ICES Annual Science Conference
Biology and Behaviour (II) (CC)
CM 1997/CC:08

An attempt to model the length-weight relationship for saithe in Icelandic waters.

by
Ásta Guðmundsdóttir
and
Björn Ævarr Steinarsson
Marine Research Institute
Reykjavík
Iceland

Abstract

In 1993 on-board scales for weighing individual fish were taken into use on Icelandic research vessels. All fish sampled for ageing have been weighted (and length-measured) since that time. These data on length and weight are used to examine length-weight relationships for saithe in Icelandic waters. The paper compares several approaches to modelling these data. The resulting models can be used to validate the data collection process in addition to obtaining biological information.

Introduction.

The main practical use for length-weight relationship at the Marine Research Institute in Iceland (MRI) is for calculating catches in numbers at age for stock assessment purposes and estimating the mean weight at age in landings and from surveys. In 1993 it was decided to weigh all fish sampled for ageing to avoid discrepancies arising from using length-weight relationships. As this data series is relatively short, length-weight relationships are used when comparing longer time series. Also when sampling huge amount of data of this kind more or less electronically, the use of good length-weight relationships for quality control is of great importance.

Until now all length-weight relationships at the MRI have been estimated based on the relationship $W=\alpha L^{\beta}$, with W=weight and L=length, by using a linear regression on a logarithmic scale for estimating α and β. As this function does not fit well over the whole range of lengths for many fish species (Lundbeck 1951) it was considered important to look through the data collected in recent years to see if more accurate relationships could be established.

Material

With the introduction of electronic on-board scales weighing of individual fish at sea became feasible. Since 1993 all fish sampled for ageing at the MRI have been length-measured and weighted by use of on-board scales. The data used for saithe in this study was collected in 1993-1997 in the Icelandic Groundfish Survey (Pálsson et al. 1997) carried out annually in March all around the island. The gear used is a bottom trawl with a 40 mm codend cover. The length is measured to the nearest centimeter and the accuracy of the scales (MAREL M2000/M74 Portable) used is ±- 1 g for fish weighting less than 3kg, ± 2g for fish between 3 and 6kg and ±5g for fish over 6 kg. Gutted weights were used to avoid to variations resulting from varying stomach content, gonad maturity and liver weight. Sampling was random, weighing every third length measured saithe in a haul.

The S-PLUS statistical package is used for programming the models (see Chambers et. al 1992).

Methods.

The relation between length and weight for saithe is shown in Fig. 1. The data is linearized using a logarithmic transformation and a model of the form

(1) $\log W = \alpha + \beta \log L$

is hypothesized (Fig. 2). When using a linear regression on log-transformed data, a correction factor is needed when backtransforming the regression function (Hayes et. al 1995). To avoid this a linear regression in the class of generalized linear models (GLM) is used with the family option. This model requires that the distribution of the data is known. By looking at Fig. 1 it seems that the variation in the data increases with length, possibly suggesting a gamma distribution. Fig. 3, shows a log-log plot of the variance versus the mean weight per length. A linear regression of the log-variance on the log-mean of the weight gives the slope of 2.12 ± 0.08, which is close to a value of 2, which is the case if a gamma density is assumed.

Model (1) was fitted, by use of GLM with the family option set as gamma distribution. This model explains about 99% of the variation in the data (R^2=0.993), but by looking at Fig. 2, there are strong indications that a linear regression is not appropriate. The bulk of the data points below log(length) of about 3.6 are below the regression line and over the regression line for the highest values of log(length). The model can be tested formally by using the decision rule:

$C_1 : \log W = \alpha + \beta \log L$

and

$C_2 : \log W \neq \alpha + \beta \log L$

and the following test statistic (see J. Neter and W. Wasserman 1974)

$$F^* = \frac{MSLF}{MSPE} = \frac{\textit{lack of fit mean square}}{\textit{pure error mean square}}$$

where

$F^* \leq F(1-\alpha; c-2, n-c)$ conclude C_1

and

$F^* > F(1-\alpha; c-2, n-c)$ conclude C_2 .

With $F^*=7.17>1.25=F(0.95;101,3010)$, C_2 is concluded, which means a linear regression is not appropriate.

