World Applied Sciences Journal 14 (10): 1599-1606, 2011
ISSN 1818-4952

The Basic of Analytical of Simple Linear Regression in Forestry Studies (Case Study: Relationship Between Basal Area and Tree Coverage of *Quercus brantii* Lindl. In Absardeh, Chahar Mahale and Bakhtiari)

[1]*S. Kahyani,* [2]*S.M. Hosseini and* [3]*R. Basiri*

[1]Faculty of Natural Resources and Marine Sciences, Tarbiat Modares University,
Noor, Mazandaran Province, I.R. Iran, P.O. Box 46414-356, Tehran, Iran
[2]Department of Forestry, Faculty of Natural Resources and Marine Sciences, Tarbiyat Modares University,
Noor, Mazandaran Province, I.R. Iran. P.O. Box 46414-356, Tehran, Iran
[3]Department of Forestry, Faculty of Natural Resources, Higher Education Complex of Behbahan, I.R. Iran

Abstract: Although there are many univariate techniques for data analysis in statistics, none of them takes into account the effects of other variables. In such a case regressional models are being used. Linear regression is the most common method of studying the linear relation between two or more variables. The regression presumptions' must be accurately considered to make reliable results of relating two variables and finding the best models. Without considering the presumptions some problems may occur. The purpose of this research is to show the correct model construction, select the best kind of model and validation for a simple linear regression with emphasis on its presumptions. In this paper the regression presumptions specially in forestry studies is arrested, because forestry studies use regression tests widely and thus the accuracy of results is completely needed. The most influence of the nullity of presumptions is the biased variance estimation, regression coefficients and coefficient of determination, also is the biased tests hypothesis and interval estimation.

Key words: Simple linear regression · Regression hypothesis · Forestry studies

INTRODUCTION

Although there are many univariate techniques for data analysis in statistics, none of them takes into account the effects of other variables. In such a case regressional models are being used. Regression analyses are using in order to Data description (Descriptive of amount of dependency between a variable and favorable factors), Prediction, Parameters estimation and Control [1]. Regression analysis mainly used to predict dependent variable values by the values of one or more independent variables [2]. In studying of linear relationship between two or much variables, linear Regression are used as the widest model [2]. The current model under using in linear Regression are: $y_i = \alpha + \beta x_i + (\varepsilon)$, That α is constant equation, β is line gradient, x_i and y_i are independent and dependent variables respectively and (ε) is regression error [3]. In order to investigate the relationship between tree cover variables in forest ecology studies, statistical regression method are used. In ecology, the regression models mainly are used for the discovery and presentation of descriptive as possible exact of relationship between variables [4]. Review of research, revealed the widespread use of simple linear regression [5-8]. In Forest ecology studies, the structural relationship between upper and low floors has been recognized important [9]. Two criteria, forest canopy [10-12] and Basal area [13] have been recognized as the common criteria for upper forest floor. Tree canopy and basal area are considered as the two important variables in most regression studies [9, 14]. The regression presumptions must be accurately considered to make reliable results of relating two variables and finding the best models. Without considering the presumptions some problems may occur. Because the effect of violation of one or more hypothesis are not visible in the final model [7].

Corresponding Author: Saleh Kahyani: M.Sc. Graduate of Forestry, Faculty of Natural Resources and Marine Sciences, Tarbiat Modares University, Noor, Mazandaran Province, I.R. Iran. P.O. Box 46414-356, Tehran, Iran. Tel.: +98 913 7025381, Fax: +98 122 6253499.

Emphasized hypotheses on simple linear regression in most statistical contexts have discussed in detail [15-18]. There are five hypotheses in relation with regression [19]. These hypotheses are:

- Model is correct, if the dependent variable (Y) have linear relationship with the independent variable (x).
- Data that used for fitting the model is a representative of the desired data.
- Errors variance is constant (is homogeneous) In other words not related to the independent variable (x) or other variables such as time.
- Errors are independent of each other.
- Errors have normal distribution. Necessary to meet these assumptions is determined based on goal of performing regression [19]. Study at the individual level (tree individuals of a species) has done on the Persian oak (*Quercus brantii* Lindl.). The purpose of this research is comparison of regression analysis method with consideration it's hypothesis and without consideration it's hypothesis for studies of forest ecology.

