Bull. Grad. School Educ. Hiroshima Univ, Part II, No.53, 2004, 49-56

Power of Parametric and Semi-parametric Models to Represent Real Life Data Situations

S. Kageyama, Satyabrata Pal[1], Subhabaha Pal[2], C. Medda[3] and T. K. Basu[3]

(Received July 16, 2004)

Real-life data situations are not easy to model because of the uncertainty and variation prevailing in the natural system and also due to the presence of the interaction of unknown factors interplaying and governing the entire real-life phenomena. Modelling fish dynamics has been considered as an indispensable area of research by the biologists, bio-researchers, applied and bio-mathematicians to understand the science behind the growth of fish. Valuable contributions in this area are available in different monographs, books and journals. The models prevalent for use are, generally, of two types – deterministic and non-deterministic. The celebrated Von Bertalanffy (VB) model, fitted deterministically, is the main thurst area of research till now to the active biology-modellers and is also considered as the most appropriate one. Non-deterministic models are widely applicable for their inherent capability to provide better representations owing to the fact that these imbibe elements of uncertainty inbuilt in their systems of operation. Efforts in this area call upon using a log transformation on a relationship representing an exponential relationship between the variables, weight and length of fish. The object of data modelling is to generate fitted values as close as possible to the observed values. Longitudinal data modelling can be achieved by parametric and nonparametric modelling. Deterministic fitting of VB equation ensures achieving precision (desired closeness) only up to a certain degree of precision. Precise modular representation of such data comes as a useful complement. This paper deals with non-linear fit (achieved through non-linear optimization), a parametric approach, of the VB model. Also introduced here is the application of distribution free approach, like, nonparametric model fitting, which does not involve parameters (which need to be estimated from the data). Such models assign more importance to data points and the span of the regions surrounding those. This paper exposes the power (in respect of achieving better representation) of non-linear fitting of the celebrated VB model and also of the nonparametric model fitting approaches (spline, loess, kernel), when these have been called upon to represent the growth dynamics of *Puntius sophore* (*Ham*), reared in experimental hoopnets. Indeed, these models are found to have more representative power many times, when compared against different performance criteria, to explain and model the real-life data situations. The focus of the paper points to the potentiality of the parametric and semi-parametric models to model many data-situations very closely and appropriately.

Keywords: Von Bertalanffy model, non-linear optimization, non-parametric models,

[1] Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia-741252, India
[2] Post Graduate Student, University of Calcutta, Ballygunge Campus, Kolkata-700019, India
[3] Indian Statistical Institute, Kolkata-700108, India

spline model, LOESS model, kernel model.

1. Introduction

The study on the aspect of growth (in respect of length) of fish over time offers a perpetual challenge to the fishery biologists as knowledge of its accurate length at specific points in the life span is essential to monitor and exploit the commercial advantages derived by rearing the animal in firms as fish stands second in the list of daily food items, consumed by human being all over the world, just after the cereal in order of importance. Over the ages, the celebrated Von Bertalanffy (VB) model, fitted deterministically, has proved to be the most ideal representation depicting the growth dynamics of fish. To capture the data characteristics, emanated from a real life data situation, a non-deterministic fit, achieved by employing non-linear optimization, has been investigated in this paper. It is found to have yielded slightly better representation when judged against a measure of closeness (agreement between observed and fitted values). Also introduced, here, are some nonparametric models, which do not possess nice parametric interpretations, but offer more precise representations, when evaluated against different measures of closeness. Indeed, closer representation of the observed data may be quite demanding in view of the fact that the generation of research data from experiments such as the one mentioned below is expensive and also such data are bound to imbibe specific features though following the underlying system from which they are created. Needless to mention that the focus of the paper is geared to answer the above observations and if the degree of closeness between the observed and fitted values is also considered as valuable, nonparametric/semi-parametric models can favorably be used to determine the values of lengths of *Puntius sophore* (*Ham*), reared in hoopnets, at time points not observed under the area of study of the experiment, as such models assign more importance to the data points and the span of the regions surrounding those. Nonparametric/semi-parametric models, spline, kernel and LOESS, have been developed and examined with respect to the data-set considered here and it has been observed that these models are very much potent as these increase the precision levels of predicted values when compared against parametric models. References of research works in these areas are, Bagenal (1978), Draper and Smith (1998), Le Cren (1951), Petrakis and Stergion (1955), Simonoff (1995) and Thisted (1988).

