ANALYZING A SERIES OF SOIL FERTILITY EXPERIMENTS FOR PREDICTION

BU-446-M

February 1973

F.B. Cady, R.L. Anderson, and D.M. Allen

ABSTRACT

This paper was given as part of the UCLA Conference on Statistical Computing, September 7 - September 11, 1971, a workshop conference where the analysis of large data sets was the focal point. A series of soil fertility experiments were individually analyzed and then combined using analysis of variance techniques. The treatment by experiment interaction was examined by measuring various site characteristics and running a regression analysis with controlled and uncontrolled variables. The data analysis problem then involved the problem of predictor variable selection using the residual sum of squares criterion. A procedure which simulates prediction was demonstrated using an alternative criterion, the prediction sum of squares (PRESS). The PRESS procedure determines whether or not to include a variable by seeing how much better the equation predicts "new" data when a potential predictor variable is included in the prediction equation. The two alternatives are compared using the data from the soil fertility experiments.

ANALYZING A SERIES OF SOIL FERTILITY EXPERIMENTS FOR PREDICTION

BU-446-M February 1973

F.B. Cady, R.L. Anderson, and D.M. Allen

Soil scientists predict crop yields given values of soil, climate and management variables. Explicit in this objective is determination of the important variables and parameter estimation of a prediction equation useful in calculating fertilizer requirements.

In many areas, corn yield is a function of nitrogen fertility. During 1962-65 in the non-irrigated western part of El Bajio area in central Mexico, there was a series of 82 experiments, each a designed study with four levels of applied nitrogen replicated in a randomized complete block design.

Before the experiments, a large number of variables that could not be controlled at a single or differing levels but could be "measured" at each site, were considered. Some, though perhaps important, were eliminated due to available resources; others were found to be unimportant in previous studies or the general literature. Other site variables, including several based on laboratory tests and not requiring field observations during the growing season, were measured but then eliminated if a sufficient range and uniform distribution of the measured values were not obtained or if extremely high associations with the retained site variables were observed.

The mean (of four replications) corn yields and measured site variables are on file at the Health Sciences Computing Facility, UCLA. Only the results from 72 experiments appear in the tables. The other ten were eliminated on the basis of poor population stands, unexpected site conditions found during the experimentation period, or extreme within-field variability, usually resulting from microclimatic environments.

The scales and indices used for the site variables were developed from past experience. For example, the drought index was calculated by summing the products of the number of days of wilting during different parts of the growing season by the estimated reductions in yield per day, based on several experiences reported in the literature. Each value of a scale involves a category of field conditions defined so that an approximate linear relationship exists between yield and the scale.

Each experiment was analyzed according to a randomized complete block design model. The experimental errors from each experiment were tested for homogeneity of error. A quadratic polynomial equation between yield and applied nitrogen was then calculated and the estimated response curves compared. Part of the observed variation among the curves is due to different levels of soil or endogenous nitrogen at the various sites, i.e. the true response surface between yield and total nitrogen available through a given time period might be the same for all sites but the responses between yield and applied nitrogen are estimated using different portions of the total available nitrogen absissa. In addition, other site variables are affecting the observed response between yield and applied nitrogen, resulting in additional variability in the individual experiment response surfaces.

Historically, a combined analysis of variance would have been calculated after examination of the experiment analyses.

Source of variation	d.f.
Sites (s)	$(s-1)$
Blocks/sites	$s(b-1)$
Applied nitrogen levels (L)	$(\ell-1)$
SxL	$(s-1)(\ell-1)$
Combined experimental error	$s(b-1)(\ell-1)$

Usually, as for the present data set, the site by applied nitrogen level, is significant and the problem of interest is to identify those characteristics of sites so that the interaction may be interpreted. More recently, agricultural scientists have been able to quantitize a number of potentially important site variables so that their measurement is practical. Consequently, in addition to the response variable, the yield of corn at controlled levels of applied or fertilizer nitrogen, the investigators in Mexico had available the following site variables that were measured but not controlled:

total soil nitrogen, percentage by weight X 100
excess moisture, 0-6 scale
drought, weighted index based on days of plant wilting
depth of rooting zone, centimeters
soil slope, percent X 10
soil texture, 1-5 scale
previous crop, 10-25 scale
hail, 0-6 scale
blight (H. Turcicum), 0-9 scale
weeds, 0-6 scale

This increased ability to measure site variables coincided with the advent of high speed digital computers and software, e.g. the BMD statistical package, that could handle adequately large multiple regression problems. By considering that three sources of nitrogen were available from the data, namely applied nitrogen, soil nitrogen and previous crop nitrogen, a model, linear in the parameters, was formulated. The relationship was believed to be approximated by a quadratic polynomial including the linear by linear interactions of the three nitrogen variables. Based on soil science knowledge, it was decided that certain site variables, namely depth of rooting zone, soil slope and soil texture would not interact with nitrogen but that, given the range of soil texture, a quadratic effect was to be expected.

