M-Estimation Of Multivariate Response Surface Models With Data Outliers

Edy Widodo[1], Suryo Guritno[2] and Sri Haryatmi[2]
[1] Doctoral Student of Mathematics UGM, [1] Department of Statistics UII
[2] Department of Mathematics UGM
edykafifa@gmail.com

Abstract:

The main objective of response surface methodology is to find the input variable settings that achieve the optimal compromise in the response variable. In general, there are three main steps in the multi-response optimization problem, namely data collection, modeling, and optimization. This paper focuses on the steps of model building using multivariate regression procedure.

Usually parameters of multivariate response surface models estimated using OLS method. However, this method is highly sensitive to outliers. Outliers can affect the results of statistical analysis, as outliers are highly likely to produce a substantial residual and often affect the model estimation. Estimates of the resulting models to be biased resulting in errors in the actual determination of the optimal point. Therefore, it takes robust response surface models against outliers. As an alternative, we proposed M-Estimation for estimating parameters in multi-response surface models.

We illustrate the proposed method using the well-known problem 'tire treads compound problem', which was originally presented by Derringer and Suich [5]. In this model, are used for three main chemical materials, such as silica (X1), silane (X2), sulfur (X3), elongation at break (Y1) and Abrasion Index (Y2). Based on this example, the performance of the OLS and M-Estimation compared, by comparing the SSE of OLS and M-Estimation. The comparison showed that the M-estimation approach produces smaller SSE. These results indicate that the parameter estimates of multivariate response surface models with the data outliers; the M-estimation has a better performance than the OLS.

Keywords: Multivariate Response Surface Model, Outlier, OLS, M-estimation, and SSE

1. Introduction

Response Surface Methodology (MPR) is a collection of statistical techniques and mathematical or useful methods to analyze the problems of some of the independent variables that affect the dependent variable or response, and aims to optimize the response (maximum, minimum, or more broadly, looking conditions around the stationary point containing ridge). MPR can be used to find a suitable function approach for predicting the response to come and determine the values of the predictor that optimize the response. MPR was first introduced by Box and Wilson [3].

In general there are three main stages in the MPR, namely: 1) data collection, through the selection of appropriate experimental design strategy, 2) estimation of the model/model development, through the selection of appropriate regression modeling methods, and 3) optimization, through the selection of the optimization method will be used to identify the arrangement of independent variables optimize response. When several responses should be analyzed simultaneously, multi-response optimization problems appear, in which the main purpose is to find the input variable settings that achieve the optimal compromise in the response variable.

Data collection, model building, and optimization are the three main steps engaged in typical multi-response optimization problems. The formation of the model is determined by the Ordinary Least Squares method (OLS), although several experimental results indicate the presence of outliers. When using OLS method, outliers can significantly affect the estimated coefficients of the response function. In addition, the presence of contaminated data, reliability and accuracy of the estimation of response surface could not be obtained because the OLS method is highly sensitive to outliers.

Single response surface robust optimization has been investigated by many authors such as Morgenthaler et al. (see [9]) and Hund et al.(see [8]), but for some response to be optimized simultaneously, required several extensions therefore proposed M-estimation of Multivariate Response Surface Model (MRSM). This paper focuses on the steps of model building and tried to estimate the model using multivariate regression procedure correctly. M-estimation approach of MRSM proposed an extension of the robust regression method is provided in multiple response problems. The reasons for selecting the M-estimator to estimate the response surface based on experimental design:

First, differences in robust regression methods can be distinguished by the breakdown point and the percentage of incorrect results (the percentage of erroneous results) so that the method can be overcome without significant impact on the estimates. However, the method with high breakdown point, it is sometimes very tricky so wrong in identifying a good point as an outlier [8].

Second, M-estimators focused on robust regression estimation for situations where there is a row matrix of predictors X with high leverage, and the only response Y that may contain outliers. In this case, the M-estimator is monotone reliable for computing a robust scale estimate and a re-descending M-estimate [12].

2. Review of Literature

Some researchers who study them are outliers Weisberg (see [13]) have proposed several graphical procedures such as normal probability plots and numerical procedures such as regression diagnostics for detecting outliers. Furthermore Wisnowskia et al. (see [15]) have studied the analysis of multiple outlier detection procedures for linear regression models, using Monte Carlo simulation to compare the performance and limitations of different approaches. In addition, Filzmoser et al. (see [6]) studied the identification of outliers in high dimensions.

