UNIVERSITY OF SOUTHAMPTON

GEOGRAPHICAL VARIATIONS IN MORTALITY: AN EXPLORATORY APPROACH

by

KELVYN JONES

Thesis submitted for the
degree of Doctor of Philosophy
September, 1980.

LIST OF CONTENTS

CHAPTER 4

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SCIENCE

GEOGRAPHY

Doctor of Philosophy

GEOGRAPHICAL VARIATIONS IN MORTALITY: AN EXPLORATORY APPROACH

by Kelvyn Jones

This thesis aims to provide a geographical contribution to the understanding of disease causation, primarily through the development of causal models of chronic disease mortality incorporating both the physical and social environments. The overwhelming impression of previous research in this field is one of conflicting findings. For example, studies examining the relationship between disease and water hardness have found positive relationships, negative relationships and no relationship whatsoever. It is contended that this failure to replicate results is a direct consequence of applying an unsuitable 'confirmatory' approach to the quantitative analysis of geographical data. It is argued also that it is necessary to adopt a more appropriate statistical methodology, that of 'exploratory' statistics, before progress can be made. After an exegesis of the exploratory approach, the commonly used technique of multiple regression is given an exploratory interpretation. Each of the assumptions of this technique is discussed, and attention focuses on the effects of breaking the assumptions and on methods of detecting and overcoming the resultant problems. This exposition is illustrated by the re-analysis of previous studies, and it is demonstrated that inappropriate methods have led some researchers to inferential error. Finally in this methodological part of the research, an examination of the analysis of ratios is undertaken; here too it is suggested that the inappropriate analysis of death rates has resulted in some researchers making incorrect inferences.

The empirical aspects of the thesis centre on the analysis of mortality variations in England and Wales. A critical appraisal of previous studies of the relationship between disease and water hardness is undertaken, and it is concluded that quantitative techniques have been poorly applied. Exploratory data analysis is then employed to develop models accounting for geographical variations in mortality experienced by the County Boroughs of England and Wales. In contrast to previous studies that have analysed these variations, no strong relationship is found between disease and water hardness. Moreover, an examination of the mortality experiences of Boroughs whose water supply has changed substantially over time also results in the conclusion that the effects of water hardness have been overestimated. Finally, the study examines the difficult problem of drawing inferences from aggregate data. Although it is concluded that much work remains to be undertaken, it is again argued that the exploratory approach may allow progress to be made towards the solution of this problem and, consequently, some guidelines for further research are outlined.

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I should like to acknowledge the advice and assistance provided by a number of people and organisations during the preparation of this thesis.

Dr. Neil Wrigley (now at the University of Bristol) supervised the first year of this research and I must thank him for his patience when the freedom of research did indeed prove to be a heady spirit. Subsequently, I have benefited from the advice of Dr. David Pinder; he did much to tease out my terse English and I hope that I have learned a great deal about presentation from him.

Financial support was provided by the Social Science Research Council and I acknowledge the help of the SSRC Survey Archive (University of Essex) and the original researchers in providing the data without which it would not have been possible to perform the empirical analyses of Chapter 7.

Attendance at a number of the postgraduate courses in Applied and Social Statistics at the University of Southampton was of great value and, in particular, Dr. J.A. John's course on 'Data Manipulation' crystallized my doubts concerning the methods of quantitative analysis that are commonly practised in geography.

My postgraduate and academic colleagues at the Universities of Southampton, Newcastle, Swansea and Reading provided encouragement and I must single out Kate Barnard, Paul Goddard, Peter Gutteridge, Dave Grafton, Robin Talbot and Pete Atkins for their willingness to listen, argue and be friends. Similarly, Paul Boagey of the Geography Library, University of Southampton, provided bibliographic assistance and was generally helpful in a number of ways.

Mrs. Goodwin typed the manuscript and I appreciate the care and time she has taken with a difficult job.

Finally, the contribution of my family should be appreciated. My parents have made many sacrifices for my continuing education and I owe them a great debt. To thank my wife, Tina, leaves me unusually lost for words but I know that without her this thesis could not have been completed.

INTRODUCTION

DEATH IS A MATTER OF MATHEMATICS

Death is a matter of mathematics.

It screeches down at you from dirtywhite nothingness
And your life is a question of velocity and altitude,
With allowances for wind and the quick,
  relentless pull
Of gravity.
Or else it lies concealed
In that fleecy, peaceful puff of cloud ahead,
A streamlined, muttering vulture, waiting
To swoop upon you with a rush of steel.
And then your chances vary as the curves
Of your parabolas, your banks, your dives,
The scientific soundness of your choice
Of what to push or pull, and how, and when.
Or perhaps you walk oblivious in a wood,
Or crawl flat-bellied over pockmarked earth,
And Death awaits you in a field-gray tunic.
Sights upright and aligned. Range estimated
And set in. A lightning, subconscious calculation
Of trajectory and deflection. With you the focal
  point,
The centre of the problem. The A and B
Or Smith and Jones of schoolboy textbooks.

Ten out of ten means you are dead.

Barry Amiel.

'The Book of Nature is written in the mathematical language. Without its help it is impossible to comprehend a single word of it'

Galileo (1564-1642).

'Since no one proposes trying to give people cancer ... the fruitful way toward better causal analysis ... is to concentrate on improving the statistical approach'

Hirschi and Selvin
(1966, 254).

INTRODUCTION

The main concern of medicine in the economically advanced countries is the growing toll from heart disease, strokes, cancer and bronchitis. In Britain ischaemic heart disease is responsible for twenty-five per cent of all deaths (that is nearly 200,000 deaths annually), while cancer accounts for almost another twenty per cent. But the incidence of death is not distributed evenly, for at many scales of analysis diseases show a marked geographical concentration. At a world scale primary carcinoma of the liver, for example, displays a distinctive pattern: it is rare in Europe, North America and Latin America, common in Africa and parts of South-East Asia and is most commonplace in Mozambique. Here it is approximately 500 times more frequent in younger age groups than it is in the United States of America. In contrast, cancer of the stomach appears rare in Africa but it has a high incidence in Japan, Chile, Iceland, Europe and North America. Intriguingly, middle-aged women in Trinidad and Tobago have a death rate from ischaemic heart disease that is over nine times that of French women, and in the same age group the death rate for both male and female Spaniards is substantially lower than that experienced by men and women living in England and Wales. Within particular countries, too, diseases often display marked geographical patterns. In Britain, northern and central Wales are said to exhibit a particularly high fatality from cancer of the stomach, while in the USA the western coastal area has been dubbed the 'stroke belt'. Patterns have also been discerned within a particular city, and several authors (for example, Girt, 1972 and Jones, 1975) have drawn attention to the high incidence of chronic bronchitis in inner-urban areas. Finally, even within a small village, Allen-Price (1960) has claimed to detect a

pattern, with houses receiving domestic water from a particular supply having an extremely high incidence of cancer deaths.

Given such observed geographical patterns, a number of questions come to mind. Are the patterns 'real' or are they brought about by chance occurrences? Do areas with a high incidence of a disease continue to suffer from this problem over a long period? Do areas with a high incidence of a particular disease have something in common? Are there other variables with a similar geographical distribution to that of the disease under study? But these questions are all subsidiary to the major issue: can we use these geographical patterns to elucidate the causes of a disease?

Although the primary goal is to establish the causes of major diseases, researchers aiming at this goal use a wide variety of approaches. Some take a microscopic route by attempting to discover why, for example, a particular cell has become cancerous; others rely on the results of experiments with animals, while further possibilities are to study individual human beings and groups of people. Each of these different avenues has its advantages and disadvantages but, for reasons discussed below, this study will adopt the group approach. The adoption of the group approach is based on three suppositions.

1. Variations in the physical and social environment determine geographical variations in chronic disease mortality.

2. An approach based on aggregate data for groups of people is a worthwhile alternative to the commonly used experimental and microscopic approaches.

3. A specifically 'exploratory' approach to data analysis, in which the researcher openly admits his ignorance, is preferable to the commonly practised 'confirmatory' approach, in which the research places heavy reliance on significance levels and hypothesis testing procedures.

Each of these fundamental suppositions will now be considered in turn.

Environment, disease and medicine

It is widely believed, both by patients and their doctors, that economically advanced countries owe their higher health standards to 'scientific' medicine and that medical technology, which is currently effective in dealing with disease, holds even greater promise in the future. The prevailing ideology of medicine is an 'engineering' one. Nature is conceived in mechanistic terms; a human being is regarded as a machine which can be taken apart, cleaned, oiled, repaired and re-assembled. Confronted with a diseased body, the approach of the physician is to intervene by chemical, surgical or electrical means to restore the patient's disordered system to normality. Partly because this particular ideology pervades much of modern medicine, resources are concentrated on large acute hospitals (where the technically interesting patients are treated) while the elderly, insane and handicapped are frequently treated in less pleasant surroundings.

Currently, heavy emphasis is placed on curative and interventionist measures rather than on preventative medicine, but this was not always so. In classical times there were two ideas concerning man's health: the one associated with the goddess Hygieia was that good health can be achieved by a rational way of life; the other, personified by the god Asclepias depended largely on the role of the physician as healer of the sick. The environmental or Hygieian viewpoint appears to have been codified by Hippocrates in his book On Airs, Waters and Places in the fifth century BC. This work was based on the assumption that bodily functions were a means to maintain a balance among four fluids - blood, phlegm, black bile and yellow bile. As long as these fluids were in equilibrium the body was healthy, but this balance could be upset by changes in environmental variables and an individual's life style. Temperature, humidity, eating,

drinking, sexual activity, work, recreation, behaviour and dress could all alter the balance. Hippocrate's book suggests, for the peripatetic physicians of the day, how to predict on the basis of the community's natural setting and social and economic functions, the endemic disease likely to be found in a town or city. Moreover, advice was given on the location of new towns, and settlers were urged to avoid marshes, damp valleys and unprotected hills.

Since at least the seventeenth century, however, the Asclepian or interventionist approach has dominated. This view of medicine was considerably reinforced by what is now called 'the germ theory' of disease. According to this theory a particular disease is the result of a specific pathogenic organism, and the aim of research is to produce a 'magic bullet' to eradicate the damaging micro-organisms. Each type of disease was thought to be the result of a different micro-organism, and in the last quarter of the nineteenth century at least twenty different pathogenic microbes were identified by eighteen different researchers (Rosen, 1958). The ideas and concepts of the germ theory remain very powerful in modern medicine. Firstly, although the theory originally applied to infections, the aim of developing a 'magic bullet' has been extended to the chronic diseases. Interferon, for example, is the latest in a long line of drugs which have been suggested as a possible cure of cancer. Secondly, the demand for replicable demonstrations of an explanation has become an essential characteristic of bacteriological research. The laboratory experiment, with its emphasis on controlled environments, is the bacteriologist's chief means of gathering data and testing hypotheses. As a result, despite the obvious need to monitor the human population, medicine and public health have become more concerned with laboratory than naturalistic research. In the United Kingdom today, the majority of research funded by the Medical Research Council is for such experimental work.

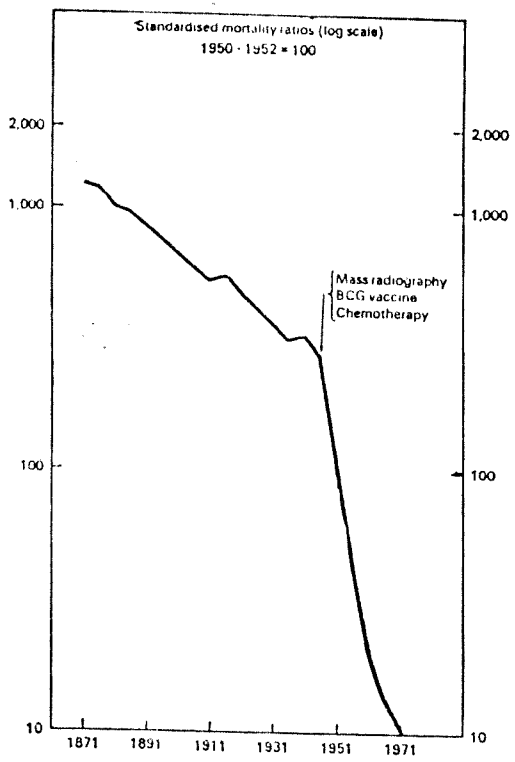
A major reason for the dominance of technological, experimental medicine is the apparent success of the approach. One of the major claims for interventionist medicine is that it produced a substantial decline in the death rate in England and Wales during the eighteenth and nineteenth centuries. Undoubtedly, Talbot Griffith's (1926) book, Population Problems in the Age of Malthus was of seminal importance in fostering this viewpoint. Griffith examined medical developments in the eighteenth century (expansion of dispensary, hospital and midwifery services and the introduction of innoculation against smallpox) and concluded that the improvements resulted in a substantial reduction in mortality. Such views on the efficacy of medicine are encouraged by present-day practitioners and Figure 1a, taken from a recent government report (DHSS, 1976), purports to show the effectiveness of medical intervention. However, in the last thirty years Griffith's ideas have come under considerable attack, especially from Habakkuk (1953) and from McKeown and his co-workers. Figure 2, which is taken from McKeown (1976), shows the rapid, continuous downward trend of mortality due to measles for children aged under 15 in England and Wales since 1850. Even though mortality from this disease is due largely to secondary infection which has only been treated by chemotherapy since 1935, eighty-two per cent of the decrease in deaths had occurred before this date. McKeown (1976, 99) concludes that:

> 'immunisation and treatment contributed little to the reduction of deaths from infectious diseases before 1935 and over the whole period since the cause of death was first registered (in 1838) they were much less than other influences'.

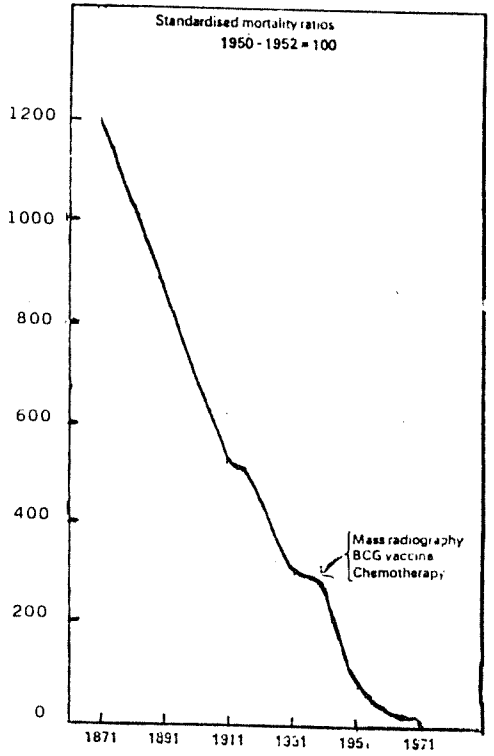
Figure 1b, which shows Figure 1a transformed from a distorting logarithmic to a linear scale, underlines this conclusion. Effective treatment of tuberculosis on a substantial scale did not begin until 1954, by which time mortality from the

FIGURE 1 MORTALITY FROM TUBERCULOSIS ENGLAND AND WALES 1871-1971

(a) logarithmic scale



(b) linear scale



SOURCES: (a) DHSS(1976)

(b) RADICAL STATISTICS HEALTH GROUP(1976)

FIGURE 2 MEASLES DEATH RATES FOR CHILDREN UNDER 15

ENGLAND AND WALES

SOURCE:MCKEOWN(1976)

disease had fallen to a fraction of its mid-nineteenth century level.

If it is to be argued that interventionist medicine played only a small part in the decline of mortality, what was it that brought about the substantial reduction in the death rate? Of undoubted importance was the progressive introduction of water purification, efficient sewage disposal and improved hygiene in the second half of the nineteenth century. McKeown (1976) believes that these essentials were supported in the twentieth century by advances in domestic, working and general environmental conditions; he places particular emphasis on improved nutrition as a factor in the declining death rate. Clearly, according to this view the successes of interventionist medicine appear somewhat illusory, while the importance of the general environment has been considerably underplayed.

Turning to mortality trends in the present century, the major failure of modern medicine has been its inability to reduce premature death in men. In several countries male death rates in middle age have been rising, while in others they have failed to decline. In England and Wales during the last eighty years the life expectancy of males has hardly changed: in 1901 the 45 year old male could be expected to live another 23 years and by 1971 this had only increased by 2 years. Coleman (1977, 12) states his view bluntly:

> 'medical research has had surprisingly little effect on life expectation. Improvements in human health in the late nineteenth and early twentieth century have been largely due to changes in the environment and modification in human behaviour'.

The theologian and philosopher Ivan Illich has been even more forthright, arguing in Medical Nemesis (1974) and Limits to Medicine (1977) that the massive expansion of modern medicine has had a largely damaging effect on peoples' health prospects. Illich bases his case partly on evidence of injury

caused by medical treatments (clinical iatrogenesis) and partly on studies which have suggested that a rise in living standards has contributed much more to the improvements in health than the advent of chemotherapy and other medical techniques. His main accusation against medicine, however, is that it 'expropriates' peoples' health by creating dependency on medical interventions and removing the essential coping capacities of individuals and small local communities.

In view of these arguments, it is not perhaps surprising that many eminent medical authorities are urging a return to the environmental approach to disease. For example, J. Higginson (1967), who is the World Health Organisation director of the International Agency for Research on Cancer, has estimated that approximately two-thirds of cancer tumours in the economically advanced societies are a result of environmental factors. Sir Richard Doll, Regius Professor of Medicine at Oxford has written that

> 'the marked differences in cancer incidence in different countries and the changes that have been noted in migrant groups when they move from one country to another are among the many pieces of evidence suggesting that most cancers are due to environmental factors. It follows that most cancers are, in principle, preventable' (Doll, 1969, 8).

The present work represents an attempt to study disease from this environmental viewpoint and, in particular, it seeks to establish which variables in the physical and social environment are associated with high mortality from chronic disease in England and Wales.

Aggregate data analysis

The second supposition that needs to be discussed is the preferred use of aggregate data to analyse the causes of mortality variations. One of the major legacies of the germ-theory approach to medicine is the current emphasis on

laboratory experimentation. If a researcher wishes to investigate the relationship between two variables, the laboratory experiment allows great measurement precision and facilitates the control of extraneous influences that may have confounded the relationship between the pair of variables under scrutiny. Moreover, because of rigorous laboratory control, the analyst can usually set one variable (the independent or explanatory variable) to a set of fixed values to determine precisely the nature of the relationship between the independent variable and the response or dependent variable. But the method suffers the severe limitation that, in the majority of instances, the classic experiment must be confined to animals other than humans and thus the human condition can only be simulated. Many reactions are now known to be species-specific and, because the same stimuli may give different results in different species, any generalisation from animal experiments must be cautious. Mindful of this problem, epidemiologists have developed a number of ways of studying disease in human groups. One such approach is based on aggregate areal data and is variously known as the 'cross-sectional' approach (Alderson, 1976), the 'ecological' approach (Susser, 1973) and even the 'geographical' approach (Lowe, 1969); in geography such an approach falls within the tradition of what Haggett, Cliff and Frey (1977) call aggregate, macro-geographical modelling. One big advantage of dealing with groups rather than individuals is that officially collected data are frequently readily available for a large number of areas. Moreover, in the case of England and Wales at least, the data are available for a considerable period of time and are reputed to be of high quality. However, the major drawback is the so-called ecological fallacy of inferring that a strong association at an areal level necessarily holds at an individual level. While this problem is undoubtedly an important one (and it will receive an extended treatment in Chapter 8) it is

possible to strengthen inferences based on aggregate data by the adoption of a number of approaches: replication, examination of aggregate results over time, and the search for confounding variables by some form of causal analysis. Each of these approaches will be briefly considered to illustrate the value of an aggregate analysis.

While the analyst of aggregate data does not have the opportunity for exact replication by which the laboratory analyst can demonstrate the consistency of a relationship, he can analyse his data in such a way as to discover whether the relationship holds for different age groups or different areas. Obviously the arguments for the validity of a relationship are strengthened when diverse approaches produce similar results. Durkheim's (1952) classic study of suicide is an excellent example of this approach. On the basis of official statistics for the nations of Western Europe, Durkheim examined the association between suicide rates and such factors as age, sex, religion, occupation, marital status and economic cycles. One of the main findings was high suicide rates in Protestant countries and low rates in Catholic ones and, to show that this association was independent of the many other differences between Catholic and Protestant countries, he separately analysed the data for different age groups and regions. In Switzerland the association with religion held for each language group, in Prussia and Bavaria for each province and, in all for 17 essentially independent comparisons, thereby considerably strengthening the arguments for the relationship.

As noted earlier, in the laboratory experiment the researcher can deliberately intervene to alter the level of the explanatory variable to assess the effect on the response variable. While deliberate intervention is not usually possible with aggregate data, by examining relationships temporally it is often possible to conduct an 'experiment of opportunity'. John Snow (1855) in his classic book On the

Mode of Communication of Cholera postulated that water supply contaminated by sewage caused the disease, and he began his studies by examining cholera deaths for administrative areas. Cholera mortality rates in the 1849 outbreak were tabulated in rank order for the districts of London and against these rates were set the companies responsible for the water supply. The table showed an obvious association between high mortality and those areas served by the Southwark and Vauxhall Company and the Lambeth Waterworks. In this way a direct relationship was established at the areal, aggregate level, but Snow went further to point out that by the time of the 1853 outbreak a change had taken place. The Southwark and Vauxhall water supply was still associated with high cholera mortality but incidence of the disease was now relatively low in the Lambeth-supplied areas. On further examination it was found the Lambeth company had moved their water intake to a point higher up the Thames, thus obtaining a supply of water free from London's sewage. Such evidence was of considerable support to Snow's theory of cholera transmission by contaminated water.

In laboratory experiments control is exercised in such a way that the relationship between two variables can be studied without the confounding influence of other variables. In aggregate analysis such rigorous control is seldom possible and the research must therefore attempt to remove the influence of confounding variables by some method of statistical analysis. An illustration of this approach may begin with a simple case: two variables x and y are found to occur in association, and we may postulate that

$$x \xrightarrow{\text{causes}} y$$

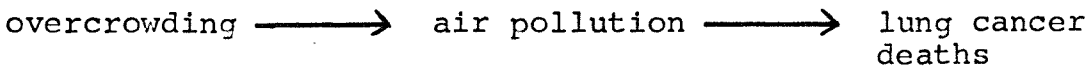
or

$$\text{air pollution} \longrightarrow \text{lung cancer deaths.}$$

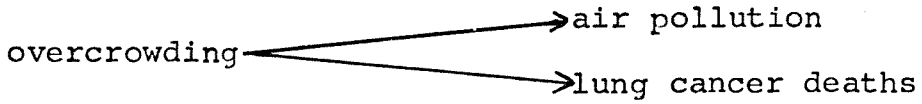
Such an observed association between two variables can occur for one of three reasons: there may be a true causal relationship between the two variables, the association may occur simply by chance, or x may co-vary with y because x and y are jointly caused by some third factor (z). Examining the latter reason for an observed relationship in more detail, and taking crowding as the variable z, it is possible to distinguish three types of relationship between the variables.

Firstly, there is the possibility that

Here both overcrowding and severe air pollution are the causes of high death rates from lung cancer[1], and if such a relationship exists, the lowering of overcrowding and air pollution levels would result in lower death rates from the disease, with the lowest rates occuring when both crowding and air pollution were reduced. The second model that can be postulated is



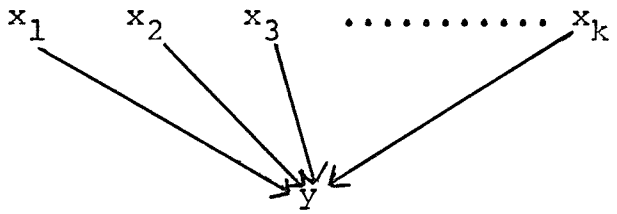
This conceptualises air pollution as a link between overcrowding and cancer deaths, and when this model is substantially correct, the reduction of both overcrowding and air pollution will result in fewer deaths.[2] The third and final model is



Here, overcrowding causes both air pollution and cancer deaths but there is no true causal relationship between pollution and death. If this is the appropriate model, lowering air pollution will have no effect on cancer deaths, and such an association, which disappears when a third variable is controlled, is termed spurious. It is obviously

essential that if there is deliberate intervention to lower death rates that decisions are not made on the basis of such false relationships.

If the relationship between two variables is not weakened, eliminated or shown to be spurious after several attempts to control for possible confounding variables, then an analyst can have more confidence in the observed relationship. The greater the number of different ways he attempts to falsify a relationship and fails the more confident he can be that the relationship is a truly causal one. One popular method of achieving such control is by using the statistical method of multiple regression. The causal model behind multiple regression is that there are k independent causes of y, the dependent variable or



Testing or attempting to disprove ideas and theories is achieved by trying to assess the independent effect of one of the explanatory by'holding constant' all other explanatory variables. While multiple regression will be the main method of research in this thesis, it must always be remembered that the process of 'holding constant' is achieved statistically by the manipulation of the observed aggregate data. Ideally, rigorous experimental manipulation and control are to be preferred, but given the current state of development of the subject, the problems of data availability and the wish to study humans not mice, in many cases there is simply no other way to investigate important questions.

Exploratory statistics

The final underlying supposition of this thesis that has to be considered is the assumed usefulness of exploratory

statistics. The exploratory approach will be considered in some detail in Chapter 1, but it is appropriate at this stage to outline the distinctive character of the approach. Traditional confirmatory statistics of the type commonly used in geography are very demanding in two ways: firstly, they are based on a number of exacting, rigid assumptions that may not exist in the real world and, secondly, they require the researcher to have a fully developed theory. Returning to the lung cancer example, a researcher must not only be able to postulate all the causes of the disease, he must also specify exactly the functional form of the relationship! Frankly, the analyst of areal mortality does not have such theories and the approach of exploratory statistics therefore deserves to be considered as a means of solving such difficulties.

Although the exploratory data analyst has some novel statistical tools in his kitbag, it is not the actual techniques that distinguish exploratory analysis from other approaches. Instead, it is a certain perspective that characterises the exploratory approach, and this perspective is based on two key principles: scepticism and openness. The exploratory analyst is sceptical of all the causal models which he postulates and estimates and, in particular, he is sceptical of using a summary measure to characterise any data set. Moreover, he openly admits that he does not know the exact form of the model relating the dependent variable to the explanatory variables, and he freely reports the steps that have been taken to build an improved model of this relationship. Because of this scepticism towards statistical summaries of data, major emphasis is placed on visual display and there are several graphical techniques which, taken together, are the data analyst's most powerful tools. Because of the'trial and error' approach to model building, the exploratory analyst may forfeit the right to test the statistical significance of the final model and, consequently, the lack of formal hypothesis testing is another distinctive feature of the exploratory perspective. In essence, the openness and scepticism principles imply a

flexible approach which is open to alternative model formulations and which emphasizes the visual representation of data.

Thesis organisation

Having considered the possibilities offered by an environmental, aggregate, exploratory approach to mortality, it is pertinent to consider the structure and presentation of the thesis. This will not be done in detail for the aim is to outline briefly the overall organisation of the work; a fuller consideration is postponed until the end of Chapter 1. Part I of the thesis is mainly concerned with detailing the exploratory approach to geographical data. Firstly, the difference between exploratory and confirmatory statistics is examined in some depth, after which the argument progresses to examine the technique of multiple regression and each of its assumptions. For each assumption a common approach is adopted: the effects of breaking the assumption are considered first and then ways of detecting and overcoming the resultant problems are discussed. The exposition is illustrated by the re-analysis of areal mortality data previously examined by other researchers and by the analysis of simulated data. The use of simulated data with known problems is designed to exemplify the usefulness of exploratory procedures. The final chapter in this technical part of the thesis is a discussion of the particular problems posed by the analysis of ratios. Previous researchers have frequently used ratios (for example, death rate) in their work but they have usually treated them as raw numbers. Regrettably, as the discussion will show, this may have unfortunate implications for the validity of their work.

Part II of this thesis is concerned with the application of methods introduced in Part I to the analysis of areal mortality data. Firstly, a critical statistical appraisal of previous studies concerning the relationship between

water hardness and heart disease is undertaken. This relationship is one of the mostly fully researched topics in epidemiology, but as the discussion will show, quantitative techniques have been poorly applied. The discussion then proceeds to an analysis of deaths from a variety of causes for the County Boroughs of England and Wales. Exploratory methods are used in order to provide an improved model of the observed geographical variations in mortality. The last chapter presents conclusions and outlines areas of further research; particular emphasis is placed on the evaluation of results from aggregate analyses.

INTRODUCTION: NOTES

1. In laboratory studies, overcrowding amongst animals has been found to lead to severe stress and disease; it has also been suggested that such stresses will occur amongst humans and that high population densities facilitate the rapid spread of infection.

2. Overcrowding tends to be associated with air pollution because areas of high density living have a relatively high density of domestic chimneys and such areas are frequently adjacent to inner-city industrial areas.

INTRODUCTION: BIBLIOGRAPHY

ALLEN-PRICE, E.D. (1960): Cancer in a Devon village
Lancet 1, 1235-1239.

ALDERSON, M. (1976): An introduction to epidemiology
MacMillan, London.

AMIEL, B. (1970): Death is a matter of mathematics
in Jones, D.L. (ed.)
War Poetry: an anthology
Pergamon, Oxford.

CATALANO, R. (1979): Health, behaviour and the community
Pergamon, New York.

COLEMAN, V. (1977): Paper doctors: a critical assessment of medical research
Temple Smith, London.

de HAAS, J.H. (1968): Geographical pathology of the major killing disorders: cancer and cardiovascular disease in
Wolstenholme, G. and O'Connor, M. (eds.) Health of Mankind
Churchill, London.

DHSS (1976): Prevention and health: everybody's business
HMSO, London.

DOLL, R. (1969): The geographical distribution of cancer
British Journal of Cancer 23, 1-8.

DUBOS, R. (1955): Second thoughts on the germ theory
Scientific American 192, 31-35.

DURKHEIM, E. (1952): Suicide
Routledge and Kegan Paul, London.

GIRT, J.L. (1972): Simple chronic bronchitis and urban ecological structure
in McGlashan, N.D. (ed.)
Medical geography
Methuen, London.

GRIFFITH, G.T. (1967): Population problems in the age of Malthus
second edition, Frank Coss, London.

HABAKKUK, H.J. (1953): English population in the eighteenth century
Economic History Review 6, 117-129.

HAGGETT, P., CLIFF, A.D. and FREY, A. (1977): Locational analysis in human geography
Arnold, London.

HARTWIG, F. and DEARING, B.E. (1979): Exploratory data analysis
Quantitative applications in social sciences
16 Sage,
Beverley Hills.

HIGGINSON, J. (1967): The role of geographical pathology in cancer research
Schweizerische Medizinische Wockenschift 18, 565-568.

HIRSCHI, T. and SELVIN, H.C. (1966): False criteria of causality in delinquency research
Social Problems, 254-268.

ILLICH, I. (1974): Medical nemesis
Calder Boyars, London.

ILLICH, I. (1977): Limits to medicine
Penguin, Harmondsworth.

JONES, K. (1975): A geographical contribution to the aetiology of chronic bronchitis
unpublished B.Sc. dissertation, Dept. of Geography, University of Southampton.

LOWE, C.R. (1969): Industrial bronchitis
British Medical Journal 31, 463-486.

McKEOWN, T. (1976): The modern rise of population
Arnold, London.

McKEOWN, T. and BROWN, R.G. (1955): Medical evidence related to English population changes in the eighteenth century
Population Studies 9, 119-141.

McKEOWN, T., BROWN, R.G. and RECORD, R.G. (1972): An interpretation of the modern rise of population in Europe
Population Studies 26, 345-382.

McKEOWN, T., RECORD, R.G. and TURNER, R.D. (1975): An interpretation of the decline of mortality in England and Wales during the twentieth century
Population Studies 29, 391-422.

POWLES, R. (1973): On the limitations of modern medicine
Science, Medicine and Man 1, 1-30.

Radical Statistics Health Group (1976): Whose priorities?
BSSRS, London.

ROSEN, G. (1958): A history of public health
M.D. publications, New York.

SNOW, J. (1885): On the mode of communication of cholera
Churchill, London.

SUSSER, M. (1973): Causal thinking in the health sciences
Oxford University Press, New York.

PART I

AN EXPLORATORY APPROACH TO GEOGRAPHICAL DATA ANALYSIS

Chapter

C H A P T E R I

INTRODUCTION TO CONFIRMATORY AND EXPLORATORY DATA ANALYSIS

'The multiple regression model is often assumed uncritically without any questioning of its assumptions. Statistical practioners should be aware of the alternative definition of regression given by the New Standard Dictionary as the 'diversion of psychic energy... into channels of fantasy ' Bibby (1977, 41).

'Even in social science, where some of the most rigid practioners of formal statistical inference are to be found, there is a growing awareness of the need for exploratory data analysis' McNeil (1977, vi).

'It would seem far better for journal reviewers to become more lenient in recommending acceptance of exploratory research clearly labelled as such and to be more suspicious of research which claims not to be of this nature' Freund and Debertin (1975, 722).

A RESEARCH PROBLEM

The relationship between water hardness and heart disease mortality is probably the most contentious issue in modern day epidemiology. For the last twenty years it has been the subject of much debate in the popular as well as the academic literature. One of the major reasons for this substantial body of research is that there is considerable disagreement over the strength and nature of the relationship between water hardness and heart disease. Consider the following quotations.

(1) 'It is now established that there is a statistical association between mortality, in particular from cardiovascular disease, and the hardness of drinking water' (Crawford, Gardiner and Morris, 1971, 327).

These workers found a negative relationship between the number of people dying from heart disease and the hardness of water (see Figure 1.1a).

(2) 'The results suggest that variations of water hardness do not materially contribute to the explanation of major differences in the extent of arterial atherosclerosis' (Strong, Correa and Solberg, 1968, 622).

These workers found no relationship (see Figure 1.1b).

(3) 'Fewer persons with soft water at home and more persons with hard water at home died from atherosclerotic heart disease' (Comstock, 1971, 6).

This author found a positive relationship (see Figure 1.1c).

The research problem now becomes 'why do we get such conflicting results?' One answer to this question is that theory is little developed in epidemiology (it is usually a

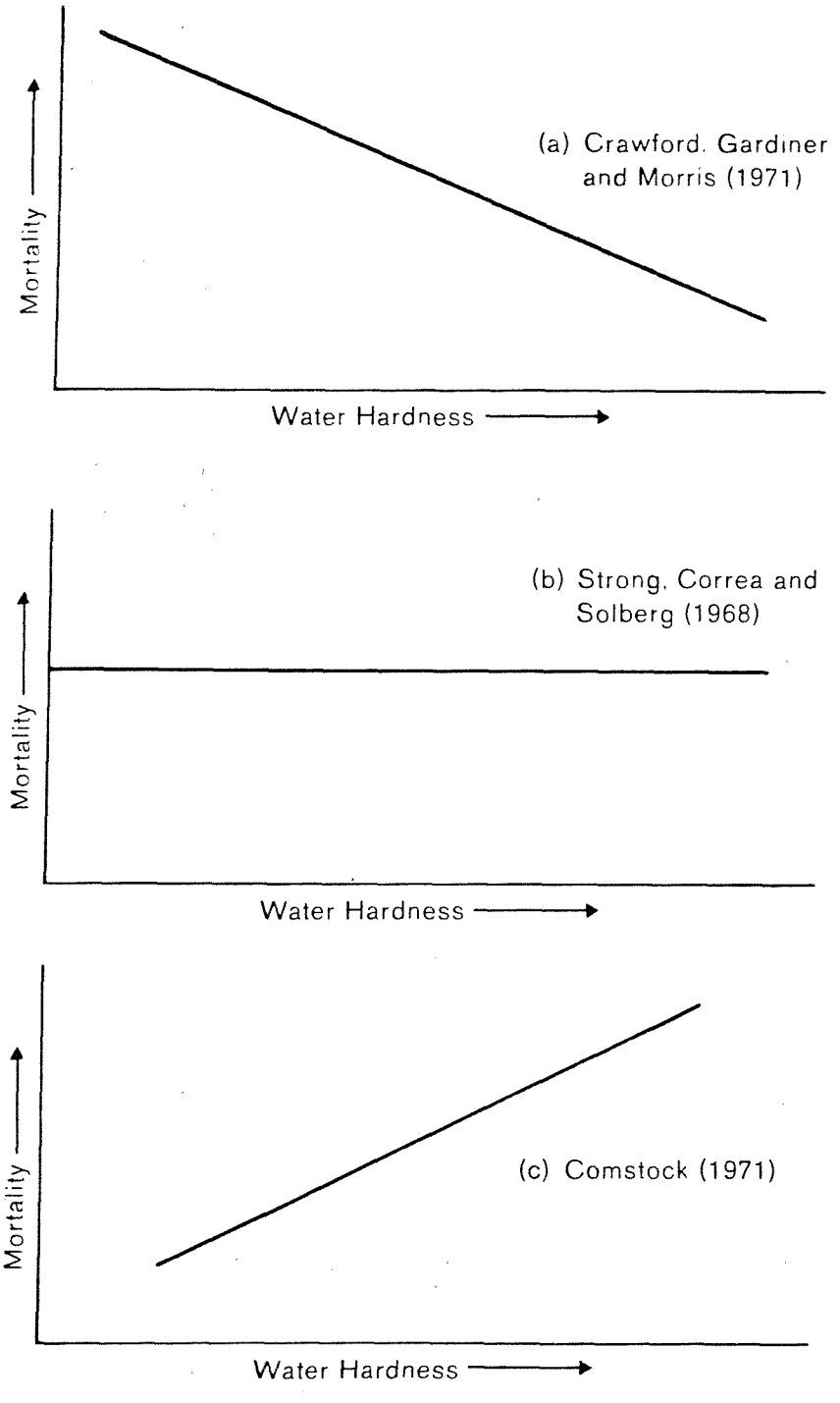


FIGURE 1.1 FUNCTIONAL RELATIONSHIPS REVEALED BY THREE STUDIES OF WATER HARDNESS AND MORTALITY

vague notion that one variable is related to another variable, everything else held constant) and the problem arises when this vague notion is translated into

$$Y = X \beta + \varepsilon \qquad (1)$$

the general linear model[1]. For, as will be illustrated later, the majority of research examining the effect of water hardness has been based on regression or correlation methods using areal mortality data.

In the present chapter, which serves as an introduction to the methodological part of the research, the discussion will begin with a basic introduction to regression analysis. In this section emphasis will be placed on the assumptions of regression modelling and a standard notation will be introduced. This will be followed by a consideration of the use of regression models in a confirmatory manner. While such an approach will be shown to have major drawbacks, another approach based on exploration will be shown to have considerable potential in geographical research. Finally, the chapter will attempt to elucidate the role of exploratory statistics in seeking an explanation of areal mortality patterns.

THE OLS REGRESSION MODEL[2]

Medical geographers and epidemiologists in their attempts to explain mortality patterns have tried to relate mortality to a set of environmental variables. Many have found it helpful to express such relationships in a functional form - in the form of an equation that describes how changes in the environmental variables will result in a change in the mortality variable. Undoubtedly, the most flexible and widely used technique for analysing such a relationship is regression analysis.

Consider the simple case of <u>one</u> explanatory variable (the independent variable) and a variable to be explained (the dependent variable). The explanatory variable is commonly denoted by x_1 and the dependent variable by y. To examine the linear functional relationship between x_1 and y, a straight line equation has to be fitted to these two variables. To do this one requires an intercept term β_o and a slope term β_1; β_o represents the value of y when x_1 is equal to zero and β_1 is the amount of change in y brought about by a change in x_1 (Figure 1.2). The β_1 or slope term is the critical term in the regression model. If it is positive, y increases as x_1 increases; if it is negative, as x_1 increases y decreases; if it is exactly zero, any change in x_1 will leave y unaffected.

The simple linear regression model can be stated verbally as:

observed value of the dependent variable = intercept term + [slope term * observed value of explanatory variable] + disturbance term

The disturbance term is included in the model because with real world observations it is unlikely that the variation in x_1 will completely account for the variation in y. Mathematically, this verbal statement of the simple linear regression becomes:

$$y_i = \beta_o + \beta_1 x_{1i} + \varepsilon_i \qquad (2)$$

where there are i = 1, 2 .... n observations.

This simple linear regression model can be easily expanded into a multiple regression equation relating a dependent variable to <u>several</u> explanatory variables:

$$y_i = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + ... \beta_m x_{mi} + \varepsilon_i \qquad (3)$$

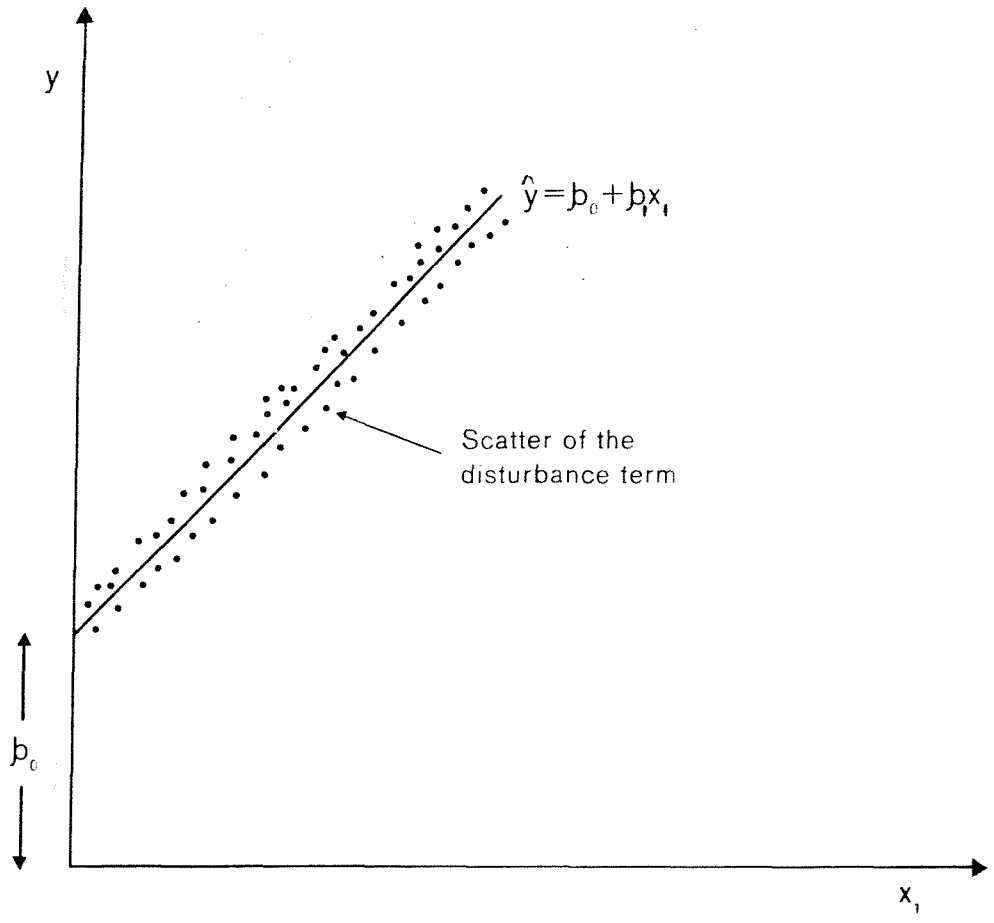


FIGURE 1.2 A LINEAR RELATIONSHIP BETWEEN TWO VARIABLES

where there are i = 1, 2 .... n observations and m explanatory variables in total. Each of the m slope terms (the partial regression coefficients) can now be interpreted as measuring the change in the dependent variable for a unit change in each associated explanatory variable holding the other variables constant. For example, in a multiple regression model with heart disease as the dependent variable and air pollution, social class and water hardness as the explanatory variables, the partial regression coefficient associated with air pollution measures the relationship between air pollution and heart disease, holding the other two explanatory variables statistically constant. The partial regression coefficients therefore potentially represent a means of assessing the importance of a particular explanatory variable in accounting for the variation of the dependent variable while also taking into account the separate influence of other explanatory variables.

The algebraic form of the multiple regression model (equation 3) can be more succinctly written in matrix terms[3]:

$$Y = X\beta + \varepsilon \qquad (4)$$

The vector Y consists of the observed dependent variable, X is a matrix of independent variables, β is a vector of unknown regression coefficients and ε is the disturbance term in vector form. (Capital letters will be used to denote matrices.)

The multiple regression model given in equations (3) and (4) is called the <u>population</u> regression model and represents the underlying true model of the relationships that exist in the real world. While there is only one population model, an infinite number of regression models can be postulated to relate a dependent variable to a set of independent variables. Consequently, in practice, a postulated model has to be estimated, evaluated and

reformulated until it approximates to the true population model. It is only when this has been accomplished that inferences can be drawn about the population model on the basis of the postulated model. To distinguish between the two types of model, the postulated model is to be written algebraically as

$$y_i = \hat{b}_o + \hat{b}_i x_{1i} + \hat{b}_2 x_{2i} + ... \hat{b}_m X_{mi} + \hat{e}_i \qquad (5)$$

and in matrix terms as

$$Y = X \hat{B} + \hat{E} \qquad (6)$$

Comparing the postulated algebraic model (5) with its population equivalent (3) it is noticeable that the population partial regression coefficients (β) have been replaced by the estimated equivalents ($\hat{b}$) and the disturbance term ε has been replaced by its estimate, the residual term $\hat{e}$. Thus, the numerical values for the regression coefficients of the postulated model ($\hat{b}$) are <u>estimates</u> of the true but unknown population values (β).

The estimation of the postulated model is usually performed on observed data according to the <u>least-squares criterion of goodness of fit</u>. Estimation is performed in such a manner that the sum of the squared differences between the fitted values of the dependent variable ($\hat{Y}$) and the actual values of the dependent variable (Y) is minimised. The differences between the fitted and actual values of the dependent variable is the residual term $\hat{E}$. Therefore, the regression coefficients are usually estimated so that the sum of the squared residuals is minimised. The formula for this estimation in matrix terms is:

$$\hat{B} = (X' X)^{-1} X' Y \qquad (7)^4$$

These particular estimates are known as the <u>unstandardised</u> partial regression coefficients and are dependent on the scale of measurement of the independent variables. If an

independent variable is first measured in metres and subsequently the units are changed to kilometres, the associated regression coefficient will increase a thousand-fold. Because of this dependence on the scale of measurement, such unstandardised regression coefficients cannot be used to compare the relative importance of different independent variables in accounting for the variations in the dependent variable. To overcome this problem, partial regression coefficients are usually estimated in standardised form:

$$\hat{B}_{(s)} = [R]^{-1}.G \qquad (8)^5$$

where $\hat{B}_{(s)}$ is the vector of the estimated standardised partial regression coefficients

$[R]^{-1}$ is the inverse of the matrix of correlations between the independent variables and

G is a vector of correlations between the dependent and independent variables.

Each standardised regression coefficient measures the change in the dependent variable (in standard-deviation units) for a unit change in each explanatory variable (in standard-deviation units), holding all other explanatory variables constant[6].

Having briefly discussed the estimation of postulated regression models according to the least-squares criterion, it is essential for later discussion that the reasons for choosing this particular criterion of fit be examined. If a large number of postulated models are calculated on the basis of different samples of data, and if these postulated models are correctly specified (that is, they have the same functional form as the population model) there will still be some variation in the estimated regression coefficients brought about by the stochastic variation inherent in real world data. However, it is obviously desirable that estimated regression coefficients are as close as possible to the true population values. In particular, in a series

of estimated models, it is desirable that the average of the estimated regression coefficients should be equal to the true population value. If a particular estimator has this desirable characteristic it is said to be unbiased (Figure 1.3). Another important characteristic for an estimator is that, when used with an infinite number of samples of data, its distribution should have the minimum spread around the true population value. An estimator having this property is said to be the minimum variance estimator. (If an estimator does not have this minimum variance it is said to be inefficient - Figure 1.3.) The least-squares estimators are chosen because they are the 'best' (minimum variance), linear, unbiased estimators (BLUE). Ordinary least-squares (OLS) regression coefficients, estimated on sample data, will provide estimates that are closer to the true population values than any other type of estimate. However, this is only the case if certain basic assumptions of OLS regression models can be fulfilled.

THE ASSUMPTIONS OF THE OLS REGRESSION MODEL

In the following discussion it is assumed that OLS regression is being used for developing and testing hypotheses about linear functional relationships on the basis of sample data[7]. Although the model is based on a sample, the assumptions are usually expressed in relation to the underlying population model. The assumptions are then stated as follows.

(1) The form of the functional relationship that is fitted to the sample data must be the functional form that exists in the population. If a linear model is fitted to the sample data, but the true population model is non-linear, it is likely that incorrect inferences will be drawn.

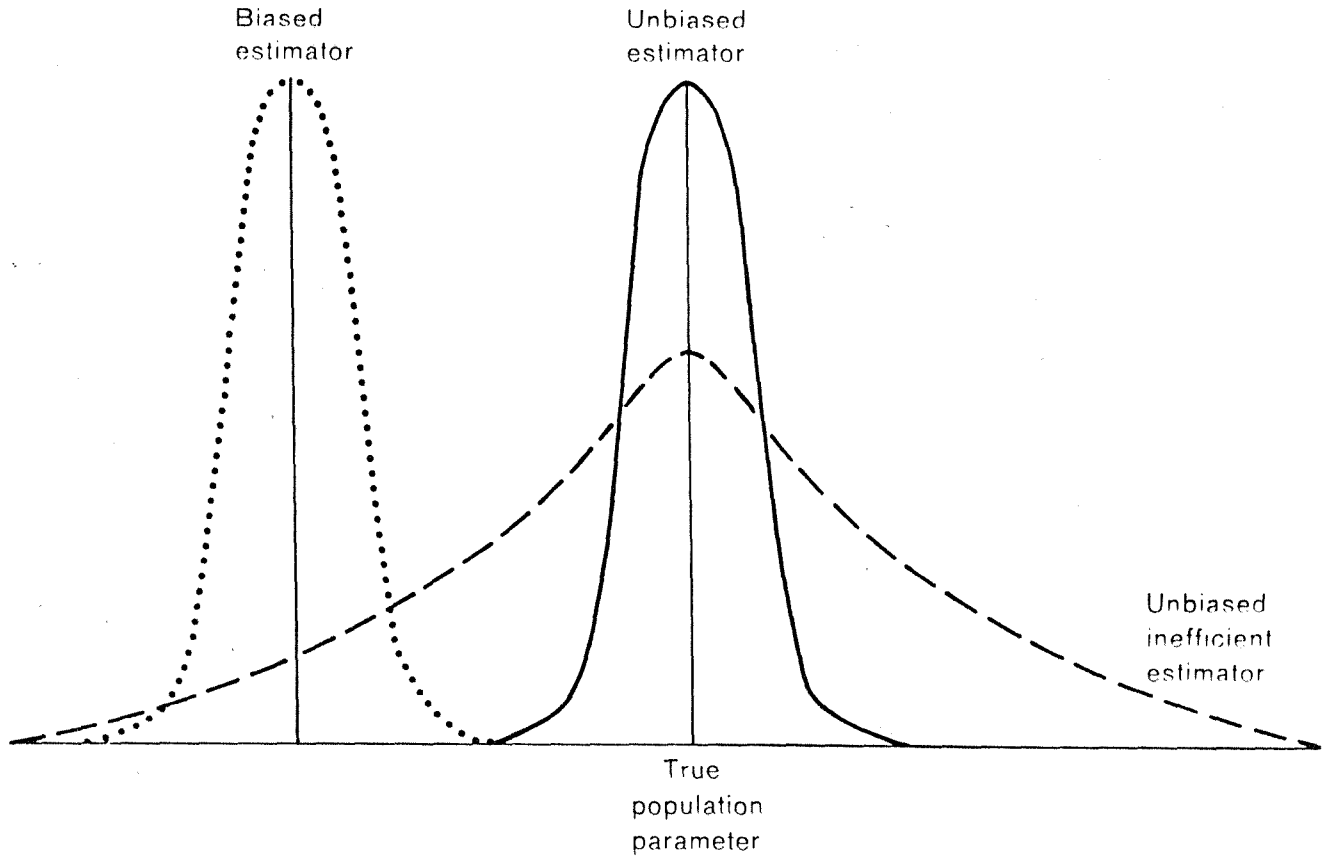


FIGURE 1.3 THEORETICAL SAMPLING DISTRIBUTIONS OF ESTIMATORS

(2) It is assumed that the mean (average or expectation) of the true, but unknown, vector of disturbances of the population model is zero. This assumption follows logically from considering the disturbance term as being the outcome of the sum of those forces acting on the dependent variable which have not been explicitly included in the regression model. If these forces are influencing the dependent variable randomly then, on average, these forces can be expected to cancel each other. That is, there is an expectation of zero. In practice, this assumption means that no important explanatory variable must be omitted from the postulated model. If this is not the case, inferences concerning the population model, drawn on the basis of the postulated model, can be expected to be incorrect.

(3) At different levels of the explanatory variables the variance or spread of the disturbance term is expected to be constant. For example, the variance of the disturbance term for high values of x_1 should equal the variance at low values of x_1. This is the assumption of homoscedasticity and if this assumption is not fulfilled the disturbances are said to be heteroscedastic.

(4) The values of the disturbance term are expected to be independent of each other. If this assumption does not hold, auto-correlation is said to be present.

(5) The data used to estimate the model should be observed with negligible measurement error.[8]

(6) There must be no exact linear relationship between any of the explanatory variables. If this assumption does not hold, multicollinearity is present in the data.

(7) The final assumption is that the disturbance term is normally distributed. This assumption is usually employed to facilitate the evaluation of a model by significance tests.[9]

Poole and O'Farrell (1971) have shown that quantitative textbooks in geography generally do not fully consider the assumptions of the regression model. Such textbooks are even less helpful if the data fails to meet these assumptions. Daniel and Wood (1971, vii) write of statistical texts:

> 'Standard texts give little guidance beyond stern warnings to be cautious'.

Consequently, one focus of the present research is the detection and correction of violations of the assumptions of the OLS regression model in order to achieve a more informative and complete analysis of a data set. But before dealing in more detail with each of the assumptions of the regression model, it is vital to distinguish between two fundamentally different approaches to regression analysis - confirmation and exploration.

A CONFIRMATORY APPROACH TO DATA ANALYSIS

Introduction

The statistical approach that is most commonly adopted by geographical researchers is that of confirmatory analysis. In essence, this traditional statistical methodology uses a sample to confirm a hypothesis that has been postulated *a priori* by the researcher. Not only is this approach the one most commonly adopted by geographical researchers it is also the approach adopted by statistical workers in general. As Gnanadesikan (1977, 196) has written:

> 'pedagogy, publications and, more generally, the codification of statistical theory and methods have been concerned almost exclusively with formal procedures such as tests of hypotheses, confidence - region estimation, and variance optimality criteria'.

In the following discussion the nature of confirmatory analysis will be considered and then the major drawbacks of the approach will be outlined. It will be shown that these drawbacks have forced researchers to use confirmatory analysis in an exploratory manner with concomitant undesirable results. Finally, the need for a truly exploratory approach will be briefly outlined.

The nature and drawbacks of confirmatory analysis

With regard to regression analysis, the standard confirmatory approach is to propose a hypothesis, collect data on this hypothesis and then fit a regression model to this data. The regression model is evaluated in terms of how much of the variation in the dependent variable can be accounted for by the variation in the explanatory variables (the coefficient of determination, R^2) and by a significance test of the overall goodness of fit of the model (the F test). Individual hypotheses concerning the relative importance of a particular explanatory variable are tested by assessing the degree to which a particular regression coefficient associated with an explanatory variable is significantly different from zero (the t test). The main advantage of such an approach to regression modelling lies in its exactness; using probability theory one can calculate exactly the probability of making an inferential error. While confirmatory statistical methods are undoubtedly 'one of the greatest intellectual products of our century' (Tukey, 1977, 7) and they can confirm a few things exactly, they can only do so under very specific circumstances. Indeed as statisticians promoted exact confirmation, their techniques became more and more inflexible.

It is possible to recognise three major drawbacks to the confirmatory approach. Firstly, the penalty paid for the exact nature of confirmatory analysis is that the data must meet a number of rigid and exacting assumptions that may or may not be fulfilled in the true underlying population model (the 'real-world'). Moreover, practitioners all too easily have neglected the crucial proviso 'if appropriate assumptions hold' and thus there has been an overemphasis on the exactness of the approach for the idealised problem.

The second major drawback of the confirmatory approach is that the a priori model or hypothesis must be well-specified. As Leamer (1978, v) has written:

> 'Traditional statistical theory assumes that the statistical model is given. By definition, non-experimental inference cannot make this assumption, and the usefulness of traditional theory is rendered doubtful'.

With regard to regression modelling, confirmatory analysis assumes that the researcher knows the exact nature of the functional relationship between the dependent and independent variables and that no explanatory variables have been omitted from the model. Indeed, the t test, the F test and the coefficient of determination which are the methods used to test hypotheses and evaluate a model, can only be legitimately used if the model has been correctly specified.[10] Medical geography rarely, if ever, has sufficient theory to specify the exact nature of the model and therefore the confirmatory approach is rendered of doubtful value to the medical geographer.

The third and final major drawback of the confirmatory approach is that practitioners of this approach fail to consider the problems of applying a large number of significance tests to the same data. Classical statistical theory does not allow a model to develop by trial and error procedure. The t and F tests associated with the regression model cannot be legitimately used if the model has been chosen after 'exploring' several alternative models on the basis of one data set.

These three major drawbacks associated with the confirmatory approach combine to form a major impasse. A researcher cannot evaluate a model unless it has been correctly specified, and if the model has not been correctly specified and has to be reformulated, then the confirmatory methods of evaluation are again rendered invalid. Faced with

this major impasse, some researchers have attempted to 'bend the rules' and use exact statistical methods in an 'exploratory' manner. Such researchers are continuing the analysis irrespective of breaking the fundamental assumptions of classical statistics.[11] At the most informal this 'exploration' with confirmatory methods involves adding or removing one or two variables from the model or trying a 'squared' version of one of the explanatory variables. At the most formal this approach becomes stepwise regression.

Stepwise regression

Stepwise regression is an automatic confirmatory approach to 'exploring' data and in such a procedure a series of different regression models are fitted to the data, each model consisting of a different set of variables. There are many varieties of stepwise regression, some adding one variable, some deleting one variable at a time, but most of these different varieties allow the computer program to automatically make the decision of which variable to include or exclude at each stage on the basis of a significance test.

The stepwise process begins with a pool of variables that are thought to be related to the dependent variable and a single cycle in a typical forward stepwise procedure involves:

(1) the examination of the variables currently included in the regression equation to see whether any should be deleted (this obviously does not occur in the first cycle); and

(2) the examination of the variables not yet included in the regression model to see whether any variable should be added.

The significance test in this procedure is a F test set at some nominal significance level (α). The stepwise process terminates when all the included variables have coefficients significant at the pre-selected level, while all the excluded variables have failed to reach this level when they

were included in the model.

The problem with such an approach to model building is that exact statistical procedures were not designed for exploratory work and consequently they can often give misleading results. As Pope and Webster (1972) have pointed out, the probability of finding at least one significant regressor, when actually there are none, increases as the number of possible regressor variables increase. If the explanatory variables are not highly inter-related, the true probability of inclusion of a variable into a model is not α but

$$\alpha^* = 1 - (1 - \alpha)^m \qquad (9)$$

where m is the number of explanatory variables. For example, if there are ten explanatory variables and α has been set to 0.05 then the upper bound[12] for the true significance level becomes

$$\alpha^* = 1 - (1 - 0.05)^{10} \qquad (10)$$

which equals 0.4013. Therefore, there is not a five per cent chance of including a variable in the model where none should be but a forty per cent chance of such an error. The conventional use of stepwise regression obviously leads to an over statement of the significance level of a set of regression coefficients. Furthermore, if a model is developed by using a stepwise procedure, the same data on which the model was developed cannot be used to test the adequacy of the model. For as Selvin and Stuart (1966, 21) write, regarding stepwise regression as 'data fishing':

> 'the fish which don't fall through the net are bound to be bigger than those which do, and it is quite fruitless to test whether they are of average size '.

This statement is reinforced when it is realised that a stepwise regression can extract 'significant' variables from

a random series (Freund and Debertin, 1975).

Finally, Wallis (1965, 450) has shown in his study of known functional relationships that

> 'stepwise multiple regression tends to pick variables that confound several independent effects and to build models that are hard to interpret in terms of the real world '.

While the unsuitability of stepwise regression for developing explanatory models will be further discussed in Chapter 2, it has to be emphasised that the results from such a confirmatory approach to exploration are frequently confusing, sometimes misleading and occasionally ridiculous. For example, in a study of perinatal mortality in the local authorities of England and Wales, Ashford and his co-workers (1973) used a stepwise regression model on eighty-eight explanatory variables and the three variables comprising the final model produced by a stepwise procedure were:

(1) standardised death rates, 1968
(2) illegitimate live births, 1960-2
(3) left wing seats in the 1966 election

Evidence enough of Wallis's (1965, 450) statement that stepwise regression results in 'models that are hard to interpret in terms of the real world'.

Confirmatory analysis : conclusions

In conclusion, because of the present lack of theory development in epidemiology and medical geography, it may be argued that it is impossible to apply confirmatory analysis sensu stricto. Moreover, if confirmatory methods are used in an exploratory manner (a task which they were not designed to do), considerable confusion may result. Indeed, if confirmatory analysis actually performs as many workers seem to think it can do -

> 'we would find one equation estimated for every phenomenon and we would have books that compiled these estimates published with the same scientific

> fanfare that accompanies estimates of the speed of light or the gravitational constant. Quite the contrary, we are literally deluged with regression equations, all offering to 'explain' the same event, and instead of a book of findings, we have volumes of competing estimates ' (Leamer, 1978, 4).

This is exactly the problem that is occurring in modern day epidemiological research. As was outlined at the beginning of this chapter, there are a large number of confirmatory studies which give conflicting results. How can we overcome this problem? One possible solution is to adopt an alternative approach to confirmation based on statistical methods which are specifically designed for exploration.

AN EXPLORATORY APPROACH TO DATA ANALYSIS

Introduction

In response to the obvious failings of the confirmatory approach to statistical analysis, statisticians have recently developed new methods of analysis that attempt an interplay between model formulation, model evaluation and the reformulation of the model on the basis of this evaluation. Regression methods are not therefore regarded as providing exact probability assessments but as a means of uncovering patterns in the data. This relatively new approach to data analysis, strongly advocated by John Tukey and his associates at Bell Telephone Laboratories and Princeton University, is called exploratory data analysis[13].

The following discussion of the exploratory approach will begin with the fundamental bases of exploration. Graphical analysis, the use of residuals, the exploratory approach to statistical assumptions and the need for cross-validation, will all be examined. Finally, the discussion will consider exploration in relation to scientific methodology and it will be shown that exploratory analysis is not, as may be supposed, an inductive approach but is essentially deductive,

having marked similarities to the Popperian view of scientific explanation.

Iteration, indication and graphical analysis

At the heart of exploratory data analysis is the realisation that frequently the first model fitted to the data will not be the final one. That is the analyst realises that he has insufficient theoretical knowledge to specify exactly the correct model and consequently several 'sweeps' will have to be made through the data in order to arrive at a final and improved model which closely approximates the population model. Because of such an iterative procedure, the analyst cannot legitimately use significance tests and the associated methods of confirmatory analysis; he must move away from exactness and move towards indication. The analyst wishes to know when there is no doubt of significance and when there is no hope of significance. This is where graphs assume their important role in exploratory statistics, for graphs are an extremely useful method of indication.

Although graphical statistical procedures have been in use for a considerable period (Royston, 1956) it was not until the early 1960s that substantive research on graphical methods was undertaken. This research into graphical methods was considerably motivated by Tukey. In a prophetic article published in 1962, Tukey wrote:

> 'procedures of diagnosis, and procedures to extract indications rather than conclusions will have to play a large part in the future of data analysis. Graphical techniques offer great possibilities in both areas' (p. 60).

The types of graph that have been developed for exploratory data analysis are specialised ones and, because most textbooks in statistics, as well as in quantitative geography, do not discuss such techniques an extensive exposition will be given in Chapter 4. The following section, however,

attempts to consider the essential role of graphs in exploratory statistics.

Graphs perform two major functions:

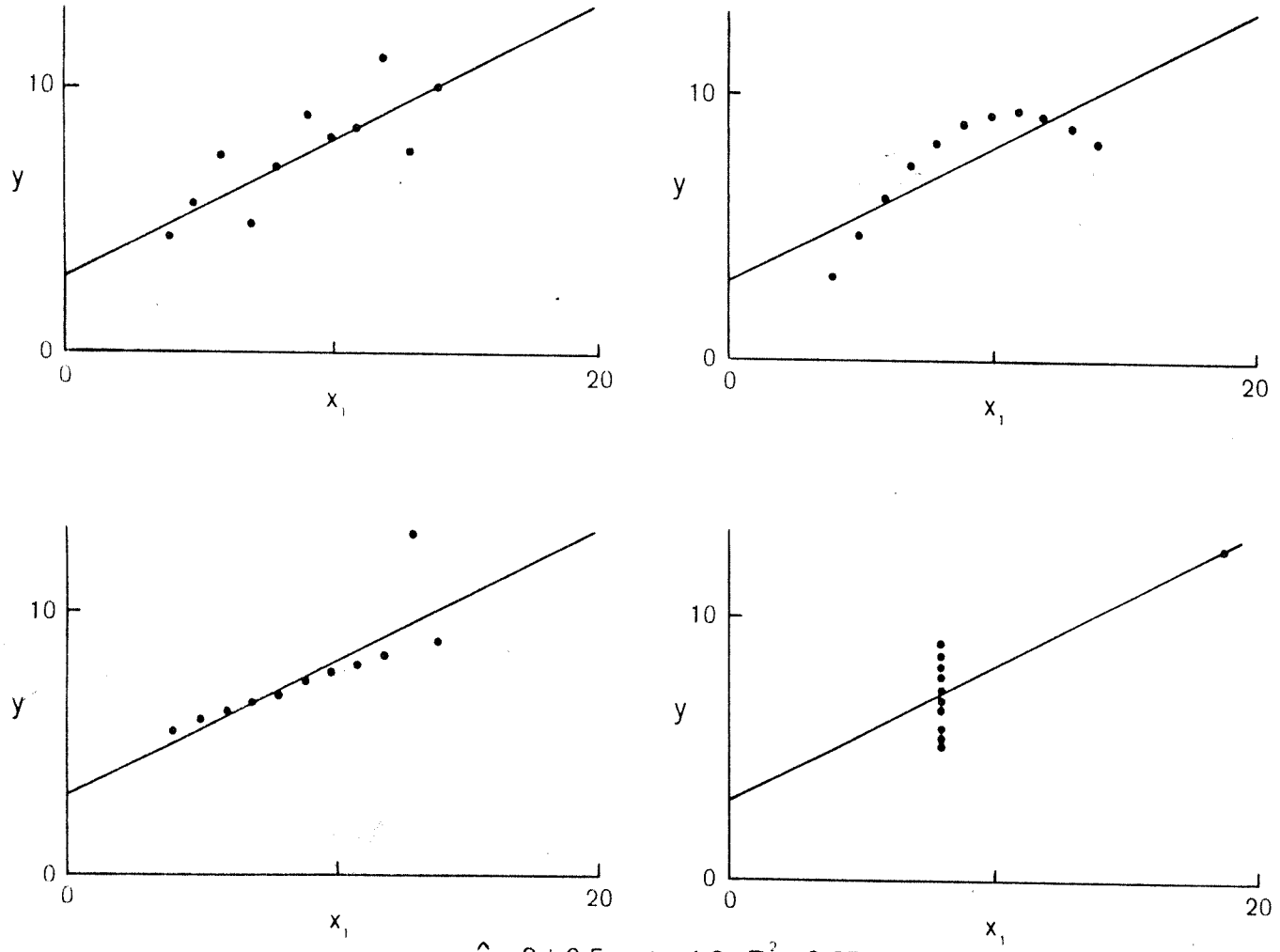
(1) they convey a greater message than a summary statistic

(2) they frequently indicate what the analyst should do next.

Graphical displays show clearly, simply and quickly 'curious' features of a data set. Moreover, graphs may reveal a particular patterning in a data set that a numerical summary statistic cannot possibly convey. An illustration of the ineffectiveness of normal summary statistics as a method of evaluating the validity of a model is given by Anscombe (1973). He constructed four data sets each with exactly the same summary statistics (the same intercept term, the same regression coefficient, the same t, F and R^2 values). An examination of Figure 1.4, however, clearly shows that each data set represents a different relationship between the dependent and independent variables. This difference is clearly not discerned by the normal summary statistics. A significant t statistic and a high R^2 does not ensure that the data has been adequately modelled. The use of graphs in this way is an essential characteristic of exploratory analysis.

As well as advising the analyst of the inadequacy of the postulated model, graphs can often indicate how to develop an improved model. For example, a linear regression of y on x_1 will yield a set of summary statistics that hold true if the relationship between y and x_1 is a linear one. A plot of the two variables will show if the variables are approximately linearly related or whether another form of relationship is more appropriate. If the latter is the case, the graph will help the analyst to select an appropriate non-linear relationship.

FIGURE 1.4 THE NEED FOR GRAPHICAL ANALYSIS



$\hat{y} = 3 + 0{\cdot}5x_1 \quad t = 4{\cdot}2 \quad R^2 = 0{\cdot}67$

Source Anscombe (1973)

Finally, graphical methods are capable of displaying large volumes of data and the same graph can often convey overall summary information as well as considerable detail. This is particularly the case when graphs of residuals ($\hat{e}$) are displayed.

Residuals

Exploratory data analysis has been likened by Tukey to numerical detective work. The detective works mainly by indication; by sifting and carefully evaluating the evidence he develops and reformulates his hypothesis of what has occurred at the scene of the crime. Continuing the analogy, the numerical detective requires powerful methods of finding clues and Tukey has likened the analysis of residuals to the use of technical aids in modern criminal investigation:

> 'residuals .... are to the data analyst what powerful magnifying glasses, sensitive chemical tests for bloodstains and delicate listening devices are to a story-book detective' (Tukey, 1977, 125).

Exploratory analysis uses residuals to develop improved models relating the dependent to the explanatory variables. The development of improved hypotheses by residual analysis is a common approach in science. Indeed, in his System of Logic, John Stuart Mill (1874) presented the 'method of residues' as the Fourth Canon of scientific analysis. He writes of residual analysis:

> 'of all the methods of investigating laws of nature, this is the most fertile' (p284).

In subsequent chapters, considerable use will be made of techniques for analysing residuals. However, before residuals combined with graphical techniques can be used to develop improved models, there is a need for a different perspective on statistical assumptions to that offered by confirmatory analysis.

Statistical assumptions

A typical example of the confirmatory approach to the assumptions of statistical procedures is given by Gupta and Hutton (1968). These authors were attempting to discover whether or not economies of scale accrue in local government services and, after a non-technical introduction to multiple regression, they write:

> 'the normality of the residuals has not been tested, but on a priori grounds, we would expect residuals comprising the sum of a large number of independent effects to be normally distributed' (p4).

This appeal to a priori grounds is only possible if the postulated model is known without doubt to have been correctly specified and, if the model has been incorrectly specified, the distribution of the residuals may give clues to this fact. (This will be demonstrated in Chapter 4.) To ignore such valuable clues is to analyse data very inadequately. However, such an approach to assumptions is a frequent hallmark of confirmatory analysis where statistical assumptions are transformed into practical assertions.

The exploratory approach to assumptions differs in two main ways from that of the confirmatory approach. Firstly if key assumptions are not met the postulated model is deemed incorrect and in need of improvement and reformulation. Therefore, an assumption is not

> 'a statement of unquestionable truth, it is a tentative statement on which initial action can usefully be based' (Leamer, 1978, 41).

Secondly, the exploratory analyst realises that some commonly invoked assumptions of statistical analysis are unlikely to be met in non-experimental data analysis. Exploratory analysts have attempted to develop other forms of estimation based on more realistic assumptions and an example of such a method of estimation will be given in Chapter 2.

Cross-validation

Finally, in this review of the fundamentals of exploratory analysis, we come to the most serious potential drawback of the methodology. The lack of a formal hypothesis grounded in strong theory when coupled with the procedure of exploratory data-mining may lead to inferential error. Exploratory procedures could capitalise on a chance result and lead the analyst to incorrect conclusions. Some workers are clearly willing to accept this problem arguing that:

> 'knowledge is gained only if the research worker is willing to take the risk of making mistakes resulting in faulty conclusions' (Mather, 1976, 2).

This argument echoes Gould's (1970, 445) exhortation to make 'good mistakes'. However, it is possible to guard against such possible inferential error by using cross-validation.

In essence, cross-validation amounts to dividing the data set into two groups; exploratory methods are used to develop a model on the basis of one data set and then confirmatory methods can be used to confront the model with the remaining data set. Returning to Tukey's analogy of numerical detective work, the criminal investigator works by indication and, by careful investigation of his data set, he presents his evidence to the judicial methods of confirmatory analysis which decides how much credence to give to each clue.

In more detail, one method of performing cross-validation is to use part of the total data set for developing the postulated model and estimating the unstandardised regression coefficients of the final model. These estimated regression coefficients are then used to predict the values of the dependent variable in the second data set. A simple correlation coefficient can then be calculated between these predictions and the actual values of the dependent variable. The resultant correlation coefficient will always be smaller

than the coefficient of determination of the model estimated on the first data set (Gardiner, 1973, 436), but if this reduction is not a substantial one the model is deemed satisfactory and is accepted. However, if the reduction is a major one it indicates that the exploratory, iterative procedure may have capitalised on a chance result and any conclusions drawn on the basis of the model developed on the first data set should be treated with appropriate caution. Cross-validation therefore gives both an appraisal of the exploratory procedures that have been used on a particular data set and an evaluation of the final model. Although this method of cross-validation was greeted in the geographical literature with enthusiasm by Curry (1967) it appears that no geographical researcher has used the method.

EXPLORATORY ANALYSIS AND SCIENTIFIC METHODOLOGY

Exploratory analysis, induction and grounded theory

Having discussed the fundamentals of exploratory analysis, this section attempts to show the relationships between exploratory methods and the methods of scientific explanation. In particular it attempts to show how exploration may lead to an understanding of the geographical variations in mortality.

At the outset, it may be thought that exploratory analysis, which has been developed for the empirical analysis of data without a fully developed theoretical framework, would be an inductive method. Indeed, according to Tukey (1962, 63)

> 'we need to give up the vain hope that data analysis can be founded upon a logico-deductive system like Euclidean plane geometry .... and to face up to the fact data analysis is intrinsically an empirical science'.

Moreover, when considering exploratory data analysis, one may be reminded of Glaser and Strauss' (1967) concept of the

grounded generation of theory. These sociological methodologists advocated the 'discovery of theory from data systematically obtained from research' (p 2) and, believing that existing methods of statistics were inadequate, they thought that

> 'the freedom and flexibility that we claim for generating theory from quantitative data will lead to new strategies and styles of quantitative analysis with their own rules yet to be discovered' (p 186, emphasis added).

Grounded theory is unashamedly empirically and inductively based and it may superficially appear that exploratory analysis is the required 'new strategy' of data analysis. However, the following discussion attempts to show that exploratory statistics is not an inductive method but a deductive way of analysing data which has marked similarities with the Popperian view of science where explanation is reached by 'conjectures and refutations'. This recognition that exploratory analysis is a deductive not an inductive method is very important; the current status of induction amongst philosophers of science is very low. As Hume pointed out for all inductive inferences, there is no logical justification for extending belief in empirical data to belief in an empirical explanation and as Harvey (1969, 37) has written:

> 'The failure of logicians and philosophers to find (or agree upon) such logical justification has led many to reject its (inductions) use entirely in the presentation of scientific knowledge'.

This basic problem of induction[14] has recently led several geographers to extol the benifits of a Popperian, deductive approach to explanation in geography. (Bird, 1975; Moss, 1977).

Box-Jenkins methodology and Popperian scientific methodology

In order to illuminate the deductive nature of exploratory analysis it is useful to examine the statistical methodology that has been proposed by Box and Jenkins (1970)

for the exploratory analysis associated not with the explanation of geographical pattern but with the forecasting and control of time-series.

The Box-Jenkins methodology (Figure 1.5) begins with a tentative or postulated model, and this model is fitted to a data set in such a way that the residual unexplained variance is minimised. A diagnostic check of the residuals from this model is then undertaken and this provides explicit information as to whether the tentative model is an appropriate one. Moreover, if this tentative model is incorrect, the method provides guidelines for finding a more appropriate one. Undoubtedly, the great strength of this exploratory approach to forecasting lies in the fact that

> 'the Box-Jenkins approach is more honest concerning what is valid to assume about the real world. Certainly their iterative scheme is a convincing summary of Popperian philosophy applied to the problem of forecasting and control' (Bibby and Toutenburg, 1977, 11-12).

Although neither the exploratory statisticians nor Popper make explicit reference to each others work, the great similarities in their approach can easily be brought out. For example, it is stressed in both exploratory statistics and Popperian methodology that one should begin with a model that should be 'tentatively entertained'. As the statisticians Tukey and Wilk (1970) write:

> 'Contemplation of raw observations with an empty mind, even when it is possible, is often hardly more beneficial than not studying them at all' (p 385).

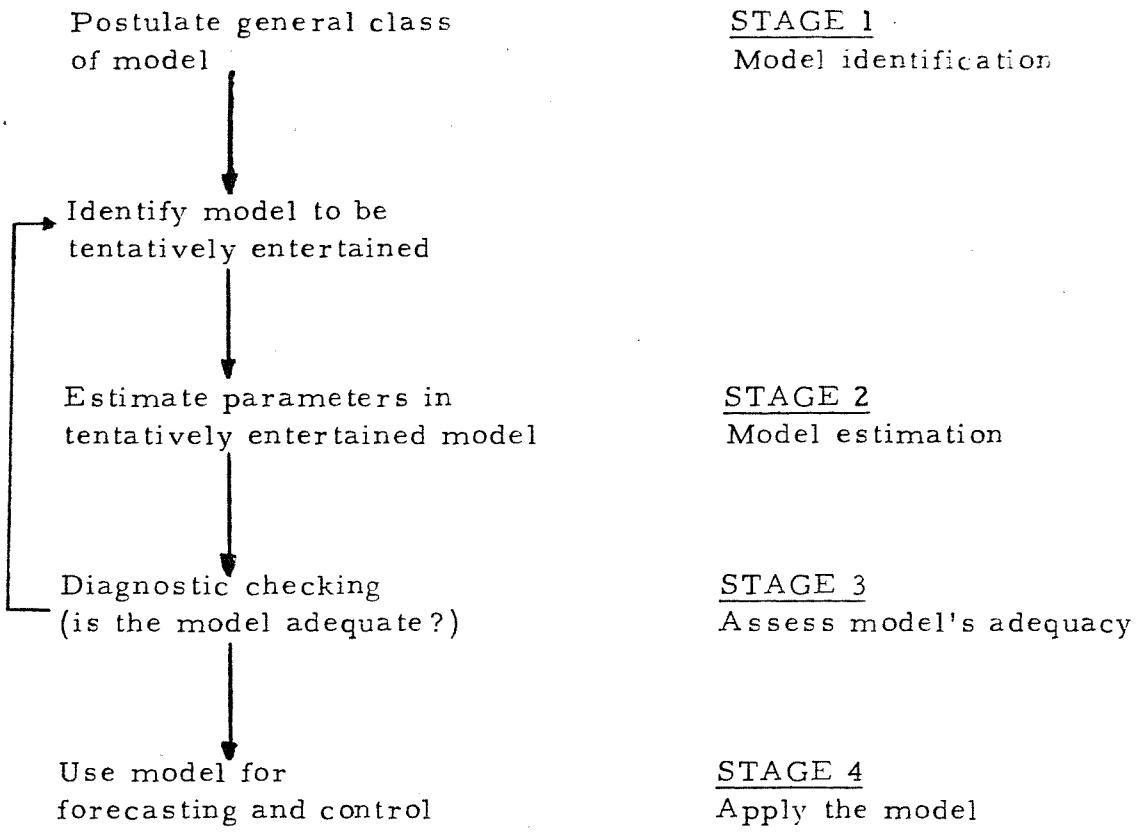
Therefore, we need

> 'some prior presumed situation, some guidance, some objectives, in short, some ideas of the model are essential' (p 372).

The statements echo Popper's (1976b, 89) view that the methods of the social sciences consist of 'trying out tentative solutions to certain problems'. Following the setting

FIGURE 1.5

THE BOX-JENKINS METHODOLOGY



Source: adapted from Makridakis and Wheelwright (1978) p. 246.

up of these tentative models or solutions we have what Popper (1976a, 51) has called the 'critical phase' of 'error elimination' and 'if an attempted solution is refuted through our criticism we make another attempt' (Popper, 1976b, 89). This in turn echoes the views of the exploratory statisticians Tukey and Wilk (1970, 373) for whom 'interaction, feedback, trial and error are all essential'. In Popperian methodology even if a tentative model has undergone several critical phases without being refuted or falsified the model is still only tentatively held as an approximation to the real world. Similarly, exploratory data analysis 'must be considered as an open-ended, highly interactive, iterative process' (Tukey and Wilk, 1970, 372). Popper has summed up his views on scientific methodology in two key concepts - conjectures and refutations - bold, but tentative models followed by critical error elimination. These concepts are also fundamental to exploratory analysis; Daniel and Wood (1971, vii) in their book on data analysis for engineers believe that a

> 'large proportion of failures in such work is due to the analyst's approach. It is not caution that is missing. Boldness in conjecture and persistence in follow up are much more important'.

Exploration and explanation

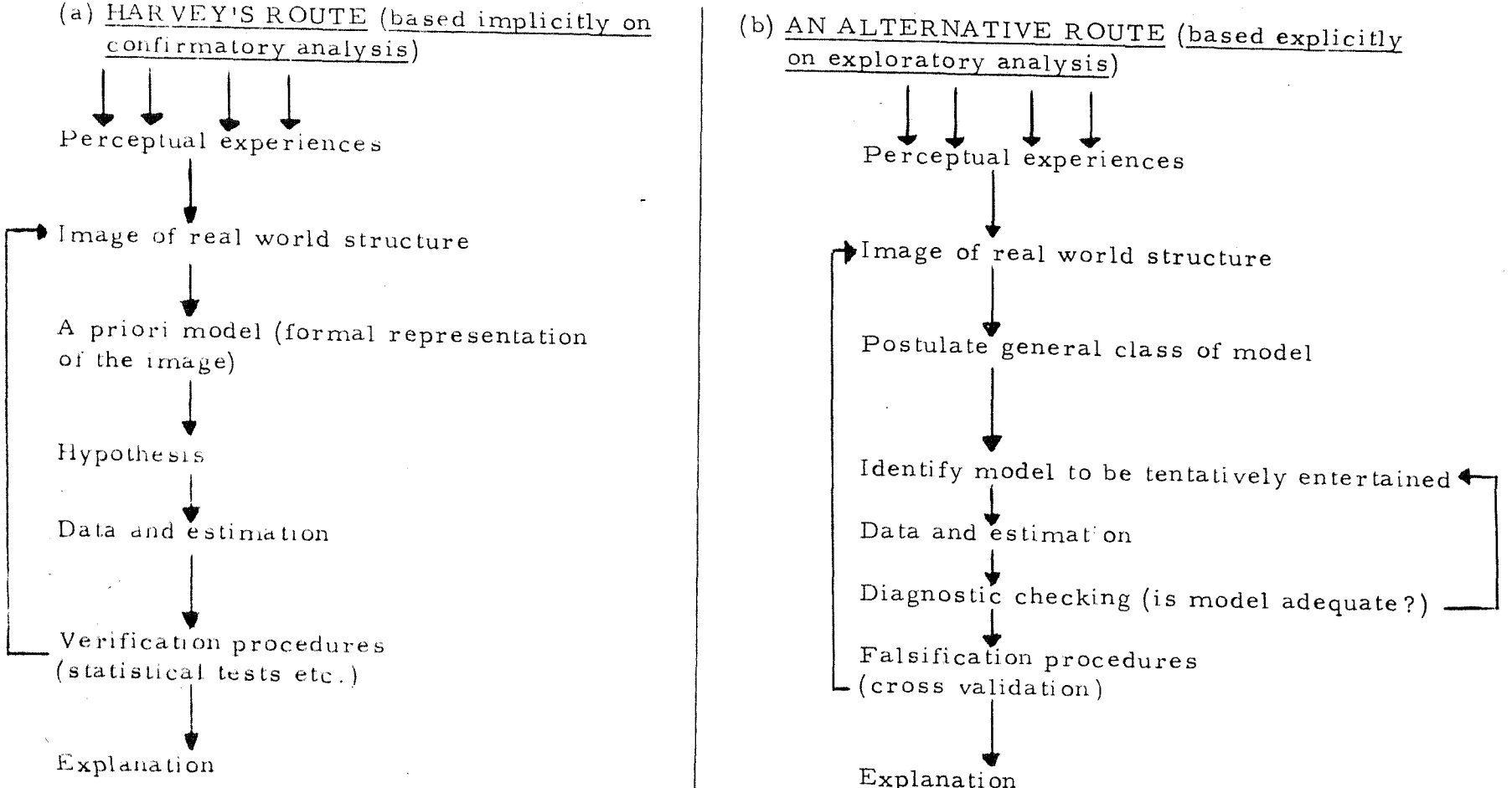
While the exploratory approach in the form of the Box-Jenkins methodology for forecasting and control has been applied to spatial series (Cliff et. al., 1975), very little attention has been given to developing an exploratory approach that attempts to explain geographical patterns. In the remainder of Part I of this work, the aim is to develop, illustrate and evaluate exploratory, explanatory methods specifically designed for the analysis of spatial data. For the moment, however, let us consider some differences between the exploratory and confirmatory approaches to scientific explanation. The confirmatory approach is well

illustrated and documented by Harvey (1969) and is presented in schematic form in Figure 1.6. Also presented in Figure 1.6 is an alternative route to explanation that results from explicitly adopting an exploratory approach. The important difference between the two approaches is that in the confirmatory route, after setting up an initial hypothesis and collecting data to estimate a model, the analyst moves immediately to explanation provided that statistical tests have verified or confirmed the initial hypothesis. However, as shown in earlier discussion, the analyst cannot apply statistical tests until the postulated model is adequate and until the statistical assumptions of the tests have been fulfilled. Therefore, in the exploratory approach, there is the important stage of the diagnostic checking of the model. It is only after the postulated model has been judged satisfactory (or has been reformulated in some way so as to make it satisfactory) that the analyst can proceed to the next stage of the analysis and attempt to falsify the model by cross-validation. (Here again the links with Popperian methodology are clear[15].)

Popper has shown that while generalisations or models are not verifiable they are in fact falsifiable (Popper, 1972; Magee, 1973). It is only when a model has passed certain hurdles that inferences can be made about the individual estimated regression coefficients of the final model. Such inferences can only be made when the exploratory methods of indication and graphical analysis have shown that the model does not break any statistical assumptions and when the final model has not been falsified by cross-validation. Finally it must be emphasised that, as in Popper's schema of scientific methodology, no model is accepted for all time as being correct. Every model must be subjected continuously to falsification procedures but, at the same time, each tentatively held model is thought to be a reasonable explanation of the real world until it has been falsified.

FIGURE 1.6

TWO ROUTES TO SCIENTIFIC EXPLANATION



Source: (a) adapted from Harvey (1969) p. 34 (b) author.

Explanation and mortality patterns

Turning now to the development of an explanation for areal mortality patterns there is little strongly developed theory to light the way. The 'image of the real world structure' (to use Harvey's term) is based on loosely conceived statements such as 'holding other variables constant, air pollution is related positively to bronchitis' together with previous empirical results which are of doubtful value (Chapter 6). In order, therefore, to put into operation the procedures outlined in Figure 1.6b it is surely reasonable to begin with a simple linear model estimated by OLS, including in this model those variables that epidemiologists have previously thought to be important in explaining mortality patterns. If any deficiencies are then found in the model by 'diagnostic checking', an improved model can be formulated. This is the subject matter of Part II of this research.

CHAPTER CONCLUSIONS

It was stated at the outset of this chapter that the epidemiological literature contains numerous studies using regression methods on areal data that have found conflicting results. It is a basic theme of this present work that these conflicting results are a direct outcome of adopting a confirmatory approach to data analysis. While this theme will be elaborated in subsequent chapters, this chapter has been concerned with the alternative approaches that can be adopted for data analysis. On reflection, it is possible to distinguish three distinct approaches - 'sanctification', 'data fishing' and 'data mining'. The first two approaches are basically confirmatory while the third approach is undeniably exploratory.

