EARL W. KEHRBERG
Purdue University

Chapter 8

Some Problems Involved in Fitting Production Functions to Data Recorded by Soil-Testing Laboratories

IF all variables were known and measurable there would be one production function for each crop. All soil types, soil conditions, and fertilizer techniques could be thrown together. However, it is practical to estimate a different production function for each technique of fertilizer application and each separate soil type. For example, row application and broadcast methods may achieve different results from the same levels and combinations of fertilizer. Evidently some variable, such as the distance the plant travels to obtain the fertilizer, is involved. Since present knowledge does not furnish good scales for the effects of and the measurement of such factors, it is necessary to limit our estimates of production functions to homogeneous classes of application situations.

Data to derive production functions of the type hypothesized are limited. The Soil Testing Laboratory has considerable information available in its records regarding the fertility status and yields of individual fields. A complete cropping history for the past year or two and information about drainage, slopes, textures, and fertilizer application are on record.[1] Controlled experiments of a design and scope to secure the necessary data would be more desirable from the viewpoint of insuring the range in data needed to estimate the production surfaces (function). Furthermore, all measurements and sampling techniques could then be supervised by trained personnel. Noncontrolled variation could be reduced to a minimum through proper experimental design. However, the soil testing data have the advantage of availability in quantity and for a period of years. Many of our laboratories collect samples each year, running into the tens of thousands. The data, therefore, warrant examination as to the possibilities for production function derivation. First, however, some attention should be paid to the specific nature of the function to be fitted.

Crop Yield Functions

While there are few problems of mathematical function selection

[1]Reference throughout this paper to specific information available in Soil Testing Laboratories is based on conditions at Purdue University. The situation varies to some extent from laboratory to laboratory.

which are peculiar to estimating crop production functions from soil laboratory records, the selection of variables is limited to those for which information is recorded. Production function estimation from soil test records is limited to the crops for which yield estimates are recorded. Corn, soybeans, and wheat are among these. Hay yields are not often recorded. Fertilizer applications are sufficiently well recorded so that the nutrient elements N, K_2O, and P_2O_5 applied the year of the crop and the year before may be estimated. Soil test results are available for P_2O_5, K_2O, and pH but not nitrogen. Hence, a production function may be derived which states that corn yield depends upon soil nutrient levels of K_2O, P_2O_5, fertilizer elements N, P_2O_5, and K_2O applied in each of two years, and pH. Further information is available to sort the data on the basis of soil type, texture, drainage, past cropping history, and other such variables as will permit a fairly homogeneous grouping. Technique of fertilizer application or machine used for application is also recorded. Plant population, soil nitrogen, and moisture are among the more important variables for which information is lacking. At the same time, certain peculiarities of the data give rise to statistical problems of deriving any specific function chosen.

Peculiarities of Soil-Testing Laboratory Data and Their Implications in Fitting Production Functions

THE SAMPLE

Soil samples are sent to the soils laboratories on a volunteer basis. These samples may not be representative of the area and/or fields from which they are drawn. McCollum and Nelson (3) have examined the possibility of fields volunteered being higher in some fertility elements and lower in others than those of a systematically drawn sample. They indicate the differences are small although statistically significant.

The fact that the average phosphorus or other nutrient level for fields in the sample is lower or higher than for the area sampled is not of major concern, however. From the standpoint of deriving a production function, equally good estimates of all portions of the production surface are desirable. Some conditions which are scarce in the soil type area may be relatively heavily represented in the best sample for deriving a production function. The observations in the sample must be representative of these various portions of the population sampled, however. That is, if fields with relatively low K_2O content are scarce in a particular area, the sample observations from such fields should be typical of the low K_2O fields. Any predictions as to yields would otherwise be meaningless when made for other farmers with those field conditions.