Furthermore, Huber (see [7]) was introduced the concept of robustness in regression. One approach that is used as a robust estimation is the M-estimator, which is based on Maximum Likelihood Estimation method (MLE). The main idea behind M-estimators is M-estimators work iteratively by changing the amount of residual squares of OLS with other functions. Cummins and Andrews (see [4]) refer to this estimator as a method of iteratively reweighted Least Squares (IRLS). This method can be applied to estimate the multivariate regression coefficients are robust in this study.

Morgenthaler et al. (see [9])) have discussed the robust response surface chemistry based on design of experiments. In addition, Hund et al. (see [8]) describe various outlier detection methods and their robustness evaluation using different experimental designs. Furthermore, Wiens and Wu (see [14]) propose a comparative study of M-estimators and presents a more optimal design than the regression model that allows. In addition, Maronna et al. (see [11]) describes the latest robust regression algorithm. Furthermore, Bashiri and Moslemi (see [2]) proposed a method of Iterative Weighting Moving Average (MAIW) to estimate the coefficients of the regression model based on the M-estimator. The aim of their research is to reduce the effect of the points wrong by considering previous data to detect outliers or the possibility of a trend in the residuals. Furthermore, Bashiri and Moslemi (see[2]) proposed a weighting method repeatedly (iterative weighting method) to modify abnormal outliers which follow the trend and residual variation that has no equal, so they have less effect on the estimated coefficients.

Many authors have presented robustness in multi-response problems, but the first to introduce the concept of robust design is Taguchi (see [12]). Furthermore, robust parameter design (robust parameter design) in the multi-response surface has been investigated among others by Myers et al. (see [10]) and Vuchkov & Boyadjieva (see [16])). This approach is often used in process improvement projects, to redesign the process in order to increase customer satisfaction by improving operational performance. In robust design, the model parameters are usually estimated by OLS.

Koksoy [10] presents the MSE as a criterion in the design of robust multi-response problem. In addition, it uses genetic algorithms and generalized reduced gradients to solve the proposed model, in which the combined array is presented as a general framework for problems in which data is collected. Furthermore, Quesada and Del Castillo [11], proposed a dual response approach to robust parameter design multivariate.

Have many robust estimators of multivariate regression models were examined as robust covariance estimator proposed by Maronna and Morgenthaler [12]. Furthermore, Koenker and Portnoy [9], proposed a multivariate regression method of M-type. Furthermore Rousseeuw et al. [23], developed a multivariate regression estimator that is efficient and useful, based on the minimum covariance determinant (MCD). Based on this procedure, the approach developed robust multivariate regression. This estimator is categorized as a high breakdown point robust algorithm, while the M-type estimators are not categorized as a high breakdown point algorithm. Although the breakdown points of this method at a high level, the efficiency of the M-type is significant to the other methods. Agullo et al. [1] have proposed an alternative robust estimator of multivariate regression is an extension of the setting least trimmed squared in multivariate methods.

3. Robust estimation of a multi-response surface

Robust estimation of regression coefficients in the multi-response problem is an important issue. Treating each response separately and apply a strong single response procedure can lead to erroneous interpretations of results. So, it is necessary to consider all responses simultaneously and estimate the variance-covariance matrix

The difference between the surface of a single response and multi-level response is a measure of the distance involved. In a single response problem using the Euclidean distance while the residual multi-response problems using Mahalanobis distance which takes into consideration the correlation between the responses. In MRSM approach, lower weights given to residual size greater distance. In each iteration, the proposed weighting function down-weights residuals by considering all responses simultaneously.

Defined variable r_{ij}, $i=1,2,3,\ldots,l$; $j=1,2,3,\ldots,p$ for residuals associated with the ith repetition of the response to the-j. Residuals for each response Y_j was first obtained by using the initials of the estimated response $\hat{Y}_j$ because $r_{ij} = Y_{ij} - \hat{Y}_{ij}$. Furthermore scaled residuals for each response expressed by sr_{ij},

$$sr_{ij} = \frac{r_{ij}-\bar{r}_j}{s_{r_j}} \tag{1}$$

Where $\bar{r}_j$ and s_{r_j} were the average residual samples residual standard deviation of the sample respectively.

