PREDICTING TRANSPORTER'S CHOICE OF MODE

Brion R. Sasaki, U.S. Office of Management and Budget

Every investment decision for transportation projects requires an extensive examination of the amount of anticipated traffic. A statistical technique, discriminant analysis, was used to determine its feasibility and applicability in estimating future traffic. Discriminant analysis is a method to statistically weigh transportation characteristics. This paper discusses an application of discriminant analysis in which travel demand is divided between transportation modes on the Ohio River. This study uses time of transit, distance of transit, annual tonnage, average shipment size, transportation rate, and handling charges as mode characteristics. An increase in the transportation rate, the most significant characteristic influencing mode choice by the user, was simulated (everything else held constant) so that a demand curve for barge transportation could be constructed.

•FEDERAL investment to support the construction and maintenance of a transportation facility requires the development of projections to estimate the traffic that will use the facility. Based on these traffic projections, an analysis can be performed to determine the benefits of the proposed investment. Hence, estimating the amount of future traffic is a key portion of the investment decision. Numerous methods such as rate comparison, linear programs, and linear regressions have been developed and implemented on this subject. A most promising method that appears to be gaining recognition is discriminant analysis.

This paper describes the basic concept and usefulness of discriminant analysis as a tool for economic research and then presents an empirical example to demonstrate these capabilities.

In discriminant analysis, a linear function is established to separate a universe into predetermined populations or groups. Then a set of observations that possess the most similar a priori characteristics is assigned to a population. To simplify the analysis for this discussion, the paper only summarizes the mathematics of the two-population case.

TWO-POPULATION CASE

The two-population case is confined to the allocation of a random sample (of attributes of the universe) into one of two populations having known probabilities (10). Assume a single variate case x_1 that has two distributed populations with known means of u_1 and u_2 and a similar standard deviation for both populations, where u_1 represents the mean of variable x_1 for population 1 and u_2 represents the mean of x_1 for population 2 (Figure 1). To allocate attributes from the random sample to the proper population requires that the means not be equal. The boundary line between the populations is the arithmetic mean Z of the total sample.

For $u_1 < u_2$, the natural method of separating permits an observation to be placed into population 2 if the value of x_1 is greater than $\frac{1}{2}(u_1 + u_2)$ and into population 1 if x_1 is less than $\frac{1}{2}(u_1 + u_2)$. In other words, if $x_1 < Z$, the random observation will be placed in population 1; if $x_1 > Z$, it will be placed in population 2.

As can be seen in Figure 1, the two populations are obviously separated. However, two types of possible misclassification exist as indicated by the area of overlap. In this area, some population 1 observations are included in population 2 and vice versa.

The misclassification occurs because the tails of each distribution overlap, and misclassification will occur whenever

$$\frac{x - u_1}{\sigma} > \frac{\frac{u_1 + u_2}{2} - u_1}{\sigma} = \frac{u_2 - u_1}{2\sigma} = \frac{\gamma}{2\sigma}$$

where $\gamma = (u_2 - u_1)$ = the distance between the means.

Increasing the distance between the two means further separates the populations and reduces the overlap. This divergence minimizes the number of misclassifications. To widen the split requires more than one variable. Let us examine a multivariate case. Assume that there exist a number of variables normally distributed by x_{iw} for i = 1, 2, ..., P and w = 1, 2, ..., n, which classifies the universe into two populations by separating the means of the two populations designated by

$$\delta_i = \bar{x}_i^1 - \bar{x}_i^2$$

To discriminate between the means, a linear function is developed that separates the two sets of variables (12).

$$Z = a_1d_1 + a_2d_2 + \ldots + a_pd_p$$

This function Z should be the maximum relative to its variance, and the variance must be proportional to

$$B = \sum_{i=1}^{P} \sum_{i=1}^{P} k_i k_m \sum_{w=1}^{M} x_{iw} x_{mw}$$

Keeping the variance constant and forming a Lagrange multiplier yield a maximum of

$$F = Z^2 - \lambda Q \sum_{i=1}^{P} \sum_{i=1}^{P} k_i d_i d_m - \lambda k_i k_m \sum_{w=1}^{n} x_{iw} x_{mw}$$

