Indian Journal of Agricultural Sciences **85** (1): 55–8, January 2015/Article

Robust analysis of agricultural field experiments

RANJIT KUMAR PAUL[1], LALMOHAN BHAR[2], SANJEEV PANWAR[3] and ANIL KUMAR[4]

Indian Agricultural Statistics Research Institute, New Delhi 110 012

Received: 6 March 2013; Revised accepted: 28 July 2014

ABSTRACT

Agricultural data generated from designed experiments are also prone to occurrence outliers. It is well known that Least Squares (LS) model can be distorted even by a single outlying observation. An outlier is one that appears to deviate markedly from the other members of the sample in which it occurs. The sources of influential subsets are diverse. Rousseeuw (1984) introduced a robust method known as Least Median of Squares (LMS) for linear regression models. By this method, the median of squares errors is minimized in order to obtain parameter estimates. It turns out that this estimator is very robust with respect to outliers. Since it focuses on the median residual, up to half of the observations can disagree without masking a model that fits the rest of the data. Therefore, the breakdown point of this estimator is 50%, the highest possible value. In the present investigation, this method is applied to analyze the data set containing outlying observations generated from agricultural field experiments. The data sets for the present investigation have been taken from Agricultural Field Experiments Information System, IASRI, New Delhi.

Key words: Agricultural experiments, Block design, Least Median of Squares estimation, Outlier

Statistical models, like all mathematical models for real phenomena, are chosen as tools to describe certain aspects of the observed phenomena. A model is, at best, correct only in an approximate sense, because the model is generally fitted under certain assumptions. Besides the appropriateness of the postulated model, standard statistical techniques are based on the assumption that the data have been sampled independently from the same distribution. However, this assumption is frequently violated in practice; one of the reasons being the presence of outliers. Observations that in the opinion of the investigator stand apart from the bulk of the data have been called "outliers", "extreme observations" "discordant observations", "rouge values", "contaminants", "surprising values", "mavericks" or "dirty data". An outlier is one that appears to deviate markedly from the other members of the sample in which it occurs. The sources of influential subsets are diverse. First, there is the inevitable occurrence of improperly recorded data, either at their sources or in their transcription to computer readable form. Second, observational errors are often inherent in the data. Although procedures more appropriate for estimation than ordinary least squares exist for this situation, the diagnostics may reveal the unsuspected existence and severity of observational errors. Third outlying data points may be legitimately occurring extreme observations. Such data often contain valuable information that improves estimation efficiency by its presence. Even in this beneficial situation, however it is constructive to isolate extreme points and to determine the extent to which the parameter estimates depend on these desirable data. Fourth, since the data could have been generated by model(s) other than that specified, diagnostics may reveal patterns suggestive of these alternatives. Consider the following example:

Agricultural data generated from designed experiments are also prone to occurrence outliers. In general, literature on outliers is very vast. A number of statistics are now developed to detect outliers in a data set following linear model. However, these statistics are developed under the assumption that the data are generated from a kind of linear model where the design matrix is of full rank. Though the concern over outliers is quite old, yet much attention is not paid on it in the field of designed experiments. Estimation of some functions of parameters of interest, e.g. treatment effects is of prime interest to an experimenter, which may be severely affected in the presence of outliers. On this line of interest, Bhar and Gupta (2001) developed some statistics for detecting outliers in designed experiments. They modified Cook statistic for its application to design of experiments, which is a follow up work of Cook (1977).

Rousseeuw (1984) proposed least median of squares estimation procedure for robust analysis of linear regression data. Paul and Bhar (2011, 2012) modified the robust methodology for its application to designed experiments. It turns out that this estimator is very robust with respect to

[1]Scientist (e mail: ranjitstat@gmail.com), Division of Statistical Genetics, [2]Principal Scientist (e mail: lmbhar@gmail.com), Division of Design of Experiments, [3]Scientist (e mail: scientist1775@yahoo.co.in), Division of Forecasting and Agricultural System Modelling

outliers. Since it focuses on the median residual, up to half of the observations can disagree without masking a model that fits the rest of the data. Therefore, the breakdown point of this estimator is 50%, the highest possible value. In the present investigation an attempt has been made to apply the essence of LMS technique to experimental data generated from agricultural field.

