Methodology for Developing Activity Distribution Models by Linear Regression Analysis

DONALD M. HILL, Senior Research Analyst, and
DANIEL BRAND, Senior Project Engineer, Traffic Research Corporation

•A PROPOSED mathematical framework for developing urban activities distribution models is described. The models distribute forecast regional totals of socio-economic variables to small zones; for example, resident population by various income levels would be distributed to traffic zones. The distribution is carried out as a function of future public policies relating to highway and rapid transit improvements, public open space, etc.

To calibrate activities distribution models, information over a historic time interval on growths and declines of the activities to be distributed is needed. Thus changes in zonal values of activities, and similar changes in the policy variables to be tested are the information with which the models are calibrated.

This paper describes a methodology for developing an activity distribution model by linear regression analysis. A simple example of the regression model is the linear equation constructed with three variables

$$\Delta R = a + b_1 \Delta Z_1 + b_2 \Delta Z_2$$

where R is the measurement of growth or decline of a land use activity; ΔZ_1 and Z_2 reflect changes in measurable and causal factors; and a, b_1 and b_2 are parameters derived by application of the least squares principle. The best values of a, b_1 and b_2 are established to minimize the expected error of estimate of ΔR by solution of the equation with known values of ΔZ_1 and ΔZ_2.

However, by the use of linear regression analysis, it is frequently argued that the model builder is seriously limited in the flexibility of the model's construction. Critics of regression analysis are quick to point out the following troublesome restrictions of regression analysis.

1. Linear relationships must exist between the dependent variable ΔR and the independent variables ΔZ_1 and Z_2.
2. The effects of the independent variables are additive and the ΔZ_1 and ΔZ_2 variables must not be interrelated with one another. Furthermore, the errors of estimate of ΔR from values of ΔZ_1 and ΔZ_2, must be normally distributed with mean zero and constant variance.

In view of these restrictions, it is argued that the advantages of regression analysis are soon canceled by the violation of one or more of the above restrictions in using a particular data set.

Evidence is presented that the above restrictions are not insurmountable obstacles in the development of a linear regression model. If any of the restrictions are violated due to the nature of the data, which appear to invalidate the construction of a linear model, then the model can be reformulated to avoid such violations. For example, the following precautionary procedures are possible:

1. Nonlinear relationships between ΔR and ΔZ variables can be linearized by breaking up the single ΔZ variable into several ΔZ variables, i.e., ΔZ_1, ΔZ_2, ΔZ_3, etc. By

Paper sponsored by Committee on Land Use Evaluation and presented at the 45th Annual Meeting.

doing so, a linear relationship will exist between ΔR and each ΔZ. Transformation of the ΔZ variable by logarithms, cosines, etc., can achieve the same results.

2. The application of factor analysis techniques can create from highly interrelated ΔZ_1 (adj) and ΔZ_2 (adj) variables which are independent of one another. In so doing, the assumption of additive effects of independent variables is confirmed. If such techniques are not available for use or not preferred, then the expected errors of estimate of ΔR which have unequal variances can be dealt with satisfactorily by suitable transformations of the ΔR and/or ΔZ variables to insure constant variance for expected errors of estimate.

Explicit analysis of locational behavior can be incorporated in the model's design. Regression models do not have to depend primarily on a blanket interpretation of past events. The model's development can be shaped in accordance with a theory of allocation of growth of activities or urban development. The researcher in the development of the model will be back and forth between the theory of the model and tests of its behavior with data. Adjustments of the theory will result to improve the model's application with empirical data. However, the theory of the model should not be warped or distorted solely to achieve a best fit to the data.

Development of the model can be achieved by applying several types of regression analysis techniques; for example: (a) ordinary least squares, (b) indirect least squares, (c) limited information-single equation method, (d) 2-stage least squares, (e) simultaneous least squares, and (f) full information maximum likelihood method.

While method (a) deals with single equation models, methods (b) to (f) deal with models formulated as systems of simultaneous equations. If single equation models are formulated, method (a) is adequate and the one to use. However, most activity distributions require models formulated as systems of equations—methods (b) to (f). The relative efficiency of each of the methods for parametric estimation is discussed in the case of simultaneous equation models.

There are distinct computational and economic advantages associated with the use of linear regression analysis. Readily available analysis methods and economical computer programs can be used by the researcher for the model's development. Also, through the economies and flexibility of regression analysis techniques, several test models can be easily evaluated. In general a great deal of knowledge and modeling experience can be gained from constructing and testing regression models.