This function can be differentiated partially with respect to k_m (m = 1, 2, ..., P). It can be simplified to obtain

$$d_m \sum_{i=1}^{P} k_i d_i = \lambda \sum_{i=1}^{P} k_i \Sigma x$$

Determining the k_i that are proportional to the estimates of the coefficient of the linear function allows the function to discriminate best between the two populations. This procedure divides the two populations by constructing an average Z-value that is equivalent in purpose to the previously discussed Z-values. This value can be obtained by

adding all x_i variables for both groups and dividing by the number of cases to yield an overall general average for each x_i. Inserting these values into equations results in the general Z-values. If the observation has a Z-value less than the average Z-value, the sample is placed in population 1. If the observation is greater than the average Z, the sample is placed in population 2.

Given the discriminant function, a demand analysis for each population can be estimated by varying only one variable for the desired population and holding everything else constant (2). This process causes that population to shift toward the other, which increases the overlap and increases the probability of observations being misclassified. The economic interpretation is that, as a particular (price) variable increases in magnitude while other variables (quantities) are held constant, demand for that population decreases.

EMPIRICAL ANALYSIS

As a demonstration of this procedure, the model predicts the mode choice of a set of users and estimates the demand for barge transportation. Data for this analysis were collected during the summers of 1970 and 1971 and adjusted to reflect future modal characteristics (3, 4, 9). The data consisted of 92 actual coal movements within the Ohio River Basin by rail and barge. Each observation consists of six characteristics of rail and barge movements. These characteristics are annual tonnage per year x_1, distance of transit x_2, time of transit x_3, average shipment size x_4, transportation rate x_5, and handling charges x_6. A Univac 1108 executive computer and a 07M biomedical (BMD) computer program were used to perform the calculations (5). One of the main features of this program is that it enters the variables in a sequential order depending on their statistical significance. In this run, the actual transportation price proved to be the most important determinant in separating the populations. The pattern of entrance of the remaining variables is given in Table 1.

Table 1 also gives the mean values of each variable for the two modes. The dissimilarity of the mean values of the variables gives some indication of their use in classifying firms by mode. Discriminant analysis bases the separation of modes of transportation on the dissimilarity of the mean values of common variables and the order of importance. Thus the larger differences between earlier entering variables assist more significantly in classifying the user correctly. In this case, the cost of transporting coal enters the analysis first and displays a wide variation between the two modes. Railroad prices exceeded barge line prices for transporting coal on the average by more than four and a half times. Also, average transit time was approximately 50 percent longer by rail than by barge. These two dissimilarities and others indicate that the modes can be fairly well separated.

The second major output of the BMD program is the mode classification printout. This output tabulates the results of the analysis. The diagonal of the matrix indicates the modes of transportation correctly classified, and all modes off the diagonal are misclassified. The results of the aggregate analysis are as follows:

Mode	Observed	Estimated
Barge	53	53
Rail	39	34

Barge movements are perfectly classified, but the rail movements are not. Five rail movements are statistically categorized as barge movements. These errors occur because the observations exhibit characteristics more common to barge than rail. Closer examination of the data reveals that all misclassified movements are actually unit train movements. Values of the critical variables (annual shipment size, average shipment size, and time of transit) for train movements exceed one deviation from the rail aver-

Figure 1. Separation of populations.

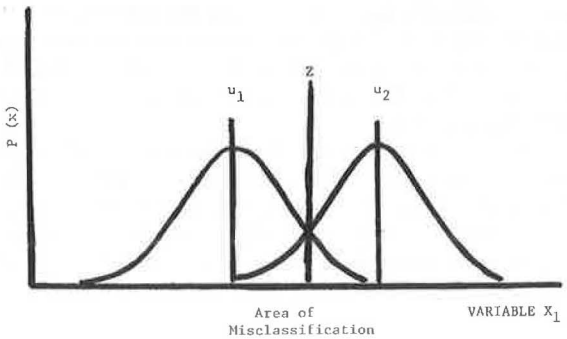


Table 1. Entrance order of variables and mean values.

