Summarizing and Interpreting Linear Programming Solutions Through Multiple Regression Analysis: Application to a Rural Economic Development Planning Model

Daniel G. Williams*

Introduction

Linear programming is useful for solving many problems involving choice. The researcher who understands the extensive programming printouts with their details of optimal activity levels and shadow prices learns a great deal from linear programming studies. But it is often difficult to explain the results to a layperson, area official or planner either because there is too much detail to absorb or because he uses other languages and concepts when dealing with the same economic or social problem.

In the past, this translation from the language of abstruse computer printouts — a form understood by the research specialist — to that of user-oriented information, has depended upon the patience, astuteness, and ability of the researcher to interpret the results and to prepare popular, readable reports. The purpose of this article is to illustrate how multiple regression analysis can be used to assist in summarizing, translating, interpreting, and reducing in volume the linear programming results into a user-oriented framework. An inferential context for the results, as well as mere summary is discussed. This regression procedure assumes that the initial economic or social problem was correctly solved by the linear programming algorithm.

No one maintains that regional growth processes are linear, technological coefficients are fixed, or that economies and diseconomies of scale do not exist. Even a nonlinear (e.g., quadratic) model would have many unrealistic economic assumptions; all models are simplifications of reality. The real question is: when the model is properly constrained to prevent excessive specialization (which might result due to the lack of industrial diseconomies of scale or the absence of a finite-size export market at a given price), do the results make sense? Are they similar to results which one would expect from a "perfectly" designed model? For the programming model explored here, this is the case (see Williams [8]). None of the research on the model has revealed any extreme sensitivity of results to changes in selected groups of constraints (except for those changes which one would have *a priori* expected). Although the literature on the subject of the validity and utility of linear models is voluminous, its discussion here is beyond the scope of this paper and will not be explored. Mention will be made later in this paper, however, on the need to limit to only a modest range the assumption of a linear relationship for the multiple regression model.

The multiple regression procedure was applied to a rural development,

*Regional Economist, Economic Development Division of Economics and Statistics Service, U.S. Dept. of Agriculture, Washington, D.C.

activity analysis, planning model (RDAAP) and was implemented in a three-county area (BMW Region), consisting of Benton, Madison, and Washington counties in northwest Arkansas.[1] The span chosen for the example used in this paper was 1960-70, but only the terminal year of that span is explicitly considered in the programming model. Plans are made as of 1960 to reach 1970 targets. The planning span is therefore neither extremely short or long run. Care should be taken, however, to modify any base-year technological coefficient whose value is expected to change substantially by the target year. Sensitivity analysis can explore the implications of any other possible coefficient changes as to whether such changes would substantially affect the optimal solution.

The RDAAP Model yields solution results for various industries, labor skills, and shipments (i.e., exports) outside the region. The model includes activities which produce manufactured products which can be used for local consumption or for export from the region. Constraints limit the amount of export (1) to a nearby market at one set of transportation costs, and (2) to a more distant market at a higher set of transportation costs. Within each of these two rings, export prices received are assumed to remain constant. This assumption is reasonable because of the relatively small size of the BMW Region compared to either of the two export rings.

The concern of the paper is: which of the export activities would planners prefer to attract to their region? There are 101 separate export activities which consist of 56 different product types (i.e., 4-digit SIC manufacturing industries). This number (56) expands to 101 because exports to the "outer ring" are differentiated from those to the "inner ring," and 45 of the 56 industries are considered to have export potential to both areas.[2]

The multiple regression analysis reduces and translates this output of 101 export activities (by type of product), and shadow prices[3] to nine or fewer[4] industry economic characteristics such as capital/output, capital/labor, value added/labor, and transport charge,[5] which are perhaps in a form more intel-

[1] Although Benton and Washington counties together were declared an SMSA after the 1970 U.S. Census of Population and, therefore, can no longer be considered rural, this application of a rural model can provide a glimpse into how an area should evolve (optimally) from a relatively more rural to more urban status.

[2] Two separate regressions could have been run, one for each of the two export rings. Sample size would be smaller for each, but all the industry characteristics would differ, in general, among all the export industries in each of the two respective groups (see footnote 5). This method, however, was not implemented because this benefit was not felt to outweigh the cost of reduced sample size.

[3] Shadow prices will be defined in a later section of this paper.

[4] In general, fewer than nine independent variables will remain if the "stepwise" rather than the "general" mutiple regression technique is used.

[5] Note that for the above 45 outer ring export industries, only the transport charge will differ among the nine characteristics compared to the identical SIC industry shipping to the inner ring.

There are no high pairwise correlation coefficients between independent variables, nor any obvious high multivariate correlations, both necessary requirements to lessen the possibility of multicollinearity. Further, the "Multicollinearity Effect" statistic measures 0.27915 compared to the multiple R^2 coefficient of .73996. This means that the proportion of the R^2 "attributed to" the entire nine variables is .73996 - .27915 = .46081; or, .46081/.73996 = 62.3 percent. (Note that the amount of R^2 attributed to any one independent variable is the R^2 value calculated with all nine explanatory factors, minus the R^2_h value, the multiple "R^2" with eight independent variables, the hth explanatory variable being deleted.) While this percentage is not extremely large, neither is it extremely small. According to Theil [7, p. 169], a 100-percent level would imply "perfect" pairwise orthogonality (i.e., linear independence) of the explanatory variables (i.e., vectors). Hence, the smaller this ratio, the less true is the "null" hypothesis. Since 62.3 percent is much larger than 0 percent, there probably is no multicollinearity problem. (continued on next page) ligible and useful to an area planner than would be details by SIC code. Instead of viewing the regression scheme, which utilizes the linear programming results, as providing a *substitute* for those results, one can view it as an adjunct to the results, both in increasing the understanding of the results and for providing a useful descriptive summary of them.