MODEL DESIGN BY LINEAR EQUATIONS

In the typical model design, one must choose a mathematical framework to describe a hypothesized set of structural relationships. This framework will comprise the variables chosen and specify the ways in which these variables are interrelated. A model framework convenient for use is a linear structural equation as follows:

$$\Delta R = b_1 \Delta Z_1 + b_2 \Delta Z_2 + \ldots + b_k \Delta Z_k + u \qquad (1)$$

Here ΔR is an urban activity variable dependent on the measurements of a number of independent variables, $(\Delta Z_1, \Delta Z_2, \ldots, \Delta Z_k)$. The parameter set $(b_1, \ldots, b_k)$, describes the relationship between the dependent variable and the independent variable set. The error term, u, occurs due to the imperfect fit of a mathematical equation to observed phenomena of urban development. It is the principle of model calibration to estimate the parameter set $(b_1, \ldots, b_k)$, so as to minimize overall the error terms, u, as well as to eliminate systematic bias in the error terms.

Eq. 1 accommodates adequately the situation where the dependent variables, ΔR, to be predicted, are not interrelated with one another. However, many model designs are premised on the occurrence of interrelationships between the dependent variables to be predicted. In accordance with this design requirement, it is desirable to formulate a framework of simultaneous linear equations; for example:

$$\Delta R_1 + a_{12}\,\Delta R_2 + \ldots + a_{1m}\Delta R_m = b_{11}\,\Delta Z_1 + \ldots + b_{1k}\,\Delta Z_k + u_1$$

$$a_{21}\,\Delta R_1 + \Delta R_2 + \ldots + a_{2m}\Delta R_m = b_{21}\,\Delta Z_1 + \ldots + b_{2k}\,\Delta Z_k + u_2$$

$$a_{m1}\Delta R_1 + a_{m2}\Delta R_2 + \ldots + \Delta R_m = b_{m1}\Delta Z_1 + \ldots + b_{mk}\Delta Z_k + u_m \qquad (2)$$

Within this framework, it is possible to account for the interrelationship between the dependent variables, $(\Delta R_1, \ldots, \Delta R_m)$, as well as accommodate the dependency of each ΔR variable on the independent variable set, $(\Delta Z_1, \ldots, \Delta Z_k)$. As in the case of Eq. 1 (which of course is a special case of equation system Eq. 2 where $a_{ij} = 0$ for $i \neq j$) the error terms, u, must account for the imperfect fit by the mathematical equation. The parameter sets $(a_1, \ldots, a_k)$ and $(b_1, \ldots, b_k)$ are estimated so that the overall errors (u_i) are minimized by the regression process of least squares.

The selection and formulation of variables in the model is critical in the model's design. The dependent variables should measure adequately the distribution which we propose to predict. The independent variables should provide adequate explanation of the distribution to be predicted, as well as retaining their separate identity with respect to one another. In particular, the following two criteria are suggested for the formulation of variables:

1. The variables formulated for incorporation into the model should be the same type. That is, variables which are changed in basically different ways by changes in definition of subregional areas and size should not be mixed in a single model. Variables will in general be of two types, i.e., point variables and aggregate variables. Point variables do not tell anything about area aggregates unless multiplied by some base quantity such as total land or total activity. Examples of point variables are densities, accessibilities, and area rate of growth. Area aggregate variables, on the other hand, refer to measurable magnitudes or quantities. Examples of aggregate variables are total population, and total employment or total land area.
2. The construction of the variables should be such that their interpretation is clear. The variables must be capable of being measured and named. Data categories assimilated to form a variable should furnish it with a logical name or explanatory description.

The formulation of variables should simplify the design of the model wherever possible. If two or more variables demonstrate similar locational characteristics and otherwise appear to cluster together due to a similarity in name and procedure of measurement, it is desirable to aggregate the variables into a single variable. Clustering or aggregating dependent variables will simplify the model design by reducing the number of estimating equations of the system. There must be one equation in the model for every dependent variable to be predicted. By aggregating dependent variables, it is possible (all else being equal) to increase substantially the predictive accuracy of the model over what might be achieved with a more complex model. Aggregation of independent variables which are highly interrelated is preferred for other reasons.

CRITERIA FOR APPLYING LINEAR REGRESSION ANALYSIS IN MODEL DESIGN

Linear regression analysis is simply defined as the estimation of the value of one variable (ΔR) from the values of other given variables (other ΔR and/or ΔZ) via a framework of some chosen linear equation. Descriptions of various regression techniques suggested for use in distributing urban activities are described hereinafter. Such regression analysis may be used provided the following criteria are met:

1. It is hypothesized in the construction of the activity distribution model that linear relationships exist between the dependent and independent variables.
2. It is hypothesized that the influences of the variables are additive. While the dependent variables are assumed to be interrelated with each other as well as being

related to the independent variables, it is desirable for the independent variables not to be interrelated with each other.

