

Comparison of theoretical derivations, simple linear regressions, multiple linear regression and principal components for analysis of fish mortality, growth and environmental temperature data

A. L. Jensen*

School of Natural Resources and Environment, University of Michigan, Ann Arbor, Michigan 48109-1115, U.S.A.

SUMMARY

Natural mortality of fish populations is difficult to estimate, and parameters for growth and environmental temperature, which are easier to estimate, have been applied to predict fish natural mortality using multiple linear regression. There are theoretical relations among all of the variables applied in the multiple linear regression, and there is high multicollinearity; the results of the multiple regression differ considerably from the theoretical relations among the variables. Simple linear regression results agree with the theoretical results but they are not as precise for prediction of mortality as multiple linear regression. A principal components analysis correctly identifies the important variables and the relations among variables but it is more complex than multiple linear regression and yet is not any more precise for predictions. A plot of the first two principal components separated the data into two groups: one was temperate water species and one was warmer water species. The analysis confirms the limitations and advantages of different data analysis methods. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: fish mortality; principal components analysis; linear regression

1. INTRODUCTION

Separation of natural mortality from fishing mortality is essential for assessment of fisheries, but it is difficult and this has led to prediction of natural mortality from other more easily estimated parameters such as those for growth. Pauly (1980) applied multiple linear regression to data for mean environmental temperature T (°C), the coefficient for growth in size K (yr^{-1}), asymptotic length L_{inf} (cm) and asymptotic weight W_{inf} (g) for 175 different fish populations to determine a statistical relation for prediction of mortality. The first observation of Pauly was in error (Gulland, 1987); analysis of Pauly's data without the first observation gives the equation

$$\log M = 0.66 \log K + 0.45 \log T \quad (1)$$

*Correspondence to: A. L. Jensen, School of Natural Resources and Environment, The University of Michigan, Ann Arbor, MI 48109-1115, USA.

Table 1. Multiple linear regression with $\log M$ the dependent variable and independent variables $\log L_{\text{inf}}$, $\log W_{\text{inf}}$, $\log K$ and $\log T$; $P < 0.0001$, $R^2 = 0.71$, $SE = 0.57$

Variable	Coefficient	SE	<i>t</i> -value	<i>P</i> -value
constant	0.379	0.527	0.72	0.47
$\log L_{\text{inf}}$	-0.520	0.280	-1.86	0.06
$\log W_{\text{inf}}$	0.081	0.090	0.91	0.36
$\log K$	0.658	0.072	9.11	<0.001
$\log T$	0.453	0.083	5.44	<0.001

which differs only slightly from the equation obtained by Pauly. In Pauly's multiple linear regression only the coefficients for $\log K$ and $\log T$ were statistically significant (Table 1), but ecological theory indicates that mortality should be related to L_{inf} , W_{inf} , K and T . All of the independent variables in the regression were correlated and the resulting predictive regression model, although useful for prediction, does not correctly identify either the variables related to mortality or the relations among the variables themselves.

Multicollinearity among independent variables does not limit the use of multiple linear regression for prediction, but multicollinearity does limit the use of multiple linear regression for identifying how variables are related and for identification of the form of the relations among independent variables (e.g. Netter *et al.*, 1996). To better understand the relations among the variables used by Pauly, and the results of different statistical methods, I compared the results of theoretical derivations, simple linear regressions, multiple linear regression and principal components analysis.

2. THEORETICAL RELATIONS

Growth and mortality of fish have been studied extensively, and the mathematical models that have been developed for growth and mortality describe observed relations well (e.g. Beverton and Holt, 1957; Ricker, 1975). Mortality of fish is described by the exponential model (e.g. Beverton and Holt, 1957; Ricker, 1975)

$$N(x) = R e^{-Mx} \quad (2)$$

where $N(x)$ is the number of individuals of age x , R is the number of recruits and M is the instantaneous natural mortality rate per individual. Derivations based on both metabolism of fish and empirical observations indicate that growth in length of fish can be described with the von Bertalanffy equation (Beverton and Holt, 1957)

$$L(x) = L_{\text{inf}} (1 - e^{-Kx}) \quad (3)$$

where $L(x)$ is length at age x , L_{inf} is asymptotic length and K is a growth parameter. Mean weight and length have been empirically related by the equation (Ricker, 1975)

$$W = aL^3 \quad (4)$$

where a is a parameter.

