J. Statistics
1997(1), 25-36

STATISTICAL METHODS FOR THE ANALYSIS OF AGGREGATE DATA ON FERTILITY DIFFERENTIALS

By

MOHAMMAD ASHRAF CHAUDHARY,

Institute of Statistics,
University of the Punjab, Lahore, Pakistan.

MUNIR AKHTAR

Department of Statistics
Islamia University Bahawalpur, Pakistan.

MOHAMMAD NAIM SAJJAD,

Department of Statistics,
Baluchistan University, Quetta, Pakistan.

ABSTARACT

This paper reviews the selected statistical procedures for the analysis of aggregate fertility change using time series of cross-sections. The strengths and weaknesses of each analytical technique and the problems encountered are presented. The discussion is primarily methodological and focuses on what questions can be answered with the use of a particular statistical procedure for thę analysis of areal unit data collected at several time points.

Key words: Time series of cross-sections, Areal-unit, Multivariate analysis, Aggregate data.

1. INTRODUCTION

The analysis in which geopolitical units serve as the basic units of analysis is usually termed as areal units analysis (Duncan et al; 1961; Hermalin 1972). The use of areas rather than individuals as the unit of analysis has been prominent in demographic research in the past. This type of analysis is influenced by the concept of population as an aggregation of individuals, the type of demographic data available and by the nature of the problem to be investigated. The rationale of using areas as the basic unit of analysis lies in the fact that the reproductive behaviour of couples is affected both by their personal characteristics and the context in which they live. This refers to the areal unit environments meaning that the areal units defined by their existing boundaries should be meaningful units of demographic, social and economic experience of the society.

The aggregate type of analysis has been frequently applied to study the relationship between fertility and economic development. several recent studies have employed areal unit analysis and successfully investigated the fertility transitions in many parts of the world. Fertility behaviour in Europe was examined with the help of areal unit analysis by, among others, Richards (1977) and Coale et al., (1979). It has also been applied in studies of fertility decline in the third world including several investigations of Taiwan [Schultz(1973); Hermalin (1971), (1976)].

2. AREAL MULTIVARIATE ANALYSIS

Areal-unit multivariate analysis has been employed in a number of ways to investigate trends and differentials of fertility and nuptiality transitions. The estimation procedures mostly rely on general linear models. As outlined by Hermalin (1975), general features of a mltivariate analysis of this type are as follows:

(i) One should have measurements for each of the areal units on a set of variables whose influences on fertility levels is to be found out. These are called independent variables.

(ii) Some sort of the measure of fertility which is assumed to be influenced by the independent variables is to be established of each areal-unit under investigations. This is called the dependent variable.

(iii) Some appropriate method of analysis which enables the investigator to appraise the relative impact of independent variable should be decided upon.

Since the analysis is at an aggregate level, the dependent as well as the independent variables should be measured at aggregate level. The choice of variables and the mode of analysis is guided by a model or conceptual forme work of the relationship one seeks to explain or interpret. The model identifies the key concepts and their relationships. The main steps in an areal multivariate analysis thus involve developing a conceptual framework, operationalizing the concepts and choosing some appropriate statistical technique. The outcome of the analysis is likely to be influenced by variations at each of these stages. For example different statistical techniques applied to the same model and data, may lead to quite different results. Similarly, differences may arise from alternative operational definitions and concepts, though the other elements remain fixed.

The question to be answered by this type of analysis is to what extent the decreasing levels of fertility is due to increasing levels of education, decreasing child mortality, family planning program inputs, and other modernizing trends. The analysis should enable the investigator to test the hypothesis that whether among areas within the country, there exists a relationship between the levels of fertility and child mortality (say) after taking into account other independent variables included in the analysis. Viewed in this way, multivariate areal analysis is becoming more and more popular as a preferred mode of analysis. Main discussion in this paper is related to the third stage of this type for multivariate analysis that is selection of an appropriate method of analysis. A brief discussion of the important statistical procedures used in several studies over the past few decades, to answer the questions raised above, is given below.

2.1 AREAL-UNIT REGRESSION ANALYSIS

Cross-sectional estimations are usually required to describe the relationship between fertility and its determinants at an aggregate level and explain the fluctuations at certain time points. The model or conceptual framework of the relationships may involve a single regression equation or a set of equations. In order to illuminate the features of statistical estimations, inferences drawn, the assumptions involved, and the problems encountered, consider a single equation regression model which may be written as:

$$Y_j = \beta_o + \sum_{i=1}^{k} \beta_i X_{ij} + U_j$$

where, j = 1, 2,..., N; the number of areal units and i = 1, 2, ..., K, the number of independent variables in the model.

Y_j = j^{th} observation of the dependent variable.

