Trends in Applied Sciences Research

ISSN 1819-3579

Academic
Journals Inc.

www.academicjournals.com

Trends in Applied Sciences Research 6 (11): 1241-1255, 2011
ISSN 1819-3579 / DOI: 10.3923/tasr.2011.1241.1255

An Alternative Multicollinearity Approach in Solving Multiple Regression Problem

H.J. Zainodin, A. Noraini and S.J. Yap
School of Science and Technology, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

Corresponding Author: H.J. Zainodin, School of Science and Technology, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

ABSTRACT

This study illustrated the procedures in selecting the best model when there are more than one independent variables. In this case, multiple regressions were used to analyze the data. First of all, all of the possible models are listed out. Then, in order to obtain the selected models, the multicollinearity test and coefficient test were carried out on all of the possible models. In this study, the alternative method was used to overcome multicollinearity, rather than the conventional method. After that, the best model was obtained by using the Eight Selection Criteria (8SC). Meanwhile, the normality test and randomness test were also carried out on the residuals of the best model. As a result, by getting the best model, the main factor that indicated the changes of percentage of body fat in men can be identified.

Key words: Multiple regressions, dummy variables, multicollinearity, alternative method, selected models, best model

INTRODUCTION

Regression analysis is a statistical technique concerning about the study of the relationship between one dependent variable and one or more independent variable (Gujarati, 1999). Researchers have made heavy use of regression analysis in business, social sciences, biological sciences and many other fields. The linear regression analysis is used to find the influence of the independent variable on the dependent variable while the multiple regressions is used to find the influence of more than one independent variables on the dependent variable. An example of the study done on multiple regressions is by Matiya *et al*. (2005) in determining the factors influencing the prices of fish and its implications on development of aquaculture in Malawi. Reliable alternative approaches are also suggested to other existing methods in order to obtain better estimates, such as, Midi *et al*. (2009) had proposed a leverage based-near neighbors in the estimation of parameters in heteroscedastic multiple regression models. Besides that, the effect of processing parameters on the microstructures and properties of automobile brake drum using multiple regression analysis was also studied by Oluwadare and Atanda (2007).

A common problem in multiple regression is multicollinearity. As Zainodin and Khuneswari (2009a) had stated that multiple regression is a regression model with more than one explanatory variable. The general form of multiple regression is shown as follows:

$$Y = \Omega_0+\Omega_1W_1+\Omega_2W_2+...+\Omega_kW_k+\text{random scatter} \quad (1)$$

where, Y is the dependent variable, Ω_0 is constant term, Ω_j is the j-th coefficient of independent variable W_j and W_j is the j-th independent variables (included the single independent variables, interaction variables, generated dummy variables and transformed variables) where j = 1, 2, ..., k. When there exist highly correlated independent variables in the model, then multicollinearity effects are said to exist. Various methods had been suggested to overcome this problem. El-Salam (2011) had proposed an estimation procedure for determining ridge regression parameter in terms of least Mean Square Error (MSE). In the presence of multicollinearity, models' parameter estimation became inaccurate. Hence, Camminatiello and Lucademo (2010) had developed an extension of the principal component logistic regression to overcome this problem. Midi *et al.* (2010) had also proposed Robust Variance Inflation Factors (RVIFs) in the detection of multicollinearity due to high leverage points which were the sources of multicollinearity.

MATERIALS AND METHODS

Multiple regressions are used to analyse the data in this study. There are four phases in the model building procedures of multiple regressions, from listing down the all possible models to carrying out the goodness-of-fit on the residual of the best model. The model building procedures are shown in Fig. 1.

All possible models: According to Fig. 1, all of the possible models have to be listed out before analysis is carried out. Zainodin and Khuneswari (2009a) stated that the number of all possible models can be calculated as follows:

$$N = \sum_{j=1}^{q} j({}^{q}C_{j}) \tag{2}$$

where, N is number of possible models and q is single independent variables which excluded the dummy variables.

Selected models

Multicollinearity test: In order to get the selected models, the multicollinearity test is carried out to remove multicollinearity source variables from each models and the procedures are shown in Fig. 2.

In this study, the alternative method is used in overcoming multicollinearity, rather than the conventional method. The multicollinearity source variables are variables with absolute correlation coefficient greater than 0.95 and they are marked with circles in the correlation coefficient matrix. There are three types of cases in the multicollinearity test and the removal steps of multicollinearity source variable are based on these three cases as follows:

Case A: The most common variable is removed first. Then, rerun the reduced model

Case B: When more than one tie exists (or with frequency two and above), the variables with the highest frequency are considered first. Then, independent variable which has the smallest absolute correlation coefficient with Y is removed. Then, rerun the reduced model

Case C: When only one tie exist (or with frequency one), the pair variables which have a higher correlation coefficient is considered first. Then, independent variable which has a smaller absolute correlation coefficient with Y is removed. Then, rerun the reduced model

