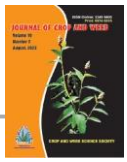
Journal of Crop and Weed, 19(2): 57-62 (2023)

ISSN- O : 2349 9400 ; P : 0974 6315



https://www.cwss.in
www.cropandweed.com

Exploration and consequences of outlier on potato field crop through factorial experiment

MD M. H. MIDDYA, M. ROY, A. SAHA, J. S. BASAK AND *A. MAJUMDER

Department of Agricultural Statistics, Bidhan Chandra Krishi Vishwavidyalaya
Mohanpur-741252, Nadia, West Bengal, India

Received : 29.01.2023 ; Revised : 18.04.2023 ; Accepted : 24.04.2023

DOI : https://doi.org/10.22271/09746315.2023.v19.i2.1703

ABSTRACT

Unwanted observations in a field experiment can lead to inaccurate outcomes and violate normality assumptions of ANOVA. Even one outlier can disrupt multi-factor analyses and yield with faulty conclusions. The present study has explored the outlier observation (observations) in factorial experiments using cook statistics. Mean shift model has been taken into consideration for constructing the test statistics. Each erroneous observation's mean differs from the rest of the observations' means in the mean shift model. Himadri (2013) has also used another test statistics namely Q_t statistics for identifying the outlying observation (or observations) in field experiment. The above test statistics are mainly used for identifying a single or multiple outliers among the set of data. These test statistics have been materialised on real experimental observation. For examining the validity of a data through cook statistics or Qt statistics, a single observation has been taken into account. After identifying the outlying observation, this observation should be deleted. Then analysis was accomplished by replacing the outlier with their estimated missing value through method of least square technique. Remarkable differences have been noticed by executing the analysis with outlying observation and without outlying observation.

***Keywords*:** Cook statistics, missing plot analysis, outlier, Q_t statistics,

Potato (*Solanum tuberosum*) is an important vegetable crop for our day-to-day life. The potato crop has been originated from Andes and Chilean highlands of South America thousand years ago. This crop was introduced in India by the Portuguese during early 17th century and later it was disseminated to north India by the British. The potato crop pertains under Solanaceae family, genus *Solanum* and species *tuberosum.* Potato contains different food component like, carbohydrate, protein, fat, and different type of nutrient and fibre. In 100g edible portion of potato mainly contains approximately 74.7g moisture,1.6g protein,0.1g fat,0.4g fibre, 22.6g carbohydrate,0.6g mineral, 0.7mg iron and 17mg Vitamin C reported by Choudhury (1996). It has great importance as a vegetable crop in our country as it contains so many food components with desired amount and different essential component for maintaining our health condition. It also considers as an economical food. Potatoes are used in various industries in India to make starch, alcohol dextrin and glucose. Different popular dried food product such as potato chips, sliced, chopped potatoes can be made from potatoes.

Potato plant is generally 0.5 to 1m tall herb with many branches. Underground stem is used for edible purpose. Odd pinnate with a profuse outer leaflet and cymose panicle is the main characteristics of this crop. Potatoes are cultivated most pat of India. The major potato growing states in India are Himachal Pradesh, Uttar Pradesh, Madhya Pradesh, West Bengal, Bihar and Assam. As it is very cheap, easy to prepare and cope with wide range of soil pH, it is considered to be an important vegetable crop. It is also called as a "poor man's friend".

Presence of nuisance observation in any set of data explicate wrong conclusion. These data points which are subjected to mislead of data are generally called as outliers. It is profoundly important to researcher to explore these types of unwanted observations otherwise it may responsible for attaining biasness in whole estimation procedure. Existence of outlier in any data contravenes the assumption of ANOVA, most importantly the normality of distribution that brings about the illegitimate inference about the whole experiment. Ultimately it will give erroneous result. Chauvenet (1863) pointed out a test for observing a single suspicious observation in linear regression model. Dixon (1950) suggested an appropriate model for studying outlier. Ferguson and Srikantan (1961) derived an idea of mean shift and variance inflation model from Dixon (1950).

**Email: anurupbckv@gmail.com*

How to cite : Middya, M.M.H., Roy, M., Saha, A., Basak, J.S. and Majumder, A. 2023. Exploration and consequences of outlier on potato field crop through factorial experiment. *J. Crop and Weed*, 19(2): 57-62.

