Neth. J. agric. Sci. 25 (1977): 103-107

An application of a multiple regression method in tropical grass breeding

A. J. P. van Wijk

National Agricultural Research Station, P.O. Box 450, Kitale, Kenya

Received: 9 March 1977

Key words: Multiple regression, selection of variables, source population, seed yield, *Setaria sphacelata*

Summary

A description and an application of a computer program are given, aimed at selecting a subset of independent variables as the best predictors of a dependent variable.

Introduction

To determine which variables contribute most to the explanation of the total variance of a dependent variable stepwise multiple regression (Snedecor & Cochran, 1967) has been widely used. The major shortcoming of this method is that only one equation is selected as the best predictor. When the independent variables are correlated, stepwise regression may ignore combinations of variables with a better fit than the selected one. Backward and forward stepwise selection may lead to different results.

Daniel & Wood (1971) discuss a computer program, the Linear Least-Squares Curve Fitting Program, which enables the user to select several subsets of variables that fit the data as well as the full equation, which contains all variables.

In this paper a brief description of the computer program is given, illustrated with a practical application for plant breeding.

Linear Least-Squares Curve Fitting Program

A dependent variable may be explained through k independent variables by fitting the data to the model

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \ldots \beta_k x_{jk} + \sigma e_j$$

$(j = 1,2, \ldots N;$

e_j uncorrelated and standard deviation of $e_j = 1)$

through the linear least squares method.

The selection of equations with a smaller number of independent variables than the full equation, fitting the data as well as the equation with all variables included, is carried out by a statistic called C_p, which estimates the ratio of the sum of squared errors in y at all N data points over the variance σ^2 of random error, according to the definition

$$C_p = RSS_p/s^2 - (N - 2p)$$

in which RSS_p = residual sum of squares in subset equation
s^2 = residual mean square in full equation
N = number of observations
p = number of parameters in subset equation.
In full equation $p = k$ when $b_0 = 0$ (k = number of variables in full equation) and $p = k + 1$ when b_0 is present.

The value of C_p represents both bias and random error.

If the C_p values of a number of subset equations are plotted against p, those for equations with small bias will be close to the line $C_p = p$, while those for equations with substantial bias will be far above the line. Bias can be reduced by adding appropriate variables to the model, but the total prediction variance will increase simultaneously. Therefore, equations with a low C_p value close to p will be candidates for selection, provided that they are the most suitable in a practical sense.

The C_p search starts with a comparison of C_p values of the variables by successively including them in equations in descending order of their corresponding t values. All variables that contribute to a minimal C_p value of a certain equation, excluding the last 2 added, constitute the 'basic set'. A search is then carried out among combinations of the remaining variables, which includes the basic set anyway. If $k > 12$, this search may be of a fractional factorial type, excluding the variables of the basic set.

Some other features of this program include, amongst other things, a study of the residuals of the fitted equation to determine if they are normally distributed, the estimation of the standard deviation from observations which are near neighbours in the predictor space by which the validity of the full model may be tested, data transformations and the calculation of the relative influence of an independent variable on the dependent variable.

The relative influence is calculated as follows:

$$\text{relative influence} = \frac{|b_i|\, w_i}{w_y}$$

where b_i = the partial regression coefficient of the i-th independent variable
w_i = the range of the i-th independent variable
w_y = the range of the observed dependent variable.

The relative influence describes the fraction of the total change in the dependent variable that can be accounted for by the accompanying total change in the i-th independent variable.

Material and methods

Seed yield was assessed over 3 consecutive harvests in 1974 and 1975 on 121 plants spaced 1 m × 1 m of *Setaria sphacelata* (Schumach.) Stapf & Hubbard cv. Nandi I, which were randomly chosen from a source population of 4000 plants of the same variety.

The following characters were recorded on the 121 plants:

- number of tillers at 3 weeks regrowth;
- time of initial head emergence (IHE), scored on a weekly base, when more than 10 heads had fully emerged, 1 being the week in which the first plant had 10 or more heads;
- erectness of the plant, expressed as the ratio between the circumference at the height of the flag leaf and at the base at IHE;
- dry weight of 15 tillers with the tip of the head emerging from the subtending leaf sheath at IHE;
- fresh weight of the plant at time of seed harvest, approximately 7 weeks after IHE;
- number of heads.

The means of the 3 repeated observations were used for the multiple regression analysis.

Results and discussion

In Table 1 the full equation with seed yield as dependent variable (y) is presented. The scatter diagram of the residuals versus the fitted y values revealed that the scatter of residuals increased with y. Therefore the dependent variable was transformed by taking the common logarithm of y. An even distribution of the residuals around the zero line was then obtained, while the squared multiple correlation coefficient increased slightly (Table 2).

