Available online at www.sciencedirect.com

ScienceDirect

Procedia Social and Behavioral Sciences 8 (2010) 549–554

ELSEVIER

Procedia
Social and Behavioral Sciences

International Conference on Mathematics Education Research 2010 (ICMER 2010)

Stepwise Multiple Regression Method to Forecast Fish Landing

Intan Martina Md Ghani[a,*], Sabri Ahmad[b]

[a,b]*Department of Mathematics, Faculty of Science and Technology, Universiti Malaysia Terengganu, Malaysia*

Abstract

There are six types of linear regression analyses that available in statistics which are simple linear regression, multiple linear regressions, logistic regression, ordinal regression, multinominal regression and descriminant analysis. Multiple linear regressions are the one of linear regression analyses that used to analyze the relationship between single response variable (dependent variable) with two or more controlled variables (independent variables). In this paper, stepwise multiple regression will use because this method is combination of forward selection and backward elimination method. The main objective in this paper is to select the suitable controlled variables in forecast fish landing. Data that has been used in this research were taken from Fisheries Annually Statistics of Department of Fisheries Malaysia. Then, the data will be analyzed by using Minitab 15 and SPSS 17.0.

Keywords**:** Forecast; Fish landing; Regression analyses; Stepwise multiple regression

1. Introduction

A statistics analysis is widely used in all aspects such as in science, medicine, fisheries (Ofuoku *et al.*, 2007) and also in social sciences (Sarker *et al.*, 2006). There are many methods in statistics and one of them is regression. There are six types of linear regression analyses which are simple linear regression, multiple linear regression, logistic regression, ordinal regression, multinominal regression and desriminant analysis. Multiple linear regression was selected to build a model of fish landing. Some method that categorized in the stepwise-type procedures which is stepwise regression also used in this paper. The main objective in this paper is to select the suitable controlled variables in forecast fish landing.

2. Literature review

Wan Nawang *et al.* (2009) have conducted a research to identify factor of youth's interest to become fishermen. There are two methods that use in their research which are factor analysis and multiple linear regression. According to the MLR that was used in their research, it shows that there are three factors that influence the respondence interest in fishermen career which are training programme, profitable and marketing. MLR also used by Ofuolu *et al.* (2007) in their research to determine of adoption of improved fish production technologies among fish farmers in

* Intan Martina Md Ghani. Tel.: +6-017-978-3143 ; fax: +09-669-466-0
E-mail address: intan_martina85@yahoo.com

1877-0428 © 2010 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.sbspro.2010.12.076

Delta States, Nigeria. Level of education, farm size, farm income and extension contact show positively correlation while age and household size shows negatively correlation between fish production technologies. MLR that applied in the research by Khamis *et al.* (2003) states that the higher of R^2 gives good result on model fitting. Combination method of Pearson correlations, multiple and simple linear regression and ANOVA was used by Sain (2006) to see if there was change in measured habitat and fish metrics occurred in relation to increased urbanization. Sain (2006) used multiple and linear regression to explain the strongest relationship between fish and habitat.

3. Methodolgy

3.1 Data

In this paper, data were taken from Fisheries Statistics at official website of Department of Fisheries Malaysia, Ministry of Agriculture & Agro-Based Industry Malaysia. The fisheries statistics include during 40 years which is from 1968 until 2007. This research is focusing on the marine fish landing in Terengganu.

3.2 Research method

Multiple linear regressions (MLR) are the method of statistics in regression that used to analyze the relationship between single response variable (dependent variable) with two or more controlled variables (independent variables). This method was selected for this research because there were more than controlled variables. In this research, response variable is marine fish landing (Y) while fishermen (X_1), fishing boat (X_2) and fishing gears licensed (X_3) were controlled variables.

There were four general steps to build forecasting model of fish landing in MLR. The general steps were checking assumptions, selecting suitable methods of MLR, interpret the output and develop equation of MLR.

