IJSA (2021), Vol. 5, No. 2, 273-283
DOI: https://doi.org/10.29244/ijsa.v5i2p273-283

Indonesian Journal of Statistics and Its Applications
Available at https://ijsa.stats.id
e-ISSN: 2599-0802

Response Surface Model with Comparison of OLS Estimation and MM Estimation*

Salsabila Basalamah[1‡] and Edy Widodo[2]

[1,2]Department of Statistics, Islamic University of Indonesia, Indonesia
‡corresponding author: salsabilabasalamah@gmail.com

Abstract

Response Surface Method (RSM) is a collection of statistical techniques in the form of experiments and regression, as well as mathematics that is useful for developing, improving, and optimizing processes. In general, the determination of models in RSM is estimated by linear regression with Ordinary Least Square (OLS) estimation. However, OLS estimation is very weak in the presence of data identified as outliers, so in determining the RSM model a strong and resistant estimation is needed, namely robust regression. One estimation method in robust regression is the Method of Moment (MM) estimation. This study aims to compare the OLS estimation and MM estimation method to get the optimal point of response in this case study. Comparison of the best estimation models using the parameters MSE and R^2_{adj}. The results of MM estimation gives better results to the optimal response results in this case study.

Keywords: ordinary least square, response surface method, mm estimation.

* Received: Jan 2021; Reviewed: Jan 2021; Published: Jun 2021

1. Introduction

Surface Response Method (RSM) is a collection of statistical techniques in the experiments and regression. In general, RSM modeling using OLS estimation. OLS estimation is an estimation that works by minimizing the amount of residual squares. However, one disadvantage of OLS estimation is it very sensitive to any assumptions in the data. If one of the classical assumptions is not fulfilled, the OLS estimation cannot produce a good model. One of the reasons this assumption has not been fulfilled is because of the outliers in the research data. The existence of outlier data not only has an impact on the accuracy of the coefficient estimation results, but outlier data also produces large residuals in the model, the variance of data becomes large, and the data interval becomes wider(Soemartini, 2007). Therefore, choosing the right method for outlier data is very important.

One of the best regression methods for outlier data is robust regression. One study (Widodo et al., 2015) states that Robust analysis with M estimation shows better results than using OLS estimation. But M estimation has a disadvantage where it is not robust enough with leverage points or outliers on predictor variables (variable *x*), this is due to the breakdown point on this estimate is 0 (Chen, 2002). The robust estimation method that is able to overcome outliers on variables *x* and *y* is MM estimation.

MM estimation method is used when the distribution of residuals is not normal and the data contain outliers which can influence the model (Soemartini, 2007). This estimation is the result of a combination of S estimation with M estimation, making MM estimation has a breakdown point of 50% and high efficiency of approximately 95%. Having a breakdown point of 50% makes estimation of MM robust against outliers on independent and response variables. Evaluation of the estimated MM can also be seen in the regression linear of resulting standard error of estimate MM with the inclusion of outliers, the results are almost the same as the method of OLS estimates by eliminating outliers, so MM estimation can be a solution for the weaknesses of OLS estimation that is not robust with outliers data.

Based on the explanation above, the aim of this study is to compare the OLS estimation method and MM estimation on RSM using a simulation case study regarding the swelling capacity of snacks.

2. Materials and Methods

2.1. Experimental Design

Optimization Experimental design is a test or a series of tests that can use both descriptive statistics and inferential statistics (Mattjik & Sumertajaya, 2006). Aside from being a test, the experimental design is also a series of activities, where the stages in the series are truly defined. This activity aims to find answers of problems studied through a hypothesis testing (Hanafiah, 2012).

Another goal of the experimental design is to change the input variable into the output variable which is the response of the experiment (Mattjik & Sumertajaya, 2006), also to obtain and/or gather as much information as is needed and useful in conducting research based on the problem to be discussed.

2.2. Box Behnken Design

Box Behnken Design (BBD) has becen developed for fitting second order polynomial response surfaces. BBD has fewer experiment numbers than CCD. So, applying this design is popular in food processes for financial reasons. In a BBD, the experimental points are situated on a hyper sphere, equidistant from the central point (Nuryanti & Djati, 2008). Figure 1 shows that BBD do not contain any points at the vertices of the cubic region delimitated by the upper and lower levels of each factor (corner points), making this methodology advantageous when these points represent expensive or physically impossible experimental conditions.