MATERIALS AND METHODS

Study Area: Absardeh forest is located in southern Zagros Mountains, 30 km from Ardal city, Chaharmahale and Bakhtiari province, Iran (31°50' to 31°57' N, 50°25' to 502°30' E and elevation 1800-2450 meter at sea level).

Research Methodology

Methodology for Determining the Regression Model Is Included:

- Verification of simple linear regression model assumptions:
- Determination of validation method of simple linear regression model:

In engineering jobs in what is common is that one-third of data are excluded to validation and two-thirds of data that remain are used for modeling [20]. In which case, it is called Threefold cross-validation. Various studies have shown that the number of classes and data grouping and its impact on value prediction errors, first, have been different. For example, some tenfold cross-validation [3], some fivefold cross-validation [3, 21] and some lack of difference between the above three types of validation, have emphasized [22]. Secondly, it has also relation to be large or small data's series. In this paper, pay attention to the set of conditions, the threefold cross-validation method was used. In this method, first, have considered whole data's on view of to be normal until consider being parametric or nonparametric regression method [23]. Then data series were randomly divided into 3 groups and repeat all the steps of determination of regression model in any order, were as follows.

Outlier Analysis: Outlier and Limit values have so great importance in the regression analysis [14]. Regression line as the average conditions is sensitive to the presence of outliers as the average specimens than it is sensitive. In this stage, multivariate outlier detection were used for investigate of outliers. In this way, mahalonobis distance square is calculated for each data through the following mathematical relationship and its significant is also clear [2, 24].

$$T_i^2 = (n-1)(X_i - \bar{X})^t \left[(X - \bar{X})^t (X - \bar{X}) \right]^{-1} (X_i - \bar{X})$$

Regression Hypothesis Testing:

- If the dependent variable (Y) have linear relationship with the independent variable (x), the model form is correct. For check this hypothesis, several methods were used. A method was using of scatter plot. Form of scatter plot is marker of the presence or absence of linear relationship [15, 19]. The second method was test of the linear relationship hypothesis that tested through the correlation significant test (t test) [12, 25]. If this coefficient is significant, meaning there is a linear relationship. The third method to test this hypothesis is drawing the error values versus estimated values ($\hat{y}$). If distribution of points has a specific trend, this hypothesis is not correct and if data have not a specific trend, this hypothesis can be accessed [15].
- Data that used for fitting the model is a representative of the desired data.
- Errors variance is constant (is homogeneous) In other words not related to the independent variable (x) or other variables such as time. To test this hypothesis, charting residuals versus independent variable were used. Lack of specific trends in the distribution of points is marker that variance of errors is constant [15].
- Errors are independent of each other. To confirm this hypothesis Durbin-Watson test was used.
- Errors have normal distribution. To confirm this assumption Kolmogorov - Smirnov test was used.

Determination of Regression Model: After reviewing the assumptions of regression, regression procedures between two variables basal area and the percentage of cross section surface of tree coverage cross section was performed and the relevant regression equation was obtained. There is two hypothesis testing in relation to simple linear regression [19].

- Significant test of independent variable coefficient (β).
- Significant test of constant coefficient (α).

Thus the above process, three regression equations were obtained that all three were validation.

In order to validation of models, estimated value of dependent variable for data that was not used in the modeling was obtained by the appropriate model, then the estimated values and observed values were compared by paired t-test [22]. If a significant difference between estimated values and observation not be seen, validated models are confirmed, otherwise, the model is removed and then for comparison are not.

Criteria Used to Validation and Selection Model: To validation and selection the best model, the various criteria of coefficient of determination and error were used [1, 26]. These criteria are shown in table 1.

Above criteria for two parts, modeling and validation were calculated. Criteria of validation section should be close to values of their equivalent criteria in modeling section until model validation is approved.

