Outliers in Designed Experiments

(AP-CESS Funded Project)

LM Bhar

Rajender Parsad

VK Gupta

Indian Agricultural Statistics Research Institute

Library Avenue, New Delhi-12

FOREWORD

Data generated from designed experiments is analyzed assuming that observations are independently and identically distributed as normal with same variance. It is also assumed that the data set contains no abnormally high or low observations, i.e., the data set contains no outlying observations. These assumptions, however, get violated during experimentation. For instance, during the experimentation, there might be an infestation of a disease or insect attack on some plots in the field, or there may be unintentional heavy irrigation on some particular block(s) or plot(s) of the experiment, or at times there may be mistakes creeping in during recording of data, etc. The presence of abnormally high or low observations may cause non-normality and/or heteroscedastic errors and may influence the conclusion drawn. It is, therefore, important to detect deviations from assumptions and suggest remedial measures. Keeping this in view, this project entitled **Outliers in Designed Experiments,** funded by APCESS fund of Indian Council of Agricultural Research, was taken at the Institute.

For identification of outliers several available statistics like Cook-statistic, AP-statistic and Q_k-statistic were applied. Special emphasis has been given to study the problem of masking effect of outliers.

Another way of tackling outliers is to use robust methods of analysis of data. The popular robust methods of estimation viz., M-estimation and Least Median of Squares (LMS) were appropriately modified so as to make them applicable in designed experiments. All these methods are illustrated with some examples.

One can also minimize the influence of outlying observations by adopting a design that is insensitive to the presence of outlying observations. Such designs are known as robust designs. A criterion has been developed to identify robust designs that are robust against the presence of any two outliers. It has been found that all binary proper variance balanced block designs are robust against the presence of any two outliers.

One significant feature of the study is the development of statistical software for handling outliers. With the development of new methodologies for tackling outliers in designed experiments, user-friendly software for implementing these new techniques will be quite useful in drawing statistically valid conclusions.

For disseminating the findings of the project, a workshop was organized at IASRI, New Delhi. It gives me pleasure to mention here that the deliberations on the findings of the project were well received by the statisticians as well as the experimenters. I would like to complement Dr. L. M. Bhar, Dr. Rajender Parsad and Dr. V.K.Gupta for undertaking this project and obtaining very useful results. It is hoped that this report will be immensely useful for the practicing statisticians and the experimenters.

(**S.D. Sharma**)
Director, IASRI, New Delhi

PREFACE

The study of outliers is now diversified to almost every field of scientific study. 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. The second one relates the robust method of estimation.

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 sub set 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.

In the present investigation an attempt has been taken to explore applicability of the available test-statistics in designed experiments. Now a number of test statistics have been developed to detect outliers in designed experiments viz. Cook-statistic, AP and Q_k-statistic. These statistics are applied to real experimental data taken from Agricultural Field Experiments Information System (AFEIS), IASRI. It has been found that many of these experiments contain outliers. The detection of influential subsets or multiple outliers is more difficult, owing to masking and swamping problems. A method has also been developed for tackling outliers in designed experiments in presence of masking and swamping. The proposed method has been illustrated with an example.

Among robust methods, M-estimation and Least Median of Squares (LMS) method have been applied to experimental data after doing suitable modifications. In M-estimation a function of errors is minimized to obtain parameter estimates and thus each observation gets different weights for estimating parameters. This function is called objective function. A good number of objective functions such as Huber's function, Andrew's function etc. are now available. Most of these objective functions involved some tuning constants. The efficiency of the M-estimation procedures depends upon how best these tuning constants are selected. For application to designed experiments the appropriate values of these constants have been proposed. A new objective function has also been proposed. All these functions have been illustrated with some examples. In LMS method median of the square errors is minimized to obtain the parameter estimates. This method has been appropriately modified for application in designed experiments and illustrated with some examples.

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. A criterion has been developed to identify robust designs that are robust against the presence of any two outliers. It has been found that all binary proper variance balanced block designs are robust against the presence of any two outliers.

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 the present study a user-friendly software has also been developed to analyze experimental data in presence of outliers.

To disseminate the findings of the project a workshop was organized on July 26, 2007 at IASRI. The findings of the project were well received by the statisticians as well as the experimenters. Prof. Aloke Dey, ISI Delhi Centre, 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.

This research was supported by the A.P.Cess fund of Indian Council of Agricultural Research (ICAR). We express our sincerest and heartfelt thanks to ICAR for sponsoring the research. We are grateful to Dr. J.P.Mishra, Assistant Director General (ES&M) for his help and support during the course of the present investigation and granting us the necessary facilities through A.P.Cess fund.

We express our deep sense of gratitude to Dr. S.D.Sharma, Director, I.A.S.R.I., New Delhi who was instrumental and supportive in taking up this study and for his encouragement during the course of study. We are also thankful to Dr. V.K.Sharma, ex-Head, Division of Design of Experiments, IASRI, New Delhi for the support provided for conducting this study.

We sincerely thank Dr. Abhishek Rathore, Scientist, IISS, Bhopal for helping us in the development of the software. Sincere thanks also goes to Dr. P. K. Batra for providing necessary data required for the present study.

We express our appreciation and thanks to the Administrative and Accounts personnel of ICAR and IASRI for their co-operation and helpful attitude in providing the facilities for the smooth functioning of the project.

Sh Jawaid Ashraf, and Sh. Amit Bharmoria, Research Associates under this project and Smt. Jyoti Gangwani and Sh. Naresh Chand, Technical Officers, deserves special thanks for their active assistance in this study. We also thankful to Sh Basudev Kole, Amit Kumar Umrao and Meenakshi Tomar for their help during the course of this investigation.

March, 2008 AUTHORS

Contents

Chapter I

Introduction and Review of Literature

1.1 Introduction

The scientists in National Agricultural Research System (NARS) conduct a large number of experiments for their research and consequently generate a huge amount of information in the form of data collected through experimentation. This information is converted into knowledge by statistically processing the data using sophisticated statistical tools. This knowledge helps in identifying most promising agricultural technologies for making recommendations to the farmers. Hence, designing of experiments and analysis of experimental data forms an integrated component for improving the quality of agricultural research.

For statistical processing of data, several assumptions are made. However, in practice there is a tendency to forget about the assumptions and to go ahead with the statistical processing of data as if the assumptions were satisfied. The assumptions, however, get violated during experimentation and as such the statistical analysis that is carried out no longer remains valid. For instance, during the experimentation, there might be an infestation of a disease or insect attack on some plots in the field, or there may be unintentional heavy irrigation on some particular block(s) or plot(s) of the experiment, or at times there may be mistakes creeping in during recording of data, etc. It is, therefore, important to detect deviations from assumptions and suggest remedial measures.

From the time when human beings started exploiting and employing the information in the collected data as an aid to understand the world they live in, there has been a concern over the unrepresentative or outlying observations in the data set. Outlier(s) in a set of data is (are) defined to be an observation (or sub-set of observations) that appears to be inconsistent with the remainder set of data. Occurrence of outlier(s) is (are) very common in every field involving data collection and outlier(s) arises from heavy tailed distributions or is simply bad data point due to error. When outliers are present in the data, the result from the analysis of such data may lead to erroneous inference. To be clearer, consider the following example.

Example 1.1: An experiment with seven chemical treatments was conducted in a randomized complete block (RCB) design with three replications at the Regional Agricultural Research Station, Nandyal, Andhra Pradesh with a view to evaluate the effect of Mepiquaet Chloride on the yield of mustard crop [net plot size: 5.60m × 2.40m]. The table below shows the data on yield in kilogram per plot for different treatments:

Table 1.1: Yield of mustard in kg/plot

Replications	Treatments						
	1	2	3	4	5	6	7
1	0.55	0.50	0.45	0.63	0.55	0.35	0.30
2	0.45	0.64	0.54	0.62	0.50	0.30	0.35
3	0.41	0.45	0.45	0.30	0.60	0.65	0.40

Analysis of the data is presented in Table 1.2. It is observed that the treatment effects are not significant at 5% level of significance.

Table 1.2: ANOVA (With original data)

Source	DF	SS	MS	F-Value	Significance Level
Treatment (adj.)	6	0.00120952	0.01410000	0.92	0.5141
Replication	2	0.08460000	0.00060476	0.04	0.9615
Error	12	0.18405714			
Total	20	0.26986667			

Note: Here DF means degree of freedom, SS means Sum of Squares, MS means mean squares error, F-value means calculated F-value and significance level means the probability at which the null hypothesis is rejected.

It was then followed by residual analysis of this data. Standardized residuals are presented in Table 1.3.

Table 1.3: Standardized Residuals

Serial No.	Replication	Treatment	Std. Residual	Serial No.	Replication	Treatment	Std. Residual
1	1	1	0.84944	12	2	5	-0.63072
2	1	2	-0.36114	13	2	6	-1.52085
3	1	3	-0.25432	14	2	7	-0.09664
4	1	4	1.20549	15	3	1	-0.53916
5	1	5	-0.00509	16	3	2	-0.68158
6	1	6	-0.89521	17	3	3	-0.25432
7	1	7	-0.53916	18	3	4	**-2.21260**
8	2	1	-0.31027	19	3	5	0.63581
9	2	2	1.04272	20	3	6	**2.41606**
10	2	3	0.50864	21	3	7	0.63581
11	2	4	1.00712				

It is observed from the table that the observation at serial number 18 and 20 stand out because of their high value of standardized residuals. These two observations seem to be influential. We carry out the analysis again after removing these two observations. The results of this analysis are presented in the Table 1.4. The dramatic effect of removing these two observations is worth noticing. The treatment effects now become significant at 5% level of significance. Removal of any other observation or pair of observations does not affect the analysis. These two observations, therefore, definitely are influential.

Table 1.4: ANOVA (After removing two observations)

Source of variation	DF	SS	MS	F- Value	Significance Level
Treatment (adj)	6	0.16439143	0.02739857	6.40	0.0054
Replication	2	0.00159048	0.00079524	0.19	0.8334
Error	10	0.04284286			
Total	18	0.20861053			

The present example clearly shows how the presence of outliers affects the analysis of the data and inferences drawn.

The presence of an outlier is often an indication of weakness in the model, the data or both. Examples of such types are contained in Atkinson (1981, 1982) and Carroll (1982). Examination of the outlier(s) allows a more appropriate model to be formulated, or enable us to assess any dangers that may arise from basing inferences on the normality assumption. This is very much the way in which outlier(s) has (have) been discussed in the statistical literature, and seems a fruitful avenue of enquiry. We begin by defining an outlier

1.2 What is an outlier ?

Daniel (1960) defines an outlier as "an observation whose value is not in the pattern of values produced by the rest of the data".

A more comprehensive definition is due to Beckman and Cook (1983). They defined the following:

Discordant observation: Any observation that appears surprising or discrepant to the investigator.

Contaminant: Any observation that is not a realization from the target distribution.

Outlier: A collective to refer to either a contaminant or discordant observation.

Influential cases: An outlier need not be influential in the sense that the result of an analysis may remain essentially unchanged when an outlying observation is removed. It is useful to regard an influential observation as a special type of outlier.

1.3 Approaches of Studying Outliers

Following Barnett (1978) and Barnett and Lewis (1984) approaches to outliers are divided into two broad categories.

- To identify the outlier(s) for further study. This forms the detection part. When an outlier is detected, the analyst is faced with a number of questions:

  Is the measurement process out of control?
  Is the model wrong?
  Is some transformation required?
  Is there an identifiable subset of observations that is important in its different behaviour?

These issues effect the interpretation and confidence in the resulting estimates and predictions. For a general discussion one may refer to Box (1979, 1980) and Cook and Weisberg (1982).

- To accommodate the possibility of outlier(s) by suitable modifications of the models and or method of analysis. The robust methods of estimation or analysis, which were created to modify least squares procedure so that the outliers do not have much influence, fall under this category. For detailed discussion on robust methods of analysis in presence of outliers, one may refer to books by Huber (1981) and Tiku *et al.* (1986).

The literature on the study of outlier(s) is very vast and much in common with almost every area like robust estimation, data analysis, each of which is important in its own right. For an excellent review a reference may be made to Beckman and Cook (1983) and Hadi and Simonoff (1993). A critical review of literature on the study of outlier(s) in designed experiments is available in Gopalan and Dey (1976), Singh *et al*. (1987), Bhar (1997), Sarkar(2002) and Nandi (2007) and references cited therein. For useful survey of literature the books by Atkinson (1985), Barnett and Lewis (1984) and Rousseeuw and Leroy (1987) may also be referred.

1.3.1 Detection of Outliers

1.3.1.1 Detection of Outlier(s) in Normal Sample Data

Presence of outlier in the data is the most serious illness to the linear statistical model and this necessitated research work in the area of detection and handling of outliers. The existence of the problem of doubtful values or outliers has been recognized for a very long time, certainly since the middle of the eighteenth century when Bernoulli (1777) questioned the assumption of identically distributed errors and condemned the widespread practice of discarding discordant observations in the absence of prior information. From this period until the middle of the nineteenth century, the main point of discussion in the literature with regard to outlying values was whether rejection is justified. Some other research workers also took the same view as Bernoulli (1777) that observations should not be rejected purely on the ground of appearing inconsistent with the remaining data. Historically objective methods for dealing with outlier were employed only after the outliers were identified through the normal inspection of the data. The first published objective test for outlier detection is due to Peirce (1852), according to which, k doubtful observations in a sample of size n should be rejected if the probability of the system of errors obtained by retaining them is less than that of the system of errors obtained by their rejection. Chauvenet (1863) developed a test for a single doubtful observation. An observation is rejected if it lies outside the lower and upper (1/4n) points of the null distribution. With this procedure the chance of wrongly rejecting a non-influential observation is 40% in large samples. Stone (1868) introduced a rejection test based on the concept of a modulus of carelessness, m, that can be expressed in the following way; a given observer in a given sampling situation makes on an average one mistake in each set of m observations taken. An observation is to be discarded if its deviation can be attributed with more probability to the observer’s carelessness than to the random variation.

The adhoc rejection tests such as Peirce (1852), Chauvenet (1863) and Stone (1868) dominated the literature until the period of First World War. These methods, however, generated much controversy and these methods were not widely used.

After the work of Pierce, Chauvenet and Stone that continued in the 20^{th} century, many formal objective techniques for the identification of outlier(s) were developed. While some authors viewed these techniques as a way of improving the estimation of mean and others prefer to consider them as tests that help us in understanding of concomitant variables.

Irwin (1925) proposed a test statistic for known population standard deviation. Tippett (1925) proposed the use of range for the detection of outliers. Grubbs (1950) proposed a statistic *that* is based on studentized residuals. Walsh (1950) proposed the use of a non-parametric test for detecting outliers. The most popular models for the study of outlier(s) perhaps are the models proposed by Dixon (1950) that are known as Dixon's location shift and scale shift model. Dixon's model for a single outlier can be described as: a location shift model is one in which out of n observations $(n-1)$ observations come from N (μ, σ^2) except the one which comes from N $(\mu+\lambda, \sigma^2)$. Likewise a scale shift model is one in which $(n-1)$ observations come from N (μ, σ^2) and the remaining one comes from N $(\mu, \lambda\sigma^2)$, where λ is some nonzero scalar quantity.

There are many more studies on detection of outliers in normal sample data (*i.e.*, sample of observations following normal distribution). In the sequel we summarize studies carried out in general linear models.

1.3.1.2 Detection of Outliers in Linear Models

The study of detection of an outlier in linear models gained special attention soon after Ferguson (1961) and Srikantan (1961) explicitly defined mean-shift and variance inflation model on using Dixon's concept. A brief description of mean-shift model and variance inflation models is given in the sequel.

Mean shift model: The justification behind this procedure is that the basic normal theory model is valid except that the expectation of at most one unknown response is shifted. In other words , in the presence of an outlier, say the i^{th} observation to be an outlier, the mean of the i^{th} observation will be shifted from μ_i to $\mu_i + c$, where c is some non-zero quantity.

Consider the following linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{1.1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is an $n \times p$ full rank matrix of known constant, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters and $\mathbf{e}$ is an $n \times 1$ vector of randomly distributed errors such that E($\mathbf{e}$) = $\mathbf{0}$ and V($\mathbf{e}$) = $\sigma^2\mathbf{I}_n$, $\sigma^2 > 0$. Here E(.) denotes the expectation and V(.) denotes the variance. $\mathbf{0}$ is a null vector and $\mathbf{I}_n$ is an identity matrix of order n. Let $\mathbf{u}_i$ denote an n-component vector with 1 in the i^{th} position and zero elsewhere. Then mean shift model for k outlying observations for the model (1.1) can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\gamma} + \mathbf{e} \tag{1.2}$$

Here $\boldsymbol{\gamma}$ is a k component vector of unknown parameters, $\mathbf{D}$ is an $n \times k$ matrix with column $\mathbf{u}_1, \mathbf{u}_2,$ $\ldots, \mathbf{u}_k$. Nonzero values of the component of the vector $\boldsymbol{\gamma}$ imply the presence of k outlying cases.

Variance inflation model: In contrast to the mean shift model, the variance inflation model for a single outlier is based on the assumption that the variance of an unknown response is larger than the remaining observations.

The well-known Cook-statistic (Cook, 1977), Q_k- statistic (Gentleman and Wilk, 1975) and AP-statistic (Andrew and Pregibon, 1978) are based on mean shift model. For an excellent study on outlying observation in mean-shift model a reference may be made to Cook and Weisberg (1982).

Removal of the individual suspected case or group of suspected cases in turn may change the result of the analysis of the data. This is the idea behind the study of the influential cases in linear models. Using this idea Gentleman and Wilk (1975) proposed the Q_k- statistic which can be described as the reduction in the residual sum of squares after deleting the k suspected outliers. For defining Q_k- statistic, rewrite the model (1.2) as

$$\mathrm{E}\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{0}\\ \mathbf{X}_2 & \mathbf{D}\end{bmatrix}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{\gamma}\end{bmatrix} \tag{1.3}$$

where $\mathbf{y}_1$ is the vector of $(n-k)$ clean observations, $\mathbf{y}_2$ is the vector of k suspected outlying observations, $\mathbf{X}_1$ has $(n-k)$ rows of $\mathbf{X}$ corresponding to the $(n-k)$ clean observations, $\mathbf{X}_2$ has k rows of $\mathbf{X}$ corresponding to the k outlying observations. After fitting the original data $\mathbf{y}$ to the model (1.3) we get the residuals as

$$\mathbf{r} = \begin{bmatrix}\mathbf{r}_1\\ \mathbf{r}_2\end{bmatrix} = (\mathbf{I}-\mathbf{H})\mathbf{y} = \begin{bmatrix}\mathbf{I}-\mathbf{H}_{11} & -\mathbf{H}_{12}\\ -\mathbf{H}_{21} & \mathbf{I}-\mathbf{H}_{22}\end{bmatrix}\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{bmatrix},$$

where $\mathbf{H}_{ij} = \mathbf{X}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'_j$, $i, j = 1,2$.

Using the above, Gentleman and Wilk statistic is obtained as

$$Q_k = \mathbf{r}'_2(\mathbf{I}-\mathbf{H}_{22})^{-1}\mathbf{r}_2 \tag{1.4}$$

F- statistic for testing the hypothesis H_0: $\boldsymbol{\gamma} = \mathbf{0}$ is

$$F_k = \frac{n-p-k}{k}.\frac{Q_k}{\mathbf{r}'_2\mathbf{r}_2 - Q_k} \quad \text{with } k \text{ and } n-p-k \text{ degrees of freedom.}$$

Cook (1977) proposed a new measure based on confidence ellipsoid for judging the contribution of each data point to the determination of the least squares estimator of the parameter vector in full rank linear regression models. If $\hat{\boldsymbol{\beta}}_{(i)}$ is the least squares estimator of $\boldsymbol{\beta}$ with the i^{th} point deleted, then the suggested measure of the critical nature of each data point is defined to be

$$D_i = \frac{(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_{(i)})}{ps^2}, \; i = 1,2,\ldots,n \tag{1.5}$$

$$= \frac{\mathbf{r}_i^2}{p}\frac{h_{ii}}{1-h_{ii}}$$

Where s^2 is an estimate of σ^2 and h_{ii} is the i^{th} diagonal element of the matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \; . \tag{1.6}$$

H is also known as "Hat" matrix.

This is known as Cook-statistic. It provides a measure of distance between $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$ in terms of descriptive levels of significance. D_i can be compared to the percentage point of an F-distribution with p and $n-p$ degrees of freedom. This measure was developed under implicit assumption that $\boldsymbol{\beta}$ is the parameter of interest. In some situations, the interest may be in q linearly independent combinations of $\boldsymbol{\beta}$, rather than the whole $\boldsymbol{\beta}$. For such situations it would be more reasonable to measure the influence of each data point on the determination of the least squares estimates of these combinations of interest. Let $\mathbf{P}\boldsymbol{\beta}$ denote the parametric combinations of interest, where $\boldsymbol{P}$ is a $q \times p$ matrix with rank q. A generalized measure of the importance of the i^{th} point is

$$\mathrm{D_i} = \frac{(\mathbf{P}\hat{\boldsymbol{\beta}} - \mathbf{P}\hat{\boldsymbol{\beta}}_{(i)})'[\mathbf{P}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{P}']^{-1}(\mathbf{P}\hat{\boldsymbol{\beta}} - \mathbf{P}\hat{\boldsymbol{\beta}}_{(i)})}{q\hat{\sigma}^2} \tag{1.7}$$

Using model (1.2) Andrews and Pregibon (1978) also proposed a statistic for detection of outlier(s) and is known as AP-statistic

$$\mathrm{AP} = \frac{|\mathbf{Z}_0'\mathbf{Z}_0|}{|\mathbf{X}_0'\mathbf{X}_0|} \tag{1.8}$$

where $\mathbf{Z}_0 = [\mathbf{X} : \mathbf{D} : \mathbf{y}]$ and $\mathbf{X}_0 = [\mathbf{X} : \mathbf{y}]$

The statistic given in (1.5) can be rewritten as

$$\mathrm{AP}_{(k)} = (1 - \frac{\mathrm{Q}_k}{\mathrm{RSS}})\left[\mathbf{I} - \mathbf{H}_{22}\right]. \tag{1.9}$$

Here RSS is the residual sum of squares obtained from fitting the full model (1.1), Q_k is given by (1.4) and $\mathbf{H}_{22}$ is as defined earlier. According to Andrews and Pregibon (1978) the quantity (1-$\mathrm{AP}_{(k)}$) corresponds to the proportion of volume generated by $\mathbf{X}_0$ attributable to the k suspected outlying observations. If this subset of observations lie far out in the factor space, it will account for a large proportion of volume of the factor space thus giving some realistic interpretation to the term outliers. Hence small values of AP-statistic correspond to the influential subset.

The relative merits of Cook-statistic and AP-statistic were examined by Draper and John (1981). They showed that a case based on AP- statistic might not be the same as based on Cook-statistic. However, two components of the AP-statistic given in (1.9) will provide considerable information not only on outlying and influential observations but also on the remoteness of the observations in the factor space. The first factor of (1.9) will be small if Q_k is large and so identifies sets of outliers. The second term $[\mathbf{I} - \mathbf{H}_{22}]$ provides a measure of the remoteness of the set of observations in the factor space, smaller value of $[\mathbf{I} - \mathbf{H}_{22}]$ indicating more remote points.

Three most important test statistics for detection of influential observations are discussed above. There are, however, a number of statistics available in the literature. Hocking and Pendleton (1983) discussed the relative merits of some commonly used influential diagnostics. Here we give a brief account of the statistics available in the literature. Hoaglin and Welsch (1978) consider the role of hat matrix in the identification of influential cases. The i^{th} diagonal element of **H**, h_{ii} reflects the role of y_i in predicting itself. Hoaglin and Welsch (1978) used the criterion $h_{ii} > \frac{2p}{n}$ to identify high influential points.

Belsley *et al*. (1980) presented several diagnostics based on the distance between the parameter estimates with full data and the parameter estimates after deleting the i^{th} observation. This statistic is known as DFBETA. Belsley *et al*. (1980) also suggested the statistic COVRATIO, which is based on ratio of the determinant of the covariance matrices of the parameter estimates from full data and the parameter estimates after deleting the i^{th} observation. Daniel and Wood (1980) introduced the weighted squared standardized distance for the case *j*, $WSSD_j$, which measures the distance of case *j* from the center of the data weighted by the relative importance of the variable.

Polasek (1984), de Gruttola *et al*. (1987) and Martin (1992) developed some methods to study the influence of outliers in regression when errors are correlated. Rousseeuw and Leroy (1987) discussed many robust techniques of outlier identification of which they preferred the one based on least median of squares (LMS) residuals. Putterman (1988) discussed the influence of outlying observation when errors follow first order autoregression. Schall and Dunne (1988) gave a comprehensive general discussion of much of the theory of an outlier and influence. In the normal general linear model $(\mathbf{y}, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{V})$ with arbitrary known variance and covariance structure three types of outliers namely distributional outlier, outlier by additive shift and transformational outlier were distinguished and test statistics associated with each type of outlier were presented.

Bian and Tiku (1997a, 1997b) proposed a method of robust estimation, which uses the Bayesian prior. They first robustified the prior distribution and then used modified maximum likelihood estimator (MMLE). Wong *et al*.(1996) considered robust estimation procedure in time series data. Moreover, their result is applicable to a wide class of symmetric and skew symmetric distributions.

1.3.1.3 Detection of more than one outlier

If the data set contains more than one outlier or influential observation, which is likely to be the case in most of the data sets, the problem of identifying such observations becomes more difficult. This is due to masking and swamping effects. Masking occurs when an outlying subset goes undetected because of the presence of another, usually adjacent, subset. Swamping occurs when "good" observations are incorrectly identified as outliers because of the presence of another, usually remote, subset. As a result consecutive application of single outlier test leads to the problem. Realizing this fact many researchers proposed test statistics for simultaneous detection of more than one outliers.

The first proposed test was the optimal block test of Murphy (1951). Tietjen and Moore (1972,1979) proposed two Grubbs type statistics for the identification of multiple outliers. Rosner (1975) developed test statistics for identification of more than one outlier that are free from masking and swamping effects. John and Draper (1978) proposed a two-stage test for the presence of two or one outlier in a two-way table by using Q_k- statistic of Gentleman and Wilk (1975). Draper and John (1980) further extended the previous work to the study of testing for three or fewer outliers in a two-way table. An aspect of design of experiments in the general regression situation when it is feared that outlier may also occur was also briefly discussed. A new statistic $F_k = \frac{(S - Q_k^*)}{S}$ was proposed by Marasinghe (1985) for detection of several outliers in linear models, where $S = (n\text{-}p)\, s^2$ and Q_k^* is the reduction in the residual sum of squares after deleting the subset of size k of observations, where n is the total number of observations, p is the total number of parameters and s^2 is error sum of squares. This statistic was incorporated into the following multistage procedure. Initially a subset of k observations is selected for testing outliers. If F_k is found to be significant, the most extreme observation in the subset as determined by the largest studentized residual is deleted. The test procedure is repeated for the (k–1) observations in the subset using the remaining sample. The procedure is stopped when a test fails to reject the non-outlier hypothesis.

Paul and Fung (1991) proposed a two-phase procedure for detecting multiple outliers in linear regression based on generalized extreme studentized residual (GESR). In phase one two sets of suspected observations are identified, one based on GESR and the other based on either D_i or least median squares (LMS) method. In the second phase list of potential outliers are prepared by the union of these two sets. GESR procedure, which is used to detect multiple outliers in **y,** controls Type 1 error rate most adequately. Hadi (1992) proposed a procedure for the detection of multiple outliers on multivariate data. The procedure is based on the division of the data set into basic and non-basic subsets using an appropriately chosen robust measure of outlyingness. Other test procedures for detection of multiple outliers that are free from masking and swamping effects are due to Davies and Gather (1993), Hadi and Simonoff (1993), Hadi (1994) and Simonoff (1988). Atkinson (1994) gave a robust method for the detection of multiple outliers based on a series of forward search method and LMS method. Each of the forward search method involves successively using least squares to fit subset of size m, m+1, … with m ranging from p, the number of parameters, to n, the number of observations. Each field yields a value of LMS criterion and a set of residuals by taking the observations with large residuals from the fit. Recently Juan and Prieto (2001) described a procedure for the detection of multivariate outliers based on angular properties of the observations.

1.3.2 Accommodation of Outliers

As mentioned earlier, to tackle the problem of outlying observations and non-normal error variances, robust method of analysis has been advocated. Robust analysis of data is such that it is insensitive to the presence of outlying observations and departure from the assumption of normality of error variance. For a detailed descriptions of different robust regression procedures one may refer to Atkinson and Riani (2000), Tsou and Cheng (2004), Cheng (2005) and Jajo (2005). In the sequel, we discuss the most commonly used robust methods of data analysis.

1.3.2.1 Breakdown Point

The performance of many robust regression methods is evaluated on the basis of breakdown point. The finite-sample breakdown point is the smallest fraction of the anomalous data that can cause the estimator to be useless. The smallest possible breakdown point is $1/n$, that is, a single observation can distort the estimator so badly that it is of no practical use to the regression model-builder. The breakdown point of the OLS (Ordinary Least Squares) is $1/n$. The breakdown point of the class of M-estimators (to be discussed shortly) is also $1/n$. Most experienced data analysts believe that the fraction of data that are contaminated by erroneous data typically varies between 1 and 10 %. This has lead to the development of high-breakdown-point estimators. Least Median of Squares (LMS) estimator has 50% breakdown point.

1.3.2.2 Robust Analysis in Linear Regression Models

Consider the linear model (1.1) again. When the observations $\mathbf{y}$ in the linear model are normally distributed, the method of least squares is a good parameter estimation procedure in the sense that it produces an estimator of the parameter vector $\boldsymbol{\beta}$ that has good statistical properties. However there are many situations where we have evidence that the distribution of the response variable is considerably non- normal, and/or there are outliers that affect the regression model.

To deal with such type of situations robust regression comes into picture. A robust regression procedure is one that dampens the effect of observations that would be highly influential if least squares were used. That is a robust procedure tends to leave the large residuals associated with outliers, there by making the identification of influential points much easier. In addition to insensitivity to outliers, a robust estimation procedure should be able to produce essentially the same results as least squares when the underlying distribution is normal and there are no outliers.

A robust procedure tries to accommodate the majority of the data. Bad points, lying far away from the pattern formed by the good ones, will consequently possess large residuals from the robust fit. So in addition to insensitivity to outliers, a robust regression estimator makes the detection of these points an easy job. Of course residuals from least square fit can not be used for this purpose, because the outliers may posses very small residuals as the least square fit is pulled too much in the direction of these deviating points.

1.3.2.2.1 M-estimator

The motivation for much of the work in robust estimation was due to Huber (1964). His class of M-estimator has been specifically designed to be insensitive to outliers and to retain high efficiency when the errors are heavier tailed than the normal, two properties not possessed by least squares.

In M-estimation actually some function of the errors is minimized to get the parameter estimates, where as in least squares, sum of squares of errors is minimized. A number of minimizing functions are proposed in the literature. For example, Huber's t-function (Huber, 1964), Andrews's wave function, Hampel's function (Andrews *et al*., 1972 and Andrews, 1974) and Ramsay's function (Ramsay, 1977). Some recent developments on M-estimation are available in Stefansky and Boos(2002) and Cui *et al*. (2004). The details of M-estimation are given in chapter III.

1.3.2.2.2 Least Median of Squares (LMS) Estimator

Rousseeuw (1984) introduced a very robust method known as Least Median of Squares (LMS) method, which has a very high break down point of around 50 %. In this method actually median of square residuals is minimized. Giloni and Padberg (2002) compared LMS regression with other regression procedures. Leontitsis and Pange (2004) considered statistical significance of LMS regression.

1.3.2.2.3 Least Trimmed (Sum of) Squares (LTS) Estimator

The least trimmed (sum of) squares estimator is found by finding the regression model parameters that satisfy.

$$\text{Minimize} \sum_{i=1}^{h} e_{(i)}^2 ,$$

where $e_{(1)}^2 < e_{(2)}^2 < \ldots < e_{(n)}^2$ are the ordered squared residuals and h must be determined. The best robust properties are obtained when $h = n/2$.

1.3.3 Robustness Aspects

Another way to deal with outliers is to adopt a robust design. A robust design means that the design is insensitive to the presence of outliers. In the recent years robust designs have been an active area of research. A substantial amount of literature is now available on the subject of robust designs. Reviews on the subject, on specific aspects have been done by Herzberg (1982), Akhtar and Prescott (1987), Srivastava et al. (1990) and Atkinson (1982). The review by Akhtar and Prescott (1987) is limited to the case of response surface designs only. Broadly speaking, a statistical procedure is said to be robust if it is not sensitive to departures from assumptions on which it is based. Several such statistical procedures are available in Huber (1981). However, the available literature reveals that the robustness of experimental designs mainly deals with the unavailability of data (missing observations). Not much work on robustness of experimental designs in the presence of outliers is available. Robustness of experimental designs in the presence of outliers was first studied by Box and Draper (1975). However, their study was confined to the case of response surface designs.

1.4 Outlier(s) in Designed Experiments

The important features of the statistics developed for identifying outliers in linear regression models are that they are developed from the point of view of whether outliers affect the parameter estimation and/or estimation of residual sum of squares for linear models in which design matrix has full column rank. The data generated from designed experiments may be prone to the occurrence of outlier(s). Though, the general set up of an experimental design is that of a linear model, yet detection and testing of outliers in this field has some problems. Firstly, in experimental designs the design matrix does not have full column rank; thus available test statistics cannot be applied as such to this setting. Secondly, in experimental designs the experimenter is interested only in a sub set of parameters. One may, therefore, be interested to see the effect of the outliers on the estimation of this subset of parameters. Unfortunately not much work in this field is available in the literature. In the following section we shall present the work done on the study of outliers in the field of designed experiments.

1.4.1 Identification

John (1978) studied the problems that arise in detecting the presence of outliers in the results from factorial experiments. He actually applied the techniques of Gentleman and Wilk (1975) and John and Draper (1978), who investigated the problem of detecting outliers in two-way table and provided a statistic Q_k which is difference between the sum of squares of residuals from the original data and sum of squares of revised residuals resulting from fitting the basic model after deleting *k*- influential observations. Ben and Yohai (1992) studied the asymptotic theory of M-estimates and their associated test for a one-factor experiment in randomized block design. They have also studied a test for treatment effects derived from M- estimates. Bhar (1997) modified Cook-statistic, Q_k- statistic and *AP*- statistic for application to the field of design of experiments. He found out the modified statistics for identification of single as well as multiple outliers using both mean shift and variance inflation models. He showed that when all the outliers belong to a particular block the above mentioned statistics do not exist. Sarker (2002) showed that the statistic also does not work for detection of a single outlier when block size is 2.

1.4.2 Robust Analysis in Designed Experiments

Not much work on robust analysis of experimental data is available in the literature. The first work in this direction is found in Lehman (1964). He obtained robust estimate of contrasts in treatment effects for experiments with one observation per cell. This estimate is based on robustly estimated treatment means. Robust treatment means are obtained by L_2 estimator. This concept of robust estimation was extended to incomplete block designs by Greenberg (1965). He also studied the asymptotic properties of these estimators. However, these estimators are not reliable as these estimators developed on the basis of data, completely ignored the model. Most of the robust methods in the modern days are developed from the model point of view.

Carroll (1980) applied so called modern robust methods to analysis for factorial experiments. He applied M-estimation procedure using Huber's function, Andrews's function and Hampel's functions to find the significance level of the different treatment contrasts.

Robustness of ANOVA with respect to interactions in some orthogonal block designs was studied by Zhang (1992). He gave a sufficient condition under which the structure of covariance of the model without interactions remains unchanged when interactions are included.

Huggins (1993) applied robust approach to the analysis of repeated measured data.

Chi (1994) provided robust methods for analysis for Cross-Over Trials. He proposed a robust procedure, combined M-estimation for analyzing cross over data with possible within- and between subject outliers. The mean squared error properties of these combined M- estimates for direct treatment effect contrasts and carryover treatment effect contrasts are examined through simulation studies. They are found to be superior to those of the generalized least squares estimates when there are possible within and between subject outliers.

Muller (1995) studied breakdown point for designed experiments. He derived the break down point for situations that often appear in designed experiments. In particular he derived the breakdown point for replicated experimental conditions and showed that a design that maximizes the breakdown point should minimize the maximal number of experimental conditions, which lie in a subspace of the parameter space.

1.4.3 Robustness Aspects

As mention earlier, Box and Draper (1975) were the first to study the robustness of experimental designs in presence of outlier(s). Let us consider the model (1.1) where Rank (**X**) = *p*. Suppose that in the model u^{th} observation was added to it an aberration *h* making it an outlier. They have defined the overall discrepancy in the estimated value of **y,** which is caused by the effect of *h* on the u^{th} observation. Box and Draper suggested that in order to make a designed experiment insensitive to outlier(s), the variance of overall discrepancy should be made minimum. The main feature of the robustness criterion of Box and Draper is that it cannot be applied to the linear model for which the design matrix is deficient in rank.

Gopalan and Dey (1976) developed a criterion of robustness on the lines similar to those given by Box and Draper (1975) in other experimental situations where the design matrix is not of full rank. Instead of taking the overall discrepancy in the estimated value of **y** they considered the discrepancy in the estimation of σ^2. They studied the robustness aspect of experimental designs by minimizing the variance of this discrepancy. The designs identified as robust against the presence of a single outlier by this method are:

a. Randomized complete block designs.
b. BIB designs.
c. All non-group divisible two associate, connected PBIB designs with $\lambda_2 = 0$.
d. All semi-regular group divisible designs.
e. All triangular PBIB designs satisfying $r + (n-4)\lambda_1 - (n-3)\lambda_2 = 0$.
f. All L_2 type PBIB designs satisfying $r + (s-2)\lambda_1 - (s-1)\lambda_2 = 0$.

Where the notations are as usual, details of these notations are available in Dey (1986). Singh *et al*. (1987) extended the results of Gopalan and Dey (1976) to find out robust designs for two-way elimination of heterogeneity. They showed that important class of variance balanced row-column designs that satisfy the property of adjusted orthogonality are robust in the presence of a single outlier.

