The Performance of Response Surface Methodology Based on the OLS and MM-Estimators for Second-Order Regression Model

Raja Rajeswari Ponnusamy

Abstract: Response surface methodology (RSM) is a set of statistical and mathematical techniques useful for developing, improving, and optimizing processes. RSM studies the relationship between responses and a set of input variables. The discussion in the study covers the use of second-order model to approximate this relationship. Analytical method and graphical method are the procedures used in solving a RSM problem. The study also presents the setting of central composite design (CCD) especially Central Composite Face Centred (CCF) in fitting a second-order model. This study proposed RSM using OLS and the MM-estimators to obtain the fitted regression models. The purpose here is to compare the performance of RSM based on OLS and robust MM-estimators in the second-order regression model. The same procedure applied to the contaminated dataset in order to find a robust regression estimator. A regression estimator is said to be robust if it is still reliable in the presence of outlier. The improvements relative to the MM method is illustrated by means of the parameter estimates for small, medium and large sample bias calculations, standard errors (SE), and root mean square errors (RMSE). A real data example analysis and simulations were employed in this study. It turns out that the performance of RSM based on MM-estimator is more efficient than the OLS-estimator in the absence of outliers for the real data analysis. Consequently, these results supported with the simulation analysis.

Keywords: Response surface methodology, MM-estimators, Robust regression estimator

I. INTRODUCTION

Response Surface Methodology (RSM) is a set of statistical and mathematical techniques useful for developing, improving, and optimizing processes. RSM can be defines as a statistical method that uses quantitative data from appropriate experiments to determine and simultaneously solve multivariate equations. It has significant roles in the design, development, and formulation of new products, as well as in the improvement of existing product designs(Myers and Montgomery, 2016).RSM is designed with a two-fold purpose. First, is to obtain the relationship between the values of some measurable response variables and the set of experimental factors (input variables) that presume to affect the response. Next is to obtain the values of the factors that yield the best value of the response. If discovering the best value of the response is beyond the available resources of the experiment, then RSM is used to at least gain a better understanding of the overall response system.

RSM is a sequential experimentation. In the first stage, the main goal is to determine whether the current conditions or levels of input factors are close to the optimum of the response surface or far from it. When the experimental region is near or within the optimum region, the second stage of the response surface study begins(Myers and Montgomery, 2016). Its main goal is to obtain an accurate approximation to the response surface in a small region around the optimum and to identify optimum process conditions.However, there is presently a widespread awareness of the dangers posed by the occurrence of outliers, which may result of keypunch errors, misplaced decimal points, recording or transmission errors, exceptional phenomenon such as earthquake or strikes, or members of different population slipping into the sample.

Outliers occur very frequently in real data, and they often go unnoticed because nowadays much data is processed by computers, without careful inspection or screening. Not only the response variable can be outlying, but also the explanatory part, leading to so called leverage points. Both types of outliers may totally spoil an Ordinary Least Square (OLS) analysis (Kutner et.al, 2013). Often, such influential points remain hidden to the user, because they do not always show up in the usual OLS residual plots.

Thus, to remedy this problem, a statistical techniques have been developed that are not so easily affected by outliers. There are the robust (or resistant) methods, the results of which remain trustworthy even if a certain amount of data is contaminated. Some people think that robust regression techniques hide the outliers, but the opposite is true because the outliers are far away from the robust fit and hence can detected by their large residuals from it, whereas the standardized residuals from OLS may not expose the outliers at all. Therefore, robust regression or methods is extremely useful in identifying outliers in RSM.

Problem Statement

This study contributes out how to investigate the response surface near the optimum yield in order to obtain the treatment combination that contributes to the maximum (or minimum) response, analyzed variable or yield. Furthermore, the formulation of RSM model, the principles of second order model and the use of Central Composite Design (CCD) in fitting second-order model for RSM will be discussed. Besides that, a better robust method is propose in RSM such as MM-estimator for better influences than the OLS in the presence of outliers. In another word it means that this study want to investigate the performance of RSM based on OLS and MM-estimator in the presence of outliers.