Masking and Swamping effects were tested by Bandre (1989) to find several outliers. (Montgomery *et al*. 2001) instructed an idea for observing outlier in linear regression model. Box and Draper (1975) opined an idea for exploring outlier in designing of an experiment. For detecting influential observation under full rank model, cook (1977) developed Cook- distance for observing disturbance observation in factorial experiment. Bhar and Gupta (2001) have constructed some procedure for examining the outlying observation in single factor experiment. Roy *et al.* (2019) experimented on bell paper for the detection of outlier through Cook (1977) statistics.Bhar *et al.* (2007) gave a details idea about the outlier detection in experimental design. A comprehensive idea for identifying suspicious observation has been propounded by Daniel (1960). John (1978) expressed his opinion about the effect of being presence of any unwanted observation in factorial experiment. Box and Mayer (1986) derived a solution for assaying data in factorial experiment.

MATERIALS AND METHODS

Two experiments were conducted at Purba Bardhaman district of West Bengal, India in 2019-20 through factorial experiment on potato crop. First one is asymmetrical factorial experiment (four different varieties along with three different seed rates) and another one is on same crop but design of experiment is symmetrical factorial experiment (three different varieties along with three different spacings). Entire experimental details are tabulated below in Table 1.

Model

Let's explain the experimental setup using generallinear model,

$y = X\theta + e;$ ..... (1)

where mean of error term is 0 and the dispersion matrix of error term is $\sigma^2 I_n$, $\sigma^2 > 0$

Where $y_{n\times 1}$ is a vector representing n observations, X is an n×p (matrix containing known constants, where full column rank is p), θ is a vector of unknown parameters with order p×1, and e is n×1 vector with an independent random variable. For the purpose of computing the least square estimate of θ, one point has been removed. The 1st step will be to ascertain the degree of inûuence of the ith data point has on the estimate $\hat{\theta}$. $\hat{\theta}_{(i)}$ specify the least square estimate of è (i-th point removed).

Cook statistics for detection of outlier

Cook (1977), gave the distance between $\hat{\theta}_{(i)}$ and $\hat{\theta}$ (D_i= the distance)

$$Di = \frac{(\hat{\theta}-\hat{\theta}_i)' [D(\hat{\theta})]^{-1}(\hat{\theta}-\hat{\theta}_i)}{\text{Rank } [D(\hat{\theta})]} \quad \text{..... (2)}$$

Di connotes (1– α)×100% confidence ellipsoid and satisfies $Di \le F_{(p,n-p,(1-\alpha))}$.

The model is the same for an experimental design d, but the rank of X is now mwhich is lesser than p.

Let $\theta = (\theta'_1\ \theta'_2)'$, where θ_1= v-component vector (containing relevant parameters) and θ_2=(p–v) component vector (contains the unwanted parameters, which are not very relevant to the experimenter). So, the model can be expressed as-

$\mathbf{y} = (X_1\ X_2)\ (è'_1, è'_2) + e$ .....(3)

where X splits into two group. From, the model, $X'Xè = X'Y$ after discarding $è_2$ and getting this expression from normal equation, it precludes only $è_1$

$C_{è1}\ è_1 = Q_{è1}$ .....(4)

$C_{è1} = X'_1\ B\ X_1$ .....(5)

$Q_{è1} = X'_1\ B\ Y$ .....(6)

$B = I_n – X_2\ (X'_2X_2)^{-1}X'_2$ .....(7)

The B (n×n) matrix is idempotent and symmetric, the $C_{è1}$(v×v matrix) is symmetric (actually it is the C matrix of the design).

Given that this linear model is intended for experimentaldesigns, It is reasonable to suppose that the vector 1 is present in the column space of X1 and X2. Thus $C_{è1}\ 1 = 0$.

We suppose that the design d we're considering here is connected, i.e., Rank ($C_{è1}$) = v–1 and for the parameter $è_1$, all the (v–1) orthonormalized contrasts are estimable.

Let $Pè_1$ be the expression for the set of (v-1) orthonormalized contrasts for the parameters $è_1$. The (v–1) × v matrix P is such that

$PP' = I_{(v-1)}$, ...(8)

$$P'P = I_v - \frac{1}{v} \quad \text{..... (9)}$$

Further more $\hat{\theta}_1$ provides the best linear unbiased estimator (BLUE) of $P\theta_1$, where $\hat{\theta}_1$ is any solution of the reduced normal equations.

