Food Science of Animal Resources

Food Sci. Anim. Resour. 2019 February 39(1):114~120
DOI https://doi.org/10.5851/kosfa.2019.e9

pISSN : 2636-0772 eISSN : 2636-0780
http://www.kosfaj.org

ARTICLE

Improving the Quality of Response Surface Analysis of an Experiment for Coffee-Supplemented Milk Beverage: I. Data Screening at the Center Point and Maximum Possible R-Square

Sungsue Rheem[1] and Sejong Oh[2,*]

[1]Graduate School of Public Administration, Korea University, Sejong 30019, Korea
[2]Division of Animal Science, Chonnam National University, Gwangju 61186, Korea

OPEN ACCESS

Received January 24, 2019
Revised January 31, 2019
Accepted February 1, 2019

***Corresponding author** : Sejong Oh
Division of Animal Science, Chonnam National University, Gwangju 61186, Korea
Tel: +82-62-530-2116
Fax: +82-62-530-2129
E-mail: soh@jnu.ac.kr

***ORCID**
Sungsue Rheem
https://orcid.org/0000-0001-7009-3343
Sejong Oh
https://orcid.org/0000-0002-5870-3038

Abstract Response surface methodology (RSM) is a useful set of statistical techniques for modeling and optimizing responses in research studies of food science. As a design for a response surface experiment, a central composite design (CCD) with multiple runs at the center point is frequently used. However, sometimes there exist situations where some among the responses at the center point are outliers and these outliers are overlooked. Since the responses from center runs are those from the same experimental conditions, there should be no outliers at the center point. Outliers at the center point ruin statistical analysis. Thus, the responses at the center point need to be looked at, and if outliers are observed, they have to be examined. If the reasons for the outliers are not errors in measuring or typing, such outliers need to be deleted. If the outliers are due to such errors, they have to be corrected. Through a re-analysis of a dataset published in the *Korean Journal for Food Science of Animal Resources*, we have shown that outlier elimination resulted in the increase of the maximum possible R-square that the modeling of the data can obtain, which enables us to improve the quality of response surface analysis.

Keywords response surface methodology, central composite design, center runs, outlier elimination, maximum possible R-square

Introduction

In food science of animal resources, for the design and analysis of experiments, response surface methodology (RSM) is frequently used. RSM is a collection of statistical methods to design experiments, model data, and optimize responses (Myers et al., 2009). Among experimental plans in RSM, central composite designs (CCD, Box and Wilson, 1951) are most popular.

A CCD consists of the three portions that are factorial runs, axial runs, and center runs. Among these portions, center runs, which are the experimental runs at the center

point, give some desirable properties to the design, and allow us to measure the amount of variation in the responses at the center point, providing a basis for the lack-of-fit test.

Thus, at the center point, a reasonable amount of variation in the responses is anticipated, which is measured as the pure error variance. This means that there should be no outliers among the responses at the center runs. If outliers, which are the observations extremely different from others, exist at the center point, it may imply that, at the runs that have produced outliers, there have occurred failures in keeping the experimental conditions homogeneous.

The bad influence of an outlier at the center point on statistical modeling is enormous! It is much worse than researchers think. It simply ruins statistical modeling, which views a response as the function plus a variation. An outlier at the center point is seriously detrimental to both estimating the function and measuring the amount of variation. If statistical modeling were cooking and raw data were ingredients, outliers at the center point would be a poisonous, toxic ingredient. They must be either corrected or eliminated. Even one outlier can ruin the statistical analysis of the data that were obtained through a research into which a lot of expenses and manpower is invested. It is possible that just one outlier makes the result of a highly cost research unreliable.

A remedy for the situation where outliers are observed at the center point is simple. It is either the correction or the elimination of such outliers. Therefore, we suggest the following: “Before fitting a statistical model, look at the responses at the center point. If the outliers are observed at the center point, examine them. If they are due to errors in measuring or typing, correct them, otherwise, delete them.” This is what we mean by data screening at the center point.