The models for growth and mortality have been applied with ecological and evolution theory to develop relations among the variables M , L_{inf} , W_{inf} and K . Maximization of reproduction with respect to age at maturity leads to the relation between mortality and the growth rate (Charnov, 1993; Jensen, 1996),

$$M = 1.50 K \quad (5)$$

which has been termed the second Beverton and Holt invariant (Charnov, 1993). Based on theoretical analyses of predation in the pelagic zone Peterson and Wroblewski (1984) obtained the following theoretical relation between dry weight (g) and mortality (yr^{-1}):

$$M = 1.92 W^{-0.25} \quad (6)$$

A theoretical relation between length (cm) and growth coefficient (yr^{-1}) obtained from metabolic relations is (Charnov, 1993; Jensen, 1996)

$$L_{\text{inf}} = c K^{-66} \quad (7)$$

where c is a parameter. Temperature is related to both growth and mortality rates (e.g. Beverton and Holt, 1957). All of the variables that Pauly (1980) applied in the multiple linear regression for prediction of M were related.

3. SIMPLE LINEAR REGRESSION

Simple linear regression was applied to examine observed relations among pairs of variables, and to compare these with the theoretical relations. Again, the first observation reported by Pauly (1980) was not included. Plots of natural mortality against each of the other variables indicated that the relation between M and K was linear, but the other relations were not linear (Figure 1). The regression of M on K gave

$$M = 0.21 + 1.47 K \quad (8)$$

with a coefficient of determination of 0.75. The slope is close to the theoretical value of 1.50. The regression of $\log M$ on $\log W$ gave

$$M = 1.51 W_{\text{inf}}^{-0.23} \quad (9)$$

which is similar to the theoretical relation given by Peterson and Wroblewski (1984); the regression was significant but the coefficient of determination was only 0.33. The separate regressions of $\log M$ against $\log L_{\text{inf}}$ and $\log T$ were also significant (Table 2), but the growth coefficient accounted for nearly twice as much of the variation in mortality as the other variables. Asymptotic length and asymptotic weight were the next best predictors and temperature was the poorest.

