Pakistan Journal of Nutrition 9 (4): 380-386, 2010
ISSN 1680-5194

Constructing the Best Regression Model for Maiwa Variety

F. Busari Abdullahi, Abubakar Usman and A.T. Cole
Department of Mathematics/Computer Science,
Federal University of Technology, Minna, Niger State, Nigeria

Abstract: As difficult as it can be to determine the plant attribute that contributes most to better yield of cereal crop named Maiwa. We use multivariate regression model to determine the contribution of Plant height (X_1); Number of leaves (X_2); Number of tillers (X_3) and Leaf's area in square feet (X_4). Four multivariate regression models were developed by dropping each attribute. A data set collected from the Institute of Agricultural Research (IAR) Ahmadu Bello University, Samaru-Zaria was used for the analysis. Using each of the models to assess the contribution of each attribute, it was discovered that the Multivariate regression model that has the best fits of the data set, when covariates are dropped one after the other is $Y = 0.02371 - 0.003111X_2 + 0.001759X_3 - 0.002503X_4$. Thus, plant height ($X_1$) is an irrelevant plant attribute for the variety-Maiwa.

Key words: Regressor, predictor, goodness-of-fit, multivariate, yield, tiller

INTRODUCTION

The statistical technique that is used to establish the existence of linear relationship between the dependent variable and the independent variables is the Regression Analysis. If there is a single independent or predicator variable is referred to as simple linear regression, while if it involved more than one independent or predictor variables we have the case of Multivariate regression or multiple regression analysis.

In many crops, especially arable crops, yield depend on some plants attributes such as plant height, number of leaves, stalk thickness, spacing variates etc. These plants attributes are referred to as the independent variables, covariates, predictors, or regressors; while the yield is the corresponding dependent variables or responses. Each of these regressors contributes to the variation in the yield of the variety.

Although the contribution varies from one crop to another, while in some crop it causes high variation in other the variation is insignificant.

According to Ogunremi (1970) the pod number or unit is an important independent variable that determines the yield in pod producing crops. Gilbert Tukers (1976) worked on sunflower and discovered that the number of heads of plant is very significant factor that determines the yield.

In fiber crops such as Kenaf, it is the plant height that determines the yield significant (Baker, 1970), while in cotton it is the number of bolls that determines the yield (Gardner and Tucker, 1976). Generally foods are produced by the leaves through photosynthesis and stored in the plant roots for plant consumption, while water required is transported through the stem from the roots to the leaves for food manufacturing. Therefore, leaves stem and roots are significant plant parts to plant attributes; hence, there is a relationship between these plant attributes and the yield.

If there is no correlation between two or more co varieties there is a possibility of having a good-fit, while if there is multicollinearity or relationship between these is the possibility of sources of variation. We shall in this paper considered the conditions of good fit in relation to the variables or attributes as proposed by Rizzi Laura (2008) and Gerald Keller and Brian Warrack (2003).

MATERIALS AND METHODS

Here we shall among other things distinguish between simple linear regression and multivariate regression model; furthermore discussed the assessing of model.

Regression analysis: A statistical technique that is used to establish the existence of linear relationship between the dependent variable and the independent variables is known as Regression analysis. It is also used to predict the value of one variable; this technique requires developing a Mathematical equation called Model.

According to Gerald Keller and Brian Warrack (2003): In developing a model it is necessary to known the nature of the relationship between dependent or response variables and each of the independent or predictor variables. This could be done by method of either Deterministic or Probabilistic models. The first model is not realistic because it does not assume the randomness of the variables involved or other external factors. The second includes the random component which measures the error of the deterministic component. The random component accounts for measureable and immeasurable variables that are not part of the model.

Corresponding Author: Abubakar Usman, Department of Mathematics/Computer Science, Federal University of Technology, Minna, Niger State, Nigeria

The Regression analysis models: The regression analysis of probabilistic form is said to be simple linear regression equation or a first order linear model if the model is written as follows:

$$Y_{ij} = \beta_0 + \beta_1 x + \varepsilon_{ij} \qquad (1)$$

Where Y_{ij} is the response, yield or dependent variable; x is the predictor, regressor or independent variable; $\$_0$ is the intercept on the y-axis; $\$_1$ is the slope or coefficient of regression; g_i is the residual error.