As the regression function is not linear, the next approach is to modify model (1) with respect to the nature of the regression function. The plot of log(mean(W)) versus log(L) (Fig. 4), indicates that a quadratic term should be included in the model. Hence the following model was fitted:

(2) $\log W = \alpha + \beta \log L + \gamma (\log L)^2$

The results of the model support that all the parameters α, β and γ are required in the model. As before the aptness of the model can be tested, the alternatives now being:

$C_1 : \log W = \alpha + \beta \log L + \gamma (\log L)^2$

and

$C_2 : \log W \neq \alpha + \beta \log L + \gamma (\log L)^2$

The same test statistic is used as before. Since $F^*=3.98 > 1.25=F(0.95,100,3010)$, C_2 is concluded, that means a quadratic regression is not appropriate. The log(mean(W)) versus log(L) plot (Fig. 4) indicates that a polynom of higher degree might be more appropriate, so a qubic term was added to model (2):

(3) $\log W = \alpha + \beta \log L + \gamma (\log L)^2 + \delta (\log L)^3$

This time too the result of the model supports, that all the parameters are required in the model. But using the same decision rule and the same test statistic as before $F^*=2.36 > 1.25=F(0.95,99,3010)$, which means a qubic polynomial is also not appropriate. When looking at Fig. 6, it it obvious that the qubic polynomial does not fit the data well for the lower part of the lengths.

As models with linear, quadratic and qubic terms are rejected due to the test $F^* > F$, there is obviously a need for a more flexible model, a model of the type

(4) $\log(W) = f(\log(L))$

Such a function can be smoothing splines, but such a model belongs to the class of the generalized additive models (GAM). By using the equivalent degrees of freedom as the smoothing parameters some fits were performed and the models compared using the anova with F test (Table 1 and Fig. 7 to 9). F^* and the corresponding F (at $\alpha=0.05$) were computed as well (Table 1). From Table 1 it can be seen that the residual deviance has decreased much relative to the error variance (*residual deviance/residual degree of freedom*) for 1 df, from the model with

log(length) to the model with s(log(length),df=3). In GAM models, the number of parameters in the model are equivalent to the number of observation minus the residual degree of freedom. Obviously there is the demand for having as few parameters in the model as possible. For all the models $F^* > F$ which means that none of the models are appropriate.

The practical use of a nonparametric model of this kind would be using a table to search in for mean weight at age for given length, instead of using a simple equation with given parameters. As another choise to keep on with parametric form is to construct the model

(5) $\log(W) = \alpha + \beta \log(L)$, *separately* for *L in each of equal intervals.*

Some fits were performed and F^* and F (at α=0.05) computed for each interval (Table 2 and Fig. 10 to 12). There is the need for 4 intervals if $F^* \leq F$ is to hold on every interval and this implies the use of a total of 8 parameters. This model lacks continuity, but on the other hand it is a parametric one and easy to use. There are more ways to divide the lengths into intervals, but here it was decided to look only at equally great intervals.

Discussion

In this paper it was decided use all data awailable, not truncate the lower or the upper part of the lengths, inspite of fewer datapoints then for the midpart. It was also decided to use a lack of fit test, although it requires that the observations are normally distributed. A natural next step would be to construct a lack of fit test for gamma distributed observations.

The models viewed in this paper, taking into account both the number of parameters and the lowest F^* number, suggest that the nonparametric generalized additive (GAM) model s(log(length),df=3) gives the best fit. But conventional piecewise linear models, based on splitting up the length intervals, might be the best choice for practical use, especially for quality control.

References

Hays, D.B., Brodziak, J.K.T., and O'Gorman, J.B 1995. Efficiency and bias of estimators and sampling designs for determining length-weight relationships of fish. Can. J. Fish. Aquat. Sci. **52** : 84-92

Lundbeck, J.: Biologisch-statistische Untersuchungen uber die deutsche Hochseefischerei. III. Das Körpergewicht und das Langen-Gewichts-Verhaltnis bei den Nutzfischen. Ber. Der dtsch. Meeresf. B XII H 3, 1951.

Neter, J., and Wasserman, W. 1974. Applied linear statistical models. Regression, analysis of variance and experimental design. Richard D. Irwin, Inc., Homewood, Illinois.

