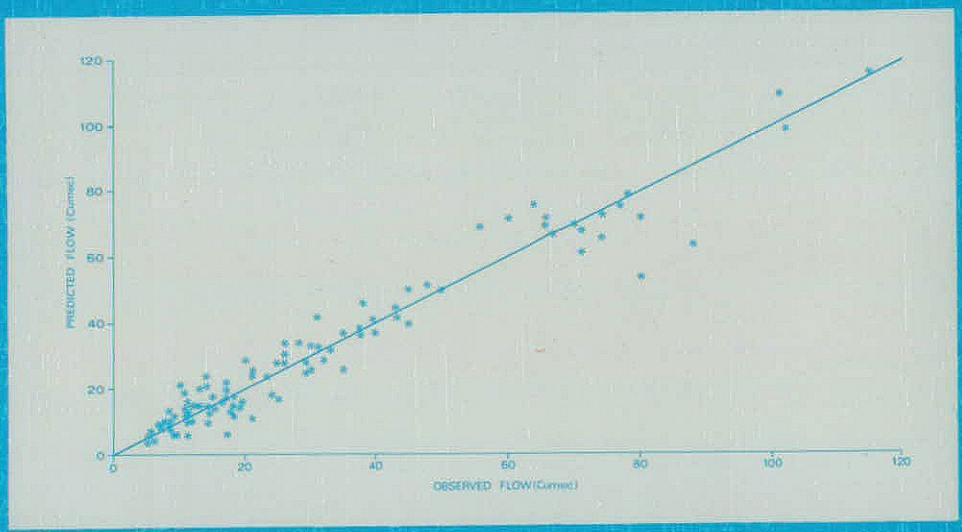
MULTIPLE REGRESSION IN HYDROLOGY

R. L. Holder



MULTIPLE REGRESSION IN HYDROLOGY

MULTIPLE REGRESSION IN HYDROLOGY

by R. L. Holder

INSTITUTE OF HYDROLOGY

WALLINGFORD

ISBN 0 948540 00 1

The Institute of Hydrology is a component establishment of the
Natural Environment Research Council

Printed in Great Britain by Galliard (Printers) Ltd, Great Yarmouth

PREFACE

There are many problems in hydrology that may be solved by multiple regression procedures. This type of analysis may be used in flood and low flow studies, for example, and in catchment modelling. Rainfall-runoff equations derived by using multiple regression procedures have been developed and used for a variety of purposes such as flow projection in times of drought and for the estimation of past flows from weather data.

As no single work of reference dealt comprehensively with the use of multiple regression in hydrology, the Department of the Environment's Central Water Planning Unit, in June 1975, commissioned one from Mr R. L. Holder of the Department of Mathematical Statistics, Birmingham University. Following the transfer of certain functions and responsibilities of the DoE Unit to the Natural Environment Research Council's Institute of Hydrology, a revised and updated book incorporating more substantial examples was prepared with the assistance of the staff of the Institute.

Because of the apparent generality of the method of relating one variable to a set of other variables, multiple regression is probably the most frequently used—and indeed misused—statistical tool. Undoubtedly, the technique is potentially very useful and it is currently the subject of much theoretical research by mathematical statisticians; but, as with any statistical procedure, it is crucially important to understand the basis, assumptions and limitations of the technique. Computer packages have taken the drudgery out of regression analysis and some allow great flexibility in the type of analysis conducted. As well as instructing the reader on the basic techniques, this book aims to educate readers to extend their use of regression beyond the standard procedures.

I am very pleased therefore that it has been possible for this Institute to publish Mr Holder's most useful addition to the hydrological literature.

J. S. McCulloch
Director, Institute of Hydrology
April, 1985

ACKNOWLEDGEMENTS

My sincere thanks to Colin Wright of the Department of the Environment, for the original idea and initial support for this book, to David Jones of the Institute of Hydrology for providing the problems and data sets for Chapter 5 as well as many helpful suggestions, to Mrs T. Carr for typing the manuscript, Mrs A. Mayho for assisting with the computing and Miss P. Binn for suggesting many improvements to the original manuscript.

R.L.H.

Birmingham,
October, 1984

CONTENTS

Chapter 1

SIMPLE LINEAR REGRESSION

1.1 Introduction

1.1.1 A problem in linear regression analysis

A study of the relationship between rainfall and run-off in a particular area may, amongst other things, have led the investigator to keep records of the annual rainfall and the annual run-off over a period of several years. An example of such records, taken from the Alwen catchment, Lewis (1957), is given in Table 1.

Table 1 Monthly rainfall and run-off for the Alwen Catchment, North Wales 1912–1915 (mm)

Year		*Jan.*	*Feb.*	*Mar.*	*Apr.*	*May*	*June*	*July*	*Aug.*	*Sep.*	*Oct.*	*Nov.*	*Dec.*
1912	rainfall	76	87	164	23	41	119	110	189	40	139	134	192
	run-off	75	66	112	25	5	18	22	126	22	64	102	161
1913	rainfall	115	74	196	140	80	96	45	67	92	126	140	119
	run-off	115	81	142	104	60	33	7	7	31	72	131	94
1914	rainfall	162	156	168	53	100	96	127	83	44	68	146	260
	run-off	125	148	132	43	44	36	31	37	24	29	114	246
1915	rainfall	185	178	42	78	67	16	138	107	44	64	112	245
	run-off	166	161	43	34	22	3	35	39	8	22	55	240

This table gives the precise details, within recording accuracy, of rainfall and run-off in the Alwen catchment between 1912 and 1915 and, as such, is the most complete statistical representation of the investigator's findings. However, some alternative statistical representation of these facts may be necessary in order to achieve some specific objective. The investigator may wish to:

(a) Summarise his data in terms of just a few pertinent numbers.
(b) Decide whether rainfall and run-off influence each other.
(c) Predict some future run-off which might be expected from a certain annual rainfall.

(d) Predict the rainfall that would be necessary to produce a certain run-off.
(e) Decide whether certain of the readings in the table are exceptional or are not of the same pattern or trend as the others.
(f) Build or complete some mathematical model relating rainfall and run-off.
(g) Make some comparison between the readings given in Table 1 and similar readings obtained from another area.

To achieve any of these objectives, the first step could be to draw a graph of annual or monthly rainfall against run-off.

Careful examination of this graph and judicious use of a ruler, a flexible curve and his own experience would help to give the investigator an answer to objectives (a), (b), (c), (d), (e) and (g). Linear regression analysis would also be helpful. However, in suggesting this further technique, it is not our intention to denigrate graphical and visual methods; indeed, it is hoped that the reader will realise that the two are complementary. Allowance for other factors, such as evaporation or month to month variation, would improve the accuracy of the relationship shown in Figure 1. More complex relationships of this type are discussed in Chapter 2.

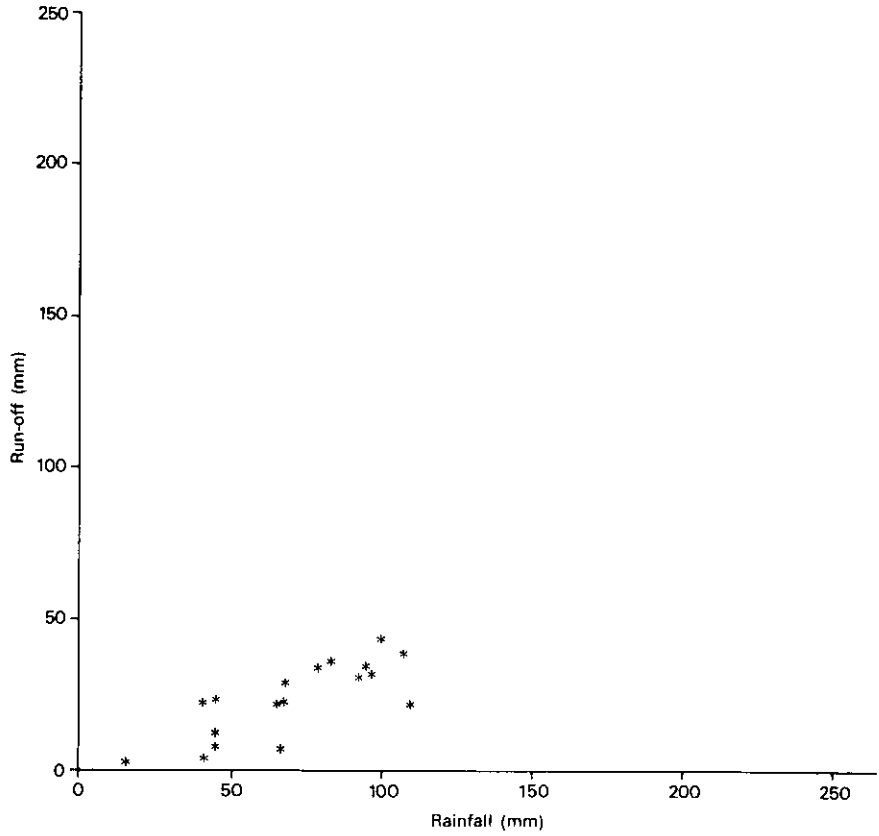


Fig. 1. Alwen catchment rainfall and run-off 1912–15.

1.1.2 Assumptions made in linear regression

Simple linear regression may be applied to problems in which a record has been made of the values of two variables, referred to as y, the dependent variable, and x, the independent variable. It is assumed that, for any such record, the mathematical model

$$y = a + bx + e \tag{1}$$

where a and b are constants and e is a variable, describes the relationship between the y reading and the x reading. By temporarily ignoring the term e, we see that a straight line (linear) relationship is assumed between y and x, with a, the intercept, and b, the slope of the graph of y plotted against x. However, if our model only allowed for readings of y and x which fell exactly on a straight line, it would be of little practical value. Inclusion of the variable e allows readings of y and x to deviate from a straight line, but assumptions are made about e so as to force these deviations to have a particular pattern. If we imagine being able to take many readings all giving the same value of x, then some y values will be greater then $a + bx$ and some less, i.e. some values of e will be positive and some negative. Most of the assumptions made in linear regression can be stated in terms of the values of e. We will assume that the arithmetic mean of the values of e is zero. We will also assume that the variance of these values of e is always the same wherever the value of x happens to fall. At a later stage, we will also need to assume that these values of e form a normal distribution.

Figure 2 indicates the type of graph one would expect if it were possible to record many values of y all with the same x value.

In practice, we will frequently have only one value of y to plot at one x position. Consequently, one of the problems we will have to consider is how to justify the above assumptions when the readings are not available in the ideal form as shown in Figure 2.

An alternative interpretation of the assumptions is as follows. If we are able to fix a value of x, then the value of y we record should be $a + bx$. However, due

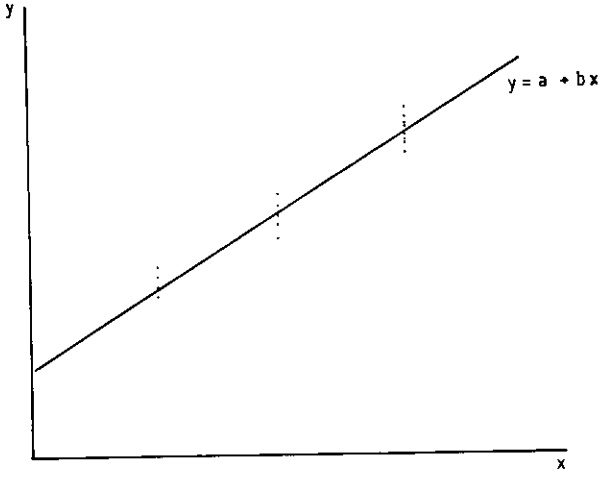


Fig. 2. Distribution of values about the regression line.

to errors, inaccuracies, uncontrollable or inexplicable variations, the reading of y that we make is $a + bx + e$, e representing the error in measurement. Then, on average, such errors should be zero, i.e. there is no consistent bias in our readings as a result of errors made. Furthermore, all readings should be made with equal precision, i.e. a given size error is equally likely to be made at any value of x. Finally, the errors of measurement should form a normal distribution.

There is one further assumption to add to both of these explanations and this is that all errors (values of e) are assumed to be independent, i.e. the magnitude of the error in one reading does not influence the magnitude of the error in another reading.

1.1.3 Interpretation of the assumptions

Let us consider the direct application of linear regression to the data of Table 1. As may be deduced from their titles, and certainly from equation (1), the variables y and x are treated differently in linear regression; they are not interchangeable. Thus, some thought must be given to which variable we call y and which we call x. Ideally, we would have one variable subject to errors and the other fixed, controlled or error free; the former would be y and the latter x.

However, with rainfall and run-off as potential y and x variables, the choice is by no means straightforward. Indeed, considering the measurement of these two variables, we would probably have to conclude that both were subject to errors; consequently, the linear regression model (1), which appears to attribute all error to one variable, is not appropriate. Models which allow both variables to be subject to error will be discussed later but for now, let us consider what circumstances might lead us to use linear regression for rainfall/run-off problems.

As is so frequently the case, it is our objective, together with some knowledge of the physical process being studied, which determines the form of the model. If we wish to predict the likely annual run-off from an annual rainfall of R, then we will need to assume that R is fixed and predict what we regard as an uncertain quantity, run-off. Thus, there is some intuitive support for assuming that the available rainfall readings are fixed and, together with some statistical reasoning, this leads us to conclude that rainfall should be treated as the independent variable x. In general, we should usually aim at taking the predicted variable to be y and the predictor to be x. In this particular example, there is a further reason for taking rainfall as the x variable in that rainfall is, to some extent, causal of run-off and hence our model may be interpreted as being of the form

$$\text{output} = \text{some function of input} + \text{error}$$

Having decided upon an x and y, we next have to consider our assumptions about the errors or inexplicable variations. Imagine drawing the best possible line to describe the points in Figure 1, as illustrated in Figure 3; the vertical displacement of each point from this straight line represents the error or inexplicable variation for that reading.

If the first assumption of errors averaging to zero is true, then this will

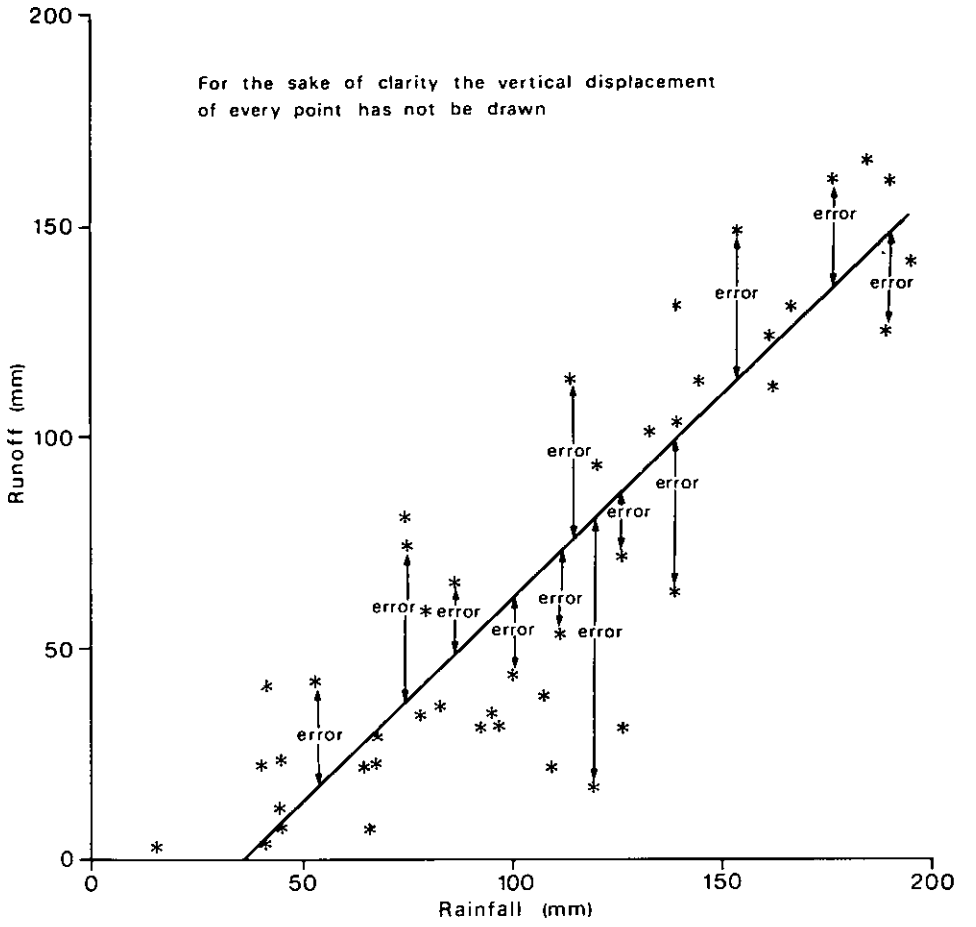


Fig. 3. Error or inexplicable variation in y readings.

usually lead to a collection of displacements which appear to have no pattern to them and few particularly outstanding values.

Similarly, reference to Figure 3 will help in assessing the second assumption, that of equal precision or error variance. This would be reflected in the graph by a similar spread of deviations about the line over the whole range of x values. If, on the other hand, points tend to group close to the line in some regions of x and are widely dispersed to either side of the line in other regions of x, then this might suggest that the precision of results varies.

The assumption that the errors form a normal distribution is not essential for all the steps in a linear regression analysis. For instance, a best fitting straight line may be obtained, and some approximate statement made about the accuracy of that line, without this assumption. However, if such an assumption can be made or arranged (see Section 3.3) a far more complete and satisfactory analysis can be accomplished. As in the case of the two previous assumptions, it is usually necessary to refer to the outcome of a linear regression analysis in order to assess the validity of this assumption (see Section

4.3). However, some knowledge of the distribution of hydrological data will be of value in detecting likely problem cases (see Section 3.2).

The assumption of independence is usually violated when there is some carry over from one reading to the next, frequently when such readings come as a sequence in time. An example might be where two run-off measurements are made over time periods which overlap or which are both affected by the same heavy rainfall or drought. Another example is where one reading contributes to the next in some way, as might happen with river flow measurements taken at stations which are close together. Problems which are more appropriately modelled as time series are considered later. However, as with the normality assumption, the choice of a best fitting straight line is not necessarily dependent on this assumption of independence being satisfied.

1.1.4 What can be achieved by using linear regression analysis?

So far, it has been suggested that linear regression analysis might help in solving problems (a) to (g) of subsection 1.1.1 and that some 'best fitting' straight line might also appear. Before plunging into a detailed description of how we might give an answer to these objectives, it would be as well to examine more specifically what it is possible to achieve using linear regression analysis.

First of all, let us suppose that the assumptions mentioned earlier are satisfied, that we have chosen a y and an x, and that we have a set of data similar to that of Table 1, namely pairs of values of y and x. We may estimate a and b in equation (1), together with their standard errors, or, alternatively, we may derive confidence intervals for a and b or for the line $a + bx$. This will give an answer to problem (a), some idea of (b), and possibly an appropriate answer to (f).

Having estimated a and b, we may use these estimates to predict a value of y corresponding to a particular x (and vice versa) by calculating

$$\hat{y} = \hat{a} + \hat{b}x \tag{2}$$

where $\hat{y}$, $\hat{a}$ and $\hat{b}$ are estimates of y, a and b respectively. Alternatively, we may derive a confidence interval for this unknown value of y. This will help to answer (c) or (d).

We may carry out tests of significance on b and/or on a in order to examine simplifications of equation (1). For example, we could test $a = 0$, or even $b = 0$. If we have several sets of similar data, then we may estimate a and b for each set and carry out tests on the similarity of the different as and bs. The first tests might help with problem (b) or (f) and the others with problem (g).

If, for each value of x recorded in our data, we estimate the corresponding value of y using equation (2) with our estimates of a and b, then the difference between the recorded and the estimated value of y is usually referred to as the residual. The set of residuals calculated from all the data contains useful information. Patterns in these residuals, when plotted against their associated x values, may indicate a poor model and may suggest the direction in which improvements might be made. The residuals may also be used to examine the validity of the assumptions mentioned in subsection 1.1.2. A residual which

was markedly different from the others would be an indication to the solution of problem (e).

1.2 The Basic Method

1.2.1 Fitting a straight line

The basic data for this section will consist of pairs of values of y and x where the identity of y and x has been established as already outlined. These pairs of values will be denoted by $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$ and model (1) will then become

$$y_i = a + bx_i + e_i \qquad (\text{for } i = 1, 2, 3, \ldots, n) \tag{3}$$

This is illustrated in Figure 4 for three pairs of points.

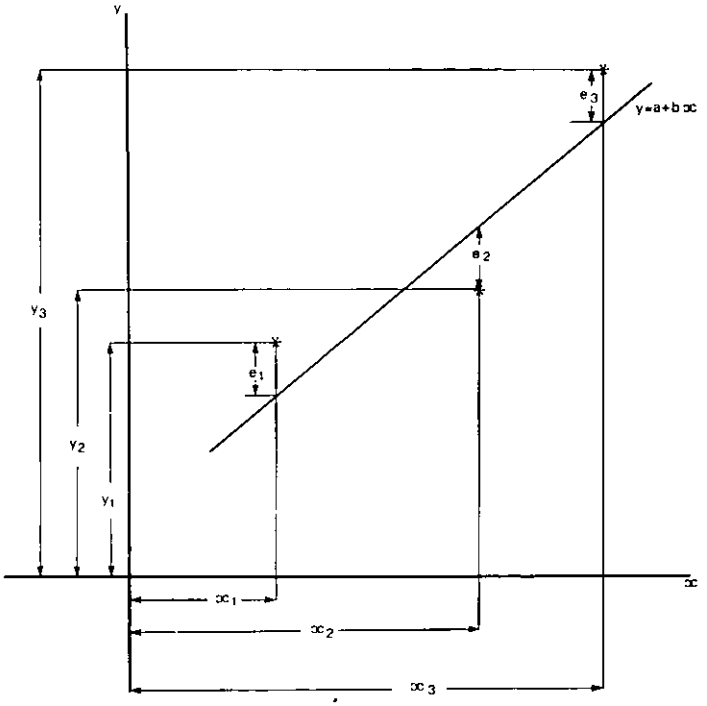


Fig. 4. Regression line and associated data.

The exact position of the line $y = a + bx$ is unknown and our problem is to make an intelligent guess at its position, given the points on the graph. There are many proposals as to how this intelligent guess should be made. We will examine one method in detail, namely least squares estimation, but we will also consider some alternatives.

The objective of least squares estimation is to choose values of the unknowns so as to minimise

$$S^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

i.e. the total vertical discrepancy of the points from the line (regardless of sign) should be as small as possible.

Solving $\partial S^2/\partial a = 0$ and $\partial S^2/\partial b = 0$ will give the values of a and b, denoted by $\hat{a}$ and $\hat{b}$, which minimise S^2. Hence, solving the equations

$$-2\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i) = 0 \qquad \text{and} \qquad -2\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)x_i = 0 \tag{4}$$

will give the estimates

$$\hat{b} = \frac{\sum_{i=1}^{n} y_1 x_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\Big/ n}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5}$$

and

$$\hat{a} = \left(\sum_{i=1}^{n} y_i\right)\Big/ n - \hat{b}\left(\sum_{i=1}^{n} x_i\right)\Big/ n = \bar{y} - \hat{b}\bar{x} \tag{6}$$

where

$$\bar{y} = \left(\sum_{i=1}^{n} y_i\right)\Big/ n \qquad \text{and} \qquad \bar{x} = \left(\sum_{i=1}^{n} x_i\right)\Big/ n$$

The first expression in equation (5) is the one usually recommended for calculation because it preserves accuracy. However, this point is only valid provided full accuracy can be retained throughout the calculation. If there are a large number of data points and y and x are relatively large values, then this may lead to $\sum_{i=1}^{n} y_i x_i$ and $(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)/n$ both being large and similar; hence, their difference may be seriously affected by the roundoff errors generated when calculating either of the large expressions. In such circumstances, the second expression in equation (5) is more satisfactory. Roundoff problems usually arise where a digital computer has been used for calculation and, under these circumstances, there is little extra hardship involved in using the alternative expression.

As sum of squares and cross products appear frequently in regression calculations, let us define the following terms:

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \Big/ n$$

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 \Big/ n$$

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\Big/ n$$

Thus, in this new notation, equation (5) becomes $\hat{b} = S_{xy}/S_{xx}$.

1.2.2 Estimates and their precision

The most we can hope to get from data which do not exactly form a straight line is estimates of a and b; if we add some new data, then almost certainly our estimates will change. By making some assumption about the variance of the variable e (see subsection 1.1.3), we can derive the variances of $\hat{a}$ and $\hat{b}$. In particular, if we assume that the variance of e (denoted by Var (e)) is σ^2, then it follows that

$$\text{Var}(\hat{a}) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}} \tag{7}$$

and

$$\text{Var}(\hat{b}) = \frac{\sigma^2}{S_{xx}} \tag{8}$$

All the quantities in expressions (7) and (8), except for σ^2, may be calculated from the data. Since σ^2 is the variance of the variable e, it is natural to use the residuals, namely

$$\hat{e}_i = y_i - \hat{a} - \hat{b}x_i \qquad (\text{for } i = 1, 2, \ldots, n) \tag{9}$$

in order to estimate σ^2. We know from equations (4) that $\sum_{i=1}^{n} \hat{e}_i = 0$ and, hence, that the arithmetic mean of the residuals is zero.

Consequently, if we use the sum of squares about the mean of the residuals as the basis of our estimate of σ^2, then that sum of squares will just be $\sum_{i=1}^{n} \hat{e}_i^2$.

The appropriate divisor is $n-2$, two degrees of freedom having been 'lost' by estimating a and b. Hence, our estimate of σ^2 will be

$$\begin{aligned} \sigma^2 &= \frac{1}{n-2} \sum_{i=1}^{n} \hat{e}_i^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2 \\ &= \frac{1}{n-2}\left(S_{yy} - \frac{(S_{xy})^2}{S_{xx}}\right) \end{aligned} \tag{10}$$

Which expression in equation (10) is chosen for calculating $\hat{\sigma}^2$ depends on two factors. If the residuals are to be calculated in any case, then it is obviously sensible to use the first expression. If they are not, then the last expression may be preferable. In evaluating the component expressions S_{xx}, S_{xy} and S_{yy}, the remarks which were made at the end of subsection 1.2.1, concerning numerical accuracy apply also in this context.

We are now able to report estimates of a and b together with estimated standard errors ($\sqrt{\text{estimated variance}}$) for those estimates. An alternative summary would be to provide confidence intervals for a and b. However, as these are probabilistic statements, they require some assumptions about the probability distribution of the variable e. The usual assumption is that e is a normal random variable; we have already assumed that its mean is zero and that its variance is σ^2. Shorthand notation for these assumptions is $e \sim N(0, \sigma^2)$.

If, in our model (3), we assume that $e_i \sim N(0, \sigma^2)$ (for $i = 1, 2, \ldots, n$) and that $e_1, e_2, \ldots, e_n$ are independent (see subsection 1.1.2), then

$$\hat{a} \sim N\left(a, \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}\right)$$

and

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{S_{xx}}\right)$$

In addition, $\sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2$, usually called the residual (or error) sum of squares, follows $\sigma^2 \chi^2_{n-2}$ (see footnote). This in turn means that

$$\frac{[\hat{a} - a]}{\sqrt{\dfrac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}} \sim t_{n-2} \tag{11}$$

and

$$\frac{\hat{b} - b}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2} \tag{12}$$

If $t(n, p)$ is defined by $p = \int_{-\infty}^{t(n,p)} f(t_n)\, dt_n$, where $f(t_n)$ is the probability density function of the t random variable with n degrees of freedom, then the $100(1 - \alpha)\%$ confidence interval for a is

$$\hat{a} \pm t(n - 2, 1 - \alpha/2) \sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}$$

and the $100(1 - \alpha)\%$ confidence interval for b is

$$\hat{b} \pm t(n - 2, 1 - \alpha/2) \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

1.2.3 Significance tests

Equations (11) and (12) may also be used to test the validity of appropriate hypotheses about a and b. For instance, with the Alwen data, we might ask whether a hypothesis of $a = 0$ is valid. If it is valid, then this would imply that the model should be

$$\text{run-off} = b \times \text{rainfall} + \text{error}$$

Each of the three random variables χ^2_n, t_n and F_{n_1,n_2} are special functions of Normal random variables which are frequently encountered in practice. The suffices are referred to as degrees of freedom and relate to the number of independent normal random variables involved in the function. Most texts on mathematical statistics give definitions of these random variables and derive their probability density functions.

so that, except for error, we would expect zero run-off when there is zero rainfall.

If our hypothesis $a = 0$ is true, then equation (11) becomes

$$\frac{[\hat{a}]}{\sqrt{\dfrac{\hat{\sigma}^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}} \sim t_{n-2}$$

Hence, if we accept the hypothesis $a = 0$ whenever

$$\frac{[|\hat{a}|]}{\sqrt{\dfrac{\sigma^2 \sum_{i=1}^{n} x_i^2}{nS_{xx}}}} < t(n-2, 1-\alpha/2)$$

and reject the hypothesis otherwise, then this will give us a $100\alpha\,\%$ significance test for this hypothesis.

We might also consider whether the hypothesis $b = 0$ is valid. If it is valid, then this would give a model

$$\text{run-off} = \text{constant} + \text{error}$$

which would imply that rainfall does not affect run-off.

If the hypothesis $b = 0$ is true, then equation (12) becomes

$$\frac{[\hat{b}]}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}} \sim t_{n-2}$$

Hence, if we accept the hypothesis $b = 0$ whenever

$$\frac{[|\hat{b}|]}{\sqrt{\dfrac{\hat{\sigma}^2}{S_{xx}}}} < t(n-2, 1-\alpha/2)$$

and reject the hypothesis otherwise, then this will give a $100\alpha\,\%$ significance test for this hypothesis.

Clearly, both of these test procedures are equivalent to accepting the respective hypotheses whenever the points $a = 0$ or $b = 0$ fall within the $100(1-\alpha)\,\%$ confidence intervals constructed for a and b.

Another hypothesis on b which may be of interest, although not to the Alwen data example, is the hypothesis $b = 1$. If we are testing a new measuring instrument and we are taking readings (y) on items where the exact result (x) is known, then fitting a straight line $y = a + bx$ will allow us to test for correct zeroing of the instrument ($a = 0$) and correct calibration ($b = 1$). This, of course, assumes that a linear relationship is appropriate.

The appropriate test procedure would be to accept the hypothesis $b = 1$ whenever

$$\frac{|\hat{b} - 1|}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} < t(n-2, 1-\alpha/2)$$

and to reject the hypothesis otherwise. Calibration experiments are discussed again later.

1.2.4 Prediction

One of the purposes of fitting a straight line to a set of data might be either to interpolate or to extrapolate. Having carried out a regression of y on x, it is usual to want to predict a value of y corresponding to a known value of x. The obvious predictor is

$$\hat{y} = \hat{a} + \hat{b}x$$

which has variance

$$\text{Var}(\hat{y}) = \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]$$

However, we must be careful to consider just what this estimate is estimating and, in particular, what its variance implies. If we were able to measure values of y repeatedly at this known value of x, then the arithmetic mean of these ys would tend to some fixed number, confusingly called the mean value of y. It is this fixed number which we are estimating and the variance represents the errors of estimation which we will make as a result of using only estimates of a and b. Our fixed number would be $a + bx$ which we could calculate exactly if only we knew a and b.