The regression of length on the growth coefficient

$$L_{\text{inf}} = 2.98 K^{-0.63} \quad (10)$$

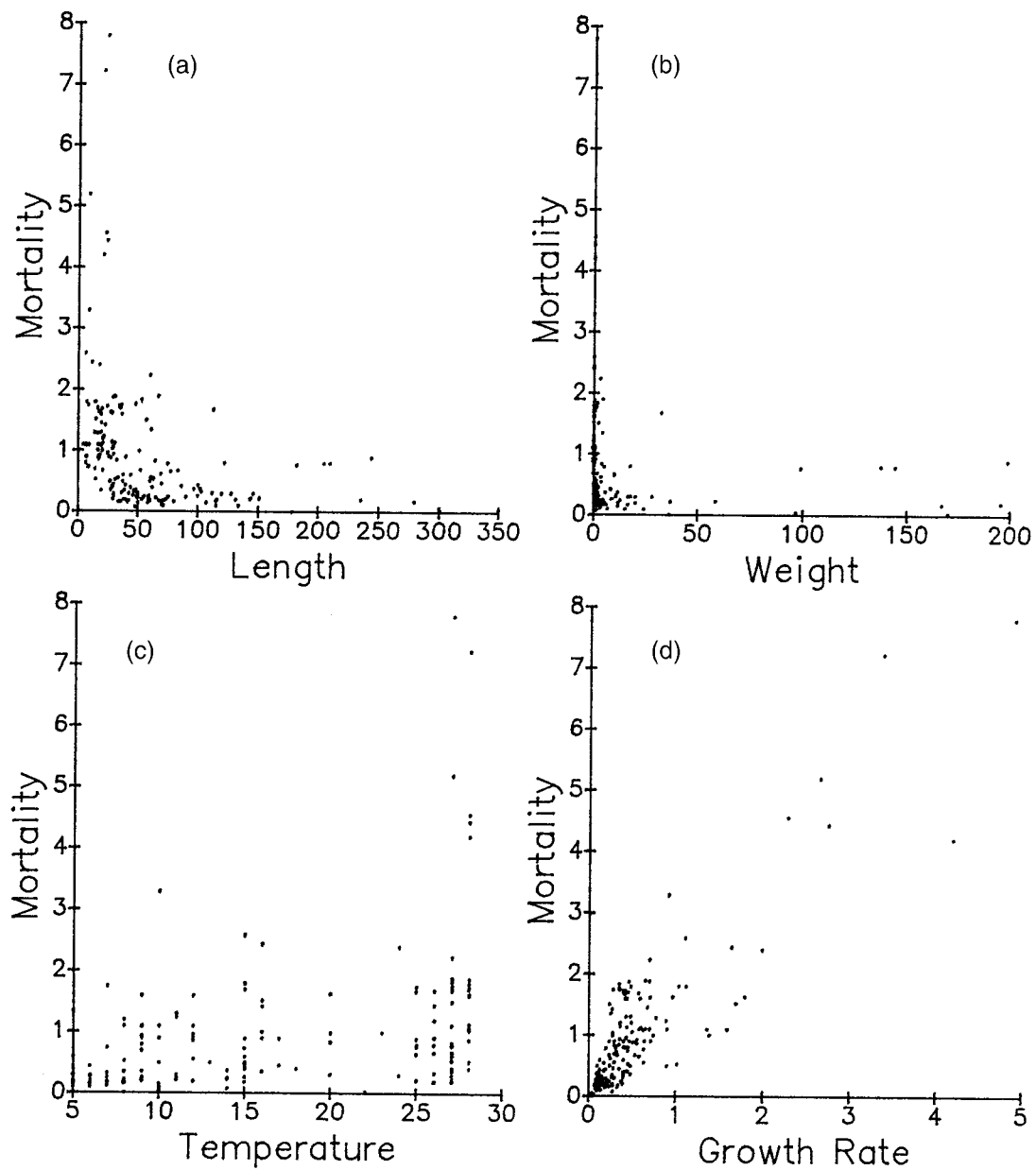


Figure 1. (a) Relation between natural mortality coefficient (yr^{-1}) and asymptotic length (cm). (b) Relation between natural mortality coefficient (yr^{-1}) and asymptotic weight (100s of grams). (c) Relation between mortality (yr^{-1}) and temperature ($^{\circ}\text{C}$). (d) Relation between natural mortality coefficient and von Bertalanffy growth coefficient

Table 2. Summary of separate simple linear regressions of $\log M$ on $\log L_{\text{inf}}$, $\log W_{\text{inf}}$, $\log K$ and $\log T$. (All were statistically significant with $\alpha = 0.05$)

Variable	Intercept	Slope	SE	R^2
$\log L_{\text{inf}}$	2.061	-0.710	0.836	0.35
$\log W_{\text{inf}}$	0.862	-0.219	0.857	0.32
$\log K$	0.498	0.940	0.620	0.66
$\log T$	-2.688	0.814	0.922	0.21

was also significant with $R^2 = 0.41$ and $P < 0.0001$, and the regression of weight on length

$$W_{\text{inf}} = 0.0092 L_{\text{inf}}^{3.027} \quad (11)$$

was significant with $R^2 = 0.97$ and $P < 0.0001$. Taken individually, the relations between pairs of variables were all significant, and where a theory existed the fitted relations were similar to those predicted by theory.