Manuscript published on 30 January 2019.
***Correspondence Author(s)**
Raja Rajeswari Ponnusamy, School of Mathematics, Actuaries and Quantitative Studies, Asia Pacific University of Technology and Innovation, Malaysia

II. METHODOLOGY

For this study purposes, the second-order model with k variables is used

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j=2}^{k} \sum \beta_{ij} x_i x_j + \varepsilon \qquad (1)$$

where x_i are input variables and β_i are regression coefficients which represents the slope of linear effect of the coded variables x_i. This model describes a hyper plane in the *k*-dimensional space of the input variables $\{x_j\}$. Model denoted in Equation (1) is the second-order response surface. A second-order model is used when the interaction terms and the quadratic terms are significant which means curvature exists in the response surface model.There are some basic assumption of RSM such that

- A structure, $y = f(\xi_1, \xi_2, ..., \xi_k)$ exist and is either complicated or unknown.
- The variables ξ_i are quantitative and continuous.
- The true function $f(\xi_i)$ can be approximated in the region of interested by a low-order polynomial.
- The design variables $x_1, x_2, ..., x_k$ are controlled and measured without error.

In RSM, the Central Composite Design (CCD) is the most popular class of region of second-order designs. In fact, the basic of CCD is derived from a sequential experiment and turns out to be an effective tool for non-sequential response surface experiment. CCD is an efficient design that is ideal for sequential experimentation and allows a reasonable amount of information for testing lack of fit while not involving a usually large number of design points. In general, there are three main types of CCD: circumscribed, face-centered, and inscribed(Myers and Montgomery, 2016). Each design has its own characteristics where the experimenter often chooses the design based on the region of interest and number of levels for each factors. In this study, Central Composite Face Centered (CCF) is chosen because it provides a cubical region and it requires three levels for each factors. The distance between axial point and the center point, α is equal to 1. For this reason, CCF is not a rotatable design.

Figure 1 presents a CCF design for two study variables: time and temperature. Both coded and natural variable level settings for time and temperature are shown in the figure. The design consists of a center point, four factorial points (corner points) and four axial points (points parallel to each variable axis on a circle of radius equal to 1.0 and origin at the center). The points in Figure 1 identify the variable level setting combinations that constitute the nine design points (experiment runs). As the value of α increases, the axial points extend beyond the faces of the square and the design region becomes more spherical (Dean and Voss, 2017).

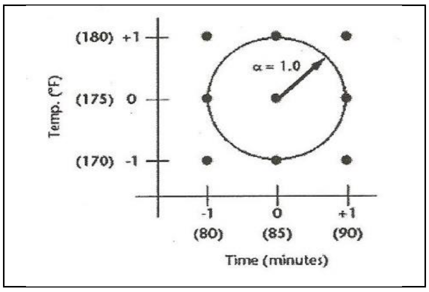


Figure 1: Two Variables CCF Design

Besides, two types of regression estimators were employed in RSM for these study which are the OLS and MM-estimators. The OLS estimator is a non-robust regression estimator because it is easily affected by outliers. The outliers will pull the least squares “fit” towards them too much where a resulting examination of the residuals will be misleading (Kutner et.al, 2013). The OLS estimator for $\boldsymbol{\beta}$ is

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \qquad (2)$$

Accordingly, to rectify this problem; MM-estimators is used so that the outliers have much less influence on the final estimates (Rousseeuw and Leroy, 2003). The MM-estimates for regression was introduced by Yohai in 1987. According to Maroonaet. al., (2006) these estimates are highly efficient and highly robust. Thus, for a given functionψ, the MM-estimator of regression solves as follows:

$$\frac{1}{n}\sum_{t=1}^{n} \psi\left(\frac{Y_t - X_t^{'}\beta_{MM}}{\hat{\sigma}_s}\right) X_t = 0 \qquad (3)$$

whereY_t be the response variable, X_t the ρ-vector of covariates, observed for $t = 1, ..., n$, $\hat{\sigma}_s$ is a scale estimate and β_{MM} is the MM-estimator.