The cost, or information lost by this data reduction (and translation) scheme can be understood by recognizing that the linear programming model "explains" 100 percent[6] of the output variability. The nine (or fewer) regression variables (industry characteristics) explain a smaller portion of the output shadow price variation. The size of the R^2 measures how "good" the reduction and translation scheme is relative to the complete linear programming results.

The Regional Planner

Why would a regional or small-area, economic development planner and his technical staff desire such regression information from a linear programming model? If the industry type were disaggregated enough to be useful for specific planning recommendations, the cost of having a full, or at least fairly complete spectrum of industries would be prohibitive. For example, since the RDAAP Model includes only 56 of about 450 4-digit SIC manufacturing industries, how can the area planner state that a given group of industries in the model is "best" for the area? The planner would wish to compare 4-digit SIC industries not included in the model.

Increasing the industry scope in the linear programming model would be desirable if there were no added expense. But the planning task really involves firms or plants, not 4-digit SIC's, and hence an extremely unwieldly and costly linear programming model involving thousands of firms would be needed to fully cover the manufacturing scope. Each firm differs, in general, with respect to its input-output coefficients and other industry economic characteristics. Although there may be an aggregation problem in applying results for 4-digit SIC's to firms and plants, such criticism would apply to any linear programming output which uses relatively more aggregated input data in a specific planning context.

The regression results can circumvent the need for additonal industry or plant activities in the model. To use this shortcut, the analyst or planner must substitute the concept of industry economic characteristics for that of industry product type. That is, is the optimal industry light or heavy, capital or labor intensive, polluting or non-polluting, highly labor productive or low, and so on?[7] As a hypothetical example, a planner would not focus on whether to attract an industry producing axe handles or one producing poultry, but

The Durbin-Watson Statistic is $d_u = 1.8211$. Since the upper "tail" statistic for $K = 9 + 1$ (i.e., includes constant term) explanatory variables, and for $n = 101$ observations is $d_u = 1.73$ (at the 1-percent significance level), one does *not* reject the hypothesis of random disturbances. Thus, there is most likely no auto, or serial correlation problem in this example.

[6] For example, using the assumption of most linear programming analyses that all coefficient values for the programming model are fixed during the time-frame studied, the linear programming algorithm yields deterministically and exactly each shadow price value. Use of stochastic programming or parametric variation in coefficients, constraints, etc. can modify these more rigid assumptions, yet the coefficient values are still fixed *conditional* upon the specific assumption(s) being simulated.

[7] Obviously, not all these questions are answered by the nine characteristics used in this analysis, but there is no conceptual reason why they could not be so answered if the proper input-output coefficients were included in the linear programming model, leading to the respective industry characteristics calculated from them.

whether to attract industry of low capital/output ratio.

There may be a few questions concerning the legitimacy or proper use of multiple regression analysis in the context of linear programming. Such questions will be examined later in this paper. However, simple and multiple regression analysis has been used on linear programming solutions by Nugent [5]. He tests two alternative hypotheses that differences (or ratios) of optimal (model) industry and shadow price levels to actual industry production and market price levels for the Greek economy (1954-61) are best explaind by either omissions or defects of the model, or by market imperfections. Optimal industry and shadow price levels are included in dependent variables to be explained alternatively by eight "market" and five "omission" independent variables. At least some of the independent variables (e.g., land/capital) were derived from input data in the linear programming model.

What is being suggested here is that while the legitimacy questions do not seem substantial and probably can be ignored in any case, for planning purposes exact precision may not be required. Multiple regression can be used as an approximate decision rule to examine quickly a large set of excluded model industries. This set of industries can be reduced, for example, to four or five individual industries by using the results of the regression analysis based on the model sample to infer shadow prices to the excluded industries.

While calculating an *exact* inferred shadow price for each excluded industry would be as time consuming and costly as creating a vastly expanded linear programming model, making approximate judgements using only some of the industry characteristics would be much simpler. If only some of the more important independent variables were used, and their values only estimated for each industry, the export shadow prices would be at least roughly determined, enabling an approximate ordinal ranking of excluded industries. The final selected small group of industries probably would be at least "near best" among the excluded model industries. Whether this approximate method is satisfactory can be checked by rerunning the linear programming model with the selected small group of industries included together with the original model industries. This method seems preferable to (1) ignoring excluded model industries altogether, or (2) including all possible industries in the linear programming model.

Objective Function

Consider the objective of maximizing gross regional product.[8] The manufacturing export constraints (rows) yield valuations in the dual problem — the shadow prices. A shadow price measures how much an objective function value will improve for a one-unit ($1 million) increase in the constraint level for that export. Similarly, the optimal solution provides negative shadow prices (reduced costs) for those exports not in the optimal basis. This cost (price) measures how much the objective function value would be worsened if these exports (individually) came in the basis at a unit level. Both these prices —

[8] Many more regional objective functions were considered. In general, the regression results for multiple R^2, partial regression coefficients, and statistical F values differ for different objectives, but to limit the scope of this paper, only gross regional product is discussed here. The results for eight alternative (including gross regional product) regional objectives in RDAAP are discussed in [11]; the more theoretical treatment of the regression procedure, however, is presented in this paper.

labeled here together as "generalized" shadow prices — ranked from positive to negative values, can indicate the relative and absolute desirability of each of the 101 export activities with respect to the given regional objective function.[9]

