VOL. 11, NO. 18, SEPTEMBER 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

RANK ORDER OUTLIER (ROO) PLOTS TO DETECT POSSIBLE OUTLIERS IN UNREPLICATED 2^K COMPLETELY RANDOMIZED FACTORIAL DESIGNS: NUMERICAL EXAMPLE

Anwar Fitrianto[1,2,3] and Low Chuan Haur[1]
[1]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia UPM Serdang, Selangor, Malaysia
[2]Laboratory of Computational Statistics and Operations Research, Institute for Mathematical Research, Universiti Putra Malaysia, Malaysia
[3]Department of Statistics, Faculty of Mathematics and Natural Resources, Bogor Agricultural University, Indonesia
E-Mail: anwarstat@gmail.com

ABSTRACT

Two-level unreplicated factorial design is very common in manufacturing industries. The design can be used to save cost since it usually needs less experimental run. But, problem appears when the experiment is done without any replication. In such kind of experiment, there are problems in identifying significant terms as well as to identify possible outlier in the data. This article discusses about the use of Pareto plot to identify significant terms for unreplicated two-level factorial experiments through numerical example. Meanwhile, the numerical example is also used to clearly describe how to create and interpret both Rank Order Outlier (ROO) and iterative ROO plot in identifying possible outlier in the experimental data.

Keywords: factorial experiment, rank order outlier, unreplicated.

INTRODUCTION

In statistics, an observation which is numerically different from the rest of the data is called an 'outlier'. Statistical outliers are unusual points in a set of data that differ substantially from the rest of the data. An outlier could be different from other points with respect to the value of one variable or, in multivariate data it could be unusual in respect of the combination of values of several variables. It occurs when the experiment run in environments with many noise variables that will affect the result of the experiments.

Researchers who often run the experiment do not possess a suitable methodologies in detecting the outliers may have a risk in misinterpreting the result. Outliers have influence on the estimation of a model that is being fitted to the data. An inaccurate conclusion can be made with the presence of outliers. Outliers can be detected by using several methods such as statistical approaches, sampling methods, graphical methods and others. The advantages of using graphical method are easy to understand and interpret as well as quick and easy to use and make visual sense.

Especially in an unreplicated two-level factorial experiment, situations will be worst when there is an outlier in the data since it is also difficult to identify this unusual response. Goupy (2006) mentioned few methods which are able to resolve this problem. The methods have been discussed previously by Goupy (1996) and Hund *et al.* (2002), which are: (i) reconstruction of the experimental design, (ii) Daniel diagram, and (iii) comparison of measured and calculated responses with least square regression. Even the suspected outlier can be identified, but there are other complications in data analysis.

In this study, the following questions have been set: (1) how is the performance of existing graphical methods in detecting outlier? (2) What is the significant factor in the experiment? These research questions lead us to specify objective of the study, namely to apply the graphical methods for detecting outliers in unreplicated 2^k completely randomized factorial designs based on numerical example.

LITERATURE REVIEW

Outlier has been a topic in statistics field for many years. Discussion about outliers includes its effects on parameter estimations in regression models, how to detect outliers as well as what are the effects in experimental designs. In experimental designs, Daniel (1960) had discussed how to locate outliers in an experimental design. According Daniel (1960), an outlier in a factorial experiment is an observation whose value is not in the pattern of values produced by the rest of the data where the definition has been accepted until today. Bhar and Gupta (2001) proposed a new criterion of detecting outlier in experimental designs which is based on average Cook-statistic. At the same year, Seheult and Tukey (2001) discussed about outlier identification and robust analysis in factorial experimental design. Meanwhile, Zhou and Julie (2003) realized the fact that in practice, experiments may yield unusual observations (outliers).

In general, there are many methods on how to identify outliers in the data, both graphical dan numerical approaches. There have been few graphical approaches to identify outliers in experimental design such as rank order outlier (ROO) plot, Daniel plot and Half-Normal plot. Recently, Goupy (2006) described how to identify an outlier and how to estimate the true value of this outlier in a two-level experimental design with at least 16 experimental runs with no replicates. The method was based on the use of a dynamic variable and the ''small effects'' of the Daniel's diagram.

Rank order outlier plot

Rank order outlier plot is a simple graphical tool that can clearly observe potential outliers through visual assessment. We can also detect outliers in ROO plot by investigating the inconsistency in the pattern of the plot. We create a dataset with the presence of the treatments, response value *y* and the rank order. The rank order need to be rearranged in ascending order. According to Sanders and Hild (2012), there are four elements to create rank order outlier plot, namely:

1. an experimental data with the presence of response variable *y*. The data is then sorted in ascending order, from the smallest to largest value of *y*.
2. create rank order of each treatment (observation) according to the value of *y*.
3. identify the most significant effects. This can be done through usual Anova or graphical approach.
4. A plot of the combination of the most significant effects with response values on the *y*-axis and rank order against *x*-axis.