III. RESULTS

A numerical example and some simulation studies are presented to illustrate the robustness of the estimators discussed in Section 3.0. For the numerical example, optimization of xanthan gum production by *X. campestric*in batch experiments was attempted using RSM and a CCF design where the simultaneous effect of the three independent variables (agitation rate (100-600 rpm), temperature (25-35 °C), time of cultivation (24-74 *h*)) were investigated for optimum xanthan and biomass production. RSM is used because it is suitable and appropriate in the study of main and interaction effects of the factors on the production of xanthan and biomass. For this study, a second ordered regression model was fitted and optimum conditions were estimated.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

In order to compare the performance of RSM based on the OLS and MM-estimator, the data was obtained from an experiment which done by Psomaset. al (2007) from Aristotle University of Thessaloniki. There are two response variables observed in this experiment which are xanthan gum (Y_1) and biomass (Y_2) production while the predictor variables or input factors are x_1 is agitation rate (rpm), x_2 is temperature (ºC) and x_3 is time of cultivation (h). It has been found that the production and the properties of xanthan gum and biomass are influenced by these three input factors. Hence, these variables are selected to find the optimized conditions for higher xanthan and biomass production using a CCF and RSM. The cultural conditions of agitation rate, temperature and time of cultivation in the xanthan and biomass production medium were varied accordingly to the experimental design as shown in Table 1.

Table 1: Levels Expressed In Coded and Natural Units.

Experimental Variables		Coded Unit (x_i)			$\Delta\xi$	$\bar{\xi}$
		-1	**0**	**1**		
Natural Unit	**Agitation Rate (rpm) (ξ_1)**	100	350	600	250	350
	Temperature ($^\circ C$) (ξ_2)	25	30	35	5	30
	Time (h) (ξ_3)	24	48	72	24	48

where $\Delta\xi$ is the increment of the experimental factor natural values corresponding to one unit of the coded variables.The coded unit is determined by the equation

$$x_i = \frac{(\xi_i - \bar{\xi})}{\Delta\xi} \qquad (4)$$

where x_i is the coded unit of the i^{th} independent variable, ξ_i is the natural value and $\bar{\xi}$ is the mean of the natural value. Based on Table 1, the coded unit for each level is -1, 0, and 1. This indicates that, the experimental design requires three levels for each factor. The OLS and MM-Estimators which are discussed earlier are used in the analysis to obtain the estimated coefficients in the second-order model. For both xanthan and biomass, all the OLS-estimated coefficients in the model are significant except for the contaminated datasets. But for MM-estimated coefficients in the model are significant for both clean and contaminated datasets. Table 2 and 3 show the second-order regression models based on OLS and MM-estimators for clean and contaminated datasets respectively.

Table 2: The Second-Order Regression Model for Clean Datasets.

Production	Estimators	Model
Xanthan	**OLS**	$\hat{y}_1 = 0.4779 + 0.1357x_1 - 0.0222x_2 + 0.1021x_3 - 0.1309x_1^2 + 0.0946x_2^2$ $-0.0519x_3^2 + 0.0173x_1x_2 + 0.0635x_1x_3 + 0.0188x_2x_3$
	MM	$\hat{y}_1 = 0.4776 + 0.1356x_1 - 0.0222x_2 + 0.1024x_3 - 0.1302x_1^2 + 0.0948x_2^2$ $-0.0523x_3^2 + 0.0172x_1x_2 + 0.0635x_1x_3 + 0.0187x_2x_3$
Biomass	**OLS**	$\hat{y}_2 = 0.2381 + 0.0596x_1 - 0.0448x_2 + 0.0366x_3 - 0.0346x_1^2 + 0.0144x_2^2$ $-0.0206x_3^2 - 0.0117x_1x_2 + 0.0083x_1x_3 + 0.0083x_2x_3$
	MM	$\hat{y}_2 = 0.2378 + 0.0596x_1 - 0.0448x_2 + 0.0366x_3 - 0.0344x_1^2 + 0.0145x_2^2$ $-0.0207x_3^2 - 0.0118x_1x_2 + 0.0082x_1x_3 + 0.0083x_2x_3$

