ANALYSIS OF COVARIANCE
AND COMPARISON OF REGRESSION LINES

J. Silk

CATMOG

20

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

ANALYSIS OF COVARIANCE AND COMPARISON OF REGRESSION LINES

by

John Silk

(University of Reading)

CONTENTS

INTRODUCTION

The *analysis of covariance* brings together features of both the
analysis of variance and *regression analysis,* and is closely allied with
techniques for the *comparison of regression lines.*

Two topics will be discussed in this introduction. As a prerequisite,
we outline the general features of the analysis of variance and regression
analysis, considering particularly the levels of measurement associated
with each. Then, we consider the purposes for which the analysis of
covariance and comparisons of regression lines may be used.

(i) Prerequisites

Classically, the *analysis of variance* is concerned with the estimation
and comparison of the mean values of a *response* or dependent variable
under different conditions. Each 'condition' is represented by a *class* or
category in the analysis. Because we may regard the value of any given
observation of the response variable as (at least in part) *dependent* upon/
the category or class in which it falls, the categories themselves may be
regarded as 'values' or 'levels' of an *independent* or *explanatory* variable.

Such 'values' are said to be measured on a *nominal* scale if there
is no intrinsic ordering of the categories. This might be the case if
the explanatory variable were aspect, as represented by the categories
'north-facing' and 'south-facing', or region, as represented by a number of
physiographic zones or administrative districts. An intrinsic ordering
of the categories, as in the case of socio-economic status, age or income
groups, is said to represent measurement on an *ordinal* scale. For example,
census volumes present much information according to the eight age groups
0 - 4, 5 - 14, ... 55 - 64, 65 or over. If these groups are numbered
1 to 8 respectively, then the numbers may be taken to represent the
ranking or ordering of the groups. Thus, the 'values' or 'levels' of a
categorical explanatory variable may represent ordered or unordered
categories.

Regression analysis concerns itself with the relationship between a
dependent or *response* variable and one or more *independent* or *explanatory*
variables. Conventionally, the assumption is made that all variables,
whether dependent or independent, are measurable on an *interval* or *ratio*
scale. Both scales allow us to determine the distance along the scale
between the attributes of any two individuals or events, as well as their
rank order. Measurement based on a ratio scale is taken to be at a higher
level than that based on an interval scale, because the former provides
measurements that can be referred to a natural origin or absolute zero
point, whereas the latter cannot. The Fahrenheit and Centigrade temperature
scales are examples of interval scales. A reading of 0 Centigrade corres-
ponds to 32 F, and of 0 Fahrenheit to -17.8 Centigrade, showing that the
zero points are arbitrarily determined in each case. Many of the variables
used in geographical research are measurable on a ratio scale, for example
distance, altitude, population and income.

In addition, it is usually assumed that the response or dependent variable is *continuous* (i.e. it can assume an unlimited number of inter-mediate values), and that the same holds for all independent or explanatory variables measured on an interval or ratio scale.

The case in which the dependent or response variable is categorized is not discussed here, but a detailed treatment may be found in another monograph in the CATMOG series (Wrigley, 1976). A technique such as the *analysis of covariance,* therefore, which combines features of the analysis of variance and regression analysis, is concerned with a response or dependent variable, measured on an interval or ratio scale, and a set of explanatory variables, of which some may be measurable on an interval or ratio scale, and others on a nominal or ordinal scale.

Familiarity with the *one-way analysis of variance model,* assuming *fixed-effects,* and with the theory underlying regression analysis, is essential, and brief reviews of both kinds of model are provided in Section II of this monograph. Many of the concepts and principles involved have also been discussed by Unwin (1975) and Ferguson (1978) in earlier contributions to the CATMOG series.

(ii) Purpose

The techniques described in this paper enable the investigator to:

1) Introduce an additional independent variable, measured on a ratio or interval scale, to adjust mean values already estimated according to the analysis of variance model.

2) Compare two or more bivariate regression lines, assuming that the variables included in the regression model are the same in each case.

The value of the procedure described in 1) becomes apparent if we briefly consider a hypothetical example (to be studied in detail in Section V). Suppose that a study is carried out to establish whether travel behaviour of the elderly differs from that of the rest of the population in a parti-cular city. Information is obtained by stratifying households into elderly and non-elderly groups, randomly selecting a number of households from each group, and asking a member of each household to keep a record of all house-hold daily travel behaviour for four weeks. One crude index of travel be-haviour is the *average number of visits made per week to the city centre,* which may be regarded as the response variable, Y. It might be found that the mean trip rate for the sample of elderly households $\bar{Y}_1$, is less than that for the non-elderly, $\bar{Y}_2$. Such a result is to be expected. However, considering the locations of households in terms of their *distance from the city centre,* X, we may find that elderly households tend to live further away from the city centre, on average, than non-elderly households. Assuming that trip rates decline with distance, it could well be that the elderly households make fewer trips *not because they are old and instrinsi-cally less mobile, but because they happen to live slightly further away from the city centre than do households in other groups.* It is desirable to isolate the influence of distance so that the difference between the two groups could be examined after allowing for this factor. Logically, the value of $\bar{Y}_1$ should be increased, and that of $\bar{Y}_2$ decreased, by appropriate amounts. Under certain conditions, described in Section V, such adjustments may be made using the analysis of covariance.

4
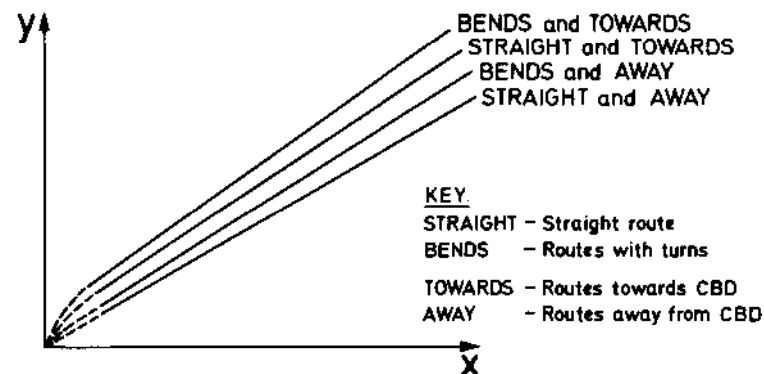
Figure 1. Hypothesized relationships between cognitive distance (Y) and objective road distance (X) (redrawn from Briggs, 1973).

The above provides an example of *confounding,* in which the influences of two or more variables on some phenomenon are difficult, or sometimes impossible, to disentangle. In this case one variable is categorical (elderly *or* non-elderly) and the other measurable on a ratio scale in terms of distance.

A number of reasons may be put forward for comparing regression lines:

a) To test *a priori* hypotheses. For instance, Briggs (1973) reasoned that distance estimates made by students would be more exaggerated in the direction of the town centre than away from it, and along routes involving bends than routes that did not. The rankings of the slope terms with respect to the four cases are illustrated in Figure 1.

b) As an aid to parsimonious description and explanation. It may be found that the values taken on by the slope terms of a number of regression lines are not significantly different. This may well be the case in Figure 2, which shows separate regression lines for north-and south-facing Arctic slopes of thickness of the active layer (Y) on daily radiation total (X) (Hannell, 1973). If it is found that the two slopes do not differ significantly, then the initial model based on two slopes and two intercepts may be reduced to another based on one slope and two intercepts. Opportunities for simplification are particularly evident if, say, a study of the relationship between two variables X and Y is conducted in ten different regions.

c) To obtain better estimates. This characteristic may be associated with the simplification process described above. A slope term of a regression relationship, for instance, may be estimated more accurately from two sets of observations than from one.

If we are not particularly interested in the precise form of the relationship between two or more variables, or in estimation problems in general, the various influences may be 'sorted out' with the aid of an analysis based on partial correlation coefficients (Weatherburn, 1962, 256-257; Blalock, 1964; Taylor 1969). Discussion of methods for explicitly

5

Figure 2. Relationship between thickness of the active layer (Y) and daily
radiation total (X) on north- and south-facing slopes (adapted
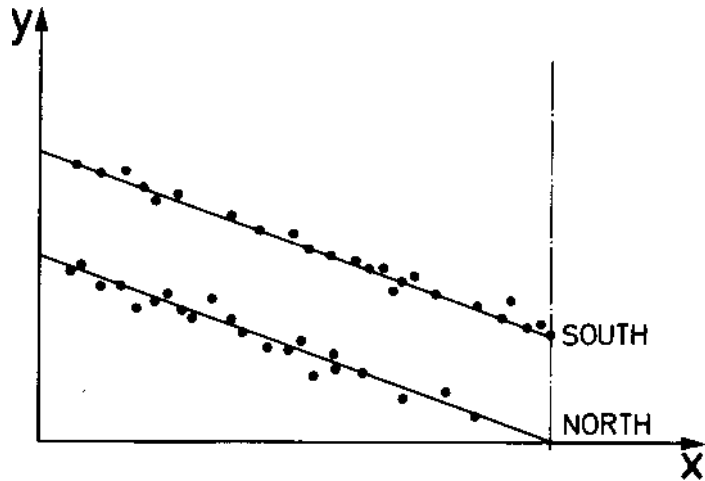from Hannell, 1973).

controlling for the influence of a third variable may be found in Davidson
(1976) and Ferguson (1978).

II THE ANALYSIS OF VARIANCE AND LINEAR REGRESSION MODELS

(i) The analysis of variance model

Suppose we are interested in the *mean* values of a response variable,
Y, in three different regions represented by the categorical explanatory
variable, A. Let the true mean values of Y for each region be represented
by the population parameters μ_1, μ_2, and μ_3, and the 'grand mean' or 'overall
mean' of measurements in all three regional populations by μ. Letting α_i
represent the difference between the overall mean, μ, and the ith regional
mean μ_i:

$$\alpha_i = \mu_i - \mu \qquad\qquad (1)$$
$$i = 1, 3$$

Rearrangement of (1), and study of Figure 3(a) shows that we may express
each regional mean as a deviation from the overall mean:

$$\mu_i = \mu + \alpha_i \qquad\qquad (2)$$
$$i = 1, 3$$

The deviation α_i is known as the *effect* of being in the ith region, compared
with the overall mean μ.
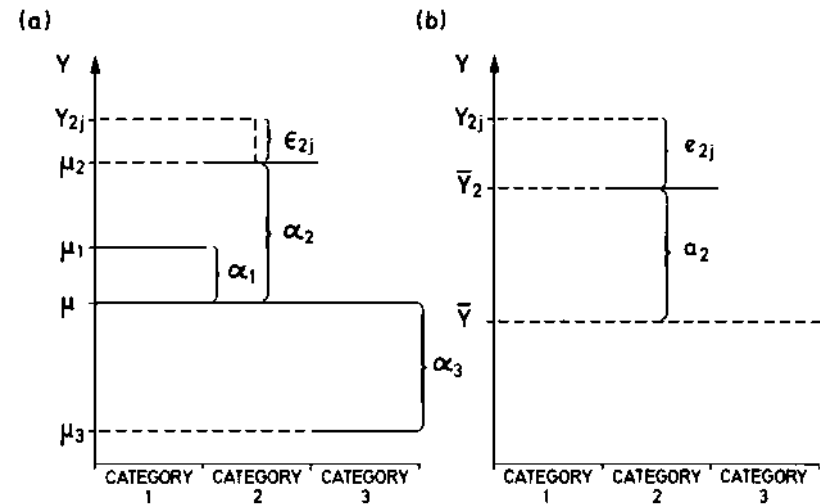
Figure 3. The analysis of variance model a) Parameters b) Estimates.

The value of the response variable, Y_{ij}, associated with any given
observation in the ith group, may be expressed as its deviation, ε_{ij}, from
the mean of the ith group. Since $\varepsilon_{ij} = Y_{ij} - \mu_i$:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

and, substituting for μ_i from (2):

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad\qquad (3)$$

so than an individual value or score may be represented in terms of two
additive deviations from μ, the first denoting the *effect* of the region
or group in which it fall (α_i), the second summarizing the combined influence
of all the relatively minor factors (or so it is hoped) which happen to
affect it (ε_{ij}). ε_{ij} is known as the population *error* or *disturbance* term.

Certain assumptions are made about the behaviour of the terms on
the right hand side of (3):

1) The true or population value of the mean for each region or group
is *fixed or constant*. It immediately follows that the values of
the overall population mean and of the effects must also be fixed.
Thus μ, the μ_i, and the α_i are *non-random* or *non-stochastic* elements.

2) The (weighted) sum of the effects is zero i.e. $\Sigma w_i \alpha_i = 0$, where
w_i is generally taken to be the number of observations in the ith
group, n_i. This assumption provides a useful check on calculations.

The remaining assumptions apply to the ε_{ij} term:

3) The ε_{ij} represent values of a *random* or *stochastic* variable, each

6

7

with expected value $E(\varepsilon_{ij}) = 0$.