Linear Influences of Variables

In the application of regression analysis to estimate the parameters of a model it is essential that there is a linear relationship between the expected value of the dependent variables and the independent variables. Fortunately, even when this condition does not apply, it is often possible to modify the original variables in some way so that the new variables meet the requirement. The modifications or transformations of data most commonly applied are the logarithmic, the square root, or the reciprocal.

One of the assumptions of the linear model is the serial independence of the error terms, u, that is, covariance $(u_i, u_{i+j}) = 0$ for all observations i and j, where $j = 0$. However, there are circumstances in which the assumption of a serially independent error term may not apply. It is possible that one may make an incorrect specification of the form of the relationship between dependent and independent variables. For example, one may specify a linear relationship between the ΔR and ΔZ variables when the true relation is quadratic. While the error term in the true relationship may be non-autocorrelated, the new quasi-error term associated with the linear relationship must contain a term in ΔZ^2. If serial correlation exists in the ΔZ-values (i.e., characteristic of time series variables), then serial correlation will occur in the quasi-error terms.

In cases of autocorrelated errors, there are three main consequences of applying straight-forward regression processes without transforming the variables affected:

1. While the estimates of the parameters will be unbiased, their error variances could be larger than those achievable by applying suitable transformations in the estimation process.
2. The estimates of the error variances associated with parameters will be understated.
3. Inefficient predictions with large errors of estimation will be obtained.

The satisfactory manner of testing for linear relationships between the dependent and independent variables is by plotting the relationships between pairs of variables on graph paper. Based on the results, a decision can be made on the value of transforming variables, so as to linearize their influences.

Additive Influences of Independent Variables

Two variables exhibiting a high degree of interrelationship are said to introduce non-additive influences on the dependent variables. Unless interactance terms descriptive of the interrelationships are introduced in the model, there occurs serious ambiguity in the calibration process in separating the influences of the two variables. This ambiguity can be reflected in large fluctuations in the parameters associated with each model derived from calibrations with different aggregations of the subregions and variable sets, etc. Also, the signs associated with parameters of the affected variables may disagree from that expected from a priori reasoning.

Nonadditivity of a particular variable, unless previously eliminated, will frequently cause heterogeneity of error variance which is associated with the estimating equation for a particular dependent variable. This should not occur as it can have a serious effect on the parametric estimation achieved by regression analysis. Regression analysis may only be validly performed provided the error variance of the estimates of the expected value of a dependent variable is constant for all values of the independent variable (i.e., homogeneity of error variance is important).

The degree of interrelationship between variables can be measured in two ways: graphical analysis by plotting pairwise relationships on graph paper, and calculation of bivariate correlation coefficients. The value of the correlation coefficient will vary between minus unity and plus unity, and in either case as it approaches its limits, a high degree of interrelationship or correlation is indicated.

If two independent variables are correlated, one of three courses may be followed: (a) eliminate the one variable considered least important to the model design, or which one believes a priori to be less important; (b) combine the two variables, provided the new aggregate variable can be named and measured; (c) substitute a scale of a variable which is natural (i.e., which experience or theory suggests is additive) to reduce and even eliminate interactance between variables. Examples of transformation by logarithms or reciprocals have been shown to reduce interrelationships.

If it is considered important to include all variables in the model, then course (b) or (c) is preferred.

If course (b) is followed, factor analysis can be useful in aggregating variables into independent, and therefore, additive influences. The basis for conducting factor analysis is a matrix of correlation coefficients describing the pairwise relationships between all variables affected. Factor analysis processes will construct factors comprising a linear function or equation of the variables whose pairwise correlations are being analyzed. The principle for constructing these factors is such that the factors are statistically independent of one another. The factors should be able to be named and associated with an aggregate influence on urban development.