However, we supposed in model (1) that any single reading of y was made up of $a + bx + e$, each reading showing some unpredictable error e from its ideal value $a + bx$ (the one we have discussed in the previous paragraph). Our estimate of any single reading will be $\hat{a} + \hat{b}x$ (as our estimate of e must be zero) but its variance will be

$$\sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right]$$

the first component for the error of the reading and the second component for the error of prediction.

An alternative presentation of this information is to give confidence intervals for $a + bx$ and $a + bx + e$, the former being for the mean value of y and the latter for a single reading of y. They are

$$\hat{a} + \hat{b}x \pm t(n-2, 1-\alpha/2)\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)} \tag{13}$$

and

$$\hat{a} + \hat{b}x \pm t(n-2, 1-\alpha/2)\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}\right)} \tag{14}$$

respectively, for $100(1-\alpha)\%$ confidence intervals.

Figures 5 and 6 show the confidence intervals (13) and (14) plotted on the same graph as the regression line, $y = \hat{a} + \hat{b}x$; these graphs give a more obvious impression of confidence interval (13) representing the precision of the regression line and confidence interval (14) representing the interval which gives some limits to the readings of y.

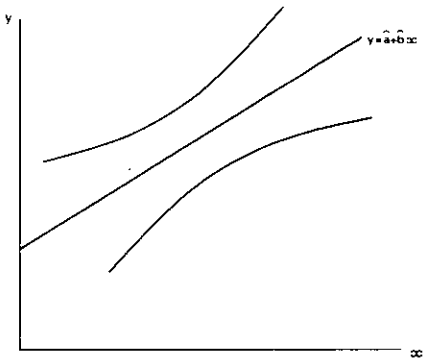


Fig. 5. Confidence interval for the mean value of y.

Figure 5 shows the loci of the confidence limits for the mean value of y. Therefore, for a fixed value of x, there is a probability of $(1-\alpha)$ of the interval defined by (13) containing the mean value of y. The whole region illustrated in Figure 5 should not be confused with the confidence region for the line $y = a + bx$, i.e. the region such that there is an overall probability of $(1-\alpha)$ of it containing $y = a + bx$.

To find the confidence region for $y = a + bx$, replace $t(n-2, 1-\alpha/2)$ in equation (13) by $[F(2, n-2, 1-\alpha)]^{1/2}$, where $1-\alpha = \int_0^{F(2,n-2,1-\alpha)} g(F)\,dF$ and $g(F)$ is the probability density function of the F random variable with 2 and $n-2$ degrees of freedom. Tables of $F(n_1, n_2, 1-\alpha)$ are widely available (for instance, Table 18 of Biometrika Tables for Statisticians, Vol. 1, Pearson and Hartley, 1972).

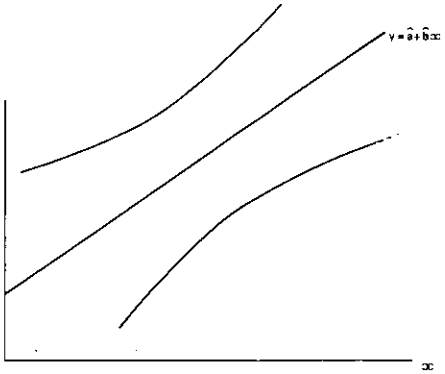


Fig. 6. Confidence or prediction interval for a single reading of y.

If, because of the nature of the variables x and y, it is necessary to predict a value of x from an observed value of y, then the natural estimate would be

$$\hat{x} = \frac{y - \hat{a}}{\hat{b}}$$

A $100(1-\alpha)\%$ confidence interval for the correct value of x would be

$$\hat{x} + \frac{(\hat{x} - \bar{x})\dfrac{t^2\hat{\sigma}^2}{S_{xx}} \pm t\hat{\sigma}\left[\hat{b}^2\left(1 + \dfrac{1}{n} + \dfrac{(\hat{x}-\bar{x})^2}{S_{xx}}\right) - \left(1 + \dfrac{1}{n}\right)\dfrac{t^2\hat{\sigma}^2}{S_{xx}}\right]^{1/2}}{\hat{b}^2 - \dfrac{t^2\hat{\sigma}^2}{S_{xx}}}$$

where $t = t(n-2, 1-\alpha/2)$.

1.3 Extensions to the Basic Method

1.3.1 Repeated observations

Let us consider the situation where, instead of a single value of y being recorded for each value of x, several independent observations of y are available. Alternatively, we could consider that, by chance, there are several values of y all with the same value of x. A notation to cope with this situation is outlined below.

Values of x	x_1	x_2	x_3	x_n
Values of y	y_{11}	y_{21}	y_{31}	y_{n1}
	y_{12}	y_{22}	⋮	⋮
	⋮	⋮		
	y_{1r_1}	y_{2r_2}	y_{3r_3}	y_{nr_n}
Number of y values	r_1	r_2	r_3	r_n

Such a situation might arise when x is a variable over which we have some control or choice and we are able to repeatedly observe values of y under identical conditions.

Our model might be

$$y_{ij} = a + bx_i + e_{ij} \qquad (\text{for } j = 1, 2, \ldots, r_i \quad \text{and} \quad i = 1, 2, \ldots, n) \tag{15}$$

which is not very different from model (3).

Let us define the following terms:

$$S_{xx}^R = \sum_{i=1}^{n} r_i(x_i - \bar{x})^2 = \sum_{i=1}^{n} r_i x_i^2 - \left(\sum_{i=1}^{n} r_i x_i\right)^2 \Big/ N$$

$$S_{xy}^R = \sum_{i=1}^{n} r_i(x_i - \bar{x})(\bar{y}_{i.} - \bar{y}_{..}) = \sum_{i=1}^{n} x_i \sum_{j=1}^{r_i} y_{ij} - \left(\sum_{i=1}^{n}\sum_{j=1}^{r_i} y_{ij}\right)\left(\sum_{i=1}^{n} r_i x_i\right) \Big/ N$$

$$S_{yy}^R = \sum_{i=1}^{n}\sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^{n}\sum_{j=1}^{r_i} y_{ij}^2 - \left(\sum_{i=1}^{n}\sum_{j=1}^{r_i} y_{ij}\right)^2 \Big/ N$$

where

$$\bar{y}_{i.} = \left(\sum_{j=1}^{r_i} y_{ij}\right) \Big/ r_i \qquad \bar{y}_{..} = \left(\sum_{i=1}^{n} \sum_{j=1}^{r_i} y_{ij}\right) \Big/ N \qquad \bar{x} = \left(\sum_{i=1}^{n} r_i x_i\right) \Big/ N$$

and

$$N = \sum_{i=1}^{n} r_i = \text{total number of } y \text{ readings}$$

Then, the least squares estimates of a and b are

$$\hat{a} = \bar{y}_{..} - \hat{b}\bar{x} \tag{16}$$

and

$$\hat{b} = \frac{S_{xy}^R}{S_{xx}^R} \tag{17}$$

The estimate of slope, $\hat{b}$, is similar to that which would have come from fitting a straight line to the pairs of points $(\bar{y}_{i.}, x_i)$ except that each point is weighted according to the number of y readings taken.

The variances of the estimates are

$$\text{Var}(\hat{a}) = \frac{\sigma^2 \sum_{i=1}^{n} r_i x_i^2}{N S_{xx}^R} \tag{18}$$

and

$$\text{Var}(\hat{b}) = \frac{\sigma^2}{S_{xx}^R} \tag{19}$$

When considering a possible estimate of σ^2, it is worth noticing the extra potential offered by data in this form. An estimate similar to (10) would be

$$\frac{1}{n-2} \sum_{i=1}^{n} r_i (\bar{y}_{i.} - \hat{a} - \hat{b}x_i)^2 \tag{20}$$

However, an alternative estimate is available by considering the variability of all the y values which have been recorded for one x value. For instance,

$$\frac{1}{r_i - 1} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i.})^2$$

would give an estimate of σ^2 from y values recorded with $x = x_i$. Using similar estimates of σ^2 from y values recorded with other x values, and combining these into a single expression, gives an estimate

$$\hat{\sigma}^2 = \frac{1}{N-n} \sum_{i=1}^{n} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i.})^2 \tag{21}$$

We might now use the estimate (20) as a measure of how well the linear model fitted the data. Formally, this may be achieved by modifying the model (15) to

$$y_{ij} = a + bx_i + L_i + e_{ij} \qquad (\text{for } j = 1, 2, \ldots, r_i \quad \text{and} \quad i = 1, 2, \ldots, n) \tag{22}$$

where $L_1, L_2, \ldots, L_n$ are unknown parameters which allow for consistent deviations from $a + bx_i$ in the means $\bar{y}_{i.}$. If we find the hypothesis $L_1 = L_2 = \cdots = L_n = 0$ to be valid, then this implies that a linear model $y_{ij} = a + bx_i + e_{ij}$ is adequate for relating y and x. Conversely, if we find the hypothesis to be unreasonable, then this implies that the linear model does not adequately explain the relationship between y and x.

An analysis of variance table provides a neat summary of the information necessary to test this hypothesis, as well as the hypothesis $b = 0$.

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*
Regression	$\hat{b}^2 S_{xx}^R$	1	$\hat{b}^2 S_{xx}^R$
Systematic departure from regression line	$\sum_{i=1}^{n} r_i(\bar{y}_{i.} - \hat{a} - \hat{b}x_i)^2$	$n-2$	(20)
Residual	$\sum_{i=1}^{n} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i.})^2$	$N-n$	(21)
Total	S_{yy}^R	$N-1$	

The column mean square has been derived from (sum of squares/degrees of freedom).

In an analysis of variance table, the total variation in the data (represented by the total sum of squares) is partitioned into a series of meaningful independent quantities. In this case,

$$\begin{matrix}\text{Total} \\ \text{variation}\end{matrix} = \begin{matrix}\text{Variation explained} \\ \text{by the regression line}\end{matrix} + \begin{matrix}\text{Variation explained} \\ \text{by } L_1, L_2, \ldots, L_n\end{matrix} + \begin{matrix}\text{Error} \\ \text{variation}\end{matrix}$$

or, in other words,

$$\begin{matrix}\text{Total sum} \\ \text{of squares}\end{matrix} = \begin{matrix}\text{Regression sum} \\ \text{of squares}\end{matrix} + \begin{matrix}\text{Systematic departure} \\ \text{from regression line} \\ \text{sum of squares}\end{matrix} + \begin{matrix}\text{Residual sum} \\ \text{of squares}\end{matrix}$$

In most properly constructed analysis of variance tables, the ratio

$$\frac{\text{Mean square due to } X}{\text{Residual mean square}}$$

will follow an F distribution, with degrees of freedom equal to those of the numerator and those of the denominator respectively, whenever X has no real effect, or role, in explaining the total variation.

Thus,

$$\frac{\text{Regression mean square}}{\text{Residual mean square}} \sim F_{1,N-n}$$

when no variation has been explained by the regression, i.e. when the hypothesis $b = 0$ is true.

Similarly,

$$\frac{\text{Systematic departure} \cdots \text{mean square}}{\text{Residual mean square}} \sim F_{n-2,N-n}$$

when there is no systematic departure from the regression line, i.e. when the hypothesis $L_1 = L_2 = \cdots = L_n = 0$ is true.

Hence, a $100\alpha\%$ significance test would lead us to accept the hypothesis $b = 0$ whenever

$$\frac{\text{Regression mean square}}{\text{Residual mean square}} < F(1, N-n, 1-\alpha)$$

Similarly, a $100\alpha\%$ significance test would lead us to accept the hypothesis $L_1 = L_2 = \cdots = L_n = 0$ whenever

$$\frac{\text{Systematic departure} \cdots \text{mean square}}{\text{Residual mean square}} < F(n-2, N-n, 1-\alpha)$$

The $100(1-\alpha)\%$ confidence intervals for a and b are

$$\hat{a} \pm t(N-n, 1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2 \sum_{i=1}^{n} r_i x_i^2}{NS_{xx}^R}}$$

and

$$\hat{b} \pm t(N-n, 1-\alpha/2)\sqrt{\frac{\hat{\sigma}^2}{S_{xx}^R}}$$

respectively, where σ^2 is given in equation (21).

1.3.2 Fitting and comparing several straight lines

If several sets of rainfall–run-off data have been collected from different sites and a linear model has proved to give a satisfactory explanation of the data, then it may prove useful to compare the estimates of a and b calculated from the data from the different sites. Some interpretation may be attached to a and b; for instance, if we interpret a as the run-off from zero rainfall and b as the proportion of rainfall appearing as run-off, then subsequent comparisons of the estimates of a and b will give some idea of the similarity of the sites in these two features.

Let us assume that there are n sites from which data have been collected and that, from site i, the data consist of r_i pairs of readings of y and x which are denoted by (y_{ij}, x_{ij}) (for $j = 1, 2, \ldots, r_i$). It will usually be sensible to fit separate straight lines to the data from each site. For site i, the model equivalent to (3) would be

$$y_{ij} = a_i + b_i x_{ij} + e_{ij} \tag{23}$$

Estimates of a_i and b_i would be derived by applying the basic method

described in Section 1.2 to the data from each site in turn. Using the above notation, this would give estimates

$$\hat{a}_i = \bar{y}_{i.} - \hat{b}_i \bar{x}_{i.}$$

and

$$\hat{b}_i = \left(\sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i.})(x_{ij} - \bar{x}_{i.}) \right) \Big/ \left(\sum_{j=1}^{r_i} (x_{ij} - \bar{x}_{i.})^2 \right)$$

where

$$\bar{y}_{i.} = \frac{1}{r_i} \sum_{j=1}^{r_i} y_{ij} \qquad \text{and} \qquad \bar{x}_{i.} = \frac{1}{r_i} \sum_{j=1}^{r_i} x_{ij}$$

Similarly, the variance of the variable e for site i, denoted by σ_i^2, would be estimated by

$$\hat{\sigma}_i^2 = \frac{1}{r_i - 2} \sum_{j=1}^{r_i} (y_{ij} - \hat{a}_i - \hat{b}_i x_{ij})^2$$

Thus, we would be able to imagine the data from each site having been condensed into three numbers, denoted by $\hat{a}_i$, $\hat{b}_i$ and $\hat{\sigma}_i^2$. In a comparison of sites, it would be sensible, from the statistical point of view, to start by comparing the values of $\hat{\sigma}_i^2$ from each of the sites.

This may be achieved by testing the hypothesis $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$. One test statistic for this is

$$M = (N - 2n) \log_e \left[\frac{\sum_{i=1}^{n} (r_i - 2)\hat{\sigma}_i^2}{(N - 2n)} \right] - \sum_{i=1}^{n} (r_i - 2) \log_e \hat{\sigma}_i^2$$

where

$$N = \sum_{i=1}^{n} r_i$$

The distribution of M is approximated by χ_{n-1}^2 whenever $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$. Thus, a $100\alpha\%$ significance test would be to accept this hypothesis whenever

$$M < \chi^2(n - 1, 1 - \alpha)$$

where $1 - \alpha = \int_0^{\chi^2(n-1,1-\alpha)} f(\chi_{n-1}^2)\, d\chi_{n-1}^2$ and $f(\chi_{n-1}^2)$ is the probability density function of χ_{n-1}^2. Tables 7 and 8 of Biometrika Tables for Statisticians, Vol. 1 (Pearson & Hartley (1972)) may be used to give values of $\chi^2(n - 1, 1 - \alpha)$ and improvements to the approximation of the distribution of M are given in the text accompanying Table 32. This test assumes that the variable e follows a normal distribution and, unfortunately, a significant value of M may indicate non-normality rather than heterogeneity of variance.

However, if the hypothesis $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$ is acceptable, then this implies that run-off readings at a fixed rainfall level show similar variability within each of the different sites and/or that the linear model is equally successful at explaining the relationship between rainfall and run-off within each of the different sites.

Being able to accept this hypothesis leads to simpler and more meaningful comparisons between the other statistics. The appropriate tests are most easily displayed in an analysis of variance table but, in order to avoid cumbersome algebraic expressions, it will be necessary to introduce some new notation. Let us define the following terms:

$$S^i_{xx} = \sum_{j=1}^{r_i} (x_{ij} - \bar{x}_{i.})^2$$

$$S^i_{yy} = \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i.})^2$$

$$S^i_{xy} = \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i.})(x_{ij} - \bar{x}_{i.})$$

[Hence, $\hat{b}_i = S^i_{xy}/S^i_{xx}$.]

$$S^o_{xx} = \sum_{i=1}^{n} \sum_{j=1}^{r_i} (x_{ij} - \bar{x}_{..})^2$$

$$S^o_{yy} = \sum_{i=1}^{n} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{..})^2$$

$$S^o_{xy} = \sum_{i=1}^{n} \sum_{j=1}^{r_i} (x_{ij} - \bar{x}_{..})(y_{ij} - \bar{y}_{..})$$

$$\hat{b}_o = S^o_{xy}/S^o_{xx}$$

where

$$\bar{x}_{..} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{r_i} x_{ij} \qquad \text{and} \qquad \bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{r_i} y_{ij}$$

The first four quantities are calculated using data from just a single site. However, although the remaining expressions are of a similar type to the first four, they involve data from all of the sites. They are calculated by ignoring the distinction of sites and using all of the data from all sites to give an 'overall' quantity (o = overall).

Finally, let us define

$$S^c_{xx} = \sum_{i=1}^{n} S^i_{xx}$$

$$S^c_{xy} = \sum_{i=1}^{n} S^i_{xy}$$

$$S^c_{yy} = \sum_{i=1}^{n} S^i_{yy}$$

$$\hat{b}_c = S^c_{xy}/S^c_{xx}$$

These four expressions are of a similar type to S^o_{xx}, S^o_{xy}, etc., in that they involve

data from all of the sites. However, they do not ignore the identity of the sites but give a combined quantity which allows certain differences that might exist between the sites to be taken into account (c = combined).

If we were able to conclude that the hypothesis $b_1 = b_2 = \cdots = b_n$ was acceptable, then $\hat{b}_c$ would represent a sensible combined estimate of the common slope. However, if, in addition, we were able to conclude that the hypothesis $a_1 = a_2 = \cdots = a_n$ was acceptable, then $\hat{b}_0$ would be a more sensible combined estimate of the common slope.

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*
Overall regression	$\hat{b}_0 S^0_{xy}$		
Difference in positions	$(S^0_{yy} - \hat{b}_0 S^0_{xy}) - (S^c_{yy} - \hat{b}_c S^c_{xy})$		$\left(= \dfrac{\text{Sum of squares}}{\text{Degrees of freedom}}\right)$
Differences in slopes	$\sum_{i=1}^{n} \hat{b}_i S^i_{xy} - \hat{b}_c S^c_{xy}$		
Residual	$S^c_{yy} - \sum_{i=1}^{n} \hat{b}_i S^i_{xy}$	$N - 2n$	
Total	S^0_{yy}	$N - 1$	

A similar procedure may be applied to this analysis of variance table as to the previous one. However, it is preferable to carry out the tests in the following order:

for $100\alpha\%$ significance tests

1. accept the hypothesis $b_1 = b_2 = \cdots = b_n$ whenever
$$\frac{\text{Difference in slopes mean square}}{\text{Residual mean square}} < F(n-1, N-2n, 1-\alpha)$$
2. if the hypothesis $b_1 = b_2 = \cdots = b_n$ has been accepted, then accept the hypothesis $a_1 = a_2 = \cdots = a_n$ whenever
$$\frac{\text{Difference in positions mean square}}{\text{Residual mean square}} < F(n-1, N-2n, 1-\alpha)$$
3. if both hypotheses $b_1 = b_2 = \cdots = b_n$ and $a_1 = a_2 = \cdots = a_n$ have been accepted, then accept that there is no linear association between y and x whenever
$$\frac{\text{Overall regression mean square}}{\text{Residual mean square}} < F(1, N-2n, 1-\alpha)$$

For more complex comparisons, such as concurrency of regression lines, the reader is referred to a more advanced text on regression analysis, such as Williams (1959) or Seber (1977).

1.3.3 Observations with unequal precision

One of the assumptions mentioned in subsection 1.1.2 was that all of the y values should be measured with equal precision, i.e. the fluctuation or variability in each y value should be the same. This will not always be true for hydrological data and methods for detecting whether this is the case, which use the data only, are given in Section 4.3.

By considering the type of data being recorded, or by using the results of previous studies, it may be possible to relate the variance of a y value to the y value itself. If it is possible, then the problem of unequal precision may be overcome by taking some transformation of the y values as described in Section 3.3.

Occasionally, the variances of the y values are known exactly. This will not usually happen when, for instance, y is run-off and x is rainfall. However, it may occur when the ys are some statistics such as the slopes of a regression line calculated on separate sets of data which are being related to some feature x measured on each of the sets of data.

Our information will then consist of the pairs of points $(y_1, x_1), \ldots, (y_n, x_n)$ together with the n variances of the y values, $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. The estimate in equation (5) will still be an unbiased estimate of the slope of the regression line. However, under these new assumptions, its variance will be $(\sum_{i=1}^{n} \sigma_i^2 (x_i - \bar{x})^2)/(S_{xx})^2$ and this is larger than the variance of an alternative estimator,

$$\hat{b} = \frac{S_{xy}^w}{S_{xx}^w} \tag{24}$$

where S_{xx}^w and S_{xy}^w are defined as follows:

$$S_{xx}^w = \sum_{i=1}^{n} w_i (x_i - \bar{x})^2$$

$$S_{xy}^w = \sum_{i=1}^{n} w_i (x_i - \bar{x})(y_i - \bar{y})$$

where

$$\bar{y} = \left(\sum_{i=1}^{n} w_i y_i \right) \Big/ \left(\sum_{i=1}^{n} w_i \right) \qquad \bar{x} = \left(\sum_{i=1}^{n} w_i x_i \right) \Big/ \left(\sum_{i=1}^{n} w_i \right) \quad \text{and} \quad w_i = \frac{1}{\sigma_i^2}$$

The corresponding estimator of a is

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

The variances of these new estimators are

$$\text{Var}(\hat{b}) = \frac{1}{S_{xx}^w}$$

and

$$\text{Var}(\hat{a}) = \frac{1}{\sum_{i=1}^{n} w_i} + \frac{\bar{x}^2}{S_{xx}^w}$$

The $100(1-\alpha)\%$ confidence intervals for a and b would be

$$\hat{b} \pm Z(\alpha/2)\sqrt{\frac{1}{S_{xx}^{w}}}$$

and

$$\hat{a} \pm Z(\alpha/2)\sqrt{\frac{1}{\sum_{i=1}^{n} w_i} + \frac{\bar{x}^2}{S_{xx}^{w}}}$$

where

$$1-\alpha = \int_{-\infty}^{Z(\alpha)} \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du$$

(The use of the normal distribution rather than the t distribution in these calculations is a direct consequence of knowing the variances of the y values.)

If repeated observations are available, and if the y readings associated with each value of x_i have variance σ_i^2, then the estimates in equations (16) and (17) will become

$$\hat{a} = \bar{y}_{..} - \hat{b}\bar{x}$$

and

$$\hat{b} = \frac{S_{xy}^{wR}}{S_{xx}^{wR}}$$

where S_{xx}^{wR} and S_{xy}^{wR} are defined as follows:

$$S_{xx}^{wR} = \sum_{i=1}^{n} w_i r_i (x_i - \bar{x})^2$$

$$S_{xy}^{wR} = \sum_{i=1}^{n} w_i r_i (\bar{y}_{i.} - \bar{y}_{..})(x_i - \bar{x})$$

where

$$\bar{y}_{i.} = \left(\sum_{j=1}^{r_i} y_{ij}\right) \Big/ r_i \qquad \bar{y}_{..} = \left(\sum_{i=1}^{n} w_i r_i \bar{y}_{i.}\right) \Big/ \left(\sum_{i=1}^{n} w_i r_i\right)$$

$$\bar{x} = \left(\sum_{i=1}^{n} w_i r_i x_i\right) \Big/ \left(\sum_{i=1}^{n} w_i r_i\right) \qquad \text{and} \qquad w_i = \frac{1}{\sigma_i^2}$$

The variances of $\hat{a}$ and $\hat{b}$ will be

$$\mathrm{Var}(\hat{a}) = \frac{1}{\sum_{i=1}^{n} w_i r_i} + \frac{\bar{x}^2}{S_{xx}^{wR}}$$

and

$$\mathrm{Var}(\hat{b}) = \frac{1}{S_{xx}^{wR}}$$

The analysis of variance table will become

Source	*Sum of squares*	*Degrees of freedom*
Regression	$\hat{b}^2 S_{xx}^{wR}$	1
Systematic departure from regression line	$\sum_{i=1}^{n} w_i r_i (\bar{y}_{i.} - \hat{a} - \hat{b}x_i)^2$	$n-2$
Residual	$\sum_{i=1}^{n} \sum_{j=1}^{r_i} w_i (y_{ij} - \bar{y}_{i.})^2$	$N-n$
Total	$\sum_{i=1}^{n} \sum_{j=1}^{r_i} w_i (y_{ij} - \bar{y}_{..})^2$	$N-1$

The methods of testing and the conclusions are similar to those described in subsection 1.2.1.

As will be seen in Section 3.3, it is a help to have repeated observations in a study where it is suspected that the variance of y may not remain constant. Initially, it is straightforward to test whether the variance of y has remained constant and then, if it has not, it is possible to allow for this even when the variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$ are unknown. Estimates of these variances may be obtained from

$$\hat{\sigma}_i^2 = \frac{1}{r_i - 1} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_{i.})^2$$

and these may be used in the preceding theory to give estimates of a and b. However, inference from the confidence intervals and the analysis of variance table should be made with caution, particularly when any of $r_1, r_2, \ldots, r_n$ are small.

Alternatively, plotting $\hat{\sigma}_i^2$ against x_i may suggest that a relationship exists between the variance of y and the variable x (e.g. $\sigma_i^2 = \alpha x_i$ or $\sigma_i^2 = \alpha x_i^2$). If such a relationship were, for instance, $\sigma_i^2 = \alpha x_i$, then w_i in equation (24) could be replaced by $1/(\alpha x_i)$ giving

$$\hat{b} = \left(\sum_{i=1}^{n} \frac{(y_i - \bar{y})(x_i - \bar{x})}{x_i} \right) \Big/ \left(\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{x_i} \right)$$

Otherwise, a transformation of y might be appropriate and this technique is described in Section 3.3.

1.4 Alternatives to Least Squares

1.4.1 Pencil and ruler

Anyone who attempts a regression analysis without plotting the data in some form is asking for trouble. A plot of y against x on graph paper will reveal the type of relationship that might exist between y and x. It will show whether y

increases or decreases with x and whether the relationship is linear or non-linear. It will suggest how strong or weak the relationship might be and, indeed, whether there is any relationship at all. It will show up points which are obviously different from the majority and it will indicate the range of y and x over which the relationship has been investigated.

Why then do we not finish the job, draw a line down the middle of the data and forget about mathematical formulae and calculations? The main reasons are that least squares is a method which is impartial, gives repeatable results and provides a framework for inference. Furthermore, if you genuinely believe that the linear regression model is the appropriate one, then least squares is the method which will give the 'best line' (i.e. the most precise estimates of a and b). Imagine being faced with a plot of points; there is frequently no natural 'middle', no person would have much confidence in someone else's straight line, and who, in any case, could quantify the precision of their straight line?

Unfortunately, least squares estimation will not necessarily give 'the right line'; it is, at best, an intelligent guess. It relies on certain assumptions and, consequently, if these are not valid, then a critical assessment by eye, which discounts some points and gives greater weight to others, may give a straight line which better suits the short term objectives that the experimenter has in mind. However, in the long term, he will probably benefit from investigating the reasons why the least squares assumptions are invalid.

1.4.2 Robust and distribution free methods

Distribution free methods of estimation and testing occupy an intermediate position between the pencil and ruler method and the method of least squares estimation. They do not require as many assumptions as the least squares method but, nevertheless, they do allow inference, as well as estimation, to be made on the slope parameter b. The assumptions usually required are that the relationship between y and x is of the form described in equation (3) and that the es are mutually independent and follow the same distribution.

A simple distribution free method of estimation is to take a pair of points, say (y_i, x_i) and (y_j, x_j), and to calculate the slope of the line joining these two points, i.e. calculate

$$\frac{y_i - y_j}{x_i - x_j} = \hat{b}_{ij}$$

This is repeated for all $n(n-1)/2$ pairs of points to give $n(n-1)/2$ separate slopes, $\hat{b}_{12}, \hat{b}_{13}, \ldots, \hat{b}_{n-1,n}$. Then, the numbers $\hat{b}_{12}, \hat{b}_{13}, \ldots, \hat{b}_{n-1,n}$ are arranged in increasing order of magnitude to give an ordered sequence denoted by $\hat{b}_{(1)} \leq \hat{b}_{(2)} \leq \hat{b}_{(3)} \leq \cdots \leq \hat{b}_{(N)}$ where $N = n(n-1)/2$. The median of this set of numbers ($\hat{b}_{((N+1)/2)}$ if N is odd and $\frac{1}{2}(\hat{b}_{(N/2)} + \hat{b}_{(N/2+1)})$ if N is even) is then taken as the estimate of the slope parameter, b.

To obtain an approximate $100(1-\alpha)\%$ confidence interval for b, the following quantities are calculated:

$$r_1 = \begin{matrix}\text{nearest} \\ \text{integer} \\ \text{to}\end{matrix} \frac{1}{2}\left(N - Z(\alpha/2)\sqrt{\frac{n(n-1)(2n+5)}{18}}\right)$$

and

$$r_2 = \begin{matrix}\text{nearest} \\ \text{integer} \\ \text{to}\end{matrix} \frac{1}{2}\left(N + Z(\alpha/2)\sqrt{\frac{n(n-1)(2n+5)}{18}}\right)$$

where $Z(\alpha/2)$ is defined in subsection 1.3.3.

The lower and upper limits of the $100(1-\alpha)\%$ confidence interval for b are taken to be $\hat{b}_{(r_1)}$ and $\hat{b}_{(r_2)}$ respectively. This approximate procedure is only valid for a relatively large n. For an exact procedure, the reader is referred to Hollander and Wolfe (1973), p. 207.

An alternative and ingenious method was developed by Daniels (1954). It is based on the fact that $y = a + bx$ may be written in the form $a = y - xb$ and that this equation may be regarded as a straight line relating a and b, with slope $-x$ and intercept y. Thus, the set of readings $(y_1, x_1), \ldots, (y_n, x_n)$ may be represented as n lines $(a = y_1 - x_1 b), (a = y_2 - x_2 b), \ldots, (a = y_n - x_n b)$ which, in pictorial form, might look like Figure 7.

Ideally, we would expect all the lines to intersect at one point which would give us our estimates of a and b. Of course, this would only occur if all the original points $(y_1, x_1), \ldots, (y_n, x_n)$ happened to fall exactly on a straight line. We will usually have to choose some region in the 'middle' of the mass of intersecting lines as containing our estimates of a and b.

As is illustrated in Figure 7, the picture will usually consist of a set of closed regions (near the middle) and a set of open regions (around the edge). A convenient score for any particular region is denoted by m and defined to be the minimum number of lines which have to be crossed to escape from that region into the nearest open region.

