ISSN : 1812-5379 (Print)
ISSN : 1812-5417 (Online)
http://ansijournals.com/ja

JOURNAL OF AGRONOMY

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Journal of Agronomy 5 (1): 32-36, 2006
ISSN 1812-5379

Modeling Oil Palm Yield Using Multiple Linear Regression and Robust M-regression

[1]Azme Khamis, [2]Zuhaimy Ismail, [3]Khalid Haron and [3]Ahmad Tarmizi Mohammed
[1]Science Studies Centre, Kolej Univesiti Teknologi Tun Hussein Onn, Malaysia
[2]Department of Mathematic, University Technology of Malaysia, Malaysia
[3]Malaysian Palm Oil Board, Malaysia

Abstract: This study shows how a multiple linear regression model can be used to model palm oil yield. The methods are illustrated by examining the time series data of foliar nutrient compositions as one of the independent variable and fresh fruit bunch as dependent variable. Other independent variables include the nutrient balance ratio and major nutrient composition. This modeling approach is capable of identifying the significant contribution of each independent variable in the improving the modeling performance. We find that the quantile-quantile plot demonstrates the existing of outlier and this directs us to use robust M-regression for removing the negative impact of outliers. Results show that robust regression in this case gives a better results than conventional regression in modeling oil palm yield.

Key words: Multiple linear regression, roubust M-regression, oil plam yield, outliers data

INTRODUCTION

Multiple linear regression is being used widely in research to determine the linear relationship between factors (Draper and Smith, 1981; Birkes and Dodge, 1993; Mokhtar, 1994). Earlier researchers believed that foliar nutrient composition had a significant correlation with oil palm yield, but the data used had not been analyzed in detail using the proposed method (Green, 1976; Muhammad *et al*., 1991; Oboh and Fakorede, 1999; Foong, 1999; Chin, 2002; Foster, 2003). The multiple linear regression method is considered as one way to understand the relationship between foliar nutrient composition and oil palm yield. In this study, the use of the foliar Nutrient Balance Ratio (NBR) in leaves is proposed as the independent variable. This modeling approach acts as the preliminary study which will further enhance the understanding of the issues and continue to motivate research in the modeling of oil palm yield.

MATERIALS AND METHODS

This study will focus our attention on the relationship between the oil palm yield and the independent variables and the error existed in the data set. Two models discussed here are the Multiple Linear Regression (MLR) and Robust M-Regression (RMR).

Multiple linear regression: Consider that the data consists of n observations on a dependent or endogenous variable y and five independent or exogenous variables N, P, K, Ca and Mg. The relationship between the dependent and independent variables is formulated as a linear model,

$$y_i = \theta_0+\theta_1 N_{1i}+\theta_2 P_{2i}+\theta_3 K_{3i}+\theta_4 Ca_{4i}+\theta_5 Mg_{5i}+\varepsilon_I \quad (1)$$

Where, θ_0, θ_1, θ_2, θ_3,θ_4, θ_5 are the regression coefficients and ε_i is the random disturbance. It is assumed that, for any set of fixed values of independent variables that fall within the range of the data, the linear Eq. 1 provides an acceptable approximation of the true relationship between the dependent and independent variables. The least square estimate $\hat{\theta}$ of θ minimizes the quadratic cost function;

$$J = \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n}\left(y_i - x_i^T\hat{\theta}\right)^2 \quad (2)$$

Equation 2 can be written in matrix notation as

$$J = (y - x\hat{\theta})^T(y - x\hat{\theta}) \quad (3)$$

The coefficient vector that minimizes this cost function can be determined by setting the first partial derivative of the cost function to zero and solving the normal equation. Then the equation will be:

$$\hat{\theta} = (x^T x)^{-1}x^T y \quad (4)$$

Let us say we are interested in calculating the mean response of y for the fixed value of x_r. So we have to

