Modelling in food technology

RADOSLAV MAVREVSKI
Department of Informatics
University Center for Advanced Bioinformatics Research
South-West University "Neofit Rilski"
66 Ivan Mihaylov Str., 2700 Blagoevgrad
BULGARIA
radoslav_sm@abv.bg https://ais.swu.bg/profile/mavrevski

Abstract: - The modelling in food technology is concerned with the question of the best model choice. The aim of the presented work is to show the basic methods for modelling and criteria for model selection in food technology, in order to develop a reliable approach for prediction of their behavior. Models in food technology, at the simplest level, are equations showing the relationship between two or more variables. Mathematical model of a process can be defined as a system of equations whose solution, given specified input data, is representative of the response to a corresponding set of inputs. Curve fitting and statistics in this work was made by Prism software. Akaike's information criteria (AIC) and Bayesian information criteria (BIC) was used in the model selection.

Key-Words: - Modelling, least squares fitting, regression, model selection, Akaike information criteria (AIC), Bayesian information criteria (BIC)

1 Introduction

Most models used in food technology are empirical. Two types of empirical model, response surface methodology (RSM) and dimensional analysis are widely used in food processing. RSM is a graphical representation of the statistical relationship between process output and independent variables, whilst dimensional analysis is a technique, which combines physical parameters that describe the problem in such a way to produce new, dimensionless, variables, and interactions. Fundamental models, or models based on theory, require knowledge of the underlying principles and mechanisms and the relationships between variables.

Modelling enables product or process knowledge to be expressed in simple statements thus reducing the apparent complexity of some problems and facilitating a solution. Has the potential to reduce the cost of experimentation, by reducing the number of experiments needed to analyses a particular problem. Mathematical modelling allows alternatives to be considered which may be difficult, or expensive to test and enables the sensitivity of a process to variables, and the design of optimal control strategies, to be studied. For example, in past, because of the importance of ensuring food at the point of consumption is free of pathogens and their toxins, considerable research has been devoted to modelling microbial growth and toxin production in foods. In the presented work, was modeled the consumption of cereals (excluding rice) per capita for France from 1995 to 2012 year. France was selected, as it is the leader in agriculture across Europe and number one in the grain.

2 Problem Formulation

2.1 Use of models in problem solving

Problem solving is something that the technologist does every day in industry. We are going to explore what is involved in problem solving by studying problems. Essentially the approach that we propose is to:

1. Defining, explaining and moving towards actually understanding what the problem;
2. Based on this understanding, and our knowledge of food technology, including any relevant product or process models we will devise a plan to investigate the problem;
3. Implementation our plan;
4. Review our findings to determine whether we have arrived at a solution.

This may be a cyclical process, where we progress to a solution in a logical way by eliminating a number of potential solutions. Questions that we might ask:

• Are there any product or process models that might be relevant?

• Has this problem occurred before? If yes, was it investigated? What was the outcome?

• Do you how anyone who might have experience of this problem?

• What do the operatives, supervisors, and managers have to say?

• Look at the data. Are there any trends?

• How do you know that the analyses or responses to questions are accurate? Could the responses or results be incorrect, invented?

• Have there been any personnel changes problems?

• Is the crisis management plan relevant?

With regression analysis we investigate the association between two or more variables to predict or estimate the value of one variable or more. In every regression problem we find two different kind of variables: a) the independent or predictor or casual or explanatory variables and b) the dependent or response variables. In an experimental research, the independent variable X is the one we can control, for instance, the price of an advertisement campaign of a product or the processing temperature of a product. The dependent variable Y is the one that is reflected from the values of the independent variable, for instance, the request of a product, the strength of a material. In a non-experimental research the discrimination between independent and dependent variables is not controled, it is random. In regression analysis there is an asymmetry in the way we handle the variables, as the one is considered as the 'result' and the other one(s) as a 'cause'. Further the dependent variable is considered as stochastic, probabilistic, while the independent variable at first is considered as non stochastic, probabilistic. For every value of the illustrative variable we consider there is a distribution of values of the dependent variable.

Simple regression is when only one predictor or independent variable is available for predicting the response of interest. Scales of measurement may be discrete and continuous, even though in practical applications, the independent variable under investigation is more often on a continuous scale. In order to choose an appropriate technique to analyze data, one must identify the type of variable under investigation. It is common in a lot of researches, the dependent variable to be a continuous variable, which we may assume after the appropriate transformation, that is normally distributed. So, the model of regression describes the mean of the normally distributed dependent variable (Y) as a function of independent variable (X). The simple linear regression model is:

$$Y_i=\beta_0 + \beta_1\chi_i+ \varepsilon_{i,}$$

Y_i = value of the response or dependent variable from the ith pair
β_0 & β_1 = two unknown parameters
χ_i = value of the independent variable from the ith pair
ε_i = random error term

Figure 1 shown mathematical modeling from a predictive point of view.