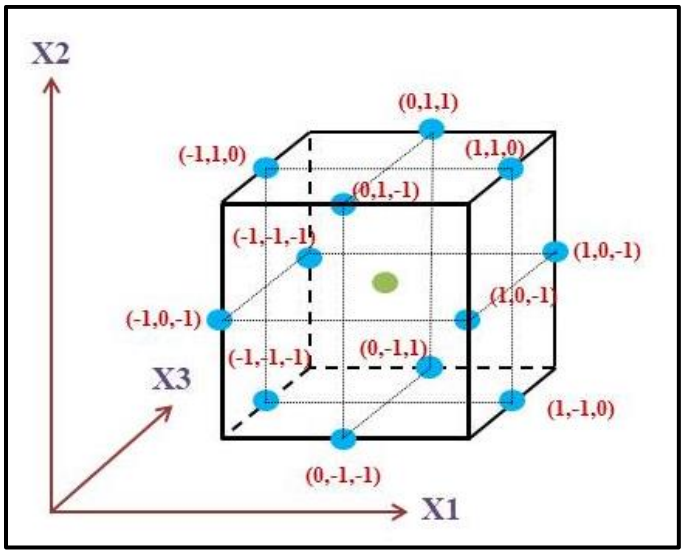

Figure 1: Box Behnken Design

BBD is not appropriate for studying factors with more than three levels (Myers et al., 2016). In other words, only triplex levels $(0,1,1)$, $(1,0,1)$ and $(1,1,0)$ are applied.

2.3. Outlier

Outlier is one or more data in a data set that very different compared to the overall data set pattern. Outlier data can be an influential observation, it means an observation that can affect the results of the estimated regression coefficients (Draper & Smith, 1998). Therefore, the action of a researcher to discard influential observations will significantly change the results of the regression equation and the conclusions produced.

The appearance of outlier in the data is caused by several possibilities, such as procedural errors in entering data, errors during measurement or analysis of data, and other causes that are really specific such as the perspective of a respondent on observations due to a reason beyond knowledge own researchers. The relation with regression analysis, an outlier data which is still forced to be analyzed can cause several things, including large residuals from the model formed; the variance in the data becomes greater; and data intervals have a wide range.

2.4. Linear Regression Analysis with OLS Estimation

Regression analysis aims to estimate and predict the value of other variables that are already known (Draper & Smith, 1998). In the regression analysis there is a prediction or estimation of the related variable value on independent variable value are more accurate because the regression results are predictive values. Therefore, the value is not necessarily right with the real value, the less deviation from the predicted value to the real value, the resulting regression equation is more precise with the real condition. The general equation is as follows:

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i,\ i = 1,2,3,\ldots,n \tag{1}$$

In general, regression analysis uses the least squares estimation method or Ordinary Least Square (OLS) to estimate the regression coefficient in the regression method. OLS estimation works by minimizing the Residual Sum of Square (SS_E) or minimizing $\sum_{i=1}^{m} e_i^2$ (Myers et al., 2016). Equation SS_E as follows:

$$SS_E = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_n x_{in})^2 \tag{2}$$

A series of equations 4 can be written in the matrix equation in equation 5:

$$X^T X\hat{\beta} = X^T Y \tag{3}$$

Then, the estimated value $\hat{\beta}$ can be found using these equations, where both segments are multiplied by the inverse of $X^T X$ who will get an equation 6:

$$\begin{aligned} &(X^T X)^{-1} X^T X\hat{\beta} = (X^T X)^{-1} X^T Y \\ &\hat{\beta} = (X^T X)^{-1} X^T Y \end{aligned} \tag{4}$$

2.5. Robust Regression Analysis with MM Estimation

Robust regression is a form of regression analysis designed to avoid some of the limitations of the parametric method. One of robust estimation method is the Method of Moment (MM) estimation. MM estimation was first introduced by (Rousseeuw & Yohai, 1984) for linear regression. The MM estimation method combines the results of S estimation and M estimation, making MM estimation has a breakdown point of 50% and high efficiency of approximately 95%. Having a breakdown point of 50% makes estimation of MM robust against outliers on independent and response variables. The evaluation of MM estimation can also be seen from the standard error produced by MM estimation with the inclusion of outliers, the results are almost the same as the OLS estimation method by removing outliers, so MM estimation can overcome the weakness of OLS estimation that is not robust with outlier data.

