applied sciences

MDPI

Article

Importance of Using Modern Regression Analysis for Response Surface Models in Science and Technology

Hsuan-Yu Chen [1] and Chiachung Chen [2,*]

[1] Africa Industrial Research Center, National Chung Hsing University, Taichung 40227, Taiwan; wakaharu37@gmail.com
[2] Department of Bio-Industrial Mechatronics Engineering, National Chung Hsing University, Taichung 40227, Taiwan
* Correspondence: ccchen@dragon.nchu.edu.tw; Tel.: +886-4-22857562

Abstract

Experimental design is important for researchers and those in other fields to find factors affecting an experimental response. The response surface methodology (RSM) is a special experimental design used to evaluate the significant factors influencing a process and confirm the optimum conditions for different factors. RSM models represent the relationship between the response and the influencing factors established with the regression analysis. Then these equations are used to produce the contour and response surface plots for observers to determine the optimization. The influence of regression techniques on model building has not been thoroughly studied. This study collected twenty-five datasets from the literature. The backward elimination procedure and *t*-test value of each variable were adopted to evaluate the significant effect on the response. Modern regression techniques were used. The results of this study present some problems of RSM studies in the previous literature, including using the complete equation without checking the statistical test, using the at-once variable deletion method to delete the variables whose *p*-values are higher than the preset value, the inconsistency between the proposed RSM equations and the contour and response surface plots, the misuse of the ANOVA table of the sequential model to keep all variables in the linear or square term without testing for each variable, the non-normal and non-constant variance conditions of datasets, and the finding of some influential data points. The suggestions for applying RSM for researchers are training in the modern regression technique, using the backward elimination technique for sequential variable selection, and increasing the sample numbers with three replicates for each run.

Keywords: response surface methodology; regression analysis; ANOVA; backward elimination method

check for updates

Academic Editor: Monika Gibis

Received: 4 May 2025
Revised: 22 June 2025
Accepted: 25 June 2025
Published: 26 June 2025

Citation: Chen, H.-Y.; Chen, C. Importance of Using Modern Regression Analysis for Response Surface Models in Science and Technology. *Appl. Sci.* **2025**, *15*, 7206. https://doi.org/10.3390/app15137206

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The industry urgently needs technology to seek optimization to improve system efficiency, increase quality, save energy, and reduce carbon emissions. A system's output or response is usually affected by several factors. Finding the optimum levels of these factors is called optimization. A simple method is to test the optimal levels of a factor and keep the other factors constant. This technique is impractical because of the interaction of other factors. To assess the effect of several factors on the response, the response surface methodology (RSM) is usually adopted to simultaneously assess the optimum conditions for several factors [1,2]. RSM has been a popular experimental design in the industry. RSM is a special experimental design used to improve the utilization of processing and develop

processes for new products. Statistical techniques help evaluate the factors influencing the process and confirm the optimum conditions for different factors [1,3].

RSM's special function is to test the experimental run for several factors with fewer samples. As the data are collected, the relationship between the system's response and its influencing factors can be established by regression analysis. Then, the effect of factors on the response is graphed. Researchers can observe the relationship through the linear or curved distributions of these figures. This graphical function led to "response surface methodology" being widely used [1–3].

The relationship between the response (y) and the input factors (x_1, x_2, …, x_k) is expressed as

$$y = f(x_1, x_2, \ldots, x_k) + \varepsilon \quad (1)$$

where y is the response, f is the unknown function in the response, x_1, x_2, …, and x_k are the independent variables, also called influencing factors, k is the number of factors, and ε is the model's error.

The mathematical model also includes linear, quadratic, and interaction effects.

If the process system involves two factors, x_1 and x_2, the form of the RSM equation is

$$y = b_o + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2 + \varepsilon \quad (2)$$

In the three-process-variable condition, x_1, x_2, and x_3, the mathematical equation of the RSM is

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{23}x_2x_3 + \varepsilon \quad (3)$$

For the four-factor cases, x_1, x_2, x_3, and x_4, the RSM equation is

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2 + b_{44}x_4^2 + b_{12}x_1x_2 + b_{13}x_1x_3 + b_{14}x_1x_4 + b_{23}x_2x_3 + b_{24}x_2x_4 + b_{34}x_3x_4 + \varepsilon \quad (4)$$

If the RSM equation includes all variables as in Equations (2)–(4), it is called the complete RSM equation.

The more factors there are, the more complex the form of RSM becomes. In the introduction of Burns et al. [4], the RSM method's development in the 1950s was described; the main advantage of this method is the reduced number of experiments. The unique feature of "response surface methodology" is deriving the optimum conditions with the graphical view provided by the fitting equation. The relationship between the independent variable and some dependent variables is established by regression analysis. This equation is an empirical model, not a theoretical one.

The experimental design is an experimental system with a combination of different levels of the influencing factor. These factors serve as the independent variables for the regression analysis. Experimental runs are a series of tests for the experiment. The response or output of experiments is the dependent variable.

Three experimental methodologies are used for response surface methodology in research: complete factorial design, Box–Behnken design, and central composite design [1,5–10].

The complete factor design (CCD) involves all related factors and levels. For example, if an experiment has three factors, each with three levels, and the experimental runs are $3^3 = 27$, the number of runs is high and the efficiency is lower than that of other methods [1,5–10].

The Box–Behnken design (BBD) adopts a specific subset of the factorial combination. The experiment ports are arranged at an equal hypersphere distance from the central point. This method is popular for evaluating factors' interaction. However, it is inappropriate for factors with some extreme points and only suitable for factors with three levels [1,5–10].

The central composite design (CCD) method is good at constructing a second-order model. It has three types of points: factorial points, a central point, and an axial point. It can be used for three or more levels and to consider some extreme points [1,5–10].

The experimental matrices in the three methods for RSM design are easily found in a textbook [1]. The number of tests required for the three test methods with three factors and three levels is 27 for the FFD method, 22 for the BBD method, and 13 for the CCD method. The replicates of the center ports for CCD may influence the sample numbers.

The criteria to evaluate the fitting agreement and predictive ability of the selection of models are essential aspects of regression analysis. The key is to find the variables' significant effect on the response. Many criteria have been used. For the fitting agreement, the criteria involve the F-value of ANOVA, coefficient of determination R^2, adjusted R^2 R^2_{adj}, lack-of-fit test, etc. The criteria to evaluate predictive ability are PRESS, predicted R-squared R^2_{pred}, and adequacy of precision. The VIF (variance inflation factor) is used to test the collinearity of these factors [1,5,11]. The "Materials and Methods" Section introduces the calculation equations and their meaning.

After evaluating the adequacy of the equation (RSM model), the final model is called an adequate regression equation. Two methods are used to determine the optimization. The first one uses the first partial derivatives of the response in terms of each factor. For example, if three factors are considered, three gradient equations of this equation are derived, and the three optimum conditions are solved with these derivative functions [1].

The second method is the graphic method, which provides the figures observed by researchers. The contour and 3D response surface plots are popular among researchers [1,3, 12,13]. The investigated technique has been illustrated in detail [14].

Accuracy and validity are the keys to finding reliable results regarding the factors influencing the response. They involve the factors that affect the response and the degree to which these influencing factors influence. However, the adequacy of the regression equations is not of concern to researchers. Many researchers use commercial software directly. Reza et al. [15] mention the importance of the model's validity and accuracy in their suggestions about the challenges associated with RSM application. The authors emphasized that the RSM model depends on the statistical technique and the importance of ensuring the accuracy and validity of the RSM model [15].

An excellent RSM textbook has been published [1]. Bas and Boyaci [2] provided an important concept of RSM. Asoo et al. [3] introduced the historical background of RSM. These papers involved science and technology, analytical methods, and process systems aplenty [16–27]. The topics of this literature included drying, extraction, fermentation, blending, mixing, and others [16–27].

The unique feature of response surface methodology is the graphical presentation of the relationship between the influential factors (variables) and the response. The contour plot represents the information in a two-dimensional form. The variation in the response under different conditions of influencing factors is indicated with the contour plot. The 3D response surface plot represents the information in three-dimensional form. The response is the z-coordinate, and the two influencing factors are the x- and y-coordinates. If a quadratic polynomial relationship does not exist, linear lines are plotted in this contour plot and a 3D response surface plot as the plateau. If the quadratic relationship is significant, two plots are presented with curves. The contour plot is usually used to observe visually the optimum levels of the influencing factors (the input variables) that can result in the maximum response [1–3].

The purpose of an RSM experiment in science and technology is to find the optimal conditions. If the RSM model is a linear equation, the contour plot and response surface plot indicate the response direction to the original design. Suppose that the RSM model is a quadratic equation. Both plots will indicate the maximum, minimum, or saddle conditions. Establishing an appropriate RSM equation is essential to ensure correct optimization conditions. The regression technique for RSM needs to be considered.

In the classical regression equation, the least squares method establishes the equation, and the estimated values of the parameter coefficients are directly used.

Modern regression techniques include considering the balance of the under- and overfitting of the model, testing the normality and the constant variance, using the criterion of predictive performance, and checking influential data points. These modern regression techniques are illustrated in Section 2, Materials and Methods.

To the best of the authors' knowledge, modern regression techniques have not been fully adopted to evaluate the correction and validity of RSM equations in research. This study collected twenty-five datasets from the literature related to RSM studies in science and technology. These datasets were used to evaluate adequate RSM equations using modern regression techniques. The issues of the RSM studies in the literature were discussed. Some suggestions for researchers who have used RSM were proposed to enhance their research ability in RSM.

2. Materials and Methods

2.1. Data Sources for the Equations of the Response Surface Methodology

Table 1 shows 25 datasets from research used to evaluate adequate RSM equations. The datasets adopted in these published papers were used to evaluate adequate RSM equations with modern regression analysis. The parameters and criteria were estimated using Sigma Plot v.14.0 (SPSS Inc., Chicago, IL, USA).

Table 1. Published data in the literature for evaluating the adequate equations of response surface methodology.

Study	Objects	No. of Data	Software	Model Evaluation	Criteria of Parameter Selection	Report Model	Plots
I. Two variables							
1. Chen et al. [28]	Poly-cornstarch-blended composite	13	Minitab Ver. 14.2	R^2, R^2_{adj} s	t-value, *p*-value	$y_{WSI} = f(x_1, x_2, x_1^2)$ $y_{WAI} = f(x_1, x_2, x_1^2)$ $y_{ML} = f(x_1, x_2, x_1^2, x_1x_2)$	Contour plots, Response surface plot
2. Milan-carrillo et al. [29]	Amaranth flour	13	Design Expert Ver. 7.0	R^2	*p*-value, Stepwise regression	Complete model	Contour plots, Response surface plot
3. Sinkhonde et al. [30]	Rubberized concrete with burnt clay brick powder	13	Not reported	Lack of fit	F-value, *p*-value	Complete model	Contour plots, Response surface plot
4. Diemer et al. [31]	Forced chicory roots	13	MOODE Ver. 12.0	R^2, R^2_{adj}	t-value, *p*-value	$y_{5\text{-}CQA} = f(x_1, x_2, x_1^2)$	Contour plots
5. Adeyanju et al. [32]	*Akara* Ogbomoso Snacks	13	Design Expert Ver. 6.0.1	R^2	*p*-value	y_{MC} = Complete model $y_{OC} = f(x_1, x_2)$ $y_{\Delta E} = f(x_1, x_2)$ y_{BF} = Complete model $y_S = f(x_1, x_2)$	Contour plots, Response surface plot
II. Three variables							
6. Nwabueze [33]	African breadfruit–corn–soy mixtures	15	Statistica	R^2	*p*-value	$Y_{TIA} = f(x_3, x_1^2)$ $Y_{phytic\ acid} = f(x_1^2)$ $Y_{tan} = f(x_3^2)$	Response surface plot
7. Bimark et al. [34]	Bioactive flavonoid compounds	20	Minitab Ver.14	Lack of fit, R^2, R^2_{adj}	*p*-value	Complete model	Response surface plot
8. Chen et al. [35]	Green asparagus juice	20	Design Expert, version not reported	R^2, R^2_{adj} Lack of fit	*p*-value	Complete model	Contour plots, Response surface plot
9. Gong et al. [36]	Defatted marigold residue	20	Microsoft Excel	R^2, R^2_{adj} Lack of fit	*p*-value, Stepwise regression	Complete model	Response surface plot
10. Chiu et al. [37]	Corn extruded with yam	15	Minitab 16	R^2, R^2_{adj} Lack of fit	*p*-value	Not reported	Response surface plot

Table 1. *Cont.*

Study	Objects	No. of Data	Software	Model Evaluation	Criteria of Parameter Selection	Report Model	Plots
11. Idrus et al. [38]	Virgin coconut oil	17	Design Expert Ver. 8.0	R^2, R^2_{adj} Lack of fit	*p*-value	$Y_{yield} = f(x_1, x_2, x_1^2, x_2^2, x_3^2)$ $Y_{FFA} = f(x_1, x_1^{2)}$ $Y_{AV} = f(x_2, x_2^{2)}$ $Y_{POV} = f(x_1, x_2, x_3, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2)$	Contour plots, Response surface plot
12. Hong et al. [39]	Pumpkin floor blends with corn	15	Design Expert Ver. 7.0	R^2, R^2_{adj} Lack of fit	*p*-value	Not reported, contour plot and response surface plots produced by complete model	Contour plots, Response surface plot
13. Wu et al. [40]	Biopolymer blend with Tapioca starch	15	Design Expert Ver.7.0	R^2, R^2_{adj}, Lack of fit	*p*-value	Complete model	Contour plots, Response surface plot
14. Yu et al. [41]	Piper nigrum microcapsules	17	Design Expert, version not reported	R^2, R^2_{adj}, Lack of fit	*p*-value	Complete model	Not reported
15. Tshizanga et al. [42]	Waste vegetable oil and eggshells	20	Design Expect Ver. 9.	R^2, R^2_{adj}, PRESS	F-value	Complete model	Contour plots, Response surface plot
16. Wu et al. [43]	Sichuan paocai	17	SPSS Ver.22.0	Lack of fit	*p*-value, Confidence interval (CI)	Complete model	Response surface plot
17. Savik and Gajic [44]	Green walnut husks	17	Design Expert 13.0.1.0	R^2, R^2_{adj}	F-value, *p*-value	Complete model	Contour plots, Response surface plot
III. Four variables							
18. Ahn et al. [45]	Seed oil	31	MINITAB Release 14	R^2	t-value, *p*-value	$Y_{EFF} = f(x_1, x_2, x_3, x_1^2, x_2^2)$	Response surface plot
19. Lee et al. [46]	Peanut sprout	31	SAS Ver. 9.0	Not reported	t-value, *p*-value	Complete model	Contour plots, Response surface plot
20. Yu et al. [47]	Peanut sprout	29	Design Expert Ver. 8.05b	R^2, Lack of fit	F-value, *p*-value	Complete model	Contour plots, Response surface plot
21. Javanbakht and Ghoreishi [48]	Lead removal from an aqueous solution	30	Design Expert Ver. 7.0.0	R^2, R^2_{adj}, R^2_{pred}	F-value, *p*-value	Complete model	Contour plots, Response surface plot
22. Yemis et al. [49]	Haskap extract and tannic acid	28	Design Expert	R^2, R^2_{adj}, PRESS, Lack of fit	F-value, *p*-value, Backward elimination	$y_{SI} = f(x_1, x_2, x_3, x_4, x_1x_2, x_1x_3, x_1^2, x_3^2, x_4^2)$	Contour plots, Response surface plot

