AN INTRODUCTION TO THE USE OF SIMULTANEOUS-EQUATION REGRESSION ANALYSIS IN GEOGRAPHY

D. Todd

(SRN 0 88094 028

4

Printed in Great Britain by Headley Brothers Ltd The Invicta Press Ashford Kent and London

CATMOG

(Concepts and Techniques in Modern Geography)

CATMOG has been created to fill a teaching need in the field of quantitative methods in undergraduate geography courses. These texts are admirable guides for the teachers, yet cheap enough for student purchase as the basis of class-work. Each book is written by an author currently working with the technique or concept he describes.

1. An introduction to Markov chain analysis - L. Collins
2. Distance decay in spatial interactions - P.J. Taylor
3. Understanding canonical correlation analysis - D. Clark
4. Some theoretical and applied aspects of spatial interaction shopping models - S. Openshaw
5. An introduction to trend surface analysis - D. Unwin
6. Classification in geography - R.J. Johnston
7. An introduction to factor analytical techniques - J.B. Goddard & A. Kirby
8. Principal components analysis - S. Daultrey
9. Causal inferences from dichotomous variables - N. Davidson
10. Introduction to the use of logit models in geography - N. Wrigley
11. Linear programming: elementary geographical applications of the transportation problem - A. Hay
12. An introduction to quadrat analysis - R.W. Thomas
13. An introduction to time-geography - N.J. Thrift
14. An introduction to graph theoretical methods in geography - K.J. Tinkler
15. Linear regression in geography - R. Ferguson
16. Probability surface mapping. An introduction with examples and Fortran programs - N. Wrigley
17. Sampling methods for geographical research - C. Dixon & B. Leach
18. Questionnaires and interviews in geographical research - C. Dixon & B. Leach
19. Analysis of frequency distributions - V. Gardiner & G. Gardiner
20. Analysis of covariance and comparison of regression lines - J. Silk
21. An introduction to the use of simultaneous-equation regression analysis in geography - D. Todd

Other titles in preparation

This series, Concepts and Techniques in Modern Geography is produced by the Study Group in Quantitative Methods, of the Institute of British Geographers.
For details of membership of the Study Group, write to the Institute of British Geographers,1 Kensington Gore, London, S.W.7.
The series is published by Geo Abstracts, University of East Anglia, Norwich, NR4 7TJ, to whom all other enquiries should be addressed.

AN INTRODUCTION TO THE USE OF SIMULTANEOUS-EQUATION REGRESSION IN GEOGRAPHY

by

Daniel Todd

Department of Geography University of Manitoba

CONTENTS

		Page
I.	INTRODUCTION	
	(i) Preface	3
	(ii) The Concept of Regression	3
	(iii) The Precedent of Econometrics	6
	(iv) The Simultaneity Aspect	7
	(v) The Case for S-E Formulations in Geography	9
II.	THE MATHEMATICAL BASIS OF S-E REGRESSION	
	(i) The Identification Problem	15
	(ii) The Consistency Problem	19
	(iii) A Question of Statistical Significance	21
	(iv) Two-Stage Least Squares	23
III.	A TWO-EQUATION MIGRATION MODEL EXAMPLE	
	(i) Conceptual Basis	25
	(ii) Calibration	26
IV.	AN EXAMPLE OF A MULTIPLE-EQUATION MIGRATION MODEL	
	(i) Background	32
	(ii) Structuring of the Model	34
	(iii) Computational Results	36
V.	CONCLUSION	39
VI.	BIBLIOGRAPHY	40
VII.	APPENDICES	42

Acknowledgements

I would like to thank Ron Johnston and Bob Haining for their helpful comments.

I. INTRODUCTION

(i) Preface

Regression analysis is a procedure whereby the effects of one or more independent variables (regressors) can be assessed on a control, or dependent, variable (regressand). By definition, the dependent variable is expected to be functionally regulated by the other variables, that is, it is 'explained' to varying degrees by the combination of independent variables. The connotation of 'explanation' has made regression analysis a very popular analytical tool in geography as, indeed, in the social sciences at large. In consequence, regression is put to catholic use, ranging from description and prediction to model building (McNeil *et al,* 1975). However, it is somewhat paradoxical that most of this popular trend towards regression applications in geography has focused on only one particular variant of regression analysis, albeit the earliest established one, and that is the classical least squares model. Other variants of regression, including functional analysis (Mark and Peucker, 1978), polynominal regressions (Unwin, 1975), regression equations using categorized variables (Wrigley, 1976) and simultaneous-equation regression models (Meyer, 1974), are much less known in geography. It is the aim of this monograph to acquaint the reader with the last of these kinds of regression, both in terms of outlining the scope for using multiple-equation regression models in geography, and also to introduce him to the established procedures for determining the solution of such models. On the part of the reader, it will be assumed that he is familiar with simple algebra and matrix presentation, and that he is aware of the utility of classical least squares regression in spatial analysis. In view of the fact that classical regression is the point of departure of this monograph, it is strongly recommended that the reader familiarize himself with the work of Ferguson (1977) in this series.

(ii) The Concept of Regression

Regression analysis, in a nutshell, is a statistical search procedure whereby the relative influences of all the factors presumed to act on a dependent variable can be isolated and gauged. As such, it is a framework from which a model can be established to infer the interaction between a dependent variable and its regressors. The term 'interaction' refers to situations where the effect of one variable depends upon another and hence are appropriate in regression analysis where a variable, appropriately called a dependent variable, is deliberately established with the sole purpose of determining the impacts of other variables upon it. On the other hand, the regressors are the variables believed to manipulate the dependent variable and they are assumed to be independent of each other (hence, their alternate term 'independent variables') so that original sources of interaction with the dependent variable can be identified. Figure 1 illustrates the conventional relationships of multiple regression whereby a one-way functional linkage is monitored between a dependent variable (represented by a circle) and a set of independent variables (enclosed in rectangles). In addition, regression allows for a proportion of the variance embodied in the dependent variable to be unexplained by the regressors included in the analysis. This unexplained variance would arise when the model is incomplete, that is, when it has

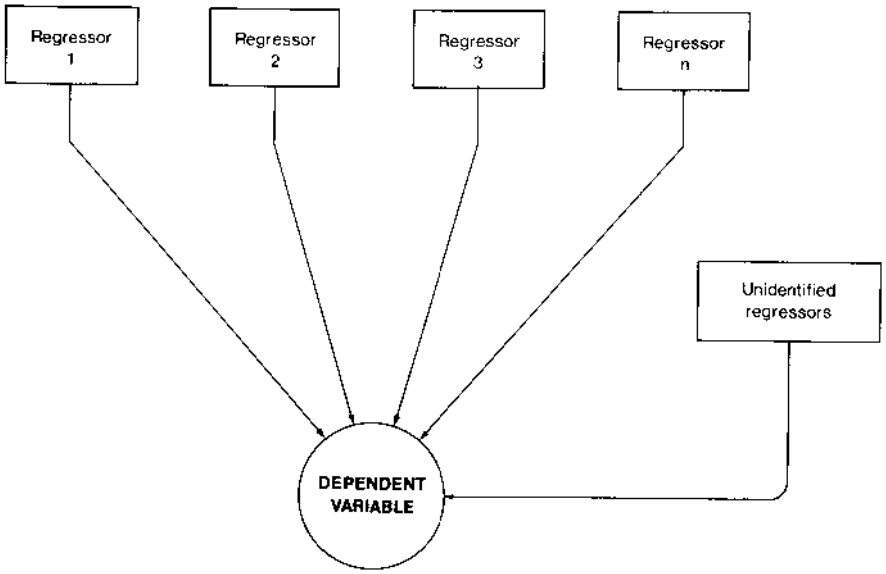

Figure 1. The multiple regression situation

failed to incorporate all of the independent phenomena which impinge on the dependent variable, and may be denoted by the generic expression 'unidentified regressors'.

Unfortunately, conventional multiple regression rests on two crucial assumptions which often prove difficult to maintain in reality. These are respectively the assumption that the regressors are truly independent of each other and the assumption of a one-way interaction linkage whereby the regressand is solely dependent on the regressor and therefore cannot contribute to the variation within the regressor. An example taken from political science will illustrate how inappropriate such assumptions may be in model construction. Alker (as simplified by Blalock, 1969, pp 60-1) proposes that four salient variables are interrelated and cannot be explained in isolation from each other. The four variables, communist vote (mnemonic of CV), polyarchy or the government of many (PO), political participation (PP), and domestic group violence (DG), are believed to be reciprocally interrelated in both negative and positive manner. Thus, a feedback loop is suggested whereby any increase in the communist vote would result in a strengthening of polyarchy which would have a depressing effect on domestic group violence as well as providing a fillip to political participation. In turn, political participation would likewise decrease violence, and this decline would serve to promote the communist vote. Over time, the political system would stabilize itself and fewer and fewer impulses would feed through the system to augment the communist vote. Figure 2 indicates the causal nature of the loop and it also introduces three further variables which are outside the feedback mechanism but which influence the constituents of the loop. It is anticipated

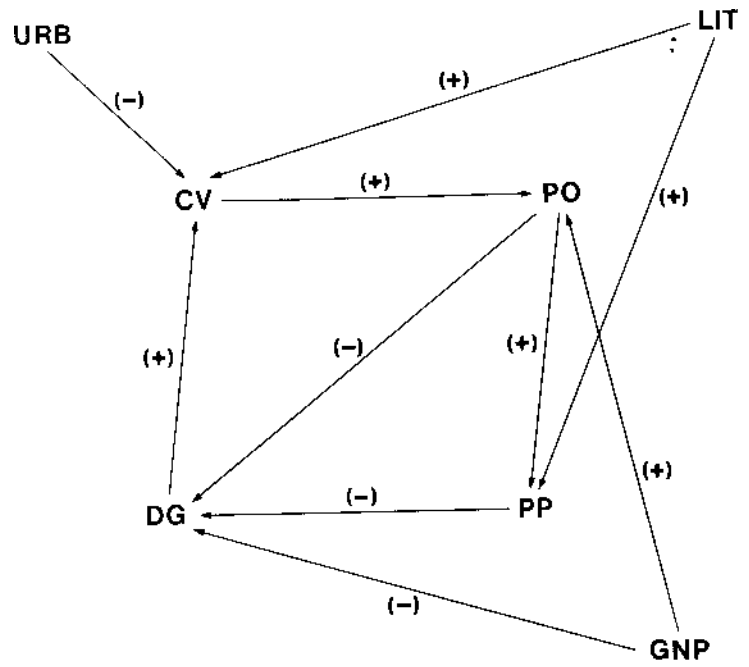

Figure 2. A causal mechanism for political variables (after Alker)

that urbanization (URB) inversely influences the communist vote, while literacy (LIT) directly influences both this variable and political participation, leaving per capita gross national product (GNP) to affect polyarchy directly and affect domestic group violence negatively.

The unsuitability of conventional multiple regression becomes apparent when we establish a series of regression equations to deal with this situation. viz:

$$\begin{aligned} CV &= a_0 + a_1URB + a_2LIT + a_3DG \\ PO &= b_0 + b_1CV + b_2GNP \\ PP &= c_0 + c_1PO + c_2LIT \\ DG &= d_0 + d_1PO + d_2PP + d_3GNP \end{aligned} \qquad (1)$$

(note that the positive signs merely denote the additive nature of the equations and do not reflect postulated directions of signs for the regression coefficients). According to the assumptions of multiple regression, each equation exists as an independent entity and is solvable as such. Thus, the equation with PO as dependent variable maintains that polyarchy can be determined by two 'givens', CV and GNP, even though one of these so-called in-

dependent variables is not independent at all, but is determined within the causal loop itself; a loop, incidentally, which must take account of the feedback effects of PO on CV. At a stroke, the twin tenets of one-way causality and independence are repudiated. In fact, as is evident from Figure 2, only three of the seven variables in the model are truly independent; the others represented in equation (1) are so interrelated in the feedback loop that they can only be solved through a simultaneous-equation procedure. Instances such as these are common in the social sciences, but it is perhaps in economic analysis more so than elsewhere that research problems have been construed to handle cases of simultaneous-equation regression. It is to econometrics, or that economics specialty concerned with statistical modelling, that we now turn.

(iii) The Precedent of Econometrics

In econometrics, situations often arise where the independent variables of a regression equation are inadequate estimators of the dependent variable because the single-equation solution ignores information about the regressors which are specified in other relationships. The classic example of this state of affairs occurs in that most basic of economic concepts, the supply-demand model. Figure 3 illustrates the well-known relationship between the positively sloping supply (S)-curve and the negatively sloping demand (D)-curve.

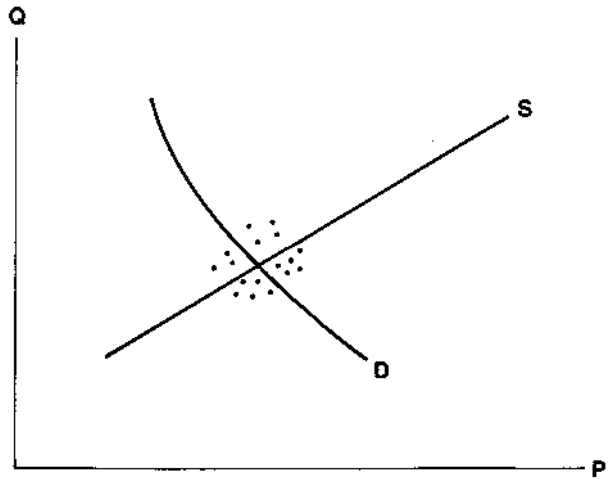

Figure 3. The supply-demand relationship

The model claims that the price of a commodity, P, is regulated by its level of demand, Q^d, while demand, in turn, is a function of supply, Q^s, itself regulated by price. This circular causal arrangement becomes clear when the supply and demand equations are spelled out under the usual assumption of equilibrium:

$$Q^d = \beta_0 + \beta_1 P \qquad \text{where } Q^d = Q^s \qquad (2)$$
$$Q^s = \alpha_0 + \alpha_1 P$$

In equilibrium, the intersection of the supply and demand curves gives the unit price and quantity of commodity supplied (and, by definition, demanded). It is apparent that neither functional equation, standing alone, would provide a true representation of the interdependence between supply and demand: the former would merely define quantity demanded as an inverse function of price whereas the latter would have supply amounting to a positive function of price. The implications of equilibrium can only be gauged when each equation takes into account the consequences of the other, that is to say, the *simultaneity bias* is accommodated in a three-equation model (the demand and supply equations are *behavioural* in that they vary according to the behaviour of economic agents, and the equilibrium assumption is an *identity* and fixed; together they comprise a model of structural equations).

