MEDWELL PUBLISHING
Journal of Animal and Veterinary Advances 8 (1): 47-50, 2009
ISSN: 1680-5593

Use of Factor Analysis Scores in Multiple Regression Model for Estimation of Body Weight from Some Body Measurements in Lizardfish

[1]Levent Sangun, [2]Soner Cankaya, [3]G. Tamer Kayaalp and [1]Mustafa Akar
[1]Department of Basic Science, Faculty of Fisheries, University of Cukurova, 01330, Adana, Turkey
[2]Department of Animal Science, Faculty of Agriculture, University of Ondokuz Mayis, 55139, Samsun, Turkey
[3]Department of Animal Science, Faculty of Agriculture, University of Cukurova, 01330, Adana, Turkey

Abstract: The aim of the study is to find out, the utility function of factor analysis scores in multiple linear regression model that were used to estimate body weight with respect to some body measurements (total length, standard length, fork length, head length, body depth, body circuit, body height) measured from Lizardfish in Iskenderun Bay. The results of the factor analysis showed that 3 factor with eigenvalues greater than 1 can be selected as explanatory variables and used to estimate body weight of Lizardfish in multiple linear regression model. The factors accounted for 98.4% of total variation in the body weight.

Key words: Lizardfish, multiple linear regression and factor analysis scores

INTRODUCTION

Information on some body measurement is essential to estimate the body measurement of fish. Multiple regression analysis has been used to interpret the complex relationships among the body weight and the some body measurement (total length, standard length, fork length, head length, body depth, body circuit, body height and etc.) of the fish (Cankaya *et al*., 2006; Akar *et al*., 2001). But, although this method is helpful for estimating the body weight of the fish, there are several reasons why the researchers are not often satisfied with results. One of them, is its poor performance when the multicollinearity is present among variables. This evidence is to present its biological interpretation may be misleading.

The specific goals of factor analysis are to provide reduce a large number of observed variables to smaller number of factors and to provide a regression equation for an underlying process by using observed variables (Tabachnick and Fidell, 2001; Keskin *et al*., 2007). Factor scores can be derived such that they are nearly uncorrelated or orthogonal. Thus, the problem for multicollinearity among the variables, which are used to estimate the body weight of the fish can be solved by using the coefficients.

The aim of the study is to find out, the utility function of factor analysis scores in multiple linear regression model that were used to estimate body weight with respect to some body measurements (total length, standard length, fork length, head length, body depth, body circuit, body height) measured from Lizardfish in Iskenderun Bay.

MATERIALS AND METHODS

Data were collected beetwen 2003-2005 from the north-eastern Mediterranean coast of Turkey. Species were caught by trawl ranging from 20-100 m. The Lizardfish were weighted with a digital balance to an accuracy of 0.01 g and measured with a precision of 0.01 cm for their total length, standard length, fork length, head length, body depth, body circuit and body height.

Linear regression analysis consists of a collection of techniques used to explore relationships between variables. The aim of the multiple regression, is to estimate $\beta = (\beta_0 + \beta_1 + \ldots + \beta_p)^t$ from the data (X_{i1}, X_{i2}, ……, $X_{ip;}$ Y_i) (Cankaya *et al*., 2006). The general expression of multiple linear regression model formed for the measurements (one dependent and p independent variables) is given in Eq. 1.

Corresponding Author: Soner Cankaya, Department of Animal Science, Faculty of Agriculture, University of Ondokuz Mayis, 55139, Samsun, Turkey

$$y_i = b_0 X_{i1}^{b_1} X_{i2}^{b_2} X_{i3}^{b_3} \ldots X_{ip}^{b_p} e_i; \quad i = 1,2, \ldots, n \tag{1}$$

where:

b_i's = Unknown parameters
e = Usually assumed to be normally distributed with mean zero and variance σ^2
Y = The dependent variable or response
$X_{i1}, X_{i2}, \ldots, X_{ip}$ = Independent variables or the predictors

