Soc. Ser. B. 40:85-93.
2) Atkinson, A.C. (1982). Regression diagnostics, transformation and constructed variables. J. Royal Stat. Soc. Ser. B. 44:1-36.
3) Ashish, S. and Srivastava, M. (1990). Regression Analysis Theory,Methods and Applications. Springer-Veralg. NY, NY.
4) Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). Regression Diagnostics. John Wiley & Sons.NY, NY
5) Cook, R.D. (1977). Detection of influential observation in linear regression. Technometrics,19:15-18.
6) Cook, R.D. (1979). Influential observation in linear regression. JASA 70:169-174.
7) Cook, R.D. and Weisberg, S. (1982). Residual and influence in regression. Chapman Hall. London.
8) Daniel, C. and Wood, F.S. (1980). Fitting Equation to Data. John Wiley & Sons. NY, NY.
9) Draper, N.R. and John, J.A. (1981). Influential observations and outliers in regression. Technometrics 23:4-26.
.0) Fellner, W.H. (1986). Robust estimation of variance components. Technometrics 28:51-60.
!1) Gray, J.B. and Ling, R.F. (1984). K-clustering as a detection tool for influential subsets in regression. Technometrics 26:305-318.
12) Hampel, F.R., Ronchetti, E.M., Rousseuw, P.J. and Stathel, W.A. (1986). Robust Statistics: The Approach Based on Influence Function. John Wiley & Sons, NY, NY.
13) Henderson, C.R. (1963) . Selection Index and Expected Genetic Advance. "Statistical Genetics and Plant Breeding". Nat'l Res. Coun. Pub. 982 - Nat'l Acad. Sci. 141-163.
14) Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. Biometrics 31:423-447.
15) Huber, P.J. (1964). Robust estimation of a location parameter. Ann. Math. Stat. 35:73-101.
16) Hocking, R.R. (1983). Development in linear regression methodology. (1959-1982). Technometrics 25:219-230.
17) Hocking, R.R. and Pendleton, O.J. (1983). The regression dilemma. Commun. Stat. 12:497-527.
18) Hoaglin, D.C. and Welsch, R.E. (1978). The Hat matrix in regression and ANOVA. Am. Stat. 32:17-22.
19) Paul, S.R. (1983). Sequential detection of unequal points in regression. The Statistician 32:417.
20) Rousseuw, P. and Leroy, A.M. (1987). Robust regression and outlier detection. John Wiley & Sons, NY, NY.
2!) Schaeffer,L.R.(1975) Dairy Sire Evaluation for milk and fat production.Guelph,Ontario: Univ .Guelph.
22) Stevans, J.P. (1984). Outliers and influential data points in regression analysis. Psychol. Bull. 95: 334.
23) Swallow,W.H. and Kianifard, F. (1996). Using robust scale estimates in detecting multiple outliers in linear regression. Biometrics 52:345-356

Table 1. Daughter's milk yield (MY) by sires and HYS.

Obs	HYS	Sire ID	MY	Obs	HYS	Sire ID	MY
1	1	A	130	21	3	E	170
	1	A	110	22	3	E	160
	1	A	150	23	3	E	180
4	1	A	141	24	3	E	210
5	1	B	150	25	3	E	203
6	1	B	140	26	4	A	140
7	1	B	159	27	4	A	162
8	1	C	116	28	4	F	186
9	1	C	145	29	4	F	190
10	1	C	155	30	4	F	150
11	2	B	110	31	5	A	151
12	2	B	141	32	6	A	170
13	2	B	160	33	6	A	134
14	2	D	130	34	6	A	120
15	2	D	168	35	6	D	190
16	3	B	124	36	6	D	160
17	3	B	150	37	6	D	200
18	3	B	170	38	6	D	196
19	3	B	180	39	6	F	193
20	3	E	156	40	6	F	170

Table 2. Measures for detecting single outlier.