The Linear Programming Model

The linear programming model (RDAAP) is of the usual form and, therefore, its mathematical structure will be only briefly sketched here.[10] In matrix notation it can be symbolically portrayed as:

$$\text{Max: } GRP = AX$$

Insert 1

$$\text{Subject to: } \quad \begin{array}{l} C_1 X \leqslant R_1 \\ C_2 X \geqslant R_2 \\ C_3 X = R_3 \end{array}$$

$$\text{and: } X \geqslant 0$$

$$\text{where} \quad \begin{array}{llllll} C_1 & d \text{ by } n & A & 1 \text{ by } n & R_1 & d \text{ by } 1 \\ C_2 & s \text{ by } n & X & n \text{ by } 1 & R_2 & s \text{ by } 1 \\ C_3 & (m\text{-}d\text{-}s) \text{ by } n & & & R_3 & (m\text{-}d\text{-}s) \text{ by } 1 \end{array}$$

$$\text{and } n > m > d+s$$

There are m row and n column activities, plus the right-hand-side column or "resource" vector. The manufacturing export contraints (101) are included among the elements of the R_1 vector. RDAAP includes a total of $m = 370$ rows, and $n = 398$ columns (excluding slack activities).

The shadow price, $(SP)_i$, for the ith $(i=1, 2, ..., 101)$ manufacturing export activity can be expressed as a partial derivative evaluated at the optimal basis solution. That is,

$$(SP)_i \Bigg|_{\substack{\text{optm.} \\ \text{basis}}} = \frac{\partial \sum\limits_{k=1}^{n} a_k x_k^*}{\partial r_i}, \text{ where } r_i \in R_1 \ (i=1, 2, ..., 101),$$

[9] Two separate regressions could have been implemented, one each for optimal basis and non-basis export industries. Since shadow prices and reduced costs are conceptually identical, such a division would not make sense, and would lead to reduced sample sizes, and higher critical F values at the same significance levels.

[10] The RDAAP Model includes service, manufacturing, and government sectors, and an agricultural sector in which technological progress is simulated by conversion of regressive farms into progressive farms. See Williams [9] and [10].

$R_1 = [r_1, r_2, \ldots, r_{101}, \ldots, r_d]'$ and $r_1, \ldots, r_{101}$ are the 101 manufacturing export constraints.

The function for the shadow price depends upon the coefficients in vectors A, R_1, R_2, and R_3, and the matrix elements in C_1, C_2, and C_3. It is a linear combination of all coefficients in the linear programming model. The linear combination is different for each of the "i" rows. Assuming the linear programming coefficents are fixed (see footnote 6), it "explains" with 100-percent accuracy the shadow price for the ith export activity.

The above equation for shadow price, $(SP)_i$, can be expressed both for export industries in the optimal basis and for those not in the optimal basis. The former values for shadow price are positive (or zero for a nonzero export level which is less than its constraint maximum, or zero for a degenerate solution). The latter values are all zero. For this latter group, the variable of interest here is not the shadow price, but rather the reduced cost, $(RC)_i$, for all i where $x_i^* = 0$ in the optimal solution (except for degenerate solutions — zero-level exports *in* the optimal basis).

$$(RC)_i \Bigg|_{\substack{\text{optm.}\\ \text{soln.}}} = \frac{\partial \sum_{k=1}^{n} a_k x_k^*}{\partial x_i^*}, \text{ evaluated}$$

at $\partial x_i^* = +1.0$, and where x_i^* is the optimal level of the export activity associated with export constraint r_i. $(RC)_i$, too, is a linear function of all matrix and vector elements, and explains with 100-percent accuracy the variation in reduced costs (see footnote 6). As indicated earlier, $(RC)_i$ can be thought of as a "negative" shadow price (or cost), and $(SP)_i$ as a "positive" shadow price.

The concern in this paper is with the generalized shadow price, $(SP)_i$ and $(RC)_i$. For simplicity, this *generalized* variable will be labeled henceforth only as $(SP)_i$. The term "shadow price" will refer always to generalized shadow price unless otherwise stated. The technique in this paper seeks to replace these two functions for the generalized shadow price (101 separate linear equations) with a single linear equation obtained from multiple regression analysis. The industry economic characteristics used in this regression are formed from, in general, complex calculations involving numerous selected elements, portions of elements, and their ratios from the elements in C_1, C_2, and C_3. This makes the estimating function less complex by ignoring the effects of elements in A, R_1, R_2, and R_3, as well as many of the elements in C_1, C_2, and C_3.

The Multiple Regression Model

The multiple regression model is of the form:

$$(SP)_i = b_0 + b_1 D_{i,1} + b_2 D_{i,2} + \ldots + b_9 D_{i,9} + e_i$$

where, $b_1, b_2, ..., b_9$ = partial regression coefficients

b_0 = constant term

$D_{i,1}, D_{i,2}, ..., D_{i,9}$ = industry economic characteristics
(assumed fixed or pre-specified in sample)

e_i = error term (assumed random)

$(SP)_i$ = generalized shadow price

$i = 1, 2, ..., 101$ (sample size of export industries)

This single estimating equation for the generalized shadow price is much simpler than the earlier 101 exact equations, but it "explains" the shadow price with less than 100-percent accuracy. Here, $(SP)_i$ is a function of the industry economic characteristics $D_{i,1}$, $D_{i,2}$, ..., $D_{i,9}$. The levels of these characteristics can be considered to be determined by a set of generally complex, nonlinear functions of the elements in c_1, c_2, and c_3, where c_1 is a subset of C_1, c_2 of C_2, and c_3 of C_3. While the industry characteristics are formed from fixed coefficients in a linear model, they are generally not linear functions of those coefficients, nor do their values change in a "linear" (constant) fashion from one manufacturing export industry to the next. Holding all other characteristics constant, the partial regression coefficient, b_t, ($t = 1, 2, ..., 9$), measures the change in the shadow price in million dollars for a unit change (units given in Table 1) in a specified industry economic characteristic.