Table 3: The Second-Order Regression Model for Contaminated Datasets

Production	Estimators	Model
Xanthan	**OLS**	$\hat{y}_1 = 47.4483 + 7.0558x_1 + 6.8979x_2 + 58.0711x_3 - 94.0716x_1^2 - 93.8461x_2^2$ $161.1219x_3^2 + 8.6674x_1x_2 + 8.7136x_1x_3 + 8.6689x_2x_3$
	MM	$\hat{y}_1 = 0.4814 + 0.1416x_1 - 0.0164x_2 + 0.1129x_3 - 0.1375x_1^2 + 0.0879x_2^2$ $-0.0338x_3^2 + 0.0246x_1x_2 + 0.0708x_1x_3 + 0.0261x_2x_3$
Biomass	**OLS**	$\hat{y}_2 = 30.5946 + 0.0596x_1 - 29.3155x_2 + 2.6403x_3 - 60.7478x_1^2 + 85.6547x_2^2$ $-47.7153x_3^2 - 0.0118x_1x_2 + 0.0083x_1x_3 + 0.0083x_2x_3$
	MM	$\hat{y}_2 = 0.2371 + 0.0596x_1 - 0.0460x_2 + 0.0345x_3 - 0.0330x_1^2 + 0.0222x_2^2$ $-0.0293x_3^2 - 0.0117x_1x_2 + 0.0082x_1x_3 + 0.0082x_2x_3$

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Based on the full second-order regression model obtained in Table 2 and Table 3 for clean and contaminated dataset respectivelyare used in RSM analysis for xanthan and biomass production. The contour and response surface plots are equally same for both OLS and MM-estimators for xantan and biomass production as in Figure 2 and Figure 3. These indications supported by the numerical results which gained in the Table 4. This table showsthat the optimum yields obtained for xanthan and biomass productions based on both estimators anticipate the same results except for OLS estimator using contaminated dataset.These imply that, the MM-estimator is more efficient than OLS-estimator when a dataset contains outliers. For OLS-estimator, the optimum yields are very large and misleading for both productions which contains contaminated dataset. But when using MM-estimator, similar results which are closer to the results as in the clean dataset was obtained (Table 4).

Table 4: Optimum Yields Using OLS and MM-Estimators for Clean and Contaminated Xanthan and Biomass Dataset

Dataset	Method	Xanthan Optimum Yield	Biomass Optimum Yield
Clean	OLS	0.615	0.2525
	MM	0.6147	0.2521
Contaminated	OLS	42.3831	28.1229
	MM	0.7214	0.251

A series of Monte Carlo Simulation were employed in this study to further assess the performance of the OLS and MM-estimators. The values of the standard errors (SE), biasedness, and root mean square errors (RMSE) of the parameter estimate in the second-order regression model were examined for small, medium and large sample size.The simulation studies yield that the MM-estimator (Table 6) performed better and more efficient in the contaminated dataset which are less biased, have smaller SE and RMSE compares to the OLS-estimator (Table 5). In addition, all these results are consistent for small, medium and large sample size.

