International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:9, No:2, 2015

Robust ANOVA: An Illustrative Study in Horticultural Crop Research

Dinesh Inamadar, R. Venugopalan, K. Padmini

***Abstract*—**An attempt has been made in the present communication to elucidate the efficacy of robust ANOVA methods to analyse horticultural field experimental data in the presence of outliers. Results obtained fortify the use of robust ANOVA methods as there was substantiate reduction in error mean square, and hence the probability of committing Type I error, as compared to the regular approach.

***Keywords*—**Outliers, robust ANOVA, horticulture, Cook distance, Type I error.

I. Introduction

DESIGN of experiment is the backbone of agricultural research experiments conducted by several researchers under research system, with a view to compare the efficacy of several well defined treatments. The data generated from these designed experiments are analyzed under certain assumptions. If any of the assumptions is violated, the conclusion drawn from the analysis may be erroneous.

Classical studies [12] revealed that the many of the past experiments conducted in different parts of India have non-normal and heterogeneous of error distribution of error variances. Apart from the problem of normality, the dataset may contain the outlying observations. A small subset (outlier) of the data can have a disproportionate influence on the estimated parameter or prediction in concern to users of regression analysis. In case of designs of experiment these small group (outlier) of data set controls the significance of treatments. Thus, the conclusion drawn from data set contaminated with outlier may be wrong. The usual way of treating these outliers is that removing it from the data set and constructing the analysis of variance but, this violates the basic principle of design of experiment that is randomization hence, concluded results are biased or inconsistent.

Since, the usual method yields biased and inconsistent results in the presence of outlier in the data set and if we remove that leads to violation of principle of design of experiment. There is need for alternative method which uses all observations of data set (including outliers) and provides unbiased and consistent results, one such method is robust analysis.

Dinesh Inamadar, Research Scholar, was with Indian Institute of Horticultural Research (ICAR), Bangalore-560089, India (e-mail: inamdardinesh74@gmail.com).

Dr. R. Venugopalan and Dr. K.Padmini are with Indian Institute of Horticultural Research (ICAR), Bangalore-560 089, India (e-mail: gopalantry@yahoo.com, veghybrid@rediffmail.com).

A. Outlier

The concept of outlier is well described in the literature on theory of regression – an outlier is an extreme observation /residuals that are larger in absolute value than others say 3 or 4 standard deviation from the means. The presence of one or more outliers is one of the causes of non-normal error terms.

In the context of designed data an outlier is defined differently. Here an outlier need not be a simple extreme value; we consider an observation as an outlier that may be responsible for the disruption of the usual pattern of the designed data [1]. Outlier is an observation whose value is not in the pattern of values produced by the rest of the data. Johnson [6] defines an outlier as an observation in a data set which appears to be inconsistent with the remainder of that set of data. Let us discuss now the effect of such outliers in the designed experiments.

II. Effect of Outliers in Experimental Designs

If some of the observations are different in some way from the bulk of the data, the overall conclusion drawn from this data set may be wrong. A number of statistics are now developed to detect outliers in a data set. Bhar and Gupta [2] developed statistics for detecting outliers in designed experiments. They modified Cook statistic for its application to design of experiments, which is a follow up work of [5].

Once an observation is detected as an outlier, the next question may arise, what to do with this outlying observation? Should we discard this observation? Deletion of observation from the existing set is not always recommended. On the other hand, robust method of estimation is advocated to dampen the effect of an outlying observation. In case of linear regression models, robust regression method is now very popular to tackle the problem of non-normal error variance and the presence of outliers. This approach is designed to employ a fitting criterion that is not as vulnerable as least squares to unusual data.

The most common general method of robust regression is M-estimation, introduced by [7]. In this method, the objective function to be minimized to get the parameter estimates is weighted according to the residual of each observation. A good number of objective functions to be minimized are proposed. Most of these functions are non-linear in nature and therefore, normal equations for solving the parameter estimates are also non-linear in parameters. Iteratively Reweighted Least Squares (IRLS) [9] methods are employed to solve these equations.

After identifying the outliers in the data set the effect of outliers on treatments is studied by deleting the outliers from

the data set and once again performing analysis of variance. But deletion of any aberrant (outlier) observation is not recommended which actually violates the principals of designs of experiment and usual information experiment is lost. Hence to overcome the violation of randomization principal of designs of experiment, we have applied robust estimation to the data set such as Huber's M-estimation and Andrew's M-estimation.