The first step in this estimation is to find the S estimate, then set the regression parameters using the M estimate. The S estimate guarantees a high breakdown point value and the M estimate makes the estimate have a high efficiency. In general, the Tukey Bisquare β function is used in both S estimation and M estimation. MM estimation uses Iteratively Reweighted Least Square (IRWLS) to find the estimated regression parameters. The MM estimation procedure can be described as follows (Soemartini, 2007):

- Estimating coefficient ($\hat{\beta}_j$), so that the residual value (e_i) is obtained from robust regression with high breakdown points.
- The residual value in the first step is used to calculate the estimated residual scale M, $\hat{\sigma}$ and also calculate the initial weight w_i.
- Then the residual value and residual scale in the second step are used in the initial iteration using the Weight Least Square (WLS) method to calculate the regression coefficient $\sum_{i=1}^{m} w_i \left(\frac{e_i}{\hat{\sigma}}\right) x_i = 0$, where w_i uses Huber or Tukey Bisquare weighting.
- The next step is to calculate the new weight w_i using the residuals from the initial iteration of WLS in the third step.

Steps 2, 3 and 4 are repeated (reiteration with the residual scale remains constant) until $\sum_{i=1}^{m}|e_i^r|$ converge, which is the difference β_j^{r+1} with β_j^n less than 10^{-4}, where r is the number of iterations.

2.6. Response Surface Method

Surface Response Method (RSM) is a collection of statistical techniques in the form of experiments and regression, as well as a collection of mathematical techniques that are useful for developing, improve, and optimizing processes (Myers et al., 2016). RSM has two objectives: to get paired independent variable that optimize the response variable and to get the model. The relationship between response (y) and the independent variable (x) can be written:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_{11} + \beta_{22} X_{22} + \beta_{33} X_{33} + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \varepsilon \quad (5)$$

where y:response variable; x_i:independent variable; ε:random residual component. In the RSM consists of two orders, namely order 1 and order 2. In order 1, the experimental design is sufficient to use a 2^k factorial design. Where 2^k factorial design each treatment has two levels of treatment. Meanwhile, in the second order using CCD trial design or using BBD.

2.7. Lack Of Fit Test

The lack of fit test is a test that sees the location of data increasing that might occur, then conducts an experiment with design points that are concentrated around the area that is suspected to be experiencing an increase. An experiment often raises produce two or more observations on the response to setting the same independent variable.

Repeated observations that occur can be checked by the lack of fit test (Myers et al., 2016).

In RSM parameter estimation to obtain an appropriate model requires lack of fit test. Order 1 equation will contain lack of fit which proves to load the optimal response point between each level of the factor being investigated, so that it will continue in order 2. While in order 2 there is no lack of fit then it can find the optimal response point (Myers et al., 2016). If the second order still contains lack of fit then move up to the next order. This means that this experiment is not a suitable trial using the RSM method.

2.8. Adjusted R-Square

Testing with adjusted R-Square or written R^2_{adj} objectively looks at the effect of adding an independent variable, is the variable able to strengthen the variation of responses. The value of R^2_{adj} obtained with (Myers et al., 2016):

$$R^2_{adj} = 1 - \frac{\frac{SS_E}{n-p}}{\frac{SS_T}{n-1}} = 1 - \frac{n-1}{n-p}(1-R^2) \tag{6}$$

3. Methods

The data used in this research is simulation data. Simulation data on the swelling snack capacity variable. The characteristics observed were frying temperature (X_1), frying time (X_2) and what percentage of African yam bean flour was used in the mixture (X_3). Observation of these three variables is to optimize capacity result of snacks swelling. The three variables consist of two levels, low and high, and have a central point. The frying temperature variable, which is at a temperature of 150^{o}C and 170^{o}C, with a center at 160^{o}C. While the frying time variable at 8 minutes and 12 minutes, with the center at 10 minutes. As well as variable lots of flour into the dough mixture that is 20% and 40%, with the center on 30%.

