Approaches to modeling of biological experimental data with GraphPad Prism software

RADOSLAV MAVREVSKI
Department of Electrical Engineering, Electronics and Automatics
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

METODI TRAYKOV
Department of Electrical Engineering, Electronics and Automatics
University Center for Advanced Bioinformatics Research
South-West University "Neofit Rilski"
66 Ivan Mihaylov Str., 2700 Blagoevgrad
BULGARIA

IVAN TRENCHEV
Department of Electrical Engineering, Electronics and Automatics
University Center for Advanced Bioinformatics Research
South-West University "Neofit Rilski"
66 Ivan Mihaylov Str., 2700 Blagoevgrad
BULGARIA

MIGLENA TRENCHEVA
Department of Finance and Accounting
South-West University "Neofit Rilski"
66 Ivan Mihaylov Str., 2700 Blagoevgrad
BULGARIA

Abstract: - Mathematical models are commonly used in biological sciences. To understand complex biological systems such as cells, tissues, or others, it is not enough to identify and characterize only individual molecules in the system. It also is necessary to obtain a thorough understanding of the interaction between molecules and different pathways. Computational models help investigators to analyze systems, develop hypotheses to guide the design of new experimental tests. Known are mathematical methods referring to different categories of biological processes. Now, modeling approaches are essential for biologists, enabling them to analyze complex physiological processes. The aim of this study is to presents a step-by-step applying non-linear regression analysis for fast and effective data analysis in the biology. To achieve this aim is used non-linear regression analysis method by GraphPad Prism software and the modeling of specific experimental data taken from available literature. Nonlinear regression is an extremely useful tool in analyzing data, but choosing a model is a scientific decision based on biology, chemistry or physiology and etc. and not be based solely on the shape of the graph.

Key-Words: - Mathematical models, fitting, model selection criteria, biological data

1 Introduction

In the past it was proposed different methods and models for describing different biological process, but little work was done on way of modeling and comparing of the models [1, 2, 3, 4].

The purpose of this study is to introduce a simple, consistent and easy-to-understand way to perform non-linear regression analysis based on user input functions using the GraphPad Prism software package. Although it is relatively easy to fit data with simple functions such as linear or logarithmic functions, data fitting with more complex non-linear functions is more difficult. This study presents an easy-to-understand step-by-step guide to applying this non-linear regression analysis method 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 of biology. The application of nonlinear regression techniques to describe experimental data is widespread across wide areas of biology.

The purpose of the biological data curve is to describe the data in the form *y = f(x)*, where *y* is the dependent variable and is measured in the experiment and *x* is controlled during the experiment and is called an independent variable whose value is fixed at X axis. *f* is a function (mathematical model) used to describe the relationship between x and y and is in the form of an equation consisting of one or more parameters. In general, better fitting of the curve means a better description of the data.

2 Problem Formulation

The problem of choosing an "optimal" model is one of the most fundamental problems in analyzing experimental data in biology.

To find the individual "optimal" models in the candidate models, are used different approximation (fitting) approaches, such as the least squares (LS) method and robust (stable) regression (RR) available in GraphPad Prism.

Frequently used in the literature, criteria for evaluating models from different classes are:

- Akaike’s information criterion (AIC) (Akaike H., 1973), **available in the latest version of Prism 6.0**;
- ayes Information Criterion (BIC) proposed by (Schwarz, G., 1978).

2.1 Least squares method

The purpose of the smallest squares method is to minimize the sum of squares of vertical deviations or (distances between Y-values) between points and the curve [5].

To find the individual “optimal” models $P^*(M_j)$ in the classes M_j, j = 1, …, n, we use least squares fitting in GraphPad Prism 6.0. Least squares fitting criterion is defined as follows:

$$F(a) = \sum_{i=1}^{n} (y_i - f(x_i, a_1, \dots, a_s))^2 \qquad (1)$$

The problem is to find $a^* = (a_1^*, \dots, a_s^*)$, such that minimizes $F(a)$.