However, not much work on these powerful methods in design of experiments is available in the literature. Carroll [3] applied this technique to un-replicated factorial experiments and Chi [4] to Cross-Over Trials. But no other work seems to be available in the literature. In case of block designs recently, [11] made necessary modifications, wherever required to apply these methods to standard block designs.

If outlier is present in the data set and we use the usual least squares method of analysis the problem that occur generally is that all the observations including the outlying observations get similar 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 have used the available functions of M-estimation that are more frequently used in the regression analysis. In block designs, we are generally interested in the estimation of some functions of sub-set of parameters. This fact was kept in mind while applying this method.

In the present work, an attempt will be made employ the foregoing developments so as to study the influence of outliers in designed experiments by employing various robust estimation procedures for handling multiple outliers in designed experiments and subsequently examine the efficiency of different procedures based on measures of variability, by applying the methods to experimental data on brinjal crop.

III. Objective of the Study

The present research is conducted to study the significance treatments of classical analysis of variance and robust analysis of variance in the presence of outliers in a designed experiment.

IV. Database

The database for this study was obtained from a concluded research project on a Vegetable crop (Brinjal, CV A. Navneet) experiments at IIHR, Bangalore. Effect of various pollination methods (treatments) on yield and its attributing characters for Rabi season (year) were considered for this study.

V. Methodology

Data consists of five treatments (pollination methods) and four replications in RCBD set up. To study the above mentioned objective we have employed classical analysis of variance for original data, to identify the outlier in the data we have employed the Cook's distance measure and for the same data we applied robust analysis of variance, there after we have compared the average error variance of both the methods. In case of case robust analysis of variance we used Huber's M-estimation and Andrew's M-estimation. Detail procedure of obtaining the analysis of variance, Cook's distance and robust analysis of variance is discussed here,

A. Analysis of Variance

A general two-way analysis of variance [10] was employed to the data set considered. Significance of treatmental effects in the presence of outliers for the original data sets was assessed based on standard F-test [10].

The model for general block design is given by

$$y = \Delta'\tau + \mu 1 + D'\theta + \varepsilon \tag{1}$$

where y is n×1 vector of observation, Δ' is an n×v incidence matrix of treatments, τ a v×1 vector of treatment effects, D' is a n×b incidence matrix θ is b×1 vector vector of block effects, I is unit vector of order n×1 and ε is a n×1 vector of errors.

We now write down

$$\mathrm{X} = [\mathrm{X}_1 \quad \mathrm{X}_2] \tag{2}$$

where X1= Δ' and X2 = [1 D']

Similarly,

$$\beta = \begin{bmatrix} \tau \\ \mu \\ \theta \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \tag{3}$$

where

$$\beta_1 = [\tau] \text{ and } \beta_2 = \begin{bmatrix} \mu \\ \theta \end{bmatrix}.$$

Now following the normal equations for estimating β , the normal equations for estimating the parameters in designed experiments are given as

$$X'WX\,\beta = X'Wy \tag{4}$$

where W is weight matrix. From (2) and (3)

$$\begin{bmatrix} X_1'WX_1 & X_1'WX_2 \\ X_2'WX_1 & X_2'WX_2 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} X_1'Wy \\ X_2'Wy \end{bmatrix} \tag{5}$$

$$X_1'WX_1\beta_1 + X_1'WX_2\beta_2 = X_1'Wy \tag{6}$$

$$X_2'WX_1\beta_1 + X_2'WX_2\beta_2 = X_2'Wy \tag{7}$$

From (7) we have

$$\beta_2 = \left(X_2'WX_2\right)^{-1}\left[X_2'Wy - X_2'WX_1\beta_1\right] \tag{8}$$

Substituting (8) in (7), we get

$$X_1'WX_1\beta_1 + X_1'WX_2\left(X_2'WX_2\right)^{-1}\left[X_2'Wy - X_2'WX_1\beta_1\right] = X_1'Wy \quad (9)$$

Or

$$\left[X_1'WX_1 - X_1'WX_2\left(X_2'WX_2\right)^{-1}X_2'WX_1\right]\beta_1 = X_1'Wy - X_1'WX_2\left(X_2'WX_2\right)^{-1}X_2'Wy \quad (10)$$

The above equation is written as

$$C\beta_1 = Q \quad (11)$$

where

$$C = \left[X_1'WX_1 - X_1'WX_2\left(X_2'WX_2\right)^{-1}X_2'WX_1\right]$$
$$Q = X_1'Wy - X_1'WX_2\left(X_2'WX_2\right)^{-1}X_2'Wy$$

Once the C matrix is calculated then the remaining analysis is as usual.