Pareto plot

We are going to detect outlier by using graphical method with SAS software. In order to do that, firstly, we need to know which factor is the most significant in the experiments; hence, we are going to construct a Pareto plot to identify them.

Pareto plot is a type of plot using the combination of bar chart displaying the percentage of categories and line graph displaying the cumulative percentage. The bars are ordered by frequency in decreasing term, which makes Pareto plot useful for deciding the most effective effect or what problem should be solved first. Normally the longest bar is relatively more significant compared to those shorter bars.

Half-normal plot (Daniel plot)

In a similar way, a disproportionately large percentage of errors or defects in any process are usually caused by relatively few problems. Pareto plot helps us to identify those significant few problems so people can target them for an action.

After knowing the most contributed or significant effect from the Pareto plot, we continue further by plotting the Half-Normal plot to have a confirmation of the most significant effect factors in the experiments. It is a simple way to find significant terms as well as to identify suspected outliers in an experimental data. The objective is similar to Daniel plot (Daniel, 1976). Daniel suggested that the examination on a normal probability plot of the estimates of the effects should be made. The effects will tend to fall along a straight line on this plot, whereas the significant effects will not lie on the straight line. The idea of plotting normal quantile plot is to compare the value from the experiment result with the value predicted to be the standard normal distribution.

The effects that are negligible are normally distributed, with mean zero and variance σ and will tend to fall along a straight line on this pool, whereas significant effects will have nonzero means and will not lie along the straight line. The Daniel plot of the effect estimates is actually based on the use of Lenth's pseudo-standard error (*PSE*) to determine the significance of effects.

DATA METHODOLOGY

Data

In this study, we focus on finding significant terms of a factorial experiment as well as to identify possible outlier in the experimental data. We use an artificial data of a 2^4 unreplicated experimental design (4 factors and 16 experimental runs) under completely randomized design. Let us say that the factors are *A, B, C* and *D* with 2 levels each (low and high) and the response variable is yield. The corresponding rank of the response variable is displayed in the last column of Table-1.

Table-1. Artificial experimental design data.

Run	Factor				Yield	Rank
	A	***B***	***C***	***D***		
1	60	6	60	6	60	6
2	30	4	30	4	30	4
3	89	12	89	12	89	12
4	29	3	29	3	29	3
5	100	13	100	13	100	13
6	85	11	85	11	85	11
7	115	14	115	14	115	14
8	75	8	75	8	75	8
9	33	5	33	5	33	5
10	23	2	23	2	23	2
11	73	7	73	7	73	7
12	10	1	10	1	10	1
13	116	15	116	15	116	15
14	83	10	83	10	83	10
15	130	16	130	16	130	16
16	79	9	79	9	79	9

METHODOLOGY

Pareto plot to detect the most contribution effects will be used. After obtaining the Pareto plot, the most contributed effects will be visualized by selecting the longest bar in the plot, but, we will recheck the model through Half-Normal plot in order to make strong evidence showing that the effects displayed in Pareto plot is the accurate contribution. After investigating both Pareto and Half-Normal plots, the most contributed effect as well as suspected outlier can be identified. The step is

continued with plotting a ROO plot by choosing response variables *y* as the *y*-axis and rank order as the *x*-axis.

ROO plot will be created by using Minitab for identifying strange pattern based on the most contributed term to the response variable. Visual assessment can be done to detect the outlier in ROO. For the next checking, iterative ROO plot will be developed to confirm whether a suspected outlier in the previous plot (if any) is obviously possible outlier.

RESULTS AND DISCUSSIONS

Analyzing data from unreplicated factorial experiment is not simple. It dues to the fact that when there is no replications, the residual error equals to zero and calculation of the *F* statistics is not possible. As the result, the significance of each term in the linear model cannot be calculated, as it can be seen in Table 2. It creates another complication since significant terms cannot be identified.

Table-2. Analysis of variance of the artificial experimental data.

Source	df	Sum of squares	Mean squares	*F*	*p*
Main Effects	4	17968.5	4492.1	*	*
2-Way Interactions	6	1788.5	298.1	*	*
3-Way Interactions	4	250.5	62.6	*	*
4-Way Interactions	1	56.3	56.3	*	*
Residual Error	0	*	*		
Total	15	20063.7			