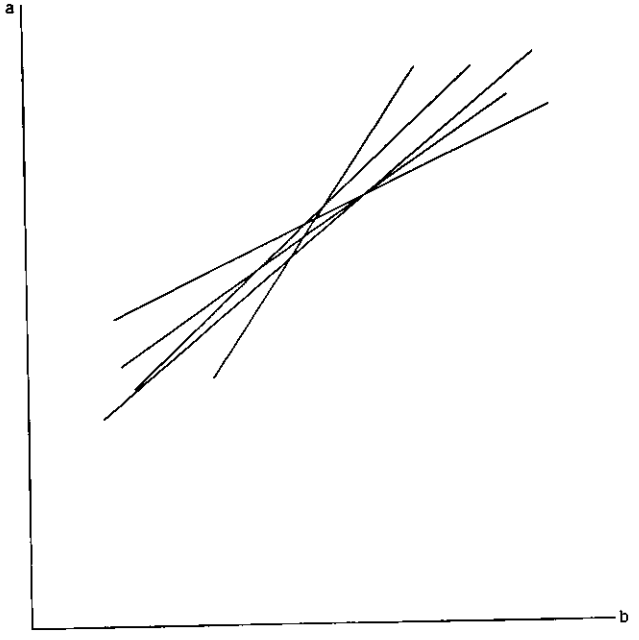


Fig. 7. Data points represented by a series of straight lines.

Thus, if one region has the largest m, then it would seem reasonable to take a and b as being in that region. However, as this does not give us unique values of a and b and as it is not necessarily true that just one region will have the largest m, it seems that a confidence interval is the natural outcome of this method of estimation.

The $100(1-\alpha)\%$ confidence region for a and b is made up of all those regions for which $m > m_0$ where the value of m_0 is calculated as follows:

for $\alpha = 0.05$,

$$m_0 \sim \text{integer (nearest)} \tfrac{1}{2}(n - 3.023\sqrt{n}) \text{ (to)}$$

for $\alpha = 0.01$,

$$m_0 = \text{integer (nearest)} \tfrac{1}{2}(n - 3.562\sqrt{n}) \text{ (to)}$$

These values of m_0 are approximations for large n. However, the former is not misleading for $n \geq 12$ and the latter for $n \geq 16$ and, in both cases, when n is below these limits, the exact value of m_0 is zero. Alternatively, the exact value of m_0 may be calculated by solving

$$\alpha = 4z \sum_{r=0}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(2rz+z)^2/2}$$

for m_0 where $z = (n - 2m_0)/\sqrt{n}$.

We may use this information to test hypotheses about a and b. For instance, in order to test the hypothesis $a = 0$, $b = 1$, we check whether the region in which this point falls has been included in the $100(1-\alpha)\%$ confidence interval. If it has, then we accept the hypothesis that $a = 0$, $b = 1$; if it has not, then we reject the hypothesis. This will give a $100\alpha\%$ significance test.

A similar use may be made of the confidence interval which was calculated by the previous method. For a $100\alpha\%$ significance test, we should accept the hypothesis on b whenever the hypothesized value of b is included in the $100(1-\alpha)\%$ confidence interval.

1.4.3 Bayesian methods

In this section on Bayesian methods, it will be more convenient to take the model relating y and x in the form

$$y_i = \alpha + \beta(x_i - \bar{x}) + e_i \tag{25}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

By comparing this model with model (3), it will be seen that $a = \alpha - \beta\bar{x}$ and $b = \beta$. Using the same notation as before, the least squares estimates of α and β are

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y} \tag{26}$$

and

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \tag{27}$$

and the variances of these estimates are $\text{Var}(\hat{\alpha}) = \sigma^2/n$ and $\text{Var}(\hat{\beta}) = \sigma^2/S_{xx}$. If the assumptions described in subsection 1.1.2 are valid, then the estimates, $\hat{\alpha}$ and $\hat{\beta}$, are independent.

Bayesian methods allow the use of information about α and β which is additional to that provided by the data. Ideally, the information about α should take the form of a distribution (called the prior distribution) which would give the possible values of α and how likely they are to occur, i.e. the prior distribution for α would be a summary of the state of knowledge about α before the data in question were available. A similar distribution should be available for β.

For example, we might assume that the prior distribution for α is Normal with mean μ_α and variance σ_α^2, i.e. our past experience suggests a tendency for α to take values centred about μ_α with the variability about that point having the characteristics of the Normal distribution. We might also assume that the prior distribution for β is Normal, but with mean μ_β and variance σ_β^2.

An objective of a Bayesian analysis is to update the prior distributions by including the information on α and β contained in the data. The resulting distribution, the 'updated prior', is called the posterior distribution and it summarises all that is known about α and β, including the information contained in the data.

In our example, where Normal prior distributions are assumed for α and β, the posterior distributions are as follows:

for α,
$$N\left(\frac{\hat{\alpha}\dfrac{n}{\sigma^2} + \mu_\alpha \dfrac{1}{\sigma_\alpha^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_\alpha^2}}, \frac{1}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_\alpha^2}}\right)$$

for β,
$$N\left(\frac{\hat{\beta}\dfrac{S_{xx}}{\sigma^2} + \mu_\beta \dfrac{1}{\sigma_\beta^2}}{\dfrac{S_{xx}}{\sigma^2} + \dfrac{1}{\sigma_\beta^2}}, \frac{1}{\dfrac{S_{xx}}{\sigma^2} + \dfrac{1}{\sigma_\beta^2}}\right)$$

However, if we want to report only a single value for α, then it would be natural to use the mean of the posterior distribution of α,

$$\frac{\hat{\alpha}\dfrac{n}{\sigma^2} + \mu_\alpha \dfrac{1}{\sigma_\alpha^2}}{\dfrac{n}{\sigma^2} + \dfrac{1}{\sigma_\alpha^2}}.$$

This is called the Bayes estimator of α and it is clearly just the weighted mean of

the least squares estimate of α and the prior mean of α. Similarly, the Bayes estimator of β is

$$\frac{\hat{\beta}\frac{S_{xx}}{\sigma^2}+\mu_\beta\frac{1}{\sigma_\beta^2}}{\frac{S_{xx}}{\sigma^2}+\frac{1}{\sigma_\beta^2}}.$$

Complete prior ignorance about a parameter is usually expressed by using a uniform prior distribution for that parameter. If uniform prior distributions are assumed for α and β, then it follows that the posterior distribution for α is $N(\hat{\alpha}, \sigma^2/n)$ and the posterior distribution for β is $N(\hat{\beta}, \sigma^2/S_{xx})$.

Thus, if nothing is known about α and β prior to collecting the data, then the Bayes estimators of α and β will correspond with the least squares estimates of α and β.

The Bayesian method has the potential to incorporate into the estimation of α and β all shades of opinion and knowledge which can be summarised in the form of a prior distribution. However, it is more likely that our prior knowledge will consist of several independent estimates of α and β which we have previously derived from similar sets of data to our present set. Thus, although we might be able to guess at the form of the distribution of these estimates, we will probably be quite unable to describe it precisely and say that, for instance, it is Normal with a particular mean and a particular variance.

Empirical Bayes methods have been derived specifically to cope with this problem. Suppose that, on $k-1$ previous occasions in comparable circumstances, data sets similar to the present one have been collected and, from each data set, estimates of α and β have been derived. Denote these estimates by $\hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_{k-1}$ and $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{k-1}$. From the present data, we may calculate least squares estimates of α, β and σ^2 as given by (26), (27) and (10). Denote these estimates by $\hat{\alpha}_k$, $\hat{\beta}_k$ and $\hat{\sigma}^2$, respectively.

Define h_α to be the larger of

$$\left(\frac{1}{k}\right)^{1/5}\sqrt{\frac{1}{k}\sum_{j=1}^{k}(\hat{\alpha}_j-\bar{\alpha})^2} \quad \text{and} \quad \left(\frac{1}{k}\right)^{1/5}\hat{\sigma}/\sqrt{n}$$

and define h_β to be the larger of

$$\left(\frac{1}{k}\right)^{1/5}\sqrt{\frac{1}{k}\sum_{j=1}^{k}(\hat{\beta}_j-\bar{\beta})^2} \quad \text{and} \quad \left(\frac{1}{k}\right)^{1/5}\hat{\sigma}/\sqrt{S_{xx}}$$

where

$$\bar{\alpha}=\frac{1}{k}\sum_{j=1}^{k}\hat{\alpha}_j \quad \text{and} \quad \bar{\beta}=\frac{1}{k}\sum_{j=1}^{k}\hat{\beta}_j$$

Now, let

$$A_j=\frac{\hat{\alpha}_k-\hat{\alpha}_j}{2h_\alpha} \qquad B_j=A_j+\tfrac{1}{2} \qquad C_j=\frac{\hat{\beta}_k-\hat{\beta}_j}{2h_\beta} \quad \text{and} \quad D_j=C_j+\tfrac{1}{2}$$

Then, an empirical Bayes estimate of α is

$$\hat{\alpha}_k + \frac{\hat{\sigma}^2}{n}\left[\frac{\sum_{j=1}^{k}((\sin A_j)/A_j)^2 + ((\sin B_j)/B_j)^2}{h_\alpha \sum_{j=1}^{k}((\sin A_j)/A_j)^2}\right]$$

and an empirical Bayes estimate of β is

$$\hat{\beta}_k + \frac{\hat{\sigma}^2}{S_{xx}}\left[\frac{\sum_{j=1}^{k}((\sin C_j)/C_j)^2 + ((\sin D_j)/D_j)^2}{h_\beta \sum_{j=1}^{k}((\sin C_j)/C_j)^2}\right]$$

For further details of this method, the reader is referred to the original paper by Clemmer and Krutchkoff (1968).

1.4.4 Linear functional relationships

It has been emphasised that the linear regression model (1) essentially assumed that error, random variation, etc. only affected the dependent variable, y. A more general, and perhaps more realistic, model might allow both y and x to be random variables.

The functional relationship model assumes that a linear relationship would exist between y and x, if y and x could have been recorded in idealised circumstances where no error was made.

Hence, the functional relationship model assumes

$$\text{ideal } y = a + b\ (\text{ideal } x)$$

However, the normal readings that we are able to take of y and x are related to the idealised ones by

$$y \text{ reading} = \text{ideal } y + e$$

and

$$x \text{ reading} = \text{ideal } x + \delta$$

where e and δ represent the errors. Thus, for our n pairs of readings, $(y_1, x_1), \ldots, (y_n, x_n)$, there will be an associated set of (unknown) errors, $(e_1, \delta_1), \ldots, (e_n, \delta_n)$, and our model will be

$$(y_i - e_i) = a + b(x_i - \delta_i) \tag{28}$$

We will also assume that both the e and the δ errors are normally distributed with variances σ_e^2 and σ_δ^2, respectively. Consequently, we are assuming that all y observations are made with equal precision, and likewise for the x observations.

Therefore, if we are studying a situation in which both y and x are subject to error and model (1) is inappropriate, then we might be obliged to use this functional relationship model. At first sight, it might seem that it will always be

better to use this model, particularly as model (1) is just a special case of model (28) with $\delta_i = 0$. However, in order to estimate a and b in model (28), more information is required than that provided by the n pairs of readings alone.

The bare minimum of information required is knowledge of either (a) σ_e^2, (b) σ_δ^2 or (c) the ratio $\lambda = \sigma_e/\sigma_\delta$.

In each case,

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

However, in case (a),

$$\hat{b} = \frac{S_{yy} - (n-1)\sigma_e^2}{S_{xy}} \tag{29}$$

in case (b),

$$\hat{b} = \frac{S_{xy}}{S_{xx} - (n-1)\sigma_\delta^2} \tag{30}$$

and, in case (c),

$$\hat{b} = \frac{(S_{yy} - \lambda^2 S_{xx}) + \sqrt{(S_{yy} - \lambda^2 S_{xx})^2 + 4\lambda^2 (S_{xy})^2}}{2S_{xy}} \tag{31}$$

If the numerator of equation (29) is negative, then take $\hat{b} = 0$. If the denominator of equation (30) is negative, then take $\hat{b} = \infty$.

In case (c), a $100(1-\alpha)\%$ confidence interval for b is given by

$$\lambda \tan\left(\tan^{-1}\left(\frac{\hat{b}}{\lambda}\right) \pm \tfrac{1}{2}\sin^{-1}\left[2t(n-2, 1-\alpha/2)X\right]\right)$$

where

$$X^2 = \frac{\lambda^2(S_{xx}S_{yy} - (S_{xy})^2)}{(n-2)[(S_{yy} - \lambda^2 S_{xx})^2 + 4\lambda^2(S_{xy})^2]}$$

In order to test the hypothesis $b = b_0$ (typically, b_0 might be 1 or 0), it is probably easiest to compute the above confidence interval and then to check whether b_0 is included in this interval. If it is, then we accept the hypothesis $b = b_0$; otherwise, we reject the hypothesis. This will provide a $100\alpha\%$ significance test.

The estimates of b given above are the maximum likelihood estimates appropriate for the three different situations. An alternative quick method of estimation is as follows:

1. Plot $(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$.
2. Divide the x axis into three parts so that approximately 1/3 of the observed x values fall in each part. (Ensure that the first and last group contain an equal number which is as close to $n/3$ as possible.)
3. Compute the arithmetic means of the x and y values in the first group (denoted by $\bar{x}_1$ and $\bar{y}_1$ respectively) and the third group (denoted by $\bar{x}_3$ and $\bar{y}_3$, respectively).
4. Estimate b by $\hat{b} = (\bar{y}_3 - \bar{y}_1)/(\bar{x}_3 - \bar{x}_1)$ and estimate a by $\hat{a} = \bar{y} - \hat{b}\bar{x}$.

A $100(1-\alpha)\%$ confidence interval for b may be formed although it requires considerably more calculation. The following table illustrates the data after being divided into three groups.

	Group 1		*Group 2*		*Group 3*	
	x values	*y values*	*x values*	*y values*	*x values*	*y values*
DATA	x_{11}	y_{11}	x_{21}	y_{21}	x_{31}	y_{31}
	x_{12}	y_{12}	x_{22}	y_{22}	x_{32}	y_{32}
	⋮	⋮	⋮	⋮	⋮	⋮
	x_{1k}	y_{1k}	x_{2m}	y_{2m}	x_{3k}	y_{3k}
Arithmetic mean	$\bar{x}_1$	$\bar{y}_1$	$\bar{x}_2$	$\bar{y}_2$	$\bar{x}_3$	$\bar{y}_3$

k should be as near to $n/3$ as possible and $m = n - 2k$.

Let us define the following terms:

$$S_{xx}^G = \sum_{i=1}^{k} (x_{1i} - \bar{x}_1)^2 + \sum_{i=1}^{m} (x_{2i} - \bar{x}_2)^2 + \sum_{i=1}^{k} (x_{3i} - \bar{x}_3)^2$$

$$S_{xy}^G = \sum_{i=1}^{k} (x_{1i} - \bar{x}_1)(y_{1i} - \bar{y}_1) + \sum_{i=1}^{m} (x_{2i} - \bar{x}_2)(y_{2i} - \bar{y}_2) + \sum_{i=1}^{k} (x_{3i} - \bar{x}_3)(y_{3i} - \bar{y}_3)$$

$$S_{yy}^G = \sum_{i=1}^{k} (y_{1i} - \bar{y}_1)^2 + \sum_{i=1}^{m} (y_{2i} - \bar{y}_2)^2 + \sum_{i=1}^{k} (y_{3i} - \bar{y}_3)^2$$

The lower and upper limits of the confidence interval for b are given as the two roots of the quadratic equation in b,

$$\frac{k}{2}(\bar{x}_3 - \bar{x}_1)^2(\hat{b} - b)^2 = \frac{[t(n-3, 1-\alpha/2)]^2}{(n-3)}(S_{yy}^G - 2bS_{xy}^G + b^2 S_{xx}^G)$$

References

Clemmer, B. A. and Krutchkoff, R. G. (1968). *Biometrika*, **55**(3), 525.

Daniels, H. E. (1954). *Ann. Math. Stat.*, **25**(3), 499.

Hollander and Wolfe (1973). *Non parametric Statistical Methods*. John Wiley & Sons.

Lewis, W. K. (1957). Investigation of Rainfall, Run-Off and Yield on the Alwen and Brenig Catchments.

Pearson, E. S. and Hartley, H. O. (1972). *Biometrika Tables for Statisticians*, Vol. 1. Cambridge University Press.

Seber, G. A. F. (1977). *Linear Regression Analysis*. John Wiley & Sons.

Williams, E. J. (1959). *Regression Analysis*. John Wiley & Sons.

Chapter 2

MULTIPLE LINEAR REGRESSION

2.1 Introduction

2.1.1 Problems for multiple linear regression analysis

An investigator may, for a variety of reasons, be interested in studying the relationship between rainfall and run-off in a particular area. Given rainfall and run-off records, he would probably find linear regression methods helpful in achieving his objectives. However, it would be foolish to suppose that, given information on rainfall only, he could hope to predict accurately the resultant run-off. Many other factors, some quantifiable, will influence the run-off in a particular area. For instance, rainfall intensity and evaporation may both influence the resulting run-off.

Thus, a realistic data base would not just consist of run-off and rainfall readings only; it would consist of readings on run-off (called the dependent variable) and readings on as many features which are liable to influence run-off (called the independent variables) as it is sensible to gather. It is to this type of data base that the technique of multiple linear regression analysis may be applied with profit. Using multiple linear regression, it may be possible to achieve objectives similar to those outlined in the sequence (a)–(g) given in subsection 1.1.1 where, instead of only rainfall, we have a whole collection of independent variables. Once we progress from studying how one or two variables influence a third, graphical techniques and visual assessment become more difficult and we have to rely much more on mathematical models. However, this does not mean that the outcome of a multiple regression analysis cannot be questioned or assessed. Applied common sense is even more vital in checking for numerical blunders, invalid assumptions, etc. when interpreting the outcome of a multiple regression analysis or considering unexpected features of the data.

2.1.2 Assumptions made in multiple linear regression

Multiple linear regression applies to problems in which records have been kept of one variable, y, the dependent variable, and several other variables

$x_1, x_2, \ldots, x_k$, the independent variables, and in which the objective requires the relationship between the variable y and the variables $x_1, x_2, \ldots, x_k$ to be investigated. For any such record, the specific mathematical relationship (model) assumed is

$$y = a + b_1x_1 + b_2x_2 + \cdots + b_kx_k + e \tag{32}$$

where $a, b_1, b_2, \ldots, b_k$ are constants and e is a variable. Thus, it is assumed that y is linearly related to each of the independent variables and that each independent variable has an additive effect on y. Therefore, at this stage, we are assuming that $x_1, x_2, \ldots, x_k$ do not interact amongst themselves in their effect on y. The variable e serves the same purpose as in the simple linear model described in subsection 1.1.2 and identical assumptions are made on e in multiple linear regression. Thus, under repeated identical conditions (that is, when values of $x_1, x_2, \ldots, x_k$ are kept constant), we expect the arithmetic mean of values of e to be zero and we expect the variance of values of e to be the same, whatever the constant values of $x_1, x_2, \ldots, x_k$.

To carry out tests of significance or to establish confidence intervals, we will need to assume that these values of e form a normal distribution and that all values of e are independent.

2.1.3 Interpretation of the assumptions

The problem of deciding which is y and which is x is more well defined in the multiple regression situation. Usually, we will want to assess the combined effect of several variables on a single variable. This may be to predict y when we know $x_1, x_2, \ldots, x_k$ or to decide which of $x_1, x_2, \ldots, x_k$ do, in fact, influence y, or we may simply want to summarise the data.

It is probably only in the latter case that there might be some doubt as to the identity of y. The type of relationship being estimated again assumes that $x_1, x_2, \ldots, x_k$ are known or error free, filling just the same role as x in subsection 1.1.3. Indeed, if it proves impossible to decide which is y amongst the variables measured, then this may indicate that multiple regression analysis is inappropriate and that some other type of correlation analysis, or principal components analysis, would be more suitable for the problem.

The assumptions about the variable e cannot be seen easily in terms of a graph, mainly because the model (32) is a hyperplane in $(k+1)$ dimensional space. However, we can use the interpretation of the simple linear model given in subsection 1.1.3 to explain this more complex situation. If we interpret $a + b_1x_1 + b_2x_2 + \cdots + b_kx_k$ as being the value of y that we expect to observe, given the conditions or situation defined by the values of $x_1, x_2, \ldots, x_k$, and if we interpret $a + b_1x_1 + b_2x_2 + \cdots + b_kx_k + e$ as being the value of y that we actually observe, then the value of e, the difference between what we actually observe and what we expect to observe, again represents the error or inexplicable variation in y.

Thus, if we knew the values of a, b_1, $b_2, \ldots, b_k$ and, hence, we could plot a graph of observed y against $y_{\text{ideal}} = a + b_1x_1 + \cdots + b_kx_k$ for each record, then we would have the situation illustrated in Figure 8.

The vertical displacement of each point from the 45° line represents the value

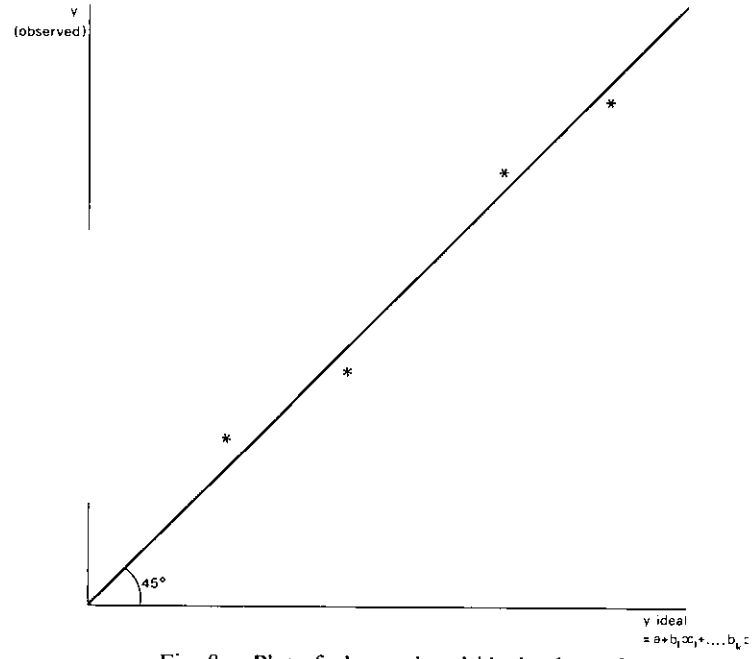


Fig. 8. Plot of observed and ideal values of y.

of e. If we were able to observe repeatedly values of y all with the same y_{ideal}, then we would obtain a vertical array of points and there should be an equal spread on either side of the line. Furthermore, if we were to repeat this procedure at a different value of y_{ideal}, then we should obtain a similar spread of points (they should be neither more nor less widely scattered). Also, these points should form a normal distribution centred on the line.

The assumption of independence has the same interpretation as in subsection 1.1.3.

2.1.4 What can be achieved by using multiple linear regression?

The quick answer is 'everything that was achieved using simple linear regression and a bit more'. Estimates of $a, b_1, b_2, \ldots, b_k$ may be derived, together with standard errors and confidence intervals. However, in multiple linear regression, there is far more scope for tests of significance and far more need for them.

Typically, for a variable x_i, we will be able to decide the following:

(1) Whether x_i has an influence on y.
(2) Whether, after allowing for the influence that other specified variables have on y, the variable x_i still gives some futher explanation of the way in which y varies.

As an example of this, let us suppose that

$$y = \text{run-off}$$
$$x_1 = \text{rainfall}$$
$$x_2 = \text{duration of rainfall}$$

Furthermore, suppose that, for the area being studied, when it rains, it rains at a constant rate. Then, we would have the relationship $x_1 = kx_2$ where k is a constant.

We would discover from our tests of significance that rainfall and duration of rainfall both influence run-off. However, when we know what the rainfall has been, the duration of the rainfall will tell us nothing further about run-off, i.e. if x_1 is known, then x_2 is redundant. Realistic practical problems are rarely this distinct, but we do have the potential to make this type of investigation in multiple regression analysis.

Having summarised our data in terms of estimates of $a, b_1, b_2, \ldots, b_k$, we may compare these estimates with similar estimates from other sets of data so as to assess the similarity of the sets of data in terms of their relationship between y and $x_1, x_2, \ldots, x_k$.

By substituting our estimates of $a, b_1, b_2, \ldots, b_k$ into model (32) (and disregarding e), we may predict y for specified values of $x_1, x_2, \ldots, x_k$. Having predicted y values at observed values of $x_1, x_2, \ldots, x_k$, we may form the 'residuals' (the differences between the predicted y values and the observed y values) just as for the simple linear model and for similar reasons.

2.2 The Basic Method

2.2.1 Fitting the model

The basic unit of data for this model will no longer be a pair of values of y and x, as in subsection 1.2.1, but $k+1$ numbers corresponding to values of $y, x_1, x_2, \ldots, x_k$. Hence, the whole data set will consist of n such basic units and will be denoted by $(y_1, x_{11}, x_{21}, \ldots, x_{k1})$, $(y_2, x_{12}, x_{22}, \ldots, x_{k2}), \ldots$ $(y_n, x_{1n}, x_{2n}, \ldots, x_{kn})$.

The model (32) would imply the relationship

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki} + e_i \qquad (\text{for } i = 1, 2, \ldots, n) \qquad (33)$$

for this set of data. However, just as it proved useful in simple linear regression to rewrite the model into the form of model (25), there are some advantages in rewriting the model (33) into the form

$$y_i = \alpha + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \cdots + \beta_k(x_{ki} - \bar{x}_k) + e_i \qquad (34)$$

where

$$\bar{x}_1 = \frac{1}{n}\sum_{i=1}^{n} x_{1i} \qquad \bar{x}_2 = \frac{1}{n}\sum_{i=1}^{n} x_{2i} \qquad \text{etc.}$$

This is the form of model usually encountered in texts on multiple regression. By comparing model (34) with model (33), we see that $b_j = \beta_j$ (for $j = 1, 2, \ldots, k$) and $a = \alpha - \beta_1\bar{x}_1 - \beta_2\bar{x}_2 - \cdots - \beta_k\bar{x}_k$.

Figure 9 illustrates how the model and data might look if plotted with $k = 2$ and $n = 3$. The shaded area represents the plane $y = a + b_1x_1 + b_2x_2$, drawn for $y, x_1, x_2 > 0$, and the large dots indicate the position of the points, (y_1, x_{11}, x_{21}), (y_2, x_{12}, x_{22}), (y_3, x_{13}, x_{23}). Hence, the lengths e_1, e_2, e_3 represent the vertical distance from each of these points to the plane.

Our problem is that, although we know the position of the points, we do not know the position of the plane; in other words, we do not know a, b_1 and b_2. The method of least squares would lead us to choose those values of a, b_1 and b_2 which minimise

$$S^2 = \sum_{i=1}^{3} e_i^2 = \sum_{i=1}^{3} (y_i - a - b_1 x_{1i} - b_2 x_{2i})^2$$

Once again, we are attempting to make the vertical discrepancy of the points from the plane (regardless of sign) as small as possible.

In the general case of n observations and k variables, we will want to choose $a, b_1, b_2, \ldots, b_k$ (or $\alpha, \beta_1, \beta_2, \ldots, \beta_k$) to minimise

$$S^2 = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - \cdots - b_k x_{ki})^2$$

$$= \sum_{i=1}^{n} (y_i - \alpha - \beta_1 (x_{1i} - \bar{x}_1) - \beta_2 (x_{2i} - \bar{x}_2) - \cdots - \beta_k (x_{ki} - \bar{x}_k))^2$$

Solving $\partial S^2/\partial \alpha = 0, \partial S^2/\partial \beta_1 = 0, \ldots, \partial S^2/\partial \beta_k = 0$ gives the following equations for the values of $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ which minimise S^2 (denoted by $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$):

$$\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

$$\sum_{i=1}^{n} (y_i - \bar{y})(x_{1i} - \bar{x}_1) = \hat{\beta}_1 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 + \hat{\beta}_2 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

$$\cdot + \hat{\beta}_k \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{ki} - \bar{x}_k)$$

$$\sum_{i=1}^{n} (y_i - \bar{y})(x_{2i} - \bar{x}_2) = \hat{\beta}_1 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) + \hat{\beta}_2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)^2$$

$$\cdot + \hat{\beta}_k \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(x_{ki} - \bar{x}_k)$$

$$\vdots$$

$$\sum_{i=1}^{n} (y_i - \bar{y})(x_{ki} - \bar{x}_k) = \hat{\beta}_1 \sum_{i=1}^{n} (x_{1i} - \bar{x}_1)(x_{ki} - \bar{x}_k) + \hat{\beta}_2 \sum_{i=1}^{n} (x_{2i} - \bar{x}_2)(x_{ki} - \bar{x}_k)$$

$$\cdot + \hat{\beta}_k \sum_{i=1}^{n} (x_{ki} - \bar{x}_k)^2$$

Simplifications may be made to the presentation of this information by:

(1) A representation using matrices.

(2) Use of the notation

$$S_{x_j y} = \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)(y_i - \bar{y}) \quad (\text{for } j = 1, 2, \ldots, k)$$

$$S_{x_j x_l} = \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)(x_{li} - \bar{x}_l) \quad (\text{for } j = 1, 2, \ldots, k \text{ and } l = 1, 2, \ldots, k)$$

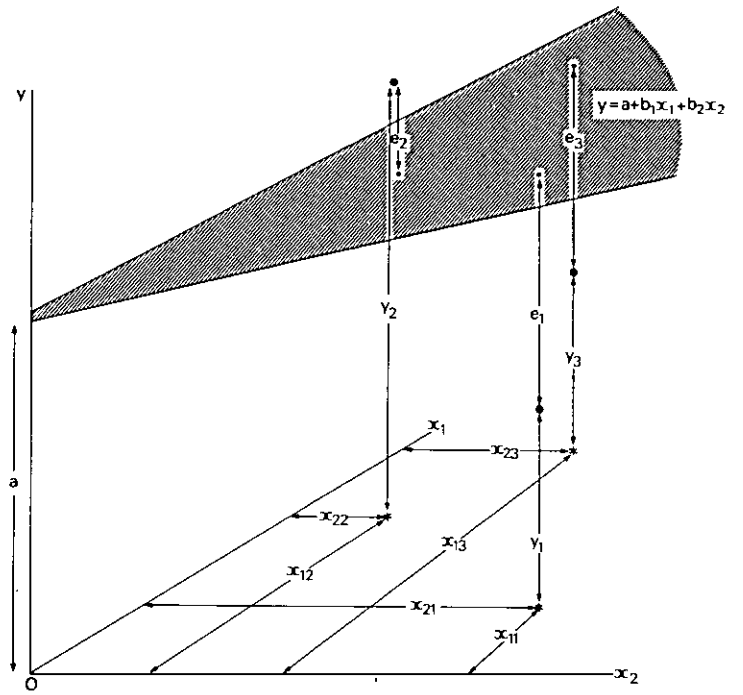


Fig. 9. Regression plane and data points.