The multivariate relationship assumed is that of the general linear model. While the size of the multiple R^2 measures how substantial is the "explanation" of the industry economic characteristics, the researcher should be cautious in applying any results outside the bounds of the sample. While the "fit" may prove satisfactory for the sample, the true relationship between shadow price and industry characteristics may be better explained by a non-linear relationship. This question has not been explored here other than to examine the plot of the standardized residuals versus the standardized predicted values of the dependent variable. Because a rough symmetry of these points was observed about the origin, the linear specification is probably satisfactory. However, it would seem prudent to infer shadow prices to only those non-included industries whose levels of industry characteristics are within or near the range of levels in the sample.

Experimental Design

The RDAAP coefficients are defined from secondary, "ruralized" data,[11] with no measure of a specific probability distribution for those coefficients. Thus, conditional upon the values for these coefficients, the programming solution is totally deterministic. This is a common assumption used in non-stochastic linear programming analyses. If one considers these coefficients as fixed, or at least pre-specified, then the industry economic characteristics calculated from them are similarly interpreted as being fixed. These nine characteristics (Table 1) are the independent variables in the regression analysis, but do not

[11] From "ruralized" data (i.e., from estimated non-SMSA national product-code industries) from the "work sheets" for the 1958 U.S. Input-Output Table by the U.S. Dept. of Commerce.

TABLE 1. RDAAP Model's Manufacturing Export Industry Economic Characteristics: Independent Variables.

Independent Variable	Units	Calculation
Transport costs per dollar of export	$10^6	Transport costs per million dollars of export
Capital/output ratio	--*	10-year capital ÷ 10th-year output
Capital/labor ratio	$100/man-hour	10-year capital ÷ 10th-year labor ("current" acct.) +
Rate-of-return ratio	--	10th-year total profits ÷ 10-year capital
Value added /output ratio	--	10th-year value added ("total" account) ÷ 10th-year output ‡
Value added/labor ratio	$100/man-hour	10th-year value added ("total" account) ÷ 10th-year labor ("total" account)
(Managerial labor)/(total labor) ratio	--	Managerial labor ("total" account) ÷ total labor ("total" account)
(Skilled labor)/(total labor) ratio	--	Skilled labor ("total" account) ÷ total labor ("total" account)
Imported input costs per dollar of output	$10^6	Direct imported input costs ("total" account) per million dollars of output ("current" account)

* Dashes indicate a pure ratio. That is, units in numerator and denominator "cancel out."

+ "Current" account refers to "current" goods production, rather than "capital" goods production.

‡ "Total" account refers to "current" vector plus associated "capital" vector (assumed to be 15 percent of total 10-year capital requirement) for the 10th year.

have to be considered as random variables.

Fixed or pre-specified independent variables can be used in regression analysis without the assumption of randomness in the independent variables [3, pp. 13-25, 109]. However, as in most economic analyses, obtaining secondary data samples would be difficult if some industry characteristics were to be pre-specified. Yet, the regression model can be viewed *as though* the independent variables are fixed (even if they are not) when either of two assumptions are made [3, pp. 25, 26]. The first assumption is that the probability statements concerning the confidence intervals and the powers of the tests are valid, but represent statements about the *conditional* distribution of the dependent variable given the independent variables.

The second, and more important alternative assumption is that the explanatory variables are considered independent random variables, each with a separate probability distribution. But each distribution is assumed not to involve the population parameters for the regression equation (i.e., constant term, partial coefficients, and variance of the error term). Again, the standard general linear model assumptions are considered to hold for the conditional distribution of the dependent variable given the independent variables. These assumptions are common ones used in economic analyses [3, pp. 25-29, 133]. The analysis in this paper will be viewed using the fixed independent variable interpretation, using the second assumption.

In short, the statistical properties, confidence intervals, unbiasedness, etc. of the regression parameters are identical whether using the fixed or random variable assumption, if, in the random variable case, the above assumptions concerning the distributions are assumed to hold. The distribution assumptions are standard ones used in most economic studies where the independent variables cannot be "controlled for" [3, p. 26]. Some possible qualifications, however, concerning the general linear model assumptions (i.e., unbiasedness) will be discussed later in this paper.

The ranges of the independent variables seem sufficient for an acceptable experimental design in that the sample standard deviations divided by their corresponding sample means range generally between 0.30 to 0.60, with a low slightly above 0.33 to a high of about 1.67. Any problem in the estimation of statistical significance was felt to reside with the lower end of this scale where the range for the independent variable might be inadequate to yield sufficient differences in the generalized shadow price. Since the covariance term between an industry characteristic and the shadow price lies in the numerator of the F statistic for the partial regression coefficient, an independent variable with a very narrow range would *ceteris paribus* tend to depress the level of F significance and, thus, the confidence in the estimated coefficient for that variable.

The above concern fortunately was unwarranted by the results since the independent variable with the lowest range — (skilled labor) / (total labor) — is one of the seven variables out of nine showing a statistically significant partial coefficient F value (at the 25-percent level). Similarly, the other independent variables with the most narrow ranges were, in general, no less statistically significant than those with larger ranges. Thus, the design of the experiment seems reasonably adequate for our purposes.