Table 1. *Cont.*

Study	Objects	No. of Data	Software	Model Evaluation	Criteria of Parameter Selection	Report Model	Plots
23. Vega et al. [50]	Fruit by-product	60	Mathematica Ver.11.1.1.0	R^2, R^2_{adj}	Not reported	Complete model	Contour plots, Response surface plot
24. Hiranpradith et al. [51]	*Centella asiatica*	30	Design Expert Ver. 13.0	R^2, R^2_{adj}, R^2_{pred}	t-value	$y_{TPC} = f(x_1, x_2, x_4, x_2^2, x_4^2)$, $y_{TFC} = f(x_1, x_2, x_3, x_4, x_1x_4, x_1^2)$	Contour plots, Response surface plot
IV. Five variables							
25. Acikel et al. [52]	*Rhizopus delemar*	46	Design Expert Ver. 7.0	R^2, R^2_{adj}	Not reported	Complete model	Contour plots, Response surface plot

2.2. Model Building of the Response Surface Methodology

The purpose of the regression analysis includes variable screening, parameter estimation, model specification, and prediction [53]. The first three categories are related to each other. Multiple regression data involves the dependent variable (response) and several independent variables. To establish the relationship between the independent and dependent variables, the significant effect of the variables on the response must be evaluated using statistical techniques for optimization in science and technology. When researchers propose dependent variables, model specifications, or model types, the estimated parameter values can be calculated using the regression technique.

A typical multiple regression model is expressed as follows:

$$y_i = b_0 + b_1 x_1 + \cdots + b_i x_i + b_j x_j + \varepsilon_i \quad (5)$$

In the quadratic regression model, ${x_i}^2$ and the integration of $x_i x_j$ are treated as the variables in a multiple regression model, such as in

$$y_i = b_0 + b_1 x_1 + b_{11} {x_1}^2 + b_{22} {x_1}^2 + b_{12} x_1 x_2 + \varepsilon_i \quad (6)$$

where ε_i is the model error.

2.2.1. The Assumptions Involved in Regression Analysis

The regression model assumes that ε_i is uncorrelated from one observation to another, ε_i has a mean of zero, ε_i has constant variance, and the x_i terms are not random. The data distribution must be normal.

Due to the basic assumption, the nonstandard conditions involved in regression analysis include [53]

1. The overfitting or underfitting of the models;
2. The non-normal condition of the datasets;
3. Heterogeneous variance;
4. Some outliers in the data;
5. Multicollinearity.

These assumptions must be checked, and these five nonstandard conditions must be remedied, as they are the first concern of modern regression analysis. The second concern of modern regression is to classify the model's performance into model fitting ability and predictive ability. Two kinds of criteria are used to introduce classic and modern concepts in regression modeling [53–56]. The predictive performance and the trade-off between bias and variance for selecting a limited number of important variables are illustrated.

2.2.2. Establishment of the Model

The key purpose of the response surface methodology is its prediction ability. When an adequate equation is established, the contour plot and 3D surface plot are plotted, and researchers make conclusions with these figures. If the selection of the regression equation is inappropriate, the conclusion and suggestion on the effect of the dependent variable (factors and levels) are meaningless.

Establishing the regression equation is called model building. This is an essential topic in regression analysis. Montgomery et al. [57] recommend these strategies:

1. Fit the complete model, which includes all variables.
2. Perform a thorough analysis of this model.
3. Transform the response (y_i) or some variables (x_i) if necessary.
4. Use the *t*-test or F-test on these individual variables.

5. Check the adequacy of the RSM model.

2.3. The Three Methods of Sequential Variable Selection

Three methods for performing sequential variable selection are forward selection, stepwise regression, and backward elimination [53,57–62].

1. Forward selection

In this procedure, all variables are selected as a single regressor with a constant term, and the variable that produces the largest R^2 is selected as the first variable. Then, the other variables are used as candidates for the second variable in terms of the constant and the first variable. The variable of the second term that produces the largest R^2 is selected as the second variable. A similar procedure is used to select the third variable, and then the procedure is continued. The selected variable's partial F-value or *p*-value is compared with the preset value. If the t-values of this variable are less than the preselected value, the procedure ends.

2. Stepwise regression

This method is a modification of forward selection. Two cutoff values of the partial F-value or *p*-value are preset. At each step, the previously selected variables in the equation are reevaluated with their partial F-value or *p*-value. If these values for a variable are larger than the cutoff *p*-value, the variable is excluded from the equation. In this method, one variable can be entered at each stage, and another can be eliminated.

3. Backward elimination

The procedures for backward elimination are as follows:

a. Fit all variables for the regression equation. Determine the t-value and *p*-value for each variable in this model [58].
b. Focus on the variable with the lowest observed t-values and its *p*-value.
c. Compare the *p*-value with a preselected significance level, usually $p < 0.05$.
d. Remove the variable if its *p*-value exceeds the preselected value [58].
e. Recompute the regression equation for the remaining variables and find the variable with the lowest t-value and highest *p*-value.
f. Repeat the backward elimination procedures of c, d, and e.
g. If no variable is dropped, the procedure ends. The regression model's selection consists of all remaining variables.
h. Perform the influential data point test, and the normality and constant variance tests.

Theoretically, the final regression models of the above three procedures—forward selection, stepwise regression, and backward elimination—should be the same. However, the selected levels of significance and the collinearity among these variables influence the selection of model variables for forward and stepwise procedures.

The disadvantage of the forward procedure is that the critical t-values are not strictly appropriate in the early stages [53]. The backward elimination procedure is recommended because this technique selects all possible explanatory variables and eliminates those of little importance to explain the variation in response y step by step [57,61–63].

The special feature of the RSM model is its multicollinearity. These derived variables of square variables and the interaction of variables, such as x_1^2, x_2^2, x_3^2, x_1x_2, x_1x_3, and x_2x_3, are prone to multicollinearity problems [56,61,63]. Rowley [64] emphasized that backward elimination is particularly useful in the collinearity problem.

In this study, a *t*-test statistic is used to assess the statistical interpretation of a variable in the regression model. A detailed explanation of this procedure has been introduced [6,63].

2.4. The Criteria for the Evaluation of RSM Equations

1. R-squared

The R^2 value is called the coefficient of determination. An R^2 value near 1.0 shows that the equation is very good at determining the relationship between the response and the independent variables.

2. Adjustable R^2

The adjusted R^2 considers the effect of the number of independent variables. Like R^2, an R^2_{adj} value closer to 1.0 indicates a regression equation's good descriptive ability.

3. Standard error of the estimated value, s

The s value indicates the actual variability in the equation with the data distribution between the response and independent variables.

4. t-value

The t-value of the variables is called the t-statistic. It is used to test the null hypothesis that the coefficients of independent variables are zero. A large t-value of the independent variable reveals that the coefficient is not zero and is valid. That is, the variable is effective.

5. *p*-value

The *p*-value of a variable coefficient is calculated from its t-value. The *p*-value serves as the probability of being incorrect to determine whether the variable coefficient is not zero. A smaller *p*-value represents a greater probability of the validity of the variable.

6. PRESS, the Predicted Residual Error Sum of Squares

This statistic is used to evaluate the predictive ability of the regression model. The calculation of PRESS is explained as follows:

To calculate the PRESS value of a regression model, for n samples, the first dataset (y_1) is removed, and the remaining n-1 datasets are used to calculate the regression equation. Then the first data point is substituted into this equation to find the predictive value $\hat{y}_{1,-1}$. The predictive error is called the first PRESS residual. In other words, the predictive error for dataset 1 is $e_{-1,-1}$. The next step is to take out the second dataset, 2, and return the data x_1 to this dataset. The second regression equation is computed with datasets without the dataset y_2. The y_2 dataset is substituted in Equation (2) to calculate the predictive value, and the predictive error for dataset 2 is $e_{2,-2}$

The procedure is repeated n times for all data and produces a set of n PRESS residuals ($e_{1,-1}$, $e_{2,-2}$, …, $e_{n,-n}$). The PRESS statistic is calculated as the sum of the squares of the n PRESS residuals. A lower PRESS value of an RSM equation indicates that this equation has better predictive ability.

7. Normality test

The normality test assesses whether the datasets are normally distributed. The regression analysis technique assumes that residuals are normally distributed about the regression line. In this study, the normality test technique used is the Kolmogorov–Smirnov method. The *p*-value calculated with this method is compared with the preset value ($p = 0.05$). Failure of the normality test reveals the inadequacy of the regression model.

8. Constant variance test

This test evaluates the constant variance of the dependent variable (response) in its population source. This study uses the Spearman Rank correlation method, and the *p*-value calculated by this method assesses the assumption of constant variance. The cutoff value is $p = 0.05$.

If the constant variance test is failed, different models with weighted values must be proposed, or the response (y_i) must be transformed to stabilize the variance.

9. Influential data point

Some statistics are used to observe influential data points. These suspicious data may be influencing data or outlier data.

a. Externally studentized residuals, t_i

The t_i value is computed with the standard error of the residual of the estimated value, where the data is not involved in the model building. Values of ± 2.0 are usually used to indicate the possibility of an outlier.

b. DFFITSi

This statistic is a criterion to reflect the prediction effect for a data point. It compares the estimated standard errors when the observed value is removed.

Usually, the cutoffs of DFFITSi are ± 2.0. The data point may be potentially influential if the criterion exceeds this threshold.

c. Cook's distance, D_i

This criterion evaluates the effect of each data point on the estimated values of the parameters in the regression model. The Di value will be more significant if a data point significantly affects the parameter values. The cutoff for the D_i value is 4 or an F-value equal to F (p, n − p, 50%), where p is the number of parameters and n is the number of data points.

2.5. The Meaning of the F-Test of the ANOVA Table

The dependent variables for a multiple regression equation are x_1, x_2, x_3, ...x_k. Suppose that one variable, xi, significantly affects the response yi by the test of a partial F-value or *p*-value of the ANOVA table. In this case, a multiple regression equation including this xi variable will be recognized as having a significant effect on the response by the F-value of the ANOVA table.

For example, if x_1 has a significant effect on the y response, two equations are proposed:

$$y_1 = b_0 + b_1 x_1 + b_i x_i + b_k x_k \quad (7)$$

$$y_2 = c_0 + c_1 x_1 + c_i x_i + c_j x_j \quad (8)$$

With the ANOVA table, both equations will significantly affect the y response with statistical tests like the F-test. However, this does not mean that other variables, such as x_i, x_j, x_k, etc., will significantly affect the response. Maybe the x_1 variables are the only significant factors.

This is the first typical misunderstanding in the application of RSM equations.

The second misunderstanding in applying RSM equations is using the sequential-model sum of squares of the response. A typical ANOVA model with a sequential-model sum of squares of the response for an RSM equation includes x_1, x_2, and x_3, which are listed in Table 2.

Table 2. Sequential-model sum of squares of response.

Source	df	SeqSS	MS	F-Value	*p*-Value
Regression (Mean)	df_m	SS_m	SS_m/df_m		
Linear	df_l	SS_l	SS_i/df_l	L_f	L_p
Square	df_s	SS_s	SS_s/df_s	S_f	S_p
Interaction	df_i	SS_i	SS_i/df_i	I_f	I_P
Residual Error	df_e	SS_e	SS_e/df_e		
Total		SS_t			

Note: df: degree of freedom; SeqSS: sequential sum of squares; MS: mean square.

The partial F-value or *p*-value is used to help the researcher conclude whether a significant effect of the linear, square, and integration term significantly affects the y response.

The trick to misusing these methods is to test the linear, interaction, and square terms with the F-value or *p*-value. For example, the square terms of an equation involving three variables x_1^2, x_2^2, and x_3^2 has a significant effect on the y response by the F-value or *p*-value of the ANOVA table, and the form of this complete equation, $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2$, may be used by researchers because the square terms have a significant effect on y.

However, besides this complete equation, other possible equations are

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{11}x_1^2 + b_{22}x_2^2 \quad (9)$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{11}x_1^2 + b_{33}x_3^2 \quad (10)$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{22}x_2^2 + b_{33}x_3^2 \quad (11)$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{11}x_1^2 \quad (12)$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{22}x_2^2 \quad (13)$$

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_{33}x_3^2 \quad (14)$$

This RSM equation has seven possible combinations (complete equation and Equations (9)–(14)). That is, if the square term significantly affects response y, the form of $b_{11}x_1^2 + b_{22}x_2^2 + b_{33}x_3^2$ is not the only possible equation for this RSM equation. All square terms (x_1^2, x_2^2, and x_3^2) must be evaluated individually.

2.6. The Effect of the Sampling Number

One advantage of response surface methodology is the small number of experimental runs for experiments. A smaller experimental sample can reduce the test cost and save time. However, the effect of the sampling number on the regression analysis was not mentioned by researchers who used this RSM technique.

