SOME PROBLEMS OF LINEAR DISCRIMINATION[*]

DWIGHT M. BLOOD[**] AND C. B. BAKER
University of Michigan and University of Illinois

TO CLASSIFY items into like groups is a methodological problem in virtually any research effort. The purpose of classification is specified by the problem of the research. The methodological objective is to develop classification criteria with which to identify a newly observed item as a member of a group reasonably homogeneous in terms relevant to the research problem. Thus it is required that the grouping be done on the basis of measurable characteristics and that the groups be defined in terms that permit the investigator to evaluate the reliability of his (new) identification.

The simplest classification problem is one in which groups can be distinguished according to one variable reflecting a characteristic in the items to be classified. Thus it might be proposed to classify a phenomenon A into classes A_1 and A_2 according to values taken by X, which measures a characteristic of A. Specifically, A_1 and A_2 will be said to constitute separable groups if the mean of X in A_1 differs significantly from the mean of X in A_2. The significance of the difference is affected by an estimate of standard error of the difference between means and the probability criterion according to which it is agreed that the difference is sufficiently unlikely to be assumed not to have occurred by chance.

Classification requirements in most problems facing the agricultural economist cannot be satisfied with such a technique. Suppose, for example, it is required that we estimate the elasticity of wheat output with respect to wheat price. To accomplish the aggregation required by this task requires that we group the various production situations in terms of opportunities producers have. Some would respond easily to price change. Others would find it impossible to respond. Logic suggests that the relevant characteristics are those which affect the slope of marginal cost

[*] Research on which this report is based was initiated as a research project of the Montana Agricultural Experiment Station, contributing to Western Regional Project No. W-16, financed with funds authorized under Title II, Research and Marketing Act of 1946. The present paper was completed in part as a report in the Research Seminar in Quantitative Economics, University of Michigan. For a more detailed formulation of the problem, see Dwight M. Blood, "Delineating Firms Sensitive to Shifts Between Wheat and Range Forage in the Northern Great Plains," Montana Agricultural Experiment Station Mimeo Cir. 84, September, 1954. A follow-up study to estimate conditions necessary to shifts in land use is reported by Terry N. Norman in "Forage-Crop Substitutions on Dryland Units Sensitive to Shift," Montana Agricultural Experiment Station Mimeo Cir. 92, March, 1956. Credit is due Professor E. H. Ward, Montana State College, for assistance in developing the project.

[**] On leave, Colorado State University.

functions of wheat output within firms in each group over the ranges of price and time specified by the problem.[1]

One might commence with the widely used "type-of-farming" areas.[2] Yet such a classification is demonstrably weak for research in the supply-response study. Boundaries are established with data reflecting decisions already made, not opportunities available. Moreover, the boundaries tend to be so drawn as to enclose geographically defined areas, whereas production opportunities may sometimes be more homogeneous with respect to nongeographic factors. That is, they may be found in noncontiguous areas. Finally, the use of a classification system that yields exclusive classes based on subjective judgment may result in a large (and unknown) percentage of misclassification.

When it is not possible to classify items on the basis of a single characteristic (variable), it is necessary to adapt some techniques to take account of the combined effects of the several variables that will distinguish the items by groups. In this paper we propose to explore three techniques to delineate production situations in the Northern Great Plains which favor (1) wheat production or (2) range forage production. The techniques are:

1. Linear multiple regression, where the dependent variable is expressed as percentage of acres in harvested wheat.
2. Linear discriminant function, which provides an index for classifying individual observations into exclusive categories.
3. Linear probability function, which provides a calculated probability of being a wheat producer for each observation, on which basis the two-way classification can be made.

An important facet of the problem is to be able to so use known information about individual ranches as to accomplish this taxonomic task with a known probability of error in misclassification. The extent to which this objective is met will influence the classification method in actual policy and production problems—for example, identifying "response-likely" firms as a basis for intensive study of the shifting process.