Selection the Best Model: For selecting the best model, usually below Criteria are used [15, 22, 25].

- Coefficient of determination of model (R^2)
- The Mean of sum of squares error (MSE)

In addition to the above criteria, R_P^2 and *PRESS* was reported. The important point that a paired t-test has been done for mean comparison of fitted values to the data with equivalent observed values of them. In order to qualitative assessment of model, equality of variance of observational data collection and estimated data were used that this operation was done through the F test [27]. Whatever the coefficient of determination is higher and errors is less, model can further estimate the amount of dependent variable and that model recommended as the best model.

Data and Experiment Condition: Total 46 plots with area of 20 R. and a rectangular shape with regard to error about 15% were measured. In each plot, the variables DBH (cm) and two perpendicular diameters of canopy cover (m) were measured. Then cross-section surface of trunk at the breast height (basal area) and cross sections canopy cover for each tree was calculated. Data (plots) according to Threefold cross-validation Method, were divided into 3 groups until be used for selection and validation model. We tried, the data randomly and equal be divided into three groups randomly and equal. First and second groups, each containing 15 data, the third group contains 16 data and generally be there 46 data. Dividing data into three groups had been done due to ensure validation data set, Because of according to Montgomery *et al*. [1] minimum number of data to evaluate the estimated error is 15.

RESULTS

Determination of Regression model in three stages was preformed based on threefold cross-validation method. For example, in the first stage, the combination of first and second groups were used for modeling and third

Table 1: Computed criteria for selection and validation of regression models

$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$
$MSE = \frac{SSE}{n}$	$R_P^2 = 1 - \frac{PRESS}{SST}$
$RMSE = \sqrt{MSE}$	$MAE = \frac{1}{n} \sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$
$MAPE = \left(\frac{1}{n} \sum_{i=1}^{n} \left\lvert \frac{y_i - \hat{y}_i}{y_i} \right\rvert \right) \times 100$	$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Table 2: Results of Shapiro-Wilk normality test for two variable of dependant and independent

Shapiro-Wilk Test	Dependent Variable (Cross Section Cover(m^2))	Independent Variable (Basal Area(cm))
Group I (sets 1,2)	$Z_{0/05,30} = 0/67, P > 0/05$	$Z_{0/05,30} = 1/07, P > 0/05$
Group II (sets 1,3)	$Z_{0/05,31} = 0/86, P > 0/05$	$Z_{0/05,31} = 1/45, P > 0/05$
Group III (sets 2,3)	$Z_{0/05,31} = 0/87, P > 0/05$	$Z_{0/05,31} = 1/27, P > 0/05$