This research was motivated by our seeing a situation where a serious outlier at the center point exists and it is overlooked, and, accordingly, a statistically insignificant model was fitted to the data. Such a situation took place in Ahn et al. (2017), whose data will be re-analyzed for the illustration of the remedy suggested in this research.

Materials and Methods

Dataset to be re-analyzed

How the elimination of outliers at center runs can improve the statistical model will be illustrated through re-analysis of a dataset described in the article entitled “Optimization of Manufacturing Conditions for Improving Storage Stability of Coffee-Supplemented Milk Beverage Using Response Surface Methodology” authored by Ahn et al. (2017). This article modeled two responses, using two factors. The two responses are the particle size and the zeta-potential of milk beverage. The two factors are the speed of primary homogenization (unit: rpm) and the concentration of emulsifier (unit: %), for which X_1 and X_2 are used as the coded factor. X_1=–1, 0, and 1 correspond to the speed of primary homogenization=5,000 rpm, 10,000 rpm, and 15,000 rpm, respectively, and X_2=–1, 0, and 1 correspond to the concentration of emulsifier=0.1%, 0.2%, and 0.3%, respectively. Among the two responses, the first response, which is the particle size, had an extreme outlier at the center point. Thus, this response is used as the Y variable in this research article.

The dataset to be re-analyzed is shown in Table 1. In this dataset, the experimental design is the CCD for two factors with an axial value of 1 and five center runs. Using this design, we can fit to the data statistical models including a full second-order model.

Statistical analysis

Data were analyzed using SAS software. SAS/STAT (2013) procedures were used for statistical modeling. Graphs were drawn using Minitab (2017).

Table 1. Experimental design in coded levels and responses

A. Data displayed in Ahn et al. (2017)					
Run			X_1	X_2	Y
1			−1	0	217.867
2			0	0	260.500*
3			0	0	186.433
4			1	1	219.767
5			0	0	181.933
6			−1	1	178.267
7			−1	−1	179.900
8			0	0	175.633
9			1	−1	179.533
10			1	0	178.367
11			0	1	182.167
12			0	0	180.333
13			0	−1	185.333
B. Data according to the standard order					
Standard order	Run	Design point	X_1	X_2	Y
1	7	1	–1	–1	179.900
2	6	2	–1	1	178.267
3	9	3	1	–1	179.533
4	4	4	1	1	219.767
5	1	5	–1	0	217.867
6	10	6	1	0	178.367
7	13	7	0	–1	185.333
8	11	8	0	1	182.167
9	2	9	0	0	260.500*
10	3	9	0	0	186.433
11	5	9	0	0	181.933
12	8	9	0	0	175.633
13	12	9	0	0	180.333
C. Data from which an outlier at the center point is deleted					
Standard order	Run	Design point	X_1	X_2	Y
1	7	1	–1	–1	179.900
2	6	2	–1	1	178.267
3	9	3	1	–1	179.533
4	4	4	1	1	219.767
5	1	5	–1	0	217.867
6	10	6	1	0	178.367
7	13	7	0	–1	185.333
8	11	8	0	1	182.167
10	3	9	0	0	186.433
11	5	9	0	0	181.933
12	8	9	0	0	175.633
13	12	9	0	0	180.333

* Outlier.

Results and Discussion

Looking at and examining the responses at the center point

First, let us look at the responses from center runs to find outliers. In Table 1A, which is in Ahn et al. (2017), displays observations according to the run number, and Table 1B displays observations according to the standard order. Which display makes it easier to find out the outliers at the center point? Obviously, the second display! Thus, we suggest presenting the data in the form of Table 1B before data modeling.

In Table 1B, runs with standard order numbers 9 through 13 are the center runs, and the responses from these runs are 260.5, 186.433, 181.933, 175.633, and 180.333. Among these values, obviously 260.5 is an outlier, and if the reason for this observation is not an error in measuring or typing, this value should be eliminated. 3D scatterplots in Fig. 1, which graphically compares the data containing an outlier and the data without an outlier, say that at the center point, 260.5 is an extreme outlier.

Now, let us see what happens if statistical models are fitted to the dataset that contains this extreme outlier.