Here simple weighted least squares analysis is done. The final configuration of the weights is treated as fixed and given a priori, a least squares weighted analysis of variance is done. This is a reasonable procedure with small sample sizes.

B. Cook's Distance

Cook statistic [5] is a distance measure, which indicates the influence of ith data point on the estimation of parameter vector. In measuring influence, it is desirable to consider both location of points in x-space and the response variable. Cook's method of measuring influence using a measure of the squared distance between the least-squares estimate based on all n points $\hat{\beta}$ and the estimates obtained by deleting the i[th] point, say $\hat{\beta}_{(i)}$. The distance measure, expressed in a general form as follows:

Consider the general linear model

$$y = X\beta + \varepsilon \quad (12)$$
$$E(e) = 0 \quad (13)$$
$$D(e) = \sigma^2 I_n, \sigma^2 > 0 \quad (14)$$

where y is an $n\times 1$ vector of observations, X is an $n\times p$ full rank matrix of known constants, θ is a $p\times 1$ vector of unknown parameters, and e is an $n\times 1$ vector of independent random variable each with zero mean and variance $\sigma^2>0$.

To determine the degree of influence, the i[th] data point has on the estimate θ a natural first would be to compute the least squares estimate of θ with point deleted. Accordingly, let $\hat{\theta}_{(i)}$ denote the least squares estimate of θ with the i[th] point deleted. An easily interpretable measure of distance $\hat{\theta}_{(i)}$ from $\hat{\theta}$ is given by

$$D_i = \frac{\left(\hat{\theta} - \hat{\theta}_{(i)}\right)'\left[D\left(\hat{\theta}\right)\right]^{-1}\left(\hat{\theta} - \hat{\theta}_{(i)}\right)}{RankD\left(\hat{\theta}\right)} \quad (15)$$

The statistics provides a measure of distance between θ and $\hat{\theta}$.

C. Robust Regression

The classical analysis of variance (ANOVA) technique is based on the principle of least squares which assumes that the underlying experimental errors are normally distributed. However, data often contain outliers due to recording or other errors. In other cases, extreme responses occur when control variables in the experiments are set to extremes. It is important to distinguish these extreme points and determine whether they are outliers or important extreme cases.

D. M-estimator

M-estimation in the context of regression was first introduced by [8] as a result of making the least squares approach robust. Although M estimators are not robust with respect to leverage points, they are popular in applications where leverage points are not an issue.

If we assume linearity, homoscedasticity, and uncorrelated errors, the maximum likelihood estimator of β is simply the OLS estimator found by minimizing the sum of squares function.

$$Min\sum_{i=1}^{n}\left(y_i - \sum x_{ij}\beta_j\right)^2 = Min\sum_{i=1}^{n}(e_i)^2 \quad (16)$$

Following from M-estimation of location, instead of minimizing the sum of squared residuals, a robust regression M-estimator minimizes the sum of a less rapidly increasing function of the residuals

$$Min\sum_{i=1}^{n}\rho\left(y_i - \sum x_{ij}\beta_j\right) = Min\sum_{i=1}^{n}\rho(e_i) \quad (17)$$

The solution is not scale equivariant, and thus the residuals must be standardized by a robust estimate of their scale $\hat{\sigma}_e$ which is estimated simultaneously. As in the case of M-estimates of location, the median absolute deviation (MAD) is often used. Taking the derivative of (17) and solving produces the score function

$$\sum_{i=1}^{n}\Psi\left(y_i - \sum x_{ij}\beta_j/\hat{\sigma}\right)x_{ik} = \sum_{i=1}^{n}\Psi\left(e_i/\hat{\sigma}_e\right)x_i = 0 \quad (18)$$

where

$$\hat{\sigma} = median\left|e_i - median(e_i)\right|/0.6745$$

with $\Psi = \rho'$. There is now a system of k+1 equations, for

which Ψ is replaced by appropriate weights that decrease as the size of the residual increases

$$\sum_{i=1}^{n} w_i\left(e_i/\hat{\sigma}_e\right)x_i = \sum_{i=1}^{n} \frac{x_{ij}\{\psi[(y_i - x_i'\beta)/s]/(y_i - x_i'\beta)/s\}/(y_i - x_i'\beta)}{s} = 0 \qquad (19)$$

j=0.1,…..,k

As

$$\sum_{i=1}^{n} x_{ij} w_{i0}\left(y_i - x_i'\beta\right) = 0$$

j=0.1,…..,k

where

$$w_{i0} = \begin{cases} \dfrac{\psi\left[\left(y_i - x_i'\hat{\beta}_0\right)/s\right]}{\left(y_i - x_i'\hat{\beta}_0\right)/s} & \text{if } y_i \neq x_i'\hat{\beta}_0 \\ 1 & \text{if } y_i = x_i'\hat{\beta}_0 \end{cases}$$