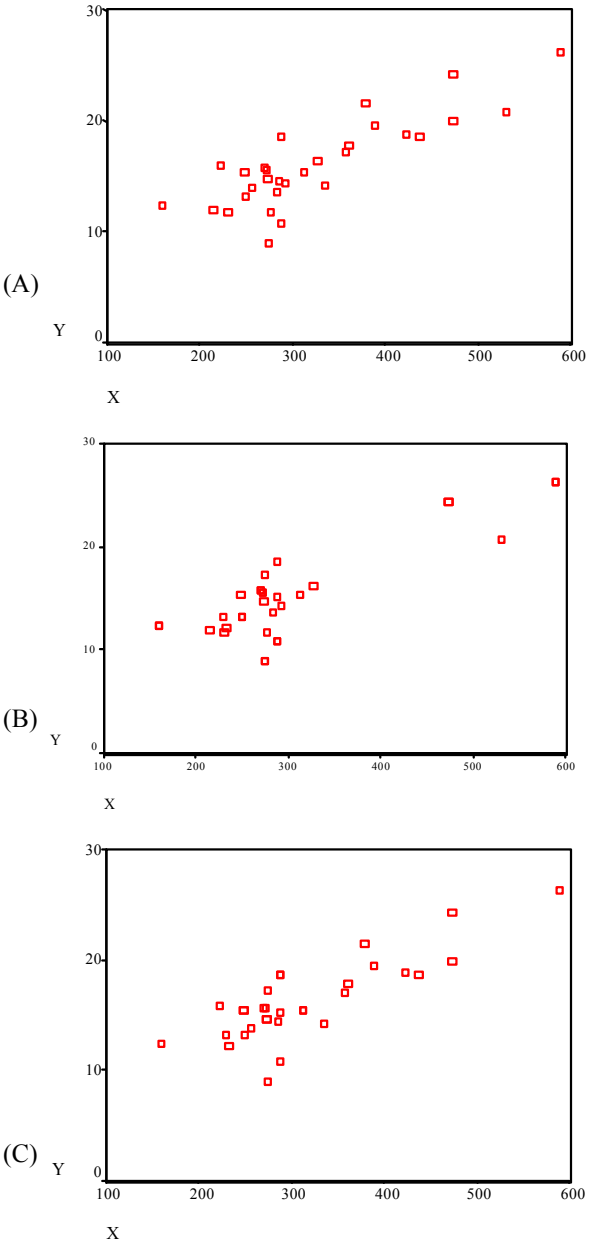

Fig. 1: Scatter plot of Independent Variable x: (Basal Area(cm)) and Dependent Variable y: (Cross Section Cover(m2)) to check the linear relationship-
A) Group I (sets 1,2), B (Group II (sets 1,3), C) Group III (sets 2,3)

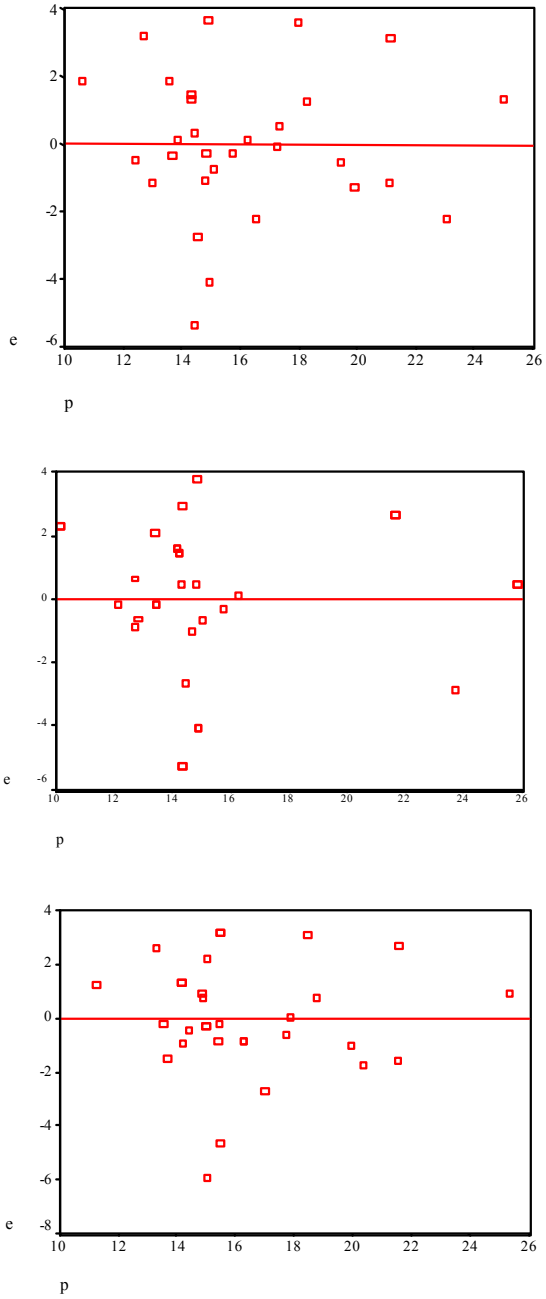

Fig. 2: Chart of error versus estimated values to check the linear relationship - A) Group I (sets 1,2), B (Group II (sets 1,3), C) Group III (sets 2,3)

