Indian Agricultural Statistics Research Institute
Library Avenue, Pusa, New Delhi – 110 012

Workshop Series **3**

OUTLIERS IN DESIGNED EXPERIMENTS

Organized at IASRI on July 26, 2007

Introduction

An outlier in a set of data is an observation (or an observation vector) that appears to be inconsistent with the remainder of the observations in that data set. Occurrence of outlier(s) is common in every field in which data collection is involved. In agricultural experiments, outlier(s) is/are likely to appear in the experimental data due to disease and or insect-pest attack on some plots in the field, or due to unintentional heavy irrigation on some particular block(s) or plot(s) of the experiment. Outlier(s) may also creep in due to transcription errors.

Presence of such abnormally high or low observations may cause a deviation from the assumptions particularly those of normality and homogeneity of observations. It is, therefore, important to detect the presence of outlier(s) along with deviations from these assumptions and suggest remedial measures.

The problem of outliers has been studied extensively in linear regression models. Approaches to study of outliers are generally divided into two broad categories: (i) to identify the outlier(s) for further study and (ii) to accommodate the possibility of outlier(s) by suitable modifications of the models and or method of analysis. The first approach relates to detection of outlier(s) while the second one relates to the study of robust methods of estimation of parameters that minimize the influence of outlier(s) on inference concerning parameters. A number of test statistics have been developed to detect outliers in linear regression models. Among them Cook-statistic is a widely used statistic. Other important test statistics for detection of outlier(s) are AP and Q_k-statistic. M-estimation procedure is a very powerful robust method of estimation used in linear regression model. In M-estimation a function of errors is minimized to obtain parameter estimates, unlike least squares method where sum of squares of errors is minimized. Each observation gets different weights for estimating parameters whereas in the usual procedure of least squares, all observations get equal weights. This function is called objective function. A good number of objective functions such as Huber's function, Andrew's function etc. are now available. Another procedure of robust estimation of parametric function is Least Median of Squares (LMS) method wherein median of the errors is minimized to obtain the parameter estimates.

Though, the general set up of an experimental design is that of a linear model, yet detection and testing of outlier(s) and application of robust methods in experimental designs need special attention because (i) the design matrix does not have full column rank (ii) interest is only in a subset of parameters rather than whole vector of parameters. Not much research appears to have been done on detection of outliers and robust methods of estimation in designed experiments. The available test statistic and robust procedures of estimation cannot be applied directly to this situation. Bhar and Gupta (2001, *Sankhya*, B63(3), 338-350) modified Cook-statistic for detecting outliers in block designs. John (1978, *Applied Statistics*, 27, 111- 119) provided some statistics for detecting outliers in factorial experiments.

One can, however, instead of taking post experimental remedial measures, take pre-experimental measures by adopting a robust design for experimentation. A robust design is insensitive to the presence of outlying observations in the sense that the inference problem on linear function of treatment effects is not affected by the presence of outliers in the experimental data. Box and Draper (1975, *Biometrika*, 62, 347-352) obtained robust regression designs in presence of a single outlier. Gopalan and Dey (1976, *Sankhya*, B38, 297-299) identified robust block designs through minimization of variance of discrepancy or bias in estimation of error variance. Bhar and Gupta (2001, *Sankhya*, B63(3), 338-350) used the minimization of average Cook-statistic to identify robust designs against presence of a single outlier. Sarker *et al.* (2005, *Metron*, 63(2), 177-191) established the

equivalence of these two criteria. All these investigations were restricted to single outlier experimental situations only. Therefore, there is a need to define a new criterion for identification of designs that are robust against the presence of more than one outlier.

The problem of outliers in linear regression models can be handled by using several statistical packages. These statistical packages are not capable of handling outliers in designed experiments. Thus with the development of new methodologies for tackling outliers in designed experiments, a user-friendly software for implementing these new techniques is also required.

In view of the above discussion, Indian Agricultural Statistics Research Institute (IASRI), New Delhi undertook a project entitled, Outliers in Designed Experiments, financed by AP-CESS fund of ICAR with the following objectives:

- To develop/identify suitable methodologies for detecting outliers in design of experiments.
- To develop/identify robust estimates of parameters of interest in designed experiments with special emphasis on M-estimation.
- To study the robustness of block designs against the presence of more than one outlier.
- To develop user-friendly software for detecting outliers and analyzing experimental data in presence of outlier(s).

Objective of the workshop

Dissemination of research findings to the stakeholders is very important for any research project. Therefore, a workshop was organized with the following objectives:

- To disseminate the research findings of this project to the stakeholders.
- To describe the applications of the theory developed in the analysis of data generated from designed experiments.
- To familiarize the participants with the application of the software for analysis of experimental data in presence of outlier(s).
- To give an exposure of the design resources server.
- To finalize the recommendations emerging from this workshop.