The equations may now be condensed into

$$\begin{bmatrix} S_{x_1y} \\ S_{x_2y} \\ \vdots \\ S_{x_ky} \end{bmatrix} = \begin{bmatrix} S_{x_1x_1} & S_{x_1x_2} & & S_{x_1x_k} \\ S_{x_1x_2} & S_{x_2x_2} & & S_{x_2x_k} \\ \vdots & & & \vdots \\ S_{x_1x_k} & & & S_{x_kx_k} \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}.$$

$$\mathbf{S}_{xy} = \mathbf{S}_{xx}\hat{\boldsymbol{\beta}} \tag{35}$$

Thus, the estimates $\hat{\beta}, \hat{\beta}_2, \ldots, \hat{\beta}_k$ are given by

$$\hat{\boldsymbol{\beta}} = \mathbf{S}_{xx}^{-1}\mathbf{S}_{xy} \tag{36}$$

which, together with $\hat{\alpha} = \bar{y}$, gives us estimates of all the parameters in model (34). ($\mathbf{S}_{xx}^{-1}$ is the matrix inverse of $\mathbf{S}_{xx}$.)

2.2.2 Estimates and their precision

If we assume that the variance of e_i is σ^2 (for $i = 1, 2, \ldots, n$), then it follows that

$$\text{Var}(\hat{\alpha}) = \frac{\sigma^2}{n}$$

However, as the estimates of $\beta_1, \beta_2, \ldots, \beta_k$ are not mutually independent, there are k^2 different variances and covariances associated with them. These are

conveniently displayed in a matrix, called the variance covariance matrix, which is denoted by

$$\begin{bmatrix} \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_3) & \text{Cov}(\hat{\beta}_1, \hat{\beta}_k) \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_2) & \text{Var}(\hat{\beta}_2) & \text{Cov}(\hat{\beta}_2, \hat{\beta}_3) & \text{Cov}(\hat{\beta}_2, \hat{\beta}_k) \\ \vdots & \vdots & & \vdots \\ \text{Cov}(\hat{\beta}_1, \hat{\beta}_k) & \text{Cov}(\hat{\beta}_2, \hat{\beta}_k) & & \text{Var}(\hat{\beta}_k) \end{bmatrix}$$

and referred to as $\mathbf{V}_{\hat{\beta}}$. It may be shown that

$$\mathbf{V}_{\hat{\beta}} = \sigma^2 \mathbf{S}_{xx}^{-1} \tag{37}$$

Equation (37) requires knowledge of σ^2 which, as in simple linear regression, will be unknown. However, also as in simple linear regression, we may estimate e_i by the ith residual.

$$\hat{e}_i = y_i - \hat{\alpha} - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2) - \cdot\ \cdot - \hat{\beta}_k(x_{ki} - \bar{x}_k) \tag{38}$$

It may easily be shown that $\sum_{i=1}^{n} \hat{e}_i = 0$ and, consequently, the arithmetic mean of the residuals is always zero. Hence, we may again base our estimate of σ^2 on $R = \sum_{i=1}^{n} \hat{e}_i^2$, called the residual sum of squares. However, in this case, the appropriate divisor will be $n - k - 1$ as $k + 1$ degrees of freedom have been 'lost' in estimating $\alpha, \beta_1, \beta_2, \ldots, \beta_k$. Hence,

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} (y_i - \bar{y} - \hat{\beta}_1(x_{1i} - \bar{x}_1) - \hat{\beta}_2(x_{2i} - \bar{x}_2) - \cdots - \hat{\beta}_k(x_{ki} - \bar{x}_k))^2$$

$$= \frac{1}{n-k-1} [S_{yy} - \hat{\beta}_1 S_{x_1 y} - \hat{\beta}_2 S_{x_2 y} - \cdots - \hat{\beta}_k S_{x_k y}] \tag{39}$$

where $S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$.

To calculate S_{yy}, $S_{x_j y}$ and $S_{x_j x_l}$, it may be easier to use the expressions

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i \right)^2 \Big/ n \\ S_{x_j y} &= \sum_{i=1}^{n} y_i x_{ji} - \left(\sum_{i=1}^{n} y_i \right) \left(\sum_{i=1}^{n} x_{ji} \right) \Big/ n \\ S_{x_j x_l} &= \sum_{i=1}^{n} x_{ji} x_{li} - \left(\sum_{i=1}^{n} x_{ji} \right) \left(\sum_{i=1}^{n} x_{li} \right) \Big/ n \end{aligned} \tag{40}$$

although remarks made in Chapter 1, and in particular in subsection 1.2.1, concerning numerical accuracy, are equally pertinent in this context.

Thus, we now have sufficient information to estimate the variances and covariances of $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ as well as the variance of $\hat{\alpha}$. It is usually more informative to study the correlations than the covariances and these may easily be derived by

$$\text{Correlation between } \hat{\beta}_j \text{ and } \hat{\beta}_l = \frac{\text{Cov}(\hat{\beta}_j, \hat{\beta}_l)}{\sqrt{\text{Var}(\hat{\beta}_j)\,\text{Var}(\hat{\beta}_l)}} \tag{41}$$

By assuming that $e_i \sim N(0, \sigma^2)$ (for $i = 1, 2, \ldots, n$), we may derive confidence intervals for the parameters $\alpha, \beta_1, \ldots, \beta_k$. In multiple regression, the residual sum of squares follows $\sigma^2 \chi^2_{n-k-1}$. Consequently, $(\hat{\alpha} - \alpha)/\sqrt{\hat{\sigma}^2/n} \sim t_{n-k-1}$ and $(\hat{\beta}_j - \beta_j)/\sqrt{\text{Estimate of Var}\,(\hat{\beta}_j)} \sim t_{n-k-1}$ where 'Estimate of Var $(\hat{\beta}_j)$' is obtained from the appropriate element of $\mathbf{V}_{\hat{\beta}}$ (given in equation (37)) after substituting $\hat{\sigma}^2$ (given in equation (39)) for σ^2.

Hence, a $100(1-\alpha)\%$ confidence interval for α is

$$\hat{\alpha} \pm t(n-k-1, 1-\alpha/2)\sqrt{\hat{\sigma}^2/n} \tag{42}$$

and individual $100(1-\alpha)\%$ confidence intervals for $\beta_1, \beta_2, \ldots, \beta_k$ are given by

$$\hat{\beta}_j \pm t(n-k-1, 1-\alpha/2)\sqrt{\text{Estimate of Var}\,(\hat{\beta}_j)} \tag{43}$$

Since $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ will almost always be correlated, there is some danger in using these separate confidence intervals, particularly when the objective is to find some 'joint' confidence region, for example, for β_1 and β_2. An assessment of the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$ (see equation (41)) would be advisable and if it proves high, then it may be as well to consider using the joint confidence region. The $100(1-\alpha)\%$ confidence region for $\beta_1, \beta_2, \ldots, \beta_k$ is defined by those values of $\beta_1, \beta_2, \ldots, \beta_k$ which satisfy

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{S}_{xx} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leq (k+1)\hat{\sigma}^2 F(k+1, n-k-1, 1-\alpha)$$

where

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad (\mathbf{A}' \text{ denotes the transpose of } \mathbf{A})$$

2.2.3 Prediction

Having estimated the unknowns in model (34), we are in a position to predict a value of y from knowledge of $x_1, x_2, \ldots, x_k$. If we have values for $x_1, x_2, \ldots, x_k$ and these are denoted by $x_{1p}, x_{2p}, \ldots, x_{kp}$, then we may predict y using

$$\hat{y} = \hat{\alpha} + \hat{\beta}_1(x_{1p} - \bar{x}_1) + \hat{\beta}_2(x_{2p} - \bar{x}_2) + \cdots + \hat{\beta}_k(x_{kp} - \bar{x}_k) \tag{44}$$

To determine the variance of $\hat{y}$, let us define $\mathbf{X}_p = [(x_{1p} - \bar{x}_1), (x_{2p} - \bar{x}_2), \ldots (x_{kp} - \bar{x}_k)]$.

Then, the variance of $\hat{y}$ is given by

$$\text{Var}\,(\hat{y}) = \sigma^2 \mathbf{X}_p S_{xx}^{-1} \mathbf{X}_p' \tag{45}$$

The problem of considering what we are actually attempting to predict with $\hat{y}$ has been discussed for simple linear regression. The distinctions made there are equally relevant in the context of multiple regression. The variance given in equation (45) only represents our uncertainty about $\alpha, \beta_1, \ldots, \beta_k$ and its use is only appropriate when we are trying to predict the mean value of y. However,

when we are trying to predict the outcome of a single reading, the variance given in equation (45) should be increased to

$$\sigma^2 + \sigma^2 \mathbf{X}_p \mathbf{S}_{xx}^{-1} \mathbf{X}_p'$$

the additional component, σ^2, being for the error of the reading.

Corresponding $100(1-\alpha)\%$ confidence intervals for the mean value or the outcome of a single reading are

$$\hat{y} \pm t(n-k-1, 1-\alpha/2)\sqrt{\hat{\sigma}^2 \mathbf{X}_p \mathbf{S}_{xx}^{-1} \mathbf{X}_p'} \tag{46}$$

and

$$\hat{y} \pm t(n-k-1, 1-\alpha/2)\sqrt{\hat{\sigma}^2 (1 + \mathbf{X}_p \mathbf{S}_{xx}^{-1} \mathbf{X}_p')} \tag{47}$$

respectively.

The confidence region for the hyperplane

$$y = \alpha + \beta_1(x_1 - \bar{x}_1) + \beta_2(x_2 - \bar{x}_2) + \cdots + \beta_k(x_k - \bar{x}_k)$$

does not have any great practical merit, mainly because of the difficulty of visually displaying such a region.

2.3 Significance Tests and the 'Best' Equation

2.3.1 General linear hypothesis

A variety of significance tests are available for studying various features of the model (34). We deal here with these tests in isolation and later on will explain how combinations of such tests may be used, for instance, to decide which is the 'best' equation.

Many tests can be constructed from one basic result which is, somewhat ambiguously, referred to as the general linear hypothesis. A hypothesis about the parameters in model (34) might, for instance, state that β_1, β_3, β_7 and β_9 are all zero (i.e. variables x_1, x_3, x_7 and x_9 are of no importance in model (34)). Thus, a general linear hypothesis might take a form which exactly specifies the values of p of the parameters in model (34).

Imagine modifying equation (34) to take account of the information contained in the hypothesis (in the above example, this would mean omitting variables x_1, x_3, x_7 and x_9 from the equation) and performing the necessary calculations to arrive at the residuals (given by equation (38)) associated with this new (smaller) model. Suppose that the sum of squares of these residuals is formed (denoted by R_H). Then, this quantity will be reasonably close to (but larger than) the residual sum of squares calculated using the full model (denoted by R) whenever the hypothesis is acceptable. In fact, it may be shown that

$$\left(\frac{R_H - R}{p}\right) \Big/ \left(\frac{R}{n-k-1}\right) \sim F_{p,n-k-1} \tag{48}$$

whenever the hypothesis (assumed in calculating R_H) is valid.

Thus, in our example, R_H would be the residual sum of squares obtained by omitting variables x_1, x_3, x_7 and x_9 from equation (34). Acceptance of the

hypothesis $\beta_1 = \beta_3 = \beta_7 = \beta_9 = 0$ whenever $[(R_H - R)/4]/[R/(n-k-1)] < F(4, n-k-1, 1-\alpha)$ would provide a $100\alpha\%$ significance test of this hypothesis.

In its most general form, a 'general linear hypothesis' will impose p functionally independent constraints on the parameters $\alpha, \beta_1, \beta_2, \ldots, \beta_k$. The result given by equation (48) will still apply, even in this very general context. Thus, a hypothesis of the form $\beta_1 = \beta_2$ (which implies $\beta_1 - \beta_2 = 0$) would fit within this framework and might well be informative if x_1 and x_2 were measuring similar quantities. R_H would be calculated by using the model

$$\begin{aligned} y_i &= \alpha + \beta_1(x_{1i} - \bar{x}_1) + \beta_1(x_{2i} - \bar{x}_2) + \beta_3(x_{3i} - \bar{x}_3) + \cdots + \beta_k(x_{ki} - \bar{x}_k) + e_i \\ &= \alpha + \beta_1[(x_{1i} + x_{2i}) - (\bar{x}_1 + \bar{x}_2)] + \beta_3(x_{3i} - \bar{x}_3) + \cdots + \beta_k(x_{ki} - \bar{x}_k) + e_i \\ &= \alpha + \beta_1(u_i - \bar{u}) + \beta_3(x_{3i} - \bar{x}_3) + \cdots + \beta_k(x_{ki} - \bar{x}_k) + e_i \end{aligned}$$

where $u_i = x_{1i} + x_{2i}$ and $\bar{u} = (1/n)\sum_{i=1}^{n} u_i$. For a $100\alpha\%$ significance test, the hypothesis $\beta_1 = \beta_2$ would be accepted whenever $(R_H - R)/[R/(n-k-1)] < F(1, n-k-1, 1-\alpha)$.

2.3.2 Initial significance tests

Having estimated all of the parameters in model (34), the most pertinent question might be 'Is there any evidence of a relationship between the y variable and any of the x variables I have used'? or, in other words, 'Can I predict y with any success, from the x variables I have used?'

This question is equivalent to considering a hypothesis $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ and, if this hypothesis were true, then model (34) would reduce to the simple form $y_i = \alpha + e_i$. The residual sum of squares resulting from such a model would be $R_H = \sum_{i=1}^{n} (y_i - \bar{y})^2$, since $\hat{\alpha} = \bar{y}$.

Then, equation (48) would lead us to accept the hypothesis $\beta_1 = \beta_2 = \cdots = \beta_k = 0$ whenever

$$\left(\frac{R_H - R}{k}\right) \Big/ \left(\frac{R}{n-k-1}\right) < F(k, n-k-1, 1-\alpha)$$

for a $100\alpha\%$ significance test.

This procedure is usually displayed in the form of an analysis of variance table.

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*
Regression	$\sum_{j=1}^{k} S_{x_j y}\hat{\beta}_j = R_H - R$	k	
Residual	$S_{yy} - \sum_{j=1}^{k} S_{x_j y}\hat{\beta}_j = R$	$n-k-1$	$\left(= \dfrac{\text{Sum of squares}}{\text{Degrees of freedom}}\right)$
Total	$\sum_{i=1}^{n} (y_i - \bar{y})^2 = R_H$	$n-1$	

The ratio

$$\frac{\text{Regression mean square}}{\text{Residual mean square}}$$

gives the same test statistic as that previously mentioned for testing the hypothesis $\beta_1 = \beta_2 = \cdots = \beta_k = 0$. The quantity

$$\sqrt{\frac{\text{Regression sum of squares}}{\text{Total sum of squares}}}$$

is called the multiple correlation coefficient between y and $x_1, x_2, \ldots, x_k$ and it takes values between 0 and 1. At one extreme, if all the residuals (38) were zero and, consequently, R was zero (i.e. the model managed to predict all the observations exactly), then the multiple correlation coefficient would be 1. At the other extreme, if the model predicted each observation as being $\bar{y}$ and, therefore, totally disregarded any contribution which might be made by $x_1, x_2, \ldots, x_k$, then R would be equal to R_H and the multiple correlation coefficient would be zero. Thus, the multiple correlation coefficient takes values between 0 and 1, a value near 1 indicating strong association (correlation) between y and $x_1, x_2, \ldots, x_k$ and a value near 0 suggesting little association (correlation) between y and $x_1, x_2, \ldots, x_k$ (or, at least, their observed values).

The next step in our analysis would probably be to enquire into the individual effect of each of the x variables on the y variable. The hypothesis $\beta_1 = 0$ would appear to consider the effect of x_1 on y, but if we test this hypothesis using the results of the general linear hypothesis (Section 2.3.1), then it takes on a special meaning. We will, in fact, be comparing the model

$$y_i = \alpha + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \cdots + \beta_k(x_{ki} - \bar{x}_k) + e_i$$

with the model

$$y_i = \alpha + \beta_2(x_{2i} - \bar{x}_2) + \cdots + \beta_k(x_{ki} - \bar{x}_k) + e_i$$

and, consequently, our test will be telling us how much better our model would be by including the term $\beta_1(x_{1i} - \bar{x}_1)$ as well as all the other terms already in the model. In other words, it will give us some idea of the additional information that x_1 can provide about y over and above that already provided by $x_2, x_3, \ldots, x_k$.

Thus, if we calculate the residual sum of squares R_H omitting the variable x_1 from our model (34), then a $100\alpha\%$ significance test for $\beta_1 = 0$ would accept $\beta_1 = 0$ whenever $(R_H - R)/[R/(n-k-1)] < F(1, n-k-1, 1-\alpha)$.

However, we also know, from Subsection 2.2.2, that

$$\frac{\hat{\beta}_1 - \hat{\beta}_1}{\sqrt{\text{Estimate of Var}(\hat{\beta}_1)}} \sim t_{n-k-1}$$

and, hence, a $100\alpha\%$ significance test for $\beta_1 = 0$ would accept $\beta_1 = 0$ whenever

$$\left| \frac{\hat{\beta}_1}{\sqrt{\text{Estimate of Var}(\hat{\beta}_1)}} \right| < t(n-k-1, 1-\alpha/2)$$

which is equivalent to checking whether $\beta_1 = 0$ falls in the confidence interval (43). (See equation (43) for the source of $\sqrt{\text{Estimate of Var}(\hat{\beta}_1)}$.)

These two test procedures are, in fact, identical; the square of the latter test statistic gives the former and it is well known that $F(1, n-k-1) = t^2_{n-k-1}$. Therefore, it is important to note that both procedures are assessing the additional information supplied by x_1 over and above that already provided by $x_2, \ldots, x_k$. Hence, in a situation in which x_1 has a strong effect on y, but it is also correlated with, say, x_2 and x_5, which in turn have a strong effect on y, it is quite likely that we would accept $\beta_1 = 0$ from the preceding test. This would simply be because x_1 provided no additional information about y after the information provided by x_2 and x_5 had been taken into account. To assess the effect of x_1 alone on y, a simple linear regression of y on x_1, as described in Chapter 1, would be appropriate.

2.3.3 Selection of variables—the 'best' equation

In the previous subsection, an example was given where x_1 supplied no further information above that already provided by x_2 and x_5. However, we might also have concluded that x_2 supplied no further information above that already provided by x_1 and x_5. This would clearly be true in the trivial case of $x_1 = x_2$.

What should we do? Should we either include x_1 and exclude x_2 or vice versa, or should we include both x_1 and x_2? In what order should we start to assess the relative importance of our variables and, furthermore, is that order crucial? As has previously been the case, our answer partly depends on what we want to find out and on what purpose we have in mind for the regression equation.

If, in fact, the requirement is to find which of $x_1, x_2, \ldots, x_k$ are associated with y, then a simple linear regression analysis of y on each of the x variables in turn will provide the answer. However, if the requirement is to predict y from available information in the form of $x_1, x_2, \ldots, x_k$, none of which are particularly costly to observe, then there is little harm in leaving most of the variables in the model. Some reduction might be made by using the technique described at the end of the subsection on initial significance tests (2.3.2), but whenever a circular conflict arises amongst a group of x variables, all should be left in the model. Occasionally it may be appropriate to quote several different models. When the requirement is to 'understand' what influences y or to predict y using only the 'significant' x variables, some more elaborate methods of selecting variables must be used. Several such methods are available and some of these are outlined in the following subsections.

2.3.4 All possible regressions

If we only require an idea of the 'best' regression equation and unlimited computer time is available, then computing all 2^k possible regression equations (either including x_1 or not, either including x_2 or not, etc.) will give a good basis from which to decide. If the multiple correlation coefficient is calculated for each regression and the resulting 2^k such coefficients are arranged in order of magnitude, then examination of those regression equations that are associated

with high multiple correlation values should give an idea of the most important factors.

2.3.5 Forward selection

For moderately large k, the 'all possible regressions' method is extremely expensive in computer time. As a cheaper alternative, 'forward selection' aims at ignoring equations which are likely to give small multiple correlations.

Step 1 is to perform the k simple linear regressions of y on x_1, y on $x_2, \ldots, y$ on x_k, as described in Subsection 1.2.2. The test statistic for $b = 0$,

$$\frac{|\hat{b}|}{\sqrt{\hat{\sigma}^2 \Big/ \left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)}}$$

is calculated for each regression and the x variable which gives the largest of these values is selected for inclusion in the 'best' equation, provided that its test statistic is significant at a specified level. As an example, suppose that this variable is x_k.

Step 2 is to compute the partial correlation coefficients between y and each of the variables not yet included in the 'best' equation, conditional on the variable already included in the best equation. In our example, these would be calculated by

$$r_{yx_j \cdot x_k} = \frac{r_{yx_j} - r_{yx_k} r_{x_j x_k}}{\sqrt{(1 - r_{yx_k}^2)(1 - r_{x_j x_k}^2)}} \tag{49}$$

where x_k is the variable included in the 'best' equation. The ordinary product moment correlation coefficients r_{yx_k}, $r_{x_j x_k}$ and r_{yx_j} are

$$\frac{S_{x_k y}}{\sqrt{S_{yy} S_{x_k x_k}}} \qquad \frac{S_{x_j x_k}}{\sqrt{S_{x_j x_j} S_{x_k x_k}}} \quad \text{and} \quad \frac{S_{x_j y}}{\sqrt{S_{yy} S_{x_j x_j}}}$$

respectively. An interpretation of the partial correlation coefficient $r_{yx_j \cdot x_k}$ is that it measures the correlation between y and x_j after both y and x_j have been corrected for the effect that x_k may have had on them. Thus, $r_{yx_j \cdot x_k}$ will give an indication of the further contribution which x_j would make in predicting y if it was included in the 'best' equation together with x_k. The variable with the largest absolute value for its partial correlation is selected for inclusion in the 'best' equation. For our example, suppose that this is variable x_{k-1}.

Step 3 is to fit the 'best' equation as it is so far and, then, to test the joint significance of all the variables included in the equation and the individual significance of the most recently included variable. If it is concluded from the first test that the model is of some value and, furthermore, from the second test, that the addition of the most recently included variable is of value in the model, then the procedure advances to step 4. However, if, from the second test, it is concluded that the most recently included variable is not of value in the model, then the procedure would stop here and the 'best' equation would be taken to be the present equation omitting the most recently included variable.

Thus, in our example, the model would be

$$y_i = \alpha + \beta_{k-1}(x_{k-1,i} - \bar{x}_{k-1}) + \beta_k(x_{ki} - \bar{x}_k) + e_i$$

and we would use the procedure described in Subsection 2.2.2 to estimate α, β_{k-1} and β_k. Then, using initial significance tests (2.3.2), we would try separately the hypotheses $\beta_{k-1} = \beta_k = 0$ and $\beta_{k-1} = 0$. Rejection of the first hypothesis would suggest that the model was of some value and rejection of the second hypothesis, as well as the first, would suggest that the addition of x_{k-1} was of value in the model (as well as x_k). However, acceptance of the second hypothesis would lead us to stop at this point and to state that the 'best' equation had been obtained by using x_k alone, i.e. the 'best' model had been given by the simple linear regression of y on x_k.

Step 4, which is similar to step 2, is to calculate new partial correlation coefficients between y and each of the variables not yet included in the best equation, conditional on the variables already included in the best equation. In general, to calculate the partial correlation coefficients of y and $x_1, x_2, \ldots, x_p$, conditional on $x_{p+1}, x_{p+2}, \ldots, x_k$, the matrix

$$\left[\begin{array}{cccc:cc} S_{yy} & S_{x_1y} & \cdots & S_{x_py} & S_{x_{p+1}y} & S_{x_ky} \\ S_{x_1y} & S_{x_1x_1} & \cdots & S_{x_1x_p} & S_{x_1x_{p+1}} & S_{x_1x_k} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ S_{x_py} & S_{x_px_1} & \cdots & S_{x_px_p} & S_{x_px_{p+1}} & S_{x_px_k} \\ \hdashline S_{x_{p+1}y} & S_{x_{p+1}x_1} & \cdots & S_{x_{p+1}x_p} & S_{x_{p+1}x_{p+1}} & S_{x_{p+1}x_k} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ S_{x_ky} & S_{x_kx_1} & \cdots & S_{x_kx_p} & S_{x_kx_{p+1}} & S_{x_kx_k} \end{array}\right]$$

is partitioned as shown and the four regions are denoted by

$$\left[\begin{array}{c:c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hdashline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array}\right]$$

Note that the variables not yet included in the model are used to form $\boldsymbol{\Sigma}_{11}$ and those already included in the model are used to form $\boldsymbol{\Sigma}_{22}$. The corrected sums of cross products between these two sets of variables occupy $\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}'$. Then, $\boldsymbol{\Sigma}_{11.2}$ is computed by

$$\boldsymbol{\Sigma}_{11.2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$$

This will be a symmetric $(p+1) \times (p+1)$ matrix whose elements are denoted, for the sake of brevity, by

$$\begin{bmatrix} a_{00} & a_{01} & a_{02} & \cdots & a_{0p} \\ a_{01} & a_{11} & a_{12} & & a_{1p} \\ a_{02} & a_{12} & a_{22} & & a_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{0p} & a_{1p} & a_{2p} & & a_{pp} \end{bmatrix}$$

The partial correlation coefficient of y and x_i, conditional on x_{p+1}, $x_{p+2}, \ldots, x_k$, is denoted by $r_{yx_i \cdot x_{p+1}, x_{p+2}, \ldots, x_k}$ and is given by

$$r_{yx_i \cdot x_{p+1}, x_{p+2}, \ldots, x_k} = \frac{[a_{0i}]}{\sqrt{a_{00} a_{ii}}}$$

It represents the correlation between y and x_i after both have been corrected for the effects of $x_{p+1}, x_{p+2}, \ldots, x_k$. The variable with the largest partial correlation is now included in the 'best' equation.

Thus, in our example, we would calculate new partial correlation coefficients between y and each of $x_1, x_2, \ldots, x_{k-2}$, conditional on x_{k-1} and x_k. The variable with the largest absolute value for its partial correlation coefficient would then be included in the 'best' equation. Suppose that this variable is x_{k-2}.

The next stage in the procedure is to go back to Step 3 with the additional variable included in the 'best' equation. In our example, this will mean that we would have to estimate $\beta_{k-2}, \beta_{k-1}, \beta_k$ and α and then test separately the hypotheses $\beta_{k-2} = \beta_{k-1} = \beta_k = 0$ and $\beta_{k-2} = 0$. If these tests suggested that x_{k-2} was of value in the model (as well as x_{k-1} and x_k), then we would again proceed to Step 4. Otherwise, we would stop with the 'best' model using only the variables x_k and x_{k-1}.

This procedure will cycle around Steps 3 and 4 until eventually it stops in Step 3 with a 'best' set of variables. The model estimated at the previous execution of Step 3 will be the 'best' equation.

2.3.6 Backward selection

This method is simpler to explain as it does not require calculation of partial correlation coefficients.

Step 1 is to fit the full regression equation with all variables included.

Step 2 is to perform the initial significance tests (Subsection 2.3.2) for each variable, i.e. to test $\beta_1 = 0$, then $\beta_2 = 0$, then $\beta_3 = 0$, etc. by computing

$$\left| \frac{\hat{\beta}_i}{\sqrt{\text{Estimate of Var}\,(\hat{\beta}_i)}} \right| \qquad (\text{for } i = 1, 2, \ldots, k)$$

If the smallest of these k quantities is less than $t(n-k-1, 1-\alpha/2)$, then the relevant variable is omitted from the equation. If not, then the equation as it stands is used as the 'best' equation. (It is, of course, necessary to fix on a value of α, preferably before starting the whole procedure.)

If a variable has been omitted in Step 2, then the procedure is to return to Step 1 with the variable omitted. The procedure is then to cycle around Steps 1 and 2 until a 'best' equation is eventually reached in Step 2.

2.3.7 Stepwise regression

Forward selection suffers from never being able to drop a variable once it has been included into the 'best' equation. Backward selection starts with all the variables in the equation and, consequently, is susceptible to rounding errors

which may arise from inverting large matrices. A compromise between these two methods would be one which performed forward selection with a 'backward' look at each stage. Stepwise regression is such a method.

Stepwise regression follows the sequence of steps outlined in forward selection except that, in Step 3, each of the regression coefficients of each of the variables included so far in the 'best' equation is tested and, for those not significantly different from zero, the corresponding variables are dropped from the 'best' equation. Thus, in our example, in the first pass through Step 3, we would not only test $\beta_{k-1} = 0$, but $\beta_k = 0$ as well. If either of these hypotheses were accepted, then the corresponding variable would be dropped from the 'best' equation.

2.4 Extensions to the Basic Method

2.4.1 Fitting and comparing several regression lines

The comparisons suggested here are a direct extension of those discussed in Subsection 1.3.2, the difference here being that, instead of having readings on only one x variable, we have readings on k x variables. Thus, if there are n sites from which we have collected data, then the data from site i will consist of r_i sets of $(k+1)$ values and will be denoted by $(y_{ij}, x_{1ij}, x_{2ij}, \ldots, x_{kij})$ (for $j = 1, 2, \ldots, r_i$).

Our model for the data from site i will be

$$y_{ij} = \alpha_i + \beta_{1i}(x_{1ij} - \bar{x}_{1i.}) + \beta_{2i}(x_{2ij} - \bar{x}_{2i.}) + \cdots + \beta_{ki}(x_{kij} - \bar{x}_{ki.}) + e_{ij} \quad (50)$$

Estimates of $\alpha_i, \beta_{1i}, \ldots, \beta_{ki}$ may be derived by applying the basic method described in Section 2.2 to each site's data in turn. The sums of squares and cross products thus defined (Subsection 2.2.1) receive an extra suffix to indicate that they relate to the data from site i, i.e. they are denoted by $S^i_{x_j y}$ and $S^i_{x_j x_l}$. A separate estimate for σ^2 (see equation (39)) will be available from each site, namely

$$\hat{\sigma}_i^2 = \frac{1}{r_i - k - 1}\left[S^i_{yy} - \sum_{l=1}^{k} \hat{\beta}_{li} S^i_{x_l y}\right] \quad (51)$$

for site i.