Some of the difference between multiple linear regression and simple linear regressions can be accounted for with a path analysis (Wonnacott and Wonnacott, 1990). In multiple linear regression, temperature and the growth coefficient were significant (Table 1). Temperature has both a direct effect on the mortality rate and an indirect effect on the mortality rate through growth, i.e. when growth is faster, fish become larger more quickly and mortality, which is size-dependent, is lower. The direct effect of temperature on mortality is the coefficient for temperature in multiple linear regression equation, 0.45. The regression of growth on temperature is

$$\log K = -2.511 + 0.53 \log T \quad (12)$$

and the indirect effect of temperature on mortality operating through growth is the product of the simple linear regression coefficient of $\log K$ on $\log T$ and the multiple linear coefficient for $\log K$ on $\log M$ or $(0.533)(0.66) = 0.35$. The sum of the indirect and direct effects is 0.81, which is the coefficient in the simple linear regression of $\log M$ on $\log T$ (Table 2).

4. PRINCIPAL COMPONENTS ANALYSIS

The results of the simple linear regressions and multiple linear regression were different because of multicollinearity, and although the path analysis accounts for the terms in the multiple linear regression it does not account for the difference between theory and the multiple linear regression results. One method suggested for analysis of data with high multicollinearity is principal components analysis, and this was applied to determine if it would untangle the relations among the variables and indicate which variables were important in determination of mortality. The principal components Z_1 , Z_2 , Z_3 and Z_4 (Table 3) were calculated using the correlation matrix of $\log K$, $\log L_{\text{inf}}$, $\log W_{\text{inf}}$ and $\log T$. The first two principal components account for nearly all of the variation (Table 4). The first component is highly correlated with $\log L_{\text{inf}}$, $\log W_{\text{inf}}$ and $\log K$, and the correlations of $\log W_{\text{inf}}$ and $\log L_{\text{inf}}$ are of opposite sign from $\log K$ (Table 4). In nature faster growth occurs with a smaller asymptotic size (e.g. Ricker, 1975), and the principal components correctly capture this relation. The first component appears related to growth (Table 4), and the second component is highly correlated

Table 3. Coefficients of equations for principal components Z_1 , Z_2 , Z_3 and Z_4 as functions of $\log L_{\text{inf}}$, $\log W_{\text{inf}}$, $\log K$ and $\log T$

Variable	Z_1	Z_2	Z_3	Z_4
$\log L_{\text{inf}}$	-0.376	0.193	0.544	5.624
$\log W_{\text{inf}}$	-0.373	0.218	0.558	-5.563
$\log K$	0.328	0.289	1.420	0.096
$\log T$	0.070	0.848	-0.752	0.121

Table 4. Correlations of principal components Z_1 , Z_2 , Z_3 and Z_4 with $\log L_{\text{inf}}$, $\log W_{\text{inf}}$, $\log K$ and $\log T$

Variable	Z_1	Z_2	Z_3	Z_4
$\log L_{\text{inf}}$	-0.96	0.22	0.17	0.09
$\log W_{\text{inf}}$	-0.95	0.25	0.18	-0.09
$\log K$	0.83	0.33	0.45	0.00
$\log T$	0.18	0.96	-0.24	0.00
% variance	63.59	28.17	7.84	0.40

with $\log T$ and appears related to environmental temperature (Table 4). I could not identify the third or the fourth component but these two components together account for little of the variation.

A principal components regression was applied to regress $\log M$ on all four of the principal components (Table 3). In a multiple linear regression with correlated independent variables interpretation of the parameters is uncertain, and in a principal components regression interpretation of the transformed independent variables (principal components) is often uncertain (e.g. Harris, 1985; Afifi and Azen, 1972; Hadi and Ling, 1998), but for Pauly's (1980) data there is little uncertainty concerning the first two components, which together account for more than 90 per cent of the variation in the data. Interpretation of the principal component regression coefficients is straightforward because the components are independent, unit free, and each of the regression coefficients has the same standard deviation. The first three components were statistically significant (Table 5). The first component, which was related to growth, was nearly twice as important as the second component, which was related to environmental temperature. The principal components analysis correctly identifies the importance of all of the variables in determination of mortality, but it is not more precise than the multiple linear regression for prediction, and it is more complex. Although the principal components analysis identified which variables were important it did not reveal the true underlying mathematical relations among variables that were obtained with ecological and evolutionary theory and confirmed with simple linear regression.

