

Multicollinearity Remedial Techniques in Model-Building

¹Noraini Abdullah and ²Zainodin Haji Jubok

School of Science & Technology, Universiti Malaysia Sabah
88400 Kota Kinabalu, Sabah, Malaysia
e-mail: ¹norainiabdullah.ums@gmail.com, ²zainodin@gmail.com

Abstract The dire need for renewable Green-Energy resources, the vitality for sustainable forest, its management and practices, just to name a few, are examples of the importance for an estimation tool. Hence, a modeling approach is developed, based on selected biomass equations adopted in Forest Science. Models using the multiple regression techniques are employed. Correlation Coefficients of variables are found to have multicollinearity effects. Illustrations on the algorithm of the remedial techniques are exemplified which focuses on the removal of absolute coefficients values of more than 0.95. Significant variables with their possible interactions are selected using statistical tests. Best model is selected based on the eight selection criteria (8SC). The best regression model without multicollinearity is found to give a better estimation with different major contributions from mensuration data.

Keywords Biomass Equation; Correlation Coefficients; Multiple Regression; Multicollinearity; Best Regression Model.

2010 Mathematics Subject Classification 62J02

1 Introduction

Foresters need model approaches to support their decision-makings, which then in turn be used such as to predict future forest stand development [1]. An allometric equation using simple linear regression, relating biomass component to independent variables like diameter at breast height (D_{bh}) and tree height (h) was developed by Wang [2] for 10 co-occurring tree species in China's temperate forests. Onyekwelu [3] estimated biomass production for energy while Noraini *et al.* [4] obtained the best model using the multiple regression technique by executing the backward elimination method of the Coefficient test. Further works of Noraini *et al.* [5] on polynomial regression models had shown an improved stem volumetric estimation when compared to the multiple regression models where data transformation had been carried out. However, the presence of multicollinearity had not been addressed; hence, Zainodin *et al.* [6] had presented an alternative approach in solving multicollinearity in regression problems based on the Microsoft Excel outputs. This paper, hence intend to further extend the alternative multicollinearity approach of Phase 2 in model-building (Noraini *et al.* [5]) on a different software output, namely the SPSS. The objective of this paper is to illustrate these multicollinearity remedial procedures, and identify the best model using the multiple regressions (MR) technique.

2 Materials and Methods

The mensuration data were collected non-destructively from 130 trees which were planted all around the city of Kota Kinabalu, Sabah using convenient sampling. The materials used in the field data measurements would include a clinometer and a girth tape. The clinometer is used to measure the height of a tree while the fiberglass tape measures the diameter indirectly, by wrapping round the tree to measure the circumference in a perpendicular plane to the stem axis, and its value is divided by PI (π) so as to estimate the diameter. The diameter measurements are taken at the respective heights of the tree, namely, at the top, middle, at breast height and at the base of tree. If the height of stem is denoted by h , then the diameter at the top (D_t) is measured at $0.9h$, the middle (D_m) at $0.5 h$, the base (D_m) at $0.1 h$, while diameter at breast height (D_{bh}) is technically approximately 1.4m above the ground level.

a) Stem Biomass Volumetric Equation

In forest science, the biomass equations used to calculate the volume of stem biomass are based on the variables which include diameter at the base (D_b), diameter at breast height (D_{bh}), diameter at middle (D_m), diameter at the top (D_t) and height of stem (h). The areas (A_b , A_m and A_t) at the corresponding stem heights (base, middle and top) are then incorporated into the stem biomass volumetric equations considered such as shown by the following equations:

$$\text{Newton's Formula: } V_N = \frac{h}{6}(A_b + 4A_m + A_t) \quad \text{Fuwape et al. [7]} \quad (1)$$

$$\text{Smalian's Formula: } V_S = \frac{h}{2}(A_b + A_t) \quad \text{Husch et al. [8]} \quad (2)$$

$$\text{Huber's Formula: } V_H = \pi h \left(\frac{D_{bh}^2}{40000} \right) \quad \text{Brack [9]} \quad (3)$$