To compare estimates of σ^2 derived from the different sites, we may formally test $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$ by calculating the test statistic

$$M = (N - (k+1)n) \log_e \left[\frac{\sum_{i=1}^{n} (r_i - k - 1)\hat{\sigma}_i^2}{N - (k+1)n}\right] - \sum_{i=1}^{n} (r_i - k - 1) \log_e \hat{\sigma}_i^2$$

where $N = \sum_{i=1}^{n} r_i$. As in the analogous test presented earlier in Subsection 1.3.2, a $100\alpha\%$ significance test on $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_n^2$ would be to accept this hypothesis whenever

$$M < \chi^2(n-1, 1-\alpha)$$

If we are able to accept this hypothesis, then we may safely proceed to tests on

the similarity of the regression lines from the n sites. For this, we need some further notation. For site i, we have already defined

$$S^i_{x_j y} = \sum_{m=1}^{r_i} (x_{jim} - \bar{x}_{ji.})(y_{im} - \bar{y}_{i.})$$

$$S^i_{x_j x_l} = \sum_{m=1}^{r_i} (x_{jim} - \bar{x}_{ji.})(x_{lim} - \bar{x}_{li.})$$

$$S^i_{yy} = \sum_{m=1}^{r_i} (y_{im} - \bar{y}_{i.})^2$$

where

$$\bar{y}_{i.} = \frac{1}{r_i} \sum_{m=1}^{r_i} y_{im} \qquad \text{and} \qquad \bar{x}_{ji.} = \frac{1}{r_i} \sum_{m=1}^{r_i} x_{jim}$$

In addition, for the data combined over all the sites, we will define

$$S^{\circ}_{x_j y} = \sum_{i=1}^{n} \sum_{m=1}^{r_i} (x_{jim} - \bar{x}_{j..})(y_{im} - \bar{y}_{..})$$

$$S^{\circ}_{x_j x_l} = \sum_{i=1}^{n} \sum_{m=1}^{r_i} (x_{jim} - \bar{x}_{j..})(x_{lim} - \bar{x}_{l..})$$

$$S^{\circ}_{yy} = \sum_{i=1}^{n} \sum_{m=1}^{r_i} (y_{im} - \bar{y}_{..})^2$$

where

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{n} \sum_{m=1}^{r_i} y_{im} \qquad \text{and} \qquad \bar{x}_{j..} = \frac{1}{N} \sum_{i=1}^{n} \sum_{m=1}^{r_i} x_{jim}$$

Then, $\hat{\beta}^{\circ}_1, \hat{\beta}^{\circ}_2, \ldots, \hat{\beta}^{\circ}_k$ will be given by

$$\begin{bmatrix} \hat{\beta}^{\circ}_1 \\ \hat{\beta}^{\circ}_2 \\ \vdots \\ \hat{\beta}^{\circ}_k \end{bmatrix} = \begin{bmatrix} S^{\circ}_{x_1 x_1} & S^{\circ}_{x_1 x_2} & & S^{\circ}_{x_1 x_k} \\ S^{\circ}_{x_2 x_1} & S^{\circ}_{x_2 x_2} & & S^{\circ}_{x_2 x_k} \\ \cdot & \cdot & & \cdot \\ S^{\circ}_{x_k x_1} & S^{\circ}_{x_k x_2} & & S^{\circ}_{x_k x_k} \end{bmatrix}^{-1} \begin{bmatrix} S^{\circ}_{x_1 y} \\ S^{\circ}_{x_2 y} \\ \cdot \\ S^{\circ}_{x_k y} \end{bmatrix}$$

These latter quantities are sums of squares, sums of cross products and regression coefficients which are derived by supposing that all the sites data were pooled into one large set and using the basic method described in Section 2.2. Thus, if we are able to conclude that the regression lines from the n sites are similar, then the best summary of the relationship between y and $x_1, x_2, \ldots, x_k$ would probably be provided by the overall estimated regression line,

$$\hat{y}_{ij} = \hat{\alpha}_0 + \hat{\beta}^{\circ}_1(x_{1ij} - \bar{x}_{1..}) + \hat{\beta}^{\circ}_2(x_{2ij} - \bar{x}_{2..}) + \cdots + \hat{\beta}^{\circ}_k(x_{kij} - \bar{x}_{k..})$$

where $\hat{\alpha} = \bar{y}_{...}$

However, we may conclude that, although the regression coefficients $\beta_{1i}, \beta_{2i}, \ldots, \beta_{ki}$ do not differ from site to site, the position parameters α_i do. To

assess this, and also to give an estimate of the appropriate regression line, we will define

$$S^{c}_{x_j y} = \sum_{i=1}^{n} S^{i}_{x_j y}$$

$$S^{c}_{x_j x_l} = \sum_{i=1}^{n} S^{i}_{x_j x_l}$$

$$S^{c}_{yy} = \sum_{i=1}^{n} S^{i}_{yy}$$

The quantities $\hat{\beta}^{c}_1, \hat{\beta}^{c}_2, \ldots, \hat{\beta}^{c}_k$ will be given by

$$\begin{bmatrix} \hat{\beta}^{c}_1 \\ \hat{\beta}^{c}_2 \\ \vdots \\ \hat{\beta}^{c}_k \end{bmatrix} = \begin{bmatrix} S^{c}_{x_1 x_1} & S^{c}_{x_1 x_2} & & S^{c}_{x_1 x_k} \\ S^{c}_{x_2 x_1} & S^{c}_{x_2 x_2} & & S^{c}_{x_2 x_k} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ S^{c}_{x_k x_1} & S^{c}_{x_k x_2} & & S^{c}_{x_k x_k} \end{bmatrix}^{-1} \begin{bmatrix} S^{c}_{x_1 y} \\ S^{c}_{x_2 y} \\ \cdot \\ \cdot \\ S^{c}_{x_k y} \end{bmatrix}$$

The appropriate regression line for site i would be

$$\hat{y}_{ij} = \bar{y}_{i.} + \hat{\beta}^{c}_1(x_{1ij} - \bar{x}_{1i.}) + \hat{\beta}^{c}_2(x_{2ij} - \bar{x}_{2i.}) + \quad + \hat{\beta}^{c}_k(x_{kij} - \bar{x}_{ki.})$$

Notice that the regression coefficients are the same for each site but that the position parameter varies from site to site.

Source	*Sum of squares*	*Degrees of freedom*
Overall regression	$\sum_{j=1}^{k} \hat{\beta}^{o}_j S^{o}_{x_j y}$	k
Difference in positions	$\left(S^{o}_{yy} - \sum_{j=1}^{k} \hat{\beta}^{o}_j S^{o}_{x_j y}\right) - \left(S^{c}_{yy} - \sum_{j=1}^{k} \hat{\beta}^{c}_j S^{c}_{x_j y}\right)$	
Difference in regressions	$\sum_{i=1}^{n} \sum_{j=1}^{k} \hat{\beta}_{ji} S^{i}_{x_j y} - \sum_{j=1}^{k} \hat{\beta}^{c}_j S^{c}_{x_j y}$	$(n-1)k$
Residual	$s^{c}_{yy} - \sum_{i=1}^{n} \sum_{j=1}^{k} \hat{\beta}_{ji} S^{i}_{x_j y}$	$N-(k+1)n$
Total	S^{o}_{yy}	$N-1$

This analysis of variance table gives the same type of information as the one in Subsection 1.3.2 and, as previously, it is preferable to carry out the tests in the following order:

for $100\alpha\%$ significance tests

(1) accept the hypothesis $\beta_{j1} = \beta_{j2} = \cdot \cdot = \beta_{jn}$ (for $j = 1, 2, \ldots, k$) whenever

$$\frac{\text{Difference in regressions mean square}}{\text{Residual mean square}} < F((n-1)k, N-(k+1)n, 1-\alpha)$$

(2) if the above has been accepted, then accept $\alpha_1 = \alpha_2 = \cdots = \alpha_n$ whenever

$$\frac{\text{Difference in positions mean square}}{\text{Residual mean square}} < F(n-1, N-(k+1)n, 1-\alpha)$$

(3) if both the above hypotheses have been accepted, then accept that there is no linear association between y and $x_1, x_2, \ldots, x_k$ whenever

$$\frac{\text{Overall regression mean square}}{\text{Residual mean square}} < F(k, N-(k+1)n, 1-\alpha)$$

In each case,

$$\text{Mean square} = \frac{\text{Sum of squares}}{\text{Degrees of freedom}}$$

2.4.2 Observations with unequal precision

We considered earlier (Subsection 1.3.3) the problem of observations with unequal precision where readings on only one x variable were available. In this subsection, we will assume that our information consists of n sets of readings of $y, x_1, x_2, \ldots, x_k$, denoted by $(y_1, x_{11}, x_{12}, \ldots, x_{1k})$, $(y_2, x_{21}, x_{22}, \ldots, x_{2k}), \ldots, (y_n, x_{n1}, x_{n2}, \ldots, x_{nk})$, together with n variances of the ys, $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$.

The estimates given by equation (36) will still provide unbiased estimates of the regression parameters $\beta_1, \beta_2, \ldots, \beta_k$; however, as in Subsection 1.3.3, alternative estimators with smaller variance are available.

Suppose that we re-define

$$S_{x_j y} = \sum_{i=1}^{n} w_i (x_{ji} - \bar{x}_j)(y_i - \bar{y})$$

$$S_{x_j x_l} = \sum_{i=1}^{n} w_i (x_{ji} - \bar{x}_j)(x_{li} - \bar{x}_l)$$

where

$$\bar{x}_j = \left(\sum_{i=1}^{n} w_i x_{ji}\right) \Big/ \left(\sum_{i=1}^{n} w_i\right) \qquad \bar{y} = \left(\sum_{i=1}^{n} w_i y_i\right) \Big/ \left(\sum_{i=1}^{n} w_i\right) \quad \text{and} \quad w_i = \frac{1}{\sigma_i^2}$$

Then, we have $\hat{\alpha} = \bar{y}$ and least squares estimates for $\beta_1, \beta_2, \ldots, \beta_k$ are given by equation (36) with $S_{x_j y}$ and $S_{x_j x_l}$ defined as above. Similarly, the variance covariance matrix is given by equation (37) with σ^2 omitted.

Thus, the $100(1-\alpha)\%$ confidence interval for β_i would be

$$\hat{\beta}_i \pm Z(\alpha/2)\sqrt{\text{Var}\,(\hat{\beta}_i)}$$

It is unlikely in practice that repeated values of y would be available with fixed values of all the variables $x_1, x_2, \ldots, x_k$ and, consequently, the n variances of the ys, $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, could not normally be estimated. Hence, no further discussion of this topic is given here.

2.4.3 Missing observations

The method of estimation described in Section 2.2 relies on there being available n complete sets of $(k+1)$ numbers $(y, x_1, x_2, \ldots, x_k)$. When, for some reason, a value for one of the x variables is not available in one of these sets, the method of estimation as described is no longer applicable. As this problem arises frequently, many attempts have been made at providing a solution which takes the form of a simple modification to the least squares method.

However, no one particular solution appears to be the best in all situations. The basic solution, with which most others are compared, is to discard all sets of observations which are not complete and then to apply the usual least squares method to the remaining data. Two of the simpler rival solutions are as follows:

(i) Use the mean value of the available observations of the variable in place of the missing value.
(ii) Select an x variable highly correlated with the variable which has the missing value. With complete data, perform a simple linear regression (as described in Section 1.2) between these two variables and use the resulting equation to predict the missing value.

It is said that solution (i) is good when using data with small correlations, the basic solution is good when moderate correlations are present and solution (ii) is best for highly correlated data. Each of the methods described can be used when several values are missing, simply by repeated application. The basic solution tends to be best when relatively few values are missing.

2.5 Special Models

2.5.1 Univariate polynomial models

Chapter 1 dealt with the problem of fitting a straight line; in contrast, this section will consider the problem of fitting a curve. If the model $y = a + bx$ is found to be inadequate in describing the relationship between y and x, then a natural extension, which introduces some curvature, would be to consider the quadratic model $y = a + bx + cx^2$. If a plot of y against x reveals two turning points, then a model $y = a + bx + cx^2 + dx^3$ might be appropriate. In general, the polynomial model

$$y = a + b_1 x + b_2 x^2 + \cdots + b_k x^k \tag{52}$$

is a model to be considered as an alternative to a straight line model (which is, in any case, only model (52) with $k = 1$).

With the usual set of data (y_1, x_1), $(y_2, x_2), \ldots, (y_n, x_n)$, then model (52) would imply a relationship

$$y_i = a + b_1 x_i + b_2 x_i^2 + \cdots + b_k x_i^k + e_i \tag{53}$$

By comparing this with model (33), it is evident that when $x_{1i} = x_i$, $x_{2i} = x_i^2$, $x_{3i} = x_i^3$, etc., the two models are identical. Hence, to fit a polynomial of degree k, it is possible to use all the techniques of multiple regression taking $x_1 = x$, $x_2 = x^2, \ldots, x_k = x^k$. Rewriting the equation (34) presents no special problems,

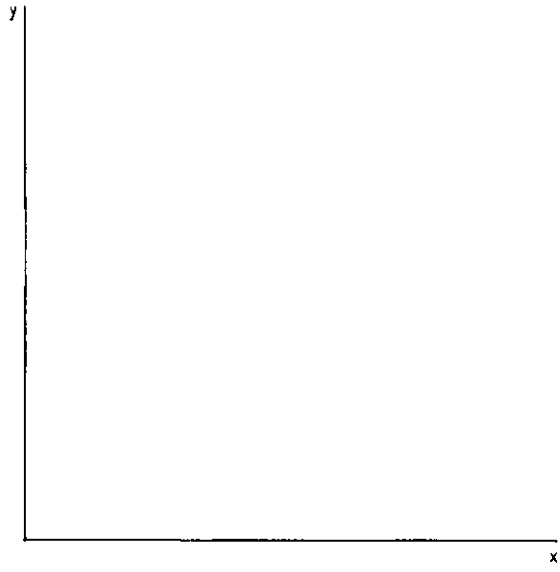


Fig. 10. Inappropriate data for fitting a quadratic model.

$\bar{x}_1$ being the mean of the x values, $\bar{x}_2$ the mean of the squares of the x values, etc. Estimates of the polynomial coefficients $b_1, b_2, \ldots, b_k$ (which equal $\beta_1, \beta_2, \ldots, \beta_k$) are given by equation (36).

Some difficulty may arise in determining $\mathbf{S}_{xx}^{-1}$ when k is particularly large or when the x values are spread in such a way that it would be difficult to envisage drawing a unique polynomial of that degree through those points. For instance, if $k = 2$ (i.e. a quadratic is being fitted) and $x_1, x_2, \ldots, x_n$ consist of n_1 values of -1 and $n - n_1$ values of $+1$, then a graph of the data would be similar to Figure 10.

One might be convinced by the argument that the best straight line should be that which passes through the mean of the y values at $x = -1$ and the mean of the y values at $x = +1$. However, there is an infinity of quadratics which would pass through these two points. This is reflected in equation (36) by $\mathbf{S}_{xx}$ being singular and, consequently, $\mathbf{S}_{xx}^{-1}$ not existing.

Numerical problems in actually finding $\mathbf{S}_{xx}^{-1}$ arise when $\mathbf{S}_{xx}$ is almost singular. Such a situation might have arisen if, in the previous example, one further y value had been available at $x = 1.000\,000\,001$. The effect of this additional information on Figure 10 is a fair reflection on the small movement $\mathbf{S}_{xx}$ would make from singularity. As most numerical routines for matrix inversion are not 100% efficient, there may be practical problems in determining $\mathbf{S}_{xx}^{-1}$.

Because $b_1, b_2, \ldots, b_k$ are associated with successively higher powers of x, a more natural order to the tests of significance is now available. For instance, investigating whether it was in fact necessary to fit a polynomial of order k rather than one of order $(k-1)$ would be a natural first step to take. This would be achieved by testing $\beta_k = 0$, as described in the second of the initial significance tests of Subsection 2.3.2. If it proved possible to accept $\beta_k = 0$, then the next step might be to refit the model with the x^k term omitted and to test

$\beta_{k-1} = 0$, and so on. Eventually, when the hypothesis $\beta_s = 0$ has been rejected, a satisfactory order of polynomial would have been reached and it would have order s.

Depending on the problem, there might be some value in considering the coefficients of the lower order terms in the model by refitting the model as

$$y_i = a + b_1 x_i + b_2 x_i^2 + \cdots + b_s x_i^s + e_i$$

and testing each of $b_1 = 0$, $b_2 = 0, \ldots, b_{s-1} = 0$ separately, using the second test procedure of Subsection 2.3.2. However, usually one wants to discover the minimum order of polynomial that it is necessary to fit and then to predict y as accurately as possible using that order of polynomial.

There is some advantage in using a set of orthogonal polynomials to rewrite the model (52). This was particularly useful when the only available calculating aid was a desk calculator as the need to invert a $k \times k$ matrix ($\mathbf{S}_{xx}$) is eliminated. However, with the availability of programmable digital computers nowadays, this advantage is less crucial.

By defining $P_r(x_i)$ as an rth order polynomial in x_i, the model (53) may be rewritten as

$$y_i = \gamma_0 P_0(x_i) + \gamma_1 P_1(x_i) + \gamma_2 P_2(x_i) + \cdots + \gamma_k P_k(x_i) + e_i \tag{54}$$

The least squares estimates of $\gamma_0, \gamma_1, \ldots, \gamma_k$ would be given by

$$\begin{bmatrix} \hat{\gamma}_0 \\ \hat{\gamma}_1 \\ \\ \hat{\gamma}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} [P_0(x_i)]^2 & \sum_{i=1}^{n} P_0(x_i)P_1(x_i) & \sum_{i=1}^{n} P_0(x_i)P_k(x_i) \\ \sum_{i=1}^{n} P_1(x_i)P_0(x_i) & \sum_{i=1}^{n} [P_1(x_i)]^2 & \sum_{i=1}^{n} P_1(x_i)P_k(x_i) \\ \\ \sum_{i=1}^{n} P_k(x_i)P_0(x_i) & \sum_{i=1}^{n} P_k(x_i)P_1(x_i) & \sum_{i=1}^{n} [P_k(x_i)]^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i P_0(x_i) \\ \sum_{i=1}^{n} y_i P_1(x_i) \\ \\ \sum_{i=1}^{n} y_i P_k(x_i) \end{bmatrix}$$

However, if it were possible to arrange for

$$\sum_{i=1}^{n} P_r(x_i)P_s(x_i) = 0 \qquad (\text{for } r, s = 0, 1, 2, \ldots, k \text{ and } r \neq s) \tag{55}$$

then we would immediately have

$$\hat{\gamma}_r = \frac{\sum_{i=1}^{n} y_i P_r(x_i)}{\sum_{i=1}^{n} [P_r(x_i)]^2} \qquad (\text{for } r = 0, 1, 2, \ldots, k)$$

When the restrictions (55) hold for the polynomials $P_0(x), P_1(x), \ldots, P_k(x)$, they are referred to as orthogonal polynomials. These restrictions enable the coefficients in the polynomials to be calculated in terms of $x_1, x_2, \ldots, x_n$. Using a set of orthogonal polynomials was particularly valuable in precomputer days as it avoided the inversion of a large matrix. However, even nowadays it is

advantageous to use orthogonal polynomials so as to avoid problems of numerical instability when inverting large, almost singular, matrices.

When the xs are equally spaced and thus $x_i = a + ib$, the problem is simplified by transforming x on to a unit interval scale using

$$X_i = \frac{x_i - \bar{x}}{b} \tag{56}$$

where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = a + \left(\frac{n+1}{2}\right)b$$

Hence, the values $X_1, X_2, \ldots, X_n$ become $-\frac{1}{2}(n-1)$, $-\frac{1}{2}(n-3), \ldots, \frac{1}{2}(n-3)$, $\frac{1}{2}(n-1)$. Rewriting model (54) using the transformed X gives

$$y_i = \alpha_0\phi_0(X_i) + \alpha_1\phi_1(X_i) + \cdots + a_k\phi_k(X_i) + e_i$$

where

$$\sum_{i=1}^{n} \phi_r(X_i)\phi_s(X_i) = 0 \qquad (\text{for } r, s = 0, 1, 2, \ldots, k \text{ and } r \neq s) \tag{57}$$

Since all problems with equally spaced x values will have the same X values (X_i will always be $i - [(n+1)/2]$, it is possible to establish $\phi_r(X_i)$ suitable for all such problems. The first six orthogonal polynomials are

$$\begin{aligned}
\phi_0(X) &= 1 \\
\phi_1(X) &= \lambda_{1n}X \\
\phi_2(X) &= \lambda_{2n}(X^2 - \tfrac{1}{12}(n^2-1)) \\
\phi_3(X) &= \lambda_{3n}(X^3 - \tfrac{1}{20}(3n^2-7)X) \\
\phi_4(X) &= \lambda_{4n}(X^4 - \tfrac{1}{14}(3n^2-13)X^2 + \tfrac{3}{560}(n^2-1)(n^2-9)) \\
\phi_5(X) &= \lambda_{5n}(X^5 - \tfrac{5}{18}(n^2-7)X^3 + \tfrac{1}{1008}(15n^4 - 230n^2 + 407)x) \\
\phi_6(X) &= \lambda_{6n}(X^6 - \tfrac{5}{44}(3n^2-31)X^4 + \tfrac{1}{176}(5n^4 - 110n^2 + 329)X^2 \\
&\quad - \tfrac{5}{14874}(n^2-1)(n^2-9)(n^2-25))
\end{aligned}$$

Furthermore, for positive values of X, tables of the values of these polynomials are available (Pearson and Hartley (1972)). For negative values of X, the following relationship enables the values of the polynomials to be calculated easily:

$$\phi_{2r}(-X) = \phi_{2r}(X)$$

and

$$\phi_{2r-1}(-X) = -\phi_{2r-1}(X) \qquad (\text{for } r = 1, 2, 3, \ldots)$$

The restrictions (57) define the polynomial $\phi_r(X)$ except for the arbitrary constant λ_{rn}. This could be taken to be unity, but tabulators usually choose the value of λ_{rn} so that the values of $\phi_r(X)$ are integers. Thus, most tables contain the values of $\phi_r(X)$ for positive X, the value of λ_{rn} and the value of $\sum_{i=1}^{n} [\phi_r(X_i)]^2$.

Using these tables, it is a simple matter to calculate the estimates

$$\hat{\alpha}_r = \frac{\sum_{i=1}^{n} y_i \phi_r(X_i)}{\sum_{i=1}^{n} [\phi_r(X_i)]^2} \tag{58}$$

and their variances

$$\text{Var}(\hat{\alpha}_r) = \frac{\sigma^2}{\sum_{i=1}^{n} [\phi_r(X_i)]^2}$$

The only inconvenience of using orthogonal polynomials occurs when rewriting the estimated equation as a polynomial in x. For the case of $k = 2$, we would have to unravel the fitted equation

$$\begin{aligned} y &= \hat{\alpha}_0 \phi_0(X) + \hat{\alpha}_1 \phi_1(X) + \hat{\alpha}_2 \phi_2(X) \\ &= \hat{\alpha}_0 + \hat{\alpha}_1 \left[\lambda_{1n} \left(\frac{x - \bar{x}}{b} \right) \right] + \hat{\alpha}_2 \left[\lambda_{2n} \left(\frac{x - \bar{x}}{b} \right)^2 - \tfrac{1}{12}(n^2 - 1) \right] \end{aligned}$$

Fortunately, it is not usually necessary to perform this step before carrying out tests of significance on the polynomial coefficients. With no prior knowledge of the order of polynomial which would describe the relation between y and x, a procedure analogous to that used with model (53) would probably be a natural first step to take. This would be achieved by, first of all, forming the analysis of variance table which is given below.

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*
Linear term	$\hat{\alpha}_1^2 \sum_{i=1}^{n} [\phi_1(X_i)]^2$	1	
Quadratic term	$\hat{\alpha}_2^2 \sum_{i=1}^{n} [\phi_2(X_i)]^2$		$\left(= \frac{\text{Sum of squares}}{\text{Degrees of freedom}} \right)$
kth Order term	$\hat{\alpha}_k^2 \sum_{i=1}^{n} [\phi_k(X_i)]^2$		
Residual	By subtraction	$n - k - 1$	
Total	$\sum_{i=1}^{n} (y_i - \hat{\alpha}_0)^2$	$n - 1$	

Then, the hypothesis 'coefficient of x^k' $= 0$ would be tested by accepting the hypothesis whenever

$$\frac{k\text{th Order term mean square}}{\text{Residual mean square}} < F(1, n - k - 1, 1 - \alpha)$$

for a $100\alpha\%$ significance test.

The procedure with model (53) suggested that if this hypothesis was accepted, then the next step would be to drop the term x^k from the model, re-estimate the parameters and then test the hypothesis 'coefficient of x^{k-1}' $= 0$. Using orthogonal polynomials, this is achieved simply by adding the kth order term sum of squares into the residual, doing likewise with the degrees of freedom and then testing the hypothesis 'coefficient of x^{k-1}' $= 0$ by accepting the hypothesis whenever

$$\frac{(k-1)\text{th Order term mean square}}{\text{New residual mean square}} < F(1, n-k, 1-\alpha)$$

for a $100\alpha\%$ significance test.

If this hypothesis was accepted, then the next step would be to add the $(k-1)$th order term sum of squares and degrees of freedom to those of the already augmented residual and to test the hypothesis 'coefficient of x^{k-2}' $= 0$. This procedure would then be repeated until a hypothesis 'coefficient of x^s' $= 0$ was rejected and then the fitted model decided on would be

$$y_i = \hat{\alpha}_0 + \hat{\alpha}_1\phi_1\left(\frac{x_i - \bar{x}}{b}\right) + \hat{\alpha}_2\phi_2\left(\frac{x_i - \bar{x}}{b}\right) + \cdots + \hat{\alpha}_s\phi_s\left(\frac{x_i - \bar{x}}{b}\right) + e_i \quad (59)$$

The new residual mean square used in the test of 'coefficient of x^s' $= 0$ would provide an estimate $\hat{\sigma}^2$ of σ^2 and it would have $(n-s-1)$ degrees of freedom.

The $100(1-\alpha)\%$ confidence intervals for $\alpha_1, \alpha_2, \ldots, \alpha_s$ are given by

$$\hat{\alpha}_r \pm t(n-s-1, 1-\alpha/2)\sqrt{\hat{\sigma}^2 \Big/ \sum_{i=1}^{n} [\phi_r(X_i)]^2} \qquad (\text{for } r = 0, 1, 2, \ldots,$$

If it is intended to use equation (59) in order to predict y for $x = x_0$, then the predicted value of y is given by

$$\hat{y} = \hat{\alpha}_0 + \hat{\alpha}_1\phi_1\left(\frac{x_0 - \bar{x}}{b}\right) + \hat{\alpha}_2\phi_2\left(\frac{x_0 - \bar{x}}{b}\right) + \cdot\ \cdot + \hat{\alpha}_s\phi_s\left(\frac{x_0 - \bar{x}}{b}\right)$$

and its variance is given by

$$\text{Var}(\hat{y}) = \sigma^2\left[\frac{1}{n} + \sum_{r=1}^{s}\left(\left[\phi_r\left(\frac{x_0 - \bar{x}}{b}\right)\right]^2 \Big/ \sum_{i=1}^{n} [\phi_r(X_i)]^2\right)\right]$$

This variance only reflects the uncertainty about $\alpha_0, \alpha_1, \ldots, \alpha_s$ and its use is appropriate when the intention is to predict the mean value of y. When the intention is to predict the outcome of a single reading of y, a further σ^2 should be added to Var $(\hat{y})$.

Corresponding $100(1-\alpha)\%$ confidence intervals for the mean value of y and the outcome of a single reading of y are given by

$$\hat{y} \pm t(n-s-1, 1-\alpha/2)\sqrt{\sigma^2\left[\frac{1}{n} + \sum_{r=1}^{s}\left(\left[\phi_r\left(\frac{x_0 - \bar{x}}{b}\right)\right]^2 \Big/ \sum_{i=1}^{n} [\phi_r(X_i)]^2\right)\right]}$$

and

$$\hat{y} \pm t(n-s-1, 1-\alpha/2)\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\sum_{r=1}^{s}\left(\left[\phi_r\left(\frac{x_0-\bar{x}}{b}\right)\right]^2 \Big/ \sum_{i=1}^{n}[\phi_r(X_i)]^2\right)\right]}$$

respectively.

However, it must be emphasised that it is dangerous to extrapolate using a polynomial model which has been formed by deleting higher powers. The order of polynomial fitted has been chosen to describe the relation between y and x within the region of x values observed and, outside of this region, there is no information on the relationship between y and x. Consequently, it is only safe to extrapolate when it is known that a certain degree of polynomial describes the relation between y and x both within the region of x values observed and over the region in which extrapolation is to be performed.

2.5.2 Multivariable polynomial models

Just as the simple model $y = a + bx$ was expanded to the polynomial model (52) so the multiple regression model (33) may be expanded to include powers of $x_1, x_2, \ldots, x_k$, products of $x_1, x_2, \ldots, x_k$ and products of powers of $x_1, x_2, \ldots, x_k$.

For example, for $k = 2$, we might have the model

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + b_{11} x_{1i}^2 + b_{22} x_{2i}^2 + b_{12} x_{1i} x_{2i} + e_i$$

(a quadratic in x_1 and x_2). However, simply by defining $X_{1i} = x_{1i}$, $X_{2i} = x_{2i}$, $X_{3i} = x_{1i}^2$, $X_{4i} = x_{2i}^2$, $X_{5i} = x_{1i} x_{2i}$, the model would become

$$y_i = a + b_1 X_{1i} + b_2 X_{2i} + b_{11} X_{3i} + b_{22} X_{4i} + b_{12} X_{5i} + e_i$$

which is identical to model (33).

Thus, the polynomial extension of model (33) simply produces a model of the same type as model (33) and it can therefore be handled by the techniques described in Subsection 2.5.1 in connection with the polynomial model for a single x variable. However, with a moderate value of k, the possible number of terms generated (even by a quadratic form) can be enormous and fitting such models is usually quite unjustifiable and frequently dangerous. Unless a large number of observations, spread over a wide region of values of $x_1, x_2, \ldots, x_k$, is available, an apparently good model may be generated, not because the correct relationship between y and $x_1, x_2, \ldots, x_k$ has been found, but because there are so many parameters in the model that there is almost a separate parameter for each observed y value.

2.5.3 Periodic regression

Many hydrological phenomena exhibit periodicity, the period being possibly annual, monthly or daily, but usually associated with time. For example, evaporation in the United Kingdom is strongly seasonal with a pronounced annual cycle.

Suppose that we have records of the value of a variable y taken at n equally

spaced time points, i.e. pairs of values (y_1, t_1), $(y_2, t_2), \ldots, (y_n, t_n)$ where $t_i = a + ib$. Then, the time scale may easily be modified to give readings at times $1, 2, \ldots, n$ by redefining the time scale as $T_i = (t_i - a)/b$.

Using this new time scale, a model with period n would be

$$y_i = \alpha + \beta \cos\left(\frac{2\pi T_i}{n}\right) + \gamma \sin\left(\frac{2\pi T_i}{n}\right) + e_i$$

$$= \alpha + \beta \cos\left(\frac{2\pi i}{n}\right) + \gamma \sin\left(\frac{2\pi i}{n}\right) + e_i \qquad (60)$$

In other words, if we had further observations, $T_i = n+1, n+2, n+3, \ldots$, then the model value of y for $T_i = k$ would be identical to the model value of y for $T_i = n+k, 2n+k, \ldots$ (except for the error terms). For instance, for $T_i = n+k$,

$$y_{n+k} = \alpha + \beta \cos\left(\frac{2\pi(n+k)}{n}\right) + \gamma \sin\left(\frac{2\pi(n+k)}{n}\right) + e_{n+k}$$

$$= \alpha + \beta \cos\left(2\pi + \frac{2\pi k}{n}\right) + \gamma \sin\left(2\pi + \frac{2\pi k}{n}\right) + e_{n+k}$$

$$= \alpha + \beta \cos\left(\frac{2\pi k}{n}\right) + \gamma \sin\left(\frac{2\pi k}{n}\right) + e_{n+k}$$

since $\cos(2\pi + \theta) = \cos\theta$ and $\sin(2\pi + \theta) = \sin\theta$ and, hence, y_{n+k} is identical to the model value for y_k (except for the error term). Thus, for a model with period n, the model values of $T_i = n+1$ onwards repeat those of $T_i = 1$ onwards.

Figure 11 gives an example of such a model with period 12,

$$y_i = 1 + 0.5 \cos\left(\frac{2\pi i}{12}\right) + 0.25 \sin\left(\frac{2\pi i}{12}\right)$$

plotted for $i = 1, 2, \ldots, 24$ (i.e. the model repeats itself after 12 values).