In order to study the effects of fertility levels on yield, the soil samples from which fertility estimates are made also must be representative of the plots or fields for which yields are estimated. This problem is of importance in soil sampling and testing as well as in deriving production functions. Any sample which is considered sufficiently representative of a field to make a soil test representative of the field should be satisfactory for production function fitting. The systematic type of field sample designed for the farmer's use is reasonably efficient in this respect.

Since no random sampling procedure is used in selecting the farmers who have soil tested, there is always an unpredictable possibility that the farmers who do have their soils tested are not representative of the entire population of farmers in the area under consideration. This situation may or may not create a problem. If the fields from which samples are sent to the soils laboratory are representative of the general soil type for which economic recommendations are to be made, even though the farmers are not, nothing is necessarily lost. However, there should be some indication of the farm management practices of fertilizer application used by these farmers in order to be able to tell other farmers how to attain the same results. It is likely that the methods of applying fertilizer are as important in determining yield response as the quantity of fertilizer used. More work is needed on this problem.

Farmers' Soil Sampling Procedures

More important is the fact that farmers take the samples. While specific directions for taking soil samples are given to the farmers, one cannot be sure that they are followed exactly. With such a large number of untrained people drawing samples, there is always the possibility of inaccurate sampling. This problem is probably resolved to some extent by the errors averaging out over the large number of samples taken.

Yield Estimates from Years Previous to Soil Test

Another peculiarity of our information is that yield estimates are made on crops raised the year previous to the soil tests. Hence, the recorded soil tests may not be the correct ones to associate with the yield data available. Unless the soil tests are reasonably stable from one year to the next, it becomes hard to distinguish soil nutrient effects from fertilizer effects in this situation. Heavy fertilizer applications can affect soil tests taken later. In the case of corn, the fertilizer applications will probably not be high enough to cause any great difficulty. Total fertility changes, in this case, would be small from one year to the next. This problem may be more serious with other crops and other fertilization techniques. It amounts to the same thing as errors of measurement of the independent variables, as discussed below.

Farmers' Yield Estimates

To complicate matters further, the yield data for crops are estimates and not necessarily actual measurements made by the farmers. Most farmers do not weigh their corn or make accurate checks on the yields. Many of these yield estimates are rough guesses by the farmers to give the soils laboratory an approximation from which to start analysis. A great deal of variability may therefore be introduced into the dependent variable — yield — by this process. If the farmers are not biased up or down in their estimates of yield, the large numbers of available observations will tend to resolve this problem. Their errors of estimation will offset each other. The importance of any possible bias would also depend on its nature, such as whether it is a constant or a relative deviation. If it were a relative amount proportional to yield level, less accurate estimates of yields at high levels would occur then at low levels. A constant overestimation or underestimation would not be too serious from the standpoint of affecting the accuracy of estimates of the additional yields that would be produced by additional amounts of fertilizer. However, the total yield estimates would be in error.

Uncontrolled and Unmeasured Variables

Many variables are unmeasurable in a cardinal sense. Texture groups, drainage groups, slopes, color, etc., can be classified but good quantitative measurements cannot be made of them. As previously mentioned, production functions can be fitted for various homogeneous groups sorted from such data. On the other hand, there are some variables which are not measurable and some, though measurable, for which measurements have not been made. These variables include moisture measurements, plant population, nitrogen test levels, and management factors.

At present there are not nitrogen tests that are universally accepted. Many soil laboratories will not record soil nitrogen except as indicated by cropping history, soil color, texture, and drainage. In deriving a production function, the data can be sorted according to these factors to achieve relatively homogeneous situations with respect to soil nitrogen. A separate function would have to be fitted to each situation.

Plant population is also uncontrolled and unmeasured in soil testing laboratory data. The number of plants per acre can usually be profitably increased as the level of fertilizer application is raised. If farmers take advantage of this situation, a relatively high correlation may exist between fertilizer application and plant population. The soil laboratory data do not indicate the extent of this problem.

At least two possibilities regarding these unmeasured factors are: (a) All or some unmeasured variables are uncorrelated with the measured independent variables and their effects may be normally and independently distributed. (b) Some or all unmeasured variables may be correlated to some degree with certain measured variables. It is the latter situation that is of concern.