Pálsson, O.K., Steinarsson, B.Æ., Jónsson, E., Guðmundsson, G., Jónsson, G., Stefánsson, G., Björnsson, H., and Schopka, S.A. Icelandic Groundfish Survey. ICES C.M. 1997/Y:??.

Statistical models in S. Chambers, J.M., and Hastie, T.J. 1992. Wadsworth & Broosk/Cole Advanced Books & Software. Pacific Grove, California.

Table 1

Analysis of Deviance Table											
Response: weight											
	Terms	Resid. Df	Resid. Dev	Test	Df	Deviance	F Value	Pr(F)	Number of parameters	F^*	F(.95,103-no.param.,3010)
1	log(length)	3111.000	19.86628						2	7.15	1.25
2	s(log(length),df=2)	3109.999	17.82163	1 vs. 2	1.001074	2.044645	374.2832	0.000000e+00	3	3.31	1.25
3	s(log(length),df=3)	3108.999	17.12785	2 vs. 3	1.000075	0.693783	127.1277	0.000000e+00	4	2.04	1.25
4	s(log(length),df=4)	3107.999	16.87737	3 vs. 4	0.999764	0.250479	45.9117	1.000000e-11	5	1.60	1.25
5	s(log(length),df=5)	3106.999	16.77795	4 vs. 5	0.999655	0.099416	18.2245	2.027245e-05	6	1.44	1.25
6	s(log(length),df=6)	3105.997	16.72993	5 vs. 6	1.002306	0.048020	8.8533	0.00292767	7	1.37	1.25

Table 2

number of intervals	F^*	F	F^*	F	F^*	F	F^*	F
2	7.32	1.34	1.80	1.39				
3	2.92	1.43	1.52	1.43	1.37	1.59		
4	0.69	1.51	1.28	1.50	1.09	1.52	1.57	2.68