Bhar (1997) and Bhar and Gupta (2001) used a different criterion of robustness to find robust designs in presence of a single outlier. He suggested the use of minimum value of average Cook-statistic as the criterion and also got the same result of Gopalan and Dey (1976) and Singh *et al.* (1987) using this criterion. He also showed that all E-optimal block designs are robust against the presence of single outlier. Sarker et al. (2005) showed that both the criteria are equivalent.

Sarker (2002) and Sarker *et al*. (2003) extended these results to the experimental situations where the interest of the experimenter is only in a subset of all possible elementary treatment contrasts rather than the complete set of all the possible elementary contrasts. They obtained the Cook-distance (single response) for the set of treatment contrasts of interest. This statistic is used to identify the outlying observation.

Sarker et al. (2005) proposed a test statistic for detection of a single outlier in block designs for diallel crosses. They have established a correspondence between two existing criteria of robustness against a single outlier i.e. minimization of average Cook-statistic and minimization of variance of discrepancy or bias in estimation of error variance. It has been shown that a proper binary balanced block design for diallel crosses is robust against the presence of a single outlier. Block designs for diallel crosses in which every line appears an equal number of times in each block are also found to be robust against the presence of a single outlier.

In most of these studies dealing with the problem of outlier(s) in block designs, the mean-shift model has been considered for studying the effect of outlier on the estimation of parameter in block design for univariate response. Bhar and Gupta (2003) studied the robustness aspects of designs under variance inflation models.

1.5 Motivation

A vast literature on outliers clearly indicates its importance in statistics. Primarily this concept was developed for an univariate sample but it started attracting people working in other fields of statistics. Mean-shift model developed by Srikantan (1961) and Ferguson (1961) becomes a strong basis for studying outliers in regression models. Various diagnostic procedures are now available in the literature for identification of outliers. Since least squares tries to avoid large residuals, the presence of a single outlier may result in a generally poor fit as the equation tries to accommodate this case at the expense of the rest of the observations. Thus large absolute value of residuals should be flagged off for further study. Most of the test-statistics for testing outliers are developed using this basic idea.

Example 1.1 clearly shows that the residuals play a significant role in identifying influential observations. For identifying outliers that have a large influence on the estimation of parameters, a number of statistics based on distance measure, which reflects the effects of deleting the suspected outlying observations, are developed.

The study of outliers is now diversified to almost every field of scientific study. The problem of outliers has been studied extensively in linear regression models. As mentioned earlier 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 square of errors is minimized. Each observation gets different weights for estimating parameters where as 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, Andersons'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 sub set 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) modified Cook statistic for detecting outliers in block designs. John (1978) provided some statistics for detecting outliers in factorial experiments. Therefore, there is a need to develop test statistic for detection of more than one outlier in experimental data. The robust method of estimation of parameters need to be modified so as to to be useful in experimental data.

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) obtained robust regression designs in presence of a single outlier. Gopalan and Dey (1976) identified robust block designs through minimization of variance of discrepancy or bias in estimation of error variance. Bhar and Gupta (2001) used the minimization of average Cook-statistic to identify robust designs against presence of a single outlier. Sarker *et al*. (2005) 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:

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

1.6 Scope of the Present Investigation

The present investigation is an attempt to give an comprehensive study of the problem of outliers. Outlier(s) may be identified for further scrutiny or it may be accommodated in such a way that its influence may be nullified. For identification of outliers several test procedures are now available in the context of linear regression model. Bhar and Gupta (2001) modified some of these statistics for application into designed experiments. All these statistics are discussed in the first chapter. The other way handling outliers in linear regression model is robust method of analysis. There are some applications of robust methods in designed experiments also available. All these methods of detecting outliers as well as robust methods of data analysis have been reviewed in the first chapter.

In the second chapter, statistics available for detecting outliers in designed experiments are applied to a large number of real experimental data sets. Statistics developed by Bhar and Gupta (2001) viz. Cook-statistic, Q_k-statistic and AP-statistic are applied to designed experiments. This chapter also deals with the problem of masking, a typical problem in case of detection of multiple outliers as described earlier. A new method of identification of multiple outliers in presence of masking in designed experiments has been proposed and illustrated with some examples.

Robust methods of analysis of experimental data is the subject matter of the Chapters III and IV. Among robust methods, M-estimation and Least Median of Squares (LMS) have been chosen for application in designed experiments. In Chapter III, we deal with M-estimation. Appropriate procedures for both estimation and testing of hypotheses have been proposed in this chapter and illustrated with some examples. LMS method has actually developed from linear regression point of view. This method has been appropriately modified for application in designed experiment. All these are discussed in Chapter IV. The modified method is illustrated with the help of real life examples.

As discussed earlier, in case of designed experiments the problem of outliers can also be handled by adopting a robust design, *i.e.*, the design is insensitive to the presence of outliers. So far robustness study was confined to the presence of a single outlier. This study has been extended for more than one outlier. Some designs that are robust against the presence of two outliers have been identified. This is the subject matter of Chapter V.

For application of the techniques developed for handling outliers in designed experiments under this project, a user friendly software has been developed. Various aspects of this software have been discussed in Chapter VI.

A dissemination workshop organized under this project. Many scientists from different parts of the country participated in this workshop. Some suggestions and recommendations have emerged from this workshop. These are enlisted in Chapter VII.

The report is concluded with a summary of the report and a list of references.

1.7 Practical/Scientific Utility

Generally data generated from experiments are analyzed without taking any care of the presence of outliers. But the presence of outliers may drastically affect the conclusions drawn from the experiment. Even a single outlier may alter the conclusions drawn. With the proper application of methodologies emerging from the present investigation, the researchers would be able to draw statistically valid conclusions. Further, the catalogue of designs identified as robust against in the presence of outliers, will be useful for designing experiments.

The presence of an outlier is often an indication of weakness in the model, the data or both. Once an observation is identified as an outlier, one would be able to critically examine their experiment. Why did the outlier occur? Is there anything wrong with the measurement process?

With such questions, scientists would be able to visualize his experiment more properly and can take appropriate measures accordingly.
The newly developed software would be helpful in to analyzing data appropriately. A mass awareness about the serious consequences of the presence of outliers would be established through the Workshop.

The proper and appropriate analysis of experiments will help in improving the research output of agricultural sciences research, which in turn will enable it to be globally competitive.

Chapter II

Detection of outliers

2.1 Introduction

We begin with the problem of detection of outliers in designed experiments. As mentioned in Chapter 1 that the statistics' developed for linear regression models can not directly be used in designed experiments because of the rank deficiency of the design matrix. Moreover, in designed experiments, we are mainly interested in the estimation of some functions of subset of parameters rather than whole set of parameters. Bhar and Gupta (2001) modified some of these statistics' for application into designed experiments. We present these statistics for detecting outliers in designed experiments in the present chapter in Section 2.2. In Section 2.3 we applied these statistics to real experimental data sets taken from Agricultural Field Experimental Information System (AFEIS), IASRI, New Delhi. However, these statistics are not free from the problem of masking. In masking, the effect of an outlier is suppressed by the presence of another outlier and therefore, if one applies single outlier detection procedure; both the outliers may remain undetected. In the context of regression analysis, there are now many statistics available for tackling the problem of masking. The approach by Pena and Yohai (1995) is quite appealing. The essence of this approach is utilized for developing a new statistic for detecting outliers in presence of masking in designed experiments in the present investigation. This is the subject matter of the Section of 2.4. Every method is illustrated with some examples.

2.2 Detection of Outliers in Designed Experiments

2.2.1 Cook-statistic

In this section we begin with Cook-statistic, given by Cook (1977, 1979). This statistic is useful in determining the degree of influence the i^{th} data point has on the parameter estimation. Consider the general linear model as mentioned in (1.1)

Cook-statistic as given in (1.5) is rewritten here as a definition

Definition 2.1: If $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{(i)}$ are the least squares estimates of $\boldsymbol{\beta}$ with and without the i^{th} data point respectively, then the Cook-statistic

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)' \left[D(\hat{\boldsymbol{\beta}})\right]^{-1} \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)}{Rank\left[D(\hat{\boldsymbol{\beta}})\right]}$$

or

$$D_i = \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)' (\mathbf{X}'\mathbf{X}) \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)}{p\hat{\sigma}^2}$$

where $\hat{\sigma}^2 = (n-p)^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.

The statistic D_i provides a measure of the distance between $\hat{\boldsymbol{\beta}}_{(i)}$ and $\hat{\boldsymbol{\beta}}$ in terms of descriptive levels of significance, because D_i is actually $100(1-\alpha)$% confidence ellipsoid for the vector $\hat{\boldsymbol{\beta}}$ under normal theory, which satisfy $D_i \le F(p, n-p, 1-\alpha)$.

Suppose, for example, $D_i \approx F(p, n-p, 1-\alpha)$ then the removal of the i^{th} data point moves the least squares estimate to the edge of the 50% confidence region for $\boldsymbol{\beta}$ based on $\hat{\boldsymbol{\beta}}$. Such a situation may be a cause for concern. For any analysis one would like each $\hat{\boldsymbol{\beta}}_{(i)}$ to stay well within 10%, say, confidence region (See Cook, 1977). Cook has also shown that this statistic can be used to assess the degree of influence for a subset of parameters as well as can be extended for more than one outlier.

Though the general set up of an experimental design is that of a linear model, yet Cook statistic cannot be applied as such for testing outliers in this field because of some problems as described earlier. Therefore there is a need to develop this statistic for experimental designs.

Cook Statistic in Designed Experiments

Consider the general linear model for an experimental design *d* (say)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon}\,;\ \mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0},\ \mathrm{D}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}_n,\ \sigma^2 > 0 \qquad (2.1)$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations. $\mathbf{X}$ is $n \times p$ design matrix with rank $m(< p)$, $\boldsymbol{\theta}$ is a $p \times 1$ vector of unknown parameters and $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector of independent random variables each with zero mean and variance $\sigma^2 (> 0)$.

Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1' \quad \boldsymbol{\theta}_2')'$, where $\boldsymbol{\theta}_1$ is a *v*-component vector containing all parameters of interest to the experimenter and $\boldsymbol{\theta}_2$ is (*p*-*v*) component vector containing the set of nuisance parameters in the model which are not of much interest to the experimenter.

Thus $$\mathbf{y} = (\mathbf{X}_1 \quad \mathbf{X}_2)\begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} + \boldsymbol{\varepsilon}, \qquad (2.2)$$

where $\mathbf{X}$ is partitioned in conformity with the parameters, $\mathbf{X}_1$ is an $n \times v$ matrix of rank v and $\mathbf{X}_2$ is an $n \times$ (p–v) matrix such that $\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$. The normal equations obtained by least squares method for estimating the parameters are given by

$$\mathbf{X}'\mathbf{X}\boldsymbol{\theta} = \mathbf{X}\mathbf{y}$$

From these equations on eliminating $\boldsymbol{\theta}_2$, we obtain the reduced normal equations involving only $\boldsymbol{\theta}_1$ as $$\mathbf{C}_{\boldsymbol{\theta}_1}\boldsymbol{\theta}_1 = \mathbf{Q}_{\boldsymbol{\theta}_1}, \qquad (2.3)$$

where
$$\begin{aligned}
\mathbf{C}_{\boldsymbol{\theta}_1} &= \mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{X}_1 \\
&= \mathbf{X}_1'\mathbf{B}\mathbf{X}_1. \\
\mathbf{Q}_{\boldsymbol{\theta}_1} &= \mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2'\mathbf{y} \\
&= \mathbf{X}_1'\mathbf{B}\mathbf{y}.
\end{aligned}$$

$$\mathbf{B} = \mathbf{I}_n - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-}\mathbf{X}_2' \qquad (2.4)$$

The matrix $\mathbf{B}$ is symmetric and idempotent.

We assume that the design d considered here is connected, i.e., all (v–1) orthonomalized contrasts for the parameters $\boldsymbol{\theta}_1$ are estimable or equivalently Rank ($\mathbf{C}_{\boldsymbol{\theta}_1}$) = v–1, and let the set of all (v–1) orthonormalized contrasts for the parameters $\boldsymbol{\theta}_1$ be given by $\mathbf{P}\boldsymbol{\theta}_1$. The ($v$–1)×$v$ matrix $\mathbf{P}$ is such that $\mathbf{PP}' = \mathbf{I}_{v-1}$, $\mathbf{P}'\mathbf{P} = \mathbf{I}_v - \frac{1}{v}\mathbf{J}_v$ and the least squares estimator of $\mathbf{P}\boldsymbol{\theta}_1$ is given by $\mathbf{P}\hat{\boldsymbol{\theta}}_1$, where $\hat{\boldsymbol{\theta}}_1$ is any solution of the normal equations (2.3).

For a connected design, the dispersion matrix of $\mathbf{P}\hat{\boldsymbol{\theta}}_1$ can be written as $D(\mathbf{P}\hat{\boldsymbol{\theta}}_1)$ = $(\mathbf{PC}_{\boldsymbol{\theta}_1}\mathbf{P}')^{-1}\sigma^2$.

Let k observations be suspected of being outliers in the sense that their expected values are shifted from the expected values of other observations. We keep k known (we take first k observations as outliers) and considered the following model:

$$\mathbf{y} = \begin{pmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2\end{pmatrix} + \mathbf{U}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{2.5}$$

where $\mathbf{U} = \begin{pmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_k\end{pmatrix}$, $\mathbf{u}_i = \begin{pmatrix}0 & 0 & \ldots & 1(i^{th}) & \ldots & 0 & 0\end{pmatrix}'$ and $\boldsymbol{\delta} = \begin{pmatrix}\delta_1 & \delta_2 & \ldots & \delta_k\end{pmatrix}'$.

The coefficient matrix of reduced normal equations $\mathbf{C}$ and adjusted treatment total vector $\mathbf{Q}$ vector are

$$\mathbf{C}_{\boldsymbol{\theta}_{1Z}} = \mathbf{X}_1'\mathbf{A}\mathbf{X}_1 - (\mathbf{X}_1'\mathbf{A}\mathbf{X}_2)(\mathbf{X}_2'\mathbf{A}\mathbf{X}_2)^{-}(\mathbf{X}_2'\mathbf{A}\mathbf{X}_1)$$

$$\mathbf{Q}_{\boldsymbol{\theta}_{1Z}} = \mathbf{X}_1'\mathbf{A}\mathbf{y} - (\mathbf{X}_1'\mathbf{A}\mathbf{X}_2)(\mathbf{X}_2'\mathbf{A}\mathbf{X}_2)^{-}(\mathbf{X}_2'\mathbf{A}\mathbf{y}),$$

where $\mathbf{A} = \mathbf{I}_n - \mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$.

As a third option we consider, a model where the k observations suspected to be outliers are actually omitted

$$\mathbf{y}_{(k)} = \begin{pmatrix}\mathbf{X}_{1(k)} & \mathbf{X}_{2(k)}\end{pmatrix}\begin{pmatrix}\boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2\end{pmatrix} + \boldsymbol{\varepsilon}_{(k)} \tag{2.6}$$

where $\mathbf{y}_{(k)}$ and $\boldsymbol{\varepsilon}_{(k)}$ have (n–k) elements and $\mathbf{X}_{(k)}$ has (n–k) rows.

Similarly the reduced normal equations for $\hat{\boldsymbol{\theta}}_1$ under this model are given by

$$\mathbf{C}_{\boldsymbol{\theta}_{1(k)}}\boldsymbol{\theta}_{1(k)} = \mathbf{Q}_{\boldsymbol{\theta}_{1(k)}} \tag{2.7}$$

where $\mathbf{C}_{\boldsymbol{\theta}_{1(k)}} = \mathbf{X}_{1(k)}'\mathbf{X}_{1(k)} - (\mathbf{X}_{1(k)}\mathbf{A}\mathbf{X}_{2(k)})(\mathbf{X}_{2(k)}\mathbf{A}\mathbf{X}_{2(k)})^{-}(\mathbf{X}_{2(k)}\mathbf{A}\mathbf{X}_{1(k)})$

$$\mathbf{Q}_{\boldsymbol{\theta}_{1(k)}} = \mathbf{X}_{1(k)}'\mathbf{y}_{(k)} - (\mathbf{X}_{1(k)}\mathbf{A}\mathbf{X}_{2(k)})(\mathbf{X}_{2(k)}\mathbf{A}\mathbf{X}_{2(k)})^{-}(\mathbf{X}_{2(k)}\mathbf{A}\mathbf{y}_{(k)}).$$

Consider the models in (2.5) and (2.6) respectively, then the following two statements hold:

(i) The information matrix $\mathbf{C}_{\theta_{1Z}}$ under model (2.5) equals the information matrix $\mathbf{C}_{\theta_{1(k)}}$ under the model (2.6)
(ii) The vector $\mathbf{Q}_{\theta_{1Z}}$ for adjusted totals for the parameters θ_1 under the model (2.5) equals the corresponding vector $\mathbf{Q}_{\theta_{1(k)}}$ under the model (2.6).

The best linear unbiased estimator (BLUE) of all ortho-normalized contrasts for $\boldsymbol{\theta}_1$ under the model (2.5) equals the BLUE of all ortho-normalized contrasts for $\boldsymbol{\theta}_1$ under the model (2.6), i.e., $\mathbf{P}\hat{\boldsymbol{\theta}}_{1Z} = \mathbf{P}\hat{\boldsymbol{\theta}}_{1(k)}$.

The difference between the estimators of contrasts of $\boldsymbol{\theta}_1$ under the model (2.2) and (2.6) can be expressed as $\mathbf{P}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_{1(k)}) = \mathbf{P}\mathbf{C}^{+}_{\theta_1}\mathbf{X}_1'\mathbf{B}\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}\mathbf{y}$, (2.8)

where **B** is as given in (2.4) and $\mathbf{V} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$.
Now following the Definition 2.1 we give the Cook-statistic for the contrasts $\mathbf{P}\boldsymbol{\theta}_1$ of $\boldsymbol{\theta}_1$ in experimental designs as:

Definition 2.2: Cook-statistic for the set of contrasts $\mathbf{P}\boldsymbol{\theta}_1$ is given by

$$D_k = \frac{(\mathbf{P}\hat{\boldsymbol{\theta}}_1 - \mathbf{P}\hat{\boldsymbol{\theta}}_{1(k)})'\left[D(\mathbf{P}\hat{\boldsymbol{\theta}}_1)\right]^{-1}(\mathbf{P}\hat{\boldsymbol{\theta}}_1 - \mathbf{P}\hat{\boldsymbol{\theta}}_{1(k)})}{\text{Rank}\left[D(\mathbf{P}\hat{\boldsymbol{\theta}}_1)\right]}$$

$$= \frac{(\mathbf{P}\hat{\boldsymbol{\theta}}_1 - \mathbf{P}\hat{\boldsymbol{\theta}}_{1(k)})'(\mathbf{P}\mathbf{C}_{\theta_1}\mathbf{P}')(\mathbf{P}\hat{\boldsymbol{\theta}}_1 - \mathbf{P}\hat{\boldsymbol{\theta}}_{1(k)})}{(v-1)\hat{\sigma}^2} \qquad (2.9)$$

Now using the fact that $\mathbf{C}_{\boldsymbol{\theta}_1}\mathbf{1} = \mathbf{0}$, we get

$$D_k = \frac{\mathbf{y}'\mathbf{V}\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{B}\mathbf{X}_1\mathbf{C}^{+}_{\theta_1}\mathbf{X}_1'\mathbf{B}\mathbf{U}(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}\mathbf{y}}{(v-1)\hat{\sigma}^2}$$

The Cook statistic D_k can also be written as

$$D_k = \frac{\hat{\boldsymbol{\delta}}'\mathbf{U}'\mathbf{B}\mathbf{X}_1\mathbf{C}^{+}_{\boldsymbol{\theta}_1}\mathbf{X}_1\mathbf{B}\mathbf{U}\hat{\boldsymbol{\delta}}}{(v-1)\hat{\sigma}^2}$$

$$= \frac{\hat{\boldsymbol{\delta}}\mathbf{U}'\mathbf{S}\mathbf{U}\hat{\boldsymbol{\delta}}}{(v-1)\hat{\sigma}^2}$$

where $\hat{\boldsymbol{\delta}} = (\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}\mathbf{y}$ and $\mathbf{S} = \mathbf{B}\mathbf{X}_1\mathbf{C}^{+}_{\theta_1}\mathbf{X}_1\mathbf{B}$. (2.10)
A convenient formula for D_k can be obtained by noting that $\mathbf{U}'\mathbf{V}\mathbf{y} = \mathbf{r}_2$ and $\mathbf{V} = \mathbf{B} - \mathbf{S}$, so that

$$D_k = \frac{\mathbf{r}_2{}' (\mathbf{U}'\mathbf{BU} - \mathbf{U}'\mathbf{SU})^{-1}(\mathbf{U}'\mathbf{SU})(\mathbf{U}'\mathbf{BU} - \mathbf{U}'\mathbf{SU})^{-1}\mathbf{r}_2}{(v-1)\hat{\sigma}^2}. \qquad (2.11)$$

Insight into D_k can be obtained by applying the spectral decomposition to $\mathbf{U}'\mathbf{SU}$. Let $0 \le \lambda_1 \le \lambda_2 \le ... \le \lambda_k$ be the eigenvalues of $\mathbf{U}'\mathbf{SU}$, then $\mathbf{U}'\mathbf{SU} = \mathbf{E}'\mathbf{\Lambda E}$

where $\mathbf{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_k)$ and $\mathbf{E}$ is orthogonal matrix. Then $D_k = \dfrac{\hat{\boldsymbol{\delta}}'\mathbf{E}'\mathbf{\Lambda E}\hat{\boldsymbol{\delta}}}{(v-1)\hat{\sigma}^2}$.

Let $\mathbf{E}\hat{\boldsymbol{\delta}} = \mathbf{g}$, thus (2.10) alternatively be written as

$$D_k = \frac{\mathbf{g}'\mathbf{\Lambda g}}{(v-1)\hat{\sigma}^2} = \sum_{m=1}^{k} g_m^2 \frac{\lambda_m}{(v-1)\hat{\sigma}^2} \qquad (2.12)$$

Distribution of D_k

A generalization of squared Studentized residuals to *k* outlying cases is given by (see also Cook and Weisberg, 1980),

$$\frac{\mathbf{r}_2{}'\mathbf{U}'\mathbf{VU}\mathbf{r}_2}{\hat{\sigma}^2} = \hat{\boldsymbol{\delta}}'(\mathbf{U}'\mathbf{VU})\hat{\boldsymbol{\delta}}.$$

The residual sum of squares under the model (2.5) is $\mathbf{y}'\mathbf{Vy} - \mathbf{r}_2(\mathbf{U}'\mathbf{VU})^{-1}\mathbf{r}_2$. Thus the additional sum of squares due to fitting the parameters δ is $\mathbf{r}_2{}'\mathbf{U}'\mathbf{VU}\mathbf{r}_2$. Assuming normality, the test statistic for testing the hypothesis $\boldsymbol{\delta} = \mathbf{0}$ is

$$t_1^2 = \frac{(\mathbf{r}_2'(\mathbf{U}'\mathbf{VU})^{-1}\mathbf{r}_2)(n-m-k)}{(\mathbf{yVy} - \mathbf{r}_2'(\mathbf{U}'\mathbf{VU})^{-1}\mathbf{r}_2)t}$$
$$= \frac{(\mathbf{r}_2'(\mathbf{U}'\mathbf{VU})^{-1}\mathbf{r}_2)(n-m-k)}{((n-m)\hat{\sigma}^2 - \mathbf{r}_2'(\mathbf{U}'\mathbf{VU})^{-1}\mathbf{r}_2)k}.$$

The null distribution of this statistic is F(*k*, *n*− *m*− *k*).

Now, $\mathbf{r}_2{}'\mathbf{U}'\mathbf{VU}\mathbf{r}_2 = \hat{\boldsymbol{\delta}}'\mathbf{U}'\mathbf{BU}\hat{\boldsymbol{\delta}} - \hat{\boldsymbol{\delta}}'\mathbf{U}'\mathbf{SU}\hat{\boldsymbol{\delta}} = \hat{\boldsymbol{\delta}}'\mathbf{U}'\mathbf{BU}\hat{\boldsymbol{\delta}} - \sum_{i=1}^{k} g_i^2\lambda_i$.

Therefore, D_k can be compared with

$$F_k = \frac{(n-m-k)\hat{\boldsymbol{\delta}}'\mathbf{U}'\mathbf{BU}\hat{\boldsymbol{\delta}} - \sum_{i=1}^{k} g_i^2\lambda_i}{((n-m)\hat{\sigma}^2 - \mathbf{r}_2'(\mathbf{U}'\mathbf{VU})^{-1}\mathbf{r}_2)k} \sim \text{F}(k, n-m-k) \text{ (See Cook and Weisberg, 1980).}$$

Computation of D_k

One goal in examining subset of *k* >1 cases is to find groups of cases that, while not individually influential when taken as a group are influential. Finding influential subsets which include cases that are individually influential may add little information because the observed influence of the subset will be due, in part to the influence of the single influential case. Conversely, finding an uninfluential subset that includes one or more cases that are singly influential would not decrease the interest in those cases. Thus good candidates for inclusion in subsets will have small distance values for k = 1, but they may well have relatively large values of residuals. Thus only a subset of all possible ${}^{n}C_k$ subsets of D_k need to be tested. Therefore, a method of subset selection should be written to find subsets of cases with large values of D_k. But in every subset, the eigenvalues of $\mathbf{U'SU}$ are needed, which is a tedious job. A realistic technique for finding influential subsets should use only the residuals and the diagonal entries of **V**. Following Cook and Weisberg (1980), we give two upper bounds for D_k for computational purpose.

Form I

Since $\lambda_k \geq \lambda_l \;\forall\, i$ =1,2, …, k, (2.9) can be approximated by,

$$D_k \leq \frac{\lambda_k}{(v-1)\hat{\sigma}^2} \sum_{i=1}^{k} g_i^2 \,. \tag{2.13}$$

Again λ_k must be replaced by an approximation that can be computed without need for obtaining $\mathbf{U'SU}$. The easiest approximation to use is $\lambda_k \leq$ $\mathbf{U'SU}$. Thus from (2.11), we get

$$D_k \leq \frac{\text{trace}(\mathbf{U'SU})}{(v-1)\hat{\sigma}^2} \sum_{i=1}^{k} g_i^2 \,. \tag{2.14}$$

Form II

For a fixed k, let $T = \max_{k \in {}^{n}c_t}$ [trace($\mathbf{U'SU}$)]

And $R^2 = \max_{k \in {}^{n}c_t} \sum_{i=1}^{k} g_i^2$

Then the bounds of D_k are given by,

$$D_k \leq \frac{\text{trace}(\mathbf{U'SU})}{(v-1)\hat{\sigma}^2} R^2 \tag{2.15}$$

and $$D_k \leq \frac{T}{(v-1)\hat{\sigma}^2} \sum_{i=1}^{k} g_i^2 \,. \tag{2.16}$$

These two bounds may combine to give another bound for D_k as

$$D_k \leq \frac{TR^2}{(v-1)\hat{\sigma}^2} \,. \tag{2.17}$$

When k = 1, all the approximations are exact. An algorithm for finding all relevant subsets with fixed k can be based on these approximations. Exact computation is required if these approximation values are larger than a selected cut-off point.

2.2.2 AP-statistic

In this section, we give AP-statistic, proposed by Andrews and Pregibon (1978). This statistic is also useful in determining the degree of influence of outliers on parameter estimation. This statistic is appropriately modified by Bhar and Gupta (2001) for application into designed experiments.

Consider again the model (2.1), i.e., $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, and assume that **X** has full column rank, and similar to the model (2.4) we also consider the model, $\mathbf{y} = (\mathbf{X} \quad \mathbf{U})\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$, where **U** is as defined earlier. Then AP-statistic is defined as,

$$AP_k = \frac{|\mathbf{Z}^{*\prime}\mathbf{Z}^*|}{|\mathbf{X}^{*\prime}\mathbf{X}^*|}$$

where $\mathbf{X}^* = (\mathbf{X} \quad \mathbf{y})$ and $\mathbf{Z}^* = (\mathbf{X} \quad \mathbf{U} \quad \mathbf{y})$ and $|.|$ denote the determinant value of a square matrix.

According to Andrews and Pregibon (1978) the quantity (1-AP$_k$) corresponds to the proportion of volume generated by **X*** attributable to the *k* outlying observations. If this subset of observations lies "far out" in the factor space, it will account for a large proportion of the volume of the space, lending some realistic interpretation to the term 'outliers'. Hence small values of AP-statistic are associated with deviant or influential observations.

However, this statistic can not be applied as such to experimental designs situation because in case of experimental designs, the matrices **X*****X** and **Z*****Z*** are singular. We, therefore, define this statistic for experimental designs suitably. We reparameterize the model (2.5) in the following way:
Since the rank of **X** is *m*, there are at most *m* linearly independent eastimable parametric functions and we let $\mathbf{P}_*\boldsymbol{\theta}$ be such a set of *m* linearly independent estimable functions, where $\mathbf{P}_*$ is an $m \times p$ matrix. Hence, every estimable linear parametric function must be a linear combination of the elements of $\mathbf{P}_*\boldsymbol{\theta}$. Hence, $\mathbf{X}\boldsymbol{\theta}$ which is estimable must be expressible as

$$\mathbf{X}\boldsymbol{\theta} = \mathbf{M}\mathbf{P}_*\boldsymbol{\theta}$$

for some $n \times m$ matrix **M** of rank *m*. so the original model (2.4) is transformed into

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \end{aligned} \qquad (2.18)$$

where $\boldsymbol{\beta} = \mathbf{P}_*\boldsymbol{\theta}$.

The model (2.18) now becomes full model and from (2.18),

$$\hat{\boldsymbol{\beta}} = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{y}.$$

It can be shown that $\mathbf{P}_*\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\beta}}$.

We now define the following two matrices $\mathbf{X}_0^* = [\mathbf{M} \quad \mathbf{y}]$ and $\mathbf{Z}_0^* = [\mathbf{M} \quad \mathbf{U} \quad \mathbf{y}]$. Now AP-statistic is defined in the context of designed experiments as,

$$AP_k = \frac{\left|\mathbf{Z}_0^{*\prime}\mathbf{Z}_0^*\right|}{\left|\mathbf{X}_0^{*\prime}\mathbf{X}_0^*\right|}.$$

After simplification it can be shown that

$$AP_k = \left|\mathbf{U}'\mathbf{V}\mathbf{U}\right|\left(1-\frac{\mathbf{r}_2'(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{r}_2}{RSS}\right) \tag{2.19}$$

where RSS is the residual sum of squares.

Following, Draper and John (1981), AP-statistic is factorized in (2.19). The first factor involves only independent variables and provides a measure of remoteness of the set of observations in the factor space; smaller value of $\left|\mathbf{U}'\mathbf{V}\mathbf{U}\right|$ indicating more remote points. The second will be small if $\mathbf{r}_2'(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{r}_2$ is large and so identifies set of outliers as in John and Draper (1978).

2.2.3 'Q_k' - statistic

In this section we present another useful statistic for testing outliers in experimental designs. This statistic is known as 'Q_k' – statistic or outlier sum of squares given by Gentleman and Wilk (1975). This statistic is modified by Bhar and Gupta (2001) for application in designed experiments. The *n*-component observations vector $\mathbf{y}$ is partitioned as $\mathbf{y} = \begin{pmatrix}\mathbf{y}_1' & \mathbf{y}_2'\end{pmatrix}'$, where $\mathbf{y}_1$ is an (*n*–*k*) component vector containing observations which are possibly suspected outliers. Correspondingly, the general linear model can be rewritten as,

$$E(\mathbf{y}) = E\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix} = \begin{pmatrix}\mathbf{L}_1\\ \mathbf{L}_2\end{pmatrix}\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\theta}.$$

Accordingly the matrix $\mathbf{X}$ is partitioned into $\mathbf{L}_1$ and $\mathbf{L}_2$. Normally, $\mathbf{L}_2$ contains *k* rows of **X**-matrix. The residuals for fitting this model by least squares are,

$$\begin{aligned}\mathbf{r} &= \mathbf{V}\mathbf{y} = [\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}']\mathbf{y}\\ &= \left[\mathbf{I}-\begin{pmatrix}\mathbf{L}_1\\ \mathbf{L}_2\end{pmatrix}(\mathbf{X}'\mathbf{X})^{-}\begin{pmatrix}\mathbf{L}_1' & \mathbf{L}_2'\end{pmatrix}\right]\mathbf{y}\\ &= \begin{bmatrix}\mathbf{I}-\mathbf{R}_{11} & -\mathbf{R}_{12}\\ -\mathbf{R}_{21} & \mathbf{I}-\mathbf{R}_{22}\end{bmatrix}\begin{pmatrix}\mathbf{y}_1\\ \mathbf{y}_2\end{pmatrix} = \begin{pmatrix}\mathbf{r}_1\\ \mathbf{r}_2\end{pmatrix}\end{aligned} \tag{2.20}$$

where $\mathbf{R}_{ij} = \mathbf{L}_i(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}_j'$, $i, j = 1,2$.

Now the model given in (2.5)can alternatively be written as

$$E(\mathbf{y}) = \begin{pmatrix}\mathbf{X} & \mathbf{U}\end{pmatrix}\begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{\delta}\end{pmatrix} = \begin{bmatrix}\mathbf{L}_1 & \mathbf{0}\\ \mathbf{L}_2 & \mathbf{I}\end{bmatrix}\begin{pmatrix}\boldsymbol{\theta}\\ \boldsymbol{\delta}\end{pmatrix}.$$

A solution of $\hat{\boldsymbol{\delta}}$ is given (2.10) as $(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{U}'\mathbf{V}\mathbf{y}$ (2.21)

Note that the matrix $\mathbf{L}_2$ can be written as

$$\mathbf{L}_2 = \mathbf{U}'\mathbf{X}.$$

Thus $\mathbf{I}-\mathbf{R}_{22} = \mathbf{I}-\mathbf{L}_2(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}_2'$

$$\begin{aligned}&= \mathbf{I}-\mathbf{U}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{U}\\ &= \mathbf{U}'(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}')\mathbf{U} = \mathbf{U}'\mathbf{V}\mathbf{U}.\end{aligned}$$

Replacing, $\mathbf{y}_2$ by 'missing value' estimate $\mathbf{y}_2 - \hat{\boldsymbol{\delta}}$ in (2.20) gives new residuals $\boldsymbol{r}^*$ whose components are

$\mathbf{r}_1^* = [\mathbf{I} - \mathbf{R}_{11} - \mathbf{R}_{12}(\mathbf{I} - \mathbf{R}_{22})^{-1}\mathbf{R}_{21}]\mathbf{y}_1, \ \mathbf{r}_2^* = \mathbf{0}$.

These are the same residuals as obtained from the model $E(\mathbf{y}_1) = \mathbf{L}_1\boldsymbol{\theta}$. Thus, the additional sum of squares due to fitting $\boldsymbol{\delta}$ as compared to model (2.5), is given by

$$'Q_k' = \mathbf{r}_2'(\mathbf{I} - \mathbf{M}_{22})^{-1}\mathbf{r}_2$$
$$= \mathbf{r}_2'(\mathbf{U}'\mathbf{V}\mathbf{U})^{-1}\mathbf{r}_2 .$$

This statistic measures the effect of an outlier and can be used to form a test-statistic as described by Gentleman and Wilk (1975) and John and Draper (1978).

2.3 Outliers in Experimental Data

Detection of outliers in experimental data was carried out. For this purpose data of block designs were taken from "Agricultural Field Experiments Information System (AFEIS)", IASRI, New Delhi. The experimental data from these experiments were investigated for the presence of any kind of problems like non-normality or heterogeneity of error variance under a project entitled 'A diagnostic study of field experiments' conducted at IASRI, New Delhi by Parsad *et al* (2004). Based on the normality and homogeneity of errors, these data were grouped into the following groups:

(i) Experiments having non-normal and heterogeneous error variance
(ii) Experiments having non-normal and homogeneous error variance
(iii) Experiments having normal and heterogeneous error variance
(iv) Experiments having normal and homogeneous error variance

Experiments from the first three groups have some problems. One of the reasons of such problems may be presence of outlier(s). We, therefore, applied test – statistic in the category (i), (ii) and (iii) of these experiments for detection of outlier(s). In most of these experiments presence of outliers is detected. A summary of the results obtained is given follows:

	Total number of experiments	Total number of experiments having at least one outlier
(i) Experiments having non-normal and heterogeneous error variance	205	125
(ii) Experiments having non-normal and homogeneous error variance	209	129
(iii) Experiments having normal and homogeneous error variance	165	118

Once outlier(s) are identified, next question may arise what to do with these outliers? One way to handle outliers is to simply discard the observations. The second way is to perform a analysis of covariance by taking one as the value of the covariate for the outlying observation and zeros for the rest of the observations. We performed both types of analysis for those experiments where outliers were found. Results obtained are same. One example is given in the sequel.

Another way and perhaps the best way is to perform a robust analysis of the data. This is one of the objectives of the present study and will be discussed in the 3rd and 4th chapters.

2.3.1 Example

An experiment with twelve treatments was conducted in a randomized complete block (RCB) design with three replications at Punjab Rao Deshmukh Krishi Vidyapeeth, Akola, Maharashtra in 1989 to study the effect of *micnelf* and other micro nutrients on the yield of groundnut crop [net plot size: 1.80m × 4.20m]. The treatment details are

T_1 = Control
T_2 = One spray of micnelf + magsulf at 20 days after sowing (DAS)
T_3 = Two spray of micnelf + magsulf at 20 DAS and 40 DAS
T_4 = Three spray of micnelf + magsulf at 20 DAS, 40 DAS and 55 DAS
T_5 = One spray of micnelf + m-o-potash at 55 DAS
T_6 = Three spray of micnelf at 20 DAS, 40 DAS and 55 DAS
T_7 = Two spray of borax at 40 DAS and 55 DAS
T_8 = Three spray of urea+ dap at 20 DAS, 40 DAS and 55 DAS
T_9 = Two spray of 0.5 ml/lit nutron at 20 DAS and 40 DAS
T_{10} = Two spray of 1.0 ml/lit nutron at 20 DAS and 40 DAS
T_{11} = Three spray of ferrus sulphate
T_{12} = Water spray

The table below shows the data on yield per plot in kilogram for different treatments:

Table 2.1: Yield of groundnut in kg/plot

Replications	Treatments											
	1	2	3	4	5	6	7	8	9	10	11	12
1	0.55	0.72	0.62	0.67	0.59	0.65	0.75	0.95	0.57	0.61	0.57	0.62
2	0.54	0.62	0.53	0.57	0.58	0.49	0.61	0.51	0.53	0.58	0.52	0.56
3	0.50	0.60	0.57	0.54	0.48	0.47	0.71	0.51	0.48	0.60	0.53	0.54

Table 2.2: ANOVA (With original data)

Source	DF	SS	MS	F Value	Significance Level
Replication	2	0.09223889	0.04611944	9.53	0.0010
Treatment (adj)	11	0.09735556	0.00885051	1.83	0.1100
Error	22	0.10649444	0.00484066		
Total	35	0.29608889			

Analysis of this data is presented in Table 2.2. It is observed that the treatment effects are not significantly different at 5% level of significance. We then computed cook statistic for each of the observations. The values of Cook-statistic for each of the observations are presented in Table 2.3. It is observed from the table that the observation at serial number 8 stands out. We tested it with the F-statistic (The table value of $F_{(11, 22)}(0.90)$ is 0.472245) and found that this observation is an outlier.