MATERIALS AND METHODS

Consider the linear regression model $Y = X\beta + \varepsilon$ (1) where Y is a $n \times 1$ vector of observations, X is a $n \times p$ matrix of explanatory variables, β is $p \times 1$ vector of parameters and ε is a $n \times 1$ vector of errors such that $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon') = \sigma^2 I$. We also assume that the rank of X is p.

Let b be any estimate of β. With n observations the residuals from this estimate are $e_i(b) = y_i - x_i'b$, ($i = 1,2,\ldots,n$), where x'_i is the ith row of X. The LMS estimate $\hat{\beta}_p$ is the value of b minimizing the median of the square residuals $e_i^2(b)$. Thus $\hat{\beta}_p$ minimizes the scale estimate

$$\sigma^2(b) = e^2_{(med)}(b) \quad (2)$$

where $e^2_{[k]}(b)$ is the k^{th} ordered squared residual. In order to allow for estimation of the parameters of the linear model the median is taken as

$$med = \text{The integer part of } (n + p + 1)/2 \quad (3)$$

The parameter estimate satisfying (2) has asymptotically, a break down point of 50%. Thus, for large n, almost half the data can be outliers, or come from some other model and LMS will still provide an unbiased estimate of the regression coefficients.

The definition of $\hat{\beta}_p$ in (2) gives no indication of how to find such a parameter estimate. Fitting an LMS regression model poses some difficulties. The first is computational. Unlike least squares regression, there is no formula that can be used to calculate the coefficients for an LMS regression. In fact, it appears that this computational complexity is inherent to all high breakdown regression estimators. Rousseeuw (1984) has proposed an algorithm to obtain LMS estimator.

The algorithm proceeds by repeatedly drawing sub samples of p different observations. For such a sub sample, indexed by $J = \{i_1, i_2, ..., i_p\}$, one can determine the regression surface through the p points and denote the corresponding vector of coefficients by β_J. We will call such a solution β_J a trial estimate. For each β_J one also determines the corresponding LMS objective function with respect to the whole data set. This means that the value

$$\underset{i = 1...n}{\text{med}} \; (y_i - x'_i \beta_J)^2 \quad (4)$$

is calculated. Finally, one will retain the trial estimate for which this value is minimal. But how many sub samples should we consider? In principle, one could repeat the above procedure for all possible sub samples of size p, of which there are ${}^nC_p$ sub samples. Unfortunately, ${}^nC_p$ increases very fast with increase in n and p, in many applications, this would become infeasible. In such cases, one performs a certain number of random selections, such that the probability that at least one of the m sub samples is good is almost 1. A sub sample is good if it consists of p good observations of the sample, which may contain up to a fraction ε of bad observations. The expression for this probability (Rousseeuw 1984), assuming that n/p is large, is

$$1 - (1-(1-\varepsilon)^p)^m \quad (5)$$

By requiring that this probability must be near 1 (say at least 0.95 or 0.99), one can determine m for given value of p and ε.

Though LMS estimator has a high break down point, yet this method did not get much popularity in designed experiments. LMS method gives parameter estimates based on clean observations only and thus outliers or distributional extreme observations cannot create any problem in parameter estimation or rather they do not have any impact on parameter estimation. One of the possible reasons why LMS method is not being used in designed experiments might be its computational difficulties. There is no exact formula for computing this estimator explicitly in linear regression models. Rousseeuw (1984) provided an algorithm for computing this estimator in linear regression models. As mentioned earlier, by this algorithm all possible subsets of size p, where p is the number of parameters in the model are fitted separately. Residuals from each of these fitted models are calculated. The median of the squared residuals for each set is calculated. The subset that gives minimum median is chosen as the final set and analysis is carried out on this sub set. Application of this algorithm to designed experiments possesses some problems. The main difficulty is the connectedness property. If we choose the size of subset as p, the design may become disconnected for some subsets or all subsets. Connectedness property is a very important property for designed experiments. If the design is connected, then all the elementary treatment contrasts are estimable, a desirable property to all experimenters. Secondly, in case of design of experiments, we are mainly interested in estimation of some functions of treatment effects. This will also be severely affected if we choose a very small subset of data for estimating the treatment effects. Combating all these problems, we propose an appropriate LMS procedure for application to designed experiments.