Due to factors that can be controlled is assumed to be constant (not random), then the correlation between subsequent responses were estimated using the scaled residuals. This estimation is used to obtain the covariance matrix$\widehat{\boldsymbol{\Sigma}}$. Covariance matrix can be affected by outliers, so it must be estimated with robust estimation using M-estimator.

It is assumed that the p responses and $r(i) = \left[s_{r_{i1}}, s_{r_{i2}}, \ldots, s_{r_{ip}}\right]; i = 1,2,3, \ldots, l$

matrix of scaled residual response at the ith repetition, then calculated the Mahalanobis distance for each response in a repetition by using the following equation

$$\boldsymbol{d}(\boldsymbol{r}(\boldsymbol{i})) = \sqrt{(\boldsymbol{r}(\boldsymbol{i}))^{\boldsymbol{T}} \widehat{\boldsymbol{\Sigma}}^{-\mathbf{1}} \boldsymbol{r}(\boldsymbol{i})} \tag{2}$$

Distribution of Mahalanobis distance squared approximated by chi-square with p degrees of freedom (Montgomery 2005). The critical point of this distribution at the confidence level α or $\left(\chi^2_{p,\alpha}\right)$ was used as the assigned weight. In other words, if the squared Mahalanobis distance is smaller than $\chi^{\,2}_{\,p,\alpha}$ is weighted 1. As for the other weights derived from the proportion of the total distance is used equation (3).

$$w_i = \begin{cases} 1 & ; if\ d(r(i)) < \chi^2_{p,\alpha} \\ \frac{\chi^2_{p,\alpha}}{\sum_{i=1}^{l} d(r(i))} & ; otherwise \end{cases} \tag{3}$$

Flowchart of the approach is illustrated in figure 1 MRSM Furthermore, the performance of the estimation approach MRSM stated sum of squared errors (SSE) were investigated by using a numerical illustration. Error for the regression coefficient θ is defined by:

$$Error = \left(\theta - \hat{\theta}\right)$$

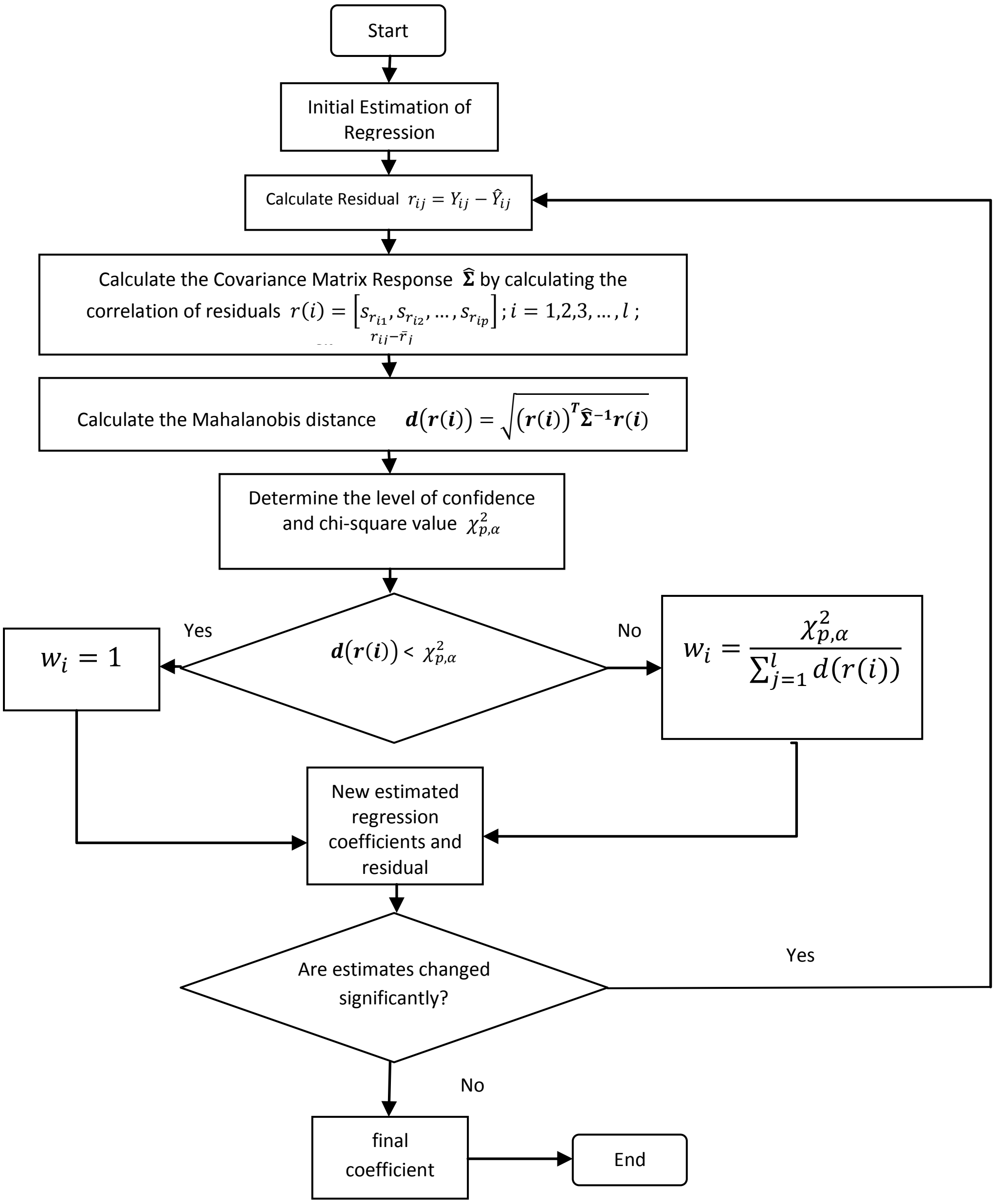
Start
Initial Estimation of Regression
Calculate Residual $r_{ij} = Y_{ij} - \hat{Y}_{ij}$
Calculate the Covariance Matrix Response $\hat{\Sigma}$ by calculating the correlation of residuals $r(i) = \left[s_{r_{i1}}, s_{r_{i2}}, \dots, s_{r_{ip}}\right]; i = 1,2,3,\dots,l;$
$r_{ij} - \bar{r}_i$
Calculate the Mahalanobis distance $d(r(i)) = \sqrt{(r(i))^T \hat{\Sigma}^{-1} r(i)}$
Determine the level of confidence and chi-square value $\chi^2_{p,\alpha}$
$d(r(i)) < \chi^2_{p,\alpha}$
Yes
$w_i = 1$
No
$w_i = \frac{\chi^2_{p,\alpha}}{\sum_{j=1}^{l} d(r(i))}$
New estimated regression coefficients and residual
Are estimates changed significantly?
Yes
No
final coefficient
End

Figure 1: Flowchart MRSM

4. Numerical illustration

Examples presented in this section are based on the experiments reported in Derringer, G., and Suich, [5]. In the real case, this study was conducted to determine the effect of several controllable factors such as hydrated silica level (X_1), silane coupling agent level (X_2), and sulfur level (X_3), and responses are elongation at break

(Y_1) and Abrasion Index (Y_2). Experimental design and data multi-response (pure data) are given in table 1.

Table 1: Experimental data of the tire tread compound problem

Experiment number	**X_1**	**X_2**	**X_3**	**Y_1**	**Y_2**
1	**-1**	**-1**	**1**	470	102
2	**1**	**-1**	**-1**	410	120
3	**-1**	**1**	**-1**	570	117
4	**1**	**1**	**1**	240	198
5	**-1**	**-1**	**-1**	640	103
6	**1**	**-1**	**1**	270	132
7	**-1**	**1**	**1**	410	132
8	**1**	**1**	**-1**	380	139
9	**-1.633**	**0**	**0**	590	102
10	**1.633**	**0**	**0**	260	154
11	**0**	**-1.633**	**0**	520	96
12	**0**	**1.633**	**0**	380	163
13	**0**	**0**	**-1.633**	520	116
14	**0**	**0**	**1.633**	290	153
15	**0**	**0**	**0**	380	133
16	**0**	**0**	**0**	380	133
17	**0**	**0**	**0**	430	140
18	**0**	**0**	**0**	430	142
19	**0**	**0**	**0**	390	145
20	**0**	**0**	**0**	390	142

$$X_1 = \frac{(phr\ silica - 1.2)}{0.5}\ ;\ X_2 = \frac{(phr\ silane - 50)}{10} \text{ and } X_3 = \frac{(phr\ sulfur - 2.3)}{0.5}$$

where

X_1, X_2, and X_3 are design levels
phr = part per hundred
$-1.633 \le X_i \le 1.633$, $i = 1, 2, 3$

Table 2 shown initials of the regression coefficients for each model (pure model) obtained using OLS.