Sample size is a criterion for ensuring the power of statistical techniques. Some complicated equations have been proposed to calculate the sample size for multiple regression [65–67]. Some easy-to-use sample size formulas have been proposed to evaluate the required sample size (n) for multiple regression equations.

1. Snee [68]

$$n \geq 2p + 20 \quad (15)$$

2. Green [69]

$$n \geq 8p + 50 \quad (16)$$

3. Khamis and Kepler [70]

$$n \geq 5p + 20 \quad (17)$$

4. Tabachnick and Fidell [71]

$$n \geq p + 104 \quad (18)$$

5. Zaarour [72]

$$n \geq 10p + 20 \quad (19)$$

where p is the number of parameters.

3. Results

3.1. Two Variables

3.1.1. Extrusion Process for Producing High-Antioxidant Instant Amaranth Flour

In Study [29], the process variables are x_1, temperature, and x_2, screw speed, and the response variables are y_{ORAC}, antioxidant capacity (ORAC), and y_{WSI}, water solubility index (WSI)—used in a central composite design including 13 runs formed by five central points.

The proposed equations for the y response are complete models; that is, $y_{ORAC} = b_o + b_1x_1 + b_2x_2 + b_{11}x_1^2 + b_{22}x_2^2 + b_{12}x_1x_2$ and $y_{WSI} = c_o + c_1x_{1+} c_2x_2 + c_{11}x_1^2 + c_{22}x_2^2 + c_{12}x_1x_2$.

Contour plots and response surface plots show the effect on y_{ORAC} and y_{WSI} of x_1 and x_2. The study presents all the curved relationships [29].

1. The y_{orac} response

The experimental data are listed in the study [29]; the multiple regression results for y_{orac} are

$$y_{orac} = 1481.845 + 24.670x_1 - 0.490x_2 - 0.0262x_1^2 + 0.0217x_2^2 - 0.0725x_1x_2 \quad (20)$$

$$(4.601) \quad (-0.10) \quad (-4.730) \quad (2.065) \quad (-2.396)$$

$$R^2 = 0.859, R^2_{adj} = 0.758, s = 122.503, PRESS = 162,811,173$$

The numeric values in parentheses below the estimated values of parameters are the t-values of the estimated values of parameters.

The normality test is passed (p = 0.285), and the constant variance test is passed (p = 0.295).

The estimated values of each independent variable and its criteria are listed in Table 3.

Table 3. The experimental values of each independent variable and its criteria for the y_{orac} response.

Coefficient	Estimated Value	Standard Error	t-Value	*p*-Value	Standard Coefficient	VIF
Constant	1481.845	571.865	2.591	0.036		
x1	24.670	5.362	4.601	0.002	8.922	186.628
x2	−0.490	4.456	−0.110	0.916	−0.108	48.234
$x1^2$	−0.0262	0.00555	−4.730	0.002	−4.705	49.088
$x2^2$	0.0217	0.0105	2.065	0.078	1.498	26.095
x1x2	−0.0725	0.0302	−2.396	0.048	−4.320	161.328

Because the x_1^2, x_2^2, and x_1x_2 variables are derived from x_1 and x_2, the five variables' variance inflation factor (VIF) is >10. This indicates the multicollinearity problems of these variables, and the backward elimination procedure is suitable for the RSM equations.

In a comparison of the t-values of x_1^2, x_2^2, and x_1x_2, the variable x_2^2 has the lowest t-value, and its *p*-value is 0.078 (>0.05). The term x_2^2 is deleted, and the regression equation is recalculated.

$$y_{ORAC} = 974.020 + 25.670x_1 + 6.138x_2 - 0.0285x_1^2 - 0.0725x_1x_2 \quad (21)$$

$$(4.039) \quad (1.673) \quad (-4.415) \quad (-2.019)$$

In a comparison of the t-values of x_1^2 and x_1x_2, the x_1x_2 variable has a lower t-value, and its *p*-value is 0.078. The variable of x_1x_2 is deleted, and the regression equation is recalculated.

$$y_{ORAC} = 2079.174 + 14.550x_1 - 1.109x_2 - 0.0285x_1^2 \quad (22)$$

$$(3.927) \quad (-1.257) \quad (-3.811)$$

The t-value of variable x_2 is −1.2557, and its *p*-value is 0.240 (> 0.05)

The variable x_2 is deleted, and the regression equation is recalculated.

$$y_{ORAC} = 1910.112 + 14.550x_1 - 0.0285x_1^2 \quad (23)$$

$$(3.818) \quad (-3.705)$$

$$R^2 = 0.597, R^2_{adj} = 0.517, s = 173.208, PRESS = 133,405,010.$$

The normality test is passed (p = 0.228), and the constant variance test is passed (p = 0.723). The final equation is called the adequate equation.

The R^2 values for the complete and final adequate equations are 0.859 and 0.597, respectively. However, the R^2 value is affected by the number of variables in the equation. The more variables are used, the higher the R^2. So, it cannot be used as the sole criterion to evaluate the fitting ability of the equation [53,54,57–60].

The PRESS value of this adequate equation is 133,405,010. This numeric value is lower than the complete model (PRESS = 162,811,173), indicating that adequate equations have better predictive ability.

The results of regression diagnostics showed that some influential data points were found. With the t_i value, the fourth and fifth data points are influential. The Cook's distance and $DFFITS_i$ of data point 6 are 1475.4 and 68.411, respectively. Further experiments should be performed to check the validity of these data points.

Figure 1 shows the contour plots for the complete and adequate equations. The difference in the equations induces a difference in the distribution of curves between the two figures. The contour and response surface plots produced with the complete equation were presented in the study [29]. The inadequate RSM equation could induce incorrect results.

2. The y_{wsi} response

The complete equation is

$$y_{wsi} = -3.679 + 0.333x_1 + 0.351x_2 - 0.00123x_1^2 - 0.00195x_2^2 + 0.00201x_1x_2 \quad (24)$$

$$(2.632) \quad (3.317) \quad (-9.347) \quad (-7.818) \quad (2.801)$$

$$R^2 = 0.961, R^2_{adj} = 0.932, s = 2.909, PRESS = 16,403.5$$

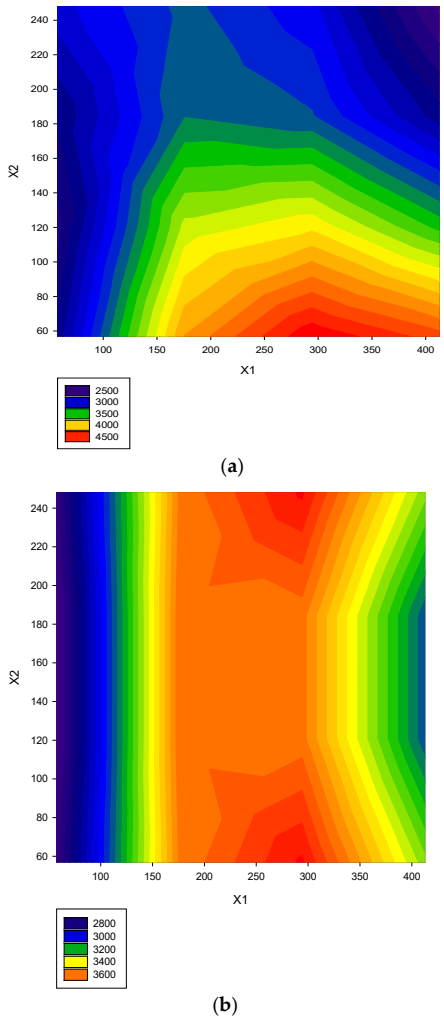


Figure 1. The contour plots for the complete and adequate equations of the y_{ORAC} response. The difference in the contour curves indicates the effect of the RSM equation on the relationship between response and influential variables. (**a**). Complete equation. (**b**). Adequate equation.

The normality test is passed (p = 0.236), and the constant variance test is passed (p = 0.723).

All the variables had a higher t-value. The variables x_1x_2 have the smallest t-value. However, the p-value is 0.026 ($p < 0.05$). So, the complete model is an adequate equation.

The results of regression diagnostics indicated some influential data points. For the tenth data point, t_i = 3.267, for the sixth data point, Di = 319.386, and DFFITSi = −42.548. The researchers showed that the runs of two data points should be performed with more replicates to find outliers or to recheck the validity of this model.

3.1.2. Compressive Strength of Rubberized Concrete

In Study [30], the experimental design was a CCD with 13 runs. The influencing factors included x_1 BCBP in %, and x_2 WTR in %. The two responses are y_{7D} (7-day compressive strength) and y_{28D} (28-day compressive strength).

The reported RSM equations for y_1 and y_2 are the complete equation; the independent variables involved are x_1, x_2, x_1^2, x_2^2, and x_1x_2 [30].

In the study, the contour and response surface plots of 7-day and 28-day compressive strength are curved distributions. The ANOVA table in the study for 7-day results showed that the p-values of x_1^2 and x_1x_2 were higher than $p < 0.05$. The ANOVA table for 28-day results indicated that the p-values of the x_1, x_1x_2, x_1^2, and x_2^2 variables were higher than the

cutoff value ($p < 0.05$). Despite the higher p-value indicated in the ANOVA table presented in the study, the complete equations are still selected and used to produce contour and response surface plots [30].

The procedure to evaluate the adequate equation of the y_1 response (7-day compression) is listed as follows:

$$1.\ y_{7D} = 25.310 + 0.458x_1 + 0.130x_2 - 0.0183{x_1}^2 - 0.0272{x_2}^2 - 0.00401x_1x_2 \quad (25)$$

$$(1.267)\quad (1.437)\quad (-0.287)\quad (-6.834)\quad (-0.303)$$

$$R^2 = 0.979,\ R^2_{adj} = 0.964,\ s = 0.664,\ PRESS = 12.070$$

The normality test is passed ($p = 0.522$), and the constant variance test is passed ($p = 0.220$).

The ${x_1}^2$ variable had the smallest t-value, and its p-value was higher than the cutoff value ($p < 0.05$). The variable was deleted.

The results of the recalculation are

$$2.\ y_{7D} = 25.352 + 0.549x_1 + 0.121x_2 - 0.0268{x_1}^2 - 0.00401x_1x_2 \quad (26)$$

$$(3.417)\quad (1.512)\quad (-7.730)\quad (-0.322)$$

$$R^2 = 0.978,\ R^2_{adj} = 0.968,\ s = 0.623,\ PRESS = 7.751$$

The x_1x_2 variable had the smallest t-value, and its p-value was larger than the cutoff value ($p < 0.05$). The variable was deleted.

The new equation is

$$3.\ y_{7D} = 25.452 + 0.509x_1 + 0.111x_2 - 0.0268{x_1}^2 \quad (27)$$

$$(5.277)\quad (1.586)\quad (-8.147)$$

$$R^2 = 0.978,\ R^2_{adj} = 0.971,\ s = 0.591,\ PRESS = 5.636$$

The normality test is passed ($p = 0.652$), and the constant variance test is passed ($p = 0.236$). No influential data point was found.

The adequate equation showed that x_1 has a curvilinear relationship with the response, and the quadratic equation x_2 variable has a linear relationship with response y_{7D}.

Compared with PRESS, the predictive ability of the adequate equation (PRESS = 5.636) is significantly improved over that of the complete equation (PRESS = 12.070).

Figure 2 shows the contour plots produced with complete or adequate equations. The distribution of each figure presents different results. When researchers use visual methods to conduct their experiments and observe the variables' effect on the response, inadequate RSM equations will induce incorrect conclusions.

The procedure to evaluate the adequate equation of the y_{28D} response (28-day compression) is listed in Appendix A.1.

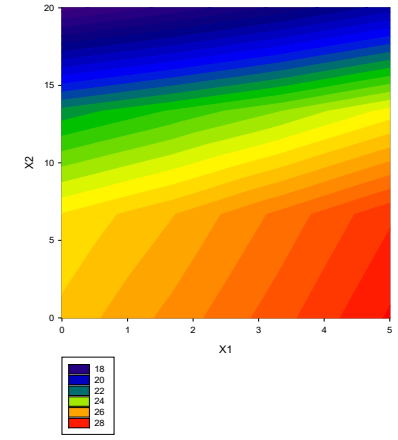
The adequate equation is

$$y_{28D} = 36.760 - 0.243x_2 - 0.0205{x_2}^2 \quad (28)$$

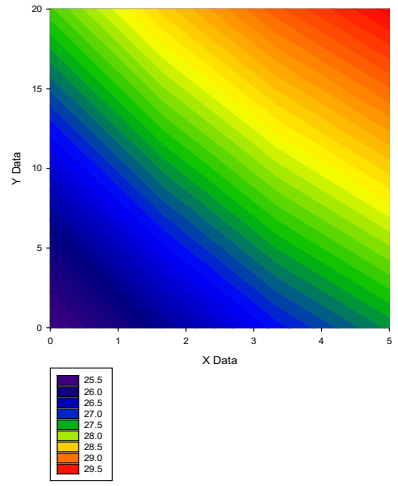
$$R^2 = 0.13,\ R^2_{adj} = 0.895,\ s = 1.604,\ PRESS = 4.544$$

The normality test is passed ($p = 0.406$), and the constant variance test is passed ($p = 0.378$). No influential data point was found.

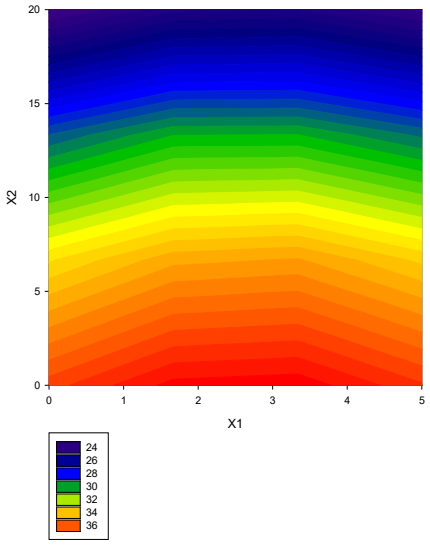
The adequate equation did not involve the x_1 variable. That is, the x_1 variable does not significantly affect y_{28D}. In Figure 2, the contour plots produced with different equations show different results. The comparison indicated the importance of producing response surface plots with adequate equations.