Hence by matrix notation

$$X' W_0 X\beta = X' W_0 y$$

where W_0 is n × n diagonal matrix of weights then one step estimator is –

$$\hat{\beta} = (X' W_0 X)^{-1} X' W_0 y$$

Here $\rho(z)$ is the function of residual, $\psi(z)$ is the derivative of $\rho(z)$ and w (z) is the weight function.

TABLE I
ROBUST CRITERION FUNCTIONS

Criterion	$\rho(z)$	$\psi(z)$	$w(z)$	Range
Least squares	½ z^2	Z	1.0	$\lvert z\rvert < \infty$
Huber's function	½ z^2	Z	1.0	$\lvert z\rvert \le t$
	$\lvert z\rvert t - 1/2t^2$	t sign(z)	$t/\lvert z\rvert$	$\lvert z\rvert > t$
Andrew's function	a[1-Cos(z/a)]	Sin(z/a)	Sin(z/a)/(z/a)	$\lvert z\rvert \le a\pi$
	2a	0	0	$\lvert z\rvert > a\pi$

VI. RESULTS

A. Seed Weight per Fruit

The result for classical analysis of variance for the character seed weight per fruit is indicates that the treatments are non-significant at 5 per cent (p=0.0722) with 5.57 as its error mean sum of square (Table II). The Cook's distance measure identified three observations as the outliers and they are fourth treatment of first replication and first and fourth treatments of third replication as outliers with distance 0.3941, 0.3262 and 0.6611 respectively (Table III).

TABLE II
CLASSICAL ANALYSIS OF VARIANCE FOR THE CHARACTER SEED WEIGHT PER FRUIT

Source	DF	Sum of Square	Mean Square	F –Value	p-value
Replication	3	2.27	0.76	0.14	0.9368
Treatment	4	63.17	15.79	2.84	**0.0722**
Error	12	66.79	5.57		
Total	19	132.23			

TABLE III
OUTLIERS IDENTIFIED FOR THE CHARACTER SEED WEIGHT PER FRUIT

Character	Replication	Treatment	Value	Cook Distance
Seed weight	1	4	0.88	0.3941
	3	1	0.48	0.3262
	3	4	10.65	0.6611

The Huber's M-estimator for the character seed yield per fruit is indicates that the treatments are significant at 5 per cent (p=0.0243) (Table IV), 38.64 per cent reduction in error mean sum square and 66.34 per cent decrease probability of committing type 1 error as compared to classical analysis of variance.

TABLE IV
ANOVA FOR HUBER'S M-ESTIMATOR FOR THE CHARACTER SEED WEIGHT PER FRUIT

Source	DF	Sum of Square	Mean Square	F -Value	p-value
Replication	3	0.37	0.12	0.0361	0.9903
Treatment	4	56.79	14.19	4.1581	**0.0243**
Error	12	40.97	3.41		
Total	19	98.14			

Average error variance =**1.75**

The Andrew's M-estimator for the character seed yield per fruit is indicates that the treatments significant at 5 per cent (p=0.0101) (Table V), 56.55 per cent reduction in error mean sum of square and 86.01 per cent decrease probability of committing type 1 error as compared to classical analysis of variance.

TABLE V
ANOVA FOR ANDREW'S M-ESTIMATOR FOR THE CHARACTER SEED WEIGHT PER FRUIT

Source	DF	Sum of Square	Mean Square	F –Value	p-value
Replication	3	7.36	2.45	1.0148	0.4201
Treatment	4	52.24	13.06	5.4005	**0.0101**
Error	12	29.02	2.42		
Total	19	88.62			

Average error variance=**1.40**

B. Seed Yield per Hour of Crossing

The classical analysis of variance for the character seed yield per hour of crossing is indicates that the treatments are non significant at 5 per cent (p=0.2898) with 4068.46 as its mean error sum of square (Table VI). The Cook's distance measure indicates that the four observations as outliers and they are first and fourth treatments of first and third replication with distance measure 0.2629, 0.34, 0.2385 and 0.6137 respectively (Table VII).