The variables measured during field data collection are shown as in equations (1), (2) and (3). However, in this paper, only models using the Newton's formula is used for illustrative purposes.

b) Multiple Regression (MR) Models

The MR models are made of a dependent variable and more than two independent variables that can be written in a general form as shown by Ramanathan [10]:

$$V = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + u \quad (4)$$

where β_0 is a constant term, β_j is the j -th coefficient of independent variable X_j . ' X_j ' is an independent variable which represents one of these types of variables, namely, single independent, interactive, generated, transformed or even dummy variables where $j = 1, 2, \dots, k$ and ' u ' as the error terms. In this work, the dependent variable is the stem volume, V , while the independent variables are the mensuration variables during field data collection. The number of possible MR models is obtained by using the

formula $\sum_{j=1}^{q=4} j \binom{q}{j}$, where q is the number of single independent variables. One of the possible models can be given by model M27 as in (5):

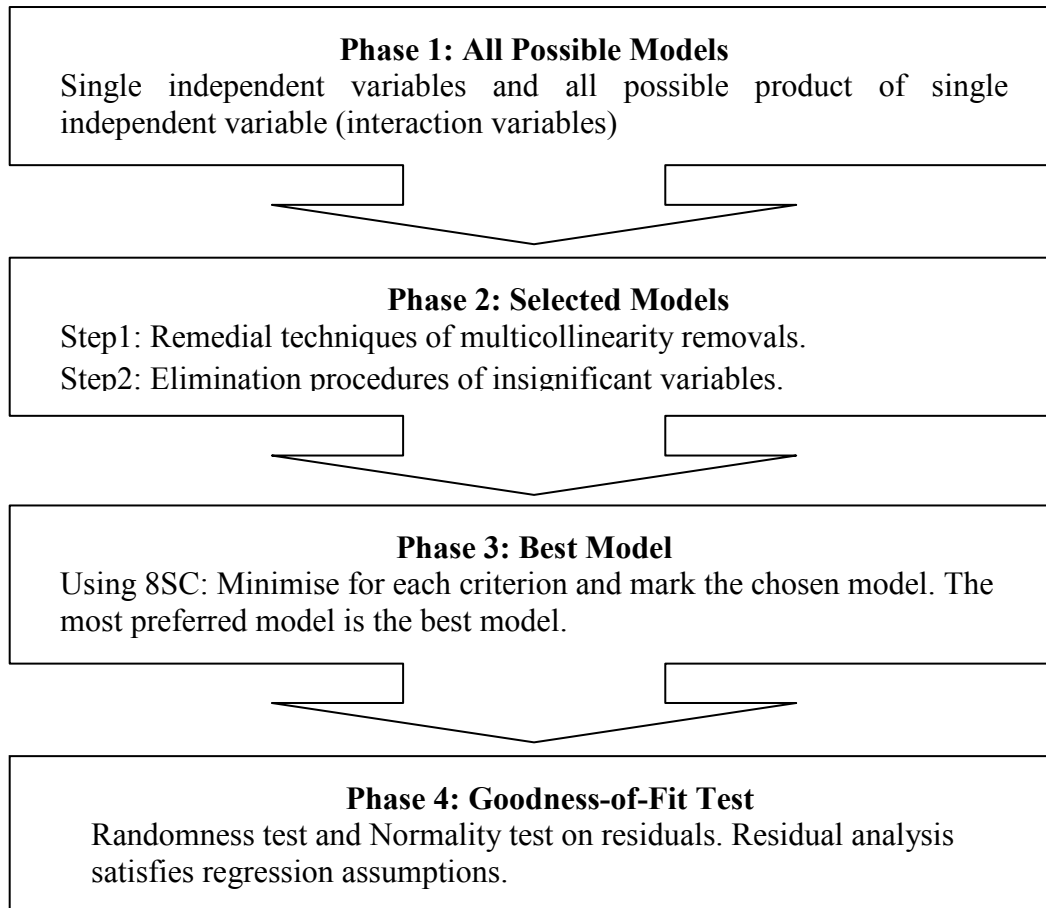
$$V = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{23} X_{23} + \beta_{123} X_{123} + u \quad (5)$$

where X_1 , and X_2 are the single independent variables, X_3 is a generated variable, X_{12} , X_{13} , X_{23} and X_{123} are the interaction variables. These variables may have undergone transformations for non-normal and nonlinear data in the form of (6):

$$\text{Transform variable} = (\text{Original value})^a \quad \text{Devore \& Peck [11]} \quad (6)$$

where ' a ' is the ladders of powers transformations for normality and linearity.