Now we consider the applicability of LMS procedure in designed experiments. As mentioned earlier that the connectedness is the main problem in designed experiments, LMS method as such cannot be applied. Therefore, this is appropriately modified for application into experimental data. The LMS method is primarily designed to tackle the problem of outliers. In case of designed experiments, generally one or two outlying observations are present in a particular data set. We, therefore, proposed LMS method in the following manner: (i) Consider the size of the subset as $n - 1$ or $n - 2$. Here, we assumed that the design remain connected after loosing one or two observations. (ii) Obtain least squares residuals for each subset. There will be in total

${}^{n}C_{n-1}$ or ${}^{n}C_{n-2}$ subsets of data. (iii) Square the residuals and obtain the median for each subset. (iv) Retain that subset which yields minimum median among all subsets. (v) Carry out usual analysis on the chosen subset.

It is well known that all Randomized Block Designs (RBD) are robust against the loss of any two observations, i e these designs remain connected even after loosing two observations. Therefore, there is no problem to apply LMS technique to RBD, by taking the size of the subset as *n*-2. There are also many block designs that are robust against the loss of one or two observations (Krishan Lal *et al.* 2001). However, this size of subset can be decreased for those designs that are robust against the loss of more than two observations. This method is applied to a number of experimental designs, taken from Agricultural Field Experiments Information System, IASRI, New Delhi. Relevant program for carrying out the analysis has been written in IML module of SAS software. For illustration consider the following example.

RESULTS AND DISCUSSION

Experimental data set 1

An experiment with 6 treatments was carried out in the randomized complete block (RCB) design with 4 replications at Mahatma Phule Agricultural University, Rahuri, Maharastra in 1987 with a view to test the validity of fertilizer adjustment equation in Groundnut. (Net plot size: 3.00m × 3.75m.). The treatments of the experiments are as follows: T_0 = Control (No fertilizer), T_1 = 25 kg/ha N + 50 kg/ha P_2O_5 T_2, = As per soil test (38 kg/ha N + 50 kg/ha P_2O_5), T_3 = 15 q/ha Target (11 kg/ha N + 16 kg/ha K_2O), T_4 = 20 q/ha Target (32 kg/ha N + 51 kg/ha P_2O_5 +31 kg/ha K_2O), T_5 = 25 q/ha Target (52 kg/ha N + 10 kg/ha P_2O_5 +56 kg/ha K_2O).

The data on yield per plot in quintals for different treatments is given in Table 1.

Table 1 Yield of sugarcane in kg/plot

Replication	Treatment			
	1	2	3	4
1	3.70	3.43	2.60	3.37
2	4.50	3.90	3.67	3.62
3	4.63	3.77	2.53	3.37
4	4.08	3.80	4.35	3.43
5	4.15	3.79	3.27	3.47
6	4.06	4.00	3.13	3.38

Analysis of this data is presented in Table 2. It was observed that the treatment effects were not significant at 5% level of significance.

LMS method

We then applied LMS-estimation procedure to the data. The result is presented in Table 3. The dramatic effect to note here is that the treatment effects are now significant at 5% level of significance.

Table 2 Analysis of variance with original data

Source of variation	DF	SS	MS	F	Significance level
Treatment	5	1.158	0.231	1.76	0.1823
Block	3	3.010	1.003	7.62	0.0025**
Error	15	1.976	0.131		
Total	23	6.145			

Table 3 Analysis of variance through LMS (Size of subset is n-2)

Source of variation	DF	SS	MS	F	Significance level
Treatment	5	0.906	0.182	4.97	0.0092**
Block	3	2.997	0.999	27.42	<.0001**
Error	13	0.474	0.036		
Total	21	4.377			

Cook-statistic was also applied to identify outlying observations, if any. It was found that observation number 15 and 16 are influential. In the final data set these two observations are actually deleted.

Experimental data set 2

An experiment with 10 treatments was carried out in the randomized block design (RBD) with 4 replications at Sugarcane Research Institute, Shahjahanapur, Uttar Pradesh with a view to find out the suitable herbicide to control weeds in sugarcane. (Net plot size: 8.00m × 5.40m.). The treatments of the experiments are as follows: T_0 = Control weeded check, T_1 = Local conventional method, T_2 = Trash mulching, T_3 = 1.0 kg ai/ha of 2, 4-D sodium salt and 0.50 kg a.i./ha of gramoxone at 3 weeks of planting followed by application of the same at 6-8 weeks of planting, T_4 = 2.0 kg ai/ha of Atrazine as pre-emergence spray, T_5 = 1.00 kg ai/ha of 2,4-D sodium salt at 8-10 weeks after planting, T_6 = 2.0 kg ai/ha of 2,4-D (Amine) as pre-emergence spray followed by spray of the same at 8-10 weeks after planting, T_7 = 2.0 kg ai/ha of Atrazine as pre-emergence spray followed by spray of glyphosate at 1.0 kg ai/ha at 6-8 weeks after planting, T_8 = 1.00 kg ai/ha of arochlor and 1.00 kg ai/ha of atrazine as pre-emergence spray, T_9 = 2.00 kg ai/ha of arochlor as pre-emergence spray.