2. Material and method

Fish samples, *Puntius sophore* (*Ham*), are collected from the localities and were kept in the experimental hoopnets of Indian Statistical Institute, Kolkata. In order to know the physico-chemical nature of the ambient medium during the experimental period, the following measurements on the temperature and Dissolved Oxygen (DO) were recorded: temperature, 29-34°C; pH 6.2-7.1; DO 8.6-10.2 ppm. The experiment was conducted from the month of January 1998 and 10 fish samples were collected from the 4th week of hatching to 48th week of development at 4 week-intervals in the year. The live length (from the tip of the snout to the end of the caudal fin) was measured in centimeters and then each fish was weighted with a precision balance in grams (Bagenal, 1978). No supplementary food was introduced during the period of the experiment. The data set comprised of the mean value of 10 fish samples on lengths and weights of 12 observations respectively, each observation being recorded after a period constituting 4 week-interval. Ultimately, weekly averages (extending from the 4th week to the 48th week) are considered for development of dynamic models.

A dynamic model based on length determines the body size (length) of fish as a function of age (time). The mathematical model developed for individual growth by Von Bertalanffy (VB) has been shown to conform to the observed growth for most of the fish species and it is regarded as a cornerstone in fishery biology. The expression of the model is given as

$$L_t = L_\infty * [1 - \exp(-K * (t - t_0))], \tag{2.1}$$

where L_∞ is the length of the fish at the end of the life-time (in fact, it is the maximum length), L_t is the length of fish at time t, K is the value of the growth parameter (curvature parameter) or growth rate, and t_0 is the time point when the fish has zero length. Biologically, this t_0 is not meaningful as the growth begins at hatching when the larva has already a certain length. The above parameters are to be estimated from a given set of data. In the following, some simple explanations are presented. If $t = 0$, then $L_0 = L_\infty * [1 - \exp(Kt_0)]$, meaning there-by that the length of the larva has a certain value at birth. It is a fact that the growth rate of fish varies with age, as age increases growth rate decreases.

NON-LINEAR FITTING

The equation (2.1) is fitted non-deterministically, or in other words, an error term is included in the equation. The equation is written as

$$y = \ell * [1 - \exp(K * (x - t))] + e,$$

where y is the fish length, ℓ is the length of the fish at the end of the life-time, x represents time, e represents error and t is a time point when the fish has zero length. The above equation has been fitted using Lavenberg Marquardt algorithm (Draper and Smith, 1998) which uses derivatives or approximations to derivatives of the sum of squared errors with respect to the parameters to guide the search for the parameters producing the smallest sum of squared errors.

SPLINE FITTING

The spline procedure uses the penalized least squares method (Simonoff, 1995; Eubank, 1988) which provides a way to balance fitting the data closely and avoiding excessive roughness or rapid variation. A penalized least squares estimate is a surface that minimizes the penalized least squares over the class of all surfaces satisfying sufficient regularity conditions. It is used to fit a nonparametric regression model. It computes the thin-plate smoothing splines to approximate smooth functions observed with noise. The spline procedure allows great flexibility in the possible form of the regression surface. It makes no assumptions of a parametric form for the model. The generalized cross validation (GCV) function is used to select the amount of smoothing. In the semi-parametric set up, a parametric part, say, a linear function is added in the model and the penalized least squares estimate is called upon. Let $\boldsymbol{x}_i$ be a d-dimensional covariate vector, $\boldsymbol{z}_i$ be a p-dimensional covariate row vector, and y_i be the observation associated with $(\boldsymbol{x}_i, \boldsymbol{z}_i)$. Assuming that the relation between $\boldsymbol{z}_i$ and y_i is linear but the relation between $\boldsymbol{x}_i$'s and y_i's is unknown, one can fit the data using a semi-parametric model as follows:

$$y_i = f(\boldsymbol{x}_i) + \boldsymbol{z}_i\boldsymbol{\beta} + e_i$$

where f is an unknown scalar-valued function that is assumed to be reasonably smooth, $e_i, i = 1, 2, ..., n$, are independent, zero-mean random errors and $\boldsymbol{\beta}$ is a p-dimensional unknown parametric column vector. This model consists of two parts. The $\boldsymbol{z}_i\boldsymbol{\beta}$ is the parametric part of the model, and the $\boldsymbol{z}_i$'s are the set of regression (covariate) variables. The ordinary least squares method estimates $f(\boldsymbol{x}_i)$ and $\boldsymbol{\beta}$ by minimizing the quantity: $(1/n)\sum_{i=1}^{n}[y_i - f(\boldsymbol{x}_i) - \boldsymbol{z}_i\boldsymbol{\beta}]^2$. The penalized least squares function is defined as

$$S_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(\boldsymbol{x}_i) - \boldsymbol{z}_i\boldsymbol{\beta}]^2 + \lambda J_2(f)$$

where $J_2(f)$ is the penalty on the roughness of f and is defined, in most cases, as the integral of the square of the second derivative of f. In fact, $J_2(f) = \int_0^1\{f''(t)\}^2dt$. The first term measures the goodness of fit and the second term measures the smoothness associated with f. The λ term is the smoothing parameter, which governs the trade-off between smoothness and goodness of fit. When λ is large, it heavily penalizes estimates with large second derivatives. Conversely, a small value of λ puts more emphasis on the goodness of fit. The smoothing parameter λ can be chosen by minimizing the generalized cross validation (GCV) function.

KERNEL FITTING

Kernel is a nonparametric smoother. For a simple regression model with one or two explanatory variables, $y_i = f(x_i) + \boldsymbol{z}_i\boldsymbol{\beta}$, a smoother $f_\lambda(x)$ is a scalar-valued function that summarizes the trend of y as a function of $\boldsymbol{x}$. It can enhance the visual perception of either a y-by-x scatter plot or a rotating plot. The smoothing parameter λ controls the smoothness of the estimate. With one explanatory variable in the model, $f_\lambda(x)$ is called a *scatter plot smoother*. For smoothing spline and kernel, the smoothing parameter λ is derived from a constant c that is independent of the units of x. For a LOESS smoother, the smoothing parameter λ is a positive constant α. Similar to the parametric regression, the R^2 value for an estimate is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}.$$

For a nonparametric smoother with a parameter λ, the fitted values can be written as $\hat{\boldsymbol{y}} = \boldsymbol{H}_\lambda\boldsymbol{y}$, where $\boldsymbol{y}$ is the $n \times 1$ vector of observed responses y_i, $\hat{\boldsymbol{y}}$ is the $n \times 1$ vector of fitted values $\hat{y}_i = \hat{f}_\lambda(x_i)$, and the smoother matrix is an $n \times n$ matrix $\boldsymbol{H}_\lambda$ that depends on the value of λ. The degrees of freedom, or the effective number of parameters, of a smoother can be used to compare different smoothers and to describe the flexibility of the smoother. The degrees of freedom of a smoother is defined as df_λ = trace $(\boldsymbol{H}_\lambda)$. With the degrees of freedom of an estimate df_λ, the mean squared error is given as $MSE(\lambda) = [1/(n - df_\lambda)]\sum_{i=1}^{n}(y_i - \hat{f}_\lambda(x_i))^2$. Cross-validation (CV) estimates the response at each x_i from the smoother that uses only the remaining $n - 1$ observations. The resulting CV mean squared error is given as

$$MSE_{CV}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_{\lambda(i)}(x_i))^2$$

where $\hat{f}_{\lambda(i)}(x_i)$ is the fitted value at x_i computed without the ith observation.