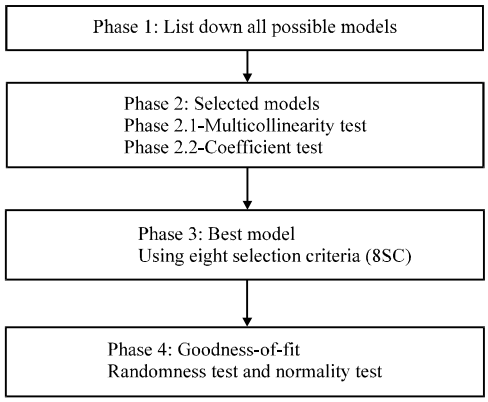

Fig. 1: Model building procedures

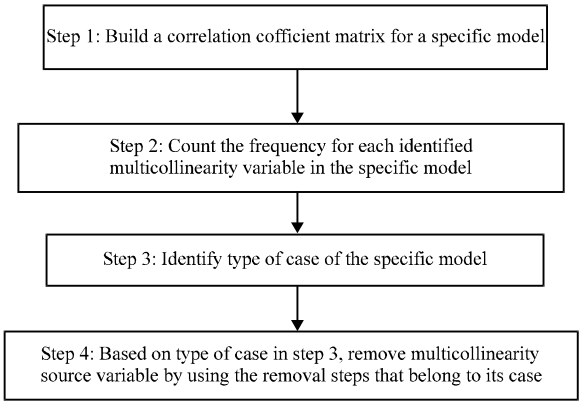

Fig. 2: Multicollinearity test procedures

Then, to get the frequency for a specific identified multicollinearity variable in the correlation coefficient matrix, the algorithm of counting the frequency is as follows:

Step 1: For each variable, draw a horizontal line until off-diagonal values
Step 2: Then, the horizontal line is continued by drawing a vertical line on the lower part values from diagonal value and circle absolute values greater than 0.95
Step 3: Lastly, among all of the values cut by both horizontal and vertical lines, count the number of times the circle (s) has appeared (the diagonal values are not considered)

Since the correlation coefficient matrix is symmetry, thus only the lower diagonal values are considered in counting the number of frequency. Thus, according to Fig. 2, after the frequency for each independent variables in a model are obtained, type of case can be identified and removal of multicollinearity source variable can be carried out. This Zainodin-Noraini multicollinearty remedial procedure is carried out to each of the possible model.

Coefficient test: After removal of multicollinearity source variables, according to Fig. 1, the next step is to perform coefficient test on the reduced model. Zainodin and Khuneswari (2009a) stated

that coefficient test is used to test the coefficient of the corresponding variables. Variables which are insignificant are eliminated subsequently. For a specific j, the hypothesis for Coefficient Test is as below:

$$H_0 : \Omega_j = 0$$

$$H_1 : \Omega_j \neq 0$$

The decision is that the null hypothesis is rejected if $|t_{cal}|$ is greater than $|t_{critical}|$ where

$$t_{cal} = [\hat{\Omega}_j - \Omega_j(H_0)]/[se(\hat{\Omega}_j)]$$

and $|t_{critical}|$ is $t_{\alpha/2,(n-k-1)}$. $se(\hat{\Omega}_j)$ is the standard error for $\hat{\Omega}_j$ and $\Omega_j(H_0)$ is the value of Ω_j under H_0 for j = 1, 2,..., k. The decision is to accept the null hypothesis. Thus, variable with the smallest $|t_{cal}|$ and is nearest to zero is eliminated from the models. The elimination process is repeated until there is no more insignificant variable in the models.

Best model: After all of the selected models are obtained, models with the same independent variables are filtered out. After that, to get the best model, Eight Selection Criteria (8SC) is carried out on the selected models which have undergone filtration. Zainodin and Khuneswari (2009b) have discussed in detail the usage of the 8SC. The Akaike Information Criterion (AIC) (Akaike, 1974) and Finite Prediction Error (FPE) (Akaike, 1969) are developed by Akaike. The Generalised Cross Validation (GCV) is developed by Golub *et al.* (1979) while the HQ criterion is suggested by Hannan and Quinn (1979). The RICE criterion is discussed by Rice (1984) and the SCHWARZ criterion is discussed by Schwarz (1978). The SGMASQ is developed by Ramanathan (2002) and the SHIBATA criterion is suggested by Shibata (1981). The Eight Selection Criteria (8SC) is presented in Table 1.

Table 1: Eight selection criteria (8SC)

No.	Criteria	No.	Criteria
1	AIC: $\left(\frac{SSE}{n}\right)(e)^{2(k+1)/n}$	5	RICE: $\left(\frac{SSE}{n}\right)\left[1-\frac{2(k+1)}{n}\right]^{-1}$
	Akaike (1974)		Rice (1984)
2	FPE: $\left(\frac{SSE}{n}\right)\left[\frac{n+k+1}{n-(k+1)}\right]$	6	SCHWARZ: $\left(\frac{SSE}{n}\right)(n)^{(k+1)/n}$
	Akaike (1969)		Schwarz (1978)
3	GCV: $\left(\frac{SSE}{n}\right)\left[1-\frac{k+1}{n}\right]^{-2}$	7	SGMASQ: $\left(\frac{SSE}{n}\right)\left[1-\frac{k+1}{n}\right]^{-1}$
	Golub *et al.* (1979)		Ramanathan (2002)
4	HQ: $\left(\frac{SSE}{n}\right)(\ln n)^{2(k+1)/n}$	8	SHIBATA: $\left(\frac{SSE}{n}\right)\left[\frac{n+2(k+1)}{n}\right]$
	Hannan and Quinn (1979)		Shibata (1981)

SSE = Sum of square error, k+1 = No. of parameters and n = No. of observations

Goodness-of-fit

Randomness test: Randomness test is used to test the randomness of residuals. The distribution of the residual can be obtained from the histogram and scatter plots of the residuals. Bin Mohd *et al.* (2007) stated that the randomness of residuals, u_i (i = 1, 2, 3,... , n), can be checked by simple correlation coefficient. The procedures are as below:

Step 1: The null and alternative hypotheses are defined as follow:

- H_0: The residuals, u_i are randomly distributed
- H_1: The residuals, u_i are not randomly distributed

Step 2: Test statistic is calculated as follows:

$$R = \frac{\frac{1}{n}\sum_{i=1}^{n} iu_i - \overline{u}\overline{K}}{S_u S_1}$$

where,

$$\overline{u} = \frac{1}{n}\sum_{i=1}^{n} u_i,\ S_u^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - \overline{u})^2,\ \overline{K} = \frac{n+1}{2},\ S_1 = \frac{n^2-1}{12}$$

and R is simple correlation coefficient and n is sample size. Since u_i are independent on i, then random variable

$$T_n = R\sqrt{\frac{(n-p)}{(1-R^2)}}$$

follows a t-distribution with degree of freedom = n-p where p = k+1 which is the number of estimated parameters.