Description and Setting of the Problem

The Northern Great Plains is a region of extremes. Operators of farms and ranches in the area face a complex array of managerial decisions.

[1] This is a static formulation of the supply response problem. However, there is no conceptual reason for not including dynamic factors which change the marginal cost functions. Changes in technology and in quality of management have been suggested by T. W. Schultz in "Reflections on Agricultural Production, Output and Supply," *Journal of Farm Economics*, Vol. 38, No. 3, August, 1956, pp. 748-762.

[2] F. F. Elliott, *Types of Farming in the United States*, U. S. Department of Commerce, Bureau of Census, 1933.

But unlike most areas, where climatic, economic, and institutional changes create diverse shifts in land use, the typical Northern Great Plains dryland operator, if confronted with any shifting alternatives at all, is ordinarily restricted to grain and/or livestock production. Instead of adding new enterprises, attempts are often made to develop intrafirm flexibility.[3]

A decision as to which type of production will be followed will affect the operation of the farm or ranch for many years to come. A new supply of range forage and a foundation breeding herd cannot be acquired overnight. Acreage allotments for grains may be based on historical land use. Adjustments in production may therefore become very "sticky." Yet, during World War I wheat acreage expanded from 56 million acres in 1914 to 74 million in 1919 (harvested basis) and remained above 60 million throughout the 1920's.[4] During World War II, plantings of wheat were increased from 53 million acres in 1942 to 84 million in 1949.

Following this, wheat acreage tapered off briefly only to increase again to 78 million acres after the outbreak of Korean hostilities.[5] It is a matter of considerable concern that important margins of transference be located between wheat and competing uses of land. Margins of transference bounding the population of wheat farms are comprised of those which separate wheat farms from farms in which land is used (1) more intensively and (2) less intensively. Prominent in the latter group are stock ranches using land to produce range forage for livestock production.

Sources of data

The observational basis for this study consists of records taken in surveys made with two random samples: one of units in a predominantly wheat producing area in northeast and northcentral Montana,[6] and one of units in a predominantly range livestock area in southeast Montana, northeast Wyoming, and western South Dakota.[7] The combined samples provide data for 274 operating units.

[3] See Emery N. Castle, "Flexibility and Diversification as a Means of Meeting Price and Yield Uncertainty in Western Kansas," *Journal of Farm Economics*, Vol. 36, No. 2, May, 1954, pp. 273-284.

[4] Warren R. Bailey and Charles W. Nauheim, "Prospective Adjustments in Wheat Farming," U. S. Department of Agriculture, B.A.E., (Mimeo report prepared for distribution at 13th Annual Agricultural Outlook Conference, Washington, D.C., October 28, 1953), p. 1.

[5] Warren Bailey and Charles Nauheim, *op. cit.*, pp. 1, 4.

[6] See Darrell F. Fienup, *Resource Productivity on Montana Dryland Crop Farms*, Montana Ag. Exp. Sta. Mimeo Cir. 66, June 1952. The universe is defined as Montana Type of Farming Areas III and IV (Northeast Montana: spring wheat, nonirrigated) and VI and VII (North-Central Montana: mixed spring and winter wheat, nonirrigated).

[7] See James R. Gray and C. B. Baker, *Organization, Costs and Returns on Cattle Ranches in the Northern Great Plains*, 1930-1952, Montana Ag. Exp. Sta. Bul. 495,

Series in the following three variables were then compiled:

X_1 = total acres operated
X_2 = acres of land leased divided by acres of land owned[8]
X_3 = total annual precipitation in inches

Series in X_1 and X_2 were obtained from the survey schedules. Data for X_3 were obtained from published reports of the U. S. Weather Bureau. We emphasize at the outset that observations were drawn from units in known classes. The purpose of the following analytical alternatives is to so use this information as to obtain a device for classifying a newly observed unit on the basis of observations as in X_1, X_2 and X_3.