One must assume the *conceptual* possibility of repeated sampling of the explanatory variables (which can be assumed fixed or random) for a meaningful interpretation of the random error term and the statistical F tests for both the whole equation (multiple R^2) and also the individual partial regression coefficients, b (and BETA), the latter measured in standard deviation units. This hypothetical sampling can be visualized as being repeated for many alternative data input sets for the given linear programming model (i.e., yielding different coefficients and industry characteristics for the programming and regression models, respectively), the only requirement being that the 101 export types (SIC's) remain in the model. The respective sets of optimal generalized shadow prices (dependent variable) will, of course, vary between each sample.

To interpret the regression model error term, it seems easier to view the independent variables as fixed or pre-specified, although as explained earlier, they are not. Therefore, in this scheme, the only industry input-output coefficients which are considered to change between samples are those which do not affect the levels of the nine characteristics. The resultant changes in shadow prices would then be assumed due to the changes in the (hypothetical) 10th and above characteristics not considered in the regression model, but which in general could change with the sampled industry input-output coefficient changes. The fact that some input-output coefficients are viewed as being fixed or pre-specified and others are viewed as random does not imply the first group to be non-stochastic and the second group to be stochastic. It merely means that some coefficients (and industry characteristics) are viewed as being pre-specified and others are not, enabling the latter to vary. In general, all coefficients (and industry characteristics) would vary for different data sets. In any case, the error term (e_i) "represents" these 10th plus characteristics, and by the Central Limit Theorem, it is assumed to be a normally distributed random variable with zero mean and constant variance.

This interpretation of the error term (e_i) as a random variable should be given further elaboration. That there is an exact but different functional (linear) relationship between each generalized shadow price and *all* linear programming model coefficients — the set of 101 generalized shadow price equations discussed earlier — may seem to cause biased estimates of the partial regression coefficients because of the apparent possibility of (1) an independent variable consisting of the ratio of two linear programming coefficients, and (2) a non-random error term correlated with the nine included independent variables. Since some of the independent variables are composed, at least in part, of ratios of input-output coefficients, each known to be linearly related with each generalized shadow price, a possible statistical bias problem could result if the dependent variable were regressed on merely the ratio of the two coefficients. However, the fact that the industry economic characteristics are much more complicated than simple ratios of two coefficients would seem to mitigate this as a possible problem, but the remote potential for bias error should be noted.

The second apparent source of bias error mentioned above can also be linked to the "excluded variable" problem. If the error term is implicitly considered to be composed of an uncountably large number of excluded industry economic characteristics, then if each of these is also implicitly assumed to be formed from manufacturing input-output coefficients known to be linearly related to each generalized shadow price, then all excluded characteristics perhaps should be included among the set of independent variables because each may seem to be in a functional relationship with the generalized shadow price. In short, since each included independent variable is also considered functionally linked to the shadow price (because of the high multiple R^2), the included and excluded variables may not be independent of one another.

There are several reasons why this apparent cause of statistical bias is not in effect. First, the industry characteristics are calculated from not all the linear programming model coefficients, but rather from only a subgroup, consisting mainly of only the manufacturing input-output coefficients. Thus, knowledge of any one, or perhaps even all (of the infinite number of) characteristics assumed to comprise the error term (and of the input-output coefficients needed to form the characteristics), would not enable one to know the values of the generalized shadow prices and, thereby, know the values for the (9) included independent variables via the high multiple R^2. The linear programming coefficients other than those of the manufacturing input-output coefficient matrix would have to be used as well. In other words, exact relations between excluded and included variables could not be calculated by this process because from the point of view of the regression model, the non-input-output, linear programming coefficients are unknown. Therefore, since knowledge of the components of the error term does not enable one to know the values for the included variables, the error term and independent variables can be considered independent.

A second, and much more important reason that such a statistical bias problem does not exist is the following. Namely that the possibility that excluded characteristics might be found functionally related to the shadow price (due to the input-output coefficients of which they are theoretically comprised) is not a valid criticism of the regression model structure. In theory [3, p. 6], if all the uncountably large, or infinite number of potential explanatory variables were included, there would be an *exact* functional relation (although unknown, immeasurably complex, and most certainly not linear) between the dependent and explanatory variables. When most of these variables, which are either unknown or *a priori* considered unimportant, are excluded and implicitly considered in the error term, their extremely large number, together with their probable offsetting effects, which yield relatively more small than large net effects, permits an assumption of a random error term, normally distributed. In other words, the relationship between excluded variables and the dependent variable is, in general, assumed to exist, but is unknown. Alternatively, we could assume in this analysis that such relationships can actually be calculated. Whether such an assumption is made, however, is irrelevant because the result is identical in either case.

General Hypothesis

The *a priori* hypothesis in this paper is that industry economic characteristics affect generalized shadow prices. Most analyses proceed from economic theory, yielding specific alternative hypotheses to be tested. This is not precisely the procedure used here because it is not clear (in the absence of much more detailed study) which set of economic or theoretical hypotheses would be associated with each alternative regional objective. The analysis in this paper was performed with eight alternative objective functions, although only one is discussed here (GRP). The same set of industry economic characteristics was used for all alternative objective functions. It was felt that these nine chosen characteristics might influence shadow prices formed from a number of alternative objectives, but the size and the direction of the effects were not hypothesized *a priori* from theory. This methodology is similar to that used by Schaffer and Tweeten [6] in their multiple regression analysis of community budgets to determine a small set of factors which might satisfactorily "explain" community net gains per employee. A theoretical explanation is attempted for each of the statistically significant independent variables in this paper, but the associated hypotheses by no means should be considered to have been explicitly derived, *a priori*, from economic theory.

