Journal of Mathematics and Statistics 9 (3): 193-197, 2013
ISSN: 1549-3644

doi:10.3844/jmssp.2013.193.197 Published Online 9 (3) 2013 (http://www.thescipub.com/jmss.toc)

A COMPARISON BETWEEN CLASSICAL AND ROBUST METHOD IN A FACTORIAL DESIGN IN THE PRESENCE OF OUTLIER

[1,2,3]Anwar Fitrianto and [1,2]Habshah Midi

[1]Department of Mathematics, Faculty of Science,
[2]Institute for Mathematical Research,
Universiti Putra Malaysia, Serdang, Malaysia
[3]Department of Statistics, Bogor Agricultural University, Bogor, Indonesia

Received 2013-05-29; Revised 2013-06-25; Accepted 2013-07-11

ABSTRACT

Analysis of Variance (ANOVA) techniques which is based on classical Least Squares (LS) method requires several assumptions, such as normality, constant variances and independency. Those assumptions can be violated due to several causes, such as the presence of an outlying observation. There are many evident in literatures that the LS estimate is easily affected by outliers. To remedy this problem, a robust procedure that provides estimation, inference and testing that are not influenced by outlying observations is put forward. A well-known approach to handle dataset with outliers is the M-estimation. In this study, both classical and robust procedures are employed to data of a factorial experiment. The results signify that the classical method of least squares estimates instead of robust methods lead to misleading conclusion of the analysis in factorial designs.

Keywords: M-Estimation, Factorial Design, Outlier, Robust

1. INTRODUCTION

In statistics, conducting an experiment is one way to obtain the data. Related to the data obtained, there are important things we need to consider, namely the presence of one or more outliers in the data. This problem has been dealt with in great detail in linear regression problems but may not get much attention in the context of experimental design. The decision to retain or discard outliers depends on the purpose of the study. Many studies have been done when we considered to keep the outlier in the data. Gentleman and Wilk (1975) and John and Drapper (1978) studied about outliers design of experiment in a two-way anova through residual analysis. Few years later, John (1978) incorporated his previous study to discuss the problems that arise in detecting the presence of one and two outliers in factorial experiments.

The presence of outlier, especially in experimental data is responsible for misinterpretation of experimental data which indicate that no abnormalities in the results where in fact it is not. The consequences of the presence of outliers are well known. Nelder (1971) noted that 1% gross error in such an experiment can result in a false inference, while 1 to 10% gross errors are rather rule than exceptions in reality. Bhar and Gupta (2001) pointed out that even a single outlier may alter the inference to be drawn from the experiment.

Our goal in this study is to show that outlier has an effect on the factorial designs, which may give missleading results. Then, a robust technique is put forward to deal with the presence of outlier in design of experiment. We will show the the performance of a robust technique of M estimator in comparison with the classical Least Squares method. The comparison of both methods will be presented using an empirical dataset.

Corresponding Author: Anwar Fitrianto, Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Serdang, Malaysia

1.1. Outliers in Design of Experiments

Many literatures discussed about outliers including how to identify outliers and how to deal with the presence of outliers. Cook (1977) introduced a statistic to indicate the influence of an observation with respect to a particular model. Related to experimental designs, Daniel (1960) had discussed how to locate outliers in an experimental design. He defined that 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. A year after, Bross (1961) had studied a strategic appraisal analysis of the problem of outliers in patterned experiments. Recent articles by Seheult and Tukey (2001) introduced a method of outlier detection and robust analysis in a factorial experimental design.

Bhar and Gupta (2001) proposed a new criterion of detecting outlier in experimental designs which is based on average Cook-statistic. Meanwhile, Zhou and Julie (2003) realized the fact that in practice, experiments may yield unusual observations (outliers). In the presence of outliers in a data, estimation methods such as ANOVA, truncated ANOVA, Maximum Likelihood (ML) and modified ML do not perform well, since these estimates are greatly influenced by outlier. Zhou and Julie (2003) verified that with robust designs, one can get efficient and reliable estimates for variance components regardless of outliers which may happen in an experiment. Their work is then followed by Goupy (2006) who described how to discover an outlier and estimate its true value. The method is based on the use of a dynamic variable and the "small effects" of the Daniel's diagram.

