

COVARIANCE ANALYSIS - A MULTIVARIATE CINDERELLA?

ALAN G. FERGUSON,
University of Nairobi

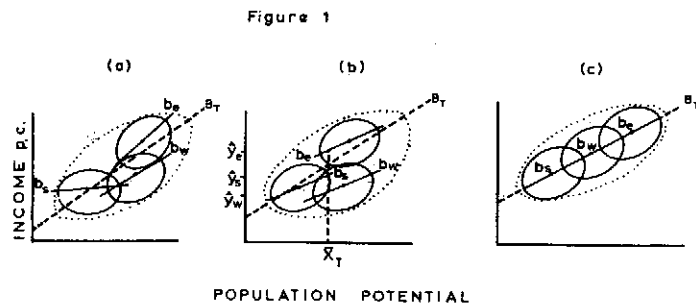
The steady improvement in data sources and the growing number of University computers with large storage capacity has allowed something of a tide of multivariate methods to wash over geographers in recent years. It is surprising, therefore, that when techniques such as principal components analysis, multiple regression, multi-dimensional scaling, etc., are in frequent use in many departments, one technique of great potential use - covariance analysis - seems to have been largely excluded from the geographer's portfolio.

The purpose of this paper is basically to provide some sort of resuscitation* for covariance analysis and to demonstrate its potential value to geographers.

Detailed documentation of the mathematics of covariance analysis may be found in several texts, two of which are listed at the end of the article. Basically, covariance analysis sets out to test whether significant differences in the parameters of the regression equation result when a set of observations (e.g., British Counties) is divided into sub-sets of treatment groups (e.g., English, Scottish and Welsh Counties). The technique is something of a hybrid, being derived from a combination of analysis of variance and multiple regression analysis.

For simplicity, the situation with one variate (dependent variable) and one covariate (independent variable) will be used as an example, although in practice we are usually confronted with a multivariate situation. The rationale for the one-covariate situation is, however, the same.

Suppose, as suggested above, British Counties are used as a data base and, hypothetically, average income per capita is regressed on population potential as $Y = \alpha + b_1 X_1 + e$ in the usual regression equation, firstly for all Counties and then separately for English, Scottish and Welsh Counties. We are now concerned with the interpretation of the regression parameters α and b_1 . If the groups are taken together, three situations are possible, shown in Figure 1 (a) - (c).



COVARIANCE ANALYSIS - A MULTIVARIATE CINDERELLA?

The total sample of all Counties is shown in Figure 1 by the dotted ellipse. The overall least-squares line is given in each figure by B_T while the individual lines for the three treatment groups are given by b_e , b_s and b_w respectively.

The three possible situations are:

- 1) The slopes of the regression lines of the treatment groups differ significantly from each other as in Figure 1 (a).
- 2) The treatment group regressions have similar slopes but if the treatment groups are superimposed on the overall mean of the covariate, \bar{X}_T , there are significant differences in the corresponding values of the dependent variable - This is equivalent to testing for significant differences between the intercept terms α_e , α_s and α_w given equal slopes - i.e., when $b_e = b_s = b_w$. In Figure 1 (b) \bar{X}_T represents the mean population potential for the whole sample and the vertical line from \bar{X}_T determines the value of the adjusted Y -means, \bar{Y}_e , \bar{Y}_s and \bar{Y}_w for each group.
- 3) There is no significant difference between the treatment groups in beta coefficient or intercept, so that each group simply reflects the overall least-squares equation, as in Figure 1 (c) - i.e., $b_e = b_s = b_w = b_T$ and $\alpha_e = \alpha_s = \alpha_w = \alpha_T$.

If significant slope differences are found (i.e., the situation in Figure 1 (a)) the second situation is not normally tested since the groups will have already been proved significantly different.