Noraini *et al.* [5] had depicted the flowchart of the four phases in model-building development of the MR models as shown by Figure 1. This paper will focus on the procedures of multicollinearity removals of Step 1 in Phase 2 of the MR model development. The Zainodin-Noraini multicollinearity remedial procedures had been exemplified by Zainodin *et al.* [6] and are based on the Microsoft Excel outputs. The Excel outputs showed the Pearson correlation coefficient table as a half matrix of the upper-lower (UL) diagonal values. This paper, however, will illustrate the multicollinearity removals using the SPSS outputs in a full matrix table as shown by the algorithms below.



Step 1: Obtain the correlation coefficient full matrix of the model in SPSS. A vertical line is then drawn downwards starting from the cells of the independent variables of the full matrix of the upper-lower (UL) diagonal values to an empty cell in a row, say labelled ‘Frequency’.

Step 2: The number of absolute values of correlation coefficients greater than 0.95 ($|r| \geq 0.95$) are then counted. This number denotes the frequency of the multicollinearity source variables.

Step 3: Each source variable and case types are then identified. Procedures in multicollinearity remedies for the case types are hence executed.

It should be noted that the algorithms of the remedial techniques in Excel by Zainodin *et al.* [6] and in SPSS are different; however, both gave similar outcomes where the MR models are used. Figure 2 shows the descriptions and remedial procedures for different multicollinearity cases (Noraini *et al.* [5]; Zainodin *et al.* [6]).

Case Types	Descriptions	Remedial Procedures
Case A	When there exist only one variable with the highest frequency of 2 or more and has the highest absolute correlation coefficient $ r \geq 0.95$ among the independent variables.	Choose the variable with the highest frequency. Remove corresponding variable which is weakly correlated with the dependent variable.
Case B	When there exist variables with highest frequency of 2 or more. Some independent variables have the same frequency of multicollinearity occurrences with high $ r \geq 0.95$. This is known as a tie.	Choose the variables with the highest multiple frequencies. Remove the independent variable with the weakest or least value with the dependent variable.

Case C	When there exists that the highest frequency 1 for all the independent variables.	Any multicollinearity source variable has the same frequency of 1. Remove the weakest with the dependent variable.
--------	---	--

Figure 2 Descriptions and Remedial Procedures for Different Multicollinearity Cases.

c) Data Definitions, Normality And Non-Linearity

The data variables are defined as shown in Table 2. Normality tests are initially carried out and transformations are then done on the non-normal data using the ladders of powers transformations. A generated variable had also been created by Noraini *et al.* [5]. The stem volume equations used for comparisons are based on the Newton's, Huber's and Smalian's formulas. In this paper, however, the multicollinearity remedial procedures are illustrated using the models of the Newton's formula.

2 Results and Analyses

Phase 1: All Possible Models

For four single independent variables taken from field data mensuration, and together with its combinations of interactions, there are 32 possible models considered in this work as shown in Table 1.

Table 1 Number of Possible Models with Interaction Combinations.

Number of Variables	Individual	Order of Interactions			
		First	Second	Third	Total
1	${}^4C_1 = 4$	-	-	-	4
2	${}^4C_2 = 6$	6	-	-	12
3	${}^4C_3 = 4$	4	4	-	12
4	${}^4C_4 = 1$	1	1	1	4
Total	15	11	5	1	32
Models	M1-M15	M16-M26	M27-M31	M32	

Table 2 shows the definition of the original data variables before any transformations are carried out. Table 3 shows the newly defined variables after transformations have been carried out on the non-normal data. The original variable (D_b) is not normal even after have undergoing the ladders of powers transformations. However, normality is attained when it is transformed into a generated variable (D_b/h) which is then used as a newly defined variable (X_3) in the regression analysis.