Index	h_{ii}	e/s	Studentized	Jackknifed	$CD^2_{(i)}$	DFFITS
1	0.2175	-0.0963	-0.1088	-0.1679	0.0270	-0.0885
2	0.2175	-0.6288	-0.7109	-1.0970	1.1710	-0.5782
3	0.2175	0.4363	0.4932	0.7608	0.5630	0.4011
4	0.2175	0.1966	0.2222	0.3429	0.1141	0.1808
5	0.2770	0.1561	0.1836	0.2832	0.1078	0.1753
6	0.2770	-0.1102	-0.1295	-0.1999	0.0542	-0.1237
7	0.2770	0.3958	0.4654	0.7180	0.6911	0.4444
8	0.3291	-0.7200	-0.8791	-1.3561	3.1580	-0.9498
9	0.3291	0.5220	0.0637	0.0982	0.0136	0.0688
10	0.3291	0.3184	0.3888	0.6000	0.6180	0.4200
11	0.2779	-0.8246	-0.9704	-1.4971	3.0221	-0.9289
12	0.2779	0.0009	0.0010	0.0015	0.0000	0.0009
13	0.2779	0.5068	0.5964	0.9201	1.1409	0.5709
14	0.3755	-0.3474	-0.4396	-0.6782	0.9678	-0.5259
15	0.3755	0.6644	0.8407	1.2970	3.5411	1.0057
16	0.2487	-0.9524	-1.0988	-1.6950	3.3291	-0.9751
17	0.2487	-0.2601	-0.3001	-0.4629	0.2478	-0.2663
18	0.2487	0.2725	0.3143	0.4849	0.2716	0.2789
19	0.2487	0.5387	0.6215	0.9588	1.0649	0.5516
20	**0.1661**	**-4.0293**	**-4.4129**	**-6.8069**	**-32.3069**	**3.376**

CFV (cut off value) for h_{ii} =2p/n =.6 ; CFV for DFFITS = $2(p/n)^{.5}$ =1.095;
CFV for CD_i^2 = 1; CFV for Studentized, Jackknifed and e_i / s_i = arbitrary value = 2.5

Table 2 (o .inued

Index	h_{ii}	e/s	Studentized	Jackknifed	$CD^2_{(i)}$	DFFITS
21	0.1661	0.2311	0.2530	0.3904	0.1062	0.1742
22	0.1661	-0.0352	-0.0385	-0.0595	0.0024	-0.0265
23	0.1661	0.4973	0.5446	0.8402	0.4922	0.3749
24	0.1661	1.2962	1.4194	2.1897	3.3432	0.9772
25	0.1661	1.1098	1.2153	1.8748	2.450[illegible]	-0.6820
26	0.3531	-0.4812	-0.5983	-0.9231	1.6287	0.1482
27	0.3531	0.1046	0.1300	0.2006	0.0769	0.4470
28	0.2681	0.4096	0.4788	0.7386	0.6996	-0.5991
29	0.2681	0.5161	0.6033	0.9307	1.1107	0.5632
30	0.2681	-0.5489	0.6417	-0.9899	1.2566	-0.5991
31	0.5000	-0.2529	-0.3577	-0.5519	1.0664	-0.5519
32	0.5000	0.2529	-0.3577	-0.5519	1.0664	0.5519
33	0.2970	-0.7173	-0.8555	-1.3197	2.5762	-0.8578
34	0.2970	-1.0901	-1.3001	-2.0057	5.9498	-0.1304
35	0.2185	0.4383	0.4958	0.7649	0.5728	0.4045
36	0.2185	-0.3605	-0.4078	-0.6291	0.3875	-0.3327
37	0.2185	0.7046	0.7970	1.2296	1.4802	0.6502
38	0.2185	0.5981	0.6765	1.0437	1.0665	0.5519
39	0.3528	0.5197	0.6460	0.9966	1.8952	0.7357
40	0.3528	-0.0927	-0.1152	-0.1778	0.0603	-0.1313

CFV (cut off value) for h_{ii} =2p/n =.6 ; CFV for DFFITS = $2(p/n)^{.5}$ =1.095;
CFV for CD_i^2 = 1; CFV for Studentized, Jackknifed and e_i / s_i = arbitrary value = 2.5

Table 3. Solution for HYS and sire effects for different cases of influential observation.

Class	Sol (1)	Sol (2)	Sol (3)	Sol (4)	Sol (5)	Sol (6)	Sol (7)
HYS_1	139.317	139.314	139.317	127.142	138.500	-0.033	-130.072
HYS_2	135.860	136.294	135.808	136.039	135.892	3.915	1.208
HYS_3	165.201	151.272	166.864	149.860	166.814	-159.738	-176.848
HYS_4	163.451	163.940	163.393	163.904	163.371	4.483	3.615
HYS_5	166.160	166.219	166.153	167.657	166.165	.360	8.243
HYS_6	166.506	166.706	166.483	147.220	164.189	2.186	-169.788
S_1	-5.660	-5.719	-5.653	-7.157	-5.665	-1.166	-26.676
S_2	5.306	4.570	5.394	5.597	5.397	-13.967	03.445
S_3	3.184	4.009	3.085	5.139	3.100	15.662	35.202
S_4	6.889	6.909	6.888	6.007	6.675	.316	-10.425
S_5	6.377	4.708	6.577	6.307	6.655	-28.278	-5.368
S_6	7.355	6.580	7.448	7.765	7.491	-12.782	4.106

Sol(1) : obs(20) =156; Sol(2): obs (20) =10 ; Sol (3): obs(20) deleted;
Sol(4) : obs (4)=20, obs(20) =10; obs (35) =30; Sol(5): obs (4,20,35) deleted.
$DFBETAS_1$ [Sol(6)]: obs (20) deleted; $DFBETAS_2$[Sol(7)]: obs (4,20,35) deleted.