TABLE VI
CLASSICAL ANALYSIS OF VARIANCE FOR THE CHARACTER SEED YIELD PER HOUR OF CROSSING

Source	DF	Sum of Square	Mean Square	F –Value	p-value
Replication	3	5898.46	1966.15	0.48	0.7001
Treatment	4	22916.31	5729.08	1.41	**0.2898**
Error	12	48821.55	4068.46		
Total	19	77636.32			

TABLE VII
OUTLIERS IDENTIFIED FOR THE CHARACTER SEED YIELD PER HOUR OF CROSSING

Character	Replication	Treatment	Value	Cook Distance
Seed yield per hour crossing	1	1	160.74	0.2629
	1	4	26.4	0.3400
	3	1	14.69	0.2385
	3	4	285.6	0.6137

The Huber's M-estimator for the character seed yield per hour of crossing is indicates that the treatments are significant at p= 0.056 with 64.84 per cent reduction in error mean sum of square and decrease in probability of committing type 1 error from 0.2898 to 0.056 as compared to that of classical analysis of variance (Table VIII).

TABLE VIII
ANOVA FOR HUBER'S M-ESTIMATOR FOR THE CHARACTER SEED YIELD PER HOUR OF CROSSING

Source	DF	Sum of Square	Mean Square	F –Value	p-value
Replication	3	601.17	200.39	0.1401	0.934
Treatment	4	17878.58	4469.6	3.1254	**0.056**
Error	12	17161.02	1430.1		
Total	19	35640.77			

Average error variance=**847.16**

The Andrew's M-estimation for the character seed yield per hour of crossing is indicates that the treatments are significant at p=0.0612 with 66.66 per cent decrease in error mean sum of square and decrease in probability of committing type 1 error from 0.2898 to 0.0612 as compared to that classical analysis of variance (Table IX).

TABLE IX
ANOVA FOR ANDREW'S M-ESTIMATOR FOR THE CHARACTER SEED YIELD PER HOUR OF CROSSING

Source	DF	Sum of Square	Mean Square	F- Value	P-value
Replication	3	1393.63	464.54	0.3424	0.7951
Treatment	4	16403.75	4100.94	3.0234	**0.0612**
Error	12	16276.64	1356.38		
Total	19	34074.03			

Average error variance =**811.13**

VII. CONCLUSION

Robust analysis of variance is the technique which uses all the observations by attaching appropriate weights to the outlying observation. The efficiency of two different robust M-estimation procedure is obtained by comparing the average error variance, in case of above two examples Andrew's M-estimation is efficient as compared to that of Huber's M-estimation. The advantages of these methods are demonstrated in a horticultural crop experiment. .

REFERENCES

[1] V. Barnnet, and T. Lewis, Outliers in statistical data, Wiley, New York, 3rd edition. 1984.
[2] L. Bhar and V. K Gupta, "A useful statistic for studying outliers in experimental designs", Sankhya, B63 (2001), pp. 338-350.
[3] R. J. Carroll, "Robust methods for factorial experiments with outliers", Appl. Stat., 29(1980), pp. 246-251.
[4] E. M. Chi, "M-estimation in cross- over trials. Biometrics",50(1994), pp. 486-493.
[5] R.D. Cook," Detection of influential observation in linear regression", Technometrics, 19(1977), pp. 15-18.
[6] R. Johnson, Applied Multivariate Statistical Analysis. Prentice Hall. 1992.
[7] P.J. Huber, "Robust estimation of location parameter",. Ann. Math. Statist.35(1964), pp.73-101.
[8] P.J. Huber, "Robust regression: Asymptotic, conjectures, and Monte carlo". Ann. Stat., 1(1973), pp. 799-821
[9] P. W. Holland and R. E. Welsch, "Robust regression using iterative reweighted least-squares. Communications in statistics- Theoryand Methods", A6(1977), pp. 813-827.
[10] D. C. Montgomery, E. A. Peck, and G. G. Vining, Introduction to Linear Regression Analysis, 3rd edition, John Wiley and Sons, Inc, New York. 2001.
[11] R.K. Paul and L.M. Bhar, "M-estimation in block design", Journal of Indian Society of agril. Stat., 65 (3) (2011), pp. 323-330.
[12] R. Parsad, V.K. Gupta, R. Srivastava, P. K. Batra, A. Kaur and P. Arya, "A diagnostic study of design and analysis of field experiments", Technical Report, IASRI, New Delhi. 2004.