'Sanctification' is using exact statistical procedures to confirm prior results. This is the confirmatory approach of textbook statistical inference. This method normally

assumes a statistical model that has been developed from strong theory and, indeed, the approach does not allow a reformulated model to be evaluated on a particular data set if that data set has been used to estimate the original model. Moreover, a model cannot be evaluated with this approach unless it has been correctly specified. The only role for this exact approach is to confirm or falsify results that are already known and therefore it assumes its importance when exploratory methods have developed a final, postulated model. Thus, confirmatory statistical tests can be used in cross-validation as a 'form of hygiene' (Ehrenberg, 1975, 323) in an attempt to ensure that a chance occurrence has not seriously affected the analysis.

'Data fishing' is attempting to explore data by confirmatory means. The term is used in a derogatory sense to

> 'indicate both the fisherman's great uncertainty over the quantity and quality of fish that appear in his net and his willingness to accept anything that shows up' (Leamer, 1978, 1).

There is no escaping the fact that such 'fishing' does take place (be it informal trial and error or formal stepwise regression) but classical statistical theory does not allow a given model to be evaluated after a sequence of trials. Furthermore, the use of exact statistical tools in such exploration may result in an inadequate explanatory model.

'Data mining' is the exploration of data using tools specifically designed for the purpose. Mining is

> 'an activity intended to bring to the surface a specific valuable commodity whose existence is likely to be relatively well established before mining commences' (Leamer, 1978, 3).

The present study appears to be the first study in geography that explicitly adopts the exploratory approach to data analysis[16]. In the subsequent chapters of this work an

attempt is made to present, extend and illustrate the methods of exploratory analysis and, moreover, to apply such methods to the analysis of areal mortality patterns. A major feature of this presentation is the critique of the usual methods designed for analysing highly related (multicollinear) data and the development of a new method of analysis for such data (Chapter 2). Chapter 3, on spatial autocorrelation, represents an attempt to develop an exploratory approach to this problem as an alternative to the confirmatory one. Chapter 4 deals with the remaining assumptions of the OLS regression model that have not been considered in the two previous chapters; methods of examining model specification are outlined and particular attention is paid to graphical analysis and cross-validation. The final chapter (Chapter 5) in the methodological part of the work is an examination of the problems associated with the analysis of ratios, and as such it represents the first major geographical discussion of these problems. Finally, the techniques which have been outlined in these five chapters are put into practice in the empirical Part II of this work.

CHAPTER 1: NOTES

1. Other researchers, in other disciplines, similarly believe that inconsistent results are an outcome of poor analysis associated with the general linear model. For example, Deegan (1975) writes of research in political science:

   'Although a number of investigators have attempted to identify empirically a process of political development, substantial controversy still surrounds a determination of the causal factors involved. It is my contention that this state of affairs is the result of inadequacies inherent in traditional techniques of causal modelling' (p 385).

2. This introduction is very cursory; considerably more detail will be found in, for example, Dhutta (1975), Johnston (1972).

3. A good detailed introduction to matrix algebra is Rogers (1971); an adequate background are the summaries presented in such textbooks as King (1969) and Davis (1973).

4. The transpose of a vector is represented by a ''''. For example if

$$A = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad A' = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

   If a vector is pre-multiplied by its transpose, the resultant value is equivalent to the sum of each element of the vector squared.

5. The calculation and use of this equation is illustrated in Chapter 2.

6. There has been a considerable debate in the statistical and econometric literature over the relative merits of standardised and unstandardised regression coefficients. However, following Ahgren and Walburg (1975), two propositions seem reasonable:

(i) in comparing different studies, the unstandardised regression coefficients are probably the most useful;

(ii) in comparing the relative contributions of different terms in a model in a single study, the consideration of standardised regression coefficients is more appropriate.

7. Other uses of OLS regression (for example, the computation of point estimates which can be used for prediction) require a less restrictive set of assumptions.

8. The formal assumption behind this statement is that the explanatory variable matrix is a set of fixed or non-stochastic elements. This assumption is frequently relaxed and the model is then given a 'conditional formulation'. To achieve this conditional formulation, the data must be measured without error.

9. The following is a checklist of assumptions that are associated with the properties of estimators (Stewart, 1976).

Property	Assumptions
unbiased estimator	1, 2, 5, 6
best linear unbiased estimator	1, 2, 3, 4, 5, 6

10. It is a common finding that when an incorrect model is fitted to a data set generated by a known functional relationship, the R^2 of the incorrect model is higher than the R^2 of the true model. This superficially problematical result is because the R^2 value only has meaning for a correctly specified model.

11. Unfortunately, although a large number of workers analyse data in this 'exploratory' manner, very few admit to such procedures in print (Freund and Debertin, 1975). However, Cornish and his co-authors (1977), in a critique of Townsend and Taylor's (1975) study on the regional identity of the north-east of England, write:

'in cross-tabulating six variables across all attitudinal variables relating to a region, Townsend and Taylor must perform over one hundred tests. Even if more than five results are significant at the 0.05 level, there is no way of discriminating which results have chance significance' (p 115).

Townsend and Taylor were obviously trying to explore a particular data set by carrying out a hundred tests in the hope that they would detect 'significant' relationships. With confirmatory techniques one cannot legitimately proceed in this manner.

12. The value given by equation (9) can be regarded as the upper bound on α^*; any dependency amongst the explanatory variables will lead to some smaller value for m.

13. Much of the debt owed to Tukey in this chapter stems not from Tukey's ingenious developments of certain tools of exploratory analysis (the stem and leaf plot, the box and whisker plot, for example), but from his more fundamental reappraisal that there is more to the analysis of data than rigid statistical inference.

14. There are two fundamental problems of induction. Firstly, there is no justification in making general statements (laws, theories, explanations) on the basis of observed, individual data. For example, even after observing that thousands of swans are white the universal, true statement 'all swans are white' cannot be made for the next observation may reveal a black swan. Secondly, induction proceeds by moving from the facts to a theory, but many philosophers have cogently argued that it is impossible to observe facts without a theory.

15. It may be argued that cross-validation is not falsifying a model but confirming it; such an approach also has parallels in the literature of the philosophy of science;

'the more sophisticated falsificationist .... recognises the importance of the role played by confirmation of speculative theories as well as the falsification of the well-established ones' (Chalmers, 1978, 33).

16. Two physical geographers have previously pointed to Tukey's ideas on exploratory analysis - Mather (1976) and Unwin (1977, 186). Tukey's ideas have also been presented at a recent conference on the teaching of quantitative methods (Cox, N.J. and Anderson, E.W. (1978), Heterodox views on teaching geographical data analysis - a paper presented at the Annual Conference of the I.B.G., Hull). At the 1979 Annual Conference, the Quantitative Methods Study Group devoted two sessions exclusively to exploratory data analysis.

CHAPTER 1 : BIBLIOGRAPHY

AHGREN, A. and WALBURG, H.J. (1975): Generalised regression analysis in Amick, D.J. and Walburg, H.J. (eds.) *Introductory multivariate analysis* McCutcham, Berkeley.

ANSCOMBE, F.J. (1973): Graphs in statistical analysis *American Statistician* 27, 17-21.

ASHFORD, J.R., READ, K.L.Q. and RILEY, V.C. (1973): An analysis of variations in perinatal mortality amongst local authorities in England and Wales *International Journal of Epidemiology* 2, 31-46.

BIBBY, J. (1977): The general linear model - a cautionary tale in O'Muircheartaigh, C.A. and Payne, C. (eds.) *The analysis of survey data* volume 2, Wiley, Chichester.

BIBBY, J. and TOUTENBURG, H. (1977): *Prediction and estimation in linear models* Wiley, Chichester.

BIRD, J.H. (1975): Methodoligical implications for geography from the philosophy of K.R. Popper *Scottish Geographical Magazine* 91, 153-163.

BOX, G.E.P. and COX, D.R. (1964): An analysis of transformations *Journal of the Royal Statistical Society* Series B 26, 211-252.

BOX, G.E.P. and JENKINS, G.M. (1970): *Time series analysis, forecasting and control* Holden-Day, San Francisco.

CHALMERS, A.F. (1978): *What is this thing called science?* Open University, Milton Keynes.

CLIFF, A.D., HAGGETT, P., ORD, J.K., BASSETT, K. and DAVIES, R. (1975): Elements of spatial structure: a quantitative approach Cambridge University Press, Cambridge.

COMSTOCK, G.W. (1971): Fatal arteriosclerotic heart disease, water hardness at home and socio-economic characteristics American Journal of Epidemiology 94, 1-10.

CORNISH, M., JACKSON, C., URSELL, G. and WALKER, R. (1977): Regional culture and identity in industrialised societies: a critical comment Regional Studies 11, 113-116.

COX, D.R. and SNELL, E.J. (1968): A general definition of residuals Journal of the Royal Statistical Series B 30, 248-275.

CRAWFORD, M.D., GARDINER, M.J. and MORRIS, J.N. (1971): Changes in water hardness and local death rates Lancet 2, 327-329.

CURRY, L. (1967): Quantitative geography 1967 Canadian Geographer 11, 265-279.

DANIEL, C. and WOOD, F.S. (1971): Fitting equations to data Wiley, New York.

DAVIS, J.C. (1973): Statistics and data analysis in geology Wiley, New York.

DEEGAN, J. (1975): Process of political development - illustrative use of a strategy for regression in presence of multicollinearity Sociological Methodology 3, 384-415.

DHRYMES, P.J. (1970): On the game of maximising $\bar{R}^2$ Australian Economic Papers 9, 177-185.

DHUTTA, M. (1975): Econometric methods South-Western, Cincinnati.

EHRENBERG, A.S.C. (1975): Data reduction Wiley, London.

FEINBERG, S.E. (1979): Graphical methods in statistics
The American Statistician 33, 165-178.

FREUND, R.J. and DERBERTIN, D.L. (1975): Variable selection and statistical significance: a sampling experiment
American Journal of Agricultural Economics 57, 721-722.

GARDINER, M.J. (1973): Using the environment to explain and predict mortality
Journal of the Royal Statistical Society Series A 136, 421-440.

GLASER, G.B. and STRAUSS, A.L. (1967): *Discovery of grounded theory*
Aldine, Chicago.

GNANADESIKAN, R. (1977): *Methods for statistical data analysis of multivariate observations*
Wiley, New York.

GOULD, P.R. (1970): Is *statistix inferens* the geographical name for a wild goose?
Economic Geography 46, 439-448.

GUPTA, S.P. and HUTTON, J.P. (1968): Economies of scale in local government
Royal Commission on Local Government Research Studies Studies 3, HMSO, London.

HARTWIG, F. and DEARING, B.E. (1979): *Exploratory data analysis*
Quantitative Applications in the Social Sciences No. 16
Sage, Beverly Hills.

HARVEY, D.W. (1969): *Explanation in geography*
Arnold, London.

HERZBERG, P.A. (1969): The parameters of cross-validation
Psychometrika 34, 1-70.

HOLMES, D.I. (1972): *Graphical methods for the analysis of data*
unpublished M. Phil. Thesis, Sheffield Polytechnic.

HOLMES, D.I. (1976): Teaching data analysis to first year students at a polytechnic
The Statistician 25, 209-211.

JOHNSTON, J. (1972): Econometric methods
McGraw Hill Kogakusha, Tokyo.

KING, L.J. (1969): Statistical analysis in geography
Prentice Hall, Englewood Cliffs.

LEAMER, E.E. (1978): Specification searches: ad hoc inference with non-experimental data
Wiley, New York.

LEINHARDT, S. and WASSERMAN, S.S. (1978): Exploratory data analysis: an introduction to selected methods in Schuessler, K. (ed.) Sociological methodology 1979
Jossey Bass, San Francisco.

LEINHARDT, S. and WASSERMAN, S.S. (1979): Teaching regression: an exploratory approach
The American Statistician 33, 196-203.

McNEIL, D.R. (1977): Interactive data analysis: a practical primer
Wiley, New York.

MAGEE, B. (1973): Popper
Fontana, London.

MATHER, P.M. (1976): Computational methods of multivariate analysis in physical geography
Wiley, London.

MILL, J.S. (1874): A system of logic
Harper, New York.

MOSS, R.P. (1977): Deductive strategies in geographical generalisation
Progress in Physical Geography 1, 23-39.

MOSTELLER, F. and TUKEY, J.W. (1977): Data analysis and regression
Addison-Wesley, Mass.

POOLE, M.A. and O'FARRELL, P.N. (1971): The assumptions of the linear regression model
Transactions of the Institute of British Geographers 52, 145-148.

POPE, P.T. and WEBSTER, J.T. (1972): The use of an F statistic in stepwise regression procedures Technometrics 14, 327-340.

POPPER, K.R. (1972): The logic of scientific discovery Hutchinson, London.

POPPER, K.R. (1976a): Unended quest: an intellectual autobiography Fontana, London.

POPPER, K.R. (1976b): The logic of the social sciences in Adorno, T. Albert, H., Dahrendorf, R., Habermas, J., Pilot, H. and Popper, K.R. The positive dispute in German sociology Heinemann, London.

ROGERS, A. (1971): Matrix methods in urban and regional analysis Holden Day, San-Francisco.

ROYSTON, E. (1956): A note on the history of the graphical presentation of data Biometrika 43, 241-247.

SELVIN, H.C. and STUART, A. (1966): Data-dredging procedures in survey analysis American Statistician 20, 20-23.

STEWART, J. (1976): Understanding econometrics Hutchinson, London.

STRONG, J.P. CORREA, P. and SOLBERG, L.A. (1968): Water hardness and atherosclerosis Laboratory Investigation 18, 620-622.

TOWNSEND, A.R. and TAYLOR C.C. (1975): Regional culture and identity in industrialised societies: the case of North East England Regional Studies 9, 379-393.

TUFTE, E.R. (1969): Improving data analysis in political science World Politics 21, 641-654.

TUKEY, J.W. (1961): Statistical and quantitative methodology in Ray, D.P. (ed.) Trends in the social sciences Philosophical Library, New York.

TUKEY, J.W. (1962): The future of data analysis
Annals of Mathematical Statistics 3, 1-67.

TUKEY, J.W. (1969): Analysing data: sanctification or detective work
American Psychologist 24, 81-91.

TUKEY, J.W. (1977): Exploratory data analysis
Addison-Wesley, Mass.

TUKEY, J.W. and WILK, M.B. (1970): Data analysis and statistics: techniques and approaches in Tufte, E.R. (ed.)
The quantitative analysis of social problems
Addison-Wesley, Mass.

UNWIN, D.J. (1977): Statistical methods in physical geography
Progress in Physical Geography 1, 185-221.

WALLIS, J.R. (1965): Multivariate statistical methods in hydrology - a comparison using data of known functional relationships
Water Resources Research 1, 447-461.

C H A P T E R 2

MULTICOLLINEARITY :

AN EXPLORATORY APPROACH

'Can we truly claim to have abided by the multicollinearity constraint .... if our 'independent' variables are nothing of the kind? And if not, does it matter? Geographers might turn their attention to this aspect of perhaps their most commonly used technique rather than being almost neurotically concerned with spatial autocorrelation. I am sure I am not alone in wishing for far more guidance on such points in my own work from the second-generation quantifiers' (Hoare, 1977, 515).

'the traditional approach of applying ordinary least squares .... uncritically is risky at best and foolhardy typically' (Wainer, 1978, 271).

'the question 'what is the best estimator?' does not admit to a unique answer .... The question, 'what is the worst overall?' does seem to have a strong candidate: least squares estimators' (Wainer, 1976, 306-307).

THE PROBLEM DEFINED

Tobler (1970, 236) has stated that the first law of geography is that 'everything is related to everything else but near things are more related than distant things'. Regression modelling, on the other hand, requires that explanatory variables are not highly related; indeed, as suggested by the alternative term for such variables, they should be 'independent'. If this is not the case the explanatory variables are termed multicollinear. Multicollinearity is a matter of degree, ranging from 'exact' multicollinearity (where there is a perfect linear correlation among the explanatory variables) to orthogonality (where there is no linear intercorrelation whatsoever). In the former case, the ordinary least-squares procedure breaks down and the analysis cannot proceed; in the latter case, multiple regression need not be performed for each of the slope terms can be estimated by a simple regression of the dependent variable. To a degree, multicollinearity is a condition that exists in most geographical relationships due to the inherent non-experimental nature of the subject. The important question is 'how severe does multicollinearity have to be before it impairs estimation?'.

There is no conclusive answer to the question but intuitively, when explanatory variables are highly related, it becomes extremely difficult to disentangle the separate effect of each explanatory variable from the combined effect of all the variables. Consider the problem of trying to estimate the separate effects of lack of food and alcoholic over-indulgence in determining the disease of cirrhosis of the liver. In a controlled experiment it would be feasible to withold food from a group of people and give large quantities of alcohol to a subset of this group. Control would be exercised so that there would no longer be a strong

inverse relationship between the amount of drink consumed and the lack of food. It is precisely this kind of controlled experiment which cannot be carried out at the geographical scale for both practical and ethical reasons. While it may be conceptually possible to separate the effect of each explanatory variable, statistical methods may be unable to achieve this in practice. Certainly, if ordinary least squares is used to estimate a model with multicollinear data, little confidence can be placed in any interpretation of the resultant estimates of the regression coefficients.

The main consequences of multicollinearity can be summarised as follows.

(1) The OLS estimates remain unbiased but they will have a large variance. With repeated estimation of a correctly specified model the average of the estimated coefficients will equal the true coefficients, but these estimates will be widely scattered around the target of the true values. In any particular application with one data set the estimated coefficients may be wildly incorrect and of course, this imprecision can be critical for substantive analysis. For example, Blalock (1963) noted that high correlations have been found between anomie and authoritarianism. He observed that, in the three instances in which these two variables were used to explain prejedice, all possible results have been found: anomie more important than authoritarianism, authoritarianism more important than anomie, and the two variables equally important. This complete failure to produce consistent results is a direct outcome of the multicollinear relationship between the explanatory variables.

(2) Some investigators, fearful of the estimation problems associated with multicollinearity, drop offending variables from the model. For example, Schwirian and La Greca (1974, 192), in their study of geographical variations in mortality, write

> 'we excluded one of the status indicators from the regression analysis. We do so to eliminate the possible problem of multicollinearity'.

Such expediency may cause problems. If the omitted variables should be included in the model, their

omission will impart bias to the remaining regression coefficients of the reduced model. This problem is further examined in Chapter 4.

(3) Regression estimates with multicollinear explanatory variables are very sensitive to particular sets of data. The addition of a few more observations can sometimes produce dramatic shifts in some of the coefficients.

In view of these possible problems and pitfalls, it is clear that researchers must be given guidance on how to recognise when multicollinearity is present and how to decide what action, if any, must be taken to overcome the problem. The discussion will begin with an example of the problems that can arise from the analysis of multicollinear data.

THE PROBLEM ILLUSTRATED

Let us suppose that an analyst believes that the number of deaths in an area is solely the result of the number of people living in that area. In order to display his statistical expertise he decides (blithely unaware of the problems of multicollinearity) not to perform a simple linear regression with the total number of people as the explanatory variable but to perform a multiple regression with three explanatory variables representing the number of people in three age groups. Let us do this for the Glamorgan local authorities for the years 1963 to 1967; the dependent variable (y) is the number of deaths from coronary heart disease for males aged 35 to 64 years and the explanatory variables are the number of males in the three age categories : 35-44 (x_1), 45-54 (x_2) and 55-64 years (x_3). Such a contrived example has been chosen so that we can state *a priori* what the results should be. All the standardised regression coefficients associated with each age category should be positive (the more males there are in the area the more should die from heart disease) and the older the age category the greater the influence this category should have

in determining the number of deaths (the 55-64 years category should have the largest standardised regression coefficient and the 35-44 years category the smallest). We can also expect a priori that the model will be severely multicollinear; an area with a large number of males in one age group can be expected to have a large number of males in the other two groups.

In order to estimate the standardised regression coefficients for this model with ordinary least squares, the following formula (first presented in Chapter 1) must be solved

$$\hat{\beta}_s = [R]^{-1} \cdot G \qquad (1)$$

where $\hat{\beta}_s$ represents the three standardised regression coefficient estimates,

$[R]^{-1}$ is the inverse of the correlation matrix for the three explanatory variables (x_1, x_2, x_3),

G represents the three correlations between the dependent variable (y) and each explanatory variable.

Let us begin with the column vector G. In the general case for m explanatory variables G can be defined as:

$$G = \begin{bmatrix} rx_1y \\ rx_2y \\ rx_my \end{bmatrix} \qquad (2)$$

where rx_my is the simple correlation (r) between the variables x_m and y. For the Glamorgan data G is calculated to be:

$$G = \begin{matrix} y \\ \begin{bmatrix} 0.9890 \\ 0.9940 \\ 0.9945 \end{bmatrix} \end{matrix} \begin{matrix} \\ x_1 \\ x_2 \\ x_3 \end{matrix} \qquad (3)$$

Thus, there is a very strong linear relationship between heart disease and each age group, and the strongest relationship is found between heart disease and the oldest age group.

In order to calculate the invese of the explanatory variable correlation matrix, $[R]^{-1}$, we must first calculate

the correlation matrix, $[R]$. In the general case for m explanatory variables, $[R]$ can be defined as

$$[R] = \begin{bmatrix} rx_1x_1 & rx_1x_2 & \cdots & rx_1x_m \\ rx_2x_2 & rx_2x_2 & \cdots & rx_2x_m \\ \vdots & & & \\ rx_mx_1 & rx_mx_2 & \cdots & rx_mx_m \end{bmatrix} \quad (4)$$

where rx_1r_m is the correlation between x_1 and x_m. For the Glamorgan data the following matrix is obtained

$$[R] = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1.0000 & 0.9969 & 0.9933 \\ 0.9969 & 1.0000 & 0.9974 \\ 0.9933 & 0.9974 & 1.0000 \end{bmatrix} \end{array} \quad (5)$$

As anticipated there is a high positive relationship between all three age groups. For the three-variable case the inverse of the correlation matrix, $[R]^{-1}$ is given by:

$$[R]^{-1} = \begin{bmatrix} \frac{rx_1x_1-(rx_2x_3)^2}{|R|} & \frac{rx_1x_3 \cdot rx_2x_3-rx_1x_2}{|R|} & \frac{rx_1x_2 \cdot rx_2x_3-rx_1x_3}{|R|} \\ \frac{rx_1x_3 \cdot rx_2x_3-rx_1x_2}{|R|} & \frac{rx_2x_2-(rx_1x_3)^2}{|R|} & \frac{rx_1x_2 \cdot rx_1x_3-rx_2x_3}{|R|} \\ \frac{rx_1x_2 \cdot rx_2x_3-rx_1x_3}{|R|} & \frac{rx_1x_2 \cdot rx_1x_3-rx_2x_3}{|R|} & \frac{rx_3x_3-(rx_1x_2)^2}{|R|} \end{bmatrix} \quad (6)$$

where rx_1x_2 is the correlation between variable x_1 and x_2 and $|R|$ is given by:

$$|R| = 1 + 2\,(rx_1x_2 \cdot rx_1x_3 \cdot rx_2x_3) - (rx_1x_2)^2 - (rx_1x_3)^2 - (rx_2x_3)^2 \quad (7)$$

This value $|R|$ is called the <u>determinant</u> of a matrix. For the Glamorgan mortality data the determinant is

$$|R| = 1 + 2\,(0{\cdot}9969.0{\cdot}9933.0{\cdot}9974) - (0{\cdot}9969)^2 - (0{\cdot}9933)^2 - (0{\cdot}9974)^2$$

$$|R| = 0{\cdot}0000312 \qquad (8)$$

The inverse correlation matrix for the Glamorgan data is:

$$[R]^{-1} = \begin{bmatrix} 166.5 & -198.2 & 32.3 \\ -198.2 & 428.1 & -230.1 \\ 32.3 & -230.1 & 198.4 \end{bmatrix} \qquad (9)$$

To obtain the standardised regression coefficients $\hat{\beta}_s$, the inverse correlation matrix, $[R]^{-1}$, has to be post-multiplied by the vector of correlations between the explanatory and dependent variables, G. Thus

$$\hat{\beta}_s = [R]^{-1}.G \qquad (1)\text{ bis}$$

For the Glamorgan data this gives

$$\hat{\beta}_s = \begin{bmatrix} 166.5 & -198.2 & 32.3 \\ -198.2 & 428.1 & -230.1 \\ 32.3 & -230.1 & 198.4 \end{bmatrix} \cdot \begin{bmatrix} 0.9890 \\ 0.9940 \\ 0.9945 \end{bmatrix}$$

$$= \begin{bmatrix} -0.21 \\ 0.65 \\ 0.55 \end{bmatrix} \qquad (10)$$

The standardised partial regression coefficient between heart disease and the 35-44 age group controlling for the other two age groups is -0.21. Similarly, the coefficient between heart disease and 45-54 age group is 0.65 and, between heart disease and the 55-64 age group it is 0.55. Such results go against *a priori* reasoning, for the spatial variation of the 45-54 age group accounts for more of the mortality variation than does the 55-64 age group. Moreover, the sign for the regression coefficient of the 35-45 age group is negative. The more people there are in this age group the fewer die from heart disease! Such implausible results are typical of severely multicollinear data. While, in this example, commonsense suggests that OLS regression is giving nonsensical results, in other research (where theoretical

underpinnings may be weaker), gross misinterpretations can easily occur.

In order to gain a deeper understanding of the problem it is instructive to examine the two extreme cases of orthogonality and exact multicollinearity. An orthogonal three-variable correlation matrix is given by:

$$[R] = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \end{array} \qquad (11)$$

The correlation between each variable and itself is perfect ($r = 1.0$), but there is no relationship whatsoever between any pair of different variables ($r = 0.0$). The determinant of such a matrix, from equation (7) is,

$$\begin{aligned} |R| &= 1+2\ (0\cdot0.0\cdot0.0\cdot0)-0\cdot0-0\cdot0-0\cdot0 \\ &= 1 \end{aligned} \qquad (12)$$

Now consider a matrix in which one explanatory has a perfect relationship with another explanatory variable (a perfect positive relationship between variables x_1 and x_2 in this case):

$$[R] = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{bmatrix} 1.0 & 1.0 & 0.0 \\ 1.0 & 1.0 & 0.0 \\ 0.0 & 0.0 & 1.0 \end{bmatrix} \end{array} \qquad (13)$$

The determinant of this matrix is

$$\begin{aligned} |R| &= 1+2\ (1\cdot0.0\cdot0.0\cdot0)-1\cdot0-0\cdot0-0\cdot0 \\ &= 0\cdot0 \end{aligned} \qquad (14)$$

The procedure for inverting a matrix illustrated above (equation (6)) cannot be performed on this matrix, for the division by a determinant equal to zero results in the value infinity. With such exact multicollinearity the correlation matrix cannot be inverted and the analysis cannot proceed.

In a similar manner, multicollinearity affects the variances of the regression coefficients. These estimated variances are a measure of how narrow or wide is the spread of the estimated coefficients around the true but unknown regression parameters. Of course we require estimates that have a narrow spread and, when multicollinearity is not a problem, OLS estimators are the estimators with the narrowest spread of all possible estimators (Chapter 1). But when multicollinearity is present the OLS estimators have a high variance. In the case of three explanatory variables (x_1, x_2, x_3) an estimate of variance associated with the regression coefficient between the dependent variable (y) and x_1 holding x_2 and x_3 constant is given by

$$\text{var}\ (\hat{b}) = \frac{s_e^2\ (rx_1x_2 - (rx_2x_3)^2)}{|R| \ .\ (N-1)\ .\ (s_y^2)} \qquad (15)$$

s_y^2 is an estimate of the variance of the dependent variable,
s_e^2 is an estimate of the variance of the disturbance term,
N is the number of observations.

An examination of equation (15) reveals that calculation of a variance estimate requires division by the determinant. When there is no correlation whatsoever among the explanatory variables the determinant will equal 1 and the variance will be at a minimum. However, as multicollinearity increases, so the determinant approaches zero and the variance of the estimated coefficients is inflated. Inevitably, therefore, the precision of the estimates will decline. In the extreme case of exact multicollinearity the determinant will be zero and the variance will be infinite.

The problem detected?

An important source of information for detecting multicollinearity is a priori expectation; a regression coefficient with an apparently incorrect sign (as in the Glamorgan data) may alert the analyst to the severity of the problem. For

situations where theory is poorly developed, a number of researchers have suggested that the simple correlation coefficients between the explanatory variables can be used as an indicator of multicollinearity. For example, Heise (1975) has implied that if any simple correlation is greater than 0.8 then the degree of multicollinearity is unacceptable. But when there are many explanatory variables, problems may arise with this approach. Two correlations of 0.6 may be as harmful as one of 0.8. Moreover, the simple correlation coefficient can, in some situations, be a very poor indicator of multicollinearity. Suppose there are three variables and ten observations on each variable:

x_1 : (1 1 1 1 1 0 0 0 0 0)

x_2 : (0 0 0 0 0 1 1 1 1 1)

x_3 : (1 1 1 1 1 1 1 1 1 1)

As x_1 plus x_2 equals x_3 there is exact multicollinearity and yet the simple correlation coefficients ($rx_1x_2 = -0.33$; $rx_1x_3 = rx_2x_3 = .59$) are not especially high. The determinant of such explanatory variables would be zero and it seems reasonable that the determinant, with its fixed range of zero to one, could be used as a measure of multicollinearity. Moreover, as can be seen from equation (7) the determinant is influenced by all the simple correlations amongst a set of explanatory variables; thus it is a single summary measure of overall multicollinearity.

Farrar and Glauber (1967) have used the determinant (in a chi-square transformation) as a confirmatory means of detecting multicollinearity. This test has been subsequently modified by Haitovsky (1969), and it is in this form that it is presented here:

$$\chi^2_h = H \log_e (1 - |R|) \qquad (16)$$

where χ^2_h is the Haitovsky chi-square statistic,
m is the number of explanatory variables,
H equals $(1+(2m+5)/6-N$
where N is the number of observations;
the number of degrees of freedom is given by $m(m-1)/2$.

The null hypothesis is that the explanatory variables are orthogonal and, if the observed χ_h^2 is lower than the theoretical value of χ_h^2 at a predetermined significance level, the null hypothesis of orthogonality can be rejected at that level. (Rockwell (1975) provides a number of applications of this test.)

Farrar and Glauber (1967) also propose a test for identifying the multicollinear variables. The test makes use of the diagonal elements of the inverse explanatory variable matrix (the so-called r_{ii} of $[R]^{-1}$).[1] These values are transformed to have an F distribution

$$FG = (r_{ii} - 1) \cdot \frac{(N-m)}{m-1} \qquad (17)$$

The null hypothesis is that a particular variable is not multicollinear with any other combination of variables. If the test statistic, FG, is greater than the tabulated value (with m-1 and N-m degrees of freedom) at some pre-determined significance level, the null hypothesis can be rejected at this level.

In order to illustrate the use of the Farrar-Glauber procedure, let us consider three data sets. The first is the Glamorgan mortality data that we suspect to be severely multicollinear. The second is taken from Preston's (1970) study of mortality patterns at an international scale, in which he attempted to elucidate the relationship between excessive male mortality and a number of explanatory variables that are listed in Table 2.1. For him, these explanatory variables were not severely multicollinear:

> 'It is widely believed that mortality factors in different countries are so closely inter-related that any attempting to isolate individual factors in a cross-sectional analysis is futile. Fortunately, however, this is not the case' (Preston, 1970 45).

Table 2.1

Explanatory variables used by Preston

x_1	daily calorie consumption per capita 1952
x_2	index of per capita income 1960
x_3	percentage of male labourforce in primary industries 1960
x_4	absolute change in x_3 since 1910
x_5	percentage of population living in localities larger than 100,000 1960
x_6	percentage of population living in localities larger than 20,000 1960
x_7	automobiles per inhabitant
x_8	average cigarette consumption per adult 1935
x_9	average cigarette consumption per adult 1945-62
x_{10}	difference between daily calories of saturated and unsaturated fat 1952

Source: Preston (1970)

Table 2.2

Testing for multicollinearity

	observed statistic	expected .05	expected .01	decision on H_o .05	decision on H_o .01
Glamorgan data					
	χ^2_h 0.2	7.81	11.34	reject	reject
FG	x_1 (35-44) 1889.0	3.42	5.66	reject	reject
	x_2 (45-54) 4946.0	"	"	reject	reject
	x_3 (55-64) 2305.0	"	"	reject	reject
Preston's data					
	χ^2_h .0009	73.0	94.0	reject	reject
FG	x_1 4.3	3.68	6.72	reject	accept
	x_2 8.1	"	"	reject	reject
	x_3 2.0	"	"	accept	accept
	x_4 2.2	"	"	accept	accept
	x_5 4.0	"	"	reject	accept
	x_6 6.8	"	"	reject	reject
	x_7 13.8	"	"	reject	reject
	x_8 4.1	"	"	reject	accept
	x_9 13.6	"	"	reject	reject
	x_{10} 5.6	"	"	reject	accept
West and Lowe's data					
	χ^2_h 184.0	7.81	11.34	accept	accept
FG	x_1 (temperature) 9.9	3.03	4.7	reject	reject
	x_2 (rainfall) 8.0	"	"	reject	reject
	x_3 (social) 6.1	"	"	reject	reject

The third set is taken from West and Lowe's (1976) examination of mortality variations for 115 county and London boroughs. The dependent variable is deaths from ischaemic heart disease and there are three explanatory variables: temperature, rainfall and a socio-economic index. West and Lowe do not discuss the degree of multicollinearity in their variables.

The results of applying the tests to the three data sets are given in Table 2.2. For the Glamorgan data the tests, as anticipated, reject the null hypothesis of no multicollinearity. For Preston's data, contrary to his opinion, the Haitovsky test rejects the null hypothesis that the explanatory variable matrix is orthogonal. Moreover, the Farrar-Glauber test indicates that a number of variables are involved in this multicollinearity. For the West and Lowe data, the tests show ambivalent results; the Haitovsky test does not detect multicollinearity, but the Farrar-Glauber tests indicate all three variables are involved in multicollinearity. Clearly, consideration must be given to the factors causing such results.

The Farrar-Glauber and Haitovsky approach to multicollinearity is undoubtedly a confirmatory one; on the basis of a sample they attempt to make inferences about a population. However, as pointed out by O'Hagan and McCabe (1975) and Maddala (1977), multicollinearity is a problem of the sample and therefore one cannot, and need not, test for multicollinearity in the population. Moreover, the significance-testing approach does not really measure the <u>severity</u> of multicollinearity; the summary measures have not been linked to the <u>effects</u> of multicollinearity and, as Willan and Watts (1978, 407) point out,

> 'the Farrar and Glauber approach often indicates rejection of the null hypothesis of orthogonality when, from a practical point of view, there is not a severe collinearity problem'.

An exploratory approach to detecting multicollinearity involves, as a first step, the use of the Haitovsky and

Farrar-Glauber procedures to indicate the problem. The second step is the application of measures capable of overcoming multicollinearity to see if this makes a substantial difference to the estimated regression coefficients. If a large difference is found between the estimated coefficients, it is highly likely that the data are multicollinear.

THE PROBLEM SOLVED?

One of the most commonly suggested solutions to the problem of multicollinearity is to acquire more data. However, this will only be successful if the multicollinearity is due to measurement errors or if the intercorrelation exists, not in the population explanatory variables, but in the particular data set being analysed. If the population explanatory variables are genuinely multicollinear, an increase in the size of the sample cannot solve the problem. Moreover, attempting to collect data specifically to reduce the problem may result in data points that are 'rare' in the population. These 'outliers' could strongly influence the estimated regression coefficients and produce meaningless results (Chapter 4).

Another common procedure intended to reduce multicollinearity is the deliberate exclusion of one or more offending variables from the regression equation. However, this solution is infrequently admitted in print (a rare example is McCally, 1966) because the removal of a multicollinear variable from a regression model, when the true model includes that variable, results in the remaining regression coefficients being biased (Chapter 4). Brown (1973, 35), for example, found that the outcome of omitting variables from a well-specified model was to 'overestimate the coefficients of interest by a factor of four'. Moreover, explanatory variables should be theoretically specified, and they cannot be deleted from theory as easily as they can be omitted from a regression equation.

Three statistical procedures have been suggested to deal with multicollinearity. They are:

(1) stepwise regression,

(2) principal components analysis, and

(3) biased estimation.

Each of these approaches will now be considered in turn.

Stepwise regression

Girt (1972, 1974) in his study of female chronic bronchitis in Leeds, believing the explanatory variables of his postulated model to be multicollinear, used stepwise regression to overcome the problem. His results were somewhat surprising.

> 'Not only was no relationship found between pollution in present residential areas and bronchitis, but the effects of the past environmental variables, in the final equation, are not as one would expect if the factors were related' (Girt, 1972, 105).

In performing stepwise regression with multicollinear data, Girt was following a number of other geographers in adopting a confirmatory approach to the problem. For example, Olsson (1965) believed that the technique had been specifically developed for the resolution of the multicollinearity problem and Morrill and Wohlenburg (1971, 11) advocated the use of the technique

> 'when there is a large number of possible explanatory variables which may be intercorrelated ..... the procedure ensures that the significant variables entering the equation are truly independent'.

Unfortunately, as we have already seen (Chapter 1) the stepwise approach to explanatory model-building is a poor one, and, as multicollinearity becomes more severe, the efficacy of the technique declines. When analysing multicollinear explanatory variables, stepwise regression cannot effectively disentangle the separate influences of different explanatory variables (Gorman and Torman, 1966). Indeed, setting different significance levels for the selection or deletion

of variables can result in a completely different final model. For example, when analysing a model with six possible explanatory variables and setting the F test at 0.05, variables one, five and six may constitute the final model; but setting the F test at 0.01 may result in the final model comprising variables one and three. As Wallis (1965, 450) in his study of known functional relationships has written:

> 'stepwise multiple regression tends to pick variables that confound several independent effects and to build models that are hard to interpret in terms of the real world'.

This, of course, is the experience of Girt quoted above, and the statement is reinforced when it is realized that a stepwise procedure can select 'significant' variables from a series of random data and produce a final model in which random data explains 99.9969 per cent of the variation of other random data (Ando and Kaufman, 1966; Mayer and Stowe, 1969).

Contrary to the hopes of a number of investigators, therefore, stepwise regression does nothing to ease the multicollinearity problem. Thus we must look elsewhere for suitable techniques.

Principal components analysis[2]

Hauser (1974), after a detailed critique of stepwise regression, suggests that geographers should use principal components to overcome the multicollinearity problem. Principal components analysis (PCA) is a technique for transforming data to produce a new set of variables (components) that are orthogonal to each other. Furthermore, this transformation is performed in such a manner that the first component absorbs the maximum possible proportion of the original data, the second component absorbs the maximum of the remaining variation in the data, and so on until the final component, which accounts for the smallest proportion of the variance in the original data. Usually PCA results in a few

large components accounting for most of the total variance of the data, and a large number of small components accounting for a small part of the total variance.

Such principal components can be used to overcome the multicollinearity problem in two ways. Firstly, an analyst can perform a PCA on the explanatory variables, discard the small components, regress the dependent variable on the retained components and re-transform the data to obtain estimates of the slope terms for the original explanatory variables. Such an approach to multicollinearity represents a form of biased estimation. While this form of estimation will be given greater consideration later, it is appropriate to point out here that recent simulation studies of the use of principal components in regression analysis have not been favourable. In particular, Mittelhammer and Baritelle (1977) have found that, even in those cases where this type of estimation performed better than OLS regression, the individual regression coefficient estimates still had large variances. Moreover, Wermuth (1975) in a large-scale simulation study, found that principal components estimators were generally inferior to other types of biased estimator.

The second approach to multicollinearity is to use the principal components as variables themselves within a regression analysis. As pointed out by Mather (1976), this is the method that is most commonly employed by geographers to tackle multicollinearity. Keeble and Hauser (1972), for example, analysed industrial growth in south-east England in this manner, while Bidot (1969) used the method to examine the determinants of local government expenditure patterns. In another variant on the general method, Kirby and Taylor (1976) have used factor analysis to produce two uncorrelated factors in an attempt to explain the variation of the 1975 EEC referendum vote in twenty-three regions of the United Kingdom. In epidemiological literature, too, PCA has been used to analyse multicollinear data. Buckatzsch (1947) used

principal components as explanatory variables in his attempt to analyse the geographical variation of infant mortality. Similarly, Gardner (1973), in a study of mortality variation in England and Wales, computed a 'socio-economic' index, based on a nine-variable PCA, which he later used in regression analysis.

Despite this widespread use of PCA, the technique has inherent drawbacks. In any analysis which attempts to break the multicollinearity deadlock, not all the principal components can be used. Therefore, the analyst has to choose the components that are to be examined in the regression analysis. Gardner (1973) decided that he would use the first principal component because it was the one which accounted for the major proportion of the variances of the explanatory variables. But this first principal component need not necessarily be the one that is the most highly related to the dependent variable. In fact there is no necessary relationship between the order of the principal components and the degree of their correlation with the dependent variable. While there are a number of statistical procedures for selecting which components should be removed from the analysis, it is perhaps the interpretability of the components that is most important. As Daultrey (1976, 45) states:

> 'to interpret the regression coefficients the principal components must be interpretable'.

If the transformed, and therefore artificial, variables can be given a specific meaning, they can be used as variables in their own right, and the method is then a defensible solution to the problem of multicollinearity. In practice, however, the unambiguous interpretation of principal components is rarely, if ever, possible (Williams, 1971)[3]. As Robson (1973, 208) has written of the geographical use of the related technique of factor analysis;

> 'one is left with the vaguest of verbal comparisons and a welter of factors designated by names which are often at once mysterious, pretentious, infelicitous, and occasionally downright dishonest'.

In many studies, the principal components become an artificial barrier between the analyst and the data being analysed and, in conclusion, one can concur with Maddala's (1977, 194) comments on the use of principal components in regression analysis:

> 'as a solution to the multicollinearity problem, thought it is often suggested, its use is very limited'.

RIDGE REGRESSION[4]

In the final part of this chapter we will discuss a technique that appears to be a promising method for the analysis of multicollinear data: ridge regression. Although this technique is of relatively recent origin (being fully outlined in the theoretical and expositional papers of Hoerl and Kennard (1970a, 1970b)[5] there is a considerable research literature on both applied and theoretical aspects of the method. The most comprehensive, if exceedingly terse, review of the theoretical aspects of the technique is Vinod's (1978b) survey. Although the technique was developed and originally illustrated with physical science data, applications have appeared in several social science subject areas, ranging from agricultural economics (Brown and Beattie, 1975), political science (Deegan 1972, 1975), and psychology (Price, 1977) to a remarkably successful prediction of the UK election results (Browne and Payne, 1975).[6] It appears that ridge regression has been discussed in four geographical papers, namely Mather (1976, 1977), Moriarty (1973) and Mather and Openshaw (1974).[7] The technique will receive a fairly extensive treatment here for, if the common problem of the misuse of statistical methods is to be avoided, the assumptions and pitfalls, as well as the possibilities, of ridge regression need to be fully considered. The technique will first be presented and illustrated and this will be followed by an evaluation of the method. This assessment of ridge regression will not only review previous research but

it will also contain some new results. In particular, a simulation study will demonstrate that ridge regression can be a considerable improvement over ordinary least squares in certain circumstances.

The technique outlined

In order to explain the technique let us first consider the method from an informal graphical viewpoint. Subsequently, we will outline the mathematical elements of the technique and illustrate the method with mortality data.

If it was possible to fit a particular multiple regression model to a large number of samples, it would obviously be desirable that as many as possible of the estimated coefficients be very close to the true coefficient: that is, the estimated coefficients should be on target with a minimum of spread around this target. As discussed in Chapter 1, OLS estimates, provided certain assumptions are fulfilled, are the estimates which are unbiased with the minimum of spread around the true value. Figure 2.1a depicts the theoretical sampling distribution of a regression coefficient for orthogonal data. The estimates are tightly grouped around the true regression value, and if the average value (or mean) of this estimator is calculated it will equal the true value (the estimator is therefore unbiased). However, Figure 2.1b depicts the theoretical sampling distribution of an estimator for a multicollinear data set. In this instance, the average or mean of the estimator is still equal to the true value but the spread or variance of the estimator is much greater than for orthogonal data. In any particular application to one multicollinear data set, the regression coefficients are likely to be very different from the true values. Therefore, when one data set is being analysed, the lack of bias is not of great value but it would be extremely useful to have an estimator with a narrow sampling distribution. Figure 2.1c shows the theoretical sampling distribution of a ridge estimator with multicollinear

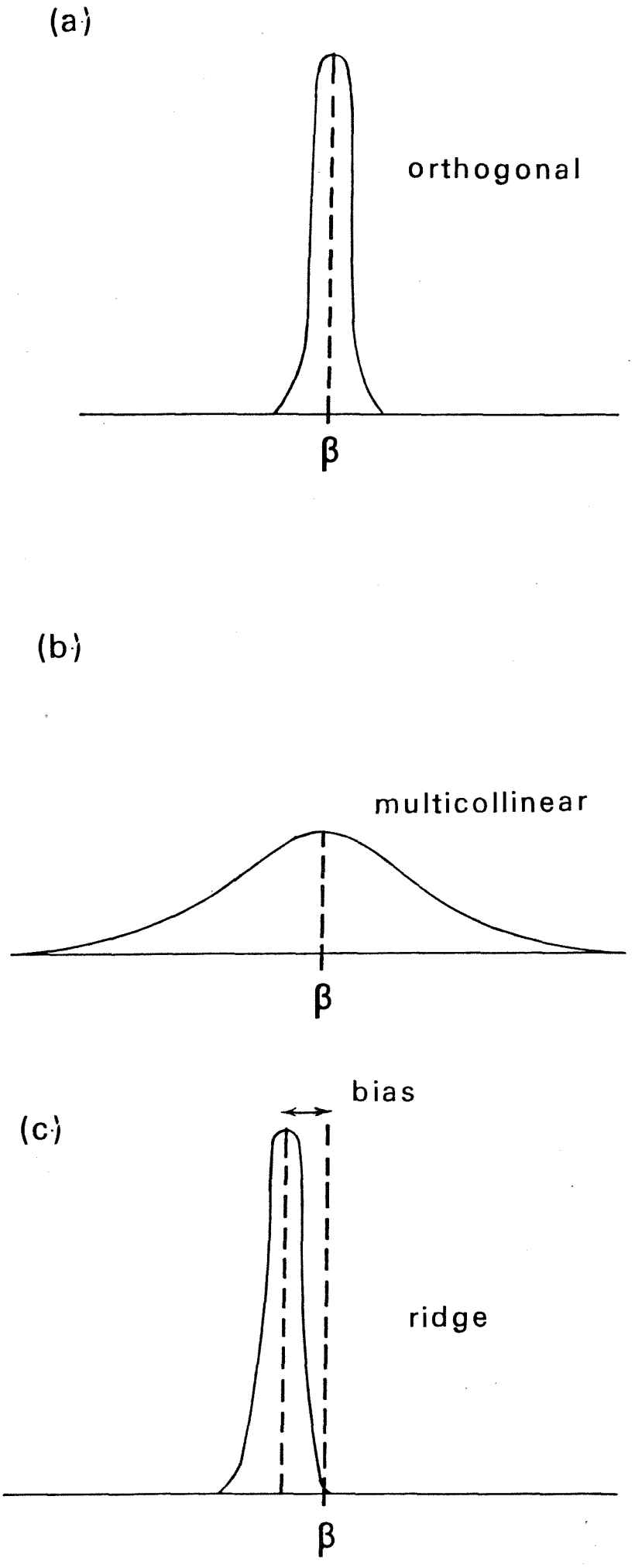


FIGURE 2.1 THEORETICAL SAMPLING DISTRIBUTIONS OF ESTIMATORS

data. The spread or variance of the estimator is much narrower than for the OLS estimator but the average or mean of the estimator is no longer equal to the true value. The ridge estimate therefore is biased, and the property of no-bias has been exchanged for a decrease in the spread of the estimator. The aim of ridge regression is therefore to trade a little bias (no one wishes the mean of the estimates to be too far from the true value) in exhange for a substantial decrease in the variance of the estimator so that in any particular application the ridge estimate is more likely to be near the true value than the OLS estimate.

More formally, the OLS estimator with orthogonal and multicollinear data is unbiased

$$E\ (\hat{B}) = \beta \qquad (18)$$

where E is the expected or average value, $\hat{B}$ is the vector of estimated regression coefficients and β is the vector of true values. Hoerl and Kennard (1970a) examined another property of the relationship between the estimated and true values, namely the 'squared distance' between the two values. They expressed this as:

$$(L_1)^2 = (\hat{B}-\beta)' \cdot (\hat{B}-\beta) \qquad (19)$$

where L_1 is the 'distance' or difference between the true and estimated coefficients. If the estimated coefficients are close in value to the true ones, $(L_1)^2$ will obviously be small. The average or expected squared distance is given by:

$$E\ (L_1)^2 = \sigma_\varepsilon^2 \sum_{i=1}^{m} (1.0/\lambda_i) \qquad (20)$$

where $E\ (L_1)^2$ is the expected squared distance, σ_ε^2 is the variance of the disturbance term and λ_i is the i'th eigenvalue associated with the explanatory variable matrix R. If there are m orthogonal explanatory variables in the R matrix there will be m eigenvalues all equal to 1.0. But as the explanatory variables become more and more multicollinear, the minimum

eigenvalue (λ_m) approaches zero. As λ_m decreases the value of $E(L_1)^2$ must increase, and the estimated vector of coefficients ($\hat{B}$) can be expected to be further from the true values (β). This squared distance value is sometimes called the Mean Square Error and the MSE of an estimator is equal to the variance of the estimator plus the bias squared. An alternative mathematical formulation for this value is

$$E(L_1)^2 = MSE = \sigma_\varepsilon^2 \text{ Trace } [R]^{-1} \qquad (21)$$

where the operator Trace refers to the sum of the eigenvalues. To recapitulate, as the explanatory variables become multicollinear, the smallest eigenvalue will approach zero, the estimated regression coefficients will be far from the true ones and the MSE (variance + bias^2) will increase. The bias of the OLS estimator, however, does not increase with multicollinearity; therefore the increase in the MSE must be accounted for by the increase in variance.

In a similar fashion, Hoerl and Kennard (1970a) considered the 'length' of the estimated coefficient vector. This is given by

$$E(\hat{B}'\hat{B}) = \beta'\beta + \sigma_\varepsilon^2 \text{ Trace } [R]^{-1} \qquad (22)$$

or

$$E(\hat{B}'\hat{B}) = \beta'\beta + \sigma_\varepsilon^2 \sum_{i=1}^{m} (1.0/\lambda_i) \qquad (23)$$

With multicollinear variables there is an increased likelihood that the estimated coefficient vector will be too 'long'. Under such a condition the least-squares solution can give individual coefficients that are too large in absolute value (Rutihauser, 1968); indeed, it is possible with multicollinear data to estimate uninterpretable standardised regression values in excess of +1 and -1 (Deegan, 1978).

Hoerl and Kennard (1970a) suggested that, with multicollinear data, an estimator is required that has both smaller MSE and shorter expected length than an OLS estimate.[8] Hoerl and Kennard have proposed the following estimator:

$$\hat{B}_s^r = \left[R + Ik\right]^{-1}.G \qquad (24)$$

where $\hat{B}_s^r$ is the standardised ridge estimator,
R is the matrix of explanatory variables,
I is an identity matrix consisting of the value one on the leading diagonal and zeros elsewhere,
k is an additive constant, ranging from zero to one,
and G is the vector of correlations between the dependent and explanatory variables.

Loosely speaking, the value k is the 'biasing' value; when k is set to zero the resulting coefficients are the OLS estimates. However, when k is greater than zero, the relationship between ridge and OLS estimators can be illustrated as follows. Recalling that

$$G = \left[R\right] \hat{B}_s \qquad (25)$$

and substituting into (24) results in

$$\hat{B}_s^r = \left[R + Ik\right]^{-1} \left[R\right] \hat{B}_s \qquad (26)$$

or $$\hat{B}_s^r = Z \hat{B}_s \qquad (27)$$

Thus the ridge estimator is a biased estimator, biased by the amount Z. (When k is zero the Z matrix will be on identity matrix and therefore $\hat{B}_s^r$ will equal $\hat{B}_s$.)

The MSE of the ridge estimator is given by

$$E\,(L_r)^2 = \text{Trace}\left[\sigma_\varepsilon^2 \; Z \; \left[R\right]^{-1} Z\right] + \beta'(Z-I)' \; (Z-I)\,\beta \qquad (28)$$

$$= \text{variance} \qquad + \text{bias}^2$$

If one examines this equation carefully and also considers the nature of the Z matrix, it will be seen that the variance term is a continuous <u>decreasing</u> function of k while the bias squared is a continuous <u>increasing</u> function of k. Given these characteristics of the ridge estimates, Hoerl and Kennard (1970a, 61) state:

> 'these properties lead to the conclusion that it is possible to move $k > 0$, take a little bias, and substantially reduce the variance, thereby improving the mean square error of estimation'.

In other words, equation (28) is the mathematical version of Figure 2.1c where we trade a small amount of bias for a substantial reduction in variance. Furthermore, Hoerl and Kennard show that, for a k greater than zero, the vector of the ridge estimates is likely to be 'shorter' than the OLS coefficient vector.

An important question is 'what value of k should be added to the explanatory variable correlation matrix?' Hoerl and Kennard have suggested that the ridge trace (a graphical method) should be employed to choose a suitable value of k. The usual procedure for the construction of this graph is to calculate a value for each ridge coefficient for all the explanatory values for several values of k in the range from zero to one, with emphasis on values less than 0.3. The values of the ridge estimates are plotted on the vertical axis against the corresponding values of k on the horizontal axis. Thus there is a curve for each explanatory variable.

Two characteristic ridge traces appear in Figure 2.2. Figure 2.2a shows the typical ridge trace for multicollinear data. It is distinguished by criss-crossing of the coefficient curves, sign changes and large changes in coefficients but, at some value of k greater than zero, the trace stabilises. Figure 2.2b, in contrast, depicts the ridge trace for orthogonal variables. Coefficient curves do not cross each other but gradually decline. Such a ridge trace obviously protrays the 'sensitivity' of the estimates to changing values of k, and it also enables the analyst to choose an appropriate value for k. According to Hoerl and Kennard the analyst should choose a value for k at the point where the ridge trace has stabilised and is behaving like 'an orthogonal system'. Obviously, an analyst does not wish to import too great a degree of bias to the estimation, and he will therefore attempt to pick a value for k as close to zero as possible. A more detailed ridge-trace rule has been proposed by Brown (1973, 29):

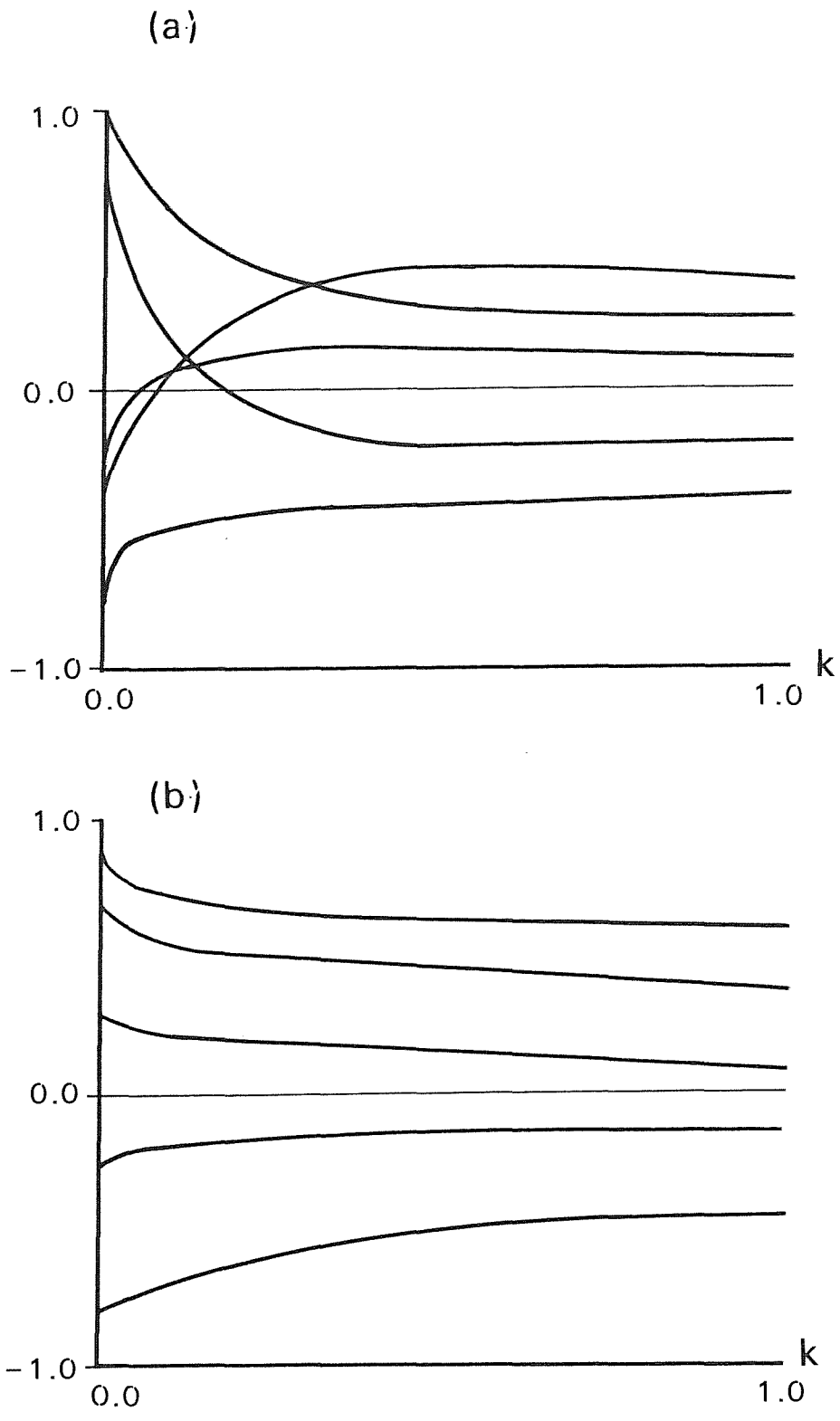


FIGURE 2.2 CHARACTERISTIC RIDGE TRACES

'select a particular value of k at the point where the last ridge estimate attains its maximum absolute magnitude after having obtained its 'ultimate' sign, where 'ultimate' sign is defined as being the sign at, say, k = 0.9'.

The technique illustrated[9]

Having discussed the mathematical background, the technique can now be applied to mortality data. Let us first consider the Glamorgan local authority data relating (y) deaths from coronary heart disease (males aged 35 to 64 years) to three 'explanatory' variables: number of males aged 35-44 (x_1), 45-54 (x_2) and 55-64 years (x_3). We can expect such data to be multicollinear and indeed, as we have seen, OLS estimation of the model results in implausible coefficients. The eigenvalues for the three explanatory variables are given in Table 2.3.

Table 2.3

Eigenvalue analysis of the Glamorgan data

	Eigenvalue	Percentage of total variation
λ max, λ_1	2.991	99.72
λ_2	0.007	0.22
λ min λ_3	0.002	0.05
total	3.000	100.00

If the explanatory variables were othogonal, each eigenvalue would be approximately equal to one. Eigenvalues are closely associated with principal components and, indeed, they measure the amount of variation accounted for by a particular component. For example, λ_1 is the eigenvalue associated with the first or major component, and this component for the Glamorgan data accounts for $\frac{2.991}{3.000} * \frac{100}{1}$ per cent of the variation, that is over 99 per cent of the variation in the explanatory variables can be accounted for by just one

component. (The value 3.000 is the sum of the eigenvalues and will always equal the number of explanatory variables.) In other words, as the first component accounts for nearly all the variation in the explanatory variable, it is unlikely that there is sufficient variation within the data for the successful estimation of three OLS regression coefficients. The sum of the reciprocal of the eigenvalues is 797.6 and, therefore, the expected MSE of the least-squares coefficients is 797.6 σ_ε^2 (from equation (20)). This is over 250 times greater than that expected from an orthogonal system ($3\sigma_\varepsilon^2$). Such a data set, with its apparent high multicollinearity, appears a suitable candidate for analysis by ridge regression. The standardised ridge regression values are given by

$$\hat{B}_s^r = [R + Ik]^{-1} . c \qquad (24)\text{bis}$$

For k = 0 (the ordinary least squares case)

$$\hat{B}_s^r = \left[\begin{bmatrix}1.0000 & 0.9969 & 0.9933\\ 0.9969 & 1.0000 & 0.9974\\ 0.9933 & 0.9974 & 1.0000\end{bmatrix} + 0.0 * \begin{bmatrix}1.0 & 0.0 & 0.0\\ 0.0 & 1.0 & 0.0\\ 0.0 & 0.0 & 1.0\end{bmatrix}\right]^{-1} * \begin{bmatrix}0.9890\\ 0.9940\\ 0.9945\end{bmatrix}$$

$$= \begin{bmatrix}-0.21\\ 0.65\\ 0.55\end{bmatrix} \qquad (29)$$

For k = 0.1

$$\left[\begin{bmatrix}1.0000 & 0.9969 & 0.9933\\ 0.9969 & 1.0000 & 0.9974\\ 0.9933 & 0.9974 & 1.0000\end{bmatrix} + 0.1 \begin{bmatrix}1.0 & 0.0 & 0.0\\ 0.0 & 1.0 & 0.0\\ 0.0 & 0.0 & 1.0\end{bmatrix}\right]^{-1} * \begin{bmatrix}0.9890\\ 0.9940\\ 0.9945\end{bmatrix}$$

$$= \begin{bmatrix}0.29\\ 0.33\\ 0.34\end{bmatrix} \qquad (30)$$

This procedure was repeated for a number of values of k in the range zero to one. The results are plotted in Figure 2.3, from which it is clear that the coefficient estimates are not stable; a small increase in the value of k shows a major change in the value of the estimates. Applying Hoerl and Kennard guidelines and Brown's rule for choosing an appropriate value of k, we can obtain the estimates presented in Table 2.4.

FIGURE 2.3 RIDGE TRACE FOR GLAMORGAN DATA

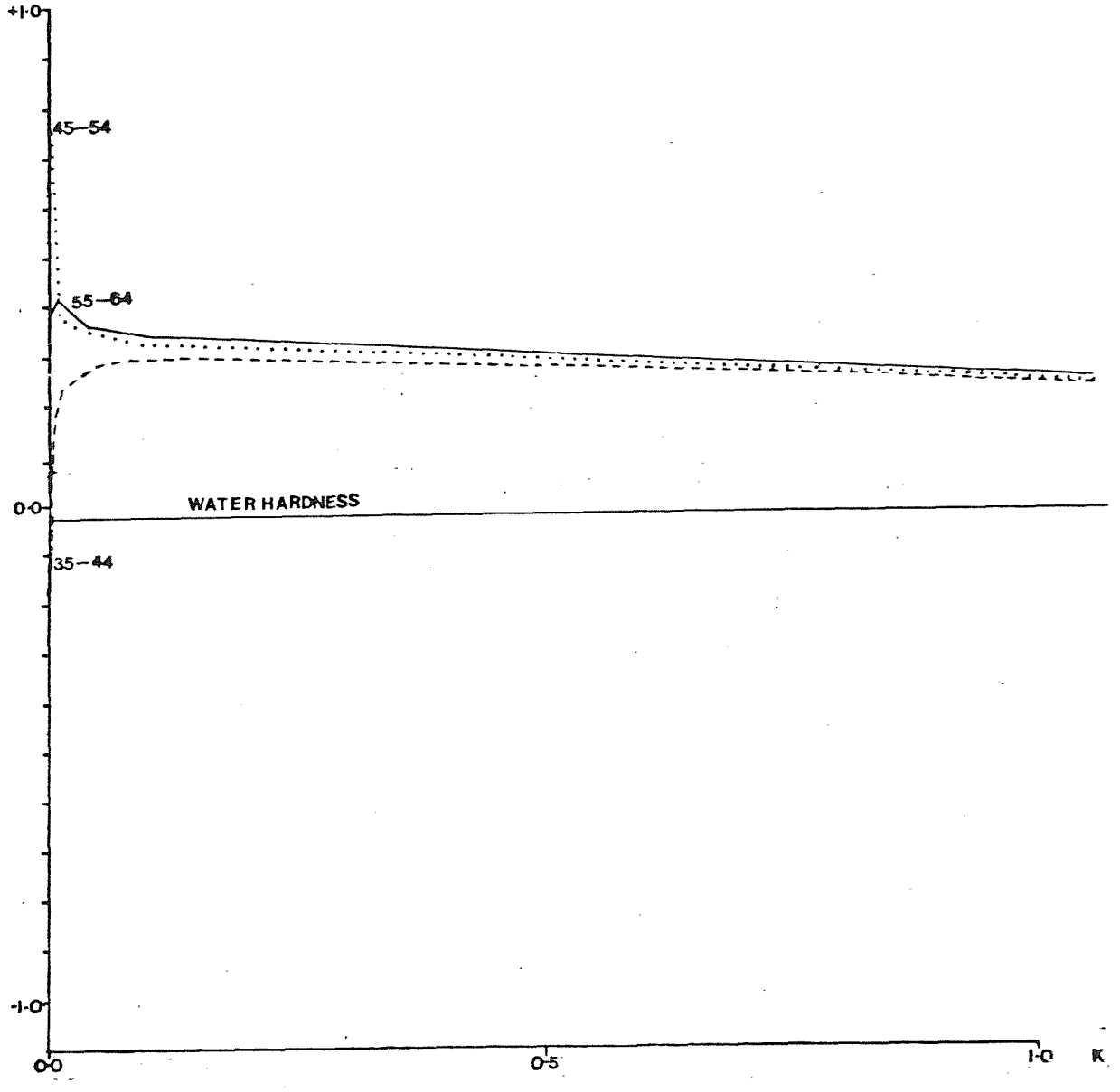


Table 2.4

Ridge trace estimates for the Glamorgan data

Explanatory variable	OLS	Hoerl and Kennard's guidelines	Brown's rule
x_1:35-44	-0.21	0.29	0.30
x_2:45-44	0.65	0.33	0.31
x_3:55-65	0.55	0.34	0.32
k	0.0	0.1	0.15

From this table it is clear that both sets of ridge coefficients are very similar, and are very different to the OLS estimates. Remembering the earlier analysis of this Glamorgan data, it was anticipated that the coefficients would all be positive, with the older age group (x_3) accounting for more of the variation in coronary heart disease than the younger age groups. While this is not the case for the OLS estimates, both sets of ridge coefficients have the expected sign and order of importance. Ridge regression undoubtedly gives a more reasonable interpretation of the Glamorgan data. While the coefficients obtained by OLS estimation are unlikely to be repeated exactly for another set of data, the ridge estimates are intuitively plausible.

The discussion can now turn to a consideration of West and Lowe's (1976) analysis of the variation of ischaemic heart disease mortality. In this study three explanatory variables (temperature, rainfall and a socio-economic index) were thought to account for the variations in death rates. Our previous examination of these variables produced ambivalent results; the Haitovsky test did not detect multicollinearity, but the Farrar-Glauber tests shows all three variables to be involved in a multicollinear relationship. An eigenvalue analysis of this data is presented in Table 2.5. As even the smallest eigenvalue accounts for twenty-two per cent of the total variation, and the sum of the

Table 2.5

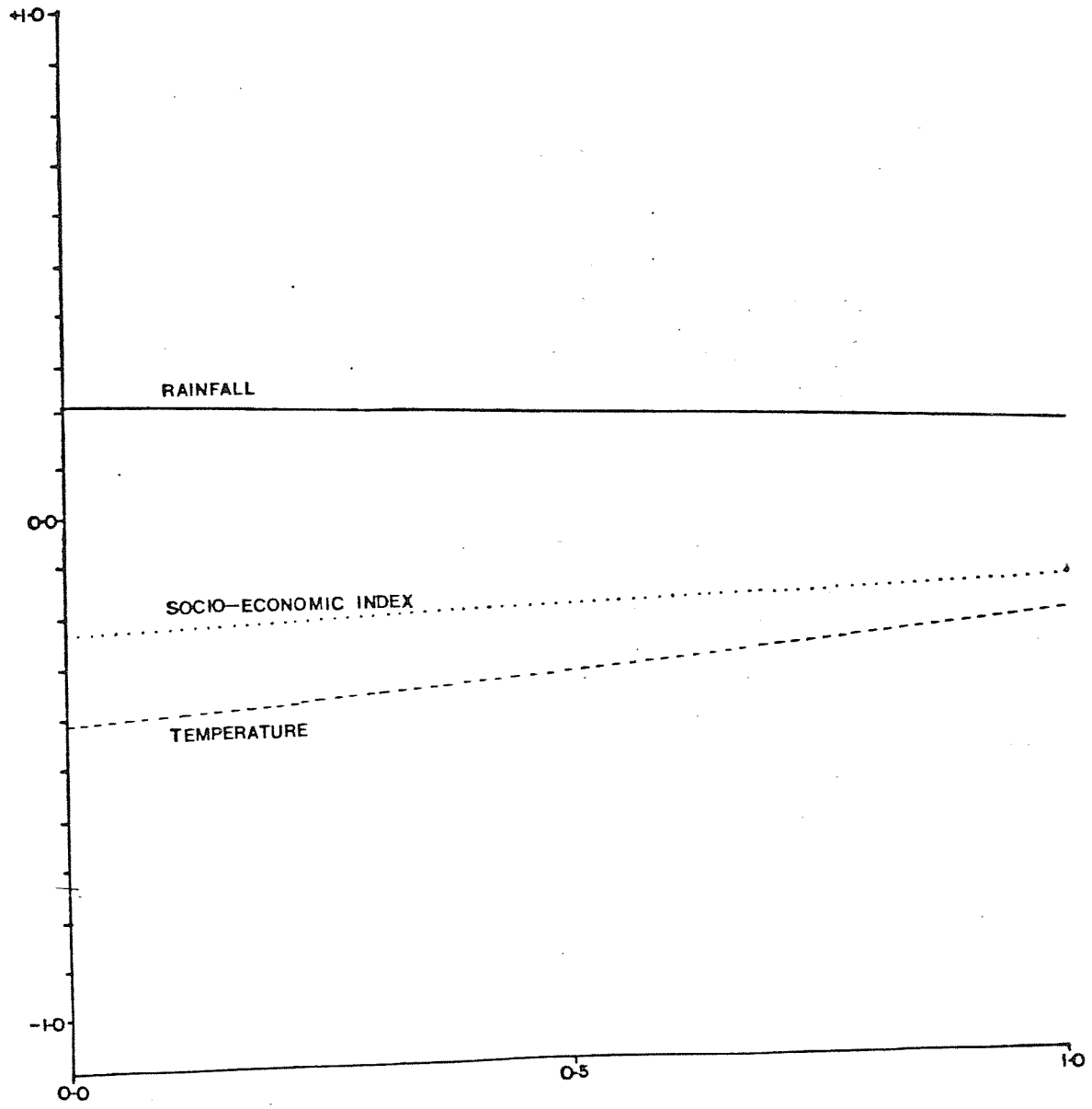
Eigenvalue analysis of West and Lowe's (1976) study

	Eigenvalue	Percentage of total variation
λmax λ_1	1.56	52
λ_2	0.78	26
λmin λ_3	0.67	22
Total	3.00	100

reciprocals of the eigenvalues is only $3.43\,\sigma^2_\varepsilon$ (compared with an orthogonal value of $3.0\,\sigma^2_\varepsilon$), it appears that the data is not severely multicollinear. The ridge trace for this model, which behaves like an orthogonal system with all the coefficients declining gradually, is shown in Figure 2.4. Both Hoerl and Kennard's guidelines and Brown's rule suggest that a value of zero is appropriate for k. The data do not appear to be multicollinear and the biased ridge estimates are unlikely to be an improvement over OLS estimates (k=0). In this particular example, while the ridge trace does not lead to improved estimation and interpretation, it does indicate to the analyst that the data is not multicollinear, a similar result to that suggested by the Haitovsky test.

The final example to be briefly considered is that of Preston (1970). Previous analysis of this data set by the Farrar-Glauber tests indicated that only some of the explanatory variables are likely to be troubled with multicollinearity. An eigenvalue analysis of this data shows similar results, for while the data appear to be 'rich' enough for the successful estimation of several coefficients, there is perhaps insufficient variation for the confident estimation of them all. (The three smallest eigenvalues only account for 1.6 per cent of the variation.) The ridge trace (although it is highly complicated with ten explanatory

FIGURE 2.4 RIDGE TRACE FOR WEST AND LOWE'S (1976) STUDY



variables and is therefore not shown) indicates that, while some coefficients behave in an orthogonal manner, others show large absolute changes in value. Only some of the coefficients, therefore, appear to be drastically altered by the introduction of bias into the estimation process. Indeed, it is those variables identified by the Farrar-Glauber analysis as being multicollinear that have undergone the greatest absolute changes. Moreover, some variables (for example cigarette smoking, with a negative OLS coefficient) which appear to be implausible at k=0 conform to a priori expectations when k is chosen according to Hoerl and Kennard's guidelines and Brown's rule.

On the basis of these three examples of ridge regression it may be concluded that the technique offers a substantial improvement over OLS when the explanatory variables of a model are highly related. In the analyses of the Preston (1970) and Glamorgan data, the relationship between the explanatory and dependent variables became more intuitively reasonable when ridge regression was used. Moreover, ridge regression appears to be a useful technique for revealing the severity of multicollinearity.[10]

An assessment of ridge regression

Despite the preceding discussion and illustrations, opinions on ridge regression are divergent and the arguments for and against the technique are most clearly seen in the debate between Conniffe and Stone (1973, 1975) and Smith and Goldstein (1975). Conniffe and Stone are critical of the use of Euclidian distance to measure the difference between the estimated $\hat{B}$ and true β. They also have reservations about its use in estimating mean square error (equations (20) and (21)). Theobald (1974), however has considered some generalisations of mean square error, and in doing so has provided a much stronger justification for ridge regression than that offerred by Hoerl and Kennard. Another criticism

of ridge regression has been the choice of k, or how much bias to introduce into the model. This particular criticism is based on two arguments.

(1) Hoerl and Kennard's proof that there exists an estimator which has smaller mean square error than OLS is dependent on k being a constant whereas, in fact, k is estimated from the data.

(2) Hoerl and Kennard's method of choosing a value of k by a graphical method seems rather vague. Moreover, there is no evidence that the 'stability' of k measures the 'correctness'of k.

Although Smith and Goldstein (1975) do not fully agree with these criticisms, they do admit that the ridge trace may or may not be a good method of choosing k and suggest that alternative methods should be examined by simulation experiments. A number of simulations of ridge regression have now been published.

The early experiments proved to be disappointing. Newhouse and Oman (1971) conducted the first simulation study of ridge estimators. Their study only examined the case of two explanatory variables and two different values for the inter-correlation between these variables. After investigating a number of different schemes for choosing k, Newhouse and Oman (1971, 15) stated that:

> 'we found that the estimator as proposed contains serious flaws, and we caution against the use of ridge analysis, as now defined, to estimate regression coefficients'.

McDonald and Galarneau (1975), in another simulation study, presented results for the three-explanatory variable case. Although ridge regression was not always better than OLS, they concluded there was sufficient improvement to warrant further investigations of ridge estimators. For the case of two explanatory variables, McDonald and Galarneau found comparable results to Newhouse and Oman and suggested that

some advantage may accrue to ridge regression when there is a large number of explanatory variables.

To examine this suggestion further we must turn away from simulation studies and look at theoretical work. Stein (1960) and James and Stein (1961) have investigated the mean square error of an estimator that shortens the length or 'shrinks' the vector of regression coefficients. In these papers it was proved that there existed an estimator that had smaller mean square error than the OLS estimator. However, this proof required that there were more than three explanatory variables. Moreover, Hocking and his co-workers (1976) have shown that the James-Stein estimator is one of a family of so-called 'biased' estimators of which ridge regression is another member. Although numerous papers have been published examining, comparing and contrasting different members of the 'biased family' (Table 2.6), the important point for ridge regression is that there is a suggestion that the method is only appropriate when there are more than three explanatory variables.[11]

Returning now to simulation studies, Hoerl, Kennard and Baldwin (1975) concluded that ridge techniques offered considerable improvement over OLS estimation and that this improvement became more pronounced both as the number of explanatory variables increased and as the multicollinearity amongst these explanatory variables became more severe. This conclusion is reinforced by a similar study by Lawless and Wang (1976). Also, in what must be regarded as the most extensive simulation to date, Dempster, Schatzoff and Wermuth (1977, 77) write of

> 'possible drastic improvements over least-squares, especially through the technique of ridge regression, especially when a high degree of correlation exists among the independent variables'.

This study examined no less than 57 estimation methods (including the James-Stein, principal component, ridge and

Table 2.6

Articles containing analytical comparisons of ridge and other non-OLS estimators

Articles	Author
Ridge/minimum variance linear unbiased estimator/Stein-Lake estimator/ minimum conditional mean square estimator/mixed regression estimator/ maximum likehood estimator.	Swamy (1973)
Ridge/shrinkage estimators.	Mayer and Willke (1973)
Ridge/Chipman's minimum mean square estimator.	Bacon and Hausman (1974)
Ridge/shrinkage estimators/James-Stein estimator/generalized inverse estimator.	Goldstein and Smith (1974)
Ridge/principal components estimator (latent-root analysis).	Hawkins (1975)
Ridge/Farebrother's minimum mean square estimator.	Farebrother (1975)
Ridge/shrinkage estimators.	Obenchain (1975)
Ridge/generalized inverse/James-Stein estimator.	Hocking, Speed and Lynn (1976)
Farebrother's minimum mean square error estimator/James-Stein estimator.	Vinod (1976b)
Ridge/shrinkage estimator.	Rolph (1976)
Ridge/principal components/generalized inverse/James-Stein estimator.	Oman (1978)
Ridge/latent root/shrunken estimator.	Gunst (1979)

OLS estimators) while consideration was given to numerous selection procedures. Ultimately, the ridge technique emerged from this comprehensive review as the most suitable method for analysing multicollinear data.

The Bayesian Perspective

The weight of simulation evidence now appears to corroborate ridge regression as a valuable technique, and further support for the method comes from a Bayesian perspective. Although not generally considered in geography, there are two major theoretical justifications for least-squares estimation, namely the frequentist justification (based on the minimum variance unbiased property of OLS estimates) and the Bayesian one. The basic step in the Bayesian approach is to try to translate *a priori* information into 'prior probabilities' about the likely value of the unknown regression coefficients. The 'posterior probabilities' of the likely values of the regression coefficients are then obtained by combining the prior probabilities with the information contained in the sample, using a method developed by Thomas Bayes in the middle of the eighteenth century.

Corresponding with these two theoretical viewpoints on the justification of OLS regression, there are two positions on what is wrong with (and what can be done about) OLS estimates and multicollinear data.

(1) The frequentist position: the variance of the estimated coefficients is too high and needs to be reduced by introducing bias into the estimation procedure.

(2) The Bayesian position: the uniform prior distribution of OLS estimation is incorrect and needs to be replaced with a realistic prior distribution (Dempster, 1973).

While Hoerl and Kennard, as originators of ridge estimation, cast their work in a frequentist mould, they were also aware of a Bayesian interpretation in that each ridge estimate can be considered as an outcome of choosing a prior normal (Gaussian)

distribution with a mean of zero. Subsequently, a number of workers have examined ridge regression from a Bayesian viewpoint and have found that it can be given an explicit formulation in Bayesian terms, thus corroborating the evidence from the frequentist viewpoint and the recent simulation experiments (Table 2.7).[12]

Choosing k

So far in our discussion we have only dealt with methods for choosing k based on the ridge trace. However, in both theoretical and simulation studies, numerous mechanical rules have been suggested to enable an 'exact' k value to be calculated. One of the simplest rules is that suggested by Hoerl, Kennard and Baldwin (1975) with

$$k = m.s_e^2 \; / \; \hat{B}' \; \hat{B} \qquad (31)$$

m is the number of explanatory variables,
s_e^2 is an estimate of the variance of the disturbance term,
and $\hat{B}$ is the vector of OLS estimated coefficients.
A more complicated estimator (RIDGM) is the one proposed by Dempster, Schatzoff and Wermuth (1977) and this particular mechanical rule has received some support from several simulation studies (Darlington, 1978, 80), and in a study by Lawless (1978) an estimator similar to RIDGM performed well in almost all the differing simulation models that were investigated. However, there is still considerable disagreement in the statistical literature over the 'best' means of choosing k. For example, Hemmerle and Brantle (1978, 119) examined 7 different estimators (including RIDGM) in a simulation study and concluded that, while ridge estimators were an improvement over OLS, 'no uniformly best procedure was apparent'. Moreover, one should not concentrate on mechanical rules to the exclusion of the ridge trace. Wichern and Churchill (1978, 310), after examining a number of different mechanical ways of choosing a value of k, have written:

Table 2.7

Articles containing a specifically Bayesian interpretation of ridge regression

Title	Author
Alternatives to least squares in multiple regression	Dempster (1973)
Ridge type estimators for regression analysis	Goldstein and Smith (1974)
Bayesian view of ridge regression	Hsiang (1975)
Bayes linear estimators	La Motte (1978)
Bayes estimates for the linear model	Lindley and Smith (1972)
Bayesian comparison of some estimators used in linear regression with multicollinear data	Oman (1978)
Choosing shrinkage estimates for regression problems	Rolph (1976)

> 'If ridge estimates are to be employed we suggest the ridge trace be used in conjunction with any mechanistic process for generating ridge estimates .... Some weight should certainly be given to sign reversals in the estimated coefficients and those are only evident from the ridge trace'.

In summary, there does not exist, as yet, an optimum and exact method for determining the correct value of k but, if one adopts an exploratory approach based on graphical analysis, then the ridge trace is a valuable and effective means of portraying the sensitivity of the estimated coefficients to introducing bias into the estimation procedure. Let us conclude this review of previous work on the subject with a quote from Vinod's (1978b, 120) comprehensive overview:

> 'In ... independent studies by investigators from diverse fields, the superiority of RR [ridge regression] over OLS is almost always noted; although there is wide disagreement over the optimum RR method'.

A simulation experiment

The discussion has so far considered previous research into ridge regression and a reasonable assessment of this work is that recent simulations have shown the method to be of considerable value when analysing multicollinear data. Moreover, the technique has received theoretical support not only from a frequentist perspective but also from a Bayesian viewpoint. However, in this concluding section of this chapter, it will be shown that previous studies have ignored a critical problem of the technique. A simulation experiment will be performed to illustrate the problem and to elucidate the conditions under which it occurs.

Ridge regression, as previously stated, tries to gain a large decrease in the variance of an estimated coefficient from a slight increase in its bias. The formula for determining the bias of the estimated coefficient is correctly given in Hoerl and Kennard (1970a) but these workers did not fully consider the impact of the true regression coefficients

on the degree of bias. If we take the simple if rather unrealistic case of two explanatory variables, the bias is given by

$$BIAS = \frac{-k}{(1+k)^2 - r_{12}^2} \left((1+k)\ \beta_1 - r_{12}\ \beta_2 \right) \qquad (32)$$

where β_1 and β_2 are the true but unknown regression coefficients and r_{12} is the correlation between x_1 and x_2. On close inspection of this equation, it will be found that bias will be increased in the following situations.

(1) Bias will be high when there is a positive correlation between the two explanatory variables and the true β values have an opposite sign and are similar in magnitude.

(2) Bias will also be high when there is a negative correlation between the two explanatory variables and the true β values are similar in both size and magnitude.

Given such results it appears that the ridge procedure can be expected to be an extremely good biased estimator in certain circumstances but a very bad biased estimator in other circumstances. Obviously, it is of vital importance to be able to differentiate between these two situations but, unfortunately, we cannot use a formula such as equation (32) because we do not know the true regression parameters. However, following Brown and Beattie (1975) and Ryan and Perrin, (1973), it may be possible to use a statistical test of the differences in MSE between OLS and ridge estimators as a guide to the bias induced by the ridge technique. The null hypothesis is that the MSE of the ridge estimates is less than or equal to the MSE of the OLS estimates and the test statistic u is given by

$$u = \frac{(\hat{B}_s - \hat{B}_s^r)' \ [\bar{R}] \ (\hat{B}_s - \hat{B}_s^r)}{m} \div \frac{\Sigma e^2}{(n-m)} \qquad (33)$$

where m is the number of explanatory variables
n is the number of observations
$\hat{B}_s$ is the vector of standardised OLS coefficients
B^r is the vector of standardised ridge coefficients
$\bar{R}^s$ is the explanatory variable correlation matrix and
$\hat{e}$ is the vector of residuals from the OLS regression.

Tables for the evaluation of the 'significance' of the statistic are given by Wallace (1972) but, if the measure is going to be used in an exploratory manner as a guide to the degree of bias, a low value of u indicates that the MSE of the ridge values is less than the MSE of the OLS coefficients.

The usefulness of this u statistic in practice has not been considered by previous researchers and given our relectance to place too much faith in a single summary statistic, it was decided to conduct a simulation study of the efficacy of the procedure. In addition, the simulations experiments can also be used to evaluate the performances of the ridge estimators with different degrees of multicollinearity. While fuller details of the simulation are given in an Appendix it is sufficient for present purposes to know that each model consisted of 100 observations for a dependent variable and nine explanatory variables. The results of three different experiments are given in Table 2.8 a,b,c and they are based on 50 replications of each model, the ridge estimates being derived by the application of Brown's rule to the ridge trace.

Table 2.8a shows the results when the explanatory variables are not multicollinear and the dependent variable is related to each of nine explanatory variables by +0.5. Clearly the mean of the OLS estimates is very close to the true values and, as shown by the standard deviations, the estimates group tightly around these mean values. The OLS estimators are accurate and precise with this orthogonal data set. With regard to the ridge procedure it will be found that the values in Table 2.8a are exactly the same as the OLS estimates for, in each of the 50 replications, the ridge trace correctly identified that the data set was orthogonal and correctly suggested that the OLS estimates were the best possible ones.

In contrast Table 2.8b shows the results of analysing severely multicollinear data.[13] In this particular experiment the pattern of multicollinearity is complex:

Table 2.8a

Comparison of OLS and ridge estimates: no multicollinearity

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
				TRUE VALUES					
	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
				OLS ESTIMATES					
mean	0.49	0.50	0.49	0.51	0.49	0.51	0.49	0.49	0.49
standard deviation	0.08	0.08	0.07	0.09	0.09	0.09	0.09	0.08	0.09
minimum	0.30	0.34	0.34	0.28	0.30	0.33	0.32	0.34	0.28
maximum	0.67	0.68	0.66	0.70	0.63	0.77	0.85	0.71	0.66
				RIDGE ESTIMATES					
mean	0.49	0.50	0.49	0.51	0.49	0.51	0.49	0.49	0.49
standard deviation	0.08	0.08	0.07	0.09	0.09	0.09	0.09	0.08	0.09
minimum	0.30	0.34	0.34	0.28	0.30	0.33	0.32	0.34	0.28
maximum	0.67	0.68	0.66	0.70	0.63	0.77	0.85	0.71	0.66

u statistic

	OLS	Ridge
mean	0.0	0.0
minimum	0.0	0.0
maximum	0.0	0.0

Table 2.8b

Comparison of OLS and ridge estimates:
positive multicollinearity, all positive true coefficients

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
				TRUE VALUES					
	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
				OLS ESTIMATES					
mean	0.51	0.50	0.51	1.72	0.47	0.43	-0.24	1.50	0.5
standard deviation	0.08	0.08	0.09	6.37	3.19	3.18	3.57	4.12	4.54
minimum	0.29	0.22	0.23	-9.31	-6.69	-6.60	-8.23	-6.64	-9.46
maximum	0.67	0.73	0.71	20.12	7.50	7.37	6.71	10.70	10.57
				RIDGE ESTIMATES					
mean	0.50	0.49	0.50	0.49	0.43	0.40	0.48	0.50	0.59
standard deviation	0.07	0.09	0.09	0.04	0.07	0.06	0.04	0.05	0.06
minimum	0.29	0.20	0.27	0.44	0.26	0.25	0.37	0.34	0.47
maximum	0.64	0.73	0.69	0.56	0.57	0.54	0.57	0.61	0.73

u statistic

	OLS	Ridge
mean	0.0	1.19
minimum	0.0	0.89
maximum	0.0	1.48

Table 2.8c

Comparison of OLS and ridge estimates:
positive multicollinearity, one negative true coefficient

	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9
					TRUE VALUES				
	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	-0.5
					OLS ESTIMATES				
mean	0.50	0.54	0.50	0.06	0.55	0.54	0.2	0.24	-0.55
standard deviation	0.07	0.09	0.08	5.07	2.58	2.59	3.43	3.54	3.69
minimum	0.39	0.35	0.34	-8.49	-4.04	-4.25	-9.99	-7.29	-9.91
maximum	0.71	0.71	0.64	11.23	6.93	7.21	8.70	5.84	6.12
					RIDGE ESTIMATES				
mean	0.49	0.51	0.49	0.16	0.09	0.07	0.17	0.16	0.11
standard deviation	0.06	0.08	0.08	0.03	0.07	0.06	0.05	0.04	0.04
minimum	0.36	0.34	0.32	0.11	-0.05	-0.06	0.01	0.03	-0.03
maximum	0.68	0.70	0.67	0.26	0.23	0.21	0.26	0.23	0.20

u statistic

	OLS	Ridge
mean	0.0	119.4
minimum	0.0	106.8
maximum	0.0	135.9

x_1, x_2 and x_3 are orthogonal, x_7 is related to x_4, x_8 is related to x_5, and x_9 is related to x_5 and x_6, and all the relationships are positive. The OLS estimates are near the true regression coefficient for the orthogonal variables but are wildly imprecise and inaccurate for the multicollinear data. Even the mean values of the regression coefficients are not close to the true values[14] and the standard deviations show that the OLS estimates have a large-spread. However, the spread of ridge estimates is much lower and they represent a considerable improvement over OLS regression. To emphasize this point one needs to consider the range of the estimated coefficients; the OLS values extend from a minimum of -9.46 to a maximum of 20.12 while these values for the ridge procedures are only 0.20 and 0.73 respectively.