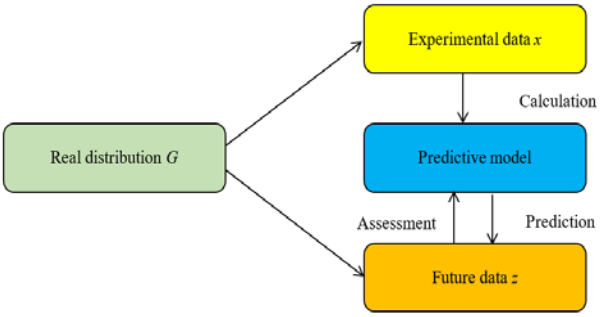


Fig. 1. Mathematical modeling for prediction

2.2 Steps of modelling

Mathematical modeling is an integral part of research in many fields. A mathematical model is a description of a system using mathematical tools. Mathematical models are used not only in the natural sciences and engineering disciplines, but they are also used in biology, economics and sociology. We usually try to determine whether the experimental data are consistent with a particular theoretical relationship and to find the model in the class model M describing this relationship. The best fit of the curve to the experimental data is quantitatively defined as the minimization of some well described criteria with respect the parameters of the models.

Usually, to find "optimal" model P_i can be applied the following steps: for each class models M_i, we find the "optimal" model P_i, using least squares or robust fit; compare models P_i, $i=1,\ldots k$, using some criteria for model selection, such as Akaike's information criteria (AIC) [1,2,3,4] and Bayesian information criteria (BIC) [2,3,4] we find

the "optimal" model *P* with respect of the applying criteria.

Finding of the individual "optimal" members P_i in the classes M_i was made by using of the well know software GraphPad Prism. GraphPad Prism combines curve fitting, basic biostatistics, and scientific graphing (http://www.graphpad.com/scientific-software/prism).

GraphPad Prism combines curve fitting, basic biostatistics, and scientific graphing. In Prism we have an opportunity to choose a model that corresponds to the experimental design from Prism's a menu of the 15 equations or enter our own equation. The results are stores with our data and graphs, along with our analysis choices so you can inspect your choices and results at any time and add notes to explain what you did and what you concluded. When we repeat an experiment, simply we replace the data, and Prism repeats all the analyses and redraws the graphs. Prism is used in the following way:

firstly, we enter the experimental data in the data table, in a column *X* enter the independent variable and in a Columns *Y* dependent enter the variable. Prism then automatically analyses an entire family of data sets at once and automatically creates the graph with curve, error bars and legend – ready to customize;

secondly, we choose suitable model from the menu Analyze->Nonlinear Regression a Prism;

thirdly, we choose fitting method (least squares or robust fit, least squares in our case). If it is suspected that experimental mistakes can lead to erroneous values whose values are way too high or too low – outliers, we must choose a robust fit using a method that is not very sensitive to violations of the Gaussian assumption [5], otherwise we choose least squares or robust fit. After that Prism automatically find optimal coefficients of the corresponding model P_i.

In the software package Prism applied in this study for nonlinear regression, there is no guarantee that the non-found solution is not local, but a global extremum. In order to have high probability for the solving to be the global minimum, we start with a large number of different initial parameter values $a_i, i = 1, \dots, k$.

After finding the individual "optimal" polynomials P_i in the chosen classes M_i we apply criteria accordingly AIC and BIC to select one of P_i named on "optimal" model *P*, according the criteria of optimal selection.

In the past have been proposed different models, but little work has been done on comparing of models. Comparing least square errors between measured and modelled data, indicates the goodness of fit of each model. However, the least squares statistic does not take into account the tradeoff between model complexity and estimation errors. Models of increased complexity can better adapt to fit to data. However, additional parameters may fit to measurement noise and not describe any important processes. Using solely the model that gives the lowest mean square error will often just lead to the largest model being selected as optimal. The "optimal" model should balance simplicity and goodness of fit.

Model selection is a process of choosing a model from a set of candidate models from different classes which will provide the best balance between goodness of fit of the data and complexity of the model.

One of the most commonly used criteria for model selection is AIC. The idea of AIC is to select the model that minimises the negative likelihood penalised by the number of parameters.

Another the most commonly used criteria BIC has the highest posterior probability [1, 3, 4, 6, 7, 8]. As it can be seen by comparing the formulas below for AIC and BIC, these two criteria differ only in that the coefficient multiplies the number of parameters. In otherwords, the criteria differ by how strongly they penalize large models. In general, models chosen by BIC will be more parsimonious than those chosen by AIC.