Interpretation of the Multiple Regression Results

The regression output can be interpreted in three senses: (1) descriptive summary only of those industries included in the model; (2) inference of probable model results to nonincluded model industries, or to industries only slightly changed from those in the model; and (3) inference of probable results for an actual planning area. These senses are in order of perhaps a decreasing statistical basis, but increasing area planning utility.

There is a fine line and, perhaps, artificial distinction between the second and third senses. The second sense refers to use of secondary data for model industries, which may, or may not be a random sample from the planning area. The lack of such a sample would at least somewhat reduce the usefulness of the results in the third sense, but would not prevent inference (sense two) to other potential industries (or for included industries with changed coefficients) which might be considered for inclusion in the model. Nie [4, p. 321] describes the two main interpretations of multiple regression analysis: summary (or descriptive), and inferential.

In both of the inferential senses, all firms or plants, as subsets of the more aggregated 4-digit SIC classifications, can be considered because of the shift in focus from industry product type to industry economic characteristics. Since plants may differ in input-output coefficients from the "average" for their main corresponding 4-digit SIC with repsect to product type, the linear programming model's selection of an optimal 4-digit SIC industry may, or may not mean that a particular plant or firm which is generally within that same SIC is also optimal.

The statistical analysis of the regression results applies easily to sense (1) above, which does not involve statistical inference. However, there may be some question as to whether this analysis yields a statistical basis for inference as in (2) or (3). In (2), it is most legitimate statistically to predict a shadow price for an (SIC) industry already included in the model, but which becomes nonincluded with only one or a few modest changes in the input-output coefficients from those in the model sample, leading to a change in one (or more) of the nine industry characteristics. The 10th plus characteristics would then be, in general, only slightly altered. For a nonincluded SIC model industry, an inference as to its shadow price would be more accurate if it too were similar to those industries already in the model. For example, if the values for the economic characteristics of the proposed industry lay in the same range as those in the model sample, the 10th and above characteristics would likely also be about average as compared to the sample. Thus, the properties of the error term should be similar, enabling the shadow price of the proposed industry to be inferred from the sample.

The results from the regression analysis, however, should be considered valid only at or near the optimal corner solution where opportunity costs implicit in the model do not change, or change only slightly. Adding too many industries to the group of model industries might lead the programming solution excessively far from its original optimal solution, and perhaps to another (new optimal) corner solution with vastly changed opportunity costs.

It seems reasonable to this author to assume that marginal additions of previously excluded industries to the linear programming model would show actual shadow prices similar to their inferred shadow prices; that the regression results for an arbitrary number of (SIC) industries (e.g., 56) would be similar to the results for one or several more industries (e.g., 57-60). If not, this would imply that the usefulness of all linear programming models would be in question — due to their extreme sensitivity — and not just the regression results.

Inference as in (3) above refers to whether the model results provide a statistical basis for planning in the actual area (BMW Region). Since all industries, including manufacturing, in the RDAAP Model were created from secondary national (albeit "ruralized") data, and do not reflect a random sample of firms (or potential firms) to the area, this question cannot be answered precisely. One could have conducted such a sample in constructing the linear programming model and, therefore, would know the confidence intervals surrounding the industry input-output coefficients in the sample. However, as is usual in formulating linear programming models, a random sample of firms was not implemented. This is not a serious problem because such a question on the "statistical" accuracy (in sense 3) would apply not only to these regression results but also to the programming results of *any* linear programming model using secondary or non-random sampled data. Therefore, the proposed regression technique has no less validity for area planning than the results from most non-stochastic linear programming analyses, and most likely has more usefulness because of the inference to nonincluded industries. Parametric programming on "questionable" input-output coefficients, or use of stochastic programming are two ways linear programming typically handles this problem, although the latter is not used in the linear programming analysis for the RDAAP Model.

Multiple Regression Results: General Regression Method[12]

Multiple R^2. The overall "explanation" of industry shadow prices by industry characteristics is reasonably successful — 74.0 percent for multiple R^2 in Table 2. Sample size is 101 and there are nine independent variables. For the overall equation, degrees of freedom are 9, 91; critical F = 2.64 for 1-percent significance; and sample F value is 28.77.

Partial Regression Coefficients. Results for the independent variables are discussed individually for those which were significant at the 25-percent level. This level was chosen in order to increase the number of results discussed. Many independent variables were found to be much more significant than at the 25-percent level.

For the partial regression coefficients (b) in Table 2, the degrees of freedom are 1, 91 and F = 1.34 for 25-percent significance. Seven of the nine industry characteristics exhibit an effect on the generalized shadow price which can be considered to be significantly different from zero. If one examines two industries differing only by a unit in the given characteristic, the b coefficient will measure, per unit of output, the difference in the contribution of each (alternatively) to gross regional product. The BETA coefficient measures the same effect but in standard deviation units. By considering the absolute value of BETA, the variables can be ranked as to their relative "impact" on the generalized shadow price.

In the following discussions of the various partial coefficients, all results and statements with respect to each of these significant variables should be considered as being precisely valid only *ceteris paribus* with respect to the levels of the other eight variables. Many results may be more general, but this question has not been systematically explored here.

[12] In this regression method, all nine explanatory variables are included. In the stepwise regression method, whose results are not discussed here, only statistically significant (assumed 25 percent) explanatory variables remain in the results. The variables are added one at a time until all (and only) statistically significant variables have been entered. The latter technique has yielded results which are, in general, very similar to those of the general regression method.

