Research Article
Application of Statistical Analysis Software in Food Scientific Modeling

[1]Miaochao Chen, [2]Kong Xiangsheng and [1]Kan Chen
[1]Department of Mathematics, Chaohu University, Hefei, 238000, P.R. China
[2]Department of Computer and Information Engineering, Xinxiang University,
Xinxiang City, Henan Province, P.R. China

Abstract: In food scientific researches, sophisticated statistical analysis problems often can be met and in this study, through SPSS statistical analysis software, the method of the curve regression model and the multiple regression model that both are common in food science has been established and the experimental results show that the method can be effectively used in the statistical analysis model of food science.

Keywords: Food science, modeling, SPSS statistical analysis software

INTRODUCTION

SPSS is a widely used program for statistical analysis in social science. It is also used by market researchers, health researchers, survey companies, government, education researchers, marketing organizations, data miners and others. The original SPSS manual (Nie *et al*., 1970) has been described as one of "sociology's most influential books" for allowing ordinary researchers to do their own statistical analysis. In addition to statistical analysis, data management (case selection, file reshaping, creating derived data) and data documentation (a metadata dictionary is stored in the data file) are features of the base software (Du and Ke, 2013).

In food science searches, the sample sources are complex, the production processes are various and there are many links of quality control measures of processing (Eke *et al*., 2013). In order to save a lot of manpower and material resources, speed up the development of new products and control the quality of products better, the important problems for food scientific researchers that must be solved are to establish the corresponding mathematical model and reveal universal natural laws. However, if we use the traditional statistical analysis method to analyze and process the experimental data and set up a mathematical model, the calculation process will be very complicated. On the basis of being engaged in experimental design and the teaching of statistical analysis for a long time, the authors of this study combine the actual problems in food scientific researches and give the introduction of the application of SPSS statistical software in the modeling of food scientific researches as follows (Jiwen *et al*., 2013).
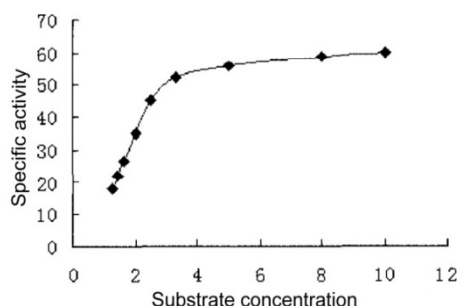

Fig. 1: The relationship of the specific activity between substrate concentration and enzyme

MATERIALS AND METHODS

In the problems of food scientific researches, sometimes two variables have a linear relationship, but in most cases, they always haven't a linear relationship but a nonlinear relationship. For example, when making a Michaelis-Menten equation and the calculation of Michaelis-Menten constants, we can measure the relationship between the specific activity of enzyme and the substrate concentration and get the 9 pairs of data as shown in Table 1. Make the scatter diagram with these data (Fig. 1) and we can see that the distribution of these points is in a curve shape (Zhao *et al*., 2011).

The basic task of curve regression analysis is to establish a curve regression equation through the actual observation data of the two correlated variables x and y so as to reveal the form of the curve relation between x and y. There are usually three ways to determine the type of the curve relationship between x and y: first, use professional knowledge as well as the known theoretical laws and practical experience; second, when

Table 1: Measured values of the research of the reasons of non-enzymatic browning in the manufacturing process of a brand of peach juice

Determination No.	Ieucoanthocyanin X1	Anthocyanin X2	Maillard reaction X3	Ascorbic content X4	Non-enzymatic browning package value y
1	0.045	0.032	0.008	2.654	9.520
2	0.063	0.013	0.007	2.891	8.960
3	0.054	0.025	0.005	3.251	8.420
4	0.062	0.012	0.004	3.218	8.320
5	0.060	0.009	0.013	2.854	7.650
6	0.053	0.016	0.011	3.264	7.850
7	0.041	0.021	0.015	2.785	7.210
8	0.074	0.011	0.008	3.216	8.470
9	0.046	0.015	0.021	3.518	9.460
10	0.065	0.021	0.018	2.920	10.910
11	0.071	0.016	0.017	3.464	12.410
12	0.094	0.012	0.011	3.217	9.760
13	0.048	0.015	0.021	2.768	8.320
14	0.057	0.021	0.017	3.276	7.870
15	0.021	0.013	0.013	2.896	7.980
16	0.095	0.021	0.011	3.217	8.120

there's no theoretical law and practical experience to be used, use the point depiction method to draw the actual measuring points on the paper with rectangular coordinates and observe the nearest function curve of the distribution trend of the actual measuring points and then use the function relation and make it matching with the curve relation; third, in the SPSS viewer, make judgment according to the squared value of the correlation coefficient and the larger the squared value of the correlation coefficient is, the better the model will be Zagorska *et al*. (2011).