Below we present the general explicate formulas for calculating of the mentioned two criteria:

$$AIC = \begin{cases} nln\left(\frac{RSS}{n}\right) + 2k, & \frac{n}{k} \geq 40 \\ nln\left(\frac{RSS}{n}\right) + 2k + \frac{2k(k+1)}{n-k-1}, & \frac{n}{k} < 40, \end{cases}$$

$$BIC = n * ln\left(\frac{RSS}{n}\right) + k * ln(n),$$

where:
n is the number of data points; *k* is the number of parameters fit by the regression plus one (because regression is "estimating" of the sum-of-squares as well as the values of the parameters); *RSS* or sum of square error is the sum of the squares of the vertical deviations from each data point to the fitted line.

The model with the lower value of AIC or BIC is defined as the "optimal" model.

3 Problem Solution

In Table 1 are shown data for the consumption of cereals (excluding rice) per capita for France from 1995 to 2012 year.

Table 1. The consumption of cereals (excluding rice) per capita for France

Consumption by years																	
1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
96.9	96.6	98.6	100.8	111.4	109.2	113.5	109.5	104.6	104.9	107.9	109.4	112.5	116.7	111.4	114.9	113.6	113.8

Figure 2 shown consumption of cereals from 1995 to 2012 year for France.

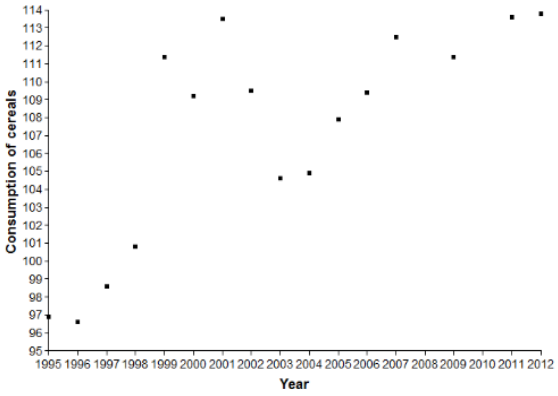


Fig. 2. Scatter plot of consumption of cereals by years

According behavior of the data from scatter plot, they data was modeled with polynomial second degree and polynomial third degree, because obviously the dependence is nonlinear. The individual optimal model in quadratic class model is *Consumption of cereals = -17.53 – 0.8198Year + 0.0004405Year*^{2} and individual optimal model in cubic class model is *Consumption of cereals = -0.4942 - 123.5Year + 0.1229Year*^{2} *+ -3.0570e-005Year*^{3} (see Figure 3.).

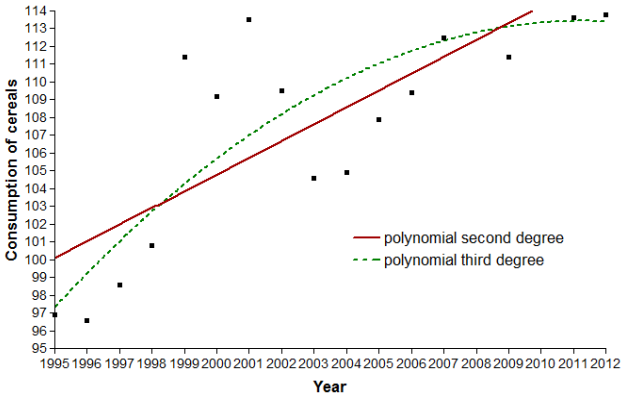
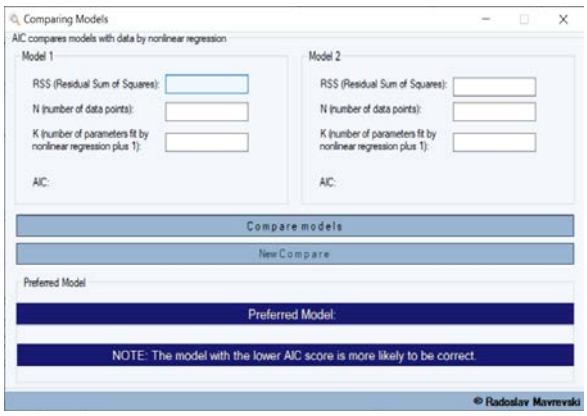


Fig. 3. Modelling with second and third degree polynomial

For calculation of the criteria values of AIC and BIC according to formulas was use a program "Comparing Models" developed by us in our previous research (see Figure 4), [6].

a)



b)

