BU-382-M also BU-233

COMBINING EXPERIMENTS TO PREDICT FUTURE YIELD DATA[1]

Foster B. Cady and David M. Allen[2]

ABSTRACT

Data from a series of fertility experiments including uncontrolled environmental variables are analyzed so that future yields may be predicted. A new criterion, the prediction sum of squares which is based on the performance of the estimated equation for predicting observations not included in the least squares estimation, is developed for selecting the best predictor variables. The procedure gives an equation with a minimal number of predictor variables and agronomically reasonable estimates of the regression coefficients.

Additional index words: data analysis, regression methodology, yield prediction, pooling of experiments.

[1] Joint contribution of the Biometrics Unit, Plant Breeding and Biometry, Cornell University, Ithaca, New York, 14850, and the Department of Statistics, University of Kentucky. Paper No. BU-382-M of the Biometrics Unit.

[2] Professor of Biological Statistics, Cornell University and Assistant Professor of Statistics, University of Kentucky.

BU-382-M

COMBINING EXPERIMENTS TO PREDICT FUTURE YIELD DATA[1]

Foster B. Cady and David M. Allen[2]

Inference to a population of yields, based on relatively few experimental data points, is a key step in agronomic research. In methodology it is stressed that the realized experiments should cover a range of soil, management and climatic conditions, resulting in a series of experiments over several locations and years. Voss et al, (5) emphasize the importance of evaluating environmental factors in interpreting data from a series of planned experiments. In the analysis of combined data, both the observed yields from the experimentally controlled treatments and measured data from the uncontrolled environmental factors are used to estimate the relationship between the response variable and the various hypothesized causal factors. This paper is chiefly concerned with the development of a new procedure for the selection of variables in an equation for predicting future yield data.

THE PREDICTION SUM OF SQUARES CRITERION

A major objective in the interpretation of data from a series of experiments is the determination of a general yield equation. Suppose a

---

[1] Joint contribution of the Biometrics Unit, Plant Breeding and Biometry, Cornell University, Ithaca, New York, 14850, and the Department of Statistics, University of Kentucky. Paper No. BU-382-M of the Biometrics Unit.

[2] Professor of Biological Statistics, Cornell University and Assistant Professor of Statistics, University of Kentucky.

nitrogen fertility study with controlled levels of applied nitrogen is conducted at each of several locations. A response model between yield and applied nitrogen can be estimated at each site but an estimated model over all sites is more important. Unfortunately, the statistical fitting of all the data to one overall model is usually very poor, as exemplified by Voss, et al (5). The same general function between, say, yield and applied nitrogen, might exist for all sites but with different levels of site variables, e.g., soil nitrogen, a different portion of the function is observed at each site. Or a general yield function might be affected through real interactions between the site variable and the experimentally controlled variable. For either case, a number of uncontrollable factors do exist and these cannot be held at constant or varying levels. However, measurement can be made of these factors called site variables.

In a series of experiments designed for a combined analysis, a relatively large number of site variables are measured, not only those that the experimenter is highly certain are of value in prediction but also questionable variables are initially included. The entire data set is then used to indicate the variables to be included in the best estimated yield equation. The addition of a variable to the prediction equation almost always increases (and never decreases) the variance of a predicted response, Walls and Weeks (6). However, failure to include an important variable may result in a biased predicted observation. An ideal procedure would select variables which are important in reducing bias without selecting those which would unnecessarily add to the variance of a predicted value.

The problem can now be stated as how one may select the predictor variables in a general yield equation where some of the variables are

measured but not experimentally controlled, and are statistically correlated with each other. The commonly used selection procedures and their drawbacks for use in combined analysis of fertility experiments are discussed by Laird and Cady (4). These procedures, in general, use the residual sum of squares as the criterion for variable selection; the stepwise procedure will include a variable in the estimated model if a statistical test for the reduction in the residual sum of squares is significant at a predetermined type I error rate. The need for the statistical test is to provide a stopping rule since the residual sum of squares, in practice, will be lowered with each additional variable.