For identifying outlier by cook statistics, let the observation belongs to first plot of a block design be assumed as an outlier. Let us consider a block design d1. The model (intra-block) for such design will be:

$\mathbf{y} = \mu 1_n + \Delta'\tau + D'\beta + e$ .....(10)

Here Δ' implies that n×v design matrix consisting treatment effects with element 0 and 1.

D' implies that n×b design matrix consisting block effects with elements 0 and 1.

μ denotes general mean,

v component vector of treatment effects is implied by the symbol τ.

b-component vector of block effects is represented by β.

$X = (X_1 X_2)$,

$X_1 = \Delta_1$, $X_2 = [1_n\ D']$, $\theta_1 = \tau$, $\theta_2 = [\mu\ \beta']'$.

let us define, matrix θ as, $\theta = I_n - D'k^{-1}D$.....(11)

θ = matrix that is symmetric and idempotent. And Cτ implies that ΔθΔ′.

Cτ is C-matrix in the block design setup. S_{11} is the diagonal element in S matrix,

Where, $S = \theta\Delta'\ C_{\tau}^{+}\Delta\theta$ ..... (12)

C_{τ}^{+} is the Moore-Penrose inverse matrix of C_{τ}

For the outlying observation, r* and t_1 are the ordinary residual and studentized residuals, respectively.

Where,= $r_1^* = y_1 - \hat{y}_1$ and $t_1 = \frac{r_1^*}{\hat{\sigma}\sqrt{v_{11}}}$, where, the matrix V's first diagonal component is v_{11}

$V = \Phi - \Phi\Delta' C_{\tau}^{+}\Delta\Phi = \Phi - S$ ..... (13)

Thus D_1 (Statistics for obtaining outlier of the first plot of the experiment) is-

$$D_1 = \frac{s_{11}}{v_{11}} \frac{t_1^2}{v-1} \text{.....(14)}$$

Q_t statistics for detection of outlier

Let the mean shift model for detecting outlier in experimental design be:

$E(y) = (X: U)\begin{pmatrix}\theta\\ \delta\end{pmatrix}$ ..... (15)

Where $U = (u_1, u_2, \ldots u_i \ldots\ldots u_t)$ and $U_{=}(0,0.\ ,1(i^{th}), 0.0.\ 0)'$, i=1,2,..... t

The normal equations for estimating parameters in equation

$[(X\ \ U)'(X\ \ U)]\begin{pmatrix}\theta\\ \delta\end{pmatrix} = (X\ \ U)'y$ .....(16)

Solution of $= \hat{\delta} = (U'V^{-}U)^{-1}U'V^{-}y$..... (17)

Where V is already mentioned in the above discussion

Qt statistic can be expressed as given below

$Q_t = r_i'(U'VU)^{-1}r_i$.....(18)

Where r_i is the i^{th} studentized residual

formula for calculating $r_i$is given below

$$ri = \frac{ei}{\sqrt{s\text{^}2(1-hj)}} \text{ ..... (19)}$$

e_i- i^{th} residual, h_j-j^{th} diagonal element of $X(X')^{-1}X'$,s^2-mean square error,X-designed matrix,X′-transpose of design matrix.

Table 1: Experimental details:

Lay out of the experiment	Type of treatment combination	Name of the crop	No of replication	No of treatment	Factor-1	Factor-2	Parameter recorded
Asymmetrical factorial RBD	4x3	Potato crop	3	12	Variety: Kufri Jyoti(v1), Kufri Pukhraj(v2), Kufri Chandramukhi(v3), kufri Ashoka(v4)	Seed rate :8 q acre^{-1}(Sr1),9 q acre^{-1} (Sr2),10 q acre^{-1} (Sr3)	Yield of potato crop (t acre^{-1})
Symmetrical factorial RBD	3^2	Potato crop	3	9	Variety: Kufri Jyoti(v1), Kufri Pukhraj(v2), Kufri Chandramukhi(v3),	Spacing:15-20cm (S1), 20-25cm (S2),25-30cm (S3)	Yield of potato crop (q acre^{-1})

RESULTS AND DISCUSSION

It is noticed that (Table 2), the 10^{th} observation located in 1^{st} replication of variety number 2, spacing number 1, displays the highest value of Cook Statistic.