In this instance the plant population, for example, may be correlated with fertility level, fertilizer application level, or method of fertilization. If such is the case, any increase in yield may not be a result of increasing one or more of the available plant nutrient supplies, but a result of increasing both the amount of plant nutrients and the plant population. A recommendation based on such an estimate will fall short when presented to and used with a lower plant population than assumed with the recommended fertilization program. This situation limits the use of the production function. However, the limitation may not be of concern in the short run. If farmers actually increase plant population and fertilizer application together, the estimate of fertilizer effects would be fairly accurate for prediction of yields under farm conditions as long as the relationship between fertility level and plant population is maintained. Unfortunately, no information as to the extent of this situation is available in the records. Some indication of the true situation might be obtained by examining data from experiments in which plant population was controlled and by making comparisons with portions of the production function derived from soil-testing records.

Fertilizer Nutrient Pounds Not Equivalent to Soil Nutrient Test Pounds

Neither the pounds of nutrients in fertilizer nor those indicated by soil tests measures the pounds of nutrients which are available to plants; both are functions of available nutrients. These variables, fertilizer elements and soil nutrient levels, may have to be considered as separate variables because the relationship between them is not known with certainty. This situation is not a serious disadvantage since there is no interest in the results of adding fertilizer nutrients to different soil fertility levels. The function and its variables in terms of the units commonly employed do not require any conversions to a common unit. The chief disadvantage is that the production function will contain more terms than would be the case if available plant nutrients such as P_2O_5 and K_2O could each be looked upon as one variable rather than several. This condition adds to the computational costs.

Correlation of Independent Variables

Perhaps one of the most troublesome problems in fitting functions to data other than those from controlled experiments is that of correlation between independent variables. Even without error of measurement of any variable, this condition is a serious limitation. It may occur because there is a correlation of independent variables in the population from which the sample is drawn or because of chance situations in sampling; therefore, difficulty results when estimating any large portion of the production surface with respect to two variables. For example, if P_2O_5 and K_2O are highly correlated, only a band on the production surface can be derived (figure 8.1). The width of this band also has an effect on the accuracy with which estimates can be made of the effects of changing amounts of K_2O or P_2O_5 within this area. Although the regression coefficients derived in the absence of measurement error will be unbiased, their variances will rise as the degree of correlation increases. Measurement errors of the dependent variables will accentuate this condition but still permit unbiased estimates.

Correlated Fertility Levels

In the event that the fertility levels found in the soil are so highly

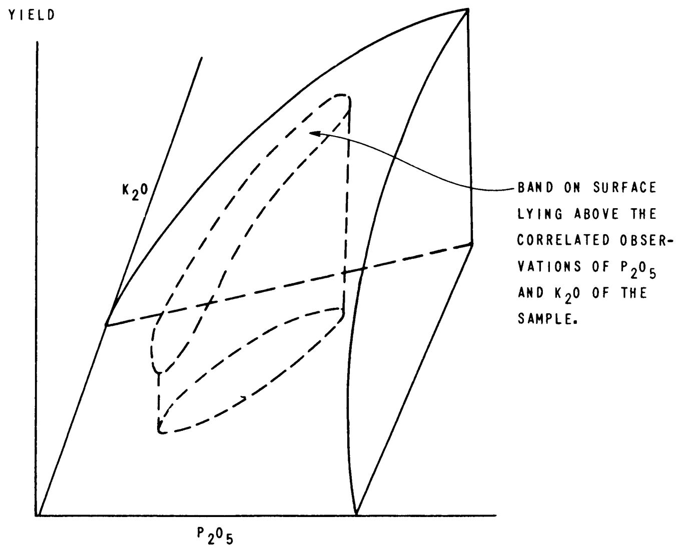

Fig. 8.1 — Correlation of independent variables.