Table 5. Results of multiple linear regression of mortality $\log M$ on the principal components

Component	Coefficient	<i>t</i> -value	<i>P</i> -value
constant	-0.545	-12.713	< 0.001
Z_1	0.759	17.653	< 0.001
Z_2	0.400	9.317	< 0.001
Z_3	0.158	3.670	< 0.001
Z_4	-0.059	-1.365	0.174

SE = 0.57 for all variables. $R^2 = 0.71$.

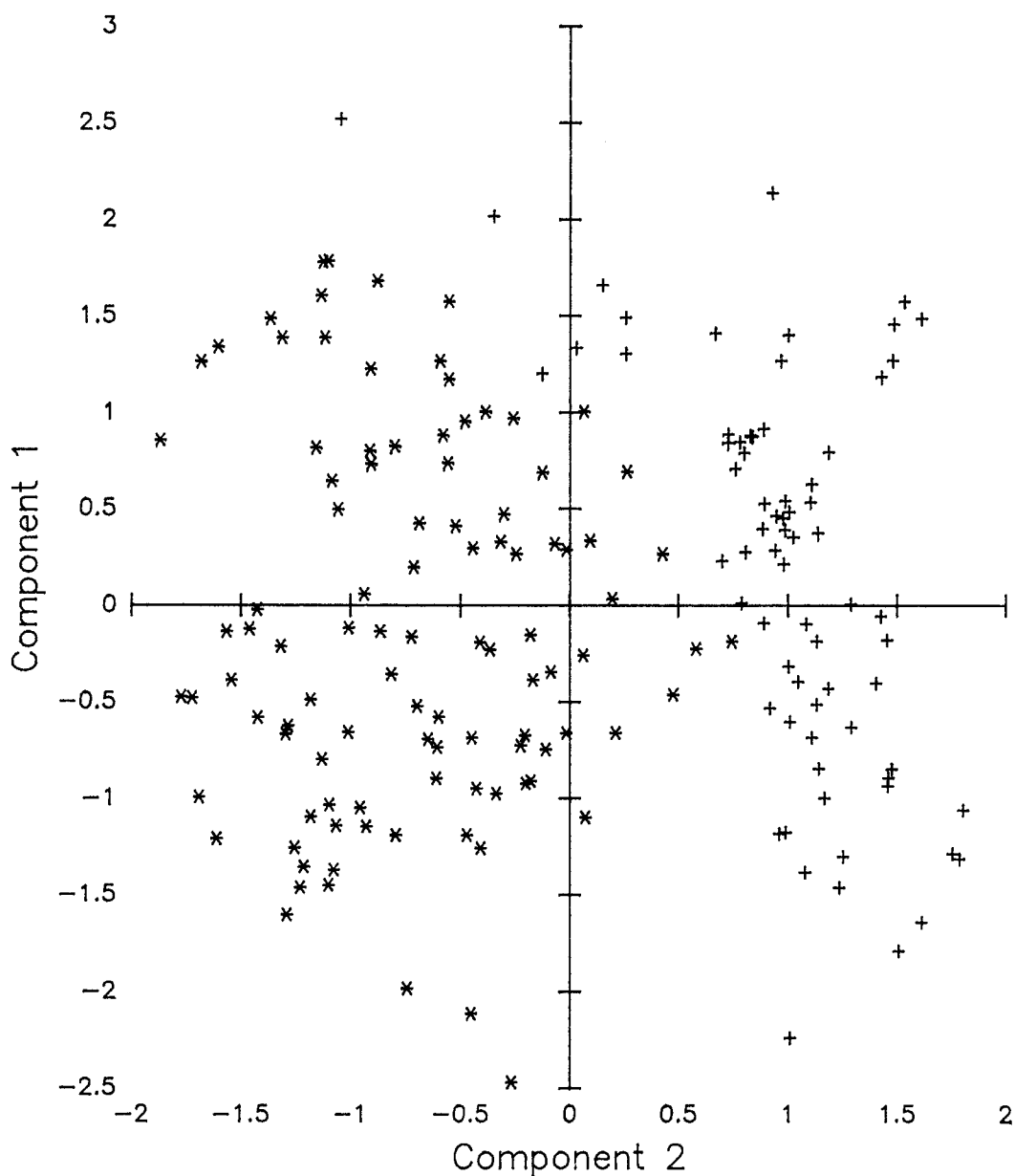


Figure 2. Values for first principal component (size) and values for the second principal component (growth) for Pauly's (1980) data. The asterisks represent temperate species

The principal components provided an interesting insight when they were applied to describe the structure of the data; a plot of the first two components showed a separation of the data into two groups (Figure 2). Individual data points were identified, and it was found that the group on the left in Figure 2 (group A) was populations from temperate waters and the group on the right (group B) was

populations from warmer waters. To confirm the existence of two groups a discriminant analysis was done using the component scores as discriminators. The discriminant functions for groups A and B were:

$$Y_A = -1.68 - 0.48 Z_1 - 2.70 Z_2 - 0.02 Z_3 + 0.23 Z_4 \quad (13)$$

$$Y_B = -2.69 + 0.69 Z_1 + 3.83 Z_2 + 0.03 Z_3 - 0.33 Z_4 \quad (14)$$

where Y_A and Y_B were the discriminant scores for groups A and B, and Z_1, Z_2, Z_3 and Z_4 were the component values which served as discriminators. Assignment of observations to groups A and B using Equations (5) and (6) gave a classification error rate of 6.32 per cent, which measures the overlap of the two populations and indicates that the group to which a population belongs can be predicted with some certainty from the principal components.