Curve regression analysis:

Step 1: Input the data in Table 1 into SPSS data editor, select "Analyze-Regression-Curve Estimation" in turn and then the dialog box of (Curve Estimation) can be opened.

Step 2: Select the variable " " and enter into the box of "Variables". Make "Y" enter into the box of "Dependents". And in the column of "Models", select all the options and press (OK) to end it.

In the column of "Models", there are 11 kinds of fitting models and the function expression that is represented by all model options is as follows.

For the option of Linear, use a linear model for fitting and the model is: $y = b_0 + b_1 x$.

For the option of Quadratic, use a quadratic polynomial for fitting and the model is: $y = b_0 + b_1 x + b_2 x^2$.

For the option of Compound, use a complex model for fitting and the model is: $y = b_0 (b_1)^x$.

For the option of Growth, use a growth model for fitting and the model is: $y = e^{(b_0 + b_1 x)}$.

For the option of Logarithmic, use a logarithmic model for fitting and the model is: $y = b_0 + b_1 \ln (x)$.

For the option of cubic, use a cubic polynomial for fitting and the model is: $y = b_0 + b_1 x + b_2 x^2 + b_3 x^3$.

For the option of s, use a s curve for fitting and the model is: $y = \exp (b_0 + b_1) /x$.

For the option of Exponential, use a logarithmic model for fitting and the model is: $y = b_0 e^{b_1 x}$.

For the option of inverse, use a hyperbolic curve for fitting and the model is: $y = (b_0 + b_1) /x$.

For the option of Power, use a power exponent model for fitting and the model is: $y = b_0 x^{b_1}$.

For the option of Logistic, use a logic model for fitting and the model is: $y = 1/ (1/u + b_0 b_1 x)$.

Multiple regression analysis: In food scientific researches, there are many independent variables that can affect the dependent variable y. For instance, the factors that affect the quality of food products include the processing temperature, the pressure of sterilization, the time of sterilization, PH value, etc. Thus, it is necessary to further discuss the regression of 1 dependent variable and several independent variables as well as the relevant problems (Wythe *et al*., 2013). The basic task of multiple regression analysis is to establish the regression equation between the dependent variable and all the independent variables according to the actual observation values of all dependent variables and independent variables so as to reveal the specific form of the relation between the dependent variable and all the independent variables and the purpose is to use the established regression equation for prediction and control. The calculation of multiple regression analysis is complex and SPSS software can make the computational process "a fool". For example, in the research of the reasons of non-enzymatic browning in the manufacturing process of a brand of peach juice, the ieucoanthocyanin (X1), anthocyanin (X2), Maillard reaction (X3) and non-enzymatic browning package value (Y) and the results are shown in Table 1 and now we make regression analysis on the relationship between Y and X1, X2 and X3:

Step 1: After input the data in Table 2 into SPSS data editor, select "Analyze-Regression Linear" in turn and then the dialog box of (Linear Regression) can be opened.

Table 2: Summary table of models

Model	R	R^2	Adjusted R^2	S.E.E.	R^2 change	F change	df1	df2	Sig. F change
					Change statistics				
1	0.821	0.674	0.651	0.86640	0.674	28.971	1	14	0
2	0.885	0.784	0.751	0.73227	0.110	6.598	1	13	0.023
3	0.939	0.882	0.853	0.56235	0.098	10.044	1	12	0.008

a predictors: (constant) X2; b predictors: (constant) X2 X1; c predictors: (constant) X2 X1 X4; d dependent variable: Y; S.E.E.: Standard error of the estimate

Table 3: Output results of the curve regression

Equation	Model summary					Parameter estimates			
	R^2	F	df1	df2	Sig.	Constant	b1	b2	b3
Linear	0.691	15.639	1	7	0.005	24.337	4.373		
Logarithmic	0.886	54.382	1	7	0.000	18.697	20.686		
Inverse	0.984	435.179	1	7	0.000	68.286	-64.850		
Quadratic	0.923	35.739	2	6	0.000	2.534	17.289	-1.190	
Cubic	0.986	115.349	3	5	0.000	-19.443	36.714	-5.509	0.264
Compound	0.592	10.172	1	7	0.015	24.459	1.119		
Power	0.809	29.638	1	7	0.001	20.755	0.551		
S	0.962	179.038	1	7	0.000	4.378	-1.787		
Growth	0.592	10.173	1	7	0.015	3.197	0.113		
Exponential	0.594	10.172	1	7	0.015	24.459	0.113		
Logistic	0.591	10.170	1	7	0.015	0.041	0.893		

Step 2: Select the left "Y" into the right "dependent (dependent variable), select "X1", "X2", "X3" and "X4" into the right "Independent" (independent variable and select "Stepwise (stepwise regression method)" in the "Method" here, Enter, Forward, Backward and other methods also can be selected.