The Empirical Analysis

The linear multiple regression equation

The first step in the analysis was to fit by least squares a linear function of the form:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3,$$

where Y = per cent of total acres operated in grain harvested, and X_1, X_2, X_3 = size in acres, ratio of acres of land leased/acres of land owned, and annual precipitation in inches, respectively.

The calculated estimates of these parameters are:

$$Y = -4.72 - \underset{(0.00018)}{0.0014X_1} + \underset{(0.033)}{0.046X_2} + \underset{(0.36)}{1.67X_3} \qquad (1)$$

$$R^2 = 0.27 \qquad \bar{X}_1 = 4{,}387 \qquad \bar{X}_2 = 35.86 \qquad \bar{X}_3 = 16.88$$

The estimates of the standard errors of the regression coefficients are represented by the figures in the parentheses immediately below the parameter estimates. Although this linear combination of independent variables accounts for only 27 per cent of the variation in number of acres of wheat harvested, it is nonetheless significant with N = 274. Although b_2 (ratio of leased-owned land) does not appear significant at the 0.05 level of confidence, the coefficients for X_1 and X_3 (total acres and precipitation, respectively) could hardly have arisen due to chance. The signs of the coefficients b_1 and b_3 are consistent with logically based expectations.

December, 1953. The universe consisted of range livestock units which received at least 50 per cent of their gross income from sale of beef cattle, and which met certain other criteria, for areas listed in the text.

[8] As defined, this variable can be used only where there exists for all observations an acreage owned greater than zero. In classifying farms in populations including wholly leased farms, this variable would need to be defined differently (e.g., leased land/all land).

Although this equation may be used as a device for predicting acreage of wheat harvested, some critical problems arise when it is used as a classificatory device. In the first place, any selected percentage of land in grain harvested to be used as a discriminating index would appear to be purely arbitrary, particularly as a means of identifying marginal firms.

Second, it is extremely difficult to formulate *any* quantitatively expressed dependent variable for regression analysis which will be satisfactory for identifying marginal firms. For example, in this equation, the dependent variable measures harvested grain only and excludes fallow land as well as land used for other crops. These exclusions might be made trivial by the assumption that land in wheat is approximately twice that yielded by the predicting equation and by introducing a constant to allow for acreage loss between planting and harvest. Nonetheless, such definitional problems are critical in making a realistic classification. Third, this formulation does not provide a convenient means of distinguishing firms susceptible to changes in classification from those that fall without question into one of the two definite categories.

The standard linear discriminant function

As an alternative classification device, the next step was to calculate a linear equation of the form:

$$Z = a_1X_1 + a_2X_2 + a_3X_3.$$

The weights a_1, a_2, and a_3 are so determined as to maximize the ratio of the variance of Z between groups to the variance of Z within groups. Thus the index Z is an optimum linear discriminator between the groups. In applying the index, a critical level of Z ($\bar{A}$ in Figure 1) is set halfway between the means of Z for the two groups. Any item—in this case firm—with a Z index higher than the critical level is classified in one group; those with Z index values below the critical level are assigned to the other.

This technique is borrowed from the biological sciences where it was developed by R. A. Fisher in 1936 for the purpose of classifying plant specimens.[9] Two representative studies are mentioned as indicative of its application in the field of economics. Durand[10] used the technique

[9] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, Vol. 7, Pt. 2, 1936, pp. 179-188. Fisher's work on discriminant functions was continued in two further papers published in the *Annals of Eugenics;* "Statistical Utilization of Multiple Measurements," Vol. 7, 1938, pp. 376-386; and "The Precision of Discriminant Function," Vol. 10, 1940, pp. 422-429.

[10] D. Durand, "Risk Elements in Consumer Installment Financing," Financial Research Program, Studies in Consumer Installment Financing 8, National Bureau of Economic Research, New York, 1941, p. 125.

to differentiate between good and bad loans on the basis of a body of financial data. Tintner[11] attempted to distinguish between prices of producers' goods and prices of consumers' goods by using information about the behavior of each class of goods throughout the business cycle.