Fitting statistical models to the original data

First, let us fit to the data the second-order polynomial regression model containing 2 linear, 2 quadratic, and 1 interaction terms by using the RSREG procedure of SAS/STAT. Results of analysis of variance for this model are shown in Table 2A.

The model in Table 2A has a very poor fit; its model p-value=0.9460 is so large that it is close to 1, whereas the p-value of an acceptable model is no larger than 0.05, and its r^2=0.1322 is so small!

Now, let us find the maximum r^2 that can be obtained through the statistical modeling of this original dataset. The r^2 of the one-way ANOVA (analysis of variance) model on 9 design points, which are designated in Table 1B, is such an r^2. This one-way ANOVA model is the fullest model among the models that can be fitted to this dataset. Table 2B displays the results from this modeling.

The model in Table 2B has also a poor fit; its model p-value=0.9622 is so high that it is near 1, and its r^2=0.3173 is still too low!

Thus, there is no way for the model displayed in Table 2A to be improved. Now, our remedy is to remove the observation with run number 2 that has the outlier 260.5 at the center point. Let us try it. Table 1C displays the dataset from which this outlier has been deleted.

(A) **(B)**

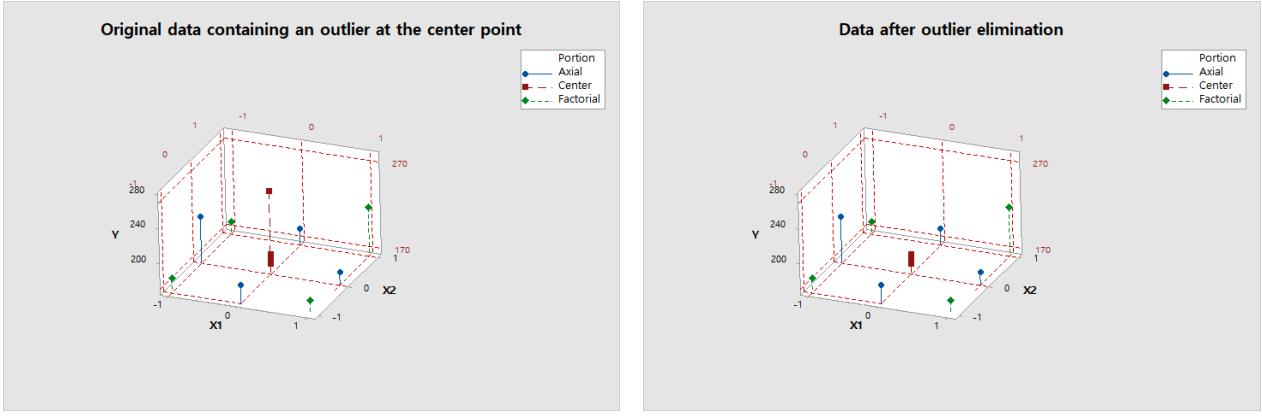

Fig. 1. 3D scatterplots of original data containing an outlier at the center point (A) and data after outlier elimination (B).

Table 2. **Analyses of variance for statistical models**

A. Analysis of variance for the 2nd-order model on the original data in Table 1A					
Model terms: X_1, X_2; X_1^2, X_2^2; X_1X_2					
Source	Degrees of freedom	Sum of squares	Mean square	F-value	p-value
Model	5	988.452	197.690	0.21	0.9460
Error	7	6,489.923	927.132		
Total	12	7,478.375			
Root MSE=30.4488		Coefficient of variation=15.80%		r^2=0.1322	

B. Analysis of variance for the one-way ANOVA model on the original data in Table 1A					
Model terms: Design points 1, 2, 3, 4, 5, 6, 7, 8, and 9					
Source	Degrees of freedom	Sum of squares	Mean square	F-value	p-value
Model	8	2,373.117	296.640	0.23	0.9622
Error	4	5,105.258	1,276.314		
Total	12	7,478.375			
Root MSE=35.7255		Coefficient of variation=18.53%		Maximum possible r^2=0.3173	