TABLE 2. Multiple Regression Results — General Regression Method * Dependent Variable: Gross Regional Product (Related Objective Function)

Multiple R^2	.73996			
F Statistic +	28.77217			

	b	BETA	F‡	25 percent significant?
Transport costs per dollar of export	- .87729	- .67089	100.507	Yes
Capital/output ratio	- .09543	- .24365	7.253	Yes
Capital/labor ratio	- .06377	- .04374	0.450	No
Rate-of-return ratio	- .03852	- .07658	0.829	No
Value added/output ratio	+.11710	+.17162	4.438	Yes
Value added/labor ratio	+1.39285	+.21038	8.902	Yes
(Managerial labor)/(total labor) ratio	- .47332	- .39072	35.139	Yes
(Skilled labor)/(total labor) ratio	- .11507	- .09174	1.433	Yes
Imported input costs per dollar of output	+.05297	+.10507	2.336	Yes
Constant	+.06674	-	-	-

* 101 observations and 9 independent variables.

+ Degrees of freedom: 9, 91. Critical F = 2.64 for 1-percent significance.

‡ Degrees of freedom: 1, 91 for all independent variables. Critical F = 1.34 for 25-percent significance; F = 2.77 for 10 percent; F = 3.96 for 5 percent; and F = 6.96 for 1 percent.

Of the significant partial coefficients, transport cost per dollar of export has both the largest impact (BETA rank) and the highest level of statistical significance — an F of 100.51 and BETA of -.67089. As export transport costs fall, GRP rises. This result seems consistent with the theoretical importance given transportation costs in the historical development of industrial location theory. For example, manufacturing export industry SIC 3141 "Shoes, except rubber," exporting to both export rings, exhibits the largest pair of export shadow prices. Although there is no necessary reason for its transport cost to be also among the lowest for all export industries (i.e., the regression coefficient measures the *partial* effect), this is, in fact, the case.

The second highest BETA and F values are seen for the (managerial labor)/(total labor) variable, with an F = 35.14 and BETA = -.39072. Because the (clerical labor)/(total labor) variable exhibits a high positive pairwise correlation with managerial labor, increases in one cannot easily be separated from the other. Because of this high correlation, the clerical labor percentage was eliminated from the list of potential independent variables. A third labor skill variable — (skilled labor)/(total labor) — is also statistically significant with an F of 1.433 (lowest of the seven significant variables), and a BETA of -.09174. It therefore ranks (BETA) seventh of the seven significant variables.[13]

In both the above cases, the b and BETA values are negative. This is easily interpretable for managerial labor since it is relatively in shorter supply[14] than other skill types. One would expect total area product to fall for an added industry which uses more (rather than less) of a scarce resource because of the higher opportunity costs resulting from industries not producing (due to the shift in this resource to the added industry). The explanation for skilled labor is probably similar, but it is not as scarce in the model as is managerial labor.

The third highest F value is shown by the value added/labor variable,[15] with an F of 8.902 and a BETA of .21038, the fourth highest. This result seems reasonable in that one would expect increased labor productivity (in value added) to be positively related to increased regional production (gross regional product). The result also can be interpreted by recognizing that value added/labor is highly pairwise correlated (.84107) with wage/labor, or "average" wage, which was deleted from this analysis because of this high correlation. Value added/labor perhaps can be considered a surrogate for wage/ labor. With this interpretation, the result tends to contradict the wisdom of the usual "shirt factory" industry type of employment, with its low average wage rate, which is often attracted to rural and southern areas of the United States. This does not seem to be the preferred type of development in view of these results.

The capital/output ratio exhibits the fourth highest F value (7.253), and the third highest BETA (-.24365). Since b (and BETA) are negative, as capital intensity (relative to output) increases, gross regional product falls. This result suggests that large capital intensive projects may not be "best" for more rural areas; rather the converse is probably true. This outcome seems consistent with those economic development theories (for poor underdeveloped countries or regions) in which cottage industry (lower capital intensity relative to both labor and output) is often suggested as the most fruitful course of development.

Value added/output is linked positively to gross regional product. That is, b (and BETA) are positive and, as the independent variable rises, the dependent

[13] A fourth skill variable — (unskilled labor)/(total labor) — was also deleted from the list of independent variables. It is highly pairwise correlated with transportation costs, value added/labor, and profits/labor; the latter also deleted from the variable list because of its high pairwise correlations with value added/labor and transport costs. Note, however, in any case, even if there were no high pairwise correlations of these labor skill variables with other potential independent variables, that not all *four* could be considered simultaneously. Since labor is divided here into four types, and the four ratios sum to 1.0 for each industry, each percentage is an *exact* linear combination of the other three.

[14] That is, shorter supply relative to its demand in the optimal solution to the linear programming model problem.

[15] The labor/output variable has been deleted from the list of explanatory variables because it is an *exact* ratio of two included variables — value added/output divided by value added/labor. Similarly, the ratio of capital/output to capital/labor would seem to equal labor/output. This is not so, however, because labor from "total" account was used for labor/output and value added/ labor, and labor from "current" account for capital/labor (see Table 1). For this same reason, the four remaining included variables are *not* functionally interrelated. That is, $(VA/O)/(VA/L_T) \neq (K/O)/(K/L_c)$.

variable also increases. F = 4.44 (fifth largest) and BETA = .17162 (ranked fifth). Because this variable is highly pairwise (.93588) correlated with wage/output, the effect of value added/output should be considered jointly with that of wage/output, the latter variable dropped from the list of independent variables because of this correlation. The interpretation of this result lies perhaps in this link between wage/output and value added/output. As value added increases, wages increase. Gross regional product is improved by a higher aggregate wage bill per unit of output. Again, the above conjecture regarding the inappropriateness of the "shirt factory" form of industrial development seems to be validated here by the results for "aggregate" wage, at least to the extent that conventional wisdom would recommend not only lower average wage levels, but also lower aggregate wage bills per unit of output.