Data processing using R 3.5.2 response optimization software uses RSM by comparing two estimation methods in getting the model, namely OLS estimation and MM estimation. Comparisons are made by looking at the MSE parameters and R^2_{adj}.

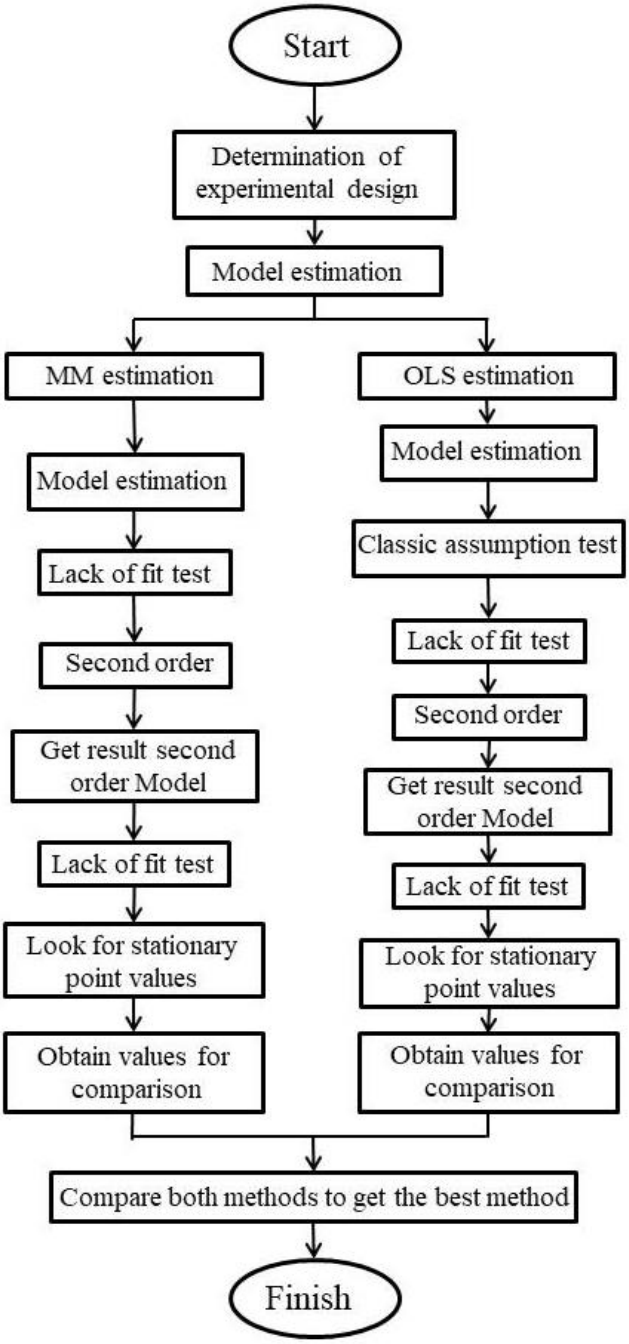

Figure 2: Flowchart

4. Result and Discussion

4.1. Normality Test

Normality test is conducted with purpose to assess the distribution of experimental data residuals, whether the distribution of residuals is normally distributed or not. Shapiro Wilk normality test is used in this normality test.

The Shapiro Wilk normality test method is an effective and valid normality test method used for small samples, with the following hypothesis:

H_0: Residual data are normally distributed.
H_1: Residual data are not normally distributed.

Test for normality with the Shapiro Wilk method using *p-value*, the decision is made based on the critical region, H_0 is rejected if the $p - value < \alpha: 5\%$. The result is $p - value = 0.01519$ it is less than $\alpha = 0.05$, so the H_0 is rejected and it means that the residual data is not normally distributed. Because the data are not normally distributed while the OLS estimation method requires that the assumption test be met, making the researcher conduct an outlier detection test.