Step 3: The null hypothesis is accepted if $|t_{critical}|$ is greater than $|T_n|$ which means that the residuals u_i are randomly distributed.

Normality test: According to Gujarati (1999) the normality of a regression model can be obtained by using the histogram of residuals and Normal Probability Plot (NPP). By plotting the histogram of residuals, the shape of the underlying probability distribution can be estimated. In the NPP, the variable of interest is normally distributed if a straight line fits the data well. Besides that, the Kolmogorov-Smirnov test and Shapiro-Wilk test are also used to test the normality of the residuals. Kolmogorov-Smirnov test is used when the number of observations is large while Shapiro-Wilk test is used when the number of observations is small. Both of these tests can be carried out by using the SPSS software. The null and hypotheses for normality test are as below:

- H_0: The residuals, u_i are normally distributed
- H_1: The residuals, u_i are not normally distributed

The decision is to accept the null hypothesis if the p-value from the SPSS output is greater than 0.05. Thus, the residuals are assumed to be normally distributed. Apart from this, some graphical plots, such as scatter plot, histogram, Q-Q plot and box plot can also used as supporting evidence for the normality test.

Data analysis

Data description: The data is obtained from Dr. A. Garth Fisher from the Human Performance Research Centre of Brigham Young University and contains the observations of 252 men (Johnson, 1996). In this study, nine variables are selected and analysed. They are the percentage of body fat using Siri's equation, abdomen circumference, adiposity index, chest circumference, hip circumference weight, density, height and neck circumference. According to Bosy-Westphal *et al.* (2005), the Siri's equation used in estimating the percentage of body fat is as follows:

$$\text{Percentage of body fat} = (495/\text{body density})-450 \quad (3)$$

where, the body density will be calculated as weight/volume. The descriptive statistics of these 9 variables are shown in Table 2.

The correlation among dependent variable, percent of body fat using Siri's equation and the other 8 independent variables is presented in Table 3. However, due to limited space, the name of the variables in Table 3 are represented by their short forms, where their full names can be referred in Table 2.

Table 2: Descriptive statistics for all 9 variables

Variables	Mean	Standard error	Median	Mode	Standard deviation	Sample variance	Kurtosis	Skewness	Range	Minimum	Maximum
Percent body fat using Siri's equation (Y)	19.1508	0.5272	19.2000	20.4000	8.3687	70.0358	-0.3338	0.1464	47.5000	0.0000	47.5000
Abdomen circumference (X_1)	92.5560	0.6793	90.9500	100.5000	10.7831	116.2747	2.2488	0.8384	78.7000	69.4000	148.1000
Adiposity index (X_2)	25.4369	0.2298	25.0500	23.7000	3.6481	13.3087	6.7125	1.5617	30.8000	18.1000	48.9000
Chest circumference (X_3)	100.8242	0.5311	99.6500	99.1000	8.4305	71.0729	0.9873	0.6816	56.9000	79.3000	136.2000
Hip circumference (X_4)	99.9048	0.4513	99.3000	98.3000	7.1641	51.3237	7.4714	1.4971	62.7000	85.0000	147.7000
Weight (X_5)	178.9244	1.8513	176.5000	184.2500	29.3892	863.7227	5.2695	1.2053	244.6500	118.5000	363.1500
Density (X_6)	1.0556	0.0012	1.0549	1.0610	0.0190	0.0004	-0.3096	-0.0202	0.1139	0.9950	1.1089
Height (X_7)	70.1488	0.2307	70.0000	71.5000	3.6629	13.4165	59.5443	-5.3850	48.2500	29.5000	77.7500
Neck circumference (X_8)	37.9921	0.1531	38.0000	38.5000	2.4309	5.9093	2.7196	0.5526	20.1000	31.1000	51.2000

Table 3: Correlation coefficient table for all 9 variables

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Y	1.0000								
X_1	0.8134	1.0000							
X_2	0.7275	0.9239	1.0000						
X_3	0.7026	0.9158	0.9118	1.0000					
X_4	0.6252	0.8741	0.8833	0.8294	1.0000				
X_5	0.6124	0.8880	0.8874	0.8942	0.9409	1.0000			
X_6	-0.9878	-0.7990	-0.7147	-0.6826	-0.6093	-0.5941	1.0000		
X_7	-0.0895	0.0878	-0.0249	0.1349	0.1704	0.3083	0.0979	1.0000	
X_8	0.4906	0.7541	0.7779	0.7848	0.7350	0.8307	-0.4730	0.2537	1.0000

Dummy transformation: Dummy variables are variables that take the values of 0 and 1 (Gujarati, 1999). Among the eight independent variables in Table 2, the latter three are transformed into dummy variables because density (X_6) and height (X_7) have negative skewness and among the other six independent variables which are highly correlated with dependent variable Y, neck circumference (X_8) has the weakest correlation coefficient value. In addition, neck circumference can also used in identifying overweight and obese patients (Ben-Noun and Laor, 2006). Therefore, it is suitable to be selected as one of the variables for this study.