This simple example contains the two leading properties which distinguish single-equation regressions from their simultaneous-equation offspring, namely, the requirement of a model consisting of several structural equations instead of just one, and the allowance for inter-dependence between equations in place of the dependent-independent variable balance of single-equation regressions wherein all 'explanation' is presumed to arise from the right-hand side of the equation. It is mandatory that both properties are present in simultaneous-equation (S-E) regression, otherwise the issue of segmentability is said to occur. In essence a model is segmentable if its equations can be subdivided into sets which can be solved without reference to the other sets. Conceivably, this could occur in the aforementioned supply-demand model if we dismissed the assumption of Say's law with its equality between supply and demand. Thus, in a state of disequilibrium demand may be a function of consumer access to credit and may not be allayed by traditional cost factors, while supply is set by the firm's desire to keep its skilled workers gainfully employed and not lose them to competitors. In both cases the price mechanism breaks down and the interdependence between supply and demand is removed. Now, the supply-demand model would consist of two equations which would not need to be solved simultaneously because they are conceptually unconnected. In reality, of course, segmentable models are much more complex than this and they usually entail the splitting up of a large number of equations into segments which explain crucial 'indicators' and which are then substituted into a key equation as explanators of the prime dependent variable under consideration (e.g. as in recursive modelling of dynamic situations).

(iv) The Simultaneity Aspect

Simultaneous-equations provide a framework for monitoring effects of variables contained in one equation as they feed back into variables contained in other equations. Their allowance for a number of variables dependent to some degree on other variables scattered throughout the equations necessitates an elaboration of terminology over and above the division between dependent and independent variables in classical least squares regression. This is accomplished when we distinguish between those variables that are truly independent and those that depend on other equations in the model for their solution. Econometricians refer to the first kind as *exogenous* variables, or those which are given and not determined by the model. On the other hand,

those which are to be explained by the model and its constituent equations are termed jointly dependent or *endogenous* variables. In fact, the number of equations in the model is limited to the number of endogenous variables for each of these variables must have its accompanying right-hand side regressors. Yet, an endogenous variable may appear on both sides of equations, either occupying the dependent variable position in the regression equation set up solely to explain its variation, or acting as a regressor in other equations of the model. Indeed, a *lagged endogenous variable* (the phenomenon measured at a previous time period and therefore known) can appear as a regressor in the equation containing the same, although unlagged, phenomenon as dependent variable. Lagged endogenous variables along with exogenous variables are lumped together as *predetermined variables.*

The distinction between endogenous and predetermined variables is essential because the former can appear anywhere in the equation system, given the interdependent nature of phenomena. The inherent interdependent aspect of endogenous variables is offset by the assumption that exogenous variables are fully independent under stochastic conditions, that is, the standard requirement of independent variables in classical least squares regression is retained. These distinctions are accommodated in the notation with y representing endogenous variables and z standing for exogenous variables. An equation system will have y_1, y_2 ... y_G endogenous variables (with the corollary of G equations) and z_1, z_2 ... z_K exogenous variables. By the same token, structural parameters for y are denoted by β and those for z are signified by γ. Thus, a simultaneous-equation model takes the following form:

$$\begin{aligned}
&\beta_{11}\,y_1 + \beta_{12}\,y_2 + \ldots + \beta_{1G}\,y_G + \gamma_{11}\,z_1 + \gamma_{12}\,z_2 + \ldots \gamma_{1K}\,z_K = 0\\
&\beta_{21}\,y_1 + \beta_{22}\,y_2 + \ldots + \beta_{2G}\,y_G + \gamma_{21}\,z_1 + \gamma_{22}\,z_2 + \ldots \gamma_{2K}\,z_K = 0\\
&\vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots \qquad\qquad \vdots\\
&\beta_{G1}\,y_1 + \beta_{G2}\,y_2 + \ldots + \beta_{GG}\,y_G + \gamma_{G1}\,z_1 + \gamma_{G2}\,z_2 + \ldots \gamma_{GK}\,z_K = 0
\end{aligned} \qquad (3)$$

where the first parameter subscript refers to the particular equation and the second identifies the variable. Equation (3) can be reformulated in standard matrix form:

$$\begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1G} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2G} \\ \vdots & & & \\ \beta_{G1} & \beta_{G2} & \cdots & \beta_{GG} \end{bmatrix} \cdot \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_G \end{bmatrix} + \begin{bmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1K} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2K} \\ \vdots & & & \\ \gamma_{G1} & \gamma_{G2} & \cdots & \gamma_{GK} \end{bmatrix} \cdot \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_K \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (4)$$

or, in shorthand as:

$$BY + \Gamma Z = 0 \qquad (5)$$

where B is capital beta and Γ is capital gamma. However, a regression version of the above equation system must take account of stochastic elements because behavioural equations cannot be divorced from random error. The stochastic element is furnished by incorporating the disturbance term, U, viz:

$$BY + \Gamma Z + U = 0 \qquad (6)$$

where $U = (u_1, u_2, \ldots u_G)$. A redrafting of equation (6) in the standard regression format after allowing for U gives:

$$BY = \Gamma Z + U \qquad (7)$$

In short, therefore, the solution of S-E regression is contingent on establishing the structural parameters of the endogenous variables in a manner that makes them dependent on the structural parameters of the predetermined variables and the disturbance term.

(v) The Case for S-E Formulations in Geography

The properties of interdependence and multiplicity of equations embodied in S-E regression can be put to useful service in the analysis of spatial systems. Four fields of spatial inquiry where this application is already underway are the realms of land use-transportation modelling, regional growth theory, modelling geomorphological pattern, and migration analysis. A brief outline of each of these fields is given in order to familiarize the reader with the various guises of structural equations.

A. Land use - allocation modelling

Many regional econometric modellers have taken inspiration for spatial components of their models from the examples provided by land use - allocation planners (Glickman, 1977). Land use - allocation models are concerned with forecasting changes in fundamental urban phenomena such as population and employment opportunities and then accommodating the spatial distribution of these growth factors throughout the subject region. These models must be cognizant of the interplay between jobs, population, public facilities and transport networks which make up the urban fabric in order to predict adequately both future requirements for land and the areas most in need of planned expansion. Consequently, models of this kind have to make allowances for feedback effects of one urban phenomenon on others, and S-E regression is an obvious candidate for fulfilling this role. The EMPIRIC model is one such allocation procedure which has been fitted to a S-E framework (Masser et al, 1971; Putnam, 1975, 1978).

In brief, this model is concerned with forecasting either population or employment in a region and distributing the anticipated growth among the zones comprising the region. A simplified version of it may be confined to predicting population and service employment based on the premises of economic base theory. Thus, service employment and population growth rates are functions of each other because any increase in population automatically calls forth more tertiary and quaternary sector jobs to service it while any upsurge in the labour market is instrumental in attracting unemployed migrants from elsewhere and so boosting population. Nevertheless, economic base theory suggests that the level of service provision of any town is a function of what the export (basic) industries can sustain from the level of their earnings. In other words, service employment expansion is directly dependent on export sectors and population growth indirectly so. The end result is a simple two-equation S-E regression with the two endogenous variables of population growth and growth in service (non-basic) employment being regulated by the predetermined variable of basic employment (and its change variable), along with complementary factors such as population potential and market potential (measures relating absolute distributions of population and purchasing power respectively to spatial accessibility), and values for population and services at

the beginning of the period of interest (time t), *viz:*

$$\Delta Pop = \alpha_1 + \alpha_2\ \Delta Nonbasic + \alpha_3\ Nonbasic_t + \alpha_4\ Pop_t + \alpha_5\ \Delta Basic + \alpha_6\ Basic_t + \alpha_7\ Mktpot_t + u_1 \quad (8)$$

$$\Delta Nonbasic = \beta_1 + \beta_2\ \Delta Pop + \beta_3\ Pop_t + \beta_4\ Nonbasic_t + \beta_5\ \Delta Basic + \beta_6\ Basic_t + \beta_7\ Poppot_t + u_2 \quad (9)$$

Variations on equations (8-9) have been tested for several metropolitan planning regions with reasonable success. Foot (1974), for instance, has conducted S-E regression procedures on approximately the same variables as those outlined in equations (8-9) for the subregion centred on the English city of Reading. Figure 4 displays the variables used by Foot along with their postulated interaction. The direction of the arrows indicates the presumed direction of the causal mechanism and clearly the two endogenous variables are

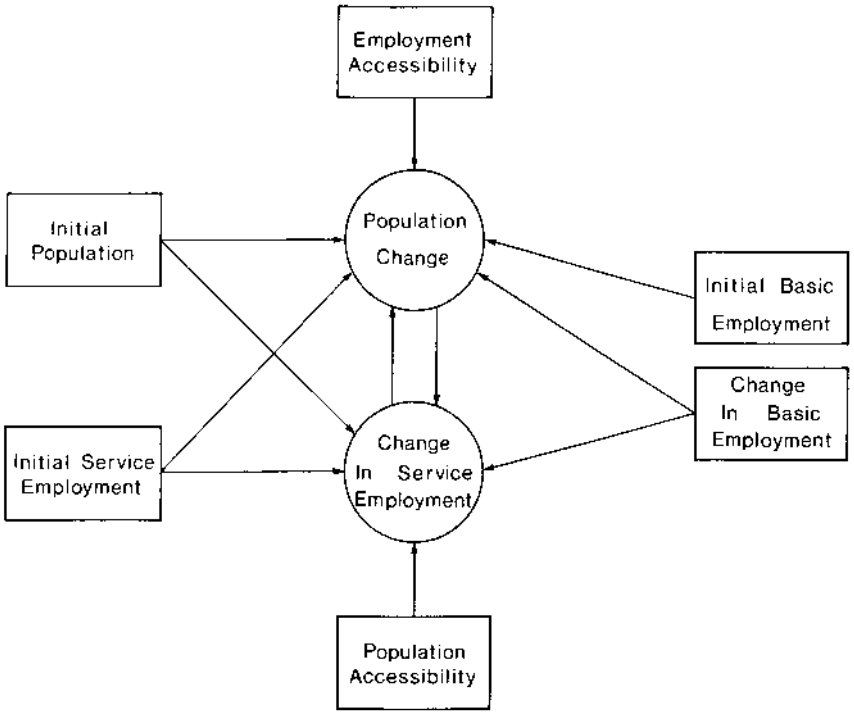

Figure 4. EMPIRIC-style model for Reading subregion (after Foot)

inextricably bound up with each other. Moreover, several of the exogenous variables act as regressors for both equations; a feature which further compounds the interdependence of population change and changes in service employment. Calibration of the model and its subsequent testing against historical

trends suggests that it is much better at estimating changes in service employment than in population. This result seems to imply that the interdependence between the two endogenous variables is not even and that population change is playing a far larger role in steering changes in services than *vice versa;* a situation which, on reflection, makes a lot of sense (because the employment multiplier of services may be more limited than that ensuing from general population accretion, which would also include basic employment).

B. Regional growth theory

Because econometricians have formulated many S-E models for national economies, it is hardly surprising that they have diverted some of that effort to the formulation of models for regional economic growth. The conceptual link between the two types is very strong; indeed, regional models are in the main merely disaggregated versions of their national counterparts. As a result, only small portions of these regional econometric models are 'spatial' in the sense of drawing on established hypotheses of spatial organization and interaction. One of the more explicitly spatial while, at the same time, remaining straightforward to outline is that of Richardson (1973) and it can be paraphrased thus:

$$y = \{ak + (1-a)\Omega\}^{\alpha} + t \quad (10a)$$

$$k = b_1A + b_2y - b_3K - b_4\hat{K} + b_5R \quad (10b)$$

$$\Omega = b_6n + b_7A + b_8P + b_9W \quad (10c)$$

$$t = b_{10}A + b_{11}k + b_{12}M + b_{13}T \quad (10d)$$

The first equation, (10a), is the standard neoclassical growth expression defining the rate of growth of regional income (y) as a function of the combined returns to scale (α) of the growth rates of capital (k) and labour (Ω), after taking into account technical progress (t). In other words, regional income expansion is dependent on scale efficiencies that improve the effectiveness of the two factors of production in the regional economic structure, labour and capital, but it is also incumbent to some extent on technological innovations which may result in whole new industries being established in the region. However, in actuality regional income growth cannot be determined so readily, because the forces which impel scale efficiencies and innovation need to be clarified. In effect, the regressors of equation (10a) cannot be taken as 'givens' and must be regarded as endogenous variables.

Accordingly, equation (10b) isolates the capital growth rate (k) and specifies that it depends directly upon an index of agglomeration economies (A - the benefits accruing from clustered location of economic activities), as well as the regional income growth rate (y) and regional comparative advantage in return on capital (R). Thus, capital expansion is closely tied in with income expansion, not only as an instigator of income growth but also as a benefactor from it. Comparative advantage refers to the region's assets as a source of profits from investments relative to other regions, and the greater the cash-flow from the region in the form of dividends and interest payments, then the greater the surplus available for further capital application. However, while k is a direct function of A, y and R, it is a negative function of the level of regional capital stock (K) and the dispersion of that stock throughout the region's urban centres ($\hat{K}$). The inverse character of these latter two terms can be rationalized as follows: existing capital stock (K) is indicative of investment opportunities that have already

materialised, so new investors are likely to look elsewhere for potential growth activities, and a high degree of existing stock dispersal (K) is not conducive to the discovery of new investment openings.