(**a**)

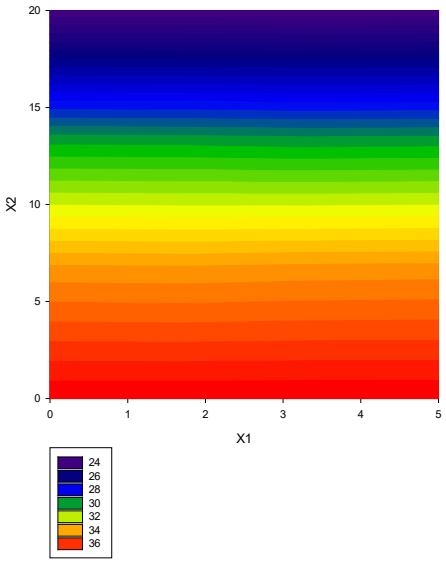


(**b**)



(**c**)

Figure 2. *Cont.*



(**d**)

Figure 2. The contour plots for the complete and adequate equations of the y_{7D} response (7-day compression) and the y_{28D} response (28-day compression). The difference in the contour curves indicates the effect of the RSM equation on the relationship between response and influential variables. (**a**). Complete equation for y_{7D} (7-day compression). (**b**). Adequate equation for y_{7D} (7-day compression). (**c**). Complete equation for y_{28D} (28-day compression). (**d**). Adequate equation for y_{28D} (28-day compression).

3.1.3. Poly-Cornstarch-Blended Biodegradable

The study used response surface methodology to evaluate the effect of x_1, amylase level, and x_2, glycerol level, on y_{WSI}, water solubility index (WSI), the y_{WAI} response, water absorption index (WAI), and the y_{ML} response, maximum load (ML), for a poly-cornstarch-blended biodegradable [28]. The experimental design is a CCD, and 13 runs were performed.

The forms of RSM equations reported in the study are

$$1.\ y_{WSI} = b_o + b_1x_1 - b_{22}x_2{}^2 - b_{12}x_1{}^2 \quad (29)$$

$$2.\ y_{WAI} = c_o + c_1x_1 - c_2x_2 - c_{11}x_1{}^2 \quad (30)$$

$$3.\ y_{ML} = d_o + d_1x_1 + d_2x_2 + d_{11}x_1{}^2 + d_{12}x_1x_2 \quad (31)$$

The variable selection method in the paper is the typical at-once variable deletion method. According to the reported results of the ANOVA tables, some variables whose *p*-value was higher than 0.05 were deleted simultaneously, and the coefficients of the remaining variables presented in the ANOVA table were used to construct these RSM models. No further calculations were performed. From the viewpoint of statistical concepts, this method is inappropriate.

The experimental data were listed in the study. In our study, the adequate regression models evaluated with the modern regression technique are

$$1.\ y_{WSI} = -3.679 + 0.627x_1 - 0.0792x_2 - 0.0446x_1{}^2 \quad (32)$$

The normality test was passed ($p = 0.791$). However, the constant variance test was failed ($p < 0.001$). There were two influential data points: the second data point, $t_i = -2{,}145$, and the seventh data point, $DFFITS_i = 3.316$. The y_{WSI} values need to be transformed to

solve the constant variance problems. The runs of the influential data need to be checked by their means and standard deviations.

$$2.\ y_{WAI} = 5.206 - 0.228x_1 + 0.00434x_2 + 0.0148{x_1}^2 \quad (33)$$

The form of the y_{WAL} equation with modern regression is the same as in the literature. The normality and the constant variance tests are passed. In the equation, the x_2 variate only has a linear relationship with y_{WAL}. However, the contour and response surface plots revealed a curved relationship in the study. The authors presented the inconsistent results of their proposed RSM equations and their response surface plots [28].

$$3.\ y_{ML} = 45.480 - 0.761x_1 - 0.158x_2 \quad (34)$$

The normality and constant variance tests were passed, and two influential data points were found: the fourth data point, $t_i = -2.217$, and sixth data point, $t_i = -2.356$.

In the research reported by the authors, the x_1 variable had a quadratic relationship with y_{MD} [28]. The contour and response surface plots in the study presented the curve distribution. However, the adequate equation indicated that both variables only have a linear relationship with the response y_{ML}. That is, an inappropriate equation will induce incorrect conclusions.

3.1.4. The Evaluation Results of the Other Literature with Two Variables

The evaluation results of the other literature with two variables are in Table 4.

Table 4. The results of the evaluation of the adequate RSM equations for two variables.

Source	Purpose	Reported Equations	Contour and Response Surface Plots	Adequate Equations	Normality Test	Constant Variance Test	Influential Data
Diemer et al. [31]	Extraction of caffeoylquinic acid x_1: temperature x_2: ethanol (%)	$y_{5\text{-CQA}} = f(x_1, x_2, x_2^2)$	Curved surface	$y_{5\text{-CQA}} = f(x_1, x_2, x_2^2)$	Passed	Passed	1st
Adeyanju et al. [32]	*Akara* ogbonoso snacks x_1: temperature x_2: time	Y_{MC} (moisture) = $f(x_1, x_2, x_1^2, x_2^2, x_1x_2)$	Curved surface	$Y_{MC} = f(x_1, x_2, x_2^2)$	Passed	Passed	No
		Y_{OC} (oil content) = $f(x_1, x_2)$	Plane	$Y_{OC} = f(x_1, x_2)$	Passed	Passed	No
		$y_{\Delta E} = f(x_1, x_2)$	Plane	$y_{\Delta E} = f(x_2)$	Passed	Passed	No
		$y_{BF} = f(x_1, x_2, x_1^2, x_2^2)$	Curved surface	$y_{BF} = f(x_2, x_2^2)$	Passed	Passed	No
		$y_S = f(x_1, x_2)$	Plane	$Y_S = f(x_1, x_2)$	passed	passed	No

In the study of Diemer et al. [31], the complete datasets of the 5-CQA (chlorogenic acid) response and two variables were listed. When these datasets were obtained using modern regression, it was found that the adequate equation is in the same form as reported in the literature. One influential factor was found. In this adequate equation, the relationship between the variables x_1 and response was linear, and that between the variables x_2 and response was quadratic. However, the response surface plot presented in the study is a curve for both variables [31].

In the evaluation of the literature data of Adeyauju et al. [32], there are five responses. For the y_{OC} and $y_{\Delta E}$ responses, the literature report has the same results as the adequate equations calculated in this study. For the y_{MC} and y_{BF} responses, the authors recommended the use of the complete model and the curve distribution in their response surface plots. However, the results of the adequate equations calculated by modern regression are different. For the response $y_{\Delta E}$, the authors reported that the x_1 and x_2 variables have a linear relationship. In this study, only x_2 variables significantly affect the y response. The authors' selection of their equations in the literature was limited to the whole linear form (x_1, x_2) or whole quadratic form (x_1, x_2, x_1^2, x_2^2, and x_1x_2) with the sequential model of the response's square. The effect of individual variables was not considered, so the results are different from modern regression. The selection of these variables in the study was carried out to use the results of the sequential model of the ANOVA table to justify the significant effects of the variables. When the linear or quadratic forms have a significant effect, all variables in the linear form (x_1 and x_2) or the whole quadratic form (x_1^2, x_2^2, x_1x_2) are accepted in these RSM models. The significant effect of each variable on response does not need to be tested individually. These problems have been illustrated in Section 2.5 of this study.

3.2. Three Variables

3.2.1. Extruded African Breadfruit–Corn–Soy

Nwabueze [33] reported the effect of three variables, x_1, feed composition, x_2, feed moisture, and x_3, screw speed, on three responses, y_{TIA}, trypsin inhibitor activity, TIA, $y_{\text{phytic acid}}$, phytic acid, and y_{tan}, tannin content, with a CCD experimental design. The replicates were performed at center points, and the total number of samples was twenty-five.

The *p*-value was used to justify the significant effect of the coefficients of the parameters for three responses. If the *p*-value was higher than 0.05, the parameter was removed. The RSM models of this response are recorded according to the estimated regression values of the remaining parameters [33].

Typical results of the ANOVA for y_{TIA} presented in the study are listed in Table 5. This table lists the *p*-values of x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, and x_3^2, and then compares them with the cutoff value, $p < 0.05$.

In Table 5, only the coefficients of b_3 and b_{11} significantly affected the response. The researchers then left these two variables alone and deleted all other variables at once. The estimated values of b_3 and b_{11} in this table were used as the final estimated values.

By the elimination-at-once method, the authors reported that the RSM equations in the study are [33]

$$y_{TIA} = -2.980433 + 0.071086x_3 + 0.00427x_1^2 \quad (35)$$

With the same technique, the other equations are

$$y_{\text{phytic acid}} = 436.2951 + 0.022895x_1^2 \quad (36)$$

$$y_{tan} = 3.51248 - 0.0000186x_3^2 \quad (37)$$

The authors' regression technique involved deleting all variables whose *p*-values were >0.05 at once, but the interaction effect of these variables was not considered.

Table 5. Estimated regression coefficients for trypsin inhibitor activity (TIA) of extruded African breadfruit–corn–soy mixtures (data source: [33]).

	Estimated Values	Standard Error	*p*-Value
b_0	−2.980433	5.183862	
b_1	0.107709	0.069684	0.1445
b_2	−0.174901	0.309534	0.5810
b_3	0.071086	0.037511	0.0789 *
b_{11}	−0.000427	0.000157	0.0168 **
b_{12}	−0.001111	0.002924	0.7098
b_{13}	−0.000500	0.000483	0.2347
b_{22}	0.004168	0.009179	0.6567
b_{23}	−0.00608	0.001728	0.7302
b_{33}	−0.000135	0.000115	0.5745

Note: * $p < 0.1$; ** $p < 0.05$.

Another question arises about the form of these equations [33]. The x_1^2 and x_3^2 variables are derived from the x_1 and x_3 variables for the polynomial equations. Suppose that the x_i^2, x_j^2, or x_ix_k variables are effective parameters. In this case, the x_i and x_j variables are validated parameters; the x_i and x_j variables should be included in this regression because the x_i^2, x_j^2, or x_ix_j variables are derived from the x_i and x_j variables.

The selection steps for y_{TIA} are listed in Appendix A.2. The adequate equations evaluated by modern regression analysis are

$$y_{TIA} = 1.574 + 0.0622x_1 - 0.00307x_3 + 0.173x_1^2 + 0.168x_3^2 \quad (38)$$

$$y_{\text{phytic acid}} = 101.215 - 1.753x_1 - 0.984x_2 + 5.186x_1^2 - 8.441x_2^2 \quad (39)$$

$$y_{tan} = 103.957 - 0.984x_2 - 8.273x_2^2 \quad (40)$$

The normality and constant variance tests were passed. y_{TIA}, $y_{\text{phytic acid}}$, and y_{tan} have one, two, and two influential data points, respectively.

The study presented three 3D response surface plots. These plots are plotted with the complete models involving x_1^2, x_2^2, and x_3^2 [33]. However, the RSM equations proposed by the authors (Equations (35)–(37)) are not complete equations. The curves of these figures did not present appropriate results for the relationship among the three responses and variables. Adequate RSM equations are essential for providing helpful information for researchers.

3.2.2. Extraction of Bioactive Components from Defatted Marigold Residue

In Study [36], the influencing factors included x_1, ethanol concentration, x_2, temperature, and x_3, time. Four responses were measured. These were y_{TPC}, total phenolics (TPC), y_{TFC}, total flavonoids (TFC), y_{ABTS}, radical scavenging activity of ABTS, and y_{DPPH}, radical scavenging activity of DPPH.

The fitting models in the study were complete equations [36]. That is, all variables (x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2) were used in their models. The four responses' contour and response surface plots were plotted with these quadratic and interaction terms. The four ANOVA tables of responses that showed the three variables' effect on the responses

were presented in the study [36]. The *p*-values indicated the insignificance of some variables. With the *p*-values, the quadratic terms of some variables had an insignificant effect on the responses. That is, the authors did not utilize the information of *p*-values of some variables to evaluate the adequacy of RSM equations.

The evaluation steps of the adequacy of RSM equations for the TPC and TFC responses are listed in Supplements S1 and S2.

The adequate RSM equations of the four responses are

$$1.\ y_{TPC} = -80.381 + 3.688x_1 + 0.219{x_1}^2 \quad (41)$$

Only the x_1 variable has a significant effect on the response y.

Two influential data points were found. For second data point, $t_i = 4.407$, with $DFFITS_i = 5.075$. For the 14th data point, $t_i = -2.312$.

The PRESS values of the complete and adequate equations are 3476.8 and 2083.706, respectively. The adequate equation has a better predictive ability than the complete equation.

$$2.\ y_{TFC} = -222.966 + 7.088x_1 + 2.165x_2 - 0.0354{x_1}^2 - 0.0272x_1x_2 \quad (42)$$

Two influential data points were found. For the second data point, ti = 2.562, with DFFITSi = 2.151. For the 14th data point, ti = −2.198. The PRESS of the complete and adequate equations was 3073.183 and 1794.138, respectively.

$$3.\ y_{ABTS} = -2.902 + 0.114x_1 + 0.0102x_2 - 0.000735{x_1}^2 \quad (43)$$

One influential data point was found. For the second run, DFFITSi = 2.798. The PRESS of the complete and adequate equation was 3.266 and 1.919, respectively.

$$4.\ y_{DPPH} = -1.254 + 0.0664x_1 + 0.00561x_2 - 0.120x_3 - 0.00511{x_1}^2 + 0.00219x_1x_3 \quad (44)$$

Two influential data points were present. For the 2nd and 14th runs, the $DEFITS_i$ values are 2.266 and −2.123, respectively. The PRESS of the complete and adequate equations was 1.865 and 0.7861, respectively. The normality and constant variance tests were passed for four responses.

In the study, the authors' reports of RSM equations for four variables were all complete models [36]. Modern regression analysis found different results. The predictive criterion, PRESS, of the four adequate equations was smaller than that of the complete equations. The adequate equation has a better predictive ability.