Table 2 Definition of Original Variables

Variables	Definition
V_N	Stem Volume (m^3): N-Newton
V_S	Stem Volume (m^3): S-Smalian
V_H	Stem Volume (m^3): H-Huber
D_t	Diameter at top of stem
D_m	Diameter at middle of stem
D_b	Diameter at the base of stem
H	Stem height (m)

Table 3 Newly Defined Variables

Original	Transformed	Newly
V_N	V_N^1	V1
V_S	V_S^1	V2
V_H	V_H^1	V3
D_t	$D_t^{3.7}$	X ₁
D_m	$D_m^{4.5}$	X ₂
D_b	D_b/h	X ₃
h	$h^{4.4}$	X ₄

Phase 2: Model Selections - Removals of Multicollinearity

This phase starts with the full model of all individual variables and their possible interactions after multicollinearity removals. Multicollinearity is a statistical phenomenon in which more than two independent variables are highly correlated, thus affecting the Sum of Square Error (SSE) of the respective models. These effects of multicollinearity can be remedied by doing model transformations or by removing the highly correlated source variables (Gujarati [12]). Absolute coefficient values that are greater than 0.75, are normally considered to cause multicollinearity effects.

Table 4 Pearson Correlation Coefficient Matrix of Model M31 (Case A)

M31	V ₁	X ₁	X ₂	X ₃	X ₄	X ₁₂	X ₁₃	X ₁₄	X ₂₃	X ₂₄	X ₃₄	X ₁₂₃	X ₁₂₄	X ₁₃₄	X ₂₃₄
V ₁	1	0.896	0.883	0.115	0.858	0.807	0.840	0.940	0.834	0.932	0.905	0.756	0.821	0.912	0.898
X ₁	0.896	1	0.916	0.218	0.640	0.903	0.968	0.979	0.901	0.907	0.785	0.872	0.885	0.979	0.916
X ₂	0.883	0.916	1	0.225	0.618	0.921	0.895	0.904	0.974	0.982	0.770	0.884	0.902	0.910	0.983
X ₃	0.115	0.218	0.225	1	-0.331	0.223	0.412	0.101	0.393	0.116	0.459	0.316	0.160	0.281	0.275
X ₄	0.858	0.640	0.618	-0.331	1	0.522	0.502	0.746	0.497	0.714	0.677	0.436	0.569	0.635	0.614
X ₁₂	0.807	0.903	0.921	0.223	0.522	1	0.908	0.904	0.927	0.924	0.689	0.987	0.992	0.936	0.954
X ₁₃	0.840	0.968	0.895	0.412	0.502	0.908	1	0.919	0.934	0.859	0.814	0.913	0.872	0.977	0.917
X ₁₄	0.940	0.979	0.904	0.101	0.746	0.904	0.919	1	0.861	0.931	0.791	0.855	0.909	0.973	0.914
X ₂₃	0.834	0.901	0.974	0.393	0.497	0.927	0.934	0.861	1	0.932	0.794	0.924	0.891	0.917	0.980
X ₂₄	0.923	0.907	0.982	0.116	0.714	0.924	0.859	0.931	0.932	1	0.773	0.871	0.926	0.912	0.977
X ₃₄	0.905	0.785	0.770	0.459	0.677	0.689	0.814	0.791	0.794	0.773	1	0.686	0.682	0.837	0.831
X ₁₂₃	0.756	0.872	0.884	0.316	0.436	0.987	0.913	0.855	0.924	0.871	0.686	1	0.968	0.921	0.933
X ₁₂₄	0.821	0.885	0.902	0.160	0.569	0.992	0.872	0.909	0.891	0.926	0.682	0.968	1	0.925	0.941
X ₁₃₄	0.912	0.979	0.910	0.281	0.635	0.936	0.977	0.973	0.917	0.912	0.837	0.921	0.925	1	0.94
X ₂₃₄	0.898	0.916	0.983	0.275	0.614	0.954	0.917	0.914	0.980	0.977	0.813	0.933	0.941	0.94	1
Freq		3	3			3	2	2	2	2		2	2	3	4
Case		B	B			B	B	B	B	B		B	B	B	A
Action Taken	move Variable X ₂₃₄ with highest frequency of type Case A .														