β_i =Regression coefficient for the i^{th} predictor variable.

X_{ij} = j^{th} observation of the i^{th} predictor variable,

U_j = Error term representing all unmeasured influences.

The regression model is based on the following assumptions.

(i) Fertility is a linear function of the predictor variables.

(ii) Fertility should not influence the predictor variables i.e., reciprocal linkage not allowed..

(iii) The population of error terms have zero means and equal variances.

(iv) The error terms are uncorrelated with one another.

(v) The error term is not correlated with any of the predictor variables.

If these assumptions are satisfied, then the method of least squares can be used to obtain the estimates of the population parameters (β's and the variance of error term). Since the regression coefficients do not depend on the level but on the pattern of change of the dependent variable in response to the independent variables, the estimates of the β's contribute to our understanding of the patterns of fertility change by revealing the relative importance of the factors affecting fertility.

If the assumptions are not satisfied by the data, the estimates will be biased and misleading. Let us assume a model in which age at marriage and education are included as predictors of fertility. If age at marriage and education are correlated, then this model will produce different estimates for the effect of education as compared to those in a model in which age at marriage is left unobserved thereby altering our understanding of the fertility process and also the influences of the other independent variables. Failure to include a variable of this sort is referred to as specification error. In order to prevent this problem, one should examine carefully both the factors assumed to be important determinants of fertility themselves and the nature of their interrelationships. If two highly correlated predictor variables are included in a model, the problem of multi-collinearity will arise and distort the estimated coefficients thereby giving misleading conclusions. It is difficult to gauge reliably the individual effects of highly intercorrelated variables [Kleinbaum et al., (1988)]. Therefore, it is advisable to include either one or the other of the correlated predictor variables. Usually, a single equation regression model is estimated where as many models of fertility can be expressed adequately only through a system of structural equations and simultaneous estimation approach gives the estimates of all the parameters involved in that case.

There are certain typical problems associated with the regression analysis based on the data resulting from small areal units. One of the problems of regression analysis using small areal units is referred to as spatial autocorrelation. One of the assumptions of this analysis that the estimates of the dependent variable form two local areas will be independent form one another is very likely to be violated in the case of small, specially adjacent areal units. In the presence of autocorrelation, least square estimates turn out to be inefficient but still unbiased. There is no practical solution to the problem of spatial autocorrelations (Hermalin, 1975).

The risk of another statistical problem referred to as "model misspecification" arises, when small areal units are used. Regression analysis requires the assumption that variables omitted from an equation but causally related to the dependent variable are uncorrelated with any of the independent variables included in the equation. It is often quite likely that these variables are influenced by the excluded variabies. If the regional variations persist after estimation which is indicated if the residuals of the estimated models are found to be correlated (spatial autocorrelation), then it means that some of the factors accounting for areal variations have not been incorporated into the model (model misspecifications). We see that the problem of misspecification and spatial auto-correlation are inter-related but conceptually these are separate problems. The least square estimates of the coefficients in the presence of specification error are biased and inconsistent. To a greater or a lesser extent, specification error plagues almost every analysis. The use of small areal units therefore, aggravate the problem by increasing the risk of error resulting from the omission of variables which operate at the regional level. Chaudhary (1986 and 1987) used areal unit regression analysis to evaluate the determinants of decline in fertility levels and rise in female age at marriage in Taiwan.

2.2 RECURSIVE PATH ANALYSIS

Recursive path analysis, a multivariate technique to estimate linear causal models has also been used frequently in areal unit analysis. It was used by Hermalin (1971) to analyze the model postulation that an area's level of socio-economic development, its demographic status in terms of age at marriage, its mortality and its level of IUD acceptance from the program, influence its level of fertility. Age at marriage and IUD acceptance were treated as endogenous variables and Ordinary Least Squares regression was considered applicable under the assumptions used.

The techniques of path analysis has close affinities with multiple regression analysis but helps make explicit the underlying assumptions and inter-relationships. Path analysis is also useful in explaining the indirect effects inherent in the model. Correlation between two variables can be decomposed into the direct effect, the indirect effects and the joint effects shared with other variables in the system (Duncan, 1961; 137-138). Path analysis is a useful multivariate technique for explaining the interrelationships in a conceptual model; under certain circumstances it can reveal the relative utility of observed correlation and lead to a more parsimonious representation of these correlation.