For Table 2.8c the pattern of multicollinearity is the same as for 2.8b but the regression coefficient associated with x_9 is -0.5 while all the other coefficients remain at +0.5. Again the OLS estimates have a wide spread and are generally unsatisfactory. Moreover, as anticipated earlier, the ridge estimates in this particular case are not an improvement over the OLS estimates. While the ridge values have a low spread they are tightly grouped around an inappropriate estimate of the true value and, as the table shows, there has been a general underestimation of all the multicollinear coefficients. Fortunately, however, the u statistics clearly reveal the problem and, for the 50 models summarised in the table, the u values were uniformly high when compared to the values of previous tables. In conclusion, therefore, ridge estimates are, in certain circumstances, a major improvement over the commonly used method of OLS regression analysis. But, in other situations, both OLS and ridge are equally poor and while no method appears universally appropriate it is at least possible to identify when ridge is likely to be the better estimator.[15]

CHAPTER CONCLUSIONS

Areal mortality data can be expected to be multicollinear; indeed, a researcher uses multiple regression because of the inter-relationships among the explanatory variables. In geography, the commonest approach to the problem has been to ignore the difficulty,but the analysis of highly multicollinear data with OLS regression is highly likely to result in inferential error. Moreover, the methods that have been used by geographers to overcome multicollinearity (omission of offending variables, stepwise regression, principal components analysis) have such major drawbacks that even the most circumspect of applications are of doubtful validity. In contrast, the judicious use of ridge regression holds considerable promise for the analysis of highly inter-related data. While this procedure originally came under attack because of its lack of theoretical base and poor performance in simulation studies, recent research has strengthened the theory and more elaborate simulation studies have provided favourable evaluations. Moreover, as the present study has shown, ridge regression is a major improvement over OLS in certain circumstances and it has proved possible to identify those situations in which the technique performs relatively badly. While the analysis of multicollinear data still requires a great deal of research, the ridge procedure is an important tool of exploratory analysis. As a number of authors have contended (Chatterjee and Price, 1977; Price, 1977 and Winer, 1978) a particularly appealing aspect of the technique is the ridge trace,for this not only alerts the analyst to the presence of multicollinearity but it also provides a graphic demonstration of the variety of possible coefficients that can be estimated from one set of data. The ridge procedure will be further illustrated in Part II of this thesis.

CHAPTER 2 : APPENDIX

Construction of models used in the simulation experiment

(1) Generating orthogonal variables

To generate three variables (x_1, x_2, x_3) which are independent of each other, all that is required is the generation of three random variables with the use of the program (G05AEF) which is available from the Numerical Algorithms Group.

(2) Generating two variables that are related to each other to a known degree

In order to generate a variable x_7 which has a pre-determined correlation with a variable x_4, first generate a vector of observations x_4 with a mean of zero, a variance of one and following a normal distribution. (This can readily be achieved by using G05AEF.) The problem is then to generate the variable x_7 as a linear function of x_4 plus an error term:

$$x_7 = \beta x_4 + \varepsilon \qquad (1)$$

This can be re-written as:

$$\text{variance } (x_7) = \beta^2 . \text{variance } (x_4) + \text{variance } (\varepsilon) \qquad (2)$$

To solve for the three unknowns in this equation (variance (x_7), variance (ε), and β^2) use is made of the following identity

$$E(\beta) = r_{x_7 x_4} \cdot \frac{\text{standard deviation } x_7}{\text{standard deviation } x_4} \qquad (3)$$

Setting the standard deviation of x_7 arbitrarily to one, the ratio of the standard deviation of x_7 to the standard deviation of x_4 is also equal to one (remembering that x_4 has been generated with a variance of one) and thus E (β) equals $r_{x_7 x_4}$, the correlation between x_7 and

x_4. Equation (2) can now be solved for the remaining unknown variance of ε. For example, if a correlation of 0.9 is required between x_7 and x_4 equation (2) becomes

$$1 = (.90)^2 \ (1) + \text{variance } (\varepsilon) \tag{4}$$

$$1 = (.90)^2 = \text{variance } (\varepsilon) = .19 \tag{5}$$

These values can now be used to generate x_7:

$$x_7 = (.9) \ (x_4 \sim N \ (0,1)) + (\varepsilon \sim N(0,.19)) \tag{6}$$

where $\sim N(0,1)$ means generate data having a normal distribution with a mean of zero and a variance of one. x_7 is thus generated so that it is related by a known degree to x_4; in a similar manner x_8 was generated so as to be related to x_4.

(3) <u>Generating two variables that are related to a third variable to a known degree</u>

In order to generate a variable x_9 which has a predetermined correlation with the variables x_5 and x_6, first generate two vectors x_5 and x_6 with a mean of zero and a variance of one. As these two variables are not related to each other by any functional relationship they can be regarded as being independent of each other. The problem is then to generate x_9 as a linear function of x_5 and x_6 plus an error term

$$x_9 = \beta_1 \ x_5 + \beta_2 \ x_6 + \varepsilon \tag{7}$$

This equation can be rewritten as

$$\text{variance } (x_9) = \beta_1^2 \cdot \text{variance } (x_5) + \beta_2^2 \cdot \text{variance } (x_6) + 2 \cdot \beta_1 \cdot \beta_2 \cdot \text{covariance } (x_5 x_6) + \text{variance } (\varepsilon) \tag{8}$$

As x_5 and x_6 are generated so as to be orthogonal the covariance term in equation (8) is zero and the equation reduces to

$$\text{variance } (x_9) = \beta_1^2 \cdot \text{variance } (x_5) + \beta_2^2 \text{ variance } (x_6) + \text{variance } (\varepsilon) \tag{9}$$

The unknown quantities in this equation are variance (x_9), β_1^2, β_2^2 and variance (ε) for x_5 and x_6 have been generated with a variance of one. To solve for these unknowns use is made of the following three identities:

$$E(\beta_1) = r\, x_9 x_5 \cdot \frac{\text{standard deviation } x_9}{\text{standard deviation } x_5} \qquad (10)$$

$$E(\beta_2) = r\, x_9 x_6 \cdot \frac{\text{standard deviation } x_9}{\text{standard deviation } x_6} \qquad (11)$$

$$R^2 = r^2 x_9 x_5 + r^2 x_9 x_6 \qquad (12)$$

For example, if a multiple correlation of 0.9 is required between the three variables x_9, x_5 and x_6, equation (12) gives

$$(0.9)^2 = r^2 x_9 x_5 + r^2 x_9 x_6 \qquad (13)$$

$$(0.81) = r^2 x_9 x_5 + r^2 x_9 x_6 \qquad (14)$$

Allowing both variables to contribute equally to determining x_9 gives

$$r^2 x_9 x_5 = r^2 x_9 x_6 = .405 \qquad (15)$$

Therefore

$$r x_9 x_5 = r x_9 x_6 = \sqrt{.405} = 0.636 \qquad (16)$$

Setting the standard deviation of x_9 arbitarily to one, equations (10) and (11) can be easily solved (remembering that x_5 and x_6 have been generated with a variance of one).

$$\beta_1 = 0.636 \qquad (17)$$

$$\beta_2 = 0.636 \qquad (18)$$

The remaining unknown of equation (9) the variance of ε can now be found as follows:

$$1 = (0.636)^2(1) + (0.636)^2(1) + \text{variance } (\varepsilon) \qquad (19)$$

$$\text{variance } (\varepsilon) = 0.191 \qquad (20)$$

x_9 can now be generated from the following equation

$$x_9 = 0.636\ (x_5 \sim N(0,1)) + 0.636(x_6 \sim N(0,1)) + (\varepsilon \sim N(0, .0191)) \quad (21)$$

Thus, x_9 is generated as a linear function of x_5 and x_6 and the degree of intercorrelation between these variables can easily be controlled.

(4) Generating a dependent variable with a known functional relationship to a set of explanatory variables

The discussion has so far considered the generation of three orthogonal variables (x_1, x_2, x_3), two variables (x_7, x_8) related to a third variable (x_4) and one variable (x_9) related to two variables (x_5, x_6). We now require a method of generating y so that it is related to the explanatory variables in a known manner. For example, if we wish to set all the true regression coefficients to 0.5 and the intercept term to 0.0, the formula for generating y is

$$y = 0.0 + 0.5x_1 + 0.5x_2 + 0.5x_3 + 0.5x_4 + 0.5x_5 + 0.5x_6 + 0.5x_7 + 0.5x_8 + 0.5x_9 \quad (22)$$

Such a dependent variable is completely determined by the explanatory variables. To introduce unexplained stochastic variation a disturbance term is added to y. This disturbance term (to satisfy the assumptions of the regression model) is generated with a mean of zero to follow a normal distribution.

CHAPTER 2 : NOTES

1. The main diagonal elements of the inverted correlation matrix are called Variance Inflation Factors. With orthogonal data, the VIF is at a minimum and is equal to 1.0. Snee (1973) suggests a maximum of four or five for the VIF before multicollinearity becomes serious.

2. Good introductions to the technique of PCA are Mather (1976, Chapter 4) and Daultrey (1976).

3. Many analysts have applied PCA to data that, instead of representing the total field of interest, are a jumbled assortment of a number of measures representing one or two facets of interest (the data which are easily available). PCA, like all techniques, is highly dependent on the data that are being examined.

4. Although the technique will be called ridge regression throughout the present work, Deegan (1979) has called the method ordinary ridge regression to distinguish it from a more general formulation (Goldstein and Smith, 1974; Hemmerle, 1975).

5. The method can be traced back in a series of articles by Hoerl (1964, 1962) in the chemical engineering literature. Hoerl called the procedure 'ridge' regression because of its mathematical similarity to the technique of 'ridge' analysis. He had used this procedure in his earlier work to analyse second-order response surface equations; the graph of such equations has a characteristic 'ridge' shape.

6. An interesting example of applying ridge regression with cross-validation is given in Snee (1978).

7. The consideration of ridge regression by Mather (1976) is a textbook introduction to the technique, together

with a summary of an application of ridge regression to a geological problem.(based on Jones, 1972). The article by Moriarty (1973) is a straightforward transfer of the Hoerl and Kennard (1970b) expository paper to a geographical situation. In both Mather (1977) and Mather and Openshaw (1974) the technique only receives one sentence. Physical geographers may be interested in the examples given by Kitanidis and Bras (1979) and Shih and Shih (1979).

8. Thus, the fitting criterion is no longer minimising the sum of the squared errors and, indeed, as k increases this value will decrease. However, the more severe the multicollinearity the further it is possible to increase k and move from the OLS estimates without an appreciable increase in the sum of the squared errors.

9. The results given in this section have been produced by a ridge regression computer program which has been written by the author. Gunst (1979) indicates the basic program structure that is required for a number of different types of biased estimation. See also Boulding and Houston (1974) and Driscoll and Reynolds (1979).

10. Three other advantages are claimed for ridge regression. Firstly, according to Marquardt (1974) ridge regression is a reasonably 'robust' to 'outliers' (Chapter 4). Secondly, Darlington (1978) claims that ridge regression's performance in relation to OLS improves as the ratio of the number of explanatory variables to the number of observations decreases. Thirdly, it has been suggested by several authors that ridge regression is a valuable method of revealing the 'sensitivity' of the regression model to the data being analysed. Examining the latter suggestion in more detail, OLS estimation does not show the sensitivity of the method to the least-squares solution. A dramatic example of such sensitivity is given by

Hoerl (1964). He considered the following set of simultaneous equations (such equations are often used to calculate unstandardised regression coefficients):

$$192.09940 = 38.86243b_1 + 51.23934b_2 + 53.503383b_3 + 50.49442564b_4$$

$$267.02003 = 51.23934b_1 + 71.22350b_2 + 74.37005b_3 + 70.181714b_4$$

$$278.83370 = 53.5033b_1 + 74.37005b_2 + 77.66275b_3 + 73.29752b_4$$

$$263.15772 = 50.49425b_1 + 70.181714b_2 + 73.29752b_3 + 69.17882b_4$$

The solution to these equations is

$b_1 = -953, 521, 374$
$b_2 = 575, 900, 539$
$b_3 = 239, 462, 717$
$b_4 = -142, 030, 294$

However, if 0.00001 is added to the left-hand side of the final equation so that it becomes 263.15773, the solution becomes

$b_1 = 1$
$b_2 = 1$
$b_3 = 1$
$b_4 = 1$

Even with such dramatic changes as these, the least-squares method does not alert the user to the sensitivity of the solution, whereas the ridge trace of the above equations would show substantial fluctuations. As another illustration of this problem Beaton, Rubin and Barone (1976) changed a set of published values beyond the last available digit by adding uniform random numbers within the range -0.5 to 0.499. These tiny variations changed most unstandardised regression coefficients drastically (for example, from -232 to + 237) but, in contrast, the ridge estimator remained

stable with respect to such deliberate variations.

11. Carmer and Hsieh (1978) have reported simulation results for a number of biased estimators with one explanatory variable; they were generally poor.

12. With ridge regression an analyst has to provide less prior information about the parameters than with ordinary Bayesian regression; ridge analysis is therefore an example of 'empirical Bayes'. Oman (1978) has shown that a number of estimators (ridge, principal component, generalised inverse, Stein) can be viewed in a Bayesian light in that they all assume different 'prior' distributions for the true regression coefficient.

13. Incidentally, the Haitovsky procedure indicated that each of the 50 models were multicollinear and the Farrar-Glauber values correctly pinpointed those variables inducing the problem.

14. From theory we should expect the mean of the estimated OLS coefficients to equal the true coefficients, for OLS procedures remain unbiased with multicollinear data. However, this is an asymptotic result while the results presented here are based on only 50 replications.

15. A large number of other simulation experiments have been performed with varying degrees of multicollinearity (positive, negative and mixed) and a variety of true regression coefficients and again it has been found that ridge estimators can be good estimators and that the u statistic offers an excellent guide to their appropriate use.

CHAPTER 2 : BIBLIOGRAPHY

ALAM, K. and HAWKES, J.S. (1978): Estimation of regression coefficients. Scandinavian Journal of Statistics 5, 169-172.

ANDO, A. and KAUFMAN, G.M. (1966): Evaluation of an ad hoc procedure for estimating parameters of some linear models. The Review of Economics and Statistics 48, 334-340.

BACON, R.W. and HAUSMAN, J.A. (1974): The relationship between ridge regression and the minimum mean squared error estimator of Chipman. Oxford Bulletin of Economics and Statistics 36, 115-124.

BALDWIN, K.F. and HOERL, A.E. (1978): Bounds on minimum mean squared error in ridge regression. Communications in Statistics Series A 7, 1209-1218.

BANERJEE, K.S. and CARR, R.N. (1971): A comment on ridge regression. Biased estimation for non-orthogonal problems. Technometrics 13, 895-898.

BEATON, A.E., RUBIN, D.B. and BARONE, J.L. (1976): The acceptability of regression solutions: another look at computational accuracy. Journal of the American Statistical Association 71, 158-168.

BELONGIA, M. (1979): Application of ridge regression with verification of new procedures. Agricultural Economics Research 31, 36-39.

BIBBY, J. and TOUTENBURG, H. (1977): Prediction and improved estimation in linear models. New York, Wiley.

BIDOT, G.B. (1969): A principal components analysis of the determinants of local government fiscal patterns. The Review of Economics and Statistics 51, 176-188.

BLALOCK, H.M. (1963): Correlated independent variables: the problem of multicollinearity Social Forces 42, 233-237.

BOULDING, J.T. and HOUSTON, S.R. (1974): A FORTRAN computer program for computation of ridge regression coefficients Educational and Psychological Measurement 34, 151-152.

BROWN, W.G. (1973): Effect of omitting relevant variables versus use of ridge regression in economic research Special Report No. 394, Oregon Agricultural Experiment Station.

BROWN, W.G. and BEATTIE, B.R. (1975): Improving estimates of economic parameters by use of ridge regression with production function applications American Journal of Agricultural Economics 57, 21-32.

BROWNE, M.W. and ROCK, D.A. (1975): Choice of additive constants in ridge regression South African Statistical Journal 9, 83.

BROWNE, M.W. and ROCK, D.A. (1978): Choice of additive constants in ridge regression South African Statistical Journal 12, 57-74.

BROWNE, P.J. and PAYNE, C.D. (1975): Election night forecasting Journal of the Royal Statistical Society Series A 138, 463-498.

BUCKATZSCH, E.J. (1947): The influence of social conditions on mortality rates Population Studies 1, 229-248.

CARMER, S.G. and HSIEH, W.T. (1978): Simulation study of 5 biased estimators for straight-line regression Communications in Statistics Series B 7, 529-548.

CHATTERJEE, S. and PRICE, B. (1977): Regression analysis by example New York, Wiley.

CONNIFFE, D. and STONE, Joan (1973): A critical view of ridge regression The Statistician 22, 181-187.

CONNIFFE, D. and STONE, Joan (1975): A reply to Smith and Goldstein
The Statistician 24, 67-68.

CRANDALL, R. (1976): On the use of stepwise regression and other statistics to estimate the relative importance of variables
Journal of Leisure Research 8, 53-58.

DARLINGTON, R.B. (1978): Reduced-variance regression
Psychological Bulletin 85, 1238-1255.

DAULTREY, S. (1976): Principal components analysis
Concepts and techniques in modern geography, No. 8, Geo Abstracts, Norwich.

DAUM, H.F. (1976): Practical limits on n and p for the use of ridge regression algorithms
unpublished Ph.D. Thesis, University of Delaware.

DEEGAN, J. (1972): The effects of multicollinearity and specification error on models of political behaviour
unpublished Ph.D. Thesis, University of Michigan.

DEEGAN, J. (1975): The process of political development: an illustrative use of a strategy for regression in the presence of multicollinearity
Sociological Methods and Research 3, 384-415.

DEEGAN, J. (1978): Occurrence of standardised regression coefficients greater than one
Educational and Psychological Measurement 38, 873-881.

DEEGAN, J. (1979): Constructing statistical models of social processes
Quality and Quantity 13, 97-119.

DEMPSTER, A.P. (1973): Alternatives to least squares in multiple regression
in Kabe, D.G. and Gupta, R.G. (eds.)
Multivariate statistical inference
New York, North-Holland.

DEMPSTER, A.P., SCHATZOFF, M. and WERMUTH, N. (1977): A simulation study of alternatives to ordinary least squares
Journal of the American Statistical Association 72, 77-104.

DRISCOLL, M.F. and REYNOLDS D.A. (1979): A computer program for biased regression
The American Statistician 33, 160.

DWIVEDI, T.D. and SRIVASTAVA, V.K. (1978): Minimum mean squared error estimators in a regression model
Communications in Statistics Series A 7, 487-494.

EFRON, B. and MORRIS, C. (1975): Data analysis using Stein's estimator and its generalisations
Journal of the American Statistical Association 70, 311-319.

FAREBROTHER R.W. (1975): The minimum mean square error linear estimator and ridge regression
Technometrics 17, 127-128.

FAREBROTHER, R.W. (1976): Further results on the mean square error of ridge regression
Journal of the Royal Statistical Society Series B 38, 248-250.

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

FEIG, D.G. (1978): Ridge regression: when biased estimation is better
Social Science Quarterly 58, 708-716.

FELDSTEIN, M.S. (1973): Multicollinearity and the mean square error of alternative estimators
Econometrica 41, 337-346.

FOMBY, T.B. and HILL, R.C. (1979): Multicollinearity and the value of *a priori* information
Communications in Statistics Series A 8, 477-486.

GARDNER, M.J. (1973): Using the environment to explain and predict mortality
Journal of the Royal Statistical Society Series A 136, 421-440.

GILBERT, C.L. (1978): Diagnosis of multicollinearity Oxford Bulletin of Economics and Statistics 40, 87-91.

GIRT, J.L. (1972): Simple chronic bronchitis and urban ecological structure in McGlashan, N.D. (ed.) Medical geography: techniques and field studies Methven, London.

GIRT, J.L. (1974): A consideration of some relationships of environment to disease in Leeds and Newfoundland: two case studies in the ecology of human disease unpublished Ph.d. Thesis, University of Leeds.

GOLDSTEIN, M. and SMITH, A.F.M. (1974): Ridge type estimators for regression analysis Journal of the Royal Statistical Society Series B 36, 284-291.

GOLUB, G.H., HEATH, M. and WANBA, G. (1979): Generalised cross-validation as a method for choosing a good ridge parameter Technometrics 21, 215-223.

GOODE, B. (1975): Ridge regression and multiple regression-applications in professional football Bulletin of the Operations Research Society of America 23, 394.

GOODNIGHT, J. and WALLACE, T.D. (1972): Operational techniques and tables for making weak MSE tests for restrictions in regressions Econometrica 40, 699-709.

GORMAN, J.W. and TOMAN, R.J. (1966): Selection of variables for fitting equations to data Technometrics 8, 27-51.

GUILKEY, D.K. and MURPHY, J.L. (1975): Directed ridge regression techniques in cases of multicollinearity Journal of the American Statistical Society 70, 769-775.

GUNST, R.F. (1979): Approach to the programming of biased regression algorithms Communications in Statistics Series B 8, 151-159.

GUNST, R.F. and MASON, R.L. (1977): Biased estimation in regression: on evaluation using mean squared error. Journal of the American Statistical Association 72, 616-627.

GUNST, R.F. and WEBSTER, J.T. (1975): Regression analysis and problems of multicollinearity. Communications in Statistics 4, 277-292.

HAITOVSKY, Y. (1969): Multicollinearity in regression analysis: comment. The Review of Economics and Statistics 51, 486-489.

HAMAKER, H.C. (1962): On multiple regression analysis. Statistica Neerlandica 16, 31-56.

HART, J.T. (1970): The distribution of mortality from coronary heart disease in S. Wales. Journal of the Royal College of General Practitioners 19, 258-268.

HAUSER, D.P. (1974): Some problems in the use of stepwise regression analysis. Canadian Geographer 18, 148-158.

HAWKINS, D.M. (1975): Relations between ridge regression and eigenanalysis of the augmented correlation matrix. Technometrics 17, 477-480.

HEISE, D.R. (1975): Causal Analysis. Wiley, New York.

HEMMERLE, W.J. (1975): An explicit solution for generalised ridge regression. Technometrics 17, 309-314.

HEMMERLE, W.J. and BRANTLE, T.F. (1978): Explicit and constrained generalised ridge estimation. Technometrics 20, 109-120.

HOARE, A. (1977): The march of the macro-men. Progress in Human Geography 1, 512-517.

HOCKING, P.R. (1976): The analysis and selection of variables in linear regression. Biometrics 32, 1-49.

HOCKING, R.R., SPEED,F.M. and LYNN, M.J. (1976): A class of biased estimates in linear regression Technometrics 18, 425-437.

HOERL, A.E. (1962): Application of ridge analysis to regression problems Chemical Engineering Progress 58, 54-59.

HOERL, A.E. (1964): Ridge analysis Chemical Engineering Progress Symposium Series 60, 67-77.

HOERL, A.E. and KENNARD, R.W. (1970a): Ridge regression: biased estimation of non-orthogonal problems Technometrics 12, 55-67.

HOERL, A.E. and KENNARD, R.W. (1970b): Ridge regression: applications to non-orthogonal problems Technometrics 12, 69-82.

HOERL, A.E. and KENNARD, R.W. (1975): A note on the power generalisation of ridge regression Technometrics 17, 269.

HOERL, A.E. and KENNARD, R.W. (1976): Ridge regression: iterative estimation of the biasing parameter Communications in Statistics Series A 5, 77-78.

HOERL, A.E. and KENNARD, R.W. (1978): Forecasting wholesale prices of meat in U.K. Journal of Agricultural Economics 29, 333-334.

HOERL, A.E. KENNARD, R.W. and BALDWIN, K.F. (1975): Ridge regression: some simulations Communications in Statistics 4, 105-123.

HOLLAND, P.W. (1973): Weighted ridge regression: combining ridge and robust regression methods Working Paper No. 11, Cambridge, Mass., National Bureau of Economic Research

HSIANG, T.C. (1975): Bayesian view of ridge regression The Statistician 24, 267-268.

JAMES, W. and STEIN, C. (1961): Estimation with quadratic loss Proceedings Fourth Berkeley Symposium in Mathematical Statistics and Probability 1, 361-379.

JONES, T.A. (1972): Multiple regression with correlated independent variables
Mathematical Geology 4, 203-218.

KADIYALA, K. (1979): Operational ridge regression estimators under the prediction goal
Communications in Statistics Series A 8, 1377-1392.

KASARDA, J.D. and SHIH, W.P. (1977): Optimal bias in ridge regression approaches to multicollinearity
Sociological Methods and Research 5, 461-469.

KEEBLE, D.E. and HAUSER, D.P. (1972): Spatial analysis of manufacturing growth in outer South-East England 1960-1967 II: method and results
Regional Studies 6, 11-36.

KENNARD, R.W. (1976): Letters to the editor
Technometrics 18, 504-505.

KEREN, G. and NEWMAN, J.R. (1978): Additional considerations with regard to multiple regression and equal weighting
Organisational Behaviour and Human Performance 22, 143-164.

KIRBY, A.M. and TAYLOR, P.J. (1976): A geographical analysis of the voting pattern in the EEC referendum June 5, 1975.
Regional Studies 10, 183-191.

KITANIDIS, P.K. and BRAS, R.L. (1979): Collinearity and stability in the estimation of rainfall runoff model parameters
Journal of Hydrology 42, 91-108.

KLEIN, G.E. (1968): Selection regression programs
The Review of Economics and Statistics 50, 288-290.

KUMAR, T.K. (1975): Multicollinearity in regression analysis
The Review of Economics and Statistics 57, 365-366.

KUPPER, L.L. and MEYDRECH, E.F. (1973): A new approach to mean squared error estimation of response surfaces
Biometrika 60, 573-579.

KUPPER, L.L., STEWARD, J.R. and WILLIAMS, K.A. (1976): A note on controlling significance levels in stepwise regression
American Journal of Epidemiology 103, 13-15.

KVALSETH, T.O. (1979): Ridge regression models of urban crime
Regional Science and Urban Economics 9, 247-260.

LaMOTTE, L.R. (1978): Bayes linear estimators
Technometrics 20, 281-290.

LAWLESS, J.F. (1978): Ridge and related estimation procedures - theory and practice
Communications in Statistics Series A 7, 139-164.

LAWLESS, J.F. and WANG, P. (1976): A simulation study of ridge and other regression estimators
Communications in Statistics Series A 5, 307-323.

LINDLEY, D.V. and SMITH, A.F.M. (1972): Bayes estimates for the linear model
Journal of the Royal Statistical Society Series B 34, 1-18.

LOTT, W.F. (1973): The optimal set of principal component restrictions of a least-squares regression
Communications in Statistics 2, 449-464.

LOWERRE, J.M. (1974): On the mean squared error of parameter estimates for some biased estimators
Technometrics 16, 461-464.

McCALLUM, B.T. (1970): Artificial orthogonalization in regression analysis
The Review of Economics and Statistics 52, 110-113.

McCALLY, S.P. (1966): The governor and his legislative party
American Political Science Review 60, 923-942.

McDONALD, G.C. and GALARNEAU, D.I. (1975): A Monte-Carlo evaluation of some ridge-type estimators
Journal of the American Statistical Association 70, 407-416.

McDONALD, G.C. and SCHWING, R.C. (1973): Instabilities of regression estimates relating air pollution to mortality Technometrics 15, 463-481.

MADDALA, G.S. (1977): Econometrics McGraw-Hill, New York.

MAHAJAN, V., JAIN, A.K., and BERGIER, M. (1977): Parameter estimation in marketing models in presence of multicollinearity application of ridge regression Journal of Marketing Research 14, 586-591.

MARQUARDT, D.W. (1970): Generalised inverses, ridge regression, biased linear estimation and non-linear estimation Technometrics 12, 591-612.

MARQUARDT, D.W. (1974): Discussion of 'the fitting of power series, meaning polynomials illustrated on band-spectroscopic data' Technometrics 16, 189.

MARQUARDT, D.W. and SNEE, R.D. (1975): Ridge regression in practice The American Statistician 29, 3-20.

MASON, R. and BROWN, W.G. (1975): Multicollinearity problems and ridge regression in sociological problems Social Science Research 4, 135-149.

MASON, R.L., GUNST, R.F. and WEBSTER, J.T. (1975): Regression analysis and the problem of multicollinearity Communications in Statistics 4, 277-292.

MASSY, W.F. (1965): Principal components regression in exploratory statistical research Journal of the American Statistical Association 60, 234-256.

MATHER, P.M. (1976): Computational methods of multivariate analysis in physical geography Wiley, London.

MATHER, P.M. (1977): Clustered data point distributions in trend surface analysis
Geographical Analysis 9, 84-93.

MATHER, P.M. and OPENSHAW, S. (1974): Multivariate methods and geographical data
The Statistician 23, 283-308.

MAYER, R.P. and STOWE, R.A. (1969): Would you believe 99.9969 % explained?
Industrial and Engineering Chemistry 61, 42-47.

MAYER, L.S. and WILLKE, T.A. (1973): On biased estimation in linear models
Technometrics 15, 497-508.

MITTELHAMMER, R.C. and BARITELLE, J.L. (1977): Two strategies for choosing principal components in regression analysis
American Journal of Agricultural Economics 59, 336-343.

MORIARTY, B.M. (1973): Causal inference and the problem of non-orthogonal variables
Geographical Analysis 5, 55-61.

MORRILL, R.L. and WOHLENBURG, E.H. (1971): The geography of poverty in the United States
McGraw-Hill, New York.

NEWHOUSE, J.P. and OMAN, S.D. (1971): An evaluation of ridge estimators
Report No. R-176-PR, Santa Monica, Rand Corp.,
(see also Econometrics 39, 402-403).

NICHOLSON, R.J. and TOPHAM, N. (1973): Stepwise regression and principal components analysis in estimating a relationship in an econometric model
Manchester School 41, 187-205.

O'HAGAN, J. and McCABE, B. (1975): Tests for severity of multicollinearity in regression analysis
The Review of Economics and Statistics 57, 368-370.

OBENCHAIN, R.L. (1975): Ridge analysis following a preliminary test of a shrunken hypothesis
Technometrics 17, 431-441.

OBENCHAIN, R.L. (1977): Classical F-tests and confidence regions for ridge regression
Technometrics 19, 429-439.

OBENCHAIN, R.L. (1978): Good and optimal ridge estimator
Annals of Statistics 6, 1111-1121.

OLSSON, G. (1965): Distance and human interaction: a migration study
Geografiska Annaler 47, 3-43.

OMAN, S.D. (1978): Bayesian comparison of some estimators used in linear regression with multicollinear data
Communications in Statistics Series A 7, 517-534.

POPE, P.T. and WEBSTER, J.T. (1972): The use of an F-statistic in stepwise regression procedures
Technometrics 14, 327-340.

PRESTON, S.H. (1970): Older male mortality and cigarette smoking
Institute of International Studies Population Monograph No. 7, Berkeley.

PRICE, B. (1977): Ridge regression: application to non-experimental data
Psychological Bulletin 84, 759-766.

ROBSON, B.T. (1973): A view on the urban scene in Chisholm, M.D.I. and Rodgers, H. (eds.)
Studies in human geography
Heinemann, London.

ROCKWELL, R.C. (1975): Assessment of multicollinearity - the Haitovsky test of the determinant
Sociological Methods and Research 3, 308-320.

ROLPH, J.E. (1976): Choosing shrinkage estimates for regression problems
Communications in statistics Series A 5, 789-802.

RUTIHAUSER, H. (1968): Once again: the least square problem
Linear Algebra and its Applications 1, 479-488.

RYAN, J.G. and PERRIN, R.K. (1973): The estimation and use of a generalised response function for potatoes in the Sierra of Peru
Technical Bulletin No. 214 Agricultural Experiment Station, North Carolina State University, Raleigh.

SCHMIDT, P. and MULLER, E.N. (1978): The problem of multicollinearity in a multistage alienation model: a comparison of ordinary least squares, maximum likelihood and ridge estimators
Quality and Quantity 12, 267-297.

SCHWIRIAN, K.P. and LaGRECA, A.J. (1974): The effect of alternative age adjustment procedures on the analysis of urban mortality problems
Social Science Quarterly 55, 189-194.

SHESKIN, I.M. (1977): The reconstitution of regression coefficients in principal components regression analysis
Discussion paper No. 55, Department of Geography, Ohio State University.

SHIH, S.F. and SHIH, W.F.P. (1979): Application of ridge regression analysis to water-resource studies
Journal of Hydrology 40, 165-174.

SMITH, A.F.M. and GOLDSTEIN, M. (1975): Ridge regression: some comments on a paper of Conniffe and Stone
The Statistician 24, 61-66.

SNEE, R.D. (1973): Some aspects of non-orthogonal data analysis part I: developing prediction equations
Journal of Quality Technology 5, 67-79.

SNEE, R.D. (1978): Validation of regression models: methods and examples
Technometrics 19, 415-428.

STEIN, C.M. (1960): Multiple regression in Olkin, I (ed.)
Contributions to probability and statistics: essays in honour of Harold Hotelling
California, Standford University Press.

STOWE, R.A. and MAYER, R.P. (1969): Pitfalls of stepwise regression analysis Industrial and Engineering Chemistry 61, 9-16.

SWAMY, P.A.V.B. (1973): Criteria, constraints and multicollinearity in random coefficient regression models Annals of Economic and Social Measurement 2, 429-450.

SWAMY, P.A.V.B. and MEHTA, J.S. (1978): Two methods of evaluating Hoerl and Kennards ridge regression Communications in Statistics Series A 7, 1133-1155.

SWINDEL, B.F. (1974): Instability of regression coefficients illustrated The American Statistician 28, 63-65.

THEOBALD, C.M. (1974): Generalisations of mean squared error applied to ridge regression Journal of the Royal Statistical Society Series B 36, 103-106.

TOBLER, W.R. (1970): A computer movie simulating urban growth in the Detroit Region Economic Geography 46, 234-240.

VINOD, H.D. (1976a): Letter to the Editor Technometrics 18, 504.

VINOD, H.D. (1976b): Simulation and extension of a minimum MSE estimator in comparison with Stein's Technometrics 18, 491-496.

VINOD, H.D. (1976c): Application of new ridge regression methods to a study of Bell system scale economics Journal of the American Statistical Association 71, 835-841.

VINOD, H.D. (1978a): Equivariance of ridge estimators through standardisation: a note Communications in Statistics Series A 7, 1157-1161 and 1169-1170.

VINOD, H.D. (1978b): Ridge regression and related techniques: a survey
The Review of Economics and Statistics 60, 121-131.

VINOD, H.D. (1978c): Ridge estimator whose MSE dominates OLS
International Economic Review 19, 727-737.

WAINER, H. (1976): Robust statistics: a survey and some prescriptions
Journal of Educational Statistics 1, 285-312.

WAINER, H. (1978): Sensitivity of regression
Psychological Bulletin 85, 267-273.

WALLACE, T.D. (1972): Weaker criteria and tests for linear restrictions in regression
Econometrica 40, 689-698.

WALLIS, J.R. (1965): Multivariate statistical methods in hydrology - a comparison using data of known functional relationships
Water Resources Research 1, 447-461.

WEBSTER, J.T., GUNST, R.F. and MASON, R.L. (1974): Latent root regression analysis
Technometrics 16, 513-522.

WERMUTH, N. (1974): Observations on ridge regression (In German)
Jahrbucher Fur Nationalokonomie und Statistik 189, 300-307.

WEST, R.R. and LOWE, C.R. (1976): Mortality from ischaemic heart disease: inter-town variation and its association with climate in England and Wales
International Journal of Epidemiology 5, 195-201.

WICHERN, D.W. and CHURCHILL, G.A. (1978): Comparison of ridge estimators
Technometrics 20, 301-311.

WICHERS, C.R. (1978): The detection of multicollinearity: a comment
The Review of Economics and Statistics 57, 366-368.

WILKINSON, R.K. and ARCHER, C.A. (1973): Measuring the determinants of relative house prices
Environment and Planning 5, 357-367.

WILLIAN, A.R. and WATTS, D.G. (1978): Meaningful multicollinearity measures
Technometrics 20, 407-412.

WILLIAMS, K. (1971): Do you sincerely want to be a factor analyst?
Area 3, 228-230.

WINER, B.J. (1978): Statistics and data analysis: trading bias for reduced mean squared error
Annual Review of Psychology 29, 647-681.

C H A P T E R 3

SPATIAL AUTOCORRELATION:

AN EXPLORATORY APPROACH

'If geographers continue to apply the usual forms of many of the basic statistical models to spatially autocorrelated data, then a very severe risk is run of reaching misleading conclusions. In addition, it means that the substantive results reported in studies to date which have not taken this problem into account should be interpreted with great caution'

Haggett, Cliff and Frey (1977, 336).

INTRODUCTION

Take a coin, toss it in the air. With an unbiased coin the probability of 'heads' is 0.5. Toss it again, the probability of heads is again 0.5. Each throw of the coin is independent of all others. This independence of observation has become the fundamental basis of statistics. But what happens when geographical observations, that is data collected over space, are used? Tobler's (1970, 236) first law of geography states that:

> 'everything is related to everything else
> but near things are more related than
> distant things'.

Most maps that geographers study display not independent values but spatial pattern. However, practitioners of so-called 'classical' statistics choose to ignore geographical reality and analyse their data as a set of independent values; such practitioners are in danger of making gross inferential errors.

This chapter considers the problem of using classical statistics for the analysis of spatially-autocorrelated data. Firstly, autocorrelation is defined and then confirmatory methods of detecting pattern are outlined. This is followed by a discussion of the effects of spatial autocorrelation on statistical methods; the t and F test and the correlation coefficient are briefly considered, while there is a more extended discussion of the effects on regression analysis. Different approaches to solving the problem are then considered and the chapter concludes with a number of empirical examples.

Before attention turns to the question of defining spatial autocorrelation one point must be made regarding the approach adopted below. Previous accounts of the problem of spatial autocorrelation have been necessarily of a highly technical nature, this chapter however, aims to present a

non-technical account in an attempt to dispel the mystique surrounding the subject. In order to achieve this aim there are no formal proofs and the reasoning is intuitive rather than mathematical.

DEFINING AND DETECTING SPATIAL AUTOCORRELATION

Autocorrelation is a special case of correlation - it literally means 'self-correlation' and can be regarded as meaning 'internal correlation'. Whereas correlation refers to the relationship between two different variables, autocorrelation refers to the relationship between the values of one variable. In general, if high values of a variable in one area are associated with high values of that variable in a neighbouring area, the data set exhibits positive spatial autocorrelation. Conversely, when high and low values alternate, the data exhibit negative autocorrelation. Of course, the geographically more common case is that of positive spatial autocorrelation.

While a visual impression of the presence or absence of autocorrelation may be obtained from a choropleth map, geographers and others have attempted to develop a confirmatory approach to the detection of spatial patterning. The null hypothesis of no patterning is first stated and then a test statistic is calculated for the sample data. The null hypothesis is either accepted or rejected on the basis of comparing this calculated statistic with an expected value of the statistic (at some pre-determined significance level) for the case when no pattern truly exists in the population data.

The test statistic (I) has the general form:

$$I = \frac{\Sigma (2) \quad w_{ij} \quad f (x_i \; x_j)}{\Sigma f (x_i)} \qquad (1)$$

The numerator is a measure of covariance among the observations and the denominator is a measure of variance. The w_{ij} values

are termed weights and they represent a measure of the 'ties' between areas. The choice of an appropriate set of weights is of crucial importance and has received considerable attention in the literature.

The early work on the detection of spatial autocorrelation (Moran, 1950; Geary, 1954) used binary weights. When there are N areas there are N by N weights in a weights matrix; each row and column of the matrix consists of ones and zeros indicating that an area does (1) or does not (0) have a common boundary (join) with another area. Such weights have two main drawbacks (Dacey, 1965).

1. Topological invariance - it is possible to have a variety of different shaped lattices yet have the same binary weights and therefore the same value of the test statistic.

2. Inflexibility - if an analyst attempts to detect spatial pattern amongst the outbreaks of an infectious disease he might wish to use weights based on flows between areas. In South Wales, for example, he may wish to construct weights so that they give emphasis to places arranged up and down the valleys as compared to areas separated by an inter-valley ridge. Binary contiguous weights do not allow such flexibility.

Cliff and Ord (1969), however, have overcome these problems by developing a generalised test in which the analyst can choose any weighting scheme he deems appropriate[1]. (A number of different schemes are illustrated in Table 3.1.) This in turn creates a new problem: what weighting matrix should one employ? The advice given by Cliff and Ord (1973, 19) is to specify the weights in accordance with the kind of spatial pattern that one is trying to detect. Frequently,

TABLE 3.1

Some Possible Weighting Schemes

Binary weights - for example

(a) a 1 if two areas have a common boundary, all other weights set to 0.

(b) a 1 if nearest neighbour in terms of air-line distance, all other weights set to 0.

General weights - for example

(a) the reciprocal of the distance between a place and its first nearest neighbour.

(b) the reciprocal of the distance between the areas multiplied by the proportion of the perimeter of each area which is in contact with that area.

however, the analyst does not have sufficient knowledge of a particular spatial pattern and its generating spatial process to choose unequivocally between different weighting schemes. Unfortunately, different weighting matrices can give completely different results. For example, consider a chess board. If one uses a rook's weighting matrix (horizontal and vertical joins) negative autocorrelation (black 'values' alternating with white) will be detected. Using a bishop's case weighting scheme (diagonal joins) positive spatial autocorrelation will be detected (black connected to black; white connected to white). The queen's case (joins in all directions), however, will show no spatial autocorrelation; the map will be patternless with equal numbers of black and white joins.

Another problem in the choice of a weighting matrix is the conflict between the 'power' of the tests and their ability to detect different types of spatial pattern. For example, if the analyst attempts to achieve complete coverage by setting all the weights to be equal, the value of the Cliff-Ord statistic will be $-(N-1)^{-1}$ irrespective of the data values. In general, as the connectivity in the weights matrix increases so the power of the test to detect pattern declines. The problem of choosing an appropriate weights matrix will be re-considered when the discussion turns to the detection of spatial autocorrelation among regression residuals.

To illustrate the application of the Cliff-Ord test, consider Figure 3.1 which shows the geographical variations in coronary heart disease in South Wales. Because the mapped values are standardised mortality ratios (SMR s) the effects of different age structures in different areas are removed from the data. Thus, an SMR of 100 represents the England and Wales average death rate, while an SMR of 130 represents a mortality experience of 30 per cent above the national average. Visually, the map shows a high degree of spatial

FIGURE 3.1 HEART DISEASE MORTALITY, MALES, S. WALES

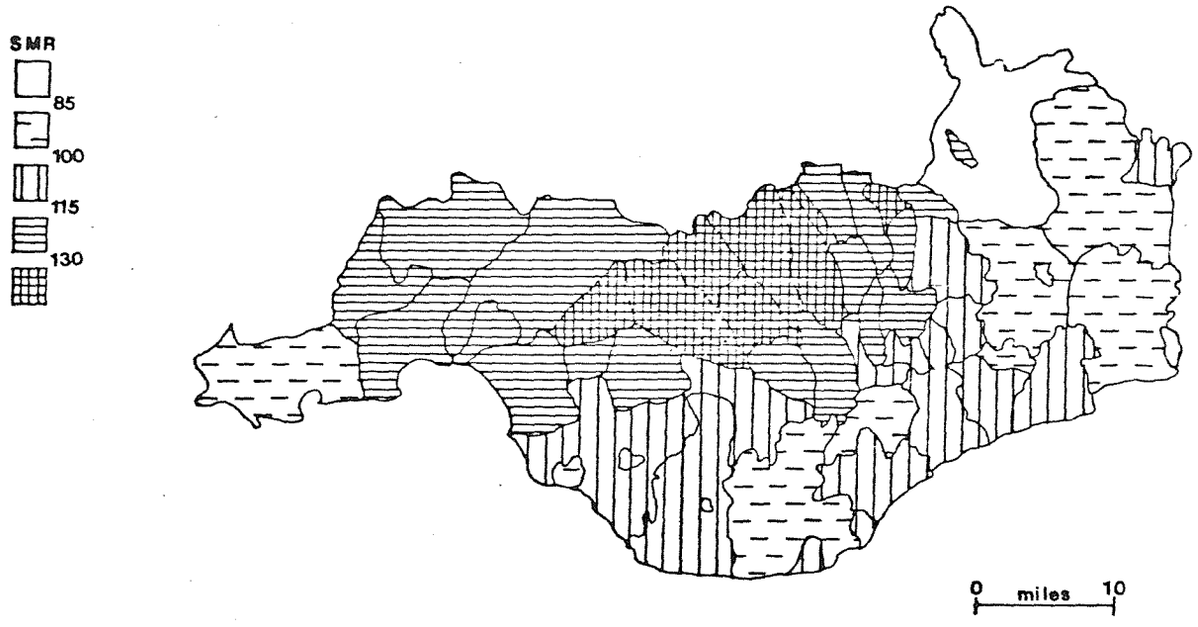
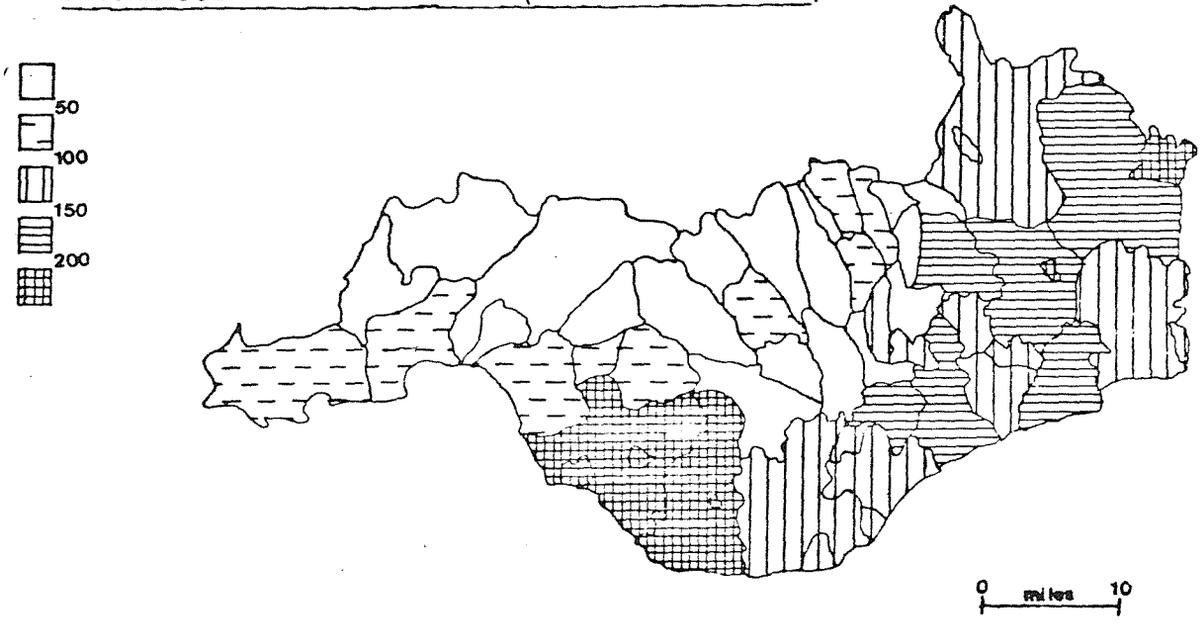


FIGURE 3.2 WATER HARDNESS (PARTS PER MILLION)



Source: HART (1970)

dependence; parts of the coalfield area are 45 per cent above the England and Wales average and many rural areas are 20 per cent below it. Applying the Cliff-Ord test with binary contiguous weights results in the rejection of the null hypothesis at the 95 per cent probability level. (The test statistic can be converted to a standard normal deviate and if this calculated value is greater than 1.96 or less than -1.96 the null hypothesis can be rejected at the 95 per cent probability level; the observed statistic for the South Wales mortality data is 6.03.)[2] Put simply, coronary heart disease has a distinct geographical pattern in South Wales and it is unlikely that such a distribution is the outcome of random processes.

THE EFFECTS OF SPATIAL AUTOCORRELATION ON CLASSICAL STATISTICS

Since methods like t tests, correlation, principal components and factor analysis are based on an assumption that is probably wrong with regard to geographical data, the following question must be posed: to what extent will the breaking of this assumption lead to inferential error? The following section attempts to answer this question by reviewing the effects of spatial autocorrelation on classical statistics. It should be noted, however, that the review will be selective because, as yet, we do not know the full consequences of non-independence for a large number of statistical procedures.

The underlying difficulty in using classical statistics with spatially patterned data is that autocorrelated observations do not contain so much 'information' as an equivalent number of independent observations. For example, annual rainfall in England and Wales is known to be highly positively spatially autocorrelated and the doubling of the present number of rain gauges in this country will not lead to a doubling of information; rain gauges that are located near

each other can be expected to give similar values. However, the critical value of a test statistic associated with a certain 'degrees of freedom' assumes that each observation is independent of every other observation. The effect of positive spatial autocorrelation, is therefore to reduce the effective degrees of freedom. While analytically it is very difficult to derive appropriate critical values of a test statistic for a given amount of spatial autocorrelation this can be readily achieved given a computer simulation experiment. Cliff and Ord (1975a) have performed such an experiment to demonstrate the effect of spatial autocorrelation on Students t.

Table 3.2 illustrates their results for two significance levels (α = 0.05 and 0.01) and a regular 7 by 7 square lattice.

TABLE 3.2

Effects of autocorrelation on empirical percentage points for Student's t.

α *	tabulated t	empirical percentage points for t level of autocorrelation					
		positive		negative		mixed	
		moderate	high	moderate	high	moderate	high
0.05	1.7	3.0	11.0	1.0	0.5	2.3	7.1
0.01	2.4	4.1	15.0	1.4	0.7	3.1	10.0

* one-tailed

Suppose an analyst wishes to calculate a t statistic for two variables for a 7 by 7 lattice to test whether or not he can reject the null hypothesis that the two variables have equal means at the 0.05 level (one-tailed). If the data are

not spatially autocorrelated the critical value of the test statistic (the percentage point) is given by published t tables and is 1.7. If the calculated statistic exceeds this value the analyst can reject the null hypothesis at the appropriate significance level. However, as the simulation experiment shows, if both variables exhibit 'moderate' positive spatial autocorrelation, the calculated statistic should be compared not with 1.7 but with the value 3.0. With such positive spatial autocorrelation the use of the published tabulated values will lead to an over-statement of the calculated statistic. When the autocorrelation for both variables is negative an understatement of the true significance of the results will occur. If both positive and negative spatial autocorrelation are present their effect is not self-cancelling for the use of the tabulated t values will lead to an overstatement of the significance of a particular result. For all three cases (positive, negative and mixed autocorrelation) the problem becomes more acute as the variables become more spatially autocorrelated. Fortunately, Cliff and Ord have been able to propose a modification of the t test that can accommodate the deleterious effects of spatial autocorrelation, provided that there are more than 25 observations. It must be acknowledged, however, that this modified test can necessitate considerable computation.

Other work has shown similar distortion of classical statistical tests. For the χ^2 test, Cliff, Martin and Ord (1975) have shown that the use of published tables will lead to the overstatement of the significance of the calculated result and that in any empirical work this 'overstatement frequently reaches gross proportions' (p 123). Hepple (1974b), on the basis of simulation results, has shown that Pearson's correlation coefficient is affected by spatial autocorrelation in the variables. Two variables can each exhibit a high degree of spatial autocorrelation, causing the

Pearson correlation coefficient to show a strong, significant relationship when, in fact, no statistically significant relationship actually exists. Such results have repurcussions for the inferential use of multivariate techniques such as factor analysis, which are based on correlation coefficients. Cliff and Ord (1975b) and Lebart (1969) have shown that, if all the variables are equi-autocorrelated, the same results will be obtained as if the independence assumption had been met. However, in practice, we can expect different variables to have different levels of autocorrelation and as Hepple (1974a, 116) has written:

> 'geographers should be more sceptical of the meaning of factor loadings and scores in human geography'.

The general impression from such work is unambiguous: the use of classical confirmatory statistics on geographical observations leads to misleading inferences.

REGRESSION AND SPATIAL AUTOCORRELATION

The effects

The majority of research on the effects of spatial autocorrelation has been concentrated on regression analysis. One of the crucial assumptions of the regression model is that the disturbance terms are independent of each other. The regression model assumes that the disturbance term for any one area is uninfluenced by the values of the disturbance term for other areas. This assumption of no autocorrelation is an important one, for if it is not satisfied the consequences for model-building and hypothesis testing are serious.

The effects of autocorrelation have been investigated analytically (Johnston, 1972) and by Monte-Carlo simulation (Hepple, 1974b; Martin, 1974). Positive autocorrelation of the disturbance term has the following effects when a model is estimated by ordinary least squares.

(1) The estimated regression coefficients will have a large variance.

(2) The variance of the disturbance term will be underestimated and, as this value is used in the calculation of the standard regression tests, this will lead to

(3) the overstatement of the significance of the model in terms of the t and F tests and

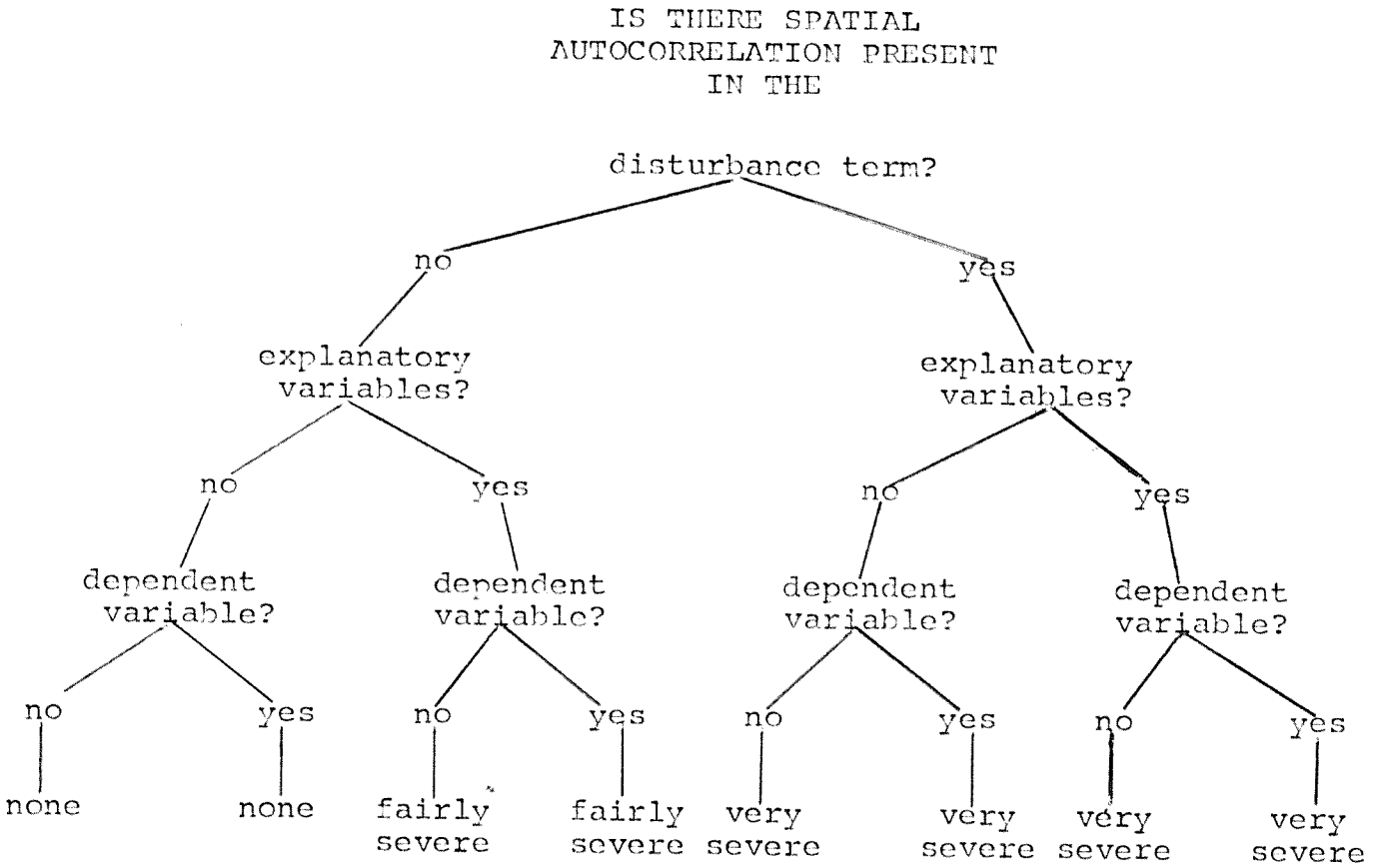
(4) the inflation of the coefficient of determination (R^2).

The seriousness of these problems depends on the extent and pattern of the spatial autocorrelation, for the degree of autocorrelation in the explanatory variables as well as the disturbance term must be taken into account. Figure 3.3 represents an attempt to order the importance of the effects in terms of which component of the regression model is spatially autocorrelated. As Johnston (1972) has pointed out, even if the disturbance term is spatially autocorrelated, provided that the explanatory variables are not, the variance of the estimators will not be seriously increased. Unfortunately, of course, in most applications the explanatory variables do exhibit spatial pattern.

Hepple (1974b) has attempted to quantify the effects of spatial autocorrelation on the formal inferential processes that are commonly applied to regression models. He found that, on a regular 25 cell lattice, the t and F values may be exaggerated 3 to 4 times with 'moderate' autocorrelation in the disturbance term. As the level of positive autocorrelation increases, the more probable it is that the researcher using confirmatory statistics will accept a relationship as being significant when it is not. Thus, the procedure for inferring from a sample to a population breaks down when the disturbance term is autocorrelated. The problem also has an insidious nature, for autocorrelated models tend to give results that 'look good' in terms of high

FIGURE 3.3

Effects of spatial autocorrelation on regression estimation



Source: Based on Griffith (1976)

t and F and R^2 values, when in fact the model is incorrect.

Detection: confirmatory and exploratory approaches

Given that spatial autocorrelation has such deleterious effects on classical statistics, it is obviously crucial that an analyst should be able to detect autocorrelation if it is present in the disturbance term. In any practical application the population disturbance term of the regression model is unknown, and any test for autocorrelation must therefore be based on the regression residuals. Unfortunately, as Theil (1965) has shown, the residuals are autocorrelated even when the population disturbance term is not (Chapter 4). To overcome this problem a test statistic is needed with a distribution theory that incorporates the pattern and levels of autocorrelation in the explanatory variables and the disturbance term. If the Cliff and Ord test (1969) illustrated previously is used for this purpose, inferential error may occur. Cliff and Ord (1972), however, have proposed a suitable test statistic.[3] The statistic is again based on a set of general weights with all the attendant flexibility and problems this entails. This test will be illustrated later.

The approach to detecting spatial autocorrelation advocated by Cliff and Ord is firmly rooted in the tradition of confirmatory analysis and consequently suffers from the drawbacks of that approach. Firstly, the Cliff-Ord statistic, unlike for example Pearson's correlation coefficient, is not bound to lie between -1 and +1 and therefore, the severity of the spatial autocorrelation cannot be assessed; an analyst can only reject or accept the null hypothesis of no autocorrelation at a particular significance level. Moreover, the value of the statistic has not been linked to the effects of spatial autocorrelation on estimation. As two geographers committed to exploratory statistics have written:

> 'considerable technical achievements have been made in the inferential theory of spatial series on irregular lattices, but much of this work has tackled the testing of the null hypothesis of zero autocorrelation, which is quite often geographically absurd' (Cox and Anderson, 1978, 31).

The second problem is the choice of weights and as Haining (1978, 204) has written:

> 'the system of using weights ... may lead to ... estimates that reflect as much our choice of weights as any intrinsic property of the underlying spatial dependence'.

If a pragmatic researcher attempts to use a variety of weights (beginning with maximum 'coverage' weights and then trying more 'powerful' weights with less coverage) to detect spatially autocorrelated residuals, he is bending the rules of the Cliff-Ord test, for no allowance for such experimentation has been made in the development of the distrubution theory of the test. Finally, and perhaps more fundamentally, it is unwise for a researcher to rely too heavily on a single summary measure to indicate the full nature of a data set.

As discussed in Chapter 1, graphs can often convey greater detail than a summary statistic and, moreover, they frequently give clues as to what a researcher should do next. This is undoubtedly the case for the exploratory approach to spatial autocorrelation, where maps of residuals are used both to indicate the presence of the problem and to suggest how the problem may be overcome. At a joint conference between statisticians and geographers, Wrigley (in a paper subsequently published in 1977) suggested that a Cliff-Ord procedure could be used to detect autocorrelation in the residuals from the particular models he was investigating. The response to this suggestion has been summarised by Unwin (1976, 120):

> 'we were often in the somewhat unexpected position of advocating a statistical approach to spatial problems whilst the statisticians urged us to draw maps'.

The use of residual maps to detect spatial autocorrelation has been considered in some detail by Thomas (1960). Although he did not point out that spatially patterned residuals would impair OLS estimation he did suggest that a residual map could indicate which variables had been omitted from a model and which areas required field investigation. Furthermore, Thomas had foreseen the division between techniques for confirmatory and exploratory analysis. For him a residual map has

> 'characteristics that make it unsuitable for hypothesis testing' (p 8-9)

but it

> 'provides the geographer with a research system in which the spatial character of the data is systematically recreated and employed in the formulation of geographic hypotheses' (p 10).

The use of residual maps to indicate spatial autocorrelation will be illustrated later.

Solving the problem

In addition to the inclusion of previously omitted variables into a reformulated model, a number of other procedures have been suggested to overcome the problem of spatially autocorrelated residuals. One such solution is to remove the dependence from the original observations by 'spatial differencing'.[4] For spatial data, differencing is achieved by subtracting the value of a particular area from the local average which has, in turn, been calculated from the observations for neighbouring areas. The procedure has been used by Lebart (1966, 1969) and Martin (1974). This is, of course, a very naive approach to the problem and has received criticism both from serial and spatial analysts. In the time-series literature, where the problem is one of temporal autocorrelation, Kadiyala (1968, 93) has stated that the differencing procedure

'does not necessarily improve the estimates, and in fact it is shown that there exist cases where it leads to less efficient estimates ... The result obtained seriously questions the usual procedure of taking the first differences rather than the original variables'.

For spatial data the difference transformation at once seeks to include and ignore the effects of spatial dependence and as Gould (1970, 444) has written: 'clearing up' so as to remove spatial autocorrelation 'represents a throwing out of the baby with the bathwater'.

Spatial differencing can also be faulted on the ground that it is a 'blanket' approach which ignores the fact that disturbance terms may appear to be autocorrelated for three different reasons. Firstly, the disturbance term may appear to be autocorrelated due to the presence of a non-linear relationship between the dependent and explanatory variables. Secondly, the omission of one or more explanatory variables can lead to apparent autocorrelation. Thirdly, the disturbance term may be autocorrelated because the model requires an 'autoregressive' structure; a spatial component needs to be directly incorporated into the regression equation. Taking each of these causes of apparent autocorrelation in turn, it is possible to propose different methods of overcoming the problem.

In order to demonstrate that a mis-specification of the form of the relationship between dependent and explanatory variables can result in an apparently autocorrelated disturbance term, let us suppose that the true relationship between y and x_1 is quadratic. That is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \varepsilon \qquad (2)$$

However, the following incorrect model has been estimated:

$$y = \hat{b}_0 + \hat{b}_1 x_1 + \hat{e} \qquad (3)$$

If there is any spatial pattern in the x_1 values then there will be spatial pattern in the residuals of the estimated incorrect model, for these residuals will contain an element of x_1^2 . Even if the disturbance term of the true model is not spatially autocorrelated the residuals of the estimated model (providing the explanatory variable has a spatial pattern) will appear to be autocorrelated. The problem of non-linearity can be tackled in a number of ways and the problem will be considered in more detail in Chapter 4.

The disturbance term in a regression model represents both true stochastic variation and the total influence of all the explanatory variables that have been omitted from the model. If an important variable has been omitted and this variable has a spatial pattern, the residuals of the estimated mis-specified model will appear to be spatially autocorrelated. As already stated, Thomas's (1960) method of mapping the residuals may suggest previously omitted explanatory variables that need to be included in a reformulated model.

Cliff and Ord (1973) found in one of their studies that the spatial autocorrelation in the regression residuals was very pervasive and apparently could not be removed by the addition of other independent variables; nor did it appear to be a result of a non-linearity. They concluded that to model this particular data set a spatially auto-regressive model was required. There are several types of 'autoregression' that can be incorporated into a model. For example, the value of a variable for a particular area may be dependent not only on the values of the explanatory variables for that area but also on its own value in neighbouring areas. Similarly, a value for an area could be dependent on a set of explanatory values for that area and also on values of the explanatory variables for the surrounding areas. Also, and more contentiously, the value of the

disturbance term in a particular area may be dependent on the values of the disturbance term in neighbouring areas, a condition that may be termed 'true' autocorrelation of the disturbance term.

The critical point about such autoregressive models is that they cannot be estimated by OLS regression. Instead a maximum likelihood method must be used. This method of estimation for spatial data has been considered by a number of workers - Cliff and Ord (1973), Hordijk (1974), Miron (1972) and Ord (1975). Despite the possibility of calculating a model with an autoregressive disturbance structure it must be noted that this form of estimation has been heavily attacked by Geary (1963). He argues that autocorrelation can only occur when the model has been mis-specified, when a non-linear form, an explanatory variable or a lagged relationship (the dependent variable being related to explanatory variables in neighbouring areas) has been ignored:

> 'In a word, the variable exists and the obvious course is to go and look for it instead of postulating the property of autocorrelation in the residuals of which no practical good can come' (p 178).

Certainly, it is difficult to conceive a true stochastic disturbance term that is dependent in space. In terms of any practical analysis, the researcher must attempt to elucidate the underlying reason for the apparent autocorrelation and take appropriate action to remedy the specific cause of the problem.

Empirical examples

The discussion so far has concentrated on the theoretical aspects of the effect of spatial autocorrelation on regression; the following section attempts to present some under-pinning of the argument by empirical examples.

Hart (1970), in his South Wales study, attempted to assess the influence of water hardness on coronary heart disease

mortality. Figure 3.1 shows the pattern of heart disease mortality after allowing for varying age distributions; the map exhibits a high degree of patterning with the coalfield areas having a higher death rate than the rural areas. Figure 3.2 is a map of water hardness for South Wales; the water can be seen to be relatively hard outside the coalfield area. Using Hart's data a regression model was estimated and it was found that the relationship between heart disease mortality and water hardness was a negative one. The usual confirmatory methods of assessing the model (the t, F values) indicated that the model was 'highly significant' and accounted for some 37 per cent of the variation in heart disease mortality. But, are the residuals from this model spatially autocorrelated? A map of the regression residuals from this model (Figure 3.4) gives an immediate impression of spatial patterning; the model has underestimated deaths in the urban coalfield areas and has overestimated them in the rural areas. Using binary continguous weights, a Cliff-Ord test, rejects the null hypothesis of no spatial autocorrelation at the 0.01 significance level. The residual map, moreover, suggests that some variables have been omitted from the model. Does, for example, working in coalmines have an effect on age of death, or is air pollution lowering the life expectancy of the urban dweller?

In fact Roberts and Lloyd (1972) have added an explanatory variable to Hart's model - rainfall.[5] The fitted model, if evaluated by confirmatory statistics, is found to be highly significant with two explanatory variables (water hardness and rainfall) explaining nearly 70 per cent of the geographical variation in heart-disease mortality. Yet, once again, the mapped residuals (Figure 3.5) indicate spatial autocorrelation. In this case however, the death rate in the urban areas has been overestimated, while the rate in the rural districts has been underestimated. If a Cliff-Ord test is applied to the residuals using binary

FIGURE 3.4 RESIDUALS FROM HART'S MODEL

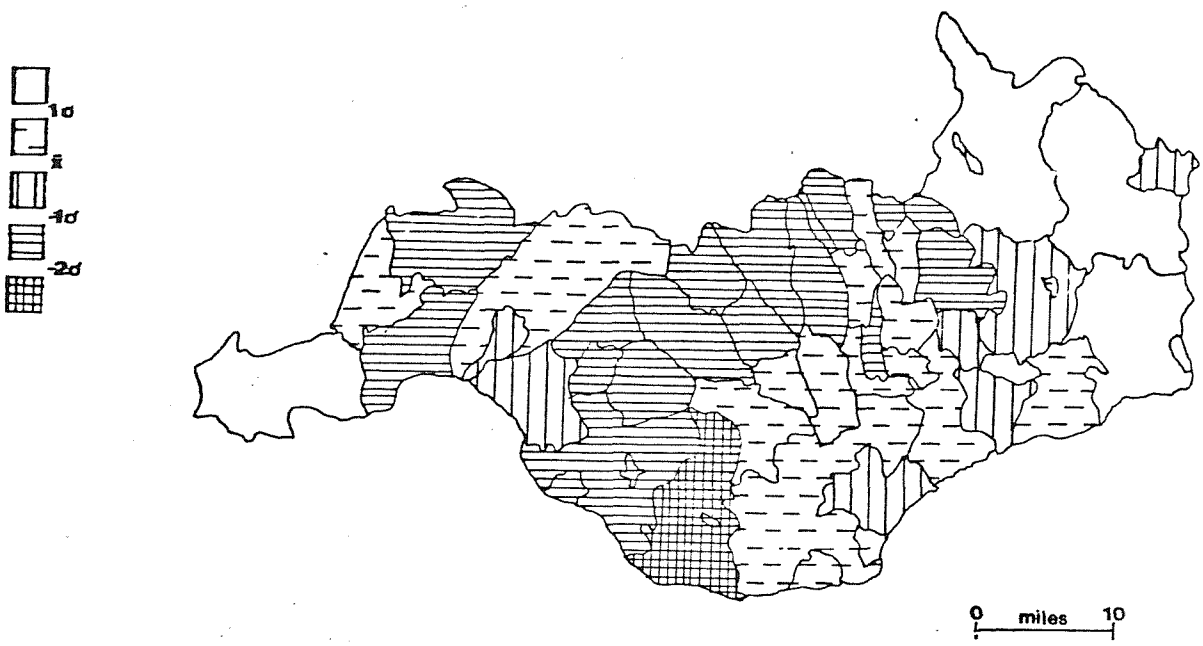
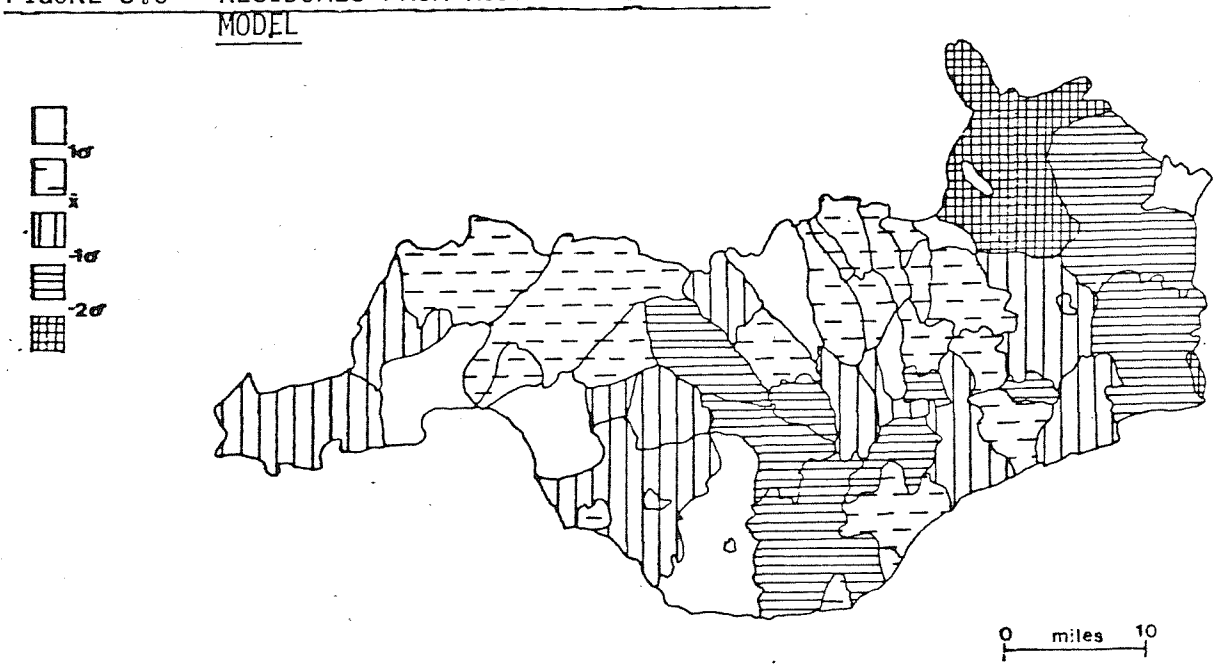


FIGURE 3.5 RESIDUALS FROM ROBERTS AND LLOYD'S MODEL



contiguous weights, the null hypothesis of no autocorrelation cannot be rejected at the 0.05 significance level. However, if we bend the rules and specify a weighting system suggested by the map pattern and assign urban-urban contiguous joins a value of one and all other joins are set to zero, the null hypothesis of no autocorrelation can be rejected at the 0.01 significance level. If we had blindly believed the original test based on binary weights we would have commited an interpretative error while the second test merely confirmed the visual impression that had been obtained from the residual map. In terms of the Roberts and Lloyd analysis, because the residuals are spatially patterned we can place no confidence in inferences from their model; we can, however, suggest that a reformulated model should be estimated that attempts to account for the differing mortality experiences of urban and rural areas.

Another investigation worthy of reconsideration is that of West and Lowe (1976), who examined the variation in ischaemic heart disease mortality rates amongst 115 county and London boroughs of England and Wales. Temperature, rainfall and a socio-economic index were used as explanatory variables and they concluded that

> 'inter-town variation in ischaemic heart disease in England and Wales may be largely due to body cooling by adverse climates' (p 200)

and suggested that

> 'there is a need to change public attitudes to wearing warm underclothing' (p 200).

Again their fitted model is a highly significant one but a map of the residuals gives an immediate impression of spatial dependence with the West Midlands conurbation being underestimated, the West Yorkshire conurbation and western London boroughs being overestimated (Figure 3.6). If a Cliff-Ord test with binary contiguous weights is carried out, the null hypothesis of no spatial autocorrelation is rejected

FIGURE 3.6 RESIDUALS FROM WEST AND LOWE'S MODEL

at the 0.01 significance level. The estimated model with such spatially patterned residuals is undoubtedly incorrect, and considerable doubt is cast on West and Lowe's statement that 'the high association between IHD mortality and temperature is a causal one' (p 200).[6]

CONCLUSIONS

The concept of spatial dependence is fundamental to geography, for if a particular spatial series has no spatial autocorrelation then it has no real spatial pattern and spatial structure; such a series could be produced by a set of random numbers. However, the geographical reality of spatial pattern is at odds with the independence assumption of classical tests and it has been shown that if classical statistics is used in inappropriate situations inferential error may result. Although such major analytical problems have been recognised in geographical analysis for at least the last decade (Cliff and Ord, 1969), examples of work continue to be published that fail to consider this crucial assumption. Indeed, the study that recognises the problem is the exception rather than the rule.[7]

With respect to regression analysis, this chapter has attempted to summarise the strong evidence that now exists for believing that serious problems of interpretation arise when regression residuals are spatially patterned. It has been suggested that the residual map is a suitable exploratory method of detecting spatial autocorrelation. If, for example, West and Lowe had used the quick and effective method of residual mapping, the shortcomings of their proposed model would have been clearly exposed. But detection of autocorrelation is not an end in itself, corrective action must be taken. Indeed, for Geary (1954, 135),

> 'If the dependent variables are found to be spatially contiguous, the fact that the remainders, after the removal of the effect of the independent variables are found to lack contiguity constituents, constitutes a *prima facie* case for regarding the independent variables as completely explaining the dependent variable'.[8]

A geographically appropriate and successful model is, therefore, one that has been reformulated and re-estimated until the map of residuals shows no systematic pattern. Further examples of this exploratory approach to spatial autocorrelation will be given in Chapter 7.

CHAPTER 3: NOTES

1. Cliff and Ord (1973) have shown that their test statistic is asymptotically normally distributed. When the number of areas is large (theoretically infinite) it becomes legitimate to convert the statistic into a standard normal deviate. However, their proof of this property assumes that no sub-area dominates the weights matrix. Sen (1976) has questioned Cliff and Ord's proof but he puts forward his own proof of this property and states that, in his experience, the number of areas under study needs to be greater than 50.

2. The significance testing can be done with regard to either of two null hypotheses.
   (a) Normality: the individual observations are the result of independent drawings from a normal population.
   (b) Randomization: whatever the underlying distribution of the population, the observed value is evaluated relative to all the possible values of the test statistic for the particular area being studied.

   In empirical work, it appears that non-normality of the data does not produce markedly different results under either assumption.

3. The statistic is asymptotically normally distributed and can therefore be evaluated as a standard normal deviate, provided that the weights are not dominated by a particular subset of areas and there is a reasonable number of areas. Hepple (1974b) has generalised the test, expressing it in a matrix form, thereby facilitating the calculation of the higher moments (leading to measures of kurtosis and skewness) and the exact distribution.

4. Interestingly, in the present context, differencing was first proposed by 'Student' to overcome the problem of spatial autocorrelation when correlating infant mortality and the death rate from tuberculosis in England, Ireland and Scotland.

5. Are Roberts and Lloyd really suggesting that the link between rainfall and heart disease is a causal one? A consideration of previous medical studies is given in Chapter 6.

6. The use of the asymptotic properties of the Cliff-Ord test with these particular examples can be criticised on a number of grounds. With regard to West and Lowe's study, some of the County Boroughs (for example, in the West Midlands) are 'joined' like a star lattice with only one articulation point. This, of course, disregards Cliff and Ord's admonition that no subarea should dominate the lattice of joins. With respect to the South Wales examples, the rural districts act like star lattices and perhaps there are insufficient observations.

7. See Granger and Newbold (1974) for a similar, if less extreme, situation occurring in econometrics.

8. While Geary rightly emphasizes the need to carefully check for spatial dependence in the residuals, his method should be applied with care. Geary achieved his stated aim by regressing the dependent variable against latitude and longitude. Whether it is particularly meaningful to do so is another matter.

CHAPTER 3: BIBLIOGRAPHY

BANNISTER, G. (1975): Population change in Southern Ontario
Annals of the Association of American Geographers 65, 177-188.

BARTELS, C.P.A. and HORDIJK, L. (1977): On the power of the generalised Moran contiguity coefficient in testing for spatial autocorrelation among regression disturbances.
Regional Science and Urban Economics 7, 83-101.

BENNETT, R.J. (1974): A review of spatial autocorrelation by A.D. Cliff and J.K. Ord
Environment and Planning 6, 241.

BERRY, B.J.L. (1971): Problems of data organisation and analytical methods in geography
Journal of the American Statistical Association 66, 510-523.

BESAG, J.E. (1974): Spatial interaction and the statistical analysis of lattice systems (with discussion)
Journal of the Royal Statistical Society Series B 36, 192-236.

BESAG, J. (1975): Statistical analysis of non-lattice data
The Statistician 24, 179-195.

CLIFF, A.D. (1969): *Some measures of spatial association in areal data*
unpublished Phd Thesis, University of Bristol.

CLIFF, A.D. (1970) : Computing the spatial correspondence between geographical patterns
Transactions of the Institute of British Geographers 50, 143-154.

CLIFF, A.D., Martin, R.L. and ORD, J.K. (1975): A test for spatial autocorrelation in choropleth maps based upon a modified χ^2 statistic
Transactions of the Institute of British Geographers 65, 109-129.

CLIFF, A.D. and ORD, J.K. (1969): The problem of spatial autocorrelation
in Scott, A.J. (ed.)
London Papers in Regional Science Volume 1, Pion, London.

CLIFF, A.D. and ORD, J.K. (1970): Spatial autocorrelation: a review of existing and new measures with applications
Economic Geography 46, 269-292.

CLIFF, A.D. and ORD J.K. (1971): Evaluating the percentage points of a spatial autocorrelation coefficient
Geographical Analysis 3, 51-62.

CLIFF, A.D. and ORD, J.K. (1972): Testing for spatial autocorrelation among regression residuals.
Geographical Analysis 4, 267-284.

CLIFF, A.D. and ORD J.K. (1973): Spatial Autocorrelation
Pion, London.

CLIFF, A.D. and ORD, J.K. (1975a): The comparison of means when samples consist of spatially autocorrelated observations
Environment and Planning Series A 7, 725-734.

CLIFF, A.D. and ORD J.K. (1975b): Model building and the analysis of spatial pattern in human geography
Journal of the Royal Statistical Society Series B 37, 297-348.

CLIFF, A.D. and ORD, J.K. (1975c): The choice of a test for spatial autocorrelation in Davis, J.C. and McCullagh, M. (eds.)
Display and Analysis of Spatial Data, Wiley, New York.

CLIFF, A.D. and ORD J.K. (1977): Large sample-size distribution of statistics used in testing for spatial autocorrelation: a comment
Geographical Analysis 9, 297-299.

COX, N.J. and ANDERSON, E.W. (1978): Teaching geographical data analysis: problems and possible solutions
Journal of Geography in Higher Education 2, 29-37.

CRUICKSHANK, D.B. (1940): A contribution towards the rational study of regional influences: group formation under random conditions
Papworth Research Bulletin 5, 36-81.

CRUICKSHANK, D.B. (1947): Regional influences in cancer
British Journal of Cancer 1, 109-128.

DACEY, M.F. (1965): A review of measures of contiguity for two and k-colour maps
Technical Report No. 2 Spatial Diffusion Study
Department of Geography, North Western University, Evanston, Illinois.

FISHER, W.D. (1971): Econometric estimation with spatial dependence
Regional and Urban Economics 1, 19-40.

GATRELL, A.C. (1977b): An introduction to spatial autocorrelation and its geographical applications
Discussion papers in geography 2, Department of Geography, University of Salford.

GEARY, R.C. (1954): The contiguity ratio and statistical mapping
The Incorporated Statistician 5, 115-145.

GEARY, R.C. (1963): Some remarks about relations between stochastic variables: a discussion document
Review of the International Statistical Institute 31, 163-181.

GOULD, P.R. (1970): Is Stastistix Inferens the geographical name for a wild goose?
Economic Geography 46, 439-448.

GRANGER, C.W.J. (1969): Spatial data and time series analysis in Scott, A.J. (ed.) London Papers in Regional Science Volume 1, Pion, London.

GRANGER, C.W.J. and NEWBOLD, P. (1974): Spurious regressions in econometrics Journal of Econometrics 2, 111-120.

GRIFFITH, D.A. (1976): Spatial autocorrelation problems: some preliminary sketches of a structural taxonomy The East Lakes Geographer 11, 59-68.

HART, J.T. (1970): The distribution of mortality from coronary heart disease in S. Wales Journal of the Royal College of General Practitioners 19, 258-268.

HAGGETT, P. CLIFF, A.D. and FREY, A. (1977): Locational Analysis in Human Geography (second edition), Arnold, London.

HAINING, R.P. (1978): The moving average model for spatial interaction Transactions of the Institute of British Geographers New Series 3, 202-225.

HEPPLE, L.W. (1974a): The impact of stochastic process theory upon spatial analysis in human geography Progress in Geography 6, 89-142.

HEPPLE, L.W. (1974b): Econometric Estimation and Model-Building with Spatial Series unpublished Phd Thesis, University of Cambridge.

HORDIJK, L. (1974): Spatial correlation in the disturbances of a linear inter-regional model Regional and Urban Economics: Operational Methods 4, 117-140.

JOHNSTON, J. (1972): Econometric methods McGraw-Hill Kogakusha, Tokyo.

KADIYALA, K.R. (1968): A transformation used to circumvent the problem of autocorrelation
Econometrica 36, 93-96.

LEBART, L. (1966): Les variables socio-economiques departmentales ou régionales I methodes statistiques d'etude
Institut d'Etude du Developpement Economique et Social, Paris.

LEBART, L. (1969): Analise statistique de la contiguité
Publications de l'Universite de Paris 18, 81-112.

MARTIN, R.L. (1974): On autocorrelation bias and the use of first spatial differences in regression analysis
Area 6, 185-194.

McCAMLEY, F. (1973): Testing for spatially autocorrelated disturbances with application to relationships estimated using Missouri county data
Regional Science Perspectives 3, 89-107.

MIRON, J. (1972): A note on the estimation of a spatially autoregressive model
Proceedings 22nd IGU Congress Montreal vol 2, 913-915.

MIRON, J.R. (1973): Spatial autocorrelation
Working Paper no 5, Department of Geography, University of Toronto.

MORAN, P.A.P. (1948): The interpretation of statistical maps
Journal of the Royal Statistical Society Series B 10, 243-251.

MORAN, P.A.P. (1950): Notes on continuous stochastic phenomena
Biometrika 37, 17-23.

ORD, J.K. (1975): Estimation methods for models of spatial interaction
Journal of the American Statistical Association 70, 120-126.

NEPRASH, J. (1934): Some problems in the correlation of spatially distributed variables
Journal of the American Statistical Association 28, 167-168.

ROBERTS, C.J. and LLOYD, S. (1972): Association between mortality from ischaemic heart-disease and rainfall in South Wales and the County Boroughs of England and Wales
Lancet 1, 1091-1093.

SEN, A. (1976): Large sample-size distribution of statistics used in testing for spatial correlation
Geographical Analysis 8, 175-184.

SEN, A. (1977): Large sample-size distributions of statistics used in testing for spatial correlation
Geographical Analysis 9, 230.

SIBERT, J. (1975): Spatial autocorrelation and the optimal prediction of assessed values
Michigan Geographical Publications, No 14, University of Michigan.

STEPHAN, F.F. (1934): Sampling errors and the interpretation of social data ordered in time and space
Journal of the American Statistical Association 29, 165-166.

'Student' (1914): The elimination of spurious correlation due to position in time or space
Biometrika 10, 179-180.

THEIL, H. (1965): The analysis of disturbances in regression analysis
Journal of the American Statistical Association 60, 1067-1079.

THOMAS, E.N. (1960): Maps of residuals from regressions: their characteristics and uses in geographic research
Report 2, Department of Geography, State University of Iowa.

TOBLER, W.R. (1970): A computer movie simulating urban growth in the Detroit region
Economic Geography 46, 234-240.

UNWIN, D.J. (1976): Patterns and processes in the plane
Area 8, 119-120.

UNWIN, D.J. and HEPPLE, L.W. (1974): The statistical analysis of spatial series
The Statistician 23, 211-227.

WEST, C.R. and LOWE, C.R. (1976): Mortality from ischaemic heart disease: inter-town variation and its association with climate in England and Wales
International Journal of Epidemiology 5, 195-201.

WRIGLEY, N. (1977): Probability surface mapping: a new approach to trend surface mapping
Transactions of the Institute of British Geographers New Series 2, 129-140.

YULE, G.U. (1926): Why do we sometimes get nonsense correlations between time series? A study in sampling and the nature of time series
Journal of the Royal Statistical Society 89, 1-69.

C H A P T E R 4

ANALYSIS OF SPECIFICATION ERRORS

'Rare is the analyst who possesses data which perfectly meet the assumptions of normality, homogeneity or linearity' Mitchell (1974, 507).