The significance testing of these situations requires the sums of squares and cross-products for the total sample and the individual treatment groups. The residual sum of squares in Y is calculated, firstly from the three within-group least-squares lines and, secondly, assuming no significant differences in slope are found, from the common pooled slope of the three groups. Significance testing is then performed using the F -distribution. These steps are clearly set out in Tatsuoka.⁴

Returning to the hypothetical example, it is interesting to consider the implications of the three situations shown in Figure 1. If significant differences resulted when the slope coefficients were compared, then there is good reason to deal with each group as a separate entity. From Figure 1 (a), one would conclude that, in England, income measured on a County basis was more responsive to population potential than in Wales or Scotland. Further analysis could proceed on a three-fold rather than single sample basis and reasons for the different responses could be sought. If no significant differences in the slope coefficients were found, the next stage would be to test whether the situation shown in Figure 1 (b) or Figure 1 (c) was the correct one. Even if the groups had equal means in X , different elevations would result in different means in Y , therefore, if significant differences were found here we would conclude that inputs of population potential had the same marginal effects on income in England, Scotland and Wales, but that population potential operated at a higher level in England than in Scotland and in Scotland than in Wales. If, however, no significant differences were found in slope or intercept, the conclusion would be that Figure 1 (c) represented the true situation and that the relationship between population potential and per capita income showed no tendency to be affected by the three-fold grouping of the British Counties. Any further analysis could proceed safely over the whole sample of Counties.

* Covariance Analysis is hardly a new technique to geographers, being used in the mid-50s by Bogue and Harris.

Moving from a hypothetical to an actual example will perhaps demonstrate the usefulness of covariance analysis to the geographer.

During current research by the author on spatial variations of activity rates in Britain, several multiple regressions were run, using the female activity rate as the dependent variable and with several combinations of independent variables. The areal base was the 61 Department of Employment sub-divisions. With several sets of independent variables, an R^2 value of over 70% resulted, obviously a fairly pleasing situation. However, the residuals from these regressions showed a high degree of spatial autocorrelation and since such high R^2 values did not suggest any obvious missing variables, it was postulated that some of the independent variables were having different effects in different parts of the country.

Consequently, a grouping programme⁵ was applied to the scores of each sub-division on the first four principal components of the input data (accounting for some 75% of its variance) until five groups of sub-divisions had been defined.

Using the five clusters of sub-divisions as treatment groups, several different combinations of independent variables (covariates) were input to a covariance analysis programme, together with the dependent variable (variate). In eighteen out of the twenty runs performed in this way, significant differences were found between the treatment groups either in slope coefficients or adjusted Y -means.

Further investigation of the regression equations at "group" level showed that in some cases considerably different factors were influencing female activity rates than was apparent from the analysis over all the 61 sub-divisions. For example, an employment structure variable which was very powerful in explaining variations in female activity rates at the national level, had very little influence in the south of the country. In general, covariance analysis provided a great deal of information on the relationships between female activity rates and the set of independent variables which would have otherwise been subsumed in the "national" analysis.

* This maximised the number of cases in the smallest group and also occurred at a major break in slope of the within group/between group distances ratio.

Table 1 shows the widely-differing results obtained when the 1968 Female Activity Rate was regressed on one of the combinations of independent variables over each group of sub-divisions. When this particular regression was carried out over the whole sample of sub-divisions, 59.5% of the variance in female activity rates was explained. However, R^2 values of between 12.1% and 79.65% resulted when the regression was run for each of the sub-groups. A covariance analysis carried out previously had shown significant slope differences among the treatment groups for this combination of variables.

TABLE 1

GROUP	INTERCEPT	B-COEFFICIENTS			$+R^2$
		X_1	X_2	X_3	
Y_1	0.16	11.67**	0.17	0.02	79.65%
Y_2	9.74	5.92	0.19	0.23	18.30%
Y_3	6.86	8.48**	0.00	0.24*	70.40%
Y_4	23.0	7.82	-0.25	-0.10	12.10%
Y_5	-7.06	6.84**	1.93**	0.02	78.11%
ALL	6.23	6.04**	0.69**	0.16**	59.53%

** - Significant at 99%. * - Significant at 95%. + - Adjusted for sample size.

X_1 = Population Density, 1968 (logged).