Hypothesized was the interaction between the other site variables and the various sources of nitrogen. Consequently the 33 independent variables as shown in Table 1 were developed and called the full model. The least squares estimation procedure was used to estimate the parameters in the full multiple linear regression model. Applied nitrogen was coded, dividing by 40. Variables AN, B^2, CA, CB, DB, HA, JA, and KA were coded, multiplying by 0.1 and A^2, BA and DA by 0.01. The symbols are defined and the estimated coefficients given in Table 1, reproduced from (3). Bothersome to one trying to interpret the estimated coefficients are the signs of the intercept and the linear effects of soil nitrogen, excess moisture, drought, depth of rooting zone, and hail. The magnitudes of the linear effects of certain variables also are not in agreement with agronomic expectation, e.g. the linear effect of blight and weeds seems large. Only looking at the linear effect can, of course, be misleading, e.g. considering all four variables involving hail and using average values for the three sources of nitrogen gives a reasonable overall estimate of the effect of hail. However, the net effect of the four blight variables is higher than agronomic expectation.

Despite the attempt to include only important site variables and depending on the estimator of experimental error to be used in hypothesis testing as discussed in (6), more than one third of the variables are not statistically significant from zero using a Type I error rate of .05. This relatively large number of non-rejections plus non-appealing signs and magnitudes of several estimates led to an attempt to reduce the full model. The stepwise regression program (5) yielded an equation with 17 variables as shown in Table 1. Even though all the variables are now significant at a type I error rate of .05, other bothersome events, such as A^2 and AN in the model but not A, now appear.

The question of comparing the reduced model with the full model arises. The R^2 for the two were nearly the same, indicating that the fit to the data was equally

well-handled by the reduced model as by the full model. Remembering that the primary objective was to develop a prediction equation, it seemed fruitful to compare the two models on the basis of how well the two estimated models would predict observations not included in the least squares estimation procedure. To some, it would appear that the full model would do better than the reduced on the basis that the extra variables must help or cannot harm since the effect of a near-zero estimate on the predicted values will be minimal. The data set was divided into halves, the full model and the reduced model previously selected by stepwise were then estimated on each half, the resulting prediction equations used to predict the other half and the squared deviations between observed and predicted added over both halves and called the prediction sum of squares. As reported in (2), the half and half procedure was repeated four times with various modifications. The average mean square, calculated by dividing the prediction sum of squares by the number of observations, for the full model was 2.01 and for the reduced model 0.74. These calculations included some data not used in the present study but the 2.01 and 0.74 are comparable to 0.38 and 0.39, respectively, the usual residual mean squares. Other divisions of the data, including estimating on n-1 observations and predicting for the n^{th}, have been calculated with the same general results. Of a surprising nature was the poor performance of the full model indicating that criteria used in arriving at a good prediction equation should be rethought.

A result not given sufficient attention by data analysts is that the variance of a predicted response cannot decrease, and usually increases, with the addition of a variable to the prediction equation. However, not including important variables gives a biased predicted value. Therefore, in an estimated prediction-equation some balance between variance and bias is desired. Looking at the previous results, it was clear that, while the full model was fine for predicting those observations used in the estimation, too many variables were included to be a good prediction equation for observations not used in the estimation procedure. To a lesser extent

the same conclusion could be made for the reduced model selected by stepwise regression. The problem here is primarily in the usual stopping rule that stops the selection procedure when a pseudo-F statistic is less than an arbitrarily chosen percentage point of the F distribution.

A natural extension to the activity of comparing models was the use of a criterion incorporating the ability of the prediction equation into the variable selection procedure. Another view would be the development of a criterion which would give different relative weights to the variance and bias of a predicted value. The C_p criterion has been used recently and is discussed in (4). We have adopted the Prediction Sum of Squares (PRESS) criterion as developed in (1) and (3). To obtain PRESS, each observation is "predicted" using all the other observations. The resulting "errors of prediction" are squared and summed to form PRESS. PRESS is appealing because it simulates prediction. It does not use an observation to aid in the "prediction" of itself. The Sequential PRESS Algorithm (SPA) presented in (1) is used to calculate PRESS for any given subset of variables and to identify the additional variable that will result in the largest reduction of PRESS.

