Evaluation of some statistical methods used for estimating relative contribution of yield factors in maize

Hayam Said Fateh

The word multivariate should say it all; these techniques look at the pattern of relationships between several variables simultaneously. As the name indicates, multivariate analysis comprises a set of techniques dedicated to the analysis of data sets with more than one variable. Several of these techniques were developed recently in part because they require the computational capabilities of modern computers. Also, because most of them are recent, these techniques are not always unified in their presentation, and the choice of the proper technique for a given problem is often difficult. The yield of maize is the integrated effect of many variables that affect plant growth throughout the season. Multivariate analysis studies may help in interpretation the results and may aid the breeder for getting better varieties and good evaluation of the agricultural practices.Three field experiments were conducted at Sides Research Station, Benisuef Governorate, Agricultural Research Center, Egypt, during 2004, 2005 and 2006 seasons to evaluate some statistical methods used for estimating the relative contribution of yield factors in maize (Zea mays L.). The N fertilizer levels applied were zero, 30, 60, 90,120 and 150 kg N / feddan. Single crosses (S.C.10, S.C. 122 and S.C. 123) and three ways crosses (T.W.C.310, T.W.C.311 and T.W.C.314) were grown. The experimental design used was a strip plot design with three replications. Four statistical procedures of relating yield components to yield namely, correlation, multiple linear regression, stepwise multiple linear regression and factorSummary -109-analysis were applied to 11 yield factors at set of 14 sample sizes, namely,10,20,30,40,50,100,150,200,250,300, 350, 400,450 and 540 observations were used. The researcher used some agronomic treatments which aimed to measure the relationship between yield factors under different degrees of variation.The most important results obtained from this investigation can be summarized as follows: 1- Correlation analysis:The results indicated that increasing sample size increased the value of correlation coefficient. Also, the smallest sample size is about (40-50) for all seasons of experimentation which corresponds to correlation of highly significant for all characters. The results obtained clearly showed that the relationship between sample size and correlation coefficient was not significant as sample size was less than 20 of all characters. Notice that when sample size equaled 40 the value of correlation coefficient become significant, but this value was less than 0.5 which corresponds to correlation of negligible. On the other hand, when the value of correlation coefficient was more than 0.5 the correlation coefficient was dependable. Hence, it can not be only depend on the level of significance but it must depend on the value of correlation coefficient with the level of significance starting with sample size of 100 plants for all characters.On the other hand, a general trend of correlation associated with large sample size. The greater sample size is about 400, which corresponds to higher correlation. It is possible to obtain the largest sample size to get great correlation coefficient for each character on the basis of any increase after this great valueSummary -110-is little or no interest. It would seem to classify these characters as follows:1-Characters with low variation between plants needed 150 (plants) samples. Number of rows / ear, number of kernels /row, plant height, number of leaves / plant and stem diameter.2-Characters with medium variation between plants needed 200(plants) samples: Ear length, ear diameter, weight of 100 kernels and leaf