4.2. Outlier Test

Outlier checking uses two methods, DFFITS and Plot Cook's Distance. Table 1 shows observations 4 and 12 which represent outlier observations using the DFFITS method.

Table 1: DFFITS Result

Observation	DFFITS	Value
4	1.39882524	>0.9428
12	2.23420063	

The results of observations are said to be outliers with the DFFITS method when values $|DFFITS| > 2\sqrt{\frac{4}{18}} = 0.9428$. Outcome of outliers observations by DFFITS in Table 1 found that there were 2 observations that contained outliers that direction to predictors and responses. We will use visualization to observe the outlier data using the Cook's Distance method, as illustrated in Figure 3.

The visualization results of Figure 3 obtained observation 4 and observation 12 crossing the red line or Cook's distance line, which means that both outliers observations are strengthened with the results of the DFFITS calculations that obtain the same results. After getting the data identified outlier researchers then conducted a study for outlier types obtained using visualization with the Cook's distance method.

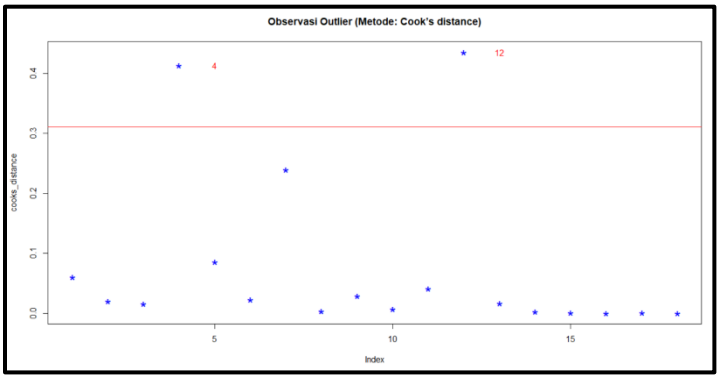

Figure 3: Plot Cook's Distance

The Cook's distance method can see how much outliers influence affects OLS estimation results, see how much outliers influence using leverage and standardized residuals as in Figure 4. Visualization results Figure 4 shows the dotted red lines are

Cook's distance limits, where observations that cross the line this is influential data in OLS estimation.

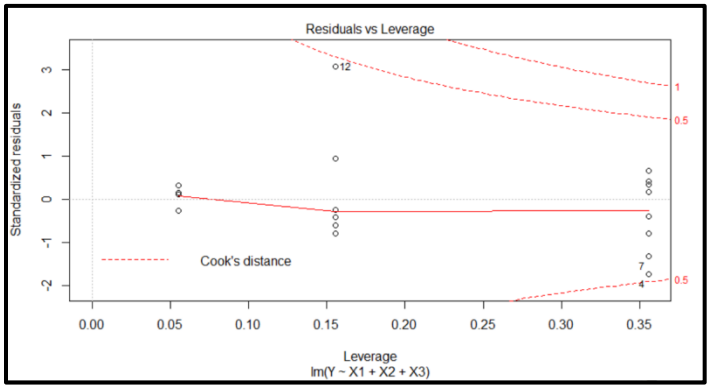

Figure 4: Plot Standardized Residual dan Leverage

Observation 12 and Observation 4 approach the line which makes this observation quite influential in OLS estimation and includes the type of Good Leverage Point.

4.3. Comparison

The comparison between OLS and MM estimation methods begin with comparing the coefficient and SE coefficient values written in Table 2. The result on Table 2 showing that MM estimation method is more stable because it has smaller error values compare with OLS estimation.