Equation (10c), meanwhile, asserts that labour expansion (0) is positively affected by the rate of natural increase in population (n), the agglomeration index (A) which presumes that most job expansion would occur wherever industrial activities concentrate, and a locational preference function (P) which would also tend to reflect the large labour markets because of the higher wage levels (W) prevalent in them. Finally, equation (10d) claims that technical progress (t) is regulated by the very fact of industrial concentration (A), increasing capital ready for investment in, amongst other things, Research and Development (k), the stature of the region's metropolis (M), and a measure of the region's propensity to innovate (T). It is apparent from Figure 5 that each of y, k, Ω and t cannot be solved without reference to each other and that a S-E approach is indispensable. However, the Richardson model is also segmentable because, as Figure 5 portrays, equation (10c) can be determined as a first step inasmuch as it does not have reciprocal relations with the prime dependent variable, regional income growth. In con-

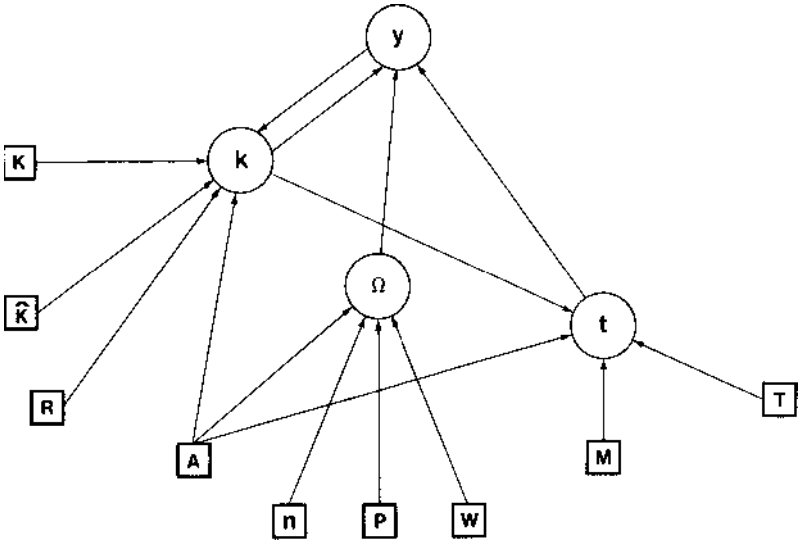

Figure 5. Richardson's regional growth model

sequence, estimates of labour expansion can be substituted into equation (10a) as a second step. However, the reciprocal tie between growth rates of capital and income necessitates the simultaneous solution of equations (10a,b) in the second step. Furthermore, even though technical progress has only a one-way link with regional income growth, the fact that t is partially regulated by k means that equation (10d) has to be omitted from the first step solution and perforce is added to the second step along with equations (10a,b). Despite its logic, the very comprehensibility of the model has acted against it

being made operational. Difficulties in establishing identity equations for such elusive spatial concepts as the index of agglomeration have kept it on the drawing boards.

C. Modelling geomorphological pattern

Geomorphologists have developed a variant of multiple regression into a fine art, namely trend surface analysis. Trend surface analysis not only provides a spatial pattern of underlying trends in physical phenomena but is also capable of decomposing the trend into universal or regional effects as opposed to purely local effects. Unfortunately, the distinction between the two trend constituents is not absolute, and a degree of overlap between regional and local determinants of trend may be expected, *a priori*. One way of accommodating this interdependence has been proposed by Agterberg and Cabilio (1969) and it calls for the application of S-E regression. An outline of the research problem of these authors will illustrate the tie between trend surface analysis and S-E regression.

In essence, they were concerned to explain the wide distribution of gold-bearing quartz veins across part of the Canadian Shield in terms of the six different rock types which comprised the country rock. After setting-up gold occurrences as the dependent variable, Agterberg and Cabilio performed a trend surface analysis which indicated both a strong regional trend in gold occurrence, along with local concentrations of occurrences as well. Therefore, gold occurrence was not only widely distributed across the region because of the ubiquitous existence of appropriate country rock, but its probability of existence was enhanced in certain localities through special factors operating in these restricted areas. It follows that gold occurrence is incumbent on region-wide lithographic factors and locally-operating lithographic factors in combination. But how can one separate the two in a regression format when they are obviously interdependent (Figure 6)? The authors did this by concocting an intermediate variable which was the outcome of regressing the local lithographic effect on the regional effect. This intermediate variable therefore has the property of redefining that part of the regional trend which stems directly from local features, which is to say, the ubiquitous element in the regional effect is removed. It is possible to establish a S-E framework whereby two equations with gold occurrence as dependent variable can be solved. One of the equations embodies the original regional effect and the local effect as regressors, whereas the other has the same local effect regressor but a 'purged' regional effect (the intermediate variable) as a substitute for the regional variable drawn from trend surface analysis. Examination of the differences in 'explanation' (% sum-of-squares reduction) between the two equations suggests that a regression based solely on trend surface results overstates the probability of gold occurrence. Thus, the removal of the interdependence between regional and local factors through a S-E regression framework provides an important corollary to the interpretation of trend surface analysis in geomorphology.

D. Migration analysis

Spatial analysis abounds with examples of multiple regression studies of interregional migration where the objectives are to determine both the factors conducive to out-migration and those responsible for attracting migrants to their eventual destinations. However, many of these studies have failed to conform to an a *priori* rationale concerning the push-pull mechanisms of.

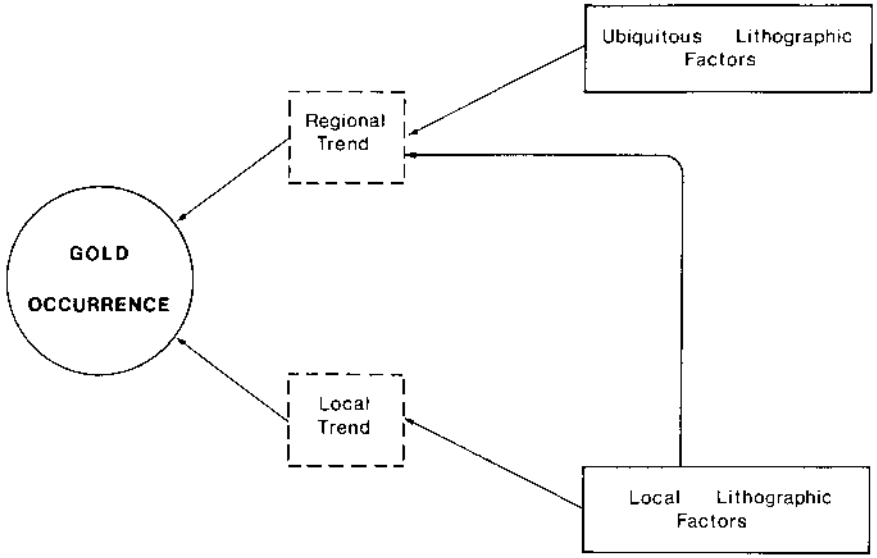

Figure 6. Relationships between the occurrence of gold-bearing rocks and their determinants

migration and some of their puzzling results are, no doubt, due to inconsistent parameter estimates stemming from classical least squares regression. In the words of Shaw (1975, p 96);

> An important question when estimating parameters via a single-equation multiple regression model is whether behavior of the dependent variable over time has influenced the independent variables of the equation. This question is particularly relevant in migration analysis as many studies relate measures of cumulative migration (e.g. $\Sigma M_{i \to j}$ from 1955-1960), to measures of independent or explanatory variables for the same period of time. In this case, parameter estimates are likely to possess a simultaneous equation bias.

Shaw is hereby alluding to the interdependence between the regressand (migration) and the regressors (causes of migration) which is technically not permissable in the standard multiple regression model. A true model of cause-and-effect must allow for the feedback effect of the dependent variable on the so-called independent variables. In order to accomplish this end, a multiple equation format is required as well as a procedure for handling the simultaneity aspect of causality and this, of course, is tantamount to requesting a S-E regression. A simple example will suffice to illustrate the case in point. Let $I_{j \to i}$ represent in-migration from a set of j places into city i for a ten year period, 1968-78. And, let ΔP_i represent population growth of city i for the same period. Now, it can be argued that the stream of in-migration into the city will be a function of the number of opportunities

available in that community and the steady increase in these openings. Glickman and McCone (1977) have found that city population 'is significantly related to many variables which might be expected to enhance employment growth' and so we can assume that APi acts as a surrogate for pull-factors inducing migrants into the city over the given time period. Therefore, an initial hypothesis can be expressed as:

$$I_{j \to i} = \alpha_o + \alpha_1 \Delta P_i + u_1 \qquad (11a)$$

But, a single-equation solution would omit the impact of the migrant stream on population growth, and instead maintain only the one-way causality that the attraction of a large and growing population encourages in-migration. Obviously the concurrency of the growth of population and on-going arrival of migrants cannot be treated as independent phenomenon. So equation (11a) has to be complemented by a reciprocal equation which is expressive of the two-way causality of population growth and migration:

$$\Delta P_i = b_o + b_1 I_{j \to i} + u_2 \qquad (11b)$$

The two equations together provide a simple S-E regression model and at a stroke remove the simultaneity bias which plagues the single equation standard multiple regression format. Several studies making use of this approach are quoted in the bibliography, and Chapters III and IV steadily expand the basic migration model to account for a multiplicity of push-and-pull factors with the intention of providing a demonstration of the successive stages involved in building up a S-E model as well as indicating the regression analyses required for calibration.

II. THE MATHEMATICAL BASIS OF S-E REGRESSION

(i) The Identification Problem

As S-E regression is primarily a device for model-building, it entails the establishment of several equations encompassing a wide spectrum of phenomena which are believed to influence, in varying degrees, the target (dependent) variables of the model. The very existence of several equations gives rise to two technical problems not directly encountered in the standard multiple regression framework usually applied in geography. These are the problems of identification and consistency; the former belongs to the initial, conceptualizing stage of model-building, whereas the latter is concerned with the technical specifications of the resultant equations.

The problem of *identification* is said to occur in those situations where any set of observed facts can be explained in a number of ways, that is, several hypotheses can be coined to account for a given set of relationships. The problematic aspect emerges when one tries to disentangle the single, 'best' hypothesis from the multiplicity of possible explanations. Thus, the object of the exercise is to determine, for each equation in a model, a set of parameters which is compatible with the preferred hypothesis for that equation. In pragmatic terms, the problem of identification summarizes to a question of whether the observations will enable the analyst to measure each and every equation in the simultaneous system. Any structural equation is.

said to be *identifiable* only if all its parameters are identifiable.

As an illustration of how the identification problem may arise, let us suppose we have an equation system of three equations concerned with determining the endogenous variables, population change in place i (ΔP_i), immigration to i from j (I_{ij}) and emigration to j from i (E_{ji}). If we concern ourselves with the first of these equations, there are three contingencies which could arise to prevent identification of the parameters in the ΔP_i equation, *viz:*

(1) the I_{ij} equation contains no variables that do not occur in the ΔP_i equation,

(2) the E_{ji} equation contains no variables that do not occur in the ΔP_i equation,

(3) there exists a linear combination of the I_{ij} and E_{ji} equations which contains no variables not occurring in the ΔP_i equation.

The first case would arise if we conceived our three-equation model as follows:

$$\begin{aligned} \Delta P_i &= f(\,(+)\, I_{ij};\, (-)\, E_{ji}) \\ I_{ij} &= f(\,(+)\, \Delta P_i;\, (-)\, E_{ji}) \\ E_{ji} &= f(\,(-)\, \Delta P_i;\, (-)\, I_{ij}) \end{aligned} \qquad (12)$$

which is to say that population change in i is a direct function of immigration into i and an inverse function of emigration from is immigration into i is positively affected by the increasing size and attractiveness of i (ΔP_i) while being deterred by reduced opportunities there as reflected by out-migration; whereas out-migration from i would occur in response to a stagnating population and would behave in a reciprocal manner with respect to in-migration. It is evident from this example that no exogenous variables are present in any of the equations and therefore the first contingency cannot be avoided. As a result, none of the parameters is estimable.

The second case would arise if equations (12) were modified by introducing an exogenous variable into the I_{ij} equation. For instance, it could be argued that immigration into i is not only a function of the dynamism of i as reflected in population growth, but it also responds positively to higher wage levels (W_i) taking effect in i's expanding job market. In consequence, the I_{ij} equation would change to:

$$I_{ij} = f(\,(+)\, \Delta P_i;\, (-)\, E_{ji};\, (+)\, W_i) \qquad (12a)$$

leaving the other two equations intact. However, this newly introduced exogenous variable will not distort the conformity of the ΔP_i and E_{ji} equations and so the second cause of nonestimability will be realized. Finally, the third case would transpire if we extended the relevance of the exogenous variable to the ΔP_i equation as well as the I_{ij} one. This is quite valid if we adopt the argument that, in promoting affluence, rising wages also facilitate extended family size and hence, greater population growth. In other words, population growth is an inverse function of emigration and a direct function of both rising affluence (W_i) and immigration. Thus, the equation system

benefits from no extra exogenous variables and the ΔP_i equation has no fewer variables than the composite of the other two, that is, the third contingency is broached.

It is clear from this simple example that a certain structural equation, in this case population change, is not identifiable whenever we are able to construct other equations (I_{ij} or E_{ji}) which contain the same explanatory variables as some or all other structural equations in the model. In essence, the structural equation is only identifiable when it can be shown that it is impossible to produce a different equation of the same prescribed form by linear combination of all equations in the model. Fortunately, two procedures are available to meet this requirement, namely the *order* and *rank* criteria for identifiability.

The order condition is manifest in two versions depending on the restrictions being applied to the model. These can either be *exclusion* restrictions or *linear homogeneous* restrictions. The former are the more common and assume that certain variables present in the model are absent from some of the equations (hence, the alternative name of 'zero restrictions'). In situations of this kind, the order condition states that for identifiability of a structural equation from within the model, the number of variables excluded from that equation must be at least equal to one less than the number of equations in the model. If we revert to the terminology of Chapter I(iv) where G represents the number of equations in the linear system, then any one equation must exclude at least G-1 of the variables appearing in the model to ensure identification. On these grounds alone, none of equations (12) is identifiable.

Linear homogeneous restrictions, as to be expected, come to bear on homogeneous equations. Simply put, if the system AX = B consisting of a matrix of coefficients (A), a vector of unknowns (X), and a vector of right-hand sides (B) is characterized by a null B, then the system is one of homogeneous equations. Equation (6) outlined in Chapter I(iv) is such a case. Homogeneous equations can only be solved (in other than a non-trivial sense) when some restrictions are placed on the unknowns (i.e. some of the unknowns, $X \neq 0$). These restrictions may state, for example, that a coefficient in one equation is equal to a coefficient in another ('linear dependence') or that the covariance between two structural disturbances is taken as zero (as is the case with the simple supply-demand model displayed in the second equation). Assigning restrictions to the parameters (including disturbances) of X to produce a relationship among the elements of that vector enables solution of the homogeneous equations by simultaneous means. This requisite of linear dependence is moulded to the order condition to state that the number of linear homogeneous restrictions (R) must not be less than the number of equations in the model less one:

$$R \geq G - 1$$

Identification based on linear homogeneous restrictions is plausible only when sufficient knowledge is known about the model's mechanisms to make *a priori* assumptions about the behaviour of the unknowns. Because this is usually a difficult process, exclusion restrictions are the more usual avenue for fulfilling the order condition.