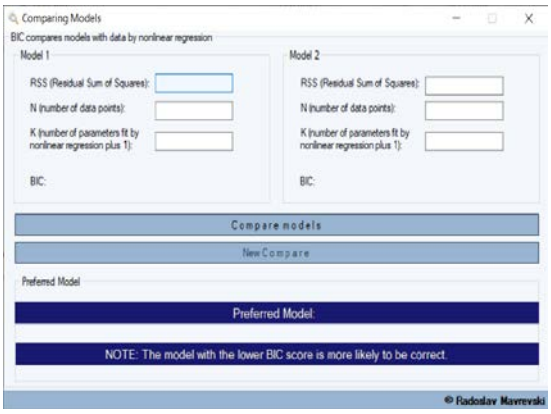


Fig. 4. Program "Comparing Models": (a) dialogue box of the program "Comparing Models" for calculating AIC; (b) dialogue box of the program "Comparing Models" for calculating BIC

Table 2. The values of the criteria for model selection

Model	**AIC**	**BIC**
Polynomial second degree	52.08	53.06
Polynomial third degree	**51.94**	**52.90**

As shown in Table 2 according both criteria AIC and BIC, optimal model is class model polynomial third degree and therefore the best model for relationships consumption of cereals - years is *Consumption of cereals = -0.4942 - 123.5Year + 0.1229Year^2 + -3.0570e-005Year^3*.

The most widely used method of parameter estimation from curve fitting is the method of least squares. A mathematical procedure for finding the best-fitting curve to a given set of points by minimizing the sum of the squares of the offsets (residuals) of the points from the curve. However, because squares of the offsets are used, outlying points can have a disproportionate effect on the fit. Ordinary least squares models are often heavily influenced by the presence of outliers. Outliers are data points which do not follow the general trend of the other observations, although there is strictly no precise definition of an outlier.

Robust regression refers to regression algorithms which are robust to outliers. Its inability to compute standard errors or confidence intervals of the parameters greatly limits the usefulness of robust regression [5, 9, 10].

Selection on the best model was made by used two commonly used model selection criteria, Akaike's Information Criterion (AIC) and Bayesian Information Criterion (BIC).

4 Conclusion

The selected model can assist in explaining the respectively relationships of the food technology and can be successfully used in predicting this relationship and save time, make no experiments and expenses. Least squares fitted methods, which can be used in fitted of different data were discussed.

Different criteria which may be used in the model selection in both nested and non-nested models, and which does not rely on *P* values or the concept of statistical significance were proposed in that study.

This study presents an easy-to-understand step-by-step way to applying regression analysis that can be applied to a function in the form $y = f(x)$ and is very suitable for rapid and reliable data analysis in different areas in food industry and technology.

Regression modeling applied in this work is one of the most widely used statistical modeling techniques in food technology for fitting a quantitative response variable *y* as a function of one or more predictor variables *x*1, *x*2, . . . , *xp*. Regression models can be used to variables either are quantitative or are indicators of factor levels. The regression analysis procedures contain diagnostic techniques for (a) identifying incorrect specifications of the model (b) assessing the influence of outliers on the fit and (c) evaluating whether redundancies (collinearities) among the predictor variables are adversely affecting the fit. Regression models are widely used because they often provide excellent fits to a response variable when the true functional relationship between the response and the predictors is unknown.

References:

[1] Akaike H. A New Look at the Statistical Model Identification. IEEE Transactions on Automatic Control, 19, 1974, No 6, 716-723.

[2] Acquah H. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. Journal of Development and Agricultural Economics, 2, 2010, No 1, 001-006.

[3] Joseph. B., L. Nicole. Methods and Criteria for Model Selection. Journal of the American Statistical Association, 99, 2004, No 465, 279-290.

[4] Zucchini W. An Introduction to Model Selection. Journal of Mathematical Psychology 44, 2000, 41-61.

[5] Motulsky H., R. Brown. Detecting outliers when fitting data with nonlinear regression – a new method based on robust nonlinear regression and the false discovery rate. BMC Bioinformatics, 7, 2006, No 1,123-143.

[6] Mavrevski R., Selection and comparison of regression models: estimation of torque-angle relationships. C. R. Acad. Bulg. Sci., 2014, 67(10), 1345-1354.

[7] Mavrevski R., M. Traykov, I. Trenchev, M. Trencheva. Approaches to modeling of biological experimental data with GraphPad Prism software. WSEAS TRANSACTIONS on SYSTEMS and CONTROL, 2018, 13, 242-247

[8] Traykov M, M. Trencheva, I. Todorin, R. Mavrevski, A. Stoilov, I. Trenchev. Risk Analysis with R Language. Proceedings of the Sixth International Scientific Conference – FMNS2015, vol. 1, ISSN 1314-0272, pp. 137 – 146.

[9] Burnham P., Anderson D. Model Selection and Multimodel Inference 2 ed., Springer-Verlag, New York, 2002.

[10] Bickel P., Zhang P. Variable Selection in Nonparametric Regression with Categorical Covariates, J. Am. Stat. Assoc. 87, 1992, 90–97.