Fig. 1, Saithe, 1993–1997

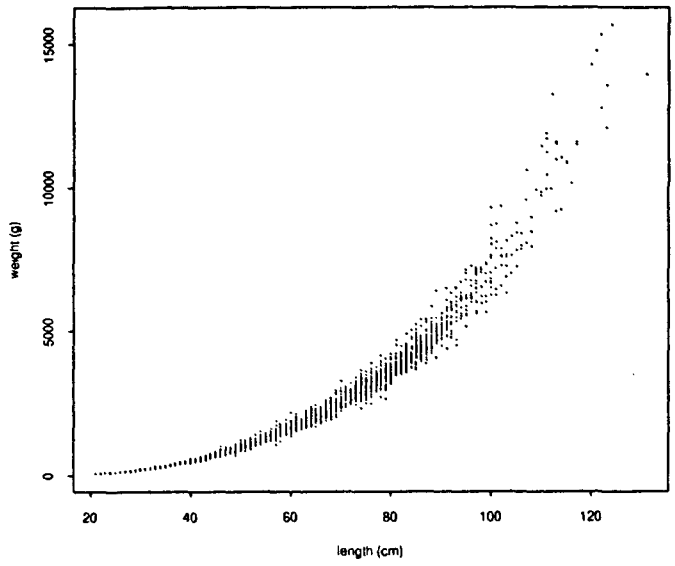

Fig. 2, a log–log transformation of the data with the regression line

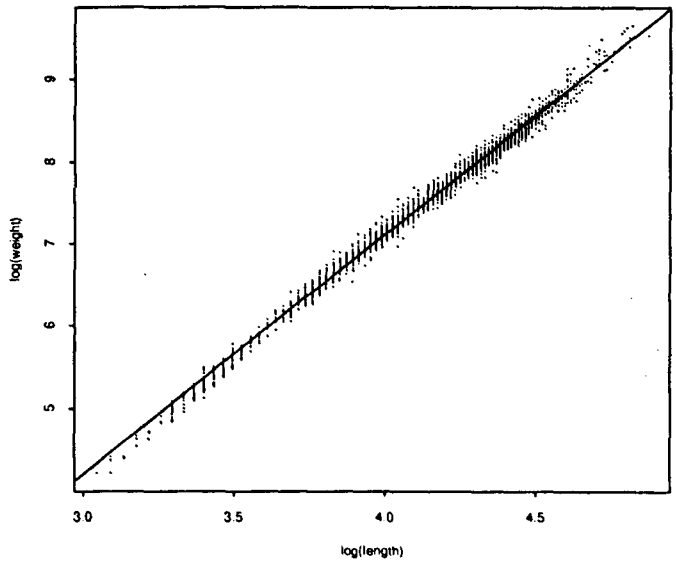

Fig. 3, log(variance) versus log(mean weight) per length

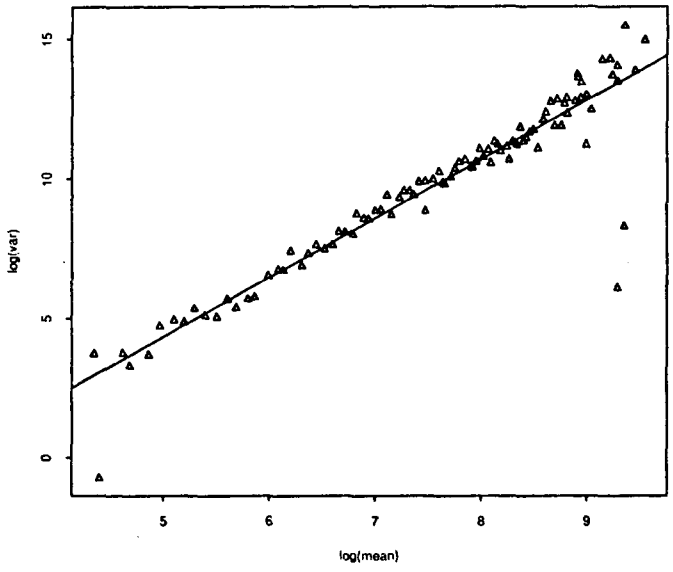

Fig. 4, log(mean weight) versus log(length) and a regression line