There are 15 terms for the $2^{k=4}$ full factorial experiment. Those terms are main effects (A, B, C, D), two levels interactions (AB, AC, AD,…, CD), three levels interactions (ABC, ABD, …, BCD), and *ABCD* interaction. Analyzing the experimental data using Minitab produces the following Pareto plot and Table 2 of estimated effects of each term. For unreplicated experimental design, Minitab uses Lenth's PSE to determine statistically significant factors and we found that main effects of *C* and *A* as the significant effects. This information is also given in the Pareto plot which reveals only *C* and *A* pass the vertical line of 17.35 as shown in Figure-1. There are no interaction terms having significant contribution to the yield. In this artificial experimental data, visual assessment of the Half-Normal plot (Figure-2) reveals that except *C* and *A*, there is no obvious gap that would indicate the presence of suspected outliers which mean that an initial inspection based on the Half-Normal plot provides no indication of an apparent outlier. Meanwhile, Table 3 displays the ordered (when the absolute values are considered) of estimated effects for the artificial experimental data. It informs us that the main effect of factor *C* has highest contribution for the response variable with an effect estimate of 54.5. Moreover, since the effect is positive, it means that higher value of yield can be obtained by setting the factor *C* at high setting.

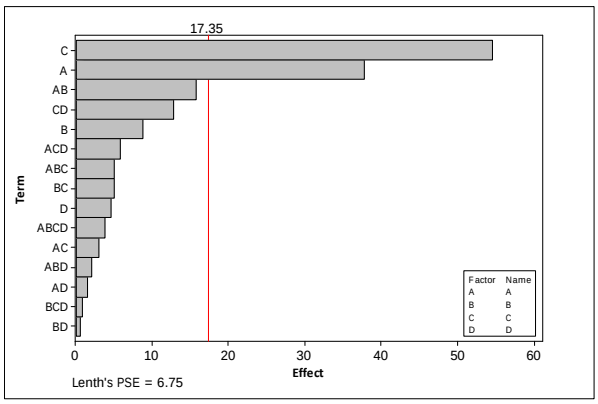

Figure-1. Pareto plot of estimated effects of the numerical example.

Table-3. Ordered estimated effects and coefficients of the artificial experimental data.

Term	Effect	Coefficient
C	54.50	27.25
A	-37.75	-18.87
AB	-15.75	-7.88
CD	12.75	6.38
B	8.75	4.37
ACD	-5.75	-2.88
BC	-5.00	-2.50
ABC	5.00	2.50
D	-4.50	-2.25
ABCD	3.75	1.88
AC	3.00	1.50
ABD	-2.00	-1.00
AD	-1.50	-0.75
BCD	0.75	0.37
BD	0.50	0.25

We also display an ROO plot in Figure-3 with the Y values are plotted against the rank based on the level of *C* for that treatment. The factor *C* determines the reference distribution of the ROO plot since it has the largest effect. The reference distribution is reflected by using cross (×) and minus (+) symbols to indicate the treatments where *C* is at low and high levels, respectively. A quick visual assessment on the associated ROO plot indicates that the twelfth largest order statistic (rank) which corresponds to observation associated with treatment 3, does not fit the

pattern of the responses where *C* was at the high level. This value is a possible outlier.

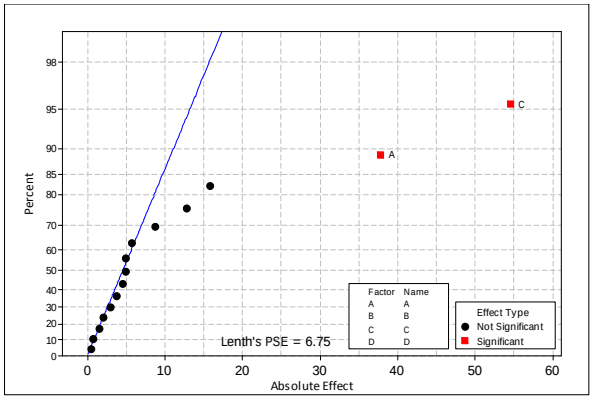

Figure-2. Half-Normal plot of the numerical example.

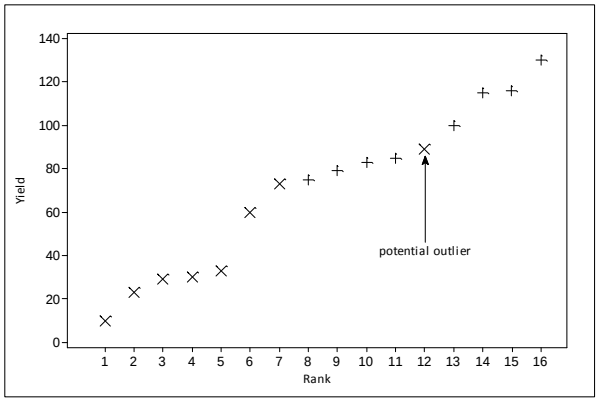

Figure-3. ROO plot for the numerical example.