2.2 Robust (stable) regression

Experimental errors may result in erroneous experimental data whose values are too high or too low. Even a single wrong value among experimental data can have an impact on the calculation of the sum of the smallest squares and lead to misleading results. One way of dealing with this problem is to make a robust regression using a method that is not very sensitive to violations of admission for normal data distribution.

Robust regression works by determining the weight for each point of the data. Weighing is done automatically and iteratively using a process called iterative weighing of the smallest squares.

If all data have normal distribution, the robust regression method and the smallest squares give almost identical results.

2.3 Model selection criteria for choice the best model available in GraphPad Prism 6.0

2.3.1 Akaike’s information criterion

One of the most commonly used criterion for model selection is AIC [1, 3, 4, 6, 7, 8, 9]. The idea of AIC is to select the model that minimizes the negative likelihood penalizing by the number of parameters. The formula for calculating AIC can be expressed as a function of the residual sum of squares (*RSS*):

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

where n is the number of data points; k is the number of the fitting parameters by the regression plus one (since regression is “an estimating” of the sum-of-squares as well as the values of the parameters); RSS, or residual sum of squares, is the sum of the squares of the vertical deviations from each data point to the graph of a curve of the “optimal” fitted model.

2.3.2 Coefficient of determination R^2

The coefficient of determination R^2 also can be used to compare regression models and is available in Graph Pad Prism. A model with a larger R^2 value means that the independent variables explain a larger percentage of the variation in the independent variable.

$$R^2 = \frac{explained\,variation}{total\,variation} \qquad (3)$$

However, this may conflict with parsimony. Usually R^2 not a good criterion for model selection from different classes. Always increase with model size “optimal” is to take the biggest model and never choose the true class model.

The comparison of interpolation utility of regression models with a different number of independent variables and a different number of parameters cannot be done by simple comparison of R^2. At least the adjusted R^2 (The adjusted R^2 always has a lower value than R^2 (unless you are fitting only one parameter).) must be used, but more sophisticated measures like an AIC.

2.4 GraphPad Prism

Prism is a powerful combination of biostatistics, curve fitting (nonlinear regression) and scientific graphing in one comprehensive program. Easily organize, analyze and graph repeated experiments; pick appropriate statistical tests and interpret the results. Seamless data analysis and full-featured graphing in one integrated program. You control the structure of your data table: store replicate determinations, related data sets, and repeated experiments all together; analyze all the data at once. Enter data directly or import ASCII files or data from other Windows programs. Menu-driven data analysis: select one of the built-in nonlinear regression models or define your own. Versatile graphing allows you to define all aspects of your graph design and save your choices with the data set. Context-sensitive help offers suggestions and background theory on data analyses, explains calculations, and lets you customize elements of a graph by clicking on them (https://www.graphpad.com/).

Place data for multiple data sets side-by-side on an organized data table, and Prism can fit them all the data sets at once. You can fit the same model separately to each data set, use global nonlinear regression to share parameter values among data sets, or fit different models to different data sets.

However, the simplicity, Prism also gives you many advanced fitting options. It can report the confidence intervals of the best-fit parameters as asymmetrical ranges (profile likelihood method), which are far more accurate than the usual symmetrical intervals. It can also automatically interpolate unknown values from a standard curve (i.e., to analyze RIA data), compare the fits of two equations using or Akaike's Information Criterion (AIC), plot residuals, identify outliers, differentially weight data points, test residuals for normality, and much more.

It was originally designed for experienced biologists in medical schools and pharmaceutical companies, especially those involved in pharmacology and physiology. Simplifies curve fitting by simply choosing an equation from a list of the most commonly used equations or introducing our own equation. Allows different fitting options such as LS and RR. Newer versions (6.0) of this product allow comparison of two equations using AIC.