LOESS FITTING

The LOESS procedure implements a nonparametric method for estimating regression and it allows great flexibility because no assumptions about the parametric form of the regression function/surface are needed. LOESS procedure is used for situations in which a suitable parametric form of the regression function is not known. Assume that for $i = 1, 2, ..., n$, the ith measurement y_i of the response y and the corresponding measurement x_i of the vector $\boldsymbol{x}$ of p predictors are related by $y_i = g(x_i) + e_i$, where g is the regression function and e_i is a random error. The idea of local regression is that at a predictor x, the regression function $g(x)$ can be locally approximated by the value of a function in some specified parametric class. Such a local approximation is obtained by fitting a regression surface to the data points within a chosen neighborhood of the point x.

In the LOESS method, weighted least squares is used to fit linear or quadratic functions of the predictors at the centers of neighborhoods. The radius of each neighborhood is chosen so that the neighborhood contains a specified percentage of the data points. The fraction of the data, called the *smoothing parameter*, in each local neighborhood controls the smoothness of the estimated surface. Data points in a given local neighborhood are weighted by a smooth decreasing function of their distance from the center of the neighborhood. In a direct implementation, such fitting is done at each point at which the regression surface is to be estimated.

3. Result and discussion

The five methods (deterministic fitting of VB equation, non-deterministic fitting of VB equation, Spline fitting, Kernel fitting and LOESS fitting) are then employed to obtain the predicted values at each data point. The original observations along with the corresponding predicted values and the values of the R^2 coefficient (Draper and Smith, 1998) with respect to each model are presented in the following table. All the five models are very good in terms of representing this real-life situation. Non-deterministic VB fit is as good as deterministic VB fit. However, it is evident that nonparametric and semi-parametric models produce superior fits. Thus, it is no denying that where precision of fitting is also a concern to reckon with, nonparametric and semi-parametric fits can be employed to model the longitudinal growth dynamics of Puntius sophore (Ham) fish. Graphs showing the fits are also annexed.

Predicted value and R^2 table

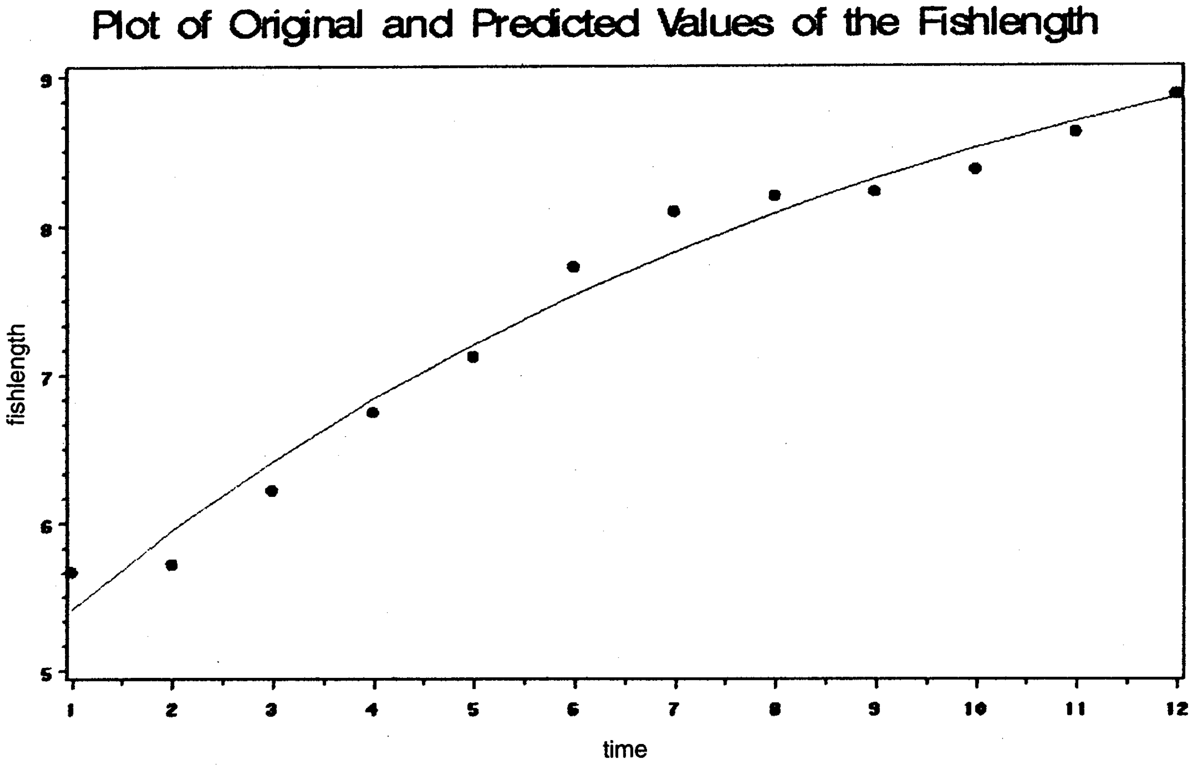
Time	Fishlength (observed)	Predicted Fishlength (Nonlinear VB)	Predicted Fishlength (Spline)	Predicted Fishlength (LOESS)	Predicted Fishlength (Kernel)	Predicted Fishlength Best Fit – Deterministic VB
1	5.675	5.4156	5.6750	5.4816	5.6824	5.577
2	5.725	5.9467	5.7250	5.8978	5.7816	6.026
3	6.225	6.4176	6.2250	6.3163	6.2280	6.438
4	6.75	6.8350	6.7500	6.7341	6.7315	6.815
5	7.125	7.2050	7.1250	7.1686	7.1531	7.161
6	7.725	7.5330	7.7250	7.5647	7.6967	7.479
7	8.1	7.8238	8.1000	7.8757	8.0656	7.770
8	8.2	8.0816	8.2000	8.1141	8.1905	8.037
9	8.225	8.3101	8.2250	8.2904	8.2406	8.283
10	8.375	8.5127	8.3750	8.4659	8.3877	8.507
11	8.625	8.6923	8.6250	8.6376	8.6246	8.714
12	8.875	8.8515	8.8750	8.8097	8.8396	8.903
R^2		0.9770	1.0000	0.9874	0.9994	0.974