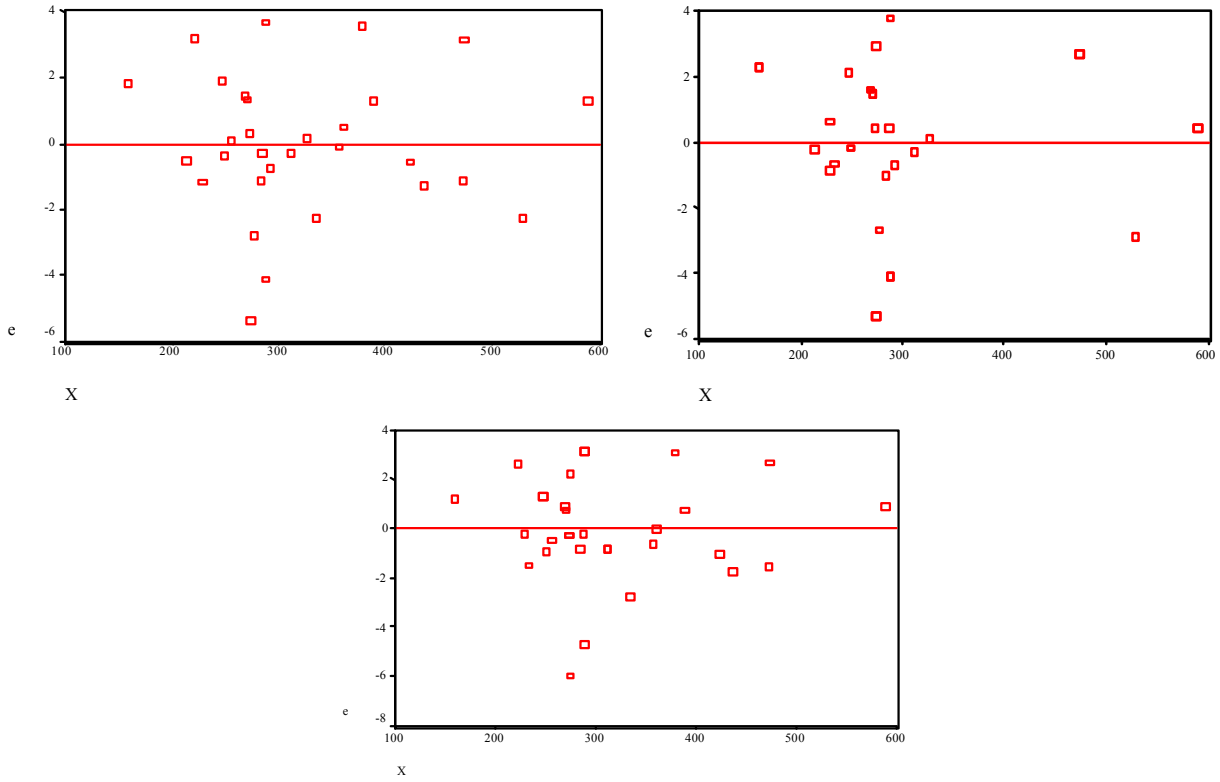

Fig. 3: Chart of error versus independent variable (Basal Area(cm)) To check homogeneity of variance of errors A) Group I (sets 1,2), B (Group II (sets 1,3), C) Group III (sets 2,3)

group used to validate. In the second stage, the combination of first and third groups were used for modeling and second group used to validate. In the third stage, the combination of second and third groups were used for modeling and first group used to validate. The results of the three groups follow:

With doing normality test, it was found that both variable data for three groups follow the normal distribution.

- With doing normality test become distinguished that data's of two variable follow for three groups of normal distribution (Table 2).
- Results analysis flung areas through the Mahalanobis square distance calculation method, as follows. A point from the first group and third group, four points for Casewise was diagnosed and for the next stage were removed.

Results of Regression Hypothesis: According to Figures 1 and 2, the independent and dependent variables in all three groups have a linear relationship. Figure 2 for all three groups revealed the distribution of points is not a linear process. Results of a significant correlation coefficient test for all three groups, confirms linear relationship.