4) The ε_{ij} are *mutually uncorrelated*, implying no correlation of such values either within or between groups or regions. This stipulation also applies where the ε_{ij} show a spatial or mapped distribution (Cliff and Ord, 1973).

5) For each group or region, the ε_{ij} are *normally distributed* with *equal variance*, σ^2. The latter assumption is also known as that of *homoscedasticity*.

Provided these assumptions are fulfilled, at least approximately, the *least squares estimators* of the population parameters have highly desirable properties, being *unbiased* and of *minimum variance* (Unwin, 1975, 19-20) (Table 1). The estimates themselves i.e. the actual numerical values yielded by applying the estimators in any particular investigation, will not generally equal the values of the unknown population parameters. This means that the *observed* or *sample* residual values, e_{ij}, will differ from those of the corresponding and unknown population disturbance terms, ε_{ij} (compare Figures 3a and 3b).

The investigator may wish to test certain hypotheses relating to the parameters, and approximate fulfillment of assumptions 3, 4 and 5 ensures that tests of significance based on the t and F distributions may be employed.

ii) The linear regression model

As for the analysis of variance model, we are interested in estimating the values of a response or dependent variable, Y. This time, however, we assume that changes in the values of some *non-categorical* explanatory variable, X, measured on an interval or ratio scale, *cause* or *produce* a change in the values of Y. If the assumption of causality seems too strong or inappropriate, we may instead speak in terms of changes in the value of Y which are associated with changes in the value of X.

The relationship between the two variables - only the *bivariate* case is considered here - is assumed to be linear, and may be expressed:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \qquad (4)$$

where α is the *intercept* or *constant* term, β the *slope* term, and ε_i a population error or disturbance term. The *population regression line,* showing the values of Y corresponding to all possible values of X within the range covered by the abscissa in the diagram, is represented by the solid line in Figure 4.

Apart from that of *linearity,* many of the assumptions underlying regression analysis are similar to those already made with respect to the analysis of variance: *fixed* population values of the intercept and slope terms, and of the values of X (the latter should also be measured with negligible error); the ε_i should be uncorrelated, and the equal variance assumption implies that their distribution should show the same spread about the regression line for all values of X.

Estimates of the population parameters are obtained from the sample observations using the least-squares estimators from Table 2 in the equation:

Table 1. Least squares estimators for the analysis of variance model

Population parameter	Estimator	
	Formula	Description
μ	$\bar{Y} = \frac{1}{N}\sum_i\sum_j Y_{ij}$	Overall sample mean
μ_i	$\bar{Y}_i = \frac{1}{n_i}\sum_j Y_{ij}$	Group or regional sample mean
α_i	$a_i = \bar{Y}_i - \bar{Y}$	Estimated effect of being in ith region or group

Error or disturbance term		
ε_{ij}	$e_{ij} = Y_{ij} - \bar{Y}_i$	Sample residual

i = regional or group membership

n_i = number of sample observations in region or group i

j = position of observation within region or group

N = total number of sample observations i.e.
$N = \sum_i n_i$

$$\hat{Y}_i = a + bX \qquad (5)$$

where Y_i denotes an estimate of Y_i and a and b are estimates of α and β respectively. The estimated or *sample regression line* usually differs from that in the population, and so therefore do the *sample residual values,* e_i, from the ε_i (Figure 4).
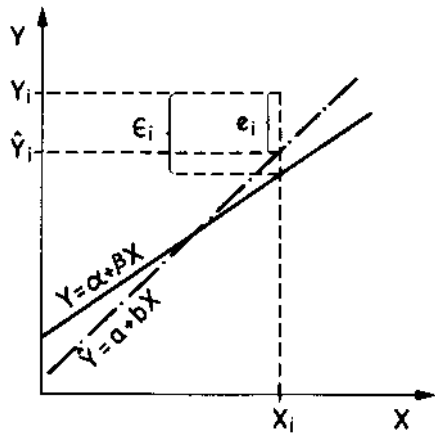
Figure 4. The linear regression model - parameters and estimates

Table 2. Least squares estimators for the bivariate linear regression model

Population parameter		Estimator	
		Formula	Description
β		$b = \dfrac{\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)/N}{\Sigma X_i{}^2 - (\Sigma X_i)^2/N}$	Sample estimate of slope term
α		$a = \bar{Y} - b\bar{X}$	Sample estimate of intercept term
Y_i		$\hat{Y}_i = a + bX_i$	Sample estimate of (mean) value of Y corresponding to X_i
Error or disturbance term			
ε_i		$e_i = Y_i - \hat{Y}_i$	Sample residual

Assumptions regarding the population disturbance term are based on the notion that any of a range of individual values of Y may be associated with a given value of X, so that the population regression line may be regarded as the locus of the *mean* values of Y corresponding to all possible

intermediate values of X within a given range. This view further emphasises the similarity with the analysis of variance model, i.e. estimation of mean values, and makes the adjustment of mean values based on regression techniques seem more 'natural' when the procedure is carried out. (N.B. The distinction between Y_i as an estimate of an individual value of Y or of a *mean* value of Y can usually be ignored, unless confidence intervals are to be constructed about the regression line Draper & Smith, 1966, 21-24.)

The form of the *analysis of covariance* model is given in Figure 5(b), and shows the influence of a categorical variable A, because there is a regression line for each of three (hypothetical) categories, combined with that of the continuous explanatory variable X. Apart from those already outlined, additional assumptions of the analysis of covariance model are

1) The regression lines for each category must be of equal slope or parallel.

2) The population disturbance terms for the observations in each category should show equal variance about each of the parallel regression lines.

3) Differences between the mean values of the covariate, X, for each category of the categorical variable A, should be 'relatively small'.

The implications of these assumptions will become clear as the discussion proceeds.

III SPECIFICATION OF MODELS FOR DUMMY VARIABLE ANALYSIS

i) Informal description of models

Before formally specifying models, it is helpful to describe the links between the analysis of variance (ANOVA), the analysis of covariance (ANCOVA), and the general process of comparing regression lines.

Consider the horizontal lines representing the means of the dependent variable Y for each of three sets of observations in Figure 5(a). This pattern of observations is one for which the three means may be compared using ANOVA. Notice that ANOVA may be regarded as analysis of values of a dependent variable, ignoring association with variables other than those whose influence may be summarised by grouping the observations. In Figure 5(a), the 'ignored variable' is X.

The residual deviations of the observations about their respective means may be scrutinised in the same way as residuals about regression (Ferguson, 1978, 14-15). We are particularly concerned with the case in which a trend within the residuals is suspected when plotted against some continuous independent variable or *covariate*, X. Figure 5(a) clearly shows such a trend for the observations within each group.

The simplest model incorporating a covariate is the analysis of covariance (ANCOVA) model, which assumes the same slope within each of the three sets of residuals (Figure 5(b)). Notice that the intercept terms of the regression lines are permitted to vary between groups, but not their slopes. In this paper, the ANCOVA model will also be known as the *Parallel model,*

(c) FULL

$\hat{Y} = (a+a_3) + (b+b_3)X$

$\hat{Y} = a+bX$

$\hat{Y} = (a+a_2) + (b+b_2)X$

(e) INTERCEPT

$\hat{Y} = a+(b+b_3)X$

$\hat{Y} = a+(b+b_2)X$

$\hat{Y} = a+bX$

(b) PARALLEL (ANCOVA)

$\hat{Y} = (a+a_3)+bX$

$\hat{Y} = (a+a_2)+bX$

$\hat{Y} = a+bX$

(d) JOINT

$\hat{Y} = a+bX$

(a) ANALYSIS OF VARIANCE (ANOVA)

$\hat{Y} = a+a$

$\hat{Y} = a+a$

$\hat{Y} = a$

Key
○ Category 1
▲ Category 2
■ Category 3
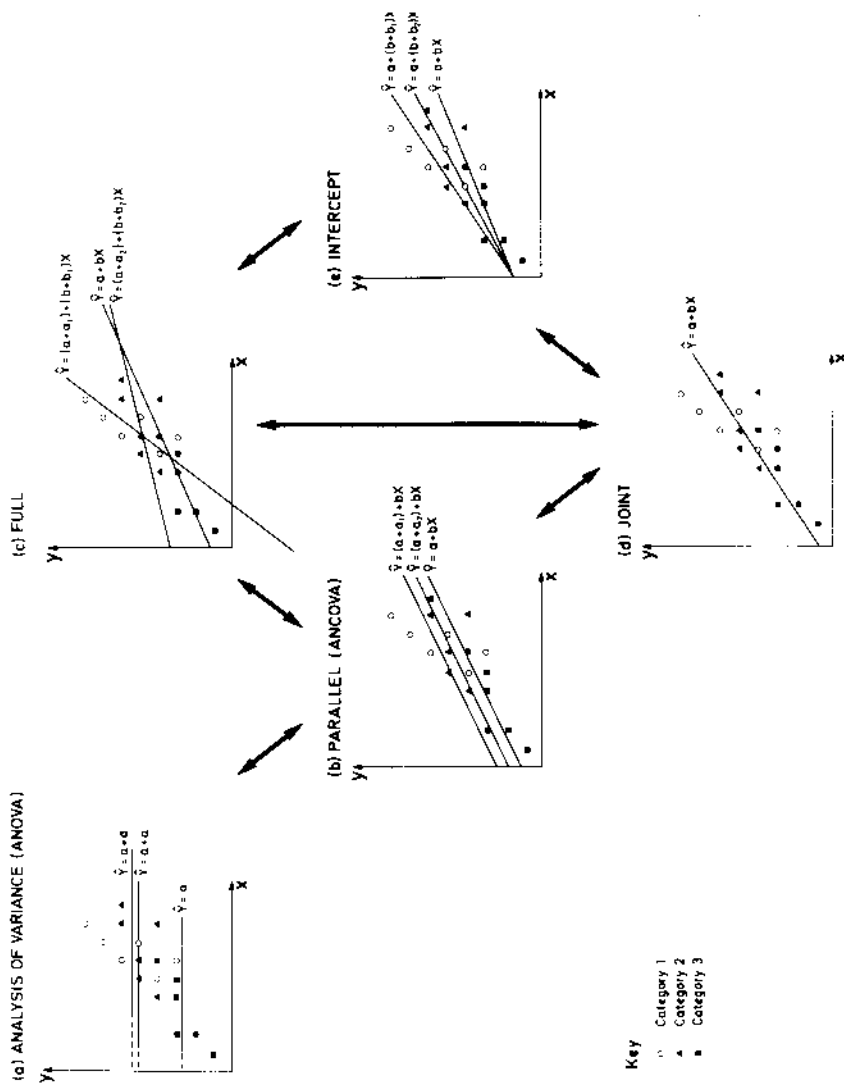
Figure 5. Relationships between models

as it represents one of four basic types of regression model which may be fitted to a set of observations. The *Intercept model* allows slopes to vary but not intercepts (Figure 5(e))   and the *Full model* leaves both slopes and intercepts unconstrained so yielding three separate regression lines as in Figure 5(c). If the classification of observations is thought to be totally irrelevant i.e. the categorical variable A is deleted, then both slope and intercept differences may be suppressed as in the Joint *model* (Figure 5(d)).

The Full model is the most complex, and simplification occurs as we move towards either the ANOVA model or the Joint model. The heavy arrows in Figure 5 show the paths along which statistical comparisons may be made.

ii) <u>Formal specification of models in dummy variable analysis</u>

<u>Analysis of variance (ANOVA)</u>

To use dummy variable analysis, the ANOVA model of (3) should be restated as:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} \qquad (6)$$
$$i = 1,k$$

in which the effect of being in the ith group is denoted by α_i. We may now respecify the ANOVA model so that each observation is expressed in terms of its deviation from the population mean of one of the groups which will be called the *anchor group*. Assuming the mth group serves as anchor:

$$Y_{mj} = \mu_m + \varepsilon_{mj}$$
$$Y_{ij} = \mu_m + \alpha_i + \varepsilon_{ij} \qquad (7)$$
$$i \neq m$$

where μ_m, μ_i are the population means of groups m and i, and α_i is the *difference* between the population means of groups m and i. In general, the α_i of (3) and the α_i of (7) will not be equal. However, the ε_{ij} are not affected, and nor therefore are the between-group and residual sums of squares (SS) or their estimates obtained from the dummy variable formulation, being the same as those provided by classical ANOVA techniques (Goldberger, 1964, 227-231). This applies to estimates obtained for the other models discussed in this section.

To carry out the analysis, assign values to observations on a set of k-1 dichotomous or dummy variables, D_i (i=1, k-1; i≠m), where k is the number of groups involved, so that:

$$D_{ipj} = \begin{cases} 1 & \text{if } p = i \\ 0 & \text{otherwise} \end{cases}$$

$$p=1,k-1; p \neq m$$

In Table 3, it is assumed k=3, m=1 so that only D_2 and D_3 are required (the choice of anchor group is quite arbitrary).