In the case at hand, the computed estimate of the discriminating equation is:

$$Z = X_1 - 63.9X_2 - 1881.48X_3 \qquad (2)$$

where variables X_1, X_2, X_3 are identical in meaning to those in equation (1).

It must be ascertained whether or not a significant difference exists between the two samples for the function Z. This test of significance is summarized in Table 1.

TABLE 1. ANALYSIS OF VARIANCE OF Z BETWEEN AND WITHIN GROUPS

Source of variation	D.F.	S.S.	M.S.	F.
Within groups	270	.00671	.000025	
Between groups	3	.00305	.001016	40.92
	273	.00976	.001041	

Referring to the F table with $n_1 = 3$ and $n_2 = 270$, the function is seen to be significant at the .001 level. Also, by definition, no other linear combination will do a better job of discriminating between these two groups with the same data. A critical zone for Z can then be established as a region within which units are sensitive to shift in classification between those producing wheat and those producing cattle. This region is obtained by calculating a mean discriminating index for each group ($\bar{a}_w = -34{,}643.8$ and $\bar{a}_c = -23{,}537.1$ for wheat and cattle respectively) and by taking the unweighted mean of these two indexes to obtain $-29{,}050.5$[12] as the critical dividing line, $\bar{A}$, between the two groups.

The results from computing a Z for each of the 274 sample members are summarized graphically in Figure 1. This diagram shows the effects of the variables used in determining the margin of reclassification. The areas on either side of $\bar{A}$ (between $\bar{A}$ and $\bar{a}_c$ and between $\bar{A}$ and $\bar{a}_w$) are the crucial areas in this example. Firms falling into these zones are "sensitive" firms, susceptible to reclassification in the opposite direction from which they are classified according to values taken by variables in the discriminant function.

[11] G. Tintner, "Some Applications of Multivariate Analysis to Economic Data," *Journal of the American Statistical Association,* Vol. 41, 1946, p. 476.

[12] In further application, the index could be made easier to use by setting this value for Z equal to 100 and making appropriate adjustments in the coefficients.

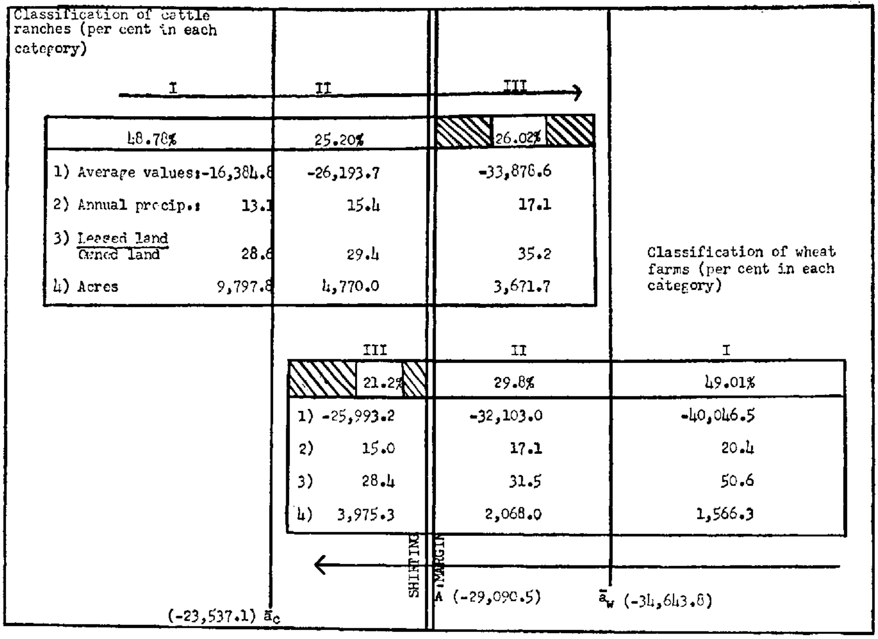

FIG. 1. CLASSIFICATION OF 274 SAMPLE RANCHES ON THE BASIS OF THEIR CALCULATED Z VALUES.