The ROO plot in Figure-4 shows an iterative investigation of effect patterns by using the two largest effects to determine the reference distribution. As we have discussed previously, main effects of factor *C* and factor *A* give largest and second largest to the response variable so that they can be used as the basic to construct an iterative ROO plot. Just like in the Figure-3, plus and cross symbols are still indicating treatments where *C* is at high and low levels, respectively. Meanwhile, circles and squares are used to indicate whether *A* is at low or high settings. Visual assessment in Figure-4 indicates that the same observation (no 12) which corresponds to experimental run no 3 does not fit with the expected pattern of the remaining observations.

So, observation 12 is possible outlier. But, it is not obvious outlier since it is produced by a single replication. It informs us that for that particular treatment, the pattern is as it was expected. Using ROO and iterative ROO plots at least they can help researcher to trace further what is happening in the experiment beside factors.

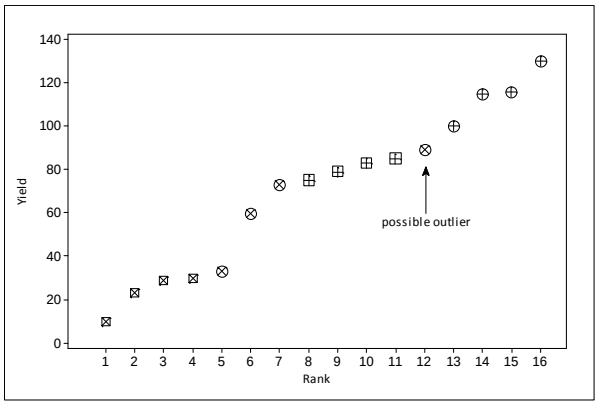

Figure-4. Iterative ROO plot for the numerical example.

CONCLUSIONS

Lenth's PSE was used to identify significant terms in analyzing unreplicated two-level factorial experiment since usual analysis of variance is not able to do. Meanwhile, through numerical example, it was clearly shown how to use ROO and iterative ROO plotsfor identifying outlier in unreplicated experimental design. We have shown that both ROO and iterative ROO plotscan identify the same observation as possible outlier. In the future work, the researcher should be able to identify outlier in unreplicated factorial design under completely randomized design or more complicated designs.

REFERENCES

Bhar, L. and V.K. Gupta, 2001. A useful statistic for studying outliers in experimental designs. Ind. J. Stat., 63, pp. 338-350.

Box. 1991. Finding bad values in factorial designs, Quality Engineering, 3(3), pp. 405-410.

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

Daniel, C. 1976. Applications of Statistics to Industrial Experimentation: John Wiley and Sons Publishers, New York.

Goupy, J. 1996. Outliers and experimental designs, Chemometr. Intell. Lab. Syst, 35, pp. 145-156.

Goupy, J. 2006. Factorial experimental design: Detecting an outlier with the dynamic variable and the Daniel's diagram. Chemometrics and Intelligent Laboratory Systems, 80, pp. 156-166.

Hamada, M and Balakrishnan. 1998. Analyzing unreplicated factorial experiment: A review with some new proposals (with comments and rejoinder), Statistica Sinica, 8(1), pp. 1-41.

Hund, E. D. Luc Massart and J. S. -Verbeke. 2002. Robust regression and outlier detection in the evaluation of robustness tests with different experimental designs, Anal. Chim. Acta, 463, pp. 53-73.

Lawson, G. 2006. Finding bad values in factorials-Revisited, Quality Engineering, 18, pp. 491-501.

Lenth, R.V. 1989. Quick and easy analysis of unreplicated factorials, Technometrics, 31(4), pp. 469-473.

Mallows, C. L. 1987. Design Data and Analysis: By Some Friends of Cuthbert Daniel: John Wiley and Sons Publishers, New York.

Montgomery, D. 2005. Design and Analysis of Experiments, 6th Edition: John Wiley & Sons Publishers, United States.

Sanders, D., and C. Hild. 2012. A graphical tool for detection of outliers in completely randomized, unreplicated 2^k and 2^{k-p} factorials. Quality Engineering, 24(2), pp. 514-521.

Seheult, A.H. and J.W. Tukey, 2001. Toward Robust Analysis of Variances. In: Data Analysis from Statistical Foundations: A Festschrift in Honour of the 75th Birthday of D.A.S. Fraser, Saleh, A.K.M.E. (Ed.), Nova Publishers, Huntington, ISBN-10: 1560729686, pp. 217-244.

Stefansky, W. 1972. Rejecting outliers in factorial designs, Technometrics, 14, pp. 469-479.

Torres, V. A. 1993. A simple analysis of unreplicated factorials with possible abnormalities, Journal of Quality Technology, 23(3), pp. 183-187.

Zhou, J. and Z. Julie. 2003. Robust estimation and design procedures for the random effects model. Canadian
J. Stat, 31, pp. 99-110.