In the Laird and Cady (4) work, a criterion was needed to evaluate the usefulness of three estimated reduced models. The commonly used criteria of R^2 and the residual mean square were not sufficiently discriminating. Since the estimated yield equations are used for predicting future observations where sets of new values for the so-called independent variables become available, it seemed reasonable to divide the data into halves, estimate the parameters of the preselected alternative models with one half, predict for the other half, and finally calculate the sum of squares between the observed and predicted responses.

It then seemed reasonable to use a similar criterion in the development of a new procedure for actually selecting the variables. This new procedure would weigh the goodness of a particular potential variable to predict observations not included in the estimation of the model parameters. The new criterion, the prediction sum of squares (PRESS), is based on the difference between Y_i, the observed response and $\hat{Y}_{(i)}$, the response predicted from an estimated equation excluding the i^{th} observation, i.e., the predicted response is based on n-1 observations. The

deviations are then squared and added over the n observations, i.e.

$$\text{PRESS} = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_{(i)} \right)^2 .$$

The sequential PRESS algorithm (SPA), utilizing the new criterion, is a procedure for selecting the predictor variables. SPA is analagous to the stepwise procedure using the residual sum of squares as the criterion. The one variable having the smallest value of PRESS is entered first. At each subsequent stage, the variable having the smallest value of PRESS in conjunction with the previously entered variable(s) is entered. For example, at the second stage, that variable among the remaining potential variables giving the largest reduction in the prediction sum of squares is selected. Allen (1) gives an efficient algorithm for this sequential procedure.

When selecting variables by SPA, PRESS typically decreases rapidly with the first few variables selected. Each of the next few variables decreases PRESS by small amounts until a minimum is obtained, after which additional variables increase PRESS. The net effect, when viewing a plot of PRESS with respect to the number of variables in the equation is a backward J shaped curve with a nearly flat bottom. Continuing to the minimum appears to reduce the biases squared relative to the variances, but the variances are known to be increasing at each stage and the biases have to be estimated, suggesting that the cut-off point could be made subjectively before the minimum is observed.

RESULTS

The procedure SPA was compared with stepwise regression using data from a series of 76 experiments in the Bajio area of Mexico that were

carried out over a four year period and described by Laird and Cady (4). Four experiments were deleted in order to have a total number divisible into groups, a facility not used in the results reported here. The full model included the 33 independent variables listed in Table 1 with the estimated partial regression coefficients. Stepwise regression using a type I error rate of 5% as a stopping rule was used and the resulting estimated partial regression coefficients of the 17 entering variables are also given in Table 1. With SPA the variables entered in the order as shown in Figure 1. The prediction sum of squares drops rapidly with the first few variables to enter the equation. The figure does not show the minimum value of the prediction sum of squares of 116.668 nor the following increase to 122.315 with all variables entered. The type of function would indicate that a cut-off for stopping variables entering the equation could be made before the minimum is reached. In this example the cut-off was made after the soil slope variable entered and the estimated partial regression coefficients are given in Table 1. The constant or mean was the single variable that most reduced the prediction sum of squares at the first stage but theoretically it could have been another variable, as exemplified later.

In practice SPA is used to select predictor variables. In this situation the experimenter would like to determine those predictor variables that will predict well future responses when new sets of independent variable values become available. The Bajio data from the first three years (228 observations) were used for selecting predictor variables by both stepwise and SPA. Using a 5% significance level for a stopping rule, 15 of the potential 33 predictor values were selected by stepwise, including K, N, N^2, CA, CB, D, DN, DB, F, G, G^2, B^2, BA, H, J, and LA. The

Table 1. The independent variables (the ranges are given in parentheses) and the estimated partial regression coefficients for the full and reduced models.