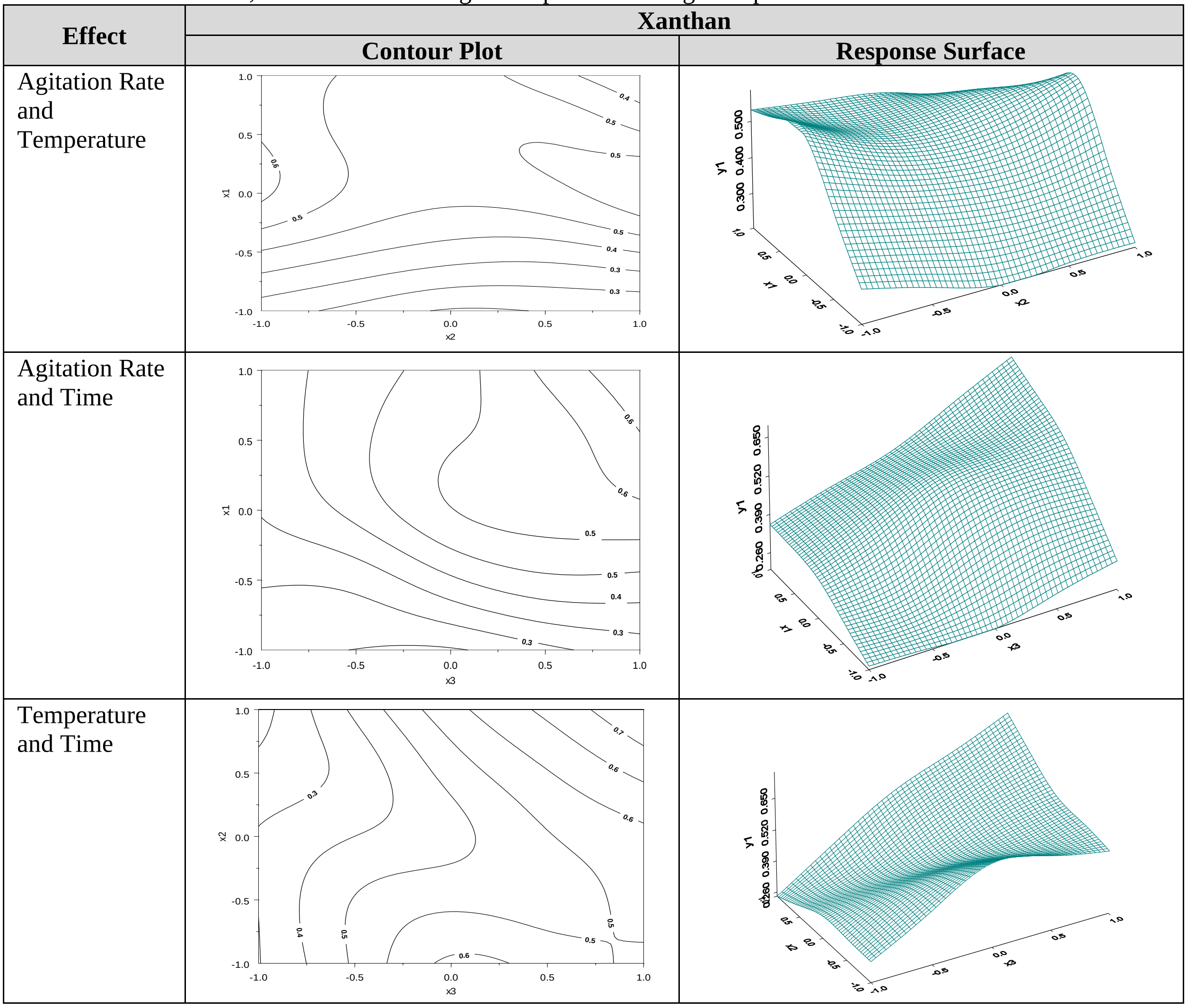


Figure 2: Xantan Contour Plot and Response Surface using OLS and MM-Estimator

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

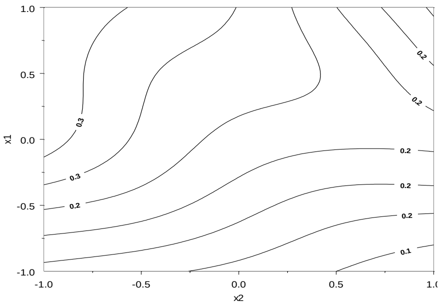
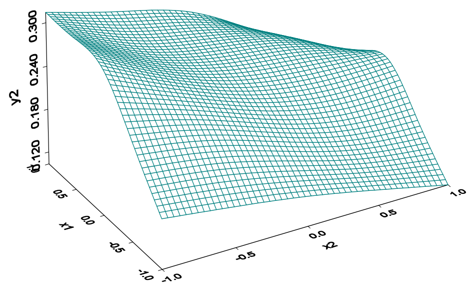
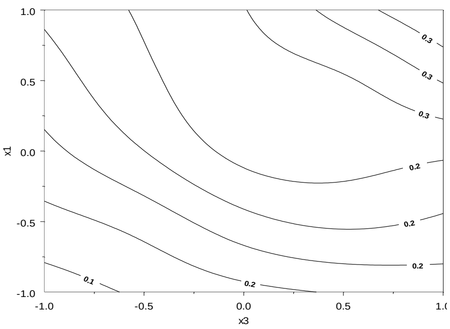
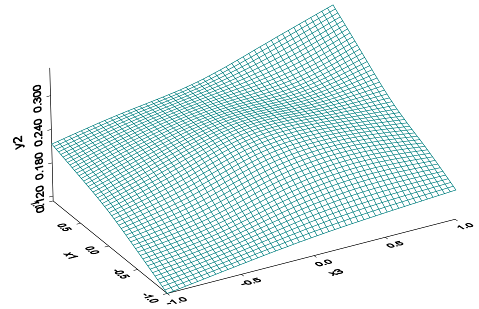
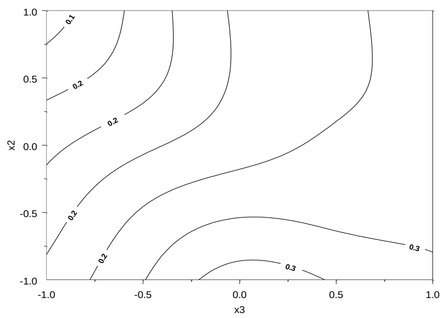
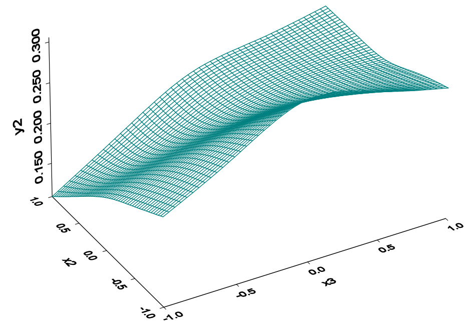
<table>
<tr><th rowspan="2">Effect</th><th colspan="2">Biomass</th></tr>
<tr><th>Contour Plot</th><th>Response Surface</th></tr>
<tr><td>Agitation Rate and Temperature</td><td> </td><td> </td></tr>
<tr><td>Agitation Rate and Time</td><td> </td><td> </td></tr>
<tr><td>Temperature and Time</td><td> </td><td> </td></tr>
</table>