As an example for illustration, taking the variables of Model M31Newton, the Pearson Correlation Coefficient matrix is given in Table 4. The multicollinearity removals algorithms are hence performed. Table 4 exhibits high correlation coefficients of the multicollinearity source variables which are highlighted in yellow. The high correlation coefficients of absolute values greater than 0.95 ($|r| \geq 0.95$) are to be removed. From Table 4, it can be seen that variable X₂₃₄ is of the Case A type that has to be removed due to multicollinearity.

Table 5 Correlation Coefficient Matrix of Model M31.1 (Case B)

M31 .1	V ₁	X ₁	X ₂	X ₃	X ₄	X ₁₂	X ₁₃	X ₁₄	X ₂₃	X ₂₄	X ₃₄	X ₁₂₃	X ₁₂₄	X ₁₃₄
V ₁	1	0.896	0.883	0.115	0.858	0.807	0.840	0.940	0.834	0.923	0.905	0.756	21	12
X ₁	0.896	1	0.916	0.218	0.640	0.903	0.968	0.979	0.901	0.907	0.785	0.872	0.885	0.979
X ₂	0.883	0.916	1	0.225	0.618	0.921	0.895	0.904	0.974	0.982	0.770	0.884	0.902	0.910
X ₃	0.115	0.218	0.225	1	-0.331	0.223	0.412	0.101	0.393	0.116	0.459	0.316	0.160	0.281
X ₄	0.858	0.640	0.618	-0.331	1	0.522	0.502	0.746	0.497	0.714	0.677	0.436	0.569	0.635
X ₁₂	0.807	0.903	0.921	0.223	0.522	1	0.908	0.904	0.927	0.924	0.689	0.987	0.992	0.936
X ₁₃	0.840	0.968	0.895	0.412	0.502	0.908	1	0.919	0.934	0.859	0.814	0.913	0.872	0.977
X ₁₄	0.940	0.979	0.904	0.101	0.746	0.904	0.919	1	0.861	0.931	0.791	0.855	0.909	0.973
X ₂₃	0.834	0.901	0.974	0.393	0.497	0.927	0.934	0.861	1	0.932	0.794	0.924	0.891	0.917
X ₂₄	0.923	0.907	0.982	0.116	0.714	0.924	0.859	0.931	0.932	1	0.773	0.871	0.926	0.912
X ₃₄	0.905	0.785	0.770	0.459	0.677	0.689	0.814	0.791	0.794	0.773	1	0.686	0.682	0.837
X ₁₂₃	0.756	0.872	0.884	0.316	0.436	0.987	0.913	0.855	0.924	0.871	0.686	1	0.968	0.921
X ₁₂₄	0.821	0.885	0.902	0.160	0.569	0.992	0.872	0.909	0.891	0.926	0.682	0.968	1	0.925
X ₁₃₄	0.912	0.979	0.910	0.281	0.635	0.936	0.977	0.973	0.917	0.912	0.837	0.921	0.925	1
Freq		3	2			2	2	2	1	1		2	2	3
Case		B	B			B	B	B	B	B		B	B	B
Action Taken	Identify variables with highest frequency with tie of Case B . Remove variable X ₁ with minimum V ₁ (0.896).													

After the first multicollinearity removal, the Pearson correlation coefficient matrix of model M31 then becomes model M31.1 as in Table 5. Table 5 shows the presence of multicollinearity source variables which are highlighted in yellow. However, it can be seen that variable X₁ is the next variable to be removed due to multicollinearity of the type Case B where there is a tie between variables X₁ and X₁₃₄. Further four iterations to remove the multicollinearity source variables of type Case B are carried until model then becomes model M31.5 as shown in Table 6 below. Table 6 shows the presence of multicollinearity of type Case C with frequency of 1.