Independent Variable	Symbol	Estimated Regression Coefficients		
		Full	Stepwise	SPA
Constant	K	-0.3170	-0.5446	1.5780
Applied nitrogen (linear) (0-3)	N	1.8410	1.8050	1.4540
Applied nitrogen (quadratic)	N^2	-0.1552	-0.1547	-0.1528
Total soil nitrogen (linear) (46-137)	A	-0.0290		0.0098
Total soil nitrogen (quadratic)	A^2	0.0150	0.0032	
A x N	AN	-0.0396	-0.0406	
Previous crop (linear) (10-25)	B	0.2220		
Previous crop (quadratic)	B^2	-0.0813	-0.0176	
B x N	BN	-0.0014		
B x A	BA	0.0771	0.0711	
Excess moisture (0-7)	C	0.1066	-0.2656	-0.2436
C x N	CN	-0.0374		
C x A	CA	-0.0217		
C x B	CB	-0.0794		
Drought (0-112)	D	0.0309		
D x N	DN	-0.0096	-0.0091	-0.0091
D x A	DA	-0.0023		
D x B	DB	-0.0259		
Depth of rooting zone (25-99)	E	-0.0054		
Soil slope (1-80)	F	-0.0124	-0.0111	-0.0086
Soil texture (linear)(2-5)	G	1.2800	1.2740	
Soil texture (quadratic)	G^2	-0.1591	-0.1630	
Hail (0-5)	H	0.5556	0.2651	-0.2737
H x N	HN	-0.0003		
H x A	HA	-0.0802	-0.0694	
H x B	HB	-0.0159		
Blight (H. Turcicum)(0-8)	J	-1.089	-0.2733	-0.2677
J x N	JN	0.0183		
J x A	JA	0.0611		
J x B	JB	0.0139		
Weeds (0-4)	L	-1.7750	-0.9231	
L x N	LN	-0.0004		
L x A	LA	0.1111	0.0757	
L x B	LB	0.0458	0.0183	

prediction sum of squares for SPA is shown in Figure 2. Again, after the first few variables, the curve is nearly flat as the minimum of 88.099 (not shown in the figure) is approached. With all variables included the prediction sum of squares is 93.819. As before, the cut-off was made after the soil slope variable, the linear effect of applied nitrogen giving the smallest prediction sum of squares at the first stage.

Rather surprising results from using these estimated models, including the full model, are shown in Table 2. Though the residual mean square using SPA is 10% higher than the stepwise, the future prediction sum of squares (as defined in the table) with SPA selected predictor variables is approximately 30% lower than the stepwise. As expected, the full model fits the data from the first three years better than either reduced estimated model as evidenced by the smallest residual mean square. However, the predictive ability of the full model is poor, more than doubling in the future prediction sum of squares when compared with the SPA estimated model.

DISCUSSION

The agronomist prefers a general yield equation with regression coefficients that are meaningful in both a qualitative and quantitative sense. The problems in interpreting estimated coefficients with correlated variables have been discussed by Voss et al, (5). As shown in Table 1, and conjectured in general, SPA selected variables yield estimates of the coefficients with proper signs and reasonable magnitudes. For example, SPA selected the linear effect of soil nitrogen; the estimated coefficient is positive, giving nearly a ton increase in yield for a field reasonably high in soil nitrogen. However, the stepwise procedure selected the quadratic effect of soil nitrogen and the interaction with applied nitrogen with estimates that cannot be interpreted quantitatively without

Table 2. Results from using the first three years of data (228 observations) for predictor variable selection and subsequently predicting for the fourth year (60 observations).

	Estimated Model		
	Full	Stepwise	SPA
Number of predictor variables	33	15	9
Residual sum of squares (228 observations)	67.521	80.134	91.806
Residual mean square	0.348	0.383	0.421
Prediction sum of squares (228 observations)	93.819	92.868	100.794
Future prediction sum of squares†	67.661	42.395	30.846

† Sixty squared deviations between the predicted responses, using the first three years of data for the predictor variable selection, and the observed fourth year responses summed over the fourth year.

considering the constant term and other variables. Using the estimated model selected by SPA, yield may be plotted as a quadratic polynomial function of applied nitrogen. Since the drought by applied nitrogen interaction is important, the plot would have a different estimated function for each of several drought levels. These curves would then be influenced by the main effects of soil nitrogen, excess moisture, slope, hail, and blight. Such a plot is Figure 3 using average values of soil nitrogen and slope, low values for hail and blight and zero for excess moisture.