Step 1: Checking assumptions

The first step is to build forecasting model by checking assumptions of data. There are four assumptions that should be check which are normality, linearity, heteroscedasticity and multicollinearity. All of the variables in this paper must be normal distribution. The normal distribution can be seen via histogram graph, plot P-P, plot Q-Q, kurtosis and skewness. If the distribution of data is not normal, so we need to use transformation.

Then, MLR should have linear relationship between response variable and controlled variables. in regression the model, we fit is a linear model ('linear model' just means 'model based on a straight line') (Andy, 2005).

The third assumption in MLR is any data should be free from heteroscedasticity. Heteroscedasticity will happen whenever there is interruption in the model that not fulfilled. If the important variables in the model are missing, hence heteroscedasticity will happen. Any model can be check whether there is heteroscedasticity or not based on Spearman's rank correlation test.

The last assumption that should be checked in research is multicollinearity. Multicollinearity means situation that has high degree of correlation between controlled variables. Any analyses can be known the present of multicollinearity by checking the value of variation inflation factor (VIF). When the value of VIF is less than 5, hence multicollinearity is not serious. While if VIF is more than 5, then multicollinearity is substantial. Multicollinearity will be more serious whenever the value of VIF is more than 10.

Step 2: Selecting suitable methods of multiple linear regression

There are three methods in MLR which are forward selection, backward elimination and stepwise regression. All three methods can be categorized into stepwise-type procedures. In this research, only stepwise regression method was applied. Stepwise regression method is a combination of forward selection and backward elimination. By referring Minitab Methods and Formulas, standard stepwise regression both adds and removes controlled variables as needed for each step. Minitab ended its procedure when all variables not in the model have p-value that are less than the specified Alpha-to-Enter value and when all variables in the model have p-value that are greater than or equal to the specified Alpha-to-Remove value. Based on Minitab StatGuide, Alpha-to-Enter is a value that determines if any of the predictors that not currently in the model should be added to the model. While Alpha-to-Remove is a value that determines if any of the predictors in the model should be removed from the model.

Step 3: Interpreting the output

From the SPSS output, we can interpret the values of Pearson coefficient, multiple coefficient of determination (R^2), multiple correlation coefficient (R) and adjusted multiple coefficient of determination (adjusted R^2). Correlation analyses is the method in regression that is used to look how far the relationship between two variables. Pearson linear correlation was applied because to check the existence of relationship between two variables in this research. According to Piaw (2006), some of the strength value (r) are as follows:

Table 1. Correlation coefficient value strength

Correlation coefficient size (r)	Correlation strength
.91 until 1.00 or -.91 until -1.00	Very strong
.71 until .90 or -.71 until -.90	Strong
.51 until .70 or -.51 until -.70	Medium
.31 until .50 or -.31 until -.50	Weak
.01 until .30 or -.01 until -.30	Very weak
.00	No correlation

Source: (Piaw, 2006)

R^2 means the proportion of the total variation in the *n* observed values of the dependent variable that is explained by the overall regression model (Bowerman *et al.*, 2005). The higher R^2, the better the model fits your data. R is the positive squared root of R^2 (Levin & Rubin, 1994).

Adjusted R^2 is a measure of the loss of predictive power or shrinkage in regression. The adjusted R^2 will explain how much variance in the outcome would be accounted for if the model had been derived from the population from which the sample was taken (Andy, 2005). The larger adjusted R^2, the better the model fits the data.

Step 4: Developing equation of multiple linear regression

In this research, the hypotheses that used:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
H_a : At least one of the β_1, β_2 and β_3 does not equal to 0

which says that

H_0 : None of the controlled variable X_1, X_2 and X_3 is significantly related to Y
H_a : At least one of the controlled variable X_1, X_2 and X_3 is significantly related to Y

The model of multiple linear regression can be represent as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \qquad (1)$$

where

Y = Response variable (Marine fish landing)
β_0 = Constant variable
β_1 = Coefficient of first control variable, X_1
β_2 = Coefficient of second control variable, X_2
β_3 = Coefficient of third control variable, X_3
X_1 = Controlled variable (Fishermen)
X_2 = Controlled variable (Fishing boat)
X_3 = Controlled variable (Fishing gears licensed)
ε = Error

4. Results and discussions

All data was analyzed using Statistical Package for the Social Sciences 17.0 (SPSS 17.0) and Minitab version 15 (Minitab 15).