Corresponding Author: Azme Khamis, Science Studies Centre, Kolej Universiti Teknologi Tun Hussein Onn, Malaysia

calculate the mean and the variance of the mean response. Let $\hat{y}_r$ be the unbiased estimator of y_r. The expectation of the mean response is given as

$$E(\hat{y}_r) = E(x_r^T\hat{\theta}) = x_r^T\theta \quad (5)$$

and the variance of the mean response is:

$$Var(\hat{y}_r) = \sigma^2 x_r^T(x^Tx)^{-1}x_r \quad (6)$$

The Nutrient Balance Ratio (NBR) in the foliar composition to the FFB yield is crucial because it will reflect on the uptake of nutrients mechanism (Foster, 2003; Fairhurst and Mutert, 1999). Thus we proposed to use NBR, Critical Leaf Phosphorus (CLP), deficiency of K and deficiency of Mg as the independent variables to estimate the FFB yield production using multiple regression analysis. Then considered the regression equation as follows;

$$\begin{aligned} y_i = {} & \theta_0+\theta_1N_{1i}+\theta_2P_{2i}+\theta_3K_{3i}+\theta_4Ca_{4i}+\theta_5Mg_{5i}+\theta_6N\text{-}P_{6i}+\theta_7N\text{-}K_{7i}+\theta_8N\text{-}Ca_{8i}+\theta_9N\text{-}Mg_{9i}+\theta_{10}P\text{-}K_{10i}+\theta_{11}P\text{-}Ca_{11i}+\theta_{12}P\text{-}Mg_{12i}+\theta_{13}K\text{-}Ca_{13i}+\theta_{14}K\text{-}Mg_{14i}+ \quad \theta_{15}Ca\text{-}Mg_{15i}+ \\ & \theta_{16}defK_{16i}+\theta_{17}\,defMg_{17i}+\theta_{18}CLP_{18i}+\theta_{19}TLB_{19i}+\varepsilon_I \end{aligned} \quad (7)$$

Where, θ_0, θ_1, …, θ_{19} are the regression coefficients and ε_i is the random disturbance, N-P is the ratio between N and P, N-K is between N and K and so on. defK is the deficiency of K in the leaf, defMg is deficiency of Mg in the leaf, CLP is Critical Leaf of P concentration and TLB is the Total Amount of Bases in the Leaf. We applied the stepwise procedure to select the significant independent variables in the regression model (Norušis, 1998).

Robust M-Regression: Outliers in a set of data that will influence the modelling accuracy as well as the estimated parameters, especially in statistical analysis as discussed by many reseasrchers (Birkes and Dodge, 1993; Mokhtar, 1994; Hampel, 1974; Andrew, 1974; Rousseeuw and Leroy, 1987; Barnett and Lewis, 1995; Khamis and Abdullah, 2004). A statistical procedure is regarded as robust if it performs reasonably well even when the assumptions of the statistical model are not true. If it is assumed that data follows a standard linear regression model, then the least squares estimates and tests perform quite well, but they are not robust when the outlier is present in the data set. In this case we are interested in using robust M-regression to model the yield data, since the quantile-quantile plot shows the existence of outliers. Peter Huber introduced the idea of M-regression in 1964.

In least squares estimation, values of $\hat{\beta}$ are chosen so that $\sum\hat{e}_i^2$ is as small as possible. In least absolute deviation estimation, values are chosen so that $\sum|\hat{e}_i|$ is as small as possible. In robust M-regression, this idea is generalized and values of $\hat{\beta}$ are chosen so that $\sum\rho(\hat{e}_i)$ is as small as possible, where, $\rho(e)$ is some function of e. In this way, least squares estimation and least absolute deviation estimation can be regarded as the special cases of M-estimation where $\rho(e) = e^2$ and $\rho(e) = |e|$, respectively.