Step 3: After pressing the button (statistic···), check "Estimates" (estimated value), "Confidence intervals (confidence interval)", "Modelfit" (test of goodness of fit of regression model), "Rsquared change (square of correlation coefficient)" as well as the residual "Durbin Watson" and then click the (Continue) button to go back to the dialog box of (Linear Regression) and press (OK) to end it.

RESULTS AND DISCUSSION

Curve regression analysis: The output results are shown in Table 3.

In the above results, the meaning of each column of data is:

- Dependent Variable: dependent variable
- Equation: the selected model used for fitting
- R Square: the squared value of the correlation coefficient
- F: F value
- df1 and df2: degree of freedom
- Sig: significance level
- Constant b0: constant term
- b1-b3: the fitted values of the undetermined coefficients in the model

The results in Table 3 are the optimal fitted results of all models and the advantages and disadvantages of all models can be compared by comparing the squared values (R Square) of the correlation coefficients. The larger the squared values of the correlation coefficients are, the better the model is. In CUB (the cubic polynomial model), Rsq = 0.986, namely the squared value of the correlation coefficient is bigger, so choosing this kind of model for fitting is the most appropriate. The model of adopting the cubic polynomial for fitting is:

$$y = -19.443 + 36.714x - 5.509x^2 + 0.264x^3$$

Multiple regression analysis: Table 2 is the summary table of models. In the table, every step of the correlation coefficient (R), the square of the correlation coefficient (R Square), the adjusted square of the correlation coefficient (Adjusted R Square), the standard error of estimate (Std. Error of the Estimate) and the change statistics (Change Stat cs) including the square of the correlation coefficient (R square Change), F value (F Change), the first degree of freedom (df), the second degree of freedom (df2) and the significant possibility (Sig. F Change) etc., have been listed. The footnote under the table shows every step of item used for prediction (including independent variables and constant terms).

Table 2 is the table of coefficient analysis. In the table, every step of constant terms and the corresponding un-standardized coefficients have been listed, including the values (B) and the standard error (Std Error) of the coefficients of constant terms and variables, Beta value of standardized coefficients, t

value and the significance level (Sig.) and the values of the undetermined coefficients of independent variables and the confidence interval of 95% of constant terms (95% Confidence Interval for B).

Based on the above information, we can obtain the multiple regression equation by the stepwise regression method:

$$y = -79.641x_1 + 207.215x_2 + 1.491x_4 + 5.484$$

and the correlation coefficient of the model i 0.939.

REFERENCES

Du, L. and Y. Ke, 2013. Research and development on food nutrition statistical analysis software system. Adv. J. Food Sci. Technol., 5(12): 1637-1640.

Eke, M.O., N.I. Olaitan and H.I. Sule, 2013. Nutritional evaluation of yoghurt-like product from baobab (*Adansonia digitata*) fruit pulp emulsion and the micronutrient content of baobab leaves. Adv. J. Food Sci. Technol., 5(10): 1266-1270.

Jiwen, G., R. Guihua, M. Wenjie, C. Huafeng and W. Shuyuan, 2013. Water quality assessment of Gufu river in three gorges reservoir (China) using multivariable statistical methods. Adv. J. Food Sci. Technol., 5(7): 908-920.

Wythe, H., C. Wilkinson, J. Orme, L. Meredith and E. Weitkamp, 2013. Food hygiene challenges in older people: Intergenerational learning as a health asset. WIT Tr. Biomed. Health, 16: 211-224.

Zagorska, J., I. Eihvalde, I. Gramatina and S. Sarvi, 2011. Evaluation of colostrum quality and new possibilities for its application. Proceeding of the 6th Baltic Conference on Food Science and Technology: Innovations for Food Science and Production (FOODBALT-2011), pp: 45-49.

Zhao, H., B. Guo, Y. Wei, B. Zhang, S. Sun, L. Zhang and J. Yan, 2011. Determining the geographic origin of wheat using multielement analysis and multivariate statistics. J. Agr. Food Chem., 59(9): 4397-4402.