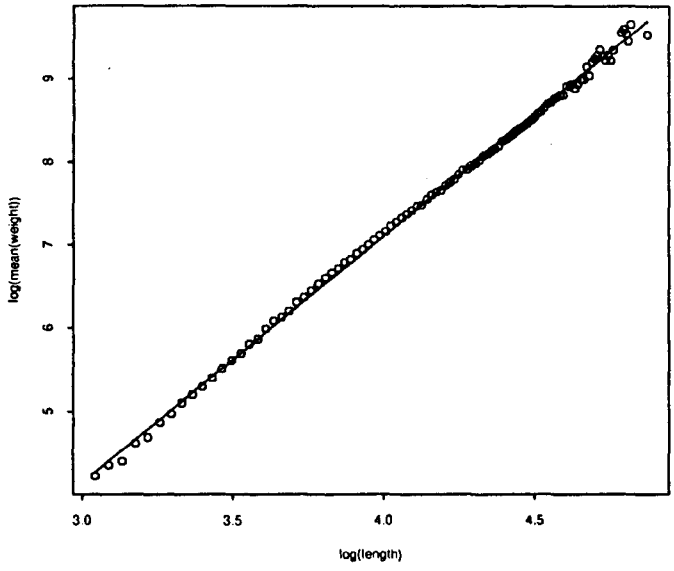

Fig. 5, the regression function is a second order polynomial

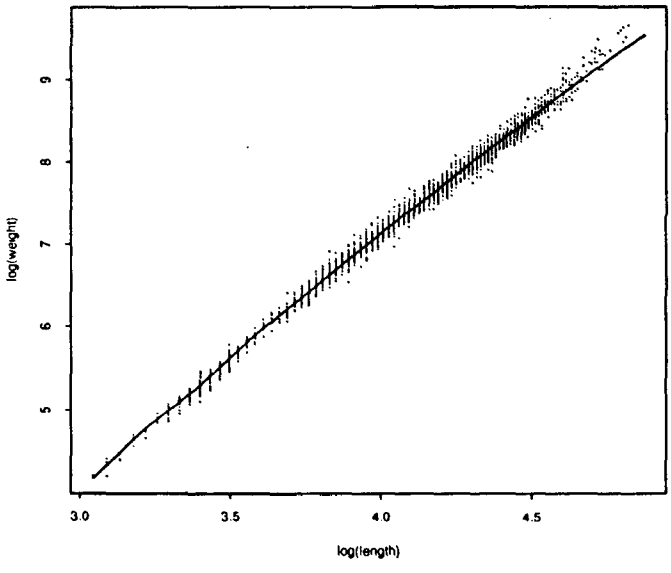

Fig. 6, the regression function is a third order polynomial

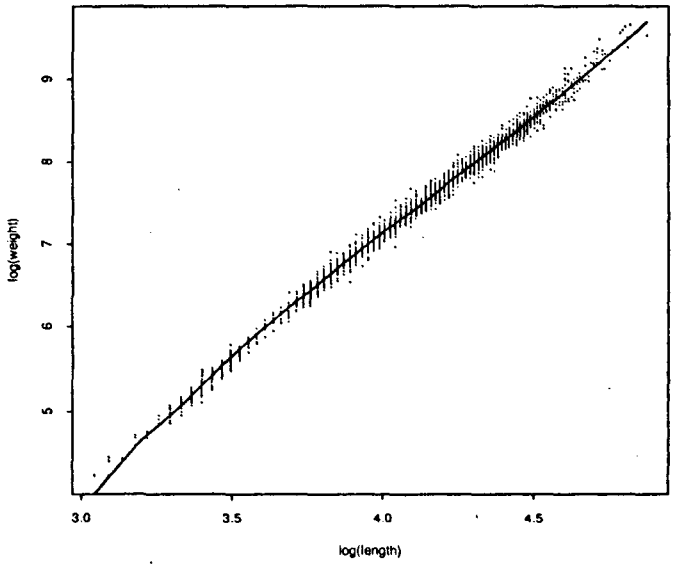

Fig. 7, smoothing splines with df=2

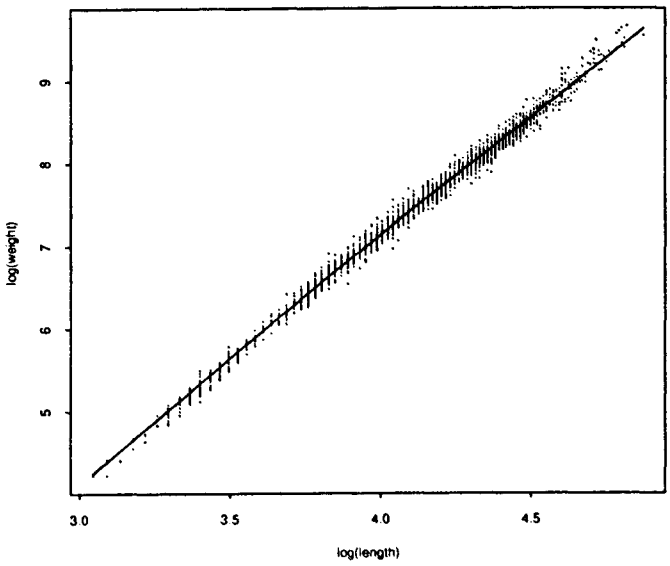

Fig. 8, smoothing splines with df=3