The transformation of independent variables into dummy variables can help to decrease the number of possible models in this study. This can be seen by using Eq. 1 and 2 if the three independent variables are not transformed into dummy variables, the number of independent variables are 8 and the number of possible models are 1024. However, if density, height and neck circumference are transformed into dummy variables, the number of possible models for 5 independent variables are 80 only.

After transformation, density (X_6), height (X_7) and neck circumference (X_8) are represented by D, H and N, respectively. The mode for Density (D), Height (H) and neck circumference (N) is 1.061, 71.5 and 38.5, respectively. For those which are less than their respective modes are denoted as 0, while for those observations which are more than their respective modes are denoted as 1. For better understanding, partly of the data of dummy variables after transformation is presented in Table 4.

Procedures in getting the best model: After transformation, according to the model building procedures in Fig. 1, all of the possible models are listed out by using Eq. 1 and 2. Since, there are five single non-dummy independent variables in this study, thus the numbers of all possible models are 80. Then, the selected models can be obtained by carried out the multicollinearity test. For illustration purpose, model M53.0.0 is considered as follows:

$$Y = f(X_1, X_2, X_3, X_5, X_{12}, X_{13}, X_{15}, X_{23}, X_{25}, X_{35}, D, H, N, X_1D, X_1H, X_1N, X_2D, X_2H, X_2N, X_3D, X_3H, X_3N, X_5D, X_5H, X_5N) \quad (4)$$

However, due to limited space, model M53.12.0, which has eliminated 12 independent variables from the parent model is considered as follows:

$$Y = f(X_1, X_2, X_3, X_5, X_{12}, X_{35}, D, H, X_1D, X_1N, X_2H, X_2N, X_3D) \quad (5)$$

Model M53.12.0. which has eliminated 12 independent variables from the parent model can be known from its model name, where 12 represents that 12 variables are eliminated in Phase 2.1 and

Table 4: Partly of the data of dummy variables after transformation

X6 (Mode = 1.061)	Density (D)	X7 (Mode = 71.5)	Height (H)	X8 (Mode = 38.5)	Neck circumference (N)
1.0708	1	67.75	0	36.2	0
1.0853	1	72.25	1	38.5	1
1.0414	0	66.25	0	34.0	0
1.0751	1	72.25	1	37.4	0
.	.	.	.	.	.
.	.	.	.	.	.
1.0271	0	70.00	0	40.8	1

zero shows that no variable is eliminated in Phase 2.2 from the parent model. For better understanding, the definition of model name is presented in Fig. 3. Besides that, the removal of multicollinearity source variables from model M53.12.0 is presented in Table 5.

The frequency tables for several cases in removing the corresponding variable from model M53.12.0 until model M53.17.0 are shown in Table 6.

In Table 5, variable X_{12} is numbered as 13 because it is the 13-th variable removed from model M53.0.0. Model M53.12.0 belongs to Case B because there exists more than one tie, where variables X_{12}, D, X_1D and X_3D has frequency of two respectively. Since variable X_{12} has the smallest absolute correlation coefficient with Y, which is 0.7505, so it is removed from model M53.12.0. Then, the analysis is rerun and a new model M53.13.0 is produced.

Besides that, for model M53.16.0, it belongs to Case C because there exists only one tie. This is due to variable X_5, X_{35}, H, X_2H has frquency of one respectively. Then, the pair variables of X_5 and X_{35} is considered first because it has a higher correlation coefficient than the pair variables of H and X_2H, which are 0.9859. After that, X_5 is removed from model M53.16.0 due to its smaller absolute correlation coefficient with Y than X_{35}, which is 0.6124. Then, the analysis is rerun and a new model M53.17.0 is produced. The same removal steps are carried out on other multicollinearity source variables according to their types of cases. The way to count frequency and the removal steps based on related types of cases. Thus, after removal of 18 variables from model M53.0.0, the correlation coefficient table for variables in model M53.18.0 is shown in Table 7.

From Table 7, it can be observed that all of the absolute correlation coefficient values (excluded the diagonal values) are less than 0.95 and thus model M53.18.0 is said to be free from multicollinearity.

Table 5: Removal of multicollinearity source variables from Model M53.12.0

					17	13				14	15	18		16	
	Y	X_1	X_2	X_3	X_5	X_{12}	X_{35}	D	H	X_1D	X_1N	X_2H	X_2N	X_3D	
Y	1.0000														
X_1	0.8134	1.0000													
X_2	0.7275	0.9239	1.0000												
X_3	0.7026	0.9158	0.9118	1.0000											
X_5	0.6124	0.8880	0.8874	0.8942	1.0000										17
X_{12}	0.7505	**0.9611**	**0.9828**	0.9126	0.8983	1.0000									13
X_{35}	0.6433	0.9152	0.9226	0.9433	**0.9859**	0.9376	1.0000								
D	-0.8081	-0.6331	-0.5320	-0.5273	-0.4903	-0.5483	-0.4922	1.0000							
H	-0.0391	0.1539	0.0393	0.1964	0.4032	0.0880	0.3301	-0.0386	1.0000						
X_1D	-0.7883	-0.5954	-0.5007	-0.4889	-0.4528	-0.5186	-0.4573	**0.9955**	-0.0221	1.0000					14
X_1N	0.4675	0.6707	0.6613	0.6740	0.6953	0.6585	0.6969	-0.3670	0.2146	-0.3403	1.0000				15
X_2H	0.0299	0.2563	0.1584	0.2916	0.5081	0.2086	0.4424	-0.0896	**0.9826**	-0.0716	0.2899	1.0000			18
X_2N	0.4729	0.6820	0.6901	0.6880	0.7130	0.6841	0.7168	-0.3659	0.2028	-0.3397	**0.9974**	0.2825	1.0000		
X_3D	-0.7948	-0.6062	-0.5051	-0.4904	-0.4575	-0.5249	-0.4607	**0.9968**	-0.0232	**0.9988**	-0.3471	-0.0734	-0.3461	1.0000	16