Huber (1973) defined the function of $\rho(e)$ as follows;

$$\rho(e) = \begin{cases} e^2 & \text{if} \quad -k \le e \le k \\ 2k|e|-k^2 & \text{if} \quad e<k \quad \text{or} \quad k<e \end{cases} \quad (8)$$

Following a suggestion of Huber's, take $k = 1.5\,\hat{\sigma}$, where, $\hat{\sigma}$ is an estimate of the standard deviation σ of the population of random errors. In order to ensure that $\rho(e)$ is a smooth function, $2k|e| - k^2$ is used instead of $|e|$. $\hat{\sigma} = 1.483(MAD)$ can also be used, where MAD is the median absolute deviations $|\hat{e}|$. The multiplier 1.483 is chosen to ensure that $\hat{\sigma}$ would be a good estimate of σ if it were the case that the distribution of the random errors was normal.

The Huber M-estimates of $\hat{\beta}$ are the values of b that minimize

$$\sum\rho(y_i-(b_0+b_1x_{i1}+\ldots+b_px_{ip})) \quad (9)$$

Where, $\rho(e)$ is the function defined in Eq. 8.

It is convenient in this case to use vector notation. The vector $\hat{\beta}$ of Huber M-estimates is defined to be the vector b that minimizes $\sum \rho(Y_i - b'x_i)$. The vector of regression coefficients, denoted by β, is first estimated by the vector of least-squares estimates. This initial estimate of β is used to calculate deviations and an estimate of σ. The algorithm is iterated in this way until a step is reached at which the improved estimate of β is the same value (or at least approximately the same value) as the previous estimate.

For example, at any step in the algorithm, let b^0 be the current estimate of β. Calculate the deviation $y_i-(b^0)'x_i$ and from this calculate $\hat{\sigma} = 1.483(MAD)$ where MAD is median absolute deviation. The y values must be adjusted to remove any large deviations. The deviation of y_i from the current estimated regression line is $e_i^0 = y_i - (b^0)'x_i$. Therefore, $y_i = (b^0)'x_i+e_i^0$, We now define $y_i^* = (b^0)'x_i+e_i^*$, where, e_i^* is the adjusted deviation obtained by truncating e_i^0 so that none of the deviations are larger than $1.5\,\hat{\sigma}$ in the absolute value. Let the improved estimate of β be the least squares estimate obtained from the adjusted data y_1^*, y_2^*, …y_n^*. Huber M-estimates are obtainable from the S-Plus package (Becker *et al*., 1988).

Data and scope: The Malaysian Palm Oil Board (MPOB) provided us with a data set taken from two of the estates in Peninsular Malaysia. The factors included in the data

set were foliar nutrient composition and Fresh Fruit Bunches (FFB) yield. The variables in foliar nutrient composition included the percentage of nitrogen concentration (N), the percentage of phosphorus concentration (P), the percentage of potassium concentration (K), the percentage of calcium concentration (Ca) and percentage of magnesium concentration (Mg). The N, P, K, Ca and Mg concentrations were considered as the independent variables and the FFB yield as the dependent variable.

The study was conducted to investigate the causality relationship between foliar nutrient composition and FFB yield. Two approaches were considered namely multiple linear regression and robust M-regression. The performances of these two models are compared using the determination coefficient, R^2. R^2 is defined as:

$$1-\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

where, y_i represent the actual observations, $\hat{y}$ the predicted values and $\bar{y}$ the mean of the observations.

RESULTS AND DISCUSSION

Multiple linear regression: The use of the major foliar nutrients composition, N, P, K, Ca and Mg was introduced to the model. The stepwise procedure in MLR was applied in order to select the significant variables in the model. For station S_1, the regression line is, FFB = -20.311+342.077 P - 14.697 Ca and the R^2 value is 0.3692. The regression equation for station S_2 is, FFB = 5.707-53.642 Mg+201.609 P - 6.298 K+3.039 N which corresponds with an R^2 of 0.422. For station S_1, the P and Ca concentrations were found to be significantly affected by FFB yield.