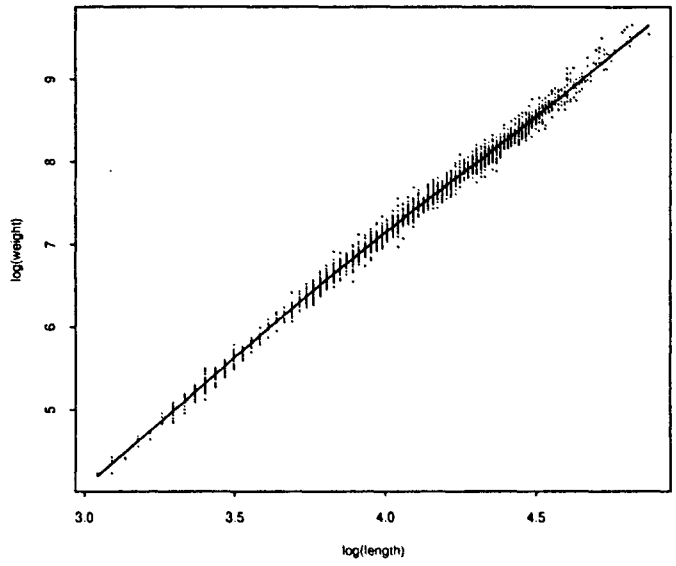

Fig. 9, smoothing splines with df=4

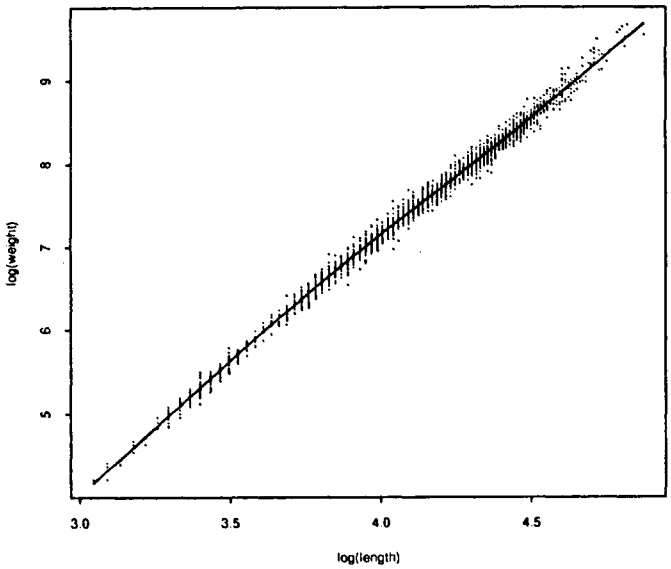

Fig. 10, the length is divided into 2 intervals

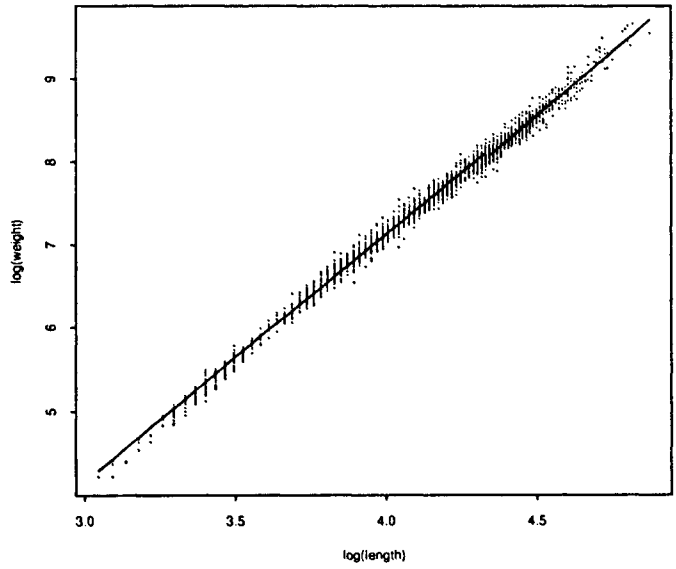

Fig. 11, the length is divided into 3 intervals

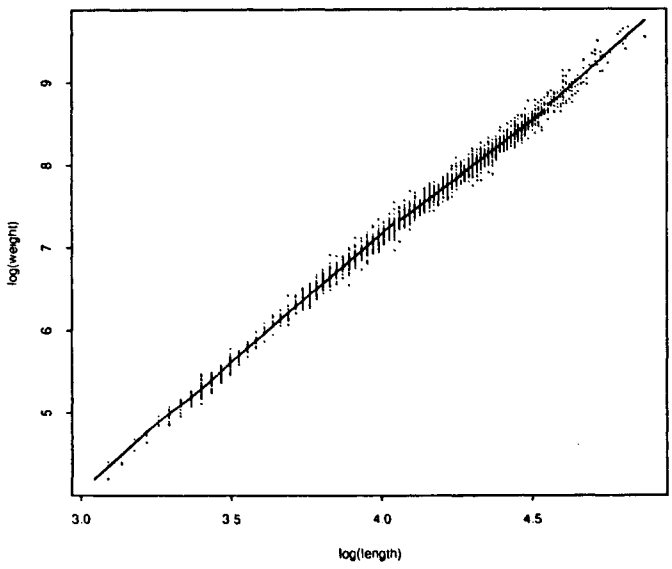

Fig. 12, the length is divided into 4 intervals