'Before we can declare a set of data to have been analysed adequately, we must have explored a variety of alternative models, we must have exposed all important anomalies in the data and we must have derived finally a model .... whose estimated parameters can be applied confidently to other models or other situations. Even though we are now well into the age of computers, not many sets of data are analysed adequately by the criteria I have suggested' Marquadt (1974, 189).

'Particularly with extensive data, the systematic plotting of residuals is likely to be the most searching way of testing and improving models' Cox (1968, 276).

INTRODUCTION

A number of procedures can and should be applied to a model during its development, and any evaluation of a model should begin with an examination of its assumptions. Such an evaluation is referred to as specification-error analysis. Specification errors in a model pose a fundamental challenge because the fitted model does not match the real world and cannot be used inferentially and consequently, the model has to be reformulated in some way. The common types of specification error are:

(1) peculiarities in the data being analysed (outliers and measurement error),

(2) an incorrectly specified disturbance term (non-normality and heteroscedasticity),

(3) the omission of important explanatory variables,

(4) disturbances which are autocorrelated,
and
(5) a model with a simultaneous equation problem.

It is the first three types of specification error that will be discussed in this chapter, for autocorrelated disturbances have already been considered in Chapter 3, and it is thought unlikely that a geographical model of mortality variation will need to be analysed as a simultaneous equation. (A model needs to be estimated as a simultaneous equation when the dependent variable affects the independent variable as well as vice-versa.)

While specification-error analysis is a very important aspect of regression modelling, it has received very little discussion from geographers. Even econometric and statistical textbooks commonly devote comparatively little space to these problems. For example, Johnston (1972), in a textbook much cited by quantitative geographers, devotes less than ten pages

to the topic of specification error (and seven of these are concerned with the one problem of heteroscedasticity). Moreover, the advice that does exist in the geographical literature is rather despairing about the problem of specification errors. For Olsson (1975, 412-413)

> 'specification errors are extremely difficult to handle .... they defy automatic detection ... There is no insurance against specification error. To say that we should proceed with an open mind, listen to our intuition and base the model in good theory is as trite as it sounds. It is nevertheless the best we can do, for to commit a specification error is to categorize wrongly. And he who categorizes wrongly cannot see the alternatives until he has seen them; what comes to the eye must first have been in the head'.

While Olsson's argument is substantially correct with regard to classical confirmatory statistical analysis, it is incorrect when exploratory data analysis is considered. This chapter attempts to demonstrate a variety of techniques for exploratory specification-error analysis and the first and major part of the discussion considers how to identify 'peculiarities' in the data, how to detect whether a model's disturbance term is homoscedastic and normal, and how to decide if a variable has been omitted from a model or an incorrect functional form of the model has been chosen. In this discussion emphasis will be placed on the use of graphical exploratory methods, although exact statistical procedures will also be considered when appropriate.

While each of the important specification errors will be treated separately, it must be remembered that many of these errors have similar effects and one 'true' specification error may result in the data apparently showing other types of error. Moreover, these different types of error can often be detected by similar means and so, in order to avoid repetition, only the most appropriate methods will be discussed in the context of a single specification error. In

the final part of the chapter the use of general statistical tests for the analysis of model specification will be examined.

OUTLIERS

It is important when analysing a set of data that the model represents the main body of that data and is not controlled by one or two extreme observations. Such extreme values are called 'outliers' and an outlier can therefore be loosely defined as an observation that 'behaves' differently from the rest of the observations. A single outlier can severely damage a statistical analysis but geographers in their research and their quantitative textbooks give scant attention to this crucial problem. In this section the effects of outliers on regression relationships are first discussed, methods for detecting outliers are then outlined and finally the problem of how to proceed when the data contain outlying observations is considered.

The effects

Research into the effect of outlying observations has shown the regression model to be extremely sensitive to such values, the outlying observations 'tilting' or 'pulling' the regression equation towards their values. Not only can the equation be dominated by one or two extreme values, but the regression summary statistics are also sensitive to them.

To illustrate the effect of a single outlier on OLS estimation, a hundred observations were generated according to the following equation:

$$y = 5.0 + 0.5x_1 + 0.5x_2 + 0.5x_3 + \varepsilon \qquad (1)$$

with the disturbance vector (ε) generated to have a mean of zero and a normal distribution.[1] An OLS regression model was fitted to this simulated data with the following results:

$$\hat{y} = 4.88 + 0.55x_1 + 0.52x_2 + 0.45x_3 \qquad (2)$$

The OLS estimates are very close to the true model values and the fitted model explains 93 per cent of the variation of the dependent variable. However, when one value of the dependent variable is changed from 14.47 to 74.47 (a mistake that may easily occur, for example, as data are punched onto computer cards), OLS estimation produces the following model:

$$\hat{y} = -1.68 + 1.66x_1 + 0.88x_2 + 0.43x_3 \qquad (3)$$

These estimates are poor and, in particular, the intercept term and the regression coefficient associated with variable x_1 depart markedly from the correct values. Moreover, this model only explains 33 per cent of the variation of the dependent variable. As Anscombe (1968, 178) has written:

> 'one sufficiently gross error in a reading can wreck the whole of a statistical analysis'.

Detection

Given the extent to which outliers can affect a regression model it is obviously extremely important to detect them. Formal statistical decision rules for the identification of outliers have been proposed by a number of workers and as Rider (1933, 2) so perceptively stated:

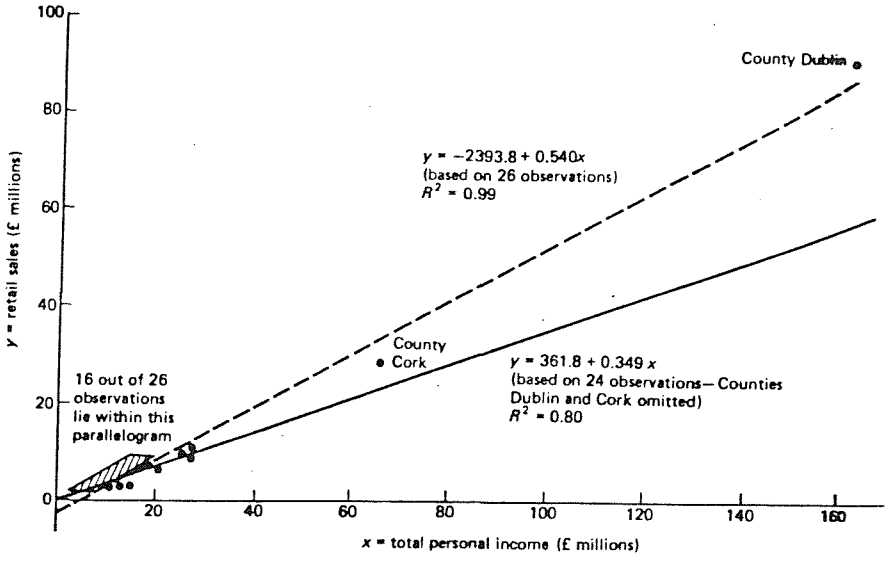
> 'devices for rejecting discordant observations could be invented without number'.

Ferguson (1961) surveys a large number of procedures that date from 1832 to 1960, and more recently Rosner (1975) and Ellenberg (1976) have proposed yet more rules. These rules are normally defined to reject an observation if a particular residual (standardised in some way) exceeds a multiple of the total residual variance of the model. Chatterjee and Price (1977, 27) have contended, however, that such rules should be applied with caution because it is the combination of the actual magnitude and the pattern of the residuals that suggests outliers. Moreover, as Anscombe (1968, 178) has argued:

> 'none of these rules has seemed entirely satisfactory, nor has any met with universal acceptance'.

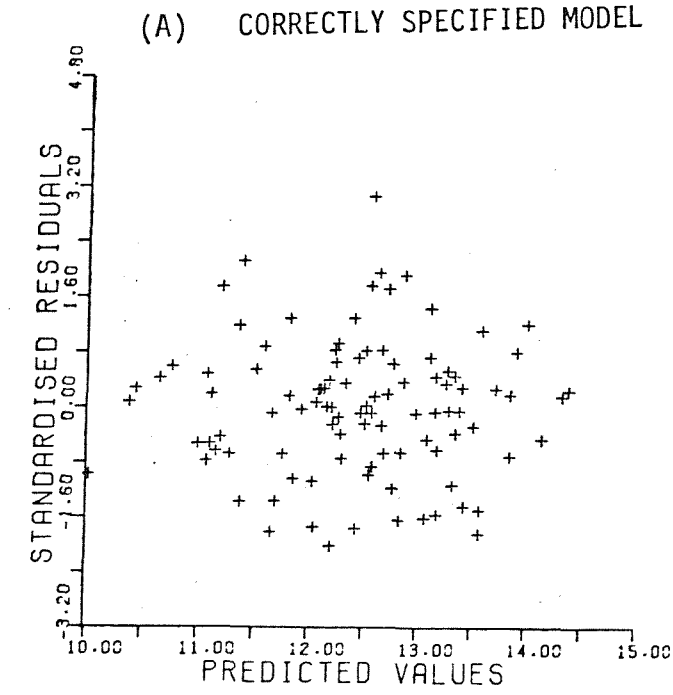
An alternative approach to detecting outliers is to use graphical methods. In a simple regression model consisting of one independent variable, an outlier may be detected by a graph of these two variables. Cliff and Ord (1973), in their re-analysis of O'Sullivan's (1968) study of the effect of total personal income on the value of retail sales, plotted the two variables (Figure 4.1) and found that the regression line was being dominated by the rather extreme counties of Dublin and Cork. Admittedly, the pattern of the data and the two outliers are immediately clear from this graph, but it has been found in other empirical analyses that the inadequacy of a model is more clearly identified in a plot of the residuals.[2] Therefore, one of the fundamental tenets of exploratory data analysis which underlies this chapter is that residual plots are more informative than graphs of raw data. While, as will be discussed later, several types of residual plots may reveal the existence of outlying points, one of the most simple yet effective methods is to plot the residuals against the predicted values of the dependent variable. In constructing such plots the residuals can usefully be scaled so that they have a mean of zero and a standard deviation of one. If this is the case, and the model has been correctly specified and there are no outlying observations, the majority of these standardised residuals will fall between plus and minus two and will be randomly distributed about zero. Figure 4.2 shows two such plots, calculated from the data used to estimate the values in equations (2) and (3). In Figure 4.2a there is no discernible pattern in the distribution of the residuals; they are not arranged in a systematic manner and they do not, in general, lie outside plus or minus two. The model appears to be correctly specified and no outlying values can be detected in the data. Figure 4.2b, in contrast, clearly shows a value that is away from the main body of the observations, and this value may be regarded as an outlier. However, such

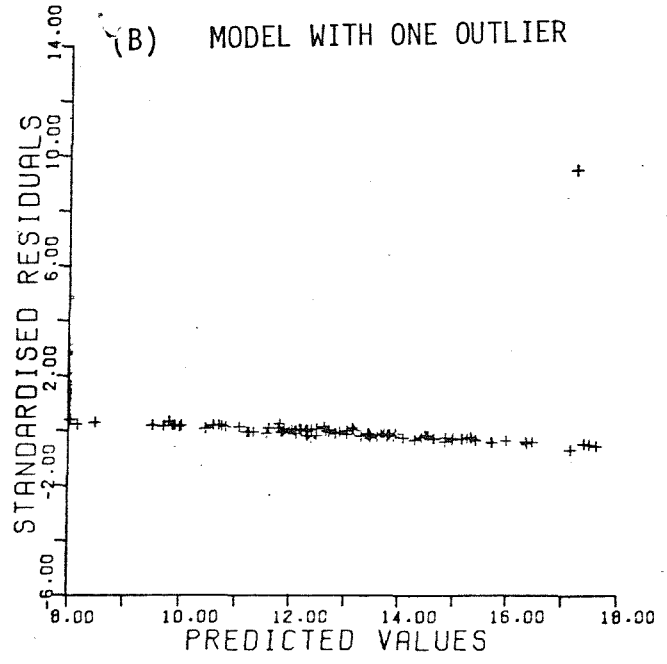
FIGURE 4.1 DETECTING OUTLIERS BY GRAPHICAL MEANS: THE CASE OF ONE EXPLANATORY VARIABLE - A SCATTERPLOT



Source: HAGGETT, CLIFF AND FREY (1977)

FIGURE 4.2 DETECTING OUTLIERS BY GRAPHICAL MEANS: THE CASE OF MORE THAN ONE EXPLANATORY VARIABLE





a pattern of residuals should not always be interpreted as the result of an incorrect observation as it could also be the result of a genuine observation with an incorrectly specified model. This leads on to the fundamental problem of how to treat an outlier once it has been discovered.

Dealing with the problem

One of the most obvious methods of dealing with an outlying point is to remove the offending value from the data and re-calibrate the model. However, as early as 1837 Hagen strongly opposed this procedure unless there was justification other than the fact that the values deviated considerably from the remaining observations. In effect, Hagan distinguished between 'true' and 'apparent' outliers and, although no other workers have explicitly adopted this categorisation, it is possible to go one step further by distinguishing three types of outlier each of which has to be treated in a different manner.

1. When a mistake has been made

If the observation has been measured incorrectly or has been mispunched on a computer card, then the observation must either be rejected or corrected. Less obviously, if an analyst can give reasons why certain observations are behaving in an abnormal way and therefore can explain why an observation has been included in a model by 'mistake', the value can also be discarded. For example, in collaborative work reported in Barnard (1978) an attempt was made to develop a model, using a large number of demographic, social and environmental variables, to account for the geographical distribution of the elderly in South Hampshire. A plot of the residuals from this model showed a number of apparently outlying observations. One of these, on further investigation, was found to be an army camp and two others represented areas with purpose-built residential homes for the elderly. Consequently, these errors, once detected were corrected by

omitting the three outlying values which the model was not specifically designed to accommodate.

Another approach that has become rather fashionable in the statistical literature for dealing with this particular type of outlier is to use so called 'robust estimators' to automatically remove the effect of such values. Huber (1973) and Andrews (1974), for example, have both developed robust forms of regression and others have suggested the use of 'Winsorized' models. 'Winsorizing' of data amounts to 'suspect' observations having their values replaced by the nearest value of a 'non-suspect' observation. But such approaches have the major drawback that the supposed outlying observations may be the result of an incorrectly specified model (see below) and, rather than suppressing such information, the researcher needs to know how to use it to develop an improved model. As Gnanadesikan (1977, 127) has written:

> 'the routine use of any robust estimate without exploring the data for the existence of specific peculiarities in them is neither wise nor necessary'.

2. When the model has been incorrectly specified

The relationship between the dependent and one independent variable may be exponential, for example, and this may show up as one or two outlying points on a residual plot. A similar pattern of residuals could also result from the residuals being heteroscedastic. Both these types of model mis-specification are considered later in this chapter when methods of dealing with such problems are discussed in detail.

3. When a 'true' outlier has been observed

There is little discussion on this difficult problem in the statistical literature but Anscombe (1973, 20) has suggested that the most preferable course is to estimate two models, one with and one without the outlying observations. Both models can then be reported but, while this action allows

an analyst to assess the effect of the outlying observations, it does not further the understanding of the model. 'True' outliers therefore remain both difficult to analyse and interpret.

An appropriate conclusion to this discussion on outliers is that, while data should be routinely checked for their presence, they should not be routinely discarded once they have been identified. Finally, it must be emphasized that the consideration of outliers has a wider importance in the analysis of model mis-specification. As Kruskal (1960, 1) has written:

> 'An apparently wild .... observation is a signal that says here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study'.

HETEROSCEDASTICITY

This basic assumption of the regression model implies that the variance of the disturbance term should remain the same, irrespective of small or large values of an explanatory variable. When this is not the case heteroscedasticity is said to be present. Since Mather and Openshaw (1974, 302) have argued that the problem of heteroscedasticity is likely to be very common in geographical research it is obviously important that we consider the effects of breaking this assumption, and how to deal with the problem once it has been detected.

At the outset it is crucial to distinguish between 'true' and apparent heteroscedasticity. A model may appear heteroscedastic for two reasons: measurement error and the omission of an explanatory variable. Commonly, as the values of an explanatory variable become large it becomes increasingly

difficult to measure accurately the value of the dependent variable. Furthermore, if a variable has been omitted from a regression model and this variable is positively related to the other explanatory variables in the model, there is likely to be an increase in the residual term as the values of the explanatory variables increase. Both these problems are not 'true' heteroscedasticity and should not be treated as such. Methods for dealing with these problems are considered later in this chapter.

The effects

Returning to the problem of 'true' heteroscedasticity, there appears to be strong empirical evidence that the most common situation in geographical research is one of increasing variance of the disturbance term with increasing values of the explanatory variable (Katona, 1954, 203). If this is the case, heteroscedasticity will have the following effects on OLS estimation.[3]

1. Although the regression estimates will not be biased they will have a large sample variance. That is they will be widely spread around the target of the true, but unknown, regression coefficients.

2. The variance of the error term will be underestimated and, as this value is used in the calculation of the standard regression tests, this will lead to (a) the overstatement of the significance of the model in terms of the t and F tests and (b) to the inflation of the coefficient of determination (R^2).[4]

Detection

Given that heteroscedasticity can have such a deleterious effect on model estimation, a method of detecting the problem is obviously required. Several procedures have been suggested for testing for heteroscedasticity, including those of Gorringe (1971) and Glejser (1969). One problem with such exact procedures is that the tests are based on unknown distributional assumptions, and parametric tests for

heteroscedasticity are known to be sensitive to departures from normality. Box (1953, 320) has likened such tests to

> 'putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port'.

Yet, while there is not a single, universally accepted test, there exists a quick and effective graphical alternative in the form of the residual plot that has already been used to detect outliers. This is well suited to detecting residuals with non-constant variance because, if the residuals are homoscedastic, a plot of the standardised residuals against the predicted values of the dependent variable will show a band of uniform width or no systematic pattern. However, a 'wedge-shaped' plot is a strong indication that the residuals are heteroscedastic.

Dealing with the problem

The most commonly used method of tackling heteroscedasticity is to transform the original model to obtain a version in which the transformed error term has constant variance. This reformulated model is then calibrated by OLS estimation and the resultant values are used to make inferences concerning the original model. The problem then becomes that of identifying the appropriate transformation. A careful examination of the diffuse literature on the topic reveals three different situations.

1. When there is *a priori* reasoning which transformation should be used.

2. When one of the explanatory variables can be shown to be inducing the heteroscedasticity.

3. When more than one variable is suspected of causing the problem.

These three distinct problems are now considered in more detail.

As an illustration of when a priori reasoning can help in the choice of a transformation, consider the following example. Let y be the number of female deaths from lung cancer in the Welsh local authorities in 1972 and x_1 be the total female population of each area. The postulated model is that the number of female death rates from this disease can be accounted for by the population of that area. (For simplicity all other explanatory variables have been ignored.) Female lung cancer deaths are rare events and it is widely assumed that such events (with a small probability of occurrence) follow a Poisson distribution. If this is correct, the mean and variance of the dependent variable will be similar, for this is a fundamental property of the distribution. (In fact, for the data analysed in Figure 4.3 the mean (0.76) was reasonably close to the variance (0.85).) As the mean of the dependent variable is expected to be related to the explanatory variable, the variance of y can be expected to be proportional to x_1 and it is therefore anticipated that the model will have heteroscedastic residuals. Such a model has been fitted to the Welsh data for 1972, and the standardised residuals are plotted against the expected values of the dependent variable in Figure 4.3a. This residual plot has a characteristic wedge shape indicating that the residuals do not have constant variance. However, it is known theoretically that the square root of a Poisson variable ($\sqrt{y}$) has a variance that is independent of the mean, and to overcome the problem of heteroscedasticity therefore, the analyst may regress $\sqrt{y}$ against x_1. The residual plot for such a model is given in Figure 4.3b, from which it can be seen that, compared with Figure 4.3a, the degree of 'fanning' has been considerably reduced, as has the coefficient of determination.

This illustration depends on the well-established result that the occurrence of rare events follows a Poisson distribution. However, there are situations when heteroscedasticity is present but there are no strong a priori

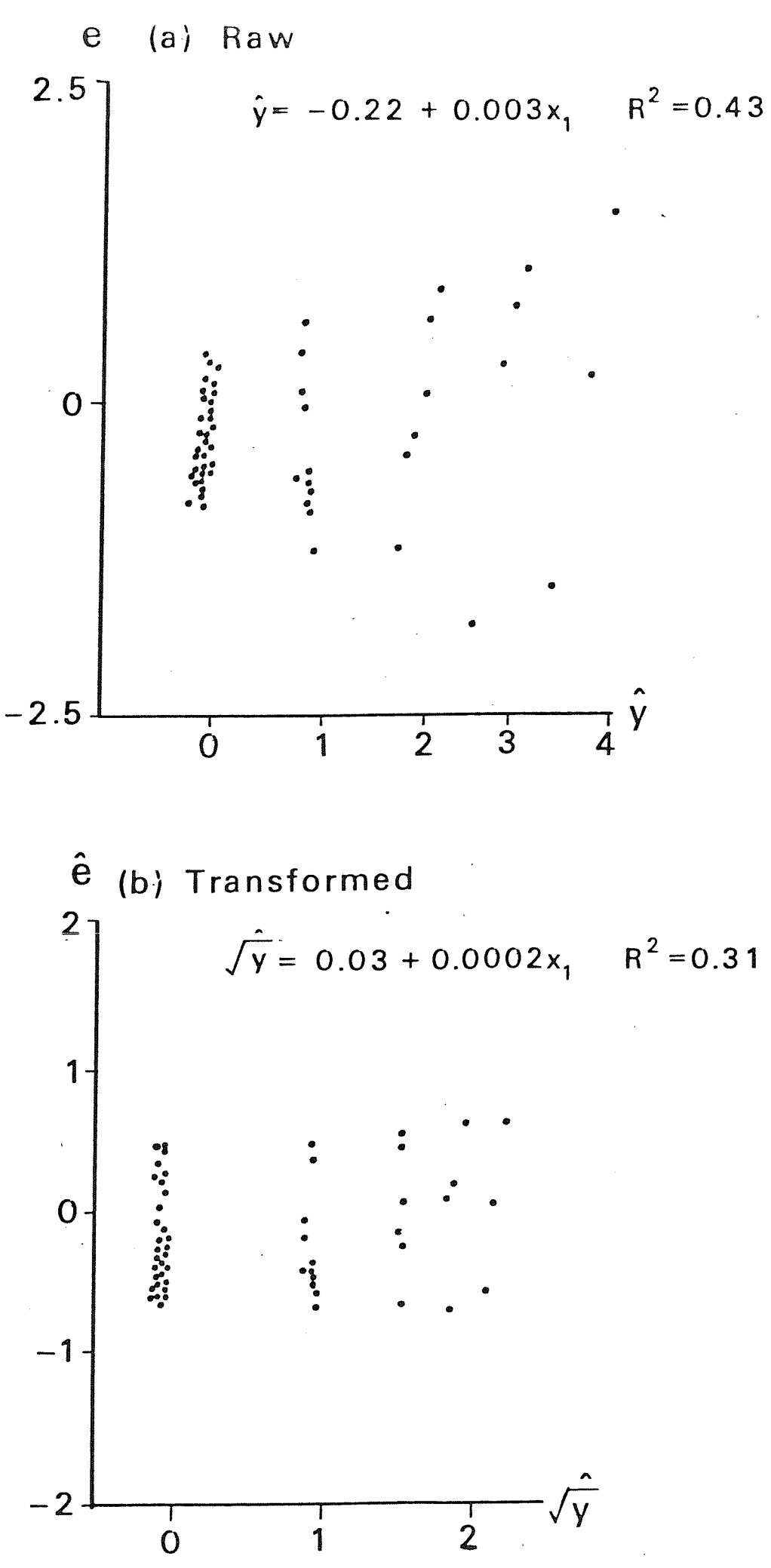


FIGURE 4.3 FEMALE LUNG CANCER DEATHS: RESIDUAL PLOTS

reasons for deciding which form the heteroscedasticity takes. In such a situation graphical analysis again helps to detect the problem and may also suggest ways of overcoming it. This approach to the problem of heteroscedasticity uses a combination of basic prior knowledge and intuition in a two-stage procedure. Firstly, OLS residuals are used to diagnose the problem and provide estimates of the residual structure; secondly, these estimates are used to select an appropriate transformation that can overcome the problem. Consider the following model:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon \qquad (4)$$

A set of data (100 observations) has been generated according to this model with β_0 set to 5.0 and all the population regression coefficients ($\beta_1, \beta_2, \beta_3$) set to 0.5. The disturbance vector (ε) has been generated so as to be proportional to x_1. An OLS regression model is fitted to these data and, as can be seen from Table 4.1, the results do not closely approximate the true population coefficients. A plot of the standardised residuals against the predicted values of the dependent variable clearly shows heteroscedasticity as a diverging band of residuals (Figure 4.4a.) Plotting these residuals against each of explanatory variables indicates that it is x_1 that is causing the problem, the other plots (Figure 4.4 c,d) showing less systematic pattern. To deal with this problem a re-formulated model is required whereby the relationship between x_1 and the disturbance term is removed by dividing equation (4) throughout by x_1 (Chatterjee and Price, 1977, 103). The results of the true, original and transformed models are all given in Table 4.1 and it can be clearly seen that the transformed version closely approximates the true model.[5]

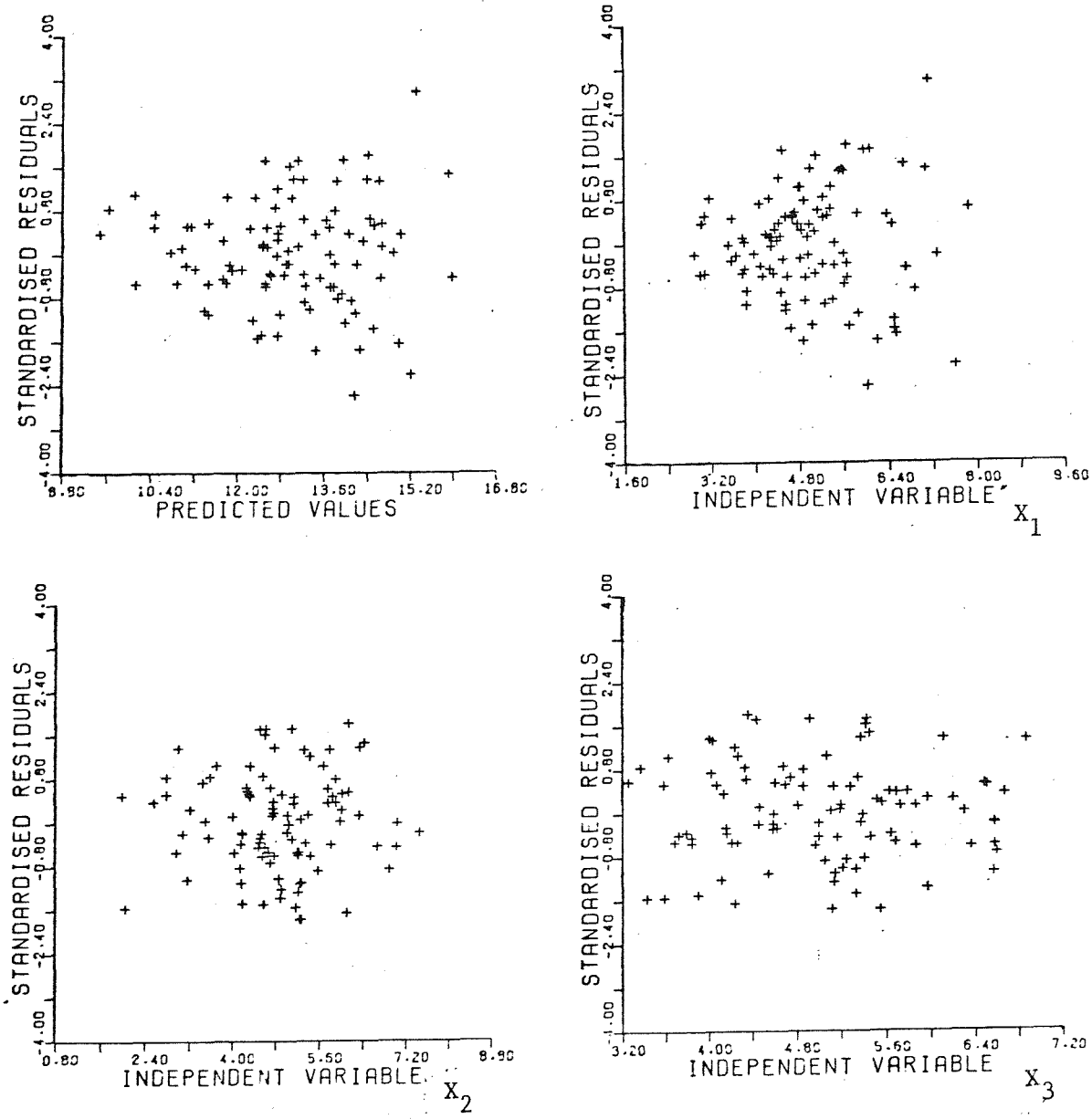
This procedure, however, assumes that the non-constancy of variance is a direct outcome of only one explanatory variable. When more than one explanatory variable is involved the difficulties of estimation become severe. One method that

Table 4.1

Estimating heteroscedastic relationships

True model	$y = 5.00 + 0.50x_1 + 0.50x_2 + 0.50x_3 + \varepsilon$
OLS estimation of original model	$\hat{y} = 2.41 + 0.76x_1 + 0.96x_2 + 0.41x_3$
OLS estimation of transformed model	$\hat{y} = 4.90 + 0.49x + 0.52x_2 + 0.45x_3$

FIGURE 4.4 RESIDUAL PLOTS ILLUSTRATING HETEROSCEDASTICITY IN VARIABLE X_1



can be used when this occurs requires several values of the dependent variable at fixed levels of the explanatory variable but, although in the experimental sciences data can be collected at certain fixed levels, it is extremely unlikely that this will be the case in non-experimental situations. Daniel and Wood (1971) have therefore considered a method of forming 'psuedo-replications' by clustering responses when the explanatory variables are approximately the same, but it must be admitted that dealing with heteroscedasticity involving more than one explanatory variable is an extremely difficult process.

This discussion has shown that heteroscedastic disturbances can cause severe problems for OLS estimation. Graphical methods can be used to detect such disturbances and can also be used as a guide to selecting appropriate transformations, thus enabling the model to be estimated by ordinary least-squares. However, the problem of estimating a model in which the disturbance term is related to more than one explanatory variable remains difficult.

One final cautionary note must be made regarding transformations. As stated at the beginning of this section, residuals may appear heteroscedastic due to the omission of an important explanatory variable or measurement error. If this is the case

> 'the blind application of .... transformations .... would make the random term homoscedastic, but the parameter estimates would still be wrong' (Koutsoyiannis, 1973, 186).

OMITTED VARIABLES AND MEASUREMENT ERROR

One of the fundamental assumptions of the regression model, as outlined in Chapter 1, is that the 'expectation' or mean value of the disturbance term is zero. A non-zero value can occur for several reasons: the model has an incorrect functional form, a variable has been omitted from consideration,

or the variables in the model have been measured with considerable error. The latter two problems will be discussed here; consideration of the models having an incorrect functional form is deferred until later.

The effects

Deegan (1974) has proposed a typology of the effects of incorrect inclusion or omission of variables from a model.

		Incorrectly Included Variables	
		No	Yes
Incorrectly omitted variables	No	correct model	Type A error
	Yes	Type B error	Type C error

A type A error occurs when an analyst correctly includes all the true (but unknown) explanatory variables but incorrectly includes other variables. With such an error no bias will be imparted to the estimated coefficients of the true model and the superfluous variables should have a regression coefficient value of approximately zero. A type B error occurs when the hypothesised model does not include a variable or variables that should have been included to produce the correct model, a situation which results in the estimated coefficients being biased. Deegan also recognises a Type C error, a combination of the A and B types. In this instance the postulated incorrect model includes some variables that have no true influence on the dependent variable and excludes variables that should be included in the true model. Such an error again results in the estimated coefficients being biased, yet the bias imparted to the model is more than just the sum of a Type A and Type B errors for even those variables only affected by a Type A error are biased when Type C errors are present.

Type B errors are particularly important in geographical and epidemiological work because some researchers deliberately

commit the error to avoid multicollinearity. For example, a researcher suspects that sulphur dioxide and smoke pollution are related to mortality. On calculating a correlation coefficient between the two variables he finds that both air pollution measures are highly related and he removes sulphur dioxide from his models. However, if both variables are truly related to mortality, this action can lead to bias. In order to demonstrate this particular difficulty and to illustrate the general problem of omitted variables a simulation study has been performed by the author.

Each model in the simulation consisted of two explanatory variables, 32 observations and was replicated 100 times. For each true model three postulated models were fitted; one of which was correctly specified, while the other two were incorrect in that one of two explanatory variables was omitted from the estimated model. Table 4.2a shows the results when the two explanatory variables are not collinear. The values given for the estimated models are the means of 100 simulated models and clearly, in this instance, the bias induced by omitting a variable is negligible. However, the table does not show that for the incorrectly specified models the spread of the estimates (as measured by standard deviations) is much wider than for the correctly specified models. Moreover, in Table 4.2b the results are given for a multicollinear model and it can be seen that incorrectly specified estimates are nearly double the true values. Clearly, if a researcher is analysing such data and deliberately omits a variable, there is a severe risk of inferential error. As a final illustration of this problem consider the model:

$$y = -0.5x_1 + 0.5x_2 + \varepsilon \qquad (5)$$

which was simulated with x_1 positively correlated to x_2. As Table 4.2c shows, omitting either explanatory variable results in the conclusion that the remaining variable is not related to the dependent variable!

Table 4.2

The effect of omitted variables on OLS estimation

(a) No multicollinearity between x_1 and x_2

True model	$y = 0.50x_1 + 0.50x_2 + \varepsilon$
Estimated model	
correct	$\hat{y} = 0.50x_1 + 0.51x_2$
incorrect	$\hat{y} = 0.49x_1$
incorrect	$\hat{y} = 0.52x_2$

(b) High positive multicollinearity between x_1 and x_2

True model	$y = 0.50x_1 + 0.50x_2 + \varepsilon$
Estimated model	
correct	$\hat{y} = 0.49x_1 + 0.52x_2$
incorrect	$\hat{y} = 0.96x_1$
incorrect	$y = 0.97x_2$

(c) High positive multicollinearity between x_1 and x_2

True model	$y = -0.50x_1 + 0.50x_2 + \varepsilon$
Estimated model	
correct	$\hat{y} = -0.51x_1 + 0.50x_2$
incorrect	$\hat{y} = 0.01x_1$
incorrect	$\hat{y} = 0.00x_2$

Dealing with the problem

Given these extremely serious effects of omitting an important variable from a model, it is obviously crucial to detect the problem. As previously noted, the omission of a variable will lead to the mean of the disturbance term being substantially different from zero, which suggests that the mean of the residuals may be inspected to determine if an important variable has been omitted. However, this solution is not perfect because a non-zero mean may also occur due to the specified model having an incorrect functional form. If a non-zero mean is found, therefore, it is probably wise to search for an incorrect functional form and, if the model appears to be correct in this respect, and if the variables are measured without substantial error, a search for the omitted variable should then be undertaken. As illustrated in Chapter 3, mapping the residual term can be an extremely useful aid in this search.

Measurement error

Measurement error, in a similar manner to the omission of a variable may result in the disturbance term of the estimated model having a non-zero mean. Moreover, if the variables in the model are observed with considerable error, the estimates of the regression coefficients will be biased and the usual confirmatory procedures of model evaluation will also be rendered inaccurate. The severity of these effects depends on a number of factors, but the most important are the correlations between the explanatory variables and the relative magnitude of the data errors (Cole, 1969, 74). Methods of overcoming the effects of measurement error have been considered by a number of econometricians and they include, in increasing complexity, Wald's (1940) 'two-group' method, Bartlett's (1949) 'three-group' method, Durbin's (1954) 'ranking' method, Reiersol's (1945) method of 'instrumental variables' and finally, maximum likelihood

estimation, (Maddala, 1977), the most general method. However, even in the simplest case where all the measurement errors are uncorrelated, overcoming their effects requires a knowledge of the ratio between the variances of the measurement errors for the explanatory variables and the variance of the disturbance term. Of course, in practice this is an impossible task and, as Hodges and Moore (1972, 186) have written with regard to methods of dealing with measurement errors:

> 'in spite of a profusion of theoretical papers about them, the methods have tended to be neglected in practice in favour of the easy-to-use least squares'.

While an analyst should undoubtedly endeavour to measure his variables with accuracy he can also take some comfort from the fact that, if measurement errors are not large compared to the true stochastic error, the effects will be slight (Chatterjee and Price, 1977, 59).

NON-NORMALITY OF THE DISTURBANCE TERM

When the assumption of normally distributed residuals is not satisfied, the consequences for OLS estimation are not serious; the estimated coefficients remain unbiased and have minimum variance. The normality assumption is usually invoked to allow confirmatory testing of the model, but as Boneau (1960) has pointed out, provided the number of observations is greater than 30, the t and F tests are only slightly affected by gross non-normality. Given these arguments it would appear unnecessary to develop methods of assessing non-normality but, as heteroscedasticity or a model with an incorrect functional form can give the impression that the residuals are non-normal, the detection of non-normality can be used as an aid to the discovery of these more fundamental types of model mis-specification.

Detection

Like all the other types of model mis-specification that are considered in this chapter, a number of formal tests

exist for the detection of non-normality (Dyer, 1974). Yet, a number of recent discussants have contended that a single summary statistic cannot convey as much information as a graphical plot and, as Seber (1977, 170) in his authorative overview of regression analysis, has written:

> 'there seems to be a general consensus
> that plots are more informative so that
> further testing may be somewhat unnecessary'.

Consequently, formal significance tests will not be discussed here; instead, emphasis will be placed on graphical procedures. In particular, the discussion will attempt to illustrate two basic 'rules' of exploratory graphical analysis.

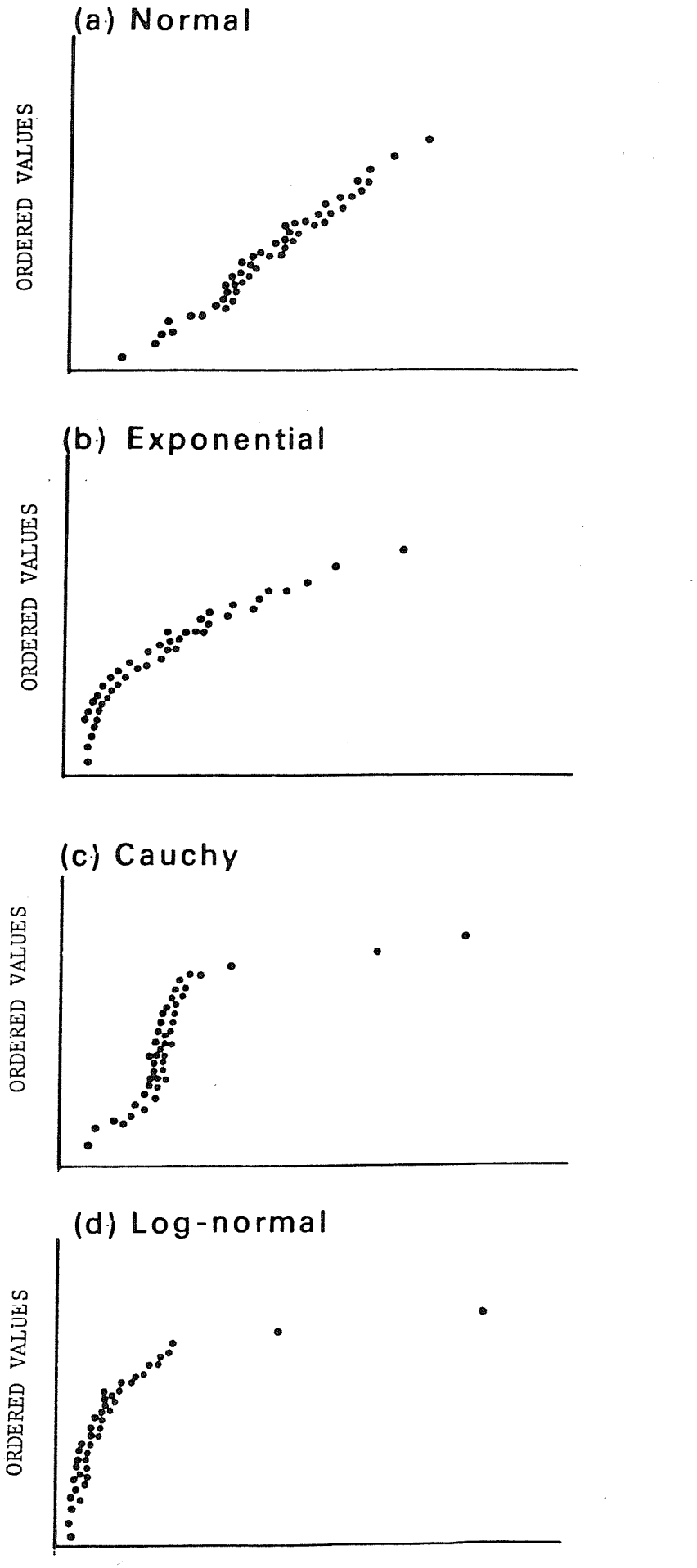
The first of these is the need to use a straight-line configuration to assess departures from an expectation. The degree of non-normality could be assessed by comparing a histogram of residuals against the characteristic 'bell-shape' of a normal distribution, but this raises a problem: into how many classes should the data be divided? Moreover, in practice it is much more difficult to use a complex curve as a reference rather than a straight line. Fortunately, the latter can be used as a means of detecting non-normality if a 'probability plot' is used, for a normal distribution will appear as a straight line on such a graph. One method of performing these plots is to use probability paper on which residuals are graphed against a 'corresponding function', such as $\frac{i-0.5}{N}$ or $\frac{i}{N+1}$ where i is a ranked residual and N is the total number of residuals. Similar plots can also be achieved by plotting the residuals against the standard normal deviates corresponding to the normal probability function. Whatever the method, an outlier is revealed by a single value being far removed from other observations; a symmetrical distribution that has longer tails than normal, or has several outliers, is shown by a plot rising at both ends; and a skewed distribution is shown by the plot being above the main body of the data at one extremity but not at the other.

The second rule of exploratory graphical analysis to be illustrated here is that the analyst must establish what is an acceptable level of departure from ideal conditions before he judges the data. The reason for this is that the residuals from a sample regression model will contain stochastic variation and it is important to get some feeling for 'normal departures' from normality. One way of setting such standards is to study simulated data with known distributional properties. Daniel and Wood (1971, 34-43) illustrated a number of probability plots of residuals from a normal distribution and concluded that a plot with only eight observations revealed almost nothing about normality; sets of 64 observations nearly always appeared straight in their central regions but fluctuated at their extremities, while sets of 384 observations seemed very stable, except for their lowest and highest points. Figure 4.5 shows a probability plot for normally distributed data as well as plots of data with other known distirbutional properties (Cauchy, log-normal and exponential, each with 50 observations).[6] It is by the study of a set of such plots that the analyst is able to develop a suitable benchmark for the use of graphical procedures.

Dealing with the problem

If the residuals are non-normal but the model is correct in all other respects, then the analyst should not be unduly concerned. But, if the model is found to be mis-specified in some way, this does require attention. Empirical work (Smith, 1972) has shown that, when a transformation is used to deal with heteroscedasticity and non-linearity, or when a previously omitted variable is included in the model, the residuals will generally appear to be more normally distributed. Thus, it appears that attempts to solve this problem should explore avenues such as these.

FIGURE 4.5 PROBABILITY PLOTS



INCORRECT FUNCTIONAL FORM

A dependent variable may be related to an explanatory variable in an infinite number of ways. While it is possible that theoretical considerations may specify a particular type of functional form, it is more common that there is no *a priori* reasoning to aid the exact choice of the relationship. For example, consider the relationship between air pollution and mortality. Is the relationship logarithmic, thereby postulating that, at high levels of air pollution, changes in the level of pollution should have less of an effect on mortality than the same changes at low pollution levels? Or should mortality be related to the square of air pollution, thus emphasizing the effect of changes at the higher values of the explanatory variable? Or should we postulate a stepped relationship - air pollution only having an effect after a certain critical threshold and up to another critical threshold?[7] The problem is that the analyst cannot specify the exact functional relationship because of a lack of theoretical knowledge. As Smith (1976, 150) has written:

> 'there are no real guidelines for specifying the mortality air-pollution relationship since we do not have a theoretical structure that yields it as a behavioural relationship'.

How, therefore, can we proceed?

The traditional approach is undoubtedly to fit a linear model and forget the problem. As Gould (1970, 441) has written, geographers are 'stuck in a linear rut'. If the traditional analyst is feeling adventurous a logarithmic relationship may be fitted (this is more likely to be the case if the researcher is a physical geographer). However, while the relationship under investigation may indeed be a linear or logarithmic one, unless the analyst has attempted to explore the nature of the functional relationship, he will not know whether the chosen relation is appropriate or not. The exploratory approach to the choice of functional form has two

aspects. Firstly, graphical procedures are used to detect non-linearity and, secondly, the same procedures are used to suggest transformations that will re-express the non-linear relationship in a more manageable linear form.

At the outset of this discussion of functional form, it must be realised that, in terms of OLS estimation, there are two types of non-linear models.

1. Those that are intrinsically linear: such models although they appear to be non-linear can be transformed to a linear model. For example, the model

$$y = \beta_o x_1^{\beta_1} \varepsilon \qquad (6)$$

can easily be transformed to a linear model by taking logarithms to the base e, that is

$$\log_e y = \log_e \beta_o + \beta_1 \log_e x_1 + \log_e \qquad (7)$$

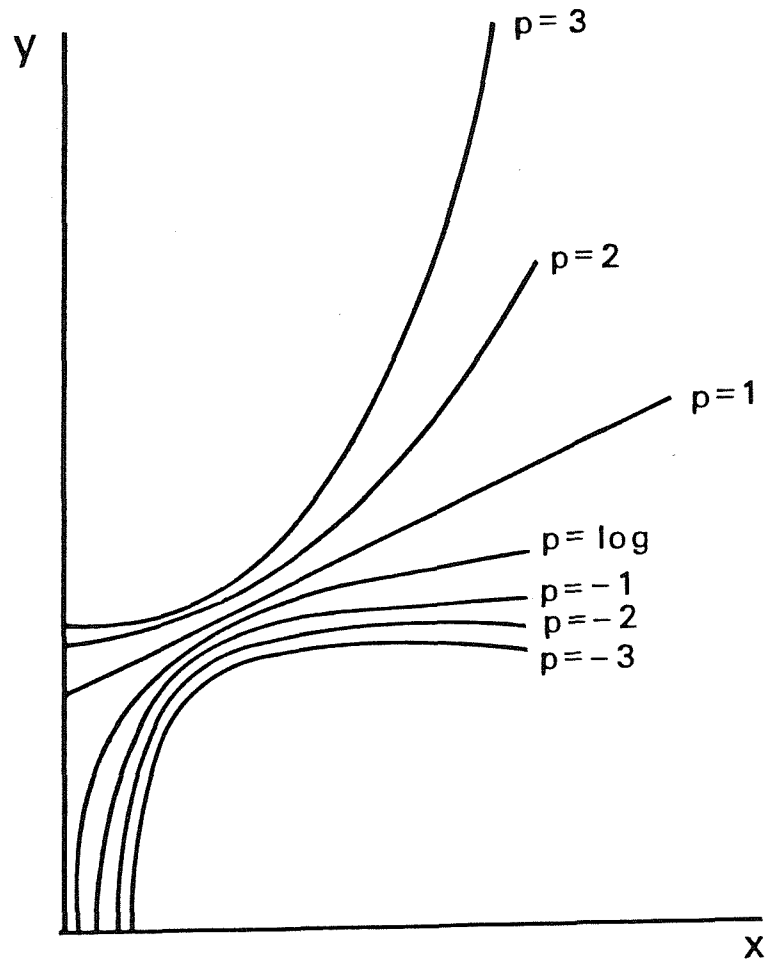
2. Intrinsically non-linear models: such models cannot be transformed to a linear model and cannot be estimated by ordinary least squares. For example, the non-linear model

$$y = \beta_o + x_1^{\beta_1} + \beta_2 x_2 \qquad (8)$$

cannot be transormed into linearity.

Intrinsically non-linear models cannot usually be solved by analytical methods and, therefore, they have to be estimated by iterative and approximate numerical methods (Gallant, 1975; Goldfeld and Quandt 1972, Mather, 1976). Such methods will not be discussed here, for although they overcome problems of non-linearity they do not overcome other problems of regression analysis such as multicollinearity. Moreover, many non-linear relationships can be approximated by simple transformations sufficiently accurately to ensure that the model can be estimated by ordinary least squares or by ridge regression. While there are a large number of transformations in general use (Hoyle's (1973) review identifies 19 commonly used transformations) the following exposition is confined to the simple set of power transformations. Following Mosteller and

FIGURE 4.6 POWER TRANSFORMATIONS



SOURCE: MOSTELLER AND TUKEY (1977)

Tukey (1977, 88) this simple set of transformations consists of raising a variable to a power (p) and, as Figure 4.6 shows, such transformations represent a wide range of functional relationships.

Graphical plots

The exploratory approach to choosing an appropriate power transformation relies on graphical analysis. At the outset it may be thought that functional form can be deduced from a scatter plot of the dependent variable against each of the explanatory variables in turn. Yet, while such a graph is extremely useful when there is only one explanatory variable in the model, it is of little use when there is more than one explanatory variable. Consider the following equation:

$$y = 10.0 - 2.0x_1 + x_2 \qquad (9)$$

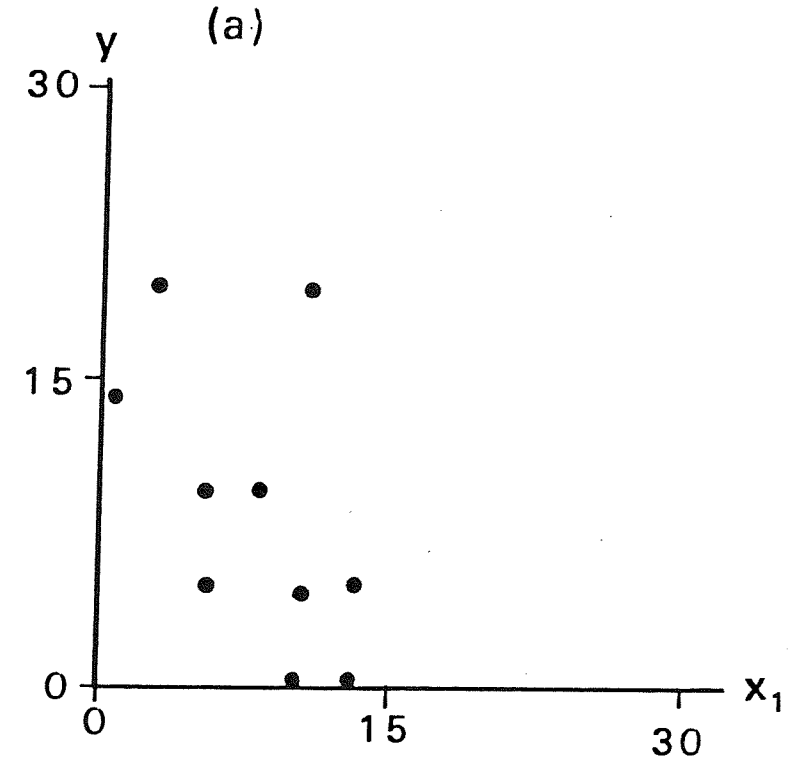
and the set of 10 observations generated from it (Table 4.3). Figure 4.7 shows two bi-variate scatter plots of this data and both the plots of y versus x_1 and y versus x_2 do not suggest a linear relationship despite the fact that equation (9) contains no disturbance term. The basic underlying problem is that the relationship is multivariate, and multi-dimensional data cannot be effectively portrayed on a 2-dimensional scatter plot. Fortunately, another approach is available: fit a linear equation to the data and use the residuals from this linear model to probe alternative functional forms.

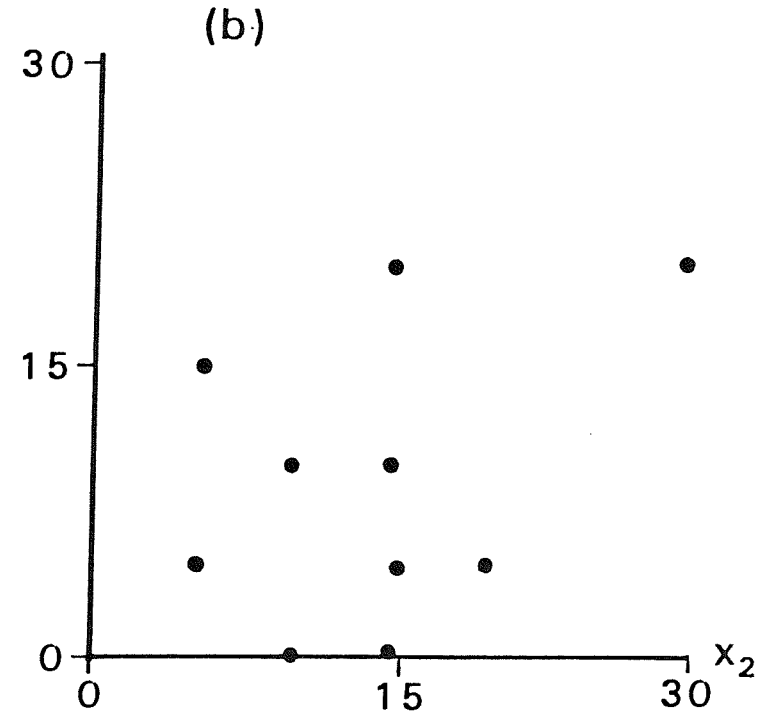
One of the simplest graphical methods of detecting an incorrect functional form is to use the plot that has already been shown to detect outliers and heteroscedasticity. However, although this type of plot is useful for detecting a mis-specified model, another type of plot is diagnostic of the actual transformation that needs to be used. Such a plot is called a partial-residual plot and was first discussed by Ezekiel and Fox (1959). Partial-residual plots have also recently been considered by Larsen and McCleary (1972) and are

Table 4.3

10 observations generated according to the linear function $y = 10.0 - 2.0x_1 + x_2$

y	x_1	x_2
0	10.0	10
5	5.0	5
15	0.0	5
10	7.5	15
20	2.5	15
5	12.5	20
5	10.0	15
0	12.5	15
10	5.0	10
20	10.0	30





$$y = 10 - 2x_1 + x_2$$

FIGURE 4.7 THE USE OF SCATTER PLOTS TO DETERMINE FUNCTIONAL FORM

closely related to Daniel and Wood's (1971) 'component plus residual' plot.

The partial residuals ($\hat{E}_p$) are defined as:

$$\hat{E}_p = \hat{E} + \hat{b}_j \, x_j \qquad (10)$$

where $\hat{E}$ is the OLS residual vector and $\hat{b}_j$ is an OLS regression coefficient associated with a particular explanatory variable x_j. Such a plot essentially examines the relationship between the dependent variable and a particular explanatory variable after the effect of the remaining explanatory variable has been removed. Mallows (1970) has suggested a minor modification so that the partial residuals are scaled to fall within the range of the associated explanatory variable. He suggests calculating

$$\hat{E}_m = \hat{E} + b_j \, (x_j - \bar{x}_j) + \bar{y} \qquad (11)$$

where $\bar{y}$ is the mean of the dependent variable and $\bar{x}_j$ is the mean of the independent variable j.

To demonstrate the use of such scaled partial residuals and to develop a 'feel' for this approach, a large number of models have been generated with known functional forms. A selection of some of these models is given in Table 4.4 and the residual plots associated with a correctly specified will be considered first. Model (a) is one in which the dependent variable is linearly related to three explanatory variables x_1, x_2 and x_3. If an analyst correctly postulates such a linear model the OLS regression estimates will be close to the coefficients of the true generating model. Figure 4.8a shows the residual plots for this model.[8] In the upper part of the diagram, a plot of the standardised residuals against each of the explanatory variables shows no systematic pattern thereby indicating that the postulated model is well-specified with no relationship between each explanatory variable and the residual variation of the dependent variable. Moreover, plotting the scaled partial residuals against each

Table 4.4

Analysing models with various functional forms

True model*	Postulated model	OLS estimated model	Specification error
(a) $y=5+0.5x_1+0.5_2+0.5x_3$	$y=b_o+b_1x_1+b_2x_2+b_3x_3+e$	$\hat{y}=4.9+0.5x_1+0.5x_2+0.5x_3$	NONE
(b) $y=5+0.5\frac{(1)}{x_1}+0.5x_2+0.5x_3$	"	$\hat{y}=5.2-0.02x_1+0.5x_2+0.5x_3$	requires a RECIPROCAL term
(c) $y=5+0.5(\log_{10} x_1)+0.5x_2+0.5x_3$	"	$\hat{y}=5.1+0.04x_1+0.5x_2+0.5x_3$	requires a LOGARITHMIC term
(d) $y=5+0.5(\text{sqrt}x_1)+0.5x_2+0.5x_3$	"	$\hat{y}=5.5+0.1x_1+0.5x_2+0.5x_3$	requires a SQUARE ROOT term
(e) $y=5+0.5x_1+0.5x_1^2+0.5x_2+0.5x_3$	"	$\hat{y}=-7.0+5.7x_1+0.5x_2+0.3x_3$	requires a SQUARE term
(f) $y=5+0.5x_1+0.5x_1^2+0.5x_1^3+0.5x_2+0.5x_3$	"	$\hat{y}=-133+47x_1+0.3x_2-2.0x_3$	requires a CUBIC term
(g) $y=5+0.5x_1+0.5x_2+0.5x_1x_2+0.5x_3$	"	$\hat{y}=-6.7+2.9x_1+2.9x_2+0.5x_3$	requires an INTERACTION term

* Each model is based on 100 generated observations

FIGURE 4.8 DETECTING MODELS WITH INCORRECT FORM BY GRAPHCIAL METHODS

(A) CORRECTLY SPECIFIED MODEL

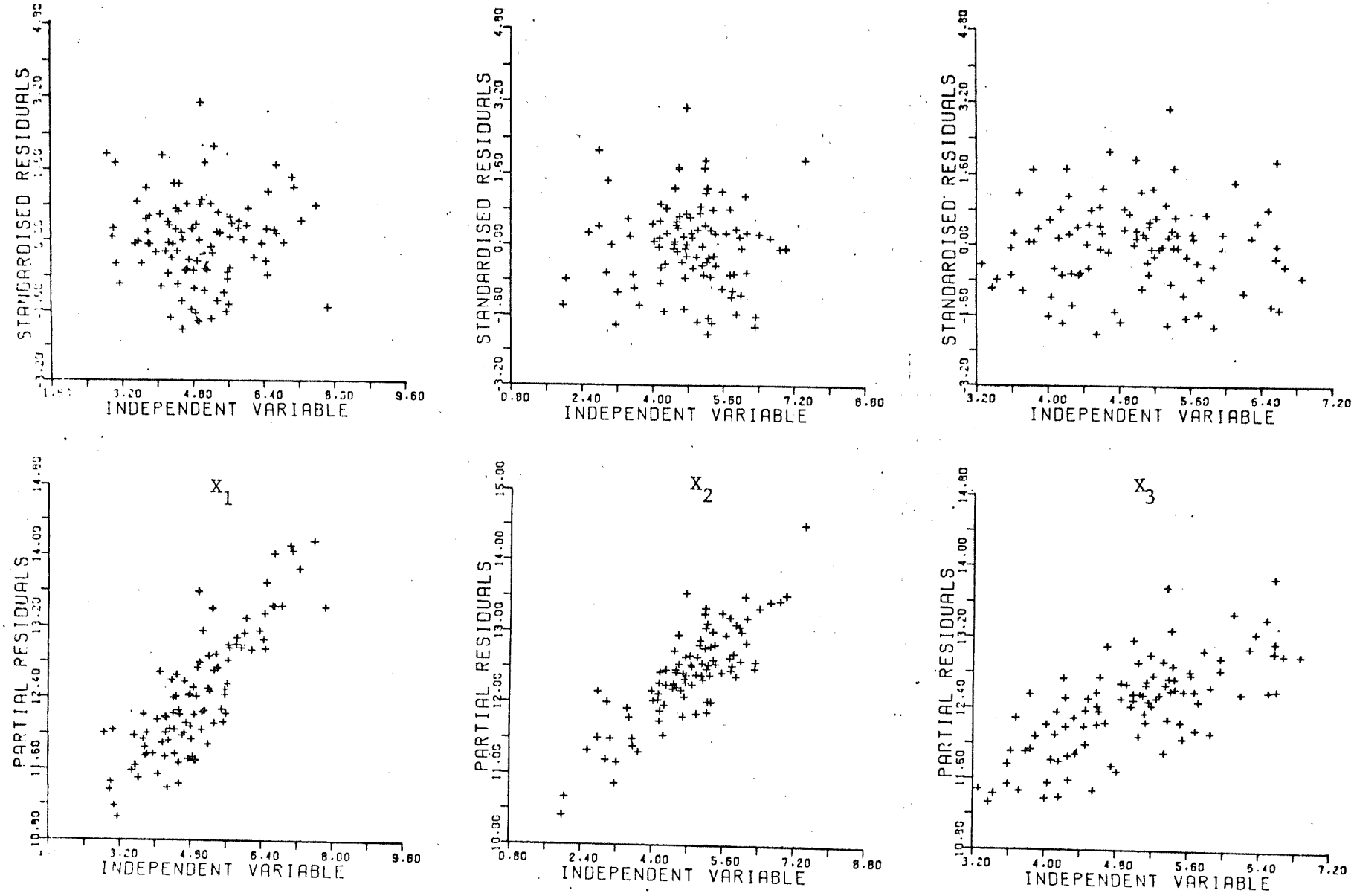


FIGURE 4.8 DETECTING MODELS WITH INCORRECT FORM BY GRAPHICAL METHODS

(B) INCORRECTLY SPECIFIED MODEL: X_1 REQUIRES A RECIPROCAL TRANSFORMATION

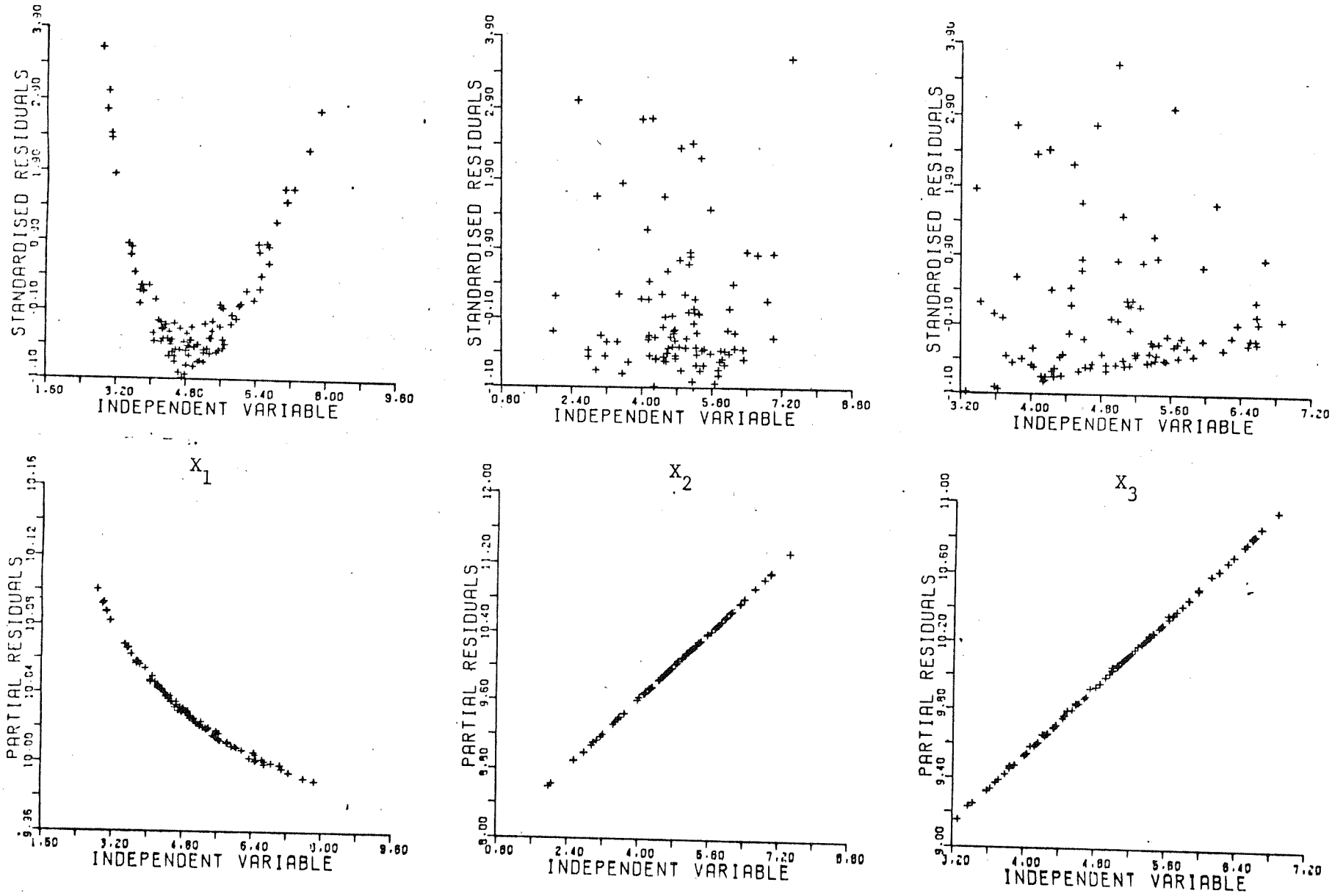


FIGURE 4.8 DETECTING MODELS WITH INCORRECT FORM BY GRAPHICAL METHODS

(C) INCORRECTLY SPECIFIED MODEL: X_1 REQUIRES A LOGARITHMIC TRANSFORMATION

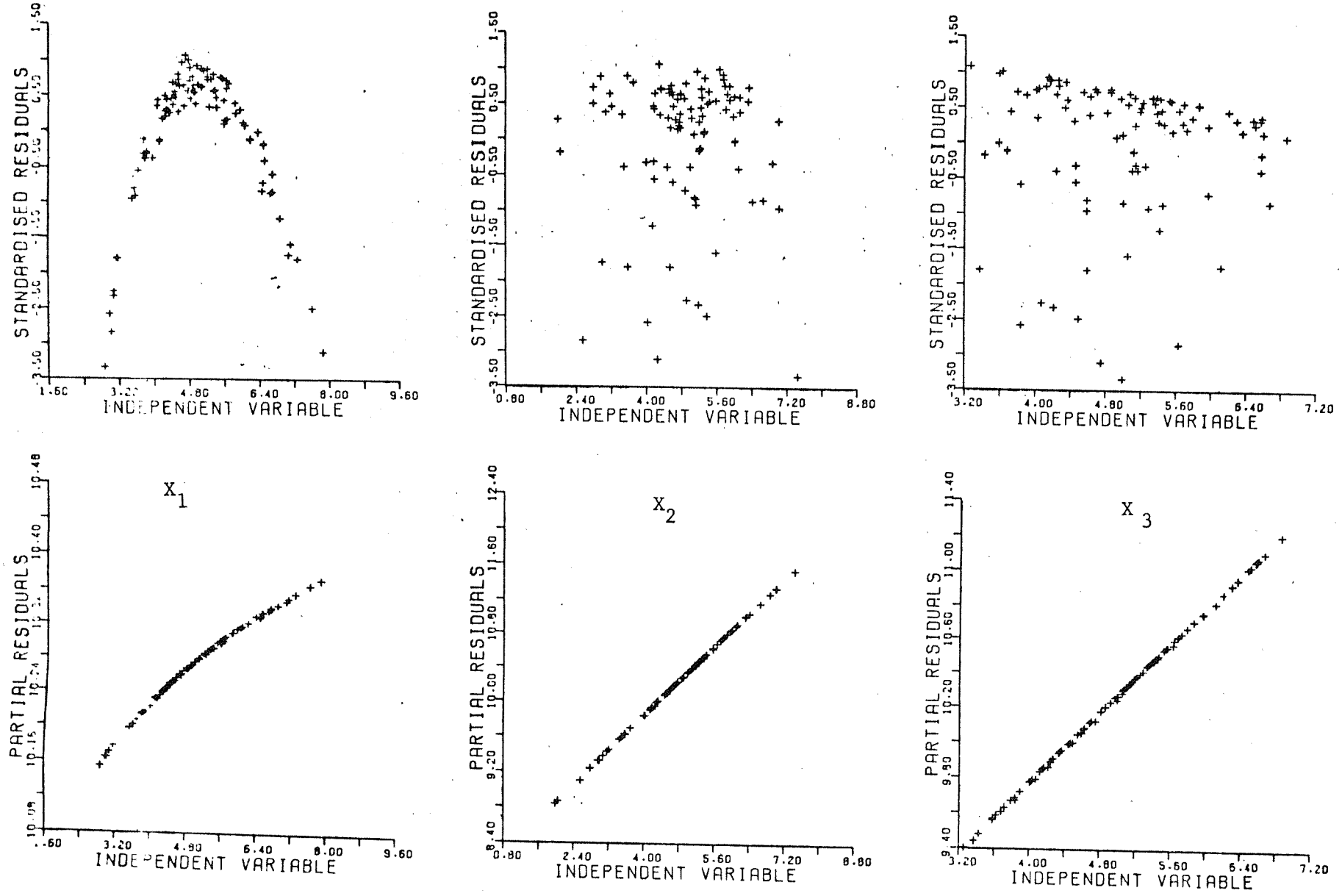
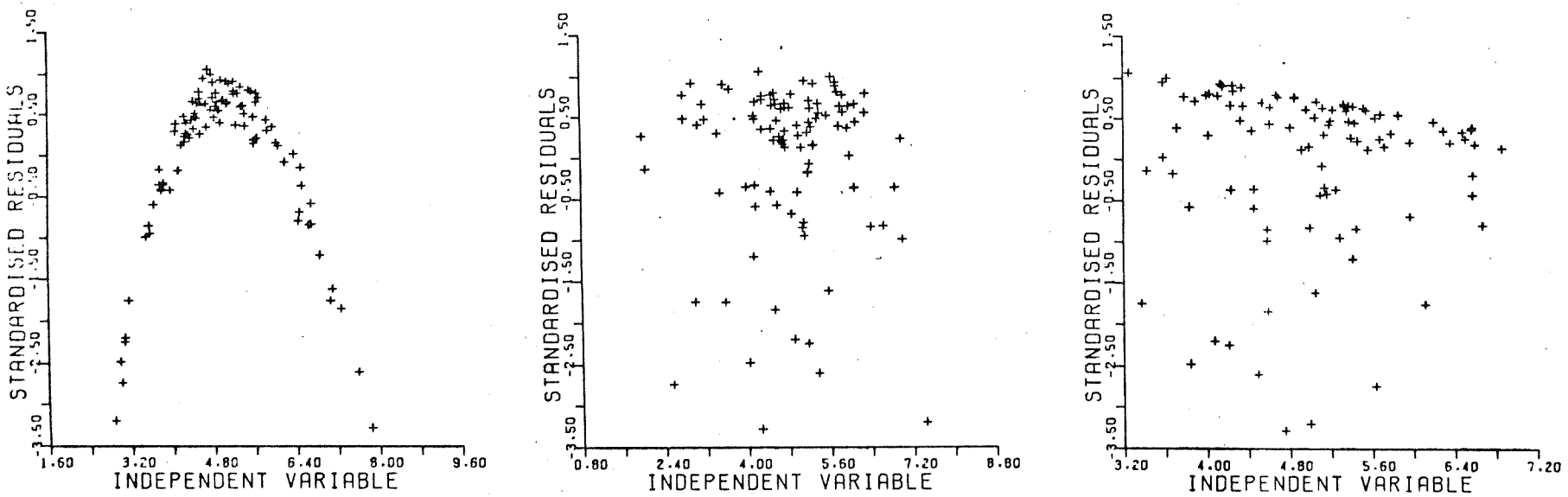


FIGURE 4.8 DETECTING MODELS WITH INCORRECT FORM BY GRAPHICAL METHODS

(D) INCORRECTLY SPECIFIED MODEL: X_1 REQUIRES A SQUARE ROOT TRANSFORMATION



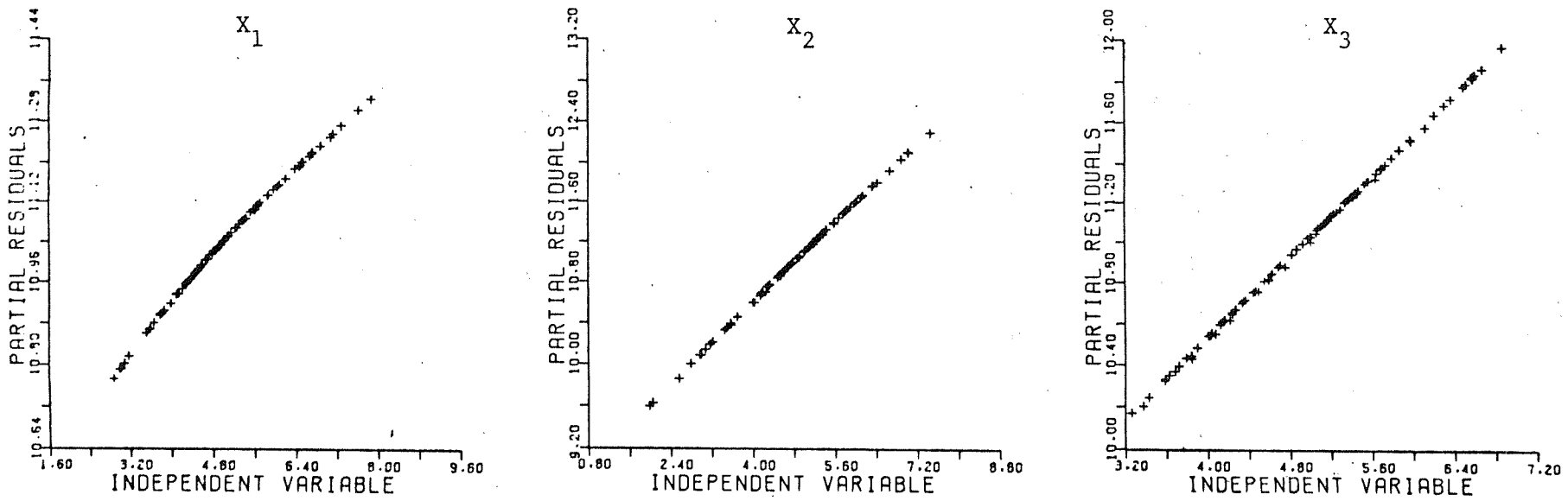


FIGURE 4.8 DETECTING MODELS WITH INCORRECT FORM BY GRAPHICAL METHODS

(E) INCORRECTLY SPECIFIED MODEL: X_1 REQUIRES A SQUARED TRANSFORMATION

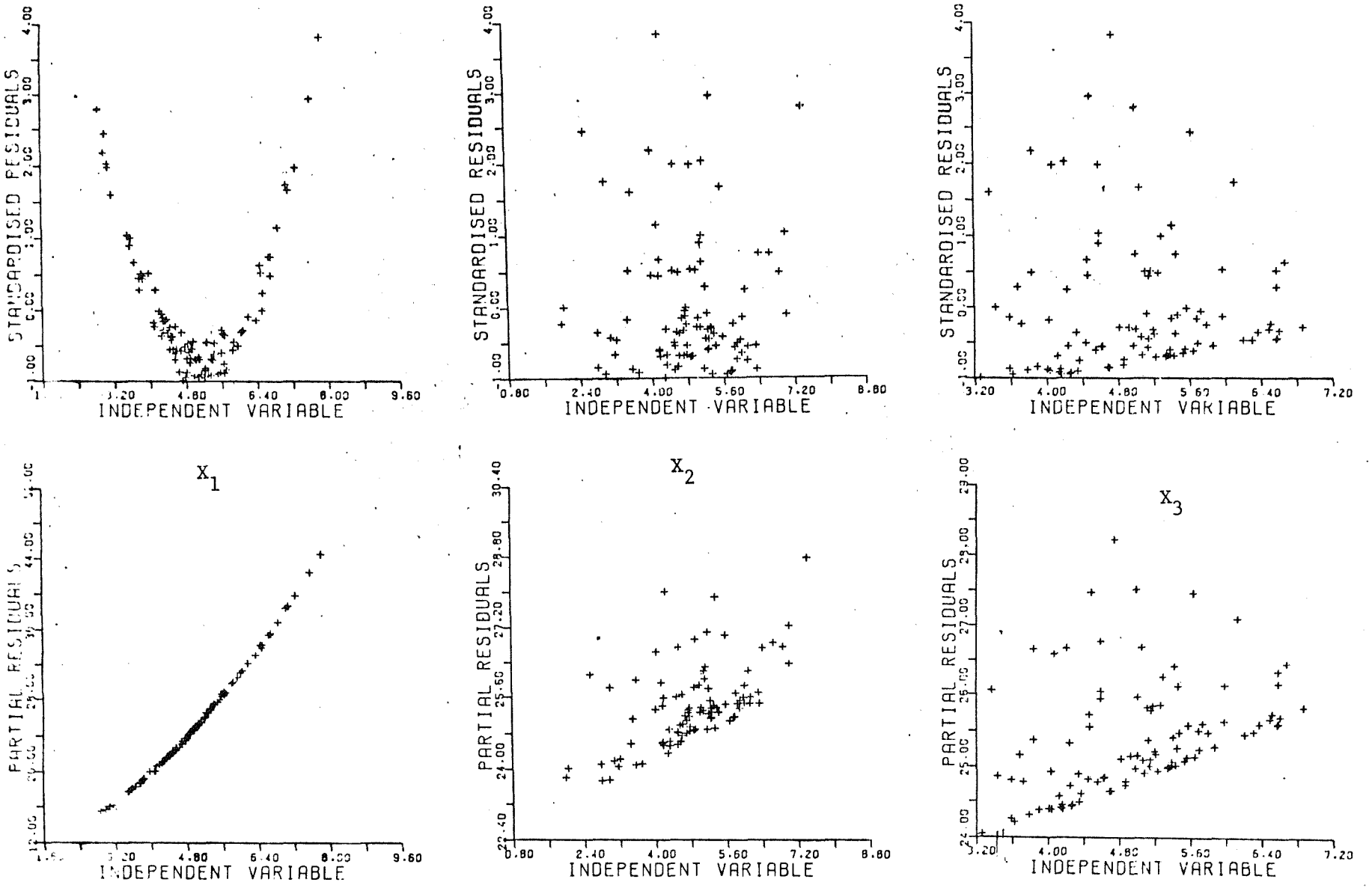


FIGURE 4.8 DETECTING MODELS WITH INCORRECT FORM BY GRAPHICAL METHODS

(F) INCORRECTLY SPECIFIED MODEL: X_1 REQUIRES A CUBIC TRANSFORMATION

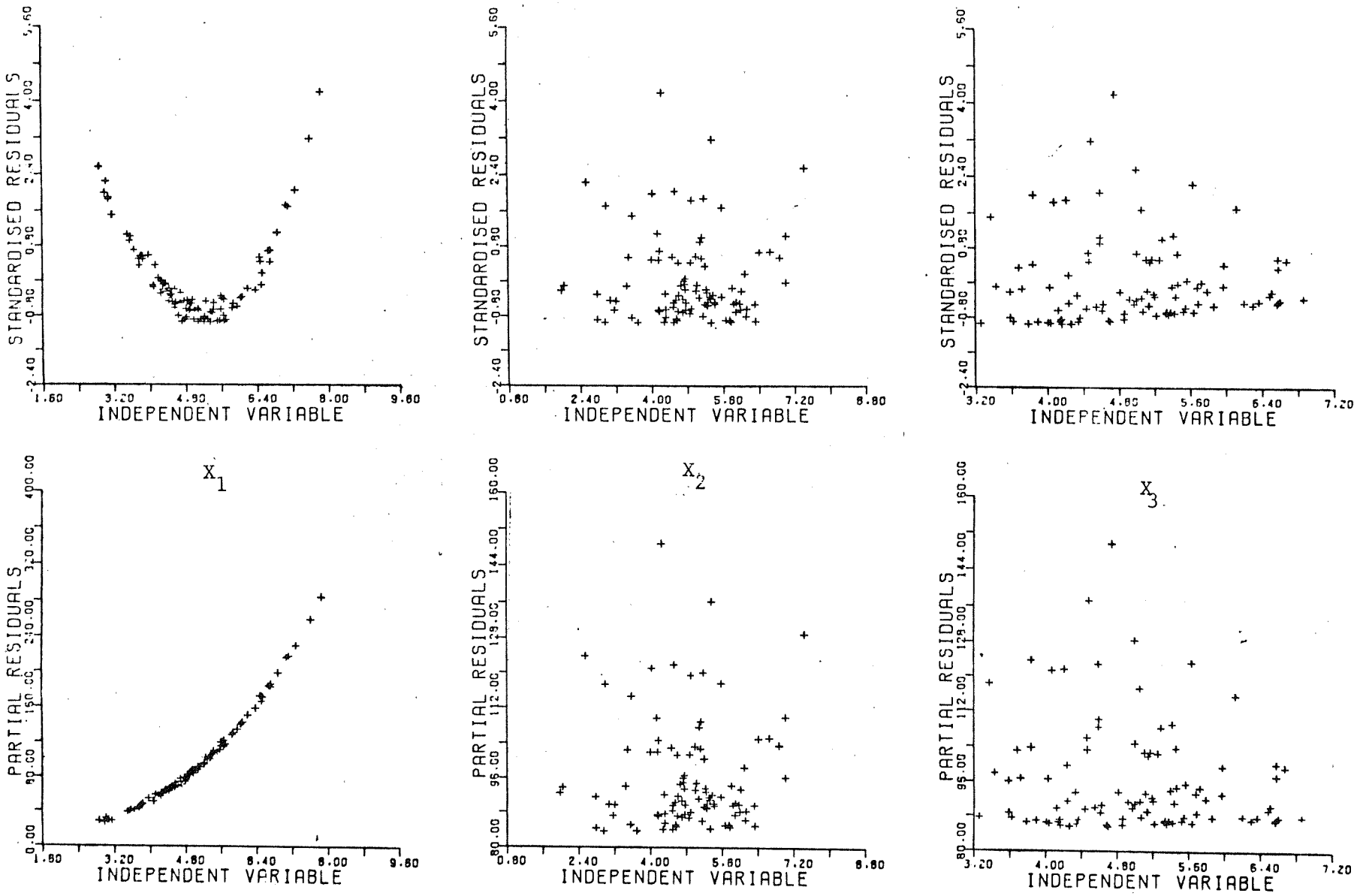
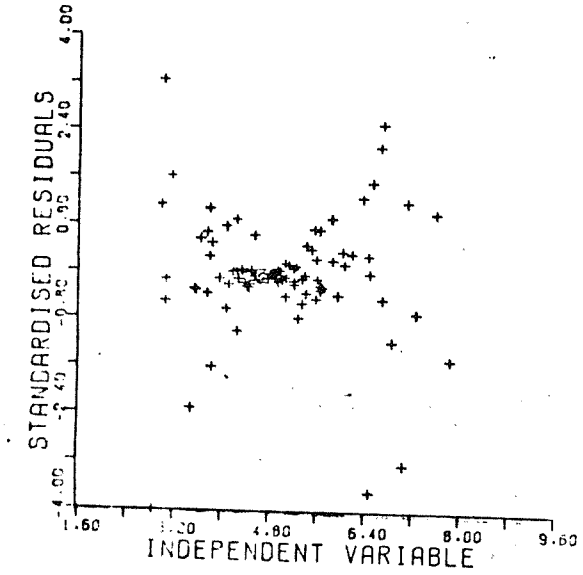
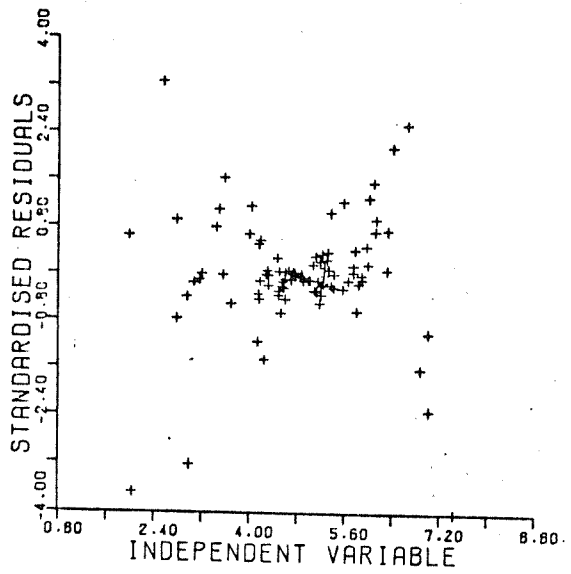
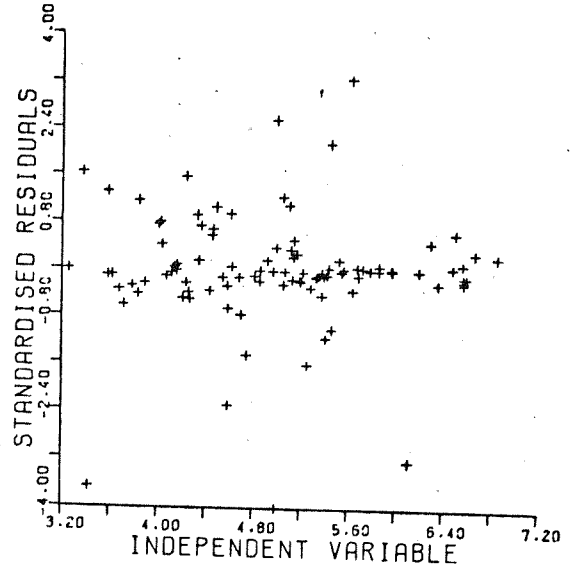


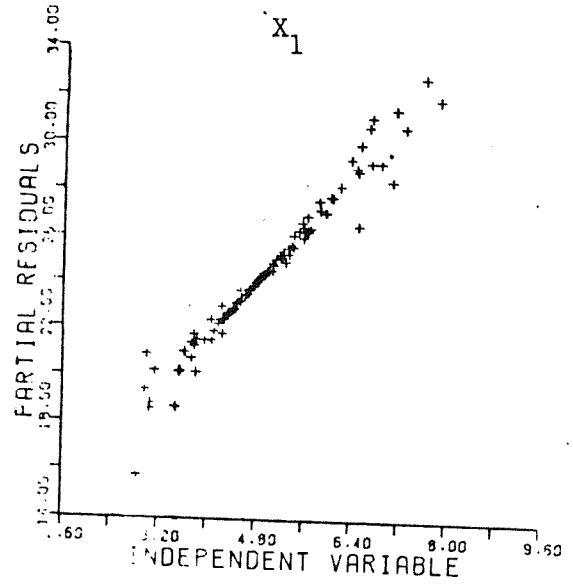
FIGURE 4.8 DETECTING MODELS WITH INCORRECT FORM BY GRAPHICAL MEHTODS

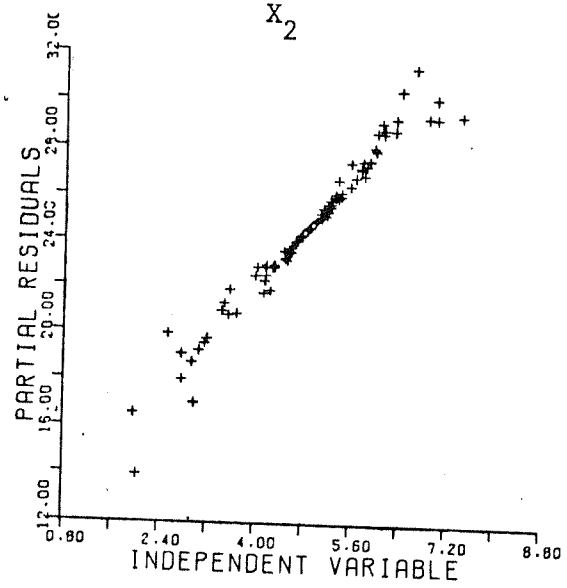
(G) INCORRECTLY SPECIFIED MODEL: X_1 AN INTERACTION TERM ($X_1.X_2$) IS REQUIRED

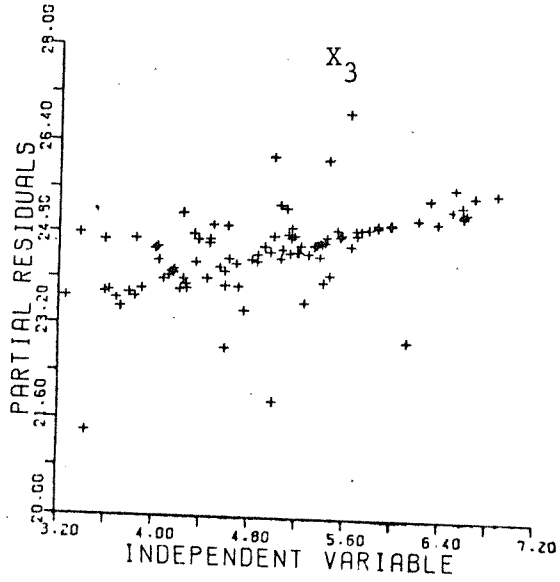












of the explanatory variables correctly reveals in the lower part of the diagram that the dependent variable is related in a linear fashion to each of the explanatory variables.