Using the 33 potential variables and the 72 experiments presented earlier, the prediction sum of squares decreases rapidly with the first few variables to enter, followed by several variables with small increases before a minimum is reached and then concluding with an increase in the prediction sum of squares. Strictly adhering to the prediction sum of squares criterion, variables would be added to the prediction equation until the minimum is reached; however, the shape of the curve resulting from plotting the prediction sum of squares against the order of the entering variables is such that a practical decision may be made to stop bringing variables into the prediction equation earlier. The resulting prediction equation from using the prediction sum of squares (PRESS) is given in the last column of Table 1.

The SPA procedure yields a prediction equation containing all main effects

except for the drought by applied nitrogen interaction. In addition, the signs and magnitudes of the estimates are agronomically reasonable leading to a straight forward interpretation of each selected variable. One cannot expect to interpret the data by viewing the estimated coefficients in the full model. This would be almost as difficult with the variables selected by stepwise regression. Interactions are undoubtedly important in a complete understanding of the basic underlying relationships involving yield determining variables. However the last column of Table 1 selects variables and gives estimates leading to a reasonable partial interpretation concerning kind and relative size of variables important in yield determination.

Which model in Table 1 will give the best predictions can only be answered with additional experiments in future years. However, one approach would use the first three years of data, determine the important variables, calculate estimates, and predict the corn yields for the fourth year using the already known numerical values for applied nitrogen and appropriate site variables. Consequently, three prediction equations, based on the full model and two reduced models selected by stepwise and SPA, were calculated using part of the data, 228 observations from the first three years, and the remaining 60 observations from the fourth year predicted. The residual mean squares based on the 228 observations were 0.35, 0.38 and 0.42 for the full, stepwise and SPA procedures, respectively. The "residual mean squares" based on the predictions of the 60 observations not used in the estimation were 1.12, 0.71, and 0.51. Again the poor performance of the full model is noticed and the PRESS criterion has given the smallest increase when predicting for observations not included in the estimation.

REFERENCES

(1) Allen, David M. 1971. The prediction sum of squares as a criterion for selecting predictor variables. University of Kentucky, Department of Statistics, Technical Report No. 23. (Submitted to Technometrics)

(2) Anderson, R. L., D. M. Allen, and F. B. Cady. 1972. Selection of predictor variables in linear multiple regression. Chapter 1, Statistical Papers in Honor of George W. Snedecor. Iowa State University Press.

(3) Cady, Foster B. and David M. Allen. 1972. Combining experiments to predict future yield data. Agron. J. 64: 211-214.

(4) Daniel, Cuthbert and Fred S. Wood. 1970. Fitting equations to data. John Wiley and Sons, Inc., New York.

(5) Draper, N. R. and H. Smith. 1966. Applied regression analysis. John Wiley and Sons, Inc., New York.

(6) Laird, R. J. and F. B. Cady. 1969. Combined analysis of yield data from fertilizer experiments. Agron. J. 61: 829-834.

TABLE 1

THE INDEPENDENT VARIABLES AND THE ESTIMATED PARTIAL REGRESSION COEFFICIENTS FOR THE FULL AND REDUCED MODELS

(Reproduced with modifications from [3].)

Independent Variable	Symbol	Estimated Regression Coefficients		
		Full	Stepwise	SPA
Constant	K	-0.3170	-0.5446	1.5780
Applied nitrogen (linear)	N	1.8410	1.8050	1.4540
Applied nitrogen (quadratic)	N^2	-0.1552	-0.1547	-0.1528
Total soil nitrogen (linear)	A	-0.0290		0.0098
Total soil nitrogen (quadratic)	A^2	0.0150	0.0032	
A × N	AN	-0.0396	-0.0406	
Previous crop (linear)	B	0.2220		
Previous crop (quadratic)	B^2	-0.0813	-0.0176	
B × N	BN	-0.0014		
B × A	BA	0.0771	0.0711	
Excess moisture	C	0.1066	-0.2656	-0.2436
C × N	CN	-0.0374		
C × A	CA	-0.0217		
C × B	CB	-0.0794		
Drought	D	0.0309		
D × N	DN	-0.0096	-0.0091	-0.0091
D × A	DA	-0.0023		
D × B	DB	-0.0259		
Depth of rooting zone	E	-0.0054		
Soil slope	F	-0.0124	-0.0111	-0.0086
Soil texture (linear)	G	1.2800	1.2740	
Soil texture (quadratic)	G^2	-0.1591	-0.1630	
Hail	H	0.5556	0.2651	-0.2737
H × N	HN	-0.0003		
H × A	HA	-0.0802	-0.0694	
H × B	HB	-0.0159		
Blight (H. Turcicum)	J	-1.0890	-0.2733	-0.2677
J × N	JN	0.0183		
J × A	JA	0.0611		
J × B	JB	0.0139		
Weeds	L	-1.7750	-0.9231	
L × N	LN	-0.0004		
L × A	LA	0.1111	0.0757	
L × B	LB	0.0458	0.0183	