The order criterion is a necessary condition for identifiability, but the rank criterion is a necessary and sufficient condition for the identification of a structural equation. The rank of a matrix is defined as the order of the largest non-zero determinant that can be obtained from the elements of the matrix (see Appendix 1 for amplification of these technical terms). It follows that the rank condition for identifiability claims that at least one non-zero determinant of order G - 1 must be capable of formation from the parameters of those variables excluded from the equation in question (though present in other equations of the model). In summing up, it is important to stress that if the rank condition is satisfied, then so is the order condition, but not *vice versa*.

Although it is the more rigorous of the two identification requirements, the rank condition presupposes technical knowledge of the determinantal structure which simply may not exist in the model-building stage. As a result, the order condition is usually applied to classify structural equations and the probability is very high that its implications will be fully endorsed by a subsequent test for the rank condition (Christ, 1966, p 322). In practice, therefore, the order criterion is usually applied as a surrogate for the sufficient condition as well as being, in its own right, the necessary condition. Accordingly, any equation that excludes less than G - 1 variables is *under-identified* in that its parameters are not estimable; any equation excluding exactly G - 1 variables is *just-identified;* and an equation deleting more than G - 1 variables is *over-identified*. Just-identified equations have an equality between the number of parameters to be estimated and the number of predetermined variables, whereas over-identified equations have more predetermined variables than parameters to be estimated. In short, whereas both of the latter two categories are estimable using S-E regression, over-identified equations can provide estimates which are not unique.

These notions of identification are better appreciated when they are applied to actual models. To this end, we have taken liberties with the elementary three-equation demographic model outlined in equations (12). As mentioned before, none of the three structural equations are identifiable in their present configuration. In terms of the order classification, they are under-identified. This becomes evident if we apply the order condition based on the exclusion restriction to the ΔP_i equation. In order to fulfil that condition, at least G - 1 variables present in the model must be omitted from the ΔP_i equation, that is, two variables present elsewhere must be absent from this particular equation. Obviously, this is not the case because the other equations mirror the variables found in that for population growth. Equations (12) need to be considerably augmented before the population growth equation becomes just-identified. One possible extension which would bring about this outcome is phrased as follows:

$$\begin{aligned} \Delta P_i &= f(\ (+)\ I_{ij};\ (-)\ E_{ji}) \\ I_{ij} &= f(\ (+)\ \Delta P_i;\ (-)\ E_{ji};\ (+)\ W_i) \\ E_{ji} &= f(\ (-)\ \Delta P_i;\ (-)\ I_{ij};\ (+)\ W_j) \end{aligned} \qquad (13)$$

which is a modification of the stage construed in equation (12a). The difference is that a new variable, W_j, wage rate in place j, has been introduced into the emigration equation so as to monitor those people leaving place i for higher wages in place j. Now, the number of variables excluded from the ΔP_i structural equation are two, W_i and W_j, which conforms to the just-identified

requirement. It is easy enough to extend the variable set by allowing for a motley of extraneous factors that influence the endogenous variables. A variable to account for religious persecution, RP_i, could be added to the emigration equation, for example, while regressors representative of choice social facilities, F_i, and political tolerance, PT_i, could be appended to the immigration equation. A possible model can be expressed as:

$$\begin{aligned} \Delta P_i &= f(\ (+)\ I_{ij};\ (-)\ E_{ji};\ (+)\ W_i) \\ I_{ij} &= f(\ (+)\ \Delta P_i;\ (-)\ E_{ji};\ (+)\ W_i;\ (+)\ F_i;\ (+)\ PT_i) \\ E_{ji} &= f(\ (-)\ \Delta P_i;\ (-)\ I_{ij};\ (+)\ W_j;\ (+)\ RP_i) \end{aligned} \qquad (14)$$

(note that W_i has been reintroduced into the ΔP_i equation for greater realism). Evidently, there are now four variables in the model excluded from the ΔP_i equation, namely, F_i, PT_i, W_j, and RP_i, so now that equation is over-identified. It would be possible to go ahead and estimate equations (13-14) by S-E methods because the ΔP_i equation of the former has a balance between the structural parameters to be estimated and the omitted predetermined variables, whereas the population growth equation of the latter has more excluded predetermined variables than there are parameters to be estimated within that particular equation. Equation (12), of course, is unestimable under the conventional stipulations of exclusion restrictions. It is essential that the identification aspect of modelling is fulfilled before the analyst can progress to the stage of estimating the parameters of structural equations.

(ii) <u>The Consistency Problem</u>

As is well known, the classical least squares regression (ordinary least squares or OLS) model is predicated on certain key assumptions concerning the regressors and the disturbance term. These can be summarized for any particular regressand, y, as follows:

$$y = BX + U \qquad (15)$$

with B being a vector of regression coefficients of K size and

(1) X is a (N x K) random matrix of sample observations on the independent variables (fixed numbers such as constant terms and dummy variables are also admissable);

(2) U is a random disturbance vector of N elements, each element being independent of the sample observations in X;

(3) the distribution of U is normal in accordance with the central limit theorem and has the properties of

(a) zero mean $E(u) = 0$

(b) constant variance $E(u^2) = \sigma^2$

(c) zero covariance $E(u_M\ u_N) = 0$

The assumptions about the disturbance term are crucial to understanding the difference between OLS and S-E estimation. The zero mean assumption signifies the absence of any systematic component in the random disturbance affecting the regressand, whereas the zero covariance assumption implies that the elements of the disturbance term are uncorrelated. Thus, the regressors are

not only independent of each other but are independent from the disturbance term too. Moreover, the elements of the disturbance term are independently (though identically) distributed as well. In fact, these assumptions of independence do not bear up in models utilizing endogenous as well as predetermined variables. By definition, endogenous variables are not independent entities within the linear system of simultaneous equations and so conflict with assumption (1) expressed above. However, the disturbance terms for the system of equations are not independent either and this can be demonstrated by reverting to a simple migration model.

Let this model be an expansion of the very basic example given in Chapter I(v)D. This is a two-equation model of population growth and in-migration where the first regressand is expanded from being merely a function of in-migration to also being a function of the exogenous variable of natural increase (NI). Nevertheless, the two equations still maintain their fundamental pattern of two-way causality, *viz:*

$$I_{j\to i} = a_0 + a_1 \Delta P_i + u_1 \tag{16a}$$

$$\Delta P_i = b_0 + b_1 I_{j\to i} + b_2 NI + u_2 \tag{16b}$$

If we focus on the in-migration equation, it is feasible to show how the determination of this equation cannot be divorced from the disturbance term of the population growth equation. The first step is to capture all the factors influencing in-migration and this is facilitated by substituting equation (16b) into (16a):

$$\begin{aligned} I_{j\to i} &= a_0 + a_1 (b_0 + b_1 I_{j\to i} + b_2 NI) + u_1 \\ &= a_0 + a_1 b_0 + a_1 b_1 I_{j\to i} + a_1 b_2 NI + a_1 u_2 + u_1 \end{aligned} \tag{17}$$

This equation indicates that the regressand $I_{j\to i}$ is being influenced by its own contribution to population growth. Thus, it becomes necessary to remove the effects of $I_{j\to i}$ from the contributions of the regressors in order to make $I_{j\to i}$, as regressand, independent from its own feedback. The procedure by which this is done is known as taking the *reduced form* of the equation. In short, the reduced form tells us how the endogenous variables are determined by the exogenous variables alone. Applying the reduced form to the in-migration equation gives us:

$$I_{j\to i} (1 - a_1 b_1) = a_0 + a_1 b_0 + a_1 b_2 NI + a_1 u_2 + u_1 \tag{18}$$

$$I_{j\to i} = \frac{a_0 + a_1 b_0}{1 - a_1 b_1} + \frac{a_1 b_2}{1 - a_1 b_1} \cdot NI + \frac{a_1 u_2 + u_1}{1 - a_1 b_1}$$

Yet although equation (18) removes the direct effects of $I_{j\to i}$ from the set of regressors, it does not remove all of the effects of in-migration from the right hand side of the equation because $I_{j\to i}$ is not independent of u_2 in equation (16b). The fact that $I_{j\to i}$ and the disturbance term, u_2, are correlated becomes apparent when we take the expectation of equation (18) so as to establish the difference between the reduced form and the expectation of the reduced form, *viz:*

$$E(I_{j\to i}) = \frac{a_0 + a_1 b_0}{1 - a_1 b_1} + \frac{a_1 b_2}{1 - a_1 b_1} \cdot NI \tag{19a}$$

$$\left[I_{j\to i} - E(I_{j\to i})\right] = \frac{a_1 u_2 + u_1}{1 - a_1 b_1}$$

and then isolate u_2 in terms of its expectations:

$$u_2 \left[I_{j\to i} - E(I_{j\to i})\right] = \frac{u_2 (a_1 u_2 + u_1)}{1 - a_1 b_1}$$

$$\begin{aligned} E\left\{u_2 \left[I_{j\to i} - E(I_{j\to i})\right]\right\} &= E \frac{\{a_1 u_2^2 + u_1 u_2\}}{1 - a_1 b_1} \\ &= \frac{a_1 Eu_2^2}{1 - a_1 b_1} + \frac{E(u_1 u_2)}{1 - a_1 b_1} \end{aligned} \tag{19b}$$

$$= \frac{a_1 Eu_2^2}{1 - a_1 b_1} \neq 0 \tag{19c}$$

In equation (19a) the disturbance term is eradicated because $E(u_1) = 0$ and $E(u_2) = 0$. However, an error factor is reincorporated when the difference between actual and expected equations is formulated. That portion of the error factor attributable to the population growth equation (i.e. disturbance u_2) is seen to be inextricably associated with the $I_{j\to i}$ term. The fact that the interaction between $I_{j\to i}$ and u_2 is not zero, as reflected in equation (19c), is indicative of the correlation between them. (Incidentally, note that the second expression in equation (19b) is eliminated because of the zero covariance assumption of disturbance terms.)

The principle of interdependence between the regressand and the disturbance term of other equations in the model applies throughout simultaneous-equation systems. Therefore, application of OLS estimates to these models would be inopportune because they would provide inconsistent results; in effect, no meaningful solution at all. The way to surmount this problem and ensure *consistent* results (i.e. a valid solution) is to utilize reduced form equations whereby the original structural equations are transformed to express the endogenous variables as functions of exogenous variables and disturbances only. Simultaneous-equation regression procedures designed to make use of the reduced form will be assessed after the problem of statistical inference of these interdependent models is mentioned.

(iii) A Question of Statistical Significance

In instances where the observations utilized in the variables of a model are representative of a large sample size then the parameters of the model accede to *asymptotic* properties. In essence, if as a sample size tends towards infinity its density tends to converge towards a given function, then that function is the asymptotic or limiting distribution of the distribution from which the sample is drawn. Should the sample have a limiting distribution which conforms to the probability limits of the normal distribution then the sample and its parameter estimates are said to be asymptotically normally distributed. As a result, the asymptotic expectation (mean) of any variably will be zero and its asymptotic variance will approach zero. These are the

requirements that guarantee consistency and, as we have seen, the condition of consistent parameter estimation is essential for the solution of S-E models. Indeed, consistency is an outcome of large-sample sizes and so it fits readily into the concern with asymptotic properties. It follows that S-E equations can make use of the asymptotic property of disturbance terms distributed approximately normally (depending on the size of sample) with zero mean and variance σ^2. Consequently, S-E parameters can be tested for statistical significance in the same manner as OLS regression coefficients, which are also expected to reflect normally distributed disturbance terms, that is, by invoking the t-test. However, a cautionary note must be sounded. This test requires that the disturbances of the variables be distributed with zero mean and constant variance. In view of the fact that the asymptotic property precludes complete normality, the t-test can be used as only a rough indicator of parameter significance in S-E regression (Christ, 1966, p 598). Similar reservations hold for the other conventional significance indices, namely the F-ratio which is used to test the significance of a whole equation and the R^2 coefficient which is the standard indicator of variance 'explanation' contributed by the equation (see Ferguson, 1977, for computation details).

Simultaneous equation models of a dynamic nature are special cases which require different statistical tests for hypothesis verification (Glickman, 1977, p 67). Models of this kind include the EMPIRIC type outlined in Chapter 1. In cases such as these, the problem becomes one of testing the validity of a range of predicted values for the key variables of population change and employment change. This involves application of one or more of three 'goodness-of-fit' tests: the mean absolute percent error (equation 20a), the root mean square error (equation 20b), and Theil's U coefficient (equation 20c) viz:

$$\text{Mape} = \frac{\sum_{j=1}^{N} \left| y_{ij}^{P} - y_{ij}^{A} \right| / y_{ij}^{A}}{N} \cdot 100 \qquad (20a)$$

$$\text{Rmse} = \left[\frac{\sum_{j=1}^{N} (y_{ij}^{P} - y_{ij}^{A})^2}{N} \right]^{\frac{1}{2}} \qquad (20b)$$

$$U = \left[\frac{\sum_{j=1}^{N} (\Delta y_{ij}^{P} - \Delta y_{ij}^{A})^2}{\sum_{j=1}^{N} (\Delta y_{ij}^{A})^2} \right]^{\frac{1}{2}} \qquad (20c)$$

where y_{ij}^{P} is the predicted value of the i^{th} endogenous variable for all observations ($j = 1 \ldots N$), y_{ij}^{A} is the actual value of this same variable during the sample period, and the delta prefix represents the change values over the period of analysis. Testing is undertaken by compiling these statistics and then comparing them across models to see which variant gives the best fit. At best, this is an imprecise procedure and so rigorous limits for accepting model validity are not imposed; for example, econometricians tend to be satisfied if their endogenous variables record mean absolute percent errors over a sample period of less than 3%.

In summing up this section, the contrast between the applicability of statistical testing for S-E regression and for standard multiple regression is quite marked. Not to put too fine a point on it, the analyst is much less comfortable using the battery of hypothesis tests in S-E regression than he is in OLS and for sound statistical reasons. The researcher must ensure a substantial sample size in order to be reasonably certain that S-E estimation is based on asymptotic normality in the underlying data, and even then, the significance tests must be interpreted with caution.

(iv) Two-Stage Least Squares

If an S-E system satisfies the conditions of identification, then the reduced form of its structural equations can be used as the basis for the derivation of consistent regression estimates. One procedure for accomplishing this end is Two-Stage Least Squares (TSLS) regression analysis. TSLS is the most straightforward means for obtaining consistent parameters in simultaneous equations and this is due to its property of estimating the equations of the model one at a time. Furthermore, some evidence exists (Namboodiri, *et al*, 1975, p 517) to suggest that TSLS is at least as good as, if not superior to, the more complicated S-E estimation procedures of three-stage least squares and limited-information maximum likelihood (the first is introduced in Chapter IV while the second is briefly outlined in Appendix 2). For these reasons, TSLS is highlighted as the most appropriate estimating procedure for the inauguration of the geographer in the ways of S-E regression.