Figure 3:Biomass Contour Plot and Response Surface using OLS and MM-Estimator

Table 5: Simulation Results using OLS estimator

Sample Size	Small			Medium			Large		
Methods	Bias	SE	RMSE	Bias	SE	RMSE	Bias	SE	RMSE
b_0	-112.120	92.300	145.225	-120.234	53.868	131.750	-124.268	33.733	128.765
b_1	6.632	4.547	8.041	6.091	2.105	6.444	6.132	1.347	6.278
b_2	5.671	7.755	9.607	6.144	4.263	7.479	6.372	2.745	6.938
b_3	5.658	3.835	6.836	5.359	1.721	5.629	5.353	1.088	5.463
b_{11}	0.576	6.967	6.991	1.518	3.579	3.888	1.637	2.475	2.967
b_{22}	0.927	6.838	6.901	1.507	3.602	3.904	1.657	2.455	2.962
b_{33}	1.330	6.948	7.074	1.572	3.649	3.973	1.680	2.565	3.066
b_{12}	1.420	6.649	6.799	1.561	3.741	4.053	1.654	2.427	2.937
b_{13}	0.990	6.930	7.000	1.493	3.720	4.008	1.557	2.442	2.896
b_{23}	1.169	7.499	7.590	1.542	3.607	3.923	1.759	2.509	3.064