Recursive path analysis is based on OLS estimation and give biased and inconsistent estimates of the equation parameters if disturbances across equations are correlated. Moreover, if the model posits reciprocal linkage between two or more endogenous variables, then the conditions for using least squares are not met because the system is no longer recursive. Of the techniques other than OLS, we discuss only Two Stage Least Squares (2SLS) which is frequently used. The main idea is to purge the endogenous variables used as explanatory variables in the system of their correlation with the disturbance terms. This is accomplished by using estimates of the endogenous variables based on a first stage regression on all or some of the exogenous variables in the model (Johnston, 1972). In order to apply this technique to any equation, the condition of identifiability must be satisfied. This requires that the number of exogenous variables excluded from any equation be at least equal to the number of endogenous variables included. Schultz (1971) used 2SLS to estimate a simultaneous equation model for Taiwan.

2.3 REGRESSION ANALYSIS OF FERTILITY CHANGE

The analysis of change overtime can be made by the comparison of regression coefficients across the equations estimated on certain time points say using single year data. However, it may not illuminate the dynamics of longitudinal components of fertility change overtime. For the analysis of change overtime, we can introduce change in the dependent as well as independent variables which can be done in a variety of ways. For example, absolute and proportionate changes can be measured, residuals from a regression of end values on initial values can be used, and regression slopes can be estimated for each unit if time series of observations are available [Duncan et al., (1961)]

The direct and absolute measures of fertility change can be regressed on the differences in the indices of predictor variables between two time points [Ting (1983)]. The independent variables showing very little change overtime can not be included as predictors in such regression analysis of change.

2.4 FACTOR ANALYSIS

Another method of multivariate analysis applicable to areal data is factor analysis. It attempts to reduce the data matrix by constructing the matrix of correlation for a smaller set of unobserved variables or latent factors.

The focus of this technique is on the mutual interdependence of a set of variables rather than on the dependence of a given variable on a number of explanatory variables. Park (1972) analyzed the fertility of Honolulu using factor analysis. He also compared the implication of factor analysis with that of a multiple regression analysis of fertility. Le Bras (1971) studied fertility differences among 89 department of France using factor analysis of this type. Factor analysis does not produce a unique solution and the technique is only exploratory in nature.

2.5 ANALYSIS OF COVARIANCE OF POOLED TIME SERIES OF CROSS-SECTIONS

Combining multiple sets of cross-sectional observations for successive time points within the period of analysis will generate data exhibiting temporal variations and can be thought of as two way analysis of covariance design. Very useful analysis can be carried out to take advantage of the information contained in the pooled data set. One of the varieties of ways may be to disregard the distinction between areal variations across the units and temporal variations within the units and naively estimate equations using all observations in the data set applying standard techniques. The data transformed as above can be utilized to estimate simple recursive systems of equations with estimates possessing all desirable statistical properties. The much larger number of observations in the pooled data permit more precise estimates than is possible in cross-sectional estimation. With this approach, the principal gain from pooling is more efficient estimation.

The multiple regression which includes as predictor variables one or more continuous variables as well as a set of dummies representing categorical variable can be used as a general method of Analysis of Covariance. The changing influence of the various socio-economic and demographic factors on fertility in Taiwan was successfully investigated with the application of this analytical procedure by Chaudhary (1990). A brief account of this procedure is given below.

This study used the socio-economic and demographic data published annually since 1961, documenting changes in almost all of the variables needed for the study of fertility trends and differentials and is related to

309 local areas of Taiwan. Pooled time series of cross-sections data set was generated by combining multiple sets of cross-sectional observations corresponding to 9 time points in the period 1961-76. The pooling resulted in 2781 (9x309) observations on each of the variables included in the analysis. Year was taken as the factor in this analysis and eight dummies were created for the time points 1961, 1963, 1965, 1966, 1968, 1970, 1972, 1974 where 1976 was taken as the reference year. Inclusion of these dummies controlled the temporal variation in addition to that explained by the covariate in the model. The covariates are Child Morality, Female Education, Child Education, Male Non-Agricultural Employment, Female Non-Agricultural Employment and Percent of Women Not Married Aged 20-24. These covariates are included in the model one by one and separate analysis of covariance tables are constructed in each case. The interactions of the covariates with the year dummies are also included to appraise the changing influence of covariates overtime of fertility in Taiwan. This analysis may be used to test the following six hypotheses:

(i) Is the interaction between 'year' and a particular covariat significant? That is, does the effect of the covariate changed from year to year?

(ii) Does the inclusion of the interaction of year and the covariate improve the fit of the model and contribute anything to the predication of fertility?

(iii) If these two hypotheses are rejected, that is the interaction between year and the covariate is insignificant then we can turn to the exploration of the simple additive model of year and covariate and proceed to test the following three hypotheses:

(iv) Is the net effect of time significant after the differences in the covriate are controlled?

(v) Is the covariate significant after controlling the affect of the time trend?

(vi) How good is the simple additive model of the effects of time and the covariate?