We carry out the analysis again after removing this observation. The results of this analysis are presented in the Table 2.4. The dramatic effects of removing this observation are worth noticing. The treatment effects now become significantly different at 5% level of significance. Removal of any other observation does not affect the analysis.

Table 2.3: Cook-statistics

Serial No.	Replications	Treatments	Cook-statistics	Serial No.	Replications	Treatments	Cook-statistics
1	1	1	0.0405781	19	2	7	0.036726
2	1	2	0.0000581	20	2	8	0.2051798
3	1	3	0.0093913	21	2	9	0.0182302
4	1	4	0.000428	22	2	10	0.0032059
5	1	5	0.0151393	23	2	11	0.001897
6	1	6	0.0270335	24	2	12	0.0048563
7	1	7	0.001993	25	3	1	0.0016231
8	1	8	**0.7569051**	26	3	2	0.0006272
9	1	9	0.0120946	27	3	3	0.0209725
10	1	10	0.0517893	28	3	4	0.002619
11	1	11	0.0263221	29	3	5	0.0135742
12	1	12	0.0093913	30	3	6	0.0107003
13	2	1	0.02597	31	3	7	0.05583
14	2	2	0.0003035	32	3	8	0.1739184
15	2	3	0.0022954	33	3	9	0.0006272
16	2	4	0.0009295	34	3	10	0.0292245
17	2	5	0.0573843	35	3	11	0.0140864
18	2	6	0.0037181	36	3	12	0.000741

Table 2.4: ANOVA (After removing observation No.8)

Source	DF	SS	MS	F Value	Significance Level
Replication	2	0.04919076	0.02459538	19.95	<.0001
Treatment	11	0.08356098	0.00759645	6.16	0.0002
Error	21	0.02588826	0.00123277		
Total	34	0.15864000			

Note: Analysis of covariance gives similar results

2.4 Detection of Outliers in Presence of Masking

In previous sections we have discussed several statistics for detection of a single outlier or an isolated influential point in designed experiments. Some statistic for detection of influential subsets of observations are also given. The detection of influential subsets or multiple outliers is more difficult, owing to masking and swamping problems. Masking occurs when one outlier is not detected because of the presence of others, swamping when a non-outlier is wrongly identified owing to the effect of some hidden outliers. Several procedures have been proposed for dealing with multiple outliers in linear regression models. Marasinghe (1985) and Kianifard and Swallow (1990) have suggested a sequential testing strategy to identify a set of *k* points, where the maximum number of outliers in the sample, *k*, is fixed in advance. Atkinson (1986), Rousseeuw and Leroy (1987) and Rousseeuw and van Zomeren (1990) have suggested the use of robust estimates with high breakdown point for the regression parameters to overcome the masking problem. These estimates are computed by using a resampling scheme. Hawkins (1980) have proposed a diagnostic procedure which is also based on a resampling scheme. Gray and Ling (1984) proposed the use of cluster analysis. Hocking (1984) has suggested that the eigenstructure of the matrix $(\mathbf{X}:\mathbf{y})'(\mathbf{X}:\mathbf{y})$ should be computed, where $\mathbf{y}$ is the vector of responses and the matrix $\mathbf{X}$ contains the explanatory variables. Pena and Yohai (1995) proposed a method to identify influential subsets by looking at the eigenvalues of an 'influence matrix'. This matrix is defined as the uncentred covariance of a set of vectors which represent the effect on the fit of the deletion of each data point. This matrix is normalized to have the univariate Cook (1979) statistics on the diagonal.

In the present section this method has been modified for application into designed experiments. In section 2.4.1 modified statistic is presented. In section 2.4.2 procedure for identifying the influential sets has been discussed. The computational aspects of this method are given in Section 2.4.4 and in Section 2.4.4 we illustrate the method with an example.

2.4.1 Development of the Statistic

Consider the general linear model for experimental designs as considered in (2.1), then the effect of deleting i^{th} data point on the set of treatment contrasts can be obtained from (2.8) as

$$\mathbf{f}_i = \mathbf{P}(\hat{\boldsymbol{\theta}}_1 - \hat{\boldsymbol{\theta}}_{1(i)}) = \mathbf{P}\mathbf{C}_{\theta_1}^{+}\mathbf{X}_1'\mathbf{B}u(u'\mathbf{V}u)^{-1}u'\mathbf{V}\mathbf{y}\,, \tag{2.22}$$

where $\mathbf{u} = \begin{pmatrix}0 & 0 & \ldots & 1i^{th} & \ldots & 0 & 0\end{pmatrix}'$.

The vector $\mathbf{f}_i$ summarizes the effect on the set of treatment contrasts of deleting the observation *i*. The individual deletion statistics identify influential points as those with large values of $\mathbf{f}_i$ in some suitable norm. For instance, Cook statistic for a set of treatment contrasts is given by $\frac{1}{(v-1)\hat{\sigma}^2}\mathbf{f}'(\mathbf{P}\mathbf{C}_{\boldsymbol{\theta}_1}\mathbf{P}')\mathbf{f}$. However, when masking is present, the $\mathbf{f}_i$ values corresponding to outliers tend to be small, and therefore they are not detected.

One of the most important types of masking situations occurs when several observations have similar effects on the least squares fit. Two observations *i* and *j* have similar effects on the set of treatment contrasts when $\mathbf{f}_i \approx \lambda\mathbf{f}_j$ for some scalar $\lambda > 0$. They will have opposite effects when $\lambda < 0$. However, this is not the only case for every masking effect. There are many situations where this condition does not hold good. That is, all types of masking does not imply proportional effects. Moreover, there are different types of proportional subsets that do not produce masking. However, this situation is particularly interesting because the standard procedures based on individual deletion will not work in this case. To detect possible sets of influential observations having similar or opposite effects on the fit, it seems sensible to look at the uncentred covariance matrix of $\mathbf{f}_i$. Let us call $\mathbf{F}$ the $(v-1)\times n$ matrix $\mathbf{F} = (\mathbf{f}_1 \quad \dots \quad \mathbf{f}_n)$ whose columns are the vectors $\mathbf{f}_i$. Then we define the n×n influence matrix **M** as

$$\mathbf{M} = \frac{1}{(v-1)\hat{\sigma}^2}\mathbf{F}'(\mathbf{P}\mathbf{C}_{\boldsymbol{\theta}_1}\mathbf{P}')\mathbf{F} \tag{2.23}$$

After doing some algebra it can easily be shown that the ij^{th} element of **M** is

$$m_{ij} = \frac{r_i r_j h_{ij}}{(1-h_{ii})(1-h_{jj})(v-1)\hat{\sigma}^2}, \tag{2.24}$$

where r_t is the t^{th} residual and h_{ij} is the ij^{th} element of the matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$.

Let **I** be an index set corresponding to a subset of k data points. Cook and Weisberg (1982) proposed to measure the joint influence of the data points with index in **I** by

$$D_I = (\mathbf{r}'_I(\mathbf{I}-\mathbf{H}_I)^{-1}\mathbf{H}_I(\mathbf{I}-\mathbf{H}_I)^{-1}\mathbf{r}_I / (v-1)\hat{\sigma}^2 \tag{2.25}$$

where the components of r_I are the least squares residuals and $\mathbf{H}_I$ the submatrix of **H** corresponding to the set **I**. A large value of D_I may be due to a single influential observation included in the set **I**. This can also be due to the sum of small individual effects of a set of observations that are masking each other. In the first case this single observation will be easily identified. A subset of individually highly influential points can cancel out others and this will lead to a small value of D_I. Therefore, we shall concentrate here on the most interesting case in which masking is due to points that can only be identified by looking at them jointly.

Let e_{ij} be the uncentred correlation coefficient between $\mathbf{f}_i$ and $\mathbf{f}_j$. This actually measures the effects on the least square fit of the i^{th} and j^{th} points. Then

$$e_{ij} = \frac{m_{ij}}{m_{ii}^{1/2} m_{jj}^{1/2}}. \tag{2.26}$$

Following Pena and Yohai(1995), suppose that there are k groups of influential observations $\mathbf{I}_1, \quad \dots, \quad \mathbf{I}_k$, such that

(a) if $i, j \in \mathbf{I}_h$, then $|e_{ij}| = 1$ (this means that the effects on the least squares fit produced by the deletion of two points in the same set $\mathbf{I}_h$ have correlation 1 or -1),

(b) if $i \in \mathbf{I}_j$ and $l \in \mathbf{I}_h$ with $j \neq h$, then $e_{ij} = 0$ ((this means that the effects produced on the least squares fit by observations i and j belonging to different sets are uncorrelated) and ,

(c) if i does not belong to any $\mathbf{I}_h$, $m_{ij} = 0$ for all j (this means that data points outside these groups have no influence on the fit)

Now, according to (a) we can split each set $\mathbf{I}_h$ in $\mathbf{I}_h^1$ and $\mathbf{I}_h^2$ such that

(i) if $i, j \in \mathbf{I}_h^q$, then $|e_{ij}| = 1$ and

(ii) if $i \in \mathbf{I}_h^1$ and $j \in \mathbf{I}_h^2$, then $|e_{ij}| = 0$

Let $\mathbf{v}_1 = (v_{11}, \ \ldots, \ v_{1n})'$, ..., $\mathbf{v}_k = (v_{k1}, \ \ldots, \ v_{kn})'$ be defined by $v_{hj} = m_{jj}^{1/2}$ if $j \in \mathbf{I}_h^1$, $v_{hj} = -m_{jj}^{1/2}$ if $j \in \mathbf{I}_h^2$ and $v_{hj} = 0$, if $j \notin \mathbf{I}_h^2$. Then, if (a) to (c) hold, by equation (2.23) the matrix **M** is

$$\mathbf{M} = \sum_{i=1}^{k} \mathbf{v}_i \mathbf{v}_i' \tag{2.27}$$

and since the $\mathbf{v}_i$'s are orthogonal the eigenvectors of **M** are $\mathbf{v}_1, \ \ldots, \ \mathbf{v}_k$, and the corresponding eigenvalues $\lambda_1, \ \ldots, \ \lambda_k$ are given by

$$\lambda_h = \sum_{i \in \mathbf{I}_h} m_{ii} \, . \tag{2.28}$$

When the matrix **M** satisfies (a) to (c), the only sets **I** with large D_I are $\mathbf{I}_h^q$, $1 \le h \le k$, q = 1, 2, and these sets may be found by looking at the eigenvectors associated with non-null eigenvalues of **M.** Equation (2.26) can also be written as

$$e_{ij} = \frac{\operatorname{sign}(r_i)\operatorname{sign}(r_j) h_{ij}}{(h_{ii} h_{jj})^{1/2}}, \tag{2.29}$$

which means that, in the extreme case that we have presented, the **H** matrix and the signs of the residuals can, by themselves, identify the set of points that are associated with masking. For real data sets, conditions (a) to (c) do not hold exactly. However, the masking effect is typically due to the presence in the sample of blocks of influential observations having similar or opposite effects. These blocks are likely to produce a matrix **M** with a structure close to that described by (a) to (c). Infact two influential observations *i* and *j* producing similar effects should have e_{ij} close to 1, and close to –1 when they have opposite effects. Influential observations with non-correlated effects have $|e_{ij}|$ close to 0. The same will happen with non-influential observations. Therefore, the eigenvectors will have approximately the structure described above. This suggests the following procedure to identify influential sets.

(a) Find the eigenvectors corresponding to the *p* non-null eigenvalues of the influence matrix **M**

(b) Consider the eigenvectors corresponding to large eigenvalues, and define the set $\mathbf{I}_h^1$ and $\mathbf{I}_h^2$ by those components with large positive and negative values respectively.

2.4.2 Procedure for detecting influential sets

It is obvious from the previous discussion that to identify influential sets, we need to look at the eigenvectors corresponding to the largest non-zero eigenvalues of the influence matrix. However, different influential subsets may have different eigenvectors. In nutshell, it is a useful to develop a strategy to look at all the eigenvectors corresponding to non-zero eigenvalues to find influential sets. In each eigenvector we must search for sets of co-ordinates with relatively large value and the same sign. When the set of influential points has many components, and the eigenvectors are standardized to norm 1, the individual value cannot be very large. Therefore, we must compare the relative value of the components to identify the elements of the set. Pena and Yohai (1995) suggested in case of regression analysis to look at the ratio between the components in decreasing order, searching for a clear cut-off point, to form a set of candidate outliers, and then to test the points in this set to identify the outliers. We also follow the line of this procedure in case of designed experiments.

2.4.3 Computational Aspects

Step 1: Identifying Sets of outlier Candidates.

A set of candidate outliers is obtained by analyzing the eigenvectors corresponding to the non-null eigenvalues of the influence matrix **M**, and by searching in each eigenvector for a set of co-ordinates with relatively large value and the same sign. The search is done in the following way.

(a) Order the co-ordinates of the eigenvector $\mathbf{v}_i$, obtaining $v_{i(1)} \le v_{i(2)} \le ... \le v_{i(n)}$, and let us call $i_{(1)}, \quad ..., \quad i_{(n)}$ the indices of the ordered co-ordinates of the eigenvector.

(b) Compute the ratios $a_j = \dfrac{v_{i(j)}}{v_{i(j-1)}}$ for $j = n, . . ., \quad n - c_1$ and $b_j = \dfrac{v_{i(j)}}{v_{i(j+1)}}$ for $j = 1, . . . , c_2$. The constants c_1 and c_2 smaller than $n/2$ and will be discussed shortly. Look for the first J_0 such that $\left|a_j\right| > k$ and the first $\mathbf{I}_0$ such that $\left|b_j\right| > k$

(c) If $i_0 > 1$ and/or $j_0 > 1$, consider the sets $J_0 = \{i_{(n)}, \quad i_{(n-1)}, \quad ..., \quad i_{(n-i_0+1)}\}$ and/or $I_0 = \{i_{(1)}, \quad i_{(2)}, \quad ..., \quad i_{(j_0-1)}\}$ as candidate outlier.

Choice of c_1 and c_2 is related to the desired breakdown point of the procedure will be smaller than $\min(c_1 / n, c_2 / n)$. Pena and Yohai (1995) suggested that c_1 and c_2 should be close to 1. This number seems to be sufficiently small to avoid numerical instability due to denominators in the ratios close to 0. In any case the ratios should be computed so that $\left|v_{i(j-1)}\right| > d$ where d is a small value. Since the candidate outlier will be further scrutinized, as outlined in step 2, taking large value for k may have more serious consequences. In case of regression analysis Pena and Yohai (1995) suggested this value to be 2.1. In case of designed experiments we took this value in the neighborhood of 2.0.

Step 2: Checking for outliers

(a) Remove all candidate outliers.
(b) Use the standard F statistic to test for groups or individual outliers. Reject sets of individual points with F statistic larger than some constant *c*.

(c) If the number of candidate outliers is larger than $n/2$, the previous procedure can be applied separately to points identified in each eigenvector.

According to Pena and Yohai (1995) in most regression applications the sample size n is much larger than p, the rank of **X**. Since we are only interested in the eigenvectors corresponding to the non-null eigenvalues, the direct computation of the eigenvalues and eigenvectors of **M** can be obtained by using spectral decomposition of the matrix **M**. However, for designed experiments this may not be a serious problem. We directly calculate the eigenvalues and eigenvectors.

2.4.4 Illustration

The test statistic developed in the previous sub-section was applied to experimental data from Agricultural Field Experiments Information System, New Delhi. It is observed that in some experiments some observations are not influential individually, but jointly with some other observations, they are influential, that is, some observations were masked by some other outlying observations and, therefore, could not be detected when diagnostic test procedure for detecting for single outlier is applied. To make the exposition clear consider the following example.

An experiment with 10 treatments was conducted in the randomized complete block (RCB) design with 4 replications at Sugarcane Research Institute, Shahjahanapur, Uttar Pradesh in 1988 to find out the suitable herbicide to control weeds in Sugarcane (net plot size: 8.00m × 5.40m.). The treatment details are

T_0 = Control weeded check
T_1 = Local conventional method
T_2 = Trash mulching
T_3 = 1.0 kg active ingredient (a.i.)/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 a.i./ha of Atrazine as Pre-emergence spray
T_5 = 1.00 kg a.i./ha of 2,4-D Sodium Salt at 8-10 weeks after planting
T_6 = 2.0 kg a.i./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 a.i./ha of Atrazine as Pre-emergence spray followed by spray of Glyphosate at 1.0 kg a.i./ha at 6-8 weeks after planting.
T_8 = 1.00 kg a.i./ha of Arochlor and 1.00 kg a.i./ha of Atrazine as pre-emergence spray
T_9 = 2.00 kg a.i./ha of Arochlor as pre-emergence spray

The table below shows the data on yield per plot in quintal for different treatments:

Table 2.5: Yield of sugarcane in q/plot

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

Analysis of this data is presented in Table 2.6. The treatment effects were not significantly different at 5% level of significance. Cook statistic for each observation is computed and values are given in Table 2.7. It is observed from the Table 2.7 that the observation number 14 stands out. We tested it with F value (The table value of $F_{(9,\ 27)}(0.95)$ is 0.472245) and found this observation is statistically influential. No other observation is found to be influential.

Table 2.6: ANOVA (With original data)

Source	**DF**	**SS**	**MS**	**F Value**	**Significance Level**
Replication	3	1.73105000	0.57701667	8.64	0.0003
Treatment (adj.)	9	0.63781000	0.07086778	1.06	0.4206
Error	27	1.80225000	0.06675000		
Total	**39**	4.17111000			

Table 2.7: Cook-Statistics

Serial No.	**Replica tions**	**Treatments**	**Cook-statistics**	**Serial No.**	**Replica tions**	**Treat ments**	**Cook-statistics**
1	1	1	0.0003126	21	3	1	0.0044407
2	1	2	0.0446265	22	3	2	0.0060796
3	1	3	0.0051954	23	3	3	0.0448183
4	1	4	0.0062219	24	3	4	0.022109
5	1	5	0.0065846	25	3	5	0.1292313
6	1	6	0.0124363	26	3	6	0.0147602
7	1	7	0.0134679	27	3	7	0.0120352
8	1	8	0.0005345	28	3	8	0.0233389
9	1	9	0.0491404	29	3	9	0.0002813
10	1	10	0.0000906	30	3	10	0.0000347
11	2	1	0.0003455	31	4	1	0.0009225
12	2	2	0.003801	32	4	2	0.0517879
13	2	3	0.043674	33	4	3	0.0048107
14	2	4	**0.3823402**	34	4	4	0.1526988
15	2	5	0.1122303	35	4	5	0.0111566
16	2	6	0.0063657	36	4	6	0.0080566
17	2	7	0.0145407	37	4	7	0.0110611
18	2	8	0.0818239	38	4	8	0.0121348
19	2	9	0.034714	39	4	9	0.1530533
20	2	10	0.0000742	40	4	10	0.0001498

We then applied our new statistic to identify the group of observations that are influential. It was found that observation number 14 and 39 are likely to be influential jointly. We then applied the present method for multiple outlier detection and found that these two observations are really influential (Value of Cook-statistic is 0.4521055). The interesting point to note here is that though the observation number 14 was detected as outlier yet observation number 39 was not. Its effect was masked by the observation number 14. It is an interesting example of masking.

The data was reanalyzed after deleting these two observations. The result is presented in table 2.8. The dramatic effect to note here is that the treatment effects are now significant at 5% level of significance. Removal of any other pair of observations does not have any effect on the analysis.

Table 2.8: ANOVA (With 2 data points deleted)

Source	DF	SS	MS	F Value	Significance Level
Replication	3	1.20704269	0.40234756	11.58	<.0001
Treatment (adj)	9	0.70698849	0.07855428	2.26	0.0519
Error	25	0.86835040	0.03473402		
Total	37	2.78238158			

2.5 Discussion

Conclusions drawn from an experiment may be misleading due to presence of some outliers as we have demonstrated through an example. Therefore, detection of outliers in the experimental data is of major concern. Once an observation is detected to be an outlier, further scrutiny of the observation is done. Firstly, it is ensured that there is no transcription error. Generally the outlying observations are deleted and usual analysis is carried out with the remaining observations. Alternatively analysis of covariance may also be performed. However, discarding an observation is not recommended since each observation carries some information and it should be exploited. As mentioned earlier, robust method of analysis is an alternative. This method has been discussed in the 3rd and 4th chapters.

The problem of multiple outliers has been studied intensively in linear regression model. Many statistics are now developed to tackle the problem of masking. In the present study we have used the concept of Ben and Yohai (1995). However, there are some statistics that are used to detect multiple outliers in presence of masking. Their applicability in designed experiments need to be explored. These statistics need to be explored.

Chapter III

Robust Analysis of Designed Experiments I:M-estimation

3.1 Introduction

In the earlier Chapter II, we have discussed about the procedures of detection of outlier(s). If there are outlier(s) present in the data, then detection of outlying observations or analysis of covariance is suggested. Each observation, however, contains some information and it should not be deleted. Alternatively, way of handling of the problem of outlying observations is robust methods of estimation of parameters or robust analysis of data. When the observations in the linear model are normally distributed and free from outlying observations, the method of least squares is a good parameter estimation procedure in the sense that it produces an estimator of the parameter vector that has good statistical properties. However there are many situations where we have evidence that the distribution of the response variable is considerable non- normal, and/or there are outliers that affect the assumptions of linear model. To deal with such type of situations robust regression comes into picture. A robust regression procedure is one that dampens the effect of observations that would be highly influential if least squares were used. That is a robust procedure tends to leave the large residuals associated with outliers, there by making the identification of influential points much easier. In addition to insensitivity to outliers, a robust estimation procedure should produce essentially the same results as least squares when the underlying distribution is normal and there are no outliers.

A robust procedure tries to accommodate the majority of the data. Bad points, lying far away from the pattern formed by the good ones, will consequently possess large residuals from the robust fit. So in addition to insensitivity to outliers, a robust regression estimator makes the detection of these points an easy job. Of course residuals from Least Squares (LS) cannot be used for this purpose, because the outliers may possess very small residuals as the LS fit is pulled too much in the direction of these deviating points. Among robust procedures, M-estimation method is most widely used.

In the present chapter the concept of M-estimation is introduced and then applied to designed experiments. In M-estimation an objective function (a function of errors) is minimized to obtain the parameter estimates. There are many objective functions of M-estimation for linear regression model available in the literature. Some of these objective functions are discussed in the present chapter and their applicability to designed experiments has been explored. In Section 3.2 we have introduced how M-estimator is used in linear regression model and listed some of the commonly used objective functions. Most of these objective functions involved some tuning constants. The efficiency of the M-estimation procedures depends upon how best these tuning constants are selected. For application to designed experiments the appropriate values of these constants have been proposed. Since parameters are estimated through robust methods, usual way of testing hypotheses may not be applicable. Appropriate testing procedures have been developed in the literature. In Section 3.2.1 we have described some of these procedures. In Section 3.3 applications of M-estimation procedures to designed experiments has been considered. The existing objective functions have been modified by suitably choosing the constants. A new objective function has been proposed. In designed experiments we are mainly interested in estimation of treatment contrasts. In view of this objective, Bhar and Gupta (2001) developed Cook-statistic for application to detection of outliers in experimental data. The proposed function is based upon Cook-statistic and, therefore, addressed the basic requirement of design of experiments. In the Section 3.3.3 we have applied the M-estimation procedures along with the newly proposed objective function to some real experimental data. A discussion on merits of the M-estimation procedure is given in Section 3.4.

3.2 M-estimator in Linear Regression (Montgomery and Peck, 2001)

Consider the linear model (1.1) again. In general, we may define a class of robust estimators that minimize a function ρ of the errors, *i.e*,

$$\underset{\beta}{\text{Minimize}} \sum_{i=1}^{n} \rho(e_i) = \underset{\beta}{\text{Minimize}} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i'\boldsymbol{\beta}), \tag{3.1}$$

where $\mathbf{x}_i'$ denotes the i^{th} row of $\mathbf{X}$.

An estimator of this type is called an M-estimator, where M stands for maximum likelihood. That is, the function ρ is related to the likelihood function for an appropriate choice of the error distribution. If the method of least squares is used (implying the error distribution is normal), then $\rho(\text{z}) = \frac{1}{2}\text{z}^2$, where z is the error term.

Generally instead of $\rho(e_i)$, the function $\rho(e_i/\sigma)$ is minimized, where σ is a scale parameter. To minimize the equation (3.1), equate the first p partial derivatives of ρ with respect to β_j (j = 0, 1, 2,…, p) to zero, yielding a necessary condition for a minimum. This gives the system of p equations

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right) = 0; \quad (j = 1, 2, \ldots, p), \tag{3.2}$$

where s is the robust estimate of scale parameter $\sigma, \psi = \rho'$ and x_{ij} is the i^{th} observation on the j^{th} regressor and $x_{i0} = 1$, s is an approximately unbiased estimator of σ.

In general the ψ function is non linear and the equation (3.2) must be solved by iterative methods. Iterative Reweighted Least Squares (IRLS) is most widely used procedure. This approach is usually attributed to Beaton and Tukey (1974).

To use Iteratively Reweighted Least Squares, suppose that an initial estimate $\hat{\boldsymbol{\beta}}_0$ is available and that s is an estimate of scale. Then write p equations in (3.2),

$$\sum_{i=1}^{n} x_{ij}\psi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{s}\right) = \sum_{i=1}^{n} \frac{x_{ij}\{\psi[(y_i - \mathbf{x}_i'\boldsymbol{\beta})/s]/\}(y_i - \mathbf{x}_i'\boldsymbol{\beta})/s}{(y_i - \mathbf{x}_i'\boldsymbol{\beta})/s} = 0 \tag{3.3}$$

as $$\sum x_{ij} w_{i0} (y_i - \mathbf{x}_i' \boldsymbol{\beta}) / s = 0 ,$$ (3.4)

where $$w_{i0} = \frac{\psi\left[\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_0\right)/ s\right]}{\left(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_0\right)/ s} \quad \text{if } y_i \neq \mathbf{x}_i'\hat{\boldsymbol{\beta}}_0$$ (3.5)

$$= 1 \quad \text{if } y_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_0$$

In matrix notations (3.4) can be written as

$$\mathbf{X'W_0X\boldsymbol{\beta}} = \mathbf{X'W_0y} ,$$ (3.6)

where $\mathbf{W}_0$ is an $n \times n$ diagonal matrix of "weights" with diagonal elements $w_{10}, w_{20},..., w_{n0}$ given by equation (3.5). We recognize equation (3.6) as the usual weighted lest-squares normal equations. Consequently the one step estimator is

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{X'W_0X})^{-1}\mathbf{X'W_0y} .$$ (3.7)

At the next step we recomputed the weights from equation (3.5) but using $\hat{\boldsymbol{\beta}}_1$ instead of $\hat{\boldsymbol{\beta}}_0$. This process is continued till the convergence criterion is met. Some robust criterion functions available in the literature are given below.

Table 3.1: Commonly used objective functions

Criterion	$\rho(z)$	$\psi(z)$	$w(z)$	**Range**
Least squares	$(1/2)z^2$	z	1.0	$\lvert z\rvert < \infty$
Huber's t function	$(1/2)z^2$	z	1.0	$\lvert z\rvert \le t$
	$\lvert z\rvert t - (1/2)t^2$	t sign(z)	$t/\lvert z\rvert$	$\lvert z\rvert > t$
Ramsay's function	$a^{-2}\begin{bmatrix} 1 - \exp(-a\lvert z\rvert) \\ (1 + a\lvert z\rvert) \end{bmatrix}$	$z\exp(-a\lvert z\rvert)$	$\exp(-a\lvert z\rvert)$	$\lvert z\rvert < \infty$
Andrews' wave function	a[1-cos(z/a)]	Sin (z/a)	Sin (z/a)/(z/a)	$\lvert z\rvert \le a\pi$
	2a	0	0	$\lvert z\rvert > a\pi$

Hampel's function	(½)z^2 a|z| -(½)a^2 (a(c|z| -(½)z^2))/(c-b)-7a^2/6 a(b+c-a)	z a sign (z) (a sign (z)(c-|z|))/(c-b) 0	1.0 a/|z| a(c-|z|)/(|z|(c-b)) 0	$\|z\| \le a$ $a < \|z\| \le b$ $b < \|z\| \le c$ $\|z\| > c$

Source: Montgomery and Peck (2001)
Note: Here $\rho(z)$ indicates the function of error i.e., z, $\psi(z)$ is the first derivative function of $\rho(z)$, $w(z)$ is the weight function and range defines the range of residuals.

Robust regression procedures can be classified on the basis of the behavior of their ψ-function. The ψ-function controls the weight given to each residual and apart from a constant of proportionality is sometimes called the influence function. For example the ψ-function for least squares is unbounded and thus least squares tends to be non robust when used with data arising from a heavy tailed distribution. The Huber's t-function (Huber, 1964) has a monotone ψ-function and does not weight large residuals as heavily as least squares. These influence functions actually redescent as the residuals becomes larger. Ramsay's function is soft redescender that is the ψ function is asymptotic to zero for large $|z|$. Andrews wave function and Hampel's function (Andrews *et al*., 1972 and Andrews, 1974) are hard redescenders, that is, the ψ-function equals zero for sufficiently large $|z|$. It is to be noted that the ρ functions associated with the redescending ψ functions are not convex, and this in theory can cause convergence problems in the iterative estimation procedure. However this is not a common occurrence.

In case of M-estimation a number of estimators of σ^2 are proposed. The commonly used estimate of the error mean square is taken as (Huber, 1973)

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} w_i \{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})/s\}^2}{n-\delta(\mathbf{X})}, \text{ where } \delta(\mathbf{X}) \text{ is the rank of } \mathbf{X} \text{ matrix.} \qquad (3.8)$$

3.2.1 Test of Linear Hypothesis

In linear model the sensitivity of the classical least squares estimates to departures from normality, such as possible presence of outliers, has led to various proposals for robust methods of estimation. Parameter estimation is usually only a first step in the analysis of data arising from a linear model. A classical least squares analysis often focused upon the analysis of variance, which tests simultaneous hypothesis in large subsets of the parameters. Since the terms in a classical analysis of variance are quadratic forms in least squares estimates, one would expect that the sensitivity of the estimates to departures from normality should be inherited by the tests. In fact, for moderate to heavy tailed distributions or in the presence of outliers, it appears that there is a loss of power in classical F test. In view of this fact, many attempts have been made to develop appropriate procedures for testing linear hypotheses. Some of these procedures are discussed in the sequel.

3.2.1.1 Test Proposed by Schrader and Hettmansperger (1979)

Consider the linear model (1.1) again. In many experimental designs, particularly analysis of variance and covariance, the main interest is in the parameters that can be summarized by the linear hypothesis as

$$H_0 : \mathbf{H}\boldsymbol{\beta} = \mathbf{h}, \tag{3.9}$$

where **H** is $(p-q) \times p$ and **h** is $(p-q) \times 1$ $(0<q<p)$. In classical theory, the maximum likelihood criterion for estimation naturally leads to likelihood ratio tests for hypothesis such as (3.9). When the likelihood function corresponds to the distribution of the errors, maximum likelihood estimates are well known to be asymptotically normal. Huber (1973) extended this asymptotic theory for M-estimate when the likelihood function and the error distribution do not match.

Define D(F) as the minimum value of $\sum \rho(r_i / \sigma)$ and D(R) as the minimum value of $\sum \rho(r_i / \sigma)$ subject to (4.9), dispersion of the residuals under the full and reduced models, respectively, where r_i is the i^{th} residual. Schrader and Hettmansperger (1979) proposed to base a test of (4.9) upon the pseudo likelihood ratio

$$\Lambda = \exp\{-D(R)\} / \exp\{-D(F)\}. \tag{3.10}$$

This is a likelihood ratio type test since the density of e_i is of the form $\exp\{-\rho(r_i / \sigma)\}$.

When Λ corresponds to the actual distribution of the e_i's, then $-2\log\Lambda$ converges in distribution to a χ^2_{p-q}, central chi-squared random variable with $p-q$ degrees of freedom. Similarly to the least squares result, Huber showed that the M-estimate $\hat{\boldsymbol{\beta}}$ is asymptotically multivariate normal with mean $\boldsymbol{\beta}$ and covariance matrix proportional to $(\mathbf{X}'\mathbf{X})^{-1}$ as in case of least square fit. The only effect of ρ on this asymptotic distribution is in the constant of proportionality, which is $\sigma^2 E\{\psi^2(e/\sigma)\} / E^2\{\psi'(e/\sigma)\}$, where $\psi(x) = d\rho(x)/dx$ and $\psi'(x) = d^2\rho(x)/dx^2$. Since the choice of ρ function has little impact on the form of the asymptotic distribution of $\hat{\boldsymbol{\beta}}$, it seems reasonable to hope that the test based upon (3.10) is also well behaved. Infact $-2\left[E\{\psi'(e/\sigma)\} / E^2\{\psi^2(e/\sigma)\}\right]\log\Lambda \to \chi^2_{p-q}$. (3.11)

The test statistic is very similar in form to the familiar F-statistic for testing H_0. Notice that $-\log\Lambda = D(\hat{\boldsymbol{\beta}}_R) - D(\hat{\boldsymbol{\beta}}_F)$ is the reduction in dispersion of the residuals obtained by passing from the reduced to the full model; hence a direct generalization of the classical F test for H_0 may be based upon the statistic

$$F_M = \lambda^{-1}\left\{D\left(\hat{\boldsymbol{\beta}}_R\right) - D\left(\hat{\boldsymbol{\beta}}_F\right)\right\}/(p-q), \tag{3.12}$$

where $\lambda = \frac{1}{2}E\left\{\psi^2(e/\sigma)\right\}/E\{\psi'(e/\sigma)\}$, $\hat{\lambda} = \frac{1}{2}(n-p)^{-1}\sum\psi^2(r_i/\hat{\sigma})/\left\{n^{-1}\sum\psi'(e/\hat{\sigma})\right\}$

and $\hat{\sigma}$ is a residual scale estimate. For small sample sizes it is better to follow the classical method and compare F_M with a critical value from a central F distribution with $p-q$ and $N-p$ degrees of freedom.

3.2.1.2 Test Based upon Quadratic Form [Huber, 1973]

Huber's asymptotic theory suggests a test of H_0 based upon the quadratic form

$$Q_M = (p-q)\left(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}\right)'\left\{\mathbf{H}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{H}'\right\}^{-1}\left(\mathbf{H}\hat{\boldsymbol{\beta}} - \mathbf{h}\right)/\hat{D}, \tag{3.13}$$

where D is the asymptotic covariance scale factor and estimated as

$$\hat{D} = \hat{\sigma}^2(n-p)^{-1}\left\{\sum\psi^2(r_i/\hat{\sigma})\right\}/\left\{n^{-1}\sum\psi'(r_i/\hat{\sigma})\right\}^2.$$

Furthermore, Huber's correction for bias in $\hat{D}$ should be employed; that is $\hat{D}$ should be multiplied by an estimate of $\left\{1 + (p/n)\operatorname{var}(\psi')/E^2(\psi')\right\}^2$. For Huber's function, this term is $\left\{1 + (p/n)(1-\mu)/\mu\right\}^2$ where $\hat{\mu} = n^{-1}\sum\psi'(r_i/\hat{\sigma})$. It has been shown that Q_M and F_M are asymptotically equivalent.

3.2.1.3 Test Based on Bickel's Pseudo Observations

Bickel (1976) proposed a robust analysis that is based on standard least squares analysis of pseudo observations. If $\hat{\boldsymbol{\beta}}_M$ is the M-estimate which can be obtained from (3.7)

$$\tilde{y}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}_M + \hat{\sigma}\psi(r_i/\hat{\sigma})/\left\{n^{-1}\sum\psi'(r_i/\hat{\sigma})\right\}^{-1}, \tag{3.14}$$

$\tilde{y}_i$ is the i^{th} pseudo value, where $\boldsymbol{x}_i'$ is the i^{th} row of **X**.

Let $\tilde{\mathbf{y}}$ be the vector $\{\tilde{y}_i\}$. Then the least squares estimates ($\hat{\boldsymbol{\beta}}_L$, say) are obtained for the model $\tilde{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}_L + \boldsymbol{\varepsilon}$. Bickel (1976) showed that $\hat{\boldsymbol{\beta}}_L$ and $\hat{\boldsymbol{\beta}}_M$ are same. He suggested that defining the pseudo values and using them in the conventional Least Squares method could obtain asymptotically correct tests. The least squares estimate of $\boldsymbol{\beta}$ based upon $\tilde{\mathbf{y}}$ is $\hat{\boldsymbol{\beta}}_L = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{y}} = \hat{\boldsymbol{\beta}}_M$. It is easy to verify that the classical F-test based upon $\tilde{\mathbf{y}}$ is Q_M.

3.2.1.4 Test Based Upon Weighted Least Squares

The most natural tests based on first deriving $\hat{\boldsymbol{\beta}}_M$ from the Iteratively Reweighted Least Squares algorithm (Holland and Welsch, 1977) and then by using the final configuration of the weights as fixed and given a priori, a least squares weighted analysis of variance could be done. This may be a reasonable procedure with small sample sizes. Asymptotic theory, however, does not support it.