Bold values are cases which are in removing the corresponding variable from model

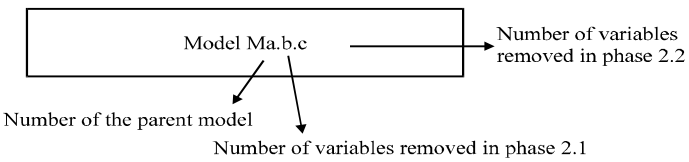

Fig. 3: Definition of model name

Table 6: Frequency tables from Model M53.12.0 until Model M53.17.0

Frequency table for model M53.12.0				Frequency table for model M53.13.0				Frequency table for model M53.14.0			
Variable	Frequency	Case	Action	Variable	Frequency	Case	Action	Variable	Frequency	Case	Action
X_1	1			X5	1			X5	1		
X_2	1			X35	1			X35	1		
X_5	1			D	2			D	1		
X_{12}	2	B	Removed	H	1			H	1		
X_{35}	1			X1D	2	B	Removed	X1N	1	C	Removed
D	2			X1N	1			X2H	1		
H	1			X2H	1			X2N	1		
X1D	2			X2N	1			X3D	1		
X_1N	1			X3D	2						
X_2H	1										
X_2N	1										
X_3D	2										

Frequency table for model M53.15.0				Frequency table for model M53.16.0				Frequency table for model M53.17.0			
Variable	Frequency	Case	Action	Variable	Frequency	Case	Action	Variable	Frequency	Case	Action
X_5	1			X5	1	C	Removed				
X_{35}	1			X35	1			H	1		
D	1			H	1			X2H	1	C	Removed
H	1			X2H	1						
X_2H	1										
X_3D	1	C	Removed								

Table 7: The correlation coefficient for variables in Model M53.18.0

	Y	X_1	X_2	X_3	X_{35}	D	H	X_2N
Y	1.0000							
X_1	0.8134	1.0000						
X_2	0.7275	0.9239	1.0000					
X_3	0.7026	0.9158	0.9118	1.0000				
X_{35}	0.6433	0.9152	0.9226	0.9433	1.0000			
D	-0.8081	-0.6331	-0.5320	-0.5273	-0.4922	1.0000		
H	-0.0391	0.1539	0.0393	0.1964	0.3301	-0.0386	1.0000	
X_2N	0.4729	0.6820	0.6901	0.6880	0.7168	-0.3659	0.2028	1.0000

Then, according to Fig. 1, the coefficient test is carried out to remove insignificant variables from the models. Therefore, further analysis is taken on model M53.18.0, where Table 8 shows the t_{cal} values for each variable in model M53.18.0.

For the hypotheses of coefficient test for model M53.18.0, $|t_{critical}|$ is $t_{0.025, (252-7-1)}$, which is 1.97. The decision is to accept the null hypothesis, where the $|t_{cal}|$ is smaller than $|t_{critical}|$, which shows that the corresponding variable of the specific coefficient has no contribution to the model. For M53.18.0, both of the corresponding variables of β_H and β_{2N}, H and X_2N have $|t_{cal}|$ which are smaller than the $|t_{critical}|$, however only one variable is eliminated in each elimination step. Thus, only variable H is eliminated due to its $|t_{cal}|$ is nearer to zero than variable X_2N. The analysis is rerun with the remaining variables and the new model is model M53.18.1. The resulting t_{cal} values after eliminated variable H are shown in Table 9.

The $|t_{critical}|$ is $t_{0.025, (252-6-1)}$, which is 1.97. The decision is to accept the null hypothesis, where the $|t_{cal}|$ is smaller than $|t_{critical}|$. Since only corresponding variables of β_{2N}, X_2N has $|t_{cal}|$ which is

Table 8: The t_{cal} values for each variable in model M53.18.0

Variables	Parameters	Standard error	t_{cal}	Decision
Constant	-41.4200	6.7449	-6.1410	
X_1	0.5595	0.0675	8.2918	Reject H_0
X_2	0.4983	0.2392	2.0831	Reject H_0
X_3	0.1818	0.0862	2.1091	Reject H_0
X_{35}	-0.0010	0.0002	-4.3792	Reject H_0
D	-7.4346	0.5905	-12.5902	Reject H_0
H	-0.2617	0.7003	-0.3737	Accept H_0
X_2N	-0.0265	0.0222	-1.1969	Accept H_0

Table 9: The t_{cal} values for each variable in model M53.18.1

Variables	Parameters	Standard error	t_{cal}	Decision
Constant	-42.5399	6.0318	-7.0525	
X_1	0.5596	0.0674	8.3065	Reject H_0
X_2	0.5571	0.1798	3.0976	Reject H_0
X_3	0.1883	0.0842	2.2355	Reject H_0
X_{35}	-0.0011	0.0002	-6.5534	Reject H_0
D	-7.4227	0.5886	-12.6106	Reject H_0
X_2N	-0.0269	0.0221	-1.2166	Accept H_0

Table 10: The t_{cal} values for each variable in model M53.18.2

	Parameters	Standard error	t_{cal}	Decision
Constant	-41.3121	5.9526	-6.9401	
X_1	0.5555	0.0673	8.2484	Reject H_0
X_2	0.5425	0.1796	3.0203	Reject H_0
X_3	0.1877	0.0843	2.2262	Reject H_0
X_{35}	-0.0011	0.0002	-6.9128	Reject H_0
D	-7.4430	0.5889	-12.6379	Reject H_0

smaller than the $|t_{critical}|$ and is nearest to zero, thus it is eliminated from model M53.18.1. The analysis is rerun with the remaining variables and the new model is model M53.18.2. The resulting t_{cal} values after eliminated variable X_2N are shown in Table 10.