The $\$_0$ and $\$_1$ are unknown parameters of the population but are estimated using the least square estimate method as stated:

$$\text{(i)}\,\beta_0 = y_{ij} - \beta_1 \overline{x}$$

and

$$\beta_1 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)_2}$$

In multivariate regression model two or more variables are assumed to be linearly related if:

$$Y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon_{ij} \qquad (2)$$

Where Y_{ij} the response in i^th^ row, j^th^ Colum; x_i the i^th^ independent variable; $\$_i$ the i^th^ coefficient of regression; $\$_0$ the intercept on the y-axis; g_{ij} the i^th^ row, j^th^ column error of the term.

The least square method could also be used to estimate the population parameters $\$_i$ by minimizing the sum of squares of error and differentiating with respect to each population parameters. The (K+1) unknown parameters shall be obtained using any convenient Mathematical methods:

$$Y_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + \varepsilon_{ij} \qquad (3)$$

$$\varepsilon_{ij} = Y_{ij} - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 \cdots\cdots - \beta_k x_k \qquad (4)$$

$$\varepsilon_{ij}{}^2 = (Y_{ij} - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 \cdots\cdots - \beta_k x_k)^2 \qquad (5)$$

$$\sum \varepsilon_{ij}{}^2 = \sum (Y_{ij} - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 \cdots\cdots - \beta_k x_k)^2 \qquad (6)$$

$$\frac{d\sum \varepsilon_{ij}{}^2}{d\beta_0} = \frac{d\sum (Y_{ij} - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 \cdots\cdots - \beta_k x_k)^2}{d\beta_0} = 0$$

i.e

$$0 = -2\sum (Y_{ij} - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 \cdots\cdots - \beta_k x_k) \qquad (7)$$

$$\frac{d\sum \varepsilon_{ij}{}^2}{d\beta_1} = \frac{d\sum (Y_{ij} - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 \cdots\cdots - \beta_k x_k)^2}{d\beta_1} = 0$$

$$0 = -2\sum x_{ik} (Y_{ij} - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 \cdots\cdots - \beta_k x_k) \qquad (8)$$

Similarly, the matrix notation could be used in obtaining the estimate of the population parameter or coefficient of regression as shown below:

$$\underline{Y} = \underline{X\beta} + \underline{\varepsilon}$$

Where:

$$\underline{Y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ y_k \end{pmatrix} \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ \beta_{kk} \end{pmatrix} \underline{X} = \begin{pmatrix} x_{11}x_{12}x_{13}\ldots x_{1k} \\ x_{21}x_{22}x_{23}\ldots x_{2k} \\ x_{31}x_{32}x_{33}\ldots x_{3k} \\ \ldots\ldots\ldots\ldots\ldots \\ x_{p1}x_{p2}x_{p3}\cdots x_{pk} \end{pmatrix} \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ .. \\ \varepsilon_k \end{pmatrix}$$

and

$$\underline{\varepsilon} \sim \text{iii } N(0, \delta_i^2),\, E(\varepsilon) = 0,\, \text{cov}(\varepsilon) = \begin{pmatrix} 1 & 0 & \ldots & \ldots & 0 \\ 0 & 1 & \ldots & .. & 0 \\ \ldots. & & \ldots & & .. \\ \ldots & \ldots & \ldots & \ldots\,\ldots & \ldots.1 \end{pmatrix} \delta^2$$

$\underline{x}$ is full rank of k+1, estimate of $\underline{\beta}$ is unbiased of β and $\hat{\delta}^2$ is biased estimate of $*^2$ while the unbiased estimate of $*^2$ is as follows:

$$E(\hat{\delta}^2) = \frac{\delta^2\,(n\text{-}r)r}{r(n - r)} = \delta^2$$

According to Rizzi Laura (2008) the assumptions that guide the linear model are related to the distribution of Error matrix; Independent variables matrix and Unknown population parameters matrix.