Table 2: The regression coefficients for each model (pure model) obtained using OLS.

	Y_1	Y_2
X_1	400.3846	139.1192
X_2	-99.6664	16.49364
X_3	-31.3964	17.88077
X_1^2	-73.919	10.90654
X_2^2	7.932689	-4.0096
X_3^2	17.30761	-3.44711
X_1* X_2	0.432752	-1.57212
X_1* X_3	8.75	5.125
X_2* X_3	6.25	7.125

To investigate the performance of the approach MRSM, some outliers are added to the above experiment. This outlier is an experiment to 8 and to 12 for all the responses. Contaminated experimental data is shown in Table 3.

Table 3: Experimentally contaminated shown in bold

Experiment number	X_1	X_2	X_3	Y_1	Y_2
1	**-1**	**-1**	**1**	470	102
2	**1**	**-1**	**-1**	410	120
3	**-1**	**1**	**-1**	570	117
4	**1**	**1**	**1**	240	198
5	**-1**	**-1**	**-1**	640	103
6	**1**	**-1**	**1**	270	132
7	**-1**	**1**	**1**	410	132
8	**1**	**1**	**-1**	80	239
9	**-1.633**	**0**	**0**	590	102
10	**1.633**	**0**	**0**	260	154
11	**0**	**-1.633**	**0**	520	96
12	**0**	**1.633**	**0**	80	263
13	**0**	**0**	**-1.633**	520	116
14	**0**	**0**	**1.633**	290	153
15	**0**	**0**	**0**	380	133
16	**0**	**0**	**0**	380	133
17	**0**	**0**	**0**	430	140
18	**0**	**0**	**0**	430	142
19	**0**	**0**	**0**	390	145
20	**0**	**0**	**0**	390	142

Response surface contaminated based on the experimental design was modeled with OLS and MRSM. Table 4 shown the results of model estimation of pure and contaminated the data, assuming that the 95% confidence level.

Table 4: Estimation Results of Data Model Pure and Contaminated

	Y_1				Y_2			
	Pure Model	OLS	Individual	MRSM	Pure Model	OLS	Individual	MRSM
Constanta	400.385	399.231	399.535	399.88	139.119	139.504	139.474	139.219
X_1	-99.666	-122.166	-121.905	-110.774	16.494	23.994	23.994	23.443
X_2	-31.396	-90.639	-89.311	-88.765	17.881	37.628	37.311	37.301
X_3	-73.919	-51.419	-51.158	-53.213	10.907	3.407	3.407	4.907
X_1^2	7.933	12.259	11.119	8.333	-4.010	-5.452	-5.340	-4.232
X_2^2	17.308	-34.615	-32.848	-30.433	-3.447	13.860	13.487	13.112
X_3^2	0.433	4.760	3.619	3.765	-1.572	-3.014	-2.902	-2.876

X_1* X_2	8.750	-28.750	-28.399	-25.777	5.125	17.625	17.625	16.643
X_1* X_3	6.250	43.750	41.716	40.332	7.125	-5.375	-5.375	-4.475
X_2* X_3	1.250	38.750	39.101	38.312	7.875	-4.625	-4.625	-4.125