Table 2. Test of normality

	Kolmogorov-Smirnov			**Shapiro-Wilk**		
	Statistic	**df**	**Sig.**	**Statistic**	**df**	**Sig.**
Marine fish landing	.080	40	.200	.973	40	.446

The normality can check by using two test as shown in Table 2 which are by Kolmogorov-Smirnov (if sample size is more than 50) and Shapiro-Wilk (if sample size is less than 50). Since the sample size in this research $(n = 40 < 50)$, so we need to use Shapiro-Wilk which gives a Sig. value equal to .446. It shows that marine fish landing is normally distributed (Sig.= .446 > 0.05).

Table 3. Correlations between response variable and controlled variables

Response variable	**Controlled variables**	**Pearson correlation**	**Sig. (2-tailed)**
Marine fish landing	Fishermen	-.429	.006
	Fishing boat	-.268	.095
	Fishing gears licensed	.630	.000
Fishermen	Fishing boat	.828	.000
	Fishing gears licensed	-.239	.138
Fishing boat	Fishing gears licensed	-.073	.655

According to Table 3, it shows that correlations between response variable and controlled variables has been determined using Pearson Correlation. From this table, it shows that the weak correlation between marine fish landing and fishermen (r=-.429), while the correlation between marine fish landing and fishing gears licensed (r=.630) is medium. The strong correlation shows between fishermen and fishing boat (r=.828). Therefore the three relationship are significant because Sig. < .05.

The strength correlation between marine fish landing and fishing boat (r=-.268), fishermen and fishing gears licensed (r=-.239) also fishing boat and fishing gears licensed (r=-.073) shows very weak. The three relationship explains that there are not significant because Sig. > .05. The negative sign in the Table 3 shows the direction of the correlation.

Table 4. Model Summary

Model	R	R^2	Adjusted R^2	Standard error of the estimate
1	.630	.397	.381	27700.45476
2	.692	.479	.451	26081.83790

The model summary is based on the Table 4. The value of R^2 is .397 (Model 1) shows that there are 39.7% (R=.630) changes in response variable (marine fish landing). It is because changes in controlled variable (fishing gears licensed). This explains that fishing gears licensed is major factor to the marine fish landing. The value of R^2 is .479 (R=.692) for Model 2 shows that there are 47.9% changes in response variable (marine fish landing) is occurred because changes in combination of two controlled variables which are fishing gears licensed and fishermen. By comparing both models, Model 2 is a better model fits to the data than Model 1. This is because the higher the value or R^2 and adjusted R^2 (Model 1= .381; Model 2= .451), the better the model fits to the data.

Table 5. Coefficients

	Model	Coefficients	Standard error	Beta	t-value	Significance value
1	Constant	16062.147	14266.689		1.126	.267
	Fishing gears licensed	33.463	6.691	.630	5.001	.000
2	Constant	88359.030	32740.887		2.699	.010
	Fishermen	-6.410	2.647	-.296	-2.421	.020
	Fishing gears licensed	29.716	6.487	.559	4.581	.000

According to Table 5, it shows coefficients. From that coefficient, we can develop two model using MLR equation such as below:

Model 1:

$$Y = 16062147 + 33.463X_3 \tag{2}$$

Model 2:

$$Y = 88359.030 - 6.410X_1 + 29.716X_3 \tag{3}$$

Based on the Model 1, the standardized coefficients for fishing gears licensed ($\beta = .630, p < .05$) is significant. This explained that only fishing gears licensed is factor to the marine fish landing. While in Model 2, the standardized coefficients for fishermen ($\beta = -.296, p < .05$) and fishing gears licensed ($\beta = .559, p < .05$) are significant. This explained that fishing gears licensed and fishermen are factors to the marine fish landing. As conclude, Model 2 is better model fits to the data than Model 1.