As the model (60) is written at the moment, the periodicity is equal to the number of observations collected. To eliminate this restriction, we may extend our model to include terms with certain other periods and then arrange for our estimation technique to select those terms with periods which best match the periodicity exhibited in the observed data.

Hence, our model would be

$$y_i = \alpha + \sum_{r=1}^{s} \left[\beta_r \cos\left(\frac{2\pi r}{n} T_i\right) + \gamma_r \sin\left(\frac{2\pi r}{n} T_i\right) \right] + e_i$$

$$= \alpha + \sum_{r=1}^{s} \left[\beta_r \cos\left(\frac{2\pi r}{n} i\right) + \gamma_r \sin\left(\frac{2\pi r}{n} i\right) \right] + e_i \qquad (61)$$

In the above model, $r = 1$ gives terms of period n, $r = 2$ gives terms of period $n/2$, and so on, until finally $r = s$ gives terms of period n/s. As phenomena with period of 2 or less are unlikely to be evident in the data, the maximum value which it is sensible to take for s is $\frac{1}{2}(n-1)$ when n is odd and $\frac{1}{2}n - 1$. when n is even.

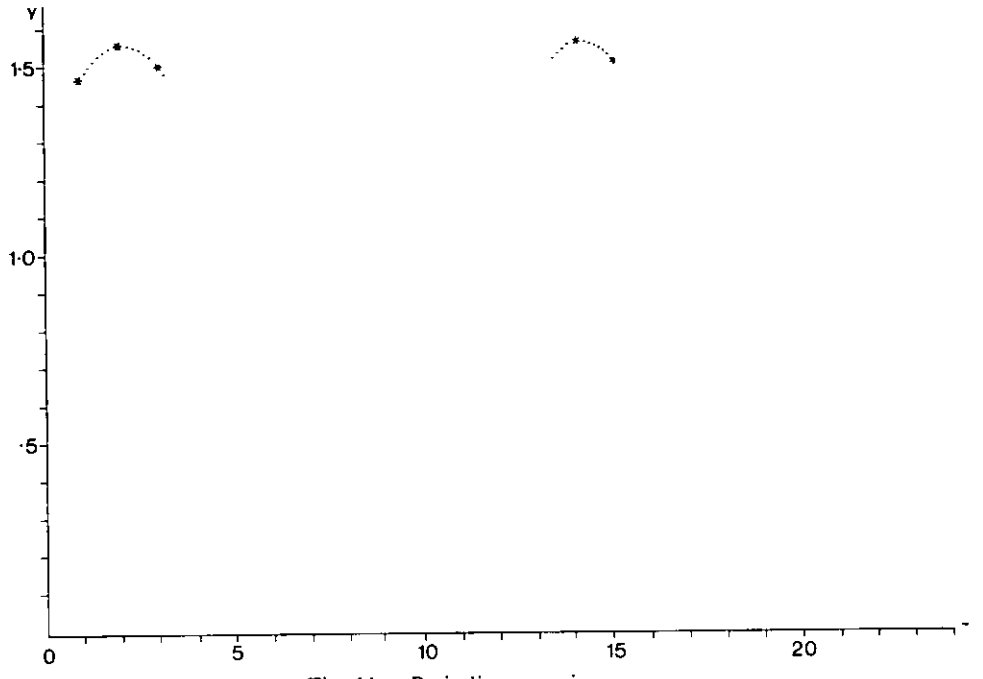


Fig. 11. Periodic regression curve.

By taking

$$x_{1i} = \cos\left(\frac{2\pi \times 1}{n} i\right) \qquad x_{2i} = \sin\left(\frac{2\pi \times 1}{n} i\right)$$

$$x_{3i} = \cos\left(\frac{2\pi \times 2}{n} i\right) \qquad x_{4i} = \sin\left(\frac{2\pi \times 2}{n} i\right) \qquad \text{etc.}$$

the form of model (61) is clearly identical to that of model (33). Hence, the methods of Section 2.2 are applicable in periodic regression. However, as with orthogonal polynomials (described in Subsection 2.5.1), the form of $\mathbf{S}_{xx}$ will simplify to be a diagonal matrix.

Since

$$\sum_{i=1}^{n} \cos\left(\frac{2\pi r}{n} i\right) = \sum_{i=1}^{n} \sin\left(\frac{2\pi r}{n} i\right) = 0 \qquad (\text{for } r = 1, 2, \ldots, s)$$

the means of all the x variables are zero.

Therefore, $S_{x_j x_l} = \sum_{i=1}^{n} (x_{ji} - \bar{x}_j)(x_{li} - \bar{x}_l)$ will become $\sum_{i=1}^{n} x_{ji} x_{li}$ which is equivalent to either

$$\sum_{i=1}^{n} \cos\left(\frac{2\pi r_j}{n} \cdot i\right) \cos\left(\frac{2\pi r_l}{n} \cdot i\right)$$

$$\sum_{i=1}^{n} \cos\left(\frac{2\pi r_j}{n} \cdot i\right) \sin\left(\frac{2\pi r_l}{n} \cdot i\right)$$

or

$$\sum_{i=1}^{n} \sin\left(\frac{2\pi r_j}{n} \cdot i\right) \sin\left(\frac{2\pi r_l}{n} \cdot i\right)$$

depending on whether x_j and x_l correspond with 'cos' or 'sin' terms in model (61) (r_j and r_l are integers between 1 and s). For $r_j \neq r_l$, the first and third of these expressions are zero and the second expression is always zero, even when $r_j = r_l$. Hence, $\mathbf{S}_{xx}$ is a diagonal matrix.

Thus, estimates of the parameters are as follows:

$$\begin{aligned}
\hat{\alpha} &= \bar{y} \\
\hat{\beta}_r &= \left(\sum_{i=1}^{n} y_i \cos\left(\frac{2\pi r}{n} i\right)\right) \Big/ \left(\sum_{i=1}^{n} \left(\cos\left(\frac{2\pi r}{n} i\right)\right)^2\right) \\
&= \frac{2}{n} \sum_{i=1}^{n} y_i \cos\left(\frac{2\pi r}{n} i\right) \\
\hat{\gamma}_r &= \left(\sum_{i=1}^{n} y_i \sin\left(\frac{2\pi r}{n} i\right)\right) \Big/ \left(\sum_{i=1}^{n} \left(\sin\left(\frac{2\pi r}{n} i\right)\right)^2\right) \\
&= \frac{2}{n} \sum_{i=1}^{n} y_i \sin\left(\frac{2\pi r}{n} i\right)
\end{aligned} \tag{62}$$

The variances of these estimates are as follows:

$$\text{Var}\,(\hat{\alpha}) = \frac{\sigma^2}{n}$$

$$\text{Var}\,(\hat{\beta}_r) = \text{Var}\,(\hat{\gamma}_r) = \frac{2\sigma^2}{n}$$

The analysis of variance table analogous to the one produced for orthogonal polynomials is given below.

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*
Terms of period n	$\frac{n}{2}(\hat{\beta}_1^2 + \hat{\gamma}_1^2)$	2	
Terms of period $n/2$	$\frac{n}{2}(\hat{\beta}_2^2 + \hat{\gamma}_2^2)$		$\left(= \frac{\text{Sum of squares}}{\text{Degrees of freedom}}\right)$
Terms of period n/s	$\frac{n}{2}(\hat{\beta}_s^2 + \hat{\gamma}_s^2)$		
Residual	By subtraction	$n - 2s - 1$	
Total	$\sum_{i=1}^{n} (y_i - \bar{y})^2$	$n - 1$	

To test the hypothesis $\beta_r = \gamma_r = 0$ (i.e. there is no periodicity of length n/r in the data), a $100\alpha\%$ significance test would accept this hypothesis whenever

$$\frac{\text{'Term of period } n/r\text{' mean square}}{\text{Residual mean square}} < F(2, n-2s-1, 1-\alpha)$$

To simultaneously test $\beta_1 = \beta_2 = \cdots = \beta_s = \gamma_1 = \gamma_2 = \cdots = \gamma_s = 0$, a $100\alpha\%$ significance test would accept this hypothesis whenever

$$\frac{\frac{n}{4s}\sum_{r=1}^{s}(\hat{\beta}_r^2 + \hat{\gamma}_r^2)}{\text{Residual mean square}} < F(2s, n-2s-1, 1-\alpha)$$

As with the orthogonal polynomial model, it is not necessary to recompute the parameter estimates each time a term is omitted. Consequently, after applying the first test procedure several times with suitable values of r and, hence, omitting certain terms from the original model, the residual sum of squares for the new model would be equal to the residual sum of squares of the original model plus the sums of squares of all terms omitted from the model. Its degrees of freedom would similarly be augmented to $n-2s'-1$ (where $s-s'$ is the number of pairs of terms which have been omitted). An estimate of σ^2 would then be provided by the new residual mean square with $n-2s'-1$ degrees of freedom.

To predict a value of y at time t, calculate $T=(t-a)/b$ and then the predicted value of y is given by

$$\hat{y} = \hat{\alpha} + \sum_{r=1}^{s}\left[\hat{\beta}_r \cos\left(\frac{2\pi r}{n}T\right) + \hat{\gamma}_r \sin\left(\frac{2\pi r}{n}T\right)\right]$$

where the summation does not necessarily include all terms as some may have been rejected as a result of the tests mentioned above. Its variance is given by

$$\text{Var}(\hat{y}) = \frac{\sigma^2}{n}(1+4s')$$

where s' is the number of pairs of terms left in the summation above.

Clearly, this variance is independent of T and, hence, the precision of estimation is the same at all points in time. This emphasises the critical dependence of estimation on having the correct model. The model (61) assumes that patterns are repeated after a definite length of time and, consequently, when that length of time is greater than the range of observed values, this method of prediction is as dangerous as polynomial prediction. Therefore, it is only reliable to extrapolate results when prior information on the existence of certain periodicity is available.

The preceding discussion deals with the case of a model containing terms with periods which are simple fractions of n, the number of readings. However, the model does not need to have its periodicity related to the number of observations. For instance, if it is known that a 12 month periodicity exists and observations have been taken monthly, then the appropriate model might be

$$y_i = \alpha + \beta_1 \cos\left(\frac{2\pi}{12}T_i\right) + \gamma_1 \sin\left(\frac{2\pi}{12}T_i\right) + e_i$$

Also, there is no reason why a model should not contain both cyclic and non-cyclic terms. For instance, a model might be of the form

$$y_i = \alpha + \beta_1 \cos\left(\frac{2\pi}{12} T_i\right) + \gamma_1 \sin\left(\frac{2\pi}{12} T_i\right) + \beta_2 x_i + \beta_3 z_i + e_i$$

where x_i and z_i are two independent variables such as temperature and rainfall.

However, for models containing terms with periods which are not simple fractions of n, the full method described in Section 2.2 would have to be used as $\mathbf{S}_{xx}$ would no longer be a diagonal matrix.

2.5.4 Dummy variables

So far, we have assumed that all our information will be quantitative, all our measurements will be numbers. However, it is often the case that some information is qualitative. For instance, we may have recorded supplementary information on the geological features of the area studied, such as the area being either permeable or impermeable. We may have recorded wind force as being either strong, medium or light.

This type of information may be included in a multiple regression model by defining dummy variables. In the first example, we would have a single dummy variable x, defined as follows:

$x = 0$ when permeable

$x = 1$ when impermeable.

The variable x would then be included in the multiple regression equation with all readings in permeable sites having $x = 0$ and all readings in impermeable sites having $x = 1$. A significant regression coefficient associated with x would indicate that permeability was of value in explaining the variable being studied.

As a result of this, we may decide to make the information concerning permeability more detailed, perhaps on a 5 point scale. We might then use a variable x taking values 1, 2, 3, 4 or 5. However, this could imply an equally spaced scale which would mean that, for instance, the difference between permeabilities 4 and 5 would be the same as the difference between permeabilities 1 and 2. It might be more satisfactory to define four dummy variables as follows:

x	y	z	w	*Permeability*
1	0	0	0	1
0	1	0	0	2
0	0	1	0	3
0	0	0	1	4
0	0	0	0	5

Thus, in areas of permeability 3, for instance, we would have $x = 0$, $y = 0$, $z = 1$, $w = 0$. The estimates of the regression coefficients associated with x, y, z and w would then give some idea of the effect of permeabilities 1, 2, 3 and 4, relative to

5, on the dependent variable being studied. For example, suppose that the regression coefficients of x, y, z and w were 0.4, 0.3, 0.2 and 0.1 respectively. Then, our model would be saying that we should add 0.4 to readings at sites with permeability 5 to get readings comparable with those at sites with permability 1, and similarly for sites with permeabilities 2, 3 and 4. It would also be saying that the ordering of permeability on a scale 1 to 5 was justified by its effect on the variable being studied. However, if the regression coefficients had been 0.4, 0.1, 0.1 and 0.4, then this would have suggested that sites with permeability 1 and 4 had a similar effect on the variable being studied and that sites with permeability 2 and 3 had a similar effect which was nearer to that of sites with permeability 5.

For phenomena whose effect on the studied variable is far less obvious than permeability, the use of dummy variables and, in particular, the study of their regression coefficients may well give insight into the relative effects of different levels or features of the phenomena.

For certain phenomena, it may be sensible to build a multiple regression model using dummy variables only. On the other hand, dummy variables may be used in conjunction with other more conventional variables to make up the regression model. No assumptions are violated by using a variable which can obviously only take two values, indeed most experimental design models are composed entirely of such variables. However, it must be emphasised that to deal with r states, levels or conditions, it is only necessary to use $(r-1)$ variables. For instance, if variables had been used in the previous problem as shown below, then $\mathbf{S}_{xx}$ would have been singular.

x	y	z	w	v	*Permeability*
1	0	0	0	0	1
0	1	0	0	0	2
0	0	1	0	0	3
0	0	0	1	0	4
0	0	0	0	1	5

2.6 Alternatives to Least Squares

2.6.1 Pencil and ruler

Although a complete assessment of the relationship between the y variable and the variables $x_1, x_2, \ldots, x_k$ cannot be made graphically, considerable insight into the relationship can often be gained from simple graphs. For instance, plots of y against x_1, y against x_2, etc. will give some indication of where there are signs of strong relationships, where there might be problems due to a poor dispersion of values of the x variable and where there are signs of nonlinearity (manifested by certain points falling a considerable way from the trend of the rest of the data or by marked curvature in the plots). Assessments of the relationships amongst the x variables may also be made graphically.

A method has recently been suggested by Andrews (1972) for displaying observations of many variables on a single two dimensional graph. Suppose that n sets of observations on variables $z_1, z_2, \ldots, z_r$ are available and denoted by $(z_{11}, z_{21}, \ldots, z_{r1}), (z_{12}, z_{22}, \ldots, z_{r2}), \ldots, (z_{1n}, z_{2n}, \ldots, z_{rn})$. For each of these sets of values, the function

$$X = z_1/\sqrt{2} + z_2 \sin t + z_3 \cos t + z_4 \sin 2t + z_5 \cos 2t + \cdot$$

is plotted (X against t) for $-\pi < t < \pi$. Thus, for n sets of observations, n periodic graphs are produced.

Probably, the inference that can most easily be drawn from the resulting graph is which of the n sets of observations are similar. A set of graphs that cluster together will suggest similarity in the sets of observations which generated those graphs.

In a regression context where we have variables $y, x_1, x_2, \ldots, x_k$, such plots may be valuable on the x variables alone or on y and $x_1, \ldots, x_k$. Similarities may be detected between certain observations and this would suggest that these observations may have arisen under similar conditions. Equally well, the technique may be used to give a concise summary of already established similarities in groups of data. This technique is usually most effective when the more important variables are used as the coefficients of the low frequency terms in the function X.

One graphical method that has been suggested for plotting three dimensional data in two dimensions is to simplify one of the variables by reducing the number of different values to say half a dozen (perhaps by grouping) and then, on a two dimensional plot of the other two variables, to relate the size (or darkness) of each point plotted to the value of the simplified third variable.

2.6.2 Robust and distribution free methods

The method of Daniels (1954) (see Subsection 1.4.2) may be extended to deal with k independent variables. Equation (33) is equivalent to

$$a = y_i - x_{1i}b_1 - x_{2i}b_2 - \cdots - x_{ki}b_k$$

(with e_i omitted). Therefore, each observation generates a hyperplane and the intersection of these hyperplanes will lead to estimates of $a, b_1, \ldots, b_k$. However, with more than one x variable, the visual appeal of this technique is lost and some of the computation can be cumbersome.

This tends to be the problem with other distribution free methods and the emphasis in recent years has been more on robust methods of estimation related to least squares. One such method was suggested by Hinich and Talwar (1975). It aims to minimise the effect of the occasional observation which is a long way from the trend of other observations, i.e. a set of values $(y_i, x_{1i}, x_{2i}, \ldots, x_{ki})$ which is associated with a large e_i (in model (33)).

The method consists of dividing the n sets of observations into $m = n/k$ separate groups and then estimating $\alpha, \beta_1, \beta_2, \ldots, \beta_k$ separately on each group

of data using the basic method described in Subsection 2.2.1. This will give rise to the following set of statistics:

Group	*Estimate*				
	α	β_1	β_2	β_3	β_k
1	$\hat{\alpha}^1$	$\hat{\beta}_1^1$	$\hat{\beta}_2^1$	$\hat{\beta}_3^1$	$\hat{\beta}_k^1$
2	$\hat{\alpha}$	$\hat{\beta}_1^2$	$\hat{\beta}_2^2$	$\hat{\beta}_3^2$	$\hat{\beta}_k^2$
⋮	⋮	⋮	⋮	⋮	⋮
m	$\hat{\alpha}^m$	$\hat{\beta}_1^m$	$\hat{\beta}_2^m$	$\hat{\beta}_3^m$	$\hat{\beta}_k^m$

A preliminary estimate of α is taken to be the median of $\hat{\alpha}^1, \ldots, \hat{\alpha}^m$ and it is denoted by $\hat{\alpha}^p$. Similarly, preliminary estimates of $\beta_1, \beta_2, \ldots, \beta_k$ are computed and denoted by $\hat{\beta}_1^p, \hat{\beta}_2^p, \ldots, \hat{\beta}_k^p$. The residuals

$$\hat{e}_i = y_i - \hat{\alpha}^p - \sum_{r=1}^{k} \hat{\beta}_r^p (x_{ri} - \bar{x}_r) \qquad (\text{for } i = 1, 2, \ldots, n)$$

are formed and a range estimate of σ is calculated using the expression $\hat{\sigma} = (\hat{\varepsilon}_{0.72} - \hat{\varepsilon}_{0.28})/1.654$, where $\hat{\varepsilon}_q$ is defined to be the value below which $100q\,\%$ of the values $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ would fall if they were arranged in increasing order of magnitude. Thus, $72\,\%$ of the values $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ would fall below $\hat{\varepsilon}_{0.72}$ and $28\,\%$ of the values would fall below $\hat{\varepsilon}_{0.28}$.

All observations whose associated residual $\hat{e}_i$ is greater than $4\hat{\sigma}$ are discarded from the data and the basic method described in Subsection 2.2.1 is then applied to the remaining data. This is particularly suitable for problems in which there are a large number of observations, some being of dubious quality. The general problems of detecting outliers (observations which are associated with large es) will be discussed in Section 4.3. The choice of $4\hat{\sigma}$ as the level at which to reject observations is somewhat arbitrary and it may be necessary to alter this to satisfy the particular requirements of the problem in hand.

2.6.3 Ridge regression and principal components regression

The basic method outlined in Section 2.2 relies on the existence of the inverse of $\mathbf{S}_{xx}$. If some of the independent variables $x_1, \ldots, x_k$ are linearly related (i.e. there is collinearity or multicollinearity), then $\mathbf{S}_{xx}^{-1}$ will not exist and, consequently, there will be no solution to equation (35).

Methods developed to cope with problems of collinearity have established themselves in their own right and, as such, they are included in this section although they are highly relevant to Section 4.3.

Principal components regression (also known as orthogonal regression) consists of selecting uncorrelated combinations of variables $x_1, x_2, \ldots, x_k$ which show maximum variation in the data (the principal components), taking these combinations as new variables $z_1, z_2, \ldots, z_k$ and performing a multiple regression analysis of y on the first few of $z_1, z_2, \ldots, z_k$. These principal

components are derived by finding the eigenvalues and eigenvectors of $\mathbf{S}_{xx}$ after it has been scaled so as to have all diagonal elements equal to unity.

This scaling of $\mathbf{S}_{xx}$ is most easily achieved at the model stage by rewriting the model

$$y_i = \alpha + \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + \cdots + \beta_k(x_{ki} - \bar{x}_k) + e_i$$

as

$$y_i = \alpha + \beta_1\sqrt{S_{x_1x_1}}\frac{(x_{1i} - \bar{x}_1)}{\sqrt{S_{x_1x_1}}} + \beta_2\sqrt{S_{x_2x_2}}\frac{(x_{2i} - \bar{x}_2)}{\sqrt{S_{x_2x_2}}} + \cdots + \beta_k\sqrt{S_{x_kx_k}}\frac{(x_{ki} - \bar{x}_k)}{\sqrt{S_{x_kx_k}}} + e_i$$

and by letting $\beta_r\sqrt{S_{x_rx_r}} = \gamma_r$ and $u_{ri} = x_{ri}\sqrt{S_{x_rx_r}}$. This gives

$$\bar{u}_r = \frac{1}{n}\sum_{i=1}^{n} u_{ri} = \frac{\bar{x}_r}{\sqrt{S_{x_rx_r}}}$$

and, after substitution, the model becomes

$$y_i = \alpha + \gamma_1(u_{1i} - \bar{u}_1) + \gamma_2(u_{2i} - \bar{u}_2) + \cdots + \gamma_k(u_{ki} - \bar{u}_k) + e_i$$

Hence, by standardising the x variables as above, the form of the model is retained but it is re-written in terms of a set of variables with the property

$$\sum_{i=1}^{n}(u_{ri} - \bar{u}_r)^2 = \frac{\sum_{i=1}^{n}(x_{ri} - \bar{x}_r)^2}{S_{x_rx_r}} = 1$$

Thus, the least squares estimates of $\gamma_1, \gamma_2, \ldots, \gamma_k$ are given by the solution of

$$\begin{bmatrix} S_{u_1y} \\ S_{u_2y} \\ \vdots \\ S_{u_ky} \end{bmatrix} = \begin{bmatrix} 1 & S_{u_1u_2} & \cdots & S_{u_1u_k} \\ S_{u_1u_2} & 1 & \cdots & S_{u_2u_k} \\ \vdots & \vdots & & \vdots \\ S_{u_1u_k} & S_{u_2u_k} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \hat{\gamma}_1 \\ \gamma_2 \\ \vdots \\ \hat{\gamma}_k \end{bmatrix}$$

where

$$S_{u_ju_l} = \frac{S_{x_jx_l}}{\sqrt{S_{x_jx_j}S_{x_lx_l}}} \quad \text{and} \quad S_{u_jy} = \frac{S_{x_jy}}{\sqrt{S_{x_jx_j}}} \quad (\text{for } j = 1, 2, \ldots, k \text{ and } l = 1, 2, \ldots, k)$$

This equation may be written as $\mathbf{S}_{uy} = \mathbf{S}_{uu}\hat{\boldsymbol{\gamma}}$.

If it is possible to solve this equation, i.e. if $\mathbf{S}_{uu}^{-1}$ exists, then the estimates of the original regression coefficients $\beta_1, \beta_2, \ldots, \beta_k$ may be determined from $\hat{\beta}_r = \hat{\gamma}_r/\sqrt{S_{x_rx_r}}$.

However, in principal components regression our first objective is to find the eigenvalues and eigenvectors of $\mathbf{S}_{uu}$. Suppose that the eigenvalues are denoted by $\lambda_1, \lambda_2, \ldots, \lambda_k$ and that their associated eigenvectors are denoted by $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ respectively. Then, these eigenvalues and eigenvectors must satisfy

$$\mathbf{S}_{uu}\mathbf{v}_r = \lambda_r\mathbf{v}_r \qquad (\text{where } \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_k)$$

and

$$\mathbf{v}_r'\mathbf{v}_r = 1$$

There are many efficient numerical techniques for finding these eigenvalues and eigenvectors, especially as $\mathbf{S}_{uu}$ is a symmetric matrix, and most computers have at least one eigenvalue routine amongst their software.

As λ_1 is the largest eigenvalue, its associated eigenvector $\mathbf{v}_1$ has the interpretation that $z_1 = (u_1, u_2, \ldots, u_k)\mathbf{v}_1$ is the linear combination of the variables $u_1, u_2, \ldots, u_k$ which shows maximum variation amongst the data. The next largest eigenvalue is λ_2 and its associated eigenvector $\mathbf{v}_2$ has the interpretation that $z_2 = (u_1, u_2, \ldots, u_k)\mathbf{v}_2$ is the linear combination of the variables $u_1, u_2, \ldots, u_k$ which, amongst those linear combinations which are uncorrelated with z_1, shows maximum variation amongst the data. The eigenvalue λ_3 and eigenvector $\mathbf{v}_3$ have a similar interpetation relative to z_2 and z_1, and so on.

Proponents of principal components regression suggest that, in most problems, the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ fall into three groups: those substantially greater than zero, those slightly greater than zero and those precisely zero (except for rounding error).

The existence of eigenvalues near zero will indicate that the inverse of $\mathbf{S}_{xx}$ probably does not exist. If the inverse of $\mathbf{S}_{xx}$ does not exist, then equation (35) cannot be solved and, hence, the basic method outlined in Section 2.2 will fail. The cause of this would be the existence of an inter-relationship amongst some of the variables $x_1, x_2, \ldots, x_k$ (for example, a relationship such as $x_1 + x_2 = x_3 + x_4$), i.e. some of the variables $x_1, x_2, \ldots, x_k$ are linearly related.

If an eigenvalue λ_r is precisely zero, then this would imply that $z_r = (u_1, u_2, \ldots, u_k)\mathbf{v}_r = \text{constant}$, the constant being $(\bar{u}_1, \bar{u}_2, \ldots, \bar{u}_k)\mathbf{v}_r$. Hence, by investigating the zero eigenvalues, it is possible to determine the nature of the relationships which exist between $x_1, x_2, \ldots, x_k$.

Potentially, there are k variables, $z_1, z_2, \ldots, z_k$, each succeeding one summarising slightly less of the variation in the data. In a principal components regression analysis, we would discard those z variables whose associated eigenvalues were nearly or precisely zero. Thus, retaining only $z_1, z_2, \ldots, z_p$, our regression equation would become

$$y_i = \delta_0 + \delta_1(z_{1i} - \bar{z}_1) + \delta_2(z_{2i} - \bar{z}_2) + \cdots + \delta_p(z_{pi} - \bar{z}_p) + e_i \qquad (63)$$

Because of the orthogonal nature of eigenvectors, estimates of the parameters $\delta_0, \delta_1, \ldots, \delta_p$ are given by the very simple equations

$$\hat{\delta}_r = \frac{1}{\lambda_r} \sum_{i=1}^{n} (y_i - \bar{y})(z_{ri} - \bar{z}_r) \qquad (\text{for } r = 1, 2, \ldots, p)$$

and

$$\hat{\delta} = \bar{y}$$

It may be shown that estimates of the regression coefficients $\gamma_1, \gamma_2, \ldots, \gamma_k$ are given by

$$\begin{bmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \\ \\ \hat{\gamma}_k \end{bmatrix} = \mathbf{v}_1\hat{\delta}_1 + \mathbf{v}_2\hat{\delta}_2 + \cdots + \mathbf{v}_p\hat{\delta}_p$$

However, in a situation in which certain of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_k$ are precisely zero, it is as well not to place any great reliance on the interpretation of the individual regression coefficients of the x variables. For example, suppose that $k = 4$ and that the fourth eigenvalue was zero, revealing that $x_1 + x_2 = x_3 + x_4$. From a principal components regression analysis, suppose that we have deduced a regression equation $y = x_1 + 2x_2 + x_3 - 2x_4$. Using the known relationship between the xs, this would give the equation $y = x_2 + 2x_3 - x_4$.

These two equations would be equally good at predicting y, but clearly the coefficients of the x variables on their own are virtually meaningless and can be varied at will.

As a further consequence of the orthogonal nature of eigenvectors, the variances of the regression coefficients are given by the simple equation

$$\text{Var}(\hat{\delta}_r) = \sigma^2/\lambda_r \qquad (\text{for } r = 1, 2, \ldots, p)$$

and an unbiased estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n-p-1}\left[\sum_{i=1}^{n}(y_i - \bar{y})^2 - \lambda_1\hat{\delta}_1^2 - \lambda_2\hat{\delta}_2^2 - \cdots - \lambda_p\hat{\delta}_p^2\right]$$

As with the orthogonal polynomial models in Subsection 2.5.1, an analysis of variance table may be established for testing the hypotheses $\delta_1 = 0$, $\delta_2 = 0$, etc.

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*
First principal component	$\lambda_1\hat{\delta}_1^2$	1	
Second principal component	$\lambda_2\hat{\delta}_2^2$		$\left(=\frac{\text{Sum of squares}}{\text{Degrees of freedom}}\right)$
pth principal component	$\lambda_p\hat{\delta}_p^2$		
Residual	By subtraction	$n-p-1$	
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$	

A $100\alpha\%$ significance test would accept $\delta_r = 0$ whenever

$$\frac{r\text{th Principal component mean square}}{\text{Residual mean square}} < F(1, n-p-1, 1-\alpha)$$

In order to predict a value of y for specified values of $x_1, x_2, \ldots, x_k$, it would be necessary to evaluate $z_1, z_2, \ldots, z_p$ and then to calculate

$$\hat{y} = \bar{y} + \hat{\delta}_1(z_1 - \bar{z}_1) + \hat{\delta}_2(z_2 - \bar{z}_2) + \cdots + \hat{\delta}_p(z_p - \bar{z}_p)$$

When the mean value of y is being predicted, this estimate would have variance

$$\sigma^2\left(\frac{1}{n} + \sum_{r=1}^{p}\frac{1}{\lambda_r}\right)$$

However, when the outcome of a single observation of y is being predicted, the above variance should be increased by σ^2.

A corresponding $100(1-\alpha)\%$ confidence interval for the mean value of y would be

$$\hat{y} \pm t(n-p-1, 1-\alpha/2)\sqrt{\hat{\sigma}^2\left(\frac{1}{n} + \sum_{r=1}^{p}\frac{1}{\lambda_r}\right)}$$

Ridge regression was introduced by Hoerl and Kennard (1970) and it involves altering equation (35) to

$$\mathbf{S}_{xy} = (\mathbf{S}_{xx} + \mathbf{W})\hat{\boldsymbol{\beta}}$$

where

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & 0 \\ 0 & w_2 & 0 \\ \vdots & \cdot & \cdot \\ 0 & 0 & w_k \end{bmatrix}$$

The reason for this modification is, once again, an attempt to overcome difficulties of collinearity, although other interpretations are available. Thus, estimates of the regression parameters are given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{S}_{xx} + \mathbf{W})^{-1}\mathbf{S}_{xy} \tag{64}$$

These estimates are no longer unbiased but, under most conditions, it will be possible to choose $\mathbf{W}$ so that these estimates have smaller mean square error (average of the squared differences between estimated and correct values of $\boldsymbol{\beta}$) than the least squares estimates.