3.3 M-estimator in Design of Experiments

In this section M-estimation procedures are modified for application in the design of experiments. As mentioned earlier, the efficiency of the M-estimation procedures is based on the value of the constants used in different objective functions. We have suitably chosen the value of the constants for their application in the design of experiments. First we have introduced the concept of block designs in brief and then apply the M-estimation in the analysis of data generated through block designs. For a detailed description of block designs one may refer to the book by Dey (1986), Nigam, Puri and Gupta (1988) and Parsad *et al*. (2000)

Block Designs

It is an arrangement of v treatments in b blocks such that j^{th} block contains k_j experimental units and the i^{th} treatment appears r_i times in the entire design, $i = 1,2,\ldots,v$ and $j = 1,2,\ldots,b$. Underlying any block design there exists a matrix **N** of order $v \times b$, whose $(i, j)^{th}$ element is n_{ij} where $n_{ij}\ (\geq 0)$ is the number of times the i^{th} treatment appears in the j^{th} block. The matrix **N** is called the incidence matrix of the design. The model for general block design is given by

$$\mathbf{y} = \mathbf{\Delta}'\boldsymbol{\tau} + \mu\mathbf{1} + \mathbf{D}'\boldsymbol{\theta} + \boldsymbol{\varepsilon} \qquad (3.15)$$

where **y** is a $n \times 1$ vector of observations, $\mathbf{\Delta}'$ is $n \times v$ incidence matrix of treatments, $\boldsymbol{\tau}$ is a $v \times 1$ vector of treatment effects, $\mathbf{D}'$ is a $n \times b$ incidence matrix of blocks, $\boldsymbol{\theta}$ is a $b \times 1$ vector of block effects, $\mathbf{1}$ is a unit vector of order $n \times 1$ and $\boldsymbol{\varepsilon}$ is a $n \times 1$ vector of errors.

Following notations generally used in block design. $\mathbf{k} = (k_1, k_2, \ldots, k_b)'$, the column vector of block sizes, $\mathbf{r} = (r_1, r_2, \ldots, r_v)'$, the column vector of replications; $\mathbf{K} = \text{diag}(k_1, k_2, \ldots, k_b)$; $\mathbf{R} = \text{diag}(r_1, r_2, \ldots, r_v)$; $\mathbf{T} = \mathbf{\Delta y}$ and $\mathbf{B} = \mathbf{Dy}$.

The reduced normal equations for estimating the linear function of treatment effects are

$$\mathbf{C}\boldsymbol{\tau} = \mathbf{Q},$$

where $\mathbf{C} = \mathbf{R} - \mathbf{NK}^{-1}\mathbf{N}$ (3.16)

and $\mathbf{Q} = \mathbf{T} - \mathbf{NK}^{-1}\mathbf{B}$. (3.17)

We also define the matrix

$$\mathbf{M} = \mathbf{\Phi}\mathbf{\Delta}'\mathbf{C}^{-}\mathbf{\Delta}\mathbf{\Phi} \qquad (3.18)$$

where $\mathbf{\Phi} = \mathbf{I} - \mathbf{D}'\mathbf{K}^{-1}\mathbf{D}$ and $\mathbf{A}^{-}$ is a g-inverse of **A**.

Now we compare this model with the linear model as given in (2.1), i.e., $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + e$. Here the rank of **X** is $p = v+b+1$. We now partition **X** matrix as

$$\mathbf{X} = [\mathbf{X}_1 \quad \mathbf{X}_2]$$

where $\mathbf{X}_1 = \mathbf{\Delta}'$ and $\mathbf{X}_2 = [\mathbf{1} \quad \mathbf{D}']$ (3.19)

Similarly, $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\tau} \\ \mu \\ \boldsymbol{\theta} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}$,

where $\boldsymbol{\beta}_1 = [\boldsymbol{\tau}]$ and $\boldsymbol{\beta}_2 = \begin{bmatrix} \mu \\ \boldsymbol{\theta} \end{bmatrix}$. (3.20)

The problem of using least squares method of analysis for a data set containing outlier(s) is that all the observations including the outlying observations get same weight and the weight is unity. But if any observation is found to be outlier then it must get some lesser weight than the clean observations. This concept is utilized in the analysis of the design of experiments. For giving appropriate weight to different observations we make use of functions available for M-estimation frequently used in the regression analysis. Now following the normal equations for estimating $\boldsymbol{\beta}$ as given in (3.6), the normal equations for estimating the parameters in designed experiments are given as

$$\mathbf{X}'\mathbf{W}\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{W}\mathbf{y}$$

$$\begin{bmatrix} \mathbf{X}_1'\mathbf{W}\mathbf{X}_1 & \mathbf{X}_1'\mathbf{W}\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{W}\mathbf{X}_1 & \mathbf{X}_2'\mathbf{W}\mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{W}\mathbf{y} \\ \mathbf{X}_2'\mathbf{W}\mathbf{y} \end{bmatrix}, \tag{3.21}$$

where $\mathbf{X}_1$ $\mathbf{X}_2$, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are as given in equations (3.20).

From (3.21), we get

$$\mathbf{X}_1'\mathbf{W}\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{W}\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_1'\mathbf{W}\mathbf{y} \tag{3.22}$$

$$\mathbf{X}_2'\mathbf{W}\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_2'\mathbf{W}\mathbf{X}_2\hat{\boldsymbol{\beta}}_2 = \mathbf{X}_2'\mathbf{W}\mathbf{y}\,. \tag{3.23}$$

From (3.23), we can get

$$\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{W}\mathbf{X}_2)^- \left[\mathbf{X}_2'\mathbf{W}\mathbf{y} - \mathbf{X}_2'\mathbf{W}\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right].$$

Substituting this $\hat{\boldsymbol{\beta}}_2$ in (3.22) we get

$$\mathbf{X}_1'\mathbf{W}\mathbf{X}_1\hat{\boldsymbol{\beta}}_1 + \mathbf{X}_1'\mathbf{W}\mathbf{X}_2(\mathbf{X}_2'\mathbf{W}\mathbf{X}_2)^- \left[\mathbf{X}_2'\mathbf{W}\mathbf{y} - \mathbf{X}_2'\mathbf{W}\mathbf{X}_1\hat{\boldsymbol{\beta}}_1\right] = \mathbf{X}_1'\mathbf{W}\mathbf{y}\,,$$

or

$$\left(\mathbf{X}_1'\mathbf{W}\mathbf{X}_1 - \mathbf{X}_1'\mathbf{W}\mathbf{X}_2(\mathbf{X}_2'\mathbf{W}\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{W}\mathbf{X}_1\right)\hat{\boldsymbol{\beta}}_1$$
$$= \mathbf{X}_1'\mathbf{W}\mathbf{y} - \mathbf{X}_1'\mathbf{W}\mathbf{X}_2(\mathbf{X}_2'\mathbf{W}\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{W}\mathbf{y}$$

The above equations are written as

$$\mathbf{C}\hat{\boldsymbol{\beta}}_1 = \mathbf{Q}\,, \tag{3.24}$$

where

$$\mathbf{C} = \left[\mathbf{X}_1'\mathbf{W}\mathbf{X}_1 - \mathbf{X}_1'\mathbf{W}\mathbf{X}_2(\mathbf{X}_2'\mathbf{W}\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{W}\mathbf{X}_1\right] \tag{3.25}$$

and $$\mathbf{Q} = \mathbf{X}_1'\mathbf{W}\mathbf{y} - \mathbf{X}_1'\mathbf{W}\mathbf{X}_2(\mathbf{X}_2'\mathbf{W}\mathbf{X}_2)^-\mathbf{X}_2'\mathbf{W}\mathbf{y}\,. \tag{3.26}$$

Treatment Contrasts

In designed experiments, the experimenter is generally interested in estimation of treatment contrasts. Let $\mathbf{P}$ be a $(v-1) \times v$ matrix of all $(v-1)$ set of orthonormalized treatment contrasts. M-estimates of this set of contrasts are

$$\mathbf{P}\hat{\boldsymbol{\beta}}_1 = \mathbf{P}\hat{\boldsymbol{\tau}} = \mathbf{P}\mathbf{C}^-\mathbf{Q}\,. \tag{3.27}$$

The variance of $\mathbf{P}\hat{\boldsymbol{\tau}}$ is given as

$$\mathrm{Var}(\mathbf{P}\hat{\boldsymbol{\tau}}) = \hat{\sigma}^2\,\mathbf{P}\mathbf{C}^-\mathbf{P}', \tag{3.28}$$

where $\hat{\sigma}^2$ is obtained from (3.8) and $\mathbf{A}'$ is a transpose of $\mathbf{A}$.

The contrasts sum of squares are given by

$$(\mathbf{P}\,\hat{\boldsymbol{\tau}})'\left(\mathbf{P}\,\mathbf{C}^{-}\,\mathbf{P}'\right)^{-}(\mathbf{P}\,\hat{\boldsymbol{\tau}})$$

The test statistic for testing the significance of $\mathbf{P}\hat{\boldsymbol{\tau}}$, i.e., $\mathrm{H}_0 : \mathbf{P}\hat{\boldsymbol{\tau}} = \mathbf{0}$ against $\mathrm{H}_0 : \mathbf{P}\hat{\boldsymbol{\tau}} \neq \mathbf{0}$ is

$$F = \frac{(\mathbf{P}\,\hat{\boldsymbol{\tau}})'\left(\mathbf{P}\,\mathbf{C}^{-}\,\mathbf{P}'\right)^{-}(\mathbf{P}\,\hat{\boldsymbol{\tau}})}{\hat{\sigma}^2} \qquad \to F_{v-1,\,error\,df}\,.$$

For testing a particular elementary treatment contrast say $\mathbf{p}_i'\,\hat{\boldsymbol{\tau}}$ where $\mathbf{p}_i'$ is the i^{th} row of contrast matrix the test statistic is

$$F = \frac{(\mathbf{p}_i'\,\hat{\boldsymbol{\tau}})'\left(\mathbf{p}_i'\,\mathbf{C}^{-}\,\mathbf{p}_i\right)^{-}(\mathbf{p}_i'\,\hat{\boldsymbol{\tau}})}{\hat{\sigma}^2} \qquad \to F_{1,\,error\,df}\,.$$

3.3.1 Development of a New Objective Function

Considerable work has been done on robust regression in the last 20 years. Despite all the advances, robust regression is not widely used in practice (see, Chatterjee and Machler, 1997) According to Chatterjee and Machler (1997) several reasons have been advanced for this state of affairs, some of them quite superficial, other substantive. Among the superficial reasons are (i) robust methods have not been presented to the statistical community vigorously. (ii) no commercial software is available for the general user, though this picture is changing. Although there is some truth to (ii), but (i) does not hold well. Robust regression methods have concentrated primarily though not exclusively, in guarding against outliers. Guarding against outliers is very important.

In standard regression analysis the linear model is fitted by the principle of least squares which is called the Gaussian paradigm. The paradigm may be stated as for a body of data, the best fit to a given model is obtained by minimizing the sum of squares of residuals. The problem associated with the application of least squares to model fitting has been discussed widely and extensively by Chatterjee and Machler (1997). The problems from the least squares fit arise due to two main causes. These may be summarized by saying that the Gaussian paradigm gives too much weight to outliers and high leverage points. This feature distorts the fitted line.

The difficulties associated with the least squares fit can be removed by modification of Gaussian paradigm. Chatterjee and Machler (1997) proposed to modify the paradigm as follows: for a given model the best fit is obtained by minimizing the sum of squares of the residuals without allowing observations which do not fit the given model or are far from the main body of the data from exercising substantial influence over the fit. The modified paradigm will generate the fitting principle for the procedure they proposed. The fitting principle they stated that the best fit for a data set to a given model is obtained by a fitting procedure that gives less weight to observations lying far from the main body of the data and also to those, which are not fitted (well) by the model.

Motivated by the work of Chatterjee and Machler (1997), we propose an objective function to be minimized for application in the designed experiments.

It is well known that outlying observations may distort the overall conclusion to be drawn. Therefore, outlying observations should be detected before analyzing any data set. In case of designed experiments, we are interested in estimating certain treatment contrasts. In view of this fact Bhar and Gupta (2001) modified Cook-statistic for application to designed experiments. Our new objective function is based on this Cook-statistic. Since it is based on Cook-statistic, the overall robust analysis will give a clear picture on the behaviour of treatment effects. We propose the function as follows:

$$\rho(z) = \left(\frac{z^2}{2}\right)\frac{h_{ii}}{(1-h_{ii})(v-1)} \qquad \text{if } [cook(i)] > p$$

$$= \left(\frac{z^2}{2}\right) \qquad \text{otherwise.}$$

The derivative of this function, i.e., ψ is given as

$$\psi(z) = z\frac{h_{ii}}{(1-h_{ii})(v-1)} \qquad \text{if } [cook(i)] > p$$

$$= z \qquad \text{otherwise.}$$

Finally weights are obtained as

$$w(z) = \frac{h_{ii}}{(1-h_{ii})(v-1)} \qquad \text{if } [cook(i)] > p$$

$$=1 \qquad \text{otherwise.}$$

where h_{ii} is the i^{th} diagonal element of **H** as given in section 2.2.1. The value of p depends on the corresponding experiment, this actually the F-value for which an observation becomes outlier. For example, in an experiment with 10 treatments and 3 replications, corresponding F-value for 9 and 18 degrees of freedom is 0.425. If the value of Cook-statistic for any observation exceeds this value, then the observation becomes outlier. Therefore value of p for this experiment is 0.425. For other experiments, this value can be calculated before applying this method.

3.3.2 Robust Testing of Hypotheses in Designed Experiments

In design of experiments, we generally test whether all treatments are equally effective, *i.e.*, we compare among the treatment effects. As mentioned earlier, many robust testing procedures are developed in regression analysis. These procedures are applied to designed experiments for testing hypotheses of interest.

3.3.2.1 Test as Proposed by Schrader and Hettmansperger (1979)

This statistic thoroughly described in sub-section 3.3.1. For testing the equality of treatment effects we use the statistic F_M as given in (3.12). In this situation substitute **H** matrix in (3.9) by **P** and **h** by **0**. If treatment effects are found to be significantly different, then the individual treatment contrasts are tested using the same statistic F_M . In this case, however, **H** is substituted by a vector **p** of treatment contrast of interest and **h** by **0**. Since this test is asymptotically equivalent to test given by Bickel (1976), therefore, we apply only Bickel's test.

3.3.2.2 Test Based on Bickel's Pseudo Observations

The testing procedure based upon Bickel's pseudo observations is as given in (3.14). Analysis of variance is performed by substituting usual observation y_i by pseudo observations $\tilde{y}_i$.

3.3.3 Robust Analysis of Experimental Data

In this M-estimation procedure have been illustrated through real experimental data retrieved from AFEIS. For performing such analyses, relevant SAS/IML codes are written and these codes are given in APPENDIX 3.1.

Examples having Clearcut Suspected Outlier

Firstly we present two examples where outlier could be suspected on the basis of the observations themselves. We study the effects of these outlying observations.

Example 3.1: An experiment with 10 treatments was conducted in the randomized complete block (RCB) design with 3 replications at Sugarcane Research Institute, Shahjahanpur in 2003 with a view to study the comparative performance of 10 manural treatments on cane yield and juice quality of sugarcane (net plot size 8.00 m × 2.70 m). The treatment details are given as follows:

T1	Full recommended dose of N through organic manure
T2	1/3 Recommended dose of N through bio-compost + 2/3 recommended dose of N through inorganic fertilizer
T3	1/2 Recommended dose of N through bio-compost + 1/2 recommended dose of N through inorganic fertilizer
T4	2/3 Recommended dose of N through bio-compost + 1/3 recommended dose of N through inorganic fertilizer
T5	1/3 Recommended dose of N through F.Y.M. + 2/3 recommended dose of N through inorganic fertilizer
T6	1/2 Recommended dose of N through F.Y.M. + 1/2 recommended dose of N through inorganic fertilizer
T7	1/3 Recommended dose of N through F.Y.M. + 2/3 recommended dose of N through inorganic fertilizer
T8	1/3 Recommended dose of N through sulphitated pressmud cakes + 2/3 recommended dose of N through inorganic fertilizer
T9	1/2 Recommended dose of N through sulphitated pressmud cakes + 1/2 recommended dose of N through inorganic fertilizer
T10	2/3 Recommended dose of N through sulphitated pressmud cakes + 1/3 recommended dose of N through inorganic fertilizer

Doses of recommended fertilizers : 120 kg/ha of F.Y.M. + 180 kg/ha of N + 90 kg/ha of P_2O_5
The data on grain yield per plot in quintals for different treatments is given in Table 3.2:

Table 3.2: Yield of sugarcane in kg/plot

Treatments	Replication		
	1	**2**	**3**
1	169.00	177.00	160.00

2	174.00	181.00	185.00
3	192.00	188.00	181.00
4	179.00	182.00	166.00
5	176.00	159.00	166.00
6	186.00	177.00	184.00
7	169.00	157.00	165.00
8	164.00	153.00	160.00
9	170.00	101.00	176.00
10	159.00	154.00	152.00

Analysis of variance was performed on the original data and results are given in Table 3.3. From the table one can observe that both the treatment effects and block effects are not significant at 5% level of significance.

Table 3.3: Analysis of variance with the original data

Source of variation	**DF**	**SS**	**MS**	**F**	**Significance level**
Treatment	9	4185.200	465.022	2.30	0.06
Block	2	602.866	301.433	1.49	0.25
Error	18	3637.800	202.100		
Total	29	8425.866			

Average variance for the set of elementary treatment contrasts is obtained as 134.82

It is suspected that observation pertaining to treatment 9 in replication 2 is low and may be an outlier. Therefore, we perform robust analysis using various objective functions. Different values of the constants used in the objective functions are
For Huber's function the value of the constant t =1.5
For Andrew's function a =1.339
For Ramsay's function a =0.3
For Hampel's function a =1.7, b =3.4, c = 8.5
Results are presented in the following tables.

Table 3.4: Analysis of variance (Huber's function: Weighted analysis)

Source of variation	**DF**	**SS**	**MS**	**F**	**Significance level**
Treatment(adj)	9	4339.029	482.114	10.12	0.00002
Block	2	156.258	78.129	1.64	0.22167
Error	18	857.515	47.639		
Total	29	5352.803			

Average variance for the set of elementary treatment contrasts is obtained as 22.759333

Table 3.5: Analysis of variance (Andrew's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	9	6604.162	733.795	11.21	0.00001
Block	2	187.213	93.606	1.43	0.26523
Error	18	1178.263	65.459		
Total	29	7969.637			

Average variance for the set of elementary treatment contrasts is obtained as 31.759333

Table 3.6: Analysis of variance (Ramsay's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	9	3798.729	422.081	9.87	0.00002
Block	2	121.449	60.724	1.42	0.26753
Error	18	769.752	42.764		
Total	29	4689.931			

Average variance for the set of elementary treatment contrasts is obtained as 23.509333

Table 3.7: Analysis of variance (Hampel's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	9	2291.184	254.576	7.77	0.00012
Block	2	83.875	41.937	1.28	0.30216
Error	18	589.751	32.763		
Total	29	2964.812			

Average variance for the set of elementary treatment contrasts is obtained as 24.84

Table 3.8: Analysis of variance (Newly developed function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	9	2398.574	266.508	8.89	0.00005
Block	2	82.740	41.370	1.38	0.27695
Error	18	539.611	29.978		
Total	29	3020.926			

Average variance for the set of elementary treatment contrasts is obtained as 19.985333

Note: Value of p = 0.425

From the above tables, one can observe that all the M-estimation procedures give similar results. The treatment effects have become highly significant now.

Analysis Based on Bickel's (1976) Proposal

The pseudo observations obtained from equation (3.14) for different M-estimation procedures and then ordinary analysis of variance was conducted as per procedure of Bickel (1976). The results are summarized in the sequel.

Table 3.9: Analysis of variance (Huber's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj.)	9	5049.601	561.067	11.03	0.00001
Block	2	181.088	90.544	1.78	0.19707
Error	18	915.612	50.867		
Total	29	6146.301			

Average variance for the set of elementary treatment contrasts is obtained as 20.23170

Table 3.10: Analysis of variance (Andrew's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	9	5604.467	622.719	10.76	0.00001
Block	2	179.408	89.704	1.55	0.23928
Error	18	1041.723	57.873		
Total	29	6825.598			

Average variance for the set of elementary treatment contrasts is obtained as 31.582317

Table 3.11: Analysis of variance (Ramsay's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	9	4379.857	486.651	9.19	0.00004
Block	2	170.513	85.257	1.61	0.22737
Error	18	953.179	52.954		
Total	29	5503.549			

Average variance for the set of elementary treatment contrasts is obtained as 21.90412

Table 3.12: Analysis of variance (Hampel's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	9	2814.824	312.758	8.01	0.00010
Block	2	108.548	54.273	1.39	0.27456
Error	18	702.828	39.045		
Total	29	3624.180			

Average variance for the set of elementary treatment contrasts is obtained as 21.030655

Table 3.13: Analysis of variance (New function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	9	2801.685	311.298	9.47	0.00003
Block	2	101.246	50.622	1.54	0.24133
Error	18	591.697	32.872		
Total	29	3494.628			

Average variance for the set of elementary treatment contrasts is obtained as 19.914703

Note: Value of p = 0.425

Here also it is observed from the above tables that all the M-estimation procedures give similar results. The treatment effects are significantly different even at 1% level of significance in robust analysis of data. We then applied Cook-statistic for identifying outlying observations, if any. It was found that the observation number 19 corresponding to treatment number 9 in the 2nd replication is an outlier. This observation was deleted and analysis of variance is obtained again. Result is presented in Table 3.14. Interestingly, analysis is found to be similar to that as obtained through robust analysis.

Table 3.14: Analysis of variance (After deleting observation No. 19)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	9	2895.196	321.688	8.11	0.0001
Block	2	110.246	55.123	1.39	0.2761
Error	17	674.420	39.671		
Total	28	3679.862			

Average variance for the set of elementary treatment contrasts is obtained as 27.182516

Example 3.2: An experiment with 4 treatments was conducted in the randomized complete block (RCB) design with 7 replications at jonal Agricultural Research Station, J.N.K.V.V., Jabalpur to study the effect of raised beds on the yield of soyabean (net plot size: 15.00m × 6.00m). The treatment details are

4 Widths of Raised Beds
T1=6m
T2=9m
T3=12m
T4=15m

The data on the grain yield per plot in quintal for different treatments are given below

Table 3.15: Yield of soyabean in quintal/plot

Replication	Treatments			
	1	2	3	4
1	14.3	13.59	11.92	12.18
2	14.08	13.22	12.56	12.06
3	14.14	11.53	12.53	12.31
4	12.97	12.55	11.32	17.76
5	14.05	13.22	12.6	11.64
6	13.39	12.3	12.75	11.09
7	13.19	14.58	12.61	11.76

Analysis of variance was performed on the original data and results are given in Table 3.15. From the table one can observe that both the treatment effects and block effects are not significant at 5% level of significance.

Table 3.16: Analysis of variance with the original data

Source of variation	DF	SS	MS	F	Significance level
Treatment (adj.)	3	7.489	2.496	1.23	0.326
Block	6	3.721	0.620	0.31	0.925
Error	18	36.418	2.023		
Total	27	47.629			

Average variance for the set of elementary treatment contrasts is obtained as 0.578073

Robust Analysis through M – estimation

We now applied different M-estimation procedures to this data and obtained the analysis of variance. We have taken the different values of the constants used in the objective functions as follows.

For Huber's function the value of the constant t=1.23

For Andrew's function a=1.5

For Ramsay's function a=0.23

For Hampel's function a=1.5, b=3.0, c=7.5

Weighted Analysis

Firstly weighted analysis of variance has been done. The results are given in the sequel.

Table 3.17: Analysis of variance (Huber's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	17.495	5.831	3.342	0.042
Block	6	6.417	1.069	0.612	0.717
Error	18	31.425	1.745		
Total	27	55.338			

Average variance for the set of elementary treatment contrasts is obtained as 0.08934

Table 3.18: Analysis of variance (Andrew's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	18.017	6.005	3.576	0.034
Block	6	5.162	0.860	0.512	0.791
Error	18	30.213	1.678		
Total	27	53.393			

Average variance for the set of elementary treatment contrasts is obtained as 0.0923

Table 3.19: Analysis of variance (Ramsay's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	20.029	6.676	3.658	0.032
Block	6	4.866	0.811	0.444	0.831
Error	18	32.853	1.825		
Total	27	57.748			

Average variance for the set of elementary treatment contrasts is obtained as 0.09453

Table 3.20: Analysis of variance (Hampel's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	17.910	5.970	3.368	0.038
Block	6	5.232	0.872	0.492	0.805
Error	18	31.907	1.772		
Total	27	55.049			

Average variance for the set of elementary treatment contrasts is obtained as 0.09143

Table 3.21: Analysis of variance (New function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	19.830	6.610	3.994	0.024
Block	6	4.588	0.764	0.462	0.824
Error	18	29.174	1.655		
Total	27	53.592			

Average variance for the set of elementary treatment contrasts is obtained as 0.08454
Note: Value of p = 0.425

It is observed from the above tables that all the M-estimation procedures give similar results. The treatment effects have become highly significant now.

Analysis Based on Bickel's (1976) Proposal
The pseudo observations obtained from equation (3.14) for different M-estimation procedures and then ordinary analysis of variance was conducted as per procedure of Bickel (1976). The results are summarized in the sequel.

Table 3.22: Analysis of variance (Huber's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	3	18.848	6.282	3.432	0.039
Block	6	5.923	0.987	0.539	0.731
Error	18	32.952	1.830		
Total	27	57.723			

Average variance for the set of elementary treatment contrasts is obtained as 0.08834

Table 3.23: Analysis of variance (Ramsay's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	3	19.511	6.503	3.712	0.030
Block	6	5.073	0.845	0.482	0.813
Error	18	31.538	1.752		
Total	27	56.122			

Average variance for the set of elementary treatment contrasts is obtained as 0.09356

Table 3.24: Analysis of variance (Andrew's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	3	18.444	6.148	3.681	0.031
Block	6	5.243	0.873	0.523	0.783
Error	18	30.065	1.670		
Total	27	53.752			

Average variance for the set of elementary treatment contrasts is obtained as 0.09287

Table 3.25: Analysis of variance (Hampel's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	3	18.203	6.067	3.496	0.037
Block	6	5.384	0.897	0.517	0.787
Error	18	31.242	1.735		
Total	27	54.829			

Average variance for the set of elementary treatment contrasts is obtained as 0.091568

Table 3.26: Analysis of variance (New function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	3	19.320	6.440	4.015	0.023
Block	6	4.954	0.825	0.514	0.790
Error	18	28.873	1.604		
Total	27	53.147			

Average variance for the set of elementary treatment contrasts is obtained as 0.086758

Note: Value of $p = 0.425$

Table 3.27: Analysis of variance (After deleting observation No. 25)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	3	14.537	4.845	12.03	0.0002
Block	6	2.108	0.351	0.87	0.5351
Error	26	23.493	26		
Total	27	53.147			

Average variance for the set of elementary treatment contrasts is obtained as 0.1193571

Here also it is observed from the above tables that all the M-estimation procedures give similar results. The treatment effects have become significant now and level of significance of block effects has been lowered significantly. We then applied Cook-statistic for identifying outlying observations, if any. It was found that the observation number 1 corresponding to treatment number 4 in the 4^{th} replication is an outlier. This observation was deleted and analysis of variance is obtained again. Result is presented in Table 3.27. Interestingly, analysis is found to be similar to that as obtained through robust analysis.

Example having Outlier but Data within Range

We now present two examples where outliers could not be detected on the basis of the data. We have to apply formal tests to detect them

Example 3.3: An experiment with 6 treatments was conducted in a randomized complete block (RCB) design with 3 replications at Sugarcane Research Station, Buralikson, Golaghat in 2000 to study the effect of green manuring on the urea use efficiency on yield of sugarcane (net plot size 5.00m × 4.50m). The treatment details are

T1	100% of recommended dose of N + 70 kg/ha of P_2O_5 as S.S.P. + 60 kg/ha of K_2O as Mur.Pot.
T2	T1 + 10 t/ha of FYM
T3	75% of recommended dose of N + 70 kg/ha of P_2O_5 + 60 kg/ha of K_2O + Green manure
T4	T3 + 5 t/ha of FYM
T5	50% of recommended dose of N + 70 kg/ha of P_2O_5 + 60 kg/ha of K_2O + Green manure
T6	T5 + 5 t/ha of FYM

The data on grain yield per plot in kg for different treatments is given in Table

Table 3.28: Yield of sugarcane in kg/plot

Treatments	**Replication**		
	1	**2**	**3**
1	84.44	77.78	80.44
2	84.00	80.00	83.56
3	87.56	84.44	94.67
4	87.11	86.67	95.11
5	84.00	88.44	84.89
6	85.78	83.11	90.22

Analysis of variance was performed on the original data and results are given in Table 3.29. From the table one can observe that the treatment effects are significant at 5% level of significance and block effects are not significant at 5% level of significance.

Table 3.29: Analysis of variance with the original data

Source of variation	**DF**	**SS**	**MS**	**F**	**Significance level**
Treatment(adj)	5	178.061	35.612	3.52	0.0429
Block	2	67.800	33.900	3.35	0.077
Error	10	101.255	10.125		
Total	17	347.116			

Average variance for the set of elementary treatment contrasts is obtained as 6.7503771

Robust Analysis through M – estimation

We now applied different M-estimation procedures to this data and obtained the analysis of variance. We have taken the different values of the constants used in the objective functions as follows.

For Huber's function the value of the constant t=1.44
For Andrew's function a =1.243
For Ramsay's function a =0.285
For Hampel's function a =1.65, b=3.3, c=8.25

Weighted analysis

Firstly weighted analysis of variance has been done. The results are given in the sequel.

Table 3.30: Analysis of variance (Huber's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	5	172.744	34.548	2.75	0.08143
Block	2	74.122	37.061	2.95	0.09840
Error	10	125.632	12.563		
Total	17	372.498			

Average variance for the set of elementary treatment contrasts is obtained as 4.5687

Table 3.31: Analysis of variance (Andrew's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	5	163.211	32.642	2.51	0.10107
Block	2	61.643	30.821	2.37	0.14371
Error	10	130.048	13.004		
Total	17	354.902			

Average variance for the set of elementary treatment contrasts is obtained as 5.4732

Table 3.32: Analysis of variance (Ramsay's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	5	174.806	34.961	2.44	0.10780
Block	2	65.624	32.812	2.29	0.15177
Error	10	143.284	14.328		
Total	17	383.714			

Average variance for the set of elementary treatment contrasts is obtained as 5.5132

Table 3.33: Analysis of variance (Hampel's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	5	140.821	28.164	2.26	0.12763
Block	2	64.055	32.027	2.57	0.12571
Error	10	124.621	12.462		
Total	17	329.497			

Average variance for the set of elementary treatment contrasts is obtained as 5.7231

Table 3.34: Analysis of variance (Newly developed function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	5	178.653	35.730	2.64	0.08982
Block	2	67.671	33.835	2.50	0.13168
Error	10	135.343	13.53432		
Total	17	381.667			

Average variance for the set of elementary treatment contrasts is obtained as 4.7689

Note: Value of p = 0.315

Analysis Based on Bickel's (1976) Proposal

The pseudo observations obtained from equation (3.14) for different M-estimation procedures and then ordinary analysis of variance was conducted as per procedure of Bickel (1976). The results are summarized in the sequel.

Table 3.35: Analysis of variance (Huber's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	5	172.020	34.404	2.89	0.07205
Block	2	74.760	37.380	3.14	0.0874
Error	10	119.045	11.904		
Total	17	365.825			

Average variance for the set of elementary treatment contrasts is obtained as 3.2657

Table 3.36: Analysis of variance (Andrew's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	5	128.117	25.623	2.35	0.11723
Block	2	48.412	24.206	2.22	0.15928
Error	10	109.036	10.903		
Total	17	285.565			

Average variance for the set of elementary treatment contrasts is obtained as 5.6572

Table 3.37: Analysis of variance (Ramsay's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	5	171.405	34.281	2.69	0.08589
Block	2	77.228	38.614	3.03	0.09359
Error	10	127.439	12.743		
Total	17	376.072			

Average variance for the set of elementary treatment contrasts is obtained as 4.1452

Table 3.38: Analysis of variance (Hampel's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	5	161.868	32.373	2.48	0.10389
Block	2	69.446	34.723	2.66	0.11849
Error	10	130.538	13.053		
Total	17	361.852			

Average variance for the set of elementary treatment contrasts is obtained as 5.0352

Note: Value of p = 0.315

Table 3.39: Analysis of variance (After deleting observation No. 7)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	5	41.323	28.264	2.54	0.1059
Block	2	54.576	27.288	2.46	0.1409
Error	9	100.002	11.111		
Total	16	195.791			

Average variance for the set of elementary treatment contrasts is obtained as 9.629853

Here also it is observed from the above tables that all the M-estimation procedures give similar results. The treatment effects have become non-significant now and level of significance of block effects has been increased significantly. We then applied Cook-statistic for identifying outlying observations, if any. It was found that the observation number 7 corresponding to treatment number 2 in the 2nd replication is an outlier. These observations were deleted and analysis of variance is obtained again. Result is presented in Table 3.39. Interestingly, analysis is found to be similar to that as obtained through robust analysis.

Example 3.4: An experiment with 4 treatments was conducted in a randomized complete block (RCB) design with 3 replications at Regional Agriculture Test & Demonstration Centre, Varanasi in 2005 to study the effect of Sulphur on yield of mustard (net plot size 5.00 m x 3.00 m). The treatment details are

4 Methods of application of Sulpher :

T1	Control
T2	20 kg/ha of Sulpher through Super as basal
T3	150 kg/ha of Gypsum as soil application before sowing
T4	2.50 kg/ha of wettable sulphur as basal

The data on grain yield in quintals per plot for different treatments is given in the following table

Table 3.40: Yield of mustard in quintal/plot

Treatments	Replication		
	1	2	3
1	1.00	1.10	1.05
2	1.15	1.20	1.15
3	1.15	1.25	1.35
4	1.35	1.15	1.20

Analysis of variance was performed on the original data and results are given in Table 3.41. From the table one can observe that both the treatment effects and block effects are not significant at 5% level of significance.

Table 3.41: Analysis of variance with the original data

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.074	0.024	3.15	0.1077
Block	2	0.001	0.001	0.08	0.9244
Error	6	0.047	0.007		
Total	11	0.122			

Average variance for the set of elementary treatment contrasts is obtained as 0.0052315

Robust Analysis through M – estimation

We now applied different M-estimation procedures to this data and obtained the analysis of variance. We have taken the different values of the constants used in the objective functions as follows.

For Huber's function the value of the constant t=1.9

For Andrew's function a=1.41

For Ramsay's function a= 0.5

For Hampel's function a=1.9, b=3.8, c=9.5

Weighted analysis

Firstly weighted analysis of variance has been done. The results are given in the sequel.

Table 3.42: (Huber's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.050	0.016	6.89	0.02270
Block	2	0.011	0.005	2.34	0.17731
Error	6	0.014	0.002		
Total	11	0.075			

Average variance for the set of elementary treatment contrasts is obtained as 0.000134

Table 3.43: (Andrew's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.047	0.015	6.43	0.02647
Block	2	0.010	0.004	1.97	0.21993
Error	6	0.015	0.002		
Total	11	0.072			

Average variance for the set of elementary treatment contrasts is obtained as 0.000146

Table 3.44: (Ramsay's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.053	0.017	7.21	0.02048
Block	2	0.009	0.004	1.83	0.23961
Error	6	0.015	0.002		
Total	11	0.077			

Average variance for the set of elementary treatment contrasts is obtained as 0.000152

Table 3.45: (Hampel's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.060	0.019	6.92	0.02248
Block	2	0.012	0.005	2.03	0.21215
Error	6	0.017	0.002		
Total	11	0.089			

Average variance for the set of elementary treatment contrasts is obtained as 0.000141

Table 3.46: (Newly developed function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.065	0.021	7.43	0.01895
Block	2	0.017	0.008	2.87	0.13349
Error	6	0.018	0.002		
Total	11	0.100			

Average variance for the set of elementary treatment contrasts is obtained as 0.000123

Note: Value of p = 0.1850000

It is observed from the above tables that all the M-estimation procedures give similar results. The treatment effects have become highly significant now. The level of significance for block effects has been lowered significantly.

Analysis based on Bickel's (1976) Proposal

The pseudo observations obtained from equation (3.14) for different M-estimation procedures and then ordinary analysis of variance was conducted as per procedure of Bickel (1976). The results are summarized in the sequel.

Table 3.47: Analysis of variance (Huber's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.045	0.014	7.09	0.02128
Block	2	0.009	0.004	2.18	0.19425
Error	6	0.013	0.002		
Total	11	0.067			

Average variance for the set of elementary treatment contrasts is obtained as 0.000124

Table 3.48: Analysis of variance (Andrew's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.060	0.019	6.79	0.02345
Block	2	0.013	0.006	2.23	0.18873
Error	6	0.018	0.002		
Total	11	0.091			

Average variance for the set of elementary treatment contrasts is obtained as 0.000138

Table 3.49: Analysis of variance (Ramsay's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.049	0.016259	7.61	0.01810
Block	2	0.009	0.004508	2.11	0.20234
Error	6	0.013	0.0021365		
Total	11	0.071			

Average variance for the set of elementary treatment contrasts is obtained as 0.000143

Table 3.50: Analysis of variance (Hampel's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.065	0.021	7.11	0.02114
Block	2	0.014	0.006	2.23	0.18873
Error	6	0.018	0.003		
Total	11	0.097			

Average variance for the set of elementary treatment contrasts is obtained as 0.000132

Table 3.51: Analysis of variance (New function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.072	0.024	7.57	0.01832
Block	2	0.019	0.009	2.92	0.13013
Error	6	0.019	0.003		
Total	11	0.110			

Average variance for the set of elementary treatment contrasts is obtained as 0.000118

Note: Value of p = 0.1850000

Table 3.52: Analysis of variance (After deleting observation No. 1)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	3	0.061	0.020	7.43	0.0273
Block	2	0.014	0.007	2.59	0.1692
Error	5	0.013	0.002		
Total	10	0.088			

Average variance for the set of elementary treatment contrasts is obtained as 0.0024383

Here also it is observed from the above tables that all the M-estimation procedures give similar results. The treatment effects have become significant now and level of significance of block effects has been lowered significantly. We then applied Cook-statistic for identifying outlying observations, if any. It was found that the observation number 1 corresponding to treatment number 1 in the 1st replication is an outlier. This observation was deleted and analysis of variance is obtained again. Result is presented in Table 3.52. Interestingly, analysis is found to be similar to that as obtained through robust analysis. Here interesting point to note that even an observation which does not appear to be very high or low can be an outlier. Above two examples are for single outlier, we now give an example with two outliers.