The upper and lower boxes represent the classification of cattle ranches and wheat ranches into three categories each on the basis of their calculated Z values. These categories, in both cases may be characterized as follows, on the basis of the data used:

Category I: Firms which can unquestionably be classified as cattle or wheat producing firms.

Category II: Firms which approach the "borderline," but still possess sufficient advantage in their own area to remain there.

Category III: "Borderline" cases which are characterized as susceptible to reclassification inasmuch as Category III for wheat (cattle) closely parallels average conditions for Category II for cattle (wheat).[13]

About one-fourth of the cattle ranches overlap into Category II of wheat farms and about one-fifth of the wheat farms overlap Category II of cattle ranches. The difference suggests that of units sensitive to shift, more have in fact shifted to wheat production than to range forage and cattle production. Over-all, the function is seen to have classified approxi-

[13] With one notable exception: the average acreage for cattle producing units which are "borderline" cases is approximately 1,500 acres greater than for wheat producing units falling into this same category.

mately 77 per cent of the 274 sample members into the group to which they belong.

One variable deserves brief additional comment. One could infer from some of the literature that yield might react as a linear function of moisture, over a limited range which has as its lower limit the minimum amount of moisture needed for plant survival.[14] If the average precipitation in cattle ranch Category I is arbitrarily raised from 13.12 to 20 inches, all other factors held constant, the value of Z would rise sufficiently to result in classification of these units as wheat farms. However, average precipitation in cattle ranch Category II need only be raised from 15.35 inches to about 17.5 inches to overlap into the wheat producing area. Although this is a rather shaky basis from which to assert that Z might possibly be a nonlinear function of the precipitation variable, it might be worthwhile to consider recomputing similar functions in this problem area with a transformation of the precipitation variable.

As shown previously the linear discriminant function provides a much more efficient and accurate classification in this illustration than does the regression equation as previously formulated. Also, the use of the discriminant function avoids the problem of defining a quantitatively expressed dependent variable, as in regression analysis. The theoretical similarities between the two techniques are, of course, substantial and have been well summarized by Kendall.[15] All that needs to be noted here is that the coefficients in both computed equations are of approximately the same relative magnitude and thus have approximately the same impact upon the outcome in each equation.[16]

The linear probability function[17]

The third linear formulation of the same three variables, known as a linear probability function, was then computed. This formulation involves the computation of a linear equation which will provide a calculated probability that any given unit belongs in a specific category.

We define a unit as a cattle or wheat ranch if the dependent variable, Y, a dummy variable, is, respectively, 0 or 1. The linear probability function is then easily obtained as the least squares regression:

$$Y = c_0 + c_1X_1 + c_2X_2 + C_3X_3.$$

[14] E.g., see O. R. Mathews and John S. Cole, "Special Dry Farming Problems," *Soils and Men*, U. S. Department of Agriculture Yearbook, 1938, p. 684.

[15] Kendall, Maurice G., *The Advanced Theory of Statistics*, New York: Hafner Publishing Co., third edition, Vol. 2, 1951, pp. 344-346.

[16] If the regression equation were transformed by adding the constant term to both sides of the equation and by dividing through by the leading term (the coefficient of X_1), then the signs would be the same for the coefficients in both equations also.

[17] This section is based on Daniel B. Suits, "Linear Probability Functions and Discriminations," discussion paper of the Research Seminar in Quantitative Economics, University of Michigan, October, 1957.

The calculated value of Y for any firm is then defined as the numerical probability that it is a wheat producing firm.

With a probability of one-half serving arbitrarily as the discriminating index, those units with a calculated probability of more than one-half would be classified as wheat ranches; those less than one-half would be classified as cattle ranches.