Table 3. Specification of models and tabulation of observations for dummy variables analysis

Y	X_0	D_1	D_2	D_3	X	D_1X	D_2X	D_3X	$(D_1X + D_2X)$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Y_{11}	1	1	0	0	X_{11}	X_{11}	0	0	X_{11}
Y_{12}	1	1	0	0	X_{12}	X_{12}	0	0	X_{12}
Y_{13}	1	1	0	0	X_{13}	X_{13}	0	0	X_{13}
Y_{14}	1	1	0	0	X_{14}	X_{14}	0	0	X_{14}
Y_{21}	1	0	1	0	X_{21}	0	X_{21}	0	X_{21}
Y_{22}	1	0	1	0	X_{22}	0	X_{22}	0	X_{22}
Y_{23}	1	0	1	0	X_{23}	0	X_{23}	0	X_{23}
Y_{24}	1	0	1	0	X_{24}	0	X_{24}	0	X_{24}
Y_{31}	1	0	0	1	X_{31}	0	0	X_{31}	0
Y_{32}	1	0	0	1	X_{32}	0	0	X_{32}	0
Y_{33}	1	0	0	1	X_{33}	0	0	X_{33}	0
Y_{34}	1	0	0	1	X_{34}	0	0	X_{34}	0

The 'variable' X_o represents the usual intercept or constant term of a regression model *and*, in the case of dummy variable analysis, membership of group m, the anchor group.

The model now becomes

$$Y_{mj} = \mu_m + \varepsilon_{mj}$$

$$Y_{ij} = \mu_m + \alpha_i D_i + \varepsilon_{ij} \tag{8}$$

and, regressing Y on D_2 and D_3, we obtain an estimating equation of the form

$$\hat{Y} = a_1 + a_2 D_2 + a_3 D_3 \tag{9}$$

in which $a_1 = \bar{Y}_1, a_2 = (\bar{Y}_2 - \bar{Y}_1)$ and $a_3 = (\bar{Y}_3 - \bar{Y}_1)$ (Goldberg, 1964, 221).

For any particular observation, say the jth observation in group 2:

$$Y_{2j} = a_1 + a_2 D_{22j} + a_3 D_{23j} = a_1 + a_2$$

because, by definition, $D_{22j} = 1$ and $D_{23j} = 0$. In general (Figure 5(a)):

$$\hat{Y}_{mj} = a_m$$

$$\hat{Y}_{ij} = a_m + a_i \tag{10}$$

$$i \neq m$$

provide estimates of the Y_{mj} and Y_{ij}, where a_m and a_i give estimates of μ_m and α_i respectively. From this point on, the notation in (10) will be used rather than that in (9).

Parallel or analysis of covariance (ANCOVA) model

The model is:

$$Y_{mj} = \mu_m + \beta X_{mj} + \varepsilon_{mj}$$

$$Y_{ij} = \mu_m + \alpha_i D_i + \beta X_{ij} + \varepsilon_{ij} \tag{11}$$

where μ_m now represents the intercept term of the anchor group, and α_i the *difference* in intercepts between the anchor group and group i.

A single independent variable, X, is added as shown in Table 3 (column 5). The regression equation takes the form:

$$\hat{Y} = a_1 + a_2 + a_3 + bX \tag{12}$$

typical observations being estimated by (Figure 5(b)):

$$\hat{Y}_{mj} = a_m + bX_{mj}$$

$$\hat{Y}_{ij} = a_m + a_i + bX_{ij} \tag{13}$$

$$i \neq m$$

It can be shown that estimates of adjusted means, normally obtained by other methods in the analysis of covariance (see Section V(i) for details of the adjustment procedure) may be obtained from the Parallel model by adding $b\bar{X}$ to the intercept term of the appropriate regression line (Silk, 1976, 10-11) so that:

$$\bar{Y}_{mA} = a_m + b\bar{X}$$

$$\bar{Y}_{iA} = a_m + a_i + b\bar{X} \qquad (14)$$

where $\bar{Y}_{mA}$ and $\bar{Y}_{iA}$ gives estimates of the adjusted means μ_{mA} and μ_{iA}.

Intercept Model

Assuming a common intercept for the regression lines in each group, but different slopes:

$$Y_{mj} = \mu + \beta_m X_{mj} + \varepsilon_{mj}$$

$$Y_{ij} = \mu + (\beta_m + \beta_i D_i)X_{ij} + \varepsilon_{ij} \qquad (15)$$

$$i \neq m$$

where μ represents the common intercept, β_m the slope term of the anchor group regression line, and β_i the *difference* in slopes between the regression lines of the anchor group and group i. The concept of an anchor intercept term is therefore carried over to accommodate slope differences. As shown in Table 3, multiplicative or 'interaction' variables $D_i X$ are specified for inclusion in the analysis, so that regression of Y on X, $D_2 X$ and $D_3 X$ yields:

$$\hat{Y} = a + b_1 X + b_2(D_2 X) + b_3(D_3 X)$$

and the estimates are given by (Figure 5(e)):

$$\hat{Y}_{mj} = a + b_m X_{mj}$$

$$\hat{Y}_{ij} = a + (b_m + b_i)X_{mj} \qquad (16)$$

$$i \neq m$$

Full Model

No constraints are placed on either slope or intercept differences in this case. The model includes the full complement of terms:

$$Y_{mj} = \mu_m + \beta_m X_{mj} + \varepsilon_{mj}$$

$$Y_{ij} = \mu_m + \alpha_i D_i + (\beta_m + \beta_i D_i)X_{ij} + \varepsilon_{ij} \qquad (17)$$

$$i \neq m$$

and estimates are obtained from (Figure 5(c)):

$$\hat{Y}_{mj} = a_m + b_m X_{mj} \qquad (18)$$

$$\hat{Y}_{ij} = a_m + a_i + (b_m + b_i)X_{ij}$$

$$i \neq m$$

Joint model

This is the simplest regression model which might be fitted to the observations, incorporating the constraints of both the Intercept and Parallel models so that slopes and intercepts of regression lines for all groups are identical:

$$Y_{ij} = \mu + \beta X_{ij} + \varepsilon_{ij} \qquad (19)$$

$$(i = 1, k)$$

All slope and intercept *difference* terms are omitted, so that regression of Y on X gives (Figure 5(d)):

$$\hat{Y}_{ij} = a + b X_{ij} \qquad (20)$$

IV STATISTICAL COMPARISON OF MODELS

Comparison may be carried out at two levels:

(i) Overall comparison between models. If any two of the models described in the previous section differ significantly in terms of the proportion of variation explained, the difference is attributed to some overall contrast between the models concerned.

(ii) Comparisons within models. For example, a significant difference between the mean values of observations in the groups of the ANOVA model may be established. However, further testing is required to establish which pairs of means differ significantly. Similarly, an overall difference between slopes of group regression lines might be detected (Intercept model) but it is quite possible that not all groups are mutually distinct.

The difference between (i) and (ii) partly reflects the general use of the F test as an overall test which does not therefore pick out specific differences or 'contrasts'. It should also be noted that the 'between-within' distinction is partly one of convenience. Comparison of equations that differ by only one term (parameter) and testing 'within' models also involves comparison of models.

i) Comparisons 'between' Models

Possible comparisons are shown in Table 4 and models are described as 'more' or 'less' complex according to the number of parameters involved. The criterion of complexity (or rather, the other side of the coin, simplicity) is useful because it seems reasonable to search for models which provide the best possible fit to the data but incorporate as few parameters as appear statistically necessary.

Table 4. Illustration of comparisons between models

		Increasing Complexity				
		J	A	I	P	F
$\hat{Y} = a + bX$	JOINT	NA	X	✓	✓	✓
$\hat{Y} = a + a_1 + a_2$	ANOVA	-	NA	X	✓	✓
$\hat{Y} = a + bX + b_1X + b_2X$	INTERCEPT	-	-	NA	X	✓
$\hat{Y} = a + a_1 + a_2 + bX$	PARALLEL	-	-	-	NA	✓
$\hat{Y} = a + a_1 + a_2 + bX + b_1X + b_2X$	FULL	-	-	-	-	NA

✓ = Permissible comparison

NA = Not applicable

X = No direct statistical comparison

Overall comparison of models therefore normally involves asking not simply whether one additional variable produces a significant increase in the multiple R^2 value, but whether several additional variables are worth including en bloc. Such comparisons will be familiar to geographers who have used trend surface analysis (Unwin, 1975, 22). The variance ratio

$$F_{k_2-k_1, N-k_2} = \frac{(R_2^2 - R_1^2)/(k_2 - k_1)}{(1 - R_2^2)/(N - k_2)} \quad (21)$$

may be used to compare a more complex model, model 2, and a less complex model, model 1, fitted to the same data and dependent variable (it is also assumed that the more complex model includes all the parameters of the less complex model). Thus R_2^2 and R_1^2 are the proportions of variation explained by the more and less complex models respectively, k_2 and k_1 the number of independent variables, *including* the constant or anchor category term, and N the total number of observations. If the ratio exceeds the tabulated value of the F distribution (one-tailed test) with $k_2 - k_1$ and $N - k_2$ degrees of freedom at a prespecified significance level such as 5%, then the difference in the proportion of variation explained by the two models merits further investigation.

If two models differ by only one term or parameter, the variance ratio reduces to

$$F_{1, N-k_2} = \frac{R_2^2 - R_1^2}{(1 - R_2^2)/(N - k_2)} \quad (22)$$

Furthermore, (22) should be used to test the overall significance of a bivariate regression (i.e. the test of $\beta = 0$), or for the overall significance of an analysis of variance for which there are only two categories, setting $R_1^2 = 0$ in each case.

ii) Sequence of Comparisons

It is important to make comparisons in the right order to establish the relative importance of slope and intercept differences. Assume the Joint model yields a highly significant F value for regression, and that comparison of the Joint and Full models reveals a highly significant difference between the two. In order to track down the source of this difference, two routes involving comparisons may now be followed.

Following the 'Intercept Route', the order of comparison is first Joint-Intercept, then Intercept-Full. Using the F test to make the Joint-Intercept comparison, a significant increase tells us that allowing slopes alone to vary is worthwhile. The Intercept-Full comparison tells us whether it is worth differentiating intercepts, given that slopes have already been permitted to vary. Taking the 'Parallel Route', the order of comparisons is Joint-Parallel, Parallel-Full.

Study of the information obtained via both routes puts us in the same position with respect to blocks of variables as someone evaluating the contribution of each individual variable to variation explained in a multiple regression equation. Individual contributions are conventionally assessed as if each variable had entered the regression equation last (Draper and Smith, 1966, 71-72). The Intercept and Parallel Routes provide 'last entry' information on the blocks representing intercept and slope differences respectively. Given a significant Joint-Full difference, an insignificant Intercept-Full comparison implies adoption of the Intercept Model, and an insignificant Parallel-Full comparison adoption of the Parallel model.

Effects of slope and intercept differences may each be found significant if entered last, implying acceptance of the Full Model and thus fitting of separate regression lines within each class. However, due to correlation between each set of effects, it is possible that neither would be judged significant if entered last. Various options are open here, including comparison within models if theoretical considerations, or a scatter diagram of the data, suggest it.

For the analysis of covariance the recommended sequence of comparisons is ANOVA vs. Parallel, Parallel vs. Full, and Parallel vs. Joint. The reasons for these comparisons, and their ordering, are discussed where they arise in the worked example of the analysis of covariance in Section V.

As neither intercept nor slope values are constrained by the Full model this amounts to saying that the regression lines fitted by this method have exactly the same intercept and slope values as lines fitted quite independently. If many groups are involved, more accurate estimates of parameters and variances will be obtained if the regression equation for each group is computed separately. Also, values of R^2 and of the standard error of estimate (s) can be obtained separately for each equation, and such information is often useful as we show in Section VI. The computationally inefficient version of the Full model is described above because its structure may be more readily compared with the other models specified in the previous section.

A further point is that tests of significance on the regression coefficients will yield different results if incorporated in the Full model rather than in separate equations, because the residual or error variance is based on pooled data in the case of the Full model, but on separate sets of observations otherwise. However, the pooled error variance is required when comparing the Full model with other models.

iv)Comparisons 'within' Models

As stated earlier, a significant overall difference between any two of the five models may be detected, but the exact source (or sources) cannot be identified using the F test. In tackling this problem, we simultaneously deal also with a technical difficulty arising because the choice of an anchor category is arbitrary (see above, P.13). Considering differences between means of groups as shown in Figure 5(a) it should be clear that choice of class 3 as the anchor category should yield regression coefficients a_1 and a_2, representing differences between the means of classes 3 and 1 and 3 and 2 respectively, which are likely to be judged significant. If classes 1 or 2 were selected as base, however, a significant difference between groups 1 and 2 is not to be expected. In analysis of variance parlance, we can only carry out tests directly on certain of the 'contrasts' because of the way the model has been set up. We may test all possible differences by obtaining estimates of standard errors from the variance-covariance matrix usually provided by a computer package program, according to the relation:

$$\text{est s.e. } (a_i - a_j) = \sqrt{\text{Var}(a_i) + \text{Var}(a_j) - 2\text{Cov}(a_i, a_j)} \qquad (23)$$

where 'est s.e.' denotes 'estimated standard error', 'Var' variance, and 'Cov' covariance. The desired t values are computed in the usual way from

$$t_{N-k_1} = \frac{a_i - a_j}{\text{est s.e.}(a_i - a_j)} \qquad (24)$$

where k_1 is the number of parameters (including the constant term) estimated for the model under consideration. A similar procedure is used to test for differences between slopes. To test whether the slope coefficient for a particular group, β_i^*, differs significantly from zero, we require the standard error of $b_i^* = b_m + b_i$, given by:

$$\text{est s.e. } (b_m + b_i) = \sqrt{\text{Var}(b_m) + \text{Var}(b_i) + 2\text{Cov}(b_m, b_i)} \qquad (25)$$

The usual (t) test against $\beta_i^* - 0$ is then carried out.