Table 1. Multiple regression with seed yield as dependent variable.

Character	Partial regression coefficient	t value	Relative influence
Tiller number	—0.074	0.4	0.05
Time of head emergence	—0.423	2.2	0.23
Erectness	17.594	1.5	0.13
Tiller weight	—1.337	0.2	0.03
Fresh weight of the plant	3.256	4.3	0.86
Head number	—0.572	0.9	0.12
Constant	—273.187		

$F^{6}_{114} = 25.1$
Residual mean square = 26948.998
Multiple correlation coefficient squared = 0.569

Table 2. Multiple regression with $\log_{10}$ seed yield as dependent variable.

Character	Partial regression coefficient	t-value	Relative influence
Tiller number	0.0003	0.8	0.08
Time of head emergence	—0.0559	2.4	0.18
Erectness	0.0120	0.9	0.05
Tiller weight	0.0077	0.9	0.09
Fresh weight of the plant	0.0029	3.2	0.45
Head number	0.0012	1.5	0.14
Constant	1.6780		

$F^{6}_{114} = 26.8$

Residual mean square = 0.0397
Multiple correlation coeficient squared = 0.586

The selection of variables is shown in Table 3. Fresh weight of the plant and time of head emergence were included in the basic set, having the highest t values. This meant that still 4 variables had to be searched in 2^4 equations.

The basic set was finally selected as the best predictor of seed yield, because of its low residual mean square compared to that of the full equation, the almost identical values of p and C_p and because of its feasibility of scoring. Both characters explained 57.1 % of the variation observed in seed yield.

The estimate of measurement error from near neighbours (0.19) agreed well with the residual mean square value (0.20) of the selected fitted equation. Replicates or near replicates indicated that there was considerable variation in the measure-

Table 3. Candidate equations with $\log_{10}$ seed yield as dependent variable.

Character[1]	p	C_p	Residual mean square
Fwplan + Timehe (basic set)	3	2.9	0.03965
Basic set + Tilnbr	4	4.0	0.03969
Basic set + Tilwei	4	4.8	0.03998
Basic set + Heanbr	4	2.6	0.03921
Basic set + Erectn	4	4.5	0.03987
Basic set + Tilnbr + Tilwei	5	5.9	0.04000
Basic set + Tilwei + Heanbr	5	4.2	0.03941
Basic set + Tilnbr + Heanbr	5	4.4	0.03951
Basic set + Heanbr + Erectn	5	4.1	0.03939
Basic set + Tilwei + Erectn	5	6.5	0.04020
Basic set + Tilnbr + Erectn	5	5.4	0.03983
Basic set + Tilnbr + Tilwei + Heanbr	6	5.8	0.03961
Basic set + Tilnbr + Heanbr + Erectn	6	5.9	0.03965
Basic set + Tilwei + Heanbr + Erectn	6	5.7	0.03958
Basic set + Tilnbr + Tilwei + Erectn	6	7.2	0.04010
Full equation	7	7.0	0.03970

[1] Fwplan = fresh weight of the plant; Timehe = time of head emergence; Tilnbr = tiller number; Tilwei = tiller weight; Heanbr = head number; Erectn = erectness.

ment of seed yield. The selected equation containing the fresh weight of the plant and time of head emergence, therefore showed little lack of fit over and above reproducibility or measurement error. Further searches for other variables and other forms of the equation would not be worthwile with these data. This would mean that random error was fitted (measurement error).

The basic set plus tiller number would have been a second choice, were it not for the cumbersome task of counting tillers, sometimes amounting to 1000 per plant.

The basic set plus head number showed the lowest residual mean square, but the C_p value differed considerably from p.

From the foregoing it is clear that fresh weight of the plant and time of head emergence were reliable predictors of seed yield, while the remaining variables except erectness, in spite of their significant correlation coefficients with seed yield ($P < 0.01$) could be left out from recording. The omission of these 4 characters, which are laborious to measure, will greatly facilitate the initial screening for seed yield in large source populations with wide variation. The same analysis has been applied for defining the factors responsible for yield of digestible organic matter (van Wijk, unpublished data).

Acknowledgments

This paper is published with the permission of the Director of Agriculture. The author wishes to thank Miss N. Shah of the Ministry of Finance and Planning, Nairobi, Kenya, for her kind assistance in making the program operable.

References

Daniel, C. & F. S. Wood, 1971. Fitting equations to data. Wiley, New York.
Snedecor, G. W. & W. G. Cochran, 1967. Statistical methods. Iowa State University Press, Ames.