C. Analysis of variance for the 2nd-order model on the data in Table 1C					
Model terms: X_1, X_2; X_1^2, X_2^2; X_1X_2					
Source	Degrees of freedom	Sum of squares	Mean square	F-value	p-value
Model	5	991.581	198.316	0.78	0.5962
Error	6	1,517.422	252.904		
Total	11	2,509.003			
Root MSE=15.9029		Coefficient of variation=8.50%		r^2=0.3952	

D. Analysis of variance for the one-way ANOVA model on the data in Table 1C					
Model terms: Design points 1, 2, 3, 4, 5, 6, 7, 8, and 9					
Source	Degrees of freedom	Sum of squares	Mean square	F-value	p-value
Model	8	2,449.393	306.174	15.41	0.0230
Error	3	59.610	19.870		
Total	11	2,509.003			
Root MSE=4.4576		Coefficient of variation=2.38%		Maximum possible r^2=0.9762	

MSE, mean square error.

Fitting statistical models to the data from which an outlier is deleted

Now, let us fit to the data in Table 1C the second-order polynomial regression model containing 2 linear, 2 quadratic, and 1 interaction terms by using RSREG procedure of SAS/STAT. Results of analysis of variance for this model are shown in Table 1C.

The model in Table 2C is better than that in Table 2A, but still unsatisfactory; its model p-value=0.5962 is still large, and its r^2=0.3952 is still small! This implies that the second-order model is insufficient to represent this dataset. But, it may be

possible for the model to be improved enough to well explain data. To check this possibility, let us also find the maximum r^2 that can be obtained through the statistical modeling of this reduced dataset. The r^2 of the one-way ANOVA model on 9 design points, which are designated in Table 1C, is such an r^2. This one-way ANOVA model is the fullest model among the models that can be fitted to this dataset. Table 2D displays the results from this modeling.

Now, the model displayed in Table 2D is satisfactory in its ability to explain the data; its model p-value=0.0230 is lower than 0.05, meeting the criterion on the p-value for an acceptable model, and its r^2=0.9762 is large enough!

This implies that the model in Table 2C can be improved by adding some proper terms to the model. Now, such an augmentation of the model to improve it in the case of a cubic central composite design, which is our case, is another topic for research. This research has been done, and a satisfactory model has been found! However, to focus on one topic at a time, the presentation of the result of this research will be given in the next article.

Conclusion

The suggestion of this research is simple: “Look at center runs before setting up the final model that explains data. If there are outliers at the center point, examine them, and they are not due to the errors in measuring or typing, just get rid of them. If the outliers are due to such errors, correct them.” This suggestion is easy to implement, and will help enhance the quality of response surface analysis in sciences including food science of animal resources.

Conflicts of Interest

The authors declare no potential conflict of interest.

Acknowledgments

This work was financially supported by the Graduate School of Public Administration at Korea University in 2018, and the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Minister of Education, Science, and Technology (NRF-2016R1A2B4007519).

Author Contributions

Conceptualization: Rheem S, Oh S. Data curation: Rheem S. Formal analysis: Rheem S. Methodology: Rheem S, Oh S. Software: Rheem S. Validation: Rheem S. Investigation: Rheem S, Oh S. Writing - original draft: Rheem S, Oh S. Writing - review & editing: Rheem S, Oh S.

Ethics Approval

This article does not require IRB/IACUC approval because there are no human and animal participants.

References

Ahn SI, Park JH, Kim JH, Oh DK, Kim M, Chung D, Jhoo JW, Kim GY. 2017. Optimization of manufacturing conditions for

improving storage stability of coffee-supplemented milk beverage using response surface methodology. Korean J Food Sci Anim Resour 37:87-97.

Box GEP, Wilson KB. 1951. On the experimental attainment of optimum conditions. J R Stat Soc Ser B 13:1-38.

Minitab. 2017. Minitab 18. Minitab Inc., State College, PA, USA.

Myers RH, Montgomery DC, Anderson-Cook CM. 2009. Response surface methodology: Process and product optimization using designed experiments. 3rd ed. John Wiley & Sons, Hoboken, NJ, USA.

SAS. 2013. SAS/STAT user's guide. Release 6.04, SAS Institute, Inc., Cary, NC, USA.