Example of Two Outliers

Here we present one examples in which data sets contain two outliers.

Example 3.5: An experiment with 5 treatments was conducted in a randomized complete block (RCB) design with 4 replications at Crop Research Centre, G.B.P.U.A.T., Pantnagar in 2006 to study the effect of different Phosphorus levels and Phosphate Solubilizing Bacteria (P.S.B.) on the productivity of urd (net plot size 4.00 m × 1.80 m). The treatments of the experiments are as follows:

Treatments:

T1	0
T2	20 kg/ha of P_2O_5 as Super
T3	40 kg/ha of P_2O_5 as Super
T4	P.S.B.
T5	20 kg/ha of P_2O_5 as Super + Phosphorus Solublizing Bacteria (P.S.B.)

The data on grain yield in quintals per plot for different treatments is given in Table 3.53:

Table 3.53: Yield of urd in quintal/plot

Treatments	**Replication**			
	1	**2**	**3**	**4**
1	1.23	1.27	1.30	1.23
2	1.32	1.50	1.35	1.38
3	1.40	1.55	1.40	1.37
4	1.32	1.38	1.55	1.43
5	1.47	1.63	1.45	1.50

Analysis of variance was performed on the original data and results are given in Table 3.54. From the table one can observe that both the treatment effect is highly significant where as block effects are not significant at 5% level of significance.

Table 3.54: Analysis of variance with the original data

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.149	0.037	8.24	0.0020
Block	3	0.039	0.013	2.88	0.0802
Error	12	0.054	0.004		
Total	19	0.242			

Average variance for the set of elementary treatment contrasts is obtained as 0.0022688

Robust Analysis through M – estimation

We now applied different M-estimation procedures to this data and obtained the analysis of variance. We have taken the different values of the constants used in the objective functions as follows.

For Huber's function the value of the constant t =2.0

For Andrew's function a =1.51

For Ramsay's function a =0.41

For Hampel's function a =2.01, b = 4.02, c =10.05

Weighted analysis

Firstly weighted analysis of variance has been done. The results are given in sequel.

Table 3.55: Analysis of variance (Huber's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.234	0.058	2.27	0.12197
Block	3	0.128	0.042	1.65	0.23023
Error	12	0.310	0.025		
Total	19	0.672			

Average variance for the set of elementary treatment contrasts is obtained as 0.000981

Table 3.56: Analysis of variance (Andrew's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.232	0.057	1.94	0.16668
Block	3	0.117	0.039	1.32	0.31349
Error	12	0.357	0.029		
Total	19	0.707			

Average variance for the set of elementary treatment contrasts is obtained as 0.00114

Table 3.57: Analysis of variance (Ramsay's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.235	0.058	1.86	0.18236
Block	3	0.114	0.038	1.21	0.34816
Error	12	0.379	0.031		
Total	19	0.738			

Average variance for the set of elementary treatment contrasts is obtained as 0.00113

Table 3.58: Analysis of variance (Hampel's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.235	0.049	2.01	0.15707
Block	3	0.114	0.037	1.53	0.25729
Error	12	0.296	0.024		
Total	19	0.735			

Average variance for the set of elementary treatment contrasts is obtained as 0.00114

Table 3.59: Analysis of variance (Newly developed function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.228	0.057	2.13	0.13963
Block	3	0.116	0.038	1.44	0.27989
Error	12	0.322	0.026		
Total	19	0.666			

Average variance for the set of elementary treatment contrasts is obtained as 0.000967
Note: Value of $p = 0.250$

It is observed from the above tables that all the M-estimation procedures give similar results. The treatment effects have become non significant now. The level of significance for block effects has been increased significantly.

Analysis Based on Bickel's (1976) Proposal

The pseudo observations obtained from equation (3.14) for different M-estimation procedures and then ordinary analysis of variance was conducted as per procedure of Bickel (1976). The results are summarized in the sequel.

Table 3.60: Analysis of variance (Huber's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.245195	0.061299	2.36	0.1119577
Block	3	0.137922	0.045974	1.77	0.2062995
Error	12	0.311688	0.025974		
Total	19	0.694805			

Average variance for the set of elementary treatment contrasts is obtained as 0.000977

Table 3.61: Analysis of variance (Andrew's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.242	0.060	1.99	0.16020
Block	3	0.136	0.045	1.49	0.26707
Error	12	0.365	0.030		
Total	19	0.743			

Average variance for the set of elementary treatment contrasts is obtained as 0.00125

Table 3.62: Analysis of variance(Ramsay's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.237	0.059	1.92	0.171
Block	3	0.123	0.041	1.33	0.310
Error	12	0.371	0.030		
Total	19	0.731			

Average variance for the set of elementary treatment contrasts is obtained as 0.00127

Table 3.63: Analysis of variance (Hampel's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.229	0.057	2.18	0.13302
Block	3	0.135	0.045	1.71	0.21789
Error	12	0.316	0.026		
Total	19	0.680			

Average variance for the set of elementary treatment contrasts is obtained as 0.00106

Table 3.64: Analysis of variance (New function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.241	0.060	2.21	0.12922
Block	3	0.130	0.043	1.59	0.24334
Error	12	0.327	0.027		
Total	19	0.698			

Average variance for the set of elementary treatment contrasts is obtained as 0.000956

Note: Value of p = 0.250

Table 3.65: Analysis of variance (After deleting observation No. 8 and 16)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	4	0.136	0.034	1.22	0.3611
Block	3	0.067	0.022	0.80	0.5202
Error	10	0.279	0.027		
Total	17	0.511			

Average variance for the set of elementary treatment contrasts is obtained as 0.017768

Here also it is observed from the above tables that all the M-estimation procedures give similar results. The treatment effects have become non-significant now and level of significance of block effects has been increased significantly. We then applied Cook-statistic for identifying outlying observations, if any. It was found that the observation number 8 and 16 corresponding to treatment number 3 in the 2^{nd} replication and treatment number 1 in the 1^{st} replication are outliers. These observations were deleted and analysis of variance is obtained again. Result is presented in Table 3.65. Interestingly, analysis is found to be similar to that as obtained through robust analysis.

Example having No Outlier

Here we present another example in which no outlier is present.

Example 3.6: An experiment with 3 treatments was conducted in the randomized complete block (RCB) design with 7 replications at Agricultural Research Station, Dharwad to asses the crop loss due to foliar disease alternaria blight in hybrid cotton (net plot size: 2.40m × 6.00m). The treatment details are

3 Plant Protective Measures:

T_0	Control(natural infection)
T_1	Supplementing artifical inoculum to create epiphytotic
T_2	Complete protection of the crop by spraying cuman l

The data on the yield per plot in kg for different treatments are given below

Table 3.66: Yield of cotton in Kg/plot

Replications	**Treatments**		
	1	**2**	**3**
1	2.37	2.30	2.55
2	2.57	2.75	2.51
3	2.64	2.70	2.89
4	2.78	2.89	2.77
5	3.14	2.90	3.07
6	2.78	2.53	2.45
7	2.83	2.74	2.58

Analysis of variance was performed on the original data and results are given in Table 3.67. From the table one can observe that the treatment effects are not significant at 5% level of significance.

Table 3.67: Analysis of variance with the original data

Source of variation	**DF**	**SS**	**MS**	**F**	**Significance level**
Treatment(adj)	2	0.0082952	0.00414762	0.23	0.8015725
Block	6	0.7059238	0.117654	6.39	0.0032546
Error	12	0.22090476	0.01840873		
Total	20	0.9351238			

Average variance for the set of elementary treatment contrasts is obtained as 0.00526

Robust Analysis through M – estimation

We now applied different M-estimation procedures to this data and obtained the analysis of variance. We have taken the different values of the constants used in the objective functions as follows.

For Huber’s function the value of the constant t=1.79

For Andrew’s function a=1.65

For Ramsay’s function a=0.47

For Hampel’s function a=1.6, b=3.20, c=8.23

Weighted Analysis

Firstly weighted analysis of variance has been done. The results are given in the sequel.

Table 3.68: Analysis of variance (Huber's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment	2	0.0082	0.0041	0.23	0.8015
Block	6	0.7059	0.1176	6.39	0.0032
Error	12	0.2209	0.01840		
Total	20	0.9351			

Average variance for the set of elementary treatment contrasts is obtained as 0.00526

Table 3.69: Analysis of variance (Andrew's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	2	0.008	0.004	0.269	0.768
Block	6	0.745	0.124	8.227	0.001
Error	12	0.181	0.015		
Total	20	0.935			

Average variance for the set of elementary treatment contrasts is obtained as 0.00442

Table 3.70: Analysis of variance (Ramsay's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	2	0.006	0.003	0.117	0.89013
Block	6	0.829	0.138	4.763	0.01045
Error	12	0.348	0.029		
Total	20	1.184			

Average variance for the set of elementary treatment contrasts is obtained as 0.010214

Table 3.71: Analysis of variance (Hampel's function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	2	0.008	0.004	0.23	0.80157
Block	6	0.705	0.117	6.39	0.00326
Error	12	0.220	0.018		
Total	20	0.935			

Average variance for the set of elementary treatment contrasts is obtained as 0.00526

Table 3.72: Analysis of variance (New function: Weighted analysis)

Source of variation	DF	SS	MS	F	Significance level
Treatment(adj)	2	0.008	0.004	0.23	0.80157
Block	6	0.705	0.117	6.39	0.00325
Error	12	0.220	0.0184		
Total	20	0.935			

Average variance for the set of elementary treatment contrasts is obtained as 0.00526
Note: Value of $p = 0.105$

It is observed from the above tables that all the M-estimation procedures give similar results. The treatment effects remain non significant.

Analysis Based on Bickel's (1976) Proposal
The pseudo observations obtained from equation (3.14) for different M-estimation procedures and then ordinary analysis of variance was conducted as per procedure of Bickel (1976). The results are summarized in the sequel.

Table 3.73: Analysis of variance (Huber's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	2	0.008	0.004	0.43	0.6604
Block	6	0.705	0.117	12.19	0.0002
Error	12	0.115	0.009		
Total	20	0.830			

Average variance for the set of elementary treatment contrasts is obtained as 0.08834

Table 3.74: Analysis of variance (Ramsay's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	2	0.008	0.004	1.72	0.2202
Block	6	0.714	0.119	49.22	<.0001
Error	12	0.029	0.002		
Total	20	0.751			

Average variance for the set of elementary treatment contrasts is obtained as 0.09356

Table 3.75: Analysis of variance (Andrew's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	2	0.008	0.004	1.72	0.2202
Block	6	0.714	0.119	49.22	<.0001
Error	12	0.029	0.002		
Total	20	0.751			

Average variance for the set of elementary treatment contrasts is obtained as 0.09287

Table 3.76: Analysis of variance (Hampel's function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	2	0.008	0.004	0.35	0.7100
Block	6	0.766	0.127	10.70	0.0003
Error	12	0.143	0.011		
Total	20	0.918			

Average variance for the set of elementary treatment contrasts is obtained as 0.091568

Table 3.77: Analysis of variance (New function: Bickel's pseudo observations)

Source of variation	DF	SS	MS	F-value	Significance level
Treatment(adj)	2	0.008	0.004	0.43	0.6576
Block	6	0.705	0.117	12.32	0.0002
Error	12	0.114	0.009		
Total	20	0.828			

Average variance for the set of elementary treatment contrasts is obtained as 0.086758
Note: Value of $p = 0.105$

Here also it is observed from the above tables that all the M-estimation procedures give similar results. The treatment effects remain non significant. We then applied Cook-statistic for identifying outlying observations, if any. It was found that there is no outlying observation present in the data. From the above tables, it is clear that analysis through Huber' function and Newly developed function and through original data analysis give same results, since there is no outlier, each observation got same weight.

3.4 Discussions

Different M-estimation procedures as available in the literature, along with the newly developed function were applied to a number of experiments taken from AFEIS, IASRI. Some of these analyses are presented in Section 3.3.3 In some of the examples treatment effects are not significant with the original data. But robust analysis revealed that the treatment effects are actually significant at 5% level of significance, where as in some examples, non-significant treatment effects become significant through robust analysis. That is, inferences to be drawn are reversed through robust analysis. Actually all these experiments contain some outlying observations. Once the outlying observations are deleted, results become similar to those obtained through robust analysis. However, from statistical point of view it is not advised to delete any observation. Because, every observation carries some information that should be exploited. Robust analysis actually gives small weights to those outlying observations, thus extracting some information from that observation.

However, a question may arise in our mind that which M-estimation procedure should we use? It is difficult to answer this question, because these procedures depend on weights and weights are determined by observations. From experiences of analyzing a good number of experiments, it is observed that Huber's function and newly developed function perform well. This is observed from the fact that the average variance of the set of elementary treatment contrasts is small for most of the experiments for these two functions comparing to other M-estimation functions. However, this is an empirical evidence true, there is no theoretical proof.

In those experiments where no outlier is present, there is a little difference between the analysis with original data and analysis through robust regression, as we have seen from the last example. Off course the levels of significance are changed a little bit. It is therefore, generally advised to carryout analysis through ordinary least squares (OLS), if we are sure that the data do not contain any outlying observation and the errors are normal. Because OLS estimates posses some good statistical properties. But in general we do not know the form of distribution of the errors in advance. It is therefore, suggested to always apply robust analysis. Even the error distribution is normal, we may not loose much efficiency.

Another point to note here is that robust analysis is also used to detect outlying observations. As mentioned in Chapter 1, robust analysis leaves the residuals of the outlying observations large and thereby helps to identify them easily. For example consider the Example 3.5. The standardized residuals obtained from different robust fits are presented in Table 3.78.

It is observed from the table that the residual for the observation 17.76 from all these functions is very high, indicating that this observation is an outlier. In fact, this observation which corresponds to treatment number 4 in replication number 4 is an outlier indicated through Cook-statistic. Deleting this observation and carrying out usual analysis gives similar results as obtained through robust analysis. Therefore, robust analysis is a very good tool for identifying outlying observation also.

Table 3.78: Standardized Residuals of Different Observations in Different Functions

Observation	Huber	Andrews	Ramsay	Hampel	New
14.30	0.630	0.110	0.703	0.721	0.724
13.59	0.769	0.216	0.785	0.511	0.767
11.92	-3.839	-3.874	-1.393	-2.798	-1.101
12.18	0.113	-0.052	-0.186	0.464	-0.389
14.08	0.187	0.079	0.285	0.219	0.345
13.22	-0.434	-0.503	0.042	-0.566	0.107
12.56	0.072	0.041	0.053	0.001	0.128
12.06	0.176	0.376	-0.387	0.346	-0.581
14.14	0.376	-0.223	1.004	0.014	1.117
11.53	-9.106	-8.835	-3.030	-7.488	-2.395
12.53	-0.194	-0.674	0.577	-0.550	0.732
12.31	1.327	0.946	0.743	0.870	0.545
12.97	-0.805	-0.236	-1.136	-0.062	-1.417
12.55	0.802	1.202	-0.426	0.841	-0.832

11.32	-1.578	-0.870	-1.650	-0.779	-1.877
17.76	**33.664**	**31.310**	**12.945**	**26.199**	**10.400**
14.05	0.554	0.392	0.421	0.497	0.481
13.22	0.085	-0.052	0.243	-0.173	0.299
12.60	0.793	0.675	0.340	0.548	0.395
11.64	-1.432	-1.100	-1.097	-0.872	-1.176
13.39	0.568	0.882	0.126	0.319	0.172
12.30	-1.217	-0.755	-0.615	-1.349	-0.496
12.75	4.909	4.882	1.800	3.478	1.602
11.09	-0.860	-0.106	-1.153	-0.628	-1.278
13.19	-3.808	-3.103	-1.636	-2.571	0.724
14.58	6.963	6.639	2.995	5.278	0.767
12.61	0.836	1.172	0.168	0.818	-1.101
11.76	-0.832	-0.098	-1.031	-0.179	-0.389

APPENDIX 3.1

SAS Code: Robust analysis of designed experiments using M-estimation

```
options ls=72 ps=2000;
data ran;
input trt blk y;
datalines;

;
proc iml;
	use ran;
		read all into X;
			mrowX=nrow(X);
			ncolX=ncol(X);
			v=max(X[,1]);
		x1=j(mrowX,v,0);
	X2=j(mrowX,1,1);
	y=X[,ncolX];
/*Create Delta matrix*/
do i=1 to mrowX;
	x1[i,X[i,1]]=1;
end;
/*Create the matrix for the nuisance factor*/
do j=2 to ncolX-1;
	order=max(X[,j]);
	D1=J(mrowX,order,0);
	intB=max(X[,2]);
/*Create the diagonal matrix of block size*/
	K=J(intB,intB,0);
	do i=1 to mrowX;
		D1[i,X[i,j]]=1;
		K[X[i,2],X[i,2]]=K[X[i,2],X[i,2]]+1;
	end;
	X2=X2||D1;
end;
x=x1||x2;
/*Calculation of parameter vector*/
betahat0=j(ncol(x),1,0);
betahat0=ginv(x`*x)*x`*y;
yhat=x*betahat0;
resid=y-yhat;
n=nrow(resid);
rank=round(trace(ginv(x)*x));
s1=median(abs(resid-median(resid)))/0.6745;
```

```
/*Calculation of the standarized residuals*/
z1=j(n,1,0);
z1=resid/s1;print z1;
a=1;
g=1.5;
          w0=j(n,1,0);
          w=j(n,n,0);
/*Calculation of weight matrix by using Huber's function */
do while(a>0.005);
          do i=1 to n;
                    z1[i,1]=abs(z1[i,1]);
                    if z1[i,1]<=g then
                    w0[i,1]=1;                    …………………….A
                    else w0[i,1]=g/z1[i,1];
          end;
          w=diag(w0);

betahat1=j(ncol(x),1,0);
          betahat1=ginv(x`*w*x)*(x`*w*y);
          yhat1=x*betahat1;
          resid2=(y-yhat1);
          s2=median(abs(resid2-median(resid2)))/0.6745;
          z2=(y-yhat1)/s2;
          a=abs((s2-s1)/s1);
          z1=z2;
          s1=s2;
end;
ww=sqrt(w);
C=(x1`*w*x1)-(x1`*w*x2)*(ginv(x2`*w*x2))*(x2`*w*x1);
Q=x1`*w*y-x1`*w*x2*(ginv(x2`*w*x2))*x2`*w*y; /*Q is the treatment total vector*/
tauhat=ginv(c)*Q;
TSS=Q`*ginv(c)*Q;
B=D1`*ww*Y;/*B is the block total vector*/
one=j(n,1,1);
cf=(one`*ww*y)*(one`*ww*y)/n;
BSS=B`*inv(k)*B-cf;
bms=bss/(intB-1);
TMS= TSS/(v-1);
totss=(y`*w*y)-cf;
ess=totss-tss-bss;
ems=ess/(n-rank);
print totss tss  bss ess;
print tms bms ems;
FT=TMS/ems;
FB=bms/ems;
pvalt=1 - probf(ft,v-1,n-rank);
```

```
pvalb=1 - probf(fb,intB-1,n-rank);
print "Huber function data set" 25;
print pvalt pvalb ft fb;
/*Calculation of the all possible elementary contrasts*/
p=j((v-1),v,0);
do i=1 to v-1;
        p[i,1]=1;
        j=i+1;
        p[i,j]=-1;
end;
print p;
conout=j((v-1),4,0);
contss=j((v-1),1,0);
do i=1 to nrow(p);
pi=p[i,];
/*Calculation of t statistics*/
contst=((pi*tauhat)`*ginv((pi*ginv(c)*pi`))*(pi*tauhat))/ems;
conpval=1 - probf(contst,1,n-rank);
conout[i,3]=contst;
conout[i,4]=conpval;
end;
varcon=j(v-1,v-1,0);
varcon=p*ginv(c)*p`*ems;
abgvar=trace(varcon)/(v-1);
print '          trt      vs' 'trt   ' '        F Value          ' 'Pr > F          ';
print varcon abgvar;
run;
quit;
```

Note: Calculation of the weight matrix for other functions is done accordingly by replacing “A” by appropriate codes.

Chapter IV

Robust Analysis of Designed Experiments II:LMS Estimation

4.1 Introduction

In Chapter III, we have presented the M-estimation for robust analysis of experimental data in details. In the present chapter we discuss another robust method of analysis of data viz. Least Median of Squares (LMS) method. We begin with the concept of LMS method and then make appropriate modification for its application in designed experiments.

In the last few decades, studies on handling of outliers in regression be classified into two groups viz. i) regression diagnostics and ii) robust regression. Both approaches are closely related and are used for identifying outliers and pointing out inadequacies of the model. They, however, proceed in a different way. Regression diagnostics first attempt to identify points that have to be deleted from data set, before applying a regression method. Robust regression tackles these problems in the inverse order, by designing estimators that dampen the impact of points that would be highly influential otherwise.

As mentioned earlier least squares (LS) model can be distorted even by a single outlying observation. The fitted line or surface might be tipped so that it no longer passes through the bulk of the data. It will introduce many small or moderate errors in order to reduce the effect of a very large error. For example, if a large error is reduced from 200 to 50, its square is reduced from 40,000 to 2,500. Increasing an error from 5 to 15 increases its square from 25 to 225(Chatterjee and Machler, 1997). Thus, a least squares fit might introduce many small errors in order to reduce a large one. A complete name for the LS method would perhaps be least sum of squares, but apparently few people have objected to the deletion of the word "sum" – as if the only sensible thing to do with *n* positive numbers would be to add them. Perhaps as a consequence of its historical name, several people have tried to make this estimator robust by replacing the square by something else, not touching the summation sign (M-estimator). One may think of replacing sum by a median as it is more robust than sum. This yields the least median of squares estimator, given by

$$\underset{\hat{\theta}}{Minimize} \;\; \underset{i}{med}\, r_i^2 , \qquad (4.1)$$

where θ is the parameter vector and r_i is the i^{th} residual. This estimator was proposed by Rousseeuw (1984) who actually materialized an idea of Hampel (1975). 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 Section 4.2, this concept is thoroughly discussed in the context of linear regression model. In Section 4.3 its applicability to designed experiments has been explored and in Section 4.4 this method is actually applied to some real experimental data. The Chapter is concluded with a section on discussion.

4.2 LMS Estimation in Linear Regression Model (Rousseeuw, 1984)

Consider the linear regression model as given in (1.1), *i.e.*, E(**y**) = $\mathbf{X}\boldsymbol{\beta}$, with **X** of full column rank *p*. Let **b** be the LS estimate of $\boldsymbol{\beta}$. With *n* observations the residuals from this LS estimate are $e_i(\mathbf{b}) = y_i - x_i'\mathbf{b}$, ($i$=1,2,…,$n$). The LMS estimate $\hat{\boldsymbol{\beta}}_p$ is the value of **b** minimizing the median of the square residuals $e_i^2(b)$. Thus $\hat{\boldsymbol{\beta}}_p$ minimizes the scale estimate

$$\sigma^2(\mathbf{b}) = e_{[med]}^2(\mathbf{b}), \tag{4.2}$$

where $e_{[k]}^2(\mathbf{b})$ 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

$$med = \text{The integer part of } (n+p+1)/2. \tag{4.3}$$

The parameter estimate satisfying (4.2) 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 can still provide an unbiased estimate of the regression coefficients. This is the maximum break down point that can be tolerated. For a higher proportion of outliers there is no longer a model that fits the majority of the data. The very robust behavior of the LMS estimate is in the contrast to that of the least squares estimate $\hat{\boldsymbol{\beta}}$ minimizing

$$S(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}), \tag{4.4}$$

which can be written as

$$S(\mathbf{b}) = \sum_{i=1}^{n} e_i^2(\mathbf{b}). \tag{4.5}$$

Only one outlier needs to be moved towards infinity to cause an arbitrarily large change in the estimate $\hat{\boldsymbol{\beta}}$, the breakdown point of $\hat{\boldsymbol{\beta}}$ is zero.

The definition of $\hat{\boldsymbol{\beta}}_p$ in (4.2) gives no indication of how to find such a parameter estimate. Since the surface to be minimized has many local minima, therefore, approximate methods may be used.

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, which we shall discuss in section 4.3.

In this algorithm all possible sample of size *p* are drawn. A regression surface is fitted to each set of observations and the median squared residual is calculated (This is merely a matter of solving set of *p* linear equations with *p* unknown parameters). The model that had the smallest median squared residual is used. Evaluating all possible subsets of *p* observations can be computationally infeasible for large data sets. When *n* is large, Rousseeuw and Leroy (1987) recommended taking random samples of observations and using the best solution obtained from these randomly selected subsets.

The second problem is that there is no theory for constructing confidence intervals for LMS regression coefficients or for testing hypotheses about them. Rousseeuw and Leroy (1987) has proposed calculating a distance measure based on LMS regression and using it to identify outliers with respect to the LMS regression. These observations are set aside and least squares regression is fitted to the remaining data. The procedure is called reweighted least squares regression.

This approach has some obvious appeal. In this approach, multiple regression line is fitted through the clean data points only, outlying observations, if any, remained outside of the fitted line. The only drawbacks of the LMS estimate are that it does not consider all the data set for estimation of parameters.

4.2.1 Computational Algorithm (Rousseeuw and Leroy, 1987)

The computation of the least median squares regression coefficients is quite difficult. It may not be possible to write down a straightforward formula for the LMS estimator. In fact, it appears that this computational complexity is inherent to all (known) affine equivariant high breakdown regression estimators, because they are closely related to projection pursuit methods. The algorithm implemented is similar in spirit to the bootstrap.

The algorithm proceeds by repeatedly drawing subsamples of *p* different observations. For such a sub sample, indexed by $\mathbf{J} = \{i_{1,} i_{2,} \ldots, i_{p,}\}$, one can determine the regression surface through the *p* points and denote the corresponding vector of coefficients by $\boldsymbol{\beta}_{\mathbf{J}}$. We call such a solution $\boldsymbol{\beta}_{\mathbf{J}}$ a trial estimate. For each $\boldsymbol{\beta}_{\mathbf{J}}$ one also determines the corresponding LMS objective function with respect to the whole data set. This means that the value

$$\underset{i=1\ldots n}{med} \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}_{\mathbf{J}}\right)^2 \tag{4.6}$$

is calculated. Finally, one will retain the trial estimate for which this value is minimal. Now the question is how many sub samples should one consider? In principle, one may repeat the above procedure for all possible sub samples of size *p*, of which there are ${}^{n}C_{p}$ sub samples. Unfortunately, ${}^{n}C_{p}$ increases very fast with increase in *n* and *p*, in many applications, this would become infeasible. In such cases, one may perform 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, assuming that *n*/*p* is large, is

$$1 - \left(1 - (1-\varepsilon)^p\right)^m. \tag{4.7}$$

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.

4.2.2 Calculation of Scale Parameter

Apart from the regression coefficients, we have to obtain a robust estimate of the scale parameter σ (the dispersion of the errors e_i) also. Rousseeuw and Leroy (1987) also proposed a robust method for calculating this parameter. According to this method an initial scale estimate s^0 (say) is calculated. This s^0 is based on the minimal median and multiplied by a finite sample correction factor (which depends on *n* and *p*) for the case of normal errors:

$$s^0 = 1.4826\left(1 + 5/(n-p)\right)\sqrt{\underset{i}{med}\; r_i^2\left(\hat{\beta}\right)} \tag{4.8}$$

the factor $1.4826 = 1/\Phi^{-1}(0.75)$ was introduced because $med_i \left|z_i\right| / \Phi^{-1}(0.75)$ is a consistent estimator of σ when the z_i are distributed like N (0, σ^2), where $\Phi^{-1}(0.75)$ is standard normal ordinate value. From an empirical study Rousseeuw and Leroy (1987) showed that this factor 1.4806 alone was not enough, because the scale estimate became too small in regression with normal errors, especially for small samples. It was not obvious at all to find an appropriate factor in practice. The behaviour of the original scale estimate through simulation, both in normal error situation and in situations where there was contamination in the response or in the explanatory variables revealed that multiplication with the factor 1+5/(*n*–*p*) gave a satisfactory solution (of course for large *n*). This preliminary scale estimate s^0 is then used to determine a weight w_i for the i^{th} observation, namely

$$\begin{aligned} w_i &= 1 \qquad \text{if } \left|r_i / s^0\right| \le 2.5 \\ &= 0 \qquad \text{otherwise.} \end{aligned}$$

By means of these weights the final scale estimate σ^* for LMS regression is calculated as

$$\sigma^* = \sqrt{\left(\sum_{i=1}^{n} w_i\, r_i^2\right) \Big/ \left(\sum_{i=1}^{n} w_i - p\right)}. \tag{4.9}$$

The main advantage of this formula for σ^* is that it is not influenced by outliers.

4.2.3 LMS Estimator in Designed Experiments

As mentioned in Section 4.2, LMS estimator has a high break down point. This method, however, did not find much favour 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. An algorithm for computing this estimator in linear regression models by Rousseeuw (1984) helped its use in regression. 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 problem is the problem of connectedness of the subset of observations. 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. Secondly, in case of design of experiments, interest is in estimation of some functions of treatment effects rather than the whole set of parameters. This may 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 in the next section that is suitable for application to designed experiments.

4.3 Application of LMS to Designed Experiments

In this section we modify the LMS procedure for its suitability and applicability in designed experiments. As mentioned in the previous section that the connectedness is the main problem in designed experiments, LMS method as such cannot be applied. Therefore, this is appropriately modified and then applied to experimental data taken from AFEIS. 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 assume that the design remains connected after losing one or two observations.

ii) Obtain least squares residuals for each subset. There are ${}^{n}C_{n-1}$ 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 Complete Block (RCB) designs 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 RCB designs, 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 (see Krishan Lal *et al*., 2001). However, this size of subset can be increased for those designs that are robust against the loss of more than two observations. This method is applied to data of a number of experiments retrieved from AFEIS. Relevant program for carrying out the analysis has been written in SAS/IML. The listing of the program is given in APPENDIX 4.1. Some of these analyses are presented in the sequel.

4.4 Illustration

Example 4.1: An experiment with 4 treatments was conducted in the randomized complete block (RCB) design with 4 replications at Regional Agriculture Test & Demonstration Centre, Azamgarh in 1987 to study the effects of different methods of application of paras on the yield of paddy (net plot size 5.00m × 3.00m). The treatment details are

T_0	Control
T_1	Paras liquid at 2.50 p.p.m.as foliar spray only once
T_2	Paras liquid at5.00 p.p.m.as foliar spray twice
T_3	Paras granules at 25kg/ha broadcasted

Table 4.1: Yield of paddy in quintal/plot

Treatments	Replication			
	1	**2**	**3**	**4**
1	6.50	6.70	6.40	5.20
2	6.90	6.00	6.50	6.60
3	6.90	5.80	6.90	7.20
4	6.30	5.90	6.20	5.80

Analysis of variance was performed on the original data and results are given in Table 4.2. From the table one can observe that both the treatment effects and block effects are not significant at 5% level of significance.

Table 4.2: Analysis of variance with the original data

Source of variation	**DF**	**SS**	**MS**	**F**	**Significance level**
Treatment	3	1.027	0.342	1.33	0.3251
Block	3	0.787	0.262	1.02	0.4295
Error	9	2.322	0.258		
Total	15	4.137			

Average variance for the set of elementary treatment contrasts is obtained as 0.129

LMS Method

We now applied LMS technique to this data set. Size of the subset is chosen as n–1. The results of analysis of variance on the chosen subset are presented in Table 4.3.

Table 4.3: Analysis of variance through LMS (Size of subset is *n*-1)

Source of variation	DF	SS	MS	F	Significance level
Treatment	3	1.581	0.527	3.27	0.0800
Block	3	1.487	0.495	3.08	0.0904
Error	8	1.288	0.161		
Total	14	4.0160			

Average variance for the set of elementary treatment contrasts is obtained as 0.086463

The dramatic change through this analysis observed is that the significant levels of both treatment and block effects have been lowered. Treatment differences are now significant at 8% level of significance. There is huge reduction in the average variance for the estimated elementary treatment contrasts. Observation number 5 corresponding to treatment number 2 in the 1st block is actually deleted in the chosen set. We then applied Cook-statistic for identification of outlier (s), if any. It was found that the observation number 5 is an outlier. In the chosen set this observation was deleted.

Example 4.2: An experiment with 8 treatments was conducted in the randomized complete block (RCB) design with 3 replications at Institute of Agriculture Science., B.H.U., Varanasi in 2005 to find out long term effect of inorganic and organic sources of nutrients on productivity and soil health of lentil crop (net plot size 10.00 m × 9.00 m). The treatment details of the experiments are as follows:

T1 =Control
T2 =120 kg/ha of N+60 kg/ha of P2O5+40 kg/ha of K2O
T3 =1/2 x A2
T4 =A3+60 kg/ha of N as foliar spray
T5 =60 kg/ha of N through F.Y.M. before sowing
T6 =2 x A5
T7 =A3+60 kg/ha of N through F.Y.M. before sowing
T8 =Farmer's Practice

The data on grain yield per plot in quintals for different treatments is given in Table 4.4:

Table 4.4: Yield of lentil in quintal/plot

Treatments	Replication		
	1	**2**	**3**
1	1.00	1.10	1.60
2	3.20	5.30	0.80
3	2.20	1.10	1.60
4	3.20	3.10	0.50
5	4.10	3.50	1.50
6	3.10	1.70	1.40
7	3.10	2.40	1.60
8	1.80	0.60	0.50

Analysis of variance was performed on the original data and results are given in Table 4.5. From the table one can observe that the treatment effects are not significant at 5% level of significance and block effects are significant at 5% level of significance.

Table 4.5: Analysis of variance with the original data

Source of variation	**DF**	**SS**	**MS**	**F**	**Significance level**
Treatment	7	12.666	1.809	1.95	0.1367
Block	2	10.155	5.077	5.46	0.0176
Error	14	13.010	0.929		
Total	23	35.833			

Average variance for the set of elementary treatment contrasts is obtained as 0.6193333

LMS Method

We then applied LMS-estimation procedure to the data. The result is presented in Table 4.6. The dramatic effect to note here is that the treatment effects are now significant at 5% level of significance.

Table 4.6: Analysis of variance through LMS (Size of subset is *n*-2)

Source of variation	DF	SS	MS	F	Significance level
Treatment	7	11.633	1.661	4.11	0.0157
Block	2	6.375	3.187	7.89	0.0065
Error	12	4.849	0.404		
Total	21	22.858			

Average variance for the set of elementary treatment contrasts is obtained as 0.2917778

We also applied Cook-statistic to identify outlying observations, if any. It was found that observation number 10 and 19 corresponding to second treatment in the second replication and third treatment in the third replication respectively are influential. In the final data set these two observations are actually deleted.

Example 4.3: An experiment with 10 treatments conducted in the randomized complete block (RCB) design with 4 replications at Sugarcane Research Institute, Shahjahanapur, Uttar Pradesh to find out the suitable herbicide to control weeds in Sugarcane. (net plot size: 8.00m × 5.40m.). The treatment details 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 below shows the data on yield per plot in kilogram in different treatments:

Table 4.7: Yield of sugarcane in kg/plot

Treatments	**Replication**			
	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

Analysis of variance was performed on the original data and results are given in Table 4.8. From the table one can observe that the treatment effects are not significant at 5% level of significance and block effects are significant at 5% level of significance.

Table 4.8: 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			

Average variance for the set of elementary treatment contrasts is obtained as 0.0335

LMS Method

We now applied LMS technique by taking subset size as n–2. The result is presented in Table 4.9. The dramatic effect to note here is that the treatment effects are now almost significant at 5% level of significance.

Table 4.9: 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			

Average variance for the set of elementary treatment contrasts is obtained as 0.0182212

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

Example 4.4: An experiment with 9 treatments was conducted in the randomized block design (RBD) with 3 replications at College of Agriculture, Nagpur To find out the suitable fungicide for control of the bacterial blight of sesamum. (net plot size: 3.30m x 2.40m.). The treatment details of the experiments are

T1 = 0.3% Copper oxy chloride
T2 = 0.25% Dithane-m-45
T3 = 0.25% thiram

T4 = Streptocycline-100 ppm
T5 = T1 + T4
T6 = T2 + T4
T7 = T3 + T4
T8 = Seed treatment with 0.3% Thiram, no foliar spray
T9 = No seed treatment,no foliar spray

The data on grain yield per plot in quintals for different treatments is given in Table 4.9:

Table 4.9: Yield of sesamum in quintal/plot

Treatments	**Replication**		
	1	**2**	**3**
1	0.06	0.02	0.04
2	0.09	0.02	0.02
3	0.01	0.02	0.02
4	0.04	0.02	0.02
5	0.02	0.03	0.08
6	0.02	0.02	0.01
7	0.04	0.04	0.02
8	0.04	0.02	0.01
9	0.02	0.01	0.01

Analysis of variance was performed on the original data and results are given in Table 4.10. From the table one can observe that the both treatment effects and block effects are not significant at 5% level of significance.