Non-linear model fitted: fishlength $= \ell * [1- \exp(-K * (\text{time}-t))]$

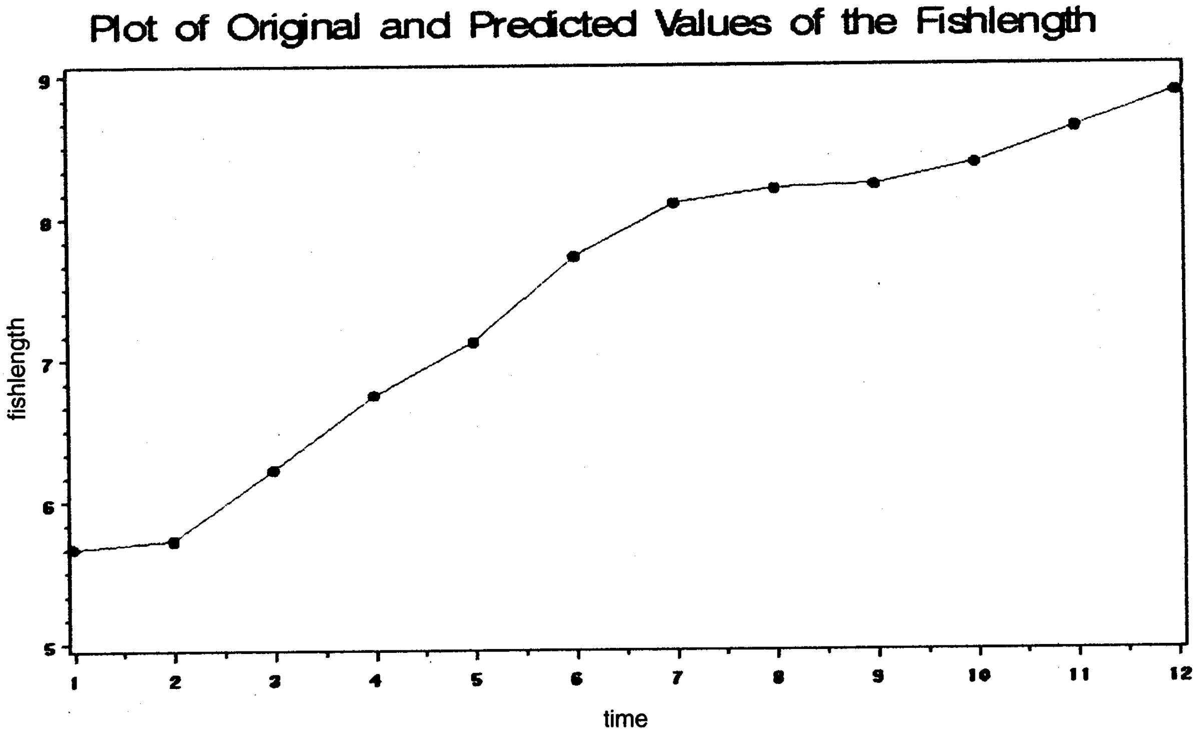
Estimates: $\ell = 10.0950, K = 0.1205, t = -5.3820$