The $|t_{critical}|$ is $t_{0.025, (252-5-1)}$, which is 1.97. Since, all of the variables have $|t_{cal}|$ that are greater than the $|t_{critical}|$, thus no variable is eliminated from model M53.18.2. Therefore, model M53.18.2 is said to be free from multicollinearity and insignificancy. Besides that, p-values can also used in eliminating insignificant variables, variables with the highest p-values and greater than 0.05 are eliminated from the model one by one. Similar procedures are carried out for other 79 possible models. Table 11 shows the summary for selected models.

All the selected models in Table 11 have filtered out models with the same independent variables, where the first appeared name of the model is taken. For example, model M53.18.2 has the same independent variables with model M57.25.3, model M75.22.3 and model M80.40.4, thus model M53.18.2 is taken to carry out the analysis. Table 12 shows the corresponding selection criteria values for each selected models.

From Table 12, model M53.18.2 is found to be the best model because it has most of the minimum values among the others in 8SC. Model M53.18.2 can be written as in the equation as follow:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{35} X_{35} + \beta_D D + u \quad (6)$$

Table 11: Summary for selected models

Selected model	Summary	k+1	SSE
M1.0.1	M1→M1.0.1	4	3094.973
M2.0.1	M2→M2.0.1	4	3821.549
M3.0.1	M3→M3.0.1	4	3890.011
M4.0.1	M4→M4.0.1	4	4330.523
M5.0.1	M5→M5.0.1	4	4312.925
M8.0.1	M8→M8.0.1	5	3027.433
M9.0.2	M9→M9.0.1→M9.0.2	4	2943.217
M10.0.1	M10→M10.0.1	5	3730.770
M12.0.2	M12→M12.0.1→M12.0.2	4	3769.550
M13.0.1	M13→M13.0.1	5	3807.426
M18.0.2	M18→M18.0.1→M18.0.2	5	2859.993
M23.0.2	M23→M23.0.1→M23.0.2	5	3635.453
M25.0.1	M25→M25.0.1	6	3725.767
M30.0.2	M30→M30.0.1→M30.0.2	6	3564.464
M35.7.1	M35→M35.1.0→...→M35.7.0→M35.7.1	5	2891.949
M44.12.2	M44→M44.1.0→...→M44.12.0→M44.12.1→M44.12.2	5	2864.029
M49.12.2	M49→M49.1.0→...→M49.12.0→M49.12.1→M49.12.2	5	3619.194
M51.12.1	M51→M51.1.0→...→M51.12.0→M51.12.1	6	3738.263
M53.18.2	M53→M53.1.0→...→M53.18.0→M53.18.1→M53.18.2	6	2835.093
M56.18.2	M56→M56.1.0→...→M56.18.0→M56.18.1→M56.18.2	6	3558.772

Table 12: The corresponding selection criteria values for each selected models

Selected model	k+1	SSE	AIC	FPE	GCV	HQ	RICE	SCHWARZ	SGMASQ	SHIBATA
M53.18.2	4	2835.0930	11.6133	11.6133	11.6162	11.8780	11.6192	12.2824	11.4318	11.6075
M73.34.4	6	2835.0930	11.7991	11.7992	11.8059	12.2049	11.8129	12.8334	11.5248	11.7861
M18.0.2	5	2859.9930	11.8086	11.8087	11.8133	12.1461	11.8182	12.6652	11.5789	11.7995
M44.12.2	5	2864.0290	11.8253	11.8253	11.8300	12.1632	11.8348	12.6831	11.5953	11.8162
M35.7.1	5	2891.9490	11.9405	11.9406	11.9453	12.2818	11.9502	12.8067	11.7083	11.9314
M71.22.3	5	2924.4800	12.0749	12.0749	12.0797	12.4199	12.0846	12.9508	11.8400	12.0656
M9.0.2	4	2943.2170	12.0562	12.0562	12.0592	12.3310	12.0624	12.7509	11.8678	12.0502
M8.0.1	5	3027.4330	12.4999	12.5000	12.5049	12.8572	12.5101	13.4067	12.2568	12.4904
M1.0.1	4	3094.9730	12.6778	12.6778	12.6810	12.9668	12.6843	13.4083	12.4797	12.6715
M56.18.2	4	3558.7720	14.5776	14.5777	14.5813	14.9100	14.5851	15.4176	14.3499	14.5704
M30.0.2	4	3564.4640	14.6009	14.6010	14.6047	14.9338	14.6085	15.4423	14.3728	14.5937
M49.12.2	5	3619.1940	14.9433	14.9433	14.9492	15.3703	14.9553	16.0272	14.6526	14.9318
M23.0.2	5	3635.4530	15.0104	15.0105	15.0164	15.4394	15.0225	16.0992	14.7184	14.9989
M66.12.4	4	3720.1710	15.2388	15.2388	15.2426	15.5862	15.2466	16.1168	15.0007	15.2312
M25.0.1	4	3725.7670	15.2617	15.2617	15.2656	15.6096	15.2695	16.1411	15.0233	15.2541
M10.0.1	5	3730.7700	15.4039	15.4040	15.4101	15.8442	15.4164	16.5213	15.1043	15.3921
M51.12.1	4	3738.2630	15.3129	15.3129	15.3168	15.6620	15.3208	16.1952	15.0736	15.3053
M12.0.2	4	3769.5500	15.4410	15.4411	15.4450	15.7931	15.4490	16.3308	15.1998	15.4334
M13.0.1	5	3807.4260	15.7204	15.7205	15.7267	16.1697	15.7332	16.8608	15.4147	15.7084
M2.0.1	4	3821.5490	15.6540	15.6541	15.6580	16.0109	15.6621	16.5560	15.4095	15.6463
M3.0.1	4	3890.0110	15.9345	15.9345	15.9385	16.2977	15.9427	16.8526	15.6855	15.9266
M5.0.1	4	4312.9250	17.6668	17.6669	17.6713	18.0696	17.6759	18.6848	17.3908	17.6581
M4.0.1	4	4330.5230	17.7389	17.7390	17.7434	18.1433	17.7480	18.7611	17.4618	17.7302