Great caution is required when interpreting tests based on comparisons 'within' models. An adequate framework presupposes sufficient knowledge to allow the investigator to specify beforehand the hypotheses to be tested, and hence those comparisons which are relevant. Where specific comparisons are not planned - perhaps because of the exploratory nature of the study - numerous inferential difficulties arise, as stated in Davis (1961) and Selvin and Stuart (1966). Many relevant papers are also listed in Snedecor and Cochran (1967, Ch. 10) and included in Kirk (1972, Section 8). Difficulties may not end here, but their treatment falls beyond the scope of this paper. For more detailed discussion of the issues in a geographic context see Hauser (1974) and Silk (1976).

V COMPARISONS IN THE ANALYSIS OF COVARIANCE - A WORKED EXAMPLE

In this section we give a simple worked example involving only two categories so that calculations may be easily followed. First, an analysis of covariance based on a conventional approach is outlined so that the logic of the adjustment procedure can be fully grasped. Then, we show how the same results may be achieved using package regression programs and dummy variable analysis. There are minor differences between the two sets of results, due to rounding error.

i) A conventional approach

Suppose a study of the relationship between average weekly trip rates to the city centre (Y) and distance from the city centre (X) has been undertaken, as suggested in the Introduction (p.4). Observations are divided into two categories, Group 1 representing the 'elderly' and Group 2 the 'non-elderly'. For simplicity, we assume an equal number of observations in each group (although this is *not* a prerequisite for analysis of covariance), with only ten households in each. These observations are shown in columns (1) and (2) of Table 5(a).

Initially confining our attention to the response or dependent variable, Y (Figure 6(a)), we carry out a one-way analysis of variance. The effect of being elderly appears to be to decrease the mean trip rate, compared with the overall mean, since $a_1 = -0.49$, and the reverse is true for the non-elderly, since $a_2 = 0.49$.[1] At the foot of Table 5(a), we give the formulae for the *total sum of squares* (TSS) and for its two components, the *between-group sum of squares* (BSS) and *residual sum of squares* (ResidSS). BSS may be regarded as the variation in Y 'explained' by the analysis of variance, just as the regression sum of squares represents the variation in Y 'explained' by an independent variable X in simple regression analysis. The deviations of the observations about their respective group means, shown in Figure 6(b), represent the individual elements making up ResidSS in this example. The diagram also shows how, in effect, an ANOVA model fits a horizontal line, at the level of the category sample mean, to the sample observations in each group. In this case, we find the proportion of variation explained is $R^2 = 0.3625$. (N.B. Conventionally, this quantity is denoted by eta in the analysis of variance, but R^2 is used here for the sake of consistency.) Testing this value for significance, we find:

Table 5. Hypothetical observations on average weekly trip rate (Y) and distance from city centre (X)

(a) Original sample observations

	(1) Y_{ij}	(2) X_{ij}	(3) Y^2_{ij}	(4) X^2_{ij}	(5) $X_{ij}Y_{ij}$
GROUP 1	4.2	1.3	17.64	1.69	5.46
(ELDERLY)	3.5	1.8	12.25	3.24	6.30
	3.2	2.1	10.24	4.41	6.72
$\bar{X}_1 = 2.38$	3.4	2.2	11.56	4.84	7.48
$\bar{Y}_1 = 2.94$	3.2	2.3	10.24	5.29	7.36
	2.6	2.5	6.76	6.25	6.50
	3.0	2.5	9.00	6.25	7.50
$a_1 = -0.49$	2.8	2.7	7.84	7.29	7.56
	2.0	2.9	4.00	8.41	5.80
	1.5	3.5	2.25	12.25	5.25
SUBTOTALS 1	29.4	23.8	91.78	59.92	65.93
GROUP 2	5.0	1.0	25.00	1.00	5.00
(NON-ELDERLY)	4.6	1.2	21.16	1.44	5.52
	4.5	1.4	20.25	1.96	6.30
$\bar{X}_2 = 1.87$	3.9	1.7	15.21	2.89	6.63
$\bar{Y}_2 = 3.92$	3.8	1.9	14.44	3.61	7.22
	3.6	2.1	12.96	4.41	7.56
$a_2 = 0.49$	3.5	2.2	12.25	4.84	7.70
	3.4	2.3	11.56	5.29	7.82
	3.7	2.4	13.69	5.76	8.88
	3.2	2.5	10.24	6.25	8.00
SUBTOTALS 2	39.2	18.7	156.76	37.45	70.63
GRAND TOTALS	68.6	42.5	248.54	97.37	136.56

$\bar{X} = 2.125$
$\bar{Y} = 3.430$

$TSS = \underset{ij}{\Sigma\Sigma}Y^2_{ij} - N\bar{Y}^2 = 248.54 - 20(3.43)^2 = 13.242$

$BSS = \underset{i}{\Sigma}n_i\bar{Y}^2_i - N\bar{Y}^2 = 240.10 - 20(3.43)^2 = 4.80$

$ResidSS = TSS - BSS = 13.242 - 4.80 = 8.442$

$R^2(ANOVA) = 0.3625$

(b) Transformed observations (translated to common mean or origin)

	(1) $y_{ij} = Y_{ij} - \bar{Y}_i$	(2) $x_{ij} = X_{ij} - \bar{X}_i$	(3) y^2_{ij}	(4) x^2_{ij}	(5) $x_{ij}y_{ij}$
	1.26	-1.08	1.5876	1.6664	-1.3608
	0.56	-0.58	0.3136	0.3364	-0.3248
	0.26	-0.28	0.0676	0.0784	-0.0728
	0.46	-0.18	0.2116	0.0324	-0.0828
	0.26	-0.08	0.0676	0.0064	-0.0208
	-0.34	0.12	0.1156	0.0144	-0.0408
	0.06	0.12	0.0036	0.0144	0.0072
	-0.14	0.32	0.0196	0.1024	-0.0448
	-0.94	0.52	0.8836	0.2704	-0.4888
	-1.44	1.12	2.0736	1.2544	-1.6128
	0.00	0.00	5.3440	3.2760	-4.0420
	1.08	-0.87	1.1664	0.7569	-0.9396
	0.68	-0.67	0.4624	0.4489	-0.4556
	0.58	-0.47	0.3364	0.2209	-0.2726
	-0.02	-0.17	0.0004	0.0289	0.0034
	-0.12	0.03	0.0144	0.0009	-0.0036
	-0.32	0.23	0.1024	0.0529	-0.0736
	-0.42	0.33	0.1764	0.1089	-0.1386
	-0.52	0.43	0.2704	0.1849	-0.2236
	-0.22	0.53	0.0484	0.2809	-0.1166
	-0.72	0.63	0.5184	0.3969	-0.4536
	0.00	0.00	3.0960	2.4810	-2.6740
	0.00	0.00	8.4400	5.7570	-6.7160

Slope of line fitted to superimposed residual values:

$b = \dfrac{\underset{ij}{\Sigma\Sigma}x_{ij}y_{ij}}{\underset{ij}{\Sigma\Sigma}x^2_{ij}} = \dfrac{-6.7160}{5.7570} = -1.167$
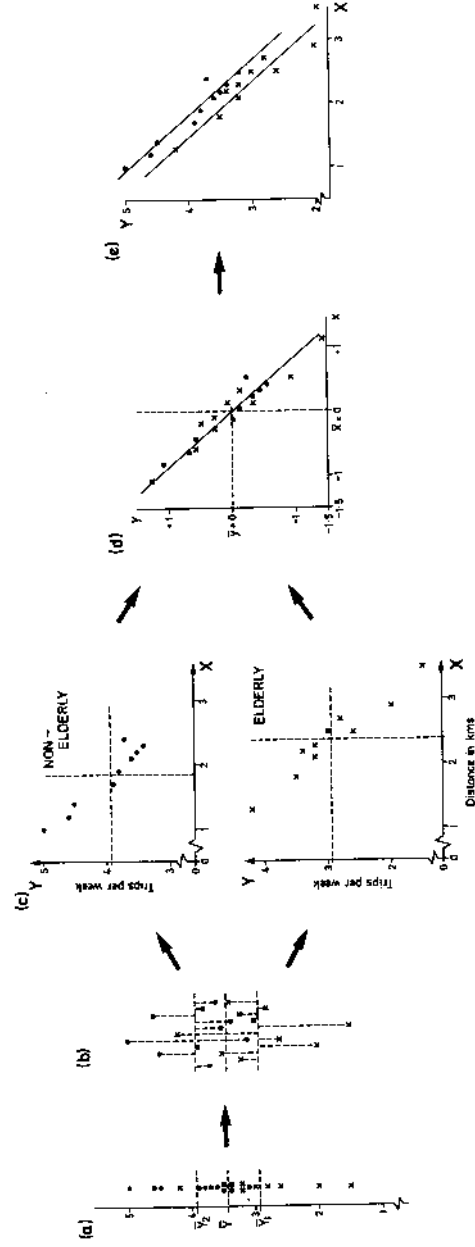
$\bar{Y}_{1A} = 3.24$ $\bar{Y}_{2A} = 3.62$

Figure 6. Steps in estimating the parameters of the analysis of covariance model.

$$F_{1,18} = \frac{0.3625 \times 18}{1-0.3625} = \underline{10.24}$$

a variance ratio significant at the 1% level.

Introducing a covariate. Because a high proportion of the total variation is still unexplained (0.6375 or 63.75%) the investigator will search for additional variables which significantly decrease this proportion and thus increase R^2. As in multiple linear regression analysis, one way is to find a variable which correlates with the sample residuals (e_{ij}) from the ANOVA model. The residual values are shown in Figure 6(b). If a continuous independent variable or *covariate* X is introduced then, as already noted, the effects of confounding a categorical variable with a continuous variable may be disentangled.

An obvious candiate for the role of covariate or 'confounding variable' in our worked example is *distance from the city centre*. If the residual values indicated in Figure 6(b) are now plotted, by group, against the appropriate distances obtained from column (2) of Table 5(a), it is quite clear that a strong negative trend exists within each set of residuals (Figure 6(c)).

The simplestmodification that may be made to the analysis of variance model by introducing a covariate is achieved if we assume that the trend within each set of residuals is the same. In our example (Figure 6(c)) this appears to be a reasonable assumption. The slope of a regression line fitted *only* to the residuals about the mean of Group 1 would not be very different from the slope of a regression line fitted *only* to the residuals about the mean of Group 2. Because of the desire to keep things as simple as possible, we first fit a line to *all* the residual values, under the assumption that the slopes for the two groups are not significantly different.

To combine or *pool* the two sets of observations, we superimpose them so that the bivariate means $(\bar{X}_1,\bar{Y}_1)$ and $(\bar{X}_2,\bar{Y}_2)$ coincide. We also ensure that the orientation of the patterns of observations remains unchanged with respect to the X and Y axes. It may be helpful to imagine each set of observations being plotted on a different transparent overlay, and the overlays then being moved, preserving orientation, until the means coincide. This entire operation may be summarized by the term 'translation to a common mean' or translation to a common origin'. The resulting pattern of observations is as shown in Figure 6(d). Notice that each observation bears the same relationship to the 'common mean' as it did to its original group mean.

The same result is achieved algebraically by first transforming the original observations of columns (1) and (2) of Table 5(a) into deviations about their respective groups means i.e. $x_{ij} = (X_{ij} - \bar{X})$, $y_{ij} = (Y_{ij} - \bar{Y})$ given in columns (1) and (2) of Table 5(b). Applying the formula for b from Table 2, we find the least squares estimate of the line fitted to the super-imposed residual values is b = -1.167 (Figure 6(d)). The observations must now be restored to their original group positions in order to adjust the mean values estimated in the analysis of variance. Figure 6(e) shows a line of slope -1.167 drawn through the bivariate mean of each group of observations. The adjustment procedure is illustrated in a simplified and enlarged version

of the diagram in Figure 7.
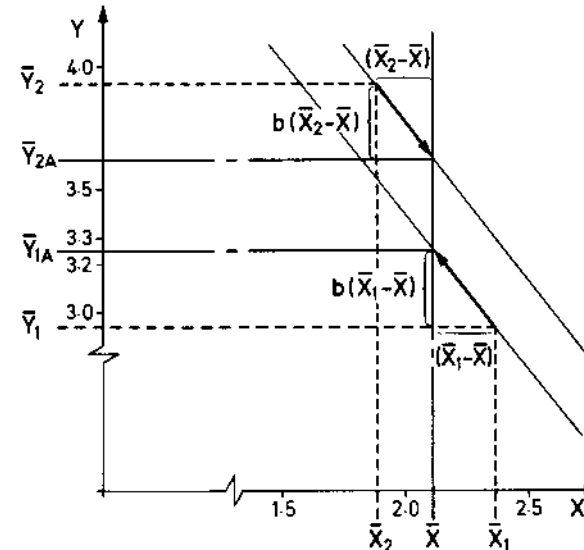
Figure 7. Obtaining the adjusted means.