Table 4.10: Analysis of variance with the original data

Source of variation	**DF**	**SS**	**MS**	**F**	**Significance level**
Treatment	8	0.003	0.00042	1.11	0.4051
Block	2	0.001	0.00060	1.58	0.2371
Error	16	0.006	0.00038		
Total	26	0.01074			

Average variance for the set of elementary treatment contrasts is obtained as 0.0002533

LMS Estimation

For this data, we applied LMS technique by considering size of the subsets as n-1 and as well as n-2. In the following tables, the results are given:

Table 4.11: Analysis of variance (LMS with *n*-1 observations)

Source of variation	**DF**	**SS**	**MS**	**F**	**Significance level**
Treatment	8	0.0027	0.00034	1.37	0.2855
Block	2	0.0003	0.00015	0.60	0.5605
Error	15	0.0037	0.00025		
Total	25	0.0068			

Average variance for the set of elementary treatment contrasts is obtained as 0.0001725

Observation corresponding to treatment number 2 in 1st replication, that is, the observation 0.09 is deleted in the final set of observations

Table 4.12: Analysis of variance (LMS with *n*-2 observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment	8	0.00175	0.000218	2.17	0.0974
Block	2	0.00064	0.000320	3.19	0.0723
Error	14	0.001408	0.000100		
Total	24	0.003800			

Average variance for the set of elementary treatment contrasts is obtained as 0.0000714

Observations corresponding to treatment number 2 in 1st replication and treatment number 5 in 3rd replication, that is, observations 0.09 and 0.08 respectively are deleted in the final set. The dramatic change in the analyses is that the treatment effects are now significant. When we applied Cook-statistic for detecting outlying observation, if any, we found that only the first observation, *i.e*., observation corresponding to treatment number 4 in 4th replication was found to be influential but not the second observation. We also obtained the residuals from LMS fit with subset size *n*-2. Using these residuals we carry out diagnostic for outlier detection and found that the above-mentioned two observations are actually influential. This is an interesting example of masking. Previously, the second outlying observation was masked and that is why it was not detected through Cook statistic. But LMS fit removes this masking. Therefore robust methods are very good tools for detection of outlying observations.

Example 4.5: An experiment with 16 treatments was conducted in the randomized complete block (RCB) design with 3 replications at Agricultural Research Institute, Patna, Bihar to evaluate mus.rock phos coated urea for N efficiency low land rain fed paddy. The treatment details of the experiments are as follows:

T1 = Control
T2 = 29kg/ha of N as Urea (4splits)
T3 = 29kg/ha of N as n.c.u (basal)
T4 = 29kg/ha of N as gyp.c.u (basal)
T5 = 29kg/ha of N as m.r.p.c.u (basal)
T6 = 29kg/ha of N as Urea(2splits)
T7 = 58kg/ha of N as Urea(4splits)

T8 = 58kg/ha of N as Urea(2splits)
T9 = 58kg/ha of N as n.c.u (basal)
T10 = 58kg/ha of N as gyp.c.u (basal)
T11 = 58kg/ha of N as m.r.p.c.u (basal)
T12 = 87kg/ha of N as Urea (4splits)
T13 = 87kg/ha of N as Urea (2splits)
T14 = 87kg/ha of N as n.c.u (basal)
T15 = 87kg/ha of N as gyp.c.u (basal)
T16 = 87kg/ha of N as m.r.p.c.u (basal)

The data on the grain yield per plot in quintal for different treatments are given in Table 4.13.

Table 4.13: Grain yield per plot in quintal of paddy

Treatments	Replication		
	1	**2**	**3**
1	3.45	3.65	3.7
2	4.9	3.4	4.4
3	3.7	3.4	4.05
4	3.8	3.6	3.8
5	4.05	3.7	4
6	3.6	3.75	3.5
7	3.7	3.7	3.65
8	4.35	4.4	4.15
9	3.1	3.9	4.3
10	4.5	4.8	4.6
11	4.15	3.65	4.2
12	3.95	3.45	4
13	4.45	4.8	4.35
14	4.1	4.15	4.2
15	4.15	4.1	4.2
16	3.2	3.7	3.8

Analysis of variance was performed on the original data and results are given in Table 4.5. From the table one can observe that the treatment effects are significant at 5% level of significance and block effects are not significant at 5% level of significance.

Table 4.14: Analysis of variance with original data

Source of variation	**DF**	**SS**	**MS**	**F**	**Significance level**
Treatment	15	5.073	0.338	3.58	0.0014
Block	2	0.240	0.120	1.27	0.2944
Error	30	2.834	0.094		
Total	47	8.148			

Average variance for the set of elementary treatment contrasts is obtained as 0.0626667

We now applied LMS technique with *n*-1 observations and result is presented in Table 4.15

Table 4.15: Analysis of variance (LMS with *n*-1 observations)

Source of variation	DF	SS	MS	F	Significance level
Treatment	15	4.985	0.332	4.479	0.0002
Block	2	0.249	0.124	1.683	0.2031
Error	29	2.151	0.074		
Total	46	7.387			

Average variance for the set of elementary treatment contrasts is obtained as 0.0502104
Deleted observation correspond to treatment number 9 in the 1st block. Here over all conclusion is not changed much. There is also no outlying observation present in the data. Therefore, deletion of a clean observation does not affect the analysis severely.

4.5 Discussion

Least Median of Squares method is a very robust method. Its breakdown point is 50%, the highest possible value. In other words, this method can tolerate even a large number of discordant observations. This is the reason why this method is very popular in linear regression models. Rank deficiency is not a problem in regression model. But this is the main hindrance for designed experiments. This may be one of the reasons why this method could not make its presence felt in the field of design of experiments. Hopefully, present investigation would give a new direction towards its application in designed experiments. We have applied this technique to a large number of real experiments; examples of some of them 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. Masked outliers cannot be detected through regular diagnostic procedures like application of Cook-statistic. Now if the data set does not contain any outlying observation, yet LMS method deletes unnecessarily one or two clean observations in the final analysis. Is it a correct procedure? Good data analyst may argue that unnecessarily we are loosing some information on some of the observations. Certainly this is true. But we do not know, in advance, whether the data set contains any outlying observation or having nonnormal errors. Application of LMS method provides a safe guard against such abnormal occurrence. Even if the data set is completely “clean”, we may not loose much by deleting one or two “clean” observation as observed in the last example.

Like many other robust regression procedures, LMS is also a very good diagnostic tool for detection of outlying observations. Residuals obtained from a robust fit are very crucial. These residuals play an important role in regression diagnostic, particularly for outlier detection, a special branch of study “robust regression diagnostics” has been developed. LMS method not only detects outlier, but is also capable of detecting masked outlier as we have seen in the 4th example in the previous section.

The main problem of application of this technique to designed experiments is the rank deficiency of design matrix. We were in a defensive mode in choosing the size of the subset as $n-1$ or $n-2$, as we know that most of the block designs, particularly RCB designs are robust against the loss of any two observations, *i.e.*, the design remains connected even after loosing two observations. Originally the algorithm for LMS was developed for a subset of p observations, where p is the total number of parameters. But in case of designed experiments, we are considering only $n-1$ or $n-2$ observation for the subset. Definitely this size is very large. One can go on decreasing this size, but he or she has to face the problem of disconnectedness.

Intuitively, we can think of relaxing the requirement of connectedness and consider the smallest possible subset. That is, one may check all possible subset of smallest possible size and keep aside the subsets that are connected. Then applying LMS procedure to these connected subsets, one can obtain a good statistical procedure. However, this needs more insights and remains an open problem.

It is, therefore, recommended that if one 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 more statistically valid results as it guards against all possible unusual happenings.

APPENDIX 4.1

SAS Code: Robust analysis of designed experiments using LMS-estimation

```
options ls=72 ps=2500;
data ran;
input trt blk oldy;
datalines;

;
proc iml;
	use ran;
		read all into A;
			n=nrow(A);
			ncolA=ncol(A);
			v=max(A[,1]);
			intB=max(A[,2]);
			p=v+intB+1;
	x1V=j(n,v,0);
	unit=j(n,1,1);
	oldy=A[,ncolA];
/*Create Delta*/
do i=1 to n;
	x1V[i,A[i,1]]=1;
end;
/*Create the matrix for the nuisance factor*/
do j=2 to ncolA-1;
	order=max(A[,j]);
	D1=J(n,order,0);
	K=J(intB,intB,0);	/*Create the diagonal matrix of block size*/
	do i=1 to n;
		D1[i,A[i,j]]=1;
		K[A[i,2],A[i,2]]=K[A[i,2],A[i,2]]+1;
	end;
	X2b=unit||D1;
end;
oldx=x1V||x2b; /*oldx matrix v trt, mean and b blk*/
med=j(n,1,0);
betahat=j(p,n,0);
y1=j(n-1,1,0);
x1=j(n-1,p,0);
y2=oldy;
x2=oldx;
newy=j(n-1,1,0);
```

```
newx=j(n-1,p,0);
j=0;

do i=1 to n;
	y2[i,]=j(1,1,0);
	x2[i,]=j(1,p,0);
	c=0;
	do	ii=1 to n;

		c=c+1;
		if ii=i then
			do;
				if c < n then
					do;
						ii=ii+1;
						x1[c,] = x2[ii,];
						y1[c,] = y2[ii,];
					end;
			end;
		else
			do;
				x1[c,] = x2[ii,];
				y1[c,] = y2[ii,];
			end;

	end;
	betahat[,i]=ginv(x1`*x1)*x1`*y1;
	yhat=oldx*betahat[,i];
	resid=oldy-yhat;
	sqrd_res=resid##2;
	med[i,1]=median(sqrd_res);
	y2=oldy;
	x2=oldx;
end;

minres=min(med);
do i=1 to n;
	if med[i,1]=minres then

	do;
		beta_est=betahat[,i];
		deleted_obs=oldy[i,];

		c=0;
		do	ii=1 to n;
```

```
                        c=c+1;
                        if ii=i then
                                do;
                                        if c < n then
                                                do;
                                                        ii=ii+1;
                                                        newx[c,] = x2[ii,];
                                                        newy[c,] = y2[ii,];
                                                end;
                                end;
                        else
                                do;
                                                newx[c,] = x2[ii,];
                                                newy[c,] = y2[ii,];
                                end;
                end;
        end;
end;
print deleted_obs beta_est;
print newx newy;
newx1v=newx[,1:v];
newx2b=newx[,v+1:p];
newd=newx[,v+2:p];
newunit=J(n-1,1,1);
newK=J(intB,intB,0); /*Create the diagonal matrix of block size*/
newK=diag(newd`*newunit);
C=(newx1v`*newx1v)-(newx1v`*newx2b)*(ginv(newx2b`*newx2b))*(newx2b`*newx1v);
Q=newx1v`*newy-newx1v`*newx2b*(ginv(newx2b`*newx2b))*newx2b`*newy;
tauhat=ginv(c)*Q;
tss=q`*ginv(c)*q;
tms=tss/(v-1);
b=newd`*newy;
GT=newunit`*newy;
CF=GT*GT/(n-1);
bss=(b`*(inv(newk))*b)-CF;
bms=bss/(intB-1);
totss=newy`*newy-CF;
ess=totss-tss-bss;
ems=ess/(n-p+1);
FT=tms/ems;
FB=bms/ems;
pvalt=1 - probf(ft,v-1,n-p+1);
pvalb=1 - probf(fb,intB-1,n-p+1);
print totss tss bss ess;
print  ems bms tms;
print pvalt pvalb ft fb;
```

```
/*Calculation of contrast matrix and the significance of the different contrasts*/
p1=j(v*(v-1)/2,v,0);  /*p1 denotes the contrast matrix*/
cnt=0;
conout=j(v*(v-1)/2,4,0);
do i=1 to v-1;
        do j=i+1 to v;
                cnt=cnt+1;
                t=j-i;
                p1[cnt,i]=1;
                p1[cnt,j]=-1;
                conout[cnt,1]=i;
                conout[cnt,2]=j;
        end;
end;
contss=j(v*(v-1)/2,1,0);
do i=1 to nrow(p1);
p1i=p1[i,];
/*Calculation of t statistics*/
contst=((p1i*tauhat)`*ginv((p1i*ginv(c)*p1i`))*(p1i*tauhat))/ems;
conpval=1 - probf(contst,1,n-p-1);
conout[i,3]=contst;
conout[i,4]=conpval;
end;
run;
print '              trt        vs' 'trt   ' '          F Value            ' 'Pr > F            ';
print conout;
quit;
```

Chapter V

Robust Designs Against Presence of Outliers

5.1 Introduction

Till now, we have discussed detection of outliers and robust methods of analysis of experimental data. Another way of minimizing the influence of outlying observations, particularly in designed experiments, is to adopt a design that is insensitive to the presence of outlying observations. Such designs are known in the literature, as robust designs. Robustness aspects of design of experiments against presence of a single outlier were first investigated by Box and draper (1975). The designs considered by them were essentially response surface designs. Gopalan and Dey (1976) initiated this study in designs for comparative experiments. Singh *et al*. (1987) identified Row-Column designs that are robust in presence of a single outlier. The robustness criterion, they considered, was based on minimization of discrepancy in estimation of error mean squares. Later on Bhar and Gupta (2001) proposed a new criterion namely minimization of average Cook statistic for identifying robust designs that are robust against the presence of a single outlier. It is well known that Cook statistic is very useful for identifying a single outlier. Using this criterion, Bhar and Gupta (2001) shown that all the designs that are shown to be robust through minimization of discrepancy in the estimation of error variance are also robust as per of minimization of average Cook statistic. They have also shown that all E-optimal designs are robust against the presence of a single outlier. Sarker (2002) has shown that all nested balanced incomplete block designs and balanced binary block designsfor diallel crosses are robust in presence of of a single outlier. Sarker *et al*. (2005) have shown that all those block designs for diallel crosses in which a line appears a constant number of times in a block are also robust against the presence of a single outlier. They have also established the equivalence of minimization of discrepancy in the estimation of error variance and minimization of average Cook statistic. Sarker *et al*. (2005) have shown that all binary balanced treatment incomplete block designs are also robust in presence of a single outlier. In earlier chapters, through examples, it has been shown that there do exit more than one outlier in experimental data. Unfortunately no work seems to have been done on the robustness of designs against the presence of more than one outlier. Therefore, in the present chapter we have developed a criterion for identifying robust designs against the presence of two outliers. At first we give in brief the results available so far in the literature.

5.2 Some Preliminaries

Consider the usual general linear model for a block design

$$\mathbf{y} = \mu\mathbf{1} + \boldsymbol{\Delta}'\boldsymbol{\tau} + \mathbf{D}'\boldsymbol{\beta} + \mathbf{e} \qquad (5.1)$$

with E($\boldsymbol{e}$) = $\mathbf{0}$ and D($\boldsymbol{e}$) = $\sigma^2\mathbf{I}_n$.

Here $\mathbf{y}$ is $n \times 1$ vector of observations, μ is the general mean, **1** is the column vector of ones, $\mathbf{\Delta}'$ is the $n \times v$ design matrix for treatments, $\mathbf{D}'$ is the $n \times b$ design matrix for blocks, $\boldsymbol{\tau}$ and $\boldsymbol{\beta}$ are the treatment effects and block effects respectively, $\mathbf{e}$ is the vector of random error terms. Also $\mathbf{\Delta}'\mathbf{1} = \mathbf{1}$, $\mathbf{D}'\mathbf{1} = \mathbf{1}$, $\mathbf{\Delta 1} = \mathbf{r}$, $\mathbf{D1} = \mathbf{k}$; where $\mathbf{r} = (r_1, \ \ldots, \ r_v)'$ and $\mathbf{k} = (k_1, \ \ldots, \ k_b)'$ are the vectors of replications and block sizes respectively. The following relations hold.

$\mathbf{N} = \mathbf{\Delta D}' = ((n_{ij}))$, $\mathbf{N1} = \mathbf{r}$, $\mathbf{N}'\mathbf{1} = \mathbf{k}$, $\mathbf{r}'\mathbf{1} = n = \mathbf{k}'\mathbf{1}$, $\mathbf{R} = \mathbf{\Delta\Delta}' = \text{diag}\ (r_1, \ \ldots, \ r_v)$ and $\mathbf{K} = \mathbf{DD}' = \text{diag}\ (k_1, \ \ldots, \ k_b)$, where the non-negative integers n_{ij} denote the number of times ith treatment appears in jth block $\sum_j n_{ij} = r_i$, $\forall i = 1, \ldots, v$; $\sum_i n_{ij} = k_j$, $\forall j = 1, \ldots, b$.

The model (5.1) can be written as,

$$\mathbf{y} = \begin{pmatrix} \mathbf{X}_1 & \mathbf{X}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix} + \mathbf{e} \tag{5.2}$$

where $\mathbf{X}_1 = \mathbf{\Delta}'$, $\mathbf{X}_2 = \begin{pmatrix} \mathbf{1} & \mathbf{D}' \end{pmatrix}$, $\boldsymbol{\theta}_1 = \boldsymbol{\tau}$ and $\boldsymbol{\theta}_2 = \begin{pmatrix} \mu & \boldsymbol{\beta}' \end{pmatrix}'$.

Thus from (5.2) we get the reduced normal equations for estimation of the linear functions of treatment effects by eliminating the effects of μ and $\boldsymbol{\beta}$ as,

$$\mathbf{C}_{\boldsymbol{\theta}1}\boldsymbol{\theta}_1 = \mathbf{Q}_{\boldsymbol{\theta}1} \tag{5.3}$$

where the matrix $\mathbf{C}_{\boldsymbol{\tau}}$ and vectors $\mathbf{Q}_{\boldsymbol{\tau}}$ are obtained as

$$\mathbf{C}_{\boldsymbol{\theta}1} = \mathbf{X}_1'\mathbf{B}\mathbf{X}_1 = \mathbf{\Delta B \Delta}'$$

$$\mathbf{Q}_{\boldsymbol{\theta}1} = \mathbf{X}_1'\mathbf{B}\mathbf{y} = \mathbf{\Delta B y}$$

where the matrix **B**, as given in (2.4), is simplified as

$$\mathbf{B} = \mathbf{I}_n - \mathbf{D}'\mathbf{K}^{-1}\mathbf{D}. \tag{5.4}$$

Single outlier

Without any loss of generality, we assume that the observation pertaining to the first treatment in the first block is an outlier. Then the incidence matrix **N** can be written as

$$\mathbf{N} = \begin{bmatrix} 1 & \boldsymbol{\varepsilon}' \\ \mathbf{f} & \mathbf{N}_0 \end{bmatrix}$$

where f is a (v–1) component (0-1) vector of remaining (v – 1) treatments in the first block, $\boldsymbol{\varepsilon}$ is a (b–1) component (0-1) vector of incidence of the first treatment in the remaining (b – 1) blocks. Now from the definition (2.1) the Cook-statistic for the set of treatment contrasts $\mathbf{P}\boldsymbol{\tau}$ is given by,

$$D_1 = \frac{\mathbf{y}'\mathbf{V}\mathbf{u}_1(\mathbf{u}_1'\mathbf{V}\mathbf{u})^{-1}\mathbf{u}_1'\mathbf{B}\mathbf{\Delta}'\mathbf{C}_\tau^{+}\mathbf{\Delta}\mathbf{B}\mathbf{u}_1(\mathbf{u}_1'\mathbf{V}\mathbf{u})^{-1}\mathbf{u}_1\mathbf{V}\mathbf{y}}{(v-1)\hat{\sigma}^2} \tag{5.5}$$

Another form of D_1 is obtained from (2.10) as

$$D_1 = \frac{\hat{\delta}_1' \mathbf{u}_1' \mathbf{S} \mathbf{u}_1 \boldsymbol{\delta}_1}{(v-1)\hat{\sigma}^2} = \frac{s_{11}}{(v-1)\hat{\sigma}^2} \hat{\delta}_1^2 \tag{5.6}$$

where s_{11} is the first diagonal element of the matrix $\mathbf{S} = \mathbf{B\Delta' C_{\tau}^{+} \Delta B}$

again, $s_{11} = \mathbf{u}_1' \mathbf{S} \mathbf{u}_1 = \mathbf{u}_1' \mathbf{B\Delta' C}_{\tau}^{+} \mathbf{B u}_1$.

Now $\mathbf{\Delta B u}_1 = \mathbf{\Delta (I - D'K^{-1}D) u}_1$, on substituting the value of $\mathbf{B}$ from (5.4) we get

$$\mathbf{\Delta B u}_1 = (\mathbf{\Delta} - \mathbf{NK}^{-1}\mathbf{D})\mathbf{u}_1 = \frac{1}{k_1}\begin{pmatrix} k_1 - 1 \\ -\mathbf{f} \end{pmatrix}.$$

Now we define $\mathbf{U}_0 = \left\{k_1(k_1 - 1)\right\}^{-1/2}\begin{bmatrix} k_1 - 1 \\ -\mathbf{f} \end{bmatrix}$

Then $s_{11} = \dfrac{k_1 - 1}{k_1} \mathbf{U}_0' \mathbf{C}_{\tau}^{+} \mathbf{U}_0$.

Thus D_1 given in (5.6) can be written as $D_1 = \dfrac{k_1 - 1}{k_1(v-1)} \mathbf{U}_0' \mathbf{C}_{\tau}^{+} \mathbf{U}_0 \dfrac{\hat{\delta}_1^2}{\hat{\sigma}^2}$. If we denote by r_1^* and t_1 the ordinary and Studentized residuals respectively, for the outlying observation, then

$r_1^* = y_1 - \hat{y}_1 = \mathbf{u}_1' \mathbf{V y}$ and $t_1 = \dfrac{r_1^*}{\hat{\sigma}\sqrt{v_{11}}}$, where v_{11} is the first diagonal element of the matrix of

$$\mathbf{V} = \mathbf{B} - \mathbf{B\Delta' C}_{\tau}^{+} \mathbf{\Delta B}. \tag{5.7}$$

Again, $v_{11} = \mathbf{u}_1' \mathbf{V u}_1 = \mathbf{u}_1'(\mathbf{B} - \mathbf{B\Delta' C}_{\tau}^{+} \mathbf{\Delta B})\mathbf{u}_1 = \dfrac{k_1 - 1}{k_1}(1 - \mathbf{U}_0' \mathbf{C}_{\tau}^{+} \mathbf{U}_0)$.

Thus we get another form of D_1 as

$$D_1 = \frac{\mathbf{U}_0' \mathbf{C}_{\tau}^{+} \mathbf{U}_0}{1 - \mathbf{U}_0' \mathbf{C}_{\tau}^{+} \mathbf{U}_0} \frac{t_1^2}{(v-1)}. \tag{5.8}$$

5.3 Robustness Against the Presence of a Single Outlier in Designs for One-way Elimination of Heterogeneity

In this section we present the results on the robustness of designs for one-way elimination of heterogeneity against the presence of a single outlier. To begin with we define a robustness criterion and then find out designs which are robust according to this criterion.

Consider the linear model of designs for one-way elimination of heterogeneity as given in (5.1).

Robustness Criterion (Bhar and Gupta, 2001)
Cook-statistic for testing a single outlier can be obtained from (2.8) as

$$D_i = \frac{s_{ii}}{v_{ii}} \frac{t_i^2}{(v-1)} \qquad (5.9)$$

Here, we assumed that the i^{th} observation is an outlier and accordingly the statistic D_i is rewritten in (5.9). Now D_i can alternatively be written as,

$$D_i = \frac{s_{ii}}{\frac{k_j - 1}{k_j} s_{ii}} \frac{t_i^2}{(v-1)} \qquad (5.10)$$

Here we assumed that the i^{th} observation occurs in the j^{th} block and k_j is the j^{th} block size.

Observe that t_i^2 is an increasing function of s_{ii} and t_i^2, being a function of residual, is an outlier measure. Very high value of t_i^2 corresponds to an outlying observation and combining with s_{ii} as in (5.10) measures the influence of an outlier. An outlier may occur in any of the n observations. A design for which the average of Cook-statistic over all possible outliers is minimum, may be termed as robust design against the presence of a single outlier.

Average of Cook-statistic
From (5.10) we get average Cook-statistic D as,

$$\text{D} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{s_{ii}}{\frac{k_j - 1}{k_j} s_{ii}} \frac{t_i^2}{(v-1)} \right) \qquad (5.11)$$

Clearly D is a weighted sum of squares of t_i. D will be minimum when the weights $\frac{s_{ii}}{\frac{k_j - 1}{k_j} s_{ii}}$ are all equal or in other words, s_{ii} are all equal.

Thus the study of robustness of a design requires the computation of the elements $s_{ii}\,; \forall\, i = 1, 2, \ldots, n$, the diagonal elements of the matrix $\mathbf{S} = \mathbf{B\Delta' C}_\tau^+ \mathbf{\Delta B}$. Note that $\mathbf{B\Delta' 1}_n = \mathbf{0}$, therefore, $\mathbf{B\Delta'}\boldsymbol{\tau}$ represents a set of treatment contrasts. Since the dispersion matrix of $\mathbf{B\Delta'}\hat{\boldsymbol{\tau}}$ is $\sigma^2\mathbf{S}$, a design will be robust if all the components of $\mathbf{B\Delta'}\hat{\boldsymbol{\tau}}$ are estimated with the same variance. The p^{th} component of $\mathbf{B\Delta'}\boldsymbol{\tau}$ is $\tau_i - \sum_i n_{ij} \frac{\tau_i}{n_{.j}}$, where p^{th} observation pertains to the i^{th} treatment in the j^{th} block and n_{ij} is as described earlier. If the least squares estimator of

$\tau_i - \sum_i n_{ij} \frac{\tau_i}{n_{.j}}$ is p_{ij}, then $p_{ij} = \hat{\tau}_i - \sum_i n_{ij} \frac{\hat{\tau}_i}{n_{.j}}$. Thus the design will be robust if and only if Var (p_{ij}) is a constant independent of *i* and *j*. These ideas are used in characterizing robust designs. For computing Var(p_{ij}) the following result is useful.

Lemma 5.1: Let $\sum_i l_i \tau_i$ be a treatment contrast and $\sum_i l_i \tau_i = \sum_i q_i Q_i$, then $\text{Var}\left(\sum_i l_i \hat{\tau}_i\right) = \sum_i l_i q_i \hat{\sigma}^2$

(i) Designs with proportional frequencies

Designs with proportional frequencies are characterized by the property $n_{ij} = \frac{n_{i.} n_{.j}}{n}$.Then it can be shown that $\text{Var}(p_{ij})/\sigma^2 = \frac{1}{n_{i.}} - \frac{1}{n}$. Thus any design with proportional frequencies cannot be robust except in the special case where all $n_{i.}$ are equal. The robustness of randomized complete block (RCB) designs follows from this.

(ii) Balanced Binary Designs

For a binary designs n_{ij} = 0 or 1. A balanced design is one which permits estimation of all elementary contrasts among the treatment effects with equal variance. It is well known that a necessary and sufficient condition for a design to be balanced is that its **C**-matrix has all the diagonal element equal and all the off diagonal elements equal. Let the off-diagonal element of the **C**-matrix of a balanced binary block design be $-\alpha$. Then it can be shown that using Lemma, $\text{Var}(p_{ij})/\sigma^2 = \frac{n_{.j} - 1}{n_{.j} v \alpha}$, which is constant if and only if $n_{.j}$ is a constant for all *j*. Since the design is balanced binary, this implies that $n_{i.}$ must also be a constant for all *i*. Hence a balanced binary design is robust if and only if it is equi-block sized and equi-replicate, *i.e*., if and only if it is a balanced incomplete block (BIB) design.

(iii) Partially balanced incomplete block (PBIB) designs with two-associate classes

Consider a two associate class PBIB design with usual parameters v, b, r, k, λ_i, n_i, p^i_{jk} (*i, j, k* = 1, 2). Details of parameters are available in Dey(1986). Let D be the class of all 2-associate class PBIB designs satisfying the following block structure:

(A): For any treatment *i*, appearing in the *j*th block, the number of first associates of treatment *i* occurring in the same block is a constant (say *g*) independent of *i* and *j*. Now on using Lemma, it can easily be shown that for a PBIB design

$\text{Var}(p_{ij})/\sigma^2 = \{(k-1)B_2 + gA_2\}/\Delta$ which is constant, where $\Delta = A_1B_2 - A_2B_1$, $A_1 = r(k-1)+\lambda_2$, $A_2 = \lambda_2 - \lambda_1$, $B_1 = (\lambda_2 - \lambda_1)p_{12}^2$ and $B_2 = r(k1) + \lambda_2 + (\lambda_2 - \lambda_1)(p_{11}^1 - p_{12}^2)$. Thus all two associate class PBIB designs having the block structure in (A) are robust. We now present some examples of two associate class PBIB designs having the block structure in (A) in sequel.

All non-group divisible two-associate PBIB class designs with $\boldsymbol{\lambda_2 = 0}$

Since $\lambda_2 = 0$ any two treatments which are mutually second associates do not occur together in any block. Thus all the treatments appearing in any block are first associates, which gives g = k–1, since the block size i fixed. Thus any two-associate class PBIB designs with $\lambda_2 = 0$ satisfies the block structure in (A).

All semi-regular group-divisible (GD) designs

It is well known that for a semi-regular GD design, $k = cm$, where c is an integer and every block contains c treatments from each group. We know that treatments belonging to any group are first associates. Thus we get $g = c - 1$. Hence such designs satisfy the block structure in (A).

All triangular PBIB designs satisfying $r + (n-4)\boldsymbol{\lambda}_1 - (n-3)\boldsymbol{\lambda}_2 = 0$

It is well known that if in a triangular design, one of the eigen-values of $\mathbf{NN'}$ matrix is $\theta = r + (n-4)\lambda_1 - (n-3)\lambda_2 = 0$, then $2k$ is divisible by n and every block of the design contains $2k/n$ treatments from each of the n rows of the association scheme. Thus designs satisfy the block structure in (A) and clearly for such designs $g = (4k/n) - 2$.

All latin-square type designs with two-constraints (L_2) satisfying $r + (s-2)\boldsymbol{\lambda}_1 - (s-1)\boldsymbol{\lambda}_2 = 0$

L_2 type PBIB designs is based on the latin-square association scheme with $v = s^2$ treatments. Further if one of the eigen values of $\mathbf{NN'}$ matrix is $\theta = r + (s-2)\lambda_1 - (s-1)\lambda_2 = 0$, then k is divisible by s and in such case every block of the design contains k/s treatments from each of the rows (columns) of the association scheme. Thus such type of designs satisfy the block structure in (A) and clearly $g = (2k/s) - 2$.

Robustness Criterion (Gopalan and Dey, 1976))

Consider the model (5.1) , in the absence of any outlier, an unbiased estimator of σ^2 is given by

$$\hat{\sigma}^2 = \frac{R_0^2}{(n-m)}$$

where $R_0^2 = \mathbf{y'y} - \mathbf{y'X}(\mathbf{XX})^{-}\mathbf{X'y}$, $\mathbf{X} = [\mathbf{1} \quad \boldsymbol{\Delta}' \quad \mathbf{D}']$.

When the i^{th} observation has added to it a quantity c, the bias or discrepancy in estimating σ^2 through R_0^{*2} is

$$d_i = \frac{c^2 a_{ii}}{(n-m)}$$

where $R_0^2 = \mathbf{y}'^* \mathbf{y}^* - \mathbf{y}'^* \mathbf{X}(\mathbf{XX})^- \mathbf{X}'\mathbf{y}^*$

$$\mathbf{y}^* = \mathbf{y} + c\mathbf{e}_i$$

$\mathbf{e}_i$ is a column vector with the uth entry unity and rest zero, and a_{ii} is the i^{th} diagonal element of the matrix $\mathbf{A} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^- \mathbf{X}'$.

Now if it is assumed that it is equally likely that c could occur with any one of the n observations, the corresponding discrepancies are $d_1,\ d_2,\ ...,\ d_n$ with average, $\bar{d} = \sum_{i=1}^{n} d_i / n = c^2 / n$. Thus it is seen that average discrepancy is fixed for a fixed number of observations. In order that no unduly large discrepancy in the estimator of σ^2 is caused by the outlier, it is desirable that d_i, i = 1, 2, …, n are made as uniform as possible. One measure of such uniformity is reflected in the variance of d_i's, given by

$$\text{Var}(d) = c^4(h - m^2 / n) / \{n(n-m)^2\}$$

where $h = \sum_i h_{ii}^2$ and h_{ii} is the i^{th} diagonal element of $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^- \mathbf{X}'$.

For fixed m and n, minimization of var (d) implies that h should be minimized, which in turn requires that all h_{ii}'s should be equal. A design for which $h_{11} = h_{22} = ... = h_{nn}$ is called a robust design against the presence of an outlier.

Utilizing this criterion Gopalan and Dey (1976) obtained identified some designs that are robust against the presence of a single outlier. Incidentally it is observed that the designs identified as robust using Bhar and Gupta (2001)'s criterion are same as that obtained by Gopalan and Dey (1976).

5.4 Robustness Against the Presence of a Single Outlier in Designs for Two-way Elimination of Heterogeneity

In this section we study the robustness of designs for two-way elimination of heterogeneity against the presence of a single outlier. We use the same criterion as developed in section (5.3) for studying the robustness of such designs.

Robustness Criterion

The Cook-statistic for testing a single outlier in designs for two-way elimination of heterogeneity can be obtained from (2.10) as

$$D_i = \frac{h_{ii}}{b_{ii} - h_{ii}} \frac{t_i^2}{(v-1)} \qquad (5.12)$$

Here, we assumed that the i^{th} observation is an outlier and accordingly the statistic D_i is rewritten in (5.12). Where b_{ii} is the i^{th} diagonal element of **B**, h_{ii} is the i^{th} diagonal element of the matrix **H**, $\mathbf{B} = \mathbf{\Phi} - \mathbf{\Phi D}_2'(\mathbf{D}_2\mathbf{\Phi D}_2')^{-}\mathbf{\Phi}$, $\mathbf{\Phi} = \mathbf{I}_n - \mathbf{D}_1'(\mathbf{D}_1\mathbf{D}_1')^{-1}\mathbf{D}_1$, $\mathbf{D}_1'$ and $\mathbf{D}_2'$ are respectively $n \times p$ and $n \times q$ matrices of full column ranks binary design matrices for rows and columns respectively, $\mathbf{H} = \mathbf{B\Delta}'\mathbf{C}_\tau^{+}\mathbf{\Delta B}$, $\mathbf{\Delta}'$ is an $n \times v$ design matrix for treatments and $\mathbf{C}_{\boldsymbol{\theta}1} = \mathbf{\Delta B \Delta}'$.

Now average Cook-statistic D is

$$\text{D} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_{ii}}{b_{ii} - h_{ii}}\frac{t_i^2}{(v-1)}\right). \tag{5.13}$$

Thus we get the same robustness criterion as developed for designs for one-way elimination of heterogeneity, i.e., all h_{ii} should be equal and, therefore, a design will be robust if and only if all the components of $\mathbf{B\Delta}'\hat{\boldsymbol{\tau}}$ are estimated with the same variance. Now the p^{th} component of $\mathbf{B\Delta}'\hat{\boldsymbol{\tau}}$ is

$$p_{ijl} = \hat{\tau}_i - \sum_i n_{ij}\frac{\hat{\tau}_i}{n_{.j}} - \sum_i n_{il}\frac{\hat{\tau}_i}{n_{.l}} \tag{5.14}$$

where the pth observation pertains to the ith treatment in the j^{th} row and l^{th} column and $\mathbf{N}_1 = ((n_{ij})) = \mathbf{\Delta D}_1'$ be the $v \times p$ treatments vs. rows incidence matrix and $\mathbf{N}_2 = ((n_{il})) = \mathbf{\Delta D}_2'$ be the $v \times q$ treatments vs. columns matrix . Thus any design will be robust if and only if $\text{Var}(p_{ijl})$ is a constant independent of *i*, *j* and *l*.

(i) Designs with proportional frequencies

Designs with proportional frequencies are characterized by the property $n_{ij} = \frac{n_{i.} \times n_{.j}}{n}$ and $n_{ij} = \frac{n_{i.} \times n_{.j}}{n}$. Then it can be shown that $\text{Var}(p_{ijl})/\sigma^2 = \frac{1}{n_{i.}} - \frac{1}{n}$, where $n_{i.} = \sum_{j=1}^{p} n_{ij}$, $n_{.j} = \sum_{i=1}^{v} n_{ij}$ and $n_{.l} = \sum_{i=1}^{v} n_{il}$. Thus any design with proportional frequencies cannot be robust except in the special case where all $n_{i.}$ are equal. The robustness of Latin Square Designs (LSD) follows from this.

(ii) Balanced Binary row-column designs

For a binary design the elements of the matrices $\mathbf{N}_1$ and $\mathbf{N}_2$ are 0 or 1 and the row sizes are *q* each and the column sizes are *p* each. It is known that a necessary and sufficient condition for a design to be balanced is that all the diagonal elements of the **C**-matrix are equal and all off diagonal elements of **C**-matrix are equal. Let the oo-diagonal element of a balanced binary design be $-\alpha$. Further we assume that between a row and column λ treatments are common which is fixed for all *j* and *l*. Then it can be shown that

$$\text{Var}(p_{ijl})/\sigma^2 = \frac{1}{v\alpha}\left\{\left(1-\frac{1}{n_{.j}}-\frac{1}{n_{.l}}\right)^2+\lambda\left(\frac{n_{.j}+n_{.l}}{n_{.j}n_{.l}}\right)^2+\frac{n_{.j}-\lambda-1}{n_{.j}^2}+\frac{n_{.l}-\lambda-1}{n_{.l}^2}\right\}$$

which is a constant if $n_{.j}$ is a constant for all j and $n_{.i}$ is a constant for all *i*. Since the design is binary, this implies that $n_{i.}$ must also be a constant for all *i*. Thus a balanced binary design is robust if and only if it is equi-row sized, equi-column sized and equireplicate and having a constant.

5.5 Robustness Aspects of Balanced Treatment Incomplete Block Designs

Sarkar *et al*.(2003) studied robustness of designs for comparing several test treatments with a control. For studying the robustness of designs for comparing several test treatments with a control treatment the result cannot be generalized for all the *n* observations. Therefore, if the outlying observation pertain to the control treatment then we take the average value of cook statistics over n_0 possible values pertaining to control treatment and if the outlying observation comes from the test treatments then we take the average value of cook statistics over all possible ($n- n_0$) values. Therefore, we can describe the robustness criterion as in case of general block designs, *i.e*, a design is said to be robust in presence of a single outlier if all the diagonal elements of the **H** matrix corresponding to the observations where control treatment appears are equal when outlying observation comes from control treatment and likewise a design is said to be robust in presence of a single outlier if all the diagonal elements of the **H** matrix corresponding to the observations where test treatments appear are equal when outlying observation comes from test treatment. Where the matrix **H** is obtained in accordance with the **H** matrix as obtained for block designs.

It is difficult to establish the equal variance of all possible treatment contrasts, therefore, they studied the particular case of variance balanced block designs for estimation of test treatments–control treatment contrasts. A block design is said to be variance balanced for the estimation of test treatment versus control treatment contrasts if it permits the estimation of these contrasts with the same variance and covariance between any two estimated test treatment versus control treatment contrasts is also same. In proper block design set up, such designs have been termed as balanced treatment incomplete block (BTIB) designs.