Table 6 Correlation Coefficient Matrix of Model M31.5 (Case C)

M31 .5	V ₁	X ₃	X ₄	X ₁₂	X ₁₃	X ₁₄	X ₂₃	X ₂₄	X ₃₄	X ₁₂₄
V ₁	1	0.115	0.858	0.807	0.84	0.94	0.834	0.923	0.905	0.821
X ₃	0.115	1	-0.331	0.223	0.412	0.101	0.393	0.116	0.459	0.16
X ₄	0.858	-0.331	1	0.522	0.502	0.746	0.497	0.714	0.677	0.569
X ₁₂	0.807	0.223	0.522	1	0.908	0.904	0.927	0.924	0.689	0.992
X ₁₃	0.84	0.412	0.502	0.908	1	0.919	0.934	0.859	0.814	0.872
X ₁₄	0.94	0.101	0.746	0.904	0.919	1	0.861	0.931	0.791	0.909
X ₂₃	0.834	0.393	0.497	0.927	0.934	0.861	1	0.932	0.794	0.891
X ₂₄	0.923	0.116	0.714	0.924	0.859	0.931	0.932	1	0.773	0.926
X ₃₄	0.905	0.459	0.677	0.689	0.814	0.791	0.794	0.773	1	0.682
X ₁₂₄	0.821	0.160	0.569	0.992	0.872	0.909	0.891	0.926	0.682	1
Freq				1						1
Case				C						C
Action Taken	Identify variables with highest frequency of 1with tie of Case C. Remove variable X ₁₂ with minimum V ₁ (0.807).									

The techniques of multicollinearity removals based on the SPSS outputs are executed on model M31 until there is no more multicollinearity source variables present in the model. Consequently, there are six multicollinearity source variables with the different case types (Case A, Case B and Case C) respectively in all the 6 iterations, and is given by the sequence:

$$M31.0 \rightarrow M31.1 \rightarrow M31.2 \rightarrow M31.3 \rightarrow M31.4 \rightarrow M31.5 \rightarrow M31.6$$

Table 7 Final Correlation Coefficient Matrix of M31.6 Newton after 6 Iterations

M31.6	V ₁	X ₃	X ₄	X ₁₃	X ₁₄	X ₂₃	X ₂₄	X ₃₄	X ₁₂₄
V ₁	1	0.115	0.858	0.840	0.940	0.834	0.923	0.905	0.821
X ₃	0.115	1	-0.331	0.412	0.101	0.393	0.116	0.459	0.160
X ₄	0.858	-0.331	1	0.502	0.746	0.497	0.714	0.677	0.569
X ₁₃	0.840	0.412	0.502	1	0.919	0.934	0.859	0.814	0.872
X ₁₄	0.940	0.101	0.746	0.919	1	0.861	0.931	0.791	0.909
X ₂₃	0.834	0.393	0.497	0.934	0.861	1	0.932	0.794	0.891
X ₂₄	0.923	0.116	0.714	0.859	0.931	0.932	1	0.773	0.926
X ₃₄	0.905	0.459	0.677	0.814	0.791	0.794	0.773	1	0.682
X ₁₂₄	0.821	0.160	0.569	0.872	0.909	0.891	0.926	0.682	1

Table 7 shows that the final correlation coefficient matrix is without the presence of multicollinearity in the model since all the absolute coefficients are all less than 0.95. The multicollinearity remedial produres in SPSS outputs have thus been demonstrated.

Elimination of Insignificant Variables and Parameters Tests

Next after the multicollinearity source variables have been removed, the least important variable is sequentially removed via the Coefficient test, judged by the size of the equivalent F–statistic or p–value. The variable with the highest p–value is thus dropped, followed by the model is refitted and then the F–statistics recalculated. The process is repeated and the iterations will end when all of the p–values are less than 0.05. This implies that the remaining variables are significant to the model.

The Global F–Test and Coefficient Test are carried out on all the MR models. The elimination procedures of the Coefficient test had been illustrated by Noraini *et al.* [4]. The effects of the variables on the stem biomass for every regression model can be verified using the p–value from the ANOVA table. At 95% level of confidence, if $p < 0.05$, H_0 is rejected. Accepting H_1 implies that the remaining independent variables have significant effects on the volume of stem biomass.