We have already suggested that mean trip rates to the city centre for the two groups may differ partly because of the differences between their average distance of residence from the city centre. Recall that these distances differ appreciably, being $\bar{X}_1$ = 2.38 km. for the elderly, and $\bar{X}_2$ = 1.87 km. for the non-elderly. We now calculate the mean trip rates expected *if there were no difference* between $\bar{X}_1$ and $\bar{X}_2$, using $\bar{X}$ as the best estimate of the mean value of all the distances under this assumption.

To make the adjustments geometrically, place the point of a pencil on the bivariate mean of group 1 in Figure 7, and move it along the group 1 regression line towards $\bar{X}$. On reaching $\bar{X}$, read off the corresponding Y value, which will be the adjusted mean, $\bar{Y}_{1A}$. Repeat to obtain the adjusted mean, $\bar{Y}_{2A}$, for group 2. Figure 7 shows that $\bar{Y}_{1A}$ lies between 3.2 and 3.3 and $\bar{Y}_{2A}$ close to 3.6.

Algebraically, the adjusted means may be obtained using the restatement of the ANCOVA or Parallel model estimating equation:

$$\bar{Y}_{iA} = \bar{Y}_i - b(\bar{X}_i - \bar{X}) \qquad (26)$$

where $\bar{Y}_{iA}$ is the adjusted mean of the ith group. The adjustment is provided by the second term on the right-hand-side of (26). Essentially, the formula states that the original mean trip rate should be altered by an amount which depends upon i) the difference between the category mean and the overall mean of the covariate, ii) the rate at which trip rates change with distance, and iii) the sign (positive or negative) of the relationship between X and Y.

Applying the formula to the original trip rates:

$$\bar{Y}_{1A} = 2.94 - (-1.167)(2.380 - 2.125) = \underline{3.24}$$

$$\bar{Y}_{2A} = 3.92 - (-1.167)(1.870 - 2.125) = \underline{3.62}$$

Notice the adjustments make sense, being *upward* for the elderly (group 1) who on average live *further* away from the city centre, and *downward* for the non-elderly (group 2) who live *closer*. We expect the elderly to make *more* trips in this case if we allow for the influence of distance, compared with the number of trips made if we take account only of the age category into which they fall. The reverse holds true for the non-elderly. However, the estimated *effects* of being in groups 1 and 2 are now:

$$a_{1A} = \bar{Y}_{1A} - \bar{Y} = 3.24 - 3.43 = \underline{-0.19}$$

$$a_{2A} = \bar{Y}_{2A} - \bar{Y} = 3.62 - 3.43 = \underline{0.19}$$

Allowance for the influence of distance from the city centre has diminished that of the categorical variable representing age, since our initial estimates of the group effects, based on a one-way analysis of variance model, were -0.49 and 0.49 respectively.

Tests of Significance. To find out whether the covariate X is worthy of inclusion in the model, we first compare the ANOVA and ANCOVA (or Parallel) models. The proportion of variation in Y explained by the ANCOVA model is obtained by adding the sum of squares (SS) explained by the ANOVA model and the regression SS explained by introducing the covariate, and dividing this total by the total sum of squared (TSS). The SS explained by X is given by:

$$\text{Regression SS} = b\Sigma\Sigma x_{ij} y_{ij} = \frac{(\Sigma\Sigma x_{ij} y_{ij})^2}{\Sigma\Sigma x_{ij}^2}$$

Taking appropriate values from the 'grand totals' row in Table 5(b), we obtain $(-6.716)^2/5.757 = 7.8348$, so that

$$R^2 = \frac{4.80 + 7.8348}{13.242} = 0.9541$$

The variance ratio used to test for the inclusion of X is given by

$$F_{1,17} = \frac{0.9541 - 0.3625}{(1-0.9541)/(20-3)} = \underline{219.11}$$

a value significant at the 1% level. We conclude that the covariate X is worthy of inclusion.•

Next, we want to know whether the assumption of parallel slopes is justified. A significant difference between Parallel and Full models implies that the assumption cannot be sustained, as the Full model specifies unequal slopes. For this comparison, we require the proportion of variation in Y

explained by adding the SS explained by the ANOVA model and the regression SS explained by introducing the covariate so that regression coefficients differ between groups, and dividing this total by TSS. The SS explained by X is now given by

$$\text{Regression SS} = \sum_i \frac{(\Sigma x_{ij} y_{ij})^2}{\Sigma x_{ij}^2}$$

Using the appropriate values from the 'subtotals' rows in Table 5(b) gives

$$\text{Regression SS} = \frac{(-4.042)^2}{3.276} + \frac{(-2.674)^2}{2.481} = \underline{7.8691}$$

and $R^2 = (4.80 + 7.8691)/13.242 = 0.9567$.

The test ratio is

$$F_{1,16} = \frac{0.9567 - 0.9541}{(1-0.9567)/(20-4)} = \underline{0.96}$$

and provides little evidence against the assumption of parallel slopes.

Finally, we compare the Parallel and Joint models to ascertain whether the differences between the group means, adjusted for the inclusion of the covariate X, are still significant. An insignificant difference implies no difference between adjusted means since the Joint model specifies no difference between groups (Figure 5(d)). This test requires the proportion of variation explained by the simple regression of Y on X and, using the 'grand totals' in Table 5(a), is found to be $R^2 = 0.9087$. Comparing the Parallel and Joint (i.e. simple bivariate) models, the test ratio is:

$$F_{1,17} = \frac{0.9541 - 0.9087}{0.0459/(20 - 3)} = \underline{16.81}$$

which is significant at the 1% level. We conclude that there is a real difference between the adjusted means and that the effects of age do exert some influence on trip frequency, over and above those exerted by distance.

ii) Approach based on dummy variable analysis

The hypothetical trip and distance data are shown in a format suitable for dummy variable analysis in Table 6. Models and derived estimating equations for ANOVA and ANCOVA are given in Table 7(a), which also shows how easily unadjusted and adjusted means may be obtained. Information for comparison of models may also be taken directly from computer package regression output, and the variance ratios for statistical testing readily calculated as shown at the foot of Table 7(b). (N.B. The value of k can also be obtained from N-q, where N is the total number of observations, and q the degrees of freedom associated with the residual SS.)

Y	X_0	X	D_1	D_2	$D_1 X$	$D_2 X$
4.2	1	1.3	1	0	1.3	0
3.5	1	1.8	1	0	1.8	0
3.2	1	2.1	1	0	2.1	0
3.4	1	2.2	1	0	2.2	0
3.2	1	2.3	1	0	2.3	0
2.6	1	2.5	1	0	2.5	0
3.0	1	2.5	1	0	2.5	0
2.8	1	2.7	1	0	2.7	0
2.0	1	2.9	1	0	2.9	0
1.5	1	3.5	1	0	3.5	0
5.0	1	1.0	0	1	0	1.0
4.6	1	1.2	0	1	0	1.2
4.5	1	1.4	0	1	0	1.4
3.9	1	1.7	0	1	0	1.7
3.8	1	1.9	0	1	0	1.9
3.6	1	2.1	0	1	0	2.1
3.5	1	2.2	0	1	0	2.2
3.4	1	2.3	0	1	0	2.3
3.7	1	2.4	0	1	0	2.4
3.2	1	2.5	0	1	0	2.5

Figure 8. Extrapolation problems

28

Table 7 (a) Estimates for ANOVA and ANCOVA models using dummy variable analysis

A. ANOVA

Model (group 1 as anchor)

$$\bar{Y}_1 = \mu_1 + \bar{\varepsilon}_1$$

$$\bar{Y}_i = \mu_i + \alpha_i D_i + \bar{\varepsilon}_i$$

Regression equation

$$\hat{Y} = 2.95 + 0.98 D_2$$

Unadjusted means

$$\bar{Y}_1 \qquad = 2.94$$

$$\bar{Y}_2 = 2.94 + 0.98 = \underline{3.92}$$

B. ANCOVA (PARALLEL)

$$\bar{Y}_{1A} = \mu_1 + \beta \bar{X}_1 + \bar{\varepsilon}_1$$

$$\bar{Y}_{1A} = \mu_1 + \alpha_i D_i + \beta X_i + \bar{\varepsilon}_i$$

Regression equation

$$\hat{Y} = 5.72 + 0.38 D_2 - 1.167X$$

Adjusted means

$$\bar{Y}_{1A} = 5.72 - 1.167(2.125) \qquad = \underline{3.24}$$

$$\bar{Y}_{2A} = (5.72 + 0.38) - 1.167(2.125) = \underline{3.62}$$

(b) Information for comparison of models

	R^2	$\underline{k}$ (No.of parameters estimated)
ANOVA	0.3626	2
ANCOVA (PARALLEL)	0.9543	3
FULL	0.9569	4
JOINT	0.9086	2

Total number of observations (N) = 20

ANOVA (Significance of grouping)

$$F_{1,18} = \frac{0.3626 \times 18}{1 - 0.3626} = \underline{10.24}$$

ANOVA vs. ANCOVA

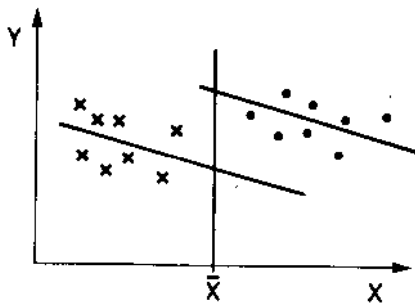$$F_{1,17} = \frac{0.9543 - 0.3626}{(1-0.9543)/(20-3)} = \underline{219.15}$$

29

iii) <u>Checking assumptions underlying the analysis of covariance</u>

We will not deal with all the possible checks, as many are common to regression analysis in general (see Ferguson, 1978, 24-30, for details), but only with those peculiar to the analysis of covariance. First, it is desirable to ensure, as far as possible, that the mean values of the covariate, X, for each group, are not very different'. If this is not the case, the problem illustrated in Figure 8 may arise. The adjusted means, $\bar{Y}_{1A}$ and $\bar{Y}_{2A}$, are both related to an overall mean value for X, $\bar{X}$, lying outside the range of observations for either group 1 or group 2. We are forced to extrapolate well beyond the limits of either group of observations to obtain the adjusted means, and can have relatively little confidence in our estimates. Where little or no control can be exercised over the values of X for each group, the assumption relating to group means must be checked after the data have been collected.

A formal test for homogeneity or equality of the variance about the (parallel) regression line fitted to each group may be unnecessary if there are equal numbers of observations in each group, as the F statistic is not very sensitive to departures from equality under these circumstances. Otherwise, this assumption may be checked using a *two-tailed* F test where there are only two groups to be considered. The test is most easily carried out by dividing the larger variance estimate (s_1^2) by the smaller variance estimate (s_2^2), and then referring to the $(\alpha/2) \times 100$ percentage point, based on appropriate degrees of freedom, in the tabulated F distribution if a test of significance at the $\alpha \times 100\%$ level is required. For example, the appropriate critical value at the 5% level is obtained if we follow this procedure and look up the critical 2½% point. Although a test for variance homogeneity would not be necessary in the case of our worked example - because of equal numbers in each group, and because the data are 'well-behaved' (Figure 6) - it is convenient to use these observations to illustrate how the test should be applied. Table 8 gives the residuals about the ANCOVA model for each group. These values were obtained as output from the package regression routine used to carry out the dummy variable analysis. The degrees of freedom associated with each residual sum of squares are n_i -1. Because there are 9 degrees of freedom for each group, we may ignore them for purposes of calculation. Hence, the F statistic is obtained directly from the total sums of squares in Table 11, and equals 0.372/0.232 or 1.60. This is smaller than $F_{9,9,0.025} = 4.03$ and so we have little reason to reject the (null) hypothesis of equal variances. If the analysis of covariance model is for some reason rejected and the analysis of variance model selected as that best fitting the data, an F test for the equality of variances may be based upon the residual sums of squares given in column (3) of Table 5(b). Degrees of freedom will be the same as for the analysis of covariance model, and the F ratio is 5.344/3.096 = 1.73, which is not significant at the 5% level.