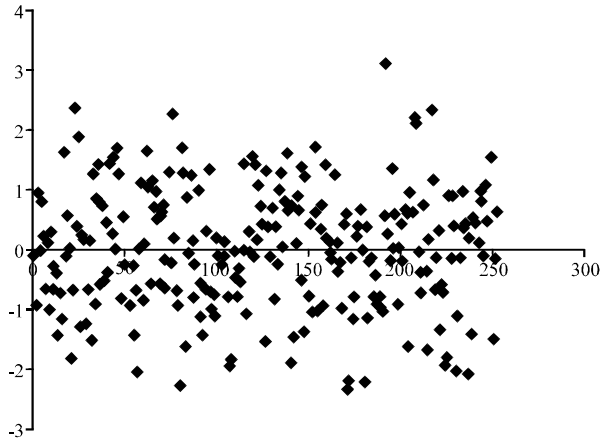

Fig. 4: Scatter plot of standardized residual

where, X_1 represents the abdomen circumference, X_2 is the adiposity index, X_3 is the chest circumference, X_{35} is the first-order interaction variables of chest circumference and weight, D represents density and u is the residual.

According to Fig. 1, after the best model is obtained, the goodness-of-fit is carried out on the residuals of the best model. In this case, the randomness test is carried out to verify the randomness of residuals. The hypothesis of randomness test is as follow:

- H_0: The observations u_i are random
- H_1: The observations u_i are not random

where, I = 1, 2,..., 252

The null hypothesis is accepted if $|T_n|$ is less than $|t_{critical}|$, where $|t_{critical}| = t_{\alpha, n-k-1}$. The calculation of T_n and the result is T_n equals to -0.0013, where k equals to 5 as can be seen in Eq. 6 that there are five independent variables in the best model. From the t-distribution table, at $\alpha = 0.05$, $|t_{critical}| = 1.65$. Since $|T_n| = 0.0013$ is less than $|t_{critical}|$, the null hypothesis is accepted and the residuals u_i are randomly distributed. Besides that, the scatter plot for the standardized residual in Fig. 4 also shows that the residuals are randomly distributed because no obvious pattern is observed.

Then, the normality test is also carried out to test the normality of the residuals in the best model. In this study, Kolmogorov-Smirnov is used to test normality since the number of observations are large, which are 252 men.

The hypothesis of normality test is shown as follow:

- H_0: The standardized residual is normally distributed
- H_1: The standardized residual is not normally distributed

The decision is that the null hypothesis is rejected if the p-value is less than 0.05. Table 13 shows the SPSS output of the Kolmogorov-Smirnov Test on the standardized residual.

Since the p-value in Table 13 is 0.2000, which is greater than 0.05, thus the null hypothesis is accepted and residuals are said to be normally distributed. Besides that, the bell-shaped

Table 13: Kolmogorov-smirnov test on standardized residual

	Kolmogorov-smirnov test		
	Statistic	Degrees of freedom	p-value
Standardized residual	0.0248	252	0.2000

Table 14: The final coefficient values of model M53.18.2

	Unstandardised coefficients			
Model M53.18.2	β	Standard error	t	p-value
Constant	-41.3121	5.9526	-6.9401	0.0000
X_1	0.5555	0.0673	8.2484	0.0000
X_2	0.5425	0.1796	3.0203	0.0028
X_3	0.1877	0.0843	2.2262	0.0269
X_{35}	-0.0011	0.0002	-6.9128	0.0000
D	-7.4430	0.5889	-12.6379	0.0000

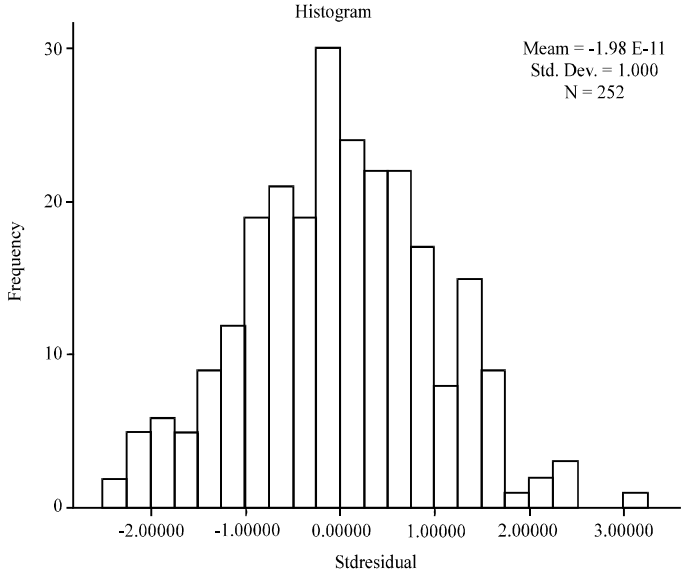

Fig. 5: Histogram of standardized residual

histogram of standardized residual in Fig. 5 also shows that the residuals are normally distributed. Therefore, the residuals of the best model are said to be random and normally distributed.