Steps in modeling with GraphPad Prism

The steps in modeling with GraphPad Prism are:

1. creating a data scatter counter (X and Y points) to check if there is any trend in that data;
2. when there is an obvious tendency in a data set, an attempt is made to find a class of model (function) expressing this trend;
3. after the selection of classes of models, we use different fitting methods for finding the best models in the given classes; finding the individual model in the given classes is usually made by various fitting methods, such as the most widely used method of least squares fitting or other such as robust fitting;
4. the selected candidate “optimal” models in each class mentioned above, are compared with some commonly used criteria for the evaluation of these models such as AIC.

3 Problem Solution

3.1 Modeling (curve fitting) with GraphPad Prism

Experimental data (see Table 1 and Figure 1) which was used to demonstrate modeling are from article by Ashton et al. [10], where they measured the length of the shell in mm to 18 female turtles and the number of eggs in each using an X-ray [11].

After entering the experimental data in Prism, the automatically built scatter plot diagram is shows that linear regression is not appropriate (see Figure 2).

Table 1. Experimental data.

Number of eggs	Length of the turtle shell (mm)
284	3
290	2
290	7
290	7
298	11
299	12
302	10
306	8
306	8
309	9
310	10
311	13
317	7
317	9
320	6
323	13
334	2
334	8

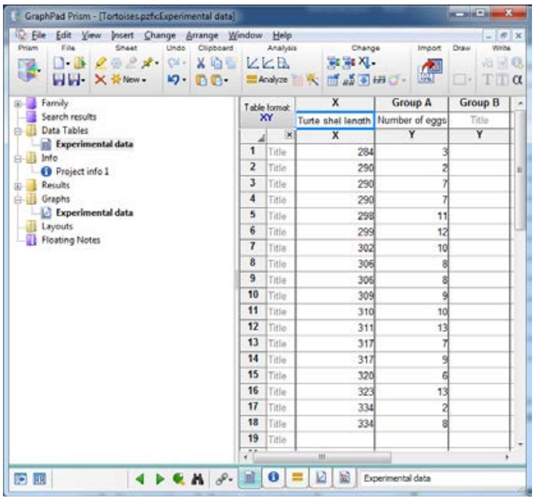

Table format: XY	X	Group A	Group B
	Turte shel length	Number of eggs	Title
	X	Y	Y
1 Title	284	3	
2 Title	290	2	
3 Title	290	7	
4 Title	290	7	
5 Title	298	11	
6 Title	299	12	
7 Title	302	10	
8 Title	306	8	
9 Title	306	8	
10 Title	309	9	
11 Title	310	10	
12 Title	311	13	
13 Title	317	7	
14 Title	317	9	
15 Title	320	6	
16 Title	323	13	
17 Title	334	2	
18 Title	334	8	
19 Title			

Fig. 1. Experimental data entered in GraphPad Prism.

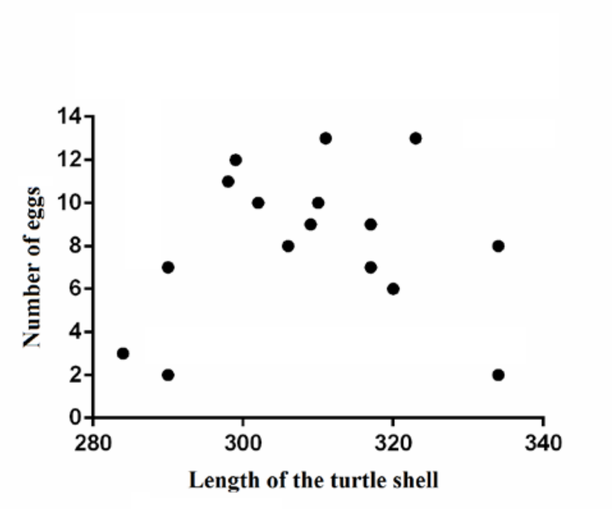

Fig. 2. Scatter plot from Prism.

In the studies in this work, the modeling of the experimental data for number of eggs - length of the turtle shell was made with the classes of polynomials of the second, third and fourth degree and the least squares fitting method (see Table 2 and Figure 3).

Table 2. Results from modelling of number of eggs - length of the turtle shell relationship.