If any of the assumptions is wrong, then there would be problems with the assumptions that relates to distribution errors; choice of regressors and the estimates of the parameters.

Rizzi Laura (2008) further stressed that if the g_{ij} is not normally distributed, then the inference procedures shall only be valid asymptotically, if the variance of g_{ij} terms are not constant or the same (homoscedasticity), then, the error are heteroscedasticity, which occurs in cross-section data and if the error terms are pair wise correlated, that is E (g_i, g_j) ... 0, I ... j this happens in time series data and the problem it causes is referred to autocorrelation.

Rizzi observed that a model is said to be a bad model if irrelevant independent variables are included; relevant independent variables excluded; incorrect functional form is used to fit the data; matrix of $\underline{X}$ has less than full column rank and the independent variables are correlation or correlated with the error term.

The last points cause multicollinearity and when there is a problem of multicollinearity, the coefficient of determination (R^2) is always high and the estimates of coefficient of regressions are always insignificant. There are always high correlation coefficients between the independent variables and high Variance Inflation Factors (VIF).

Multicollinearity can be remedied by removing one or more independent variables if they are observed to cause multicollinearity or increase the sample size or transform the equation model.

As a result of this, Rizzi Laura (2008) suggested a prediction criterion called Ameriya Prediction Criterium (PC). This is used to evaluate the goodness of fit of a model:

$$\text{Pr ediction criterium} = \frac{\text{RSS}\left[n+(k-1)\right]}{\left[n-(k-1)\right]} \qquad (9)$$

Where, RSS is the regression sum of square, n is the observations, (K-1) is the number of independent variables in the model.

Assessing the model: According to Gerald Killer and Brian Warrack (2003), Models are assessed using the following methods; Standard error of estimate; t-test of the slope, that is $\$_i$; Coefficient of determination, that is R^2; F-ratio or p-value in ANOVA.

If the standard error is large then the error will also be large, which implies that the model is poorly fitted, but if the standard error is small this implies that some of the errors will be close to zero, this implies a good-fitted model.

The standard error is estimated using the relation:

$$\text{S tan dard error} = \sqrt{\frac{\text{SSE}}{n-2}}$$

Where SSE is the sum of square of error, n is the number of observations.

If the error is normally distributed the test statistics is student t- distributed with V = n-2 degrees of freedom. This is used to prove if there exists a linear relationship between two or more covariates or variables. Similarly a large value of F- indicates that a significant proportion of the variation in Y- is explained by the regression equation and that the model is valid, while a small value of F- implies that most of the variation in Y- is unexplained.

Data analysis: The experiment was conducted after the planting and successful germination of the seed and maturity stage of the plant. Sampled were examined and measure of twenty of the samples height of the plant (X_1), umber of leaves (X_2), number of tillers (X_3) and leaf's area in square feet (X_4). It should be noted that this attributes are indeed essential for the yield.

Data presentation: The data for this study is secondary data obtained from the Institute of Agricultural Research (IAR), A.B.U, Samaru, Zaria.

We developed four models dropping one plant attribute each for the subsequent Multivariate regression models.

Model 1: In this model all the four predictor variables were used and the model is

$$Yij = 0.01692 + 0.002595X_1 - 0.0003013X_2 - 0.001687X_3 - 0.002303X_4$$

From the equation, the intercept on Y-axis is 0.01692 and for every increase of one unit of the plant height there is an increase of about 0.002595 of the yield. But for every increase in the number of leaves of Maiwa variety there is a decrease of about 0.0003013 in the yield. There is an increase of about 0.001687 in the yield of Maiwa when the number of tillers is increased by one unit and a decreased of about 0.002303 was observed in the yield when the leaf square area is increased by one unit.

The coefficient of determination R^2 is 0.1087, this implies that about 11% of the variation in the Maiwa yield is explained by the regression model while the remaining percentages is unexplained due to environmental factors. This gives the p-value of 0.7658 which is greater than " = 0.05, thus not significant and R^2 shows the data does not fit the model.