1.2. Linear Model of a Factorial Experiment

Usual general linear model of an experimental design is written as follows Equation (1):

$$Y = X\theta + \varepsilon \tag{1}$$

where, $Y_{n\times 1}$ is a vector of response variable, $X_{n\times p}$ is the design matrix of nonstochastic constant, $\theta_{p\times 1}$ is vector of parameters to be estimated and $\varepsilon_{n\times 1}$ is vector of errors with zero expectation, $E(\varepsilon) = 0$ and covariance matrix $V(\varepsilon) = \sigma^2 I$. In standard ANOVA, the underlying regression estimator is the least squares estimator, where parameters are chosen to minimize the regression sum of squares.

1.3. The use of Cook's Distance

There are many articles in the literatures that discuss outlier detections. In this article, we consider to employ Cook's Distance which was developed by Cook (1977). Cook's Distance is one of the important methods in statistics to identify outlier or influential observation. It is used for assessing influence in regression models. Cook's Distance usually denoted by D_i, identifies cases with unusual values that have considerable influence on a numerical analysis. Cook distance of the i-th observation is based on the differences between the predicted responses from the model constructed from all of the data and the predicted responses if the i-th observation is eliminated. Fox (1997) suggested a cut-off value of 4/(n-k-l) for detecting influential cases where n is the number of observations and k is the number of predictor (factor).

In linear regression model, Cook's distance, D_i is defined as:

$$D_i = \frac{\left(\hat{\theta}_{(i)} - \hat{\theta}\right)' (X'X)\left(\hat{\theta}_{(i)} - \hat{\theta}\right)}{p \times \hat{\sigma}^2} \tag{2}$$

But since our model here is based on linear model in a design of experiment, we can simplify the Equation (2) above become:

$$D_i = \frac{e_i^2}{p \times \hat{\sigma}^2}\left(\frac{h_i}{(1-h_i)^2}\right) \tag{3}$$

where, $H = X(X'X)^{-1}X'$, $h_i = x_i'(X'X)x_i$ and p = number of predictors in model plus one.

It can be seen from the Equation (3) that D_i is calculated using leverage values and standardized residuals. It considers whether an observation is influential with respect to all fitted values. The template is used to format your paper and style the text. All margins, column widths, line spaces and text fonts are prescribed; please do not alter them. Your paper is one part of the entire proceedings, not an independent document. Please do not revise any of the current designations.

1.4. Robust M Approach

Robust linear models are useful for filtering linear relationships when the random variation in the data is not normal or when the data contain significant outliers. The main purpose of robust regression is to provide resistant (stable) results in the presence of outliers.

Many robust methods have been developed to rectify the problem of outliers. In this study we employ the M

estimators and incorporate this method in linear model two-way experimental designs. It is well known that the least squares estimation method optimize the fit of the model by minimizing the sum of the squared deviations between the actual and predicted Y values, $\Sigma(y-\hat{y})^2$. The method can be represented as Equation (4):

$$\min \sum_{i=1}^{n} \varepsilon_i^2 \tag{4}$$

Huber (1973) and Huber (1981) developed a robust estimator called M-estimator, which are based on the idea of replacing the squared residuals, ε_i^2, with another function of the residuals, given by Equation (5):

$$\min \sum_{i=1}^{k} \rho(\varepsilon_i) \tag{5}$$

where, ρ is a symmetric function with a unique minimum at zero. In general, a sensible ρ-function should have the following properties:

$\rho(\varepsilon) \geq 0$,
$\rho(0) = 0$,
$\rho(\varepsilon) \geq \rho(-\varepsilon)$ and
$\rho(\varepsilon_i) \geq \rho(\varepsilon_i')$ for $|\varepsilon_i| > |\varepsilon_i'|$

Two procedures commonly used to solve the non-linear normal equations for the M-estimates are the Newton-Raphson and the Iteratively Re-weighted Least Squares (IRLS). Practically, the most widely used procedure is the IRLS. In IRLS, the initial fit is calculated and then a new set of weights is calculated based on the results of the initial fit. The iterations are continued until a convergence criterion is met.