X_2 = % of Males in Socio-Economic Groups 1 to 4, 1966.

X_3 = % of Female employment in manufacturing accounted for by high female-employing industries, 1968.

Covariance analysis, then, provides an important supplement to conventional regression analysis by testing in effect whether the observation base chosen (e.g., sets of nations, regions, sample points) is affecting the relationship between dependent and independent variables and, alternatively, whether more effective analysis could be performed using different bases. Its potential use in geography is relatively unrestricted - whenever two or more interval scale variables are being considered on two or more nominal scales, covariance analysis may be relevant. Thus, the technique may be shared - e.g., by the biogeographer investigating the relationship between soil depth and pH on several different lithologies, and the economic geographer studying the suburbanization of population in different national or regional contexts.

It was noted in the opening paragraph that examples of the use of covariance analysis in geography were very thin on the ground. Since the pioneering work in this field by Bogue and Harris and Kitagawa and Bogue in the mid-50s there have been very few instances of the use of the technique. Dogan used covariance analysis in his study of French election data and King¹⁰ employed the technique in his analysis of the spacing of urban settlements in U.S.A. These references and a few others are classified by Greer-Wootten¹¹. Two examples of the use of covariance analysis in research on the journey to work and the perception of the public transport system in Dublin are forthcoming from O'Farrell and Markham^{12,13}.

The paucity of examples suggests that there may be particular problems associated with the use of covariance analysis. This is to some extent true. Firstly, it is most effective when a large number of cases is involved so that subdivision of the original sample will not result in a calamitous loss in degrees of freedom. Secondly, in common with all techniques which use the F-distribution as a significance test, a multivariate normal distribution is required for most rigorous use. Thirdly, it is a technique which tends to be costly in computer time and storage.

Notwithstanding these problems, analysis of covariance is a technique which seems to have been largely overlooked by the geographer and it is hoped that this paper has demonstrated its intrinsic value sufficiently to promote it from the Cinderella status it occupied among research techniques currently in use by geographers.

NOTES

1. Bogue, D.J. and Harris, D.I. (1954) *"Comparative Population and Urban Research via Multiple Regression and Covariance Analysis"*. Scripps Foundation, Chicago: Population Research and Training Centre.
2. Tatsuka, M.M. (1971). *Multivariate Analysis: Techniques for Educ. and Psych. Research*". Chapter 5 John Wiley.
3. Blalock, H.M. (1960) *"Social Statistics"*, Chapter 20, pages 359-382, McGraw-Hill.
4. *Ibid.*
5. "Master Cluster" - a programme originally appearing in Hitchin, D. (1969) *"Cluster Analysis"*, Mimeo, University of Sussex.
6. BMD X82 in *"Biomedical Computer Programmes"*, X-Series Supplement, W.J. Dixon (Ed.) University of California Press.
7. Bogue and Harris, *op. cit.*
8. Kitagawa, E. and Bogue, D.J. (1955), *"The Suburbanisation of manufacturing Activities within Standard Metropolitan Areas."* Scripps Foundation, Chicago.
9. Dogan, M. (1969), "A Covariance Analysis of French Election Data" in Dogan, M. and Rokkam, S. (Eas.). *"Quantitative Ecological Analysis in the Social Sciences"*. Chapter 11, pages 285-298, M.I.T. Press.
10. King, L.J. (1961) *"Multivariate Analysis of the Spacing of Urban Settlements in the U.S."*. A.A.A.G. Vol. 51 pages 222-233.
11. Green-Wootten, B. (1972), *"A Bibliography of Statistical Applications in Geography."* A.A.G. Technical Paper, No. 9.
12. O'Farrell, P.N. and Markham, J. (forthcoming). *"The Journey to Work: A Behavioural Approach"*.
13. O'Farrell, P.N. and Markham, J. (forthcoming). *"Commuter Perceptions of the Public Transport System."*