In contrast, let us now consider the residual plots associated with an incorrectly specified model. Model (b) of Table 4.4 represents an equation in which the dependent variable is linearly related to x_2 and x_3, but is related in a reciprocal fashion to variable x_1. If an analyst incorrectly postulates that the dependent variable is linearly related to all three explanatory variables, the OLS regression estimate associated with x_1 will be substantially different from the true generating model. Fortunately, the residual plots associated with this model clearly reveal the problem. The most striking features of Figure 4.8b are the plots associated with x_1. The standardised residuals when plotted against x_1 show a distinct curved band indicating that there remains a systematic relationship between this particular explanatory variable and the dependent variable after a linear model has been fitted. The partial residual plots also reveal a non-linear relationship between the explanatory variable x_1 and the dependent variable.

A selection of other models estimated with an incorrect functional form is given in Table 4.4 and residual plots for these models are displayed in Figure 4.8 (c to g). Examining the table reveals that model mis-specification can result in gross errors of estimation, for example, when the true model contained a cubic relationship, (model f) the strength of the associated regression coefficient was over-estimated by nearly a hundredfold. Moreover, an examination of the graphs of Figure 4.8 will show that, in each case, the residual plots are an excellent means of detecting model mis-specification.[9]

Box-Cox transformations

The examples of Table 4.4 are rather naive in that the true regression coefficients relating the explanatory variables

to the dependent variable are all of the same sign and magnitude (+ .5). In practice it has been found that the selection of an appropriate transformation can be aided considerably by the combined use of partial residuals and Box-Cox transformations. In their original study, Box and Cox (1964) suggested that the dependent variable, y, should be transformed to have the form $\frac{y^{\lambda}-1}{\lambda}$ and, in order to choose an appropriate transformation, a large number of models each with the different value of lambda (λ) should be fitted.[10] The model with the smallest residual variance should then be chosen as being the model with the most appropriate form. In principle this procedure can be generalised so that, in addition to the dependent variable being transformed, all the explanatory variables can be similarly treated, each to a possibly different but suitable value of lambda. However, the search procedure cannot be used efficiently with more than two or three variables and, even in this rather simple case, the computations become very cumbersome and time consuming (Maddala, 1977, 317). The novel approach to Box-Cox transformations that is suggested here is to use the procedure to explore the functional relationship between the dependent variable and the partial residuals. This approach can be summarised as follows.

(1) Calculate an OLS regression model

(2) Take one explanatory variable and calculate the associated partial residuals.

(3) For a reasonable range of values of lambda, regress the partial residuals on the associated explanatory variables transformed to the different values of lambda.

(4) Plot the 'maximised log-likelihood' (L-max) against the trial series of lambda and choose that value of lambda for which L-max is at a maximum. This is the required power transformation for the explanatory variable, for at this value the residual sum of squares of the model will be at a minimum.

(5) Repeat steps 2 to 4 for each explanatory variable to ascertain which explanatory variable requires the most 'powerful' transformation (that is the variable requiring the highest value of lambda to make the relationship linear). Transform this particular explanatory variable to the appropriate value and leave the remaining explanatory variables untransformed.[11]

(6) Repeat steps 2 to 5 until the Box-Cox procedure suggests that the dependent variable is linearly related to each of the explanatory variables.

As a demonstration of this procedure consider three models (b,c,d) from Table 4.4. For all three models the dependent variable is related to x_2 and x_3 in a linear fashion, but it is related to the reciprocal of x_1 in model (b), to the logarithm of x_1 in model (c) and to the square root of x_1 in model (d). Figure 4.9 illustrates the plots of L-max against lambda for each model and each explanatory variable. For x_2 and x_3 in all three models L-max reaches a maximum when lambda equals 1, thus correctly indicating that no transformation is required and that the relationship between x_2, x_3 and y is a linear one. However, when the partial residuals associated with x_1 are analysed for model (b), L-max reaches a peak when lambda equals -1.0. This indicates, again quite correctly, that x_1 of the model requires a reciprocal transformation. Similarly, for model (c), which requires a logarithmic transformation of x_1, and for model (d), which requires a square-root transformation, L-max reaches a peak when lambda equals 0.0 and 0.5 respectively. Thus, in both cases, the Box-Cox procedure has correctly identified the appropriate transformation. Moreover, when the appropriate transformation was applied to the data and the models re-analysed by the Box-Cox technique, L-max was found to peak when lambda equalled zero, thereby indicating that no further transformation was required and that the models were now correctly specified.

The role of transformations

The discussion so far has concentrated on the selection

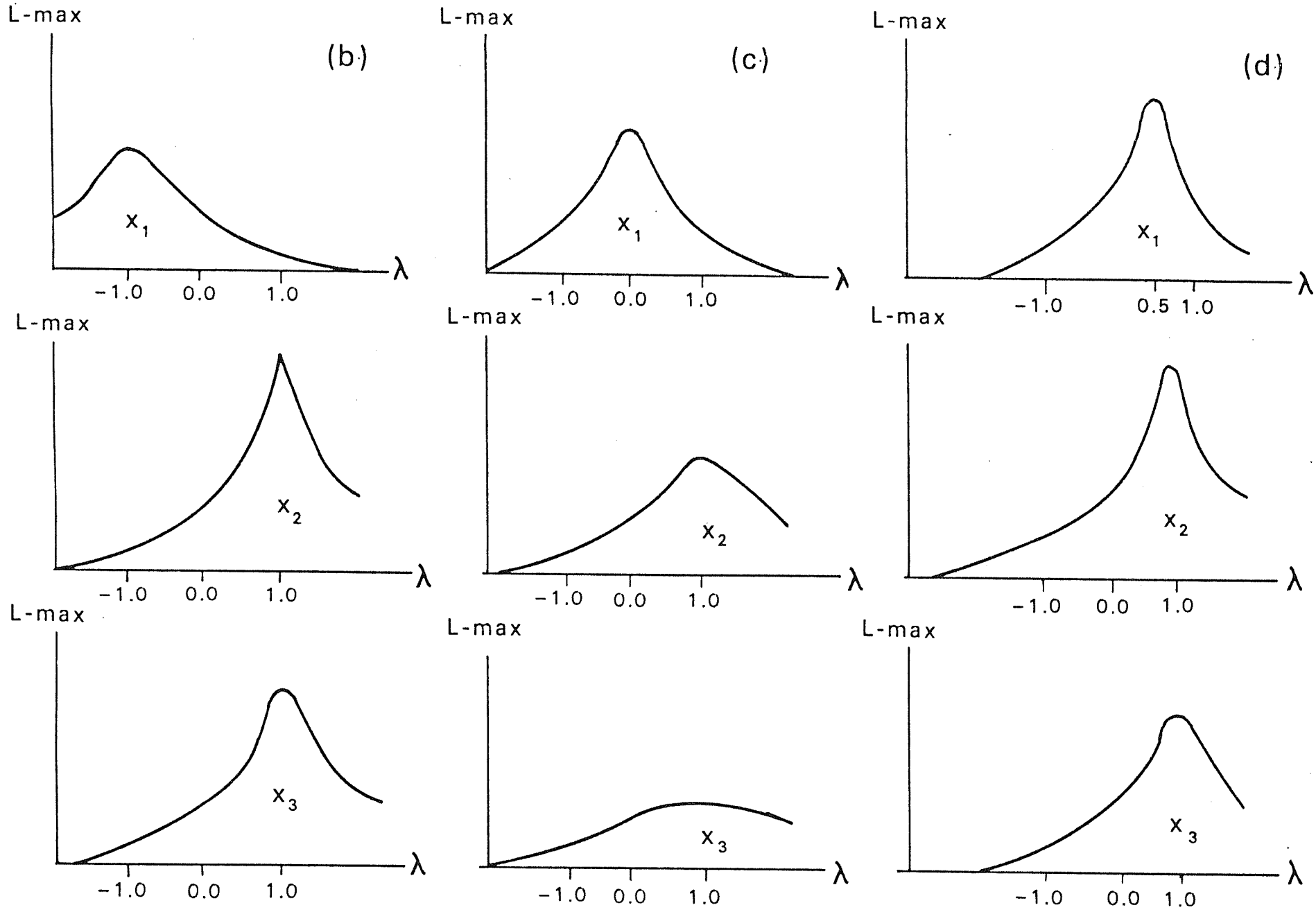


FIGURE 4.9 USING BOX-COX TRANSFORMATIONS AND PARTIAL RESIDUALS

of an appropriate transformation, but many geographers are reluctant to use transformations in their work. Their objections are usually based on two grounds: a feeling that transformed scales are somehow 'unnatural' and an unease that their use involves an unacceptable ad hoc element, if indeed it does not verge upon statistical cheating. Mitchell (1974, 522) has issued the following warning:

> 'transformations only rarely have a theoretical basis and must be used with great care if inferences are to be made about processes or patterns associated with spatial activities or man-environment relationships'.

Gould (1970, 442) has also been concerned about the uncritical use of transformations:

> 'Too often we end up relating the value of one variable to the log of another, with the square root of the third, the arc sin of a fourth, and the log of a fifth. Everything is normal, statistically significant at the one percent level - except that we have not the faintest idea what it means'.

Undoubtedly, such abuses of transformations do occur. Morgan (1976, 14), for example, in an attempt to explain the geographical pattern of residential differentiation uses seven independent variables. Population size, proportion Negro, proportion immigrants and population growth were made 'better approximate normal' by taking logarithms of their values. But the assumption of normality, as discussed above, only refers to the distribution of the disturbance term of a multiple regression model. Moreover, making each variable normal by 'logging' does not necessarily mean that the relationship between each explanatory variable and the dependent variable is transformed to an intrinsically linear one. Indeed, the true relationship between the variables may be a linear one and such an indiscriminate use of a transformation may lead to an incorrectly specified model.

Hopefully, in contrast to such work, the use of transformations as proposed here will lead to correctly specified

models. There are several arguments in favour of adopting such an approach. Firstly, as Kendall and Stuart (1966, 85) have contended, there is no reason why the quantity measured, rather than some function of it, should be best suited to the assumptions of the model. Secondly, non-linearity is not unusual or un-natural; many scientific relationships in the natural sciences take a non-linear form, and such relationships can often be more easily estimated after transformation. For example, Boyle's law in physics has the form:

$$P = {}^{C}/_{V} \qquad (12)$$

where P is the pressure of a gas, V is its volume and C is a constant which depends on its temperature and type. This relationship, as can be seen from Figure 4.10(a) is non-linear. However, it is possible to transform it to produce a linear form (Figure 4.10b):

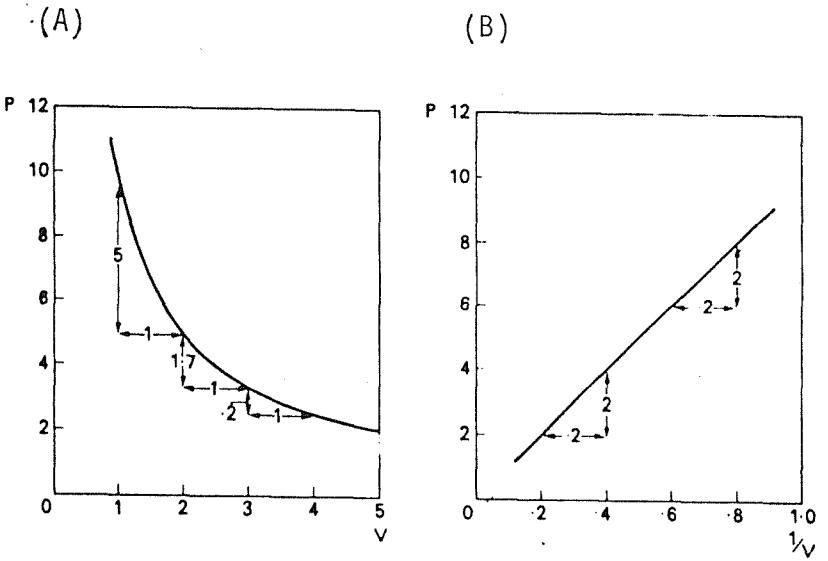
$$P = C. ({}^{1}/_{V}) \qquad (13)$$

While equation (12) cannot be estimated by linear methods, equation (13) represents a linear model and it is undoubtedly easier and simpler to perform a transformation rather than to use specifically non-linear methods of estimation. Thirdly, transformations are neither unusual nor un-natural. The often cited case of pH, a logarithmic transformation of hydrogen ion long accepted as a useful measure, serves as a reminder that 'naturalness' may reflect convention as much as reality. Fourthly, as argued by Evans, Catterall and Rhind (1975, 7)

> 'if a relationship between two variables is more meaningfully expressed after the use of two different transformations what further justification is required?'

Finally, when the choice of models and transformations is not based on theoretical reasoning, some authors, for example Curry (1967) have wondered whether the whole procedure is a 'game of numbers'. Curry himself has concluded that this quandary can only be resolved by the use of cross-validation

FIGURE 4.10 THE USE OF A TRANSFORMATION IN ESTIMATING BOYLE'S LAW



SOURCE: EHRENBURG (1975)

as first discussed in Chapter 1. Given the use of cross-validation, the judicious use of transformations appears to be an indispensable part of regression modelling.

In this discussion of functional form it has been shown that there is little in epidemiological theory to guide the choice of an exact functional relationship between variables. However, it is hoped that a procedure of fitting a simple linear model followed by careful examination of residuals coupled with cross-validation will aid the researcher in approximating the correct functional form. If researchers continue to use linear models without exploring non-linear relationships, there is no doubt that they will run the risk of drawing incorrect inferences from their work.

RAMSEY'S SPECIFICATION ERROR TESTS

So far the discussion has focused on exploratory and graphical methods of evaluating a model, but a confirmatory approach based on statistical tests is also available. In particular, Ramsey (1968) has proposed a set of general tests that can be used when more than one mis-specification is present in a model. Here the three types of error that can be detected by the Ramsey procedures will be outlined, the need for transformed residuals will be considered and, finally, the actual tests will be presented and illustrated.

The errors

Ramsey's procedures enables the analyst to distinguish between three groups of specification errors.

Group A — errors resulting in the disturbance term of a particular model having a non-zero mean. This can occur for a number of reasons: a model may have an incorrect functional form, a variable may have been omitted, or the model may need to be estimated as a simultaneous equation.

Group B — errors occurring when the model is heteroscedastic.

Group C errors occurring if the distribution of the disturbance term is not the same as that assumed in the model specification.

As with all analyses of model mis-specification, Ramsey's tests involve the examination of the model's disturbance term. The least-squares residual vector ($\hat{E}$) provides one approximation of the true disturbance vector, but such residuals have some undesirable characteristics in that, even when the disturbances are truly homoscedastic, the least-squares residuals are not.[12] More formally, although statistics based on OLS residuals are asymptotically unbiased and normally distributed, their co-variance matrix will, in general, differ substantially from that of the corresponding matrix based on the true, but unknown, disturbance terms. Moreover, this difference between the OLS residuals and the true disturbance term depends on the values of the explanatory variables in the model. Obviously, a new type of residual vector is required that has a scalar co-variance matrix when the true disturbance term exhibits such a property. The BLUS residuals form such a vector (BLUS stands for Best Linear Unbiased Scalar covariance matrix). Theil (1971, 202-213) defines this vector and describes its properties. BLUS residuals are defined in such a manner that they have a scalar co-variance matrix, whatever the values of the explanatory variables, provided that the fitted model contains no specification error.

For the Ramsey procedures the null hypothesis (H_o) is that the BLUS residuals are normally distributed with a mean of zero and a scalar co-variance matrix; H_o assumes, therefore, that the model is correctly specified. There are three alternative hypotheses (H_1, H_2 and H_3) and all three are expressed in terms of the distributional properties of the BLUS residuals. Models which give rise to Group A errors have BLUS residuals that are normally distributed with a scalar co-variance matrix and a non-zero mean (H_1). Group B errors result in BLUS residuals that are normally distributed with a mean of zero and a non-scalar co-variance matrix (H_2).

Group C errors result in BLUS residuals that are not normally distributed, but they have a scalar co-variance matrix and a mean of zero (H_3). A particular model may suffer from no specification error, one type of error or any combination of the three error types.

The tests

Ramsey (1974) has suggested three tests to examine the null and alternative hypotheses.[13]

1. Regression specification error test (RESET)

The BLUS residuals are regressed against a polynomial (usually of second or third order) of the BLUS transformed dependent variables. Under the null hypothesis of no specification error the regression coefficients of the estimated model should not be significantly different from zero. Thursby (1975) has carried out a simulation study of the effectiveness of this procedure to detect whether an explanatory variable has been omitted from the model. In general, he found Ramsey's test to be extremely powerful, but the ability of the procedure to detect mis-specification decreased as the importance of the omitted explanatory variable increased. This is unlikely to be a troublesome problem in empirical analysis, for it can be hoped that the most important variables will have already been included in the model.

2. Bartlett's M specification error test (BAMSET)

This test is designed to test the null hypothesis against the alternative H_2 and therefore it requires an examination of the co-variance structure of the BLUS residuals. A simulation study of the small-sample power of this test has been conducted by Ramsey and Gilbert (1972), who concluded that it was a reasonably powerful test against the null hypothesis of no specification error.

3. The Shapiro-Wilk test for normality (WSET)

To examine the disturbance term for Group C errors, Ramsey proposes the use of the Shapiro-Wilk test. This test can detect deviations from normality due to either skewness or kurtosis and is consequently superior to the commonly used tests such as the chi-square test. Shapiro, Wilk and Chen (1968) have shown that the small-sample power of the test compares favourably with a number of alternative tests. For a sample size greater than fifty, WSET needs to be replaced by the analogous D statistic of D'Agostino (1971).

Applications

The use of the Ramsey tests has apparently, been entirely confined to econometricians. Ramsey (1968) examined eight variations of an original model and, despite low sample sizes (most of the models had less than 20 observations) the results indicated that

> 'the tests were sensitive to the variance alternatives and the tests enabled one to select one model in preference to the others (Ramsey, 1974, 44).

Gilbert (1969) provides an example of the use of specification error tests to discriminate between a very large number of alternative model formulations for the demand for money. Other econometric applications of the tests are to be found in Lee (1972), Loeb (1973), Ramsey (1969, 1972) and Ramsey and Zarembka (1971).

While the tests do not appear to have been used by either geographers or epidemiologists, they have been used by an econometrician in the analysis of the effects of air pollution on mortality. Using socio-economic, demographic and pollution data for a sample of 50 Standard Metropolitan Statistical Areas in the USA, Smith (1976) postulated 36 different models to account for the variation in mortality rates amongst these areas. While several of the models were 'significant' using conventional tests (t and F), the null hypothesis of no specification error was rejected at the 0.10 probability level with at least one of the tests for all the models; according to the Ramsey tests all 36 models were mis-specified. In order to develop an improved model, Smith used a procedure developed by Park (1966) to accommodate a heteroscedastic disturbance term. He found his results 'somewhat unsettling' (p 77), for the final model did not show a strong relationship between air pollution and mortality. This result directly contradicts the conclusions that had been reached by a number of workers using

broadly similar SMSA data (Lave and Seskin, 1976; McDonald and Schwing, 1973; Koshal and Koshal, 1973) but these researchers had not attempted to detect specification errors in their model.

Against this background, it is instructive to apply the Ramsey procedures to Hart's (1970) model and West and Lowe's (1976) model (see Chapters 2 and 3). In both cases this evaluation results in the rejection of the null hypothesis of no specification error at the 0.05 level[14] and it seems, therefore, that the models are mis-specified. Thus, despite the conclusions reached by these authors, their results cannot be used to make valid inferences about the geographical variations of heart disease mortality.

Although advantages can be claimed for Ramsey tests, it must be acknowledged that they are firmly rooted in the tradition of confirmatory statistics; consequently, they enjoy the advantage of exactness and the disadvantage of inflexibility which is characteristic of such an approach. In particular, the distribution theory of the tests does not allow their repeated application to one body of data. (Smith's (1976) work, discussed above, broke this fundamental rule of confirmatory statistics by exploring 36 different models with one data set[15].) Moreover, having detected a mis-specification the tests give little indication of what action the analyst should take. The tests cannot distinguish between a model having an incorrect functional form and one in which an important explanatory variable has been omitted. Similarly, the tests can only detect the presence of hetero-scedasticity, they give little clue to the nature of the problem and how it may be overcome. However, given their effectiveness in detecting a mis-specified model, one can suggest that the tests can be used as a final hurdle that a model must pass before it is deemed acceptable and worthy of interpretation. A model should be fitted, developed and improved by exploratory methods, and then its appropriateness

should be tested by the Ramsey procedures. Such a suggestion is of course a variant on the method of 'cross-validation' considered in Chapter 1, and the application of this approach will be further considered in Chapter 7.

CONCLUSIONS

This discussion of specification errors has been extensive and involved with each error being treated as a separate entity to avoid confusion and repetition. Consequently, it is necessary not only to summarise the major points of the argument but also to relate significant aspects of each specification error to each other.

Outliers or discrepant observations can have a major effect on a regression model, but such values can easily be detected by a number of graphical plots; standardised residual plots, partial residual plots and probability plots will all show outliers as points away from the main body of the data. If a genuine mistake has been made the 'outlying' observation can be removed from the data; but if the outlier is a consequence of a mis-specified model (non-linearity or heteroscedasticity) other corrective action must be taken.

Heteroscedasticity results in the OLS estimates of a model having a high degree of imprecision. Heteroscedastic residuals have a distinctive wedge-shaped pattern when examined by residual plots, and commonly they will also appear to be non-normally distributed when plotted on normal probability graphs. While in a number of simple cases a model with non-constant variance can be estimated by ordinary least squares after a transformation has been performed, at present there is no suitable estimation procedure for models when the pattern of the heteroscedasticity is complex. Heteroscedasticity can also appear to be a problem in a model as a result of measurement error and the omission of an important explanatory variable.

The effect of omitting a variable depends on the relative importance of that variable and on the degree and pattern of its relationship to the variables already included in the model. One method of detecting an omitted variable is to map the residuals from the model, but it must be remembered that a model with an incorrect functional form may also display a distinctive map pattern. The model with an

important omitted variable can also be expected to have non-normal residuals that do not have a mean of zero.

Non-normality of residuals should not be a major problem in itself, but if a probability plot shows the residuals to be non-normal this may indicate that the model is heteroscedastic or has an incorrect functional form.

OLS estimates of a model with an incorrect functional form are unlikely to reflect the true population parameters, but it has been found that an automatic procedure originated by Box and Cox is, when used in conjunction with graphical procedures, an effective means of detecting and overcoming the problem of non-linear relationships.

General specification error tests offer a confirmatory approach to the detection of models with incorrect functional forms, heteroscedastic disturbance terms and disturbance terms that are not normally distributed. It is suggested that such tests are especially valuable when used in cross-validation.

In contrast to the rather despairing views of some geographers in response to the problem of specifying a correct model, it is hoped that the methods outlined in this chapter will aid researchers in their attempt to build an improved model. One lacuna, however, remains. What should the analyst do when his models appear to suffer from more than one specification error? The basic 'rule' of exploratory analysis is to deal with the most prominent form of mis-behaviour before dealing with other forms of mis-specification. For example, if a model is suspected of having a genuine outlier and a non-linear relationship, then the problems should be treated in that order. Similarly, Kruskal (1968) has ranked the benefits of transformation in the following order of importance:

(1) to simplify and make a relationship linear,
(2) to stabilise variance and
(3) to improve normality.

Fortunately, by happy coincidence, it appears that it is generally possible to find transformations which simultaneously satisfy, or nearly satisfy, the requirements of a model, provided that the analyst does not strive to achieve any one of them too rigorously (Kendal and Stuart, 1966). As Smith (1972) has pointed out, the aim of the analysis is to achieve approximate linearity, homogeneity and normality by means of a simple transformation.

Finally, it must be admitted that the methods outlined in this chapter have been used on artificial data. For example, the models in Table 4.4 which were used to demonstrate the usefulness of partial residual plots had only one variable with an incorrectly specified form. Moreover, all these models were generated to be orthogonal so that multicollinearity was not a problem. In real data, however, it can be anticipated that several mis-specifications are likely to occur, and more than one assumption of the regression model may be broken in any model. Clearly the confrontation of these problems is of great importance, and the use of procedures that have been outlined in this chapter on real data will, therefore form the subject of Part II of this thesis.

CHAPTER 4 : NOTES

1. All the simulations reported in this chapter were performed by the author on the ICL 1900 computer at Southampton University, using random number generators from the NAG library.

2. Moreover, as will be shown later, a simple graphical plot of the dependent variable against each of the explanatory variables of a model can often give a misleading picture of the overall model.

3. They are the same deleterious effects as for positive spatial autocorrelation.

4. Mather (1976, 83) has conducted a simulation experiment to demonstrate the effects of heteroscedasticity. He generated a model with 150 observations following the equation

$$y = 1.5 + 2x_1 + 3x_2 + \varepsilon$$

with the residual vector (ε) having a variance proportional to x_1. In thirty replications of this model he found the following results:

OLS estimates of a heteroscedastic model

True	Mean	Maximum	Minimum
1.5	1.97	14.78	-15.51
2.0	2.08	6.20	- 0.55
3.0	2.74	5.62	0.20

Although the mean values of the coefficients are close to the true values (for heteroscedasticity does not lead to bias) the spread of the estimates is very large. Of course, in any particular application of the OLS regression model where only one set of data is available, the property of no bias is of little comfort to the researcher.

5. This particular transformation is equivalent to the application of the method of weighted least squares (WLS). If in a heteroscedastic model the variance of the disturbance term increases with increasing values of an explanatory variable, the greater variance of the disturbances at high values of the explanatory variable gives a less accurate indication of where the true regression line lies than it does at low values of the explanatory variable. It therefore seems intuitively reasonable to assign less importance to 'high' disturbances in determining the regression coefficients. This can be achieved by assigning different weights to each disturbance term. If it is suspected that the disturbances are proportional to one explanatory variable x_1, the WLS estimates are obtained from minimising a weighted sum of squared residuals

$$\Sigma \frac{1}{x_1} \cdot e_i^2$$

instead of minimising the value

$$\Sigma\ e_i^2$$

as in OLS estimation. It can be shown, however, that WLS estimation yields identical results to a correctly transformed model estimated by ordinary least squares.

6. These plots were produced by using the statistical computer language GENSTAT and the CDC 7600 computer of the University of London Computer Centre.

7. Models involving a critical threshold will not be examined in detail here. They can, however, be detected by the graphical procedures outlined in this chapter and can be estimated by using 'dummy' variables (Johnston, 1972).

8. These residual plots were produced by a computer program developed by the author and are suitable for use with CALCOMP plotter software.

9. The final model (g) in Table 4.4 contains an interaction term ($x_1 \cdot x_2$). This concept of a dependent variable being related to the interaction of two explanatory variables is quite common in epidemiology. For example, many researchers believe that bronchitis is in part, the outcome of an interaction or synergism between smoking and air pollution. As is shown by the results presented in Table 4.4, if this synergism is not taken into account it is unlikely that the model will be correctly estimated by ordinary least squares. Figure 4.8(g) shows residual plots from this model, and comparison of these plots with those from other models reveals that the residuals of a model requiring an interaction term have a distinctive pattern. To examine such a pattern more fully, it is possible to make use of another type of graphical procedure. Standardised residuals can be plotted against the product of a pair of explanatory variables; if an interaction term is required the residual terms will be correlated with the product and there will be a trend in the data. To take account of such an interaction a new term, which is simply a product of two synergistic variables, has to be included in a reformulated model. For further details, together with a large number of examples, see Allison (1977).

10. The advantage of the transformation $(y^{\lambda}-1)/\lambda$ as compared with the simple power transformation (y^{λ}) is that the former is continuous when λ equals zero because, as this condition is approached, $\frac{y^{\lambda}-1}{\lambda}$ approaches the logarithm of y. When λ in the Box-Cox transformation is -1.0 y undergoes a reciprocal transformation. Similarly if λ equals 0.5, 2.0, 3.0, y undergoes a square root, squared and cubic transformation respectively. When λ equals 1 the dependent variable remains untransformed.

11. It is a basic rule of exploratory data analysis that the most fundamental problem is dealt with first. If an estimated model contains only one incorrect functional relationship it is possible, if the mis-specification is gross enough, that the Box-Cox procedure will indicate that several explanatory variables require transformation. However, by re-estimating a model with the gross mis-specification of one variable overcome by transformation, it may be possible to establish whether the other explanatory variables require transformation. Figure 4.9 was calculated by a modified version of Chang's (1977) program.

12. Such heteroscedasticity is usually not gross enough to affect residual plots but, because of rigid and demanding assumptions, it can make statistical tests inexact.

13. Ramsey (1968) originally proposed four tests to detect specification errors but, as a result of a later simulation experiment, he rejected one of the tests (Ramsey, 1974).

14. A program for performing the Ramsey tests is available from the Computer Institute for Social Science Research, Michigan State University, East Lansing, Michigan 48824.

15. In addition, the tests are not independent under the null hypothesis of no specification error, and the joint distribution is at present unknown.

CHAPTER 4 : BIBLIOGRAPHY

AHGREN, A. and WALBURG, H.J. (1975): Generalised regression analysis in Anick, D.J. and Walburg, H.J. (eds.) Introductory multi-variate analysis McCutchen, Berkeley.

AITKEN, A.C. (1935): On least squares and linear combinations of observations Proceedings of the Royal Society of Edinburgh 55, 42-48.

ALLISON, P.D. (1977): Testing for interaction in multiple regression American Journal of Sociology 83, 144-153.

ANDREWS, D.F. (1971): A note on the selection of data transformations Biometrika 58, 249-254.

ANDREWS, D.F. (1974): A robust method for multiple linear regression Technometrics 16, 523-531.

ANDREWS, D.F. and PREGIBON, D. (1978): Finding outliers that matter Journal of the Royal Statistical Society Series B 40, 85-93.

ANDREWS D.F. and TUKEY, J.W. (1973): Teletypewriter plots for data analysis can be fast : 6-line plots, including probability plots Applied Statistics 22, 192-202.

ANSCOMBE, F.J. (1967): Topics in the investigation of linear relations fitted by the method of least-squares Journal of the Royal Statistical Society Series B 29, 1-52.

ANSCOMBE, F.J. (1968): Statistical analysis, special problems of, I. Outliers International Encyclopaedia of the Social Sciences Volume 15, 178-182, MacMillan and Free Press, London.

ANSCOMBE, F.J. (1973): Graphs in statistical analysis. American Statistician 27, 17-21.

ANSCOMBE, F.J. and TUKEY, J.W. (1963): The examination and analysis of residuals. Technometrics 5, 141-160.

BARNARD, K.C. (1978): The residential geography of the elderly : a multiple-scale approach. unpublished Ph.d. thesis, University of Southampton.

BARTLETT, M. (1949): Fitting a straight line if both variables are subject to error. Biometrika 5, 207-242.

BEHNKEN, D.W. and DRAPER, N.R. (1972): Residuals and their variance patterns. Technometrics 14, 101-111.

BOHRNSTEDT, G.W. and CARTER, T.M. (1971): Robustness in regression analysis in Costner, H.L. (ed.) Sociological Methodology 118-146 Jossey-Bass, San Fransisco.

BONEAU, C.A. (1960): The effects of violations of assumptions underlying the t test. Psychological Bulletin 57, 49-64.

BOX, G.E.P. (1953): Non-normality and tests on variances. Biometrika 41, 318-335.

BOX, G.E.P. and COX, D.R. (1964): An analysis of transformations (with discussion). Journal of the Royal Statistical Society Series B 26, 211-243.

BOX, G.E.P. and TIDWELL, P.W. (1962): Transformation of the independent variables. Technometrics 4, 531-550.

CHANG, H.S. (1977): Functional forms and the demand for meat in the United States. The Review of Economics and Statistics 59, 355-359.

CHATTERJEE, S. and PRICE, B. (1977): Regression analysis by example. Wiley, New York.

CLIFF, A.D. and ORD, J.K. (1973): Spatial autocorrelation. Pion, London.

COCHRAN, W.G. (1970): Some effects of errors of measurement on multiple correlation
Journal of the American Statistical Association 65, 22-34.

COLE, R. (1969): Data errors and forecasting accuracy
in Mincer, J. (ed.)
Economic forecasts and expectations
Bureau of Economic Research, Washington.

COLLET, D. and LEWIS, T. (1976): The subjective nature of outlier regression procedures
Applied Statistics 25, 228-237.

COX, D.R. (1968): Notes on some aspects of regression analysis
Journal of the Royal Statistical Society Series A 131, 265-279.

CURRY, L. (1967): Quantitative geography, 1967
Canadian Geographer 11, 265-279.

D'AGOSTINO, R.B. (1971): An omnibus test for normality for moderate and large-size samples
Biometrika 58, 341-348.

DANIEL, C. and WOOD, F.S. (1971): Fitting equations to data
Wiley, New York.

DEEGAN, J. (1970): Problems in the examination of residuals
Unpublished Curriculum Development Project.

DEEGAN, J. (1972): The effects of multicollinearity and specification error on models of political behaviour
Unpublished Phd Thesis, University of Michigan, Michigan.

DEEGAN, J. (1974): Specification error in causal models
Social Science Research 3, 235-259.

DEEGAN, J. (1976): The consequences of model mis-specification in regression analysis
Multivariate Behavioural Research 11, 237-248.

DHRYMES, P.J., HOWREY, E.P., HYMANS, S.H., KMENTA, J., LEAMER, E.E., QUANDT, R.E., RAMSEY, J.B., SHAPIRO, H.J., and ZARNOWITZ, V. (1972): Criteria for evaluation of econometric models Annals of Economic and Social Measurement 1, 291-324.

DOLBY, J.L. (1963): A quick method for choosing a transformation Technometrics 5, 317-326.

DRAPER, N.R. and HUNTER, W.G. (1969): Transformations: some examples revisited Technometrics 11, 23-40.

DRAPER, N.R. and SMITH, H. (1966): Applied regression analysis Wiley, New York.

DURBIN, J. (1954): Errors in variables Review of the International Statistical Institute 22, 23-32.

DURBIN, J. (1970): Testing for serial correlation in least squares regression when some of the regressors are lagged dependent variables Econometrica 38, 410-421.

DYER, A.R. (1974): Comparison of tests for normality with a cautionary note Biometrika 61, 185-189.

EHRENBURG, A.S.C. (1975): Data reduction Wiley, London.

ELLENBERG, J.H. (1976): Testing for a single outlier from a general regression Biometrics 32, 637-645.

EVANS, T.S., CATTERALL, J.W. and RHIND, D.W. (1975): Specific transformations are necessary Census Research Unit Working Paper No. 2 Dept. of Geography, University of Durham.

EZEKIEL, M. and FOX, K.A. (1959): Methods of correlation and regression analysis 3rd edition Wiley, New York.

FEDER, P.I. (1974): Graphical techniques in statistical data analysis - tools for extracting information from data Technometrics 16, 287-299.

FERGUSON, T.S. (1961): Rules for rejection of outliers Revue de l'Institut International de Statistique 29, 29-43.

FISHER, F.M. (1966): A priori information and time-series analysis North-Holland, Amsterdam.

GALLANT, A.R. (1975): Non-linear regression American Statistician 29, 73-81.

GERSON, M. (1975): The techniques and uses of probability plotting The Statistician 24, 235-257.

GILBERT, R.F. (1969): The demand for money: an analysis of specification error unpublished Phd. Thesis, Michigan State University.

GLEJSER, H. (1969): A new test for heteroscedasticity Journal of the American Statistical Association 64, 316-323.

GNANADESIKAN, R. (1977): Methods for statistical data analysis of multivariate data analysis Wiley, New York.

GOLDFELD, S.M. and QUANDT, R.E. (1972): Non-linear models in econometrics North Holland, Amsterdam.

GORRINGE, P.A. (1971): Detecting heteroscedastic disturbances unpublished MA Thesis, Dept. of Economics, University of Manchester.

GOULD, P.R. (1970): Is statistix inferens the geographical name for a wild goose Economic Geography 46, 439-448.

GUDGIN, G. and THORNES, J.B. (1974): Probability in geographical research: applications and problems. The Statistician 23, 157-177.

HAGEN, G.H.L. (1837): Grandzuge der Wahrschein lichkeitsrechung Ernst and Korn, Berlin.

HAGGETT, P. CLIFF, A.D. and FREY, A. (1977): Locational analysis in human geography 2nd edition, Arnold, London.

HART, J.T. (1970): The distribution of mortality from coronary heart disease in S. Wales Journal of the Royal College of General Practitioners 19, 258-268.

HOCKING, R.R. (1972): Mis-specification in regression American Statistician 28, 39-40.

HODGES, S.D. and MOORE, P.G. (1972): Data uncertainties and least-squares regression Applied Statistics 21, 185-195.

HOGG, R.V. (1979): Statistical robustness: one view of its use in applications today American Statistician 33, 108-115.

HOYLE, M.H. (1973): Transformations - an introduction and a bibliography International Statistical Review 41, 203-223.

HUBER, P.J. (1973): Robust regression: asymptotics, conjectures and Monte Carlo Annals of Statistics 1, 799-821.

IRWIN, G.A. and MEETER, D.A. (1969): Building voter transition models from aggregate data Mid-West Journal of Political Science 13, 545-566.

JOHNSTON, J., (1972): Econometric methods, Mc-Graw Hill Kogakusha, Tokyo.

KATONA, G. (1954): Contributions of survey methods to econometrics Columbia University Press, New York.

KENDAL, M.G. and STUART, A., (1966): The advanced theory of statistics volume 3 Griffin, London.

KOSHAL, R.K. and KOSHAL, M. (1973): Environments and urban mortality - an econometric approach Environmental Pollution 4, 247-259.

KOUTSOYIANNIS, A. (1973): Theory of econometrics: an introductory exposition of econometric methods
MacMillan, London.

KRUSKAL, J.B. (1968): Statistical analysis, special problems II, Transformations
International Encyclopaedia of the Social Sciences
MacMillan and Free Press, volume 15, 182-193.

KRUSKAL, W.H. (1960): Some remarks on wild observations
Technometrics 2, 1-4.

LARSEN, W.A. and McCLEARY, S.J. (1972): The use of partial residual plots in regression analysis
Technometrics 14, 781-790.

LAVE, L.B. and SESKIN, E.P. (1976): Air pollution and human health
John Hopkins Press, Baltimore.

LEE, K.W. (1972): An international study of manufacturing production functions
unpublished Phd. Thesis, Michigan State University, Michigan.

LOEB, P.D. (1973): Specification errors and econometric models
unpublished Phd. Thesis
Rutgers University.

McCALLUM, B.T. (1972): Relative asymptotic bias from errors of omission and measurement
Econometrica 40, 757-758.

McDONALD, G.C. and SCHWING, R.C. (1973): Instabilities of regression estimates relating air pollution to mortality
Technometrics 15, 463-481.

MADDALA, G.S. (1977): Econometrics
Mc-Graw Hill, New York.

MALLOWS, C.L. (1970): A note on plotting partial residuals
Bell Telephone Laboratories Technical Memorandum.

MALLOWS, C.L. (1979): Robust methods - some examples of their use
American Statistician 33, 179-195.

MARQUADT, D.W. (1974): Discussion on Beaton, A.E. and Tukey, J.W.
Technometrics 16, 147-189.

MATHER, P.M. (1976): Computational methods of multivariate analysis in physical geography
Wiley, London.

MATHER, P.M. and OPENSHAW, S. (1974): Multivariate methods and geographical analysis
The Statistician 23, 283-308.

MITCHELL, B. (1974): Three approaches to resolving problems arising from assumption violation during statistical analysis in geographical research
Cahiers de Géographie de Québec 18, 507-523.

MORGAN, B.S. (1976): Social status segregation in comparative perspective: the case of the United Kingdom and United States
Dept. of Geography Occasional Paper No. 3.
King's College, London.

MOSTELLER, F. and TUKEY, J.W. (1977): Data analysis and regression
Addison Wesley, Mass.

NELDER, J.A. (1968): Regression, model-building and invariance
Journal of the Royal Statistical Society Series A 131, 303-315.

O'SULLIVAN, P.M. (1968): Accessibility and the spatial structure of the Irish economy
Regional Studies 2, 195-206.

OLSSON, G. (1970): Explanation, prediction and meaning variance: an assessment of distance interaction models
Economic Geography 46, 223-233.

OLSSON, G. (1975): Birds in egg
Michigan Geographical Publication No. 15, Dept. of Geography, University of Michigan, Ann Arbor.

PARK, R.E. (1966): Estimation with heteroscedastic error terms
Econometrica 34, 888.

RAMSEY, J.B. (1968): Tests for specification errors in classical linear least squares regression analysis unpublished Phd. Thesis, University of Wisconsin.

RAMSEY, J.B. (1969): Tests for specification errors in classical linear least squares analysis Journal of the Royal Statistical Society Series B 31, 350-371.

RAMSEY, J.B. (1972): Limiting functional forms for market demand curves Econometrica 40, 327-341.

RAMSEY, J.B. (1974): Classical model selection through specification error tests in Zarembka, P. (ed.) Frontiers in Econometrics, Academic Press, New York.

RAMSEY, J.B. and GILBERT, R. (1972): Some small sample properties of tests for specification error Journal of the American Statistical Association 67, 180-186.

RAMSEY, J.B. and ZAREMBKA, P. (1971): Specification error tests and alternative functional forms of the aggregate production function Journal of the American Statistical Association 66, 471-477.

RAO, P. (1971): Some notes on mis-specification in multiple regression American Statistician 25, 37-39.

REIERSOL, O. (1945): Confluence analysis of instrumental sets of variables Almquist and Wiksell, Stockholm.

RIDER, P.R. (1933): Criteria for the rejection of observations Washington University Studies No. 8, Washington.

ROSNER, B. (1975): On the detection of many outliers Technometrics 17, 221-229.

SCHLESSELMAN, J. (1971): Power families: a note on the Box and Cox transformation Journal of the Royal Statistical Society Series B 33, 307-311.

SEBER, G.A.F. (1977): Linear regression analysis
Wiley, New York.

SHAPIRO, S.S. and WILK, M.B. (1965): An analysis-of-variance test for normality
Biometrika 52, 591-611.

SHAPIRO, S.S., WILK, M.B. and CHEN, H.J. (1968): A comparative study of various tests for normality
Journal of the American Statistical Association 63, 1343-1372.

SMITH, J.H. (1972): Families of transformations for use in regression analysis
American Statistician 26, 59-61.

SMITH, V.K. (1976): The economic consequences of air pollution
Ballinger, Cambridge, Mass.

SOCKLOFF, A.L. (1976); Analysis of non-linearity via linear regression
Review of Educational Research 46, 267-291.

STIGLER, S.M. (1977): Do robust estimators work with real data?
Annals of Statistics 5, 1055-1098.

THEIL, H. (1971): Principles of econometrics
Wiley, New York.

THORNBY, J.I. (1972): A robust test for linear regression
Biometrics 28, 533-543.

THURSBY, P. (1975): Tests for omitted variables and incorrect functional form in regression analysis
unpublished Phd. Thesis, University of North Carolina.

TUKEY, J.W. (1957): On the comparative anatomy of transformations
Annals of Mathematical Statistics 28, 602-632.

TUKEY, J.W. (1968): The true purpose of transformations
Biometrics 24, 1041.

TUKEY, J.W. (1977): Exploratory data analysis
Addison-Wesley, Reading, Mass.

WALD, A. (1940): The fitting of straight lines if both variables are subject to error
Annals of Mathematical Statistics 11, 284-300.

WEST, R.R. and LOWE, C.R. (1976): Mortality from ischaemic heart disease - inter-town variation and its association with climate in England and Wales
International Journal of Epidemiology 5, 195-201.

WICKERS, M.R. (1972): A note on the use of proxy variables
Econometrica 40, 759-761.

WILK, M.B. and GNANADESIKAN, R. (1968): Probability plotting methods for the analysis of data
Biometrika 55, 1-17.

WITHEY, J.R. and COLLINS, B.T. (1976): A statistical assessment of the quantitative uptake of vinyl chloride monomer from aqueous solution
Journal of Taxicology and Environmental Health 2, 311-321.

WOOD, F.S. (1973): The use of individual effects and residuals in fitting equations to data
Technometrics 15, 677-695.

WOOD, J.T. (1974): An extension of the analysis of transformations of Box and Cox
Applied Statistics 23, 278-283.

ZAREMBKA, P. (1974): Transformation of variables in econometrics
in Zarembka, P. (ed.)
Frontiers in econometrics
Academic Press, London.

C H A P T E R 5

PERCENTAGES, RATIOS AND INBUILT RELATIONSHIPS

'the discovery of possible interrelations between diseases by an examination of their death rates .... has not been without fascination for more than one investigator. Personally, I have considered the problem more than once, but always failed to make progress owing to the existence of spurious correlations which I did not see how to meet'.

Pearson, Lee and Elderton (1910)

INTRODUCTION[1]

It is quite common in studies of areal mortality for one variable to be divided by another to eliminate the effect of the latter on the former. For example, deaths in a particular age group can be divided by the number of people in that age group to eliminate the effect of population size. A more complex ratio of this kind is the Standardized Mortality Ratio used in Howe's (1963) National Atlas of Disease. Technically, population size in this example is being used as a 'deflator' variable. While such ratios are undeniably useful in comparative studies, they induce a number of problems when used in multivariate analysis. As Chayes (1971, vii) has written:

> 'The process of ratio formation often imposes on its end products, interrelations very different from those characterizing the raw variables'.

Such problems should undoubtedly attract the attention of the geographical researcher. Indeed, for Unwin (1977, 189):

> 'perhaps the most interesting technical problem encountered in regression analysis in physical geography concerns the use of data which are proportions of some fixed total'.

Unfortunately, despite consideration of the problems inherent in the analysis of closed (percentage) and ratio data in a large number of disciplines, there has been little discussion of the problem in geography.[2] As Evans (1975, 195) has stated:

> 'Little attention is paid to the nature of data in human geography. Most such data are expressed in percentage or ratio form, yet human geographers have as yet paid little attention to the experiences of geologists in ... correlating ... these'.

This chapter will attempt to provide the much-needed, geographically oriented overview of the literature on the statistical analysis of closed and ratio data. This overview is an attempt to present this literature in as non-technical a manner as possible. To achieve this aim no proofs are given and the reasoning is intuitive rather than mathematical. The discussion will begin with the specific problem of correlations between percentages and will broaden to consider ratios, regression models and principal components analysis. Finally, a distinction will be drawn between reducible and irreducible ratios and it will be shown that this distinction has considerable importance for the analysis of areal mortality data.

PERCENTAGES, CLOSED DATA, CORRELATION AND INBUILT RELATIONSHIPS

Percentages of a fixed total cause considerable difficulties when used in the calculation of correlation coefficients. Table 5.1 shows raw and percentage data for the distribution of males and females in the wards of the Rhondda in 1971. As the row totals of the raw data sum to a different total for each ward the raw data are said to be 'open', while the percentage data, which sum to a fixed total for each ward, represent a 'closed' array. When the open data are used and the number of males is plotted against the number of females there is, as one would expect, a high positive relationship (Figure 5.1a). However, when the closed data are used a perfect negative relationship is formed (Figure 5.1b). In fact, when any two-category classification is adopted and closed data are used there will always be a perfect negative correlation between the values. This is what Evans (undated) has called the reductio ad absurdam of correlation with closed data.

In a three-way classification of closed data similar problems arise. For example, consider the division of the Gross Domestic Product of the EEC into three categories -

No.	WARD	RAW DATA (figures in hundreds)			PERCENTAGE DATA	
		Total	Male	Female	Male	Female
1	Ferndale	104	50	54	48.1	51.9
2	Llwynypia	77	37	40	47.9	52.1
3	Pentre	57	27	30	47.9	52.1
4	Penygraig	76	37	39	48.5	51.5
5	Porth	109	53	56	48.6	51.4
6	Trealaw	82	39	43	47.7	52.3
7	Treherbert	84	41	43	48.7	51.3
8	Treorchy	84	40	44	47.6	52.4
9	Tylorstown	86	42	44	49.0	51.0
10	Ynyshir	63	30	33	47.9	52.1
11	Ystrad	68	33	35	49.0	51.0

TABLE 5.1 Male and female populations of the Rhondda Wards (1971)

Source: Registrar General (1972), 1971 Census of England and Wales: County Report – Glamorgan, HMSO

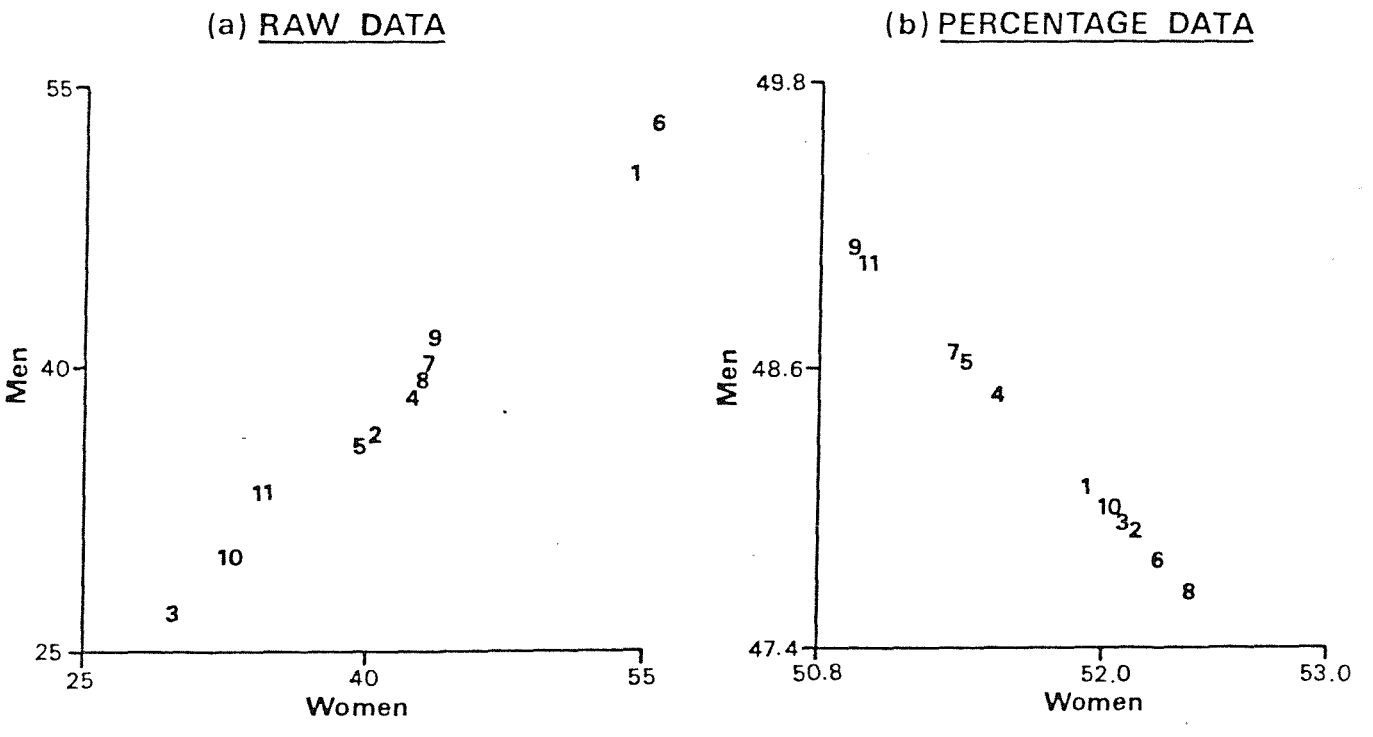


FIGURE 5.1 Relationships between male and female populations of the Rhondda wards (1971)

agriculture, mining plus manufacturing and 'other' activities (Table 5.2). In such a three-way classification, once any two variables are given the third is fixed. If 14% of the GDP is produced by agriculture and 36% is produced by mining and manufacturing, then 50% of the GDP must be accounted for by the remaining category. Such a set of percentage data is called a ternary (three-variable) closed array. One property of such an array is that even when all three of the correlations between the original raw variables are strongly positive, at least two of the three correlations of their ternary equivalents will be negative. In this particular case, if each member country of the EEC is classified according to a three-fold division of sources of GDP, positive correlations are found for the open data, while negative correlations are found between the closed variables (Figure 5.2). In general, providing the variances of the variables being correlated are large, negative correlations are more likely with a three-way classification and closed data than are positive relationships. 'True' positive correlations will be weakened and perhaps even turned into an apparently negative correlation. It is this propensity for pre-determined or inbuilt relationships, termed 'spurious' correlations by Pearson (1897), that is at the core of the closed data problem.

Further understanding of this problem can be gained from considering the calculation of the correlation between two variables (A,B) of a closed array consisting of three variables (A,B and C). The Pearson product-moment correlation coefficient between A and B (r_{AB}) can be calculated from:

$$r_{AB} = \frac{S_C^2 - S_A^2 + S_B^2}{2S_A S_B} \qquad (1)$$

where S_A is the standard deviation of the closed variable A. It is apparent from this equation that although the correlation only involves variables A and B, the correlation coefficient's

TABLE 5.2

Gross Domestic Product in the EEC (1973)

No.	Country	G.D.P. Total	Source of G.D.P.					
			Raw data*			Percentage data		
			Agric.	Mining + Manufact-uring	Other	Agric.	Mining + Manufact-uring	Other
1	United Kingdom	153	5	47	101	3	31	66
2	France	198	12	71	115	6	36	58
3	Belgium	35	1	12	22	4	34	62
4	Luxembourg	2	0	1	1	4	43	53
5	W. Germany	351	11	154	186	3	44	53
6	Italy	138	11	47	80	8	34	58
7	Irish Republic	6	1	2	3	15	29	56
8	Netherlands	60	3	10	38	5	31	64
9	Denmark	28	2	8	18	8	28	64

* Thousand million dollars

Source: Geographical Digest (1976) George Phillip and Son, London, 66-69

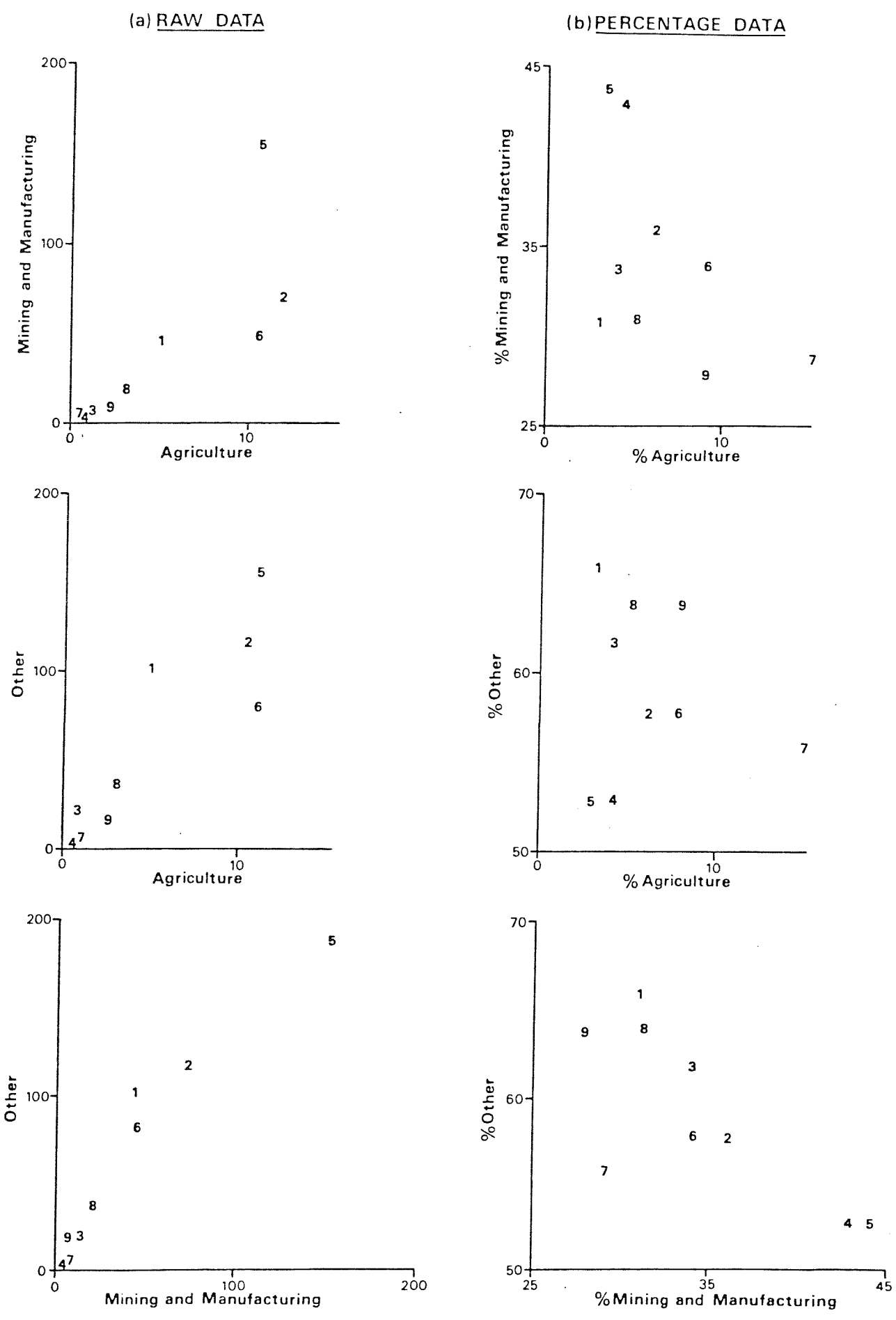


FIGURE 5.2 Relationships between sources of Gross Domestic Product in the EEC (1973)

magnitude and sign is in part determined by the variance of variable C. If variables A and B have a large spread or variance then the correlation between A and B is likely to be negative, even when A and B are random variables. Similarly, if the variances of A and B are relatively small a positive relationship is to be expected.

From the above discussion it can be seen that increasing the number of categories from two to three has decreased the relative importance of the inbuilt relationship - the perfect negative correlation of the two-way classification of closed data is no longer a foregone conclusion with a ternary closed array. This is a general result. As Chayes (1971) has shown, if one assumes that each of the categories of a classification has equal variance, the 'spurious' correlation (ρ) when the true correlation is zero can be calculated from the following formula:

$$\rho = \frac{-1}{m-1} \qquad (2)$$

where m is the number of categories. For example, if there are twenty categories each with the same variance and the data are in percentage form, the expected or inbuilt correlation between each pair of variables is:

$$\frac{-1}{m-1} = \frac{-1}{19} \simeq -0.05 \qquad (3)$$

This formula can also be useful when assessing the importance of a correlation between closed variables. When evaluating the importance of the correlation between two samples, the usual null hypothesis is that the population correlation between the two variables is zero. But while this is correct for freely varying data, this is obviously not the case for closed data. A correlation of -0.1 between two of a set of four equally variable percentages may mean a positive effect, since it should be compared not with a null expectation of zero but with a null expectation of -0.33 (that is -1/(4-1)). It is this procedure of calculating the

expected inbuilt relationship, and using this value as the null hypothesis that has characterized the work of geologists examining this problem. Yet, in practice, this procedure is complicated by the assumption of equal variance amongst the variables, for such equal variance is not commonly encountered in real data. Moreover, even if only two classes of a set of classes are being compared, the variances of all the classes need to be equal if the expected inbuilt relationship is to be calculated by equation (2).

Considerable research has been conducted on how to proceed when the variances are unequal but, as yet, the results have been rather disappointing. A discussion of this research is given in Jones (1978). The mechanics of the various procedures that have been developed will not be discussed here, since it is sufficient for the purpose in hand to know that the use of percentage data leads to inbuilt relationships.

RATIOS, CORRELATION AND INBUILT RELATIONSHIPS

An inbuilt relationship can also occur with ratios, providing that either the numerator or denominator of one ratio is related to the numerator or denominator of another ratio.[3] Table 5.3 shows the types of relationship that can occur between ratios and gives approximate formulae for calculating the null correlation.[4] These formulae are easy to use providing the means and variances of the original variables are known. For example, if each mean ($\bar{X}$, $\bar{Y}$, $\bar{Z}$) of the original variables is one, and if the standard deviation of the three variables is also one ($s_x = s_y = s_z = 1$) the null correlation (ρ) between two ratios with a common denominator is

$$\rho = \frac{1.0 * 1.0 * 1.0}{(1.0 + 1.0) * (1.0 + 1.0)}$$

$$\rho = \frac{1}{4} = 0.5 \qquad (4)$$

Table 5.3

Inbuilt Correlations With Ratio Variables

Type of Relationship	Variables	Approximate null correlation	Estimated null correlation *
Correlation of a ratio with its numerator	$\frac{X}{Y}$; X	$\frac{\bar{Y}s_x}{\sqrt{(\bar{Y}^2s_x^2 + \bar{X}^2s_y^2)}}$	+0.71
Correlation of a ratio with its denominator	$\frac{X}{Y}$; Y	$\frac{-\bar{X}s_y}{\sqrt{(\bar{Y}^2s_x^2 + \bar{X}^2s_y^2)}}$	-0.71
Correlation between two ratios with a common numerator	$\frac{X}{Y}$; $\frac{Y}{Z}$	$\frac{\bar{X}\ \bar{Z}s_y^2}{\sqrt{(\bar{Y}^2s_x^2 + \bar{X}^2s_y^2)\ \ (\bar{Y}^2s_z^2 + \bar{Z}^2s_y^2)}}$	+0.5
Correlation between two ratios with a common denominator	$\frac{X}{Y}$; $\frac{Z}{Y}$	$\frac{\bar{X}\ \bar{Z}s_y^2}{\sqrt{(\bar{Y}^2s_x^2 + \bar{X}^2s_y^2)\ \ (\bar{Y}^2s_z^2 + \bar{Z}^2s_y^2)}}$	+0.5
Correlation between two ratios, the denominator of one being the numerator of the other	$\frac{X}{Y}$; $\frac{Y}{Z}$	$\frac{-\bar{X}\ \bar{Z}s_y^2}{\sqrt{(\bar{Y}^2s_x^2 + \bar{X}^2s_y^2)\ \ (\bar{Y}^2s_z^2 + \bar{Z}^2s_y^2)}}$	-0.5

* Assuming that each variable has a mean ($\bar{}$) and a standard deviation (s) of one

Source : adapted from Chayes (1971) pp 14-16

Other expected null correlations when the means of the variables and the standard deviations are one are also given in Table 5.3.

While the formulae in Table 5.3 are undoubtedly useful it must be remembered that they are only approximations. They are good approximations providing that the coefficients of variation (the standard deviation/mean) for the variables are small. Chayes (1971) in a Monte-Carlo experiment showed that when the coefficient of variation of a variable is greater than 0.5 care must be exercised in the use of the approximate formulae. If this is the case, probably the wisest decision is to proceed as follows.

1. Determine the mean and standard deviation of the component variables of the ratios, and generate a large number of sets of values for each raw variable having their means and variances, and in such a manner that the variables are uncorrelated.

2. Form the desired ratios and calculate the average correlation for the large number of sets of data that have been generated. This average correlation is the inbuilt relationship or null correlation and the observed correlation value of the ratios can be evaluated against it.

Using modern computers this procedure can be carried out quickly and easily (Chayes, 1973).

Atchley, Gaskins and Anderson (1976) have carried out such a simulation experiment. They examined the sample correlations between two uncorrelated variables X_1 and X_2 after they had been 'deflated' by a third uncorrelated variable X_3. They found that when the coefficients of variation of the three variables were equal, the inbuilt correlation between the two ratio variables was 0.5. That is the value expected from the Chayes formula and given in Table 5.3. However, when the coefficient of variation of the numerator variable was set to be small in comparison to the coefficient of variation of the denominator, the correlation between the two ratio variables exceeded 0.99, even when the

raw variables were independent. Obviously, in such a situation, inferences concerning the magnitude and sign of the correlation coefficient have to be drawn with great care. This statement applies with equal force regardless of whether the analysis is in the statistical or intuitive, graphical sense.

A rather complicated example of the correlation analysis of ratio variables is Galle, Grove and McPherson's (1972) study of the effects of living density on mortality. This study attempted to investigate the following model[5]:

$\log\left(\frac{P}{A}\right)$ is dependent on $\log\left(\frac{P}{R}\right)$, $\log\left(\frac{R}{H}\right)$, $\log\left(\frac{H}{S}\right)$ and $\log\left(\frac{S}{A}\right)$

where P = population
A = area
R = rooms
H = housing units
S = housing structures

There are obviously several ratios in this model which have common numerators or denominators and the following relationships were found between the dependent and each of the independent variables[6].

$r_{01} = 0.15$

$r_{02} = -0.56$

$r_{03} = 0.74$

$r_{04} = 0.72$

For convenience the dependent variable is termed 0, and the independent variables of the model are numbered sequentially from left to right, 1 to 4. Following Schuessler (1973) the positive relationships between the dependent variable (0) and the independent variables 1 and 4 are to be expected, since they have respectively a common numerator and a common denominator. Whilst the dependent variable and independent

variable 2 have no common term, variables 1 and 2 can be expected to have a negative relationship (the numerator of one ratio is the denominator of the other) and as r_{01} is positive, r_{02} will tend to be negative. This expectation is confirmed. Similarly, on the expectation that both r_{03} and r_{02} are negative, r_{03} can be expected to be positive. Again this prediction is confirmed. This is not to argue that the conclusions drawn from this study are incorrect, but such a possibility should be considered in the light of these inbuilt relationships[7].

RATIO DATA AND REGRESSION ANALYSIS

The discussion so far has centred around ratio variables and correlation coefficients. Far less explicit work has been done on regression coefficients, yet it seems reasonable to expect the same problem of inbuilt relationships. The problem can perhaps be most clearly seen when one considers three variables X, Y and Z. The relationship of X on Y is required, but the relationship is thought to be complicated by the variable Z. For example, if the relation of social class (X) to deaths (Y) is required the analyst may wish to control for the number of people in each area (Z). So the regression equation relating the number of people of low social class per thousand population to the number of deaths per thousand population is calculated. The problem is that relating X/Z to Y/Z is clearly, in part, relating the reciprocal of Z to itself. If all the values of X, Y and Z are positive and there is no relationship between X and Y then, whenever Z is relatively small, both X/Z and Y/Z will tend to be large. Moreover, in attempting to allow for the effect of Z by forming the ratios Y/Z and X/Z the analyst is implicitly assuming that the relationship between Y and Z, and X and Z is linear, and that when Z is zero Y and X are also zero. This, of course, need not necessarily be the case. Finally, in the ordinary least-squares regression model it is assumed that the variance of the error term is constant. But

Table 5.4

Inbuilt Regression Values With Ratio Variables

Type of Relationship	Variables	Approximate Null Correlation: First variable regressed on second	Approximate Null Correlation: Second variable regressed on first	Estimated null values*	
Regression of a ratio with its numerator	$\frac{X}{Y}$; X	$\frac{1}{\bar{Y}}$	$\frac{\bar{Y}^3 s_x^2}{\bar{Y}^2 s_x^2 + \bar{X}^2 s_y^2}$	1.0	0.5
Regression of a ratio with its denominator	$\frac{X}{Y}$; Y	$\frac{-\bar{X}}{\bar{Y}^2}$	$\frac{-\bar{X}\bar{Y}^2 s_y^2}{\bar{Y}^2 s_x^2 + \bar{X}^2 s_y^2}$	-1.0	-0.5
Regression between two ratios with a common numerator	$\frac{Y}{X}$; $\frac{Y}{Z}$	$\frac{\bar{Z}^2 s_y^2}{\bar{X}\ (\bar{Z}^2 s_y^2 + \bar{Y}^2 s_z^2)}$	$\frac{\bar{X}^3 s_y^2}{\bar{Z}\ (\bar{X}^2 s_y^2 + \bar{Y}^2 s_x^2)}$	0.5	0.5
Regression between two ratios with a common denominator	$\frac{X}{Y}$; $\frac{Z}{Y}$	$\frac{\bar{X}\bar{Z} s_y^2}{\bar{Y}^2 s_z^2 + \bar{Z}^2 s_y^2}$	$\frac{\bar{X}\bar{Z} s_y^2}{\bar{Y}^2 s_x^2 + \bar{X}^2 s_y^2}$	0.5	0.5
Regression between two ratios, the denominator of one being the numerator of the other	$\frac{X}{Y}$; $\frac{Y}{Z}$	$\frac{-\bar{X}\bar{Z}^3 s_y^2}{\bar{Y}^2 (\bar{Z}^2 s_y^2 + \bar{Y}^2 s_z^2)}$	$\frac{-\bar{X}\bar{Y}^2 s_y^2}{\bar{Z}\ (\bar{Y}^2 s_x^2 + \bar{X}^2 s_y^2)}$	-0.5	-0.5

* Assuming that each variable has a mean and a standard deviation of one

Source : adapted from Chayes (1971) pp 60-62

if this homoscedasticity assumption is met in the raw data, then it will generally not be met in the regression model with ratio data. The model will suffer from the problem of heteroscedasticity and the OLS estimates will lose their minimum variance property (Chapter Four)[3].

Table 5.4 lists Chayes' approximate formulae for calculating the value of the pre-determined regression coefficients. When the coefficient of variation of the variables are considered to be too high for the approximation to be accurate, a Monte-Carlo simulation can be performed in a similar fashion to that described previously for the analysis of ratio correlations.

An example of the use of ratios in regression analysis is the study by Schwirian and La Greca (1971). They analyzed the effects of a number of environmental, social and demographic variables on the crude mortality rate of 73 census tracts in Columbus, Ohio. It can be clearly seen from Table 5.5 that their analysis is a web of inbuilt relationships. (One might also suspect that the variables are highly multicollinear.) However, Schwirian and La Greca did not discuss the problems of ratio data, and even the subsequent criticism and extension of the analysis failed to consider the inherent difficulties of the regression analysis of ratio data (Poston, 1974; Schwirian and La Greca, 1974). (This example will be reconsidered when some guidelines are proposed for the regression analysis of ratio data.)

PRINCIPAL COMPONENTS, RATIOS, CLOSED DATA AND INBUILT RELATIONSHIPS

A principal components analysis (PCA) of a set of data can either be based on the correlation coefficients of the observed data or on the variance-covariance matrix. If the correlation coefficients are used with closed or ratio data their values will be to some extent pre-determined. Similarly, a PCA on such data will result in the component structure

Table 5.5

Variables used by Schwirian and La Greca (1971)

Dependent variable :	total number of deaths	/ total population
Independent variables :	number of blacks	/ total population
	number of foreign born	/ total population
	number of migrants	/ total population
	number of people aged 65 or over	/ total population
	total population	/ total number of households
	number of owner occupied houses	/ total number of households
	number of 'sound' houses	/ total number of households
	number of 'recent' houses	/ total number of households
	number of 'overcrowded' houses	/ total number of households
	number of unemployed males	/ total number of males
	number of professional males	/ total number of males
	median number of persons to all housing units	
	median school years of adults	
	median family income	
	median age of all males	
	median age of all females	
	median value of all owner occupied units	

being partly pre-determined. In an analagous fashion the components based on the variance-covariance matrix will also be artificial. For example, if closed data are analysed each row of the variance-covariance matrix must always sum to zero. Therefore, the relationship between two of a set of closed variables must be influenced by the other closed variable and, as Johnston (1977, 36) has written, 'since we cannot separate out this part, we cannot interpret .... the component loadings properly'.[9]

Atchley, Gaskins and Anderson (1976) have compared the results of a PCA on raw and ratio data. They generated a correlation matrix of six variables which contained highly significant positive and negative relationships as well as independent variables. Five of the variables were then divided by the sixth variable and both the five 'raw' and the five ratio variables were subjected to a PCA. The 'deflator' variable was chosen so as to be highly correlated with the other five variables, and in such a manner that it had a greater standard deviation than the other variables. It was found that the results of a PCA on the raw and ratio data differed greatly; there were large changes in the magnitude and sign of all the coefficients on the various components and, for the ratio data, the importance of the first axis was greatly increased.

Table 5.6 shows the results of a PCA on housing variables. The input data consisted of correlation coefficients for both raw and ratio data for each housing variable for the 83 County Boroughs of England and Wales as constituted in 1961. To obtain the ratio data, the raw housing variables have been deflated by the total number of households. Comparing the two analyses it is apparent that the results are very different. Over half the component loadings have changed sign. With the raw data all the variables contribute equally to the first component, which accounts for 90 per cent of the total variation of the original data, but with the ratio data the

Table 5.6

Principal components analysis of the correlation matrix of raw and ratio data

Raw data

Variable	Principal Components I	II	III	IV	V	VI
Shared households	-.4	.9	-.2	-.1	.0	.0
Households with 1½ persons per room	-.4	.1	.8	.4	.1	.0
Households with exclusive use of hot water	-.4	-.3	-.3	.2	.3	-.7
Households with exclusive use of fixed bath	-.4	-.2	-.3	.1	.4	.7
Households with exclusive use of W.C.	-.4	-.2	-.2	.2	-.8	.1
Local Authority housing	-.4	-.2	.2	-.8	.0	-.1
% of variance accounted for	90	6	3	1	0	0

Ratio data

Variable/Total Households	Principal Components I	II	III	IV	V	VI
Shared households	-.2	.6	-.4	-.5	-.5	-.2
Households with 1½ persons per room	-.4	-.3	-.5	-.5	-.5	.0
Households with exclusive use of hot water	.6	.1	-.3	.2	-.3	-.6
Households with exclusive use of fixed bath	.6	.0	-.4	-.2	.0	.7
Households with exclusive use of W.C.	.4	-.2	.5	-.7	.1	-.2
Local Authority housing	.0	-.7	-.3	.1	.6	-.2
% of variance accounted for	35	23	20	12	8	3

Source : Registrar General (1962) 1961 Census of England and Wales : County Reports H.M.S.O.

first principal component accounts for much less of the total variation. Furthermore, it is highly likely that the two sets of components would be given two different substantive interpretations by a researcher.

In conclusion, the results produced by analysts who have performed a PCA, regression and correlation analysis on ratio data without taking into account inbuilt relationships must be regarded with caution. If these analyses had been performed on raw data the interpretations would probably differ. A major question is, therefore, which form should the data take: raw or ratio? The next section is an attempt to answer this question in terms of regression analysis.

REDUCIBLE AND IRREDUCIBLE RATIOS AND GUIDELINES FOR RESEARCH

The problems of inbuilt relationships when using ratio variables in a statistical analysis raise the issue of what alternative procedures can be used. Obviously, the effect of one variable may be 'controlled' by including that variable as an explanatory variable in the regression model as well as by deflation. Both procedures yield measures that have been corrected for the effect of a disturbing variable; rarely, however, will the measures give equivalent results. According to Tufte (1969, 652) there appears little to recommend the analysis of ratios:

> 'Ratio correlations can be misleading, and the questions they are designed to answer can be more usefully framed as regression problems'.

Moreover, as Schuessler (1974, 395) has pointed out

> 'deviations from regression have a practical advantage, since they are more amenable to statistical analysis than ratios'

and, as Evans (undated) has shown, regression analysis is a much more flexible method than the use of deflation.[10]

However, while regression analysis is generally more flexible than deflation, the choice between the two methods must be dictated by the problem in hand, not technical

considerations. As Yule pointed out in 1910, if the hypothesis under consideration concerns the component parts of the ratios, a conclusion based on the ratio correlations will be incorrect. But, if the hypothesis concerns the ratios per se a conclusion based on the correlation of ratios will be appropriate. (See also Kuh and Meyer, 1955; Schuessler, 1974.) As Maddala (1977, 267) has written

> 'if our interest is in fact the relationship between X/Z and Y/Z, there is no reason why this correlation need be called spurious'.

The central issue is therefore the conceptual status of the ratios, that is whether the particular variables are irreducible ratios (for example, population density) or whether a ratio has been computed simply to remove the effect of a particular variable. In the latter case it is the raw data which are of crucial importance, whereas in the former case it is the combination of the numerator and denominator together which has to be considered. Unfortunately in many geographical applications researchers seem to be using ratios simply to remove the effects of a certain variable; this practice can obviously lead to inferential error.

This crucial distinction between reducible and irreducible ratios also has important repercussions for those researchers who have attempted to develop approximate formulae for calculating inbuilt correlations between ratios. Such work can be seen as an attempt to consider hypotheses about component parts of ratios when only ratios are used in the analysis. Surely, when the hypothesis is not in terms of the ratios, a regression analysis based on the raw variables is to be preferred to calculating deflated variables, performing an analysis on these ratios and interpreting the results in terms of the expected null correlation? Moreover, when the hypothesis is in terms of the ratios per se the approximate formulae have no role whatsoever to play in the analysis. Therefore, the only remaining use for the methods of calculating expected inbuilt relationships that have been extensively

considered in this chapter, is when the hypothesis is in terms of the component parts of the ratios but the data are only available in ratio form.

If one accepts the above arguments, three research guidelines can be proposed; they are illustrated in terms of the study of Schwirian and La Greca (1971).

1. Specify, in advance, whether the hypothesis is in terms of the ratio or component variables. Considering the variables listed in Table 5.5, total population, for example, has been used to deflate a number of variables, not to create an important ratio but simply to control for population size. Because of this, the hypothesis should not be in terms of the ratios.

2. If the hypothesis is in terms of the ratios then include the ratios in the model. With regard to Schwirian and La Greca's study, only one variable appears to be irreducible - household density (total population / total number of households.) Following Calhoun's (1962) study of the effects of high living densities on rats, a number of investigations have studied the effect of this variable on human mortality. Attention does not focus on the component parts of the ratio, so the ratio itself should be included in the regression model.

3. If the hypothesis is in terms of raw variables, include the particular variables of interest in the model. Of the variables listed in Table 5.5 only household density should remain as a ratio variable. With regard to the other ratio variables, the numerator variable may be included in the regression model and both total population and total number of households should also be included as 'control' variables.

The use of these guidelines in empirical research will be illustrated in the second part of this thesis.

CONCLUSIONS

Analysing ratio data visually and statistically, without considering the problem of inbuilt relationships, may lead the unwitting investigator to accept as meaningful a relationship that exhibits only random variation, and to reject as random a relationship that truly departs from randomness. Geographers and epidemiologists have in the past failed to consider inbuilt relationships and the fundamental difference between irreducible ratios and the inclusion of a ratio merely to control for a particular variable. Unfortunately, this failure may be critical for the meaningful interpretation of their analyses.

CHAPTER 5 : NOTES

1. This chapter is a modified version of Jones (1978).

2. Evans (undated) has produced an undergraduate practical handout on the subject and Norcliffe (1977) gives a brief textbook introduction to the problems of analysing closed data. The problems of using factor and principal components analysis on such data have been alluded to by Johnston (1976) and are more fully considered by the same author in 1977.

   Wrigley's (1973) discussion is specifically concerned with the difficulties of analysing a dependent variable that is a percentage and will not be considered here. The mapping of ratio variables has recently been discussed in two publications of the Census Research Unit at the University of Durham (Visvalingham, 1976; Visvalingham and Dewdney, 1977.)

3. The use of the word 'related' is deliberate; the problem of spurious correlation occurs not only when the same variables are used to form the ratios but also when either the denominators or numerators are highly related to each other. Rangarajan and Chatterjee (1969) consider the appropriate null value for

$$\frac{X_1}{X_2} \quad ; \quad \frac{X_2}{X_4}$$

   where X_2 and X_4 are not the same variable but are highly related.

4. Benson (1965, 37, 39) gives a more complete table, illustrating a wider range of possible ratios that may occur together with approximations for the extent of their inbuilt correlation.

5. If a logarithmic transformation of a ratio is performed, in effect one is replacing ratios by differences. For example, if

$$A = \log X$$
$$B = \log Y$$
$$\text{and } C = \log (X/Y)$$
$$\text{then } C = A - B$$

6. This model is an example of a 'chain relative' where the numerator of one ratio is the denominator of another (Meifeld, 1927)

7. As will be argued later, Galle, Grove and McPherson's study should not be criticised for failing to study inbuilt relationships, for the underlying hypothesis of the model clearly relates to the ratios not the component parts of the ratios.

8. When the residuals from a regression model are heteroscedastic, deflation is commonly used as a means of overcoming the problem. This can certainly be a valid and useful procedure, providing that the analyst makes inferences, not from the deflated relationship, but from the results transformed back into the form of the original model.

9. Butler (1976) has suggested that it is possible to assess the effects of inbuilt relationships by comparing the results of a PCA performed on ratio data with those obtained from a PCA on the 'spurious' correlations that can be predicted from Chayes' and Kruskal's (1966, 1970) procedure. The most reasonable interpretation of the PCA on the 'spurious' correlations is that these are the results that would have occured in the absence of any meaningful departure from randomness between any pair of variables; that is, these principal components are entirely a result of the data taking a closed form. However, Butler (1976, 35) himself admits that 'further work in this area is definitely warranted'.

10. While a ratio allows for the numerator varying with the denominator in strict proportion according to a linear relationship passing through the origin, regression can accommodate non-linear relationships that do or do not pass through the origin (Evans, undated).

CHAPTER 5 : BIBLIOGRAPHY

A more detailed bibliography on the subject of percentages, ratios and inbuilt relationships is given in Jones (1978).

ATCHLEY, W.R., GASKINS, C.T. and ANDERSON, D. (1976): Statistical properties of ratios I. empirical results Systematic Zoology 25, 137-148.

BENSON, M.A. (1965): Spurious correlation in hydraulics and hydrology Proceedings of the American Society of Civil Engineers, Journal of the Hydraulics Division 91, 35-43.

BUTLER, J.C. (1974): Analysis of correlations between percentages Journal of Geological Education 74, 56-61.

BUTLER, J.C. (1976): Principal components analysis using the hypothetical closed array Journal of the International Association of Mathematical Geology 8, 25-36.

CALHOUN, J. (1962): Population density and social pathology Scientific American 206, 139-148.

CHAYES, F. (1971): Ratio Correlation University of Chicago Press, Chicago.

CHAYES, F. (1973): Determining null values of ratio correlation by simulation Annual Report of the Director of the Geophysical Laboratory 1630, 681-682.

CHAYES, F. and KRUSKAL, W. (1966): An approximate statistical test for correlations between proportions Journal of Geology 74, 692-702.

CHAYES, F. and KRUSKAL, W. (1970): An approximate statistical test for correlations between proportions: some corrections Journal of Geology 78, 380.

DARROCH, J.N. (1969): Null correlation for proportions Journal of the International Association of Mathematical Geology 1, 221-227.

DARROCH, J.N. and RATCLIFF, D. (1970): Null correlation for proportions - II Journal of the International Association of Mathematical Geology 2, 307-312.

EVANS, I.S. (1975): Discussion following a paper given by Cliff, A.D. and Ord, J.K. Journal of the Royal Statistical Society Series B 37, 341-342.

EVANS, I.S. (undated): Percentages, ratios and inbuilt relationships Unpublished Teaching Handout, Department of Geography, University of Durham.

CALLE, O.R., COVE, W.R., and McPHERSON, J.M. (1972): Population density and pathology: What are the relations for man? Science 176, 23-30.

GALTON, F. (1897): Note to the memoir by Professor Karl Pearson F.R.S. on spurious correlation Proceedings of the Royal Society of London 60, 498-502.

HOWE, G.M. (1963): National atlas of disease mortality in the United Kingdom Nelson, London.

JOHNSTON, R.J. (1976): Residential area characteristics: research methods for identifying urban sub-areas - social area analysis and factorial ecology in Herbert, D.T. and Johnston, R.J. (eds.) Spatial processes and form Wiley, London.

JOHNSTON, R.J. (1977): Principal components analysis and factor analysis in geographical research: some problems and issues South African Geographical Journal 59, 30-44.

JONES, K. (1978): Percentages, ratios and inbuilt relationships in geographical research: an overview and bibliography Discussion Paper No. 2, Department of Geography, University of Southampton.

KUH, E. and MEYER, J.R. (1955): Correlation and regression estimates when the data are ratios Econometrica 23, 400-416.

KUNREUTHER, H. (1966): The use of the Pearsonian approximation in comparing deflated and undeflated regression estimates Econometrica 34, 232-234.

MADANSKY, A. (1964): Spurious correlation due to deflating variables Econometrica 32, 652-655.

MADDALA, G.S. (1977): Econometrics McGraw-Hill, New York.

NEIFELD, M.R. (1927): A study of spurious correlation Journal of the American Statistical Association 22, 331-338.

NORCLIFFE, G.B. (1977): Inferential statistics for geographers: an introduction Hutchinson, London.

PEARSON, K. (1897): Mathematical contributions to the theory of evolution - on a form of spurious correlation which may arise when indices are used in the measurement of organs Proceedings of the Royal Society of London 60, 489-498.

PEARSON, K., LEE, A. and ELDERTON, E.M. (1910): On the correlation of death rates Journal of the Royal Statistical Society 73, 534-539.

POSTON, D.L. (1974): An examination of urban mortality using age-adjusted death rates
Social Science Quarterly 55, 182-188.

RANGARAJAN, C. and CHATTERJEE, S. (1969): A note on the comparison between correlations of original and transformed variables
American Statistician 23, 28-29.

REED, J.L. (1921): On the correlation between any two functions and its application to the general case of spurious correlation
Journal of the Washington Academy of Science 11, 449-454.

SCHUESSLER, K. (1973): Ratio variables and path models in Goldberger, A.S. and Duncan, O.D. (eds.)
Structural equation models in the social sciences
Seminar Press, New York.

SCHUESSLER, K. (1974): Analysis of ratio variables: opportunities and pitfalls
American Journal of Sociology 80, 379-396.

SCHWIRIAN, K.P. and La GRECA, A.J. (1971): An ecological analysis of urban mortality rates
Social Science Quarterly 52, 574-587.

SCHWIRIAN, K.P. and La GRECA, A.J. (1974): The effect of alternative age adjustment procedures on the analysis of urban mortality patterns
Social Science Quarterly 55, 190-194.

TUFTE, E.R. (1969): Improving data analysis in political science
World Politics 21, 641-654.

UNWIN, D.J. (1977): Statistical methods in physical geography
Progress in Physical Geography 1, 185-221.

USLANER, E.M. (1976): The pitfalls of per capita
American Journal of Political Science 20, 125-133.

USLANER, E.M. (1977): Straight lines and straight thinking: can all of those econometricians be wrong? American Journal of Political Science 21, 183-191.

VISVALINGHAM, M. (1976): Chi-square as an alternative to ratios for statistical mapping Working Paper 8, Census Research Unit, Department of Geography, University of Durham.

VISVALINGHAM, M. and DEWDNEY, J.C. (1977): The effects of the size of areal units on ratio and chi-square mapping Working Paper 10, Census Research Unit, Department of Geography, University of Durham.

WRIGLEY, N. (1973): The use of percentages in geographical research Area 5, 183-186.

YULE, G.Y. (1910): On the interpretation of correlations between indices or ratios Journal of the Royal Statistical Society 73, 644-647.

ZODROW, E. (1974): Note on closure correlation Canadian Journal of the Earth Sciences 11, 1616-1619.

PART II

APPLYING THE EXPLORATORY APPROACH : THE RELATIONSHIP BETWEEN DISEASE AND WATER HARDNESS

Chapter

C H A P T E R 6

DISEASE AND WATER HARDNESS: A CRITICAL EXAMINATION OF THE STATISTICAL EVIDENCE

'Here again, another group is in hot pursuit of the water problem and, if anything, the water is even more muddy ... It would appear that there must be some important and useful information in all these data, but its true significance remains obscure' Ogglesby (1966, 186).

INTRODUCTION

The relationship between water hardness and heart disease mortality is probably the most contentious issue in modern epidemiology. It has been the subject of a large number of original articles, reviews and even editorials in such high-circulation journals as the Lancet and the British Medical Journal. The reasons for such a considerable literature are threefold. Firstly, the excess number of deaths in soft-water areas as compared to hard-water areas may be very large; secondly, prophylactic measures may be taken to remove this excess and thirdly, there is a considerable disagreement over the strength of the relationship between water hardness and heart disease. If the relationship is truly a cause-and-effect one, it has been estimated that the excess number of deaths of men aged 45 to 64 years in England and Wales is of the order of 10,000 per year. Taking other age groups, women and other countries into account reveals a considerable excess of deaths and it is argued that this possible excess could be reduced by changing or artificially hardening the water supply. Moreover, unlike other possible causes of heart disease (such as smoking, high-cholesterol diet, lack of exercise) it is thought that the water supply should be readily amenable to control and change. This argument is, of course, based on the supposition that the relationship between water hardness and heart disease is a true, causal one. There is, however, considerable disagreement in the medical literature over the exact nature of the relationship, and some researchers have even argued that no relationship exists.