correlated that it is not possible to separate the individual effects of nutrients, it is reasonable to use a function of the pair of nutrients as a single variable. Of course, the production function furnishes insufficient information to cover a situation where it is necessary to predict yield with these variables in some other relationship to each other than found in these data. Correlation of soil fertility levels would not be as serious as correlation of fertilizer applications in this respect. If the fertility levels of the soil are highly correlated in the population there could be a justification of a combined soil fertility level for a given soil type. On the other hand, failure to achieve a range in fertilizer application prevents analysis of shifts in the kind and amount of fertilizer nutrients that would be profitable under a variety of agronomic and economic circumstances. Hence, if possible, a wide range in the amount of any factor of production in the sample is desirable in order to examine its effects on yield, but for best results it has to be independent of other production factors in its variation over this range.

Correlated Fertilizer Application Levels

Many farmers may already be following fertilizer recommendations of the extension personnel and soils laboratory. Farmers with a given soil fertility situation may therefore be applying essentially the same fertilizer combinations and amounts; hence, the correlation between fertilizer elements applied may be high. The effects of adding a particular plant nutrient are then difficult to assess and the possibility of substituting one fertilizer element for another will be missed. There is no other solution to this situation unless it is possible to locate soil test records with relatively wide uncorrelated variation of the independent variables and to include them in the sample.

Errors in Measurement of Independent Variables

From a statistical viewpoint another possible problem is the inexact measurement of independent variables (1). The methods of function fitting usually used (least squares regression) depend upon an assumption that the independent variables are measured without error, if unbiased estimates of the regression coefficients are to be obtained (4). Such a condition rarely (if ever) is met in practice. If the errors are relatively small, the bias may be small. Insofar as the variability of laboratory tests on a particular soil sample is concerned, this condition is probably the case. However, the sample comes from an entire field from which a number of subsamples are systematically selected (5). These subsamples are mixed and the result is associated with the yield given for the field. Hence, sampling variability enters the estimate of the independent variables. Similar problems arise with pH tests. Fertilizer applications would not be a problem in this light if the fertilizer and soil nutrient levels were always evenly distributed. The rate per acre would then be the actual rate that went with the yield in question, starting from a particular level of soil fertility. If the distribution is uneven, however, a situation exists where some parts of the field may be receiving a much higher rate and some much lower. The combinations of these average to the rate used or recorded, but these do not necessarily give the same yield response as if the average rate were evenly applied. Similarly, if the fertilizer rates may be assumed equally, original fertility will vary throughout the field with similar implications.

There are methods of weighted regression which can be used to overcome the errors of measurement problem (6). Present methods do not account at the same time, however, for both errors of measurement and errors in the equation, i.e., omission of independent variables.

The measurement problem is most serious when the substitution rates between various nutrient elements are desired. In order to derive the substitution rates, unbiased estimates are needed for the function's parameters. These are very difficult to obtain from the situation involving serious measurement errors of the independent variables.

A further complication can arise if these measurement errors are combined with correlation between independent variables (2). Often, estimates result which may look reliable by use of the standard significance tests of the regression coefficients and examination of standard deviations. Unfortunately, the results can be unreliable.

References Cited

1. EZEKIEL, M., 1941. Methods of Correlation Analysis. Wiley, New York, 2nd ed., pp. 366-71.

2. KOOPMANS, T. C., 1950. Statistical inference in dynamic economic models. Cowles Commission. Monograph No. 10, Wiley, New York, pp. 258-65.

3. McCOLLUM, R. E., and NELSON, W. L., 1954. How accurate is a summary of soil test information. Proc. Soil Sci. Soc. Amer. 18:287-92.

4. OSTLE, B., 1954. Statistics in Research. Iowa State College Press, Ames, p. 126.

5. PURDUE UNIVERSITY AGRICULTURAL EXTENSION SERVICE, 1954. How to take an accurate soil sample. Mimeo. AY - 14.

6. TINTNER, G., 1952. Econometrics. Wiley, New York, pp. 121-53.