The operations involved in TSLS can be summarized in the following manner.

(1) Select one endogenous variable from those in the model to act as the target dependent variable.

(2) (First stage of TSLS) Determine the OLS estimates of the reduced form equations for the remaining endogenous variables in the equation using all the predetermined variables in the model (i.e. in all equations, not just the one of interest).

(3) Replace the observed values for these remaining endogenous variables by their estimated values from step (2).

(4) (Second stage of TSLS) Perform OLS regression of the target dependent variable on a set of variables made up of the reduced form parameter estimates for the remaining endogenous variables (step 3) and the original, observed values for those predetermined variables present in the equation of interest.

(5) Repeat steps 1-4 for each target dependent variable in the model.

In mathematical terms, the first step involves the selection of one equation out of the linear system. Let us suppose that y_1 is the target regressand and, as determined in Chapter I(iv), is drawn from:

$$BY + \Gamma Z + U = 0 \qquad (21)$$

to give:

$$y_1 = \overset{**}{B}\overset{*}{Y} + \overset{**}{\Gamma} Z + u_1 \qquad (22)$$

where y_1 is a vector of observations on a target dependent variable,
$\overset{*}{Y}$ is a matrix of observations for the other endogenous variables appearing in the target equation,

$\overset{*}{B}$ is a vector of regression parameters for $\overset{*}{Y}$,

$\overset{*}{Z}$ is a matrix of observations for those predetermined variables present in the target equation,

$\overset{*}{\Gamma}$ is a vector of regression parameters for the $\overset{*}{Z}$,

and u_1 is a vector of disturbances.

After this step is accomplished, the second step (and, formally, the *first stage* of TSLS) is the estimation of the other endogenous variables acting as regressors in the target equation (i.e. the $\overset{*}{Y}$). This is accomplished by taking each of the $\overset{*}{Y}$, in turn, and regressing it against all of the predetermined variables in the equation system, and not just those contained within the target equation (i.e. the right hand side for each $\overset{*}{Y}$ equation is made up of the Z vector as opposed to being confined to $\overset{*}{Z}$). In other words, OLS is performed on the reduced form of all the equations in the system with the exception of the target one, viz:

$$\hat{\overset{*}{Y}} = \pi Z + V \qquad (23)$$

where $\hat{\overset{*}{Y}}$ are the OLS estimates of the set of endogenous variables present as regressors in y_1, π is a matrix of reduced form parameters corresponding to the $\overset{*}{Y}$, and V is the matrix of reduced form disturbances. Replacement of $\overset{*}{Y}$ by the $\hat{\overset{*}{Y}}$ estimates satisfies the third step and sets the scene for the *second stage* of computation (step 4). Fulfillment of this second stage is incumbent on the regression of y_1 against the reduced form estimates of $\overset{*}{Y}$ and the observed values for the predetermined variables in the target equation, $\overset{*}{Z}$, *viz*:

$$y_1 = \hat{\overset{**}{B}}\hat{\overset{**}{Y}} + \hat{\overset{**}{\Gamma}}\overset{**}{Z} + \bar{u}_1 \qquad (24)$$

with $\bar{u}_1$ being a disturbance vector modified by the addition of an error component obtained in producing the $\hat{\overset{*}{Y}}$ estimates and $\hat{\overset{*}{B}}$ and $\hat{\overset{*}{\Gamma}}$ are the second stage regression coefficients. Equation (24) ensures that the target regressand, y_1, is approximately purged of its correlation with the disturbance term. TSLS is repeated on all the other endogenous variables in the model (y_2, y_3 ... y_G) until all regressands are approximately purged of association with their disturbance terms (u_1, u_2 ... u_G).

An embellishment of the procedure can be achieved by adapting the expanded migration model (equations 16a-b) to the five steps. To this end, equations (16a) and (16b) can be rewritten respectively as:

$$y_i = \beta_1 Y_p + u_i \qquad (25a)$$

$$y_p = \beta_2 Y_i + \gamma_1 Z_n + u_p \qquad (25b)$$

with the subscripts referring to the previously identified variables. Taking (25a) as our point of departure, we compute the first-stage by regressing Y_p on Z_n:

$$\hat{Y}_p = \pi_1 Z_n + v_p \qquad (26)$$

and substitute the reduced form estimate, $\hat{Y}_p$ from equation (26) into the Y_p vector of the original equation. The second stage is brought to fruition when

equation (27) is solved:

$$y_i = \hat{B}_1 \hat{Y}_p + \bar{u}_i \qquad (27)$$

Attention is diverted to the population growth equation in order to effect the fifth stage because this is the only other endogenous variable in the model. Briefly, a repetition of the TSLS procedure can be accomplished in the following formulations:

$$\begin{aligned} y_p &= \beta_2 Y_i + \gamma_1 Z_n + u_p \\ \hat{Y}_i &= \pi_1 Z_n + v_i \qquad (28) \\ y_p &= \hat{\beta}_2 \hat{Y}_i + \hat{\gamma}_i Z_n + \bar{u}_p \end{aligned}$$

The next two chapters provide worked examples of elaborated versions of such a migration model. However, the extra variables required to accommodate the variability of the real world complicates the algebra of TSLS estimation and so this simple example here is presented merely to enable the reader to 'cut his teeth' on the organization of S-E regression and is not intended to be taken seriously as a realistic model of urban in-migration (for example, the under-identification of the population growth equation would prevent estimation).

In sum, TSLS is an equation-by-equation procedure for handling S-E regression which has the advantage of producing consistent parameter estimates for all equations that are just-identified, or, as is the more usual case, over-identified. The property of consistency implies that the TSLS regression coefficients converge in probability to the true parameter values as the sample size of the observation approaches infinity.

III. A TWO-EQUATION MIGRATION MODEL EXAMPLE

(i) Conceptual Basis

The intent of this chapter is to illustrate the compilation of a simple two-equation migration model and go on to display the procedures undertaken in applying TSLS regression to its operationalization. The chapter following this one will review a complex, several equation system developed to examine processes of inter-urban migration in the United States. Therefore, a sequential element is attempted whereby the calibration procedures are detailed in this chapter and the modelling aspects of S-E systems emphasized in the one after. In so doing, the intricacies involved in organizing mathematical calibration are covered in the less complex model, thus avoiding unnecessary obfuscation in the fourth chapter where conceptual interdependence is stressed.

As intimated above, a migration model is hereby established consisting of two endogenous variables and hence equations, along with seven predetermined variables. This model is formulated as follows:

$$y_1 = \beta_{11}y_2 + \gamma_{11}z_1 + \gamma_{12}z_2 + \gamma_{13}z_3 + \gamma_{14}z_4 + \gamma_{15}z_5 + u_1 \quad (29a)$$

$$y_2 = \beta_{21}y_1 + \gamma_{21}z_1 + \gamma_{22}z_2 + \gamma_{23}z_6 + \gamma_{24}z_7 + u_2 \quad (29b)$$

and the terms are identified in Table 1. Notice that two of the predetermined variables are common to both equations and that the two endogenous variables are present in each other's equations. This latter factor effectively precludes segmentability (Chapter I (iii)) and earmarks the two equations for S-E consideration. Verbally, we can state that equation (29a) claims net out-migration to be some function of the level of average family income as well as the effects of four exogenous variables and previous migration rates (i.e. a lagged endogenous variable). In turn, equation (29b) asserts that average family income is bound up with the level of out-migration and is otherwise affected by four predetermined variables. The rationale for this interaction between migration and income derives from the neo-classical argument of regional growth theory, which maintains that people move in response to differentials in wages. It follows then that places with low income will suffer a net migration loss whereas those with high incomes will act as magnets to migrants and so will realize population gains. Accordingly, income levels (represented above by average family income) are key instigators of migration movements. But migrants will continue to flow into high wage areas until those places find themselves with a surfeit of labour. Under pure market conditions, employers will respond to the glut of potential workers by cutting wages and substituting labour for capital in the production process. At the same time, the areas losing population will gradually find themselves with a shortage of labour and wage rates will start to rise if only to persuade would-be migrants to stay at home. Eventually, wages will equalize across all places and migrants will cease to move in order to gain financially (i.e. assuming that they do not change occupations, that frictional and structural unemployment are insignificant, and that unions do not intrude to regulate wage rates). The predetermined variables merely complement that rationale: the lagged migration factor is attempting to monitor the momentum generated by prior migration streams while the distance-to-urban centre variable is a crude indicator of the deterrent effect on spatial separation from migrant origins to potential destinations. Both are believed to influence the causal mechanism linking income and migration. However, certain exogenous variables influence only part of the causal mechanism, that is to say, population change is expected to influence migration directly (but only indirectly affect income) as, indeed, is the pressure of a high man-land ratio of rural occupance. On the other hand, the value of farm products sold is an obvious contributor to average family income in rural districts. The two shift-share variables (see Table 1 for definition) are divided between equations with the differential shift being assigned to the migration equation (with the expectation that local comparative advantage, or the lack of it, will either contain or boost respectively the 'push' factor in migration), and the proportionality shift being allocated to the income equation (under the premise that areas experiencing above-average growth rates in employment will have spin-offs in terms of higher wages). Figure 7 provides a diagrammatic view of the variable relationships.

(ii) Calibration

The model is operationalized for the 114 local government areas which cover the 'prairie' component of the province of Manitoba. That region, along with other primarily agricultural parts of North America, has experienced a

TABLE 1: VARIABLE IDENTIFICATION FOR WORKED MIGRATION MODEL

y's	z's	u's
1. net out-migration, 1966-71	1. lagged net out-migration, 1961-66	1. disturbance term for y1
2. average family income, 1971	2. distance to nearest urban centre	2. disturbance term for y2
	*3. employment differential shift, 1961-71	
	4. population change, 1961-71	
	5. man-land ratio, 1971	
	6. average value of farm products sold, 1971	
	*7. employment proportionality shift, 1961-71	

* Shift-share analysis is concerned with differentiating the local rate of employment growth from the overall rate. As such, it assumes that, *ceteris paribus,* localities should perform at the same rate as Manitoba as a whole. Any difference between the two rates is attributable either to the industry mix of the locality (the proportionality shift) or locational factors peculiar to the locality (the differential shift). The formulae used in concocting z_3 and z_2 are taken from Richardson (1969).

steady improvement in farm productivity at the expense of the labour force. Rural dwellers have quit the prairies in response to the diminishing job opportunities. Moreover, the fairly rapid growth of urban economies with their added attraction of generally higher wage levels has hastened the rural depopulation trend in many cases. Given this particular context, we can postulate the direction of signs for the regressors as they effect the two regressands in equations (29a,b).

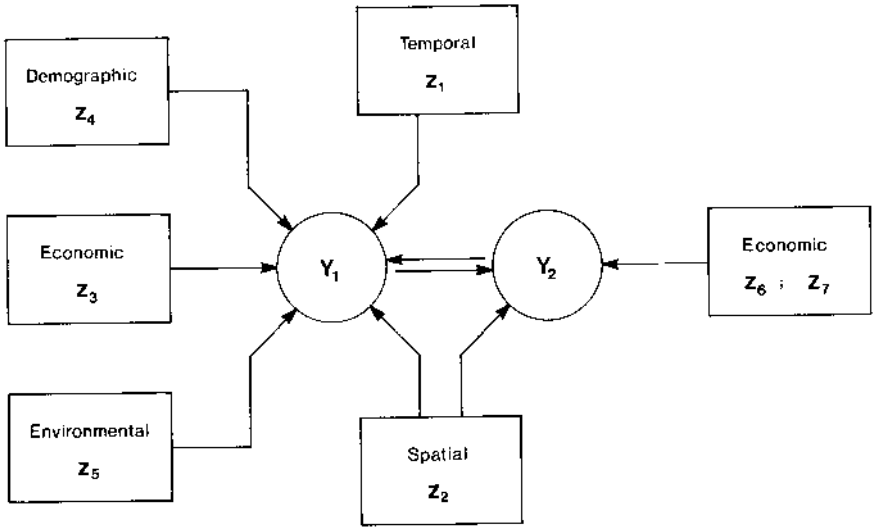

Figure 7. Variables impinging on net out-migration and average family income

TABLE 2: POSTULATED RELATIONSHIPS FOR TWO-EQUATION MIGRATION MODEL

Regressand:	y_1	y_2
Regressors:	$y_2 < 0$	$y_1 > 0$
	$z_1 > 0$	$z_1 > 0$
	$z_2 > 0$	$z_2 < 0$
	$z_3 < 0$	$z_6 > 0$
	z_4 ?	$z_7 < 0$
	$z_5 > 0$	

Table 2 intimates what we might expect, *a priori*. Thus, income is expected to be inversely associated with net out-migration for logical reasons. Likewise, the sign for the differential shift should be negative in order to display the comparative disadvantage of local opportunities. In view of the overall record of migration loss in rural Manitoba, lagged out-migration, distance to centres and the man-land ratio should all have positive signs: the first to indicate the on-going character of rural exodus; the second to suggest that residents occupying isolated districts will contribute proportionately more to the migration stream; and the third to pick up the inefficiency of small holdings. A degree of uncertainty surrounds population change; less people in rural areas may lead to a tapering-off of out-migration but it is equally plausible to surmise that in a similar situation more people would

feel compelled to leave because of the collapsing viability of declining agricultural service towns and villages. Meanwhile, the income regressand is presumed to be directly associated with out-migration, following from the proposition that a declining workforce is commensurate with increased labour scarcity and higher wages. Similarly, the previous migration stream should have taken pressure off the surfeit of labour. It is thought, however, that rural isolation is not favourable to economic diversification so the distance-to-centre variable is expected to be negative in sign (as is the proportionality shift factor because of the general lack of dynamic sectors in rural areas). Finally, value of farm sales is assumed to make up a large proportion of the household income in regions such as Manitoba. Verification of these relationships awaits calibration of the model.