Phase 3: Best Model via the Eight Selection Criteria (8SC)

Best model is one which fulfills the least value of most of the criteria as described in Ramanathan [10]. These 8SC’s have also been detailed in Table 3 by Noraini *et al.* [5]. The Wald test is used to test the removal and eliminated variables of the biomass regression model, by checking the effect of the omitted independent variables on the volumetric stem biomass. The 8SC is based on the (k+1) estimated parameters. The best model is found to be of model M26.2.2 of the Newton’s formula, chosen from the selected models based on the 8SC.

Phase 4: Residual Analyses using the Goodness-of-Fit

The goodness-of-fit tests comprises of the randomness test and normality test so as to ensure that the MR assumptions are not violated. Randomness test is based on the standardized residuals scatter plot so as to determine that the residuals are randomly distributed and normality test is on the Kolmogorov-Smirnov statistics for $n > 50$.

From the t–distribution table, the value of $|t_{\text{critical}}|$ is 1.980 at the significant value of 0.05. Since $|T_n| = 1.8597 \times 10^{-06}$ is less than $|t_{\text{critical}}| = 1.980$, therefore the null hypothesis is accepted at 5% significance level, and the residuals are then said to be at random. The Kolmogorov-Smirnov statistics gives a value of 0.0475 at 0.200 significance level ($p > 0.05$), hence, implying that the standardized residuals are significantly randomly and normally distributed as indicated in Table 8.

Figure 3 and Figure 4 show the scatter plot and histogram of the standardized residuals which hence support the evidence of the model's goodness-of-fit.

Table 8 Goodness-of-fit Tests on Standardized Residuals M26.2.2Newton

Randomness Test	$T_{\text{calculate}}$	$T_{\text{critical}}(0.025,130)$		Decision
	1.8597×10^{-06}	1.980		
Normality Test	Kolmogorov-Smirnov Statistics			
	Statistic	df	Sig.value	Normal
Standardized Residuals	0.0475	130	0.200	

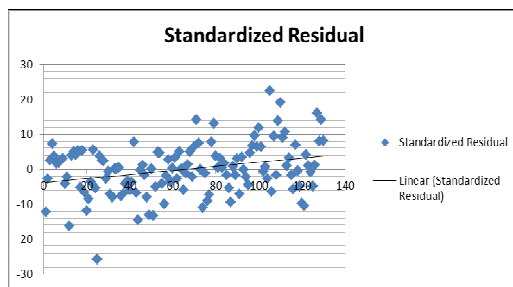


Figure 3 Standardized Residual Scatter Plot

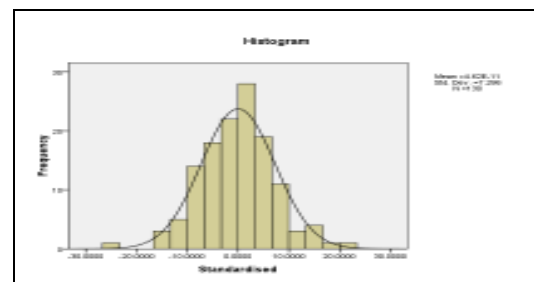


Figure 4 Histogram of Standardized Residuals

Table 8. Final Correlation Coefficient Matrix of Best Model M26.2.2Newton

Model M26.2.2 Newton	Unstandardized Coefficients		t	P-value
	B	Std. Error		
(Constant)	-0.227	0.023	-9.661	8.973×10^{-17}
3	-4.113	0.295	-13.960	3.989×10^{-27}
12	3.282	0.420	-7.807	2.181×10^{-12}
14	0.241	0.140	17.050	3.416×10^{-34}
23	-3.313	1.388	-2.388	0.018
24	0.323	0.024	13.363	1.035×10^{-25}
34	1.923	0.049	39.507	9.651×10^{-72}