Table 6: Simulation Results using MM-estimator

Sample Size	Small			Medium			Large		
Methods	Bias	SE	RMSE	Bias	SE	MSE	Bias	SE	MSE
b_0	0.111	2.351	2.353	0.027	0.968	0.968	-0.002	0.655	0.655
b_1	-0.005	0.153	0.153	-0.001	0.065	0.065	0.001	0.043	0.043
b_2	-0.006	0.146	0.147	0.005	0.064	0.064	0.002	0.041	0.041
b_3	0.000	0.153	0.153	-0.002	0.064	0.064	-0.001	0.043	0.043
b_{11}	-0.007	0.149	0.150	0.003	0.062	0.062	0.002	0.042	0.042
b_{22}	0.002	0.152	0.152	-0.002	0.066	0.066	-0.003	0.044	0.044
b_{33}	-0.006	0.149	0.149	-0.002	0.065	0.065	0.000	0.043	0.043
b_{12}	-0.001	0.144	0.144	-0.003	0.067	0.067	0.000	0.041	0.041
b_{13}	-0.003	0.148	0.148	-0.001	0.065	0.065	0.001	0.042	0.042
b_{23}	0.005	0.151	0.151	-0.003	0.066	0.066	-0.002	0.043	0.043

IV. CONCLUSION

In this study, the performance of RSM using the OLS-estimator and MM-estimator has been studied for clean and contaminated (with outliers) datasets. Both methods have been employed on a numerical example that anticipates the same results for the data without outliers. The results gain for yield optimum in RSM using OLS-estimator did not differ much from those estimates by MM-estimator for data without outliers. However, when the example numerical data has been modified to contain several outliers, the result obtained was the performance of RSM using MM-estimates performed better than the OLS-estimates. Besides, a series of simulation studies were then conducted in order for us to have a more general picture on the performance of both methods under different conditions. The simulation studies yield that in the presence of outliers in a dataset, the MM-estimators performed better which are less biased with smaller SE and RMSE compares to the OLS-estimators. This result is consistent for small, medium and large sample size. In the future, researches may want to apply the same techniques using other robust regression estimators that are available. This study can be higher order of multiple regression model using circumscribed and inscribedin the RSM.

REFERENCES

1. Dean, A. and Voss, D. (2017). *Design and Analysis of Experiments.* 2nd Edition. Springer.
2. Kutner, M.H., Nachtsheim, C.J., Neter, N., and William, U. (2013). *Applied Linear Regression Models.*5th Edition. McGraw-Hill.
3. Maroona, R., Martin, D., and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. John Wiley and Sons, Inc.
4. Myer R.H., and Montgomery D.C. (2016). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. 4th Edition. Canada: John Wiley and Sons.
5. Psomas, S.K., Liakopoulou-Kyriakides, M. and Kyriakidis, D.A. (2007). Optimization Study of Xanthan Gum Production Using Response Surface Methodology. *Biochemical Engineering Journal*. Vol.35. 273-280.
6. Rousseeuw, R. J. and Leroy, A. M. (2003). *Robust Regression and Outlier Detection*. New York: Wiley and Sons, Inc..

AUTHORS PROFILE

Raja Rajeswari Ponnusamy is working in School of Mathematics, Actuaries and Quantitative Studies, Asia Pacific University of Technology and Innovation, Malaysia.