Entrance Order of Variable	Description	Mean Value	
		Rail	Barge
x_5	Transportation rate, dollars	3.26	0.72
x_3	Time in transit, hours	92.08	62.2
x_2	Haul distance, miles	145.7	159.5
x_4	Average shipment size, tons	1,551	9,017
x_6	Handling charges	0.36	0.29
x_1	Annual tonnage/year	53,638	44,583

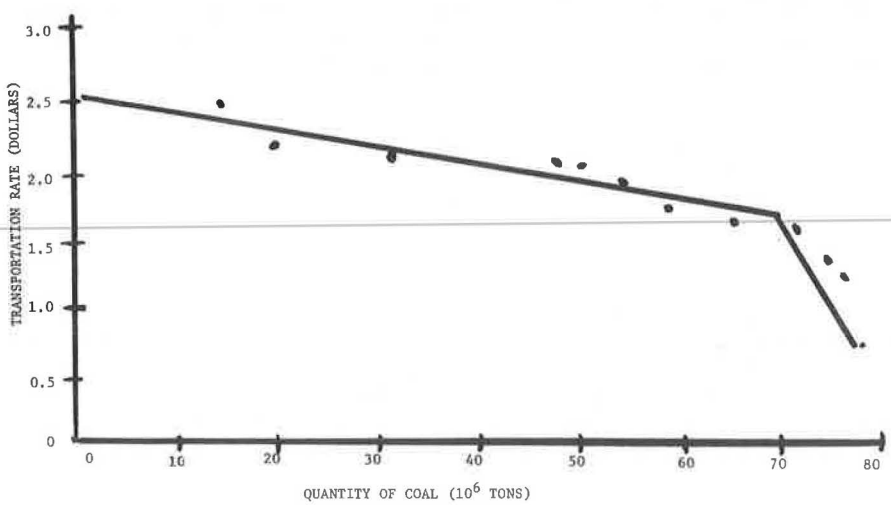
Note: 1 mile = 1.6 km; 1 ton = 907 kg.

Table 2. Demand schedule for barge transportation.

Price Increase	Classification of Firms		Barge Quantity (tons)	Price Increase	Classification of Firms		Barge Quantity (tons)
	Barge	Rail			Barge	Rail	
0.72	53	0	76,628,908	1.72	33	20	58,476,417
1.12	53	0	76,628,908	1.92	28	25	54,418,975
1.22	51	2	76,352,896	2.07	23	30	47,530,267
1.32	49	4	74,950,461	2.12	18	35	31,264,948
1.52	43	10	71,176,824	2.22	15	38	20,093,363
1.62	39	14	65,565,723	2.52	12	41	10,495,132

Note: 1 ton = 907 kg.

Figure 2. Demand curve for barge transportation.



age for those variables. In fact, the numerical values of these variables approach the barge mean values. Hence, the combination of these factors places these rail movements into the barge group. Because the model only misclassifies 5 percent of the sample and those movements can be explained, this method appears to be quite acceptable in predicting the mode of transportation a user will select.

DEMAND ANALYSIS

Manipulation of the data and the model enables a simulated demand curve for the barge transportation to be derived. This method uses a basic economic technique in which barge prices are altered while everything else is held constant. Implementation of this method indicates the responsiveness of barge demand to the alteration in prices. Plotting the demand for barge transportation at different prices produces a simulated demand for barge transportation (8).

Shifting the barge transportation price toward rail average transportation price reduces the difference between the two transportation rates, and each set of modal characteristics begins to more closely resemble the other. Consequently, the overlap of modal characteristics results in barge users being classified as rail demand. These misclassifications are economically interpreted as a decrease in demand for barge transportation because of the increased transportation price. Continuing to raise the price of barge transportation will eventually result in all anticipated barge users being allocated as rail demand.