When the dependent (Y) values are plotted against the independent (X) values, the curve cannot be represented by a straight line every time, that is, the relationship may be curvilinear. In order to get a linear curve, we transform X and Y values into a logarithmic value. If we take the logarithm of the Eq. 1, it can be defined as:

$$\ln y_i = \ln b_0 + b_1 \ln X_{i1} + b_2 \ln X_{i2} + \ldots + b_7 \ln X_{ip} + \ln e_i \tag{2}$$

or

$$Y = a + b_1 z_{i1} + b_2 z_{i2} + b_3 z_{3i} + \ldots + b_p z_{ip} + \delta_i \tag{3}$$

where, Y = lny, z_{i1}, z_{i2}, ..., z_{ip}, respectively $\ln x_{1i}$, $\ln x_{2i}$, ..., $\ln x_{ip}$; b_1, b_2, ..., b_p and $a = \ln b_0$ are the parameters and X_{ip} (p = 1, 2, …, 7) are the independent variables, which are total length (cm), standard length (cm), fork length (cm), head length (cm), body depth (cm), body circuit (cm) and body height (cm), respectively.

$\delta_i \sim (0, \sigma^2)$ or $e_i \sim (1, e^{\sigma^2})$ is the random error (Draper and Smith, 1981; Gunst and Mason, 1980; Kleinbaum *et al*., 1998). In the multiple regression analysis, the following t-test statistics is benefited in order to test the importance of regression coefficients which is given in Eq. 4.

$$t_j = \frac{b_j - \beta_j}{\sqrt{var(b_j)}} \sim t_{\alpha(n-p-1)}; \; j = 1, 2, \ldots, p \tag{4}$$

where, var (b_j) is the diagonal member of matrix $s^2(X'X)^{-1}$ and also s^2 is the Mean Square of residual (MS) which obtains from the ANOVA. Of course, we test null hypothesis such as $\beta_j = 0$.

One of the important problems for usage of multiple linear regression analysis is the presence of multicollinearity among the used predictor variables to estimate the body weight of the fish. Multicollinearity is a statistical term for the existence of a high degree of linear correlation amongst two or more explanatory variables in a multiple regression model. In the presence of multicollinearity, it will be difficult to assess the effect of the independent variables on the dependent variable (Anonymous, 2008). To detection of multicollinearity, tolerance or the Variation Inflation Factor (VIF) should be calculated by means of the following equations:

$$\text{Tolerance} = 1 - R^2, \; \text{VIF} = 1/\text{Tolerance} \tag{5}$$

The largest VIF value among all independent variables is used as indicator of the severity of multicollinearity. A maximum VIF value in excess of 10 is taken as an indication that multicollinearity may be unduly influencing the least squares estimates in multiple linear regression (Neter *et al*., 1989).

To overcome the limitations of the multiple linear regression method, the usage of its method based on factor scores which were estimated in factor analysis can be preferred rather than this classical method for conditions in which varying degrees of multicollinearity are present among the examined variables.

The goals of factor analysis are to determine the number of fundamental influences underlying a domain of variables, to quantify the extent to which each variable is associated with the factors and to obtain information about their nature from observing which factors contribute to performance, on which variables (Tinsley and Brown, 2000). This allows numerous intercorrelated variables to be condensed into fewer dimensions, called factors.

The basic factor analysis equation can be presented in matrix form as:

$$Z = \lambda F + \varepsilon \tag{6}$$

where:

Z = A p×1 vector of variables
λ = A p×m matrix of factor loadings
F = A m×1 vector of factors
ε = A p×1 vector error (Sharma, 1996)

In our study, the correlation matrix of variables was used to obtained eigenvalues. In order to facilitate interpretation of factor loadings (l_{ik}), VARIMAX rotation was used. Factor coefficients (c_{ik}) were used to obtain factor scores for selected factor (Keskin *et al*., 2007). Factor scores can be derived such that they are nearly uncorrelated or orthogonal. Thus, the problem for multicollinearity among the variables which are used to estimate the body weight of the fish can be solved by using the coefficients. The factor number equals the number of Eigenvalues of the population correlation matrix that are greater than unity (Tinsley and Brown, 2000). Therefore, the factors with eigenvalues >1 were employed in multiple regression analysis (Sharma, 1996).

All the computational work was performed to estimate Body Weight (BW) with respect to some body measurements (Total Length (TL), Standard Length (SL), Fork Length (FL), Head Length (HL), Body Depth (BD), Body Circuit (BC) and Body Height (BH)) measured from Lizardfish in Iskenderun Bay by means of MINITAB and SPSS statistical package programs.

RESULTS AND DISCUSSION

The descriptive statistics for Lizardfish traits are given in Table 1.