Heterogeneity of error variance, caused by nonadditivity, will usually be reflected by a relationship of the error variance to the mean (m) or expected value of the dependent variable for a particular independent variable. The choice of a suitable variable transformation will frequently depend on the relationship between the error variance

TABLE 1

SUMMARY OF TRANSFORMATIONS[a]

Variance in Terms of Mean m	Transformation	Approximate Variance on New Scale in Absence of Heterogeneity
m	$\sqrt{x}$ or $\sqrt{x + 1/2}$	0.25
$\lambda^2 m$	for small integers	$0.25\lambda^2$
$\lambda^2 m^2$	$\log_e x$, $\log_e (x + 1)$	λ^2
	$\log_{10} x$, $\log_{10} (x + 1)$	$0.189\lambda^2$
$2m^2/(n - 1)$	$\log_e x$	$2/(n - 1)$
$m(1 - m)/n$	$\sin^{-1}\sqrt{x}$ (degrees)	821/n
	$\sin^{-1}\sqrt{x}$ (radians)	0.25/n
$km(1 - m)$	$\sin^{-1}\sqrt{x}$ (radians)	0.25k
$\lambda^2 m^2(1 - m)^2$	$\log_e [x/(1 - x)]$	λ^2
$(1 - m^2)^2/(n - 1)$	$1/2 \log_e [(1+x)/(1-x)]$	$1/(n - 3)$
$m + \lambda^2 m^2$	$\lambda^{-1} \sinh^{-1} (\lambda\sqrt{x})$, or	
	$\lambda^{-1} \sinh^{-1} (\lambda\sqrt{x + 1/2})$	0.25
	for small integers	
$\mu^2(m + \lambda^2 m^2)$	$\lambda^{-1} \sinh^{-1} (\lambda\sqrt{x})$, or	
	$\lambda^{-1} \sinh^{-1} (\lambda\sqrt{x + 1/2})$	$0.25/\mu^2$
	for small integers	

[a]Bartlett, M. S. "The Use of Transformation," Biometrics 3, 39-52, 1947.

and the mean of observations. This relationship is usually determined by empirical analysis with subregional data.

Table 1 gives transformations that have been found to have practical value.

SCOPE OF MODEL DESIGN

The design of the activity distribution model (1, 2) is based on a combination of deductive and inductive reasoning based on observations of urban development patterns. It represents an iterative procedure in which the analyst begins with general observations of subject matter; develops a hypothesis or theory of the causal system which explains the behavior of his subject matter; tests this hypothetical structure for its power to explain the observed data of his field, in this case urban development; studies carefully the discrepancies between the explanation provided by his hypothetical structure and the observed data; revises his hypothetical structure on the basis of these discrepancies; tests the structure again; etc. The analyst is thus back and forth between his theoretical explanation of the causal system and his observation of all possible aspects of the subject matter on urban development. His goal in this iterative process is to reduce the discrepancies between theory and observation to a minimum.

Identification of Equation Systems

The problem of identification in a system of causally interrelated variables is connected with making an empirical estimation of the system from observed data. The problem only exists for systems of simultaneous equations, and does not occur when the area of study can be fully explained by a single equation. Each equation in the system will be designed to explain one dependent variable of the system in terms of those causes which exert a direct or approximate influence on it. These causal variables include both other dependent variables, and independent variables.

The essential meaning of identification can now be stated. Any particular equation in our system is identified if it is sufficiently different from all of the other equations, i.e., in its form, the variables included in it, and any restrictions on the values which its parameters can take. By "sufficiently different" we mean that it must be impossible to arrive at an equation which "looks like" the particular equation we are testing by any linear combination of other equations in the system, or of all of the equations including the one being tested.

Sample Identification Problem. Suppose that our system consists of two dependent variables, ΔR_1, ΔR_2, and three independent variables, ΔZ_1, ΔZ_2, ΔZ_3. Suppose that we are assuming linear relations, and that we have as yet no clear ideas about structure specification. We might then simply put all variables in the system into each equation.

$$\begin{aligned} &(a)\quad a_{11}\Delta R_1 + a_{12}\Delta R_2 + b_{11}\Delta Z_1 + b_{12}\Delta Z_2 + b_{13}\Delta Z_3 = u_1 \\ &(b)\quad a_{21}\Delta R_1 + a_{22}\Delta R_2 + b_{21}\Delta Z_1 + b_{22}\Delta Z_2 + b_{23}\Delta Z_3 = u_2 \end{aligned} \qquad (1)$$

The a's and b's are constant coefficients or parameters, and the u's can be treated here as either constant terms or as random disturbances. We can assume that Eq. (a) is supposed to explain ΔR_1 and that Eq. (b) is intended to explain ΔR_2. Let us further assume that the system we are analyzing is represented by a sample of observed data.

$$\left[\Delta R_{it}\right], \left[\Delta Z_{jt}\right] \quad (i = 1, 2;\ j = 1, 2, 3;\ t = 1, 2, \ldots, T) \qquad (2)$$

We now attempt to use these data to estimate the parameters of our system (1) above. But since the two equations look exactly alike, when we apply our observed data to the estimation of parameters we get exactly the same result for each equation. There is no way of distinguishing the behavior of one part of the system from that of the other using empirical methods.