The p-values for all the independent variable are not significant , hence the t-test are also insignificant since the R^2

is less than 0.75, the variables are independent of one another. This implies that, there is no multicollinearity among the independent variables which is justified by the Variance Inflation Factors (VIF) for the coefficient which is less than 5.

Model 2: We further dropped leaves area (X_4) and obtained the model below:

$$Y_{ij} = 0.008794 + 0.005130X_1 - 0.0003071X_2 + 0.001505X_3$$

From the model, we deduced that for every unit increase in the plant height there is an increase of 0.005130 increases in the yield, but a decrease of 0.0003071 was observed for every increase of one unit in the leaves area. Similarly an increase of about 0.001505 was noted for every unit increase in the number of tiller.

The coefficient of determination R^2 obtained for the model was 8.75%, that implies about 9% of the total variation could be explained by the regression model, while the remaining 91% was due to error. The p-values are 0.6803 and then t-test of independent variable are insignificant. None of the independent variables makes a significant contribution, since the R^2 is less than 0.75 there is no problem of multicollinearity. The correlation matrix shows that there a strong relationship between the number of leaves and the number of tillers, the variance inflation factors which is one of the litmus test for multicollinearity existence is less than 5 for the coefficient (Rizzi Laura, 2008). This justified the non-existence of multicollinearity; hence the data set fits the model best.

Considering when the number of tillers (X_3) is dropped in the regression, the following model $Y_{ij} = 0.00342 + 0.006399x_1 - 0.000627x_2 - 0.001546x_4$ from the result we have the coefficient of determination R^2 equals 0.0439 which is 4.39%, this shows that about 4.4% of the total variation was explained by the regression while, the remaining percentages was due to error. The p-value was also found to be 0.8636; both the p-value and the t-test for the coefficient were not significant.

Since the R^2 is less than 0.75 which shows that multicollinearity is not a problem. This is justified by the variance inflation factor value for the coefficient of the independent variables which is less than five. Hence there is no multicollinearity.

Model 3: When the number of leaves (X_2) was dropped in the regression, the effect of the leaves resulted in the following model:

$$Y_{ij} = 0.002021 + 0.005261X_1 + 0.000453X_3 - 0.002427X_4$$

The intercept on the y-axis shifted to 0.002021, while the R^2 equals 0.489. This value indicates that about 5% of the total variation was explained by the regression and that there is no multicollinearity; since R^2 is less than 0.75. In support of this claim, the variance inflation factor shows values that are less than five for each of the regressors, although the p-value and the t-test shows that there is no linear relationship between the yield and the regressors of the model, the standard error is large to have discarded the model and there is multicollinearity. It was observed that from the model that for every one unit increase in the plant height, number of tillers and the leaf 's square area, there are increase of 0.005261 and 0.00453 with a decrease of about 0.002427 in the yield respectively.

Model 4: When plant height (X1) was omitted in the regression, the following model was obtained:

$$Y_{ij} = 0.02371 - 0.0003111X_2 + 0.001759X_3 - 0.002503X_4$$

The intercept shifted to 0.02371 on the y-axis. This led to increase of about 0.001759 in the yield due to increase of one unit in the number of tillers. But an increase in one unit of the number of leaves led to decrease of 0.003111 of the yield. Similarly an increase of one unit of the square area of the leaf led to a decrease of about 0.002503 in the yield. The coefficient of determination R^2 obtained for the model was 0.1068. This shows that about 11% of the total variation was due to the regression; also for R^2 less than 0.75 and the VIF less than five for each of the coefficient of the independent variables, there is no multicollinearity. The p-value of about 0.6016 and the t-test showed that there is no linear relationship between the yield and the independent variables. Rizzi Laura (2008) further stressed that there three ways of verifying the presence of multicollinearity. High R^2 with insignificant estimated coefficient, if R^2 >0.75; High correlation coefficient between regressors; High Variance Inflation Factors (VIF).