First group ($r = 0.88$, $p < 0.001$), Second group ($r = 0.86$, $p < 0.001$),
Third group ($r = 0.81$, $p < 0.001$).

Run the second hypothesis (Data used for model fitting is representative of the desired data). In this study, random sampling was used. For the implementation of this hypothesis samples were randomly selected and data collection in three groups was classified as random.

Stability or homogeneity of variance of from Figure 3, showed that such data do not have a specific process has been fulfilled in this case.

Hypothesis of independence of errors was investigated through Durbin-Watson test and independence of errors was proved for the three groups.

Table 3: Regression model with test of model coefficients

		Third group	Second group	First group
T test	Constant coefficient	$t = 6.51, P < 0.001$	$t = 3.33, P < 0.002$	$t = 3.73, P < 0.001$
	Variable coefficient	$t = 6.13, P < 0.001$	$t = 9.04, P < 0.001$	$t = 6.67, P < 0.001$
Model		$y = 8.4+0.03x$	$y = 4.34+0.04x$	$y = 5.81+0.03x$

Table 4: Paired sample t test for evaluation of credit models

Third group	Second group	First group
$t_{0.05,13} = -0.55$, $P > 0.05$	$t_{0.05,14} = -1.18$, $P > 0.05$	$t_{0.05,15} = 1.59$, $P > 0.05$

Table 5: A summary of statistics derived from models obtained

Models	R^2(%)	R_P^2(%)	*MSE*	*PRESS*	*t* test (value*P*)	*F* test (value*P*)
Group I	62	54.2	4.5	149.2	0.99	0.27
Group II	74	70.7	5.5	189.1	0.99	0.31
Group III	60	54.6	2.2	67.8	0.99	0.13

Table 6: Summary of statistics of estimation error for the models obtained

Models	*MAE*	*SSE*	*RMSE*	*MAPE*	*t* test (value*P*)	test *F* (value*P*)
Group I	1.9	77.1	2.4	11	0.13	0.13
Group II	1.8	65.8	2.1	13.6	0.26	0.37
Group III	1.6	74.1	2.2	7.2	0.11	0.30

Table 7: Ranked statistics values from model

Models	R^2(%)	R_P^2(%)	*MSE*	*PRESS*	Ranks Mean	Ranks Mean
Group I	2	1	2	2	7	1.75
Group II	3	2	1	1	7	1.75
Group III	1	3	3	3	10	2.5

Table 8: Ranked values of estimation error statistics

Models	*MAE*	*SSE*	*RMSE*	*MAPE*	Ranks Mean	Ranks Mean
Group I	1	1	1	2	5	1.25
Group II	2	3	3	1	9	2.25
Group III	3	2	2	3	10	2.5

Third group that have this model: $y = 8.4 + 0.03x$, $R^2 = 0.60$, $p < 0/.0001$, Selected as the best model.

(First group: $Dw_{0.05} = 1.74$, $p<0.05$),
(Second group: $Dw_{0.05} = 1.84$, $p<0.05$),
(Third group: $Dw_{0.05} = 1.80$, $p<0.05$).

Investigation of Hypothesis of normal errors for all three groups confirmed this assumption.

(First group: $Z_{0.05} = 0.61$, $p>0.05$),
(Second group: $Z_{0.05} = 0.82$, $p>0.05$),
(Third group: $Z_{0.05} = 0.68$, $p>0.05$).

Simple linear regression model and validate them for different groups have come in Tables 3 and 4.

Selection the Best Model: In order to select the best simple linear regression model between the two variables In order to select the best simple linear regression model between the two variable Basal area (independent) and variable percentage of tree canopy (dependent), a summary of fit statistics are presented in Table 5.

Finally, statistics of estimation error for the models obtained for each group are presented in Table 6.

Based on Tables 7 and 8, each model offers a rank for itself. For the highest values of error, the lowest rating and for the highest values of determination coefficients, the highest rank was allocated. Best model, based on higher average rank was selected.