Table 6. Stepwise regression results

Step	1	2
Constant	16062	88359
X_3	33.5	29.7
t-value	5.00	4.58
p-value	.000	.000
X_1		-6.4
t-value		-2.42
p-value		.020
S	27700	26082
R-Sq	39.69	47.94

R-Sq(adj)	38.11	45.13
Mallows Cp	5.8	2.1

According to Table 6, it presents the result of stepwise regression using Minitab 15. There are two steps that used to select the controlled variables. In the step 1, X_3 indicates smallest p-value than Alpha-to-Enter (p=.000<.05). Therefore X_3 is the first variable that enters into the model. In the step 2, the p-value for X_1 shows .020 which is smallest than .05. Therefore X_1 is the second variable that needs to enter into the model. X_3 is retained in the model with the coefficient 29.7, t-value is 4.58 and p-value is .000. After the second step, there are no controlled variables that enter and remove the model.

For the marine fish landing output, the value of S decreases from step 1 (S=27700) to step 2 (S=26082), R-Sq and R-Sq(adj) in the step 2 (R-Sq=47.94, R-Sq(adj)=45.13) is higher than step 1 (R-Sq=39.69, R-Sq(adj)=38.11)and Mallows Cp in step 2 is closer to number of controlled variables than step 1. Hence, these statistics indicates step 2 which containing controlled variables X_3 and X_1 is provides better fits to the data. The final model shows only two controlled variables by using stepwise regression and produced regression model such as follows:

$$Y = 88359 + 29.7X_3 - 6.4X_1 \tag{4}$$

5. Conclusions

This paper using multiple linear regressions (MLR) to built forecasting model for fish landing. Based on the output from SPSS 17.0, there are two model was built. By comparing two models using SPSS 17.0, Model 2 is a better model fits to the data than Model 1. This is because the value of R^2 and adjusted R^2 in Model 2 (R^2=.479, adjusted R^2=.451) is higher than Model 1 (R^2=.397, adjusted R^2=.381). In the stepwise regression method shows that only two controlled variables that were selected using Minitab 15 which are fishermen and fishing gears licensed. This explained that only fishermen and fishing gears licensed are factors to the marine fish landing.

References

Andy, F. (2005). Discovering statistics using SPSS: (and sex, drugs and rock ‘n’ roll). London: SAGE.

Bowerman, B.L., O’Connell, R.T. and Koehler, A.B. (2005). Forecasting, time series and regression. 4th ed. United States of America: Brooks/Cole Thomson Learning Inc.

Khamis, A., Ismail, Z. & Shabri, A. (2003). Pemodelan harga minyak sayuran menggunakan analisis regresi linear berganda. *Matematika*, Jilid 9, bil. 1, 59-70.

Levinm R.I. and Rubin, D.S. (1994). Statistics for management. 6th ed. Englewood Cliffs, N.J.: Prentice Hall.

Ofuoku, A.U., Olele & Emah, G. (2007). Determinants of adoption of improved fish production technologies among fish farmer in Delta States, Nigeria. *Journal of Fisheries International*, 2(2):147-151.

Piaw, C.Y. (2006). Asas statistik penyelidikan (Buku 2). Kuala Lumpur: McGraw-Hill.

Sain, R.L. (2006). Characterizing how fish communities and physical habitat structure are affected by urbanization in an East Tennessee Watershed, Master thesis, University of Tennessee, Knoxville.

Sarker, M.A., Chowdhury, A.H. & Itohara, Y. (2006). Entrepreneurships barriers of pond fish culture in Bangladesh- A case study from Mymensingh district. *Journal of Social Sciences*, 2(3):68-73.

Wan Nawang, W.M.Z., Mamat, I. & Mohd Isa, A.M. (2009). Faktor peramal minat belia untuk menjadi nelayan: Satu kajian di Mukim Kuala Besut, Terengganu. *Jurnal Teknologi*, 50(E):29-52, Universiti Teknologi Malaysia.