Considering these three factors and the discussions so far on the variable Maiwa, it is clearly seen that there is no problem of multicollinearity.

We use Rizzi Laura Prediction Criterion (PC) as stated in equation (16) to investigate the goodness-of-fit of the models discussed above as follows:

The general model that contains all the independent variables used:

$$Yij = 0.01692 + 0.002595X_1 - 0.0003013X_2 + 0.001687X_3 - 0.002303X_4$$

RSS = 0.002942, R^2 = 0.1087, n = 20, k = 4 ÷k-1 = 3

$$PC_1 = \frac{0.002942(20+3)}{20-3} = \frac{0.002942 \text{ x } 23}{17} = 0.00398$$

When X_1, is excluded the resultant model:

$$Y_{ij} = 0.02371 - 0.0003111X_2 + 0.001759X_3 - 0.002503X_4$$

RSS = 0.002949, R^2 = 0.1068, n = 20, k = 3

$$PC_2 = \frac{0.002949(20+2)}{20-2} = \frac{0.002949 \text{ x } 22}{18} = 0.00360$$

When X_2, is excluded the resultant model:

$$Y_{ij} = 0.002021 + 0.005261X_1 + 0.000453X_3 - 0.002427X_4$$

RSS = 0.003140, R^2 = 0.0489, n = 20, k = 3

$$PC_3 = \frac{0.003140(20+2)}{20-2} = \frac{0.003140 \text{ x } 22}{18} = 0.00384$$

When X_3, is excluded the resultant model:

$$Y_{ij} = 0.00342 + 0.006399X_1 - 0.000627X_2 - 0.001546X_4 \quad 12$$

RSS = 0.003156, R^2 = 0.0439, n = 20, k = 3

$$PC_4 = \frac{0.003156(20+2)}{20-2} = \frac{0.003156 \text{ x } 22}{18} = 0.00386$$

When X_4, is excluded the resultant model:

$$Y_{ij} = 0.008794 + 0.005130X_1 - 0.0003071X_2 + 0.001505X_3$$

RSS = 0.003012, R^2 = 0.0875, n = 20, k = 3

$$PC_5 = \frac{0.003012(20+2)}{20-2} = \frac{0.003012 \text{ x } 22}{18} = 0.003681$$

It follows that $PC_2<PC_5<PC_3<PC_4<PC_1$, this shows that the best model was obtained when X_1 (Plant height) was excluded in the regression for it has the smallest PC of 0.00360. Hence, Y_{ij} = 0.002021 + $0.005261X_1$ + $0.000453X_3$ - $0.002427X_4$ is the best model that fits the data set and has the best goodness- of -fit? Therefore plant height is an irrelevant plant attribute for this variety (Maiwa).

RESULTS AND DISCUSSION

The result of the Multivariate regression analysis model as extracted from Instat Graph statistic software output are presented as Appendix 1. Using all the require conditions to assess the goodness of fit of the model proposed by Rizzi Laura (2008) and Gerald Keller and Brian Warrack (2003) the best model was obtained when the plant height was dropped or excluded as an independent variable in the model. Thus,

$$PC_2<PC_5<PC_3<PC_4<PC_1$$

This explains that plant height is an irrelevant plant attributes of Maiwa variety. The best model therefore dependent on number of leaves, leaf 's square area and tillers in that order even though some of these attributes contributed insignificantly to the models goodness-of-fit.

Appendix 1:

Model 1 (MAIWA)

Multiple regression results: What equation fits the data the best?

[A:Y] = 0.01692 + 0.002595*[B:X1] -0.0003013*[C:X2] + 0.001687*[D:X3] -0.002303*[E:X4]

How good is the fit?

R^2 = 10.87%. This is the percent of the variance in A:Y explained by the model.

The P-value is 0.7658, considered not significant. The P-value answers this question:

If there were no linear relationship among the variables, what is the chance that R^2 would be that high (or higher) by chance.

Since P is high, the rest of the results will be of little interest.