The parameter estimates for this formulation are:

$$\begin{aligned} Y &= -.28 - \underset{(.0000043)}{.000028X_1} + \underset{(.00077)}{.0018X_2} + \underset{(.007)}{.05X_3} \\ R^2 &= .31 \end{aligned} \tag{3}$$

Adding the constant term to both sides and dividing the entire equation through by the coefficient of X_1, we obtain

$$-27{,}845 = X_1 - 63.92X_2 - 1{,}881.48X_3 \tag{3a}$$

The coefficients of X_1, X_2 and X_3 are now *identical* with those of the discriminant function showing that the relative weights assigned the variables are exactly the same for the two functions. The critical value of −27,845 differs from the −29,090.5 assigned by the discriminant function. Hence, the discriminating index for the linear probability function would tend to classify more operators in the wheat category than does the discriminant function. The difference in the discriminating margins arises because the critical level of Z in the discriminant function is established empirically by the samples. The discriminating value of the index yielded by the probability function is set *a priori*. In general, they will not agree. Suits discusses in detail the possible reasons for differences in the final classification in comparing the two techniques. He also sets forth the foundation of the equivalence of the coefficients in both types of equations.[18]

In comparison with the computed regression equation, it will be noted that the coefficient of X_2 (ratio of leased land to owned land) is now significant at the 0.05 level of confidence, and that the R^2 has increased by a slight amount. On the basis of the standard errors we may conclude that this linear formulation may be used to predict the probability that any ranch chosen randomly from our population would be a wheat ranch.

The important thing to note, however, is that despite a slightly different assignment, the probability formulation is for all intents and purposes identical to the linear discriminant function. The probability formulation also avoids the difficulty that arises in defining a relevant dependent variable in ordinary regression analysis such that an objective classification can be made. Moreover, there is more meaning in a calculated prob-

[18] Daniel B. Suits, *op. cit.*

ability number for many areas of economic analysis than in a number calculated for the sole purpose of making a dichotomous classification.

Summary and Conclusions

We have attempted to examine a classification problem which, on the surface, seemed rather intractable. Three related tools of discrimination were applied in a rather summary and cookbook fashion, and the results compared:

1. *Linear regression:* this technique involved the computation of a value for the dependent variable that could be used to classify items into two groups. In the formulation presented, the equation served to aggregate the acres of wheat harvested for all of the 274 ranches, and on the basis of three independent variables, predict how many should in turn be parceled back to each one. But any classification, in this example at least, would have to be purely arbitrary as between cattle and wheat ranches.

2. *Linear discriminant function:* this method provides a device with which a new observation can be taken at random and placed either in the wheat box or the cattle box based on an empirically determined function.

3. *Linear probability function:* this method provides a calculated probability that any given observation would be a wheat farm. Having obtained this information, we could proceed to classify observations into two groups on the basis of the calculated probability that each observation would be a wheat farm. Since the calculated probability that any given observation will be a wheat farm is meaningful in and of itself, this formulation is clearly preferable to the linear discriminant function.

No pretense is made that the foregoing analysis constitutes an adequate treatment of the real classification problem—that of delineating firms sensitive to shifts between wheat and range forage production in the Northern Great Plains. For example, one might criticize the inclusion of objective and behavioral factors in the same discriminating equation. Perhaps the analysis might be improved by using only these factors which the farmer has to take as given and which cannot be changed in a substantive fashion by his own behavior.[19] The shortcomings of the analysis should not detract from the purpose of the paper, however, which was to demonstrate the application of these classificatory devices in the field of agricultural economics. Within the general problem area, a fruitful area of investigation might be that of adapting these and related techniques to the task of locating "boundary situations" and defining "type-of-farming" areas. This approach might provide the basis for selecting noncontiguous areas within a class with demonstrable criteria.

[19] Yet it might be argued that in many real situations operators in the Northern Great Plains can over fairly long periods of time do little more about size of firm and leased acreages than they can about precipitation.