Deterministic VB (Best Fit):
Estimates: $\ell = 11, K = 0.0667, t = -8.5114$

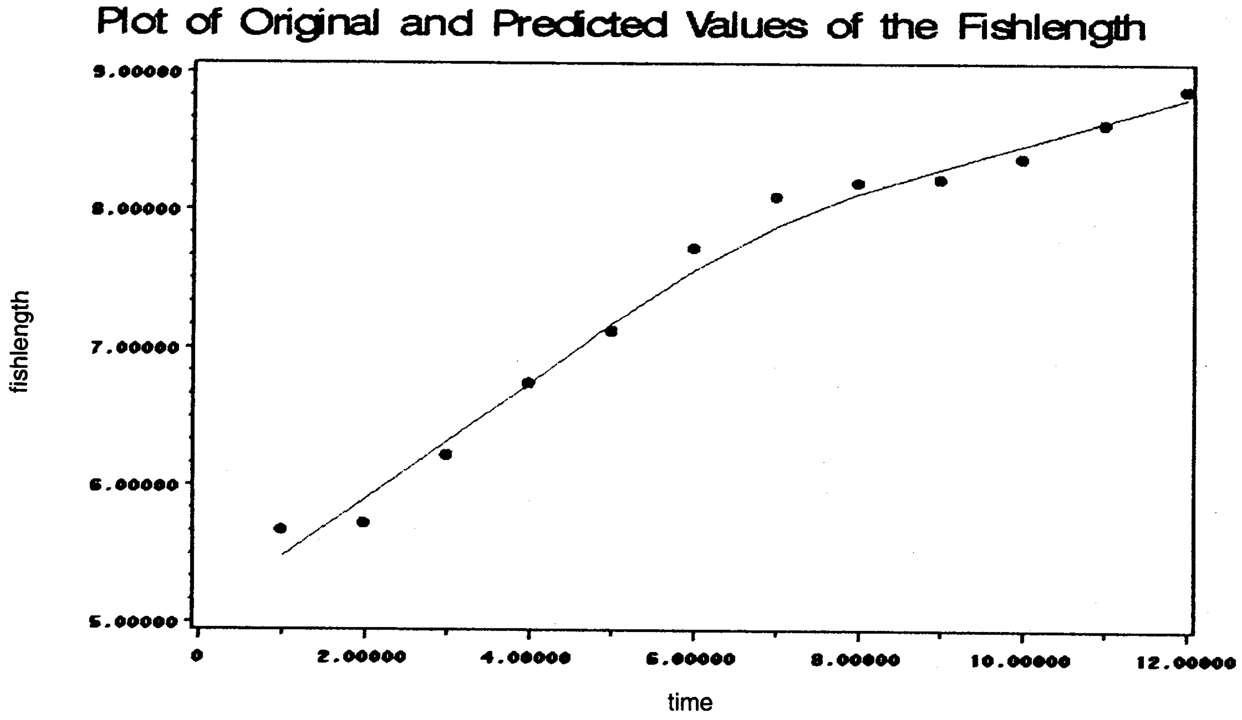
Scatter Plot (non-linear VB fitting)