The residual analysis plays an important role in judging the adequacy of the model. The analysis is needed to justify whether the model's performance indicates that the error is normally distributed to ensure that the interpretation of the model is valid. The normality assumption in error must be followed to make sure that the model and individuality tests are valid. Fig. 1 presents the scatter plots of error distribution and normal probability for both the S_1 and S_2 stations. Residual analysis was performed individuality to investigate the distribution of error modeling. We found that the error distribution for all stations was scattered within the mean (zero). The normal probability was also plotted and the results are shown in Fig. 1. It is found that all the plots form approximately straight lines, which clearly indicate that the error is normally distributed.

The second stage in this study is applying the MLR to the Major Nutrient Composition (MNC) and the Nutrient Balance Ratio (NBR) as independent variables. For station

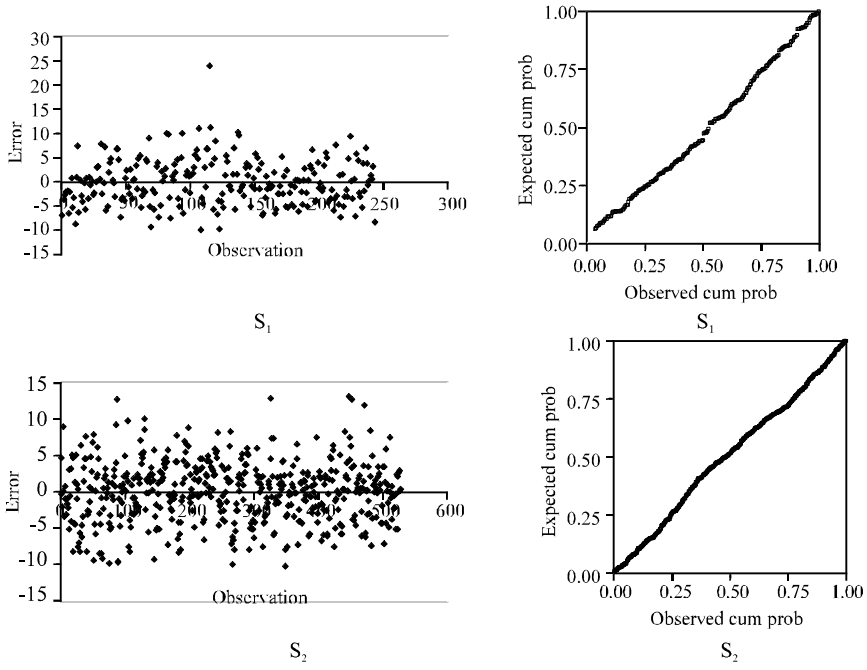

Fig. 1: The error distribution and normal probability plots for the MLR model for stations S_1 and S_2

S_1 the variables in the models were not changed, but the R^2 values were decreased from 0.392 to 0.379. The regression equation for S_2 is, FFB = 4.150+130.174 P-4.285 K - 0.952 N-P +1.990 N-Mg and it corresponds with an R^2 value of 0.433. This shows a slight improvement in the R^2 value. In station S_2, the variables included in the model were also changed from N, P, K and Mg to N, K, the N-P ratio and the N-Mg ratio.

Generally, by introducing the NBR into the regression model, only a slight improvement can be achieved in model accuracy. Actually the MNC information is sufficient for the purpose of improving the model, whereas the NBR information does not provide additional accuracy due to its complex interpretation.

Robust M-regression: The purpose of introducing robust M-regression to this study is to improve the fit of the model by eliminating some of the data values by treating them as outliers in the data. The quantile-quantile plot in Fig. 2 proved the existing of outlier observations in the data set. For station S_1, observation 240, 242 and 243 is considered as outliers data. Observations 30, 522 and 527 as shown in Fig. 2 in station S_2 are outlier's data. The presence of outliers in the data set had an influence on the fit of the model and therefore on the overall performance. This result is similar to that of the MLR regression. The R^2 values for stations S_1 and S_2 were