Transformed data for all traits were explored for normality by using Kolmogorow-Smirnov normality test in SPSS (10.0 V) and were normally distributed ($p>0.05$).

Bivariate correlations displaying the relationship among all morphological characters considered are given in Table 2.

There were positive relationships among all body measurements and the body weight of Lizardfish (Table 2). The highest correlation was predicted between standard length and fork length (0.99), while the lowest correlation was between body depth and total length (0.52) ($p<0.01$). When multiple linear regression is used to analyze a data set, as the magnitude of the relationships among the independent variables (SL and FL, r = 0.99) increases, less and less reliance can be placed on the results generated by an ordinary least squares solution. The standard error, t-values, p-values and VIF values for each regression coefficient (β_i) based on the results of multiple regression analysis was given in Table 3.

Table 3 shows that the fork length, body depth and head length were found to be insignificant. Moreover, there was multicollinearity between standard length and fork length due to the largest VIF values (41.6 and 32.1) of the two traits of Lizardfish. This results indicated that the standard errors inflate (for example the value of standard error for FL is 0.34, while the value of the mean is 0.24), resulting in unstable parameter estimates.

The result of factor analysis presented that the first-three of ten factors were selected as independent variables for multiple regression model because three factors have eigenvalues >1. Because eigenvalues presented variances and that standardized variable contributes to principal component extraction is 1, a component with the eigenvalue less than 1 is not as important (Keskin *et al*., 2007; Tabachnick and Fidelli, 2001). The selected three factors explain 85.6% of total variation in factor analysis (Table 4). Factor 1-3 accounted for 31.5, 29.9 and 24.2% of the variation (Var) in all variables, respectively. Moreover, the first factor accounted for 36.8% (2.206/5.989) of the variation in the solution; the second factor accounted for 34.9%

Table 1: Descriptive statistics for examined traits of lizardfish

				95% C.I. for mean	
Traits	n	Mean	SD	Lower bound	Upper bound
BW	43	56.46	17.403	51.108	61.819
TL	43	19.74	2.226	19.054	20.425
SL	43	18.23	1.939	17.634	18.827
FL	43	16.73	1.873	16.149	17.302
HL	43	2.71	0.292	2.624	2.804
BD	43	2.61	0.341	2.509	2.719
BC	43	2.10	0.210	2.035	2.165
BH	43	8.37	1.078	8.036	8.699

BW: Body Weight; TL: Total Length; SL: Standard Length; FL: Fork Length; HL: Head Length; BD: Body Depth; BC: Body Circuit; BH: Body Height; CI: Confidence Interval

Table 2: Bivariate correlation for some body traits and body weight of lizardfish

Traits	BW	TL	SL	FL	HL	BD	BC	BH
BW	1.00							
TL	0.74**	1.00						
SL	0.97**	0.73**	1.00					
FL	0.96**	0.72**	0.99**	1.00				
HL	0.87**	0.53**	0.88**	0.87**	1.00			
BD	0.80**	0.52**	0.70**	0.69**	0.64**	1.00		
BC	0.80**	0.60**	0.76**	0.73**	0.63**	0.72**	1.00	
BH	0.91**	0.55**	0.81**	0.81**	0.77**	0.85**	0.73**	1.00

**p<0.01

Table 3: Results of multiple regression analysis

Traits	Coefficients	SE	t-value	p-value	VIF
Constant	-2.98	0.35	-8.53	<0.001	
TL	0.29	0.09	3.43	0.002	2.6
SL	1.14	0.41	2.80	0.008	41.6
FL	0.22	0.34	0.65	0.517	32.1
HL	0.27	0.15	1.81	0.079	5.6
BD	0.17	0.10	1.75	0.089	3.9
BC	0.24	0.11	2.18	0.036	3.0
BH	0.72	0.12	5.75	<0.001	5.9

S = 0.043, R-Sq = 98.5%, R-Sq(adj) = 98.2%

Table 4: Results of factor analysis

	Factor score coefficients (c_{ik})			Rotated factor loadings (l_{ik}) and communalities			
Variables	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3	Communality
TL	-0.326	-0.002	0.391	0.285	-0.263	0.732	0.687
SL	0.468	0.752	1.623	0.710	-0.385	0.590	1.000
FL	0.351	-0.018	-0.435	0.719	-0.384	0.533	0.971
HL	0.604	-0.035	-0.750	0.804	-0.388	0.270	0.871
BD	-0.390	-0.805	-0.055	0.287	-0.851	0.290	0.891
BC	-0.197	-0.133	0.151	0.341	-0.557	0.514	0.691
BH	0.034	-0.518	-0.380	0.507	-0.735	0.285	0.879
Variance				2.206	2.090	1.693	5.989
Var. (%)				0.315	0.299	0.242	0.856