In practice, it is necessary to decide on some value for $\mathbf{W}$. To start with, we will assume that $w_1 = w_2 = \cdots = w_k = w$ and make use of a 'ridge trace'. A ridge trace involves evaluating $\hat{\boldsymbol{\beta}}$ in equation (64) for a range of values of $w(\geq 0)$ and plotting the individual coefficients against w on a single graph. A typical example with $k = 5$ is shown in Figure 12.

A point where the curves are beginning to flatten out, such as $w = 0.4$, would be taken as a reasonable value of w by the proponents of ridge regression. It is usual to take as small a value of w as possible so as to keep $\hat{\boldsymbol{\beta}}$ close to the least squares estimate (36).

An alternative method is to make use of the principal components mentioned earlier in this section. For this method, the regression equation is written as in equation (63) and the ridge regression technique is applied to this model. Hence, using an analogous notation, we obtain the equation

$$\mathbf{S}_{zy} = (\mathbf{S}_{zz} + \mathbf{W})\hat{\boldsymbol{\delta}} \tag{65}$$

A suggested choice for w_i is σ^2/δ_i^2 (for $i = 1, 2, \ldots, k$).

Since neither σ^2 nor $\delta_1, \delta_2, \ldots, \delta_k$ are known, it is necessary to obtain first estimates of these parameters. These initial estimates may be obtained from the usual principal components regression analysis. New estimates of $\boldsymbol{\delta}$ are formed by solving equation (65) and these estimates are used to calculate new values for $w_1, \ldots, w_k$ which, in turn, are used in equation (65) to re-estimate $\boldsymbol{\delta}$. The procedure continues until values of w_i stabilise.

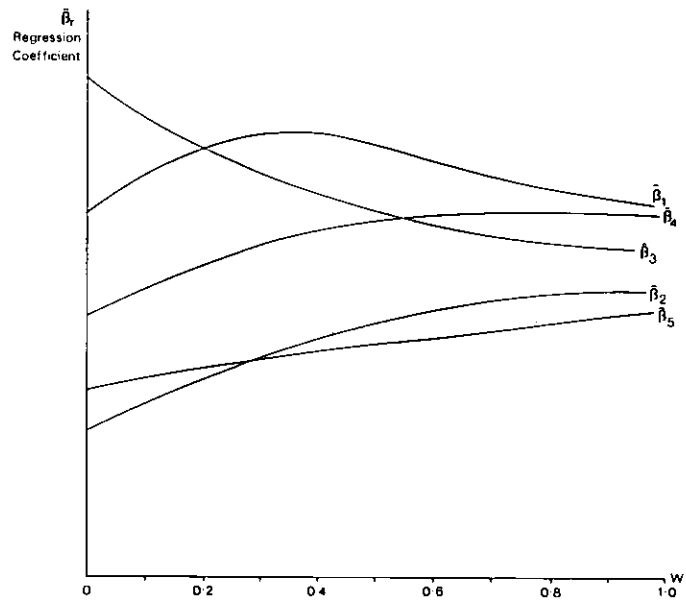


Fig. 12. A typical ridge trace.

However, as the estimates obtained are biased estimates and the bias is a function of the unknown regression parameters, it is not a straightforward matter to proceed further with confidence intervals or tests of significance on the parameters.

2.6.4 Bayesian methods

There is an interesting link between ridge regression and certain Bayesian approaches to the problem.

If we assume that $\beta_1, \beta_2, \ldots, \beta_k$ have normal prior distributions in which β_i has mean zero and variance σ_i^2, and that the β_is are independent, then the posterior distribution of $\boldsymbol{\beta}$ has mean $(\mathbf{S}_{xx} + \mathbf{W})^{-1}\mathbf{S}_{xy}$ where $\mathbf{W}$ is defined in equation (64) but with the added restriction that $w_i = \sigma^2/\sigma_i^2$, the ratio of the error variance to the 'uncertainty variance' in the prior distribution of β_i (for $i = 1, 2, \ldots, k$).

Two additional pieces of information are available from the Bayesian approach. The first is that the variance covariance matrix of the posterior distribution of $\boldsymbol{\beta}$ is given by $\sigma^2(\mathbf{S}_{xx} + \mathbf{W})^{-1}$. The second is an extension to equation (64). If the prior distribution of β_i is known to have mean β_i^0, and $\boldsymbol{\beta}_0' = [\beta_1^0, \beta_2^0, \ldots, \beta_k^0]$, then the posterior mean becomes $(\mathbf{S}_{xx} + \mathbf{W})^{-1}(\mathbf{S}_{xy} + \mathbf{W}\boldsymbol{\beta}_0)$.

2.6.5 Functional relationships

By generalising the models of Subsection 1.4.4, we obtain an overall model

$$\text{ideal } y = a + b_1(\text{ideal } x_1) + b_2(\text{ideal } x_2) + \cdots + b_k(\text{ideal } x_k)$$

and

$$\begin{aligned} y \text{ reading} &= \text{ideal } y + e \\ x_1 \text{ reading} &= \text{ideal } x_1 + \delta_1 \\ x_2 \text{ reading} &= \text{ideal } x_2 + \delta_2, \text{ etc.} \end{aligned} \tag{66}$$

If we introduce a variable x_0 which always takes the value unity (and, hence, x_0 reading = ideal x_0) and, furthermore, we denote y by x_{k+1}, then our model becomes $\sum_{r=0}^{k+1} b_r \text{ ideal } x_r = 0$ where $b_0 = a$ and $b_{k+1} = -1$.

Thus, in general, we may consider the model (66) as being of the form

$$\sum_{r=1}^{k} b_r \text{ ideal } x_r = 0$$

where $x_r = \text{ideal } x_r + e_r$.

If we denote the variance of e_r by σ_{rr} and the covariance between e_r and e_s by σ_{rs}, then the matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{1k} \\ \sigma_{12} & \sigma_{22} & \sigma_{2k} \\ \vdots & \cdot & \cdot \\ \sigma_{1k} & \sigma_{2k} & \sigma_{kk} \end{bmatrix}$$

is the variance covariance matrix of $e_1, e_2, \ldots, e_k$.

Suppose that we have n sets of observations on $x_1, x_2, \ldots, x_k$ which are denoted by $(x_{11}, x_{12}, \ldots, x_{1k})$, $(x_{21}, x_{22}, \ldots, x_{2k}), \ldots, (x_{n1}, x_{n2}, \ldots, x_{nk})$ and that $\mathbf{S}_{xx}$ is defined to be

$$\mathbf{S}_{xx} = \begin{bmatrix} S_{x_1x_1} & S_{x_1x_2} & S_{x_1x_k} \\ S_{x_1x_2} & S_{x_2x_2} & S_{x_2x_k} \\ \vdots & \cdot & \cdot \\ S_{x_1x_k} & S_{x_2x_k} & S_{x_kx_k} \end{bmatrix}$$

where

$$S_{x_rx_s} = \sum_{i=1}^{n} (x_{ri} - \bar{x}_r)(x_{si} - \bar{x}_s)$$

Then, estimates of $\mathbf{b} = [b_1, b_2, \ldots, b_k]'$ are given by the latent vector corresponding to the smallest latent root of $|\mathbf{S}_{xx} - \lambda\boldsymbol{\Sigma}| = 0$.

However, as it is extremely unlikely that $\boldsymbol{\Sigma}$ would be fully known in a hydrological problem, no further details of this method are given here.

References

Andrews, D. F. (1972). *Biometrics*, **28**, 125–36.

Daniels, H. E. (1954). *Ann. Math. Stat.*, **25**(3), 499.

Hinick, M. J. and Talwar, P. P. (1975). *J. Am. Stat. Ass.*, **70**, 113–19.

Hoerl, A. E. and Kennard, R. W. (1970). *Technometrics*, **12**, 55–67.

Pearson, E. S. and Hartley, H. O. (1972). *Biometrika Tables for Statisticians*, Vol. 1. Cambridge University Press.

Chapter 3

BEFORE A MULTIPLE REGRESSION ANALYSIS

3.1 What to Include and Why

3.1.1 Why is the analysis being conducted?

At some stage it will be advisable to consider just why a multiple regression analysis is being attempted and precisely what are the objectives. This salutary exercise is best carried out before any analysis is performed, for the following reasons:

(a) It will probably influence the manner in which the analysis is conducted.
(b) It will anticipate problem areas and some precautions may then be taken before the analysis is attempted.

We have set out already some objectives (Subsections 1.1.1 and 2.1.1) which might lead an investigator to apply a simple or multiple linear regression analysis and, for the purposes of the rest of this section, these are best condensed into:

(1) Prediction of y for specified values of $x_1, x_2, \ldots, x_k$ which are within the region of observed values.
(2) Prediction of y for specified values of $x_1, x_2, \ldots, x_k$ which are outside the region of observed values.
(3) Investigation into which of the variables, $x_1, x_2, \ldots, x_k$, influence y and the nature of any relationships that may exist.

Objective 1 is probably the easiest to tackle since the basic methods outlined in Sections 1.2 and 2.2 are usually sufficient. There may be some gain in precision from using the methods of Section 2.3 to eliminate variables but this gain must be balanced against the problems mentioned below in association with Objective 3. The methods described in Sections 2.4, 2.5 and 2.6 may, of course, be used if they are appropriate.

Objective 2 is far more hazardous. Although the methods of Sections 1.2 and 2.2 again form the basis for estimation, some assessment of the stability of the estimated relationship should be made before any reliance is placed on prediction. In such a situation, it is usually unwise to attempt to eliminate variables.

Objective 3 will require using the methods of estimation given in Sections 1.2 or 2.2 together with the techniques of elimination of variables given in Section 2.3. However, there is the danger that an important variable may not have been observed and this would consequently distort the estimated relationship and any conclusions drawn from it. Another problem is that high correlations between the x variables may lead to misinterpretation, particularly with methods from Section 2.3.

3.1.2 Which independent variables should be used?

When considering which independent variables to measure, it is worth aiming for the following ideals, even though it is rarely possible to achieve them.

The first ideal is that variables which are not highly correlated should be selected wherever possible since high correlations lead to problems of estimation due to singularity of $\mathbf{S}_{xx}$ and to problems of interpretation when using the methods of Section 2.3. If some of the x variables are measuring a similar quantity, then it will usually be better to replace them with a single variable formed by a simple combination of them. Alternatively, a principal components analysis as mentioned in Subsection 2.6.3 may suggest a combination of these variables which is worth using on its own. When the x variables naturally fall into groups according to the nature of the measurement being made, e.g. climatic variables, surface drainage variables, seasonality factors, etc., it may be better to perform a separate principal components analysis on each of these sets. This might then lead to a single combined climate variable, a single surface drainage variable, etc. and each of these may then be used in the regression model.

The second ideal is that variables should be selected in such a way so as to ensure that the regression parameters have some 'physical' interpretation, as well as a statistical one. This will give interpretation to any tests mentioned in Section 2.3 and, in particular, it may enable the investigator to see why certain variables might be eliminated from the regression equation. Equally well, it may enable the investigator to insist that certain variables are retained in the model because of a known causal relationship between the dependent variable and the independent variable in question.

The third ideal must be to include all the important variables, i.e. to get the model right! This is particularly important with Objective 2, where a meaningful relationship must be established within the experimental region before it is at all likely to be valid outside that region. As already mentioned, elimination of variables from a regression equation, using techniques described in Section 2.3, is liable to lead to error when an important variable has been omitted. In particular, if the omitted variable happens to be causally related to y, and if some of the included variables are correlated to the omitted one and consequently also correlated to y, then quite misleading inferences might be drawn about the effect of the included variables. These inferences would, at best, only be valid within the experimental region.

In contrast, it is quite often useful to include a variable which is known to be unrelated to the dependent variable y. Confirmation of its redundancy, from the analysis of Sections 2.2 and 2.3, is some check on both numerical and

inferential procedures. The converse might suggest either numerical inaccuracy or that an important variable has been omitted and that spurious relationships are being generated as a consequence of this.

Instead of including all independent variables as values of $x_1, x_2, \ldots, x_k$, it may be useful to use particular values of one or more of these variables to break the data into groups. For instance, consider the simple case of a model with two independent variables, x_1 and x_2, where x_1 is causally related to y and x_2 is associated with y, but not causally (a nuisance variable). By grouping the data in such a way that, within a group, all values of x_2 are identical, it will be possible to perform separate regressions of y on x_1 for each of the groups. The stability of the regression equation in different circumstances may then be assessed from the variability of the fitted regressions from group to group. Information about the effect of x_2 on y will be temporarily lost, but this should not be important when it is only a nuisance variable. The idea may be extended to grouping the data in such a way that values of two or more variables are constant within one group.

In the simple case of two independent variables, our overall model would be $y = a + b_1x_1 + b_2x_2$. When the model $y = a + b_1x_1$ is used for each group, this should lead to b_1 being the same for each group, and to a (which is $a + b_2x_2$ in the overall model) varying from group to group. However, if b_1 does vary from group to group, then this might mean that the effect of x_1 is influenced by the level of x_2 and that a model of the form $y = a + b_1x_1 + b_2x_2 + b_3x_1x_2$ might be more satisfactory.

This raises another major problem on independent variables; having decided which variables are to be included, how should these variables be introduced into the model? The possibilities are enormous. For example, just from two variables, we might have the model $y = a + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1/x_2 + b_5 \log x_1$, and many more terms could have been included. With n observations of y, it is not difficult to dream up $(n-1)$ artificial variables (like those given in the above example) which would lead to a residual sum of squares of zero, i.e. a perfect fit of model to data. However, it is unlikely that there would be any meaning in such a model and, perhaps more important, it is quite possible that such a model would be quite inadequate with a new set of data. In other words, the model would provide no insight into the relationship between the dependent variable and factors influencing it.

Selection of the types of functions of variables to be introduced into the model must be made bearing in mind the practical interpretation of those functions. Methods of choosing a suitable transformation of a variable will be given in Section 3.3, but these should be treated with caution as they will lead to a model which is only valid within the region of observed values. When an attempt is made to include powers and products of powers of x variables (such as x_1^2, $x_1^2x_2x_3$, $x_1^5x_4$, etc.), this must be done in an ordered manner and with restraint. The remarks made earlier in Subsection 2.5.1 about the order of testing for polynomial models are also applicable in this context. It is usually sensible to begin by enquiring into the contribution made by the high order terms and then to eliminate these terms successively until a satisfactory model is reached. A danger of including too many terms in the model is that an artificially small residual sum of squares could be produced simply because of

the large number of parameters. However, from significance tests, it might appear that all of the included terms were vitally important.

The inclusion of products of certain of the variables implies some joint (interactive) effect of these variables on y. Thus, it may be informative to use grouping of the data to examine the relationship between these variables. For example, inclusion of the term x_1x_2 would imply that x_1 and x_2 had some joint effect on y. By grouping the data according to x_2, it would be possible to examine the nature of the relationship between x_2 and the regression coefficient of x_1. If it is found that they are linearly related, then inclusion of the term x_1x_2 would be justified (giving an overall model $y = a + b_1x_1 + b_2x_2 + b_3x_1x_2$). But if, for instance, a quadratic relationship is revealed, then it would be appropriate to include a term $(a + bx_2 + cx_2^2)x_1$ (giving an overall model $y = a + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1x_2^2$).

3.2 The Distribution of the Dependent Variable

3.2.1 Requirements of least squares

In Subsections 1.1.2 and 2.1.2, it was stated that, for estimation purposes, the errors e_i must have constant variance and that, for inferential purposes, they must follow a normal distribution. (The assumption of independence will be discussed in Section 3.4). As we are regarding the xs as fixed, this implies that y must have constant variance and follow a normal distribution. However, for a particular dependent variable such as run-off, it will usually never be possible to say whether either of these assumptions is true. We may be able to speculate on the possibility, advance a few theories, carry out tests, examine residuals, etc., but, at best, we will only ever be able to say that we are 'pretty sure' or that 'it's a reasonable assumption'.

Because of the uncertainty of the situation, it is important to consider whether these assumptions are crucial and to investigate the circumstances under which the method of least squares is liable to give misleading estimates and inference. This type of investigation examines the 'robustness' of the particular method, in this case, least squares.

Concerning estimation, least squares estimates remain sensible estimates regardless of the normality assumption. However, as has already been mentioned (Subsections 1.3.3 and 2.4.2), when the variances of the e_is are not constant, the estimates may be relatively imprecise, although still unbiased, and the variances of the estimates will be incorrectly estimated. In the case of simple linear regression when it is assumed that the e_is have variance σ^2, the variance of the slope estimate is given by equation (8), $\text{Var}(\hat{b}) = \sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2$, whereas it should be $\sum_{i=1}^{n}\sigma_i^2(x_i - \bar{x})^2/[\sum_{i=1}^{n}(x_i - \bar{x})^2]^2$ when $e_1, e_2, \ldots, e_n$ have variances $\sigma_1^2, \ldots, \sigma_n^2$. Thus, if large values of x are associated with large variances, then the latter will exceed the former and hence equation (8) will underestimate the true variance of the slope estimate. However, the variances will need to be quite noticeably different before these weighted methods show an appreciable gain in the precision of the estimates.

Although an inconvenience as far as estimating the model is concerned, a relationship between an x variable and the variance may, in itself, be an

important feature of the process being studied. Discovering this feature may be just as informative as actually fitting the model. Consequently, it is worth considering the reasons why the variances are changing before either ignoring the problem or trying to use some transformation to eliminate it.

One problem associated with the distribution of the e_is is that of asymmetry, i.e. positive and negative errors of the same magnitude not being equally likely. This problem has an unpleasant effect both on least squares estimates and on most other estimation procedures. It is usually advisable to reduce asymmetry as much as possible by a transformation of the data.

Another problem is that of outliers, i.e. values of e which are exceptionally large and more frequent than the normal assumptions would suggest. Outliers may be produced by gross errors of observation or by undetected changes (or instability) in the phenomena being studied. In attempting to detect the presence of outliers, it is easy to confuse their possible existence with that of asymmetry. Consequently, it is advisable to eliminate this problem whenever possible and techniques which help to detect outliers are given in Section 4.3 (see also Subsection 2.6.2).

Frequently, the omission of an important variable from the model is the cause of apparent variation in the variances of the e_is, asymmetry or even outliers, as discussed in Subsection 3.1.2. However, the same problems may also appear when a linear model is inappropriate or when polynomial terms which should have been included have been omitted.

Concerning tests of significance and confidence interval statements, the assumption of normality for the e_is is more crucial. However, although this implies that y must be normally distributed (as the independent variables are regarded as fixed), it appears that if either y or some of the x variables are near normally distributed, then the tests of significance are not misleading. In other words, the tests of significance appear to be insensitive to non-normality in y whenever the xs themselves come from a near normal distribution. On the other hand, if the xs do not come from a near normal distribution and if some x values are very different in magnitude from the remainder, then the tests of significance are very sensitive to non-normality in y.

If the variances of the e_is are not constant and the basic method of Section 2.2 is applied, then σ^2 will be incorrectly estimated and, consequently, the variances of the regression coefficients will be incorrectly determined. This will lead to errors in tests of significance, the magnitude of these errors depending on the relative magnitudes of the variances of the e_is. Typically, with ratios of 3:1, a nominal 5% test of significance may correspond with only 15% significance. Thus, although unequal variances do not have a serious effect on the regression coefficient estimates, they may seriously distort any inference drawn from tests of significance, in particular, in any of the variables selection procedures described in Subsections 2.3.3 to 2.3.7.

3.2.2. Evidence to justify or question the assumptions

It would be unwise to assume that, for instance, log (run-off) always follows a normal or near normal distribution just because in a few studies such an assumption has been justified. There are no simple rules which govern the

distribution of a quantity in all contexts. Nevertheless, relevant past experience is extremely useful in assessing the possible validity of the least squares assumptions and, wherever possible, reputable studies similar to the one being undertaken should be examined for supporting evidence and counter evidence.

If the variable being studied either is actually calculated from the average of several other variables or could conceptually be regarded as such, then the central limit theorem may be of help. Roughly, this states that the distribution of the arithmetic mean of a set of random variables tends to the normal distribution as more and more random variables are included in that mean, provided that the values of one of the random variables do not dominate those of the others. For example, the distribution of rainfall values will usually become nearer to the normal distribution as the time base is increased. On a daily basis, there are many zero readings and the data is not normally distribution. However, average monthly rainfall calculated as the arithmetic mean of the twelve monthly rainfalls in a year might have a nearly normal distribution provided that it did not, for instance, always only rain in November, or never rain in June. Furthermore, the average rainfall of 48 monthly figures might have a distribution which was nearer to the normal than the distribution of the 12 month average.

If there are obvious restrictions in the range of values that the variable can take, or if certain intermediate values are impossible, or if the variable is discrete, then it may be prudent to question the assumption of a normal distribution. By their nature, many of the variables studied in hydrology, such as river flow, give positive values only and this contrasts with a normal random variable which potentially can take values from $-\infty$ to ∞. However, provided that the majority of values are well above zero (case (a) in Figure 13), this problem may not be of practical significance. On the other hand, if a large proportion of values are just above zero (case (b)), then the assumption of normality may be quite untenable.

Another possibility is that the variable can only take a set of discrete values, for example, 0, 1, 2, 3 and 4. The distribution of this variable will consist of five spikes as shown in Figure 14(a) and this is a long way from the shape of the normal distribution. Equally well, the measuring apparatus which provides the values of our variable may only work in a series of relatively wide steps. A

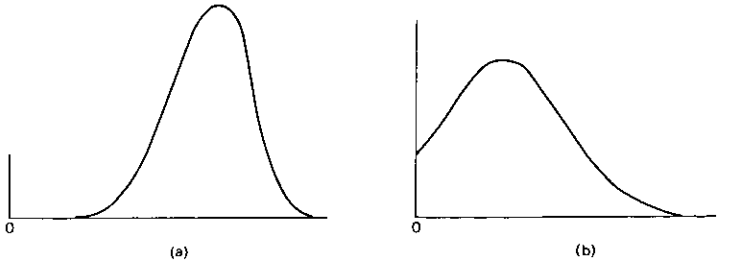


Fig. 13. Distribution of non-negative variables

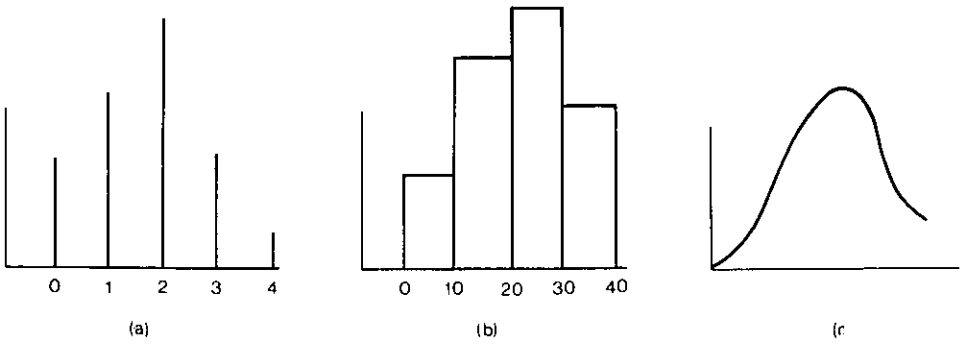


Fig. 14. Discrete, discontinuous and truncated distributions.

typical shape for the distribution of this variable is given in Figure 14(b). Finally, Figure 14(c) shows the shape of a distribution which might result from a variable where values above a certain level could not be recorded. By considering the nature of the reading being made and any instruments being used, it is possible that these gross departures from normality might be detected.

The existence of a lower limit for values of a variable might cause asymmetry in the distribution. For instance, if some values are a long way above the lower limit, but most are not far from the lower limit, then the distribution might be similar to that shown in Figure 15(a).

Figure 15(b) shows another type of asymmetry which has been caused by the distribution of the variable (shown by the continuous line) being composed of a mixture of two distributions (shown by the dotted lines). For example, this might occur in river flow measurements when flow is maintained by artifical means in dry weather (lower distribution) and there is run-off in wet weather (higher distribution). In such a situation, asymmetry might be overcome by dividing the data into dry and wet weather data and fitting separate models to the two sets.

Consideration of the nature of the readings may also help in assessing the assumptions of constant variance. Probably, the situation which can most easily be detected is where the variance increases (or decreases) with the mean. Typically, this may be the result of instrument error increasing with the

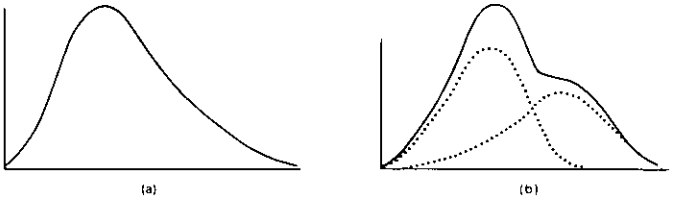


Fig. 15. Asymmetric distributions.

magnitude of the readings or the phenomena being much more variable at high mean levels. An example of the latter would be a situation where there are extremely high levels of run-off (or river flow) which then produce quite different physical phenomena from those experienced at medium and low levels. A transformation may help to overcome this problem but, again, it might be more satisfactory to divide the data into two groups and accept that the two sets of data relate to different phenomena.

When the readings are counts of the occurrence of some phenomenon, it may be that a Poisson distribution is a reasonable model for the readings. A direct consequence of this assumption is that the mean and variance are equal, i.e. a rise in the counts means a more variable count. Again, a transformation might help in this situation.

There is also the possibility of visually assessing the assumption of constant variance. Plotting y against x in the simple linear regression situation was advocated in Subsection 1.4.1 for a variety of reasons. Figure 16 indicates some of the possible outcomes of such an exercise.

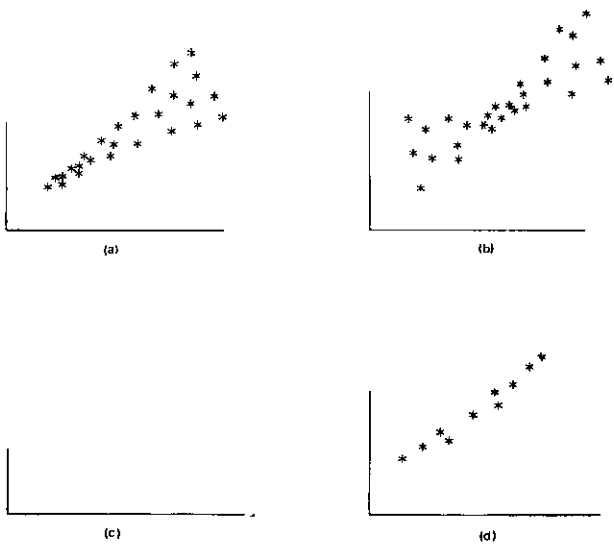


Fig. 16. Some problem sets of data.

Outcome (a) may have been caused by the variance of y increasing with x and outcome (b) by the variance of y increasing as x values become more extreme. For clarity, both graphs show a fairly uniform density of points over the region of x values. However, in practice the density might vary (reflecting the relative scarcity of observations at various x values) and it is easy to misinterpret density varying with x as variance varying with x.

Outcomes (c) and (d) are a reminder of other problems, nonlinearity and outliers respectively. Again, each may be assessed with a graph, but consideration of the nature of the readings is also important. Existence of an asymptote or a maximum (or minimum) value of y may invalidate the use of a linear model, particularly when readings are taken near these limits.

Assessment of the combination of values which the independent variables actually take may suggest instability in the process or phenomenon being studied which, as a consequence, might cause a wild y reading (an outlier). Where several independent variables have been recorded, a plot of y against each x in turn, as suggested in Subsection 2.6.1, may arouse certain suspicions which might be confirmed by analysis of the residuals (see Section 4.3).

3.2.3 Tests of the assumptions

Only when the data consists of several values of y, all with the same set of values for the independent variables, will it be possible to apply any of the above tests of the assumptions before a regression analysis. Although this may occur in simple linear regression, it is unlikely to happen in multiple regression. However, it may be possible to form groups of y values with similar values for the independent variables and, provided that not too much confidence is placed on the outcome of the tests of the assumptions, some useful information may be gained from them.

If it is possible to form n groups of y values with r_i in the ith group, then the test of equality of variances given in Subsection 1.3.2 may be applied to these data, but note must be taken of the sensitivity of this test to non-normality.

There are many tests for assessing normality which use either just a single sample of y values or several samples of y values. As non-normality can take many forms, it is not a simple matter to recommend just one of these tests.

In the case where just a single sample of y values is tested for normality, the values of y are denoted by $y_1, y_2, \ldots, y_n$ and $\bar{y} = \sum_{i=1}^{n} y_i/n$. A test to detect asymmetry is based on the statistic

$$\sqrt{b_1} = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^3}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{3/2}}$$

However, a test to detect deviations from the 'normal' shape in symmetric distributions is based on the statistic

$$b_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{2}}$$

An alternative to the above is the statistic

$$a = \frac{\sum_{i=1}^{n}|y_i - \bar{y}|}{\sqrt{n\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

These tests are designed to pick out distributions where tails are too long or peaks are too fat.

The distributions of $\sqrt{b_1}$, b_2 and for a for normal samples (and hence the significance points) are given in Table 34, Pearson and Hartley (1972).

An alternative test which makes use of normal order statistics was proposed by Shapiro and Wilk. Details of this test and necessary tables are given in Tables 15–18, Pearson and Hartley (1971). This test appears to have good properties and the practical advantage that tables are available for its use with samples of only three or more y values, whereas $\sqrt{b_1}$ is tabulated for $n > 25$, b_2 for $n > 50$ and a for $n > 11$.

Pearson and Hartley (1971) also give details of combined tests of normality using several samples of y values.

Graphical methods may also be used for assessing normality. The following method requires the use of normal probability paper. The values $y_1, y_2, \ldots, y_n$ are arranged in increasing order of magnitude, $y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)}$, and the quantity $[(i - \frac{1}{2})/n]$ is used as an estimate of the distribution function at $y_{(i)}$. The scales of normal probability paper are arranged so that a plot of $y_{(i)}$ against $100[(i - \frac{1}{2})/n]$ (for $i = 1, 2, \ldots, n$) will produce a set of points which, roughly, form a straight line whenever $y_1, y_2, \ldots, y_n$ come from a normal distribution.

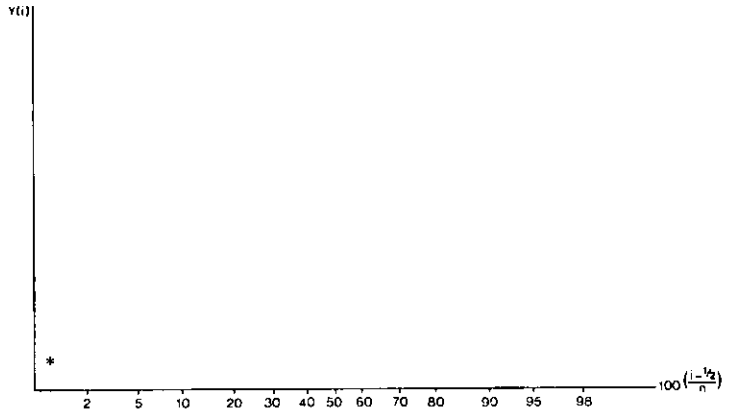


Fig. 17. Use of normal probability paper.

This is illustrated in Figure 17. The points will rarely fall exactly on a straight line but a visual impression of linearity–nonlinearity (normality–non-normality) may be formed from the graph. There is some controversy about the use of $[(i - \frac{1}{2})/n]$ and some statisticians prefer $(i - 0.3)/(n + 0.4)$ but this is only crucial when the plot is being used for estimation purposes. Table 9, Pearson and Hartley (1971), may be used for a similar graphical assessment when normal probability paper is not available.