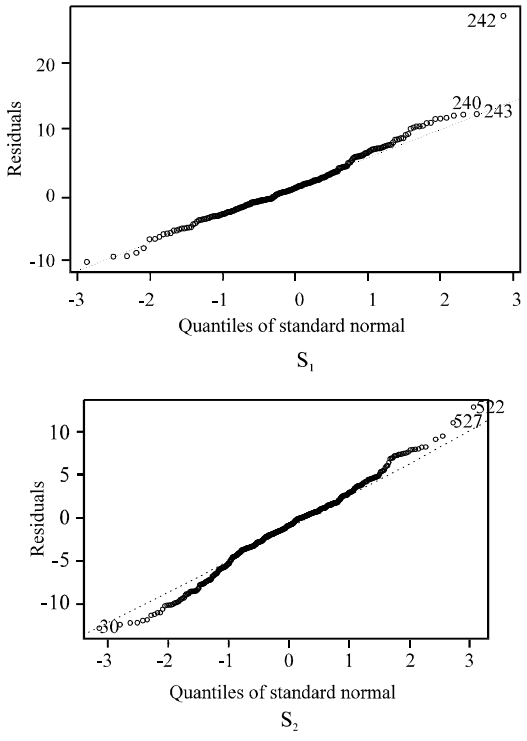

Fig. 2. The quantile-quantile plot for stations S_1 and S_2

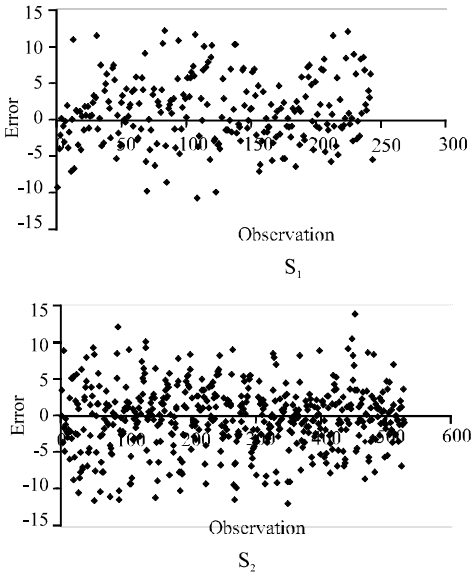

Fig. 3: The error distribution plot for the RMR model for stations S_1 and S_2

recorded as 0.571 and 0.598, respectively. The regression equation for station S_1 using the RMR model is given as, FFB = -16.790+331.546 P - 5.466 K - 19.296 Ca. Meanwhile, the regression equation for S_2 station is given as, FFB = -5.5279+329.027 P – 6.7802 Ca –31.283 Mg.

The R^2 value corresponds to the variance explained by the independent variables in the model. For example, the concentrations of P, Ca and Mg explain about 59.80% of the variance in the model and the rest can be are explained by unobserved variables at station S_2. As with the MLR model, the distribution of the errors also investigated. The distribution error plot for both stations is shown in Fig. 3. Generally, the error was scattered randomly along the mean line (y = 0). Those plots are similar to the plots of the MLR model. Thus the conclusion can be made that the model is valid.

CONCLUSIONS

Now discussion will be focused on the performance of the MLR(MNC) and MLR(NBR) models. From this, we can see whether the inclusion of NBR, TLB, deficiency of K or deficiency of Mg as independent variables improves. The differences in model performance between the MLR and RMR models will be discussed.

Table 1 gives the values of R^2 for stations S_1 and S_2. Compare the R^2 values between the MLR (MNC), MLR (NBR) and RMR models. The results shows that modeling using the MLR(MNC) and MLR(NBR) models is comparable. The last column in the table represents the R^2

Table 1: The R^2 values for MLR(MNC), MLR(NBR) and RMR models

	MLR(MNC)	MLR(NBR)	RMR
Station S_1	0.392	0.379	0.571
Station S_2	0.422	0.433	0.598

values for the RMR model. By using the RMR method, the R^2 values were increased by 45.66% from 0.392 to 0.571 and by 41.71% from 0.422 to 0.598, at stations S_1 and S_2, respectively. Therefore, we can deduce that the RMR method managed to increase the accuracy level of oil palm estimation.