3.2.3. Corn Extrudate Fortified with Yam

Chiu et al. [37] studied the optimization of the extrusion characteristics of corn–yam extrudates. Their variables were x_1, yam flour contents, x_2, moisture content, and x_3, screw speed. The four responses included y_{BD}, bulk density, y_{RER}, radial expansion ratio, y_{WAI}, water absorption index, and y_{HD}, hardness. The authors reported that their RSM equations were quadratic polynomial models. Then, the quadratic polynomial equations were used to make the contour and response surface plots, and the effects of variables on these responses were observed in the two types of plots.

The authors used the coefficient of determination R^2 and lack of fit as criteria to evaluate significant effects for all quadratic equations. However, their ANOVA table in the study showed an insignificant effect of some variables at $p < 0.01$ and $p < 0.05$ [37]. That is, the authors misunderstood the meaning of a significant test.

The modern regression technique evaluates the adequate equation. The experimental data were listed in the study [37]. The results are listed as follows:

1. y_{BD}, bulk density

The complete equation is

$$y_{BD} = 0.0449 - 0.00192x_1 + 0.00657x_2 - 0.000122x_3 + 0.0000221{x_1}^2 + 0.0000599\ {x_2}^2$$
$$(-2.833)\quad (3.258)\quad (-0.559)\quad (2.543)\quad (1.104)$$
$$+ 0.000000483{x_3}^2 + 0.000113x_1x_2 + 0.000000501x_1x_3 - 0.0000201\ x_2x_3, \quad (45)$$
$$(1.392)\quad (5.395)\quad (0.200)\quad (-4.795)$$
$$R^2 = 0.996,\ {R^2}_{adj} = 0.988,\ s = 0.002,\ PRESS = 1.031 \times 10^{-4}$$

The adequate equation is

$$y_{BD} = -0.0127 - 0.00171x_1 + 0.00825\ x_2 + 0.000177\ x_3 + 0.0000205{x_1}^2$$
$$(-3.741)\quad (6.206)\quad (2.975)\quad (2.374)$$
$$+ 0.000113x_1x_2 - 0.0000201x_2x_3, \quad (46)$$
$$(5.586)\quad (-4.787)$$
$$R^2 = 0.993,\ {R^2}_{adj} = 0.988,\ s = 0.002,\ PRESS = 7.71 \times 10^{-5}$$

The normality test was failed (p = 0.047), and the constant variance test was passed (p = 0.281). The influential data were the 8th data point ($DFFITS_i$ =2.338) and the 13th (t_i = −3.737).

The adequate equation only involved the ${x_1}^2$ variable. In other words, ${x_1}^2$ was the only quadratic variable that influenced the y_{BD} response. ${x_2}^2$ and ${x_3}^2$ did not significantly affect the y_1 response. An inappropriate equation could induce an incorrect result when producing response surface plots. The datasets did not pass the normality test, and two influential data points were found. Further study needs to be performed.

2. y_{RER} radial expansion ratio

The complete equation is

$$y_{RER} = 4.103 - 0.0494x_1 - 0.0487x_2 - 0.0136x_3 + 0.000658{x_1}^2 + 0.000547{x_2}^2$$
$$(-8.868)\quad (-2.937)\quad (7.534)\quad (9.923)\quad (1.104)$$
$$-0.0000218{x_3}^2 + 0.000701x_1x_2 + 0.00000750x_1x_3 - 0.0000362\ x_2x_3, \quad (47)$$
$$(-7.643)\quad (4.087)\quad (0.547)\quad (-1.058)$$
$$R^2 = 0.996,\ {R^2}_{adj} = 0.989,\ s = 0.014,\ PRESS = 0.011$$

The adequate equation is

$$y_{RER} = 4.081 - 0.0469x_1 - 0.0442x_2 + 0.0134x_3 + 0.000651{x_1}^2$$
$$(-12.553)\quad (-12.228)\quad (7.861)\quad (9.196)$$
$$-0.0000221{x_3}^2 + 0.000701x_2x_3, \quad (48)$$
$$(-7.796)\quad (4.106)$$
$$R^2 = 0.994,\ {R^2}_{adj} = 0.990,\ s = 0.014,\ PRESS = 0.007$$

The normality test was passed (p = 0.442), and the constant variance test was passed (p = 0.620). Two influential data points were found, the fourth (t_i = 2.966, $DFFITS_i$ = 4.494) and sixth (t_i = 2.657, $DFFITS_i$ = 2.366).

The results of modern regression indicated that the x_1^2 and x_3^2 variables were valid, and only the x_2 variables had a linear relationship with the response y_{RER}.

3. y_{WAI}, water adsorption index

The complete equation is

$$y_{WAI} = 6.754 + 0.0528x_1 + 0.329x_2 _ 00232x_3 - 0.00148x_1^2 - 0.0178x_2^2$$
$$(0.834) \quad (1.750) \quad (-1.135) \quad (-1.826) \quad (-3.523)$$
$$+ 0.0000128x_3^2 + 0.000750x_1x_2 + 0.00000501x_1x_3 _ 0.000801\ x_2x_3, \quad (49)$$
$$(0.396) \quad (0.385) \quad (0.0321) \quad (2.055)$$
$$R^2 = 0.937,\ R^2_{adj} = 0.882,\ s = 0.156,\ PRESS = 1.278.$$

The adequate equation is

$$y_{WAI} = 2.730 + 0.570\ x_2 - 0.00422\ x_3 - 0.0173x_2^2 \quad (50)$$
$$(3.561) \quad (-3.397) \quad (-3.044)$$
$$R^2 = 0.906,\ R^2_{adj} = 0.773,\ s = 0.176,\ PRESS = 0.696.$$

The normality test and constant variance tests were passed. One influential data point was found, the fifth (t_i = 2.539).

According to the modern regression, the x_1 variables did not significantly affect the y_{WAI} response. Only the x_2 variable had a quadratic relationship with the y_{WAI} response.

4. y_{HD}, hardness

The complete equation is

$$y_{HD} = 1.958 - 0.196x_1 + 0.253x_2 + 0.00638x_3 + 0.00811x_1^2 - 0.00148\ x_2^2$$
$$(-1.968) \quad (0.856) \quad (0.198) \quad (6.366) \quad (-0.186)$$
$$-0.00000750x_3^2 + 0.00281x_1x_2 _ 0.0000450x_1x_3 _ 0.000338\ x_2x_3 \quad (51)$$
$$(-0.147) \quad (0.919) \quad (-0.184) \quad (-0.551)$$
$$R^2 = 0.988,\ R^2_{adj} = 0.967,\ s = 0.245,\ PRESS = 4.543$$

The adequate equation is

$$y_{HD} = 3.812 - 0.171x_1 + 0.167\ x_2 - 0.00375\ x_3 + 0.00814x_1^2 \quad (52)$$
$$(-4.216) \quad (9.764) \quad (-2.743) \quad (8.137)$$
$$R^2 = 0.986,\ R^2_{adj} = 0.980,\ s = 0.193,\ PRESS = 0.938$$

The normality test and constant variance tests were passed. One influential data point was found (ninth, t_i = −3.102, $DFFITS_i$ = −2.403).

For the y_{HD} response, x_2 and x_3 have a linear relationship, and only x_1 has a quadratic form (curves).

The authors used the complete models to produce the contour and response plots [37]. However, the ANOVA tables in their report indicated that some parameters did not significantly affect the response. In comparing the adequate equations with the complete equations proposed by the authors, the importance of using regression analysis correctly cannot be overstated.

3.2.4. The Adequate Equations of the RSM in the Other Literature

Table 6 lists the results of evaluating the adequate RSM equations for the three variables studied based on the literature.

Table 6. The results of the evaluation of the adequate RSM equations for the three variables.

Source	Purpose	Reported Equations	Contour and Response Surface Plots	Adequate Equations	Normality Test	Constant Variance Test	Influential Data
Bimakv et al. [34]	CO_2 extraction of bioactive flavonoid compounds x_1: temperature x_2: pressure x_3: flow rate	y_{ER} extract ratio = complete equation	Curved surface	Complete equation	Passed	Passed	1st, 13th
Chen et al. [35]	Enzymatic clarification of asparagus juice x_1: temperature x_2: pH x_3: enzyme concentrations	$y_{clarity} = f(x_1, x_2, x_3, x_2x_3, x_1^2, x_2^2, x_3^2)$	Curved surface	$y_{clarity} = f(x_1, x_2, x_3, x_2x_3, x_1^2, x_2^2, x_3^2)$	Passed	Passed	8th, 9th
		$y_{DPPH} = f(x_1, x_3, x_1x_2, x_1x_3, x_1^2, x_2^2, x_3^2)$	Curved surface	$y_{DPPH} = f(x_1, x_2, x_3, x_1x_2, x_1x_3, x_1^2, x_2^2, x_3^2)$	Failed ($p = 0.001$)	Passed	16th
Idrus et al. [38]	Extraction of virgin coconut oil x_1: coconut milk x_2: fermentation time x_3: refrigeration time	$y_{yield} = f(x_1, x_2, x_1^2, x_2^2, x_3^2)$	Curved surface	$y_{yield} = f(x_1, x_2, x_3, x_2x_3, x_1^2, x_2^2, x_3^2)$	Passed	Passed	13th
		$y_{FFA} = f(x_2, x_1^2)$	Curved surface	$y_{FFA} = f(x_1, x_2, x_3, x_1^2, x_2x_3)$	Passed	Passed	1st, 5th, 6th
		$y_{AV} = f(x_2, x_1^2)$	Curved surface	$y_{AV} = f(x_1, x_2, x_3, x_2x_3, x_1^2)$	Passed	Passed	No
		$y_{POV} = f(x_1, x_2, x_3, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2)$	Curved surface	$y_{POV} = f(x_1, x_2, x_3, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2)$	Passed	Passed	3rd, 6th, 12th, 15th
Hong et al. [39]	Pumpkin flour with corn x_1: pumpkin x_2: moisture x_3: screw speed	Not reported	y_{RER}: Curved surface	$y_{RER} = f(x_1, x_3, x_1x_2)$	Passed	Passed	15th
		y_{RER} (radial expansion ratio)	y_{BD}: Curved surface	$y_{BD} = f(x_1, x_2, x_3, x_1^2)$	Passed	Passed	no
		y_{BD} (bulk density) y_{WAI} (water adsorption index) y_{HD} (hardness)	y_{WAI}: Curved surface	$y_{WAI} = f(x_1, x_2, x_3, x_1^2, x_2^2, x_3^2)$	Passed	Passed	3rd, 12th
			y_{HD}: Curved surface	$y_{HD} = f(x_1, x_2, x_3, x_1^2)$	Passed	Passed	8th
Wu et al. [40]	Maleic anhydride content in biopolymer blends x_1: Tapioca starch content x_2: maleic anhydride content x_3: screen speed	y_{TS} (tensile strength) = complete equation	y_{TS}: Curved surface	$y_{TS} = f(x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_3^2)$	Passed	Passed	1st
		y_{EL} (Elongation) = complete equation	y_{EL}: Curved surface	$y_{EL} = f(x_1, x_2, x_3, x_1x_2, x_2x_3, x_2^2, x_3^2)$	Passed	Passed	2nd, 3rd, 10th, 7th, 11th
		y_{WA} (water ability) = complete equation	y_{WA}: Curved surface	$y_{WA} = f(x_1, x_2, x_3, x_1x_2, x_1^2, x_2^2)$	Failed ($p = 0.02$)	Passed	2nd, 6th, 11th
Yu et al. [41]	Pipernigrum microcapsules x_1: wall materials x_2: wall concentration x_3: air temperature	y_{EFF} (efficiency) = complete equation	Not reported	$y_{EFF} = f(x_1, x_2, x_3, x_1x_3, x_2x_3, x_2^2, x_3^2)$	Passed	Failed ($p = 0.044$)	11th

Table 6. *Cont.*

Source	Purpose	Reported Equations	Contour and Response Surface Plots	Adequate Equations	Normality Test	Constant Variance Test	Influential Data
Tshizanga et al. [42]	Biodiesel production x_1: temperature x_2: oil ratio x_3: catalyst loading	y_{BY} (biodiesel yield) = complete equation	Curved surface	$y_{BY} = f(x_2, x_2^2)$	Passed	Passed	no
Wu et al. [43]	Cryoprotectants for direct vat set starters x_1: skim milk powder x_2: sucrose x_3: L-proline or glycerol	y_{SICC} = complete equation	Curved surface	$y_{SICC} = f(x_1, x_2, x_3, x_2x_3, x_1^2, x_3^2)$	Passed	Passed	2nd, 3rd, 6th, 7th, 10th, 11th
		y_{61} = complete equation	Curved surface	$y_{61} = f(x_1, x_2, x_3, x_2x_3, x_1^2, x_2^2, x_3^2)$	Passed	Passed	2nd, 3rd, 6th, 7th, 10th, 11th
Savik et al. [44]	Antioxidant cellulose from walnut husks x_1: UAE time x_2: temperature x_3: MWP time	y_{TAC} = complete equation	Curved surface	$y_{TAC} = f(x_1, x_2, x_1^2)$	Passed	Passed	No

The RSM model reported by Bimakr et al. [34] was a complete equation. The modern regression results are the same. Two influential data points were found, which required further study.

Two responses were studied for the enzymatic clarification of green asparagus juice [35]. With the ANOVA tables, the variables with a p-value > 0.05 were deleted at the same time in this study. Their RSM equations were proposed with the remaining variables and estimated values. After checking using modern regression, the $y_{clarity}$ response had the same form as the RSM equations and the y_{DPPH} was different [35].

Idrus et al. [38] reported the aqueous extraction of virgin coconut oil. The affecting factors were screened with the p-values in their ANOVA tables, which were significant ($p \leq 0.05$). The RSM equations were then established with the remaining variables. In the study [38], only the proposed equations of the y_{pov} response have the same results as our study.

Hong et al. [39] investigated four physicochemical properties of a pumpkin flour blend with corn. The RSM equations were not reported in the study. The contour and response surface plots showed that all variables had a curvilinear effect on the response. However, the results of modern regression in Table 6 indicated that the variables of the adequate equations of the four responses were not complete models. The influence factors for y_{RER} (radial expansion ratio) were x_1, x_3, and x_1x_2; no quadratic relationship existed. The affecting factors for the y_{HD} response (hardness) were x_1, x_2, x_3, and x_1^2. Only the x_1 factor had a curvilinear relationship with y_{HD}. The coefficients of variables and a significant test with p-values were presented in a literature table. However, the authors did not use this information to select adequate models [39].