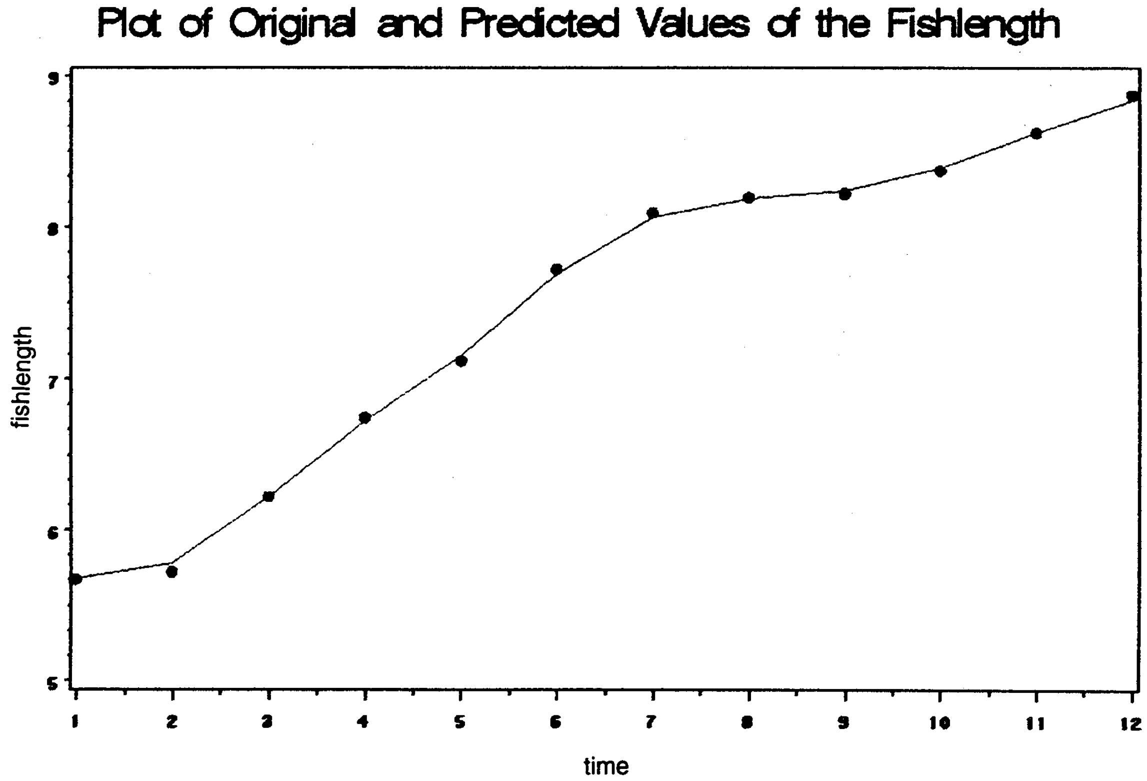
Spline Scatter plot



LOESS Scatter Plot



Kernel Scatter Plot



References

Bagenal, T. (1978). *Methods for Assessment of Fish Production in Fresh Waters.* Third Edition. Blackwell Scientific Publishing, Oxford, London.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis.* Third Edition. Wiley, New York.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression.* Marcel Dekker, New York.

Le Cren, E. D. (1951). The length weight relationship and seasonal cycle in gonal weight and condition in Perch (Perea Fluviatilis). *J. Anim. Ecol.* **20**, 201-219.

Petrakis, G. and Stergion, K. L. (1995). Weight length relationship of 33 fish species in Greek waters. *Fisheries Research* **21**, 465-469.

Simonoff, J. (1995). *Smoothing Methods in Statistics.* Springer, New York.

Thisted, Roland A. (1988). *Elements of Statistical Computing.* Chapman and Hall, New York.