Table 4. H matrix for multiple outliers (Entries are rounded values of $100h_{ij}^{*}$).

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	22	22	23																																					
2		22	23																																					
3	22	23	24																																					
4				34																																				
5					28	28	28																																		
6					28	28	28																																		
7					28	28	28																																		
8								34	32	32																														
9								32	33	33																														
10								32	33	33																														
11											30	28	27																											
12											28	28	28																											
13											27	28	28																											
14														38	37																									
15														37	39																									
16																26	25	24																						
17																25	25	24																						
18																24	24	25																						
19																			49																					
20																			17	17	17	18	18																	
21																			17	17	17	17	17																	
22																			17	17	18	19	18																	
23																			18	17	19	21	21																	
24																			18	17	18	21	20																	
25																								27	27	27														
26																								27	27	26														
27																								27	26	28														
28																											35	35												
29																											35	36												
30																													27	27	27									
31																													27	27	26									
32																													27	26	28									
33																																50	49							
34																																49	50							
35																																		30	30					
36																																		30	30					
37																																				45				
38																																					22	22	22	
39																																					22	25	25	
40																																					22	25	25	

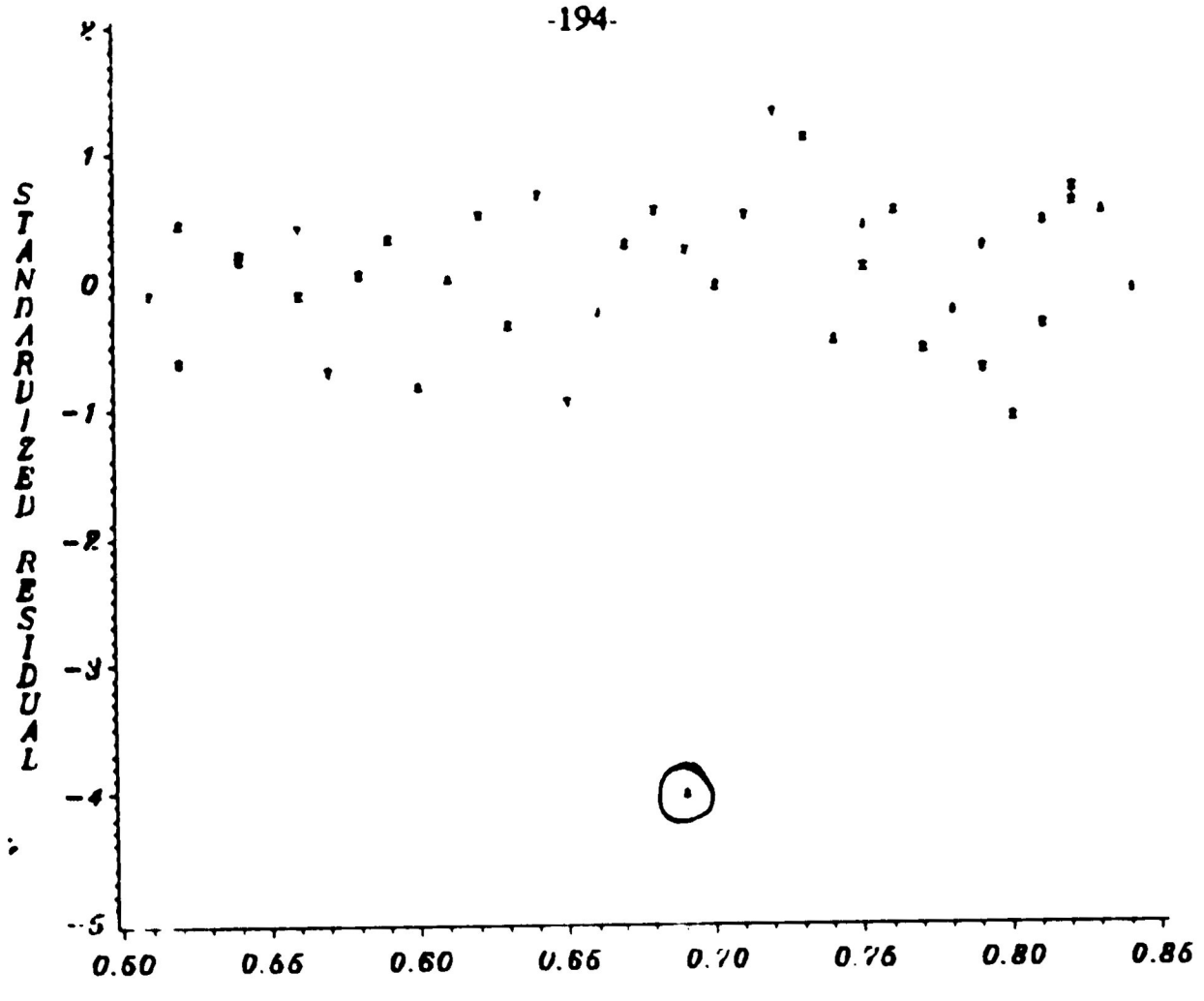

FIG 1. DIAGNOSTIC OUTLIER BY PLOTTING STANDARDIZED RESIDUAL VS PROBABILITY PLOT (RANKIT PLOT)

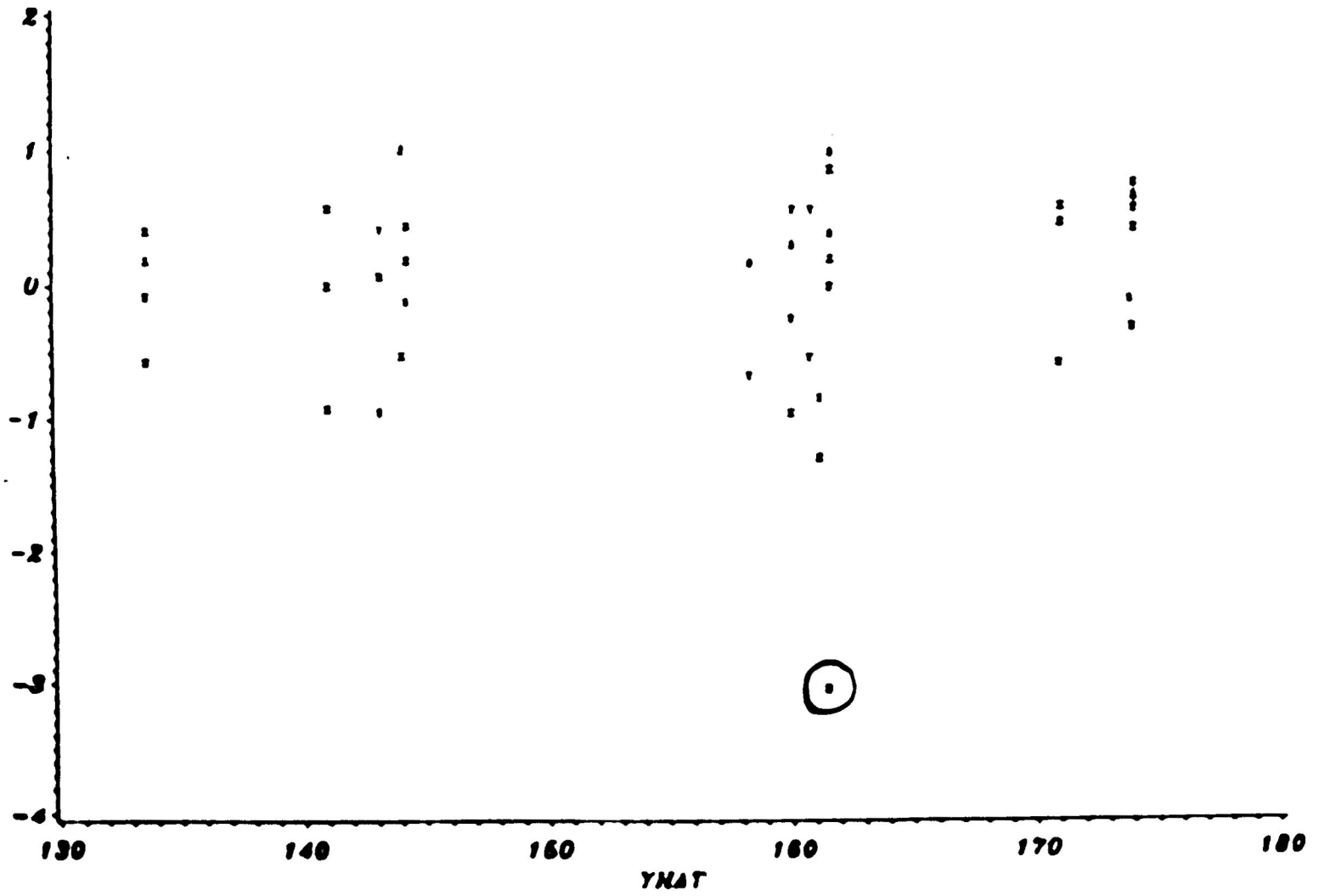

FIG 2. DIAGNOSTIC OUTLIER BY PLOTTING DEFITS VS PREDICTED VALUE