REFERENCES

- Afifi AA, Azen SP. 1972. *Statistical analysis: a computer oriented approach*. Academic Press: New York.
- Beverton RJH, Holt SJ. 1957. *On the Dynamics of Exploited Fish Populations*. United Kingdom Ministry of Agriculture and Fisheries, Fisheries Investigations (Series 2). **19**: 1–533.
- Charnov EL. *Life history invariants*. 1993. Oxford University Press: New York.
- Gulland JA. Natural mortality and size. 1987. *Marine Ecology – Progress Series* **39**: 197–199.
- Hadi AS, Ling RF. 1998. Some cautionary notes on the use of principal components regression. *The American Statistician* **52**: 15–19.
- Harris RJ. 1985. *A primer of multivariate statistics*. Academic Press: New York.
- Jensen AL. 1996. Beverton and Holt life history invariants result from optimal trade-off of reproduction and survival. *Canadian Journal of Fisheries and Aquatic Science* **53**: 820–822.
- Netter J, Kutner MH, Nachtsheim CJ, Wasserman W. 1996. *Applied linear statistical models*. Irwin: Chicago.
- Pauly D. On the interrelationships between natural mortality, growth parameters, and mean environmental temperature in 175 fish stocks. 1980. *Journal Consiel International Exploration Mer*. **39**: 175–192.
- Peterson I, Wroblewski JS. 1984. Mortality rates of fishes in the pelagic ecosystem. *Canadian Journal of Fisheries and Aquatic Science* **41**: 1117–1120.
- Ricker WE. 1975. *Computation and interpretation of biological statistics of fish populations. Bulletin 191*, Department of the Environment. Ottawa, Canada.
- Wonnacott TH, Wonnacott RJ. 1990. *Introductory Statistics* (5th edn). John Wiley: New York.