This study found that when using statistical approaches such as regression model, the accuracy of the estimation is around 84 to 86% and the error of estimation is about 14 to 16%. Even though the proposed method has the ability to increase accuracy, there must be some areas we can explore to find the best model. The major goal of our study is to find the best model with the highest accuracy of estimation. This study found that robust M-regression is the best alternative tool in improving modeling performance when studying the causality relationship between the foliar nutrient composition and the FFB yield.

REFERENCES

Andrew, D.F., 1974. A robust method for multiple linear regression. Technometrics, 16: 523-551.

Barnett, V. and T. Lewis, 1995. Outliers in Statistical Data. John Wiley and Sons, England. 3rd Edn., pp: 584.

Birkes, D. and Y. Dodge, 1993. Alternative Methods of Regression. John Wiley and Sons, Inc. NY. 1st Edn., pp: 228.

Becker, R.A., J.M. Chambers and A.R. Wilks, 1988. The New S Language: A programming environment for data analysis and graphics. Wadsworth and Brooks/Cole: Pacific Grove, CA. 1st Edn., pp: 702.

Chin, S.A., 2002. Narrowing the yield gap in oil palm between potential and realization. The Planters, 78: 541-544.

Draper, N.R. and H. Smith, 1981. Applied Regression Analysis. John Wiley and Sons, New York, 2nd Edn., pp: 709.

Fairhurst, T.H. and E. Mutert, 1999. Interpretation and management of oil palm leaf analysis data. Better Crops Intl. 13: 48-51.

Foong, F.S., 1999. Impact of moisture on potential evapotranspiration, growth and yield of palm oil. Proc. 1999 PORIM Intl. Palm O il Cong. (Agric.), pp: 265-287.

Foster, H., 2003. Assessment of Oil Palm Fertilizer Requirements. In 1st Edn.,Thomas Fairhust and Rolf Hardter, Oil Palm Management for Large and Sustainable Yields. PPI, PPIC and IPI.

Green, A.H., 1976. Field Experiments as a Guide to Fertilizer Practice. In Corley, R.H.V., J.J. Hardon and B.J. Wood, 1976. Oil Palm Research: Developments in Crop Science (1). Elsevier Scientific Publishing Company, Netherlands. 1st Edn.,. pp: 235-261.

Hampel, F.R., 1974. The influence curve and its role in robust estimation. J. Am. Stat. Asso., 69: 383-393.

Huber, P.J., 1973. Robust estimation of a location parameter. Ann. Math. Stat., 35: 73-101.

Khamis, A. and M. Abdullah, 2004. On robust environmental quality indices. Pertanika J. Sci. Technol., 12: 1-10.

Mokhtar A., 1994. Analisis Regresi. Dewan Bahasa Dan Pustaka, Kuala Lumpur. 1st Edn., pp: 241.

Muhammad, A.T., Z.Z. Zakaria, M.T. Dolmat, H.L. Foster, H.A. Bakar and K. Haron,1991. Relative efficiency of urea to sulphate of ammonia in oil palm: Yield response and environmental factors. Proc. 1991 PORIM Intl. Palm Oil Conf. Agric., pp: 340-348.

Norušis, M.J., 1998. SPSS® 8.0. Guide to Data Analysis. Prentice Hall, New Jersey. 1st Edn., pp: 553.

Oboh, B.O. and M.A.B. Fakorede, 1999. Effects of weather on yield components of the oil palm in forest location in Nigeria. J. Oil Palm Res., 11: 79-89.

Rousseeuw, P.J. and A.M. Leroy, 1987. Robust Regression and Outlier Detection. Wiley, New York. 1st Edn., pp: 352.