This chapter represents an attempt to provide a critical appraisal of conflicting evidence and, in particular, attention will be paid to the use of statistical procedures by both supporters and opponents of the water theory. The discussion will begin with an outline of the early development

of the 'water story' by considering the areal mortality studies that were reported in the 1960s. The development and extension of the water theory will then be discussed in terms of studies based on clinical surveys, post-mortem findings and bulk and trace elements in the water supply. This will be followed by an outline of the arguments against the acceptance of the water theory and, finally, a critical review of the statistical bases of this work will be presented.

Before beginning the water story it is appropriate to consider some definitions of both water hardness and the diseases under study. The two main components of water which make it 'hard' are calcium and magnesium. Two distinct types of hardness can be recognised: temporary hardness (due, in the main, to bi-carbonates in the water) and permanent hardness (which is primarily caused by sulphates and chlorides). As well as these 'bulk' elements, water in different parts of the country also differs in terms of trace elements such as lead, cadmium, lithium and vanadium. Both bulk and trace elements in water have been associated with mortality from a number of diseases.

Although the water theory has been extended to a number of other diseases, research has generally been concentrated on the cardiovascular group of diseases - the disease of the heart and blood—supplying vessels. Unfortunately, the International Classification of Diseases has been revised periodically and the definition of what constitutes this cardiovascular group of diseases has also changed (Table 6.1). In 1971 this broad group of diseases accounted for about half the deaths in England and Wales. Ischaemic heart disease is now the major sub-category in the group and this disease is the result of an inadequate reception of oxygen-rich blood by the heart muscle (myocardium); this disease is simply known to the layman as a 'heart attack'. Cerebrovascular disease, or vascular

Table 6.1 Classifications of cardiovascular disease

Date	Disease (major categories only)	England and Wales percentage of total deaths	
		Male	Female
1951	All diseases of the circulatory system	35	37
	arteriosclerotic and degenerative heart disease	26	26
	hypertensive disease	4	4
	disease of arteries	2	3
	chronic rheumatic heart disease	1	3
	other diseases of the heart	1	1
	diseases of veins and other diseases of the circulatory system	0	0
	rheumatic fever	0	0
	Vascular lesions of the central nervous system	10	15
1961	All diseases of the circulatory system	36	38
	arteriosclerotic and degenerative heart disease	28	25
	hypertensive heart disease	2	2
	other hypertensive disease	1	1
	diseases of arteries	0	0
	chronic rheumatic heart disease	1	2
	other diseases of the heart	2	3
	diseases of veins and other diseases of the circulatory system	1	1
	rheumatic fever	0	0
	Vascular lesions of the central nervous system	11	17
1971	All diseases of the circulatory system	50	54
	ischaemic heart disease	29	22
	cerebrovascular disease	11	18
	other forms of heart disease	4	6
	diseases of arteries, arterioles and capilliaries	3	4
	hypertensive disease	1	2
	chronic rheumatic heart disease	1	2
	diseases of veins and other diseases of the circulatory system	1	1
	active rheumatic fever	0	0

Source: Registrar General (1951, 1961, 1971)

lesions of the central nervous system as it used to be known, is the result of either a substantial reduction in blood flow to some part of the brain or of intra-cranial bleeding. This disease is more widely known as apoplexy or stroke. Hypertensive disease is the result of abnormally high blood pressure, while chronic rheumatic heart disease is the result of a heart failure induced by a 'fever' which can occur during recovery from a throat infection.

THE WATER STORY: AREAL MORTALITY STUDIES

Negative relationships

The story of the investigation of the relationship between water hardness and cardiovascular-disease (CVD) mortality begins in 1957 with a Japanese agricultural chemist, Jun Kobayashi. He plotted on a map of Japan the degree of acidity of river water, compared this map visually with a map of death rates from apoplexy (strokes) for the Japanese prefectures, and found a marked association. Kobayashi observed a negative relationship between water hardness and deaths from apoplexy; the greater the water hardness the lower the death rate from this disease. These findings were re-examined by the American epidemiologist Henry Schroeder who was in Japan at this time. Using statistical tests of areal differences he substantiated Kobayashi's negative association for apoplexy while he also found a negative relationship between water hardness and heart disease.

In the United States, meanwhile, researchers were discovering a very marked geographical distribution of deaths from CHD (coronary heart disease). Some parts of the country had a death rate from this disease that was twice as high as other areas. These high rates were predominantly along the east, west and Gulf coasts, while the low rates were found in the western plains and some areas of the mid-south. In 1960 Schroeder, on returning from Japan, searched

for a possible association between water supply and CVD mortality in the United States. Using data on a state basis, he found a strong, negative relationship between water hardness and mortality from cardiovascular disease for both men and women. Moreover, support for these findings came from the same researcher's study of white male mortality for the 163 largest Metropolitan areas of the United States. In this investigation he found that water hardness, magnesium and calcium concentration were all negatively correlated with CHD mortality.

In 1961 a team of British researchers led by the late Lady Crawford carried out a parallel study based on data for the County Boroughs of England and Wales. Similar results were found to Schroeder's in the United States, the main difference being that the relationship was much stronger with calcium in the England and Wales study and no substantial correlation was found with the concentration of magnesium in drinking water.

Following these early studies, a large number of investigations have been carried out in a number of different countries and at different scales. Muss (1962), for example, studied two areas of New York that differed in their water supply. He found that the hard-water area had a 10 per cent lower CVD mortality than the soft-water area and for him these results tended to 'confirm the national studies' (p 1371). The results of a Swedish study were published in February, 1965; a substantial negative correlation between calcium in the water and 'degenerative heart disease' was found (Biorck et al). For the Netherlands, Bierstehen reported in 1967 a negative relationship between water hardness, calcium and CVD mortality in women but not in men. In April 1969, Anderson and his co-workers reported that the soft-water counties of Ontario had higher CHD death rates than the hard-water counties. More recently, for Canada as

a whole, Neri et al (1972) have discovered that lower death rates from cardiovascular and non-cardiovascular disease tend to prevail in hard-water areas. Meanwhile both Schroeder (1966) and the British team (1968) were re-examining their results with more recently published data. Again both investigations found a substantial negative correlation between CHD mortality and water hardness.

Conflicting results

The results that have been mentioned above have all shown a negative association between some element of water hardness and some type of cardiovascular disease. However, it is important to stress that a number of other studies using areal mortality data have produced differing results. In 1964, Lindemann and Assenzo studied the death rates (1950-1959) for the counties of Oklahoma, and their results showed a slight positive relationship between water hardness and CHD mortality. For Ireland, Mulachy (1964) examined the 15 largest cities and found no relationship between hardness and CVD mortality. The same author (1966) extended his study to include the urban areas of Northern Ireland (bringing the number of observations to 28) and again he found no substantial relationship. Furthermore, the hypothesis that cardiovascular disease mortality is inversely related to water hardness did not receive support from studies in Washington, of major Australian cities, and of three different water-supply areas in Los Angeles.

EXTENDING THE WATER STORY

Introduction

Despite, or possibly because of such conflicting findings, researchers have continued to examine the water hypothesis, developing and extending the research in a number of ways. Whereas all the studies mentioned above have concentrated on areal relationships, a number of

investigators have attempted to use individual data drawn from autopsies and clinical surveys. In particular, some researchers have tried to use post-mortem findings to elucidate possible mechanisms relating water hardness to cardiovascular disease, while other researchers have used such data to determine whether sudden death from heart disease is more prevalent in soft- rather than hard-water areas. Meanwhile, investigations based on clinical surveys of individuals have attempted to discover if the 'risk' factors (such as cholesterol levels associated with heart disease) are lower in hard- than in soft-water areas. Finally, there are researchers who have extended the water story by attempting to determine exactly which water component may be related to cardiovascular disease. These workers have used autopsies, clinical surveys and areal mortality studies.

Post-mortem findings, narrowed arteries and sudden death

Working on the basis that water hardness truly does affect cardiovascular disease, researchers have attempted to examine the relationship between the main processes of the disease and water hardness. In ischaemic heart disease, narrowing of the arteries deprives the heart muscle of an adequate blood supply and a part of the muscle may die (infarct). Normally after such an infarction has occurred the heart muscle heals, but death may occur either suddenly from an electrical disturbance (an arhythmia) or from mechanical failure of the heart. An arhythmia occurs when the electrical activity of the heart becomes unbalanced and such a disturbance is particularly likely to occur in the first few hours after an infarction. Mechanical failure of the heart is a much slower process, with death occurring because of gradual degeneration of the heart muscle.

A number of studies have attempted to find a relationship between the early stages of the disease (narrowing of the arteries) and water hardness. Strong and his co-workers

(1968), in a major international study, compared over 23,000 sets of arteries from 14 countries but did not discover a relationship with water hardness. In this country, Crawford and Crawford (1967) compared the coronary arteries of two groups of necropsies from both soft-water Glasgow and hard-water London. In one group the men had died as a result of accidents, while in the other group death had been brought about by a heart attack. The results were puzzling: although they found that there were more depositions and narrowed arteries in the accidental death group for men aged 30-44 in soft-water Glasgow, this did not hold for men aged 44-69 years of age. Moreover, those men that had died of heart attacks in Glasgow had less narrowing and deposition of the arteries than a similar group in London.

On the basis of such evidence, a number of the proponents of the water hypothesis have suggested that the link must be between water hardness, the heart muscle and lethal arhythmias. Such deaths tend to be rapid; indeed, the sufferer often dies before he can be admitted to hospital and thus the death has usually to be certified by a coroner. It has therefore been suggested that if water hardness is genuinely related to fatal arhythmias, 'sudden death' should be higher in soft- rather than hard-water areas. Such a hypothesis has been examined by a number of workers, but inconclusive results have again been found. Anderson and his co-workers (1969) defined 'sudden death' as death certified by a coroner and found that such deaths were higher in soft-water areas of Ontario than in the hard. In England and Wales, Crawford and Clayton (1973) studied 1,795 deaths in twelve towns; six of these towns had a hard-water supply while the other six had a soft-water one. They found that 'sudden deaths' (defined as 'death within one hour of the first heart attack') were more common in the soft-water towns. However, another group of Canadian research workers have re-examined Anderson's original study in Ontario and they concluded that

> 'in the opinion of the present writers, the Ontario statistics so far examined, taken as a whole, hardly sustain Anderson's hypothesis .... Our statistical analysis suggest that in Canada generally, water hardness does not have the value as a predictor of CHD mortality rates that it has been shown to have in the United States, England etc.' (Neri, Hewitt and Mandel, 1971, 104).

Yet again the results from a number of investigations are not conclusive and obviously further research is required before the hypothesis that water hardness is related to sudden death can either be rejected or accepted.

Clinical surveys and risk factors

Other researchers have attempted to extend the water story by trying to discover whether certain 'risk' factors such as cholesteral level, heart rate and blood pressure are higher in areas with soft water supplies. Such research has been conducted on the basis of clinical surveys of morbidity. Elwood (1971), together with a group based at the MRC epidemiology unit at Cardiff, have examined 243 men in hard-water areas and 357 men in the soft-water areas of South Wales, both samples being considered representative of the general male population. They found no important differences between the men from the two areas in terms of blood pressure, cholesterol and sugar levels. However, Stitt (1973) and his co-authors found differences in blood pressure, cholesterol and heart-rate levels between male executive civil servants living in six hard and six soft-water English towns. Similarly, studies in other countries have added to the confusion. For example, Masironi (1976) and his co-workers found average blood pressure amongst the villagers of New Guinea to vary inversely with water calcium; yet Bierenbaum (1973, 1975) and his co-authors found, in one study, cholesterol to be higher in the hard water areas, while in another study they found no significant difference. Obviously, such results have not elucidated the water story

and again more investigations are required before any decision can be made as to whether variations in risk factors are determined by differences in water supply.

Water components: bulk and trace elements

Another facet of the water hypothesis that has been examined is the relationship between different components of the water supply and cardiovascular disease. Researchers have looked at both the bulk elements in drinking water (such as calcium, magnesium and sodium) as well as trace elements such as cadmium, lithium and cobalt. For the bulk elements, areal mortality studies have found the major relationship to be with magnesium and potassium in the United States, but with calcium in Britain. In relation to these elements, two hypotheses have been proposed to account for the relationships.

(1) Bulk elements obtained from drinking water may be an important addition to the elements obtained from other sources such as food.

(2) Some bulk elements may be deposited on the inside of water pipes thus preventing the uptake of toxic elements by drinking water.

With regard to the first hypothesis the evidence is again controversial. Dauncey and Widdowson (1972) examined 157 men in hard-water London and 196 men in soft-water towns. Using urinary excretion as an estimate of the amounts of calcium, magnesium and potassium that had been absorbed by the body they found no significant differences between the areas. Similarly, Anderson (1972) on the basis of post-mortem findings did not detect any differences between three Canadian cities with differing water supplies. However, on the basis of autopsies Anderson (1973) and his co-workers found that residents of soft-water areas have lower magnesium concentrations in the heart muscle than do residents of hard-water areas. Finally, Chipperfield (1976) and her co-authors, in a study of autopsies in Hull and Burnley,

found magnesium concentration in heart muscle to be higher in the soft-water area. This, of course, is a directly opposite result to that found by Anderson on the basis of his autopsy results.

In relation to the second hypothesis relating bulk elements to cardiovascular disease, soft water tends to corrode metal pipes while water that has a high calcium content may prevent corrosion by forming a protective shield of insoluble calcium salts on the inside of the pipe. While a number of toxic elements may be taken up by running water, attention has been concentrated on the uptake of lead from domestic water pipes. Plumbosolvency is not a problem that has been solved in this country and, indeed, a recent study of Glasgow by Beattie and his colleagues (1972) found that the lead content of water drawn from domestic kitchen taps was up to eighteen times the WHO acceptable limit. Crawford and Morris (1967) in a study of nine soft-water and nine hard-water towns, found a high concentration of lead in the drinking water of the soft-water towns. Moreover, as well as lead concentrations being higher in soft drinking water, it is also known for rats that there is a greater absorption of lead by the body when dietary calcium is low. Unfortunately, studies of human populations have again produced conflicting results. In a study of six hard and six soft-water towns, a pathological examination of human ribs found a higher concentration of lead in the soft-water towns (Crawford and Crawford, 1969), whereas Elwood (1976) and his colleagues, in their study of deaths in Caernarvonshire, found no association between cardiovascular disease and water lead. Those proponents of the water hypothesis who have favoured the importance of lead have tended to regard it as an 'enhancing' factor which does not account for all the association between water hardness and cardiovascular disease but, with such conflicting results, it is obviously

impossible to assess the true contribution of lead as a cause of cardiovascular mortality.

With regard to trace elements, cadmium and lithium in particular have been associated with cardiovascular disease. Schroeder (1969) has pointed out that the only trace element to reproduce in rats the pathological elements of hypertensive disease in man is cadmium. One possible routeway for cadmium to be absorbed by man is through drinking water that has passed through galvanised pipes. If lemonade is left in a galvanised bucket for a few hours and is then drunk the result can be acute and fatal cadmium poisoning. It has also been found that soft water is more capable of dissolving cadmium than hard water. In the United States, areal mortality studies by Voors (1970, 1971) and Blachly (1969) have implicated lithium as the water component that is related to cardiovascular disease. However, taking studies in the United Kingdom, Sweden and the USA as a whole, no clear pattern of association between trace elements and disease has been found and indeed, no trace elements have been consistently found at higher concentrations in hard-water than in soft-water areas.

Infant mortality and cancer

The final extension of the water hypothesis that will be considered in this review is the examination of the association between drinking water and other diseases. Two broad groups of mortality have been suggested to have links with water hardness: infant mortality (deaths from children under one year of age) and deaths from malignant neoplasms (cancer). Crawford (1972) and her co-workers were able to find relationships between water hardness and infant mortality in the United Kingdom; in general the softer the water the greater the infant death rate. However, research in other countries (for example, Allwright, 1974) has found no such relationship for infant mortality and, in this country, no

relationship between water hardness and anencephaly (children born without a fully developed central nervous system) has been found for south-west Lancashire (Fielding and Smithells, 1971).

The study of cancer deaths has also provided a profusion of conflicting results. For example, Page, Harris and Epstein (1976) in a study of Louisiana, examined the relationship between all cancer deaths, deaths from gastric cancer, deaths from cancer of the urinary tract and the percentage of people living in an area who drank water which had been piped from the Mississippi. They found a relationship between all three types of death and water supply. Similar findings were reported for New Orleans (which also extracts water from the Mississippi) and this study by the Environmental Defense Fund suggested that if New Orleans changed its source of water 'in the long run over 50 premature deaths from cancer among white males alone would be averted annually' (quoted in De Rouen, 1975). However, when this study was re-analysed, De Rouen did not find a consistent relationship between cancer deaths and drinking water.

In summary, those studies that have investigated the relationship between disease and various bulk and trace elements of drinking have found conflicting results. Moreover, it has so far proved impossible to resolve this conflicting evidence concerning the role of water hardness in determining disease patterns.

OPPOSITION TO THE WATER STORY

The preceding discussion has been largely concerned with telling the story of the search for the association between water and chronic disease but, in this section, the views of the opponents to the theory will be examined.

Quality of data

It has been argued that the basic data that have been used by epidemiologists are very crude and that measurement error may have led to inferential error. With regard to death certification, it has been contended that doctors in different areas tend to classify diseases in different ways resulting in apparent geographical patterns of certified death. However, a number of studies have examined death certification in detail and have concluded that the apparent patterns accurately reflect true variations. For example, Crawford and her co-authors (1977), on the basis of the examination of individual death certificates and other information obtained from the doctor certifying death, found that there were differences between soft-water and hard-water areas in terms of hypertensive heart disease, ischaemic heart disease and cerebrovascular disease mortality. Similar criticisms have also been made of the quality of the water-supply data. Frequently, studies are based on the examination of concentrations of bulk and trace elements as they leave the water—supply works and not as they reach consumers. Moreover, many areas receive supplies from a number of variable sources and it may be that 'average' concentrations do not adequately reflect this variety. However, several investigations have been based on the examination of water supplies in the home, and some of these studies have found a negative association between water hardness and cardiovascular disease.

Non-specificity of findings and anomalous results

Another criticism that has been made of the water hypothesis is the non-specificity of the findings. For example, McCabe (reported in Winton and McCabe, 1970) in a correlation study of 135 areas with well-defined water supplies found negative correlations between water hardness and heart disease, cancer, cirrhosis of the liver, stomach and duo-

denal ulcers, strokes, congenital malformations, diseases of infancy and gastritis. Similarly, for the United States as a whole, inverse relationships have been observed between water hardness and cancer of the respiratory tract, congenital malformations and motor vehicle accidents! However, as the supporters of the water hypothesis point out, these associations with other diseases are not as marked as with cardiovascular disease.

Since Turner's (1962) study of England and Wales, the opponents of the water hypothesis have invariably noted the anomalous results for Birmingham. This county borough receives extremely soft water from Welsh reservoirs yet it has low rates of cardiovascular mortality similar to neighbouring county boroughs with hard water. The supporters of the hypothesis have replied that cardiovascular disease is multi-factorial in its aetiology and, in England and Wales at least, it is remarkable how few exceptions can be found to the rule of soft water: high cardiovascular disease.

Lack of causal links

One of the major flaws in the argument for accepting a causal relationship between water hardness and disease is undoubtedly the lack of any cogent theoretical explanation of the associations found. In relation to the importance of bulk and trace elements to the human body, the contribution of drinking water is thought to be slight. For example, Hollingworth (1956) has estimated that the calcium contribution from drinking water forms only about a tenth of the average intake of dietary calcium, so that it seems unlikely that this small contribution forms an essential part of the total intake. However, it may be that calcium from water supplies modifies the absorption of certain dietary components. For example, Anderson and his co-authors (1971) studied the compositions of tea infusions and found that when calcium is high there is a substantial reduction in the 'oxidate' content

of the drink. (It also appears that tea is the single richest source of oxidate in the UK diet.) Moreover, other proponents of the water hypothesis have contended that the degree of water hardness may determine its capacity to extract trace elements from rocks, soil, water pipes and even cooking utensils. However, such speculations have not been validated in all the studies that have been undertaken to examine them.

Lack of control variables

Another argument favoured by those researchers who do not support the water hypothesis is that in areal mortality studies the influence of other possible causal variables has not been properly controlled. This has been raised in particular by those researchers who believe that the association between water hardness and CVD mortality is non-causal, working through a common link with climate. For example, Roberts and Lloyd (1972), in a study of the local authorities in South Wales and the English and Welsh County Boroughs, suggested that the association between ischaemic heart disease and water hardness is a secondary one, the true association being between IHD mortality and rainfall (sic). Dudley and his co-authors (1969) had reported for the USA that the amount of variation of CHD mortality explained by water hardness was less than that explained by a climatic comfort index and West, Lloyd and Roberts (1973, 39) have written:

> 'we believe .... that climatic rather than drinking water differences account for inter-town variations in IHD mortality in England and Wales'.

They also argue that their findings are so important that there is a need to bring influence to bear on the

> 'appropriate public attitudes e.g. to wearing warm underclothing and adequately protective outer clothing'.

Of course, such studies implicating climate have as little

theoretical justification as the water hardness studies and, moreover, in relation to this latter study it was shown in Chapters 3 and 4 that the model used by West and his co-authors is mis-specified and cannot be used to make valid inferences. This statistical criticism of a previous study leads to the final and major argument against the uncritical acceptance of the water hypothesis.

A STATISTICAL CRITIQUE OF PREVIOUS STUDIES

Introduction

It is a major tenet of the present study that the statistical methods used by epidemiologists to test hypotheses concerning disease and water quality have not always been applied correctly. This is a major criticism of twenty years research carried out by a large number of investigators and it needs substantial support. In order to assess the statistical credentials of the work, a large number of studies were examined; these were identified by using the medical citation journal Indexus Medicus, searching under the terms 'water supply' and 'water softening' for each year between 1958 and 1978. In all 99 articles directly concerned with the water hypothesis were identified. Of these, 26 were 'reviews' illustrating the contentious nature of the subject; but, because attention is to be concentrated on studies reporting 'original' results, such review articles will not be considered here.

Subjective studies

Of the 73 studies reporting original results, 13 have no statistical basis whatsoever. For example, Barritt (1972, 186) argues on the basis of commonsense:

> 'it is common knowledge among village communities that well water is more wholesome than river water or piped supplies'

and he implicates nitrate as a causal factor in disease because nitrate goes into the body yet never appears in urine.

Also in this group of studies are those based solely on visual comparison. For example, Kobayashi's original study is based merely on the visual comparison of two maps - acidity of river water and deaths from apoplexy in Japan. Similarly, Takahashi (1967) compared maps of cerebrovascular death rates on a regional scale with a geological map of Europe and concluded that calcium is an important variable in determining deaths from this disease. Such studies, of course, are highly subjective and do not take into account a whole range of other variables that could co-vary with water hardness. Such studies, whether they support or refute the water hypothesis are of little real value in forming a valid judgement concerning that hypothesis.

Statistical studies

The remaining 60 papers, based on statistical methods and original results can be classified into three groups: those based on 'matched' controls (23), those based on correlation analysis or simple regression with one independent variable (23) and those based on multiple regression analysis (14). Each of these broad groups will now be examined in turn.

'Matched' controls

The method of 'matched' controls has been mainly used by those researchers who have worked on an individual basis (clinical surveys, post-mortem findings), but a few studies based on areal mortality data have also attempted to use the method.[1] Basically the approach consists of 'matching' individuals or areas on a number of characteristics so that they only differ in terms of hardness of water supply. But the fundamental problem with a match—controlled study is that the controls are never fully matched, a problem that can be clearly seen in an exchange of letters in the Canadian Medical Association Journal. Anderson (1975)

and his co-workers examined 64 males who had died of accidents and found that 44 men who had lived in the soft-water areas had lower magnesium concentrations in the heart muscle than 20 men who had lived in a hard-water area. Vobecky and Shapcott (1975, 925) argued that, before any conclusions could be made on the basis of these results, the individuals would have to be matched on a large number of characteristics because those living in a hard-water area could differ in a large number of ways from those living in soft-water areas. Anderson and his co-authors would have liked to have done this but they found it an impossible task:

> 'matching after the event would demand very large numbers, and the problems inherent in matching before the event boggle the mind - this was, after all, an autopsy study' (p 925).

Another study that exemplifies the problem of such an approach is Crawford and Crawford's (1969) investigation based on the studies of ribs removed from bodies of men who had proviously lived in either soft-water Glasgow or hard-water London. They write:

> 'In general, the lead values in the soft-water area are higher than those in the hard-water area ... These differences are highly significant, ($P > 0.001$)' (p 700).

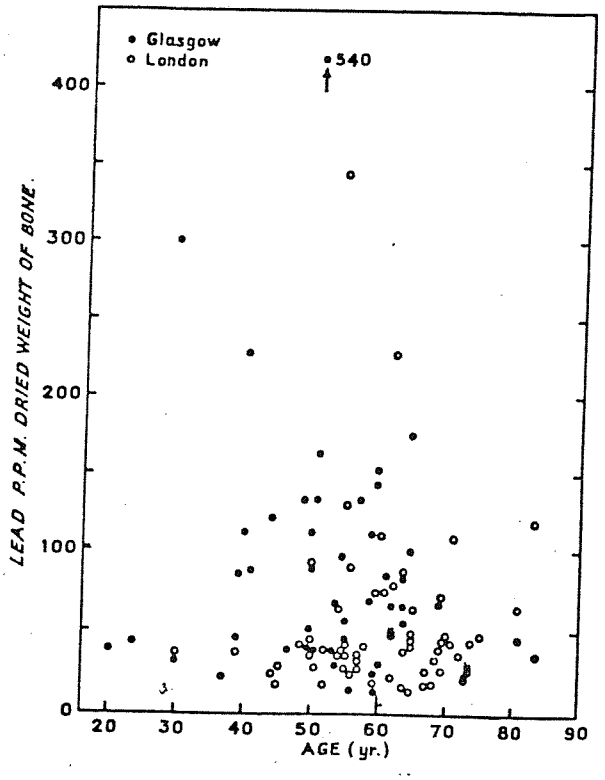
However, when a scatter graph of these relationships is examined the range of values for both hard and soft-water areas is considerable (Figure 6.1), and as the authors correctly point out

> 'there are, of course, many factors on which we have no information e.g. the amount of industrial exposure, dietary habits and contact with lead paints in childhood and these may explain some of the very high values found in individuals in both areas' (p 700).

They concluded that

> 'the question of lead being involved in disease processes in cardiovascular disease cannot be answered with the present data' (p 701).

FIGURE 6.1 SCATTER PLOT OF LEAD CONTENT OF BONES AGAINST AGE AT DEATH: SOFT-WATER GLASGOW, HARD-WATER LONDON



SOURCE: CRAWFORD AND CRAWFORD (1969)

Unfortunately, all the other studies in this group suffer to a lesser or greater degree from similar problems of matching and it is not surprising, therefore, that such confusing results have been obtained. Moreover, such studies pose considerable problems in terms of sample size; as it is very costly and difficult to perform clinical surveys and studies based on autopsy findings, only a small number of subjects are normally measured. For example, Morgan's (1972) study of tissue copper and lead was based on only 101 observations and Crawford and Clayton's (1973) study of ribs used only 55 observations. Similarly, Comstock (1971) in an analysis of so-called matched controls (that were not in fact matched, for example, by either income or occupation) used only 189 observations. But, as has been pointed out by Neri, Hewitt and Schreiber (1974), even if there was a difference in death rates as high as twenty per cent between the extremes of high and soft-water areas, 2000 observations would be needed to detect such a difference at conventional significance levels. It is therefore hardly surprising that Comstock failed to detect a 'significant' relationship between water hardness and death from heart disease.

In conclusion, it is very difficult to match observations in terms of similar characteristics so as to control for those particular characteristics; this is especially the case when there are few observations. A detailed examination of the 23 studies in this category revealed that no study had adequately controlled for all the variables that could have accounted for differences in disease and one must, therefore, regard their findings with suspicion.

Correlation analysis

The majority of studies that specifically examined areal mortality data have used correlation coefficients, and 23 such studies have been identified. The main problem of adopting such an approach is that other variables that may

be causal factors also co-vary with water hardness. For example, in England and Wales a map of soft water mirrors a map of heavy industry, air pollution and low socio-economic status. Unless such factors are adequately controlled in some way, it is impossible to study the specific relationship between water hardness and disease. It does not matter how significant the correlation between hardness and disease is found to be, those studies that have used such an approach (or simple regression with one independent variable) are practically worthless and add nothing to our knowledge concerning the water hypothesis.

Multiple regression analysis

The final category of studies is that based on partial correlation and multiple regression analysis using OLS estimation. There are 14 examples of this work. Although this form of analysis allows variables to be statistically controlled, it is based on a number of fundamental assumptions that have to be met before it can be used to make valid inferences. As discussed in earlier chapters, some of the fundamental assumptions of multiple regression are that no important variables are omitted, the model is not multicollinear, the model has the correct functional form and the error term is neither spatially autocorrelated nor heteroscedastic. Table 6.2 outlines whether or not these 14 studies have attempted to perform some sort of specification error analysis on the fitted model. Although this table is very generous (in that if a study merely acknowledges a problem without attempting to overcome it, an asterisk is still awarded) there are very few researchers who appear to be aware of the pitfalls of multiple regression? For example, Morton (1976) and his co-workers' model of congenital malformations in South Wales used only water elements as independent variables, although a study in this field that completely ignores social data must suffer from the problem of omitted variables. At the very least, such a

TABLE 6.2

Studies Using Multiple Regression or Partial Correlation to Examine the Water Hypothesis

Authors	Explanatory variables		multi-collinearity	hetero-scedasticity	functional form	spatial auto-correlation	used graphical plots
	social	physical					
Biorck et al. (1965)		*	*				
Crawford, Gardner and Morris (1972)		*					
Crawford, Gardner and Sedgwick (1972)	*	*	*				
De Rouen (1975)	*	*					
Dudley et al. (1969)	*	*	*				
Elwood et al. (1974)		*	*				
Morris et al. (1962)		*					
Morton et al. (1976)		*					
Page et al.(1976)	*	*		*			
Roberts and Lloyd (1972)		*					
Voors (1970)		*					*
Voors (1971)		*					
West et al.(1973)	*	*					
Gardner (1973)	*	*	*		*	*	*

problem should be investigated and not ignored. Similarly, Roberts and Lloyd (1972) used only rainfall and water hardness as independent variables and Page (1976) and his co-authors openly admit in their study:

> 'that there remain the problems of mis-specification in the simple linear models and of omission of variables' (p 56).

But, instead of cataloguing specific errors in these multiple regression analyses, let us turn to the study which, according to Table 6.2, is the most statistically rigorous investigation of the water hypothesis: that of Gardner (1973). Gardner is a statistician and the results presented in this article are the outcome of collaborations with the medical researchers of the MRC Social Medicine Unit. Not only is this investigation a statistically aware one, it is also very important in terms of the water hypothesis, for the results show a closer association between death rates and water indices than has been found in any other research project. It may also be added that this research won for Gardner in 1971, the Frances Wood Memorial Prize of the Royal Statistical Society.

The study is based on both routinely published data and specially collected data for the 83 County Boroughs of England and Wales. In all, over a hundred variables were collected to reflect the social and physical environment; the mortality data were based on age, sex, and cause-specific death rates for the main causes of death in England and Wales for 1951 and 1961. These data were divided into two groups, and only those towns that had a population of 80,000 or over in 1961 were used for regression modelling. The study was undoubtedly exploratory and correlation analysis was used in the 'initial screening' of some 5,000 pair-wise relationships. Independent variables were chosen for the regression model if they had consistent associations with cause-specific death rates and if they had been implicated in previous epidemiological research. The chosen environmental variables were however,

correlated with each other and in an attempt to overcome the multicollinearity problem, Gardner used principal components analysis to derive a 'social factor score'. (This was in fact the first or major component of the PCA.) For the original model, four explanatory variables ('social factor score', air pollution, latitude and water calcium) were regressed on each death rate in turn. In addition to using linear models, attempts were also made to transform both the dependent and independent variables, but broadly similar results were obtained. The residuals from the original models were compared with the other environmental variables that had been omitted from the model and, on this basis, rainfall was included in the model as the fifth regressor variable. The residuals from this reformulated model were mapped, but they appeared to offer little clue to any variables that had been inadvertently omitted from the study. Finally, the estimated parameters of the model were used to predict the death rates in the twelve county boroughs with a population of less than 80,000; this did not suggest that the models were fundamentally incorrect.

In this investigation, therefore, some graphical procedures have been used, multicollinearity of the independent variables has been recognised, non-linear models have been examined, and a form of cross-validation has been employed. While this approach corresponds to a considerable degree with the exploratory approach that has been developed in this thesis, nonetheless Gardner's study suffers from a number of drawbacks. Firstly, and most importantly, the form and choice of the variables used in the study can be criticised. The principal components analysis, which was used to construct the 'social factor score', was based on the correlations between percentage and ratio variables and, as discussed in Chapter 5, this can induce a number of problems of interpretation. Moreover, each time a regression model was fitted, death rates not absolute numbers of death were used and, as

was also shown in Chapter 5, the use of ratios merely to control for population size is a questionable procedure. In relation to the choice of explanatory variables, no occupational variables were included in the model. As it is known that workers in different occupations suffer differing disease incidence (see Table 6.3), the effects of such variables should have been explored. Moreover, the inclusion of a non-causal variable such as latitude in the regression model results in a number of difficulties. Latitude may have 'captured' the variation in the dependent variable that could have been explained by a number of causal variables, and the use of such a 'spatial' variable prevents the effective use of residual mapping as a means of detecting omitted variables. With regard to the problem of multicollinearity, Gardner used only the first principal component in the regression analysis but, as was discussed in Chapter 2, this procedure is known to be of doubtful validity. One final criticism that can be made of the study is that significance levels were attached to each regression coefficient without a warning that these may be misleadingly high because of the data-mining procedures that had been used by Gardner (see Chapter 1).

CHAPTER CONCLUSIONS

Some workers appear to be in no doubt of the importance of the water factor in human disease. Schroeder (1974, 310) after fifteen years of research has written:

> 'Although the water factor in no way accounts for all the mortality from arteriosclerotic heart disease, it probably constitutes a serious hazard to human health in persons exhibiting arterio sclerosis'.

Masironi (1970, 695) felt sufficiently confident to write that:

> 'the fact that consistent trends have been demonstrated by analysing data obtained by various investigators following different geographical and demographic approaches, and using different geochemical indicators, indicates that such trends cannot be simply a matter of statistical chance'.

TABLE 6.3

SMR s for England and Wales by Cause and by Selected Occupational Groups, Males aged 15-64, 1959-63

Disease	Occupational Groups					
	Miners and Quarrymen	Gas, Coke and Chemical Workers	Glass and Ceramic Workers	Furnace, Forge Foundry and Rolling Mill Workers	Professional Technical Workers, and Artists	Administrators and Managers
Stomach Cancer	149	125	90	121	51	71
Lung Cancer	80	110	128	140	55	50
Strokes	105	75	83	107	85	44
Coronary Disease	98	93	80	90	96	95
Bronchitis	156	99	133	147	34	30

A value of 100 represents average mortality for all occupations from the disease. Underlined values are of doubtful validity.

Source: Registrar General (1971).

In contrast to these statements, this present review does not accept or reject the hypothesis that there is a water factor influencing the development of cardiovascular disease. This opinion is based on the fact that studies using different approaches have found conflicting, inconsistent results and, as argued above, the statistical methodology of the studies can be severely criticised. Indeed, even the study that Sharrett and Feinleib (1975, 22) regard as the 'most extensive investigation of the water-hardness story', that of Gardner and his colleagues, can and has been criticised on a number of points.

Given these conclusions about previous research, how should the water hypothesis be investigated further? One way forward is to undertake a prospective study, examining a large number of individuals living in different water-supply areas over time and in such a way as to control for the influences of other variables. Such an investigation is presently underway and in twenty towns of England and Wales 7,500 individuals are being studied. Data are to be collected on such variables as diet and detailed analyses of domestic tap water are to be carried out (Powell, 1977, personal communication). One problem of this approach is that it is a very costly, time-consuming undertaking; the researcher obviously has to wait for the individual subjects to develop cardiovascular disease or die from it. But an alternative approach is to perform yet another areal-mortality study using an improved statistical methodology.

This chapter has been primarily concerned with statistical criticism of the 'hit and run' variety (Bross, 1960) - a number of studies have been criticised but little has been put in their place. Moreover, it is important to show not only that the studies used inadequate methodology but, also that, if appropriate methods had been used, different conclusions would have been drawn. This obviously requires that a study should be conducted using a more

rigorous methodology. If, with a more rigorous approach, water hardness is found to be highly related to cardiovascular disease, the results from a prospective study may be able to pinpoint the possible mechanisms or processes linking the two variables. However, if the improved methodology shows no relationship, it may be that considerable research expertise has been misdirected over the last twenty years because of the spurious results found by inadequate studies. In the remainder of this thesis such an areal-mortality study will be undertaken.

Before proceeding to this study, however, a general point must be made. This chapter has concentrated exclusively on the relationship between water hardness and disease because this subject well illustrates the varied approaches of epidemiologists and their equally varied conclusions. However, a similar story of conflicting results associated with inadequate methodology could have been told for other variables and their relationships with disease. For example, Wall (1976, 106), reviewing studies that have investigated the effects of air pollution on human health, concluded that although

> 'almost all studies, irrespective of methodology, have confirmed the negative impact of air pollution upon human health .... many aspects of the effects .... remain to be resolved'.

Moreover, as discussed in Chapter 4, the one study (Smith, 1977) that used specification-error analysis to develop an improved model did not find a substantial relationship relating disease and air pollution, although other researchers using broadly similar data (but no specification-error analysis) found a strong relationship between the variables.

Similarly, those studies that have investigated the effects of high-density living on human health and behaviour have reached conflicting conclusions. Roncek (1975) has reviewed the literature on density and human behaviour

providing a methodological critique similar in intent if different in scope to the present review. Considering the importance of density variables he concluded that

> 'recent research into the effects of these variables has received widespread attention in the major scientific journals. Yet it is questionable whether the methodological quality of the research warrants this and whether the conclusions of these studies are based upon strong enough evidence to be scientifically acceptable' (p 843).

In particular, he criticises the choice of variables used in these studies and the uncritical use of multiple regression. Obviously, an improved methodology is required for studying other possible causes of disease as well as water hardness.

CHAPTER 6: NOTES

1. As an example of the match-controlled approach to the study of areal mortality, consider the study of Meyers (1975). He examined the differences in mortality rates between two cities - Brisbane (soft water) and Queensland (hard water) and found, contrary to expectations, that the death rates were higher in the hard-water city. Of course, the cities are not truly matched and they will differ in a large number of ways besides water supply. As Meyer himself admits:

> 'it is difficult to confirm that a single environmental factor plays a dominant role' (p 101).

This is particularly the case when the analyst considers only a single variable!

2. Interestingly, only one study, that of Neri, Mandel and Hewitt (1972) specifically raised the problem of spatial autocorrelation:

> 'when data relate to geographical areas ... they are not likely to satisfy the condition of being independently drawn sample points' (p 931).

Although this study noted the problem, no action was taken to overcome this major difficulty.

CHAPTER 6 : BIBLIOGRAPHY

ALLEN-PRICE, E.D. (1960): Uneven distribution of cancer in West Devon. Lancet 1, 1235-1238.

ALLWRIGHT, S.A., COULSON, A., DETELS, R. and PORTER, C.E. (1974): Mortality and water hardness in three matched communities in Los Angeles. Lancet 2, 860-864.

ANDERSON, T.W. (1972): Serum electrolytes and skeletal mineralisation in hard and soft water areas Journal of the Canadian Medical Association 107, 34-37.

ANDERSON, T.W.,HEWITT, D., NERI, L.C., SCHREIBER, G. and TALBOT, F. (1973): Water hardness and magnesium in heart muscle. Lancet 2, 1390-1391.

ANDERSON, T.W. and Le RICHE, W.H. (1971): Sudden death from ischaemic heart disease in Ontario and its correlation with water hardness and other factors. Canadian Medical Association Journal 105, 155-160.

ANDERSON, T.W., Le RICHE, W.H. and MACKAY, J.S. (1969): Sudden death in ischaemic heart disease: correlation with hardness of local water supply New England Journal of Medicine 280, 805-807.

ANDERSON, T.W., NERI, L.C., SCHREIBER, G.B., TALBOT, F.D.F. and ZDROJEWSKI, A. (1975): Letter to the editor Canadian Medical Association Journal 113, 925.

ANDERSON, W., HOLLINS, J.G. and BOND, P.S. (1971): The composition of tea infusions examined in relation to the association between mortality and water hardness Journal of Hygiene (Cambridge) 69, 1-15.

BARRITT, N.W. (1972): Heart disease and hard water Lancet 2, 186.

BEATTIE, A.D., MOORE, R.W., DEVENAY, W.T., MILLER, A.R. and GOLDBERG, A. (1972): Environmental lead pollution in an urban soft-water area British Medical Journal 2, 491-493.

BERG, J.W. and BURBANK, F. (1972): Correlations between carcinogenic trace elements in water supplies and cancer mortality Annals of the New York Academy of Science 199, 249-264.

BIERENBAUM, M.L., FLEISCHMANN, A.I. and DUNN, J.P. (1969): Serum Lipids in hard and soft water communities Israel Journal of Medical Science Journal 5, 657-660.

BIERENBAUM, M.L., FLEISCHMANN, A.I., DUNN, J.P., HAYTON, T., PATTISON, D.C. and WATSON, P.B. (1973): Serum parameters in hard and soft water areas Lancet 1, 122.

BIERENBAUM, M.L., FLEISCHMANN, A.I., DUNN, J.P. and ARNOLD, J. (1975): Possible toxic water factors in coronary heart disease Lancet 1, 1008

BIERSTEHEN, K. (1967): Drinkwater zachtheid in Sterfte Tijdsche Soc. Geneesk 45, 658-661.

BJÖRCK, G., BOSTROM, H. and WIDSTROM, A. (1965): On the relationship between water hardness and death rates in cardiovascular disease Acta Medica Scandinavaca 178, 239-252.

BLACHLY, P.H. (1969): Lithium content of drinking water and ischaemic heart disease New England Journal of Medicine 281, 682.

BOSTROM, H. and WESTER, P.O. (1967): Trace elements in drinking water and death rates in cardio-vascular disease. Acta Medica Scandinavio 181, 465-473.

BROSS, I.D.J. (1960): Statistical criticism Cancer 13, 394-400.

CHIPPERFIELD, B., CHIPPERFIELD, J.R., BEHR, C. and BURTON, P. (1976): Magnesium and potassium content of normal heart muscle in areas of hard and soft water Lancet 1, 121-122.

CLAYTON, D.G., CRAWFORD, M.D. and MORRIS, J.N. (1973): Cardiovascular disease in hard and soft water areas
Lancet 1, 613-614.

COMSTOCK, G.W. (1971): Fatal arteriosclerotic heart disease, water hardness at home, and socio-economic characteristics
American Journal of Epidemiology 94, 1-10.

CORREA, P. (1972): Atherosclerosis and the geochemical environment
Annals of the New York Academy of Science 199, 217-218.

CRAWFORD, M.D. (1972): Hardness of drinking water and cardiovascular disease
Proceedings of the Nutritional Society 31, 347-353.

CRAWFORD, M.D. and CLAYTON, D.G. (1973): Lead in bones and drinking water in towns with hard and soft water
British Medical Journal 2, 21-23.

CRAWFORD, M.D. and CLAYTON, D.G. (1973): Sudden death and water quality
British Journal of Preventive and Social Medicine 27, 68.

CRAWFORD, M.D., CLAYTON, D.G., STANLEY, F. and SHAPER, A.G. (1977): An epidemiological study of sudden death in hard and soft water areas.
Journal of Chronic Disease 30, 69-80.

CRAWFORD, M.D. and CRAWFORD, T. (1969): Lead content of bones in a soft and a hard water area.
Lancet 1, 699-701.

CRAWFORD, M.D., GARDNER, M.J. and MORRIS, J.N. (1968): Mortality and hardness of local water supplies
Lancet 1, 827-831.

CRAWFORD, M.D., GARDNER, M.J. and MORRIS, J.N. (1971): Cardiovascular disease and the mineral content of drinking water
British Medical Bulletin 27, 21-24.

CRAWFORD, M.D., GARDNER, M.J., MORRIS, J.N. (1971): Changes in water hardness and local death rates
Lancet 2, 327-329.

CRAWFORD, M.D., GARDNER, M.J. and MORRIS, J.N. (1972): Water hardness, rainfall and cardiovascular mortality. Lancet 2, 1396-1397.

CRAWFORD, M.D., GARDNER, M.J. and SEDGWICK, P.A. (1972): Infant mortality and hardness of local water supplies. Lancet 1, 998-992.

CRAWFORD, M.D. and MORRIS, J.N. (1967): Lead in drinking water. Lancet 2, 1087-1088.

CRAWFORD, T. and CRAWFORD, M.D. (1967): Prevalence and pathological changes of ischaemic heart disease in a hard water and in a soft water area. Lancet 1, 229-232.

DAUNCEY, M.J. and WIDDOWSON, E.M. (1972): Urinary extraction of calcium, magnesium, sodium and potassium in hard and soft water areas. Lancet 1, 711-714.

DAVIES, B.E. and PINSENT, R.J.F.H. (1975): Minerals and morbidity. Cambria 2, 85-93.

De ROUEN, T.A. (1975): The New Orleans drinking water controversy: a statistical perspective. American Journal of Public Health 65, 1060-1062.

DINGLE, J.H., OGLESBY, P., SEBRELL, W.H., STRAIN, W.H. WOLMAN, A. and WILSON, J.R. (1964): Water composition and cardiovascular health. Illinois Medical Journal 125, 25-31.

DUDLEY, E.F., BELDIN, R.A., JOHNSTON, B.C. (1969): Climate, water hardness and coronary disease. Journal of Chronic Disease 22, 25-48.

ELWOOD, P.C. ABERNETHY, M. and MORTON, M. (1974): Mortality in adults and trace elements in water. Lancet 2, 1470-1472.

ELWOOD, P.C., BAINTON, D., MOORE, F., DAVIES, D.F., WAKLEY, E.J., LONGMAN, M. and SWEETNAM, P. (1971): Cardiovascular surveys in areas with different water supplies. British Medical Journal 2, 362-363.

ELWOOD, P.C., CHADD, M.A., BURR, M. and HAYMAN, L.M. (1972): Blood clotting and fibrinolysis in areas with different water supplies. British Journal of Preventive and Social Medicine 26, 246-248.

ELWOOD, P.C., ST. LEGER, A.S., MOORE, F., MORTON, M. (1976): Lead in water and mortality Lancet 1, 748.

FEDRICK, J. (1970): Anencephalus and the local water supply Nature 227, 176-177.

FIELDING, D.W. and SMITHELLS, R.W. (1971): Anencephalus and water hardness in S.W. Lancashire British Journal of Preventive and Social Medicine 25, 217-219.

FODAR, J.G., PFEIFFER, C.J. and PAPEZIK, V.S. (1973): Relationship of drinking water quality (hardness-softness) to cardiovascular mortality in Newfoundland Canadian Medical Association Journal 108, 1369-1373.

GARDNER, M.J. (1973): Using the environment to explain and predict mortality Journal of the Royal Statistical Society (Series A) 136, 421-440.

GARDNER, M.J. (1976): Soft water and heart disease in Lenihan, J. and Fletcher, W.W. (eds.) Environment and Man Blackie, Glasgow.

GARDNER, M.J., CRAWFORD, M.D. and MORRIS, J.N. (1969): Patterns of mortality in middle-age in the county boroughs of England and Wales British Journal of Preventive and Social Medicine 23, 133-140.

GREENBERG, B.G. (1963): Is soft water dangerous? Journal of the American Medical Association 184, 85-86.

HADDEN, D.R. (1974): Soft water and ischaemic heart disease, a study of two towns in N. Ireland Ulster Medical Journal 43, 45-46.

HAMILTON, E.I. (1973): Lead in drinking water British Medical Journal 2, 664-665.

HART, J.T. (1970): The distribution of mortality from coronary heart disease in South Wales Journal of the Royal College of General Practitioners 19, 258-268.

HERNBERG, S. (1973): Health hazards of persistent substances in water
WHO Chronicle 27, 192-193.

HEYDEN, S.J. (1976): The hard facts behind the hard water theory and ischaemic heart disease
Journal of Chronic Disease 29, 149-157.

HILL, M.J. (1973): Bacteria, nitrosiamines and cancer of the stomach
British Journal of Cancer 28, 562-567.

HOLLINGSWORTH, D.F. (1956): Nutritional requirements and food fortification: dietary supplies of calcium and iron
Chemistry and Industry 24, 1510-1512.

IBRAHIM, M.A. (1977): Drinking water and carcinogenesis: the dilemmas
American Journal of Public Health 67, 719-720.

JOOSSENS, J.U. (1973): Salt and hypertension, water hardness and cardiovascular death rate
Triangle 12, 9-16.

KNOX, E.G. (1973): Ischaemic heart disease, mortality and dietary intake of calcium
Lancet 1, 1465.

KNOX, E.G. (1974): New etiologies for ischaemic heart disease
American Heart Journal 88, 809-811

KOBAYASHI, J. (1957): On geographic relationships between the chemical nature of river water and death rate from apoplexy
Berichte des Ohara Instituts für Landwirtshaftliche Biologie 11, 12-21.

LINDEMANN, R.D. and ASSENZO, J.R. (1964); Correlations between water hardness and cardiovascular deaths in Oklahoma counties
American Journal of Public Health 54, 1071-1077.

LOWE, C.R., ROBERTS, C.J. and LLOYD, S. (1971): Malformations of central nervous system and softness of local water supplies
British Medical Journal 2, 357-361.

MASIRONI, R. (1969): Trace elements and cardiovascular diseases
Bulletin of the WHO 40, 305-312.

MASIRONI, R. (1970): Cardiovascular mortality in relation to radioactivity and hardness of local water supplies in the USA
Bulletin of the WHO 43, 687-697.

MASIRONI, R. (1974): Trace elements in relation to cardiovascular diseases
WHO Offset Publication Number 5, Geneva.

MASIRONI, R., KOIRTYOHANN, S.R., PIERCE, J.O., SCHAMSCHULA, R.G. (1976): Calcium content of river water, trace element concentrations in toenails, and blood pressure in village populations in New Guinea
The Science of the Total Environment 6, 41-53.

MEYERS, D. (1975): Ischaemic heart disease and the water factor- a variable relationship
British Journal of Preventive and Social Medicine 29, 98-102.

MILLER, R.W. (1976): Carcinogens in drinking water
Pediatrics 57, 462-464.

MORGAN, J.M. (1972): Tissue copper and lead content in ischaemic heart disease
Archives of Environmental Health 25, 26-28.

MORRIS, J.N., CRAWFORD, M.D. and HEADY, J.A. (1961): Hardness of local water supplies and mortality from cardiovascular disease in the County Boroughs of England and Wales
Lancet 1, 860-862.

MORRIS, J.N., CRAWFORD, M.D. and HEADY, J.A. (1962): Hardness of local water supplies and mortality from cardiovascular disease
Lancet 2, 506-507.

MORTON, W.E. (1971): Hypertension and drinking water constituents in Colorado
American Journal of Public Health 61, 1371-1378.

MORTON, W.E. (1974): Can epidemiology elucidate the water story?
American Journal of Epidemiology 100, 85-86.

MORTON, M.S., ELWOOD, P.C. and ABERNETHY, M. (1976): Trace elements in water and cogenital malformations of the central nervous system in S. Wales
British Journal of Preventive and Social Medicine 30, 36-39.

MULACHY, R. (1964): The influence of water hardness and rainfall on the incidence of cardiovascular and cerebrovascular mortality in Ireland
Journal of the Irish Medical Association 55, 17-18.

MULACHY, R. (1966): The influence of water hardness and rainfall on cardiovascular and cerebrovascular mortality in Ireland
Journal of the Irish Medical Association 59, 14-15

MUSS, D.L. (1962): Relationship between water quality and deaths from cardiovascular disease
Journal of the American Water Works Association 54, 1371-1378.

NERI, L.D., HEWITT, D. and MANDEL, J.S. (1971): Risk of sudden death in soft water areas
American Journal of Epidemiology 94, 101-104.

NERI, L.C., HEWITT, D. and SCHREIBER, G.B. (1974): Can epidemiology elucidate the water story?
American Journal of Epidemiology 99, 75-88.

NERI, L.C., HEWITT, D., SCHREIBER, G.B., ANDERSON, T.W., MANDEL, J.S. and ZDROJEWSKY, A. (1975): Health aspects of hard and soft waters
American Water Works Association 67, 403-409.

NERI, L.C., MANDEL, J.S., and HEWITT, D. (1972): Relation between mortality and water hardness in Canada
Lancet 1, 931-934.

OGGLESBY, P. (1966): Further views on the water hypothesis
Illinois Medical Journal 127, 183-186.

PAGE, T., HARRIS, R.H. and EPSTEIN, S.S. (1976): Drinking water and cancer mortality in Louisianna
Science 193, 55-57.

PETERSON, D.R., THOMPSON, D.J. and NAM, J.M. (1970): Water hardness, arteriosclerotic heart disease and sudden death
American Journal of Epidemiology 92, 90-93.

Registrar General (1951, 1961, 1971): Statistical review of England and Wales Part 1, Medical
HMSO, London.

Registrar General (1971): Decennial Supplement, England and Wales, 1961 Occupational Mortality Tables
HMSO, London.

REID, D.D. (1973): Arteriosclerotic disease in relation to the environment in
Howe, G.M. and Loraine, J.A. (eds.) Environmental Medicine,
Heinemann, London.

ROBERTS, C.J. and LLOYD, S. (1972): Association between mortality from ischaemic heart disease and rainfall in South Wales and in the County Boroughs of England and Wales
Lancet 1, 1097-1093.

RONCEK, D.W. (1975): Density and crime: a methodological critique
American Behaviourial Scientist 18, 843-860.

SCHROEDER, H.A. (1958): Degenerative cardiovascular disease in the Orient: hypertension
Journal of Chronic Disease 8, 312-333.

SCHROEDER, H.A. (1960): Relationships between mortality from cardiovascular disease and treated water supplies - variations in states and 163 largest municipalities of the United States
Journal of the American Medical Association 172, 1902-1908

SCHROEDER, H.A. (1960): Relations between hardness of water and death rates from certain chronic and degenerative diseases in the United States
Journal of Chronic Disease 12, 586-591.

SCHROEDER, H.A. (1966): Municipal drinking water and cardiovascular death rates
Journal of the American Medical Association 195, 81-85.

SCHROEDER, H.A. (1969): The water factor
New England Journal of Medicine 280, 836-837.

SCHROEDER, H.A. (1974): Cardiovascular mortality, municipal water and corrosion
Archives of Environmental Health 28, 303-311.

SHACKLETTE, H.T., SAUER, H.I., MIESCH, A.T. (1972): Distribution of trace elements in the environment and the occurrence of heart disease in Georgia
Bulletin of the Geological Society of America 83, 1077-1082.

SHAPER, A.G. (1974): Soft water, heart attacks and stroke
Journal of the American Medical Association 230, 130-131.

SHARRETT, A.R. (1977): Water hardness and cardiovascular disease. Elements in water and human tissues
Science of the Total Environment 7, 217-226.

SHARRETT, A.R. and FEINLEIB, M. (1975): Water constituents and trace elements in relation to cardiovascular diseases
Preventive Medicine 4, 20-36.

SMITH, V.K. (1977): The economic consequences of air pollution
Ballinger, Cambridge, Mass.

STEWART, A. (1975): Fact and fancy about drinking water
American Journal of Public Health 65, 1111.

STITT, F.W., CLAYTON, D.G., CRAWFORD, M.D. and MORRIS, J.N. (1973): Clinical and biochemical indicators: disease among men living in hard and soft water areas
Lancet 1, 122-126.

STOCKS, P. (1973): Mortality from cancer and cardiovascular disease in the county boroughs of England and Wales classified according to sources and hardness of water supplies
Journal of Hygiene (Cambridge) 71, 237-252.

STRONG, J.P., CORREA, P. and SOLBERG, L.A. (1968): Water hardness and atherosclerosis
Laboratory Investigation 18, 620-622.

TAKAHASHI, E. (1967): Geographic distribution of mortality rate from cerebrovascular disease in European countries
Tohoku Journal of Experimental Medicine 92, 345-378.

TURNER, R.C. (1962): Radioactivity and hardness of drinking waters in relation to cancer mortality rates
British Journal of Cancer 16, 27.

VOBECKY, J. and SHAPCOTT, D. (1975): Ischaemic heart disease and elements in water
Canadian Medical Association Journal 113, 922-925.

VOORS, A.W. (1970): Lithium in the drinking water and atherosclerotic heart death: epidemiological argument for protective effect
American Journal of Epidemiology 92, 164-171.

VOORS, A.W. (1971): Minerals in municipal water and arteriosclerotic heart disease
Americal Journal of Epidemiology 93, 259-266.

WALL, G. (1976): Some contemporary problems in research on air pollution
Progress in Geography 8, 95-131.

WEST, R.R., LLOYD, S., and ROBERTS, C.J. (1973): Mortality from ischaemic heart disease - association with weather
British Journal of Preventive and Social Medicine 27, 36-40.

WINTON, E.F. and McCABE, L.J. (1970): Studies relating to water mineralization and health
Journal of the American Water Works Association 62, 26-30.

WOLF, H. (1976): Softened water need not be a danger
American Water Works Association Journal 16, 15-23.

C H A P T E R 7

MORTALITY VARIATIONS AMONG THE COUNTY BOROUGHS OF ENGLAND AND WALES

'Our most urgent needs are those numerical indices of places of health which are a better Judge of Ayres than the conjectured notions we commonly read and talk of; our aim, concerned with prevention, must be to observe and measure the influence of each factor that promotes disease' William Petty (1623-1687) quoted in Reid (1958).

'An apparently wild .... observation is a signal that says here is something from which we may learn a lesson, perhaps of a kind not anticipated beforehand, and perhaps more important than the main object of the study' Kruskal (1960, 1).

INTRODUCTION

The principal aim of this chapter is to attempt to disprove that a relationship exists between water hardness and heart disease. In particular, the study will re-examine the mortality patterns previously analysed by Gardner (1973) to see whether the observed strong negative relationship becomes spurious when additional explanatory variables are controlled and a different statistical methodology is employed. To achieve this objective requires a well-specified model of mortality variation, but it would be foolhardy to claim that it is possible to use this model to determine fully the causes of disease. Some causal variables may have no discernible effect at this aggregate level and therefore the true influence of such variables can only be determined by a study specifically designed to disprove such relationships. For example, a researcher may suggest that the frequently observed relationship between cigarette smoking and mortality is spurious with the true causal association being between mortality and the constant raising of a hand to the mouth. These competing explanations cannot be evaluated at an aggregate level (not least because of a lack of data) but the water hypothesis, which derives much of its support from aggregate areal studies, can possibly be falsified at this level. Consequently, while we cannot prove which variables cause disease we can, in a truly Popperian fashion, attempt to disprove the water hypothesis. If the strength of the relationship remains the same in this re-analysis, greater faith can be placed in the association and the hypothesis can be 'tentatively entertained' and subjected to further falsification (Chapter 1). If the relationship weakens substantially, it may be permissible to regard it as spurious. It must be emphasised that either result is important, and it is on this basis that the study is offered as a contribution to the understanding of disease causation.

The chapter begins with a consideration of the dependent variable (areal mortality); here the accuracy of data on death certification is discussed as is the geographical distribution of disease. The argument then proceeds to the explanatory variables, at which point a summary of the evidence for believing that such variables are truly causal is provided. The various stages of an exploratory analysis are then outlined as a prelude to the presentation and evaluation of the analytical results. Finally, an 'experiment of opportunity' is conducted to examine the effect of changing water supplies on cardiovascular mortality.

MORTALITY : THE DEPENDENT VARIABLE

Accuracy of death certification

This study is based on routine official mortality statistics and, because it is obviously important that the dependent variable be measured as accurately as possible, we need to examine how these data are registered and collected. Mortality data are derived from the medical certificate of the cause of death issued by a general practitioner, together with the information given to the registrar by an informant who previously was acquainted with the person who has died. In the first part of the death certificate the practitioner records the disease or condition directly causing death, giving antecedent conditions if appropriate. In the second part he indicates other significant conditions contributing to the death, but which were not related to the disease or condition actually causing death. The informant is expected to provide the following information about the deceased: the name, sex, marital status, the date and place of birth, his or her final occupation and usual address and, finally, the date and place of death. A copy of the form is sent to the Office of Population Censuses and Survey (previously the Office of the Registrar General) where

the data are collated by trained staff. The underlying cause of death is coded according to a standard manual prepared by the World Health Organisation and the address, sex, marital status and place of death of the deceased are also classified and coded. These data are then used to produce extensive annual reports, in addition to which it is possible to approach the Registrar General to obtain specific unpublished information.

All routine data-collection systems are open to the challenge that the material is not completely accurate, and it may be suggested that there is a potential error at each stage in the following chain of events between the diagnosis of the cause of death by the clinician and the publication of mortality data:

diagnosis → certification → coding → processing → publicatio

However, it is generally thought that the greatest inaccuracy occurs with the identification of the true cause of death. Alderson (1965) compared the information available from a detailed case history with the wording on the death certificate for 2,243 adult deaths. For 90 per cent of the cases he found the death certificates to be an accurate representation of the true cause of death. While this figure may be regarded as reasonably satisfactory it must be added that the accuracy of mortality data can be further improved by excluding deaths over a certain age limit. Such elderly deaths are likely not to be the result of a single underlying condition but an outcome of a complex interaction of morbid conditions as the total body system gradually degenerates. Moreover, deaths in the younger age groups are more likely to have their certificates based on autopsies and coroners' reports, thereby potentially further improving the accuracy of the data. Consequently, following Gardner (1973), deaths over the age of 75 will be excluded from the analysis in this chapter and, in order to replicate the analysis, deaths will be divided into two age categories (45-64 and

65-74) and two time periods (1948-54, 1958-64 for the younger age group; 1950-54, 1958-64 for the older age group). These particular time periods have been chosen to allow for the collection of a sufficient number of deaths to facilitate a meaningful statistical analysis; but they are not long enough to mask temporal changes in either the accuracy of the County Borough populations (based on Census data for 1951 and 1961) nor the degree and geographical distribution of the mortality patterns. Finally, while the analysis will be conducted for a number of different causes of death, a study will also be made of mortality from all causes. This is thought to be a very accurate statistic because while there are approximately 500,000 deaths per year only about 100 persons disappear and possibly die without a death certificate being obtained (Alderson, 1976).

In summary, the observed number of deaths from different diseases can be expected to give a reasonably accurate guide to the true mortality pattern and, as Howe (1970, 3) has written

> 'the extent of this error must not be exaggerated. In the younger age groups it is relatively small, and it is only in old age when several different physiological systems are collapsing at the same time that the error is appreciable.'

Indeed, for one researcher (Murray, 1962) the England and Wales mortality data are 'excellent' (p 130), 'accurate' (p 134) and 'reliable' (p 134). Moreover, any possible remaining inferential error can be alleviated by replicating the analysis for different age groups and different time periods.

Geographical pattern

The mapping of medical data has received considerable research attention by medical geographers. Howe (1970), for example has mapped Standardised Mortality Ratios for the Metropoliton Boroughs, the County Boroughs and the 'urban'

and 'rural' districts of the United Kingdom counties for two time periods (1954-58 and 1959-63). The SMR is a means of overcoming the problem of areas having different age structures and while a SMR of 100 represents the national mortality rate a value of 200 indicates that an area has a mortality rate which is twice the national average. Similar maps for England and Wales have also been presented by Murray (1962) for the period 1948-57. Not only have geographers produced such maps, they have also engaged in various studies to determine the optimum method for portraying the data. For example, Armstrong (1969) has urged medical geographers to determine appropriate class intervals by calculating standard deviations about a mean, while Forster (1966) and Hunter and Young (1971) have suggested that medical maps should be drawn not on an areal base but on a demographic one. With this latter example each administrative area is drawn not in terms of its areal extent but in relation to the size of the population; those areas with the largest absolute population having the greatest visual impact.

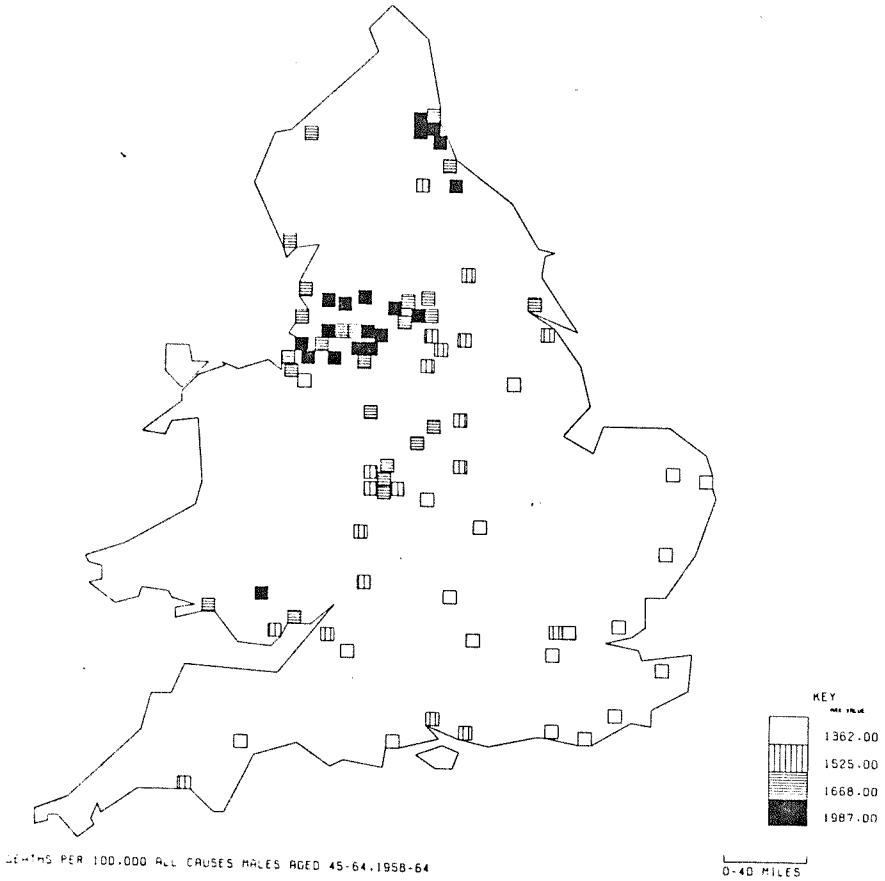
The maps presented in Figures 7.1 to 7.4, however, have a number of features that distinguish them from previous research. Firstly, and most importantly, they show the geographical pattern for a specific age range and consequently they can be expected to be a more reliable estimate of the true underlying pattern than maps such as those of Howe (1970), which are based on SMRs for all age groups. Secondly, the class intervals of the maps have been chosen by the automatic technique of nested means. To calculate such values the mean of the data is first calculated and the data are divided into two groups above and below the mean. The means are then determined for each of these groups and the three means are then adopted as class intervals. While the Armstrong (1969) method of using standard deviations is appropriate for normal (Gaussian) data, it is not a particularly useful statistic for mortality rates which can be expected to

be skewed (Jones, 1978). The nested mean, in contrast, will produce class intervals with a geometric pattern for highly skewed data, while for uniformly distributed and normally distributed values it will produce, respectively, equal-interval classes and intervals based on standard deviations. Standard deviations, therefore, are a special case of nested means and the latter technique is a more general method and is to be preferred. The third distinctive feature of the present maps is the decision not to use a demographic base. Such maps are useful when the number of deaths in some areas are small and unreliable (for they give greater visual prominence to large population areas) but such cartograms are frequently visually grotesque, whilst the loss of contiguity results in a failure to convey a clear spatial impression of the data. It is thought that the number of deaths in each County Borough is sufficiently large not to warrant the use of cartograms and consequently each County Borough is represented by a square of identical size. The final feature that distinguishes the present maps from previous efforts to map mortality patterns in this country is that they have been produced by automatic computer methods.[1]

Before proceeding to describe the geographical variations in mortality displayed by these maps two important points must be emphasised. The maps are based on administrative areas and, while true 'regions' of specific disease mortality should exist within the general areas enclosed by the administrative boundaries, they are unlikely to be co-extensive with them. Secondly, the maps should not be used to make general inferences; they cannot be used to make statements such as 'northern England has a worse mortality rate than southern England'. The intricate system of administrative areas in Yorkshire and Lancashire may give a false impression of the true overall pattern of mortality, and inferences should therefore be confined to the County Boroughs per se (Haggett, Cliff and Frey, 1977, 351).

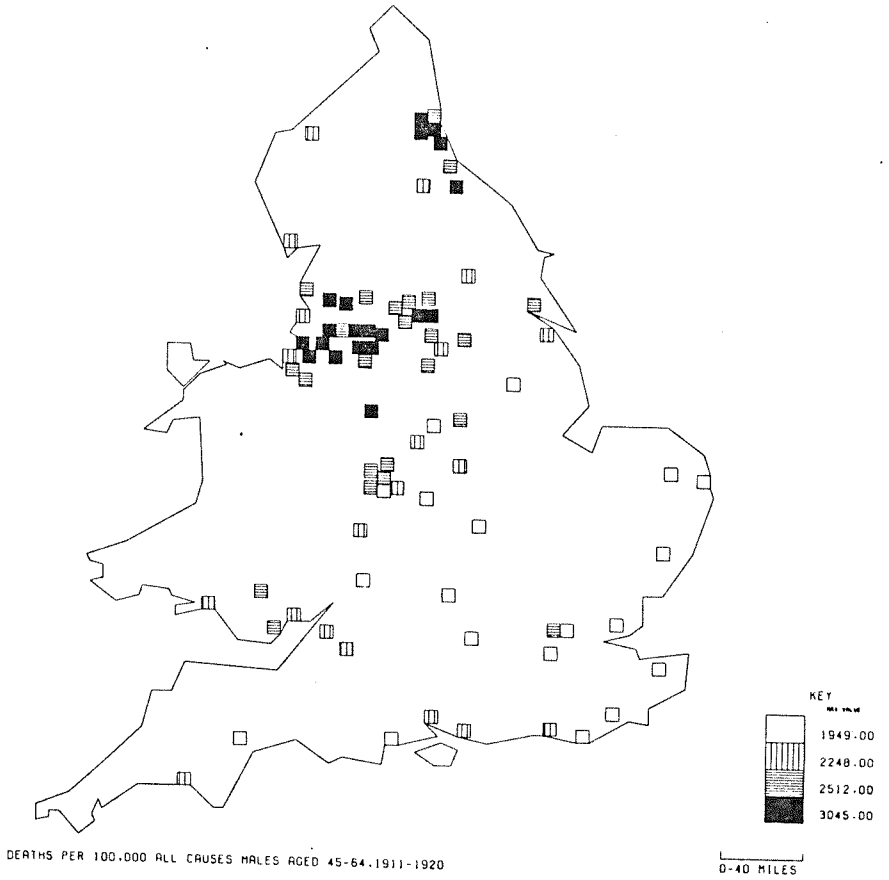
A large number of maps could have been presented, but a selection (Figures 7.1 to 7.4) has been chosen to convey the significant features of the geographical variations of mortality amongst the County Boroughs. In particular, a number of comparisons are made for deaths for all causes to determine whether the geographical pattern remains consistent when different time periods and different age groups are examined. If the maps of overall mortality do not show a strong and consistent geographical pattern it may be concluded that the distributions are random; if this is the case it would be foolhardy to attempt to explain the pattern displayed. But if the pattern remains relatively constant then it is essential to discover what forces have determined the distribution of mortality from all causes. In addition to overall mortality, the figures also examine the geographical pattern for the five diseases that account for most of the deaths in England and Wales (lung cancer, bronchitis, coronary heart disease, vascular lesions of the central nervous system and cardiovascular disease). If each of these diseases has a distinctive pattern it can be suggested that each possesses a specific aetiology. Moreover, if this is the case, it can also be suggested that the explanatory variables will be related in a different manner to each cause-specific disease, and it is therefore essential that a separate regression model be developed for each type of mortality.

Figure 7.1a can be expected to be a highly accurate and meaningful map of mortality variation because it is based on deaths from all causes for men in middle age (45-64 years). It is intended to use this particular map as a base against which comparisons are to be made for different time periods, age groups and causes.[2] The distribution of male mortality rates for all causes 1958-64 (Figure 7.1a) displays a striking geographical pattern, with the northern and western County Boroughs having generally high rates of mortality.



DEATHS PER 100,000 ALL CAUSES MALES AGED 45-64, 1958-64

FIGURE 7.1 (A) ALL CAUSES, 1958-64



DEATHS PER 100,000 ALL CAUSES MALES AGED 45-64, 1911-1920

(B) ALL CAUSES, 1911-1920

The five County Boroughs with the highest rates (Salford, Manchester, Liverpool, Middlesborough and Merthyr Tudful) are all towns in the north and west, while the five lowest rates (Croydon, Eastbourne, Reading, Norwich and Ipswich) are located in the south and east. Salford, with the worst mortality of any County Borough, had a rate that was twice as high as the lowest County Borough, Ipswich. In Salford 2 per cent of males aged 45-64 died annually while the comparable rate for Ipswich was only 1 per cent.

Figure 7.1b examines death from all causes for the period 1911-1920. Comparison with Figures 7.1a reveals that the average annual death rate per 100,000 fell from 2248 in 1911-20 to 1525 in 1958-64.[3] Moreover, this improvement was accompanied by a contraction of the range of death rates experienced. Compared with the 1958-64 values quoted earlier (2 and 1 per cent), the highest annual rate (Liverpool) in 1911-1920 was over 3 per cent, while the lowest rate (Croydon) was 1.6 per cent. Important though these trends were, however, the contrasts between Figures 7.1a and b should not be overstressed. The most remarkable feature of the two maps is, in fact, the similarity of the geographical distributions, and it is clear that the pattern of male mortality (45-64 years) changed little during the first half of this century. Indeed, four of the five towns with the highest rates in 1958-64 (Liverpool, Manchester, Salford and Middlesborough) were among the worst ten County Boroughs in 1911-20, while each of the five towns with the lowest rates in 1958-64 were in the lowest ten for 1911-1920.

Figure 7.2a displays another map of overall mortality but, in comparison with 7.1a, it examines geographical variations for an older age group (65-74 years as opposed to 45-64 years). Contrasting the two maps, it can be seen that Tyneside shows a lower (but still relatively high rate) for the older age group than the younger equivalent. Moreover, Merthyr Tudful and Swansea have higher rates for the older

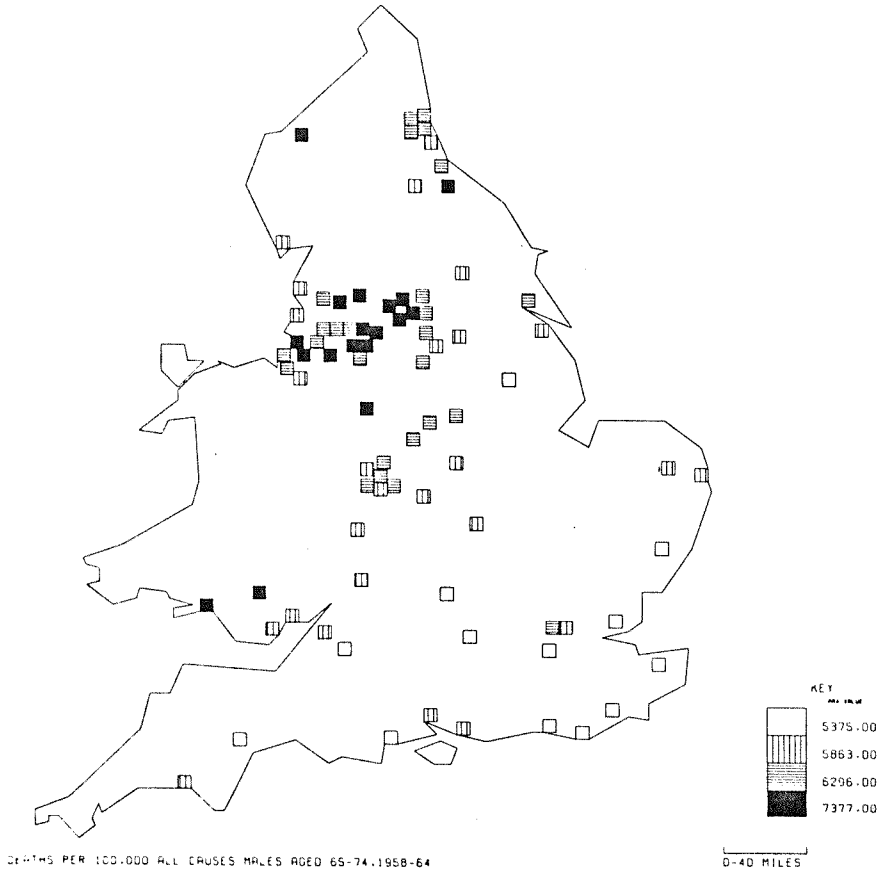
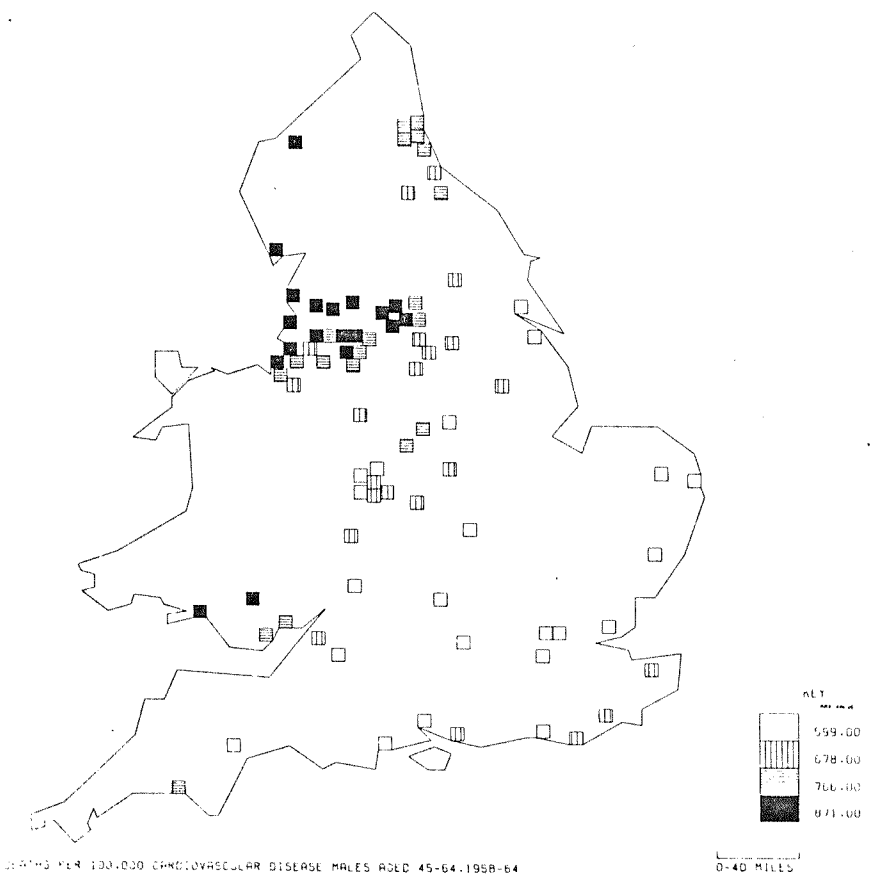


FIGURE 7.2 (A) ALL CAUSES, 65-74 YEARS



(B) CARDIOVASCULAR DISEASE

age group. But again the contrasts should not be over-emphasised: both maps have very similar geographical distributions, with those County Boroughs to the north and west of a line from the Severn Estuary to the Humber having a generally unfavourable mortality experience for both age groups.

In summary, therefore, mortality maps for different age groups and time periods reveal a remarkably consistent pattern, with the highest rates being found in Merseyside, south-east Lancashire and the West Riding of Yorkshire, while the lowest occurred in East Anglian and southern County Boroughs. The West Midlands conurbation experienced average rates in Figures 7.1a, 7.1b and 7.2a while only the Tyneside conubation and the South Wales County Boroughs displayed a variable mortality rate in these comparisons.

The discussion so far has only considered death from all causes, and it remains to be seen whether each type of disease will display a distinctive pattern, thereby suggesting that each possesses a specific aetiology. Figure 7.2b examines the geographical distribution of cardiovascular mortality. As this disease accounts for more deaths than any other, it can be expected that Figure 7.2b will closely resemble Figure 7.1a. While the geographical distributions of cardiovascular and overall mortality do indeed show many similarities, there are some differences: Tyneside and the West Midlands conurbation show relatively lower levels of cardiovascular mortality and, conversely, the South Wales Boroughs show relatively higher rates. The differences between overall and cause-specific mortality become more pronounced when one examines two sub-categories of cardiovascular disease - vascular lesions of the central nervous system (strokes) and coronary heart disease. A comparison of vascular lesions of the CNS (Figure 7.3a) and all-causes mortality (Figure 7.1a) reveals that the Merseyside conurbation has relatively low mortality from strokes,

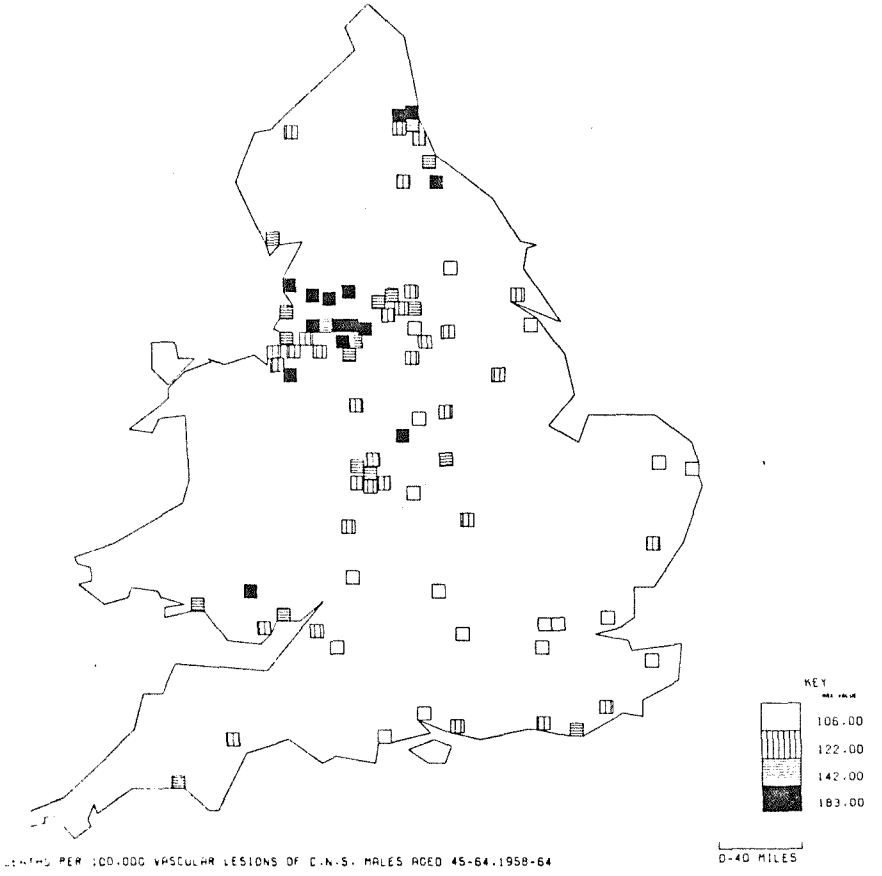
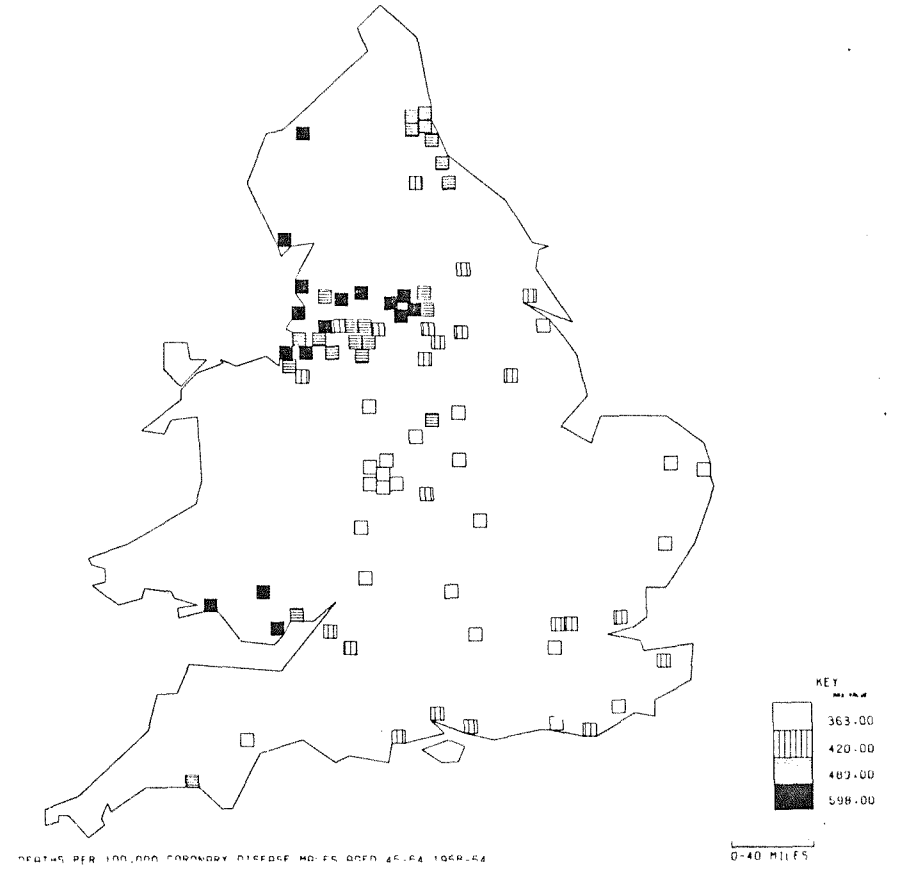


FIGURE 7.3 (A) VASCULAR LESIONS OF THE CENTRAL NERVOUS SYSTEM

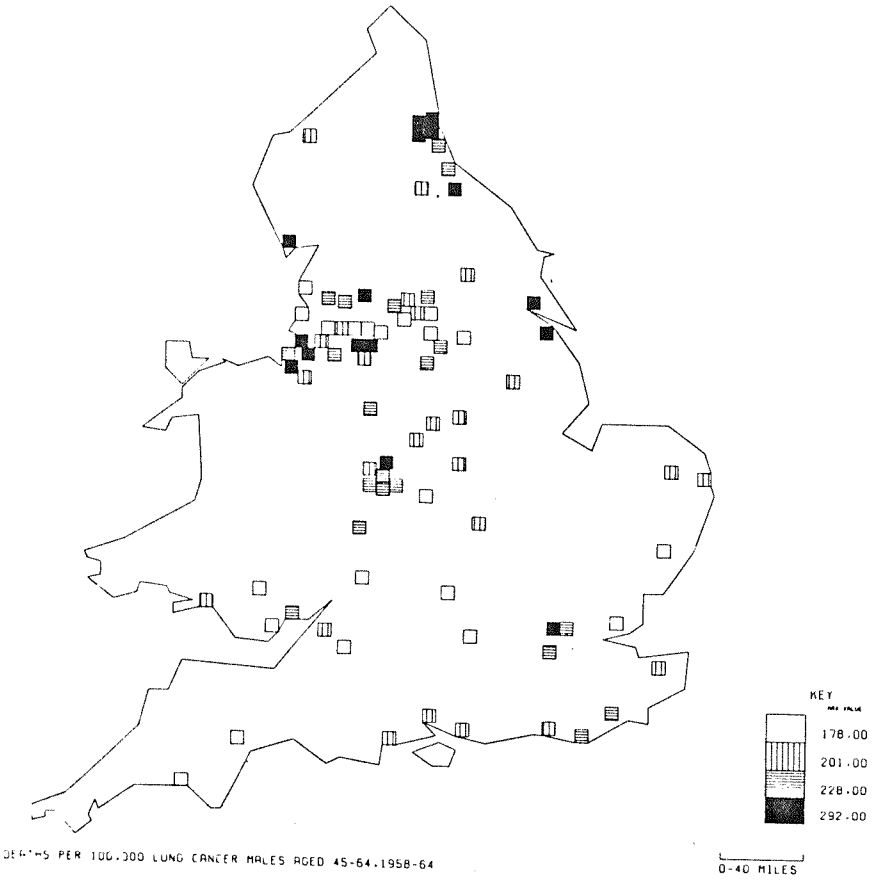


(B) CORONARY HEART DISEASE

Tyneside displays a more mixed pattern and a number of 'northern' towns (Barnsley, Derby, Grimsby, York) have uncharacteristically low rates. For coronary heart disease (Figure 7.3b) the contrast with overall mortality is even more marked. In general, the lowest rates are found not in southern England but in 'middle' England and East Anglia, with Burton, Walsall, Gloucester, Dudley, Wolverhampton, Great Yarmouth and Ipswich being among the ten County Boroughs with the most favourable rates. However, it also must be stressed, as with overall mortality, that cardiovascular disease exhibits strong geographical patterning. Uniformly low rates are found for the West Midlands conurbation, relatively high rates are experienced by Tyneside, while the West Yorkshire and Glamorgan County Boroughs suffer very high rates from heart-disease mortality.