Information about the error and SSE for each method are given in Table 5

Table 5: Results of calculation errors and SSE of each method

	Y_1			**Y_2**		
	OLS	Individual	MRSM	OLS	Individual	MRSM
Constanta	1.331716	0.7225	0.255025	0.148225	0.126025	0.01
X_1	506.25	494.5731	123.3877	56.25	56.25	48.2886
X_2	3509.733	3354.147	3291.202	389.944	377.5249	377.1364
X_3	506.25	518.0631	428.7384	56.25	56.25	36
X_1^2	18.71428	10.1506	0.16	2.079364	1.7689	0.049284
X_2^2	2695.998	2515.624	2279.203	299.5322	286.7604	274.2005
X_3^2	18.72293	10.1506	11.10222	2.079364	1.7689	1.700416
X_1* X_2	1406.25	1380.048	1192.114	156.25	156.25	132.6643
X_1* X_3	1406.25	1257.837	1161.583	156.25	156.25	134.56
X_2* X_3	1406.25	1432.698	1373.592	156.25	156.25	144
SSE	11475.75	10974.02	**9861.337**	1275.033	1249.199	**1148.61**

Table 5 shows that the procedure in multivariate data containing outliers MRSM has the smallest SSE for the estimated coefficients on the entire response when compared with OLS and robust estimators individualized approach.

5. Conclusion

The results showed that the multivariate data containing outliers MRSM approach has performance or better efficiency when compared to the robust procedures individually and the OLS method. However, for wider application still needs further research using other measures of performance.

References

[1] Agullo J., Croux C. and Van Aelst S., (2008). *The multivariate least trimmed squares estimator*. Journal Multivariate Analysis 99: 311-338.

[2] Bashiri M., Moslemi A., (2011). *A robust moving average iterative weighting method to analyze the effect of outliers on the response surface design*. International Journal of Industrial Engineering Computations 2: 851-862.

[3] Box, G.E.P. dan Wilson, K.B., 1951, *On the experimental attainment of optimum conditions,* Journal of the Royal Statistical Society Series B, 13, 1-45.

[4] Cummins D J., Andrews C W., (1995). *Iteratively reweighted partial least squares: A performance analysis by Monte Carlo simulation*. Journal of Chemometrics; 9: 489-507.

[5] Derringer, G., and Suich, R., (1980) *Simultaneous optimization of several response variables.* J Qual Technol 12:214 – 219

[6] Filzmoser P., Maronna R., Werner M., (2008). *Outlier identification in high dimensions.* Computational Statistics & Data Analysis 52:1694-1711.

[7] Huber P J., (1981). *Robust Statistics.* New York: John Wiley & Sons.

[8] Hund E., Massart D L., Smeyers-Verbeke J., (2002). *Robust regression and outlier detection in the evaluation of robustness tests with different experimental designs.* Analytica Chimica Acta; 463: 53–73.

[9] Koenker R., Portnoy S., (1987). *L-estimation for linear models*. Journal of the American Statistical Association 82: 851–857.

[10] Koksoy O., (2008). *A nonlinear programming solution to robust multi-response quality problem.* Applied Mathematics and Computations 196: 603-612.

[11] Maronna R A., Martin, R D., Yohai V J., (2006). *Robust Statistics: Theory and Methods.* New York: John Wiley and Sons.

[12] Maronna R A., Morgenthaler S., (1986). *Robust regression through robust covariance*. Communications in Statistics-Theory and Methods 15: 1347–1365.

[9] Morgenthaler S., Schumacher M M., (1999). *Robust analysis of a response surface design.* Chemometrics and Intelligent Laboratory System 47: 127-141.

[10] Myers R H., Montgomery D C., Vining G G., Borror C M., Kowalski, S M., (2004). *Response surface methodology: A retrospective and literature survey*. Journal of Quality Technology 36: 53-77.

[11] Quesada G M., Del Castillo E., (2004). *A dual-response approach to the multivariate robust parameter design problem.* Technometrics 46: 176-187.

[12] Taguchi G., (1986*).* *Introduction to Quality Engineering*. Quality Resources, White Plains, NJ.

[13] Weisberg S., (1985). *Applied Linear Regression, 2nd Edition*, New York: John Wiley & Sons, Inc.

[14] Wiens D., Wu E K H., (2010). *A comparative study of robust designs for M-estimated regression models.* Computational Statistics and Data Analysis 54:1683-1695.

[15] Wisnowskia J W., Montgomery D C., Simpson J R, (2001). *A comparative analysis of multiple outlier detection procedures in the linear regression model*. Computational Statistics & Data Analysis 36: 351-382.

[16] Vuchkov I., Boyadjieva L., (2001). *Quality Improvement with Design of Experiment. In.* A. Keller, editor, Kluwer Academic Publishers: The Netherlands.