DISCUSSION AND CONCLUSION

In Statistics for each type of hypothetical pre-test conditions provided that users are required to observe them. Lack of attention to the pre-assumptions, will violation of their potential difference and finally this issue influencing the characteristics of regression equation [7]. The most influence of the nullity of presumptions is the biased variance estimation, regression coefficients and coefficient of determination, also is the biased tests hypothesis, standard error of regression coefficient and interval estimation. In such circumstances the amount of estimated regression coefficients are correct, but assumption tests and estimated distance is not true. Model efficiency comes down when the oblique occurs [7] this means that the results taken from the label violated the hypothesis is pressed [24]. Therefore, the analyst should first be inform the assumptions of regression and then be able to implement these assumptions before the regression to test. There are a series of different tests to determine significant regression assumptions [7, 15, 24, 28]. Order of applying these tests is important because it violated one of the hypotheses may be the next test to discredit [7]. Appropriate tests for regression assumptions in this paper are recommended. For determine the accuracy of the first assumption of regression (linear relationship between two variables) different methods including charts and hypothesis testing are presented. The concept of correlation is built based on the linear relationship, therefore, as a precise method can be used for this work [7, 15, 24, 28]. If you violate this assumption, regression coefficient firstly, gives non-oblique estimation of real values and predicted values Secondly among other coefficients will not have lower variance, predicted values will not follow the normal distribution and standard error of regression coefficients are developed oblique [7]. Results of correlation in this study ($r = 75$, $P<0.001$) indicated that there was significant linear relationship between two variables. Violations of the second regression assumption (data used for model fitting is representative of the desired data) cause oblique problem in regression coefficients is the biased regression coefficients and correlation coefficient, also is the biased tests hypothesis and interval estimation [28]. Random selection of data and random data to three sets of evidence on this assumption is correct. Violations of the third regression assumption (be constant error variance) are result in lack of unbiasedness regression coefficients but the other, standard error of regression coefficients will not be correct [7, 28]. In this paper, homogeneity of variance test and the proposed diagram Neter *et al*. [29] (diagram of remains against independent variables), Both of them are confirmed the accuracy of constant of variance. Violations of the fourth regression assumption (Errors is independent of each other) and violations of fifth regression assumption (Normal distribution of errors) cause low efficiency of regression coefficients (variance of estimated regression coefficients does not reaches the minimum value) assumption tests of estimation interval will lose your credit [18]. Of course, violations of the fifth assumption if sample size is large cannot create serious problem [7]. Errors in this study had a normal distribution. In this study, none of the regression assumptions were not violated. Thus assumption tests, estimate of coefficients, standard error of estimate of coefficients, correlation coefficients and determine and estimated interval of Submitted model are validated.

ACKNOWLEDGEMENT

We thank from all the organs of natural resources of Chaharmahal & Bakhtiari province, Province Environmental Protection Bureau, Department of Agriculture and Higher Education Complex of Behbahan and Department of Forestry of Tarbiat Modares University that with special consideration helped us until this research project to be done better.

REFERENCES

1. Montgomery, D.C., E.A. Peck and G.G. Vining, 2001. Introduction to linear regression Analysis, 3rd edn., Wiley, New York, NY.
2. Reimann, C., P. Filzmoser, R.G. Garrett and R. Dutter, 2008. Statistical Data Analysis Explained, John Wiley & Sons, Ltd. ISBN: 978-0-470-98581-6
3. Zhang, P., 1993. Model selection via multifold cross-validation. Annals of Statistics, 21(1): 299-311.
4. Longman, R.H.C., C.J.F. Ter. break and O.F.F. Van Tongeren, 1987. Data analysis in Community and Landscape ecology. C. Ombridge university press, Wageningen, pp: 299.
5. Williams, M. and T. Gregoire, 1993. Estimating weights when fitting linear regression models for tree volume. Department of forestry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0324, USA, 1725-1731.
6 Williams, M.S. and H.T. Schreuder, 1996. Prediction of gross tree volume using regression models with non-normal error distributions. J. Forest Sci., 42(4): 419-430.