From Table 8, the model equation for model M26.2.2Newton can be written as:-

$$M_2 \quad V_1 = -4 X_3 + 3 X_1 + 0 X_1 - 3 X_2 + 0 X_2 + 1 X_3 \quad (7)$$

Equation (7) shows that model M26.2.2 has a single independent variable (X_3) and five first order interactions (X_{12} , X_{14} , X_{23} , X_{24} and X_{34}). The other single independent variables (X_1 , X_2 and X_4) are insignificant, but contribute as interaction variables in the model. Substitution of the defined variables back into equation (7), equation (8) is then obtained:

$$M_{26.2.2} \quad V_N = -0.227 - 4.113 \frac{D_b}{h} + 3.282 D_t^{3.7} D_m^{4.5} + 0.241 D_t^{3.7} h^{4.4} - 3.313 D_m^{4.5} \frac{D_b}{h} + 0.323 D_m^{4.5} h^{4.4} + 1.923 D_b h^{3.4} \quad (8)$$

It can be seen from equation (8) that the generated variable constituting of the diameter at the base (D_b) has a positive direct relationship, while the stem height, h is inversely proportional to the stem volume. These variables (D_b and h) also contribute significantly as interactions towards the stem volume with the other transformed single independent variables. Hence, it can be said that variables, D_b and h , gave major contributions towards the stem volume in the model since only the generated variable remains as a single independent variable while other variables contribute significantly as interactions. The best model using the Newton's formula (M26.2.2) with the mensuration variables can then be used for prediction and estimation.

3 Discussions and Conclusions

Comparisons of models using 8SC have indicated that the Newton's formula gives a better estimation based on the least values of SSE. Remedial techniques in minimizing multicollinearity effects are applied to obtain a robust model where the best model has the least SSE of the 8SC and any insignificant variables would have been eliminated. Using the Newton's volume formula, the diameter at base (D_b) and the stem height (h) give major contributions while significant contributions are from the other independent variables to the volumetric stem biomass. It should be noted that the best model obtained is hence free from multicollinearity effects and also free from non-contributing independent variables.

References

- [1] Hasenauer, H. (eds.). Concepts within Tree Growth Modelling. *Sustainable Forest Management*. New York: Springer. 2006.
- [2] Wang, C. Biomass allometric equations for 10 co-occurring tree species in Chinese temperate forests. *Forest Ecology and Management*. 2006. 222: 9–16.
- [3] Onyekwelu, J.C. Above-ground biomass production and biomass equations for even-aged *Gmelina arborea* (ROXB) plantations in south-western Nigeria. *Biomass and Bioenergy*. 2004. 26: 39 – 46.
- [4] Noraini, A., Zainodin, H.J. & Nigel, J. J. B. 2008. Multiple regression models of the volumetric stem biomass. *WSEAS Transactions on Mathematics*. 2008. 7: 492-502.
- [5] Noraini, Abdullah, Amran Ahmed & Zainodin, H.J. An improved volumetric estimation using polynomial regression. *Journal of Science & Technology (UTHM)*. 2011. 3(2): 29-42.
- [6] Zainodin, H.J., Noraini A. & Yap, S.J. An alternative multicollinearity approach in solving multiple regression problem. *Trends in Applied Science Research*. 2011. 6(11): 1241-1255.
- [7] Fuwape, J.A., Onyekwelu, J.C. & Adekunle, V.A. J. Biomass equations and estimation for *gmelina arborea* and *nauclea diderrichii* stands in akure forest reserve. *Biomass and Bioenergy*. 2001. 21: 401-405.
- [8] Husch, B., Beers, T.W. & Kershaw, J. A. Jr. *Forest Mensuration*. 4th Edition. New York: John Wiley & Sons. 2003.
- [9] Brack, C. Updating urban forest inventories: An example of the DISMUT model. *Urban Forestry & Urban Greening*. 2006. 5: 189-194.
- [10] Ramanathan, R. *Introductory Econometrics with Applications*. 5th Ed. Ohio: South-Western Thomson Learning. 2002.
- [11] Devore, J. L. & Peck, R. *Statistics. The Exploration and Analysis of Data*. 6th Edition. USA: Wadsworth, California. 1993.
- [12] Gujarati, D.N. *Basic Econometrics*. 3rd Ed. USA: Mac-Graw Hill Inc. 1995.