Testing these hypotheses requires the estimation of the following four models:

$$Y_{jt} = \beta_{10} + \beta_{11} T_1 + \beta_{12} T_2 + \ldots + \beta_{18} T_8 + U_{1jt} \qquad (1)$$

$$Y_{jt} = \beta_{20} + \beta_{21} x_{jt} + U_{2jt} \qquad (2)$$

$$Y_{jt} = \beta_{30} + \beta_{31} T_1 + \beta_{32} T_2 + \ldots + \beta_{38} T_8 + \beta_{39} X_{jt} + U_{3jt} \qquad (3)$$

$$Y_{jt} = \beta_{40} + \beta_{41} T_1 + \beta_{42} T_2 + \ldots + \beta_{48} T_8 + \beta_{49} X_{jt} + \beta_{410} (T_1 X_{jt}) + \beta_{411} (T_2 X_{jt}) + \ldots + \beta_{417} (T_8 X_{jt}) + U_{4jt} \qquad (4)$$

where j = 1,2, ..., N, the number of local areas and t = 1, 2, ..., K, the number of time points, β's denote the regression coefficients, $T_1, T_2, \ldots\ldots, T_8$, are dummy variables representing the years: 1961, 1963, 1965, 1966, 1968, 1970, 1972, and, 1974 and X_{jt} is the observation of the covariate corresponding to the j^{jt} local area at the time point t. The terms $T_1 T_{jt}, \ldots\ldots, (T_8\ T_{jt})$ represent the interactions of the covariate and year dummies and U_{jt} is the error term representing all unmeasured influences.

This techniques takes into account the time factor more explicitly in the study of the transition of the fertility behavior. Moreover, this procedure is distinct from the areal unit regression analysis discussed earlier because instead of analyzing the single year data or overtime charges based on two points of time, this statistical model incorporates time dummies to illuminate the time path of changing relationships.

REFERENCES

[1] **Choudhary, M.A.,** Diamond, I.D., and Casterline, J. B., (1990), "On Cross-Sectional Analyses of the Decline of Fertility in Taiwan", *Pakistan Journal of Statistics,* 6(3) A, 75-82.

[2] **Choudhary, M.A.,** (1987), "Social & Dèmographic Determinants of Rise in Female Age at Marriage in Taiwan: 1961-76", *Pakistan Economic and Social Review,* XXV, No. 2, 73-88.

[3] **Choudhary, M.A.,** (1986), "An Areal Unit Regression Analysis of the Determinants of Fertility Differentials in Taiwan: 1961-76", *Pakistan Journal of Statistics,* 2(3) B, 79-90.

[4] **Coale et.ab.,** (1979), "*Human Fertility in Russia Since the 19th century",* Princeton University Press, Princeton, New Jersey.

[5] **Duncan et. al., (1961),** *"Statististical Geography: Problems in Analyzing Areal Data",*Glencoe, Illinios: The Free Press.

[6] **Hermalin, A. I., (1971),** "Taiwan: Appraising the Effect of a Family Planning Program Through an Areal Analysis", *Taiwan Population Studies*, Working paper No.14, Ann. Arbar, University of Michingan.

[7] **Hermalin, A. I., (1975),** "*Regression Analysis of Areal Data",* in C. Chandrasekaran and A. Hermalin, eds., *Measuring the Effect of Family Planning Programs on Fertility.,* Paris :OECD, pp.245-299.

[8] **Hermalin, A.I., (1976),** "Spatial Analysis of Family Planning Program Effects in Taiwan", Paper Presented at the Seventh Summer Seminar in Population, East-West Population Institure, Honolulu, Hawaii, June 1976.

[9] **Jhonston, J., (1984),** "*Economertric Methods*",New York: McGraw-Hill.

[10] **Kleinbaum, D.G., Kupper, L.L., and Muller, K.E. (1988),** "*Applied Regression Analysis and Other multivariate Methods",* PWS-KENT Publishing Company Boston.

[11] **Park (1972),** "Multivairate Analysis of Areal Fertility in Honolulu", East-West Centre.

[12] **Richards, T. (1977),** "Fertility Decline in Germany: An Econometric Appraisal", *Population Studies,* Vol. 31, No. PP. 537-553.

[13] **Schultz., T.P. (1971),** "*The Effects of Population Polices: Alternative Methods of Statistical Inference",* Santa Monica, Rand (p. 468).

[14] **Schultz, T.P. (1973),** "Explanation of Birth Rate Changes Over Sapce and Time: A Study of Taiwan", *Journal of Political Economy,* Volume 32, No. 2, pp. S238-s273.

[15] **Ting, Y.T. (1983),** "*The Transitions of Family Limitations Practice in Taiwan, 1961-1980:* An Areal Unit Analysis", Ph. D. Dissertation Department of Sociology, University of Michingan.