The assumption of parallel or equal slopes is tested by comparing the analysis of covariance (Parallel) model with the Full model, as described in the previous section.

Table 8. Residuals about the <u>analysis of covariance model- worked example</u>

Group 1		Group 2	
e_{1j}	e^2_{1j}	e_{2j}	e^2_{2j}
0.000	0.000	0.065	0.004
-0.117	0.014	-0.102	0.010
-0.067	0.004	0.032	0.001
0.250	0.063	-0.218	0.048
0.167	0.028	-0.085	0.007
-0.200	0.040	-0.052	0.003
0.200	0.040	-0.035	0.001
0.233	0.054	-0.018	0.000
-0.333	0.111	0.398	0.158
-0.133	0.018	0.015	0.000
	0.372		0.232

N.B. Values are correct to 3 decimal places only

VI <u>COMPARISON OF REGRESSION LINES BY DUMMY VARIABLE ANALYSIS - A PRACTICAL</u>

APPLICATION: STOCKING AND ELWELL (1976)

i) <u>Introduction</u>

In their paper on 'Rainfall erosivity over Rhodesia', Stocking and Elwell (1976) are primarily concerned with the estimation of 'erosivity', defined by them as the ability of rainfall to cause erosion'. The measure of erosivity actually employed, EI_{30}, represents the product of kinetic energy and the maximum sustained rainfall intensity lasting for 30 minutes, recorded in units of joules/mm/m²/hour. Their study provides a very good example of work in which values of a variable, erosivity (Y), which are not generally available, or which are relatively time-consuming and expensive to establish, are estimated from values for which fairly detailed records are already available - in this case, mean annual rainfall (X). One of the aims of the study was to produce an erosivity map for Rhodesia as an aid to soil conservation planning and the prediction of storm soil losses.

Analysis of 8000 storms permitted direct estimation of erosivity for 33 recording points over the country, but a far greater number of points was

required in order to construct the erosivity map. Visual inspection of available data suggested that the association between long-term mean annual rainfall and annual erosivity was a fairly strong one, and so a regression analysis was carried out in order to calibrate the relationship between the two variables.

It was thought that the relationship might vary according to the region in which the recording station was located. Observations were placed in one of four regions - highveld, middleveld, lowveld or eastern districts (Figure 9). Because the original data were not published in Stocking and Elwell's paper, 'best guesses' of these values, given with the dummy variable tableau in Table 9, were derived by enlarging one of the figures from their paper (see Figure 10) using a Plan-Variograph and a suitably graduated transparent overlay. Errors in estimating the original values, and in reproducing the regression analysis, proved to be minor.
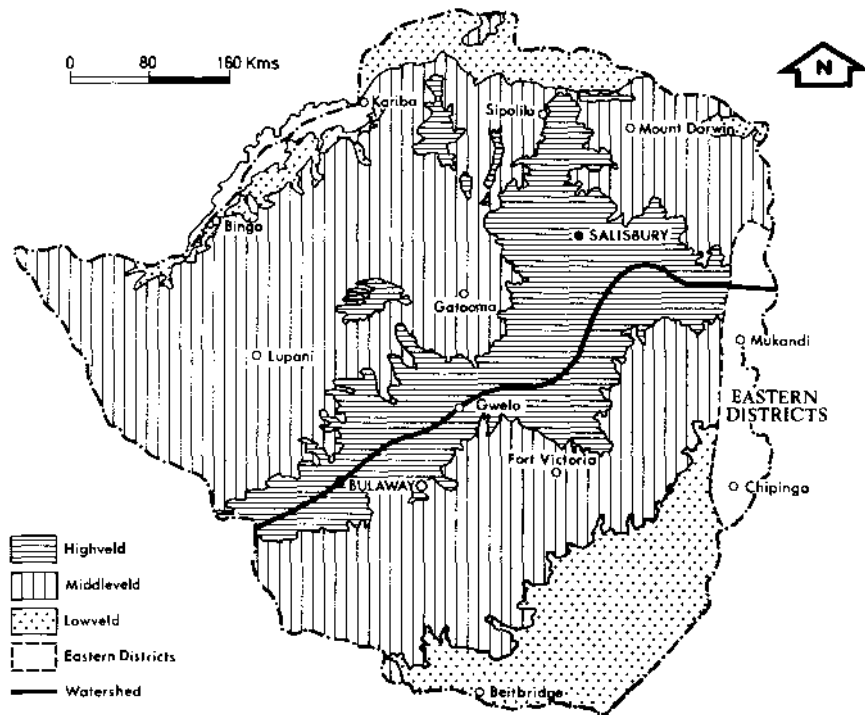

Figure 9. Regional division of Rhodesia for regression of erosivity on rainfall (adapted from Stocking and Elwell, 1976).

Table 9. Dummy Variable Tableau for Erosivity and Rainfall Data

Y	X_o	X	D_1	D_2	D_3	D_4	D_1X	D_2X	D_3X	D_4X	
6560	1	775	1	0	0	0	775	0	0	0	
6910	1	550	1	0	0	0	550	0	0	0	
6830	1	520	1	0	0	0	520	0	0	0	
6540	1	485	1	0	0	0	584	0	0	0	
7080	1	510	1	0	0	0	510	0	0	0	
7970	1	595	1	0	0	0	595	0	0	0	
7875	1	615	1	0	0	0	615	0	0	0	
8200	1	545	1	0	0	0	545	0	0	0	
8300	1	653	1	0	0	0	653	0	0	0	
8800	1	770	1	0	0	0	770	0	0	0	
9050	1	590	1	0	0	0	590	0	0	0	
9380	1	637	1	0	0	0	637	0	0	0	HIGHVELD
9250	1	663	1	0	0	0	663	0	0	0	
10150	1	825	1	0	0	0	825	0	0	0	
10380	1	795	1	0	0	0	795	0	0	0	
10970	1	595	1	0	0	0	595	0	0	0	
10100	1	573	1	0	0	0	573	0	0	0	
11700	1	825	1	0	0	0	825	0	0	0	
11725	1	808	1	0	0	0	808	0	0	0	
12100	1	705	1	0	0	0	705	0	0	0	
12050	1	715	1	0	0	0	715	0	0	0	
12200	1	730	1	0	0	0	730	0	0	0	
12100	1	795	1	0	0	0	795	0	0	0	
13050	1	850	1	0	0	0	850	0	0	0	
13540	1	845	1	0	0	0	845	0	0	0	
4750	1	520	0	1	0	0	0	520	0	0	
6200	1	455	0	1	0	0	0	455	0	0	
6170	1	703	0	1	0	0	0	703	0	0	
7080	1	536	0	1	0	0	0	536	0	0	
7380	1	538	0	1	0	0	0	538	0	0	
7500	1	580	0	1	0	0	0	580	0	0	
7120	1	610	0	1	0	0	0	610	0	0	
7725	1	607	0	1	0	0	0	607	0	0	
7620	1	660	0	1	0	0	0	660	0	0	
7975	1	655	0	1	0	0	0	655	0	0	MIDDLEVELD
8280	1	598	0	1	0	0	0	598	0	0	
8700	1	598	0	1	0	0	0	598	0	0	
8480	1	620	0	1	0	0	0	620	0	0	
9100	1	602	0	1	0	0	0	602	0	0	
9100	1	655	0	1	0	0	0	655	0	0	
10875	1	714	0	1	0	0	0	714	0	0	
11125	1	660	0	1	0	0	0	660	0	0	
11200	1	752	0	1	0	0	0	752	0	0	
13610	1	690	0	1	0	0	0	690	0	0	
14225	1	692	0	1	0	0	0	692	0	0	
2500	1	418	0	0	1	0	0	0	418	0	
2950	1	430	0	0	1	0	0	0	430	0	
3940	1	300	0	0	1	0	0	0	300	0	
3750	1	305	0	0	1	0	0	0	305	0	LOWVELD

32

33

Table 9 (continued)

	X_0	X	D_1	D_2	D_3	D_4	D_1X	D_2X	D_3X	D_4X
4680	1	368	0	0	1	0	0	0	368	0
4920	1	378	0	0	1	0	0	0	378	0
5900	1	389	0	0	1	0	0	0	389	0
8500	1	574	0	0	1	0	0	0	574	0
9200	1	559	0	0	1	0	0	0	559	0
9980	1	647	0	0	1	0	0	0	647	0
10350	1	604	0	0	1	0	0	0	604	0
12080	1	690	0	0	1	0	0	0	690	0
5700	1	710	0	0	0	1	0	0	0	710
8170	1	718	0	0	0	1	0	0	0	718
9350	1	727	0	0	0	1	0	0	0	727
9640	1	767	0	0	0	1	0	0	0	767
9800	1	757	0	0	0	1	0	0	0	757
10200	1	865	0	0	0	1	0	0	0	865 EASTERN
11900	1	912	0	0	0	1	0	0	0	912 DISTRICTS
11650	1	953	0	0	0	1	0	0	0	953
13350	1	1025	0	0	0	1	0	0	0	1025
13160	1	1052	0	0	0	1	0	0	0	1052
13390	1	1118	0	0	0	1	0	0	0	1118

ii) Results

One major difference between the results reported by Stocking and Elwell and those given here arises because they regressed rainfall upon erosivity, and then used the equations to estimate erosivity from rainfall, whereas we have regressed erosivity (Y) upon rainfall (X). The latter procedure is the correct one, because the regression line of X upon Y is *not* the same as that of Y upon X, unless there is perfect correlation between the two variables (Davies, 1961, 151-152).

Simple bivariate regression equations for each of the four regions are given in Table 10(a). All the equations, as shown by the t values, are statistically significant at the 1% level or better. For stations in the middleveld and lowveld regions, both the slopes (21.74 and 21.87) and intercepts (-4817.88 and -3756.33) are fairly similar. The slope terms for the highveld and eastern districts regions are similar (13.22 and 14.79), but the intercepts very different (738.31 and -2339.07). None of the individual regional regression lines, except that for the highveld, closely resemble the regression line (Joint model) estimated with respect to observations in all the regions.

Stocking and Elwell used a method for comparing the 4 regional regression lines with the joint regression line, but it appears that they were in fact able only to test for differences between intercepts, and not for those between slopes. The parameters of the models upon which the comparisons given here are based were estimated using dummy variables D_2, D_3 and D_4, and multiplicative variables D_2X, D_3X and D_4X in Table 9. Thus, D_1 and D_1X were deleted, and the intercept and slope values for the

---

Table 10. Results For Four Regions

(a) Estimation of parameters

COEFFICIENTS

MODEL	a	a_2	a_3	a_4	b	b_2	b_3	b_4	s	R^2	k
JOINT	311.38	-	-	-	13.23	-	-	-	1570	0.6176	2
PARALLEL	-1617.31	-58.40	304.91	-2382.49	16.69	-	-	-	1610	0.6827	5
INTERCEPT	-2193.83	-	-	-	17.42	-	-	-	1596	0.6884	5
FULL	738.31	-5556.19	-4494.63	-3077.37	13.22	8.52	8.65	1.57	1584	0.7076	8

(t values e.g.: 0.16, 1.37, -2.79)

Bivariate Regression Equations

Region	Equation	t values	s	R^2
Region 1 (Anchor) Highveld	$\hat{Y} = 738.31 + 13.22X$	4.85	1727	0.4998
Region 2 Middleveld	$\hat{Y} = -4817.88 + 21.74X$	3.74	1865	0.4382
Region 3 Lowveld	$\hat{Y} = -3756.32 + 21.87X$	6.72	1463	0.8118
Region 4 Eastern Districts	$\hat{Y} = -2339.06 + 14.79X$	6.64	1044	0.8354

(b) Comparison of models

Joint vs. Parallel: $F_{3,63} = \dfrac{(0.6827-0.6176)/(5-2)}{(1-0.6827)/(68-5)} = 4.34^{**}$

Joint vs. Intercept: $F_{3,63} = \dfrac{(0.6884-0.6176)/(5-2)}{(1-0.6884)/(68-5)} = 4.82^{**}$

Parallel vs. Full: $F_{3,60} = \dfrac{(0.7076-0.6827)/(8-5)}{(1-0.7076)/(68-8)} = 1.69^{\dagger}$

Intercept vs. Full: $F_{3,60} = \dfrac{(0.7076-0.6884)/(8-5)}{(1-0.7076)/(68-8)} = 1.31^{\dagger}$

** significant at 1% level

† not significant

(c) Variance-covariance matrix for parallel model

$N = 68$

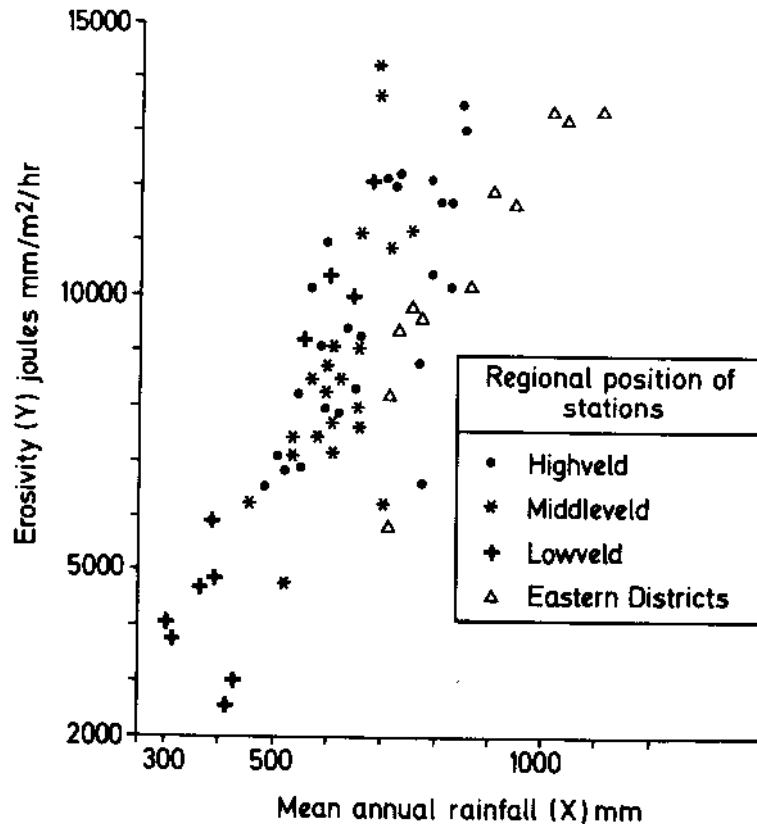|	X	a_2	a_3	a_4
X	3.0282			
a_2	171.1236	242998.5		
a_3	626.6155	139111.5	449409.6	
a_4	-588.4730	70446.8	-18609.3	453745.0

Figure 10. Erosivity and rainfall observations for comparison of regressions problem (adapted from Stocking and Elwell, 1976).