Sum-of-squares [0.002942], SD of residuals [0.01401], R^2 [0.1087], Adjusted R^2 [0.1290], Multiple R [0.3297], F ---- 0.4573

Which variable(s) make a significant contribution?

Variable	t ratio	P-value	Significant?
(Constant)	0.4276	0.6750	No
B:X1	0.1781	0.8611	No
C:X2	1.003	0.3316	No
D:X3	1.044	0.3130	No
E:X4	0.5979	0.5588	No

Each P-value compares the full model with a simpler model omitting one variable. It tests the effect of one variable, after accounting for the effects of the others. Is multicollinearity a problem?

Variable	VIF	R^2 with other X
B:X1	1.15	0.1275
C:X2	2.58	0.6126
D:X3	2.68	0.6271
E:X4	1.21	0.1702

Each R^2 quantifies how well that X variable is predicted from the other X variables (ignoring Y). VIF is calculated from R^2.

All R^2 values are low (<0.75). The X variables are independent of each other. Multicollinearity is not a problem.

Model 2 (MAIWA)

Multiple regression results: What equation fits the data the best?

[A:Y] = 0.008794 + 0.005130*[B:X1] -0.0003071*[C:X2] + 0.001505*[D:X3]

How good is the fit?

R^2 = 8.75%. This is the percent of the variance in A:Y explained by the model.

The P-value is 0.6803, considered not significant. The P-value answers this question:

If there were no linear relationship among the variables, what is the chance that R^2 would be that high (or higher) by chance?

Since P is high, the rest of the results will be of little interest.

Sum-of-squares [0.003012], SD of residuals [0.01372], R^2[0.0875]' Adjusted R^2 [-0.0837], Multiple R [0.2957] F ------- 0.5111

Which variable(s) make a significant contribution?

Variable	t ratio	P-value	Significant?
(Constant)	0.2416	0.8122	No
B:X1	0.3755	0.7122	No
C:X2	1.044	0.3119	No
D:X3	0.9680	0.3474	No

Each P-value compares the full model with a simpler model omitting one variable. It tests the effect of one variable, after accounting for the effects of the others. Is multicollinearity a problem?

Variable	VIF	R^2 with other X
B:X1	1.05	0.0469
C:X2	2.58	0.6122
D:X3	2.59	0.6134

Each R^2 quantifies how well that X variable is predicted from the other X variables (ignoring Y). VIF is calculated from R^2.

All R^2 values are low (<0.75). The X variables are independent of each other. Multicollinearity is not a problem.

Model 3 (MAIWA)

Multiple regression results: What equation fits the data the best?

[A:Y] = 0.003242 + 0.006399*[B:X1] -6.272E-05*[C:X2] -0.001546*[E:X4]

How good is the fit?

R^2 = 4.39%. This is the percent of the variance in A:Y explained by the model.

The P-value is 0.8636, considered not significant. The P value answers this question:

If there were no linear relationship among the variables, what is the chance that R^2 would be that high (or higher) by chance?

Since P is high, the rest of the results will be of little interest.

Sum-of-squares [0.003156]; SD of residuals [0.01404]; R^2 [0.0439]; Adjusted R^2 - [0.1353]; Multiple R [0.2096]; F ----- 0.2450

Which variable(s) make a significant contribution?

Variable	t ratio	P-value	Significant?
(Constant)	0.08661	0.9321	No
B:X1	0.4521	0.6573	No
C:X2	0.3211	0.7523	No
E:X4	0.4074	0.6891	No

Each P-value compares the full model with a simpler model omitting one variable. It tests the effect of one variable, after accounting for the effects of the others. Is multicollinearity a problem?

Variable	VIF	R^2 with other X
B:X1	1.07	0.0694
C:X2	1.09	0.0793
E:X4	1.16	0.1397

Each R^2 quantifies how well that X variable is predicted from the other X variables (ignoring Y). VIF is calculated from R^2.

All R^2 values are low (<0.75). The X variables are independent of each other. Multicollinearity is not a problem.