Theorem 5.1. All BTIB designs that are binary with respect to test treatments are robust against the presence of a single outlier.

This result has been illustrated with the following example.

Example 5.1: Let us consider the following BTIB design with parameters v = 4, b = 4, k = 4, r = 3, r_0 = 4, λ = 2, λ_0 = 3.

Block	**Treatments**			
1	0	1	2	3
2	0	1	2	4
3	0	1	3	4
4	0	2	3	4

Below are given the observations and diagonal elements of the **S** matrix corresponding to the observations of the above mentioned design.

Observation	Diagonal element of S matrix
1	0.2045454
2	0.2651515
3	0.2651515
4	0.2651515
5	0.2045454
6	0.2651515
7	0.2651515
8	0.2651515
9	0.2045454
10	0.2651515
11	0.2651515
12	0.2651515
13	0.2045454
14	0.2651515
15	0.2651515
16	0.2651515

The example clearly shows that the diagonal elements of S matrix corresponding to the observations coming from control treatment are equal and also the diagonal elements corresponding to the test treatments are equal.

5.6 Robustness of Block Designs for Diallel Crosses against the Presence of a Single Outlier

In agricultural experiments especially in plant breeding trials, block designs for diallel crosses are extensively used. Outlier(s) is (are) likely to appear in diallel cross experiments. The difference of block designs for diallel crosses with usual block designs is that in mating designs the observations are recorded on the crosses that are made up of two distinct lines whereas the experimenter is interested in comparing the contrasts of general combining ability (*gca*) effects of the lines. Therefore, handling of outlier(s) in case of designs for diallel cross experiments is not same as that of usual block designs. Hence, there is a need to propose some statistic to detect an outlier in diallel cross experiment. Sarker *et al*. (2005) studied the robustness aspects of designs for diallel crosses. Their results are summarized below.

Theorem 5.2: All proper binary balanced block designs for diallel crosses are robust against the presence of a single outlier.

Theorem 5.3: Randomized complete block designs for complete diallel crosses and block designs for diallel crosses obtainable from Family 5 designs of Das, Dey and Dean (1998) are robust against the presence of a single outlier.

5.7 Robustness of Nested Block Designs Against the Presence of a Single Outlier

Heterogeneity in the experimental material is the most important problem to be reckoned with in the statistical designing of experiments. In general block designs, the blocking is done to control one source of variability in the experimental material. But there are some experimental situations where more sources of variations may be present and cannot be controlled by ordinary blocking alone. Nested block designs have been developed to deal with the experimental situations where one factor of variability is nested within another factor of variability.

A nested block design is defined as a design with two systems of blocks where the second system is nested within the first. To be clearer, let us consider the following experimental situations:

Experimental Situation 5.1: In animal experiments, generally littermates (animals borne in same litter) are experimental units within a block *i.e.* litters are blocks. However, animals within the same litter may differ in their initial body weight. Body weight can be taken as another blocking factor. Then we have a system of blocks nested within a block.

Experimental Situation 5.2: Consider a field experiment conducted using a block design and harvesting is done block wise. To meet the objectives of the experiment, the harvested samples are to be analyzed for their content in the laboratory by different technicians at the same time or by a technician over different periods of time. Therefore, to control the variation due to technicians or time periods, this is taken as another blocking factor.

As mentioned earlier, the experiments for the above and similar experimental situations should be conducted using a nested block design.

Sarker (2002) studied robustness aspects of nested block designs.

Definition 5.1: A connected nested block design is said to be sub-block variance balanced if and only if all the non-zero eigen values of the corresponding **C**-matrix are equal.

Sarker proved the following result.

Theorem 5.4: All proper nested binary sub block balanced designs are robust against the presence of single outlier.

5.8 Robustness of Experimental Designs Against the Presence of More Than One Outlier

Recall the Cook-statistic from (2.6) or any k outliers

$$D_1 = \frac{\mathbf{r}_2'(\mathbf{U}'\mathbf{B}\mathbf{U}-\mathbf{U}'\mathbf{S}\mathbf{U})^{-1}(\mathbf{U}'\mathbf{S}\mathbf{U})(\mathbf{U}'\mathbf{B}\mathbf{U}-\mathbf{U}'\mathbf{S}\mathbf{U})^{-1}\mathbf{r}_2}{(v-1)\hat{\sigma}^2} \tag{5.15}$$

Now by applying the spectral decomposition to $\mathbf{U'SU}$, we get $\mathbf{U'SU} = \mathbf{E'\Lambda E}$, where $0 \le \lambda_1 \le \lambda_2 \le ... \le \lambda_k$ be the eigenvalues of $\mathbf{U'SU}$, $\mathbf{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_k)$ and $\mathbf{E}$ is orthogonal matrix. Then $D_k = \dfrac{\hat{\boldsymbol{\delta}}'\mathbf{E}'\mathbf{\Lambda E}\hat{\boldsymbol{\delta}}}{(v-1)\hat{\sigma}^2}$.

Let $\dfrac{1}{\hat{\sigma}}\mathbf{E}\hat{\boldsymbol{\delta}} = \mathbf{t}*$. It can be seen that the elements of $\mathbf{t}*$ are actually Studentized residuals. Thus (6.14) can alternatively be written as

$$D_k = \frac{\mathbf{t}'*\mathbf{\Lambda t}*}{(v-1)} = \frac{1}{(v-1)}\sum_{m=1}^{k}\lambda_m t*_m^2 . \qquad (5.16)$$

<u>Robustness Criterion</u>

A design for which the average Cook-statistic over all possible set of k outliers is minimum, may be termed as robust against the presence of k outliers. The average of Cook statistic is given by

$$\text{D} = \frac{1}{(v-1)\ {}^nC_k}\sum_{*}\ \sum_{m=1}^{k}\lambda_m t*_m^2 \qquad (5.17)$$

* This summation is over all possible subsets of size k.

Here, D is a weighted sum of squares of t^* and D will be minimum when the weights, i.e., all λ's are equal. Thus to show that a design is robust against presence of k outliers, we have to show that the eigenvalues of every $k \times k$ matrix $\mathbf{U'SU}$ are same. The computation of eigenvalues of every $k \times k$ matrix $\mathbf{U'SU}$ is very difficult for any value of k. We, therefore, concentrate our study of robustness for the case of two outliers.

For $k = 2$, D becomes

$$\text{D} = \frac{1}{(v-1)\ {}^nC_2}\sum_{*}\ \sum_{m=1}^{2}\lambda_m t*_m^2$$

and therefore, average of Cook-statistic for all possible set of 2 observations (${}^nC_2$ in number) will become weighted sum of squares of t_i^2. This average will be minimum when the weights λ's are equal for all the sets. Thus the study of robustness of a design requires the computation of eigenvalues of each of the 2×2 matrices $\mathbf{U'SU}$. Then we have the following result.

Theorem 5.5: All binary proper variance balanced block designs are robust against the presence of any two outliers.

Proof: Recall from Chapter 2 that the matrix **S** is given as $\mathbf{S} = \mathbf{BX}_1'\mathbf{C}_{\theta_1}^{+}\mathbf{X}_1\mathbf{B}$. Now for a block design setup $\mathbf{B} = \mathbf{I}_n - \mathbf{D}'\mathbf{K}^{-1}\mathbf{D}$, $\mathbf{X}_1 = \mathbf{\Delta}'$. For a variance balanced block design $\mathbf{C}_{\theta_1}^{+} = \dfrac{1}{\gamma}\mathbf{I}$, where

γ is the unique eigenvalue of C-matrix. For proper designs matrix **K** would be some constant × a diagonal matrix. Thus each of the 2×2 matrices $\mathbf{U'SU}$ would be same and hence all eigenvalues of $\mathbf{U'SU}$ would be same. Hence the result.

Remark 5.1: For more than two outliers, it seems difficult to identify robust designs using average Cook-statistic. It may not be possible to derive any condition by which we can show that the eigenvalues of every $k \times k$ matrix $\mathbf{U'SU}$ are same. It is, therefore, needed to search for other robustness criteria. Criterion of Gopalan and Dey (1976) may be explored for extending for more than one outlier.

5.9 Discussion

In the present investigation, we have made an attempt to develop test statistic for detection of outlier(s) in the experimental data and to identify robust designs against the presence of a single outlier or more than one outlier. It may be noted that both the criteria of robustness are dependent on design matrix **X** alone and don't involve the observation vector **y**. As a result the value of F-statistic for studying the significance of treatment effects may get affected in the presence of outlier (s) even in case of the designs identified as robust according to the above criteria. Therefore, besides detection of outlier (s) and identification of robust designs in presence of outlier (s), it is essential to develop some estimation/analytical procedures so that inference on the parameters of interest does not change. One way to deal with such situations is to develop robust estimation procedure of estimation of treatment contrasts. These techniques are discussed in the third and fourth chapters. However, one may also think of either deleting the observation(s) identified as outlier(s) or carrying out the analysis of covariance. Bhar (1997) has shown that the reduced normal equations under the analysis of covariance model or with suspected observation(s) deleted from the model are same. Therefore, both the above alternatives are same and one may use either of them. This has been illustrated with the help of the following example:

Example 5.2: Nigam and Gupta (1979) A manurial trial with six levels of Farm Yard Manure (FYM) was carried out in a randomized complete block (RCB) design with four replications at the Central Experimental Station, Sagdividi, Junagadh with a view to study the rate of decomposition of organic matters in the soil and its synthetic capacity in soil on cotton crop. Table 5.1 below shows the data on yield per plot in kg. for different levels of FYM without and with one artificially introduced outlier shown in the parenthesis.

Table 5.1: Yield of Cotton in kilogram (kg)

Level of FYM	Replication 1	Replication 2	Replication 3	Replication 4
1	6.90	4.60	4.40	4.81
2	6.48	5.57	4.28	4.45
3	6.52	7.60(10.60)	5.30	5.30
4	6.90	6.65	6.75	7.75
5	6.00	6.18	5.50	5.50
6	7.90	7.57	6.80	6.62

* The figure in the parenthesis is an artificially introduced outlier.

The results of analysis of variance of the original data are given in the following table.

Table 5.2: ANOVA with Original Data

Source	DF	Sum of Squares	Mean Squares	F-value	Prob.>F
Replication	3	6.1212	2.0404	4.36	0.0214
Treatment	5	15.4483	3.0896	6.60	0.0020
Error	15	7.0262	0.4684		
Total	**23**	**28.5957**			

It can be seen that the treatment as well as replication effects are significantly different at 5 % level of significance.

An artificial outlier is introduced in the 9th observation. The results of the Analysis of Variance in the presence of a single outlier are given as below:

Table 5.3: ANOVA with artificially introduced outlier

Source	DF	Sum of Squares	Mean Squares	F-value	Prob.>F
Replication	3	8.8337	2.9446	2.26	0.1239
Treatment	5	17.8208	3.5642	2.73	0.0601
Error	15	19.5837	1.3056		
Total	**23**	**46.2382**			

It is clear from the analysis that when one outlying observation is there in the data, the mean square error has increased approximately 2.8 fold. One can also see that neither the treatment effects nor replication effects are significantly different at 5% level of significance. Therefore, it follows that inspite of using a robust design against the presence of a single outlier, the inference on parameters of interest may change in the presence of outlier(s) in the experimental data.

To remove the effect of outlier, we used the two approaches viz. (i) analysis of covariance (ii) deleting the outlying observation. The results obtained from these two approaches are given as below:

Table 5.4: Analysis of Covariance in the presence of a single outlier

Source	DF	Sum of Squares	Mean Squares	F-value	Prob.>F
Replication	3	5.5687	1.8562	5.31	0.0118
Treatment	5	15.7441	3.1488	9.01	0.0005
Covariate	1	14.6935	14.6935	42.07	<0.0001
Error	14	4.8902	0.3493		
Total	**23**	**46.2382**			

Table 5.5: Analysis of Variance with missing data

Source	DF	Sum of Squares	Mean Squares	F-value	Prob.>F
Blocks	3	5.5687	1.8562	5.31	0.0118
Treatment	5	15.7441	3.1488	9.01	0.0005
Error	14	4.8902	0.3493		
Total	**22**	**26.2387**			

One can observe that the results with analysis of covariance and by deleting the outlying observation are same. It can also be observed that both the treatment effects and block effects are significantly different at 5% level of significance. Therefore, these approaches may be able to take care of presence of outlier(s) in the experimental data. Therefore, one has to adopt a two pronged strategy. First use a design that is robust against the presence of outliers. It will help in minimization of average Cook-statistic and discrepancy in the estimation of error variance. Secondly use detection and handling of outliers procedures in the analysis of experimental data.

Chapter VI

Development of Software

6.1 Introduction

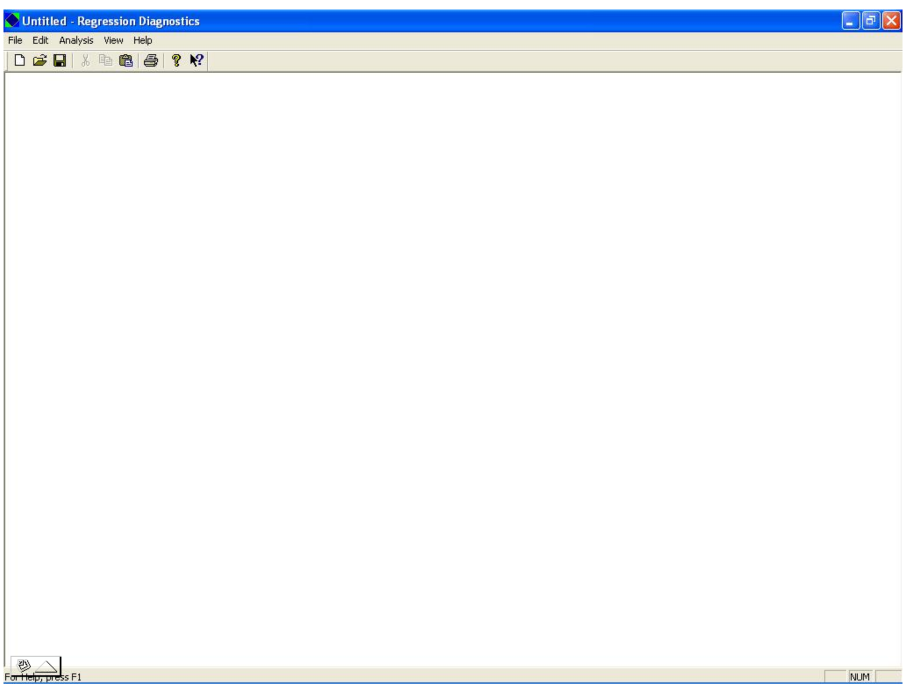
The problem of outliers in linear regression models can be handled by using several statistical packages. These statistical packages are not capable of handling outliers in designed experiments. Thus with the development of new methodologies for tackling outliers in designed experiments, a user-friendly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

The software is written in visual C++ language. The software has the following features:

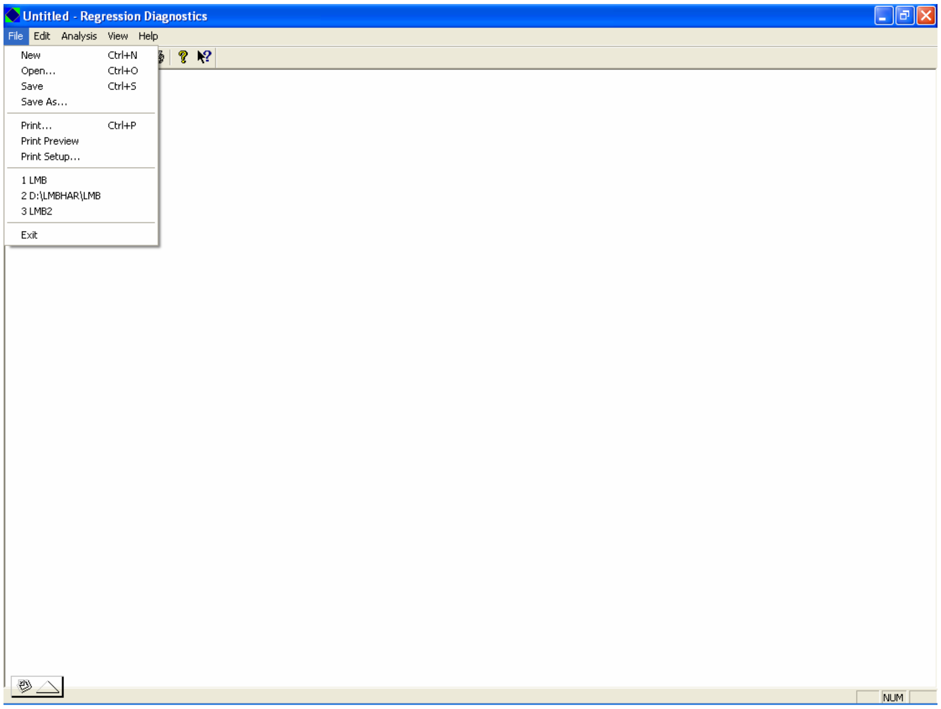
- It can identify outliers in experimental data.
- It can directly apply the robust methods of estimation for analyzing the data. Here one has two options: M-estimation (Huber's function) or LMS method.
- It has option to analyze the data after deleting the outlying observations.

6.2 Working with the software

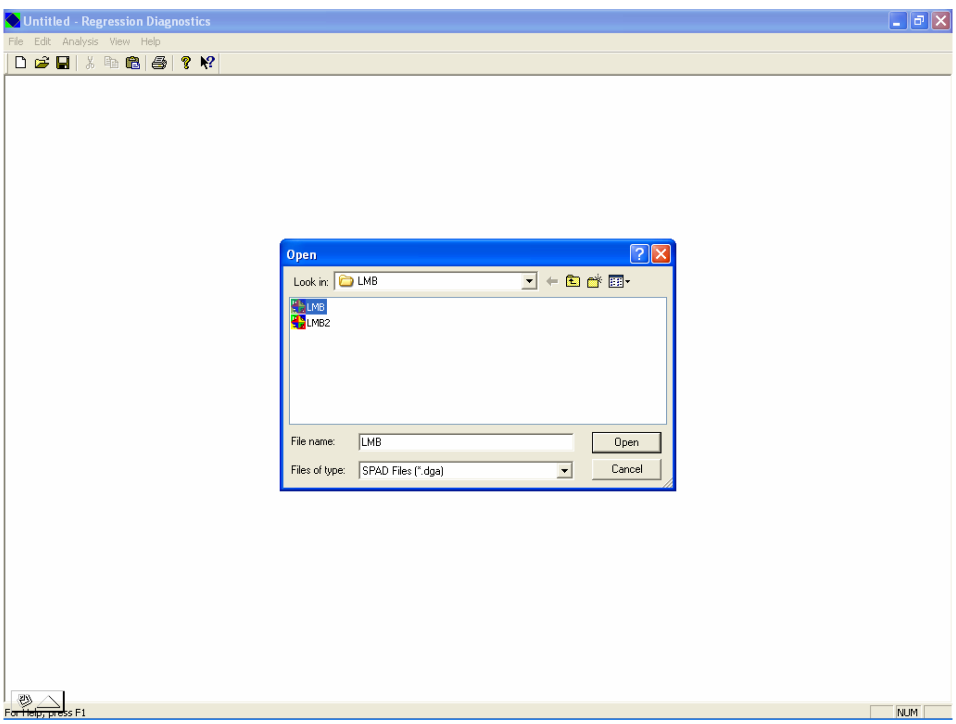
- After clicking the software icon, we will get the opening page like following:
- It has five tool bars: File, Edit, Analysis, View and Help



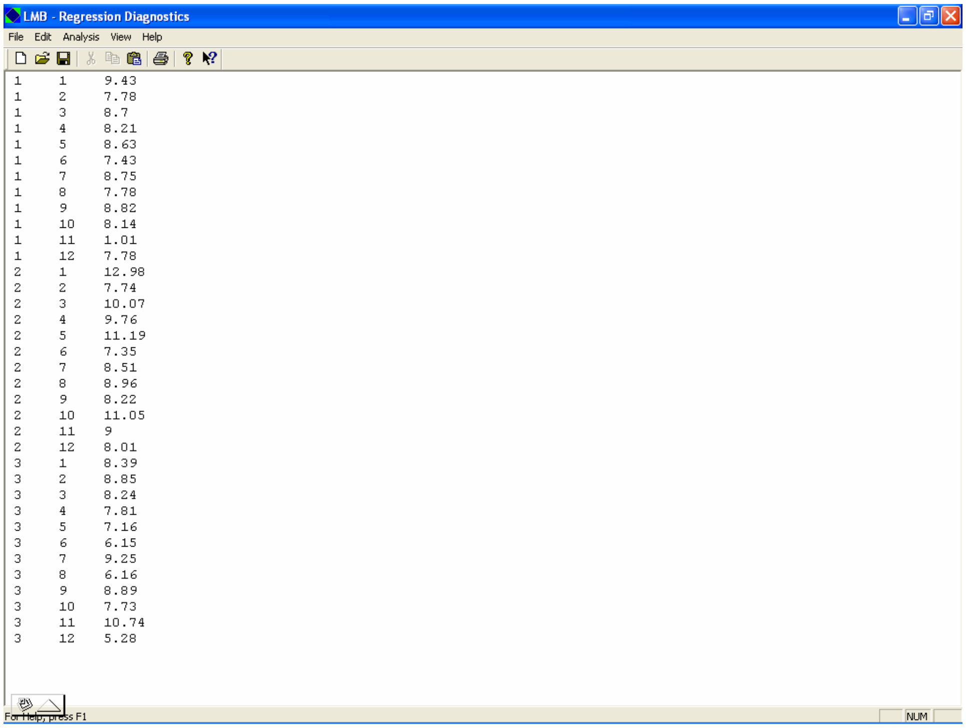
- By clicking the file menu, we will get a page as follows
- If we want to enter data from the key board, then we click New, otherwise we open an existing file by clicking Open



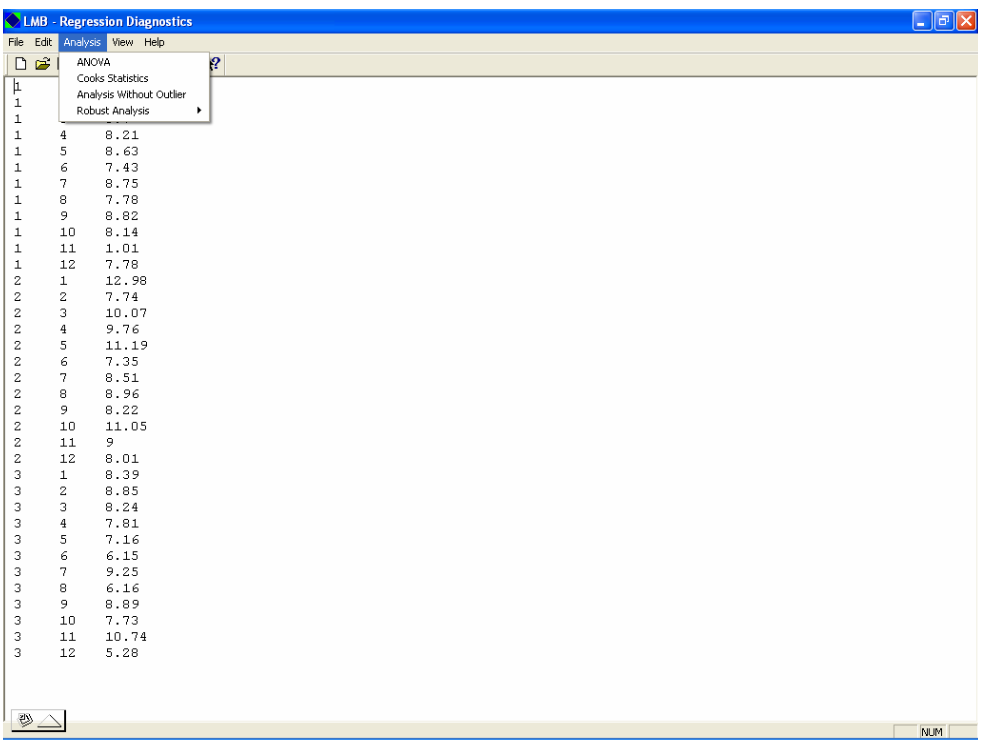
- By clicking Open, we get a page as follows:
- The page is showing two existing files.



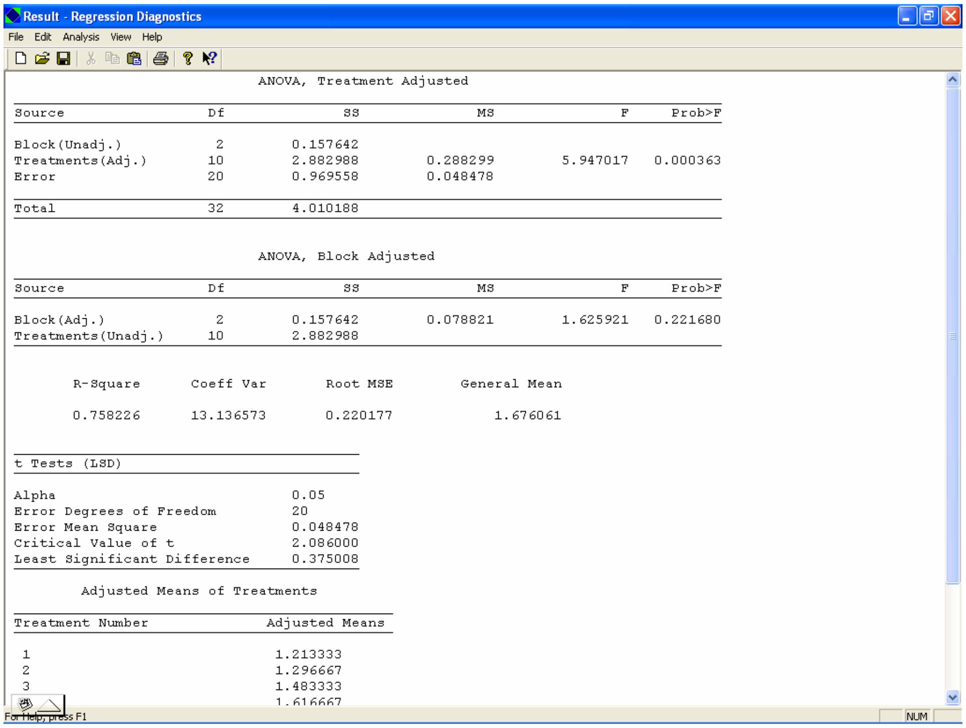
- By clicking LMB1 we get an data file as follows:
- First column is showing the replication number, second column is the treatment number and third column is the response data.
- Data has to be fed like this only.



- By clicking the analysis menu, we get a box for analysis as follows:



- By clicking ANOVA we get the result window as follows:



ANOVA, Treatment Adjusted

Source	Df	SS	MS	F	Prob>F
Block(Unadj.)	2	0.157642			
Treatments(Adj.)	10	2.882988	0.288299	5.947017	0.000363
Error	20	0.969558	0.048478		
Total	32	4.010188			

ANOVA, Block Adjusted

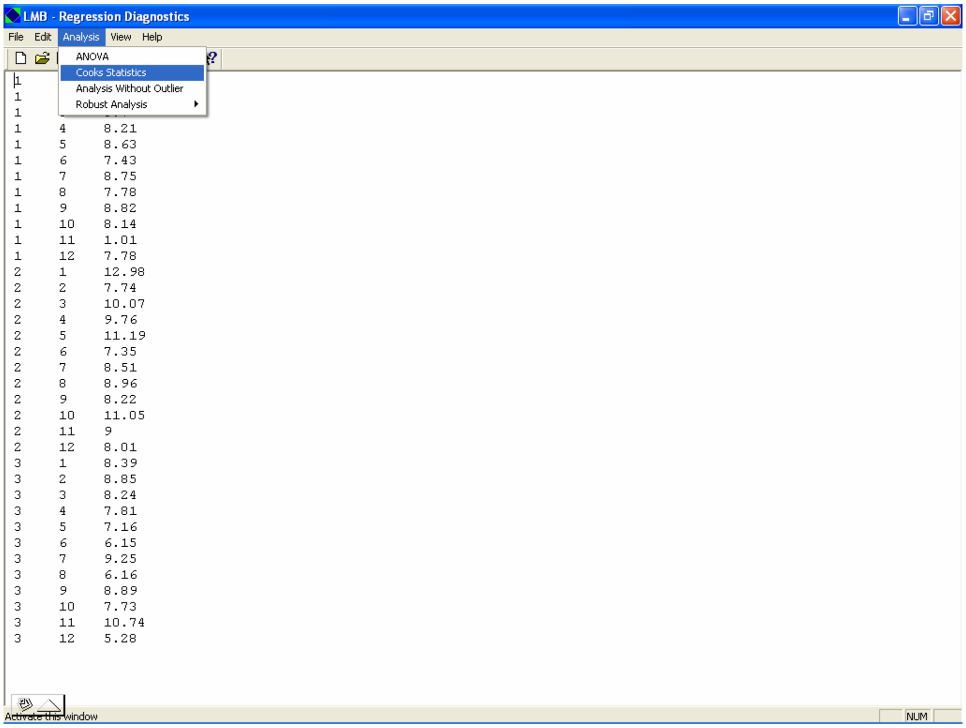
Source	Df	SS	MS	F	Prob>F
Block(Adj.)	2	0.157642	0.078821	1.625921	0.221680
Treatments(Unadj.)	10	2.882988			

R-Square	Coeff Var	Root MSE	General Mean
0.758226	13.136573	0.220177	1.676061

t Tests (LSD)	
Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.048478
Critical Value of t	2.086000
Least Significant Difference	0.375008

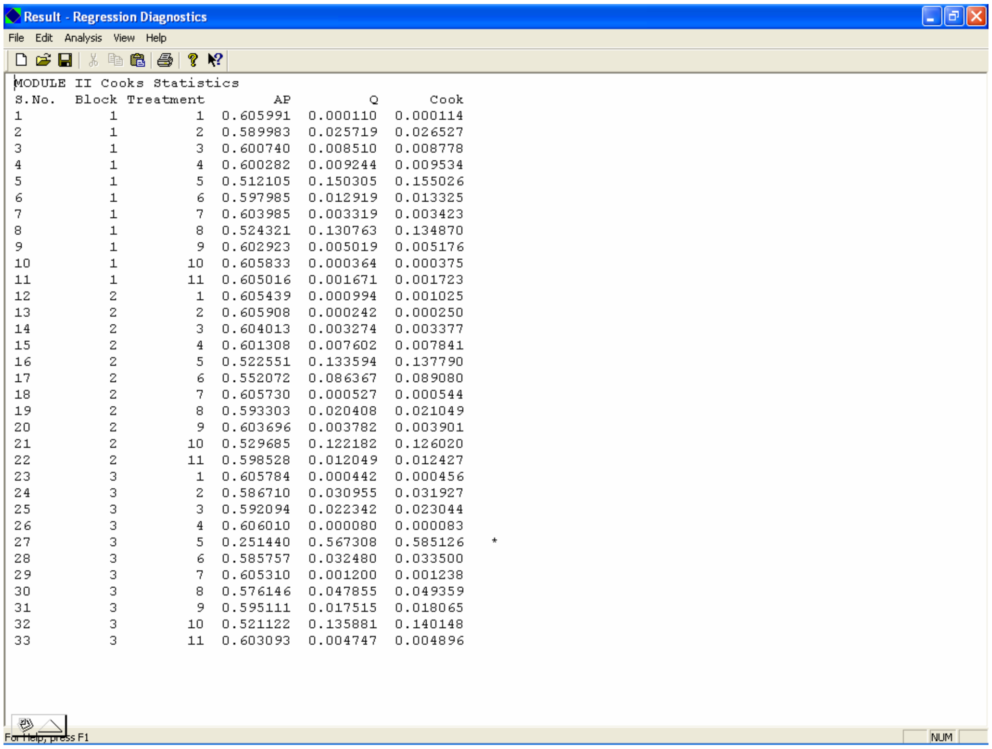
Adjusted Means of Treatments

Treatment Number	Adjusted Means
1	1.213333
2	1.296667
3	1.483333
	1.616667



1		
1		
1		
1	4	8.21
1	5	8.63
1	6	7.43
1	7	8.75
1	8	7.78
1	9	8.82
1	10	8.14
1	11	1.01
1	12	7.78
2	1	12.98
2	2	7.74
2	3	10.07
2	4	9.76
2	5	11.19
2	6	7.35
2	7	8.51
2	8	8.96
2	9	8.22
2	10	11.05
2	11	9
2	12	8.01
3	1	8.39
3	2	8.85
3	3	8.24
3	4	7.81
3	5	7.16
3	6	6.15
3	7	9.25
3	8	6.16
3	9	8.89
3	10	7.73
3	11	10.74
3	12	5.28

- By clicking Cook statistics we get Cookstatistic for each observation.
- This command also gives the AP-statistic and Q statistics for each observation.
- It also identify the outlying observation; for example it put a star corresponding the outlying observation number 27.