The Table 4 shows the data on yield per plot in kilogram in different treatments.

Analysis of this data presented in Table 5 showed that the treatment effects were not significant at 5% level of significance.

LMS method

We now applied LMS technique by taking subset size as n-2. The result is presented in Table 6. The dramatic effect to note here is that the treatment effects are now almost significant at 5% level of significance.

Cook statistic was also applied for outlier detection and found that observation 14 and 39 corresponding to

Table 4 Yield of sugarcane in kg/plot

Replication	Treatment			
	1	2	3	4
1	2.52	2.77	2.32	2.31
2	2.82	2.77	2.38	2.14
3	2.42	2.52	2.44	2.38
4	2.67	3.69	2.30	2.13
5	2.50	3.21	1.90	2.51
6	3.01	3.05	2.46	2.79
7	2.65	2.64	2.35	2.21
8	2.62	2.53	2.47	2.52
9	2.18	2.47	2.15	2.66
10	2.57	2.82	2.26	2.35

Table 5 Analysis of variance with original data

Source of variation	DF	SS	MS	F	Significance level
Treatment	9	0.637	0.071	1.06	0.4206
Block	3	1.731	0.577	8.64	0.0003**
Error	27	1.802	0.067		
Total	**39**	4.171			

Table 6 Analysis of variance through LMS (Size of subset is *n*-2)

Source of variation	DF	SS	MS	F	Significance level
Treatment	9	0.707	0.0786	2.26	0.0519
Block	3	1.207	0.402	11.58	<.0001
Error	25	0.868	0.035		
Total	37	2.782			

treatment number 4 in 2nd replication and treatment number 9 in 4th replication are really influential. Incidentally, in the final data set these two observations are actually deleted.

Least Median of Squares method is a very robust method. Its breakdown point is 50%, the highest possible value. That is, this method can tolerate even a large number of discordant observations. We have applied this technique to a large number of real experiments; two examples are presented in the previous section. It is observed that if the data contains outlying observations, then LMS method out rightly rejects these outlying observations while selecting the final set. This is an advantageous procedure for a contaminated sample. Moreover, if the data set contains masked outliers, then LMS method is a very good method for estimating parameter effects that have very good statistical properties. It is therefore recommended that if it is sure that the data set does not contain any outlying observation or not having nonnormal errors, ordinary least square analysis is the best. If, however, nothing is known about the data, LMS method with the subset size of *n*-1 or *n*-2 can always give us good result; because it will guard against all possible unusual happenings.

REFERENCES

Bhar L and Gupta V K. 2001. A useful statistic for studying outliers in experimental designs. *Sankhya* **B**, **63**: 338–50.

Carroll R J. 1980. Robust methods for factorial experiments with outliers. *Applied Statistics* **29**: 246–51.

Chi E M. 1994. M-estimation in cross-over trials. *Biometrics* **50**: 486–93.

Cook R D. 1977. Detection of influential observation in linear regression. *Technometrics* **19**: 15–8.

Hampel F R. 1975. Beyond location parameters: robust concepts and methods. *Proceedings of the 40th session of the ISI* **46**: 375–91.

Huber P J. 1964. Robust estimation of a location parameter. *Annals of Mathematical Statistics*. **35**: 73–101.

Krishan Lal, Gupta V K and Bhar L. 2001. Robustness of designed experiments against missing data. *Journal of Applied Statistics* **28**: 63–79.

Paul R K and Bhar L. 2011. M-Estimation in block designs. *Journal of the Indian Society of Agricultural Statistics* **65**(3): 323–30.

Paul R K and Bhar L. 2012. Robust analysis of Block Designs: A new objective function. *International Journal of Agricultural and Statistical Sciences* **8**(**1**): 243–50.

Rousseeuw P J. 1984. Least median of squares regression. *Journal of the American Statistical Association* **79**: 871–80.