DEMAND ANALYSIS FOR AGGREGATE DATA

Simulating the barge transportation price (positively) for the aggregate data results in a truncated demand curve. This curve consists of inelastic and elastic sections. The truncated point (kink) connects the two linear sections, which forms a simulated demand curve for barge transportation. The inelastic section stretches from the initial price of $0.72 to a total price of $1.67. The next point on the demand curve represents the unitary elasticity point. Points above the $1.68 level display an elasticity coefficient greater than one. From Table 2, these points can be identified and plotted (Figure 2).

Within the inelastic section of the barge demand curve, the simulation technique estimates that 8.82 million tons (8.0 Mg) of coal will be moved by rail. However, the majority (72 percent) of users remain with the barge mode. This implies that the barge users find it economically more advantageous to absorb the additional barge costs than to switch modes. Continuing to increase barge prices eventually results in the unitary elasticity point. In the aggregate case, the kink lies between the $0.90 and $1.00 increase in the average barge rate. Estimation through graphic technique yields a value of $0.95 ($1.67 transportation rate) for the kink point. All positive values, increases above the kink point, for barge prices are considered part of the elastic portion of the demand curve, for the demand for barge transportation in this section displayed an elasticity coefficient greater than one. Increasing barge transportation price continues until the demand for the barge mode reaches zero. Thus, this procedure enables the derivation of the elastic section of the demand curve. Extrapolating the elastic section of the demand curve estimates the last point of the demand curve at $2.65. Figure 2 shows that barge transportation price did not become equivalent to the average transportation price of rail before the barge elasticity coefficient exceeded unity. In fact, the barge simulated transportation price only attained 51 percent of the average transportation price of rail before the kink point occurred.

In conclusion, discriminant analysis has been demonstrated to be able to predict the observed behavior of users. This method also permits other monetary and nonmonetary variables to be included in the analysis to determine mode choice and provides a statistical technique for estimating the sensitivity of mode choice to each of the modal characteristics.

ACKNOWLEDGMENT

This work was completed while the author was employed as an economist with the U.S. Army Engineer Institute for Water Resources. The views expressed are solely the author's and do not necessarily represent the views of the Corps of Engineers or Office of Management and Budget.

REFERENCES

1. T. W. Anderson. An Introduction to Multivariate Statistical Analysis. John Wiley and Sons, 1958.
2. L. G. Antle and R. W. Haynes. An Application of Discriminant Analysis to the Division of Traffic Between Modes. Institute for Water Resources, U.S. Army Corps of Engineers, Rept. 71-2, 1971.
3. C. A. Berry, J. Cannelli, and B. Sasaki. IWR-ORD Transportation Mode Data Study. Ohio River Division, U.S. Army Corps of Engineers, 1970.
4. C. A. Berry, J. Cannelli, J. Dworkin, L. Elliott, L. Kuhn, and B. Sasaki. IWR-ORD Transportation Mode Study. Ohio River Division, U.S. Army Corps of Engineers, 1971.
5. W. J. Dixion, ed. BMD Biomedical Computer Programs. Univ. of California Press, Berkeley, 1968.
6. D. A. Gansner, D. W. Seegrist, and G. S. Walton. A Technique for Defining Subareas for Regional Analysis. Growth and Change, Vol. 2, No. 4, Oct. 1971.
7. M. G. Kendall and A. Stuart. The Advanced Theory of Statistics, 2nd Ed. Hafner, 3 vols., 1968.
8. L. N. Moses, ed. Cost-Benefit Analysis for Inland Navigation Improvements. Institute for Water Resources, U.S. Army Corps of Engineers, Rept. 70-4, 3 vols., 1970.
9. B. Sasaki. A Regional Model of the Future Demand for Transportation: The Case of Barge Transportation. Institute for Water Resources, U.S. Army Corps of Engineers, Paper 74-P3, 1974.
10. G. W. Snedecor and W. G. Cochran. Statistical Methods, 6th Ed. Iowa State Univ. Press, 1968.
11. M. M. Tatsuoka. Multivariate Analysis: Techniques for Educational and Psychological Research. John Wiley and Sons, 1971.
12. G. Tinter. Econometrics. John Wiley and Sons, 1965.