area.3-Character with high variation between plants, needed 250 (plants) samples. Number of kernels/ ear.The best model fitted to the relation between sample sizesand correlation coefficients of yield factors in maize over all seasons was quadratic.The quadratic model worked well for describing the relation between sample size and correlation coefficient of yield factors. The results show that minimum sample size ranged from 50 to 70 to get correlation coefficient ranging from 0.4074 to 0.4840. All correlation coefficients were highly significant at 1% level of significance. Also, the minimum sample size differed from each other according to yield factors in maize which more clearly reflected the sample size effects of yield factors. On the other hand, with high variation between plants of any character needed more observations. Also, maximum sample size ranged from 350 to 425 get correlation coefficient ranging from 0.627 to 0.739. Correlation coefficients were highly significant at 1% level of significance. The maximum sample size differs fromSummary -111-each other attributed to yield factor. Hence, with high variation between plants of any character more observations are needed. 2- Multiple linear regression analysis:Results of multiple linear regression indicate that increasing sample size increased the value of R2and adjusted R2and decreased the VIF value and SE of all variables. The smallest sample size is about 200, which corresponds to VIF value of all variables except number of kernels/ row, and number of kernels / ear in three seasons. A highest VIF denotes that there is high collinearity or multicollinearity between predicted variables. On the other hand, other variables having a VIF value less than 10.0 of all sample sizes means that there is no collinearity. Meanwhile, VIF of some characters were affected by sample size. VIF value greater than 10.0 was evidence confirmed a part of the instability of the regression coefficients due to interrelation among the explanatory variables which means thatthere is collinearity or multicollinearity. Thus, the sample sizeshould be adequate to provide a high probability of detecting assignificant an effect size of given magnitude if each an effectactually exists. Also, a sample size that is too small will not allowto properly addressing research questions with proposedstatistical analysis. Meanwhile, large samples need to makeaccurate statistical conclusions. Hence, larger sample is needed inorder to obtain a more precise estimate. It is clear that thecharacters with a high variation between plants needed a largesample while the characters with a low variation between plantsneeded a small sample size. These results point out that numberof leaves and stem diameter with a low variation between plantsSummary -112-and other characters had high variation between plants. These results help in planning appropriate selection of sample size for improving maize crop. 3- Stepwise multiple linear regression analysis:It can be indicated that increasing sample sizes increased the value of R2, adjusted R2and number of accepted variables. Standard error (SE) decreased by increasing sample size. Furthermore, the accepted variables had no collinearity where VIF value less than ten. Number of accepted variables differs from sample size to other. Meanwhile, accepted variables were affected by sample size. Also, sample sizes can be classified into three sizes. These sizes were small which ranged from 20 to 40, medium which ranged from 50 to 200 and large which ranged from 250 to 540 observations. At small sample size, number of accepted variables was two (ear length and stem diameter) in the three seasons. At medium size, six variables were accepted in the first and second seasons, but in third season five variables were accepted. At large sample size, number of accepted variables were seven, six and five in the first, second and third season, respectively. Accepted variables differed from sample size to other. On the other hand, number of kernels / row and weight of 100 kernels were removed for all sample sizes in the first season. In the second season weight of 100 kernels was removed for all sample size. Also, number of rows / ear, number of kernels / row and weight of 100 kernels were removed for all sample size in the third season. Although these variables are most important variables of yield components the stepwise analysis removed these variables Hence, these variables had high multicollinearity.Summary -113-The instabilities were likely due to linear relationships among yield components a condition referred to as multicollinearity. The degree of collinearity between predictor variables was the most important factor influencing the selection of authentic variables. On the other wise, stepwise analyses accepted the morphological characters. These variables gave little relative contribution of grain yield. It yields R-squared values that are badly biased high. It gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large). Also, variables having insignificant correlation were removed by

using stepwise analysis at small sample size. On the other hand, the variables of yield components which had significant correlation were removed by using stepwise analysis for all sample size. Stepwise methods will not necessarily produce the best model if there are redundant predictors (common problem). Models identified by stepwise methods have an inflated risk of capitalizing on chance features of the data. 4- Factor analysis:The results indicate that increasing sample size increased communality (h2) value of all characters. A communality value of 0.6 seems high and its effect is greater. Communality value less than 0.6 seems low and had no effect. When communality is high, it plays a greater role in the interpretation of the factor .Also, number of row/ear, number of kernels/row, number of kernels/ear and weight of 100 kernels are the most variables which were effective in independent structure at small sample size less than 50 observations .Hence, these characters are the most important of grain yield and did not need to the large sample size for factorSummary -114-analysis or estimation of independent structure. On the other hand, the other characters needed a large sample size for factor analysis and estimation of independent structure. Furthermore, factor analysis must be used when sample size is small. Hence, the factor analysis is better than multiple and stepwise regression at small sample size .Similarly, when sample size of 400 all characters had highly communality and representative in independent structure. Also the results indicate that the optimum sample size was 400 plants which gave the largest value of communality for all variables. Increasing sample size increased eigen value and variance ratio. It was observed that the values of variance ratio are greater than the values of eigen value. On the other hand, when sample size is greater than 400 the values of eigen value and variance ratio became fixed. Factor analysis were constructed using the principal factor analysis technique to establish the dependent relationship between morphological characteristics and yield components .Use for factor analysis by plant breeders has the potential of increasing the comprehension of the causal relationships of variables , and can help to determine the nature and sequence of trials to be selected in a breeding program .