Table 2: List of cook statistic valuesfor yield of three Potato varietieswith three different spacing.

Variety	R1			R2			R3		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
V1	0.049445	0.031645	0.012962	7.91E-05	0.007911	7.91E-05	0.0535	0.007911	0.011016
V2	**0.38765***	0.026205	0.000316	0.046332	0.023405	0.018278	0.1665	7.91E-05	0.023405
V3	0.021278	0.037598	0.001142	0.00062	0.02856	0.010645	0.01463	0.00062	0.018762

Note. *Significant at 95%i.e. (1-á) x100% confidence level.

Table 3: List of Qt statistic of yield of Potato in relation to three different spacing in three varieties.

variety				R2			R3		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
V1	0.59333	0.37973	0.15554	0.00094	0.09493	0.00094	0.64175	0.09493	0.13218
V2	**4.65177**	0.31446	0.00379	0.55597	0.28085	0.21933	1.99137	0.00094	0.28085
V3	0.25533	0.45116	0.01370	0.00744	0.34271	0.12774	0.17559	0.00744	0.22514

Table 4: Mean values with associated ranks and error mean squares of an ANOVA table, both before and after elimination of a significant outlier from the experiment on yield of Potato

Variety no.	Treatment means (Actual)	Treatment means (Removing outlier)	Spacing	Treatment means (Actual)	Treatment means (Removing outlier)	EMS (Actual)	EMS (Removing outlier)
V1	149.44(2^{nd})	149.444(2^{nd})	S1	144.77(3^{rd})	142.347(3^{rd})	27.0925	10.0537
V2	158.88(1^{st})	156.458(1^{st})	S2	158.33(1^{st})	158.333(1^{st})		
V3	144.88(3^{rd})	144.888(3^{rd})	S3	150.11(2^{nd})	150.111(2^{nd})		

Table 5: List of Cook statistic of yield of potato in relation to three different seed rate in four varieties.

variety	R1			R2			R3		
	Sr1	Sr2	Sr3	Sr1	Sr2	Sr3	Sr1	Sr2	Sr3
V1	2.96E-05	0.31174	0.203356	0.19143	0.03838	0.35833	0.304291	0.052655	0.061039
V2	0.282482	0.084025	0.19143	0.19143	0.007836	0.32691	**0.62923***	0.132277	0.014903
V3	0.254665	0.34244	0.10031	0.091987	0.003804	0.268394	0.030896	0.11347	0.022179
V4	0.001043	0.34244	0.157813	0.168659	0.04962	0.12744	0.366411	0.030896	0.052655

Note.* Significant at 95% i.e. (1-á) x100% confidence level

Table 6: List of Qt statistic of yield of Potato in relation to three different seed rates in four varieties.

variety	R1			R2			R3		
	Sr1	Sr2	Sr3	Sr1	Sr2	Sr3	Sr1	Sr2	Sr3
V1	0.00026	2.79392	1.82254	1.71566	0.34397	3.21147	2.72715	0.47190	0.547055
V2	2.5317	0.753062	1.71566	1.71566	0.07022	2.92987	**5.63944**	1.18551	0.13356
V3	2.28239	3.06906	0.89901	0.82442	0.03409	2.40543	0.27690	1.01695	0.198776
V4	0.00934	3.06906	1.41437	1.51157	0.44470	1.14216	3.28389	0.27690	0.47190

Table 7: Mean values with associated ranks and error mean squares of an ANOVA table, both before and after elimination of a significant outlier from the experiment on yield of Potato..

Variety no.	Treatment means (Actual)	Treatment means (Removing outlier)	Spacing	Treatment means (Actual)	Treatment means (Removing outlier)	EMS (Actual)	EMS (Removing outlier)
V1	14.9889(3^{rd})	14.9889(3^{rd})	S1	**14.45833**(3^{rd})	**14.6925**(3^{rd})	0.114672	0.09153
V2	**15.778** (1^{st})	**16.09** (1^{st})	S2	15.76667(1^{st})	15.76667(1^{st})		
V3	14.4889(4^{th})	14.4889(4^{th})	S3	14.96667(2^{nd})	14.96667(2^{nd})		

The tabulated value of F for degrees of freedom 8 and 16 at the 95% confidence level, (1-α) x100% is 0.311915. Thus, the specified value of Cook Statistic is significant and considered to be potential outlier.