The final statistically significant partial coefficient is that for the imported input cost per dollar of output — an F of 2.336 (sixth largest), and a BETA of .10507 (also ranked sixth). Thus, as the percentage of an industry's imports relative to industry output increases, gross regional product rises. This finding for the import cost perhaps can be interpreted by considering that the result tends to follow the advice of Albert O. Hirschman [2], who suggests for an underdeveloped country, a planner should pursue a policy of attracting industry in which merely the "finishing touches" are put on the disassembled imported goods before re-exporting. In other words, this implies that a very large percentage of the total value exported should consist of goods (inputs) which were previously imported.

Conclusion

The results can be viewed in three alternative ways: (1) descriptive or summary only; (2) inference to excluded model industries; and (3) inference to area industries. The first interpretation is perhaps on relatively more secure theoretical ground than the two inferential senses, but it is these latter two which probably will be of most use to small-area economic planners. This author does not feel that the qualifications necessary to accept as valid the results in the two inferential senses, are much different than assumptions made in most empirical planning work. This regression technique is suggested here as an additional regional, or small-area planning tool.

The analysis given here involves a small-area, linear programming, economic planning model. The regression method presented in this paper might perhaps be applicable to studying the results of other types of optimization models (e.g., nonlinear). This question, however, has not been explored here. Other applications might be made to promote optimal individual farm management, crop selection, or agricultural land use, etc. The main problem would be one of devising appropriate "explanatory" factors which are easily interpretable and which could be hypothesized to exhibit some effect on the objective function. Multiple regression analysis may be able to "translate" this programming printout into a more meaningful form. There is a resemblance between the technique used here and a methodology labeled Response Surface Analysis [1]. Because the two seem to be only superficially similar, response surface analysis has not been discussed in this paper.

The regression technique used has shown that industry can be viewed as to its economic characteristics as well as to its product type (SIC) in a linear programming model. Some specific results indicate that gross regional product increases, *ceteris paribus,* with industries exhibiting low export transport costs, low capital intensity relative to output, low percentages of managerial and skilled labor, high percentage of imported inputs, and high value added per unit of output and per unit of labor.

Some of the results — such as low transport costs and more efficient use of scarce resources (e.g., managerial and skilled labor) — concur with the conventional "planning wisdom." But many do not. For example, it is often felt that rural areas should attempt to attract low-wage industry. The results here for both value added/labor (and average wage) and value added/output (and aggregate wage) suggest that such a policy may be in error with respect to maximizing gross regional product. It is also often suggested that a large infusion of capital is necessary for economic growth. While investment capital is obviously needed for regional growth, this should not imply (as often as it does) that capital intensity (e.g., capital/output) need be sizeable for most individual industries. The negative relation between capital/output and gross regional product tends to imply the converse. In addition, some economic development strategies for less developed regions have stressed the need for industrial complexes, with their implied low import requirements. These complexes have been suggested as being "best" for the planning area, so that it can have more "balanced growth" and independence from other areas. The results here suggest that higher import requirements are associated with larger increases in gross regional product.

REFERENCES

1. Hill, William J. and William G. Hunter. "A Review of Response Surface Methodology: A Literature Survey," *Technometrics*, 8, (4), November 1966, 571-590.

2. Hirschman, Albert O. *The Strategy of Economic Development*. New Haven: Yale Univ. Press, 1958.

3. Johnston, J. *Econometric Methods*. New York: McGraw-Hill, Inc., 1963.

4. Nie, Norman H., C. Hadlai Hull, Jean G. Jenkins, Karin Steinbrenner, and Dale H. Bent. *Statistical Package for the Social Sciences, Second Edition*, New York: McGraw-Hill, Inc. 1975.

5: Nugent, Jeffrey B. "Linear Programming Models for National Planning: Demonstration of a Testing Procedure," *Econometrica*,38 (6), November 1970, 831-855.

6. Schaffer, Ron E. and Luther G. Tweeten. *Economic Changes from Industrial Development in Eastern Oklahoma*. Oklahoma State Univ., Agricultural Experiment Station, Bulletin B-715, July 1974.

7. Theil, Henri *Principles of Econometrics*. New York: John Wiley & Sons, Inc., 1971.

8. Williams, Daniel G. "Objective Function Tradeoff Curves in a Rural Economic Development, Activity Analysis Planning Model," *The Annals of Regional Science* (forthcoming issue).

9. ________. "Agricultural Census Data as a Source of Linear Programming Vectors," *Agricultural Economics Research*, 30, (2), April 1978, 34-37.

10. ________. "Structural Details of a Linear Programming, Rural Economic Development Planning Model: Northwest Arkansas," Working Paper No. 7907, Economic Development Division, Economics, Statistics, and Cooperatives Service, U.S. Dept. of Agriculture, June 1979.

11. ________. *Use of Multiple Regression Analysis to Summarize and Interpret Linear Programming Shadow Prices in an Economic Planning Model*. U.S. Dept. Agr., Econ. Stat. & Coop. Serv., Tech. Bul. No. 1622, May 1980.