Polynomial models	Parameters	Mean value	Standard error
Second degree $a + bx + cx^2$	**a**	**-899.90**	**270.30**
	b	**5.86**	**1.75**
	c	**-0.009**	**0.003**
Third degree $a + bx + cx^2 + dx^3$	a	-5108	6870
	b	46.72	66.67
	c	-0.14	0.22
	d	0.00014	0.00023
Fourth degree $a + bx + cx^2 + dx^3 + ex^4$	a	-90867	179895
	b	1160	2334
	c	-5.56	11.35
	d	0.012	0.025
	e	-0.000009	0.00002

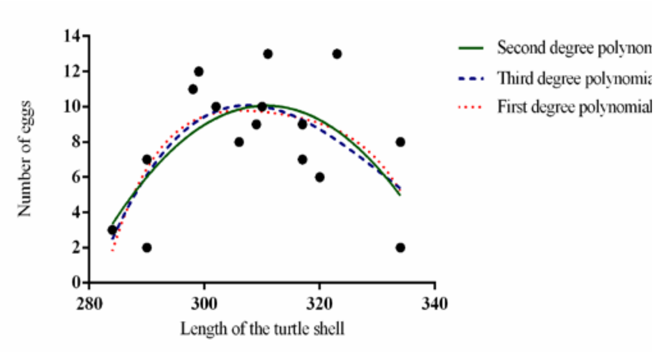

Fig. 3. Number of eggs - length of the turtle shell (mm) curves obtained by fitting with second, third and fourth degree polynomials by the least squares method in GraphPad Prism.

3.2 Comparison of class models (second, third and fourth degree polynomials) with the AIC criterion and adjusted R^2 in GraphPad Prism

AIC

GraphPad Prism 6.0 does not provide information about the AIC values for both models and calculates and results in a AIC difference report ΔAIC = AIC$_A$ - AIC$_B$, where A is the simpler model (with fewer parameters), and B is the more complex model (with more parameters). When the more complex model has a low AIC$_B$ value and is preferable, Prism give an account the difference in ΔAIC with positive value. When the simpler model has a lower AIC$_A$ value, and is preferable, Prism give an account the difference in ΔAIC with negative value (see Table 3).

Table 3 Comparison of class with the AIC criterion in GraphPad Prism.

Polynomial models	**ΔAIC** **(AIC differences)**	**Preferred model** **(more likely to be true)**
Second and Third degree Polinomyals	**-3.446**	**Second degree polynomial**
Second and Fourth degree Polinomyals	**-7.770**	**Second degree polynomial**

Adjusted R^2

A quick and easy way to compare models would seem to be to choose the one with the smaller adjusted R^2 (See Table 4).

Table 4 Comparison of class with the adjusted R^2 criterion in GraphPad Prism.

Polynomial models	Second order polynomial	Third order polynomial	Fourth order polynomial
Aadjusted R^2	**0.3583**	0.3305	0.2914

Figure 4 shown "optimal" class model according by AIC and adjusted R^2 model selection criteria in Prism.

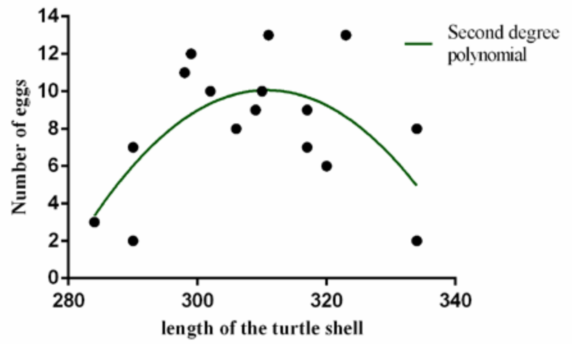

Fig. 4. Optimal curve of the number of eggs - length of the turtle shell (mm) relationship obtained by fitting with second degree polynomial: $-899.9 + 5.857x - 0.009425x2$·

In the first part of the plot it is not surprising and it is easy to imagine why the big tortoises will have more eggs. The drop in the number of eggs over 310 mm length of the turtle shell is an interesting result and it can be assumed that egg production drops in these turtles as they age. Generally, we should not use the curve for values outside the range of observed X-values, as extrapolation with polynomial regression is very likely to give erroneous results. For example, extrapolation with a second degree polynomial may indicate that turtles with a bottom length of less than 279 mm or greater than 343 mm may have a negative number of eggs.