Result - Regression Diagnostics

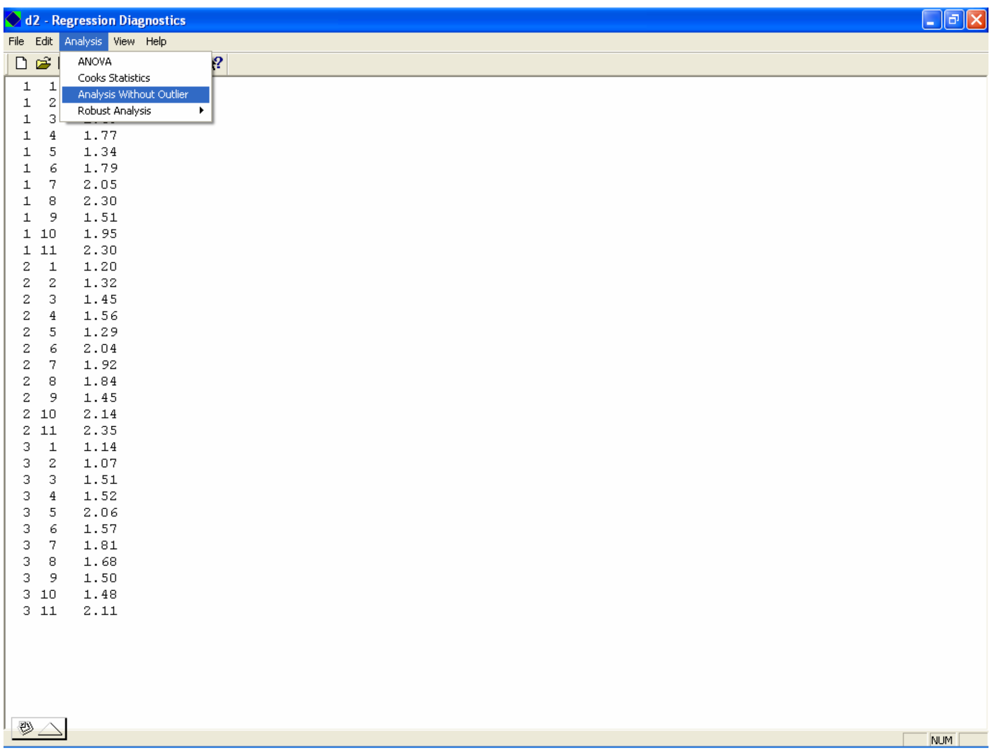
File Edit Analysis View Help

MODULE II Cooks Statistics

S.No.	Block	Treatment	AP	Q	Cook	
1	1	1	0.605991	0.000110	0.000114	
2	1	2	0.589983	0.025719	0.026527	
3	1	3	0.600740	0.008510	0.008778	
4	1	4	0.600282	0.009244	0.009534	
5	1	5	0.512105	0.150305	0.155026	
6	1	6	0.597985	0.012919	0.013325	
7	1	7	0.603985	0.003319	0.003423	
8	1	8	0.524321	0.130763	0.134870	
9	1	9	0.602923	0.005019	0.005176	
10	1	10	0.605833	0.000364	0.000375	
11	1	11	0.605016	0.001671	0.001723	
12	2	1	0.605439	0.000994	0.001025	
13	2	2	0.605908	0.000242	0.000250	
14	2	3	0.604013	0.003274	0.003377	
15	2	4	0.601308	0.007602	0.007841	
16	2	5	0.522551	0.133594	0.137790	
17	2	6	0.552072	0.086367	0.089080	
18	2	7	0.605730	0.000527	0.000544	
19	2	8	0.593303	0.020408	0.021049	
20	2	9	0.603696	0.003782	0.003901	
21	2	10	0.529685	0.122182	0.126020	
22	2	11	0.598528	0.012049	0.012427	
23	3	1	0.605784	0.000442	0.000456	
24	3	2	0.586710	0.030955	0.031927	
25	3	3	0.592094	0.022342	0.023044	
26	3	4	0.606010	0.000080	0.000083	
27	3	5	0.251440	0.567308	0.585126	*
28	3	6	0.585757	0.032480	0.033500	
29	3	7	0.605310	0.001200	0.001238	
30	3	8	0.576146	0.047855	0.049359	
31	3	9	0.595111	0.017515	0.018065	
32	3	10	0.521122	0.135881	0.140148	
33	3	11	0.603093	0.004747	0.004896	

For Help, press F1

NUM



d2 - Regression Diagnostics

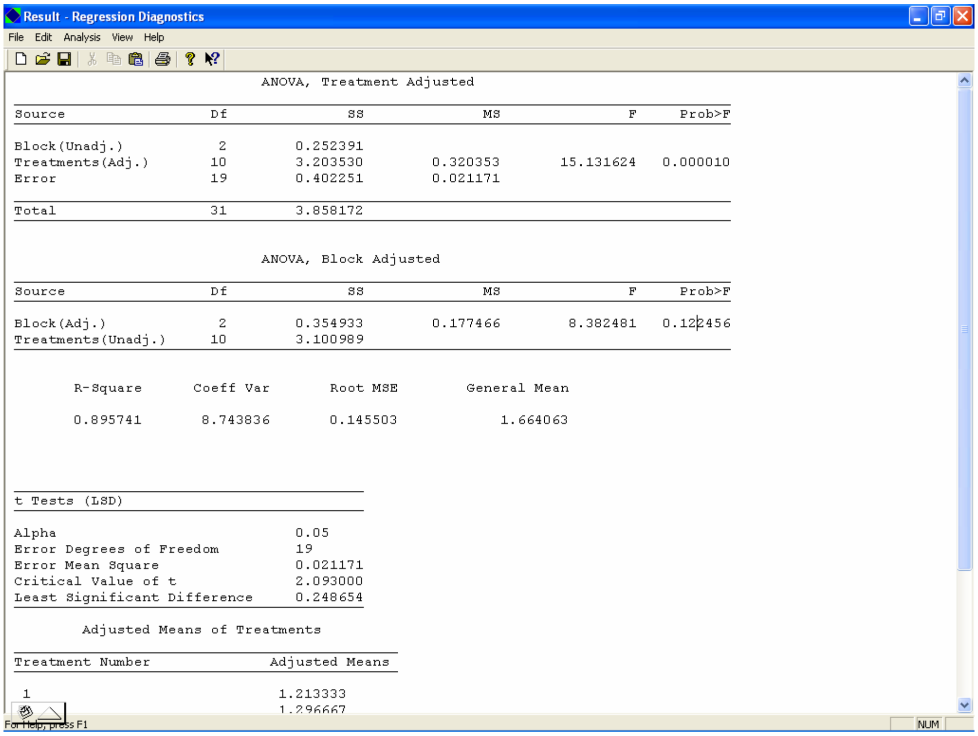
File Edit Analysis View Help

ANOVA
Cooks Statistics
Analysis Without Outlier
Robust Analysis

1	1	
1	2	
1	3	
1	4	1.77
1	5	1.34
1	6	1.79
1	7	2.05
1	8	2.30
1	9	1.51
1	10	1.95
1	11	2.30
2	1	1.20
2	2	1.32
2	3	1.45
2	4	1.56
2	5	1.29
2	6	2.04
2	7	1.92
2	8	1.84
2	9	1.45
2	10	2.14
2	11	2.35
3	1	1.14
3	2	1.07
3	3	1.51
3	4	1.52
3	5	2.06
3	6	1.57
3	7	1.81
3	8	1.68
3	9	1.50
3	10	1.48
3	11	2.11

NUM

- If we want delete this observation and reanalyze the data, we have to click the command Analysis without outlier
- It results into another ANOVA without outlying observations as follows



ANOVA, Treatment Adjusted

Source	Df	SS	MS	F	Prob>F
Block(Unadj.)	2	0.252391			
Treatments(Adj.)	10	3.203530	0.320353	15.131624	0.000010
Error	19	0.402251	0.021171		
Total	31	3.858172			

ANOVA, Block Adjusted

Source	Df	SS	MS	F	Prob>F
Block(Adj.)	2	0.354933	0.177466	8.382481	0.122456
Treatments(Unadj.)	10	3.100989			

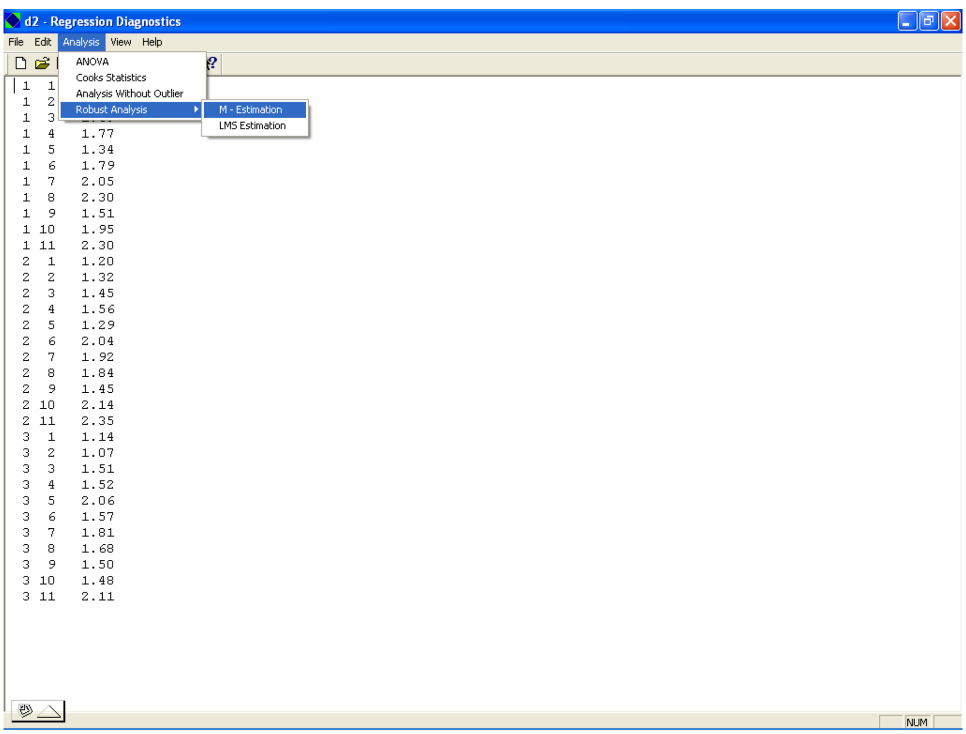
R-Square	Coeff Var	Root MSE	General Mean
0.895741	8.743836	0.145503	1.664063

t Tests (LSD)	
Alpha	0.05
Error Degrees of Freedom	19
Error Mean Square	0.021171
Critical Value of t	2.093000
Least Significant Difference	0.248654

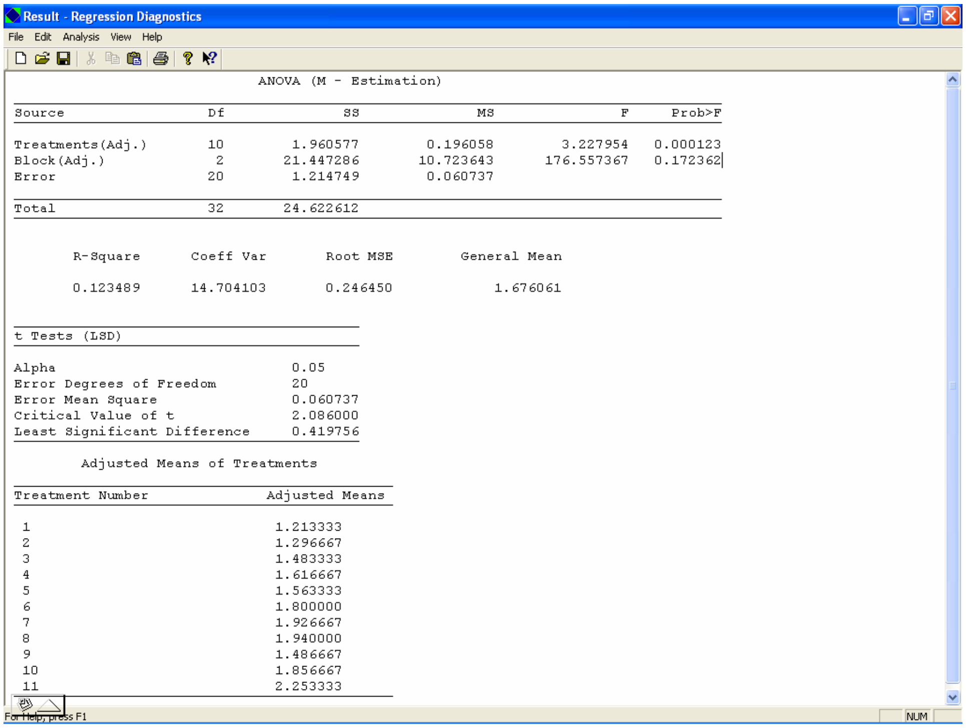
Adjusted Means of Treatments

Treatment Number	Adjusted Means
1	1.213333
	1.296667

- Finally we opt for the robust analysis of the data.
- We have to click Robust analysis and get two options, M-estimation and LMS method.



- By clicking M-estimation we get an ANOVA for M-estimation procedure as follows



ANOVA (M - Estimation)

Source	Df	SS	MS	F	Prob>F
Treatments(Adj.)	10	1.960577	0.196058	3.227954	0.000123
Block(Adj.)	2	21.447286	10.723643	176.557367	0.172362
Error	20	1.214749	0.060737		
Total	32	24.622612			

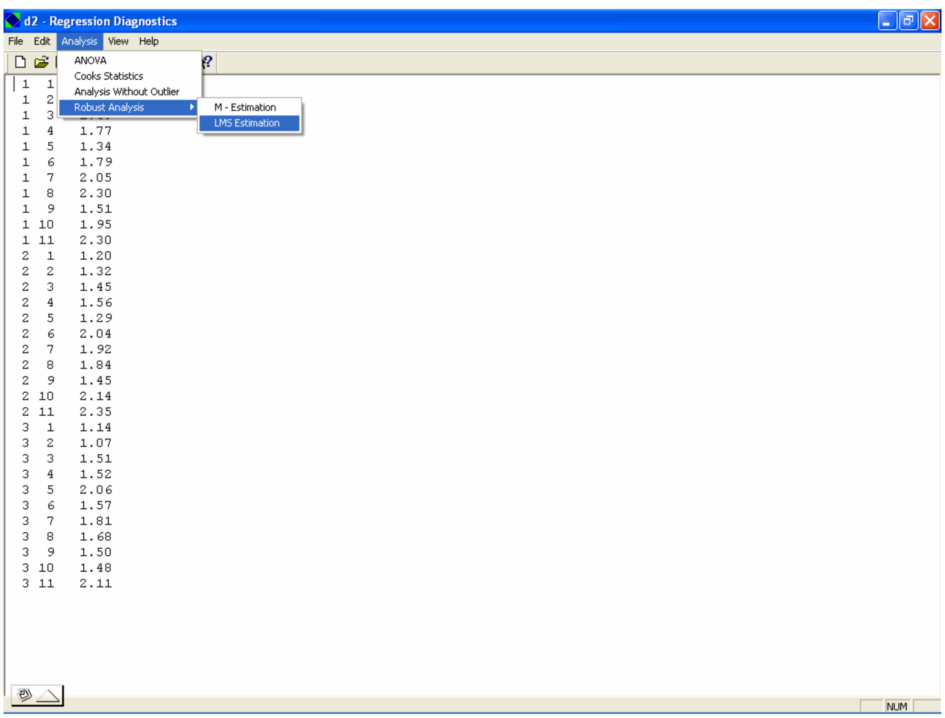
R-Square	Coeff Var	Root MSE	General Mean
0.123489	14.704103	0.246450	1.676061

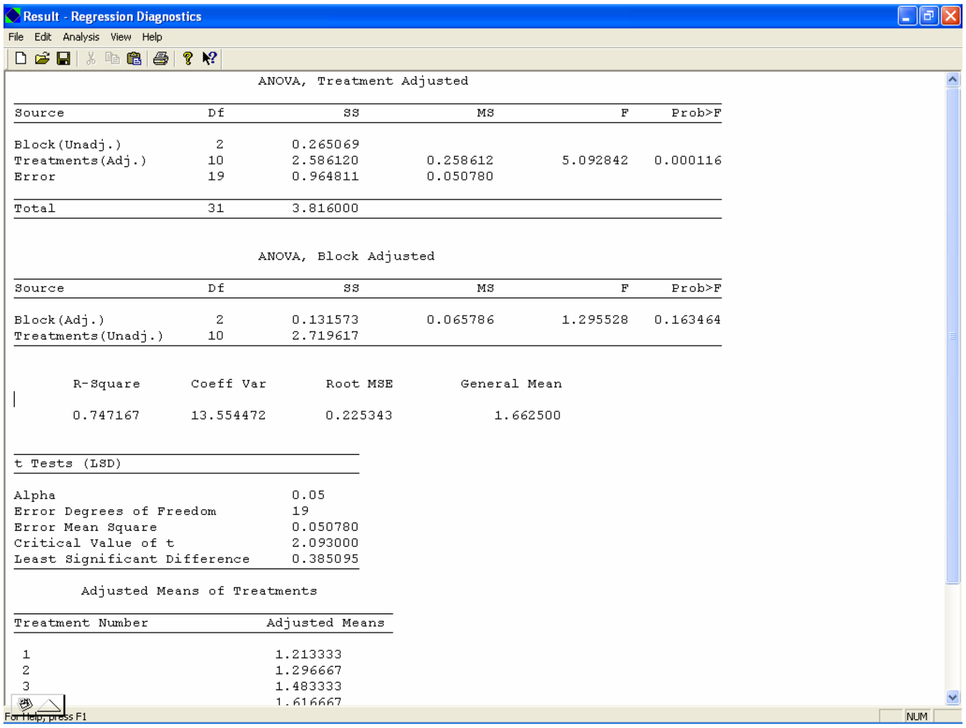
t Tests (LSD)	
Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.060737
Critical Value of t	2.086000
Least Significant Difference	0.419756

Adjusted Means of Treatments

Treatment Number	Adjusted Means
1	1.213333
2	1.296667
3	1.483333
4	1.616667
5	1.563333
6	1.800000
7	1.926667
8	1.940000
9	1.486667
10	1.856667
11	2.253333

- By clicking LMS we get The ANOVA obtained through LMS method.





Result - Regression Diagnostics

File Edit Analysis View Help

ANOVA, Treatment Adjusted

Source	Df	SS	MS	F	Prob>F
Block(Unadj.)	2	0.265069			
Treatments(Adj.)	10	2.586120	0.258612	5.092842	0.000116
Error	19	0.964811	0.050780		
Total	31	3.816000			

ANOVA, Block Adjusted

Source	Df	SS	MS	F	Prob>F
Block(Adj.)	2	0.131573	0.065786	1.295528	0.163464
Treatments(Unadj.)	10	2.719617			

R-Square	Coeff Var	Root MSE	General Mean
0.747167	13.554472	0.225343	1.662500

t Tests (LSD)	
Alpha	0.05
Error Degrees of Freedom	19
Error Mean Square	0.050780
Critical Value of t	2.093000
Least Significant Difference	0.385095

Adjusted Means of Treatments

Treatment Number	Adjusted Means
1	1.213333
2	1.296667
3	1.483333
	1.616667

For Help, press F1 NUM

Chapter VII

Dissemination Workshop

7.1 Introduction

In order to popularize the research findings of the project among the end users, a dissemination workshop was organized at IASRI on 26th July, 2007. Forty five scientists of NARS participated in the workshop. Dr. NN Goswami, former Vice-Chancellor, Chandra Sekhar Azad University of Agriculture and Technology, Kanpur was the chief guest.

7.2 Objective of the Workshop

Dissemination of research findings to the stakeholders is very important for any research project. Therefore, a work shop 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 outliers.
- To give an exposure to the design resources server.
- To formalize the recommendations emerging from this workshop.

7.3 Programme of the Workshop

The participants of the workshop included many eminent scientists actually engaged in the 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 the following salient achievements of the project were presented:

7.3.1 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) 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.

7.3.2 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 methods have been applied to the real life experimental data. The application of these methods improves the credibility of the inferences drawn.

7.3.3 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.

7.3.4 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 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.

7.4 Recommendations of the workshop

7.4.1 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!

7.4.2 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 effort 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 outlier presence. Simulation studies may be carried out to obtain the null distribution of the test statistic.

7.4.3 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.

List of Participants

1	Dr. NN Goswami	Former Vice-Chancellor, Chandra Sekhar Azad University of Agriculture and Technology, Kanpur
2	Dr. Rajendra Prasad	Ex-National Professor, Indian Council of Agricultural Research
3	Dr. Aloke Dey	Professor, Indian Statistical Institute, Delhi Centre, New Delhi
4	Dr. VK Gupta	National Professor, Indian Council of Agricultural Research
5	Dr. SD Sharma	Director, Indian Agricultural Statistics Research Institute, New Delhi
6	Dr. V.K.Sharma	Head, Div. of Design of Experiments, IASRI, New Delhi
7	Dr. V.K.Bhatia	Head, RCMU, IASRI, New Delhi
8	Dr. Ranjana Agrawal	Head, Div. of Forecasting Techniques, IASRI, New Delhi
9	Dr. Madhuban Gopal	National Fellow, IARI, New Delhi
10	Dr. Rajender Parsad	National Fellow, IASRI, New Delhi
11	Dr. L.M.Bhar	IASRI, New Delhi
12	Dr. P.K.Batra	IASRI, New Delhi
13	Dr. A.P. Mishra	National Research Centre on Rapeseed-Mustard, Bharatpur
14	Dr. R. Srivastava	IASRI, New Delhi
15	Dr. Aloke Lahiri	IASRI, New Delhi
16	Dr. Krishan Lal	IASRI, New Delhi
17	Dr. Seema Jaggi	IASRI, New Delhi
18	Sh. M.R. Vats	IASRI, New Delhi
19	Dr. Cini Verghese	IASRI, New Delhi
20	Dr. Anil Kumar	IASRI, New Delhi
21	Sh D.K.Sehgal	IASRI, New Delhi
22	Smt Rajinder Kaur	IASRI, New Delhi
23	Sh O.P.Khanduri	IASRI, New Delhi
24	Sh N. K. Sharma	IASRI, New Delhi
25	Dr. S. M. G. Saran	IASRI, New Delhi
26	Dr. A. K. Paul	IASRI, New Delhi
27	Dr. Ajit Kumar Gupta	National Research Centre for Agroforestry, Jhansi
28	Dr. A. Dhandapani	NCIPM, New Delhi
29	Dr. R.C.Agrawal	NBPGR, New Delhi
30	Dr. G. K. Jha	Div. of Economics, IARI, New Delhi
31	Sh. Amrender Kumar	IASRI, New Delhi
32	Smt Joyti Gangwani	IASRI, New Delhi
33	Sh Naresh kumar	IASRI, New Delhi
34	Sh P. K. Mitra	IASRI, New Delhi
35	Dr. A. K. Singh	IARI, New Delhi
36	Dr. Ramasubraniam, V.	IASRI, New Delhi
37	Sh S.B.Lal	IASRI, New Delhi
38	Sh Ranjit Kumar Paul	IASRI, New Delhi
39	Sh Basudev Kole	IASRI, New Delhi

40	Dr. D.K.Yadav	IARI, New Delhi
41	Sh Bishal Gurung	IASRI, New Delhi
42	Sh Sanjay Prasad	IASRI, New Delhi
44	Dr. K.K.Biswas	IARI, New Delhi
45	Sh Amit Kumar Umrao	IARI, New Delhi

Summary

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 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 square of errors is minimized. Each observation gets different weights for estimating parameters where as 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 sub set 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.

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. However, this study is so far confined to identify robust designs against presence of a single outlier.

With this view in mind the present study has been taken to investigate thoroughly the problem of outliers in designed experiments. Both detection and accommodation of outliers have been considered in the present investigation. Problem associated with outliers has been discussed with

some examples in the first chapter. A thorough review of the subject is also presented in the first chapter along with the scope of the present investigation. The practical utility of the present investigation is also discussed in this chapter.

Detection of outliers in designed experiments has been considered in the second chapter. For detecting outliers in designed experiments Bhar and Gupta (2001) provided three statistics viz., Cook-statistic, AP-statistic and Q_k-statistic. These statistics are applied to real experimental data taken from Agricultural Field Experiments Information System (AFEIS), IASRI. It has been found that many of these experiments contain outliers. Actually these experimental data were investigated for the presence of any kind of problems like non-normality or heterogeneity of error variance under a project entitled 'A diagnostic study of field experiments' conducted at IASRI. Based on the normality and homogeneity of errors, these data were grouped into several groups like non-normal and heterogeneous error variance etc. Statistics for detecting outliers were applied to these data sets. The results obtained is summarized in a table. Once outlier(s) are identified, next question may arise what to do with these outliers? One way to handle outliers is to simply discard the observations. The second way is to perform an analysis of covariance by taking one as the value of the covariate for the outlying observation and zeros for the rest of the observations. Both types of analysis for those experiments where outliers were found were carried out. Outlier detection method has been illustrated with an example.

The detection of influential subsets or multiple outliers is more difficult, owing to masking and swamping problems. Masking occurs when one outlier is not detected because of the presence of others, swamping when a non-outlier is wrongly identified owing to the effect of some hidden outliers. Pena and Yohai (1995) proposed a method to identify influential subsets by looking at the eigenvalues of an 'influence matrix'. This matrix is defined as the uncentred covariance of a set of vectors which represent the effect on the fit of the deletion of each data point. This matrix is normalized to have the univariate Cook (1979) statistics on the diagonal. This method has been modified for application in designed experiments and procedure for identifying the influential sets has been discussed. The proposed method has been illustrated with an example.

Another way to tackle the problem of outliers is to perform a robust analysis of the data. A robust procedure tries to accommodate the majority of good data points. Bad points, lying far away from the pattern formed by the good ones. Among robust procedures, M-estimation method is most widely used. In the third chapter the concept of M-estimation is introduced and then applied to designed experiments. Generally, in M-estimation an objective function (a function of errors) is minimized to obtain the parameter estimates. There are many objective functions of M-estimation for linear regression model available in the literature. Some of these objective functions are discussed in the present chapter and their applicability to designed experiments has been explored. Most of these objective functions involved some tuning constants. The efficiency of the M-estimation procedures depends upon how best these tuning constants are selected. For application to designed experiments the appropriate values of these constants have been proposed. For testing the hypotheses appropriate robust testing procedures are available in the literature. Some of these procedures have been discussed in this chapter. The existing objective functions have been modified by suitably choosing the constants. A new objective function has been proposed. The proposed function is based upon Cook-statistic and, therefore, addressed the basic requirement of design of experiments. All these functions along with the newly developed function have been illustrated with some examples.

In chapter 4 another robust method of analysis of data viz. Least Median of Squares (LMS) method has been introduced. The concept of this method as it is developed from linear regression model context is presented in this chapter. It is well known that least squares (LS) model can be distorted even by a single outlying observation. The fitted line or surface might be tipped so that it no longer passes through the bulk of the data. In least square method sum of square of errors is minimized to obtain the parameter estimates. It known that sum is not robust. In contrast to sum, in LMS method median of the square errors is minimized to obtain the parameter estimates. 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. Rousseeuw (1984) has proposed an algorithm to obtain LMS estimator. However, this algorithm cannot be applied directly in designed experiments. This method has been appropriately modified for application in designed experiments and illustrated with some examples.

There is yet another way of minimizing the influence of outlying observations, particularly in designed experiments is to adopt a design that is insensitive to the presence of outlying observations. Such designs are known in the literature, as robust designs, robust in the sense that the outlying observation does not have any impact on the estimation of parameters. Robustness of experimental designs against missing observations or any other disturbance has been studied extensively in the literature. There is a little work on robustness against outliers is available in the literature. Moreover, this study is confined to the presence of a single outlier. In the present chapter this study has been extended for more than one outlier. A general criterion for identifying robust designs against the presence of any t outliers has been developed in this chapter. However, identification robust designs using this criterion is mathematically intractable. Therefore, this criterion has been applied to identify robust designs that are robust against the presence of any two outliers. It has been found that all binary proper variance balanced block designs are robust against the presence of any two outliers.

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 required. A software has been developed for analyzing experimental data in presence of outliers. Various aspects of this software have been discussed in the 6^{th} chapter.
The report is concluded with a summary.

Under this study, a dissemination workshop was organized on 26^{th} July, 2007 at IASRI. Many renowned personalities working both in the field of statistics and field experimentations from different parts of the country have participated in this workshop. Salient achievements of the study have been discussed in the workshop. A number of recommendations and suggestions emerged from the discussion. These are presented in the 7^{th} chapter.

References

Akhtar, M. and Prescott, P. (1987). A review of response surface designs. *Pakistan J. Statist.* **B3**, 11-26.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics*, **16**, 523-531.

Andrews, D.F. and Pregibon, D. (1978). Finding the outliers that matter. *J. Roy. Statist. Soc.*, Ser B, **40**, 87 – 93.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). *Robust estimation of location*. Princton University Press, Princton, N. J.

Atkinson, A. C. (1981). Robustness, transformations and two graphical displays for outlying and influential observations in regression. *Biometrika*, **68**, 13 – 20

Atkinson, A. C. (1982). Regression diagnostics, transformations and constructed variables (with discussion). *J. Roy. Statist. Soc.*, Ser B, **44**, 1 – 35.

Atkinson, A. C. (1985). *Plots, transformations and regression : An introduction to graphical methods of diagnostic regression analysis*. Claredon press: Oxford,

Atkinson, A. C. (1986). Masking unmasked. *Biometrika*, **73**, 533-541.

Atkinson, A. C. (1994). Fast very robust methods for the detection of outliers. *J. Amer. Statist. Assoc.*, **89**, 1329-1339.

Atkinson, A. and Riani, M. (2000). *Robust diagnostic regression analysis*. 1st edition, New York: Springer.

Barnett, V. (1978). Study of outliers: purpose and model. *Appl. Statist.*, **27**, 242-250.

Barnett, V. and Lewis, T. (1984). *Outliers in statistical data*. New York: John Wiley.

Beckman, R. J. and Cook, R.D. (1983). Outlier……….s (with discussion). *Technometrics*, **25**, 119 – 163.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression diagnostics*. John Wiley, New York.

Bernoulli, D. (1977). The most probable choice between several discrepant observations and the formation therefrom of the most likely induction, in C. G. Aallen (1961), *Biometrika*, **48**, 3-13.

Ben, M. G. and Yohai, V. J. (1992). Robust analysis of variance for a randomised block design. *Comm. Statist. Theory Methods*, **21**, 1779 – 1798.

Bhar, L. (1997). *Outliers in experimental designs*. Ph.D. Thesis. I.A.R.I. New Delhi.

Bhar, L. and Gupta, V.K. (2001). A useful statistic for studying outliers in experimental designs. *Sankhyā*, **B63**, 338-350.

Bian, G. and Tiku, M. L. (1997a). Bayesian inference based on robust priors and MML estimators: Part I, Symmetric Location-scale distributions. *Statistics*, **29**, 81 – 90.

Bian, G. and Tiku, M. L. (1997b). Bayesian inference based on robust priors and MML estimators: Part II, Skew Location-scale distributions. *Statistics*, **29**, 317 – 345.

Bickel, P. J. (1976). Another look at robustness. *Scand. J. Statist.*, **3**, 145-148

Box, G. E. P. (1979). Strategy of scientific model building. In: *Robustness in statistics*, eds. R. L. Lawner and G. N. Wilkinson, New York: Academic Press.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc.*, Ser A, **143**, 383 – 430.

Box,G.E.P. and Draper,N.R. (1975). Robust designs. *Biometrika*, **62(2)**, 347-352.

Carroll, R. J. (1980). Robust methods for factorial experiments with outliers. *Applied Statistics*, **29**, 246-251.

Carroll, R. J. (1982). Two examples of transformations when there are possible outliers. *Appl. Statist.*, **31**, 149 – 152.

Chatterjee, S. and Machler, M. (1997). Robust regression: A weighted least squares approach. *Comm. Statist. Theory Methods*, **26**, 1381 – 1394.

Chauvenet, W. (1863). Method of least squares. *A mannual of spherical and practical autonomy*, **11**, 5th Ed., New York: Dover.

Cheng, T. C. (2005). Robust regression diagnostics with data transformations. *Comput. Statist. Data Anal.* **49**, 875—891

Chi, E. M. (1994). M-estimation in cross-over trials. *Biometrics*, **50**, 486-493.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19(1)**, 15-18.

Cook, R. D. (1979). Influential observations in linear regression. *J. Amer. Statist. Assoc.*, 74, 169 – 174.

Cook, R. D., Holschuh, N. and Weisberg, S. (1982). A note on alternative outlier model. *J. Roy. Statist. Soc.*, Ser B, **44**, 370 – 376.

Cook, R. D., and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman – Hall.

Cui, H., Xumin, H. and Ng, K. W. (2004). M-estimation for linear model with specially correlated errors. *Statist. Prob. Letter*. **66**, 383-393.

Daniel, C. (1960). Locating outliers in factorial experiments. *Technometrics*, **20**, 385-395.

Daniel, C. and Wood, F. S. (1980), *Fitting equations to data* (2nd ed.). John Wiley: New York.

DAS A., DEY A., and DEAN, A.M. (1998) Optimal designs for diallel cross experiment, *Probability and Statistics Letters*, **36**, 427-436.

Davies, L. and Gather, U. (1993). Identification of multiple outliers. *J. Amer. Statist. Assoc.*, **88**, 782 – 801.

de Gruttola, V., Ware, J. H. and Lous, T. A. (1987). Influence analysis of generalized least squares estimators. *J. Amer. Statist. Assoc.*, **82**, 911 – 917.

Dey, A. (1986). *Theory of Block Designs*. Wiley Eastern Limited: New Delhi

Draper, N. R. and John, J. A. (1980). Testing for three or fewer outliers in two-way tables. *Technometrics*, **22**, 9 –15.

Draper, N. R. and John, J. A. (1981). Influential observations and outliers in regresssion. *Technometrics*, **23**, 21 – 26.

Dixon, W. J. (1950). Analysis of extreme values. *Ann. Mathe. Statist.*, **21**, 488 – 506.

Ferguson, T. S. (1961). On the rejection of outliers. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. I), Berkeyleyand Los Angles: University of California Press. 253 – 287.

Gentleman, J. E. and Wilk, M. B. (1975). Detecting outliers in two-way table: 1. Statistical behaviour of residuals. *Technometrics*, **17**, 1 –14.

Giloni, A and Padberg, M (2002). Least trimmed squares regression, least median squares regression, and mathematical programming. *Math. Comput. Modelling*, **35**, 1043-1060.

Gopalan, R. and Dey, A. (1976). On robust experimental designs. *Sankhya*, **B38**, 297 - 299.

Gray, J. B. and Ling, R. F. (1984). K-clustering as a detection tool for influential subsets in regression. *Technometrics*, **26**, 197-208.

Greenberg, V. L. (1965). Robust estimation in incomplete block designs. Unpublished Dissertation, University of California, Berkley.

Grubbs, F. E. (1950). Procedures for detecting outlying observations in samples. *Technometrics*, **11**, 1-21.

Hadi, A. S. (1992). Identifying multiple ouyliers in multivariate data. *J. Roy. Statist. Soc.*, Ser B, **54**, 761 – 771.

Hadi, A. S. and Simonoff, J. S. (1993). Procedures for the identification of multiple outliers in linear models. *J. Amer. Statist. Assoc.*, **88**, 1264 – 1272.

Hadi, A. S. (1994). A modification of a method for the detection of outliers in multivariate samples. *J. Roy. Statist. Soc.*, Ser B, **56**, 393 – 396.

Hampel, F. R. (1975). Beyond location parameters: robust concepts and methods. *Proceedings of the 40th session of the ISI*, **46**, 375-391.

Hawkins, D. M.(1980). *Identification of outliers*. Chapman and Hall, London.

Herzberg, A. M. (1982). The robust designs of experiments: A review. *Sardica*, **8**, 223-228.

Hoaglin, D. C. and Welsch, R. (1978). The hat matrix in regression and anova. *Amer. Statistician*, **32**, 17-32.

Hocking, R. R. (1984). Duiscussion of Gray and Ling Paper. *Technometrics*, **26**, 321-323.

Hocking, R. R. and Pendleton, O. J. (1983). The regression dilemma. *Comm. Statist. Theory Methods*, **12**, 497 – 527.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least squares. *Comm. Statist. Theory Methods*, **6**, 813-827.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73-101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo, *Ann. Stat.*, **1**, 799-821.

Huber, P. J. (1981). *Robust statistics*. John Wiley, New York.

Huggins, R. M. A. (1993). A robust approach to the analysis of repeated measures. *Biometrics*, **49**, 715-720.

Irwin, J. O. (1925). On a criterion for the rejection of outlying observations. *Biometrika*, **17**, 238-250.

Jajo, N. K. (2005). A review of robust regression and diagnostic procedures in linear regression. *Acta Math. Appl. Sin. Engl. Ser.* **21,** 209-224

John, J. A. (1978). Outliers in factorial experiments. *Appl. Stat.*, **27**, 111 – 119.

John, J. A. and Draper, N. R. (1978). On testing for two outliers or one outlier in two-way tables. *Technometrics*, **20**, 69 – 78.

Juan, J. and Prieto, F. J. (2001). Using angles to identify concentrated multivariate outliers. *Technometrics*, **42**, 311-322.

Kianifard, F. and Swallow, W. (1990). A Monte Carlo comparison of five procedures for identifying outliers in linear regression *Comm. Statist. Theory Methods*, **19**, 1913-1938.

Krishan Lal, V. K. Gupta and Bhar, L.(2001), Robustness of designed experiments against missing data. *Journal of applied statistics*, **56**, 107-116.

Lehmann, E. L. (1964). Robust estimation in analysis of variance. *Comm. Statist. Theory Methods*, **12**, 957 – 966.

Leontitsis, A. and Pange, J. (2004). Statistical significance of the LMS regression. *Math. Comput. Simulation*, **64**, 543-547.

Marasinghe, M. G. (1985). Amultistage procedure for detecting several outliers in linear regression. *Technometrics*, **27**, 395-399.

Martin, R. J. (1992). Leverage, influence and residuals in regression models, when observations are correlated. *Commun. Statist. – Theor. Meth.*, **21**, 1183 – 1212.

Montgomery, D. C. and Peck, E. A. (1982). *Introduction to linear regression analysis*. John Wiley: New York.

Muller, C. H. (1995). Breakdown points for designed experiments. *J. Statist. Plan. Inference*, **38**, 125-140.

Murphy, R. B. (1951). *On tests for outlying observations*. Unpublished Ph. D. Thesis, Princeton University.

Nandi, P. (2007). Design and analysis of multi-response esxperiments. Unpublished Ph.D. Thesis, IARI, New Delhi.

Nigam, A. K. and Gupta, V. K. (1979). *Handbook on analysis of agricultural experiments*. IASRI, Library Avenue, New Delhi-12.

Nigam, A. K., Puri, P. D. and Gupta, V. K. (1988). *Characterizations and analysis of block designs*. Wiley Eastern limited: New Delhi

Rajender Parsad, V.K. Gupta and O.P. Khanduri (2000). Cataloguing and construction of variance balanced block designs: Computer algorithms for construction. IASRI Publication, New Delhi

Parsad, R., Gupta, V. K., Srivastava, R., Batra, P. K., Kaur, A. and Arya, P. (2004). *A diagnostic study of design and analysis of field experiments*. Technical Report, Indian Agricultural Statistics Research Institute, New Delhi.

Paul, S. R. and Fung, K. Y. (1991). A generalized extreme Studentized residual multiple – outlier detection procedure in linear regression. *Technometrics*, **33**, 339 – 348.

Pena, D. and Yohai, V.J. (1995)- The detection of influential subsets in linear regression by using an influence matrix, *J. Roy. Statist. Soc.*, B, **57**, 145-156.

Peirce, B. (1852). Criterion for the rejection of doubtful observations. *Astronomical Journal*, **2**, 161-163.

Polasek, W. (1984). Regression diagnostics for general linear regression models. *J. Amer. Statist. Assoc.*, **79**, 336 – 340.

Puterman, M. L. (1988). Leverage and influence in autocorrelated regression models. *Appl. Statist.*, **37**, 76-86. Corrigendum, **40** (1991), 237.

Ramsay, J. O. (1977). A comperative study of several robust estimates of slope, intercept, and scale in linear regression. *J. Amer. Statist. Assoc.*, **72**, 608-615.

Rosner, B. (1975). On the detection of many outliers. *Technometrics*, **17**, 221 – 227.

Rousseeuw, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.*, **79**, 871-880.

Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: John Wiley.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *J. Amer. Statist. Assoc.*, **85**, 633-651.

Sarker, S. (2002). Studies on outlier(s) in designed experiments. Unpublished Ph. D. Thesis, IARI., New Delhi

Sarker, S. and Gupta, V. K. and Parsad, R. (2003). Robust block designs for making test treatment- control treatment comparisons against the presence of an outlier. *J. Ind. Soc. Agril. Stat.* **56**(1), 7-18.

Sarker, S., Parsad, R. and Gupta, V.K. (2005). Outliers in designs for diallel crosses. *Metron*, **63** (2), 177 – 191.

Schrader, R. M. and Hettmansperger, T. P. (1979). Robust analysis of variance based upon a likelihood ratio criterion. *Biometrika*, **67**, 93-102

Schull, R. and Dunne, T. T. (1988). A unified approach to outliers in the general linear model. *Sankhya*, Ser B, **50**, 157 – 167.

Simonoff, J. S. (1988). Detecting outlying cells in two-way contingency table via back ward stepping, *Technometrics*, **30**, 339-345.

Singh, G., Gupta,S. and Singh,M. (1987). Robustness of row-column designs. *Statistics & Probability Letters*, **5**, 421-424.

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

Srivastava,R., Gupta, V.K. and Dey, A. (1990). Robustness of some designs against missing observations, *Comm. Statist. Theory Methods*., **19**, 121-126.

Stefansky, L. A. and Boos, D. D. (2002). The calculus of M-estimation. *Amer. Statist.* **56**, 29-38.

Stone, E. J. (1868). On the rejection of discordant observations. *Monthly Notices of the royal Astronomical Society*, **28**, 165-168.

Tietjen, G. L. and Moore, R. H. (1972). Some Grubbs – type statistics for the detection of several outliers. *Technometrics*, **14**, 583 – 597.

Tietjen, G. L. and Moore, R. H. (1979). Correction to "Some Grubbs – type statistics for the detection of several outliers"(Vol.14. pp. 383 – 597). *Technometrics*, **21**, 396.

Tiku, M. L., Tan, W. Y. and Balakrishnan, N. (1986). *Robust inference*. New York: Marcel Dekker.

Tippett, L. H. C. (1925). On the extreme individuals and the range of samples taken from a normal population. *Biometrika*, **17**, 364-387.

Tsou, T. S. and Cheng, K. F. (2004). Parametric robust regression analysis of contaminated data. *Comm. Statist. Theory Methods* **33**, 1887-1898.

Walsh, J. E. (1950). Some nonparametric tests of whether the largest observations of a set are too large or too small. *Ann. Math. Statist*., **21**, 583-592.

Wong, W. K., Tiku, M. L. and Bian, G. (1996). Time series models with nononormal innovations, symmetric location- scale distributions. Preprint (Seen with the courtesy of authors)

Zhang, Z. (1992). The robustness of ANOVA with respect to interactions in some orthogonal block designs. *Comm. Statist. Theory Methods* **21**, 233-240.