Suppose, next that we do more work on the theory of our system, and arrive at a specification which excludes ΔZ_2 and ΔZ_3 as variables from (a) and ΔZ_3 from (b). Let us call the new equations (c) and (d). Now the two equations "look different" from each other. We have restricted b_{12}, b_{13} and b_{23} to zero. This is the most common kind of restriction which aids identification. But is there still any danger of getting those two equations mixed up in empirical estimation? Suppose we test by making a linear combination of (c) and (d). Thus suppose we form ℓ(c) + m(d) where ℓ and m are arbitrary multipliens. The resulting equation has the form

$$a_1\Delta R_1 + a_2\Delta R_2 + b_1\Delta Z_1 + b_2\Delta Z_2 = v \qquad (3)$$

This is different from the new specification we have made for (c), for it excluded both ΔZ_1 and ΔZ_2, but it is no different from our new specification for (d). In our new system (c) is completely distinguishable empirically from the rest of the system, but (d) is not. Therefore (c) is identified, and (d) is not identified.

Now suppose that our theoretical specification had removed ΔZ_2 and Z_3 from (a) and ΔZ_1 from (b), giving equations (e) and (f). Suppose we make a linear combination ℓ(e) + m(f),

$$a_1\Delta R_1 + a_2\Delta R_2 + b_1\Delta Z_1 + b_2\Delta Z_2 + b_3\Delta Z_3 = w \qquad (4)$$

This form does not look like either (e) or (f), and both equations in our system are fully distinguishable and hence identified.

In conclusion, the main basis for identification is the inclusion of only the main causal variables in each equation, and the exclusion of irrelevant variables, both dependent and independent. But there are other bases for obtaining distinguishability of one equation from all others, and these include cases like the following. It might be that there is a natural restriction that two parameters in the equation have a preordained ratio to each other, or that one or more parameters have preordained values, indicated by theory, or arrived at by separate studies. Sometimes a nonlinearity in an equation may insure identifiability, or even a specification of differences in the variances of the random components in particular equations may achieve this.

A necessary, but not sufficient, condition of an identified system of m equations is that in each equation, at least m - 1 of the variables are restricted, usually by setting them to zero. This is known as the "order" condition of identifiability. If fewer than m - 1 variables are restricted in any equation, the system is said to be under identified, and cannot be solved by the parameter estimation programs. If more than m - 1 variables in some equations and at least m - 1 variables in all equations are restricted, the system is said to be over identified. This will usually be the case with activity distribution models.

Methods of Identifying a Model. By and large, the identification of the system of simultaneous equations which comprise the model will be determined by a priori reasoning in support of a particular theory of urban development. These are, however, empirical tests which can be applied as a guide in choosing an appropriate identification for the model.

Tests of Model Design

The testing of the model is usually carried out by regression processes, such as least squares (LS) or maximum likelihood (ML). Their purpose is to make the best possible tests and estimates of the structural parameters associated with variables of the model. In doing so, a complete separation is sought between the systematic part of the relationships and the random part. Generally, testing can profitably begin with an examination of our estimates of the random component.

An examination is conducted of nonsystematic residuals of the equation which the estimation process may have produced. If these reveal any trend, cycle or sawtoothed behavior then the model design (i.e., its identification) is on this basis rejected. It is concluded that the model does not contain all of the systematic forces which affect the dependent variable being explained, or it may contain some forces which should not be there.

Next, one examines the standard errors of the parameters attributable to variances associated with the observed data and conducts accompanying t-tests of significance. Here one tests again the model design, this time to see which variables test out as significant and as causes affecting dependent variables. But these tests can only be suggestive rather than rigorous, if our residual has already tested to be nonrandom and containing systematic elements.

In making the tests of significance of parameters (and hence of the associated causes), the model design can be open to two types of error. First, the test may reject a design which is really appropriate. This is the well-known Type I error. It can arise because the source of data is not complete or adequately representative of subregional development patterns. Application of more representative data, with an appropriate level of significance can reduce this danger.

A second kind of error which one may make is to accept a design which is false. This is the Type II error. Some other identification of the model is correct, but the one chosen has produced estimates which happen to fali into the range of acceptance for the model. Here we have an identification error which could slip by the tests.

Finally, one tests the results at this stage through reapplying to them one's knowledge of the subject matter. On the basis of general observation of the pattern of development, and of the tests of the primary model design on this basis, one achieves concepts about the sizes and signs of the parameters associated with the variables of the model. If the regression tests produce results which are markedly different from expected, one must take this as a rejection of the model design, or otherwise as some combination of data error and error in the model's identification. Consequently, it is such rejections which lead the analyst forward in the iterative process of model testing.