The initial *technical* step is to determine the level of identifiability of each equation. Because the rank condition is usually intractable at this stage, onus is put on the order condition. Basically, this can be paraphrased to mean that the number of zs excluded from an equation must be greater or equal to the number of ys in that equation minus one. Equation (29a) excludes two zs and contains two ys and equation (29b) excludes three zs for its two ys, so both of them are over-identified. Therefore, we can proceed with TSLS estimation. Precision would require transformation of the raw data in Appendix 3 in order to satisfy the three basic prerequisites of regression systems: to ensure linearity; to approximate normality in the disturbance terms; and to stabilize the variance of the disturbance terms. However, the most versatile transformation, that of the logarithmic, is precluded from our analysis because of the presence of negative observations in several of the variables. Eradication of the negative character of the shift and population change variables in order to perform a logarithmic conversion would be logically nonsensical. Accordingly, the variables are used in their original form even though this fails to conform to the usual practice of a preliminary check of the basic assumptions of the linear model. This action is excused in the name of *heuristic* exposition.

The first of the five steps for TSLS estimation has already been accomplished with the selection of the two target equations, those for the y_1 and y_2 regressands. Commencing with y_1, the *first stage* (second of the five steps) of TSLS entails OLS estimates of y_2's reduced form, which is to say, y_2 regressed against all the predetermined variables in the model. These values, represented by $\hat{y}_2$, are duplicated in the fourth column of Appendix 3; the estimated parameters of that reduced form equation are specified below:

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	R^2
$\hat{y}_2$	-0.163	-0.217	-1.431	-0.007	-0.098	0.071	-0.966	0.41

Usually, the equation estimated in the first stage is of no particular interest to the modeller, but it may be used to complement the second stage results in a spatial sense as will be later demonstrated. The third step of computation requires the substitution of y_2 into equation (29a) and the fourth step is the subsequent regression of y_1 on $\hat{y}_2$ and $z_1 \ldots z_5$. Results of this *second stage* regression are outlined in the top half of Table 3.

Implementation of the fifth step requires a repeat of the first four steps on equation (29b), the specification for the average family income regressand. As can be seen from Figure 7, this dependent variable has a reciprocal relationship with the other endogenous variable, net out-migration,

TABLE 3: TSLS RESULTS FOR EQUATIONS (29)

			Estimated Coefficient	Standard Error	t-Statistic
Dependent Variable	y_1	(constant)	535.4	974.4	0.549
Regressors	y_2		-0.061	0.157	-0.388
	z_1		0.312	0.150	2.078 *
	z_2		4.047	3.569	1.134
	z_3		0.095	0.383	0.248
	z_4		-0.157	0.057	-2.788 **
	z_5		-3.003	1.342	-2.239 *
	R^2	.22			
	F	5.15	($F_{0.99}$ (6, 107) ∿ 2.9)		
	SE	300.43			
			$t_{0.90}$ (107) ∿ 1.29		
			$t_{0.95}$ (107) ∿ 1.66		
			$t_{0.99}$ (107) ∿ 2.36		
Dependent Variable	y_2	(constant)	5997.7	296.5	20.225 **
Regressors	y_1		2.408	1.168	2.062 *
	z_1		-2.297	0.895	-2.566 **
	z_2		-29.654	9.053	-3.276 **
	z_6		1.506	0.794	1.898 *
	z_7		1.525	0.326	4.676 **
	R^2	.09			
	F	2.03	($F_{0.95}$ (5, 108) ∿ 2.3)		
	SE	1127.66			
			$t_{0.90}$ (108) ∿ 1.29		
			$t_{0.95}$ (108) ∿ 1.66		
			$t_{0.99}$ (108) ∿ 2.36		

and shares the friction of distance variable with it. However, it is also presumed to be regulated by two exogenous variables, z_6 and z_7, which do not directly affect y_1. The second step for TSLS solution of y_2 is completed with the OLS estimation of y_1's reduced form, *viz:*

	z_1	z_2	z_3	z_4	z_5	z_6	z_7	R^2
$\hat{y}_1$	0.181	0.231	-1.238	-0.119	-0.156	0.195	-1.447	0.26

and the actual estimates of $\hat{y}_1$ are presented in the second column of Appendix 3. Substitution of $\hat{y}_1$ into equation (29b) and the resultant second stage regression parameters are displayed in the bottom half of Table 3.

In addition, Table 3 provides the usual significance statistics and they indicate a low degree of model explanation (R^2 values) although they also suggest that the net out-migration equation is significant at the 99% level and that for average family income is significant at about 90% (F-statistics). With respect to the individual coefficients, those with t-values commensurate with 95% significance are marked with a single asterisk whereas those significant at 99% are denoted by two asterisks. Clearly, the leading regressor for net out-migration is 4, population change. The uncertainty surrounding this variable at the hypothesis stage is eradicated as it has a strong inverse association with out-migration. In other words, as the population base declines in rural districts, the tempo of out-migration is enhanced. Two other variables make meaningful contributions to out-migration; the lagged migration variable, which conforms to expectations, and the man-land ratio, which does not. Instead of suggesting that the pressure of farmers on the land leads directly to out-migration, this latter variable implies that residents of low density farming areas are most likely to migrate. Presumably this is a reflection of low quality land incapable of supporting a large population in the first place and equally incapable of improvement to an extent sufficient to provide existing dwellers with an adequate living. In contrast, all of the variables are significant for the average family income equation. Three of the five coefficients conform to *a priori* postulates, but the lagged migration variable and proportionality shifts have signs opposite to those predicted. Apparently the previous rural depopulation has failed to promote the income position of the families left behind and this is in direct contradiction to the outcome proposed by neo-classical growth theorists. Moreover, the behaviour of the proportionality shift index indicates that, even in relatively depressed rural areas, strongly competitive industries contribute a disproportionate amount to the regional economy and, ultimately, to family incomes.

It is possible that the aforementioned incorrect signs (in an a *priori* sense) for regression coefficients are due to the occurrence of multicollinearity. This phenomenon denotes the presence of linear relationships in the regressors which, when stated simply, implies that the so-called independent variables are far from being orthogonal. Indeed, multicollinearity can be defined as the opposite extreme to orthogonality, the preferred condition of the regressors, and the greater the departure from orthogonality of these variables then the greater the presence of linear relationships (Farrar and Glauber, 1967). If multicollinearity becomes pronounced, efficient regression coefficients cannot be estimated. In order to obviate this danger, the regressors must be soreened prior to analysis and cases of obvious intercorrelation should be suppressed. This procedure is easier said than done in S-E situations where the regressors include endogenous variables which, b7

their very nature, monitor interdependence between equations.

If we assume that each of the disturbance terms for the two equations approximates zero and that the regressors capture most of the variation in the model (hardly the case in our example given the low R^2s), it is possible to gain a spatial insight into the workings of the feedback system. This is accomplished by gauging the differences between the original observations for the ys and the estimated observations, $\hat{y}$s, derived from the reduced-form (first stage) equations. In brief, the $\hat{y}$s (Appendix 3) show how far the regressands can be explained by the predetermined variables alone. Any disparity between y and $\hat{y}$ can be attributed to the effects of the endogenous regressors excluded from the reduced form. Therefore, those estimates which approximate the original observations are suggestive of rural districts which function largely outside the simultaneity bias of the model. In the case of y_1, that would be tantamount to saying that income has little or no feedback on the level of out-migration, whereas for y_2 it implies that variations in income can be accounted for by the independent variables without any recourse to the impact of migration. Conversely, spatial units with substantial discrepancies between original and estimated observations (i.e. residuals) are those most dependent on feedback effects of the jointly determined migration and income variables. The residuals most pertinent here would be few in number given the assumption of high R^2s and, hence, high regression explanation. Isolation of substantive residuals (regardless of whether they are positive or negative) would provide a mappable picture of those areas of Manitoba in which income and out-migration operate as an interdependent mechanism.

IV. AN EXAMPLE OF A MULTIPLE-EQUATION MIGRATION MODEL

(i) Background

The simple migration model described in the previous chapter serves to demonstrate the interdependent nature of factors affecting migration. Yet migration is a highly complex phenomenon which is influenced by a gamut of variables ranging quite literally from conditions of physical climate to those of social climate. A realistic migration model should be capable of accommodating all of these diverse regulators. Unfortunately, the addition of a large number of *explanatory variables* (endogenous and predetermined combined) tends to undermine the computational efficiency of TSLS. In fact, the presence of more than 20 predetermined variables will make solution of reduced-form equations by TSLS very cumbersome. There are two alternatives to this predicament. The first is to reduce a large number of explanatory into a manageable set by means of *principal components analysis* (Daultrey, 1976), and the second is to replace TSLS by a different computational technique. *Three-Stage Least Squares* (3SLS) is one such technique and it has the advantage of not only giving more efficient parameter estimates than TSLS, but also of performing the regression simultaneously on all equations in the model rather than on one at a time. Nonetheless, a penalty is paid for this greater efficiency and that is a more complioated mathematical procedure than that required for TSLS. In brief, 3SLS involves the application of *generalized least squares* (GLS) estimation to equations that have already been subjected to TSLS estimation.

The main difference between GLS and OLS is that the former dispenses with the assumptions of homoscedasticity and zero autocorrelation which are the cornerstones of the OLS treatment of the disturbance term. It may be

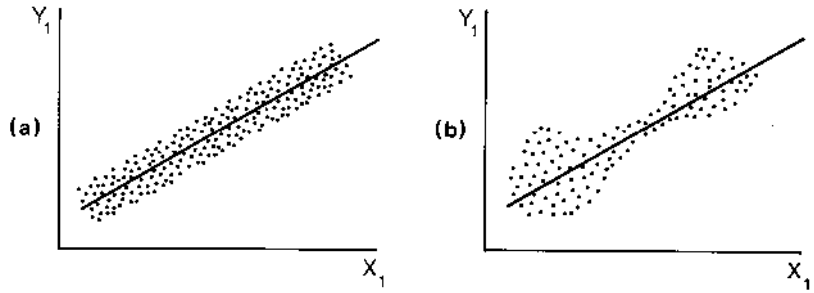

Figure 8. (a) Homoscedasticity and (b) Heteroscedasticity

recalled from Chapter II (ii) that the disturbance term in OLS is assumed to have constant variance, $E(u^2) = \sigma^2$, and zero covariance, $E(u_M u_N) = 0$. The former assumption is homoscedasticity, which means that the variances of the residuals are equal. If the variances are not equal, they are said to be heteroscedastic and this is symptomatic of systematic variation in the disturbance term (Figure 8). When homogeneity of variance is violated then the regression line has different errors across scale values on the X axis and this signifies that an improper form of regression equation has been fitted. Such heteroscedastic outcomes are not unusual in model-building where equations may be successively formulated on a 'trial-and-error' basis. Zero autocorrelation, on the other hand, refers to the property of zero covariance in the disturbance terms. In the absence of this property, autocorrelation occurs and the residuals are no longer independent of each other. The severity of this problem depends on the degree of interdependence between adjacent residuals. It can be a particular nuisance in time-series analysis when the residuals are representative of sequential time elements which are causal progressions from each other (then autocorrelation is renamed serial correlation), but residual association is also common in spatial situations where contiguous activities spill-over to influence geographically adjoining observations. Because GLS can handle these two problems which plague OLS, it emerges as a much more versatile technique than classical least squares regression.

Details of GLS computational procedures are not recounted here; suffice it to say that the crux of the technique is based on the assumption that the degree of association between the elements of the disturbance term is known (i.e. a covariance matrix of non-zero residuals can be concocted). In 3SLS that is usually not the case (and incidentally is the reason why GLS is not more popular than OLS in single-equation regression) and instead an approximation of it is derived from the residuals of TSLS estimation. The third stage (i.e. after TSLS) is to compute GLS parameter estimates for the structural equations. In sum, the main technical distinction between TSLS and 3SLS is that the simpler technique uses OLS to purge the stochastic component from the endogenous variables whereas the three-stage procedure utilizes GLS for the same purpose. The extra complexity may be justified in large equation systems and 3SLS is the computational foundation for the multiple-equation migration model developed by Greenwood (1973) and to which we now turn.

(ii) Structuring of the Model

Greenwood constructs a five-equation, two-identity model aimed at determining the interaction between migration and urban change. He uses 1950-60 data for 100 American SMSAs with individual populations in excess of 250 000. The equations and their accompanying rationale are debated in turn:

Out-migration

$OM = f\ (\underline{IM}, \underline{\Delta\ INC}, \underline{\Delta\ EMP}, \underline{\Delta\ UNEMP}, INC50, UNR50, CLF50, EDU50, AGE60, e_1)$

(endogenous variables underlined)

i.e. Out-migration is some function of in-migration, income change, employment change, unemployment change, 1950 median city income, 1950 city unemployment rate, 1950 civilian labour force, median education level, 1960 median age and a stochastic element. All variables in the model are log transformed.

The idea that out-migration and in-migration are inextricably bound up with one another is to capture the symmetry of interurban flows; people abandoning low income cities in favour of high-income cities should ensure almost a perfect inverse association between OM and IM for any given city, *a priori*. Moreover, people respond to anticipated future gains and so the pace of the change in income and employment opportunities becomes important for gauging the migrant's expected future returns (hence the inclusion of Δ INC, Δ EMP and Δ UNEMP with hypothesized -, -, and + signs). Increasing age and inadequate education should act as deterrents to out-migration.

In-migration

$IM = f\ (\underline{OM}, \underline{\Delta\ INC}, \underline{\Delta\ EMP}, \underline{\Delta\ UNEMP}, INC50, UNR50, CLF50, e_2)$

i.e. In-migration is some function of out-migration, income change, employment change, unemployment change, 1950 median city income, 1950 city unemployment rate, 1950 civilian labour force and a stochastic element.

The reciprocity with OM is apparent when the income and employment variables are assigned opposite signs inasmuch as they now act as 'pull' factors attracting migrants into a given city.

Income change

$INC = f\ (\underline{OM}, \underline{IM}, \Delta\ EDU, \Delta\ GOVT, DEW, DNS, e_3)$

i.e. Income change is some function of out-migration, in-migration, education change, local government expenditure change, dummy east-west variable (the former area assumed to be stagnating while the latter is expanding), dummy north-south variable (again, a dichotomy between dynamic and not-so-dynamic areas with the south fitting the first slot and the north the second slot), and a stochastic element.

The seeming paradox of change in income being attributed to both out- and in-migration occurs because migration in whatever form eventually affects labour demand as well as supply. Thus, out-migration reduces labour supply and can upgrade the wages of the workers remaining, as the neo-classicists argue, but emigration also removes consumers from the urban market place

which can reduce demand for goods and services and soon result in a negative employment multiplier. The men thrown out of work will continue to saturate the potential labour force and undermine the bargaining position of those still gainfully employed. Therefore it cannot be taken for granted that out-migration will compel wage rises for non-migrants. By the same token, in-migration creates an upsurge in demand for city goods and services and this is translated into more jobs, which eliminate pressures on the labour force and so prevent wages from dropping. In short, income change may be affected either positively or negatively by both out- and in-migration. The predetermined variables monitor the better chances of more educated personnel in gaining higher wages, the role of government spending in stimulating the economy, and the geographic dichotomies between prosperous cities in the West (e.g. California) and South (e.g. Texas) and the older, rather run-down centres of the East and North (e.g. Upper Great Lakes, New England).