4 Conclusion

In this study is introduce a simple, consistent and easy-to-understand way to perform non-linear regression analysis based on user input functions using the GraphPad Prism software package.

The compromise between the quality of approximation and the simplicity of the model under these criteria does not provide a solid basis for solving the problem of curve fitting, as these criteria ignore both the statistical adequacy and the reliability of the conclusions about the problem. For this, the assessment needs to be made on the basis of more than one model selection criteria. In addition, if all candidate fitting models are poor, the model selection criteria will not give a warning about this. Therefore, it is important to correctly formulate candidate models.

In the general case if you want to select the "optimal" model with the smallest mean square error, the AIC and / or R^2 criteria are very appropriate. They will choose the optimal model that has the relatively same error for each point in the experimental data point. These criteria report a compromise between the complexity of the model (number of parameters) and the accuracy [1, 3, 7].

The comparison of interpolation utility of regression models with a different number of parameters cannot be done by simple comparison of R^2. At least the adjusted R^2 must be used, but more sophisticated measures like an AIC is strongly advised.

A disadvantage of the used polynomial models is that in many cases they are not suitable for the biological interpretation and are worthless outside the range of observed data, i.e. cannot be used to predictions beyond this range.

Nonlinear regression is an extremely useful tool in analyzing data [8, 9]. The goal of nonlinear regression is to fit a model to our data. Nonlinear regression is one of the most powerful and useful features in Prism. Can to fit any model to our data to plot a curve and to determine the best-fit values of the model parameters. Prism can compare models and answer of question "For each data set, which of two equations (models) fits best?". Choosing a model is a scientific decision. You should base your choice on your understanding of biology, chemistry or physiology and etc. The choice should not be based solely on the shape of the graph. Letting a program choose a model for our can be useful if our goal is to simply create a smooth curve for simulations or interpolations. In these situations, we don't care about the value of the parameters or the meaning of the model. We only care that the curve fit the data well and does not wiggle too much. We suggest to avoid this approach when the goal of curve fitting is to fit the data to a model based on biological principles. Don't use a computer program as a way to avoid understanding your experimental system, or to avoid making scientific decisions.

Acknowledgements

This work is partially supported by the project of the Bulgarian National Science Fund, entitled: Bioinformatics research: protein folding, docking and prediction of biological activity, NSF I02 /16 12.12.14.

References:

[1] H. Acquah, Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship, J. Dev. Agric. Econ., **2**, 2010, 1-6.

[2] SJ. Ahn, Geometric Fitting of Parametric Curves and Surfaces, JIPS **4**, 2008, 153-158.

[3] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control, **19**, 1974, 716–772.

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

[5] ID. Coope, Circle fitting by linear and nonlinear least squares, J. Optim. Theory. Appl. **76**, 1993, 381-388.

[6] S. Konishi, G. Kitagawa, Information criteria and statistical modeling, New York: Springer Science and Business media, 2008.

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

[8] C.M Hurvich, C. Tsai, Regression and time series model selection in small samples, Biometrika, 76, 1989, 297-307.

[9] C.M. Hurvich, J.S. Simonoff, C-L. Tsa, Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, Journal of the Royal Statistical Society, 60, 1998, 271-293.

[10] K. G. Ashton, R. L. Burke, J. N. Layne, Geographic variation in body and clutch size of gopher tortoises, Copcia, 49, 2007, 355-363.

[11] N. Stoeva. The Right of the Personal Data Protection - Nature and Guarantees. Proceedings of the International Scientific Seminar “Intellectual Property in Bulgaria - Perception, Awareness and Behavior” Trencheva, T. (compl.), Za bukvite-O Pismeneh, Sofia, 2018, pp. 89-104.