During this process of iterative revisions of the identification, there is always the danger of warping the theory, and hence design, to make the model fit the particular data source. This is a real trap, and no doubt one could fall into it. But there is a defense against it. The defense lies in carefully preserving the strength, logic and realism of the model's design. It is only when the observed data, and the discrepancies or residuals between observed data and the systematic explanation, reveal some clearly relevant but hitherto unsuspected force or omitted force that the identification should be revised. Design should never be altered merely to get a good statistical fit when the theoretical underpinning of such alterations is weak, illogical, and unrealistic.

When the scientific process has reached a terminal stage, one should have minimal identification errors, and hence the estimates of standard errors of estimates should be realistic. During the process one has resisted rejecting a good theory on the basis of statistical tests, while at the same time one has been even more resistant to warping a design solely to get good statistical fits. The systematic model should be in agreement without general observations and knowledge about the subregional development. And finally, the residuals should be in a purely random sequence, with mean zero and constant variance.

The test of successful estimation of the true model comes partly in its explanatory power, and partly in its predictive power. If one has found satisfactory causal explanation of development, and if the model is performing in a known way, one should be able to make satisfactory predictions.

REGRESSION PROCESSES

Development of the model can be achieved by applying several types of regression analysis methods.

Ordinary Least Squares

One applies ordinary least squares to a single equation in a model (3, Chap. 4, pp. 106-138), i.e.,

$$\Delta R = B\Delta Z + u \qquad (1.1)$$

where

ΔR = vector of dependent variables;
ΔZ = vector of independent variables;
B = parameter associated with independent variables; and
u = residual error.

If, however, there are two or more dependent variables in each equation one does not know which dependent variable to select as the primary dependent variable of an equation, i.e.,

$$A\Delta R = B\Delta Z + u \tag{1.2}$$

where A = parameters associated with the dependent variables. The remaining dependent variables are always correlated with the error term in the equation because of the simultaneous nature of the equations in the model. Therefore, ordinary least square estimators are always biased (estimate does not equal true value) and they will also be inconsistent—for increasing numbers of sample observations, the estimates continue to be biased (3, Chap. 6, pp. 148-150).

For these reasons ordinary least squares is considered to be an unsuitable estimation method for dealing with systems of simultaneous equations. On the other hand, when dealing with a single equation containing one dependent variable, it is the method to use.

Indirect Least Squares

In the situation where a system of simultaneous equations is exactly identified, this is the proper estimation method to use. The other simultaneous estimation methods to be mentioned below always provide identical estimators to the indirect least squares method for the case of exact identification (exactly m - 1 of the parameters are set equal to zero where m is the number of dependent variables in total). The indirect least squares method is less complicated than the other methods, hence it provides definite computation economies.

The procedure (4, Chap. 4.4, pp. 135-137) is to estimate the parameters of the reduced form equations by application of the ordinary least squares method. A reduced form equation has only one dependent variable which is defined as the primary dependent variable, i.e.

$$\Delta R = D\Delta Z + u \tag{2.1}$$

By deciding that a certain number of the parameters in each equation of the simultaneous equation system are zero, the reduced form equations are converted into a simultaneous system where each equation contains one or more dependent variables, i.e., multiply (2.1) by A to obtain

$$A\Delta R = AD\Delta Z + Au \tag{2.2}$$

Write B = AD; therefore (2.1) is converted into a simultaneous system

$$A\Delta R = B\Delta Z + u$$

To recap, the exact number of parameters per equation which are set equal to zero is m - 1.

Limited Information Estimation Methods

Limited Information Single Equation Method (LISE) or Least Variance Ratio Method (LVR). This is a limited information maximum likelihood approach. It is a maximum likelihood approach (4, Chap. 6.2, pp. 166-167) in that the logarithmic likelihood function for the dependent variable is defined, i.e.

$$L(\alpha) = \tfrac{1}{2} \log AWA' - \tfrac{1}{2} \log \alpha M \alpha' + k - \tfrac{1}{2} \log \text{determinant } W \qquad (3.1)$$

where

$$\alpha = [A, B]$$

$$M = \begin{bmatrix} M_{\Delta R \Delta R} & M_{\Delta R \Delta Z} \\ M_{\Delta Z \Delta R} & M_{\Delta Z \Delta Z} \end{bmatrix}$$

$$M_{\Delta R \Delta R} = M = \frac{1}{T} \sum_{t=1}^{T} \Delta R_t^2 \text{ etc., (T = number of observations)}$$

$$W = M_{\Delta R \Delta R} - M_{\Delta R \Delta Z} M^{-1}_{\Delta Z \Delta Z} M_{\Delta Z \Delta R}$$

Next, the function is maximized to yield uniquely the ratios of the parameters associated with the dependent variables of each equation. By setting one of the parameters equal to unity the remaining parameters are defined. The parameters of the independent variables are determined by solving a mathematical identity of dependent variable parameters and values of both dependent and independent variables.