'anchor region', highveld, estimated by the coefficients of X_9 (the constant term) and X respectively. Notice that the regression coefficients estimated for the four simple bivariate equations may also be estimated from the coefficients for the Full model.

Table 10(b) shows that both the Parallel and Intercept models explain a significantly greater proportion of the variation in erosivity than the Joint model. Because the Full model does not represent a significant improvement on either of these two models, multicollinearity between the slope and intercept 'effects' is evident. Neither set of effects is significant if entered last. As stated earlier, it is not possible to compare the Parallel and Intercept models statistically. Nevertheless, some criterion for choosing between them is desirable. An appropriate measure in

this case is the standard error of estimate, s, obtained from the square root of the residual mean square estimated for the regression model in question, and reflecting the accuracy with which values of the dependent variable are estimated. These values are also given in Table 10(a). Using this criterion, the Intercept model fares slightly better than the Parallel model, their s values being 1610 and 1596 respectively. However, the value of 1584 for the Full model is lower still. This value cannot be taken as it stands, since the Full model specifies four separate regression lines, each with its own standard error of estimate, but it may be a useful indicator. It proves to be so here, as the standard errors with respect to the individual lines are *less* than 1584 for all regions except the middleveld (Table 10a). Despite the statistically insignificant differences between Parallel and Full and Intercept and Full models, a criterion reflecting accuracy of estimation - which is the main purpose of Stocking and Elwell's analysis - seems to indicate that a separate regression line should be used for each region. The price we pay for relative accuracy in the highveld, lowveld and eastern districts regions is a rather unreliable equation for the middleveld region (s = 1865, Table 10a).

Stocking and Elwell substituted values of mean annual rainfall into their regression equations and tested the differences between estimates yielded by the 'all points' (i.e. Joint model) and regional grouping lines for significance. Using the Parallel model, based on four regions, we can carry out similar tests with respect to any given pair of adjusted means - here, these tests amount to comparing any given pair of regression lines because slope is held constant. Differences involving the regression line for the anchor group may be assessed directly from package regression output, as the t values for the coefficients a_2, a_3, and a_4 in Table 10(a) refer to the Highveld - Middleveld (t = 0.12), Highveld - Lowveld (t = 0.45) and Highveld - Eastern Districts (t   3.54) differences respectively. Only the latter difference is found to be significant. For other comparisons, estimates of standard errors must be derived from the appropriate variance-covariance matrix in Table 10(c). Applying equation (23) to the Middleveld-Eastern Districts difference:

$$\text{est s.e. } (a_2 - a_4) = \sqrt{242998.5 + 453745.0 - 2(70446.84)}$$
$$= \sqrt{555849.82} = \underline{745.55}$$

and

$$t_{63} = \frac{a_2 - a_4}{\text{est s.e. } (a_2 - a_4)} = \frac{-58.40 - (-2382.49)}{745.55} = \underline{3.12}$$

a value significant at the 1% level. Any other difference may also be tested for significance, bearing in mind the cautionary comments already made about unplanned comparisons in Section IV(iv).

As a result of their analysis, Stocking and Elwell suggest that it would be justifiable to use only two regression lines - one for eastern districts and one for the rest of Rhodesia. This proposition may be tested if we regress Y on X, D_4 and D_4X only, so that the anchor group represents the 'rest of Rhodesia'. The pattern of results is very much the same as that obtained with respect to the original four-region classification,

multicollinearity between slope and intercept effects again being present (Table 11a,b).

Table 11. Results for two regions

(a) Estimation of parameters

MODEL	COEFFICIENTS						
	a	a₂	b	b₂	s	R²	k
JOINT	311.38	-	13.23	-	1727	0.6176	2
PARALLEL	-1301.26	-2312.13	16.25	-	1589	0.6810	3
INTERCEPT	-1586.61	-	16.71	-2.76	1587	0.6818	3
FULL	-1501.64	-837.43	16.57	-1.78	1599	0.6822	4

		s	t values	R²
Bivariate Regression Equations	Region 1 (Anchor) Rest of Rhodesia $Y = -1501.64 + 16.57X$	1672	9.84	0.6384
	Region 2 Eastern districts $Y = -2339.07 + 14.79X$	1044	6.64	0.8299

(b) Comparison of models - two region case only

Joint vs. Parallel $F_{1,65} = $ 12.94**

Joint vs. Intercept $F_{1,65} = $ 13.01**

Parallel vs. Full $F_{1,64} = $ 0.24 †

Intercept vs. Full $F_{1,64} = $ 0.08 †

(c) Comparison of models - two region vs. four region case

2-region Full vs. 4-region Full $F_{4,60} = $ 1.31†

2-region Parallel vs. 4-region Full $F_{5,60} = $ 1.08†

2-region Intercept vs. 4-region Full $F_{5,60} = $ 1.06†

** significant at 1% level

† not significant

Comparing the Full model based on four regions with the Parallel, Intercept and Full models based on two regions - eastern and the rest of Rhodesia - no significant differences can be found, thus supporting the proposition (Table 11c). Comparison of standard errors produces yet another unequivocal result, since the value for the 'rest of Rhodesia' is relatively high (s = 1672) and exceeded only by that for the individually derived equation for the middleveld (s 1865, Table 10a).
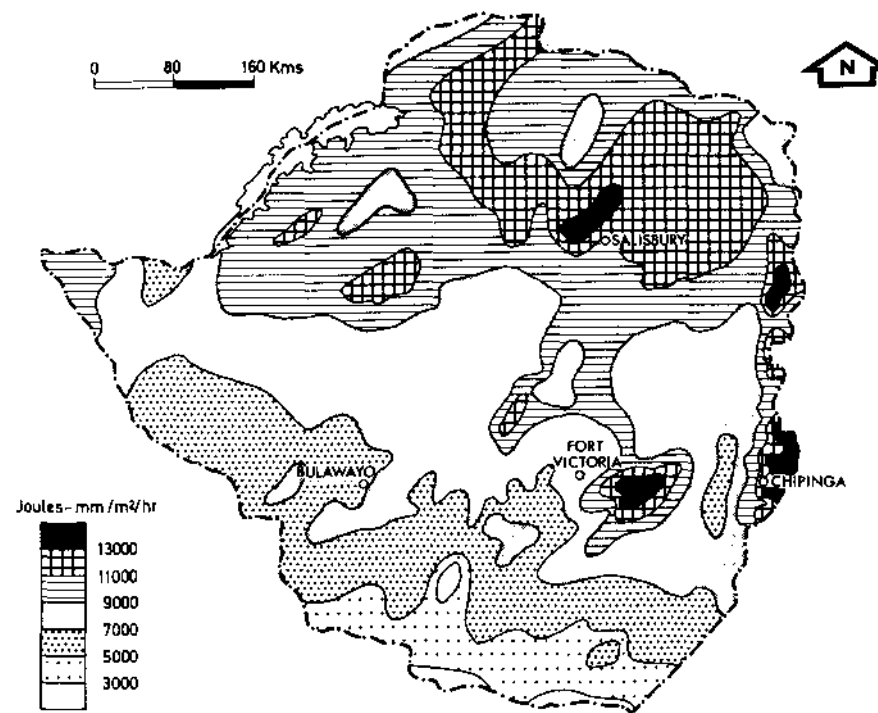
Figure 11. Mean annual erosivity over Rhodesia estimated from regression equations (redrawn from Stocking and Elwell, 1976).

Stocking and Elwell applied the four separate regression equations of Table 10(a) to the mean annual rainfall map of Rhodesia, converting the isohyets to isopleths of erosivity. The resulting 'erosivity map' is reproduced in Figure 11. Remember that the extreme high and low erosivity values have been exaggerated by Stocking and Elwell because of their regression of X on Y rather than Y on X.

If one of the Parallel or Intercept models for the four-region case had proved superior in terms of accuracy of estimation, how would the erosivity estimates have been obtained? Taking the Parallel model as an example, the estimating equation for the anchor region highveld, according to Table 10(a), is:

$$Y = a + bX = -1617.31 + 16.69X$$

and that for the middleveld is given by:

$$Y = a + a_2 + bX = -1617.31 + (-58.40) + 16.16X = -1675.71 + 16.69X$$

Corresponding to a mean annual rainfall of, say, 600mm, we obtain erosivity estimates of

$$Y = -1617.31 + 16.69 \times 600 = \underline{8397}$$

and

$$Y = -1675.71 + 16.69 \times 600 = \underline{8338}$$

Estimates from other models are easily obtained by adding terms in the appropriate rows in Table 10.

VI EXAMPLES OF ANALYSIS OF COVARIANCE AND DUMMY VARIABLE ANALYSIS IN GEOGRAPHY

As already indicated in the Introduction, analysis of covariance and related techniques may be used to fulfill a number of objectives. A list of broad headings, with examples, is given in Table 12. Any given study may pursue more than one of these objectives, and the examples briefly discussed below may appear under more than one heading.

One important goal is *estimation* in the sense of calculating an estimate of a value of Y which corresponds to a given value of X. This goal is of great interest to planners, and to managers of natural resources, because of its relevance for prediction and forecasting. Silk (1976, 29-34) provides a detailed example showing how the relationship between commercial trip generation and size of establishment in the London Borough of Haringey varies across SIC groups, and we have shown earlier how Stocking and Elwell (1976) compared regression lines in order to produce an erosivity map of Rhodesia estimated from mean annual rainfall and regional location. It is interesting that Stocking and Elwell were unable to carry out a fully satisfactory analysis because neither the analysis of covariance nor dummy variables was used. There is no doubt that geographers could, and should, pay greater attention to the use of many familiar techniques, such as regression and the analysis of variance, for estimation purposes. Transportation researchers lead the way in this regard at the moment e.g. Bayliss and Edwards (1970), Douglas and Lewis (1971), Heathington (1972), Douglas (1973), and opportunities exist in geography as shown by O'Sullivan (1969), Chisholm (1971) and Chisholm and O'Sullivan (1973).

Hypothesis-testing represents a more familiar objective to geographers, because of the emphasis placed on significance testing in our statistics texts and courses. Here, we are thinking primarily of *a priori* hypotheses i.e. those developed before the data are analyzed, or even before they are collected. The ranking of slope terms in the regression of perceived upon actual distances by Briggs (1973) has already been mentioned (p. 5, and see Figure 1). Despite our familiarity with hypothesis-testing, comparison of the slopes of two or more regression lines, or indeed testing against any null hypothesis other than $\beta = 0$, is rare. Many potential applications exist in the field of behavioural geography, e.g. distance perception studies, and in general urban and social geography where distance-decay curves, rent and population density gradients and the like are frequently studied.