After obtaining the Cook Statiatic, we also examine that the value is corrected or not, through Qt statistic. Then we check (Table 3) the maximum value of Qt statistic possesses in which observation. We have found that the 10^{th} observation is highest, so it is considered to be influential.

Table 4 shows that the error mean square of analysis of actual observation is greater than the error mean square of analysis after removal of outlier. Thus, it is ascertained that outlier removal enhances the efficacy of the experiment. It is also notified that rank of the variety no 2 and spacing number 1 has not been altered its original position. It is noticed that (in Table 5), the 16^{th} observation located in 3^{rd} replication of variety number 2, seed rate 1, displays the highest value of Cook Statistic. The tabulated value of F for 11 and 22 degrees of freedom at 95% i.e. (1-α) x100% confidence level is 0.380. Thus, the specified value of Cook Statistic is significant and considered to be potential outlier. After getting out the cook statistic, we also examine that the value is corrected or not, through Qt statistic. Then we check (in Table 6) the maximum value of Q_t statistic possess in which observation. We have found that the 16^{th} observation is highest, so it is considered to be influential. Table7Shows that the error mean square of analysis of actual observation is greater than the error mean square of analysis after removalof outlier. Thus, it is ascertained that outlier removal enhances the efficacy of the experiment. Additionally, it is noted that the positions of variety No. 2 and seed Rate No. 1 have not changed from their initial positions.

CONCLUSION

Detection of outlier in field experiment improves the efficiency of the experiment and also removes the additional pseudo influence of theoutlier.In this experiment Table 2 identifies the 10th observation of variety number 2(Kufri Pukhraj), spacing number 1(20-25cm), as an outlier with the highest Cook Statistic value. Table 3 confirms its influence through the highest value of Q_t statistic. Outlier removal showed an improvement in experimental efficacy, without altering the rank of variety and spacing. Similarly, in Table 5, the 16th observation of variety number 2 (Kufri Pukhraj), seed rate1(8 q $acre^{-1}$), is identified as a potential outlier using Cook Statistic, with Table 6 confirming its influence through the Q_t statistic. Again, outlier removal improved experiment efficacy by reducing error mean square while maintaining the rank of variety and seed rate. In future these methods can be applied, successfully in any field experiments for any crop under study. The performance of treatments under consideration can be judged with extra precision.

REFERENCES

Bendre, S.M. 1989. Masking and swamping effects on tests for multiple outliers innormal sample. *Communications in statistics-Theory and Methods*, **18**: 697-710.

Bhar, L.and Gupta, V.K. 2001. A useful statistic for studying outliers inexperimentaldesigns. *Sankhya B*, **63**: 338-350.

Bhar, L., Parsad, R. and Gupta, V.K. 2008. Outliers in Designed Experiments,Project report, IASRI, New Delhi

Box, G.E.P. and Meyer, R.D. 1986. Analysis of unreplicated factorials allowing for possibly faulty observations. University of Wisconsin, Report No.3.

Chauvenet, W. 1960. A manual of spherical and practical autonomy (Vol. II, 5thed.). New York: Dover.

Choudhury, B. 1996. Vegetables. National Book Trust, New DelhiIndia, pp.24-25

Daniel, C. 1960. Locating outliers in factorial experiments. *Technometrics*, **2**: 149-156.

Dixon, W.J. 1950. Analysis of extreme values. *Annals of Mathematical Statistic*, **21**: 488-506.

Ferguson, T.S. 1961. On the rejection of outliers. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1). Berkeley and Los Angles: University of California Press. 253-287.

John, J.A. 1978. Outliers in factorial experiments. *Applied Statistics*, **27**: 111-119.

Montgomery, D.C., Wisnowski, J.W. and Simpson, J.R. 2001. A comparative analysis of multiple outlier detection procedures in the linear regression model, *Computational Statistics & Data Analysis*, **36**: 351–382.

Roy, H. S. 2013. A Study on Outliers in Factorial Experiments. M.Sc (Ag) Thesis, Indian Agricultural Research Institute, New Delhi, pp. 42-45

Roy, M., Saha, A., Basak, J., Saha, M., Middya, M., Roy, S. and Majumder, A. 2022. Detection and impact ofoutliersonbell pepper (*Capsicum annuum* L. var. *grossum* Sendt.) in field condition. *J. Crop and Weed*, **18(1)**: 187-195.