Programme of the workshop

The participants of the workshop included many eminent scientists actually engaged in field of experimentation and some renowned statisticians from various institutions. Among those who participated in the workshop were Dr. NN Goswami, former Vice-Chancellor, Chandra Sekhar Azad University of Agriculture and Technology, Kanpur, Dr. Rajendra Prasad, Ex-National Professor, Indian Council of Agricultural Research, Dr. Aloke Dey, Professor, Indian Statistical Institute, Delhi Centre, New Delhi and Dr. SD Sharma, Director, Indian Agricultural Statistics Research Institute, New Delhi. Dr. VK Gupta, National Professor and Co-investigator of the project made a presentation describing the motivation for taking up the project; introducing outliers, their presence in the experimental data, and also ways to handle outliers. Dr. Rajender Parsad, National Fellow and Co-investigator of the project made a presentation on Diagnostics in Designed Experiments. Through real life examples, he demonstrated that the assumptions of normality and homogeneity of error variances may be violated due to presence of outliers in the experimental data. Dr. LM Bhar, Principal Investigator gave comprehensive presentation on the salient findings of the project.

In brief following salient achievements of the project were presented:

Identification of outliers(s)

One statistic that has strong intuitive appeal for identification of outliers is the Cook distance. This measure was introduced in the context of linear regression wherein it measured the squared distance between the estimated parameters using the full set of data and the estimated parameters obtained after deleting an observation. The distance is obtained for all the observations by deleting one observation at a time. The observation giving the largest distance may be tested for being an outlier. Although Cook-statistic has strong intuitive appeal, its application to designed experiments is not straight forward. Therefore, Cook statistic is suitably modified for making it applicable for detecting outliers in designed experiments.

Cook-statistic, however, has a limitation that it is not capable of handling the problem of masking (the effect of an outlier is suppressed by the presence of another outlier). If one applies single outlier detection procedure, both the outliers may remain undetected. In the context of regression analysis, Pena and Yohai (1995, *Journal of Royal Statistical Society*, B57, 145-156) developed a statistic that takes care of the masking effect and also enables one to detect outliers. In the present investigation, this statistic has been appropriately modified for designed experiments. This statistic takes care of the masking effect.

The modified test statistic was used to detect outliers in experimental data obtained from Agricultural Field Experiments Information System (AFEIS), IASRI, New Delhi. Mainly those experiments were selected for detection of outliers which were found having data with non-normal and/ or heterogeneous errors as identified in another investigation namely A Diagnostic Study of Design and Analysis of Field Experiments. At least one outlier is detected

in 372 experiments out of 579 tried. This confirms the presence of outliers in the experimental data.

Robust methods of analysis

Once outlier(s) has/have been detected then the next question is, "what do we do with the outlier(s)?" One commonly used practice is to remove the outlier(s) and analyze the remaining data. But every observation generated contains some information about the parameters of interest and a lot of resources are spent on its collection. Therefore, we need to develop robust methods of estimation of parameters of interest. For application to designed experiments various M-estimation procedures like Huber's function, Andrew's function etc. have been modified by changing their tuning constants. Actually, for each M-estimation method, a different objective function is used and these objective functions are bounded by some constants known as tuning constants. Determination of these constants is subjective and depends on the type of data being analyzed as well as the experience of the analyst. A new objective function that determines weights for the observations using Cook-statistic has been proposed.

Least Median of Squares (LMS) method has been modified for application in designed experiments.

Robust methods of estimation available in the literature as well as modified have been applied to the real life experimental data. The application of these methods improves the credibility of the inferences drawn.

Robust designs

A new criterion based on modified Cook statistic for identifying robust designs against presence of more than one outlier has been developed. Using this criterion, all binary variance balanced block designs have been shown to be robust against the presence of two outlying observations.

Software developed

Graphic user interface based software has been developed for analyzing experimental data in the presence of outlying observations. The software has the following features:

- It can identify outliers in experimental data.
- It can directly apply the robust methods of estimation for analyzing the data. Here one has two options: M-estimation (Huber's function) or LMS method.
- It has option to analyze the data after deleting the outlying observations.

After the presentation Dr. Aloke Dey, gave his remarks on the project and the findings. He was appreciative of the efforts made in this project. He also felt that the findings of the project should be published in reputed journals. The findings of the project were well received by the statisticians as well as the experimenters.

Dr. Rajender Parsad demonstrated 'Design Resources Server' designed and developed by National Fellow and National Professor at IASRI. This server is available at Institute's web site at www.iasri.res.in/design. Participants appreciated the usefulness and importance of this server for strengthening the status of experimentation in NARS.