7. Marshall, P., T. Szikszai, V. Le May, and A. Kozak, 1995. Testing the distributional assumptions of least squares Linear regression. J. The Forestry Chronicle, 2(71): 213-218.
8. Williams, M.S., 1997. A regression technique accounting for heteroscedastic and asymmetric errors. J. Agricultural, Biological and Environmental Statistics, 2(1): 108-129.
9. Mitchell, J.E. and S.J. Popovich, 1996. Effectiveness of basal area for estimating canopy cover of ponderosa pine. J. Forest Ecology and Management, 95: 45-51.
10. Mitchell, J.E. and P.N.S. Bartling, 1991. Comparison of linear and nonlinear overstory-understory models for ponderosa pine. J. Forest Ecology and Management, 42: 195-204.
11. Bunnell, F.L. and D.J. Vales, 1989. Comparison of methods for estimating forest overstory cover: differences among techniques. Canadian J. Forest Res., 20: 101-107.
12. Barbour, M.G., J.H. Burk, W.D. Pitts, F.S. Gilliam and M.W. Schwartz, 1999. Terrestrial Plant Ecology (3th edition), An important of Addison Wesley Longman Incorporation, pp: 649.
13. Uresk, D.W. and K.E. Severson, 1989. Understory-overstory relationships in ponderosa pine forests, Black hills, south Dakota. J. Range Management, 42: 203-208.
14. Philip, M.S., 1994. Measuring Trees and Forests (2th Edition), CAB International, pp: 310.
15. Dowdy, Sh., S. Weardon and D. Chilko, 2004. Statistics for research. john Wiley & Sons, Inc., pp: 241-260.
16. Chan, Y.H., 2004. Biostatistics 201:Linear Regression Analysis, Basic Statistics For Doctors. Singapore Med. J., 45(2): 55-61.
17. Chatterjee, S., and A.S. Hadi, 1988. Sensitivity analysis in linear regression. John Wiley & Sons, New York, NY, pp: 315.
18. Kmenta, J., 1986. Elements of econometrics. Second Edition. Macmillan Publishing Company, New York, NY., pp: 786.
19. Helsel, D.R. and R.M. Hirsch, 2002. Statistical Methods in Water Resources. USGS publication, USA, pp: 510.
20. Witten, I.H. and E. Frank, 2000. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, San Francisco, CA.
21. Breiman, L. and P. Spector, 1992. Submodel selection and evaluation in regression: the X-random case. International statistics review, 60(3): 291-319.
22. Feng, C.X., Z.G. Yu and A. Kusiac, 2006. Selection and validation of predictive regression and neural network models based on designed experiments. IIE Transactions, 38: 13-23.
23. Dytham, C., 1999. Choosing and using statistics, Blackwell Science Ltd.
24. Hair, J.F., R.E. Anderson, R.L. Tatham and W.C. Black, 1995. Multivariate data analysis (4th edition). Prentice Hall, upper Saddle River, New Jersey, USA, pp: 745.
25. Zar, J.H., 1999. Biostatistical analysis(4^{th} Ed). Prentice HALL International. Inc., USA. pp: 993.
26. Draper, N.R., and H. Smith,1998. Applied Regression Analysis, 3^{rd} edn., Wiley, New York, NY.
27 Feng, C.X., Yu. Z. and J.H. Wang, 2004. validation and data splitting in predictive regression modeling of honing surface roughness data. International J. Production Res., 43(8): 1555-1571.
28. Cohen, J., P. Cohen, S.G. West and L.S. Aiken, 2003.Applied multiple regression/correlation analysis for the behavioral sciences 3rd edn. Lawrence Erlbaum Associates, Publishers. New Jersey, London.
29. Neter, J. and W. Wasserman, 1974. Applied linear statistical models. Richard D. Irwin, Inc., Homewood, IL, pp: 842.