Wu et al. [40] used RSM to evaluate the effects of extrusion variables and maleic anhydride content on biopolymer blends. They used the complete models to describe the RSM equations and to present the curve distribution of the contour plot and response surface plots. Regression results of a significant effect of variables on response have been reported in the study [40]. However, these statistical results were not used to assess whether the variables had a significant effect. All variables were used as the affecting factors of the response. For the y_{TS} response, only x_3 has quadratic terms. The quadratic terms are x_2^2 and x_3^2 for y_{EL}, and x_1^2 and x_2^2 for y_{WA}. The complete model involved all factors in

an appropriate equation. The influential data points were found for three responses. The normality test was failed for the y_3 response.

Yu et al. [41] studied the factors affecting piper nigrum microcapsules with spray drying. The results of ANOVA and the statistics of the model were presented in a table in the study [41]. The F-values and *p*-values showed that six variables did not significantly affect the response. However, the complete equation was reported in the study. The results of the modern regression showed that the x_1 term did not have a quadratic form with a response. The constant variance test was failed ($p = 0.644$). The transformation of y_{EFF} could be performed for the recalculation of the RSM models.

Tshizanga et al. [42] reported on optimizing biodiesel production from wastes. The authors proposed complete equations involving all variables and produced curve contour and response surface plots. The regression results of our study showed that only the x_2 variable affected the response. Two variables, x_1 and x_3, did not significantly influence the response.

In a study of cryoprotectants for direct vat set starters in Sichuan paocai, Wu et al. [43] used RSM to determine the optimization. The two responses were y_{SICC} (*L. plantarum* SICC) and y_{Y61} (*B. subtilis* Y61). The ANOVA table and the *p*-values for each coefficient were listed in the study [43]. The *p*-values of some parameters indicated an insignificant effect on the response. However, the complete models were proposed and used to produce a curved relationship for contour and response surface plots. The results of modern regression analysis indicated that the influence factors for y_{SICC} were x_1, x_2, x_3, x_2x_3, x_1^2, and x_3^2. The variable x_2 only has a linear relationship with y_{SICC}. Influence factors for y_{Y61} were x_1, x_2, x_3, x_2x_3, x_1^2, x_2^2, and x_3^2. There was no interaction effect on y_{Y61} for the x_1x_2 and x_1x_3 terms.

Savic and Gajic [44] reported the optimization of antioxidants and cellulose from walnut husks. The reported y_{TAC} equation was a complete model. The study's results reported that the interaction between x_1 and x_3 was statistically insignificant ($p > 0.05$) and could be excluded from the equation. However, the term of x_1x_3 remained in their model [44]. Through the use of the experimental data in the study, the adequate model with modern regression included the variables of x_1, x_1^2, and x_2. The x_3 variable did not have a significant effect on the response.

3.3. Four Variables

3.3.1. Haskap Extract and Tannic Acid

Yemis et al. [49] investigated the effect of four variables, x_1, polyphenol-rich haskap extract, x_2, tannic acid, x_3, temperature, and x_4, time, on C. sakazakii inactivation (y_{SI}). The CCD included 28 runs. The statistics included PRESS and lack of fit. The response y_{SI} was transformed as a logarithmic reduction.

The significant effect of the variables on the logarithmic response was evaluated using the backward elimination method. The reported RSM model included the variables of x_1, x_2, x_3, x_4, x_1x_2, x_1x_3, x_1^2, x_3^2, and x_4^2. x_2^2 and other interaction terms were excluded. Contour and response surface plots were produced with this RSM equation.

The datasets listed in the study were obtained and evaluated using modern regression. The results indicated the validity of this reported RSM model. The original y_{SI} value could not pass the constant variance test, so the authors transformed these y_{SI} responses into a logarithmic form.

After the logarithmic response was used, the normality and constant variance tests were passed. One influential data point was found.

The RSM equation of Yemis et al. [49] is adequate. It proves that a correct regression analysis technique obtained an effective RSM model for further analysis.

3.3.2. Microencapsulation of Seed Oil

Ahn et al. [48] investigated microencapsulation efficiency with four variables: x_1, soy concentration, x_2, milk protein isolate ratio, x_3, soy lecithin concentration, and x_4, homogenizing pressure. The reported RSM equations involved the following variables: x_1, x_2, x_3, x_1^2, and x_2^2. The variable selection method used was to screen the variable with its p-value < 0.05 in the ANOVA table simultaneously. The x_3^2 variable was excluded. However, the response surface plots presented in the study showed a curved relationship between the x_3 variables and response y_{EFF} (efficiency) [48]. The reported equation and the response surface plots have inconsistent results.

The selection of effective variables for response in this study is outlined in Supplement S3. The final result with the modern regression included the following variables: x_1, x_2, x_3, x_1x_3, x_2x_3, x_1^2, and x_2^2. Compared with the reported variables, x_1x_3 and x_2x_3 were significant factors in the response. However, two variables were excluded from the study.

The normality test and constant variance test were passed. The t_i and DFFITSi criteria produced two influential data points.

3.3.3. Extraction of Total Phenolic and Flavonoid Content

Hiranpradith et al. [51] studied the factors affecting the maximization of y_{TPC}, total phenolic content, and y_{TFC}, total flavonoid content. The influencing variables included x_1, ethanol concentration, x_2, ultrasonic power, x_3, extraction time, and x_4, solvent volume. The authors reported that the influencing factors for y_{TPC} are x_1, x_2, x_4, x_2^2, and x_4^2, and those for y_{TFC} are x_1, x_2, x_3, x_4, x_1x_4, and x_1^2. The screening method of variables was to remove the variable terms with p-values > 0.05 at once. Some response surface plots in the study were inconsistent with the reported RSM equations.

The modern regression results differed from the reported RSM models in the study. Differences in statistical methods could explain the inconsistent results.

The adequate y_{TpC} was

$$y_{TpC} = -11.504 + 1.046x_1 + 1.238x_4 - 0.00829x_1^2 \quad (53)$$

The influencing factors included x_1, x_4, and x_1^2. Only the x_1 variable had a curvilinear effect on the response y_{TpC}.

The normality test was failed (p = 0.004), and the constant variance test was passed. Three influential points were found with the criteria of t_i and EFFITS.

The adequate y_{TFC} was

$$Y_{TFC} = -5.764 + 0.741x_1 + 0.564x_2 \quad (54)$$

Only the variables of x_1 and x_2 significantly affected the response y_{TFC}. No quadratic terms were found for x_1, x_2, x_3, and x_4. The normality test was passed, and the constant variance test was failed (p = 0.002). Two influential data points were found. Further regression analysis must be performed to remedy the violation of the assumption of constant variance.

3.3.4. The Regression Results of the Other Literature

The results of the RSM models of modern regression are listed in Table 7.

Table 7. The results of the evaluation of the adequate RSM equations for the four variables.

Source	Purpose	Reported Equations	Contour and Response Surface Plots	Adequate Equations	Normality Test	Constant Variance Test	Influential Data
Lee et al. [46]	Microencapsulation of peanut sprout	Complete equation	curved surface	$y_{yield} = f(x_1, x_2, x_4, x_1x_2)$	Passed	Passed	4th 17th 30th
Yu et al. [47]	Yield of resveratrol content in peanut sprout	Complete equation	curved surface	$y_{YRC} = f(x_1, x_2, x_3, x_4, x_1x_3, x_2^2, x_3^2)$	Passed	Passed	18th 19th
Javanbakht and Ghoreishi [48]	Lead removal from aqueous solution	Complete equation	curved surface	$y_{LRC} = f(x_1, x_2, x_3, x_4, x_1x_4, x_2^2, x_3^2)$	Failed ($p = 0.023$)	Passed	7th 23rd
Vega et al. [50]	Natural food colorants from wild fruits	Complete equation	curved surface	$y_{TAC} = f(x_1, x_2, x_3, x_1x_3)$	Failed ($p = 0.011$)	Passed	30th 34th 46th

Lee et al. [46] reported the factors that influence the optimization of the microencapsulation of peanut sprouts. The influencing variables were x_1, water/oil ratio, x_2, first emulsifier, x_3, water/oil/water ratio, and x_4, second emulsifier. The response y was the yield of microencapsulation. The t-value and *p*-value of each variable were listed in the ANOVA table in the study [46]. The *p*-values of some variables were >0.05. However, the reported RSM equation was a complete model in the study. The results of modern regression revealed that only the variables x_1, x_2, x_4, and x_1x_2 were influence factors.

A study of optimizing germination conditions to improve the resveratrol content yield of peanut sprout was performed by Yu et al. [47]. The affecting variables were x_1, soaking temperature, x_2, soaking time, x_3, germinal temperature, and x_4, germinal time. In the study, the RSM model was a complete equation, and it was used to produce the contour and response surface plots. The F-value and t-value of each variable have been listed in the ANOVA table in the study. Some variables did not significantly affect the response with *p*-values > 0.05. However, the authors proposed a complete model [47]. The adequate equation evaluated by the modern regression only involves the variables x_1, x_2, x_3, x_4, x_1x_3, x_2^2, and x_3^2.

Javanbakht and Ghoreishi [48] studied the optimization of lead removal from aqueous solutions. The response, y_{LRC}, was the lead removal capacity. The influencing factors were x_1, pH, x_2, temperature, x_3, lead ion concentration, and x_4, adsorbent dose. The reported RSM equation was a complete equation. The curves of contour and response surface plots were presented. However, the ANOVA table in the study indicated that only six variables had a lower *p*-value ($p < 0.05$). The adequate regression evaluated with modern regression indicated that significant variables were x_1, x_2, x_3, x_4, x_1x_4, x_2^2, and x_3^2. The normality test was failed ($p = 0.023$), and the constant variance test was passed.

Vega et al. [50] studied optimization for wild *Myrtus communis* L. fruit by-products as a natural colorant source. The response was y_{TAC} (total anthocyanin content). The influencing factors were x_1, pH, x_2, ultrasound power, x_3, time, and x_4, solid/liquid ratio. The authors reported a complete equation and used this model to produce contour and response surface plots. In our study, the adequately evaluated model with modern regression only involved four variables (x_1, x_2, x_3, and x_1x_3). No quadratic terms exist in this adequate equation. The normality test was failed ($p = 0.011$), and the constant variance test was passed.

3.4. Five Variables

Acikel et al. [52] assessed the optimization of medium components for lipase production. Their research considered five variables: x_1, sucrose, x_2, molasses sucrose, x_3, yeast

extract, x_4, sunflower oil, and x_5, Tuken−80. The authors proposed a complete equation, which included twenty variables for y_{LA}, lipase activity, and y_{BC}, biomass concentration. The contour and response surface plots were produced with two complete equations. The surface figures of the curve were presented for all variables [52].

The results of the modern regression technique are listed as follows:

$$y_{LA} = f(x_1, x_2, x_3, x_4, x_5, x_1x_2, x_1x_3, x_1x_4, x_1x_5, x_2x_4, x_3x_4, x_3x_4, x_3x_5, x_4x_5, {x_1}^2, {x_2}^2, {x_4}^2) \quad (55)$$

The ${x_3}^2$ and ${x_5}^2$ variables did not significantly affect y_1. The normality and constant variance tests were passed, and there were no influential data points.

$$y_{BC} = f(x_1, x_2, x_3, x_4, x_5, x_1x_2, x_1x_3, x_1x_4, x_1x_5, x_2x_4, x_3x_4, x_3x_5, x_4x_5, {x_1}^2, {x_2}^2) \quad (56)$$

The ${x_3}^2$, ${x_4}^2$, and ${x_5}^2$ variables did not significantly affect y_{BC}. The constant variance tests were passed. However, the normality test failed. Further studies are needed to treat the non-normality problem.

4. Discussion

This study collected twenty-five datasets related to research to check the adequacy of RSM models. Only some papers reported an adequate equation to express the relationship between the response and influencing factors [34,49]. All datasets are adopted from the literature. The original experimental data are listed in the studies. The common issues in the application of RSM in the literature are listed in Table 8.

Table 8. The common issues in the application of RSM in the literature.

Issue		References
1.	They adopted a complete model and then plotted contour and 3D response surface plots to present the optimization of these variables, without using the information of the coefficient value, t-value, and *p*-value for each variable in the ANOVA table.	[29,30,36,37,39,40,43,44,46–48,52]
2.	The at-once variable deletion method was used to delete variables whose *p*-value was higher than the preselected value (usually $p < 0.05$).	[28,51]
3.	Some variables were deleted with the at-once variable deletion method. However, the contour and response surface plots were produced with the complete equations.	[31,33,35,38,45].
4.	The ANOVA table of the sequential model was misused to keep all variables in the linear or square term without significant testing for each variable.	[32,42]
5.	Datasets did not pass the normality test.	[35,40,48,50,52]
6.	Datasets failed the constant variance test.	[28,41,51]
7.	Influential data points were found.	[28,29,31,33–41,43–52]
8.	There were three replicates for each run in the experiment. However, only the mean of each run was used for the regression calculation.	[28,31,37,39,40,43,49].

Most papers adopted a complete model and then plotted contour plots and 3D response surface plots to present the optimization of these variables. In the literature,

the ANOVA tables included the coefficient value, t-value, and p-value for each variable. However, researchers did not use this information to screen the adequate equation [29,30,36,37,39,40,43,44,46–48,52].

Some papers used the at-once variable deletion method. After the first regression calculation, the ANOVA table of regression results showed each parameter's variables, coefficient values, standard error, t-value, and p-value. If the p-value of variables is higher than the preselected value (usually $p < 0.05$), these variables are deleted simultaneously. The equation is then proposed using the first regression calculation's remaining variables and their coefficient values. This method is incorrect for model building [28,51].

Some studies proposed the reported RSM equations using the at-once variable deletion method. However, they still produced the contour and response surface plots with the complete equations [31,33,35,38,45].