Parsimony, at least in terms of description, may be achieved because of a reduction in the number of regression terms required. Hart and Salisbury (1965) examined the relationship between village population change and distance from the nearest urban centre for a sample of 400 settlements in the American Midwest. They found that a single overall regression line

40

Table 12. Classification of objectives of analysis of covariance studies illustrated by geographic examples

Objective	Examples	Topics
Estimation	Silk (1976)	Commercial trip generation from factories in Haringey, London.
	Stocking & Elwell (1976)	Erosivity as a function of mean annual rainfall in Rhodesia.
Hypothesis-Testing	Briggs (1973)	Relationships between perceived and actual distances for a student sample in Columbus, Ohio.
Parsimony	Hart & Salisbury (1965)	Association between village population change and distance from nearest urban centre for 400 sample settlements in the American Midwest.
	O'Farrell & Markham (1974)	See text
	Silk (1976)	See text
Correction for Confounding	O'Farrell & Markham (1974)	Differences in perceived travel time to and from work by car- and bus-users corrected for differences in average travel times to/from work for the two groups.
	Trenhaile (1974)	See text
Classification	King (1961)	After regression of a measure of settlement spacing on selected variables, introduction of three different schemes: i) Non-central vs. central places. ii) Regional classification based upon two categories - 'near-level land' and 'other'. iii) Regional classification by five generalised farming types.

41

(Joint Model) could be substituted for the nine regression lines - one for each individual state. Similarily, O'Farrell and Markham (1974) were able to substitute one regression equation for the two initially obtained in a study of the relationship between actual and perceived waiting times with respect to car and bus-users in Dublin. Silk (1976, 29-33) showed how nine separate regression lines could be reduced to a 'hybrid' model consisting of three regression lines with different slopes but a common intercept, plus a free-standing regression line representing one 'deviant' category.

Correction for confounding is a rarely mentioned objective in the geographic literature, although examples may be found in O'Farrell and Markham (1974) and Silk (1976, 19-23). Trenhaile (1974) concluded that any difference between shore platform gradients developed on chalk bedrock compared with those developed in lias sandstones and shales was statistically insignificant if allowance were made for variations in tidal range. Relationships between platform gradient and other variables, such as fetch, were analyzed in the same way.

Last, but not least, the technique may be used as an aid to *classification*. No matter whether a set of categories has been derived *a priori* or represents the fruit of opportunism as research proceeds, it is possible to construct and test statistically classification schemes based on regression models. Most of the papers in Section B of the references provide good examples of this particular use of covariance analysis, and may also be regarded as *exploratory* in the sense that identification of, and statistical testing for, empirical regularities are based on the same data set (Hauser, 1974). Particularly clear discussions of this process of exploration may be found in King (1961), Kariel (1963) and Yeates (1965).

Two comments on procedure may be made here. First, almost all studies concentrate upon comparison of Joint, Parallel and Full models. Admittedly, there are more situations in which we should be pleased to discover that a number of separate regression lines have a common slope, but considerable simplification is also possible if a common intercept can be identified. Second, geographers invariably approach analysis of covariance by way of regression analysis. Essentially, an analysis of variance is carried out on the residuals about a single regression plane. Although the analysis is carried out 'backwards' in terms of the scheme of comparisons outlined previously, there is no reason why the whole process should not be carried out taking the Joint model as its starting point.

A list of studies which employ *dummy variable analysis* is given in Section C of the references.

VII COMPUTER ROUTINES

A number of standard computer package programmes is available for carrying out the analysis of variance and covariance. The Statistical Package for the Social Sciences (SPSS) provides a useful discussion of these techniques as well as the programs in Chapter 22 of the manual (Nie et al, 1975); the library of Biomedical (BMD) Computer Programs is another widely available alternative (Dixon, 1968, 285-304; 597-605; 705-718). To implement the BMD programs on one of the ICL 1900 series of computers, it is

necessary to refer to the appropriate Numerical Algorithms Group (NAG) NIMBUS manual. NAG routines for the analysis of variance are also available.

Routines which automatically provide the output required to compare regression lines do not appear to be generally available. The author used two ALGOL programmes, ASA6 and ASA9, written by members of the Department of Applied Statistics at the University of Reading, to check results obtained using dummy variables and multiple regression routines. Summary information on residual sums of squares and degrees of freedom for 'between' comparisons in the bivariate case is given by ASA6, and in the multivariate case by ASA9. The SPSS manual does, however, provide a full discussion of dummy variable analysis, with examples (Nie et al; 1975, Chapter 21).

Standard multiple regression routines may be used to carry out all the comparisons 'between' and 'within' models described earlier. If such routines are employed a print-out of the variance-covariance matrix is usually available, thus enabling 'within' comparisons to be made between any given pair of intercepts or slopes. The writer exclusively used multiple regression routines, because of the great flexibility possible if dummy variables are employed, e.g. combining of existing categories or creation of new categories.

References to computer program descriptions are given in Section A.

VIII FURTHER EXTENSIONS AND CONCLUSION

The scope of the technique may be extended in various ways. First, by adding one or more categorical independent variables; second, by adding one or more continuous independent variables - this was the strategy adopted by Yeates (1965) to compare multiple regression equations between six radial sectors in Chicago; third, by adding both categorical and continuous variables.

All three extensions, but particularly the first and the last, may require a large number of observations, first, because two-(or higher) way analysis of variance, or its dummy variable equivalent, are expensive in terms of degrees of freedom and, second, to ensure that there are no empty cells in the cross-classification. It is worth noting that the effects in an n-way analysis of variance may be estimated using dummy variable technique if we need to deal with a cross-classification scheme with unequal numbers of observations in the cells, a situation which the conventional estimation procedure cannot handle. Further discussion of these issues may be found in Goldberger (1964, 227-231), and worked examples in Snedecor and Cochran (1967, ch. 14).

There are many branches of geography within which the analysis of covariance, and the associated techniques for comparison of regression lines, may be profitably employed. Technically, there are few obstacles to their more widespread use since the underlying statistical theory is well-developed, and a number of different computing procedures is available.

BIBLIOGRAPHY

A. Technical papers and computer programs

Blalock, H.M. (Jr) (1964), *Causal inferences in nonexperimental research.* (Chapel Hill: University of North Carolina Press).

Blalock, H.M. (Jr) (1972), *Social statistics* (International student edition). (New York: McGraw-Hill), ch. 20.

Cohen, J. (1968), Multiple regression as a general data-analytic system. *Psychological Bulletin,* 70, 426-443.

Davidson, N. (1976), *Causal inferences from dichotomous variables.* Concepts and techniques in modern geography,9. (Norwich: Geo Abstracts Ltd.).

Davies, O.L. (ed) (1961), *Statistical methods in research and production.* (Edinburgh: Oliver and Boyd).

Dixon, W.J. (ed) (1968), *Biomedical computer programs.* (California: University of California Press).

Draper, N.R., and Smith, H. (1966), *Applied regression analysis.* (New York: Wiley).

Ferguson, R. (1978), *Linear regression in geography,* Concepts and techniques in modern geography, 15. (Norwich: Geo Abstracts Ltd.).

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

Goldberger, A.S. (1968), *Topics in regression analysis.* (New York: MacMillan), ch. 8.

Hauser, D.P. (1974), Some problems in the use of stepwise regression techniques in geographical research. *Canadian Geographer,* 18, 148-158.

Jennings, E. (1967), Fixed effects analysis of variance by regression analysis. *Multivariate Behavioural Research,* 2, 95-108.

Johnston, J. (1963), *Econometric methods* (International student edition). (New York: McGraw-Hill), ch. 8.

Kerlinger, F., and Pedhazur, E.J. (1973), *Multiple regression in behavioral research.* (New York: Holt, Rinehart and Winston), chs. 6-11.

Kirk, R.E. (1972), *Statistical issues: a reader for the behavioral sciences.* (Belmont, California: Brooks/Cole).

Nie, N., Bent, D.H., and Hull, C.H. (1975), *Statistical package for the social sciences* (SPSS). (New York: McGraw-Hill).

Selvin, H.C., and Stuart, A. (1966), Data dredging procedures in survey analysis. *The American Statistician,* June 20-23.

Silk, J.A. (1976), A *comparison of regression lines using dummy variable analysis.* Geographical Paper, 44. (Reading: Department of Geography, University of Reading).

Snedecor, G.W., and Cochran, W.G. (1967), statistical *methods.* (Ames, Iowa: Iowa State University Press).

Suits, D.B. (1957), The use of dummy variables in regression equations. *Journal of the American Statistical Association,* 52, 548-551.

Taylor, P.J. (1969), Causal models in geographic research. *Annals of the Association of American Geographers,* 59, 402-404.

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

Weatherburn, C.E. (1962), *A first course in mathematical statistics.* (Cambridge University Press), (2nd edition).

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

B. Applications of analysis of covariance

Briggs, R. (1973), Urban cognitive distance, 361-390 in: Downs, R.M., and Stea, D. (eds), *Image and environment.* (London: Arnold).

Carey, L., and Mapes, R. (1972), *The sociology of planning* (London: Batsford), ch. 4.

Dogan, M. (1969), A covariance analysis of French electoral data. 285-298 in Dogan, M., and Rokkan, S. (eds), *Quantitative ecological analysis in the social sciences.* (Cambridge Mass: MIT Press).

Doornkamp, J.C., and King, C.A.M. (1970), *Numerical analysis in geomorphology,* (London: Arnold).

Garner, B.J. (1966), *The internal structure of retail nucleations.* Northwestern University Studies in Geography, 12. (Evanston, Illinois: Department of Geography, Northwestern University).

Hart, J.H., and Salisbury, N.E. (1965), Population changes in Middle Western villages: a statistical approach. *Annals of the Association of American Geographers,* 55, 140-160.

Kariel, H.G. (1963), Selected factors areally associated with population growth due to net migration. *Annals of the Association of American Geographers,* 53, 210-223.

King, L.J. (1961), A multivariate analysis of the spacing of urban settlements in the United States. *Annals of the Association of American Geographers,* 51, 222-233.

O'Farrell, P.N., and Markham, J. (1974), Commuter perceptions of public transport work journeys. *Environment and Planning,* A6, 79-100.

Trenhaile, A.S. (1974), The geometry of shore platforms in England and Wales. *Transactions,* Institution of British Geographers, 62, 129-142.

C. Applications of dummy variables and comparisons of regression lines

Bayliss, B.T., and Edwards, S.L. (1970), *Industrial demand for transport* London: H.M.S.O.

Blaikie, P.M. (1973), The spatial structure of information networks and innovative behaviour in the Ziz valley, Southern Morocco. *Geraphiska Annaler,* 55B, 83-105.

Cheshire, P. (1973), Regional unemployment differences in Great Britain. in Cheshire, P. (ed) *Regional papers II.* National Institute for Economic and Social Research, (Cambridge: Cambridge University Press).

Chisholm, M. (1971), Freight transport costs, industrial location and
            regional development. 213-244 in Chisholm, M., and Manners, G.
            (eds), *Spatial policy problems of the British economy.*
            Cambridge:(Cambridge University Press).

Chisholm, M., and O'Sullivan, P. (1973), *Freight flows and spatial aspects
            of the British economy.* (Cambridge: Cambridge University Press).

Douglas, A.A., and Lewis, R.J.(1971), Trip generation techniques:
            household least-squares regression analysis. *Traffic Engineering
            and Control,* 12, 477-479.

Douglas, A.A. (1973), Home-based trip models - a comparison between category
            analysis and regression analysis procedures. *Transportation,*
            2, 53-70.

Hannell, F.G. (1973), The thickness of the active layer on some of Canada's
            arctic slopes. *Geografiska Annaler,* 55A, 177-184.

Heathington, K.W. and Isibor, E. (1972), The use of dummy variables in trip
            generation analysis. *Transportation Research,* 6, 131-142.

Knos, D.S. (1968), The distribution of land values in Topeka, Kansas.
            269-289 in Berry, B.J.L., and Marble, D.F. (eds), *Spatial
            analysis.* (Englewood Cliffs: Prentice-Hall).

Lee, T.H. (1963), Demand for housing: a cross section analysis. *Review of
            Economics and Statistics,* 45, 190-196.

O'Sullivan, P. (1969), *Transport networks and the Irish economy.* (London:
            Weidenfield and Nicholson, London School of Economics).

Orcutt, G.H., Greenberger, M., Korbel, J., and Rivlin, A.M. (1961),
            Micro-analysis of socio-economic systems: a simulation study.
            (New York: Harper and Row).

Silk, J.A. (1976), *A comparison of regression lines using dummy variable
            analysis.* Geographical Paper, 44, (Reading: Department of
            Geography, University of Reading).

Starkie, D.N.M. (1970), The treatment of curvilinearities in the calibration
            of trip-end models. In *Urban Traffic Model Research.*
            London: Planning, Transportation Research and Computation Ltd.
            (PATRAC).

Starkie, D.N.M., and Johnson, D.M. (1975). *The economic value of peace and
            quiet.* (Farnborough: Lexington Books/D.C. Heath).

Stocking, M.A., and Elwell, H.A. (1976), Rainfall erosivity over Rhodesia.
            Transactions, Institute of British Geographers (New Series),
            1(2), 231-245.

Vickerman, R.W. (1974). A demand model for leisure travel. *Environment and
            Planning,* 6, 65-77.

Yeates, M.N. (1965), Some factors affecting the spatial distribution of
            Chicago land values, 1910-1960. *Economic Geography,* 41(1),
            57-70.