Table 2: Parameter Estimation Comparison Result

Predictor	*Coefficient*		SE *Coefficient*	
	OLS Estimation	MM Estimation	OLS Estimation	MM Estimation
Constanta	0.49652	0.456977	0.04443	0.014645
X_1	0.10340	0.103530	0.03572	0.009109
X_2	0.01960	-0.022585	0.03572	0.009632
X_3	0.10190	0.102045	0.03572	0.008968
X_1^2	-0.18005	-0.099974	0.06861	0.020115
X_2^2	0.24495	0.114343	0.06861	0.018541
X_3^2	-0.15155	-0.071549	0.06861	0.019425
X_1X_2	-0.01300	-0.012951	0.03993	0.009815
X_1X_3	0.10325	0.103237	0.03993	0.009807
X_2X_3	0.01600	0.016053	0.03993	0.009816

Table 3: Optimal Value Result

Model	Optimal Response	Optimal X_1	Optimal X_2	Optimal X_3
OLS Estimation	0.54%	164.26ºC	9.91 minutes	34.79 %
MM Estimation	0.62%	174.11ºC	10.11 minutes	47.38 %

The results of optimization with OLS and MM estimation methods have been done. The optimal response point for the OLS and MM estimation is written in Table 3. The optimization results of capacity swelling have an higher optimal value in MM estimation than the OLS estimation. Both of these methods have an optimal response value at the saddle point. The optimal point direction is seen from the canonical results. The canonical results of both methods in Table 4.

Table 4: Canonical Result

Model	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	Canonical Form
OLS Estimation	-0.2195717	-0.1122608	0.2451825	Saddle Point
MM Estimation	-0.13969580	-0.3226315	0.1147789	Saddle Point

The next step is to determine the best method for the experimental data identified as containing outliers. Comparisons were made by looking at the results of MSE and R^2_{adj} in each method. The MSE value is use because it can measure how good the estimate is, where the smaller the MSE value, the better the estimation result.While R^2_{adj} is also used as a comparison because it gives a picture of the quality of a regression line equation and provides a description of the ability to explain the model in solving problems.

Table 5: Comparison Results

Parameter	Method	
	OLS Estimation	MM Estimation
RSE	0.1129	0.05328
SSE	0.10197128	0.02271
MSE	0.01274641	0.0028387584
R^2_{adj}	66.86%	90.04%

The comparison results using MSE and R^2_{adj} values (Table 5) found that the MM estimation method is better or more accurate in the estimated parameters generated in RSM optimization in this case study, because it has an MSE value of 0.0028, this result is smaller than the MSE results on OLS estimation which has a result of 0.0127. MM estimation became the best estimate in this experiment because the ability to

explain the model in solving problems was 90.04%, while the OLS estimation had the ability to explain the model in solving problems by 66.86%.

5. Conclusion

The results comparison between OLS and MM estimation in the RSM case study shows that the best optimization method is MM estimation method, because all standard errors in MM estimation have smaller value compared to OLS estimation. MM estimation is also the best method seen from MSE and . The MM method is more accurate in parameter estimates because it has a MSE value of 0.0028, this result is smaller than MSE in the OLS estimation of 0.0127 and MM estimation can explain the model in solving problems by 90.04%, greater than the OLS estimation which can explain the finishing model the problem is 66.86%. MM estimation has the optimum point of response at 0.62%, while the optimal point for OLS estimation is 0.54%.

References

Chen, C. (2002). Paper 265-27 Robust regression and outlier detection with the ROBUSTREG procedure. *Proceedings of the Proceedings of the Twenty-Seventh Annual SAS Users Group International Conference*.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (Vol. 326). John Wiley & Sons.

Hanafiah, K. A. (2012). Rancangan percobaan teori dan aplikasi edisi ketiga. *PT Raja Grafindo Persada, Jakarta*, *260*.

Mattjik, A. A., & Sumertajaya, I. (2006). Perancangan percobaan. *IPBPres, Bogor*.

Myers, R. H., Montgomery, D. C., & Anderson-Cook, C. M. (2016). *Response surface methodology: process and product optimization using designed experiments*. John Wiley & Sons.

Nuryanti, D., & Djati, H. (2008). Metode permukaan respon dan aplikasinya pada optimasi eksperimen kimia. *Risalah Lokakarya Komputasi Dalam Sains Dan Teknologi Nuklir*, 373–391.

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis* (pp. 256–272). Springer.

Soemartini. (2007). *Pencilan (Outlier)*. Bandung: UNPAD.

Widodo, E., Guritno, S., & Haryatmi, S. (2015). Response Surface Models with Data Outliers through a Case Study. *Applied Mathematical Sciences*, *9*(37): 1803–1812.