The application of the method requires the user to know the specification of the single equation being estimated (i.e., which parameters are zero), and the independent variables appearing in the remaining equations which are assumed to have non-zero parameters. The detailed specification concerning the parameters of dependent variables in remaining equations is assumed unknown. Hence, only limited information needs to be known to obtain the estimators.

Two-Stage Least Squares

The basic idea (3, Chap. 9.5, pp. 258-260) of the 2-stage least squares (TSLS) is to select a dependent variable in each equation of the system and set its parameter equal to unity; i.e., rewrite $A\Delta R = B\Delta Z + u$

$$\Delta R_1 = A_2 \Delta R_2 + B_2 \Delta Z + u \qquad (3.2)$$

Next replace the remaining dependent variables by their estimates based on ordinary least squares regression between each dependent variable and all independent variables in the model,

$$\Delta R_2 = \Delta Z (\Delta Z' \Delta Z)^{-1} \Delta Z' \Delta R_2 \qquad (3.3)$$

Finally ordinary least squares is applied to the selected dependent variable, the regression estimates of the remaining dependent variables, and the independent variables in each single equation.

There is a basic similarity between LISE and 2-stage least squares as they both make use of all the independent variables in the model in order to estimate the parameters of a single equation, but do not require a detailed specification of the dependent variables in the remaining equations of the model. Both methods are consistent estimating methods. For large numbers of sample observations, both methods provide unbiased estimates of the parameters. It is reported that for special cases with a small number of observations, the 2-stage least squares method may provide more efficient estimators than LISE—estimators with smaller limiting variance (5).

Full Information Method (FI)

This method implies the use of full information concerning the specification of the simultaneous equation system. The FI methods are anticipated to provide the most efficient estimators of all the methods. There are two different techniques which comprise the FI method, simultaneous least squares and maximum likelihood.

Simultaneous Least Squares (SLS)

SLS (6) is a distribution free method of estimation (no assumption is made about the distribution of residual error). The method is the simultaneous equation counterpart of ordinary least squares. It takes completely into account the simultaneous interactions of all dependent variables in the system:

$$A\Delta R = B\Delta Z + u \tag{4.1}$$

It is a least squares method in that the sum of the squared deviations between observed and estimated dependent variables are minimized; i.e. minimize

$$E^2 = \sum_{t=1}^{T} \sum_{i=1}^{N} u^2_{sit}$$

where

$$u_s = A^{-1} u$$

Maximum Likelihood Technique (ML)

Complete information on the simultaneous system is taken into account (4, Chap. 5, pp. 143-162). The likelihood function for the dependent variables, conditional upon the values of the independent variables, is determined for the complete model. By assuming that the residuals of the estimating equations are multivariate normally distributed, the logarithmic likelihood function is defined,

$$L(\alpha, \sigma) = \log \det B - \tfrac{1}{2} \text{trace } (\alpha' \sigma^{-1} \alpha M) + k - \tfrac{1}{2} \log \det \sigma$$

where

det = determinant;
α = [A, B];
σ = non-singular covariance matrix of residual error u; and
trace matrix R = $\sum_i r_{ii}$ (sum of diagonal elements).

Maximizing the logarithm of the function with respect to the parameters of the model and its residuals lead to difficult estimating equations.

There are two assumptions involved in the use of ML, which may restrict its application. The first is the assumption that the residual errors are multivariate normally distributed. While the distributions of the residuals are probably bell-shaped and may be asymptotically normal (a property of large samples), the assumption of normality is not closely met with small samples of data. The second assumption (7) concerns the optimal properties of structure estimation. If the residual errors are normally distributed, both maximum likelihood and least square techniques lead to identical results which are linear unbiased estimates. However, where the residuals are non-normal then the ML and LS estimators are quite different. Nevertheless, the LS estimates are still the best linear unbiased estimates.

In conclusion, SLS is preferred to ML because of its distribution free properties and secondly because of anticipated computer economies. The computation economies are achieved by using a truncated procedure of SLS. This gain will, of course, be at the small expense of loss of accuracy in estimation. In truncated SLS the results are accepted after two or three stages of the recursive procedure of SLS estimation.