As demonstrated, the SPA procedure can be used for analyzing and interpreting an existing set of data. However, a more important role of SPA is the selection of predictor variables in situations where the estimated model will be used for predicting yields at a later date with new sets of independent variable values. It is assumed that the new values are in the same range as those used in the predictor variable selection stage. For example, with soil test recommendations, a series of experiments can be conducted in a given area over a period of time. Based on these experiments, a prediction equation is desired which can be used in future years with soil test values, knowledge of probabilities involving climatic variables and farmer information on management variables. Now the emphasis in the variable selection procedure primarily is the selection of those variables which will do well in predicting future observations; the secondary emphasis is the selection of those variables which will do the best in reducing the residual sum of squares for that particular set of data, which is important if the interpretation of the existing set of data is the sole objective.

Most of the variables selected by SPA were also selected by the

stepwise procedure. In fact, the ordering of the first few variables is the same for the two procedures and curves for the residual sum of squares analogous to those in Figures 1 and 2 are similar for the first few variables. However, the residual sum of squares will in practice continue to decrease as more variables enter the equation. In addition the differences between the two procedures in the variables entering during the earlier stages can be important. The linear effect of soil nitrogen is an example in the data set presented here. The two procedures were also compared with data from Voss et al, (5) with 29 potential predictor variables. SPA selected the linear effect of plant population, a variable not selected by the stepwise procedure which selected several interactions in the prediction model. Of the ten variables selected by PRESS, nine were linear effects of site variables, the other being the interaction between drought and applied nitrogen.

A recommendation could be made for using the stepwise procedure but with an earlier cut-off point, e.g., using smaller significance levels than .05 or perhaps not using any hypothesis testing. It is believed that the stepwise procedure as described by referenced textbooks, e.g. Draper and Smith (3), historically is too deeply associated with hypothesis testing to make any real break with standard use. Also, it is believed that the prevailing practice is to use the larger significance levels in variable selection so that a so-called important variable would not be eliminated from the equation. This practice for including a large number of variables in an equation is based, in part, on the feeling that extra variables can't do any harm. With different approaches, it has been shown in both the Laird and Cady (4) and the present study that extra variables can indeed be detrimental for prediction purposes.

The SPA procedure, based on the prediction sum of squares criterion, is not associated with any hypothesis testing. The cut-off point is determined by the minimum value of the criterion. With small data sets, the minimum is rapidly reached as variables are entered and the increase in the criterion immediately begins after the minimum is reached. An example is given by Allen (1). With large data sets, such as the one used in this study, a practical recommendation is to use a cut-off before the absolute minimum is reached. This is arbitrary, of course, but the decision is not difficult with the rapid decrease in the prediction sum of squares before reaching the flat portion of the curve.

REFERENCES

1. Allen, David M. 1971. The prediction sum of squares as a criterion for selecting predictor variables. Submitted to Technometrics.

2. Colwell, J. D. and K. M. Stackhouse. 1970. Some problems in the estimation of simultaneous fertilizer requirements of crops from response surfaces. Australian J. of Exp. Agr. and Anim. Husb. 10: 183-195.

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

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

5. Voss, R. E., J. J. Hanway and W. A. Fuller. 1970. Influence of soil, management, and climatic factors on the yield response by corn (Zea Mays L.) to N, P and K fertilizer. Agron. J. 62:736-740.

6. Walls, R. C. and D. L. Weeks. 1969. A note on the variance of a predicted response in regression. Amer. Statist. 23:24-26.

FIGURE AND TABLE LEGEND

Figure 1. Prediction sums of squares (PRESS) when variables are selected using the SPA procedure and data from four years. Letters on the abscissa are identified in Table 1.

Figure 2. Prediction sums of squares (PRESS) when variables are selected using the SPA procedure and data from the first three years. Letters on the abscissa are identified in Table 1.

Figure 3. Predicted yield response as a function of applied nitrogen for several levels of drought.

Table 1. The independent variables (the ranges are given in parentheses) and the estimated partial regression coefficients for the full and reduced models.

Table 2. Results from using the first three years of data (228 observations) for predictor variable selection and subsequently predicting for the fourth year (60 observations).