Table 5: Results of multiple regression analysis based on the result of factor analysis

Predictors	Coefficients	SE	t-value	p-value	VIF
Constant	-107.80	3.486	-30.92	<0.001	
Factor 1	07.27	0.446	16.33	<0.001	2.4
Factor 2	-11.99	0.831	-14.43	<0.001	4.5
Factor 3	7.26	0.279	26.02	<0.001	4.2

S = 2.268, R-Sq = 98.4%, R-Sq(adj) = 98.3%

(2.090/5.989) of the variation in the solution, the third factor accounted for 28.3% (1.693/5.989) of the variation in the solution.

After orthogonal rotation, the factor loadings were presented the relationship between examined variables and corresponding factors (Table 4). Here, it was seen that there were high correlation between standard, fork and head length traits of the fish and Factor 1; body depth and body height traits were high correlated with Factor 2 and total length trait was high correlated with Factor 3. The highest values of communalities indicate that the variances of variables were efficiently reflected by factors in multiple regression analysis. Factor score values for the three factors, which were obtained by means of factor score coefficients given in Table 4, were used as independent variables in the regression analysis to determine significant factor/s on body weight of Lizardfish (Table 5).

It was found that all of the selected factors had significant effect on body weight. The factors also explained 98.4% of the variance in the body weight of Lizardfish. Moreover, because VIF values for the factors were smaller than 10, the problem of multicollinearity presented in Table 3 was solved (Table 5).

CONCLUSION

Due to the fact that this study evaluated classical multiple linear regression and the regression analysis based on factor scores building model and drawing conclusions in the presence of multicollinearity among the examined predictor variables, this paper provides an introduction to a variety of multiple linear regression methods.

REFERENCES

Akar, M., L. Sangun and M. Baylan, 2001. A study about some quantitative traits for specie of Serranus hepatus, XI. Aquaculture Symposium, Hatay, Turkey, 1: 360-367 (in Turkish). ISBN: 975-7989-11-8.

Anonymous, 2008. Multicollinearity. http://en.wikipedia.org/wiki/Multicollinear, Last Access.

Cankaya, S., G.T. Kayaalp, L. Sangun, Y. Tahtali and M. Akar, 2006. A comparative study of estimation methods for parameters in multiple linear regression model. J. Applied Anim. Res., 29: 43-47. ISBN: 0971-2119.

Draper, N.R. and H. Smith, 1981. Applied Regression Analysis. 2nd Edn. New York: John Wiley and Sons, Inc., pp: 709. ISBN: 0-471-02995-5.

Gunst, R.F. and R.L. Mason, 1980. Regression analysis and its application, a data-oriented approach. New York: Marcel Dekker, Inc., pp: 402. ISBN: 0-824-76993-7.

Keskin, S., I. Daskiran and A. Kor, 2007. Factor analysis scores in a multiple linear regression model for the prediction of carcass weight in akkeci kids. J. Applied Anim. Res., 31: 201-204. ISBN: 0971-2119.

Kleinbaum, D.G., L.L. Kupper, K.E. Muller and A. Nizam, 1998. Applied Regression Analysis and Multivariable Methods. 3rd Edn. Duxbury Press, pp: 798. ISBN: 0-534-20910-6.

Neter, J., W. Wasserman and M.H. Kutner, 1989. Applied Linear Statistical Models. 2nd Edn. Boston, MA: Irwin Inc., pp: 667. ISBN: 0-256-07068-7.

Sharma, S., 1996. Applied Multivariate Techniques. New York: John Wiley and Sons, Inc., pp: 493. ISBN: 0-471-31064-6.

Tabachnick, B.G. and L.S. Fidell, 2001. Using Multivariate Statistics. 4th Edn. New York: Allyn and Bacon, Inc., pp: 996. ISBN: 0-321-05677-9.

Tinsley, H.E.A. and S.D. Brown, 2000. Handbook of Applied Multivariate Statistics and Mathematical Modeling. New York: Academic Press, pp: 721. ISBN: 0-12-691360-9.