Section 2.5 introduced the misuse of the ANOVA table of the sequential-model sum of squares of the response. As the linear or square term significantly affects the response, all variables in the linear or square terms were accepted as affecting factors. This incorrect result was found in two studies [32,42].

If the RSM models are inappropriate, the equations' contour plots and 3D response surface plots are incorrect, and the conduct of the optimization conditions of these variables is meaningless for researchers.

For the twenty-five studies related to science and technology, some datasets did not pass the normality test [35,40,48,50,52], and some failed the constant variance test [28,41,51]. These datasets need to be transformed to correspond to the basic assumption of modern regression. The datasets of Yemis et al. [49] indicated the non-normal condition, and the authors used the logarithmic transformation to solve the problem. Yang et al. [73] emphasized that the homoscedasticity assumption plays a more critical role than normality in the validity of ANOVA in checking the linear regression models. Departures from the homogeneous assumption will induce serious incorrect results [73].

Influential data points are usually found in the datasets of these studies. They may be outliers or influencing data points in the literature. One research article states that a data point seriously influenced the coefficients' values, and the Cook's distance was very high [29]. In the study of Sinkhonde et al. [30], the authors reported the results of checking influential data points with some criteria.

The cause of the presence of influential data points could be experimental errors or the selection of the form of RSM models. Experimental errors may be due to sample preparation, instrument performance, or an operator's mistake. Different forms of regression equations, such as some nonlinear equations, could be used to improve the RSM model's fitting ability. Replicates of the experimental run at the same level could help the researcher to assess the significant difference between a data point and other data points under the same experimental conditions.

Most datasets of experiments did not have replicates for each run; only one data point was available at each level and factor. This makes it difficult to justify the correctness of these data points further. Some research reports three replicates for each case [28,31,37,39,40,43,49]. However, only the mean of each run is used to perform the regression analysis. Through the use of the mean value instead of all replicates, the results of the regression coefficient values are the same, but the statistics of the statistical test are different. For example, there are 17 runs with three replicates for each run in the study of Wu et al. [43]. As the mean of each run is used only, the sample size is 17. If all the original data are used, the sample number is increased to 51, and the degree of freedom for the residues is increased significantly. Then the statistical test of power could be improved significantly. The influencing points or outliers

could be justified if some influential data points were found. Three replicates for each run could provide some evidence for assessing the datasets.

One advantage of RSM is the small number of experiments, which reduces the time and cost spent. However, a smaller number of samples becomes a disadvantage of the RSM method for evaluating an adequate equation. In comparing the required data numbers with some empirical equations (Equations (15)–(19)), the sample numbers of the RSM equation used in experiments were limited. In the experimental design, three replicates for each run could provide enough sample numbers to perform the regression analysis.

Fifteen papers, 60% of the total literature, used Design Expert software. However, only one study [49] obtained an appropriate equation. Bimark et al. [34] used Minitab Ver. 14 software and proposed an adequate RSM equation.

Based on the results of this study, some suggestions can be proposed for the utilization of RSM in experimental design:

1. Training in the modern regression technique
   Receiving regression analysis training could enhance researchers' ability to propose adequate RSM equations
2. The backward elimination technique has been proven helpful for sequential variable selection. This method could be incorporated into commercial software to help researchers establish an adequate RSM equation.
3. Increasing the sample numbers to correspond to the minimum sample requirement is very important. This could enhance the power of the statistical test. Three replicates for one experiment run are recommended. All the data points with these replicates could be used to check the influential data point and decide whether it is an outlier or an influential point.

5. Conclusions

This study collected twenty-five research datasets from the literature to evaluate their adequate RSM equations with modern regression analysis. The results of this study indicated some common issues in establishing RSM models. Most papers used a complete model to express the relationship between the response and the influential variables, producing contour and response surface plots. When researchers observe these plots to optimize these variables, the conclusions of the experiments may be incorrect. The ANOVA tables included the coefficient value, t-value, and p-value for each variable in the literature. However, researchers did not use this information to screen the important influencing variables. Some researchers used the at-once variable deletion method. That is, as the p-value of these variables is higher than the preselected value, these variables are deleted simultaneously. Some RSM equations were proposed using the at-once variable deletion method, and contour and response surface plots were produced with the complete equations. Some papers misused the sequential model of the ANOVA to accept all variables in linear or square terms, as these terms significantly affect the response. Actually, all variables need to be tested individually. Some datasets did not pass the normality test. Some datasets failed the constant variance test. Influential data points are found in most of the literature in this study.

The suggestions for applying RSM for researchers are to enhance training in modern regression, use the back elimination method to evaluate the influencing variables in the RSM models, and increase the sample size with three replicates in each run. An adequate RSM model can optimize the influencing variables for response in science and technology.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/app15137206/s1, Supplement S1: The evaluation of the adequate equation for the TPC response [36]; Supplement S2: The evaluation of the adequate equation for the TFC response [36]; Supplement S3: The evaluation of the adequate equation for the response [45].

Author Contributions: Conceptualization, H.-Y.C. and C.C.; methodology, H.-Y.C. and C.C.; software, C.C.; formal analysis, H.-Y.C.; investigation, H.-Y.C. and C.C.; data curation, H.-Y.C.; writing—original draft preparation, H.-Y.C. and C.C.; writing—review and editing, H.-Y.C. and C.C.; visualization, C.C. supervision, C.C.; project administration, C.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors would like to thank the Ministry of Science and Technology of the Republic of China for financially supporting this research.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Appendix A.1. The Procedure to Evaluate the Adequate Equation of the y_{28D} Response (28-Day Compression) [30]

The results for the complete model with regression analysis are as follows:

$$1.\ y28_D = 36.213 + 0.059x_1 - 0.343x_2 - 0.202{x_1}^2 - 0.0157{x_2}^2 + 0.00147x_1x_2 \quad \text{(A1)}$$

$$(1.115) \quad (-1.414) \quad (-1.205) \quad (-1.498) \quad (0.0422)$$

$$R^2 = 0.928,\ R^2{}_{adj} = 0.877,\ s = 0.1.739,\ PRESS = 121.931,$$

The normality test is passed (p = 0.397), and the constant variance test is passed (p = 0.504. Delete the x_1x_2 variable and recompute:

$$2.\ y28_D = 36.176 + 1.074x_1 - 0.339x_2 - 0.202{x_1}^2 - 0.0157{x_2}^2 \quad \text{(A2)}$$

$$(1.299) \quad (-1.641) \quad (-1.288) \quad (-1.69)$$

$$R^2 = 0.928,\ R^2{}_{adj} = 0.892,\ s = 1.627,\ PRESS = 7.751$$

Delete the ${x_1}^2$ variable and recalculate:

$$3.\ y28_D = 36.596 + 0.0654x_1 - 0.243x_2 - 0.0205{x_2}^2 \quad \text{(A3)}$$

$$(0.238) \quad (-1.217) \quad (-2.184)$$

$$R^2 = 0.913,\ R^2{}_{adj} = 0.884,\ s = 1.686,\ PRESS = 4.874$$

Delete the x_1 variable and recalculate:

$$4.\ y28_D = 36.760 - 0.243x_2 - 0.0205x^2 \quad \text{(A4)}$$

$$(-1.279) \quad (-2.295)$$

$$R^2 = 0.13,\ R^2{}_{adj} = 0.895,\ s = 1.604,\ PRESS = 4.544$$

The normality test is passed (p = 0.406), and the constant variance test is passed (p = 0.378). No influential data point was found.

Appendix A.2. The Selection of Adequate Variables for y_{TIA} [33]

$$1.\ y_{TIA} = 1.527 - 0.0622x_1 + 0.120x_2 - 0.307x_3 + 0.170{x_1}^2 - 0.0923{x_2}^2$$
$$(-0.732)\quad (1.409)\quad (-0.362)\quad (2.202)\quad (1.195)$$
$$+0.165{x_3}^2 + 0.0687x_1x_2 - 0.136x_1x_3 + 0.0937x_2x_3 \quad \text{(A5)}$$
$$(2.113)\quad (0.619)\quad (-1.228)\quad (0.665)$$
$$R^2 = 0.523,\ R^2_{adj} = 0.237,\ s = 1.398,\ PRESS = 3.303.$$

The x_1x_2 variable was removed due to the failure of the significance test. The equation was recalculated.

$$2.\ y_{TIA} = 1.527 - 0.0622x_1 + 0.120x_2 - 0.0307x_3 + 0.170{x_1}^2 + 0.0923{x_2}^2$$
$$(-7.746)\quad (1.437)\quad (-0.369)\quad (2.245)\quad (1.219)$$
$$+0.165{x_3}^2 - 0.136x_1x_3 + 0.0737x_2x_3 \quad \text{(A6)}$$
$$(2.175)\quad (-1.252)\quad (0.678)$$
$$R^2 = 0.511,\ R^2_{adj} = 0.267,\ s = 1.398,\ PRESS = 2.812.$$

x_1x_3 was deleted, and a new equation was obtained:

$$3.\ y_{TIA} = 1.527 - 0.0622x_1 + 0.120x_2 - 0.0307x_3 + 0.170{x_1}^2 + 0.0.0923{x_2}^2$$
$$(-0.759)\quad (1.460)\quad (-0.375)\quad (2.282)\quad (1.258)$$
$$+0.165{x_3}^2 - 0.136x_1x_3 \quad \text{(A7)}$$
$$(2.211)\quad (-1.272)$$
$$R^2 = 0.497,\ R^2_{adj} = 0.290,\ s = 1.303,\ PRESS = 2.583.$$

x_1x_3 was deleted, and a new equation was obtained:

$$4.\ y_{TIA} = 1.527 - 0.0622x1 + 0.120x_2 - 0.0307x_3 + 0.170{x_1}^2 + 0.0923{x_2}^2$$
$$(-0.746)\quad (1.436)\quad (-0.309)\quad (2.244)\quad (1.218) \quad \text{(A8)}$$
$$+0.165{X_3}^2$$
$$(2.174)$$
$$R^2 = 0.449,\ R^2_{adj} = 0.266,\ s = 0.308,\ PRESS = 2.655.$$

${x_2}^2$ was deleted, and a new equation was obtained:

$$5.\ y_{TIA} = 1.574 - 0.0622x_1 + 0.120x_2 - 0.0307x_3 + 0.173{x_1}^2 + 0.168{x_3}^2 \quad \text{(A9)}$$
$$(-0.737)\quad (1.418)\quad (-0.364)\quad (2.255)\quad (2.185)$$
$$R^2 = 0.404,\ R^2_{adj} = 0.247,\ s = 0.120,\ PRESS = 2.783.$$

x_2 was deleted, and a new equation was obtained:

$$6.Y_{TIA} = 1.574 - 0.0622x_1 - 0.0307x_3 + 0.173{x_1}^2 + 0.168{x_3}^2 \quad \text{(A10)}$$

$$(-0.719) \quad (-0.355) \quad (2.200) \quad (2.132)$$

$$R^2 = 0.241, R^2_{adj} = 0.209, s = 0.321, PRESS = 2.856.$$

Although the linear terms of x_1 and x_3 are insignificant ($p > 0.05$), the quadratic terms ${x_2}^2$ and ${x_3}^2$ are significant in the model. So, the linear terms of x_1 and x_3 are hierarchically added to the quadratic equation.

The normality test is passed ($p = 0.130$), and the constant variance test is failed ($p = 0.246$). One influential data point was found in the sixth run, $t_i = -2.346$.

References

1. Myers, R.H.; Montgomery, D.C.; Anderson-Cook, C.M. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*; John Wiley & Sons: Hoboken, NJ, USA, 2016.
2. Baş, D.; Boyacı, İ.H. Modeling and optimization I: Usability of response surface methodology. *J. Food Eng.* **2007**, *78*, 836–845. [CrossRef]
3. Asoo, H.R.; Alakali, J.S.; Ikya, J.K.; Yusufu, M.I. Historical background of RSM. In *Response Surface Methods-Theory, Applications and Optimization Techniques*; IntechOpen: London, UK, 2024.
4. Bruns, R.E.; Scarminio, I.S.; de Barros Neto, B. *Statistical Design-Chemometrics*; Elsevier: Amsterdam, The Netherlands, 2006.
5. Bezerra, M.A.; Ferreira, S.L.C.; Novaes, C.G.; Dos Santos, A.M.P.; Valasques, G.S.; da Mata Cerqueira, U.M.F.; dos Santos Alves, J.P. Simultaneous optimization of multiple responses and its application in Analytical Chemistry—A review. *Talanta* **2019**, *194*, 941–959. [CrossRef] [PubMed]
6. Dejaegher, B.; Vander Heyden, Y. Experimental designs and their recent advances in set-up, data interpretation, and analytical applications. *J. Pharm. Biomed. Anal.* **2011**, *56*, 141–158. [CrossRef] [PubMed]
7. Yolmeh, M.; Jafari, S.M. Applications of response surface methodology in the food industry processes. *Food Bioprocess Technol.* **2017**, *10*, 413–433. [CrossRef]
8. De Oliveira, L.G.; de Paiva, A.P.; Balestrassi, P.P.; Ferreira, J.R.; da Costa, S.C.; da Silva Campos, P.H. Response surface methodology for advanced manufacturing technology optimization: Theoretical fundamentals, practical guidelines, and survey literature review. *Int. J. Adv. Manuf. Technol.* **2019**, *104*, 1785–1837. [CrossRef]
9. Szpisják-Gulyás, N.; Al-Tayawi, A.N.; Horváth, Z.H.; László, Z.; Kertész, S.; Hodúr, C. Methods for experimental design, central composite design and the Box–Behnken design, to optimise operational parameters: A review. *Acta Aliment.* **2023**, *52*, 521–537. [CrossRef]
10. Olabinjo, O.O. Response surface techniques as an inevitable tool in optimization process. In *Response Surface Methods-Theory, Applications and Optimization Techniques*; IntechOpen: London, UK, 2024.
11. Meloun, M.; Militký, J. Detection of single influential points in OLS regression model building. *Anal. Chim. Acta* **2001**, *439*, 169–191. [CrossRef]
12. Bhattacharya, S. Central composite design for response surface methodology and its application in pharmacy. In *Response Surface Methodology in Engineering Science*; IntechOpen: London, UK, 2021.
13. Anderson, M.J.; Whitcomb, P.J. *RSM Simplified: Optimizing Processes Using Response Surface Methods for Design of Experiments*; Productivity Press: New York, NY, USA, 2016.
14. Rodrigues, A.C. Response surface analysis: A tutorial for examining linear and curvilinear effects. *Rev. Adm. Contemp.* **2021**, *25*, e200293. [CrossRef]
15. Reza, A.; Chen, L.; Mao, X. Response surface methodology for process optimization in livestock wastewater treatment: A review. *Heliyon* **2024**, *10*, e30326. [CrossRef]
16. Bezerra, M.A.; Santelli, R.E.; Oliveira, E.P.; Villar, L.S.; Escaleira, L.A. Response surface methodology (RSM) as a tool for optimization in analytical chemistry. *Talanta* **2008**, *76*, 965–977. [CrossRef]
17. Nwabueze, T.U. Basic steps in adapting response surface methodology as mathematical modelling for bioprocess optimisation in the food systems. *Int. J. Food Sci. Technol.* **2010**, *45*, 1768–1776. [CrossRef]
18. Nwabueze, T.U.; Iwe, M.O. Residence time distribution (RTD) in a single screw extrusion of African breadfruit mixtures. *Food Bioprocess Technol.* **2010**, *3*, 135–145. [CrossRef]
19. Khuri, A.I. Response surface methodology and its applications in agricultural and food sciences. *Biom. Biostat. Int. J.* **2017**, *5*, 155–163. [CrossRef]
20. Weremfo, A.; Abassah-Oppong, S.; Adulley, F.; Dabie, K.; Seidu-Larry, S. Response surface methodology as a tool to optimize the extraction of bioactive compounds from plant sources. *J. Sci. Food Agric.* **2023**, *103*, 26–36. [CrossRef]
21. Madamba, P.S. The response surface methodology: An application to optimize dehydration operations of selected agricultural crops. *LWT—Food Sci. Technol.* **2002**, *35*, 584–592. [CrossRef]