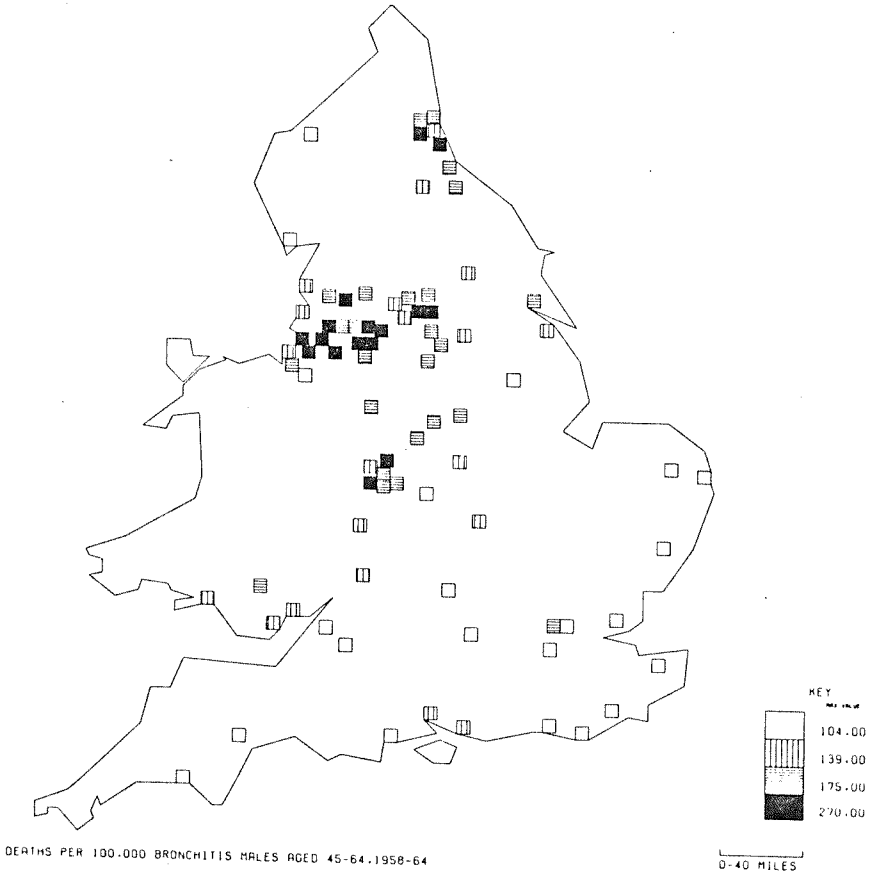
Lung cancer also shows a different spatial pattern to that of overall mortality when the same age group and time period are compared (Figures 7.4a and 7.1a). Many towns in Lancashire and the West Riding of Yorkshire have relatively low lung-cancer rates and, most unusually, one southern County Borough (West Ham) is ranked the ninth highest of all County Boroughs. The map of bronchitis (Figure 7.4b), however, displays a pattern that is similar to overall mortality, with the highest rates being concentrated in the County Boroughs of Lancashire and the West Riding of Yorkshire. But even with this disease there are some contrasts with overall mortality; to take one prominent example, the West Midlands conurbation has relatively high rates, with very high rates being experienced in Walsall and Dudley. Moreover, the differences in the mortality rate between towns are particularly large; when all deaths are considered the highest rate is generally twice the lowest, but with bronchitis the difference between the worst (Salford) and the best (Eastbourne) is a factor of five.

Consideration of all the maps in Figures 7.1 to 7.4 leads to the conclusion that there are considerable differences in death rates among the County Boroughs, differences that can be found for separate age groups, two time periods and



DEATHS PER 100,000 LUNG CANCER MALES AGED 45-64,1958-64

FIGURE 7.4 (A) LUNG CANCER



DEATHS PER 100,000 BRONCHITIS MALES AGED 45-64,1958-64

(B) BRONCHITIS

specific causes of death. Moreover, these differences appear non-random. County Boroughs close to each other tend to have similar values and, if we are to improve our understanding of disease causation, it is essential to discover what forces have determined these geographical patterns. In addition, while some towns have generally unfavourable rates for all the major diseases, some County Boroughs have high rates for one disease and low rates for another. For example, West Ham has the third lowest incidence of death from strokes, Wigan has the tenth lowest rate for lung cancer, Merthyr Tudful has the third lowest rate for lung cancer and yet these same towns are ranked ninth, eighth and first for cardiovascular mortality. These contrasts suggest that to some extent each disease has its own aetiology and an important task, therefore, is to determine which explanatory variables are associated with which disease.

EXPLANATORY VARIABLES

Having considered the dependent variable in some detail, attention now turns to the explanatory variables. A critical reading of a vast epidemiological literature has revealed a large number of variables that have been suggested to account for the geographical variations in mortality, and the following discussion attempts to describe and evaluate the evidence for believing that each variable truly causes disease. Sources of data for every explanatory variable are also considered and, because of the need to examine changes over time and allow for 'lagged' relationships, there is a discussion of the collection of suitable data for 1931, 1951 and 1961.

Density and overcrowding

The reforming medical officers of the last century attributed a large number of deaths to overcrowding. For the medical officer of The Strand, writing in 1856

> 'the main cause to which we must attribute the high mortality is the close packing and overcrowding which exists throughout the district .... Overcrowding and disease mutually act and re-act upon each other'
> (Quoted in Wohl, 1973, 612).

While such early writings dealt in the main with infectious disease, the idea that morbidity and mortality are causally related to high density has been extended to include chronic disease. A great deal of the support for this view has come from laboratory experiments with animals. In 1962 Calhoun published a well-known study examining the effects of overcrowding on rats. In his experiments the rats were provided with adequate food and water but kept at a much higher population density than in their natural habitat. Calhoun found that they suffered an increased mortality rate and exhibited aggressive behaviour and sexual aberrations. Subsequent experiments have produced similar results for other animals, with monkeys, horses and shrews having an increased mortality rate, while high levels of density supposedly lead to homosexuality in fish and reduced fertility in elephants (Galle, Gove and McPherson, 1972). It appears that there is strong evidence for accepting a true causal relationship between mortality and overcrowding for animals but for man the results are much less conclusive. Boots (1979) has provided a comprehensive review of recent research and has identified 18 studies which have used multiple regression to examine the relationship at the aggregate level. Unfortunately, he is forced to make a familiar evaluation of this research

> 'Ecological studies are marked by an almost bewildering array of findings. Almost every possible combination of outcomes has been identified by one study or other and it is possible to marshall sufficient studies to support virtually any opinion concerning the relationship between density and disorders'
> (Boots, 1979, 33).

This is exactly the same problem that has troubled researchers enquiring into the water hypothesis, and Boots suggests that this bewildering array of results is a direct outcome of poor regression analysis.

> 'Often the existence of high multicollinearity ... leads to unstable and imprecise beta weights and partial correlation coefficients ... which makes their interpretation difficult, or sometimes impossible. Most often this problem is addressed by simply ignoring its existence.
>
> Finally there are some general problems associated with the selection of a regression technique. Most studies adopt linear regression procedures without considering the possibility of non-linear alternatives .... Similarly, at the conclusion of the analysis there is a singular disregard of residuals' (Boots, 1979, 45).

Again these are familiar arguments and obviously the relationship between density and disease needs to be thoroughly examined by exploratory techniques.

In the present study there are three measures of density and overcrowding. They are

(1) population density, which is the population of a County Borough divided by its area,

(2) persons per room, and

(3) the number of people living in a County Borough who live in overcrowded households as defined by the Census (that is over 1.5 persons per room)

The first two measures are irreducible ratios, and as a consequence of the discussion in Chapter 5 they are included in the model because interest focuses on these variables as ratios per se. Data have been derived from the Registrar General's Census Reports (1931, 1951 and 1961) for the three variables.

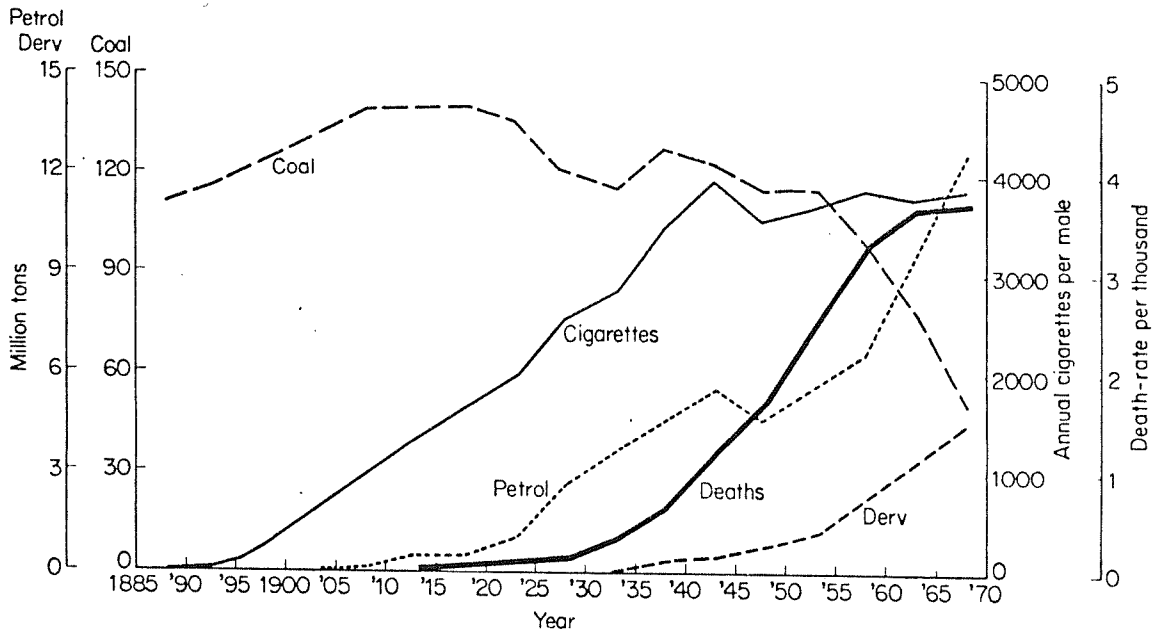
Cigarette smoking

Perhaps the variable that is most commonly associated with an increased risk of mortality is cigarette smoking. However, it must be remembered that the studies implicating

this variable are of relatively recent origin and there are still a number of researchers who vehemently argue that the relationship has not been proven beyond reasonable doubt.

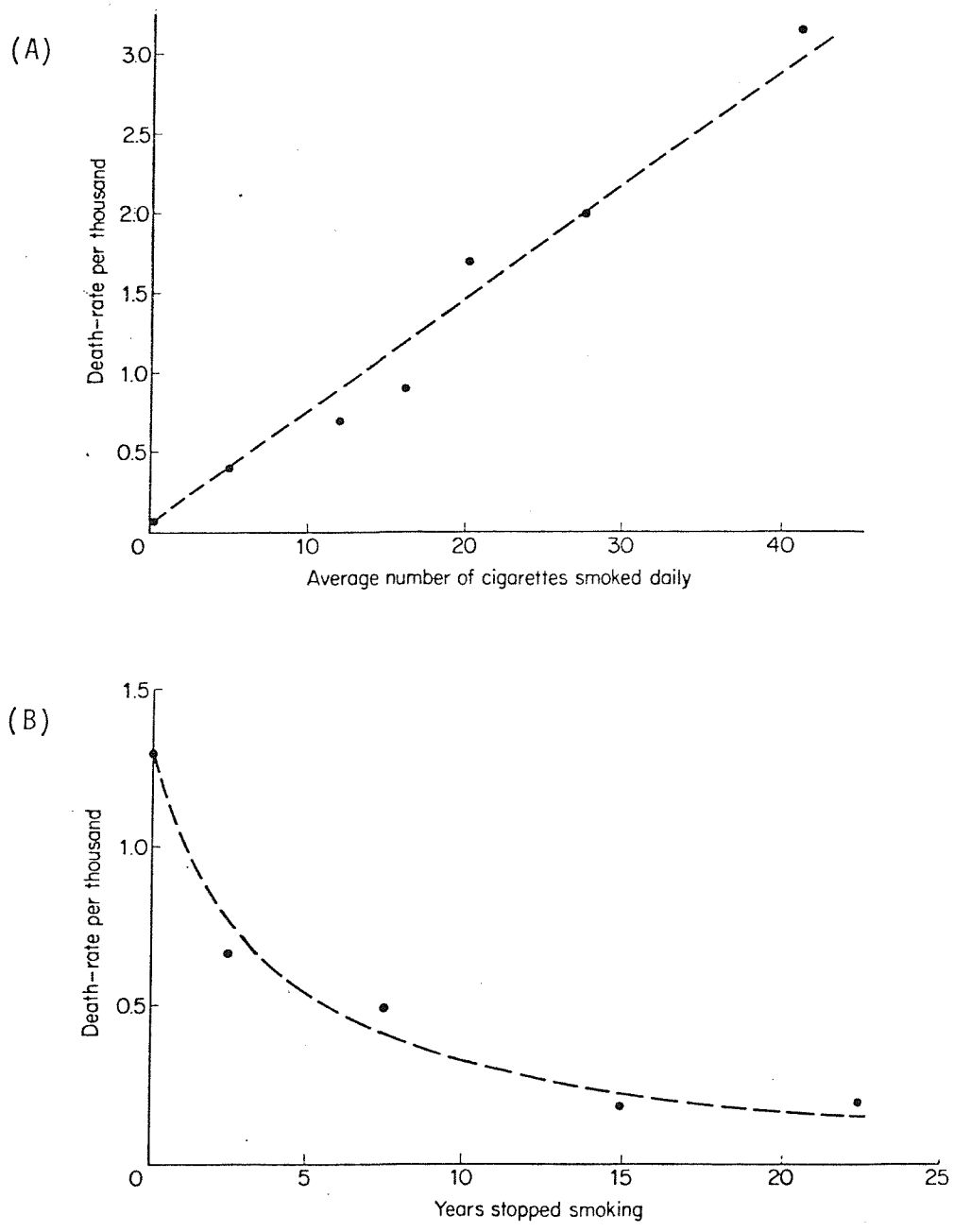
During the first half of the twentieth century the death rate from lung cancer rose steadily and by the late 1940's the mortality statistics suggested that there had been a dramatic rise in the number of deaths from this cause in the United Kingdom. Careful thought was given at this time to the possible factors that could have been associated with the observed rise and, in particular, attention was focused on atmospheric pollution and cigarette smoking. As Figure 7.5 indicates, it did not appear that atmospheric pollution from coal could have induced the rise, but the graph does indicate that increased consumption of petrol, diesel fuel and cigarettes (allowing for a twenty-year time lag) may have been the cause. In order to test the relationship between cigarette smoking and early death, Doll and Hill (1950, 1952, 1954, 1964) performed their classic studies on the aetiology of lung cancer. In their earlier work they arranged for patients with suspected lung cancer to be interviewed in hospital wards and a selection of patients with other conditions were also questioned. The results showed that the patients suffering from lung cancer tended to be heavy smokers and this finding was not associated with any of the other factors (occupation, social class, urban/rural residence, living near a gas works, use of different forms of heating, or history of previous respiratory illness) which were examined. Doll and Hill subsequently extended their study by conducting a prospective analysis of 60,000 doctors. In 1951 they wrote to all practitioners on the medical register in the United Kingdom and asked them to complete a brief questionnaire on smoking habits. Over the next ten years the Registrar General provided notification when a doctor died, and Figure 7.6a shows the strong linear relationship found between daily cigarette consumption and

FIGURE 7.5 LUNG CANCER MORTALITY (MALES AGED 60-64) AND ATMOSPHERIC POLLUTION IN THE PRESENT CENTURY



SOURCE: WALLER (1967)

FIGURE 7.6 LUNG CANCER AND CIGARETTE SMOKING



SOURCE: DOLL AND HILL (1964)

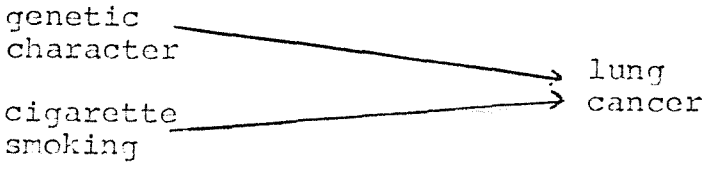
death rates from lung cancer standardised for age. An even more crucial point in their argument was the examination of the mortality among ex-smokers. As Figure 7.7b shows, there was a high mortality due to lung cancer for those doctors who had just given up smoking, with a lower mortality for those who had not smoked for twenty years. (Doctors who had not smoked for twenty years had a lung cancer mortality rate equivalent to non-smokers.) Following these particular studies by Doll and Hill there have been a large number of analyses which have implicated cigarette smoking as a causal variable in disease. These studies have also suggested that tobacco consumption is not only related to lung cancer but also to other diseases such as bronchitis and cardio-vascular disease.

Some researchers, however, have suggested that there is evidence for not accepting cigarette smoking as a cause of disease, one of the major protagonists in this debate being P.J.R. Burch. He argues that the studies that show a relationship are based on inadequate statistical research and for him

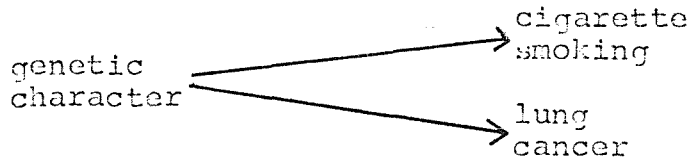
> 'the application of rigorous methods of statistical inference remains conspicuously absent from much of the literature' (Burch, 1973, 437).

Moreover, he points out that some studies find a <u>negative</u> relationship between smoking and disease and in particular, he cites 6 studies that have found such an association for Parkinson's disease. However, his major criticism is that numerous analyses have failed to take into account the genetic character of the people under study. He suggests that researchers have failed to distinguish between two models:

(1)



(2)



In the first model, both genetic character and cigarette smoking are related to disease, but in the second the association between smoking and lung cancer is spurious, with a person's genetic character pre-disposing him both to smoke and to die at an early age from cancer.

Clearly, despite the acceptance by the British government of the detrimental consequences of cigarette smoking, and their warning published on cigarette packets, the relationship still requires further testing and falsification. While it is not feasible to test Burch's two models at the aggregate level (it can be suggested that this can be adequatedly achieved only by the analysis of twins) it is at least possible to test whether the relationship holds at the aggregate level when a large number of other explanatory variables are held constant. Moreover, it may also prove interesting to determine which particular diseases have the strongest relationship with cigarette smoking. Unfortunately, routine data are not published on the geographical variations of cigarette smoking, but Gardner (1973) was able to derive crude levels of cigarette consumption for each County Borough in 1952 from the Tobacco Research Council and the Tobacco Manufacturers' Standing Committee. It has, however, not proved possible to collect comparable data for 1931 and 1961.

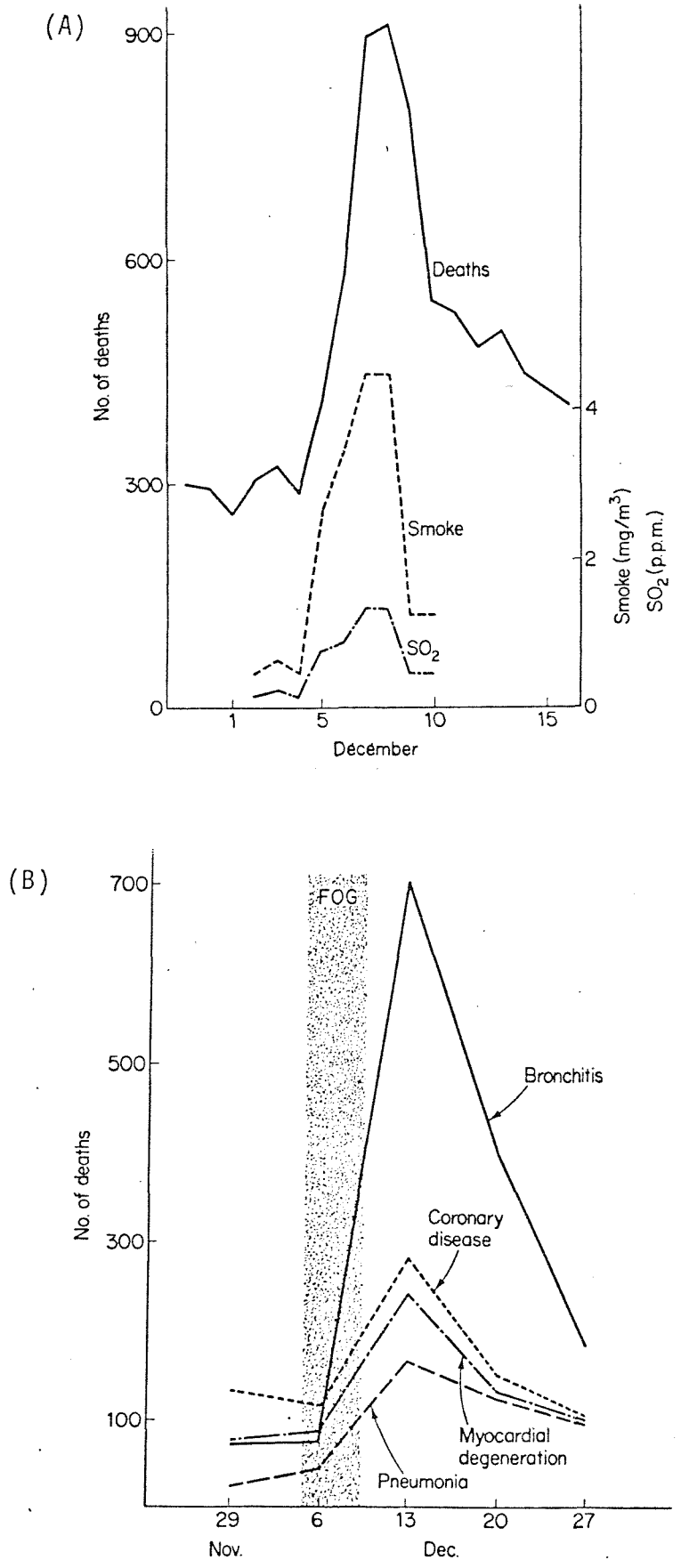
Air pollution

Air pollution has been implicated as an important factor in determining disease by a large number of studies, but it is important to distinguish between the short-term episodes of extremely high air pollution and the long-term effects of living in a polluted environment. The effects of

high air pollution are dramatically illustrated by an analysis of the London smog for 5-8th of December, 1952. Figure 7.7 presents a clear indication of the link between the level of atmospheric pollution and the number of deaths recorded daily during the smog. In particular, there was a major rise in deaths from bronchitis and, apart from motor-vehicle accidents and suicides, every other specific cause showed some rise.

While this example strongly suggests that pollution is a major immediate determinant of death, it does not provide evidence that atmospheric pollution is an initial cause of disease. Someone who is killed by a rise in the smoke level is likely to be already gravely ill. In this country a number of workers have examined the long-term effects of air pollution, among them Lave and Seskin (1970) who have studied data for the County Boroughs. They suggest (p 726) that reducing the level of air pollution to that enjoyed by the least-polluted area would lower the average bronchitis mortality rate by 70 per cent. Furthermore, they have claimed that similar, less extreme, improvements could be achieved for mortality from lung cancer, stomach cancer and cardiovascular disease. However, such conclusions are based on a model with only two explanatory variables (air pollution and population density or social class) and their use of regression analysis displays a conspicuous disregard for the assumptions of this technique. Moreover, Lave and Seskin's (1976) identical study of the effects of air pollution on the basis of SMSA data for the USA has been criticised by Smith (1976) for its lack of specification-error analysis. As discussed in Chapter 4 Smith, with a well-specified model, did not find a relationship between air pollution and disease, but his work can be criticised for its use of percentage data. Again there is conflicting evidence based on an inappropriate statistical methodology and there is a genuine need to undertake a study based on exploratory data analysis.

FIGURE 7.7 MORTALITY AND AIR POLLUTION: LONDON SMOG, DECEMBER, 1952



SOURCE: REPORTS ON PUBLIC HEALTH AND MEDICAL SUBJECTS NO. 95 HMSO (1954)

Air pollution data are very inadequate for this country, and this is particularly the case for 1951 and 1961. Lave and Seskin's study was based on data published by Stocks (1959) but unfortunately this is only available for 53 of the 83 County Boroughs. However, data have been presented by Daly (1959) for all the County Boroughs. His two indices (domestic and industrial fuel consumption) are based on the amount of coal used for these purposes in 1952. While these measures may appear very crude there are three reasons why they should be employed in this study. Firstly, there is simply nothing else available. Secondly, there is a strong association between the data of Daly and Stocks and a scatter plot shows a reasonably linear relationship. Finally, Daly's indices are highly correlated with mortality (for example, the correlation between the index of domestic fuel consumption and male all-causes death rate for 1958-64 is 0.74). Furthermore Gardner, Crawford and Morris (1969, 135) imply that this relationship is independent of other social indicators, and it therefore remains to be discovered whether the relationship holds when additional explanatory variables are used and when a different statistical methodology is employed.

Social class, diet and proxy variables

Stevenson, the 'inventor' of the Registrar General's social classification of 1921, which is the basis of the classification in use today, specifically developed his class categories for the analysis of mortality variations. In particular, he criticised previous studies for defining class solely in income terms, and ignoring habits and hygiene. Consequently, he designed his classification so that the highest social class had the lowest mortality, and there was a fairly regular increase, class by class, so that the lowest group had the highest mortality. This relationship, as Table 7.1 shows, is still to be found today. In particular, this relationship holds for different age groups

Table 7.1

Social class and mortality

(a) Mean annual death rate per 100,000 men by age groups and social class 1961.

Age group	Social class I	II	III	IV	V
45-64	535	545	708	734	1119
55-64	1699	1820	2218	2202	2912
65-74	4606	5160	6347	5702	6715

(b) SMR by disease and social class, 1961.

Disease	Social class I	II	III	IV	V
Bronchitis	28	50	97	116	194
Stomach cancer	62	63	101	114	163
Lung cancer	53	72	107	104	148
Stroke	86	89	101	98	135
Coronary disease	98	95	106	96	112

I : professional
II : intermediate
III: skilled
IV : partially skilled
V : unskilled

Source : Registrar General (1971)

and, in general terms, for different diseases, with the gradient for bronchitis being particularly marked.

It is important to emphasise that in Stevenson's original work the notion of social class was supposed to represent other variables such as diet and hygiene, and it is in this context that the variable is used in the present study. In particular, it is intended to use social class as a 'proxy' for diet. Unfortunately, while it has been suggested that diet accounts for a substantial part of the variation in mortality rates, data are not available for the County Boroughs. Moreover, the value of social class as a surrogate for diet cannot be determined, for again data are not published in a suitable form. However, as Table 7.2 illustrates, diet is related to income, and income is closely related to social class. Examining the relationship between income and diet in more detail, the highest income group has a relatively high consumption of meat, dairy produce, fresh fruit and fresh and frozen green vegetables. Conversely there is a relatively low consumption of sugar, potatoes and bread by this high-income group.

With regard to the relationship between diet and mortality, it has been suggested that high cholesterol diets (meat, fat and dairy produce) lead to a high blood cholesterol in humans and consequent premature death. The earliest evidence for this relationship comes from Russian laboratory studies of rats but, once again, there is conflicting evidence on the effect of diet on humans and some studies have failed to find any relationship.

For example, Groen et al (1962) have compared the mortality experiences of vegetarian Trappist monks and similarly cloistered Benedictines and found the incidence of cardiovascular disease to be the same in both orders, but much less than males outside monastries. Consequently, they suggested that something other than diet differentiates these

Table 7.2

Annual averages (1970) of household consumption of selected foods according to income class*

Income in pounds per week	oz's per person per week					
	Butter	Lard	Sugar	Potatoes	fresh vegetables	fresh fruit
+ 69	7	2	15	37	16	41
45-68	7	2	14	41	14	32
27-44	6	2	16	51	14	24
14-26	6	2	17	57	12	19
less than 14	5	2	18	56	11	17

* Not all groups are given : households with less £14 per week and no wage earner, and pensionable households are not included in this table.

Source : National Food Survey Committee (1973) Household food consumption and expenditure HMSO, London.

celibate men from the majority of the population! In Britain, according to Wynn and Wynn (1980) the theory that excessive consumption of saturated fats is the cause of ischaemic heart disease does not explain the distribution of the disease by class or by regions. On the basis of rates for regions and social class, they argue that the poorest socio-economic groups have the worst ratios for heart disease but the lowest consumption of fats, while Scotland has a relatively low level of fat intake but a high heart-disease mortality. Such simplistic analyses can be easily discredited, for other variables besides diet co-vary with social class. For example, people in the lowest social groups have the highest cigarette consumption (Todd, 1969) and therefore the detrimental effects of tobacco smoking may outweigh the beneficial effects of a low-fat diet for the lower social classes. Consequently, a more detailed analysis is required in which the effects of social class and cigarette smoking are considered simultaneously.

As a final example of the conflicting evidence on the importance of diet, one can compare the recommendations of two groups of dietary and medical experts (The Guardian, June 12th, 1980). The American Heart Association on the basis of thousands of animal studies and a great deal of epidemiological research, have been urging Americans to adopt a low-cholesterol diet since 1961, while the Food and Nutrition Board of the National Academy of Sciences, on the basis of the same evidence, write of 'no significant connection between cholesterol intake and cholesterol concentration in the body'. Clearly there is a need for an exploratory analysis into the effects of diet and, while it is impossible to achieve this in the present study, it is at least possible to test whether the relationship between social class and disease exists when occupation, smoking, air pollution and high-density living are statistically controlled.

While McCallum (1972) and Wickers (1972) have argued that it is preferable to use a poor surrogate variable than to ignore an unmeasured variable altogether, it is obviously also desirable that a surrogate should be chosen with care. The National Food Survey Committee (1973, 30) have considered the relationship between income and diet in Britain and they found

> 'the most striking feature of these analyses according to income level is the similarity between the dietary patterns of all groups except those at the extremes of the income range'.

Consequently, at first sight it would appear legitimate to include in our model a variable either representing high or low social class. However, because high social class can be expected to be positively related to the other explanatory variables but negatively related to the dependent variable, mortality, bias problems can be anticipated with ridge regression (Table 7.1 and Chapter 2).[4] Consequently, the decision has been made to exclude high social class and employ instead data relating to Social Class V (for 1951) and socio-economic group 11 (unskilled manual workers) for 1961. Unfortunately, social class data are not available for 1931.

Occupation

There is considerable evidence that working in certain occupations leads to a high risk of suffering from a particular disease. For example, Acheson et al (1968) have found a remarkably high rate of nasal cancer for woodworkers, while the increased risk of pneumonoconiosis amongst coalminers is recognised by the payment of injury benefit to those suffering from this crippling disease. As illustrated in Chapter 6 (Table 6.3), deaths from bronchitis are 50 per cent above the average for miners, quarrymen, furnacemen, forge, foundry and rolling-mill workers, while gas, coke and chemical workers have relatively high rates

from stomach and lung cancer. This may appear to be convincing support for the effects of the occupational environment, but there is also some evidence to the contrary. It appears from Table 7.3 that bronchitis is much higher for those men who work in a 'dusty' environment, as compared to those who work on the land. However, from the highly similar pattern displayed by married women (classified by their husbands' occupation) it can be suggested that these large differences owe little to direct occupational effects and must be attributed to more general factors that are shared by both husbands and their wives. Moreover, Wynn and Wynn (1980, 501) in their analysis of disease in women have reached similar conclusions and write that with regard to the cause of disease,

> 'it is likely that 90 per cent of the causation lies within the home environment, shared by men and their wives: food and drink, the shared packet of cigarettes'.

Clearly the evidence concerning this particular explanatory variable is again conflicting, and the separate influence of occupation needs to be assessed while controlling for such variables as air pollution and cigarette smoking. Data on occupations for the County Boroughs were derived from the Censuses of 1931, 1951 and 1961 and an attempt was made to obtain comparable figures for each date. As Armstrong (1972) has shown, considerable effort must be expended upon the early censuses but fortunately, since 1931, the classification of occupations adopted by the Registrar General has remained generally similar and, providing that broad groups are adopted (Table 7.4), it is possible to derive comparable statistics.

Unemployment

Unemployment has been suggested by a number of researchers to have both a direct and indirect effect on human health. Brenner (see Moore, 1980), on the basis of time-series data

Table 7.3

Bronchitis SMRs England and Wales
for males and females 15-64, 1950-1953

Occupation	Men	Married Women
Coal miners	135	175
Face workers	200	190
Iron and steel workers	158	213
Agricultural workers	53	82
Farmers	31	52

Source : Registrar General (1957) Decennial Supplement, 1951, HMSO, London

Table 7.4

Occupation variables 1931, 1951 and 1961

1931	1951	1961
Mining and quarrying occupations	Mining and quarrying occupations	Miners and quarrymen
Makers of coal, gas, coke and by-products plus workers in chemical processes	Coal, gas etc, workers in chemicals	Gas, coke and chemical workers
Makers of bricks, pottery and glass	Workers in ceramics, glass, cement etc.	Workers in bricks, pottery, glass cement etc.
Furnacemen plus foundry workers plus smiths and forge workers	Furnacemen plus rolling and tube-mill workers plus smiths and forgemen	Furnacemen, forge, foundry and rolling-mill workers

Sources : Registrar General's Classification of Occupations 1931, 1951, 1961.

relating unemployment levels to morbidity, suggests that the stress associated with being involuntarily unemployed leads to direct damage to mental and physical health. Others, however, have argued that the link with unemployment is indirect, with those people who are out of work being forced both to eat on inadequate diet and to live in poor housing conditions. For example, Hart (1976, 3) has attributed the relatively high mortality of South Wales to

> 'the effect of gross social deprivation between the wars'.

He pointed out that prior to 1971 the infant mortality rate in the South Wales valleys had been for many years consistently 25 per cent in excess of the England and Wales rate, but that since 1971 there had been a dramatic and continuing fall to below the national figure. He argued that this improvement must be due to the fact that most of the post-1971 births were to mothers who themselves were born after the period of the Depression and high unemployment. Hart, therefore, contends that unemployment and poor social conditions not only affect one generation but also the one which follows.

In the present study, data are available on average pre-war unemployment (Gardner, 1973) and it is possible to derive figures for the number of people 'out of work' in 1951 and 1961 from the Census. Areas of high unemployment tend to have high proportions of men engaged in particular occupations and a high proportion of people of low social class, and it remains to be seen, therefore, whether the relationship exists when these variables are statistically controlled.

Housing conditions

Housing conditions have been suggested by a number of workers to be a causal factor in determining disease. For example, Turner (1964) has suggested that dampness would encourage the growth of fungi of the species aspergillum and cladosporium and this could result in respiratory catarrh

which could, in turn, lead to other respiratory diseases such as bronchitis. Moreover, Girt (1972), in his study of female bronchitics, found 'damp' housing to be an important determinant of their distribution in Leeds. However, as discussed in Chapter 2, his study was based on stepwise regression and must therefore be regarded with suspicion. For the present study, it was not possible to derive a measure of the number of houses with damp conditions in the County Boroughs. Consequently, a proxy variable was chosen (number of households lacking exclusive use of fixed bath) and comparable figures have been derived from the Census for 1951 and 1961. Unfortunately, the 1931 Census did not collect such data.

Water Hardness

As discussed in Chapter 6, many studies have found conflicting evidence for the effect of water hardness, but for the County Boroughs Gardner (1973) has claimed a strong negative association between disease and water hardness. Using the same data as Gardner, it is possible to derive measures of total hardness for 1931, 1951 and 1961 for all the County Boroughs. Moreover, for all but ten Boroughs, data are available for magnesium, sodium and calcium concentration.

Control variables

The dependent variables in this study are the absolute number of deaths for different age groups, diseases and time periods. Because it has been decided to analyse absolute numbers and not death rates, control variables for 'population size' must be included in the model. In order to specify the model as accurately as possible it was decided to use five year age groups (45-49, 50-54, 55-59, 60-64, 65-69, 70-74) to control for differing number of people in each County Borough. Similarly, because the number of households lacking exclusive use of a fixed bath has been

included in the model, the total number of households has also been included to act as a control.

EXPLORATORY PROCEDURES

Table 7.5 outlines an exploratory framework that can be used to develop models that account for geographical variations in mortality. There are two important features that distinguish this approach from the confirmatory one. Firstly, the exploratory researcher is sceptical of the initial model he specifies and, consequently, he does not leap from 'calibration' to 'interpretation'. Instead, he concentrates on searching for weaknesses in the model (unfulfilled assumptions) using a variety of methods specifically designed for specification-error analysis. Secondly, acknowledging that exploratory procedures can 'capitalise on a chance result', the exploratory researcher employs some form of cross-validation as a safeguard against model mis-specification. The researcher only proceeds to interpretation when the model has been fully evaluated and deemed to be correct. This of course the exploratory procedure that was first considered in Chapter 1, but Table 7.5 differs from Figure 1.6b in that the techniques discussed in Chapters 2, 3 and 4 have been incorporated into the overall framework. Thus, ridge regression is used to calibrate a multicollinear model, normal probability and partial-residual plots are used to evaluate a model, while the Ramsey tests are used to cross-validate a model.

In order to illustrate how this procedure performs in practice, let us select one disease (bronchitis mortality, males, 65-74 years, 1958-64) and apply these exploratory techniques.[5] Stage 1, that of examining the literature and collecting data, has already been discussed and consequently it requires no further consideration. Stage 2 is the specification of the model, the variables used in the initial model being listed in Table 7.6. An examination of this

Table 7.5

Developing a model by exploratory procedures

STAGE

1 EXAMINE LITERATURE AND COLLECT DATA

2 SPECIFY MODEL

3 CALIBRATE MODEL

(a) OLS estimation if no multicollinearity present
(b) Ridge estimation if multicollinearity present

4 SPECIFICATION-ERROR ANALYSIS

(a) Normal probability plot
(b) residual plots
(c) partial residual plots
(d) map of residuals
(e) partial residuals: Box-Cox transformations

5 MODEL EVALUATION AND IMPROVEMENT

If no problems detected go to STAGE 6
If problems detected take corrective action
(a) omit outlier
(b) search for omitted variable(s)
(c) transform data
Return to STAGE 2

6 CROSS-VALIDATION

(a) evaluate model with significance tests on another body of data
(b) predict values for the dependent variable of the second body of data
(c) perform Ramsey's specification-error analysis

If model successfully cross-validated go to STAGE 7
If model fails go to STAGE 4 and analyse second data set

7 MODEL INTERPRETATION

table reveals that the chosen variables relate to 1961 with the exception of water hardness (the main variable under study) and pre-war unemployment selected to represent past social conditions). It was decided in this initial model to ignore the possibility of 'lagged' relationships for the other explanatory variables because it was thought that the inclusion of variables for 1951 and 1931 would increase the number of explanatory variables to an unmanageable size. Moreover, if any such 'lagged' relationships are of major importance in determining mortality variations, their omission from the model should be signalled by specification-error analysis. While the functional form relating the control variables (number of males aged 65-69 years and 70-74 years, number of households) to the dependent variable can be expected to be linear, there is insufficient a priori information to specify the nature of the relationship between each of the 15 explanatory variables of Table 7.6 and the dependent variable. Consequently, it was decided to specify a linear relationship for all the explanatory and control variables as an initial approximation, and subsequently to evaluate the appropriateness of this choice by an analysis of the residuals of the initial model.

At Stage 3 the specified model is calibrated and, following the discussion in Chapter 2, it can be suggested that the model should be estimated by ordinary least squares if no multicollinearity is present, and by ridge regression if this fundamental assumption is not met. In the present analysis it can be anticipated that the model will be multicollinear: County Boroughs with a large number of people in one age group (65-69) can be expected to have a large number of another age group (70-74), a large number of households, and a large number of people in socio-economic group 11. Similarly, water hardness in 1931 can be expected to be closely related to hardness in 1961, while industrial air pollution can be anticipated to co-vary with domestic air

Table 7.6

Bronchitis mortality 65-74 years, 1958-64 : estimating the initial model

Variable	OLS	Ridge
Number of males aged 65-69 years, 1961	- 1.46	0.16
Number of males aged 70-74 years, 1961	+ 1.30	0.17
Number of households, 1961	0.91	0.11
Pre-war unemployment	0.27	0.18
Miners and quarrymen, 1961	0.03	0.03
Gas, coke and chemical workers, 1961	0.03	0.03
Bricks, pottery, glass and cement producers, 1961	- 0.03	0.01
Furnace, forge, foundry, rolling workers, 1961	- 0.04	0.01
Socio-economic group 11, 1961	0.07	0.10
Persons per room, 1961	- 0.04	- 0.03
Persons per acre, 1961	- 0.03	0.00
No. of people living in households with more than 1.5 persons per room, 1961	- 0.14	0.05
No. of households lacking exclusive use of fixed bath, 1961	0.30	0.18
Cigarette consumption	0.27	0.16
Water hardness, 1931	0.00	- 0.01
Water hardness, 1961	- 0.02	- 0.02
Domestic air pollution	- 0.03	0.02
Industrial air pollution	- 0.00	0.01
R^2	0.98	0.97
u	0.0	0.05
Determinant	0.000 000 001	0.98

pollution. In fact the correlation coefficient relating the two age variables is 0.9, while the equivalent statistic for the water variables and air pollution measures is 0.9 and 0.5. Such a model with highly inter-related variables will be badly estimated by ordinary least squares and, as Table 7.6 demonstrates, a number of the OLS coefficients have an incorrect a priori sign. For example, according to the OLS estimates, an increase in air pollution, persons per room, and persons per acre will all result in an unexpected decrease in the number of deaths from bronchitis. Moreover, the two standardised regression coefficients associated with the age variables are both in excess of 1.0 and are therefore completely uninterpretable. Such results can only occur with multicollinear data (Chapter 2) and the value of the determinant (0.000000001) provides further evidence that this model cannot be satisfactorily estimated by ordinary least squares.

The model was subsequently re-calibrated with a ridge procedure and the values in Table 7.6 were obtained when the bias parameter, k, was 0.01 (a value suggested by applying both Hoerl and Kennard's (1970) guidelines and Brown's (1973) rule). A comparison of the ridge and OLS estimates reveals that all the ridge coefficients (with the exception of persons per room) have the correct a priori signs and none are in excess of 1.0. Moreover, while the coefficient of determination of the ridge model is slightly lower than the OLS value (0.97 as compared to 0.98), the u statistic (0.05) for the biased model does not suggest that too much bias has been introduced into the estimation.

It is tempting to summarily conclude, on the basis of the high coefficient of determination, that the model is a good one and proceed to interpretation, noting that pre-war unemployment, lack of exclusive use of fixed bath and cigarette consumption are the three most important determinants of bronchitis mortality. However, as discussed in Chapter 1,

the coefficient of determination is on inadequate measure of the validity of a model, and consequently an exploratory specification-error analysis needs to be undertaken. As Stage 4 of Table 7.5 reveals, a number of different exploratory procedures can be applied to a model. However, as discussed in Chapter 4, we must be alert to the fact that many specification errors have similar effects, and one 'true' specification error may result in the data apparently showing other types of error.

Figure 7.8, which is a normal probability plot of the residuals from the ridge model, indicates immediately that something is wrong with the fitted model. While most of the residuals are reasonably close to a straight line, implying that the majority of the residuals are normally distributed, one residual, that for Manchester, is detached from the main body of the data. The standardised residual for Manchester is 4.21, and this suggests that the actual bronchitis mortality is much higher in Manchester than the model predicts. Moreover, when it is recalled that a value in excess of 2.0 should be regarded with suspicion, it can be suggested that this 'outlier' will have a major and detrimental effect on model estimation (Chapter 4).

Such an outlying residual may be the outcome of a number of factors: the residual may be a true outlier or it may be the result of the model requiring a non-linear relationship, an interaction term or another explanatory variable. In examining the suggestion that an important variable has been omitted, we may consider the spatial distribution of the residuals. Figure 7.9a examines the bronchitis mortality rate for the 65-74 age group in 1958-64. In effect this is the spatial pattern after the variations in population size have been controlled, and a familiar and distinctive pattern is revealed with the highest rates generally being found in the northern County Boroughs. Indeed, the three County Boroughs with the most unfavourable rates

FIGURE 7.8 INITIAL MODEL: NORMAL PROBABILITY PLOT

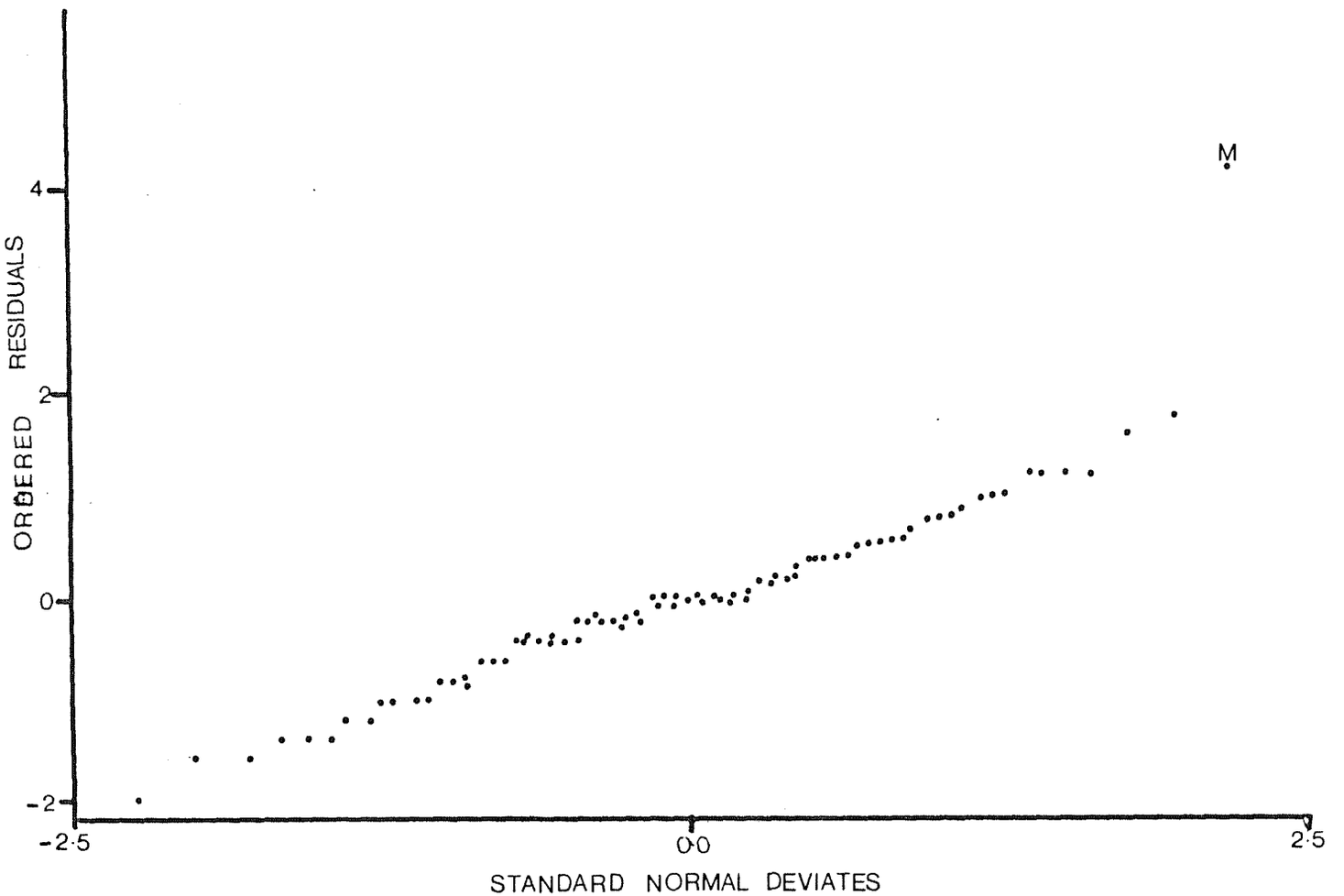
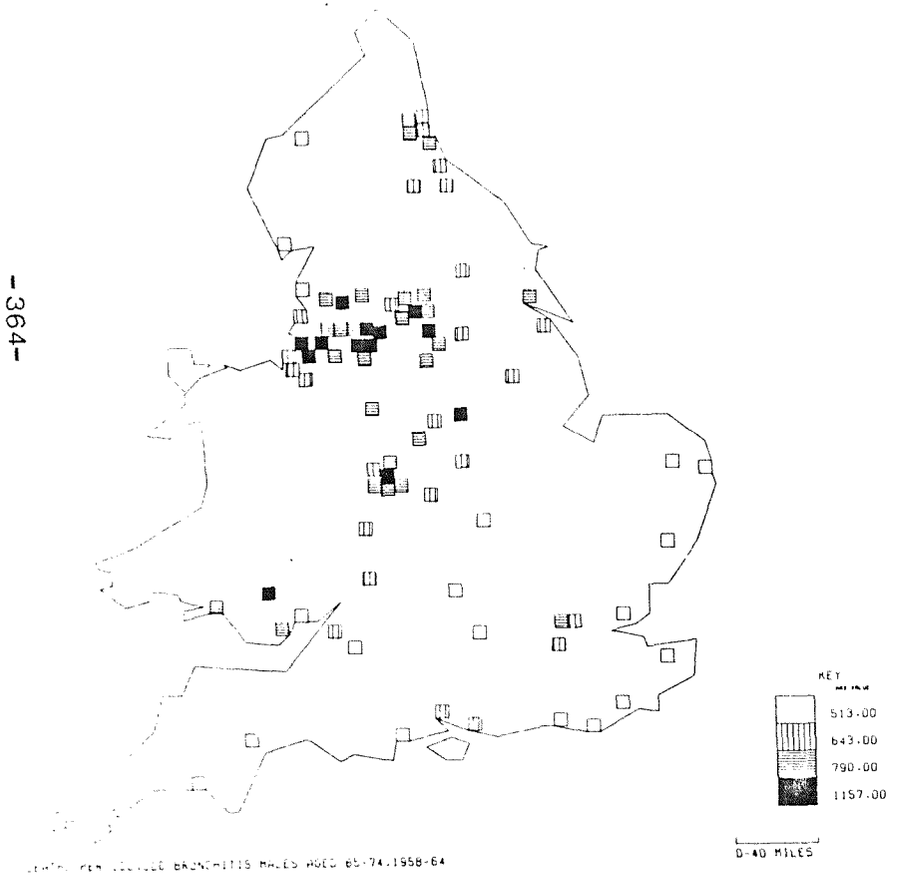
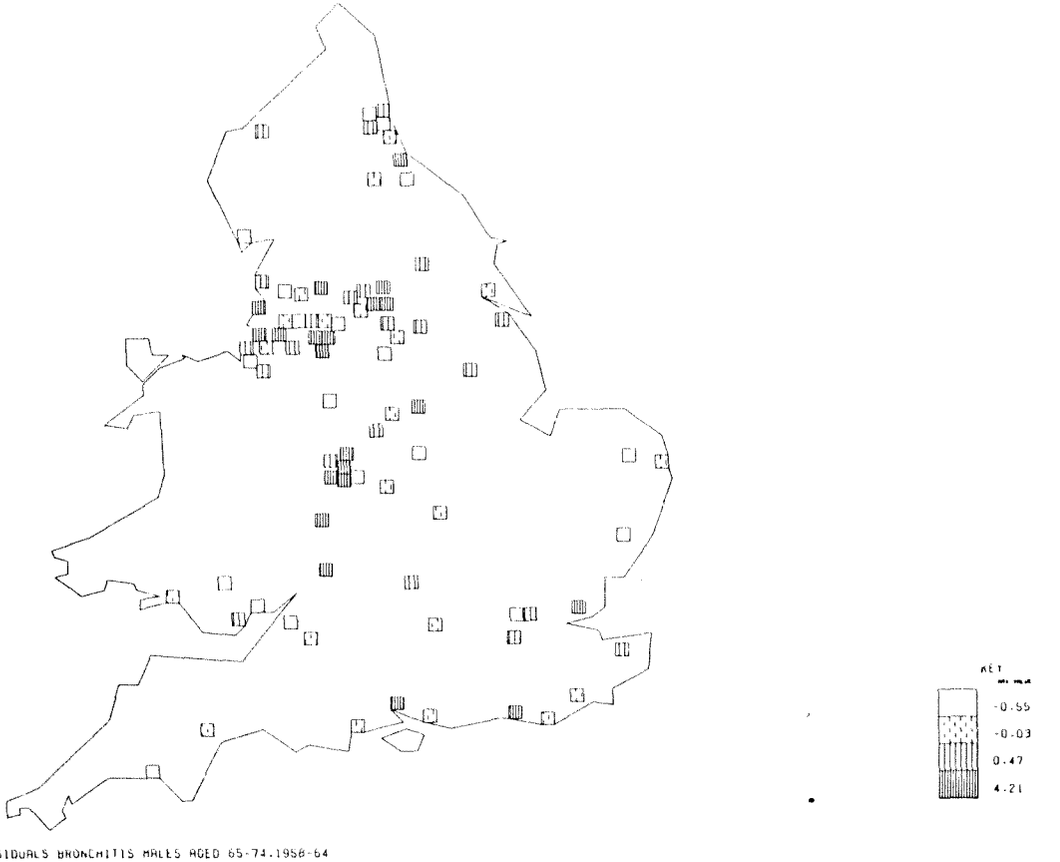


FIGURE 7.9(A) BRONCHITIS DEATH RATES



DEATH PER 100000 BRONCHITIS MALES AGED 65-74.1958-64

(B) RESIDUALS FROM INITIAL MODEL

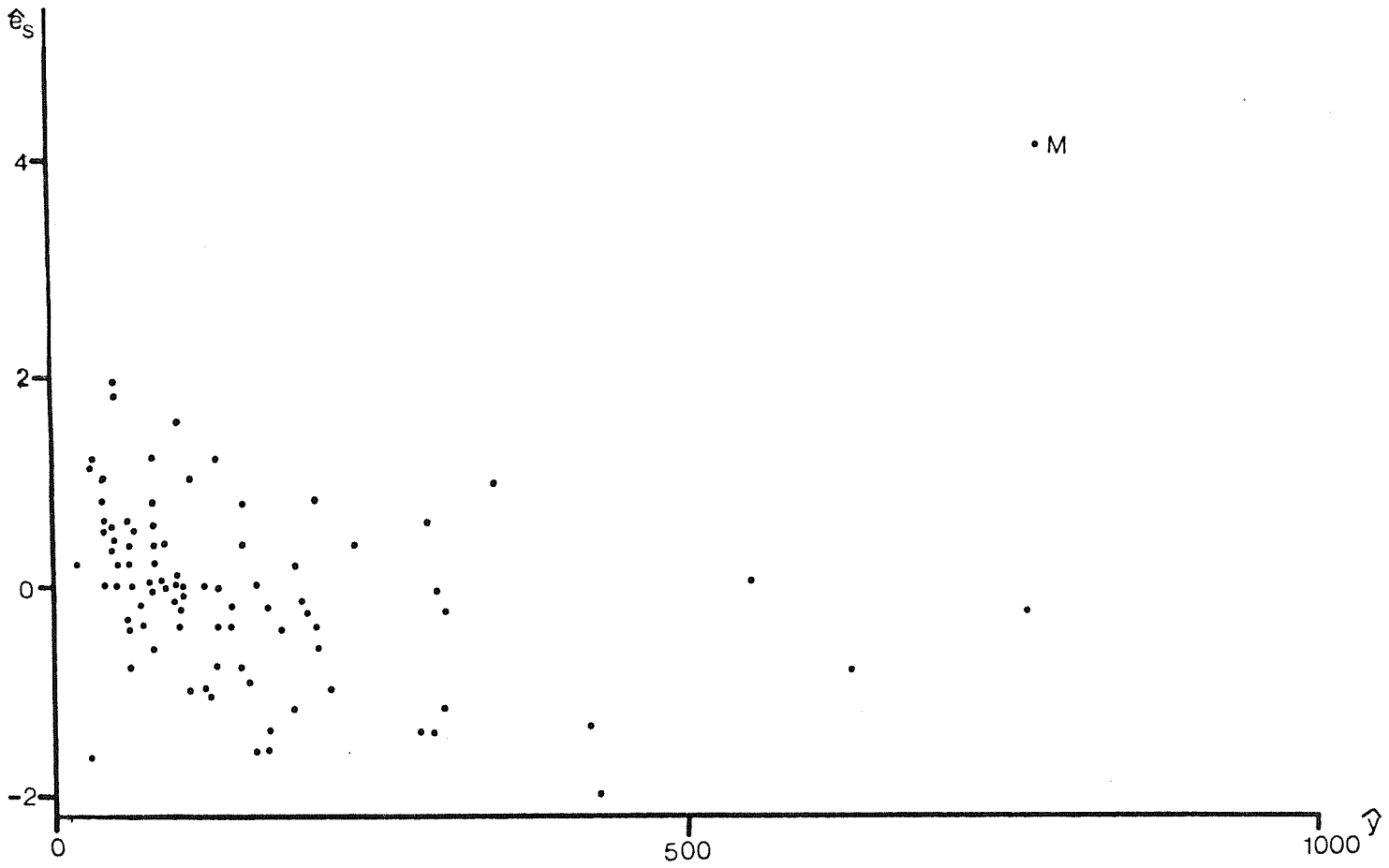


RESIDUALS BRONCHITIS MALES AGED 65-74.1958-64

for the 65-74 age group (Salford, Manchester, Oldham) are also the towns with the three highest rates for the 45-64 age group in 1958-64 (Figure 7.9a and 7.4b). The residuals (Figure 7.9b) from the initial model, however, do not exhibit such strong patterning; both northern and southern County Boroughs show positive residuals (model under prediction) and negative residuals (model over prediction). These results suggest that the explanatory variables do indeed explain some of the variation in the distribution of mortality in addition to that accounted for by the control variables. Unfortunately, however, there appears to be some remaining spatial patterning in these residuals, with the West Midlands, Manchester and West Yorkshire conurbations all generally showing positive residuals. Yet even a careful examination of the distribution of the residuals does not suggest which specific variable has been inadvertently omitted from the model. Moreover, the residual map does not resembly any 'lagged' version (1931 and 1951) of the explanatory variables already included in the model.

As discussed in Chapter 3, spatially autocorrelated residuals can be obtained from an incorrectly specified model which requires a non-linear or interaction term. Figure 7.10 examines this suggestion with a plot of the standardised residuals against the predicted values of the dependent variable. While some may see a curved band of residuals (thereby indicating the need for a non-linear or interaction model) others may see a band of roughly uniform width with one outlier for Manchester (denoted by M in Figure 7.10). In order to explore the choice of functional form in more depth, an analysis based on Box-Cox transformations of the partial residuals was performed. Following the procedure detailed in Chapter 4 (with the exception of using ridge instead of OLS estimates), the results suggest that domestic air pollution should be transformed by λ of 1.6 and cigarette consumption transformed by λ of 1.4. However, when these transformations are performed and the

FIGURE 7.10 INITIAL MODEL: RESIDUAL PLOT



re-specified model estimated by ridge regression, the standardised residual for Manchester remains stubbornly in excess of 4.0.

Reviewing the exploratory procedures that have been employed so far (Stage 5), it appears that Manchester has a level of bronchitis mortality that is very much in excess of that predicted by the fitted model. It does not appear, however, that this outlying value is a result of a non-linear relationship between the dependent and explanatory variables. Moreover, while the residuals display some geographical pattern, it is not possible to suggest which important explanatory variable has been omitted from the model. The evidence therefore seems to suggest that the large residual for Manchester is either the result of an incorrect or mispunched value for bronchitis in this County Borough, or that the residual represents a 'true' outlier. If either of these possibilities is true, the value for Manchester can be omitted and the model re-estimated. However, there is strong evidence that this large residual is neither an incorrect value nor a true outlier for similarly large residuals are found for Manchester when different age groups and time periods are analysed (3.52 for 45-64 age group, 1958-64: 3.68 for 45-64 years, 1948-54; 4.14 for 65-74 years, 1950-54). This consistent underprediction cannot be easily dismissed and requires some form of explanation.

Following exploratory principles it was decided to fit an alternative model which included an interaction term (domestic air pollution multiplied by cigarette consumption). The inclusion of this interaction term was made on the basis of prior knowledge and suggestions from the exploratory procedures discussed above. Firstly, previous researchers (Lambert and Reid, 1970) have found that bronchitis incidence in England and Wales is highest for people who both smoke cigarettes and live in high-pollution areas. Secondly, the Box-Cox analysis has demonstrated that bronchitis

may be related in a non-linear manner to cigarette consumption and domestic air pollution. Finally, an examination of the explanatory variables reveals that Manchester has both a relatively high level of domestic air pollution and cigarette consumption.

A reformulated model which included the interaction term was calibrated and subjected to specification-error analysis. Figure 7.11 examines a normal probability plot of the residuals and, as this approximately conforms to a straight line, there is no suggestion of outliers or non-normality. Indeed, the residual for Manchester is 0.14 and therefore the underestimation is much less than in the original model. A map of the residuals (Figure 7.12) also reveals that the reformulated model is well specified, for there is no obvious spatial patterning. Moreover, a residual plot (Figure 7.13), the Box-Cox analysis and the partial residuals confirm still further the suggestion that bronchitis mortality is linearly related to all the explanatory variables and that the model is well-specified.

The OLS and ridge estimates for the reformulated model are given in Table 7.7. While we cannot proceed to interpret these results because the model has not been cross-validated, it is interesting to compare them with the results for the original model (Table 7.6). Some of the OLS coefficients have changed dramatically, and now the greater the number of males aged 70-74, the fewer die from bronchitis! The ridge coefficients have changed much less drastically, but the differences are still of interest. Pre-war unemployment, number of households lacking exclusive use of fixed bath, and cigarette consumption are all less important in determining bronchitis mortality than in the original model. Moreover, the most important variable in determining bronchitis mortality (apart from the controls) is the interaction term, and given this result it is perhaps not surprising that the exploratory procedures revealed specification error.

FIGURE 7.11 REVISED MODEL: NORMAL PROBABILITY PLOT

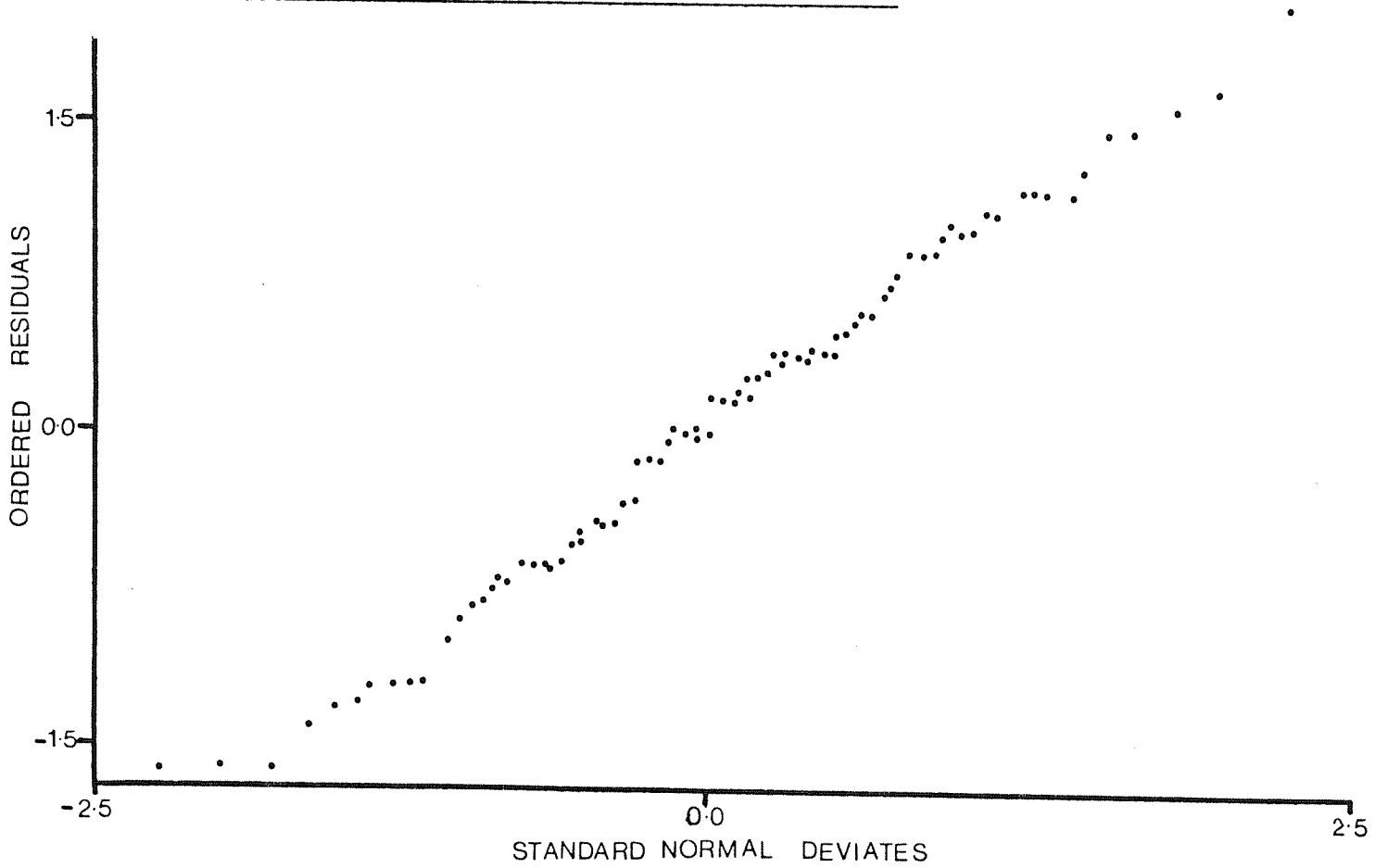


FIGURE 7.12 RESIDUALS FROM THE REVISED MODEL.

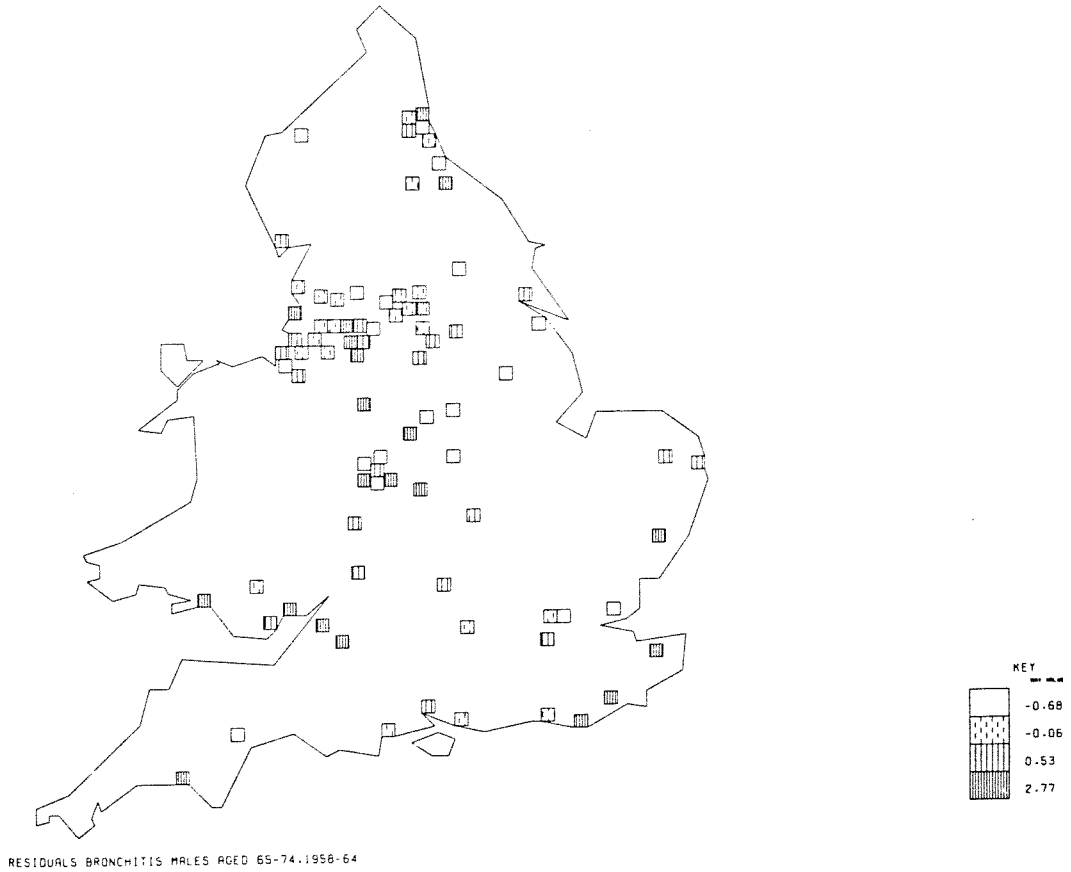


FIGURE 7.13 REVISED MODEL: RESIDUAL PLOT

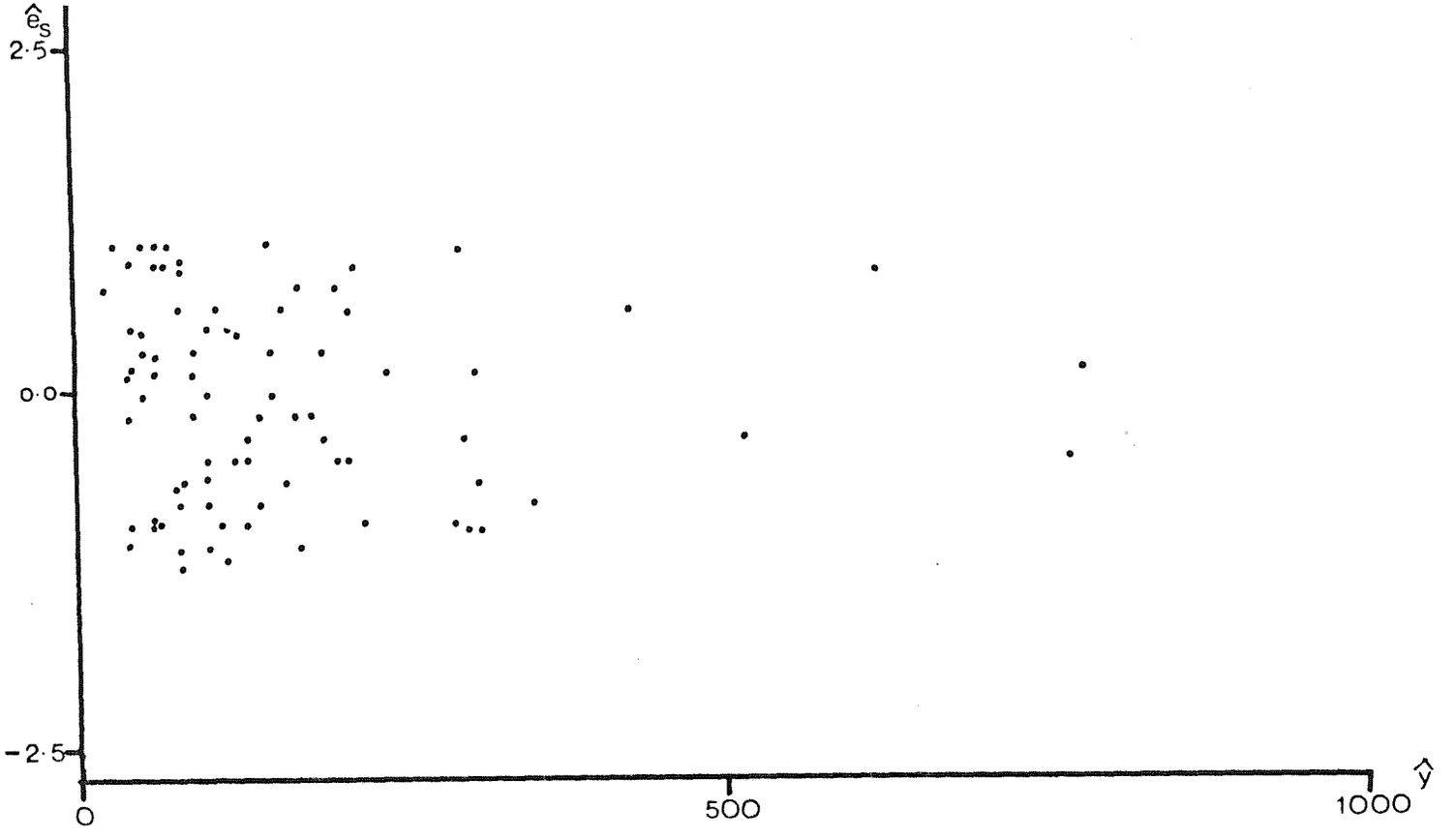


Table 7.7

Bronchitis mortality 65-74 years, 1958-64 : estimating the revised model

Variable	OLS	Ridge
Number of males aged 65-69 years, 1961	0.46	0.16
Number of males aged 70-74 years, 1961	- 0.01	0.17
Number of households, 1961	0.02	0.11
Pre-war unemployment	0.08	0.14
Miners and quarrymen, 1961	0.02	0.03
Gas, coke and chemical workers, 1961	0.02	0.03
Bricks, pottery, glass and cement producers, 1961	- 0.03	0.01
Furnace, forge, foundry, rolling workers, 1961	- 0.02	0.01
Socio-economic group , 1961	0.10	0.10
Persons per room, 1961	- 0.00	- 0.01
Persons per acre, 1961	- 0.04	0.00
No. of people living in households with more than 1.5 persons per room, 1961	- 0.14	0.05
No. of households lacking use of fixed bath, 1961	0.28	0.14
Cigarette consumption	- 0.08	0.12
Water hardness, 1931	- 0.01	- 0.01
Water hardness, 1961	- 0.00	- 0.02
Domestic air pollution	- 0.04	0.02
Industrial air pollution	0.01	0.02
Cigarettes * air pollution	0.37	0.15
R^2	0.99	0.97
u	0.0	0.06
Determinant	0.000 000 001	0.99

Finally, with regard to Table 7.7, it must be added that the coefficient of determination is satisfactorily high, while the u statistic is pleasingly low. Moreover, only one coefficient (persons per room) has an incorrect *a priori* sign and even this is closer to zero than in the original model.

Before the model can be interpreted, some form of cross-validation (Stage 6) must be carried out. There are many different ways of cross-validating a set of results, and in earlier chapters we have discussed three such methods. In the first, the data are split into two groups, and the regression equation fitted on one data set is used to predict the values of the dependent variable of the second data set. If there is a high correlation between the predicted and actual values (relative to the coefficient of determination of the fitted model) the analyst proceeds to interpretation; if there is a relatively low correlation between these values, he must use exploratory procedures to determine how the second data set differs from the first. In the second method of cross-validation, the data are divided into two parts, a model is developed on one data set and unimportant variables with a low regression coefficient are omitted from the model. This model is then evaluated by the usual test procedures (*t* and F) to see whether the remaining variables are important determinants of the variation in the dependent variable of the second data set. The third method, introduced in Chapter 4, also divides the data into two parts and the model developed on one set is tested for specification error by the application of the Ramsey tests to the other set.

In the present analysis, however, a deliberate decision has been made not to divide the data set but to estimate the model with all 83 County Boroughs. This decision has been made so that the model is estimated as accurately as possible: the variance of the estimates depends not only on model specification and multicollinearity but also on the difference between the number of explanatory variables and the number of observations. Although, as Darlington (1978) points out, ridge regression deals with this problem more

effectively than ordinary least squares, it was decided that at least 83 observations were needed to estimate successfully coefficients associated with 15 explanatory variables, 3 control variables and an intercept term. This, of course, precludes the use of the methods of cross-validation previously described. However, it can be contended that as the Ramsey tests have not been used in model exploration and improvement, they are an essentially independent means of testing the reformulated model for specification error.

The proposed method of cross-validation is, therefore, to evaluate the model developed for all the 83 County Boroughs by the three Ramsey specification-error tests. If no specification error is found the analysis may proceed to interpretation; if the model is found to be mis-specified it may be possible to use the details of the Ramsey tests to suggest those aspects of the model that require improvement. In the present analysis, BLUS residuals were calculated for the bronchitis mortality ridge model, and the probability level for the rejection of the null hypothesis of no specification error was set to 0.01. Even with this severe test of model adequacy, the Ramsey procedures did not reject the null hypothesis (Table 7.8). Consequently it can be concluded that the model does not suffer from incorrect functional form and omitted variables (RESET), heteroscedastic residuals (BAMSET), and non-normal residuals (WSET). Given these highly satisfactory results, interpretation (STAGE 7) is now permissible and the relative importance of each explanatory variable in determining mortality variations will be considered in detail in the next section.

Table 7.8

Cross-validation with Ramsey specification-error tests : bronchitis mortality 65-74 years, 1958-64

Test	Probability level	Critical values	Calculated value	Decision
RESET (F)	0.01	3.52	0.86	cannot reject null hypothesis
BAMSET (χ^2)	0.01	9.21	69.45	cannot reject null hypothesis
WSET (D)	0.01	- 3.24 1.37	0.03	cannot reject null hypothesis

Reviewing the use of exploratory procedures in building and developing regression models, three important conclusions can be drawn. Firstly, without ridge estimation it would have been impossible to develop a meaningful regression equation; certainly, ordinary least squares cannot be used to evaluate the true contribution of the multicollinear variables that have been suggested to account for geographical variations in mortality. Secondly, although the exploratory procedures were not highly suggestive of an improved model, they did at least clearly indicate that the original model was considerably in error and required reformulation. Finally, if no exploratory technique had been employed and the analysis moved from 'calibration' to 'interpretation', there would have been a critical inferential error. Indeed, the true importance of the air pollution * cigarette consumption interaction as the major determinant of variations in bronchitis mortality would not have been appreciated without adopting an exploratory perspective.

RESULTS AND DISCUSSION

Similar exploratory analyses have been performed for each disease, age group and time period and the results from the well-specified models are given in Tables 7.9 to 7.15. In particular, the values given in Table 7.9 are central to the objective of this chapter for they give the standardised regression coefficients relating water hardness to mortality. Clearly, while most of the coefficients are negative (except for two associated with lung cancer) they are also very close to zero, thereby implying that geographical variations in water hardness account for very little of the geographical variations in mortality. For example, the standardised ridge coefficient relating hardness in 1961 to cardiovascular male mortality (65-74 years) for the period 1950-54 is - .032, and thus there is less than a .1 per cent chance ($-.032^2$) that a death has occurred because of the water hardness in a County Borough. Also displayed in Table 7.9 are the OLS coefficients derived by Gardner for the relationship between mortality and water calcium. While the two sets of coefficients are not strictly comparable,[6] it is interesting to note that all the coefficients derived in the present re-analysis are much smaller than Gardner's and, indeed, the ridge coefficients for cardiovascular disease (which he claimed was related to water hardness) are smaller than his coefficients for lung cancer, a disease for which he did not claim a relationship.

Another way of comparing the present study with that of Gardner's is to examine the affirmation that

> 'if the death rates throughout England and Wales were at the level of those in the hard-water areas it can be calculated that some 10,000 men fewer would die annually in middle age alone' (Gardner, 1973, 437).

Table 7.9

Relationships between water hardness and male mortality

Disease	Age Group	Period	Standardised regression coefficients: water hardness 1931	water hardness 1951/61	calcium Gardner (1973)
All causes	45-64	1948-54	-.009	-.016	-.23
	45-64	1958-64	-.005	-.019	-.24
	65-74	1950-54	-.014	-.017	-.43
	65-74	1958-64	-.003	-.019	-.30
Cardiovascular disease	45-64	1948-54	-.021	-.032	-.30
	45-64	1958-64	-.011	-.031	-.29
	65-74	1950-54	-.007	-.032	-.44
	65-74	1958-64	-.013	-.035	-.35
Bronchitis	45-64	1948-54	-.022	-.021	-.14
	45-64	1958-64	-.012	-.008	-.26
	65-74	1950-54	-.046	-.004	-.28
	65-74	1958-64	-.014	-.015	-.20
Lung cancer	45-64	1948-54	-.009	+.007	-.02
	45-64	1958-64	+.006	-.008	-.08
	65-74	1950-54	+.002	-.009	-.11
	65-74	1958-64	-.001	+.004	-.21
Strokes	45-64	1948-54	-.013	-.032	?
	45-64	1958-64	-.005	-.005	?
	65-74	1950-54	-.002	-.038	?
	65-74	1958-64	-.007	-.021	?
Coronary heart disease	45-64	1948-54	-.023	-.034	?
	45-64	1958-64	-.019	-.004	?
	65-74	1950-54	-.023	-.008	?
	65-74	1958-64	-.009	-.024	?

This is undoubtedly a major claim for 10,000 represents about 13 per cent of all deaths in England and Wales (45-64 years) for 1961. Let us begin this analysis by 'transfering' supplies from West Hartlepool (which had harder water than any other County Borough) to Birmingham, which had the softest water of any County Borough in 1961. This would result in a hardening of Birmingham's water by 435 ppm in 1931 and by 318 ppm in 1961. The unstandardised ridge coefficients relating total hardness in 1931 and 1961 to overall male mortality (45-64 years) are - 0.0141 and - 0.0544 respectively. Therefore, the change in hardness would result in a decrease in the predicted deaths from 2014 by (435 multiplied by - 0.014) plus (318 multiplied by - 0.0544). That is a total of 23 deaths would be avoided annually but this represents only a 1 per cent decline in Birmingham's death rate for the 45-64 age group. Transfering the hardest water to the borough with the softest water is the case most favourable to Gardner's arguments and when one considers that many towns already have relatively hard water, and would therefore experience even less of a decline in mortality, Gardner's claim must be seen as a gross over-estimation.

The evidence in this re-analysis strongly suggests that the commonly observed relationship between water hardness and disease is largely spurious: when the common association between water hardness, mortality and other explanatory variables is adequatedly controlled the individual contribution of water hardness is minimal. Obviously, if less than .1 per cent of mortality variations can be attributed to hardness, 99.9 per cent must be accounted for by other explanatory variables plus random variation, and Tables 7.10 to 7.15 identify those variables that have an association with mortality.

For each disease, age group and time period, variables with the standardised regression coefficients that are in

excess of .05 are given in rank order in the tables. (The control variables have been excluded from these lists.) Also given in the tables are the coefficients of determination (R^2) for each model, and a striking result is the degree to which the models account for mortality variation. No coefficient is below 0.96 while many are above 0.98, thereby indicating that the explanatory and control variables account for at least 96% of the geographical variations in mortality in the County Boroughs. Another general trend that can be found in the tables is that, while many of the explanatory variables are similarly ranked for each age group and time period for a particular disease, there are differences between the diseases. For example, the rankings of both vascular lesions of the central nervous system and coronary heart disease are very consistent for the two age groups and two time periods, but these rankings are quite different to those for lung cancer and bronchitis.