Employment change

$EMP = f\ (\underline{OM}, \underline{IM}, \underline{NATINC}, INC50, \Delta\ EDU, \Delta\ GOVT, DEW, DNS, e_4)$

i.e. Employment change is some function of out-migration, in-migration, natural increase in the labour force, 1950 median income, education change, local government expenditure change, the two geographic dummy variables and a stochastic element.

The confusion surrounding the effects of migration on income are not replicated to the same extent with employment. On the whole, in-migration is expected to boost urban employment positively for most people will refrain from moving to a new place to be unemployed. Similarly, out-migration will have a tendency to reduce employment levels. Nevertheless, there is evidence to suggest that this view may be too simplistic. Cebula (1974), for example, has shown that there is a strong tendenoy in the United States for people to move to those areas that provide easy access to public goods (including blacks in search of generous welfare benefits). Therefore, the postulated relationships between employment change and in- and out-migration are questionable. The other variables are surrogates for factors which stimulate labour supply and, as in the income equation, the dummy variables monitor the incidence of regional comparative advantage.

Unemployment change

$\Delta\ UNEMP = f\ (\underline{OM}, \underline{IM}, \underline{NATINC}, DEW, DNS, e_5)$

i.e. Unemployment change is some function of out-migration, in-migration, natural increase in the labour force, the dummy geographic variables and a stochastic element.

Structural unemployment may exist in a city and be immune to the normal demand stimulation policies for employment creation. Unemployment change can thus be legitimately regarded as independent of employment change. The causes of structural unemployment (e.g. educational deficiencies, ethnic discrimination) may be such as to mitigate against out-migration, although, *a priori*, one might suppose that abysmal job prospeots will act to promote out-migration. Likewise, the lack of success elsewhere by poorly-equipped migrants may induce a large number of them to return home and even though this means continued unemployment it may be regarded as being better to be surrounded by one's family and friends than to be ensconced in a strange city. As a result, the impact of migration on unemployment changes is unpredictable.

Change in labour force

Δ CLF = Δ EMP + Δ UNEMP

An identity claiming that change in labour force is the sum of changes in levels of employment and unemployment.

Natural increase in labour force

NATINC = Δ CLF + OM - IM

A second identity stating that the combination of the change in labour force and the net migration gain is indicative of the natural increase in the labour force. The two identities act to close the system of equations.

An S-E procedure is required to estimate the five structural equations (all of which are over-identified).

(iii) Computational Results

Although the number of predetermined variables in his model is not excessive, Greenwood prefers to apply 3SLS rather than TSLS because of the former's more efficient parameter estimates and its ability to handle heteroscedasticity and autocorrelation. The actual 3SLS estimates derived for the U.S. case are presented in Table 4. Note that the R^2 values are for OLS equivalents of the specified equations and are included only as a crude inference of goodness-of-fit variations among the equations. Regression coefficients statistically significant at 90% are denoted by one asterisk while those significant at 95% or better are distinguished by two asterisks. Interdependence between the two kinds of migration comes out clearly on the employment and unemployment change variables, yet all equations have at least one endogenous variable acting as significant regressor. The feedback basis of the model is thus vindicated.

The conceptual implications of these results are summarized in Table 5 in which expected and observed relationships are compared. Underlined relationships are those designated as statistically significant whereas those marked with a tick violate expected signs on coefficients. Only three directions of signs fail to conform to expectations, of which only one is statistically significant: the other thirty-two abide either by *a priori* specifications or clarify previously uncertain associations. The high R^2s, although probably symptomatic of multicollinearity, nevertheless suggest that three of the five structural equations are 'explained' reasonably well. In view of relationships found significant from t-tests, each equation can be paraphrased to embody the following rationale:

(1) Out-migrants are attracted away from places of high unemployment at the initial period to places with a large urban labour force; the probability of moving is a direct function of the migrant's level of schooling at the beginning period. Correspondingly, migrants are repulsed by high and rising income levels in places of origin, and advancing age acts to discourage movement.

(2) In-migration into a city is a direct function of out-migration from other places, increasing incomes at the point of destination, and a large and expanding labour force there. It is deterred by a high unemployment rate in the potential place of destination.

TABLE 4: GREENWOOD'S 3SLS ESTIMATES (in logarithmic form)

Equation	(1) OM	(2) IM	(3) Δ INC	(4) Δ EMP	(5) Δ UNEMP
Constant	6.431**	-3.427**	0.117	0.689	2.064**
Regressors:					
OM		0.318**	-0.023	-0.224**	-0.399**
IM	0.139		0.043**	0.232**	0.192**
ΔINC	-1.247*	2.166**			
ΔEMP	0.165	2.843**			
ΔUNEMP	0.476	-0.598			
NATINC				-0.011	0.047**
INC 50	-0.545**	0.212		-0.065	
UNR 50	0.339**	-0.439**			
CLF 50	0.791**	-0.546**			
EDU 50	1.288**				
AGE 60	-1.212**				
ΔEDU			0.462**	-0.715	
ΔGOVT			0.033	0.071	
DEW			0.045**	0.129**	-0.104
DNS			-0.003	0.172**	0.172**
R^2	0.96	0.93	0.46	0.72	0.20

(3) Positive income change is a direct function of steady in-migration, improving education levels of the populace and a distinct comparative advantage enjoyed by some regions of the country. Accordingly, the neo-classical presumption of declining wages accompanying in-migration is not substantiated.

(4) Increasing employment opportunities are closely associated with gains from in-migration and regional comparative advantage while being countermanded by out-migration and the rising level of educational attainment. This latter phenomenon is unexpected and might reflect the presence of an over-educated population for the level of skills required in the job market.

(5) Increasing unemployment goes hand-in-hand with in-migration, an increasing labour force and regional disadvantage. The significant NATINC factor emphasizes the tendency of high unemployment among young entrants into the labour force. However, it is mitigated by out-migration, so that both in- and out-migration may be occurring concurrently in depressed cities; a tinge of realism which confounds the neo-classical argument.

We can do no better than quote from Greenwood to obtain a synopsis of the relevance of these parameters.

TABLE 5: EXPECTED AND OBSERVED MODEL RELATIONSHIP

Equation:		(1) OM	(2) IM	(3) Δ INC	(4) Δ EMP	(5) Δ UNEMP
Coefficient						
OM	(E)		>0	?	<0	?
	(O)		>0	<0	<0	<0
IM	(E)	>0		?	>0	?
	(O)	>0		>0	>0	>0
INC	(E)	<0	>0			
	(O)	<0	>0			
EMP	(E)	<0	>0			
	(O)	>0✓	>0			
UNEMP	(E)	>0	<0			
	(O)	>0	<0			
NATINC	(E)				>0	?
	(O)				<0✓	>0
INC 50	(E)	<0	>0		<0	
	(O)	<0	>0		<0	
UNR 50	(E)	>0	<0			
	(O)	>0	<0			
CLF 50	(E)	>0	>0			
	(O)	>0	>0			
EDU 50	(E)	>0				
	(O)	>0				
AGE 60	(E)	<0				
	(O)	<0				
EDU	(E)			>0	>0	
	(O)			>0	<0✓	
GOVT	(E)			>0	>0	
	(O)			>0	>0	
DEW	(E)			?	?	?
	(O)			>0	>0	<0
DNS	(E)			?	?	?
	(O)			<0	>0	>0

> The findings of this study do not suggest that out-migration encourages greater income growth such that regional income differentials are narrowed through interregional migration. Rather, to the extent that it does influence income growth, out-migration appears to depress such growth, as well as to depress employment growth. Out-migration does however, tend to relieve unemployment in sending localities (Greenwood, 1973, p 108).

Greenwood concludes by asserting that migration studies based on OLS produce inconsistent results precisely because of their inability to take into account the simultaneity bias. Models of migration cannot be divorced from S-E solution frameworks.

V. CONCLUSION

This monograph has focused on S-E regression, a variant of regression analysis much neglected by geographers. Simultaneous-equation regression is capable of handling causal mechanisms which stray across several equations and, because of this asset alone, its neglect in geography has been unwarranted. Comprehension of spatial mechanisms increasingly requires a systems view of phenomena. Indeed, geographers find that the spatial variables they deal with have appendages that reverberate through sectoral and temporal dimensions and these many layers of linkages and interdependence have to be monitored if only to record their feedback effects on spatial phenomena. Simultaneous-equation regression is a major tool for operationalizing models with systems perspectives and this has been demonstrated in the monograph with cases drawn from land-use transportation studies, geomorphology, regional growth, and migration; all are prime areas of concern to geographers.

The monograph concentrated on TSLS as the means for implementing S-E regression. TSLS is the simplest of the S-E estimation techniques, yet this simplicity does not detract from its usefulness and Blalock (1971, p 156), for one, is of the opinion that TSLS is wholly appropriate for dealing with over-identified equations. In his words, 3SLS and full-information maximum likelihood methods (Appendix 2) .. seem more sensitive to specification errors that arise in instances where theoretical foundations for the model are weak. It would appear as though two-stage least squares is entirely adequate for less advanced fields such as political science and sociology, given the presence of relatively poor measurement procedures and the very tentative nature of existing theories'. It scarcely needs to be added that geography fits equally well into Blalock's category of less advanced theoretical disciplines (relative to economics) and perforce, it follows that TSLS is a fitting technique for many geographical research designs. Nevertheless, TSLS like any other S-E modelling procedure cannot be operationalized without good conceptual frameworks. The very fact that S-E regression is more involved than single-equation regression calls for careful specification of variable interaction. In itself, this can only serve to promote stronger theoretical foundations in geography. Better model formulation is only one attribute of S-E regression and others can be paraphrased as:

(1) it enables a formal distinction between truly independent and jointly dependent variables;

(2) it provides a means for handling feedback between discrete equations within an extended model format;

(3) its more complex manifestations can accommodate the fraught issues of heteroscedasticity and autocorrelation; and

(4) it builds on functional relationships made familiar through classical least squares procedures and, hence, well known to geographers.

This last attribute can only serve to boost the attractiveness of S-E regression. Indeed, geographers should regard it as a complementary tool to the familiar OLS multiple regression analysis. The respected econometrician,

Christ, is convinced that our sister discipline of economics has gained substantial benefits from the addition of S-E methods to its inventory of regression techniques. He states (Christ, 1960, p 845) that

> it is not yet clear that the (ordinary) least squares method for structural estimation is dead. It is now clear, however, that even for small samples (ordinary) least squares will not do as well as simultaneous equation methods. The important task ahead is to learn more about how to decide which estimation method is likely to be best for any given actual econometric problem.

We can do no better in conclusion than to urge geographers to follow suit in their own discipline and match the relevant problem context with the appropriate regression instrument.

VI. BIBLIOGRAPHY

A. Econometric foundations

Christ, C.F. (1960) A symposium on simultaneous equation estimation. *Econometrica,* 28, 835-845.

Christ, C.F. (1966) *Econometric models and methods.* (John Wiley, London).

Goldberger, A.S. (1964) *Econometric theory.* (John Wiley, New York).

Kmenta, J. (1971) *Elements of econometrics.* (MacMillan, New York).

Schildernick, J.H.F. (1977) *Regression and factor analysis applied in econometrics.* (Martinus Nijhoff, Leiden).

B. The scope of regression

Farrar, D.E. and D.R. Glauber (1967) Multicollinearity in regression analysis: the problem revisited. *Review of Economics and Statistics,* 49, 92-107.

Ferguson, R. (1977) *Linear regression in geography.* Concepts and techniques in modern geography, 15, (Geo Abstracts Ltd. Norwich).

McNeil, K.A., F.J. Kelly and J.T. McNeil (1975) *Testing research hypotheses using multiple linear regression.* (Carbondale, Southern Illinois University Press).

Mark, D.M. and T.K. Peucker (1978) Regression analysis and geographic models. *The Canadian Geographer,* 22, 51-64.

Namboodiri, N.K., L.F. Carter and H.M. Blalock (1975) *Applied multivariate analysis and experimental design.* (McGraw-Hill, New York).

Unwin, D.J. (1975) *An introduction to trend surface analysis.* Concepts and techniques in modern geography, 5, (Geo Abstracts Ltd. Norwich).

Wrigley, N. (1976) *Introduction to the use of logit models in geography.* Concepts and techniques in modern geography, 10, (Geo Abstracts Ltd. Norwich).

C. Pertinent modelling examples

Blalock, H.M. (1969) *Theory construction.* (Prentice-Hall, Englewood Cliffs).

Blalock, H.M. ed (1971) *Causal models in the social sciences.* (Aldine-Atherton, Chicago).

Cliff, A.D. and J.K. Ord (1975) Model building and the analysis of spatial pattern in human geography. *Journal of the Royal Statistical Society,* Series B, 37, 297-328.

Daultrey, S. (1976) *Principal components analysis.* Concepts and techniques in modern geography, 8, (Geo Abstracts Ltd. Norwich).

Glickman, N.J. and W.W. McHone (1977) Intercity migration and employment growth in the Japanese urban economy. *Regional Studies,* 11, 165-181.

Putnam, S.H. (1975) *An empirical model of regional growth.* Monograph Series 6, (Regional Science Research Institutue, Philadelphia).

Putnam, S.H. (1978) *Urban residential location models.* Studies in applied regional science, 13, (Martinus Nijhoff, The Hague).

Richardson, H.W. (1969) *Regional economics.* (Weidenfeld and Nicolson, London).

Richardson, H.W. (1973) *Regional growth theory.* (John Wiley, New York).

Shaw, R.P. (1975) *Migration theory and fact.* Bibliography series, 5, (Regional Science Research Institute, Philadelphia).

D. S-E spatial applications

Agterberg, F.P. and P. Cabilio (1969) Two-stage least-squares model for the relationship between mappable geological variables. *Journal of the International Association of Mathematical Geology,* 1, 137-153.

Cebula, R.J. (1974) Migration and the Tiebout hypothesis: an analysis according to race, sex and age. *Journal of the American Statistical Association,* 69, 876-879.

Foot, D.H.S. (1974) *A comparison of some land-use allocation/interaction models.* Geographical paper, 31, (University of Reading, Reading).

Greenwood, M.J. (1973) Urban economic growth and migration: their interaction. *Environment and Planning,* 5, 91-112.