During the discussions, it was felt that before analysis of the experimental data, one should check for the presence of outlying observations. If outliers are found, appropriate measures should be taken as discussed in the workshop. Dr. Madhuban Gopal was very appreciative about the workshop and felt that dissemination workshops should be organized more frequently.

Recommendations of the workshop

Roust designs

- An experiment, in any field of agricultural sciences, should be conducted using designs that are robust against the presence of outlier(s). It is known that all binary, balanced block designs, many two-associate class partially balanced incomplete block designs, variance balanced row-column designs that satisfy the property of adjusted orthogonality, nested balanced incomplete block designs, proper binary balanced block designs for diallel crosses are robust against the presence of a single outlier. Binary balanced block designs have also been shown to be robust against the presence of two outliers. Therefore, the experimenters should adopt these designs for their experimentation whenever outliers(s) are suspected in the data to be generated. In some experimental data sets, more than two outliers may also be present. There is, therefore, a need to investigate the robustness of designs against the presence of two or more outliers in the experimental data by suitably defining appropriate criteria of robustness. Efforts should be made to evolve new robustness criteria, if required.
- Two criteria of robustness *viz.* minimization of average Cook-statistic and minimization of variance of discrepancy or bias in the estimation of error variance are equivalent in the presence of a single outlier. It would be of interest to study if this holds for more than one outlier case also!

Analytical techniques for outliers

- Before analyzing the experimental data, the data should always be subjected to diagnostic checks for the validity of assumptions involved in the analysis including the presence of outliers. If no outlier is detected, one should go ahead with usual analysis with the original data. On

the other hand if an outlier is detected, then further probing is required. Serious efforts should be made to ensure that there are no transcription errors or human error. The actual randomized layout of the design should also be looked into to locate for trends among the observations arising from nearby plots. If the extreme observations are due to human error, then non-statistical appropriate checks should be applied for its correction. If the experimenter is satisfied that the outlying observation(s) is(are) not due to transcription or recording errors, then the usual analysis of data may be carried out after deleting the outlier(s) or adopting analysis of covariance on the original data by defining pseudo-auxiliary variables.

- It may not always be desirable to delete any observation that is detected to be outlying because every observation contains useful information, more so when it is a true realization from the distribution from which other observations have come. In such a situation, robust methods of estimation of parameters of interest may be employed. Some of the robust methods of estimation useful in case of experimental data are M-estimation and Least Median of Squares (LMS). Application of M-estimation needs some special skills since it involves tuning constants. A proper choice of these constants gives efficient results. If it is known that the data contains only one or two outlying observations, then one can apply LMS method.
- For block size two Cook statistic and other statistics cannot be applied to detect outliers because of some mathematical problems. Therefore, some efforts are required for development of the procedure of detection of outliers in the experimental data generated from block designs with block size two.
- Several test statistics have been developed for detecting the outliers. But for the situation when more than one outlier is present in the data, the null distribution of the test statistic is not known. In such a situation, it is not possible to test the null hypothesis regarding the presence of outliers. Simulation studies may be carried out to obtain the null distribution of the test statistic.

Multi-response experiments

- Many experiments are conducted in NARS in which several responses are observed from a given experimental unit. Such experiments are known as multi-response experiments. Outlier(s) in multi-response experiments is/ are likely to appear. In the multi-response experiments the problem of outliers has not been studied in the literature. The problem is difficult in the sense that the outliers in multi-response experiments have to be defined appropriately. It may happen that the entire response vector is an outlier. It may also happen that the sub-vector of responses is an outlier but the whole response vector is not. This is the problem concerning a single outlying response vector. The problem becomes more difficult when there are more than one outlying response vectors. Hence, there is a need to make some serious efforts to develop test statistic for detecting outlying response vectors. For handling of outlier(s), robust procedures of estimation of treatment contrasts in the presence of outlying response vector(s) needs to be developed.
- Designs for multi-response experiments that are robust in presence of outlier(s) need to be identified.

General recommendations

- The participants strongly felt that such dissemination workshops should be organized more frequently for the benefit of scientists of NARS.
- The students at IASRI in the discipline of Agricultural Statistics, particularly the Ph.D. students, should be encouraged to take one minor in the discipline of agricultural sciences depending upon their research interest so that they get a first hand feel of the real problem.

LM Bhar
Rajender Parsad
VK Gupta

For Further Information Contact:
Director, I.A.S.R.I., Library Avenue, Pusa, New Delhi 110 012
Phone: 91-11-25841479; Fax: 91-11-25841564; E-mail: director@iasri.res.in

Published on behalf of the Director, Indian Agricultural Statistics Research Institute
(Indian Council of Agricultural Research), New Delhi 110 012