DISCUSSION

This study showed that model M53.18.2 is the best model, where the equation is shown as in Eq. 6 to represent the factors that affect the percentage of body fat in men. Table 14 shows the final coefficient values of model M53.18.2.

As can be seen from Table 14, the positive coefficient values show that the percentage of body fat in men by using the Siri's equation (Y) will increase if the corresponding variables increase, while the negative coefficient values show that the percentage of body fat in men (Y) will decrease if the corresponding variables decrease. Thus, the increment in abdomen circumference (X_1),

adiposity index (X_2) and chest circumference (X_3) will cause increment on percentage of body fat by Siri's equation (Y) in men. However, the increment in Density (D) and first-order interaction variables of chest circumference and weight (X_{35}) will cause decrement on percentage of body fat by the Siri's equation (Y) in men. This increment in Density (D) is found to bring the most decrement or influence (β = -7.4430) but a very minor change ($\beta = -1.1\times10^{-3}$) on the percentage of body fat in men.

CONCLUSION

As a conclusion, the body density is found to be the main factor that contributed negatively in estimating the percentage of body fat in men, followed by the positive relationships of the other main factors, namely, the abdomen circumference, adiposity index and chest circumference. The interaction variable between the chest and the body weight only caused a very minor negative effect on the percentage of body fat. It is also suggested that further analysis can be carried out by including the Brozek's equation, which is also used in estimating percentage of body fat in human. Comparisons can then be made on the efficiency of both the Siri's equation and Brozek's equation in estimating the percentage of body fat.

REFERENCES

Akaike, H., 1969. Fitting autoregressive models for prediction. Ann. Inst. Stat. Math., 21: 243-247.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control, 19: 716-723.

Ben-Noun, L. and A. Laor, 2006. Relationship between changes in neck circumference and cardiovascular risk factors. Exp. Clin. Cardiol., 11: 14-20.

Bin Mohd, I., S.C. Ningsih and Y. Dasril, 2007. Unimodality test for global optimization of single variable functions using statistical methods. Malaysian J. Math. Sci., 1: 205-215.

Bosy-Westphal, A., S. Danielzik, C. Becker, C. Geisler and S. Onur *et al.*, 2005. Need for optimal body composition data analysis using air-displacement plethysmography in children and adolescents. J. Nutr., 135: 2257-2262.

Camminatiello, I. and A. Lucademo, 2010. Estimating multinomial logit model with multicollinear data. Asian J. Math. Stat., 3: 93-101.

El-Salam, M.E.F.A., 2011. An efficient estimation procedure for determining ridge regression parameter. Asian J. Math. Stat., 4: 90-97.

Golub, G.H., M. Heath and G. Wahba, 1979. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics, 21: 215-223.

Gujarati, D.N., 1999. Essentials of Econometrics. 2nd Edn., McGraw-Hill, New York, USA.

Hannan, E.J. and B.G. Quinn, 1979. The determination of the order of an autoregression. J. R. Stat. Soc. Ser. B, 41: 190-195.

Johnson, R.W., 1996. Fitting percentage of body fat to simple body measurements. J. Stat. Edu., Vol. 4, No. 1.

Matiya, G., Y. Wakabayashi and N. Takenouchi, 2005. Factors influencing the prices of fish in central region of malawi and its implications on the development of aquaculture in malawi. J. Applied Sci., 5: 1424-1429.

Midi, H., A. Bagheri and A.H.M.R. Imon, 2010. The application of robust multicollinearity diagnostic method based on robust coefficient determination to a non-collinear data. J. Applied Sci., 10: 611-619.

Midi, H., S. Rana and A.H.M.R. Imon, 2009. Estimation of parameters in heteroscedastic multiple regression model using leverage based near-neighbors. J. Applied Sci., 9: 4013-4019.

Oluwadare, G.O. and P.O. Atanda, 2007. Effect of processing parameters on the microstructures and properties of automobile brake drum, J. Applied Sci., 7: 2468-2473.

Ramanathan, R., 2002. Introductory Econometrics with Applications. 5th Edn., South-Western, Thomson Learning, Ohio, USA.

Rice, J., 1984. Bandwidth choice for nonparametric kernel regression. Ann. Statistics, 12: 1215-1230.

Schwarz, G.E., 1978. Estimating the dimension of a model. Ann. Stat., 6: 461-464.

Shibata, R., 1981. An optimal selection of regression variables. Biometrika, 68: 45-54.

Zainodin, H.J. and G. Khuneswari, 2009a. A case study on determination of house selling price model using multiple regression. Malaysian J. Math. Sci., 3: 27-44.

Zainodin, H.J. and G. Khuneswari, 2009b. Model-Building approach in multiple regressions. J. Karya Asli Lorekan Ahli Matematik, 2: 1-14.