Greenwood, M.J. (1975) A simultaneous-equations model of urban growth and migration. *Journal of the American Statistical Association,* 70, 797-810.

Greenwood, M.J. (1978) An econometric model of internal migration and regional economic growth in Mexico. *Journal of Regional Science,* 18, 17-31.

Glickman, N.J. (1977) *Econometric analysis of regional systems: explorations in model building and policy analysis.* (Academic Press, New York).

Masser, I., A. Coleman and R.F. Wynn (1971) Estimation of a growth allocation model for north-west England. *Environment and Planning,* 3, 451-463.

Meyer, D.R. (1974) Use of Two-stage Least Squares to solve simultaneous equation systems in geography. in: *Proceedings of the 1972 meeting of the IGU commission on quantitative geography,* ed M.Yeates, (McGill-Queen's University Press, Montreal).

Muth, R.F. (1971) Migration: chicken or egg? *Southern Economic Journal,* 37, 295-306.

Okun, B. (1968) Interstate population migration and state income inequality. *Economic Development and Cultural change,* 16, 297-313.

E. Computer program

Cooper, J.P. and G.A. Curtis (1976) *Econometric Software Package, Users Manual.* (University Graduate School of Business, Chicago).

VII. APPENDICES

Appendix 1

In order to understand the rank criterion it is necessary to use some simple matrix concepts. The determinant is a crucial constituent of the rank criterion, as indeed it is of many linear algebraic concepts including the well-known characteristic equation for an eigenvalue problem. Consider two simultaneous equations with two unknowns, X_1 and X_2:

$$A_{11}X_1 + A_{12}X_2 = B_1$$

$$A_{21}X_1 + A_{22}X_2 = B_2$$

or, in matrix form:

$$\begin{bmatrix} A_{11} A_{12} \\ A_{21} A_{22} \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$

Solving these equations for the Xs gives:

$$X_1 = \frac{B_1 A_{22} - A_{12} B_2}{A_{11} A_{22} - A_{12} A_{21}}$$

$$X_2 = \frac{A_{11} B_2 - B_1 A_{21}}{A_{11} A_{22} - A_{12} A_{21}}$$

which are the *determinantal equations* for the unknowns X_1, X_2. Note that the denominators are the same and are obtained as the product of the elements on the principal diagonal of the 2 x 2 matrix of A parameters minus the product of the elements on the secondary diagonal. This procedure is known as taking the *determinant* of matrix A and is represented as IA I. The procedure can be

extended by means of Cramer's rule (see any introductory text) to find the determinants of any number of simultaneous equations. It is worth stating that the value of the determinant of an orthogonal matrix is +1 or -1, but it is zero if the matrix is afflicted with multicollinearity. Thus, the rank criterion requires that variables excluded from the equation in question must have among them at least one which is truly independent.

The determinant is also affected by the order condition. The order of a matrix is simply its size according to the number of rows and columns present and so the order of the determinant refers to the number of rows and columns of the matrix of simultaneous equations (in the example given above the order is two). In turn, the rank of a matrix is defined as the order of the largest non-zero determinant that can be obtained from the elements of. the matrix. As the determinant is incumbent on an equal number of rows and columns for this formulation, the rank of the matrix is the largest square matrix which can be ascertained from the underlying matrix. In sum, identification of any one equation from a system of simultaneous equations requires that the variables excluded from the equation in question can be formed into a matrix which fulfils the rank criterion.

Appendix 2

Maximum likelihood estimation is an alternative to least squares estimation in regression analysis. The objective of estimation is to make inferences about the values of unknown population parameters from sample data which are normally distributed. Consider the normal distribution of a single variable, X:

$$f(X; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(1/2\sigma^2)(X-\mu)^2}$$

which is dependent upon two parameters, the universal mean (μ) and standard deviation (σ). The likelihood function, L, or the probability of obtaining the sample, X is:

$$L = \prod_{i=1}^{N} f(X_i; \mu, \sigma)$$

where there are i=1...N independent observations. Substitution of the normal density function into L gives:

$$L = \frac{1}{(2\pi)^{N/2} \sigma^N} e^{-(1/2\sigma^2) \sum_{i=1}^{N} (X_i - \mu)^2}$$

and taking natural logarithms for both sides:

$$\log L = \log (2\pi)^{\frac{-N}{2}} + \log \sigma^{-N} + \log e^{-(1/2\sigma^2) \sum_{i=1}^{N} (X_i - \mu)^2}$$

$$= -\frac{N}{2} \log 2\pi - N \log \sigma - 1/2\sigma^2 \sum_{i=1}^{N} (X_i - \mu)^2$$

Next, the maximum of μ and σ can be obtained by setting the partial derivatives equal to zero and solving:

$$\frac{\partial (\log L)}{\partial \mu} = 1/\sigma^2 \sum_{i=1}^{N} (X_i - \mu)^2 = 1/\sigma^2 \left(\sum_{i=1}^{N} X_i - N\mu\right) = 0$$

$$\frac{\partial (\log L)}{\partial \sigma} = \frac{-N}{\sigma} + 1/\sigma^3 \sum_{i=1}^{N} (X_i - \mu)^2 = 0$$

and the maximum likelihood estimators for μ and σ^2 are:

$$\hat{\mu} = 1/N \sum_{i=1}^{N} X_i = \bar{X}$$

$$\hat{\sigma}^2 = 1/N \sum_{i=1}^{N} (X_i - \bar{X})^2$$

which are independent as the underlying population is normal. The fact that $\hat{\mu}$ and $\hat{\sigma}^2$ are uncorrelated is a property of key interest to equation formulations where there are some doubts about the independence of the regressors. If it can be assumed that the disturbance term is normally distributed, then maximum likelihood estimators can be applied to estimating the reduced form of S-E models. Indeed, the limited-information and full-information maximum likelihood methods have been devised to do just that. They differ in that the former utilizes only the restrictions imposed on the particular equation being estimated whereas the latter makes use of restrictions imposed on all structural equations in the model. Their computation is quite complex and so is not dealt with here (see Goldberger, 1964). Moreover, in view of the fact that geographers are not generally familiar with maximum likelihood methods, the monograph has concentrated on the least squares alternatives (TSLS and 3SLS) for S-E regression estimation. A brief introduction to maximum likelihood in linear spatial models is provided in Cliff and Ord (1975).

APPENDIX 3 DATA SET

y_1	$\hat{y}_1$	y_2	$\hat{y}_2$	z_1	z_2	z_3	z_4	z_5	z_6	z_7
590	326	5833	5959	633	13	-224	-504	46.96	100	357
512	449	4495	5126	402	21	175	-787	51.21	79	-319
168	367	6340	7117	640	3	-504	-60	13.88	247	758
118	258	5739	5662	284	8	81	-126	17.68	54	-174
62	24	8053	6926	69	10	-463	437	22.47	69	758
23	230	6907	6047	462	7	-144	-58	15.32	123	211
235	115	7008	6930	194	10	-486	-279	25.15	91	798
375	304	5762	5294	382	22	3	-301	60.74	159	46
561	280	4670	5345	445	27	-92	-563	67.94	100	155
659	608	5430	5000	1248	25	-79	-706	63.21	186	115
1313	543	5887	4954	306	35	132	-3567	80.41	281	-201
297	261	9314	5135	183	31	26	-575	93.24	76	-46
299	634	4946	5070	564	40	60	-622	37.82	192	-133
162	239	5931	5317	134	28	18	-150	55.60	85	-2
263	274	4877	4744	250	35	85	-1113	108.60	126	-111
579	502	5831	4950	512	30	129	-482	51.56	143	-207
297	324	5607	4647	259	40	124	-461	90.76	190	-171
474	444	5026	5121	420	38	-24	-471	22.08	202	130
54	144	5659	5440	183	21	-35	-105	75.30	84	71
232	402	4524	5020	177	41	81	-352	58.74	95	-142
372	313	5047	5471	256	26	-18	-295	43.91	47	24
247	327	4289	4967	384	33	23	-390	75.59	153	11
139	236	4672	4985	197	41	-2	-177	102.22	78	30
180	250	4553	4820	113	45	25	-219	115.29	101	-26
328	293	5490	4986	247	40	-8	-407	79.48	122	44
201	117	4169	4980	163	27	48	-218	100.16	85	-40
96	280	6178	4911	0	51	18	-142	84.08	82	-17
265	241	6330	4948	234	30	75	-280	89.85	145	-91
237	126	6557	5378	123	19	97	839	42.18	102	-172
236	246	7260	4962	162	35	53	-329	80.49	87	-48
251	463	5387	5724	448	11	84	-349	27.81	40	-244
-92	-20	7332	8631	138	5	-1112	529	16.51	93	1795
102	320	6665	7229	553	3	-514	-200	13.90	49	724
-64	79	7273	7032	0	12	-555	-4667	129.17	88	821
289	249	5039	5507	591	15	-85	-174	35.93	72	155
461	405	5325	5100	499	14	184	-573	55.94	141	-315
82	105	5927	6206	172	14	-287	186	66.38	160	531
330	806	6318	5648	1954	1	-239	-565	35.87	2237	328
585	-94	8392	8692	0	2	-1043	852	8.81	63	1661
471	295	6084	5002	1044	18	-119	-275	53.45	134	281
157	115	4170	5339	141	19	16	-108	73.01	75	10
47	81	5922	5917	14	6	-27	-111	56.12	102	55
378	360	5357	5395	319	17	83	-288	47.45	131	-110
257	233	4340	5222	111	17	156	-495	62.52	91	-238
126	115	5183	5384	203	10	66	-83	78.69	96	-92
29	246	5913	5479	148	24	9	-130	41.35	74	-4
2515	604	5616	4765	373	12	532	-2079	44.76	68	-1040
191	52	4709	5305	127	14	47	-121	94.81	106	-64
425	230	6110	5309	290	38	-190	-496	75.52	130	1410
150	207	6406	5661	214	25	-124	-143	35.75	72	240
261	429	6181	5218	230	38	23	-363	53.78	143	-30

Appendix 3 continued ...

y_1	y_1	y_2	y_2	z_1	z_2	z_3	z_4	z_5	z_6	z_7
0	62	6581	5826	116	11	-96	10	62.69	77	186
283	289	5287	5258	250	22	67	-344	60.49	94	-82
93	-103	7880	7465	91	13	-856	331	30.74	159	1571
105	144	5685	5624	91	10	43	-580	66.32	60	-92
107	150	8584	5715	43	14	27	-49	42.66	13	-78
1	136	5689	5875	69	20	-134	-88	32.23	83	261
147	250	4509	5200	118	30	72	-172	41.97	20	-117
161	99	3655	5324	175	20	-20	-232	81.02	71	69
180	259	4239	5156	314	28	10	-232	51.35	26	27
155	255	4839	5337	198	23	46	-296	56.05	39	-82
99	164	4245	5574	157	16	-6	-104	54.52	39	20
389	376	3816	5143	394	27	72	-315	37.40	56	-95
453	362	4526	5012	584	21	59	-374	66.82	163	-32
250	257	4585	5232	139	20	145	-352	58.46	96	-259
274	335	4575	5303	200	25	133	-370	35.23	37	-257
430	342	4879	5352	385	24	-5	-736	47.57	9	-5
228	529	3820	4505	457	30	363	-422	50.22	31	-662
389	262	5147	5586	186	17	27	-270	42.60	63	-48
190	365	5785	4958	155	37	148	-217	45.60	66	-259
271	382	4239	5423	208	34	-18	-443	35.80	72	24
343	576	4627	4871	649	30	182	-609	20.14	32	-341
219	443	5271	5009	325	38	94	-112	42.27	69	-163
108	216	5511	5673	222	33	-203	-132	19.14	51	392
89	350	5741	5299	236	41	-49	-107	16.26	17	91
669	639	4574	4554	670	48	112	-584	40.92	119	-150
675	624	7237	5463	455	26	174	-459	8.24	20	-490
469	442	5880	5534	339	60	-322	3	17.63	30	563
430	354	5485	5708	239	25	-28	-434	3.41	15	-5
122	432	5828	5228	306	53	-120	-39	14.60	45	213
137	352	4336	4856	103	47	83	-221	53.24	60	-115
272	364	5260	4743	257	44	76	-478	66.57	106	-74
299	433	5511	4871	140	52	95	-220	34.03	44	-190
206	256	5447	5175	118	27	84	-348	69.18	89	-123
262	331	4655	5042	241	35	50	-388	58.17	99	-49
345	342	4559	5160	148	44	-14	-200	41.93	63	37
151	325	3913	4863	240	37	106	-231	66.73	65	-166
241	348	4830	4995	321	33	79	-291	55.45	30	-133
130	294	3829	5103	273	33	28	-169	38.57	21	-4
290	386	5319	5042	140	41	84	-427	43.59	39	-145
263	361	5215	4871	372	41	16	-402	55.68	63	27
382	587	3875	4436	482	54	136	-569	70.12	108	-215
1539	584	4367	4439	560	72	-58	-453	32.30	28	142
430	400	3667	4939	385	44	49	53	11.62	10	-91
745	372	4577	4691	506	57	39	2919	44.86	177	-16
454	559	5746	5308	639	3	257	-783	49.04	106	-496
341	441	4903	4785	444	13	349	-530	66.68	103	-596
99	382	4626	5057	366	23	142	-382	66.43	91	-231
118	-96	5533	4748	266	12	77	-239	35.57	50	145
214	299	8362	5431	314	19	15	-1331	47.47	61	-29
547	451	4420	5368	450	26	38	-99	14.68	67	-61
234	467	3473	4783	391	23	334	-211	34.97	17	-611
197	351	4626	4870	308	37	114	-62	45.69	25	-187
112	224	5533	5337	155	23	53	-203	52.95	30	-97

Appendix 3 continued ...

y_1	$\hat{y}_1$	y_2	$\hat{y}_2$	z_1	z_2	z_3	z_4	z_5	z_6	z_7
201	265	5910	5378	236	18	87	-233	36.05	42	-151
392	407	4075	4851	442	20	261	-480	38.85	35	-453
1539	468	3763	4837	560	35	122	-305	24.57	41	-222
784	382	6910	5924	325	22	-83	-319	10.16	7	43
-94	253	6575	5798	114	24	-55	50	15.84	14	57
337	450	5592	5472	155	26	156	-690	11.30	8	-369
273	394	5586	5340	399	25	64	-445	10.74	8	-154
322	636	3877	4715	573	19	463	-601	10.55	5	-921
261	268	6027	5562	192	28	-24	-38	14.85	0	3