ROBUSTREG procedure in SAS provides two linear tests to asses a particular effect. The first test is a robust version of the F test, which is named to as the ρ (rho) test (SASI, 2008). Under H_0, $S_n^2 \sim \lambda \chi_q^2$, where λ is the standardization factor, which is equal to:

$$\lambda = \frac{\int \psi^2(s) d\Phi(s)}{\psi'(s) d\Phi(s)}$$

Meanwhile, according to (SASI, 2008), S_n^2 can be written as:

$$S_n^2 = \frac{2}{a}[Q_1 - Q_0]$$

And:

$$Q_0 = Q\left(\hat{\theta}(0)\right) = \min\left\{Q(\theta) \mid \theta \in \Omega_0\right\},$$

$$Q_1 = Q\left(\hat{\theta}(1)\right) = \min\left\{Q(\theta) \mid \theta \in \Omega_1\right\},$$

where, Φ is cummulative distribution function obtained from standard normal distribution.

The second linear test is a robust version of the Wald test, which is named to as R_n^2 test. It uses a test statistic of:

$$R_n^2 = n\left(\hat{\theta}_{i1}, \hat{\theta}_{i2}, \ldots, \hat{\theta}_{iq}\right) H_{22}^{-1} \left(\hat{\theta}_{i1}, \hat{\theta}_{i2}, \ldots, \hat{\theta}_{iq}\right)^T$$

where, $\frac{1}{n}H_{22}$ is the q×q block, corresponds to $\left(\hat{\theta}_{i1}, \hat{\theta}_{i2}, \ldots, \hat{\theta}_{iq}\right)$, of the asymptotic covariance matrix of the M estimate $\hat{\theta}_M$ of θ in p-parameter linear model (SASI, 2008). In design of experiment, null hypothesis both ρ and R_n^2 tests specify no significant contribution of a particular effect on response variable. When H_0 of no effects is correct, the R_n^2 has chi-squares distribution with q degrees of freedom χ_q^2.

1.5. Empirical Results

To illustrate the comparisons between classical and robust approach in dealing with outlier in factorial experiments, we provide an empirical example. In this example we consider a famous dataset discussed by Daniel (1960) **Table 1**. The analysis is conducted by SAS release 9.2. For data without any outliers (clean data), we employ PROC GLM, meanwhile the contaminated data will be analyzed using PROC PROBUSTREG.

We now apply the classical Least Square (LS) approach to the clean data since we knew that the LS is always better in dealing with 'clean' observations. From **Table 2 and 3**, it is clear that a single outlier has nullified the main effect of chemical B to the response variable. In addition, the presence of an outlier has also reduced the usual goodness-of-fit measurement of R^2. When there is no outlier in the data, both chemical A and B account for about 88.61% of the variability of the response variable. But, it is reduced to 71.14% when there is an outlier in the data.

Table 1. Hypothetical two-way experimental data as mentioned in Daniel (1960)

	B			
A	b_1	b_2	b_3	b_4
a_1	35	32	40	37
a_2	29	29	36	34
a_3	25	29	40 (20)	30
a_4	29	25	35	25
a_5	22	20	29	29

Table 2. ANOVA table of the clean data

Source	df	SS	MS	F	p
A	4	328	82.00	12.00	0.000
B	3	310	103.33	15.12	0.000
Error	12	82	6.83		
Total	19	720			

Table 3. ANOVA table of the modified data

Source	df	SS	MS	F	p
A	4	368	92.00	5.47	0.010
B	3	130	43.33	2.57	0.103
Error	12	202	16.83		
Total	19	700			

Table 4. Cook's distance for clean and modified data

			Cook's distance	
Index	A	B	Clean	Modified
1	a_1	b_1	0.183	0.033
2	a_1	b_2	0.020	0.033
3	a_1	b_3	0.081	0.008
4	a_1	b_4	0.000	0.008
5	a_2	b_1	0.020	0.000
6	a_2	b_2	0.000	0.008
7	a_2	b_3	0.081	0.008
8	a_2	b_4	0.020	0.000
9	a_3	b_1	0.081	0.033
10	a_3	b_2	0.020	0.206
11	a_3	b_3	0.183	0.668
12	a_3	b_4	0.081	0.033
13	a_4	b_1	0.183	0.132
14	a_4	b_2	0.081	0.008
15	a_4	b_3	0.183	0.297
16	a_4	b_4	0.081	0.074
17	a_5	b_1	0.020	0.000
18	a_5	b_2	0.081	0.074
19	a_5	b_3	0.081	0.008
20	a_5	b_4	0.183	0.033