Turning now to a consideration of the effects of each explanatory variable that has been included in the models, we can begin our discussion with occupational variables. In general, these are not important determinants of mortality variation. In particular, none of the standardised ridge coefficients associated with 'mining and quarrying' exceed .05, even though a number of County Boroughs (Barnsley, Nottingham, Stoke, Wigan, Merthyr Tudful) had more than 5,000 such workers in 1951. Similarly, none of the coefficients associated with 'makers of bricks, pottery, glass and cement' exceed .05, and again this is despite the fact that some County Boroughs (for example, St. Helens, Stoke, Birmingham, Croydon, Manchester) had a large number of such workers. In contrast, 'gas, coke and chemical workers' and 'furnace, forge, foundry and rolling-mill workers' gave rise to coefficients that exceed .05 for bronchitis, lung cancer and strokes. Moreover, the associations between these occupational variables and bronchitis

Table 7.10

Standardised regression coefficients exceeding .05 : all causes mortality*

Age group : 45-64
Period : 1948-54

Rank	Variable
1	Cigarette consumption
2	Pre-war unemployment
3	Lack of exclusive use of fixed bath, 1951
4	Social class V, 1951
5	Domestic air pollution * cigarettes
6	More than 1.5 persons per room, 1951

R^2 : 0.98

Age group : 65-74
Period : 1950-54

Rank	Variable
1	Cigarette consumption
2	Lack of exclusive use of fixed bath, 1951
3	Social class V 1951
4	Domestic air pollution * cigarettes
5	Pre-war unemployment
6	More than 1.5 persons per room, 1951

R^2 : 0.98

Age group : 45-64
Period : 1958-64

Rank	Variable
1	Pre-war unemployment
2	Cigarette consumption
3	Domestic air pollution * cigarettes
4	Lack of exclusive use of fixed bath, 1961
5	Socio-economic group II, 1961
6	More than 1.5 persons per room, 1961

R^2 : 0.98

Age group : 65-74
Period : 1958-64

Rank	Variable
1	Cigarette consumption
2	Pre-war unemployment
3	Lack of exclusive use of fixed bath, 1961
4	Domestic air pollution * cigarettes
5	Socio-economic group II, 1961
6	More than 1.5 persons per room, 1961

R^2 : 0.98

* excluding control variables

Table 7.11

Standardised regression coefficients exceeding .05

: coronary heart disease mortality

CORONARY HEART DISEASE

Age group : 45-64
Period : 1948-54

Rank	Variable
1	Cigarette consumption
2	Pre-war unemployment
3	More than 1.5 persons per room, 1951
4	Social class V 1951
5	Domestic air pollution * cigarettes
6	Lack of exclusive use of fixed bath, 1951

R^2 : 0.97

Age group : 65-74
Period : 1950-54

Rank	Variable
1	Cigarette consumption
2	Pre-war unemployment
3	More than 1.5 persons per room, 1951
4	Social class V, 1951
5	Lack of exclusive use of fixed bath, 1951
6	Domestic air pollution * cigarettes

R^2 : 0.96

Age group : 45-64
Period : 1958-64

Rank	Variable
1	Cigarette consumption
2	Pre-war unemployment
3	More than 1.5 persons per room, 1961
4	Socio-economic group 11, 1961
5	Domestic air pollution * cigarettes
6	Lack of exclusive use of fixed bath, 1961

R^2 : 0.96

Age group : 65-74
Period : 1958-64

Rank	Variable
1	Cigarette consumption
2	Pre-war unemployment
3	More than 1.5 persons per room, 1961
4	Socio-economic group 11, 1961
5	Domestic air pollution * cigarettes
6	Lack of exclusive use of fixed bath, 1961

R^2 : 0.96

Table 7.12

Standardised regression coefficients exceeding .05

: vascular lesions of the central nervous system (strokes)

VASCULAR LESIONS OF THE CENTRAL NERVOUS SYSTEM

Age group : 45-64
Period : 1948-54

Rank	Variable
1	Cigarette consumption
2	Lack of exclusive use of fixed bath, 1951
3	Pre-war unemployment
4	Domestic air pollution * cigarettes
5	Social class V, 1951
6	More than 1.5 persons per room, 1951

R^2 : 0.97

Age group : 65-74
Period : 1950-54

Rank	Variable
1	Cigarette consumption
2	Lack of exclusive use of fixed bath, 1951
3	Pre-war unemployment
4	Domestic air pollution * cigarettes
5	More than 1.5 persons per room, 1951
6	Social class V, 1951
7	Furnacemen, rolling and tube-mill workers, foundry workers, smiths and forgemen, 1951

R^2 : 0.97

Age group : 45-64
Period : 1958-64

Rank	Variable
1	Cigarette consumption
2	Lack of exclusive use of fixed bath, 1961
3	Pre-war unemployment
4	Domestic air pollution * cigarettes
5	Socio-economic group II, 1961
6	More than 1.5 persons per room, 1961

R^2 : 0.97

Age group : 65-74
Period : 1958-64

Rank	Variable
1	Cigarette consumption
2	Lack of exclusive use of fixed bath, 1961
3	Pre-war unemployment
4	Domestic air pollution * cigarettes
5	More than 1.5 persons per room, 1961
6	Socio-economic group II, 1961

R^2 : 0.97

Table 7.13

Standardised regression coefficients exceeding .05 : Lung cancer

LUNG CANCER

Age group : 45-64
Period : 1948-54

Rank	Variable
1	Domestic air pollution * cigarettes
2	Social class V, 1951
3	Cigarette consumption
4	More than 1.5 persons per room, 1951
5	Pre-war unemployment
6	Lack of exclusive use of fixed bath, 1951
7	Gas, coke and chemical workers

R^2 : 0.98

Age group : 65-74
Period : 1950-54

Rank	Variable
1	Domestic air pollution * cigarettes
2	Social Class V, 1951
3	Cigarette consumption
4	Pre-war unemployment
5	Lack of exclusive use of fixed bath, 1951
6	More than 1.5 persons per room, 1951
7	Gas, coke and chemical workers, 1951

R^2 : 0.97

Age group : 45-64
Period : 1958-64

Rank	Variable
1	Domestic air pollution *
2	Socio-economic group II, 1961
3	Cigarette consumption
4	Pre-war unemployment
5	More than 1.5 persons per room, 1961
6	Lack of exclusive use of fixed bath, 1961
7	Gas, coke and chemical workers, 1961

R^2 : 0.97

Age group : 65-74
Period : 1958-64

Rank	Variable
1	Pre-war unemployment
2	Cigarettes
3	Domestic air pollution * cigarettes
4	Lack of exclusive use of fixed bath, 1961
5	More than 1.5 persons per room, 1961
6	Socio-economic group II, 1961
7	Gas, coke and chemical workers, 1961

R^2 : 0.98

Table 7.14

Standardised regression coefficients exceeding .05 : Bronchitis

BRONCHITIS

Age group : 45-64
Period : 1948-54

Rank	Variable
1	Domestic air pollution * cigarettes
2	Pre-war unemployment
3	Gas, coke and chemical workers, 1951
4	Cigarette consumption
5	Social class V, 1951
6	Lack of exclusive use of fixed bath, 1951
7	More than 1.5 persons per room, 1951
8	Domestic air pollution

R^2 : 0.95

Age group : 65-74
Period : 1950-54

Rank	Variable
1	Domestic air pollution * cigarettes
2	Cigarette consumption
3	Lack of exclusive use of fixed bath, 1951
4	Gas, coke and chemical workers, 1951
5	Pre-war unemployment
6	Social class V, 1951
7	Furnacemen, rolling and forgemen

R^2 : 0.96

Age group : 45-64
Period : 1958-64

Rank	Variable
1	Domestic air pollution * cigarettes
2	Pre-war unemployment
3	Lack of exclusive use of fixed bath, 1961
4	Cigarette consumption
5	Socio-economic group II, 1961
6	Gas, coke and chemical workers, 1961
7	Furnace, forge, foundry, rolling-mill workers 1961

R^2: 0.96

Age group : 65-74
Period : 1958-64

Rank	Variable
1	Domestic air pollution * cigarettes
2	Pre-war unemployment
3	Lack of exclusive use of fixed bath, 1961
4	Cigarette consumption
5	Socio-economic group II, 1961
6	More than 1.5 persons per room, 1961
7	Furnace, forge, foundry, rolling-mill workers 1961
8	Gas, coke and chemical workers, 1961

R^2 : 0.97

Table 7.15

Standardised regression coefficients exceeding .05 : cardiovascular disease

CARDIOVASCULAR DISEASE

Age group : 45-64
Period : 1948-54

Rank	Variable
1	Cigarette consumption
2	Pre-war unemployment
3	Social class V, 1951
4	Lack of exclusive use of fixed bath, 1951
5	Domestic air pollution * cigarettes
6	More than 1.5 persons per room, 1951

R^2 : 0.98

Age group : 65-74
Period : 1950-54

Rank	Variable
1	Cigarette consumption
2	Lack of exclusive use of fixed bath, 1951
3	Pre-war unemployment
4	Social class V, 1951
5	Domestic air pollution * cigarettes
6	More than 1.5 persons per room, 1951

R^2 : 0.97

Age group : 45-64
Period : 1958-64

Rank	Variable
1	Cigarette consumption
2	Pre-war unemployment
3	Socio-economic group 11, 1961
4	Lack of exclusive use of fixed bath, 1961
5	Domestic air pollution * cigarettes
6	More than 1.5 persons per room, 1961

R^2 : 0.97

Age group : 65-74
Period : 1958-64

Rank	Variable
1	Cigarette consumption
2	Pre-war unemployment
3	Lack of exclusive use of fixed bath, 1961
4	Socio-economic group 11, 1961
5	Domestic air pollution * cigarettes
6	More than 1.5 persons per room, 1961

R^2 : 0.97

and lung cancer are consistent, being found for different age groups and time periods. Given such results it is tempting to conclude that workers in such dusty, industrial environments have an increased risk of developing diseases of the chest and lungs. However, because of the aggregate nature of the study and the poor quality of the air pollution data, an alternative must be considered. It can be suggested that there are variations in air pollution brought about by chemical, coke and metal production which are not fully captured by Daly's pollution indices. Consequently, it is arguable that it is general air pollution and not occupational environment that causes increased mortality levels. Unfortunately, it is not possible to distinguish between these competing explanations in this analysis, and it can be suggested that a study is required which examines individuals within a particular city (Jones, 1975).

It is perhaps not surprising that 'persons per acre' is not strongly related to disease. Many County Borough boundaries are overbounded while others are underbounded, and 'persons per acre' is therefore a very poor measure of population density. Examining the other two overcrowding variables, it can be suggested that if there is a relationship between density and disease, it involves a threshold, for while the coefficients associated with 'persons per room' never exceeds .05, the variable 'more than 1.5 persons per room' exceeds this value for each disease age group and time period. Moreover, while this latter variable usually ranks lowest of all variables included in the tables, it is the third most important determinant of coronary heart disease, a disease which has often been associated with high stress. It must be admitted however, that the present analysis is really a rather inadequate test of the density and overcrowding hypothesis, for within any County Borough it can be expected that there will be considerable variations in density and crowding. However, the results at least

suggest that the hypothesis should be 'tentatively entertained' until it can be subjected to further falsification. The remaining household variable, 'lack of exclusive use of fixed bath', which has been chosen in this study as a surrogate for general household conditions, is usually one of the least important variables included in the tables. But for vascular lesions of the CNS it is the second most important explanatory variable for each age group and time period. Virtually nothing is known about the epidemiology of this particular disease and this result remains puzzling.

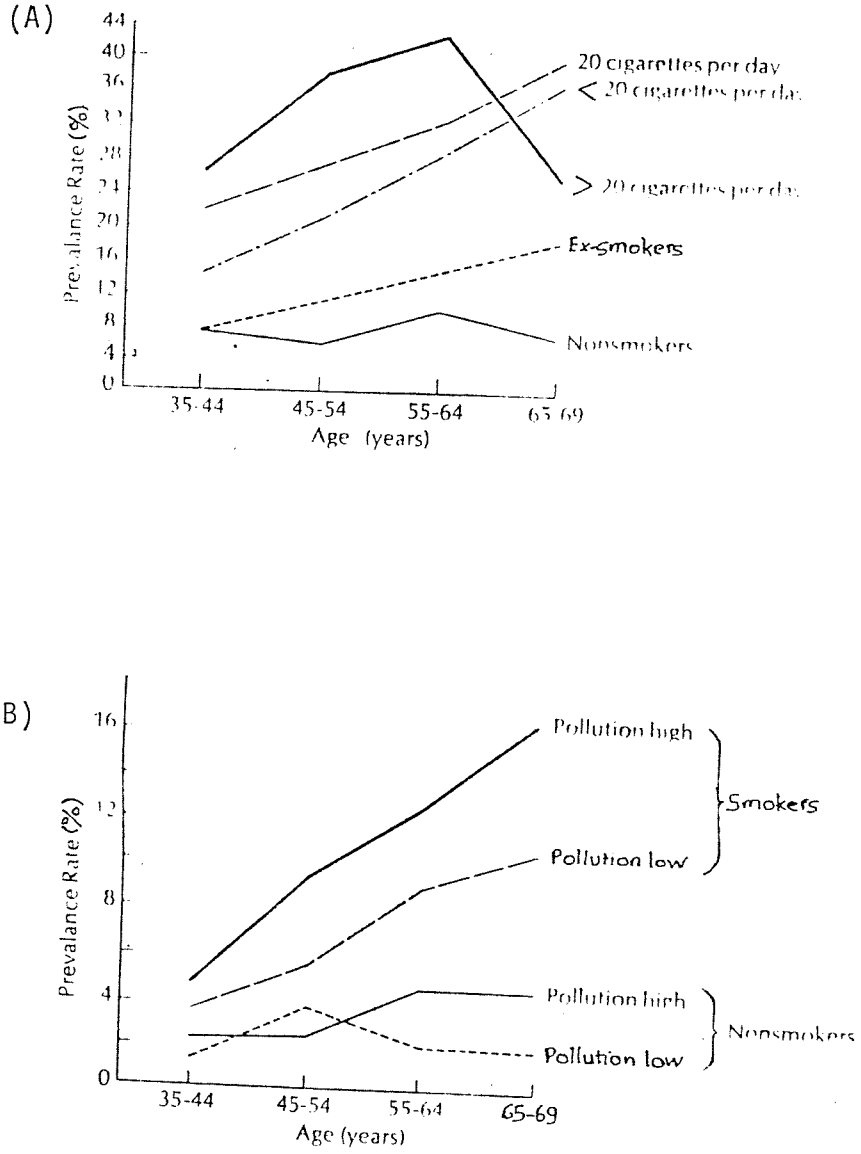
The standardised coefficients associated with low social class (Social class V, 1951; SEG 11, 1961) and pre-war unemployment exceed .05 for each disease, age group and time period. Indeed, with regard to overall mortality, pre-war unemployment is second only to cigarette smoking in its explanatory power. While this study strongly suggests that social class and unemployment have an effect on mortality that is independent of occupation, smoking and air pollution, it is not possible to determine what aspect of social class really accounts for disease variation. It is still an open question whether the diet of the lower social classes (high consumption of potatoes, cereals and sugar) or some other aspect of the way of life of the lower social classes (such as exercise) is the true cause of disease.

In general, neither the domestic nor industrial air-pollution has a substantial association with disease; only one standardised regression (that for bronchitis, 45-64 years, 1948-54) exceeds .05. However, the interaction variable (domestic air pollution * cigarettes) exceeds .05 for every disease, age and time period, and for bronchitis and lung cancer this variable is generally the most important determinant of their geographical distribution among the County Boroughs of England and Wales. This is an important and interesting result because it suggests that for these two diseases the combined effect is greater than the individual

effects of cigarette and air pollution. Examining the epidemiological literature it is possible to derive further support for this synergistic effect. Lambert and Reid (1970) undertook a cross-sectional study based on individual questionnaires for over 9,000 people in England and Wales. As Figure 7.14a demonstrates, the prevalence of persistent cough and phlegm in males increases for older age groups and for those who smoke. Moreover, Figure 7.14b clearly demonstrates a synergistic effect for bronchitis, with the highest incidence being found for individuals who are cigarette smokers and live in high-pollution areas, while individuals who are non-smokers and live in areas with low air pollution have the lowest rates. Lambert and Reid's study is based on a completely different methodology to the present analysis (individual not aggregate, morbidity not mortality) and yet it produces similar results for the same disease; this gives much greater credence to the finding that air pollution and cigarettes interact to cause disease. Returning to the tables, it can be seen that while the interaction variable is the most important determinant of lung cancer and bronchitis, cigarettes on their own are the most important determinant of mortality from all causes, cornonary heart disease, strokes and cardiovascular disease. For mortality from all causes 45-64 years 1958-64, the standardised ridge coefficient for cigarette smoking is +0.19 and this implies that this variable accounts for 3.6 per cent ($+0.19^2$) of the variation in the dependent variable. Of course the majority of the variation is accounted for by the control variables (age groups, number of households) and, when it is realised that cigarette smoking accounts for thirty-six times more variation than water-hardness, the substantial effect of cigarette consumption can be fully appreciated.

Finally, it must be stressed that neither of the water hardness variables (total hardness 1931, 1951/61) appear in

FIGURE 7.14 BRONCHITIS, AIR POLLUTION AND CIGARETTES



SOURCE: LAMBERT AND REID (1970)

Tables 7.10 to 7.15. This is because none of the standardised coefficients associated with these variables exceed .05 and it is clear, therefore, that the effect of water hardness is less than the effect of any of the six or seven other variables that do appear in the tables. Unfortunately, it is impossible in this study to make substantial progress towards evaluating the specific effects of these other explanatory variables, but it has proved possible to put the water hardness theory to the test and demonstrate that the effects of hardness have been overestimated. Moreover, because this test has been based on Gardner's work which, as Chapter 6 has shown, is probably the most thorough study available, even greater suspicion must be cast on the less-comprehensive studies that have found strong associations between water quality and disease. These far-reaching conclusions are the most important results of this exploratory analysis.

FURTHER ANALYSES

Although the models on which Tables 7.9 —— 7.15 are based appear well-specified, it was decided to investigate the relationship between mortality and a number of other explanatory variables in order to ascertain whether a causal variable has been inadvertently omitted. In particular, the effects of various elements of the water supply, climatic and geographical variables were estimated by fitting a number of ridge regression models to the available data.

Gardner (1973) in his study of the County Boroughs found that the strongest association between water supplies and mortality was between calcium and cardiovascular disease. Unfortunately, as previously noted, there are 10 towns (Bury, Canterbury, Dewsbury, Eastbourne, Hastings, Tynemouth, Wakefield, Wigan, Worcester, Merthyr Tudful) for which detailed information on water supplies is not available; in the main

analysis, therefore, total hardness (which is closely related to calcium) was used. However, another set of models has been calibrated in which three additional variables measuring various bulk elements of the water supply have been included. These additional variables are calcium parts per million, magnesium ppl0m, and sodium ppl0m, and all are for the period 1958-63. The results for cardiovascular disease are given in Table 7.16 and while the strongest relationships are found for calcium, even these coefficients are not sufficiently large to conclude that water supplies are a major determinant of mortality. Indeed, if the water supplies of the County Borough with the highest calcium content (Ipswich) are transferred to the town with the lowest content (Plymouth) there is less than a 1 per cent decrease in all causes mortality for men aged 45-64 years.

Table 7.16

Cardiovascular disease and water supplies

Cardiovascular disease	Calcium	Magnesium	Sodium	Hardness 1931	Hardness 1951/1961
45-64 1948-54	-.018	+.001	+.011	-.008	-.011
45-64 1958-64	-.016	-.004	-.009	-.006	-.009
65-74 1950-54	-.015	-.004	+.001	-.001	-.008
65-74 1958-64	-.010	+.010	-.011	+.002	-.014

A number of researchers have suggested that climate is an important determinant of mortality variations. For example, Roberts and Lloyd (1972) have suggested that rainfall is causally related to mortality while Gardner (1973) found strong positive associations between rainfall and cardiovascular disease, but strong negative relationships for

lung cancer. Moreover, West and Lloyd (1976) have suggested that another climatic variable, temperature, is an important determinant of heart disease mortality among the County Boroughs of England and Wales. While each of these studies has been evaluated and criticised in previous chapters, another study, that of Dudley et al (1969) can be similarly reproached for inappropriate use of statistical techniques. This study considered a total of 35 explanatory variables, but a 'climatic comfort' index (combining levels of relative humidity and temperature was found to be the variable most highly correlated with mortality from coronary heart disease. Unfortunately, however, the investigation was based on percentage data and the researchers used the unsuitable technique of stepwise multiple regression with variables which were highly likely to be severely multicollinear.

In order to assess the effects of climatic variables on mortality variations, the same measures (average sunshine, 1921-50; average temperature, 1921-50; average rainfall, 1916-50) used by Gardner have been included in the model. The results for cardiovascular disease are given in Table 7.17, together with estimates of the effects of geographical position. The latter, it must be noted, has been included because Gardner (1973) found strong negative associations between latitude and cancer of the lung and strong positive associations between latitude and cardiovascular disease. Moreover, according to Jones and Bourne (1977, 223)

> 'Several studies ... have suggested that the geographic position of a community within Great Britain may have an independent significance in determining mortality rates'.

While they did not elaborate on a causal mechanism to link latitude, longitude and mortality (nor did they suggest how to lower mortality rates), it is still interesting to see whether these relationships held when a large number of explanatory variables are statistically controlled. The same

data for latitude and longitude as previously used by Gardner (1973) have therefore been re-analysed, and it is on this material that the standardised ridge coefficients relating to location are based.

Table 7.17

Cardiovascular disease, climate and location

Cardiovascular disease	Standardised ridge coefficients				
	Sunshine	Temperature	Rainfall	Latitude	Longitude
45-64 1948-54	-0.001	-0.013	+0.003	-0.010	-0.008
45-64 1958-64	+0.009	-0.005	-0.005	-0.016	-0.006
65-74 1950-54	+0.007	-0.008	-0.008	-0.013	+0.014
65-74 1958-64	+0.010	+0.011	+0.011	+0.008	-0.002

Examination of Table 7.17 reveals that, with respect to climatic and locational variables, the associations with cardiovascular mortality are small, while there is also little consistency between time periods and age groups. Consequently this evidence strongly suggests that neither climate nor geographical position have a substantial effect on mortality variations in the County Boroughs of England and Wales.

Finally, in this multiple regression analysis there is a need to consider the effects of differential migration on the relative levels of mortality in each County Borough. For example, if people retire early due to poor health and migrate from industrial County Boroughs to live in coastal retirement areas, this would result in these retirement areas having a relatively high mortality rate, while the rate in the industrial and inland County Boroughs would be depressed. However, the retirement areas have very low rates of mortality

and it can therefore be suggested that the results of Tables 7.9 to 7.15 are in error because the fit and able males (perhaps of a high social class) have migrated to the coastal areas thereby giving a spuriously low mortality rate for these areas.

Unfortunately, since Hill's (1925) classic study of urban and rural mortality, which established that the rates were to some extent affected by migration, there has been little explicit examination of this important problem. To tackle the problem directly requires a longitudinal study, following people through time, and all that can be achieved in the present study is an indirect solution. The topic has been approached in two ways: firstly by recalibrating several models while omitting a number of coastal retirement areas and, secondly, by re-estimating models with a dummy variable representing areas of high outmigration. Applying the first strategy, Blackpool, Southport, Eastbourne, Hastings, Southend, Great Yarmouth were omitted from the calculations, but this did not substantially affect the ridge coefficients associated with each explanatory variable. Following the second strategy the dummy variable (Gujatari, 1970) was set to one for the twenty highest areas of outmigration (Barnard, 1978), while all other areas were set to zero. The results of this exercise show that the standardised ridge coefficients associated with the dummy variable were always close to zero, an outcome which again adds weight to the conclusion that the results in Tables 7.9 to 7.15 have not been unduly affected by differential migration.

In summary, an attempt has been made to falsify the results of Tables 7.9 to 7.15 by examining a number of variables that have been deliberately omitted for the original models and by considering the effects of differential migration. As the models have been evaluated against these and have survived, greater faith can be placed in the inferences that have been made on the basis of the tables.

It therefore appears from this evaluation that the estimated weak association between water hardness and disease is not a spurious one.

AN 'EXPERIMENT OF OPPORTUNITY'

Although this study has found the association between water hardness and mortality to be weak, an essential aspect of Gardner's (1973) study which has not been previously considered in Chapter 6 or in this chapter deserves explicit attention. As a result of earlier studies suggesting a strong negative relationship, there has been a discussion of the feasibility of conducting a

> 'controlled experiment of changing the character of the water supply to one of each of a number of pairs of towns, each pair being matched as far as possible in all relevant ways' (Gardner, 1973, 437).

As an alternative to the massive difficulties and expense of performing such a trial, Gardner attempted an 'experiment of opportunity' by examining those County Boroughs which had experienced a substantial change in the water supply in the 30 years prior to 1961. Eleven County Boroughs were identified that had undergone a change of water hardness of approximately 50 parts per million during this time period; in six the water had become softer while in five the hardness had increased. Gardner attempted to discover whether these changes had resulted in changes in death rates and, in particular, he examined the changes in average annual death rates between 1948-54 and 1958-64 for cardiovascular and non-cardiovascular diseases. For cardiovascular mortality at age 45-64 years the death rate had increased over the period and the hypothesis was, therefore, that any changes would be above the average rise in the towns where the water supply had softened, and below the average where the water had been hardened. For the 65-74 age group the cardiovascular

death rate had fallen and consequently the hypothesis was that the decrease would be greatest in the towns that had experienced harder water and least in the towns which had received softer supplies. Defining 'relative risk' as the ratio of the death rate in 1948-54 to the rate in 1958-64,

> 'Tests of the statistical significance of the results have been applied to the logarithms of relative risk, assuming these are normally distributed within each group. The means and standard deviations under the null hypothesis were estimated from the 'no-change' group which is numerically dominant and normal deviate tests were then carried out for each of the 'hardened' and 'softened' groups of towns. For cardiovascular disease seven out of the eight changes in average death rates were in the hypothesised direction, although only two reached the nominal level of statistical significance using a two-tailed test. It may be argued that a one-tailed test is appropriate in this case, but the previous statement is unaffected by doing this. Taking the eight normal deviates as the outcomes of independent tests of the effect of changing water hardness on cardiovascular mortality, and combining them the overall significance of the results is high ($P > 0.02$). For the non-cardiovascular death rates no significant differences are found'
> (Gardner, 1973, 439).

Initially, these appear to be impressive results, for Gardner has been able to do more than merely examine the water-hardness/disease relationship by statistically controlling for other variables. Indeed, by examining changes over time he has approached the laboratory experiment, where the effects of deliberately manipulating the explanatory variable are carefully observed. The relationship therefore appears to have passed a critical test and, according to Gardner (1973, 439) 'it would seem that the technical problems of increasing the hardness of soft water should be explored'. Crucially, and regrettably however, this study does not conform to standard experimental practice, because Gardner has not controlled for other variables that may also

have changed over this period. Moreover, from an exploratory viewpoint, summary statistics and statistical tests based on the 'ideal conditions' of normality and independence should be treated with caution and there is a genuine need for the data for the two periods to be more closely scrutinised.

Ranking the cardiovascular death rates for the 45-64 age group reveals considerable changes in rank between the two periods. While it is true that some County Boroughs do not change rank (East Ham, 75th; Ipswich, 83rd; Kingston-upon-Hull, 63rd; Norwich, 80th; Southampton, 72nd), the average change of rank is over 10 positions. Indeed, some County Boroughs dropped in rank by over 20 positions (Bolton: 8th in 1948-54, 39th in 1958-64; Exeter: 32nd and 78th; Gloucester: 33rd and 74th; Gloucester: 33rd and 74th; Great Yarmouth: 49th and 81st; Grimsby: 35th and 62nd; Tynemouth: 9th and 32nd) while others gained in rank by at least 20 positions (Bootle: 52nd in 1948-54, 13th in 1958-64; Burton: 62nd and 38th; Wakefield: 42nd and 22nd; Wallasey: 31st and 10th; Merthyr Tudful: 40th and 1st). Only one of these very large changes (that for Burton) involves a town that has undergone a substantial change in its water supply and, moreover, Canterbury and Leicester, which experienced harder water, displayed a decrease in rank which is contrary to the hypothesis of soft water: heart disease. Similarly, for the 65-74 cardiovascular mortality rates, although Leicester and Sunderland changed water supplies over the period they did not experience a change in rank, while the introduction of softer water was accompanied by a decrease in rank for Derby.

It is possible to draw a number of conclusions from this examination of changes in rank. Firstly, mortality rates appear to be quite variable over relatively short periods of time. (This does not appear to be the result of ratios based on small numbers, for numerically the number of cardiovascular deaths is quite large.) Secondly, the largest

variations in rank are for towns that have not changed their water supply. Thirdly, while most of the towns that have changed their water hardness have also experienced a change in death rate that is in accordance with the hypothesis of soft water: heart disease, there are several exceptions to this relationship. In summary it appears that, while water hardness may have a slight effect on heart disease, Gardner's (1973) study has ignored other changes that have occurred in the County Boroughs which have resulted in major changes of rank.

Over the period 1948-54 to 1958-64 there have been numerous changes in the level of possible explanatory variables for each County Borough, and it can be argued that some of these changes are linked to changes in mortality. For example, Table 7.18 shows the Standardised Mortality Rate for three periods by social class. While such comparisons are fraught with difficulties,[7] a clear pattern can be seen in the table, with the differences between the social classes becoming more pronounced over time. In contrast, for cardio-vascular disease there is evidence that the differences between the social classes have diminished over time (Blaxter, 1976) and, obviously, any study that attempts to investigate

Table 7.18

Male SMRs by social class
1930-32, 1949-53, 1959-63 : all causes

	Social class				
Date	I	II	III	IV	V
1930-32	90	94	97	102	111
1949-53	86	92	101	104	118
1959-63	76	81	100	103	143

Source: Registrar General (1971) Decennial supplement: occupational mortality HMSO, London.

mortality changes must take into account the social class composition of the County Boroughs. With regard to the Boroughs that have changed their water supply, Roberts and Lloyd (1972) have argued that the 'harder' and 'softer' towns are different in terms of social class. Indeed, on examining the data, towns receiving a softer water supply (Sunderland, South Shields, Derby, Coventry, Burton, Bristol) had a higher proportion of males of lower social class in 1951 and 1961 than did towns which received harder water (Lincoln, Leicester, Canterbury, Bournemouth, Birkenhead). Yet the greatest change (1951-61) in the number of people of low social class occurred in the second group of towns. While it is impossible with this evidence to evaluate the separate influences of relative levels of low social class and changes in class composition on mortality rates, it can at least be argued that ignoring these variables may lead to errors of interpretation.

Other variables associated with social class and disease have also changed over this period. For example, the studies of the Tobacco Research Council (Todd, 1969) have revealed a substantial decrease in the number of smokers of high social class, while among males classified as Social Class V, little decrease has been discernible. Moreover, researchers engaged in the General Household Survey have found that there are proportionally more ex-smokers in the South-East and East Anglia than in other parts of Great Britain. Ample evidence can therefore be presented to demonstrate that Gardner's study, which examines changes in water hardness to the exclusion of other explanatory variables must be regarded with great caution.

In order to assess separately the contribution of changes in water hardness the effect of other changes must be controlled in some way. One method of achieving this statistically is to perform a multiple regression analysis employing the following variables.

Dependent variable : the difference for each County Borough in the number of deaths between the two time periods (1948-54 and 1958-64)

Explanatory variables : the differences in the levels of the variables that have been previously used as explanatory variables in this study.

It was decided to investigate the deaths from cardiovascular disease previously analysed by Gardner and to concentrate on male deaths in the 45-64 age group. The latter choice was made because, as was established earlier in this chapter, data for this group are thought to be more reliable indicators of cause-specific mortality than data relating to older age groups. Additionally, it was decided to examine changes in the explanatory variables 1931 to 1951 as well as those for 1951 to 1961, the intention being to allow for 'lagged' relationships. The control variables were easily measured, as they were simply the differences in the number of people in each group (45-49, 50-54, 55-59, 60-64) and the change in the number of households 1951 to 1961. Similarly, the changes in the occupation, overcrowding, unemployment and total water hardness variables presented no difficulty. However, because of the lack of suitable data, changes in low social class and in the provision of exclusive use of fixed baths can only be measured for 1951-1961. Even more unfortunately, values for the important variables of air pollution and cigarette consumption are only available at one point in time.

Bearing in mind these deficiencies, a ridge regression model was fitted to the data. The results were that the unemployment and social class variables had reasonably substantial standardised coefficients (0.16 for 1931-51 and 0.13 for 1951-61), but the coefficients associated with

water hardness were again close to zero (-0.008 for 1931-51 and -0.019 for 1951-61). However, these results should be treated with caution, for closer inspection reveals that they are derived from an inadequate model; the residuals appear to be spatially autocorrelated and the Ramsey RESET test rejects the null hypothesis of no specification at the 0.01 probability level. On the basis of previous results it can be suggested that the model is mis-specified due to the omission of air pollution and cigarette smoking data, and it is therefore impossible to specify an improved model due to the lack of suitable information.[8]

The fitted model is clearly mis-specified but, while definite conclusions cannot be drawn, it is perhaps possible to make some conjectures. The omitted variables (changes in air pollution and cigarette consumption) can be anticipated not to be highly related to changes in water hardness, and consequently the degree of bias imparted to the regression coefficients associated with the water-supply variables should be slight (Chapter 4). If this is the case, it can be speculated that the inclusion of the omitted variables would not drastically alter the values for the water hardness coefficients, and again the conclusion would be drawn that hardness is not a major determinant of mortality variations.

In spite of the problems of data availability, it is possible to draw some conclusions from this 'experiment of opportunity'. Firstly, mortality rates are quite variable over relatively short periods of time, and for the 83 County Boroughs it is not uncommon for a town to experience a change of rank of over 20 positions in less than 10 years. Secondly, Gardner's use of statistical summary measures does not convey an overall picture of the changes of mortality rates in relation to changes in water hardness. A number of towns have changed rank in the opposite direction to the hypothesis of soft water:high mortality and moreover, the

Boroughs which exhibited the greatest change in rank were not those which experienced major changes in water supply. Thirdly, it can be claimed that because Gardner examined only changes in water hardness to the exclusion of changes in other possible explanatory variables, his results must be regarded with suspicion. Finally, and most speculatively, it can be suggested that when account is taken of changes in the level of a number of explanatory variables, changes in water hardness are not related to changes in mortality.

CONCLUSIONS

In many respects this chapter represents the culmination of previous work in this thesis, for an attempt has been made to meet the criticisms in Chapter 6 concerning the relationship between water hardness and mortality, by the use of methods developed in Chapters 1 to 5. Without the ridge procedures introduced in Chapter 2, it would have been impossible to analyse the multicollinear data which were needed to build a realistic model. Without the insights provided by the exploratory methods discussed in Chapters 3 and 4, the inadequacy of the original model would not have been revealed. Without the fundamental distinction between irreducible and reducible ratios that was established in Chapter 5 there would have again been the inferential error that is so characteristic of epidemiological research.

With regard to the empirical findings of this chapter, two conclusions can be drawn. Firstly, by adopting a different and, it is argued, a more appropriate statistical methodology, results have been obtained which diverge from those obtained by Gardner (1973) using the same data. As his study is probably the most comprehensive and statistically aware analysis of the water hardness/disease relationship, these divergent results cast strong suspicion on the findings of the numerous other studies that have been undertaken to examine this relationship. Secondly, the present study does

not provide strong support for the widely advocated water hypothesis. Even if this hypothesis is viewed in the most favourable light (by comparing the results for hardness, geographical position and climate), the negative and consistent relationship must be seen as a weak one. In comparison, the air pollution/cigarette interaction variable, cigarette consumption per se, pre-war unemployment levels and the other variables listed in Table 7.10 to 7.15, have been shown to be much more influential. On this evidence the major determinants of mortality variations are not variations in the physical environment, but are instead aspects of man's 'way of life'. This conclusion is arguably most unfortunate: water hardness can be readily controlled by adding calcium carbonate or by mixing and transfering supplies, but human habits, attitudes and circumstances are much more difficult to alter radically. Even more regretably, the results relating to the influence of pre-war unemployment suggest that, at least to some extent, present day mortality rates may be the outcome of past conditions. If this is the case, improvements made today will not have the desired effect until a considerable period of time has elapsed. Moreover, only after a considerable length of time will it be possible to evaluate fully any prophylactic measures that have been taken.

Reviewing the evidence on the effects of water hardness, some may make the comment that if Schroeder on his return from Japan had performed a detailed exploratory analysis instead of relying on a test of areal differences and correlation analysis of percentage data, a great deal of research effort may have been re-focused on other determinants of mortality variations. Such a comment would be rather churlish and unfair, however, for in the last twenty years there has been considerable technical and conceptual progress. Technically, the modern computer allows the analyst to rapidly calibrate, evaluate and reformulate a model, whereas

the researcher working by hand or with early computers would be tempted to make inferences on the basis of the first model calibrated. Similarly, there have been major conceptual changes, for the exploratory researcher is no longer willing to rely on summary statistics based on unrealistic assumptions. The combined result of these changes, in spite of poor data availability and the highly aggregate nature of the study, is that it has been possible to make a contribution to the epidemiology of disease.

CHAPTER 7 : NOTES

1. All the maps in this chapter have been drawn directly onto 35mm microfilm by using the facilities of the University of London Computer Centre and the program (CHORMAP) developed by Margaret Jeffery of the Department of Geography, London School of Economics.

2. Only male mortality will be analysed in this chapter, for males are more exposed to occupational hazards than females, and thus maps of male death rates reflect the total complex of environmental conditions. Moreover, men in general die earlier than women, and an equivalent map to Figure 7.1a for women reveals that the highest female death rate amongst the County Boroughs is lower than the lowest male death rate for the same age group and time period.

3. The average death rate is given by the second highest value in the key to the maps: this is the first nested mean.

4. Social class I is positively related to most of the variables included in the model but it can be anticipated to be negatively related to mortality when population size is controlled. On the basis of the simulations of Chapter 2, it can also be suggested that the inclusion of this variable will bias all the regression coefficients of the model. Indeed, when both Social Class I and V were included in the model, not only was there an unexpected positive effect for Social Class I but the u statistic was over sixteen times higher than when this variable was omitted. It is important to stress that similar problems are not expected to occur when water hardness is included in the

model. While Gardner (1973) found hardness to be negatively related to mortality, it is also generally negatively related to the other explanatory variables that constitute the model.

5. In terms of the actual application of the exploratory procedures, the calculations have been performed by Fortran computer programs written by the author for the ICL 1900 computer at the University of Southampton.

STAGE	COMPUTER PROCEDURES
1:	Read tape of Gardner's (1973) data (supplied by the SSRC Survey Archive, University of Essex); read data derived from Census; convert reducible ratios to raw values and store all data in a file.
2:	Read file and analyse data with RIDGE program; inspect ridge plots, choose value of k, re-run ridge to derive ridge coefficients (standardised and unstandardised) and residuals.
3:	Read residuals with NEST program and calculate nested means; input residuals and nested means to CHORMAP at the University of London Computer Centre.
4.	Read residuals with NORM program (written by the author in GENSTAT and run at ULCC) and draw normal probability plots.
5.	Input unstandardised regression coefficients into PALT and calculate partial residuals; draw graphs on CALCOMP plotter.
6.	Input unstandardised regression coefficients into BOX-COX program which is a modified version of Chang's (1977) program.
7.	Cross-validate results with program RAMS which has been produced by the Computer Institute for Social Science Research, Michigan State University.

6. Rigorous comparisons are not possible for three reasons. Firstly, as discussed in Chapter 1, an analyst should not use standardised regression coefficients to compare different models. Secondly, dependent variables differ between the models, for whereas Gardner used death rates the present study analyses the number of deaths. Thirdly, Gardner's coefficients relate to calcium whereas the present study uses total water hardness because calcium values are not available for all the County Boroughs. An analysis using the available figures for calcium is discussed later.

7. At each time period a different classification of occupations has been adopted. While the 1959-63 mortality rates have been adjusted for groups I and V to the 1950 classification, this has not been done for groups II, III and IV and no adjustment at all has been performed for the 1930-32 rates.

8. Another model was calibrated which included cigarette consumption and an interaction variable (domestic air pollution * cigarettes) measured at one time point. Unfortunately, this did not improve the specification of the model nor did the inclusion of measures of the explanatory variables for 1961 result in a well-specified model.

CHAPTER 7 : BIBLIOGRAPHY

ACHESON, E.D., COWDELL, R.H., HATFIELD, E. and MacBETH, R.G. (1968): Nasal cancer in woodworkers in the furniture industry
British Medical Journal 2, 587-601.

ALDERSON, M.R. (1965): The accuracy of certification of death
unpublished M.D. thesis, University of London.

ALDERSON, M.R. (1974): Central Government Routine Health Statistics
Heinemann, London.

ALDERSON, M.R. (1976): An introduction to epidemiology
MacMillan, London.

ARMSTRONG, R.W. (1969): Standardised class intervals and rate computation in statistical maps of mortality
Annals of the Association of American Geographers 59, 382-390.

ARMSTRONG, W.A. (1972): The use of information about occupation
in Wrigley, E.A. (ed.) Nineteenth Century Society
Cambridge University Press, Cambridge.

BARNARD, K.C. (1978): The residential geography of the elderly: a nutiple scale approach
unpublished Phd. Thesis, Department of Geography, University of Southampton.

BLAXTER, M. (1976): Social class and health inequalities
in Carter, C.O. and Peel, J. (eds.) Equalities and inequalities in health
Academic Press, London.

BOOTS, B.N. (1979): Population density, crowding and human behaviour
Progress in Human Geography 3, 13-63.

BROWN, W.G. (1973): Effect of omitting relevant variables versus use of ridge regression in economic research
Special Report No. 394
Oregon Agricultural Experiment Station.

BURCH, P.J.R. (1978): Smoking and lung cancer: the problem of inferring cause
Journal of the Royal Statistical Society Series A 141, 437-477.

CALHOUN, J.B. (1962): Population density and social pathology
Scientific American 206, 139-148.

CHANG, H.S. (1977): Functional forms and the demand for meat in the United States
The Review of Economics and Statistics 59, 355-359.

DALY, C. (1959): Air pollution and causes of death
British Journal of Preventive and Social Medicine 13, 14-27.

DARLINGTON, R.B. (1978): Reduced-variance regression
Psychological Bulletin 85, 1238-1255.

DOLL, R. and HILL, A.B. (1950): Smoking and carcinoma of the lung
British Medical Journal 2, 739-748.

DOLL, R. and HILL, A.B. (1952): A study of the aetiology of carcinoma of the lung
British Medical Journal 2, 1261-1285.

DOLL, R. and HILL, A.B. (1954): The mortality of doctors in relation to their smoking habits - a preliminary report
British Medical Journal 1, 1451-1455.

DOLL, R. and HILL, A.B. (1964): Mortality in relation to smoking. Ten year's observation of British doctors
British Medical Journal 1 1399-1410.

DUDLEY, E.F., BELDIN, R.A. and JOHNSON, B.C. (1969): Climate, water hardness and coronary heart disease Journal of Chronic Disease 22, 25-48.

FORSTER, R. (1966): Use of a demographic base map for the presentation of areal data in epidemiology British Journal of Preventive and Social Medicine 20, 156-171.

GALLE, O.R., GOVE, W.R. and McPHERSON, J.M. (1972): Population density : what are the relationships for man? Science 176, 23-30.

GARDNER, M.J. (1973): Using the environment to explain and predict mortality Journal of the Royal Statistical Society Series A 136, 421-440.

GARDNER, M.J. (1976): Softwater and heart disease? in Lenihan, J. and Fletcher, W.W. (eds.) Environment and man, Blackie, Glasgow.

GARDNER, M.J., CRAWFORD, M.D. and MORRIS, J.N. (1969): Patterns of mortality in middle and early old age in the County Boroughs of England and Wales British Journal of Preventive and Social Medicine 23, 133-140.

GIRT, J.L. (1972): Simple chronic bronchitis and urban ecological structure in McGlashan, N.D. (ed.) Medical geography : techniques and field studies Methuen, London.

GROEN, J.J., TIJONG, K.B. KOSTER, M., WILLEBRANDS, A.F. VERDONCK, G. and PIERLOOT, M. (1962): The influence of nutrition and ways of life on blood cholesterol and the prevalence of hypertension and coronary heart disease among Trappist and Benedictine Monks American Journal of Clinical Nutrition 10, 456-470.

GUJATARI, D. (1970): Use of dummy variables in testing for equality between sets of coefficients in linear regression : a generalisation The American Statistician 24, 18-23.

HAGGETT, P., CLIFF, A.D. and FREY, A. (1977): Locational analysis in human geography
Arnold, London.

HART, J.T. (1976): The distribution of needs and resources in the National Health Service - data from S. Wales
paper presented to Regional Aspects of the Re-organisation of the National Health Service Conference, Regional Studies Association.

HILL, A.B. (1925): Internal migration and its effects upon the death rates : with special reference to the county of Essex
HMSO, London.

HOERL, A.E. and KENNARD, R.W. (1970): Ridge regression : biased estimation of non-orthogonal problems
Technometrics 12, 55-67.

HOWE, G.M. (1970): National atlas of disease mortality in the United Kingdom
Nelson, London.

HUNTER, J.M. and YOUNG, J.C. (1971): Diffusion of influenza in England and Wales
Annals of the Association of American Geographers 61, 637-653.

JONES, D.R. and BOURNE, A. (1977): A model of the National Health Service for structural analysis and planning
Socio-Economic Planning Sciences 11, 221-231.

JONES, K. (1975): A geographical contribution to the aetiology of chronic bronchitis
unpublished BSc. dissertation
Department of Geography, University of Southampton.

JONES, K. (1978): Percentages, ratios and inbuilt relationships in geographical research: an overview and bibliography
Discussion Paper No. 2, Department of Geography, University of Southampton.

KOSHAL, R.K. and KOSHAL, M. (1976): Air pollution and cancer: an econometric approach. Clean Air (Australia) 10, 19-22.

KRUSKAL, W.H. (1960): Some remarks on wild observations. Technometrics 2, 1-4.

LAMBERT, P.M. and REID, D.D. (1970): Smoking, air pollution and bronchitis in Britain. Lancet 1, 853-857.

LAVE, L.B. and SESKIN, E.P. (1970): Air pollution and human health. Science 169, 723-733.

LAVE, L.B. and SESKIN, E.P. (1976): Air pollution and human health. John Hopkins Press, Baltimore.

McCALLUM, B.T. (1972): Relative asymptotic bias from errors of omission and measurement. Econometrica 40, 757-758.

MOORE, J. (1980): Preventive mental health - now! Radical Community Medicine 2, 21-22.

MURRAY, M.A. (1962): The geography of death in England and Wales. Annals of the Association of American Geographers 52, 130-149.

National Food Survey Committee (1973): Household food consumption and expenditure: 1970 and 1971. HMSO, London.

REID, D.D. (1958): Environmental factors in respiratory disease. Lancet 1, 1237-1239.

ROBERTS, C.J. and LLOYD, S. (1972): Association between mortality from ischaemic heart disease and rainfall in S. Wales and the County Boroughs of England and Wales. Lancet 1, 1091-1093.

SCRIPTER, M.W. (1970): Nested-means map classes for statistical maps. Annals of the Association of American Geographers 60, 385-393.

SMITH, V.K. (1976): The economic consequences of air pollution
Ballinger, Cambridge, Mass.

STOCKS, P. (1959): Cancer and bronchitis mortality in relation to atmospheric deposit and smoke
British Medical Journal 1, 74-79.

TODD, G.F. (1969): Statistics of smoking in the United Kingdom
Tobacco Research Council, London.

TURNER, W.C. (1964): Air pollution and respiratory disease
Proceedings of the Royal Society of Medicine 57, 618-620.

WALLER, R.E. (1967): Bronchi and lungs: air pollution in
Raven, R.W. and Roe, F.J.C. (eds.) The prevention of cancer
Butterworths, London.

WEST, R.R. and LOWE, C.R. (1976): Mortality from ischaemic heart disease - inter-town variation and its association with climate in England and Wales
International Journal of Epidemiology 5, 195-201.

WICKERS, M.R. (1972): A note on the use of proxy variables
Econometrica 40, 759-761.

WOHL, A.S. (1973): Unfit for human habitation in
Dyos, H.J. and Wolff, M. (eds.) The Victorian City : images and realities
Routledge and Kegan Paul, London.

WYNN, M. and WYNN, A. (1980): Are there really diseases of affluence?
New Society 53, 500-501.

C H A P T E R 8

ANALYSING DISEASE/ENVIRONMENT RELATIONSHIPS: PROBLEMS AND PROSPECTS

'This book has no beginning and no end. It is a tapestry woven at dawn and undone at dusk .... And yet, even though work has no beginning and no end, a sentence must sometimes start and sometimes it must end' Olsson (1975, vii, 502).

INTRODUCTION

A fundamental supposition of this thesis is that there is a need to analyse disease patterns not in terms of the laboratory experiment but in terms of the human population. Accepting this supposition and making use of present knowledge and readily available data has resulted in an aggregate, ecological approach to the study of disease/environment relationships. This research has also adopted the notion of Hirschi and Selvin (1966, 254) that

> 'since no one proposes trying to give people cancer ... the fruitful way toward better causal analyses ... is to concentrate on improving the statistical approach'.

The first part of this final chapter attempts to summarise what has been achieved by adopting an exploratory statistical approach. While this résumé will obviously deal with the material that has been presented in Part I of this thesis, it will also provide suggestions for further research. The second, and more substantial section considers the problem of aggregation, which has particular importance for the empirical results of Chapter 7.

EXPLORATORY ANALYSIS: OVERVIEW AND FURTHER RESEARCH

Confirmatory and exploratory analysis

The dominant approach to quantitative geographical data analysis is that of confirmatory hypothesis testing. Mosteller and Tukey (1977, 25) point out that a 'caricature of one recipe' for performing a hypothesis test is:

> 'Apply a significance test to each result, believe the result implicitly if the conventional level of significance is reached, believe the null hypothesis otherwise. Such a complete flight from reality and its uncertainty is fortunately rare'.

But, regrettably, it is exactly this type of approach that is used (or more correctly, reported) by geographers. Indeed, an examination of those geographical articles that use statistical procedures will reveal a heavy reliance on significant tests with those results exceeding the 99.9 per cent commonly given four asterisks as an acknowledgement of their 'significance' (see for example, Cliff and Ord, 1973, 49). But, as discussed in Chapter 1, such 'significant' results are often only found after considerable trial and error and re-application of the tests. If this is the case, the usual significance levels have little meaning and, indeed, with traditional statistics it is not possible to re-apply a hypothesis test to the same body of data. Moreover, while significance tests appear rigorous and precise, they are often based on assumptions that are demanding and unrealistic such as normality, independence and no geographical patterning. In contrast, the exploratory approach recognises the need for trial and error and the value of graphical plots, and it is sceptical about summary measures and statistical assumptions. In many ways exploratory statistics is a radical departure from the dominant approach in quantitative geography, and it is contended in this thesis that it is a substantial improvement over the consensus view.

With regard to the future use and continued development of exploratory analysis it is possible to make two general points. Firstly exploratory procedures appear to offer an interesting and valuable approach to the problem of teaching statistics to geographers. Exploratory techniques with their emphasis on graphical display may have considerable appeal for the student who fears mathematical symbols and resists the rote learning of hypothesis-testing procedures. The more sophisticated student who reacts unfavourably to the simple definite answers of significance tests may be stimulated by the exploratory procedures refusing to deliver

neat answers but instead revealing the true complexity of the real world. Moreover, exploratory procedures by their trial-and-error approach offer an opportunity to 'learn by doing'. Interestingly, the Open University has recently adopted an exploratory textbook (Erickson and Nosanchuk, 1977) as a course text.

Secondly, if the explanatory approach is to be adopted and improved by other researchers there is a genuine need for wider discussion and informed, critical evaluation.[1] Until recently, while there have been a number of general discussions and presentations of exploratory statistics, there has not been a comprehensive review specifically written for geographers. However, Cox and Jones (forthcoming) present such a much-needed overview,[2] and it is to be hoped that this will provoke a critical examination of present practices and of possible alternatives.

In the rest of this section an attempt is made to summarise the major arguments concerning exploratory statistics that have been presented in this thesis and to outline areas for future research.

Multicollinearity

Chapter 2 dealt with the widespread problem of multicollinearity; the commonest approach is to ignore it with the likely outcome of inferential error. Some geographers, it is true, have advocated a stepwise solution, while others have suggested the use of principal components analysis. Yet, both procedures have marked disadvantages and their use with multicollinear data is not to be advocated. Ridge regression, in contrast, has much to recommend it and this graphical exploratory procedure appears not only to indicate multicollinear data, but it also provides estimates that are an improvement over least-squares. Even so, it must be acknowledged that a considerable amount of research into the

technique is still required. For example, the simulation work reported in Chapter 2, only used estimates derived by examining the ridge trace. There are a large number of automatic methods of calculating the ridge estimates and a thorough investigation is needed to establish which procedures perform best and under what conditions. Moreover, the reliance on a single summary statistic ($_U$) to detect severe bias is not in the spirit of exploratory analysis, and further research is required into the relationships between, on the one hand, bias, magnitude and sign of the regression coefficients and, on the other hand, magnitude and sign of the inter-relationships between the explanatory variables.[3] Finally, although it has not been previously considered as a solution to multicollinearity, the 'jacknife' procedure deserves to be investigated as a possible supplement to the ridge technique. This simple procedure consists of dropping one observation at a time and re-estimating the regression model so that there are as many sets of regression coefficients as there are observations. The mean or median of the regression coefficients associated with a particular explanatory variable is then calculated and it is this value that is interpreted as being the most appropriate estimate of the true regression parameter. While preliminary results are encouraging, the jacknife requires to be evaluated with simulated data of known properties before it can be recommended as an effective way of overcoming multicollinearity.

Spatial autocorrelation

Chapter 3 dealt with the common geographical problem of spatial autocorrelation. Despite the important work of Cliff and Ord (1973) in this field, many geographers continue to violate the assumption of no autocorrelation and this can be critical for the interpretation of their results. Two important and distinctive aspects of the content of Chapter 3 were the criticism of autoregressive modelling and the

emphasis placed on graphical means of detecting autocorrelation. It was argued that it is impossible to have a true random disturbance term that displays autocorrelation. Consequently, autocorrelation can only result from a mis-specified model and those researchers who have developed autoregressive models to analyse their data should have concentrated their efforts on improving the specification of the model. In relation to the second distinctive feature of Chapter 3, that of the detection of autocorrelation, it was argued that because of difficulty in specifying weights for the Cliff and Ord test, and because of the inability to make repeated applications of the test, a simple map of the residuals was to be preferred to a confirmatory significance test. To support this argument it is obviously necessary that the human eye can detect a spatial pattern in which adjacent areas have a similar value. Preliminary work suggests that the spatial pattern must be 'obvious' to the human eye before autocorrelation seriously affects estimates, but further research is required in this area. Tukey's (1977) robust method of separating time-series data into 'smooth and rough' or 'pattern and noise' could usefully be transferred to the analysis of areal data. An application of this method, which is based on running medians, to some of the data presented in Cliff and Ord (1973) produced a successful enhancing of the spatial pattern that was said to have existed in the data (Jones, 1979). While such research is still in its infancy, the general impression created so far is that a visual assessment of the degree of autocorrelation is sufficient to allow a decision to be made about the adequacy of a model.

Specification-error analysis

In essence, the fourth chapter dealt with those assumptions of the regression model which had not been considered in the previous chapters. While a great deal of material was covered in Chapter 4 it is perhaps possible to

emphasize a number of points which are of particular importance. Firstly, it was a general theme that it is indeed possible to detect mistakes or model mis-specification by a thorough analysis of residuals. Secondly, considerable emphasis was placed on graphical procedures as a means of detecting outliers, heteroscedasticity and incorrect functional form. Thirdly, the informed use of transformations was advocated as an appropriate means of estimating non-linear and heteroscedastic models. Finally, and perhaps most importantly, a method was suggested for overcoming the difficult problem of the choice of functional form. The suggested procedure, based on Box-Cox transformations and partial residuals, appears on the basis of the simulations in Chapter 4 to be a valuable one, yet much research remains to be undertaken. In particular, further simulations need to be performed for cases in which the model has more than one incorrect functional form. Moreover, the ad hoc decision in Chapter 7 to use ridge estimates to calculate the partial residuals with multicollinear data, needs to be examined in a further simulation experiment.

Ratios

The use of ratios, proportions and percentages is widespread in geography. Indeed, Haynes (1978) has shown that among a sample of quantitative human geography papers, 76 per cent used proportions and 24 per cent used densities. Unfortunately, most geographers analyse ratios as if they were simple numbers and do not take into account the special characteristics of such variables. Chapter 5 dealt with one serious problem of ratios, that of inbuilt relationships. As a further illustration of the difficulty, consider Table 8.1, the first part of which shows the correlation coefficients calculated by King (1969, 177) for the relationships between different crops in the counties of Ohio in 1940; each variable is the proportion of agricultural land on which a particular crop is being grown. Table 8.1b represents

Table 8.1

Correlations between proportions of cropland under different crops: 88 counties of Ohio, 1940

(a) calculated correlation coefficients

Corn				
Wheat	+ .46			
Oats	- .33	+ .08		
Soybeans	+ .01	+ .13	+ .38	
Hay	- .45	- .64	- .32	- .60
	Corn	Wheat	Oats	Soybeans

(b) inbuilt correlation

Corn				
Wheat	- .34			
Oats	- .23	- .17		
Soybeans	- .17	- .13	- .09	
Hay	- .40	- .29	- .20	- .15
	Corn	Wheat	Oats	Soybeans

Source: Evans and Jones (forthcoming)

the inbuilt correlations which would be obtained by using proportions if there was no relationship whatsoever between the crops.[4] Comparing the two parts of the table it is clear that the King's observed correlation between hay and corn is essentially an outcome of analysing proportions rather than absolute values. Similarly, the strength of the negative correlations between hay and oats, hay and wheat, and corn and oats are a direct result of analysing ratios.

More fundamentally, King can be accused of ignoring the distinction between irreducible and reducible ratios. For the former the ratio itself is the variable of interest while, in the latter, the ratio has been merely formed to control for the denominator variable. This crucial distinction (which has been largely neglected by the literature on the analysis of ratios) results in the conclusion that reducible ratios should not be analysed by regression analysis. The model should be specified in terms of absolute numbers and the denominator variable included as an independent variable. Not to proceed in this way can have serious consequences for the interpretation of the resultant model.

Given the inherent problems of analysing ratio data and their widespread use in geography it is essential that researchers be made aware of these difficulties. There has been occasional recognition of the special properties of ratios, but this recognition has usually come as an afterthought rather than belonging to the core of a work. For example, after many chapters using ratios and percentages as the main illustrations, Johnston (1978) concludes with a section suggesting that there may be a problem with such data. However, two recent works provide a full discussion of the geographical analysis of ratios. In Jones (1978) there is an extended treatment of the techniques for calculating the degree of inbuilt correlation, together with a computer program implementing the procedures. In addition, in Evans and Jones (forthcoming) there is a discussion of the

problems of mapping ratio data and analysing regression models with ratios as the dependent variable. Here it may be noted that when mapping ratios there is a major difficulty when small numbers are involved. For example, if the national death rate from a particular disease and a particular age group is 0.1, a small community may have widely fluctuating death rates over time. If there are 10 people in that age group and one person dies, the community has an average death rate, but if two people had died the rate would be twice the national average and if no people had died the death rate would have been zero. Clearly, little reliance can be placed on ratios with small numbers but, fortunately, it is possible to use a chi-square measure to overcome this difficulty.[5] Evans and Jones have also discussed the analysis of models with a ratio dependent variable; such models can be expected to suffer from nonsensical values and heteroscedasticity. It is recommended that estimation is not performed by ordinary least squares, but the ratio should be transformed and estimated by weighted least squares. In conclusion, as Evans and Jones have written:

> 'If the ratios can reasonably be avoided, they should be: this may save embarrassment'.

AGGREGATE DATA ANALYSIS

Ecological fallacy

Having briefly reconsidered the major aspects of Part I of this thesis, the discussion now turns to a problem which was cursorily discussed in the Introduction and has important consequences for the results presented in Chapter 7 of this thesis.

The difficulty of inferring individual characteristics from aggregate data is widely recognised and is usually known as the ecological fallacy. The classic paper on the subject is that of Robinson (1950) which demonstrated that an aggregate-level regression coefficient need not equal the

corresponding individual coefficient. This argument was supported by empirical examples of discrepancies between individual and aggregate coefficients, some of which even had reversed signs. One of the examples that Robinson used to illustrate his paper was the relationship between being negro and being illiterate. Using the nine Census Divisions of the United States the correlation between the percentage of population aged over 10 who were negro in 1930 and the percentage of the population aged over 10 who were illiterate in 1930 was 0.946. However, at the individual level the correlation was 0.203, and there was, therefore, less than a 5 per cent chance (0.203^2) that a black person was illiterate, despite the fact that black people were spatially concentrated in the same area as illiterates. In another example, the aggregate correlation at the Census Division level was -0.619 between percentage illiterate and percentage foreign born, but at the individual level a more plausible weak positive association (+0.118) was established.

Accepting the results of the Robinson study for the moment, it appears that the results of Chapter 7 should be regarded with strong misgivings. For example, although it has been found that areas with high cigarette consumption have high rates of mortality from bronchitis, at the individual level this does not mean that people who die from bronchitis are smokers. Even if the claim is made that the major causal variable under study (water hardness) is a variable that to some extent affects all the individuals in the area, the reader may remain suspicious. Even if it is claimed, following Allardt (1969) and Baldwin, Bottoms and Walker (1976), that the deliberate use of the ecological fallacy is an important speculative tool for 'clearing the ground' and making tentative statements about individuals, suspicion will probably remain. The examination of changes over time and different age groups and diseases may allay some of these doubts but, such is the influence of Robinson's

arguments, aggregate data analysis is held in low regard even by the many geographers who continue to use the method. However, recent research into the problems of the ecological fallacy has revealed that the problems of aggregate causal models need to be considered in terms of the models' specification.

Aggregation and mis-specification

Analysing the same data as used by Robinson, the regression coefficient relating illiteracy to being foreign born is 0.07 at the individual level. But when, at the state level, the percentage of a state's population that is illiterate is related to the percentage of a state's population that is foreign born, the regression coefficient is an unlikely -0.29. Ignoring the problems of analysing such ratio data (at the state level the variables have common denominators) and concentrating on the specification of the model, it can be suggested that the Robinson model is mis-specified and omits important explanatory variables. For example, it is arguable that a model of illiteracy should assess the influence of non-foreign minority groups and the influence of educational provision. In the 1930's the foreign-born population was concentrated in the north-east and north-central states where the most extensive public school systems were to be found. Conversely, the negro population was heavily concentrated in areas with relatively low school attendance. Robinson's foreign-born variable, therefore, measures the combined effects of minority-group illiteracy rates and the quantity of education services on the states' illiteracy rates. As demonstrated in Chapter 4, the omission of important variables from a regression model may lead to bias, and it can be suggested that the implausible result found by Robinson is the result of specification error and not aggregation bias. Hanushek and Jackson (1977) have re-analysed the 1930 data used by Robinson but in their model, percent illiterate is related to percent foreign born,

the percentage of a state's population aged 7-13 enrolled in schools (as a measure of educational provision), percent black, percent Mexican and percent Indian. In this model the partial regression coefficient associated with percent foreign born is +0.12. This estimated coefficient has the correct a priori sign and is very close to the individual level of coefficient of 0.07. Similarly, Hanushek, Jackson and Kain (1974, 91) have also shown how improving the specification of macro-relationships can reduce the bias of a regression coefficient estimated with aggregate data. While the accusation can be made that such work suffers from the problems of analysing multicollinear ratio data, this research does demonstrate that the results found by Robinson may be the outcome of a mis-specified model and not aggregation bias.

Langbein and Lichtman (1978), in their recent review of aggregate modelling, have suggested that aggretate variables do not always result in poor estimates and, in some cases, aggregate analysis is to be preferred to analysis at the individual level. It appears that the effects of aggregation depend both on the type of aggregation and the degree to which the model is correctly specified. For example, if both the individual and aggregate models are correctly specified and grouping has occurred on the basis of the explanatory variables, the aggregate model can be expected to be very close to the individual model. As an illustration of this result, Langbein and Lichtman (1978) estimated the effect of students' race and parents' education on students' verbal scores. At the individual level with 3,077 students the estimated model was

$$\begin{matrix}\text{students'}\\ \text{verbal}\\ \text{score}\end{matrix} = .38 \begin{matrix}\text{students'}\\ \text{race}\end{matrix} + 0.9 \begin{matrix}\text{parents'}\\ \text{education}\end{matrix} + \varepsilon \qquad (1)$$

When the data were aggregated into 30 groups according to the values of the explanatory variables, the resultant aggregate model was very similar to the individual model:

$$\begin{matrix}\text{students' verbal score} \\ \text{students' race} \\ \text{parents' education}\end{matrix} \quad = \quad .35 \quad + \quad 0.7 \quad + \quad \varepsilon \qquad (2)$$

Moreover, if both the individual and aggregate models are mis-specified it is possible, though not likely, that the aggregate model is closer to the true, underlying model than the individual model. (This can result when the aggregation has occurred in such a way that the relationship between the omitted and included variables has been decreased.) Most importantly, Langbein and Lichtman demonstrate that specification error is more deleterious than aggregation bias and consequently, it is always better to estimate parameters from a correctly specified aggregate model than an incorrectly specified individual model, whatever procedure has been used to aggregate the data. Finally, the researcher may benefit from aggregation when both the individual and aggregate models are mis-specified but the aggregate model is better specified through the inclusion of more relevant variables.

Langbein and Lichtman (1978, 31) are forthright in stating that

> 'If individual data are less complete or less reliable than a complimentary set of aggregate data, using the aggregate data may be preferable even though individual level behaviour is at issue'.

Clearly, according to this view, aggregate studies should not be summarily dismissed. As Johnston (1978, 263) writes, referring to an excellent aggregate study of voting behaviour (Crewe and Payne, 1976),

> 'Ecological correlations are very valuable in many circumstances, of course, as long as we are careful in their interpretation. They often offer strong clues as to an individual relationship, for which more detailed data are unavailable'.

However, before hastily concluding that it is legitimate under certain circumstances to make inferences from aggregate

data, we must consider in more detail another aspect of the aggregation problem.

Modifiable areas

A well-known example of the problem of modifiable areas is that given by Yule and Kendall (1965). Taking yields per acre of wheat and potatoes in the 48 counties of England and Wales in 1936, they found a correlation of 0.22. However, when they grouped the data in pairs (Figure 8.1) according to their order in the official Agricultural Statistics, the correlation coefficient for the 24 modified observations was 0.30. Continuing the grouping process until 12, 6, and 3 'counties' were formed, they obtained correlation coefficients of 0.58, 0.76 and 0.99.

Unlike the ecological fallacy, where the problem arises from the aggregation of individuals, the modifiable-area problem is concerned with the differing results that can be obtained from different types of aggregation. While the Yule and Kendall example illustrates a 'scale' problem in that different size groups have different correlation coefficients, it is also possible to have a variety of correlation coefficients for the same size groups. For example, if the grouping of the original county data had been performed not on their order in Agricultural Statistics (which roughly approximates to an ordering based on contiguity) but on an alphabetical basis it is likely that the resultant correlation coefficients would be different from that obtained by Yule and Kendall. As an illustration of the problem, consider the relationship between the percentage vote for the Republican candidates in the congressional elections of 1968 and the percentage of population over the age of 60 as recorded in the 1970 United States Census for the 99 counties of Iowa (Openshaw and Taylor, 1979). At the level of these 99 counties the correlation coefficient is 0.3466, but when the data are

FIGURE 8.1 DIFFERENT LEVELS OF AGGREGATION FOR THE YULE AND KENDAL DATA

SOURCE: THOMAS AND ANDERSON (1965)

grouped to form 6 areas the results differ according to the type of aggregation (Table 8.2). The lowest correlation (0.2651) is obtained for the 6 Congressional Districts and the highest correlation coefficient (0.8624) is for areas grouped according to their urban/rural character.

Table 8.2

Some effects on the correlation coefficient of different areal arrangements of the Iowa counties into six zones

Alternative combinations of counties	r
6 Republican-proposed Congressional Districts	0.4823
6 Democrat-proposed Congressional Districts	0.6274
6 Congressional Districts	0.2651
6 Urban/Rural Regional Types	0.8624
6 Functional Regions	0.7128
99 Iowa counties	0.3466

Source: Openshaw and Taylor (1979)

Table 8.3 shows this aggregation problem in all its pervesity. To derive this table, Openshaw and Taylor attempted to find the maximum and minimum correlation that could be obtained for the Iowa data when the counties were aggregated into a fixed number of groups. When the chosen number of groups was 6, it proved possible to obtain the complete range of correlation coefficients. As the degree of aggregation decreased the range of possible correlations became narrower but, at all levels of aggregation, a wide variation was found. As a final illustration of the problem, consider the Cliff and Ord (1969) study which found an R^2 of 0.4051 when a regression model relating milch cows to rainfall was fitted for the 25 counties of Ireland. Openshaw (1977), using an

Table 8.3

Maximum and minimum value of the correlation coefficient for the Iowa data

Number of groups	Grouping based on contiguity	
	minimum	maximum
6	-.999	.999
12	-.984	.999
18	-.936	.996
24	-.811	.979
30	-.770	.968
36	-.745	.949
42	-.613	.891
48	-.548	.886
54	-.405	.823
60	-.379	.777
66	-.180	.709
72	-.059	.703

Source: Openshaw and Taylor (1979)

automatic zoning algorithm, was able to estimate a model with an R^2 of 0.996 for a ten-zone aggregation, but he also obtained a value of 0.0 for another ten-zone aggregation. If the Irish data had been collected and published for ten areas instead of 25, any result between these two extremes could have been found, depending on the precise nature of the grouping chosen.

While it should be emphasized that it is not always possible to aggregate the data in such a way as to get extreme

variations in the results, it is obviously worrying that such variations can occur. As Yule and Kendall (1965, 312) have written :

> 'our correlations will accordingly measure the relationship between variates for all the specified units chosen for the work. They have no absolute validity independent of those units but are relative to them'.

The investigator of the relationship between water hardness and heart disease has no control over the areas for which data are published and he may observe, by chance, a strong positive relationship between the variables. While, if the data had been collected and published for a different set of areas, he may have observed a strong negative relationship. Clearly this is a critical problem for aggregate geographical analysis; indeed, it has been described as 'the most serious problem facing spatial study' (Openshaw, 1979, 11). What can be done to overcome these problems of aggregation? It is possible to suggest three different approaches.

(1) Abandon the aggregate approach and work only at the individual level.

(2) Continue to work at the aggregate level but conduct investigations at a number of different scales.

(3) Continue to work at the aggregate level but adopt an exploratory approach to the difficulties of the aggregation problem.

Each of these proposed approaches is now considered in turn.

Studies of individuals

One approach to the aggregation problem is to admit that little can be achieved by the analysis of aggregate data and that research should be concentrated on individuals. This is now the approach of the team (based at the Royal Free Hospital) that originally analysed the County Borough data, and they are currently undertaking a mammoth study of 7,500 men aged 40-59 in 22 towns in Britain. For each person in the survey, a questionnaire is administered by a

nurse; height, weight, blood pressure and lung function are measured, a blood sample is taken and the respondent undergoes an electrocardiogram. The questionnaire is a lengthy one and consists of some 150 questions dealing with parental history, occupation, current health, current medical treatment, diet, consumption of alchohol, smoking and exercise. The blood sample is analysed for the levels of 16 different elements and 8 haematological indices. The electrocardiogram is viewed on an oscilloscope, transformed to a magnetic tape and is then sent to Glasgow for computer analysis. In addition, the domestic water supply of 10 per cent of the respondents is analysed for 33 different elements by a team of researchers from the Medmenham Water Research Centre. In order to follow the progress of the 7,500 men, their general practitioners have been requested to provide information when the sample individuals attend the doctor for a defined list of cardiovascular diseases. Finally, the Office of Population, Censuses and Surveys will provide the death certificate of each person in the sample when he dies.

Such an approach avoids the aggregation problem, but it suffers from a number of other drawbacks. While there are some relatively minor problems (such as the cost of study and the time, effort and research expertise required to complete the project) the major difficulty facing the project must again be that of statistical analysis. If the researchers aim to achieve more than mere description and cross-tabulation, and if they wish to perform some sort of multivariate analysis to establish whether any observed relationship is spurious, the problems will be severe. As with most survey data, the dependent variable is likely to be either dichotomous (died from heart disease, did not die from heart disease) or polychotomous (suffers from no/slight/moderate/severe heart disease). When a polychotomous dependent variable is estimated by ordinary least squares it is usual to find nonsensical values of the fitted dependent

variable and heteroscedastic residuals. Consequently, the OLS estimates are no longer the minimum variance estimates and, in any particular application, the estimated coefficients may be very different from the true coefficients (Chapter 4). Moreover, the commonly used methods of evaluating a model (the t and F tests and the coefficient of determination, R^2) are rendered inappropriate when the dependent variable has a dichotomous or polychotomous form.

The analysis of such so-called 'qualitative' or categorical data is currently a major area of statistical research and, in geography, Wrigley (1975) has pioneered the statistical solutions to the problems. However, while the problems of nonsensical values and heteroscedasticity may be overcome by the use of maximum-likelihood log-linear models, these improved procedures suffer from a number of other drawbacks. As Wrigley (1977, 44-45) points out, such estimation procedures are sensitive to multicollinearity and no methods have, as yet, been developed to overcome the problem. Similarly, there are problems with the analysis of residuals from a maximum likelihood model. For example, with a dichotomous dependent variable there will be two residuals for each observation. In locations where the probability of having heart disease is high the residuals will be either small and positive or large and negative. In locations where the probability of having heart disease is low the residuals will either be small and negative or large and positive. These highly non-normal residuals from an intrinsically non-linear model would certainly prohibit the usual Cliff-Ord test for spatial autocorrelation and, this is yet another field in which, as Wrigley (1976, 30) writes: 'much work remains to be done'.

Another approach to qualitative data analysis, that of Goodman, suffers similar problems to the procedure advocated by Wrigley. In Goodman's approach (see Upton, 1978) both the dependent and independent variables are categorical and

the analysis proceeds by fitting a series of models to a set of cross-tabulated data. The approach is unable to deal with multicollinear data (which reveals itself by near empty cells in the cross tabulation) and it is impossible to assess and detect the effects of spatial autocorrelation, primarily because residuals refer to particular cells in the cross tabulation and not to geographical areas. Moreover, fitting the model is accomplished by a procedure based on significance tests which is highly reminiscent of the stepwise procedure which was severely criticised in Chapter 1.

The analysis of categorical data is currently receiving research attention after a considerable period of neglect. While it is to be hoped that improved techniques will be developed, the current user of these methods with individual survey data faces severe problems. Consequently, while the analyst of individual data avoids aggregation problems he is confronted with other difficulties which are not encountered in terms of aggregate data.

Finally, in relation to the mammoth study of the Royal Free Hospital research team, it must be stressed that, ironically, the project is a testament to the value of aggregate analysis. If the original strong negative relationships had not been found for the County Boroughs, it is unlikely that the individual study would have been undertaken. Moreover, it is interesting to speculate whether this individual study would have been conducted if the original study had found the observed weaker relationships that have been reported in Chapter 7. Consequently, it is argued that further investigation is still required at the aggregate level before research effort is concentrated at the more expensive and time-consuming individual level.

Changing scale

A second possible approach to the problems of aggregation is to conduct studies at several different scales. If similar

results are found at a variety of scales, greater credence can be given to observed relationships than if the associations are only found at one scale of analysis. Moreover, given the previous discussion of aggregation and specification error, it is obviously desirable not only to change scales but also to formulate a well-specified model. To achieve these requirements the author is currently assembling a comprehensive data set for the 168 local authorities of Wales. As constituted in 1971 these local authorities are markedly different from the County Boroughs and they represent a wider range of environmental conditions. Data have been collected on cause-and age-specific deaths and, for some of the local authorities, improved measure measurements on air pollution (based on moss-bag observations) and domestic water (based on a detailed analysis of a sample of households) are available. It is intended to apply the same procedures as used in Chapter 7 to this data but, while the similarities and contrasts between the two levels of analysis may prove to be interesting, it must be admitted that this approach only tackles the scale problem and avoids other difficulties inherent in aggregation. At both the County Borough and Welsh Local Authority levels, different results might be found if the data were to relate to a different set of areal units to that adopted by the Registrar General. Consequently, at both scales, the aggregation problem needs to be approached in a more direct fashion.

An exploratory approach

It is unlikely that the aggregation problem will be solved by a neat statistical solution. The number of possible aggregations is too large and it is extremely difficult to define criteria by which the appropriateness of a particular aggregation can be statistically assessed. If one views the problem from the rigorous standpoint of confirmatory statistics, it appears that no solution is in sight. However, a trial-and-error, exploratory procedure has been suggested

by Openshaw (1977, 1979). This approach promises a way of examining the aggregation problem in relation to a particular data set and it also provides a means of assessing the severity of the problem.

The usual approach in regression analysis is to regard the zones or areas as fixed and the principle problem, therefore, is to estimate the regression coefficients relating the dependent to the independent variables. However, Openshaw's Cappadocian approach is to set the regression coefficients of the model to a pre-determined value and manipulate the zoning system so as to achieve the required result. Adopting this approach, whereby models are fitted by selecting an appropriate zoning system, the visual representation of the zones becomes an important exploratory tool. As an illustration of this approach, a geographer may state the hypothesis that water hardness is strongly related to heart disease mortality in a negative manner. The zoning system is then manipulated so as to achieve such a relationship. If a strong negative result cannot be found, whatever the degree and type of aggregation, the hypothesis is rejected. If such a result is obtainable, the analyst can consider under what circumstances strong negative relationships are found. He may consider whether such values are only obtained when the model is mis-specified and he may examine the residual spatial autocorrelation of the fitted model. Clearly, if no problems can be detected with the chosen model, greater faith can be placed in the model. Moreover, because the aggregate zones can be mapped, the analyst can 'see' what types of zones lead to particular results. For example, are strong negative relationships found only when the bulk of the zones are aggregated to form a 'super-zone'? Are strong negative relationships produced by a particular aggregation related to the independent variables or do the zones suggest a variable that may have been excluded from the model? The mapping of the zones

therefore provides a visual representation of the interactions between the model, base data, scale and aggregation problems. Having explored the possibilities for one fixed model, the analyst can proceed to specify other models and could, for example, attempt to find the zonal arrangement producing the strongest positive result. This approach (like much of exploratory analysis) is unlikely to produce a final, definite conclusion. It does, however, demonstrate the aggregation problem to an analyst and it may possibly provide insights into why and how varying results have been obtained.

Openshaw (1977) has considered how a computer program may be developed for his approach but, unfortunately, there is as yet no published full-scale illustration of the method. It is intended in the future to use this exploratory zoning approach in the analysis of both the County Borough data and in the analysis of the mortality data for the Welsh local authorities. In particular, it is hoped to link this approach with the analysis of specification errors (as presented in Chapter 4) in order to elucidate the interrelationships between model mis-specification and the aggregation problem.

Further research

The three approaches that have been outlined to overcome the problem of aggregation assume that the problem is genuine and pervasive. But, in fact, very little is known about how and in what way aggregation leads to incorrect inferences. Moreover, given the possible severe and widespread nature of the difficulty, it is essential that the aggregation problem be thoroughly researched. A critical reading of the literature on the topic suggests that previous research has erred in concentrating on correlation coefficients and realistic data.

Openshaw and Taylor (1979), Openshaw (1977), Robinson (1950) and Yule and Kendall (1965) have all used correlation

coefficients to illustrate aggregation problems, but it is known that such measures are affected by aggregation when the unstardised regression coefficients are not. For example, Blalock (1964, 103) examined the relationship between the difference in income for whites and blacks (the dependent variable) and the percentage of the total population that was black (the independent variable). For 150 counties in the south of the United States of America, he found a correlation coefficient of 0.54 and a regression coefficient of 0.26. A series of experiments was then performed on the data: the counties were aggregated by three different methods (randomly, by maximising the variance of the independent variable, and by geographical proximity) and for each method they were aggregated so that there were 75, 30, 15 and 10 modified areas. As Table 8.4 clearly shows, the correlation coefficient varies markedly from 0.26 to 0.95, while the regression coefficient only varies from 0.18 to 0.36.

Table 8.4

Aggregation experiments on correlation and regression coefficients

	Correlation Coefficient					Regression Coefficient				
	Level of aggregation									
Method of grouping	150	75	30	15	10	150	75	30	15	10
Random	.54	.67	.61	.62	.26	.26	.36	.31	.27	.18
Maximising variance	.54	.67	.84	.88	.95	.26	.26	.26	.26	.26
By proximity	.54	.63	.70	.84	.81	.26	.27	.28	.28	.34

Source: Blalock (1964, 103)

Moreover, the range of regression coefficients is even narrower (0.26 to 0.34) when the data are aggregated by geographical

proximity. Aggregation may have little effect on unstandardised regression coefficients and research is required into the circumstances in which it is possible to achieve such results. In addition, given that spatial autocorrelation can have deleterious effects on correlation coefficients (Chapter 3), further research into the aggregation problem should be conducted with unstandardised regression coefficients.

Those studies that have, in fact, used regression coefficients have unfortunately not isolated the aggregation problem by examining simulated data with known properties. For example, Openshaw (1979), in his illustration of the aggregation problem, re-analysed the data for Ireland presented by Cliff and Ord (1969). The major problem in analysing real-world data is that the model may be mis-specified. For example, if an important variable has been excluded from a model, any type of aggregation that maximises the influence of the omitted variable may lead to very inaccurate estimates. But, if the model had been correctly specified, the effect of aggregation may be considerably reduced. Certainly, recent geographical research which examines the aggregation problem has ignored the problems of specification errors and this casts considerable doubt on the value of such work.

In summary, there is a real need for experiments to be conducted with unstandardised regression coefficients on simple, generated data with known properties. Until such research is undertaken the true nature of the problem will remain obscure and the worth of aggregate studies (such as Chapter 7) in making valuable, correct causal inferences will remain uncertain. However, it is hoped that the preceding discussion will have emphasised four points. Firstly, Robinson's (1950) paper, which is generally used to denigrate aggregate studies, contains a serious flaw. As Hanushek and his co-authors (1974, 90) write of Robinson's work:

'had he considered a more complete and accurately specified model, his empirical findings would have been much different and his conclusions, relating to the appropriate use of aggregate data, would have been much more limited and much less severe'.

Secondly, many other researchers have confused the effects of specification error and aggregation bias. Consequently, and again quoting Hanushek and his co-authors (1974, 100):

'It is simply not true, however, that any simple correlation using microdata is superior to the coefficient estimates from a similar, but well-specified, multivariate aggregate model. Multivariate models usually are more interesting in terms of behaviour content, and they often have better (less biased) coefficient estimates'.

Thirdly, the exploratory approach of Openshaw (1977, 1979) deserves to be used in a full-scale analysis and, in particular, the procedures must be used in conjunction with specification error analysis. Fourthly, and finally, it is essential that further research be conducted into the aggregation problem because, as Openshaw and Taylor (1979, 18) write after their experiments:

'our general feeling is that the modifiable areal unit problem is much more complex than has previously been believed'.

CHAPTER CONCLUSIONS

The exploratory approach to data analysis offers many opportunities. Such an approach, which is sceptical about what can be assumed about data and is honest in reporting the steps taken to arrive at a final analysis, deserves to be widely practised. This thesis has contributed to exposing the inherent contradictions of confirmatory analysis, to demonstrating exploratory methods of overcoming the multicollinearity problem, to suggesting ways of choosing an appropriate functional form and to indicating the pitfalls of analysing ratios. It is contended that exploratory

procedures need continued publicity and informed evaluation because they may represent a considerable improvement over the techniques that are commonly used by geographers.

Finally, it is appropriate to return to the issue of water-hardness and disease. While exploratory procedures in Chapter 7 indicated that there is only a weak negative association between mortality and hardness of water, it would be unrealistic and misleading to deny that much remains to be done. In particular, the difficulties presented by the ecological fallacy and the modifiable-area problem pose a crucial question: are the results of Chapter 7 of substantive value or are they merely the outcome of analysing data for a particular set of areas? At present it is impossible to answer this question and, for any progress to be achieved, there is an urgent need for considerable applied and theoretical research into the problems of aggregation. Re-iterating the words of Hirschi and Selvin (1966, 254) neatly summarises the way ahead:

> 'since no one proposes trying to give people cancer ... the fruitful way toward better causal analyses ... is to concentrate on improving the statistical approach'.

CHAPTER 8 : NOTES

1. An approach that has many similarities with the exploratory perspective was introduced into geography by McCarty (1956). However, his ideas have not continued to influence quantitative research. Indeed, as displayed by the difference between the first (1965) and second editions (1977) of Haggett's Locational Analysis in Human Geography there is now little discussion of McCarty's work.

2. In general, they discuss the differences between exploratory and confirmatory analysis and, more particularly, they direct attention to those attitudes and procedures (graphical displays, residual analysis, robustness and transformations) which appear most useful in exploratory work.

3. The usefulness of the ridge technique also needs to be explored in relation to trend surface analysis for it can be expected that complicated surfaces will be estimated with multicollinear data.

4. The inbuilt correlations were calculated by using the formula of Mosimann (1962).

5. The chi-square measure has been widely used in collaborative work. Barnard (1978) found that the use of this statistic resulted in differing interpretations of the spatial pattern of the elderly in South Hampshire from that given by the analysis of ratios. Similarly, the chi-square statistic resulted in a different interpretation of maps of foundation collapse (Clark, 1980), and maps of census data for Reading (Jones and Kirby, 1980). A re-analysis of the County Borough data did not, however, lead to a new interpret-

ation (the absolute number of deaths is relatively large) but an analysis of Welsh local authority data showed some striking differences to that of Howe (1963).

CHAPTER 8 : BIBLIOGRAPHY

ALKER, H.J. (1969): A typology of ecological fallacies in Dogan, M. and Rokkan, S. (eds.) Quantitative ecological analysis in the social sciences MIT, Cambridge, Mass.

ALLARDT, E. (1969): Aggregate analysis: the problem of its informative value in Dogan, M. and Rokkan, S. (eds.) Quantitative ecological analysis in the social sciences MIT, Cambridge, Mass.

BALDWIN, J., BOTTOMS, A.E. and WALKER, M.A. (1976): The urban criminal Tavistock, London.

BARNARD, K.C. (1978): The residential geography of the elderly: a multiple-scale approach unpublished Ph.D. thesis, University of Southampton.

BLALOCK, H.M. (1964): Causal inferences in non-experimental research University of North Carolina, Chapel Hill.

CLARK, M.J. (1980): Property damage by foundation failure in Doornkamp, J.C. and Gregory, K.J. (eds.) Atlas of drought in Britain 1975-76 Institute of British Geographers, London.

CLARK, W.A.V. and AVERY, K.L. (1976): The effects of data aggregation in statistical analysis Geographical Analysis 8,428-438.

CLIFF, A.D. and ORD, J.K. (1969): The problem of spatial auto-correlation in Scott, A.J. (ed.) Studies in regional science Pion, London.

CLIFF, A.D. and ORD, J.K. (1973): Spatial autocorrelation Pion, London.

COX, N.J. and JONES, K. (forthcoming) Exploratory data analysis in Wrigley, N. and Bennett, R.J. (eds.) Quantitative geography in Britain Routledge and Kegan Paul, London.

CRAMER, J.S. (1964): Efficient grouping: regression and correlation in Engle curve analysis
Journal of the American Statistical Association 59, 233-250.

CREWE, I. and PAYNE, C. (1976): Another game with nature: an ecological regression model of the British two-party vote ratio in 1970.
British Journal of Poplitical Science 6, 43-81.

ERICKSON, B.H. and NOSANCHUK, T.A. (1977): Understanding data
McGraw-Hill, Toronto.

EVANS, I.S. and JONES, K. (forthcoming): Ratios in Wrigley, N. and Bennett, R.J. (eds.)
Quantitative geography in Britain
Routledge and Kegan Paul, London.

HANNAN, M. and BURSTEIN, L. (1974): Estimation from grouped observations
American Sociological Review 39, 374-392.

HANNAN, M.T., FREEMAN, J.H. and MEYER, J.W. (1976): Specification of models for organizational effectiveness
American Sociological Review 41, 136-143.

HANUSHEK, E.A. and JACKSON, J.E. (1977): Statistical methods for social scientists
Academic Press, New York.

HANUSHEK, E.A., JACKSON, J.E. and KAIN, J.F. (1974): Model specification, use of aggregate data, and the ecological fallacy
Political Methodology 1, 87-106.

HAYNES, R.M. (1978): A note on dimensions and relationships in human geography
Geographical Analysis 10, 288-291.

HIRSCHI, T. and SELVIN, H.C. (1966): False criteria of causality in delinquency research
Social Problems 13, 254-268.

HOWE, G.M. (1963): National atlas of disease mortality in the United Kingdom
Nelson, London.

JOHNSTON, R.J. (1978): Multivariate statistical analysis in geography
Longman, London.

JONES, K. (1978): Percentages, ratios and inbuilt relationships in geographical research: an overview and bibliography
Department of Geography, University of Southampton, Discussion Paper 2.

JONES, K. (1979): Taking the rough with the smooth: robust alternatives to trend surface analysis
Seminar paper presented to the Department of Geography, University of Newcastle-upon-Tyne.

JONES, K. and KIRBY, A.M. (1980): Mapping census data for enumeration districts.
Working Paper 2, Accessibility to urban resources project, Reading.

KING, L.J. (1969): Statistical analysis in geography
Prentice-Hall, New Jersey.

LANGBEIN, L.I. (1977): Schools of students: aggregation problems in the study of student achievement
Evaluation Studies Review Annual 2, 270-298.

LANGBEIN, L.I. and LICHTMAN, A.J. (1978): Ecological Inference
Quantitate Applications in the Social Sciences Number 10
Sage, Beverly Hills.

LICHTMAN, A.J. (1974): Correlation, regression and the ecological fallacy: a critique
Journal of Interdisciplinary History 4, 417-433.

MECKSTROTH, T.W. (1974): Some problems in cross-level inference
American Journal of Political Science 18, 45-66.

McCARTY, H.H. (1956): Use of certain statistical procedures in geographical analysis
Annals of the Association of American Geographers 46, 263.

MOSIMANN, J. (1962): On the compound multinomial distribution, the multivariate beta distribution and correlation among proportions Biometrika 49, 65-82.

MOSTELLER, F. and TUKEY, J.W. (1977): Data analysis and regression Addison-Wesley, Mass.

OLSSON, G. (1975): Birds in egg Michigan Geographical Publication Number 15 Ann Arbor, Michigan.

OPENSHAW, S. (1977): Geographical solution to scale and aggregation problems in region-building, partitioning, and spatial modelling Transactions of the Institute of British Geography, New Series 2, 459-472.

OPENSHAW, S. (1978): An empirical study of some zone design criteria Environment and Planning Series A 10, 781-794.

OPENSHAW, S. (1979): Appropriate methods for testing hypotheses derived from studies of spatially aggregated data, paper presented to session on Aspects of Exploratory Data Analysis, IBG, Manchester.

OPENSHAW, S. and TAYLOR, P.J. (1979): A million or so correlation coefficients: three experiments on the modifiable areal unit problem in Wrigley, N. (ed.) Statistical methods in the spatial sciences Pion, London.

ROBINSON, W.S. (1950): Ecological correlation and the behaviour of individuals American Sociological Review 15, 351-357.

SHIVELY, W.P. (1969): Ecological inference: the use of aggregate data to study individuals American Political Science Review 63, 1183-1196.

SHIVELY, W.P. (1974): Utilizing external evidence in cross-level inference
Political Methodology 1, 61-74.

SMITH, K.W. (1977): Another look at the clustering perspective on aggregation problems
Social Methods and Research 5, 289-316.

THOMAS, E.N. and ANDERSON, D.L. (1965): Additional comments on weighting values in correlation analysis of areal data
Annals of the Association of American Geographers 55, 492-505.

TUKEY, J.W. (1977): Exploratory data analysis
Addison-Wesley, Mass.

UPTON, G.J.G. (1978): The analysis of cross-tabulated data
Wiley, Chichester.

WELSH OFFICE (1975): Report of a collaborative study of certain elements in air, soil, plants, animals and humans in the Swansea/Neath/Port Talbot area, moss bag study of atmospheric pollution across South Wales.
HMSO, Cardiff.

WRIGLEY, N. (1975): Analysing multiple alternative dependent variables
Geographical Analysis 7, 187-195

WRIGLEY, N. (1976): An introduction to the use of logit models in geography
Concepts and Techniques in Modern Geography 10, Geo Abstracts,
Norwich.

WRIGLEY, N. (1977): Probability surface mapping
Concepts and Techniques in Modern Geography 16, Geo Abstracts.
Norwich.

YULE, G.U. and KENDALL, M.G. (1965): An introduction to the theory of statistics
Fourth impression of fourteenth edition, Griffin, London.