Model 4 (MAIWA)

Multiple regression results: What equation fits the data the best?

[A:Y] = 0.002021 + 0.005261*[B:X1] + 0.0004531*[D:X3]- 0.002427*[E:X4]

How good is the fit?

R^2 = 4.89%. This is the percent of the variance in A:Y explained by the model.

The P-value is 0.8432, considered not significant. The P-value answers this question:

If there were no linear relationship among the variables, what is the chance that R^2 would be that high (or higher) by chance?

Since P is high, the rest of the results will be of little interest.

Sum-of-squares [0.003140]; SD of residuals [0.01401]; R^2 [0.0489]; Adjusted R^2 - [0.1295]; Multiple R [0.2211]; F ----- 0.2740

Which variable(s) make a significant contribution?

Variable	t ratio	P-value	Significant?
(Constant)	0.05509	0.9567	No
B:X1	0.3670	0.7184	No
D:X3	0.4321	0.6714	No
E:X4	0.6302	0.5375	No

Each P-value compares the full model with a simpler model omitting one variable. It tests the effect of one variable, after accounting for the effects of the others. Is multicollinearity a problem?

Variable	VIF	R^2 with other X
B:X1	1.11	0.0975
D:X3	1.13	0.1139
E:X4	1.20	0.1694

Each R^2 quantifies how well that X variable is predicted from the other X variables (ignoring Y). VIF is calculated from R^2.

All R^2 values are low (<0.75). The X variables are independent of each other. Multicollinearity is not a problem.

Model 5 (MAIWA)

Multiple regression results: What equation fits the data the best?

[A:Y] = 0.02371 -0.0003111*[C:X2] + 0.001759*[D:X3] - 0.002503*[E:X4]

How good is the fit?

R^2 = 10.68%. This is the percent of the variance in A:Y explained by the model.

The P-value is 0.6016, considered not significant. The P-value answers this question:

If there were no linear relationship among the variables, what is the chance that R^2 would be that high (or higher) by chance?

Since P is high, the rest of the results will be of little interest.

Sum-of-squares [0.002949]; SD of residuals [0.01358]; R^2 [0.1068]; Adjusted R^2 - [0.0607]; Multiple R [0.3268]; F ------- 0.6378

Which variable(s) make a significant contribution?

Variable	t ratio	P-value	Significant?
(Constant)	2.356	0.0316	Yes
C:X2	1.087	0.2932	No
D:X3	1.160	0.2631	No
E:X4	0.7006	0.4936	No

Each P-value compares the full model with a simpler model omitting one variable. It tests the effect of one variable, after accounting for the effects of the others. Is multicollinearity a problem?

Variable	VIF	R^2 with other X
C:X2	2.50	0.5992
D:X3	2.51	0.6022
E:X4	1.10	0.0935

Each R^2 quantifies how well that X variable is predicted from the other X variables (ignoring Y). VIF is calculated from R^2.

All R^2 values are low (<0.75). The X variables are independent of each other Multicollinearity is not a problem.

REFERENCES

Baker, E.F.I., 1970. Kenaf and Roselle in Western Nigeria. World crops Nov/Dec, 1970, PP: 380-386.

Gerald Keller and Brian War Rack, 2003. Statistics for Management and Economics. Zcurt Hinrich Brooks Kole Thomas learnig, pp: 602-695.

Gibert Tukers, 1976. Growth, yields and yields components of safflower as affected by source, rate and time of application of Nitrogen. Agronomic J., 59: 54-56.

Ogunremi, E.A., 1970. Relationships Between Yeilds and some Agronomic Characters of sugar-cane in southern Nigeria. Nig. Agri. J., PP: 1-8.

Rizzi Laura, 2008. Specification error, Multicollinearity and Qualitative covariates. Www site net Nov 19, 2008.

Gardner, B.R. and T.C. Tucker, 1976. Nitrogen effects on cotton: Vegetative and fruiting characteristics. Soil science America Proc. Vol. 31, No.6, Nov-Dec. 1967; pp: 780-791-56.