Table 5. Robust linear test for the A effect

Test	Test statistic	λ	df	χ^2	p
ρ	10.747	0.7977	4	13.47	0.0092
R_n^2	39.321		4	39.32	<0.0001

Table 6. Robust linear test for the B effect

Test	Test statistic	λ	df	χ^2	p
ρ	10.7636	0.7977	3	13.49	0.0037
R_n^2	24.8532		3	24.85	<0.0001

To verify that the observation of third row and third column of the modified data is an outlier, we employ the Cook's distance approach. The result is displayed in **Table 4**. The presence of a single outlier in the data inflates the Cook's distance from 0.183 of the clean data to 0.668. The Cook's distances indicate that cases 11 are an influential observation. The presence of this outlier has made the effect of chemical B insignificant. This result has huge impact in the analysis and as a result in applied science, especially in industry.

We used PROC ROBUSTREG of SAS Release 9.2 and employ the robust M to rectify this problem. In comparison with the classical LS, the M estimator produces better results in dealing with the outlier.

By using the M estimator, as we can see in **Table 5 and 6**, we discovered that both chemicals A and B significantly contribute to the response variable with p values of the test statistics are equal to 0.0092 and 0.0037, respectively. From the results we can conclude that the robust M estimator has proven to reduce the effects of outlier on the analysis and lead to significant conclusion of the chemical B and the response.

2. CONCLUSION

In this study we enlightened the importance of employing a robust method in the experimental designs, especially for the factorial experiments to reduce the effects of outliers on the analysis. The numerical example indicates that in the presence of even a single outlier has large effect on the LS procedures. However, the M procedure is less affected by outlier. It can improve the analysis and nullify the effects of outlier. The results of the analysis clearly show that robust approach correctly identifies the significant factors in the presence of outlier.

3. REFERENCES

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

Bross, D.J., 1961. Outliers in patterned experiments: A strategic appraisal. Technometrics, 3: 91-102.

Cook, R.D., 1977. Detection of influential observation in linear regression. Technometrics, 19: 15-18.

Daniel, C., 1960. Locating outliers in factorial experiments. Technometrics, 2: 149-156. DOI: 10.1080/00401706.1960.10489889

Fox, J., 1997. Applied Regression Analysis, Linear Models and Related Methods. 9th Edn., Sage Publications, Thousand Oaks, ISBN-10: 080394540X, pp: 597.

Gentleman, J.F. and M.B. Wilk, 1975. Detecting outliers in a two-way table: I. Statistical behavior of residuals. Technometrics, 17: 1-14. DOI: 10.1080/00401706.1975.10489265

Goupy, J., 2006. Factorial experimental design: Detecting an outlier with the dynamic variable and the Daniel's diagram. Chemomet. Intell. Laboratory Syst., 80: 156-166. DOI: 10.1016/j.chemolab.2005.05.005

Huber, P.J., 1973. Robust Regression: Asymptotics, Conjectures and Monte Carlo. Ann. Stat., 1: 799-821.

Huber, P.J., 1981. Robust Statistics. 1st Edn., Wiley, New York, ISBN-10: 0471418056, pp: 308.

John, J.A. and N.R. Drapper, 1978. On testing for two outliers or one outlier in two-way tables. Technometrics, 20: 69-78. DOI: 10.1080/00401706.1978.10489618

John, J.A., 1978. Outliers in factorial experiments, J. Royal Stat. Soc.. Series C, 27: 111-119.

Nelder, J.A., 1971. A Statistician's Point of View. In: Mathematical Models in Ecology, Blackwell, Oxford, pp: 367-373.

SASI, 2008. SAS/Stat 9.2 User's Guide. 1st Edn., SAS Institute Inc., Cary, ISBN-10: 1607642514, pp: 232.

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.

Zhou, J. and Z. Julie, 2003. Robust estimation and design procedures for the random effects model. Canadian J. Stat., 31: 99-110. DOI: 10.2307/3315906