FACTOR ANALYSIS PROCESS

Variables which possess high statistical association are grouped together in clusters called factors. In particular, the intercorrelations among all the variables under study constitute the basic data for factor analysis (8).

All variables are assumed to be in standardized form, i.e., each has a mean value of zero and a variance of unity. It is the object of factor analysis to represent a variable in terms of several underlying factors, by a simple mathematical model of the linear form,

$$\Delta_{Rj} \text{ or } \Delta Z_j = a_{j1} F_1 + a_{j2} F_2 + \ldots + a_{jf} F_f$$

Naming the Factors

The factors are not named by the process, and this anonymity must be removed before the statistical association indicated by factor analysis can be evaluated against the planner's a priori knowledge of cause and effect relationships. The variables which are most closely associated with (those which supposedly make the most significant contribution to) each cluster should help in naming the factor.

Significance of Factors

The relative importance of each factor is indicated by its eigenvalue, which represents a measure of the total contribution of the factor to the variances of all the variables being analyzed. Eigenvalues for all factors are produced by the technique. An eigenvalue of unity or greater is considered to indicate a significant factor. Experience with our prototype activities distribution models (1, 2) has shown, however, that there are a small number of factors with very high eigenvalues, a few more with eigenvalues of unity or more, and a large number of factors which have eigenvalues less than unity. The latter, strictly speaking, are considered little more than statistical "noise," whose contribution to the variances of the variables will generally be insignificant.

Selection of Factors

The factor analysis process provides for specifying the number of factors to be used. Normally the process discards all factors with an eigenvalue less than unity. In some instances, the arguments for using unity as a cutoff are marginal, and in some cases a factor with an eigenvalue less than one may be significant. With a small number of input variables, a factor with an eigenvalue of less than one could make a significant contribution to the variance of the variables. In such cases, one may specify the number of factors required.

Regardless of which cutoff option is employed, the eigenvectors associated with eigenvalues of the factors are computed and are normalized so that the squared eigenvector coefficients associated with each factor add to unity or less. The normalized eigenvector coefficients associated with each factor (known as factor loadings) are produced in array form.

Structure of Each Factor

The construction of factors is established by a regression procedure, based on the array of factor loadings. Each factor is presented as a linear function of the variables. An array is produced which indicates for each factor the statistical importance of each variable in its construction.

Factor Rotation

There is a possibility that several factors will look very much alike, and possess similar eigenvalues. In order to sharpen the picture of the system as much as possible, a varimax method of rotation is utilized in factor analysis processes. The rotation should maximize large factor loadings and minimize small ones, and the distinction between factors should be much sharper in the rotated than in the unrotated case. Both the unrotated and rotated arrays are true shadows of the same shape taken in different lights. Traditionally, the multiplicity of true shadows offered by factor analysis has deterred investigators from using the method as a "proof." The cautious investigator has assured himself that he uses it (in moderation) only to stimulate his insight into a mass of data; that is, to prompt a review of his logic.

It is emphasized that use of factor analysis is always subject to demand for a logical explanation of clustering. However, it seems intuitively attractive with the large amount of data available in computer-size models to suppose that the surface of the factor shape is sufficiently regular that the maxima found by rotation, if not the best view, is at least one of the good views.

REFERENCES

1. Hill, D. M., Brand, D., and Hansen, W. B. Prototype Development of a Statistical Land Use Prediction Model for the Greater Boston Region. Highway Research Record 114, pp. 51-70, 1966.
2. Irwin, N. A., and Brand, D. Planning and Forecasting Metropolitan Development. Traffic Quarterly, Oct. 1965.
3. Johnston, J. Econometric Methods. McGraw-Hill, New York.
4. Cowles Commission for Research in Economics. Studies in Econometric Method.
5. Nagar, A. L. The Bias and Moment Matrix of the General k-Class Estimators of the Parameters in Simultaneous Equations. Econometrika, Vol. 27, pp. 575-595, 1959.
6. Brown, T. M. Simultaneous Least Squares: A Distribution Free Method of Equation System Structure Estimation. International Economic Review, Vol. 1, No. 3, Sept. 1960.
7. David, F. V., and Neyman, J. Extension of the Markoff Theorem on Least Squares. Statistical Research Memoirs, Vol. II, London: Dept. of Statistics, Univ. of London, Univ. College, pp. 105-116, Dec. 1938.
8. Harman, H. H. Modern Factor Analysis. Univ. of Chicago Press, 1960.