22. Koç, B.; Kaymak-Ertekin, F. Response surface methodology and food processing applications. *Gida J. Food* **2010**, *35*, 63–70.
23. Said, K.A.M.; Amin, M.A.M. Overview on the response surface methodology (RSM) in extraction processes. *J. Appl. Sci. Process Eng.* **2015**, *2*, 8–17.
24. Malekjani, N.; Jafari, S.M. Food process modeling and optimization by response surface methodology (RSM). In *Mathematical and Statistical Applications in Food Engineering*; CRC Press: Boca Raton, FL, USA, 2020; pp. 181–203.
25. Kidane, S.W. Application of response surface methodology in food process modeling and optimization. In *Response Surface Methodology in Engineering Science*; IntechOpen: London, UK, 2021.
26. Tirado-Kulieva, V.A.; Sánchez-Chero, M.; Yarlequé, M.; Aguilar, G.F.V.; Carrión-Barco, G.; Santa Cruz, A.G.Y. An overview on the use of response surface methodology to model and optimize extraction processes in the food industry. *Curr. Res. Nutr. Food Sci.* **2021**, *9*, 745–754. [CrossRef]
27. Istiqomah, A.; Saputra, O.A.; Firdaus, M.; Kusumaningsih, T. Response Surface Methodology as an Excellent Tool for Optimizing Sustainable Food Packaging: A Review. *J. Biosyst. Eng.* **2024**, *49*, 434–452. [CrossRef]
28. Chen, Y.D.; Peng, J.; Lui, W.B. Composition optimization of poly (vinyl alcohol)-/cornstarch-blended biodegradable composite using response surface methodology. *J. Appl. Polym. Sci.* **2009**, *113*, 258–264. [CrossRef]
29. Milán-Carrillo, J.; Montoya-Rodríguez, A.; Gutiérrez-Dorado, R.; Perales-Sánchez, X.; Reyes-Moreno, C. Optimization of extrusion process for producing high antioxidant instant amaranth (*Amaranthus hypochondriacus* L.) flour using response surface methodology. *Appl. Math.* **2012**, *3*, 1516–1525. [CrossRef]
30. Sinkhonde, D.; Onchiri, R.O.; Oyawa, W.O.; Mwero, J.N. Response surface methodology-based optimisation of cost and compressive strength of rubberised concrete incorporating burnt clay brick powder. *Heliyon* **2021**, *7*, e08565. [CrossRef] [PubMed]
31. Diemer, E.; Chadni, M.; Grimi, N.; Ioannou, I. Optimization of the accelerated solvent extraction of caffeoylquinic acids from forced chicory roots and antioxidant activity of the resulting extracts. *Foods* **2022**, *11*, 3214. [CrossRef] [PubMed]
32. Adeyanju, J.A.; Abioye, A.O.; Adekunle, A.A.; Ibrahim, T.H.; Oloyede, A.A.; Akinwusi, D.E. Process optimization of deep-fat frying variables and effects on some quality characteristics of *akara* Ogbomoso snacks produced from cowpea. *Food Res.* **2024**, *8*, 502–507. [CrossRef]
33. Nwabueze, T.U. Effect of process variables on trypsin inhibitor activity (TIA), phytic acid and tannin content of extruded African breadfruit-corn-soy mixtures: A response surface analysis. *LWT—Food Sci. Technol.* **2007**, *40*, 21–29. [CrossRef]
34. Bimakr, M.; Rahman, R.A.; Ganjloo, A.; Taip, F.S.; Salleh, L.M.; Sarker, M.Z.I. Optimization of supercritical carbon dioxide extraction of bioactive flavonoid compounds from spearmint (*Mentha spicata* L.) leaves by using response surface methodology. *Food Bioprocess Technol.* **2012**, *5*, 912–920. [CrossRef]
35. Chen, X.; Xu, F.; Qin, W.; Ma, L.; Zheng, Y. Optimization of enzymatic clarification of green asparagus juice using response surface methodology. *J. Food Sci.* **2012**, *77*, C665–C670. [CrossRef]
36. Gong, Y.; Hou, Z.; Gao, Y.; Xue, Y.; Liu, X.; Liu, G. Optimization of extraction parameters of bioactive components from defatted marigold (*Tagetes erecta* L.) residue using response surface methodology. *Food Bioprod. Process.* **2012**, *90*, 9–16. [CrossRef]
37. Chiu, H.W.; Peng, J.C.; Tsai, S.J.; Tsay, J.R.; Lui, W.B. Process optimization by response surface methodology and characteristics investigation of corn extrudate fortified with yam (*Dioscorea alata* L.). *Food Bioprocess Technol.* **2013**, *6*, 1494–1504. [CrossRef]
38. Idrus, N.F.M.; Febrianto, N.A.; Zzaman, W.; Cuang, T.E.; Yang, T.A. Optimization of the aqueous extraction of virgin coconut oil by response surface methodology. *Food Sci. Technol. Res.* **2013**, *19*, 729–737. [CrossRef]
39. Hong, F.L.; Peng, J.; Lui, W.B.; Chiu, H.W. Investigation on the physicochemical properties of pumpkin flour (*Cucurbita moschata*) blend with corn by single- screw extruder. *J. Food Process. Preserv.* **2015**, *39*, 1342–1354. [CrossRef]
40. Wu, C.Y.; Lui, W.B.; Peng, J. Optimization of extrusion variables and maleic anhydride content on biopolymer blends based on poly (hydroxybutyrate-co-hydroxyvalerate)/poly (vinyl acetate) with tapioca starch. *Polymers* **2018**, *10*, 827. [CrossRef] [PubMed]
41. Yu, Y.; Wei, R.; Jia, X.; Zhang, X.; Liu, H.; Xu, B.; Xu, B. Preparation of piper nigrum microcapsules by spray drying and optimization with response surface methodology. *J. Oleo Sci.* **2022**, *71*, 1789–1797. [CrossRef]
42. Tshizanga, N.; Aransiola, E.F.; Oyekola, O. Optimisation of biodiesel production from waste vegetable oil and eggshell ash. *S. Afr. J. Chem. Eng.* **2017**, *23*, 145–156. [CrossRef]
43. Wu, L.; Yang, Z.; Zhang, Y.; Li, L.; Tan, C.; Pan, L.; Gao, H. Optimization of the cryoprotectants for direct vat set starters in Sichuan paocai using response surface methodology. *Foods* **2025**, *14*, 157. [CrossRef]
44. Savić, I.M.; Savić Gajić, I.M. Extraction and characterization of antioxidants and cellulose from green walnut husks. *Foods* **2025**, *14*, 409. [CrossRef]
45. Ahn, J.H.; Kim, Y.P.; Lee, Y.M.; Seo, E.M.; Lee, K.W.; Kim, H.S. Optimization of microencapsulation of seed oil by response surface methodology. *Food Chem.* **2008**, *107*, 98–105. [CrossRef]
46. Lee, Y.K.; Ahn, S.I.; Kwak, H.S. Optimizing microencapsulation of peanut sprout extract by response surface methodology. *Food Hydrocoll.* **2013**, *30*, 307–314. [CrossRef]
47. Yu, M.; Liu, H.; Yang, Y.; Shi, A.; Liu, L.; Hu, H.; Wang, Q.; Yu, H.; Wang, X. Optimising germinated conditions to enhance yield of resveratrol content in peanut sprout using response surface methodology. *Int. J. Food Sci. Technol.* **2016**, *51*, 1754–1761.

48. Javanbakht, V.; Ghoreishi, S.M. Application of response surface methodology for optimization of lead removal from an aqueous solution by a novel superparamagnetic nanocomposite. *Adsorpt. Sci. Technol.* **2017**, *35*, 241–260. [CrossRef]
49. Yemiş, P.G.; Yemiş, O.; Öztürk, A. Optimization of haskap extract and tannic acid combined with mild heat treatment: A predictive study on the inhibition of cronobacter sakazakii. *Foods* **2025**, *14*, 562. [CrossRef] [PubMed]
50. Vega, E.N.; González-Zamorano, L.; Cebadera, E.; Barros, L.; da Silveira, T.F.; Vidal-Diez de Ulzurrun, G.; Tardio, J.; Lazaro, A.; Camara, M.; Fernansez-Ruiz, V.; et al. Wild *Myrtus communis* L. Fruit by-product as a promising source of a new natural food colourant: Optimization of the extraction process and chemical characterization. *Foods* **2025**, *14*, 520. [CrossRef] [PubMed]
51. Hiranpradith, V.; Therdthai, N.; Soontrunnarudrungsri, A.; Rungsuriyawiboon, O. Optimisation of ultrasound-assisted extraction of total phenolics and flavonoids content from centella asiatica. *Foods* **2025**, *14*, 291. [CrossRef] [PubMed]
52. Açıkel, Ü.; Erşan, M.; Açıkel, Y.S. Optimization of critical medium components using response surface methodology for lipase production by *Rhizopus delemar*. *Food Bioprod. Process.* **2010**, *88*, 31–39. [CrossRef]
53. Myers, R.H. *Classical and Modern Regression with Applications*, 2nd ed.; Duxbury Press: Monterey, CA, USA, 1990.
54. Ryan, T.P. *Modern Regression Methods*; John Wiley & Sons: Hoboken, NJ, USA, 2008.
55. Wilcox, R.R.; Keselman, H.J. Modern regression methods that can substantially increase power and provide a more accurate understanding of associations. *Eur. J. Personal.* **2012**, *26*, 165–174. [CrossRef]
56. Marinoiu, C. Classic and modern in regression modelling. *Econ. Insights—Trends Chall.* **2017**, *69*, 41–50.
57. Montgomery, D.C.; Peck, E.A.; Vining, G.G. *Introduction to Linear Regression Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2021.
58. Allen, M.P. *Understanding Regression Analysis*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2004.
59. Berger, D.E. *Introduction to Multiple Regression*; Claremont Graduate University: Claremont, CA, USA, 2008.
60. Rawlings, J.O.; Pantula, S.G.; Dickey, D. *Applied Regression Analysis*; Springer: New York, NY, USA, 1998.
61. Dielman, T.E. *Applied Regression Analysis for Business and Economics*, 4th ed.; Duxbury/Thomson Learning: Pacific Grove, CA, USA, 2005.
62. Mendenhall, W.; Sincich, T. *Regression Analysis. A Second Course in Statistics*, 12th ed.; Prentice Hall: Hoboken, NJ, USA, 2012.
63. Chowdhury, M.Z.I.; Turin, T.C. Variable selection strategies and its importance in clinical prediction modelling. *Fam. Med. Community Health* **2020**, *8*, e000262. [CrossRef]
64. Rowley, E.K. Comparison of Variable Selection Methods. Ph.D. Thesis, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, 2019.
65. Kelley, K.; Maxwell, S.E. Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant. *Psychol. Methods* **2003**, *8*, 305–321. [CrossRef]
66. Bonett, D.G.; Wright, T.A. Sample size requirements for multiple regression interval estimation. *J. Organ. Behav.* **2011**, *32*, 822–830. [CrossRef]
67. Hanley, J.A. Simple and multiple linear regression: Sample size considerations. *J. Clin. Epidemiol.* **2016**, *79*, 112–119. [CrossRef] [PubMed]
68. Snee, R.D. Validation of regression models: Methods and examples. *Technometrics* **1977**, *19*, 415–428. [CrossRef]
69. Green, S.B. How many subjects does it take to do a regression analysis? *Multivar. Behav. Res.* **1991**, *26*, 499–510. [CrossRef] [PubMed]
70. Khamis, H.J.; Kepler, M. Sample size in multiple regression: 20+ 5k. *J. Appl. Stat. Sci.* **2010**, *17*, 505–517.
71. Tabachnick, B.G.; Fidell, L.S. *Using Multivariate Statistics*; Allyn & Bacon/Pearson Education: Boston, MA, USA, 2012.
72. Zaarour, N. A simple relationship between the sample size and the number of independent variables. *J. Bus. Econ. Stat.* **2024**, *2*, 1–22.
73. Yang, K.; Tu, J.; Chen, T. Homoscedasticity: An overlooked critical assumption for linear regression. *Gen. Psychiatry* **2019**, *32*, e100148. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.