Detecting an outlier in a sample of n values is equivalent to detecting a special type of non-normality. Thus, not surprisingly, the tests for non-normality mentioned earlier in this subsection also tend to be used to detect the presence of outliers. Another test worth mentioning is based on the statistic

$$u = \frac{y_{(n)} - y_{(1)}}{\sqrt{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right) \Big/ (n-1)}}$$

where $y_{(n)}$ and $y_{(1)}$ are the largest and smallest y values respectively. Table 29c, Pearson and Hartley (1972), gives the distribution of u from which significance points may be determined. One advantage of using this criterion is that it points to a particular value as being the outlier.

If it is suspected that more than one value is grossly in error, i.e. there is more than one outlier, then the above test may be re-applied to the sample with the first outlier omitted. However, it is quite possible that the presence of several outliers will be missed by this method. A test which copes with this problem is called Grubbs Test. It involves ordering the observations, $y_{(1)} \leq y_{(2)} \leq y_{(3)} \leq \cdots \leq y_{(n)}$, and calculating the statistic

$$L_k = \frac{\sum_{i=1}^{n-k} (y_{(i)} - \bar{y}_k)^2}{\sum_{i=1}^{n} (y_{(i)} - \bar{y})^2}$$

where

$$\bar{y}_k = \frac{1}{n-k} \sum_{i=1}^{n-k} y_{(i)} \qquad \text{and} \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_{(i)}$$

This quantity is designed to be used for testing whether the k largest values are outliers and tables of the distribution of L_k are given in Tietjen and Moore (1972). A simple modification allows for a test of whether the k smallest values are outliers. In the same paper, a statistic is also given for testing whether the k values furthest from the mean (above or below) are outliers.

A recent book which covers the topic of outliers from regression models as well as outliers in general is Hawkins (1980).

3.3 Transformations

3.3.1 Variance stabilising transformations

Most variance stabilising transformations exploit a known or an observed relationship between the mean and the variance of the dependent variable. As already mentioned in Subsection 3.2.2, it may be possible, by considering the nature of the variable, to anticipate a relationship between its mean and its variance. Alternatively, plotting y against x may give some empirical evidence of a relationship.

If we are in the fortunate position of having repeated y values under similar conditions, i.e. all with the same set of values for the independent variable(s) (as described in Subsection 3.2.3), then we will be able to estimate the mean and the variance of each group of data and plot these estimates against each other. From the resulting graph, we will be able to assess a possible relationship between the mean and the variance. Thus, if we have n groups of y readings, the readings in any one group being taken under similar conditions, and we denote the readings in the ith group by $y_{i1}, y_{i2}, \ldots, y_{ir_i}$, then we may calculate

$$\bar{y}_i = \frac{1}{r_i} \sum_{j=1}^{r_i} y_{ij} \qquad \text{and} \qquad s_i^2 = \frac{1}{r_i - 1} \sum_{j=1}^{r_i} (y_{ij} - \bar{y}_i)^2$$

and plot s_i^2 against $\bar{y}_i$.

Alternatively, as in Subsection 3.2.3, it may be possible to form groups of y values with similar values for the independent variable(s) and, provided that some caution is exercised, to proceed as above. A particular danger is that an important variable may have been omitted from the independent variables studied.

Thus, we might be in the position to assume that the mean of y, denoted by μ, and the variance of y, denoted by σ_y^2, are related by

$$\sigma_y^2 = f(\mu) \tag{67}$$

where the form of f is known. For any transformation $z = g(y)$, we have the approximate relationship

$$\sigma_z^2 = \left[\frac{dg(\mu)}{d\mu}\right]^2 f(\mu)$$

where σ_z^2 is the variance of $z = g(y)$. We may now choose the function g so as to make σ_z^2 independent of μ. This would be achieved by choosing a function g which satisfied

$$g(\mu) \propto \int \frac{d\mu}{\sqrt{f(\mu)}}$$

For example, if a plot of estimated standard deviation s_i against $\bar{y}_i$ produced a straight line through the origin, then we might assume the form of relationship (67) to be

$$\sigma_y^2 = (b\mu)^2$$

Our choice of transformation to stabilise the variance would have to satisfy

$$g(\mu) \propto \int \frac{d\mu}{b\mu} = \frac{1}{b} \log_e \mu$$

Thus, by taking the transformation $z = \log_e y$, the variance of our transformed variable would be approximately

$$\sigma_z^2 = \left[\frac{1}{\mu}\right]^2 [b\mu]^2 = b^2$$

The table on page 84, taken from Bartlett (1949), summarises some of the more usual transformations.

3.3.2 Transformations to normality and linearising transformations

When the phenomenon that we wish to study as the dependent variable does not have a natural underlying measurable scale, we may only be able to arrange different states of the phenomenon in order. If there are n different states of the phenomenon being studied, then these could be arranged in order and the numbers 1, 2, 3, . . ., n (their rank) associated with the different states. For example, the three weather conditions, dry, drizzle, heavy rain, might be replaced by the numbers, 1, 2, 3. However, when y can only take three possible values, the distribution will not bear much resemblance to the normal distribution.

Distribution of y	*Variance in terms of mean, μ*	*Transformation*	*Approximate variance on new scale*
Poisson	μ	$\sqrt{y}$	0.25
Binomial (proportion)	$\mu(1-\mu)/n$	$\sin^{-1}\sqrt{y}$	$1/4n$
Negative binomial	$\mu+a\mu^2$	$\frac{1}{\sqrt{a}}\sinh^{-1}\sqrt{ay}$	0.25
Empirical	$a\mu$	$\sqrt{y}$	$1.4a$
Empirical	$a\mu^2$	$\log_e y$	a
		$\log_{10} y$	$0.189a$
Empirical	$b\mu+a\mu^2$	$\frac{1}{\sqrt{a}}\sinh^{-1}\sqrt{ay}$	$0.25b$

The use of expected normal scores assumes an underlying normally distributed measurement for the phenomenon. The n values of y (namely 1, 2, 3, ..., n) correspond to n values of the underlying measurement, the smallest of the underlying measurements corresponding to $y=1$, the next smallest to $y=2$, etc. Thus, if we replace $y=1$ by the mean value of the smallest observation in a sample of n values from a normal population, then we will be getting somewhere near to the underlying normal scale. The mean value of the ith smallest observation in a sample of n values from a normal population is

$$z_i = \frac{n!}{(i-1)!\,(n-i)!}\int_{-\infty}^{\infty} X\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,[\Phi(X)]^{i-1}[1-\Phi(X)]^{n-i}dX$$

where

$$\Phi(X) = \int_{-\infty}^{X}\frac{1}{\sqrt{2\pi}}e^{-u^2/2}\,du$$

A transformation to normality is to replace $y=i$ by z_i. Tables of z_i are readily available, for example, Table 28, Pearson and Hartley (1972). The table below sets out the transformation for $n=8$.

y	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Transformed y ($=z_i$)	−1.424	−0.852	−0.473	−0.153	0.153	0.473	0.852	1.424

A more comprehensive set of transformations to normality was suggested by Johnson (1949). Use of these transformations usually requires a knowledge of the mean and variance of y, together with $\sqrt{b_1}$ and b_2, as defined in Subsection

3.2.3. Thus, if we have n values of y, denoted by $y_1, y_2, \ldots, y_n$, taken under similar conditions, then we would need to calculate the following quantities:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$\sqrt{b_1} = \frac{\frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^3}{\left(\frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2\right)^{3/2}}$$

$$b_2 = \frac{\frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^4}{\left(\frac{1}{n}\sum_{i=1}^{n} (y_i - \bar{y})^2\right)^{2}}$$

There are three transformations which are as follows:

$$\text{SB: } z = \gamma + \delta \log(x/(1-x)) \qquad (0 < x < 1)$$

$$\text{SU: } z = \gamma + \delta \sinh^{-1}(x) \qquad (-\infty < x < \infty)$$

$$\text{SL: } z = \gamma + \delta \log(x) \qquad (0 < x < \infty)$$

where $x = (y - \xi)/\lambda$ and γ, δ, ξ and λ are unknown.

These unknowns are estimated using $\bar{y}$, s^2, $\sqrt{b_1}$ and b_2, and the particular transformation which best suits a particular problem is chosen on the basis of $\sqrt{b_1}$ and b_2. The methods of estimating γ, δ, ξ and λ are given in Pearson and Hartley (1971) and Tables 34 and 35 help with the calculation. Alternatively, a computer program is given in Hill, Hill and Holder (1976).

A power transformation which achieves symmetry, but not necessarily normality, may be used when repeated y values, under similar conditions, are available. The p and $(1-p)$ quantiles (those points such that $100p\%$ and $100(1-p)\%$ respectively of sample values fall below them) are determined and denoted by $y_{(p)}$ and $y_{(1-p)}$. The median y_m is also determined. The transformation is y^λ where λ is the solution of

$$\left(\frac{y_{(p)}}{y_m}\right)^\lambda + \left(\frac{y_{(1-p)}}{y_m}\right)^\lambda = 2$$

A suggested choice for p is 0.01. (It is usual to exclude the solution $\lambda = 0$.)

Now, let us consider a situation in which a linearising transformation on the dependent variable might be appropriate. If the dependent variable y is a proportion (or percentage), then its values will be constrained to lie between 0 and 1 (or 0 and 100). Thus, a relationship between y and any independent variables is unlikely to be linear as a linear relationship would not naturally give values constrained to lie between two limits. Consequently, either a

nonlinear model must be used or the proportion must be transformed to a new variable which, potentially, would be able to take values anywhere within the range $(-\infty, \infty)$. One such transformation, the probit transformation, replaces y by z where z is given by

$$\Phi(z) = y$$

The original use of this transformation was in connection with toxicological investigations. A certain dosage was given to a set of animals and the proportion killed was recorded. This was repeated at dosage levels $d_1, d_2, \ldots, d_n$ giving the proportion killed as $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$. A possible theoretical model relating the proportion killed p_i to the dosage d_i is

$$p_i = \int_{-\infty}^{a+bd_i} \frac{1}{\sqrt{2\pi}} e^{-1/2u^2}\, du$$

Thus, if $\Phi(z_i) = \hat{p}_i$, i.e. the probit transformation is applied to the $\hat{p}_i$s, then this would suggest the model

$$z_i = a + bd_i + e_i \tag{68}$$

which is a linear regression relationship between the transformed proportions and the dosages. Unfortunately, the transformed variables do not have constant variance. The variance of z_i is

$$\frac{p_i(1-p_i)}{r_i\left[\frac{1}{\sqrt{2\pi}} e^{-z_i^2/2}\right]^2}$$

where r_i is the number of animals which are given dosage d_i.

The extension of equation (68) to a multiple regression model would be achieved by measuring other variables (besides dosage) which might affect the response of the animal. For instance, weight (w) might be included to give the relationship

$$z_i = a + bd_i + cw_i + e_i$$

In general, the probit model may be used to transform any proportion which is based on a count of the number of occurrences of a particular phenomenon against the number of opportunities that phenomenon had to occur. However, it is important that the outcome at each opportunity is independent of previous outcomes.

An alternative to the probit transformation is the logit transformation. For the logit transformation, the transformed variable z is given by

$$z_i = \log_e\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right)$$

and the variance of z_i is given by

$$\frac{1}{r_i p_i(1-p_i)}$$

The logit transformation has the advantage over the probit transformation of being, computationally, a simpler function to deal with. Furthermore, the model $z_i = a + bd_i + e_i$ implies the theoretical relationship

$$p_i = \frac{1}{1 + e^{-(a+bd_i)}}$$

the logit model, which is mathematically easier to handle than the probit model. A sketch of the logit model is given in Figure 18.

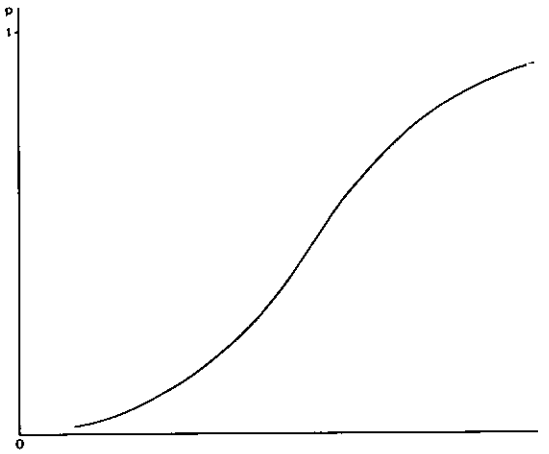


Fig. 18. Logistic curve.

With both the probit and logit transformations, the variance of the transformed variable is a function of p_i and, hence, of the unknown parameters a and b. Thus, it is not possible to calculate the weights w_i, as defined in Subsection 1.3.3, without knowing a and b. Various approximate methods exist for overcoming this problem, such as using first estimates of a and b to calculate the weights, then estimating a and b by using weighted linear regression as described in Subsection 1.3.3, using these new estimates of a and b to recalculate the weights, and so on. However, in these circumstances, the method of least squares differs from the more general method of estimation called maximum likelihood estimation. Fitting a probit model by maximum likelihood estimation is described in Finney (1964). However, the computer package GLIM (Baker and Nelder (1978)) enables maximum likelihood estimates to be computed for unknown parameters in logit models and probit models, as well as many other linear models.

3.3.3 Box–Cox transformations

Box and Cox (1964) suggested the transformations

$$z = \frac{y^\lambda - 1}{\lambda} \qquad (\lambda \neq 0)$$

$$z = \log_e y \qquad (\lambda = 0)$$

to help satisfy the requirements of normality, constant variance and additivity, when fitting linear models (such as model (33), for example). The suggested procedure is to choose a particular value of λ and to calculate the residual mean square $\hat{\sigma}^2$ (given by equation (39) for model (33)) using the transformed variable z as the dependent variable. Then, the quantity $L_b(\lambda)$ is calculated by

$$L_b(\lambda) = (n-k-1)\left(\left(\frac{\lambda-1}{n}\right)\sum_{i=1}^{n}\log_e y_i - \tfrac{1}{2}\log_e \hat{\sigma}^2\right) \tag{69}$$

By varying λ and repeating the above procedure, a graph of $L_b(\lambda)$ plotted against λ may be produced. An example of such a graph is shown in Figure 19.

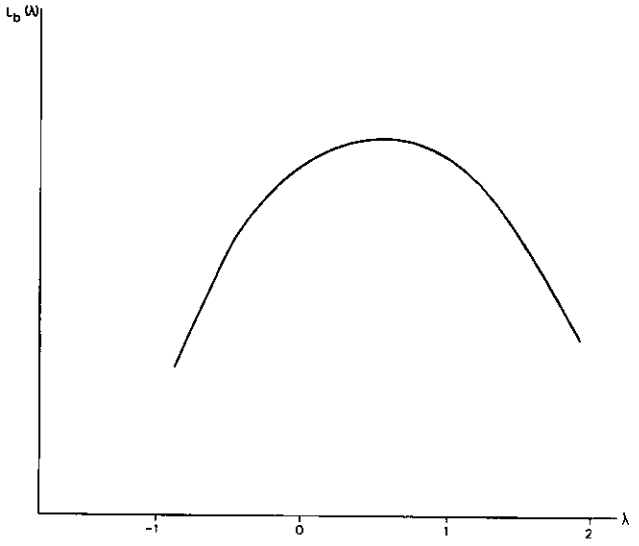


Fig. 19. Box–Cox transformation.

The best choice of λ is taken to be the value which maximises $L_b(\lambda)$. However, if $\lambda = 0$ is the best choice, then $\log_e y$ is taken to be the best transformation.

There will be some trouble with these transformations whenever it is possible for y to be negative. An alternative set of transformations

$$z = \frac{(y+\lambda_2)^{\lambda_1} - 1}{\lambda_1} \qquad (\lambda_1 \neq 0)$$

$$z = \log_e(y+\lambda_2) \qquad (\lambda_1 = 0)$$

was suggested to cope with this problem. The procedure is again to calculate the residual mean square $\hat{\sigma}^2$ with specified values of λ_1 and λ_2 and then to evaluate

$$L_b(\lambda_1, \lambda_2) = (n-k-1)\left(\left(\frac{\lambda_1-1}{n}\right)\sum_{i=1}^{n}\log_e(y_i+\lambda_2) - \tfrac{1}{2}\log_e \hat{\sigma}^2\right)$$

The problem now is to choose λ_1 and λ_2 to maximise $L_b(\lambda_1, \lambda_2)$, but this may be

achieved either graphically or by using some numerical algorithm of maximisation.

It might be useful to be able to test the hypothesis $\lambda = \lambda_0$; in particular, testing the hypothesis $\lambda = 0$ may well be informative. In the case of the simpler one-parameter transformation, a $100\alpha\%$ significance test would reject the hypothesis $\lambda = \lambda_0$ whenever

$$2(L_b(\lambda_{\max}) - L_b(\lambda_0)) > \chi^2(1, 1-\alpha)$$

where $\lambda_{\max}$ is the value of λ which maximises $L_b(\lambda)$.

Transformations need not only be used for the dependent variable. For instance, it might be appropriate to transform $x_1, x_2, \ldots, x_k$ to $x_1^{r_1}, x_2^{r_2}, \ldots, x_k^{r_k}$ to achieve a strong relationship between dependent and independent variables. Similar methods to those just described may be used but there is also a clever iterative scheme which cuts out some computation.

Suppose that the correct transformation for x_1 is thought to be somewhere near to $x_1^{r_1^*}$ (r_1^* could be 1 as a preliminary guess). Then, a Taylor series expansion of $x_1^{r_1}$ would give

$$x_1^{r_1} = x_1^{r_1^*} + (r_1 - r_1^*)x_1^{r_1^*} \log_e x_1 + \cdot$$

Thus, if the variables $x_1^{r_1^*}$ and $x_1^{r_1^*} \log_e x_1$ are included in the regression model, then the estimated regression coefficients of those variables may be used to calculate an estimate for the correct power of x_1. This estimate of r_1 would be given by

$$r_1^* + \frac{\text{estimated regression coefficient of } x_1^{r_1^*} \log_e x_1}{\text{estimated regression coefficient of } x_1^{r_1^*}}$$

This estimate of r_1 may then be substituted for r_1^* and the regression may be repeated to obtain an even better estimate of r_1, and so on.

Similarly, including

$$x_1^{r_1^*}, x_2^{r_2^*}, \ldots, x_k^{r_k^*} \quad \text{and} \quad x_1^{r_1^*} \log_e x_1, x_2^{r_2^*} \log_e x_2, \ldots, x_k^{r_k^*} \log_e x_k$$

in the regression equation would enable improved estimates of $r_1, r_2, \ldots, r_k$ to be computed from preliminary guesses $r_1^*, r_2^*, \ldots, r_k^*$.

A similar technique may be used to determine an appropriate transformation for the dependent variable, but such an exercise should be undertaken with extreme caution.

3.4 Autocorrelation in Multiple Regression

3.4.1 Possible causes and consequences

One of the assumptions made in Subsections 1.1.2 and 2.1.2 is that the errors, $e_1, e_2, \ldots, e_n$, are mutually independent. There are several types of dependent variable for which it would not be immediately apparent that this assumption had any validity. For example, one such variable, of particular relevance in hydrology, is the dependent variable which represents the state of a certain phenomenon in time. The readings $y_1, y_2, \ldots, y_n$ may be daily, monthly,

annual, etc. values of that phenomenon and it is quite likely that y_2 (for instance) will be strongly affected by y_1. However, before we dismiss all time phenomena as inappropriate for regression analysis, it is important to consider the exact implications of the assumption being made.

The error terms $e_1, e_2, \ldots, e_n$ represent the deviations of the observed values of y from their true or model (or expected) value. It is these deviations which we want to assume are independent from one reading to the next. We might automatically assume that two successive y values are correlated because the phenomenon being studied gives similar values from one day to the next, but this does not necessarily violate the required independence assumption. Two successive days readings may be similar purely because their model values are similar and they may still have independent error terms. However, when, for instance, a reading on day 1 which is above its true or model value means that the reading on day 2 will also be above its model value, then the assumption of independence is invalid and the errors are said to be autocorrelated.

The joint use of the terms 'model' and 'true' raises an important point. If we do not get the model exactly right (i.e. the model value is no longer the true value), then the error terms will contain a component which is the part that the model failed to explain. It is then quite likely that deviations from the model in previous observations would give us some idea of the deviation to be expected in the present observation. This would mean that our independence assumption was invalid, but the real cause of this would be that the model was incorrect. Thus, it is important to consider the deviation from the actual model being fitted and not from some ideal model that we would like to be fitting but which is unknown to us. In assessing this problem, it might be easiest to consider whether we have been able to observe all causal variables and whether they have been sensibly included in the model. If they have not been, then it is likely that, unless those particular variables happen to remain constant over the time period studied, the deviations from the model used will not be independent.

If the methods of Sections 1.2 and 2.2 are applied when the errors are not independent, then the consequences are similar to those arising from unequal error variances (see Subsection 3.2.1). In the case of positive correlation between successive y values (i.e. positive errors tending to be followed by positive errors and negative errors tending to be followed by negative errors), simple linear regression leads to the variance of the regression coefficient being underestimated and, to make matters worse, to the estimate of variance (given by equation (10)) also underestimating σ^2. These inaccuracies are reflected in the tests of significance. Typically, with a serial correlation of 0.3 between successive errors and a sample of $n = 11$ pairs of y and x values, what should be a 5% significance test will actually correspond to a real significance of anything between 1.4% and 14.6%. The only good result is that the estimates derived using Section 1.2 or 2.2 are still unbiased when the errors are not independent.

3.4.2 Transformations

To overcome the problem of autocorrelated errors, we will need to assume some model for these errors. Suppose that $y_1, y_2, \ldots, y_n$ are arranged in order

of time. Then, we might assume that

$$e_r = \rho e_{r-1} + \eta_r \qquad (\text{for } r = 2, 3, \ldots, n)$$

where $\eta_r \sim N(0, \sigma_0^2)$ and the ηs are mutually independent. Thus, each error is directly associated with the preceding error. This is called a first order autoregressive process and ρ is called the first order autocorrelation (or serial correlation) coefficient.

In such a situation, it would be sensible to modify $y_1, y_2, \ldots, y_n$ to $y_1\sqrt{1-\rho^2}$, $y_2 - \rho y_1$, $y_3 - \rho y_2, \ldots, y_n - \rho y_{n-1}$. Then, by rearranging the original model (33), we would have the model

$$y_1\sqrt{1-\rho^2} = a\sqrt{1-\rho^2} + \sum_{l=1}^{k} b_l x_{l,1}\sqrt{1-\rho^2} + e_1\sqrt{1-\rho^2}$$

and

$$y_r - \rho y_{r-1} = a(1-\rho) + \sum_{l=1}^{k} b_l(x_{l,r} - \rho x_{l,r-1}) + (e_r - \rho e_{r-1})$$
$$(\text{for } r = 2, 3, \ldots, n)$$

By defining x_1 always to be 1, we may eliminate the need to retain the separate constant term a. The coefficient b_1 will serve the same purpose and we may now omit terms involving a from the above equations.

If we now define $Y_1, Y_2, \ldots, Y_n$ and $X_1, X_2, \ldots, X_k$ to be

$$Y_1 = y_1\sqrt{1-\rho^2} \qquad X_{l,1} = x_{l,1}\sqrt{1-\rho^2}$$

and

$$Y_r = y_r - \rho y_{r-1} \qquad X_{l,r} = x_{l,r} - \rho x_{l,r-1} \qquad (\text{for } r = 2, 3, \ldots, n \text{ and } l = 1, 2, \ldots, k)$$

then our model will become

$$Y_i = \sum_{l=1}^{k} b_l X_{l,i} + w_i \qquad (\text{for } i = 1, 2, \ldots, n)$$

where $w_i \sim N(0, \sigma_0^2)$. This latter step assumes that the phenomenon being studied has been observed for a long period of time. If this is not the case, then Y_1 should be ignored and only $Y_2, Y_3, \ldots, Y_n$ should be used.

The regression coefficients $b_1, b_2, \ldots, b_k$ may now be estimated using the standard methods of Section 2.2 because $w_1, w_2, \ldots, w_n$ are independent.

Clearly, ρ plays a crucial part in this transformation and it may not be obvious what value it should be given. A value frequently chosen is $\rho = 1$, which implies that the differences of successive values are included in the model and that Y_1 disappears from the model. However, this does not necessarily confirm its use in every situation.

One suggested method for choosing ρ is to apply the basic method of Section 2.2 to calculate the residuals $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ using equation (38). An estimate of ρ is then given by

$$\hat{\rho} = \left(\frac{1}{n-1}\sum_{i=1}^{n-1} \hat{e}_i \hat{e}_{i+1}\right) \Big/ \hat{\sigma}^2$$

where $\hat{\sigma}^2$ is given by equation (39).

A more complicated model for the errors is

$$e_r = \rho_1 e_{r-1} + \rho_2 e_{r-2} + \cdots + \rho_p e_{r-p} + \eta_r$$

where again $\eta_r \sim N(0, \sigma_0^2)$ and the η_rs are mutually independent.

If $\rho_1, \rho_2, \ldots, \rho_p$ are known, then the previous method may be extended by defining Y_r and $X_{l,r}$ to be

$$Y_r = y_r - \rho_1 y_{r-1} - \rho_2 y_{r-2} - \cdots - \rho_p y_{r-p}$$

and

$$X_{l,r} = x_{l,r} - \rho_1 x_{l,r-1} - \rho_2 x_{l,r-2} - \cdots - \rho_p x_{l,r-p}$$

However, it is extremely unlikely that $\rho_1, \rho_2, \ldots, \rho_p$ will be known.

A straight forward method which allows $\rho_1, \rho_2, \ldots, \rho_p$ to be estimated from the data is, first of all, to apply the basic method of Section 2.2 to the data and to estimate the residuals $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$. From our model, we know that

$$\begin{aligned} y_n &= a + b_1 x_{1,n} + b_2 x_{2,n} + \cdots + b_k x_{k,n} + e_n \\ &= a + b_1 x_{1,n} + b_2 x_{2,n} + \cdots + b_k x_{k,n} \\ &\quad + \rho_1 e_{n-1} + \rho_2 e_{n-2} + \cdots + \rho_p e_{n-p} + \eta_n \\ y_{n-1} &= a + b_1 x_{1,n-1} + b_2 x_{2,n-1} + \cdots + b_k x_{k,n-1} \\ &\quad + \rho_1 e_{n-2} + \rho_2 e_{n-3} + \cdots + \rho_p e_{n-p-1} + \eta_{n-1} \qquad \text{etc.} \end{aligned}$$

Thus, by defining additional variables $x_{k+1}, x_{k+2}, \ldots, x_{k+p}$ to be

$$\begin{aligned} x_{k+1,i} &= \hat{e}_{i-1} \\ x_{k+2,i} &= \hat{e}_{i-2} \\ &\cdots \qquad (\text{for } i = p+1, p+2, \ldots, n) \\ x_{k+p,i} &= \hat{e}_{i-p} \end{aligned}$$

and substituting $\hat{e}_{i-r}$ for e_{i-r} (for $r = 1, 2, \ldots, p$ and $i = p+1, p+2, \ldots, n$) in the previous formula, we may perform the basic method of Section 2.2 on the $n-p$ observations $y_{p+1}, y_{p+2}, \ldots, y_n$ and the independent variables $x_1, x_2, \ldots, x_{k+p}$ in order to estimate $\rho_1, \rho_2, \ldots, \rho_p$. Although this is not the best of methods available, it has the advantage that it is conceptually simple to grasp and easy to implement.

Although strictly out of place in this section, it is convenient to mention now the problem frequently encountered in hydrology that some of the independent variables are previous values of the dependent variable. Thus, in these circumstances, the model would really be

$$\begin{aligned} y_i &= a + b_1 y_{i-1} + b_2 y_{i-2} + \cdots + b_p y_{i-p} \\ &\quad + b_{p+1} x_{1i} + b_{p+2} x_{2i} + \cdots + b_{p+k} x_{ki} + e_i \end{aligned}$$

It may be shown that, for large samples (large n), the estimates of $a, b_1, b_2, \ldots, b_{p+k}$, derived by the application of the basic method of Section 2.2, have similar properties to those of the estimates of regression parameters in the conventional model (33), provided that the errors $e_1, e_2, \ldots, e_n$ are independent. In small samples, these estimates of $a, b_1, b_2, \ldots, b_{p+k}$ are biased.

However, even for large samples, if the problem of autocorrelation arises, then the estimates of $b_1, b_2, \ldots, b_{p+k}$ begin to become suspect. In particular, they are no longer consistent or unbiased estimates. As the estimated residuals are also inconsistent, there is little opportunity for using them to correct for autocorrelation by the methods described previously. Also, in the case of a first order autoregressive process for the errors, the estimated residuals will show less autocorrelation than is actually present. Thus, certain tests which are discussed in Section 4.3 will underestimate the effect of autocorrelation, i.e. they will be biased towards accepting the hypothesis of independent errors.

References

Baker, R. J. and Nelder, J. A. (1978). *The GLIM System, Release 3.* Numerical Algorithms Group, Oxford.

Bartlett, M. S. (1947). *Biometrics*, **3**, 39–52.

Box, G. E. P. and Cox, D. R. (1964). *J. R. Stat. Soc., Series B*, **26**, 211–46.

Finney, D. J. (1964). *Probit Analysis.* Cambridge University Press.

Hawkins, D. M. (1980). *Identification of Outliers.* Chapman and Hall.

Hill, I. D., Hill, R. and Holder, R. L. (1976). *J. R. Stat. Soc., Series C*, **25**, 180–9.

Johnson, N. L. (1949). *Biometrika*, **36**, 149–76.

Pearson, E. S. and Hartley, H. O. (1972). *Biometrika Tables for Statisticians*, Vol. 1. Cambridge University Press.

Pearson, E. S. and Hartley, H. O. (1971). *Biometrika Tables for Statisticians*, Vol. 2. Cambridge University Press.

Tietjen, G. L. and Moore, R. H. (1972). *Technometrics*, **14**, 583–97.

Chapter 4

AFTER A MULTIPLE REGRESSION ANALYSIS

4.1 Some Preliminary Checks

4.1.1 Examining the form of the regression equation

Before making use of the fitted multiple regression model and before carrying out any elaborate checks on residuals (see Section 4.3), it is as well to apply a few simple checks to the model itself.

The effect of some of the independent variables on the dependent variable may be known. For example, it may be that a rise in the value of one of the independent variables produces a rise in the value of the dependent variable. If this is the case, then a check should be made that it is reflected in the regression equation by the regression coefficient associated with that independent variable being positive. A negative regression coefficient would suggest that a rise in the independent variable produces a fall in the dependent variable.

There will be exceptions to this pattern which will usually occur when there are strong interrelationships between some of the independent variables. Because of the nature of the phenomena being studied, a change in value of one independent variable may imply a change in values of the other independent variables. In this situation, the joint effect of the highly related variables on the independent variable must be considered. Although this may be laborious, it is particularly important as the existence of highly correlated independent variables may lead to problems and inaccuracies in inverting the matrix $\mathbf{S}_{xx}$ (see equation (35)) which, consequently, might lead to a nonsensical fitted regression model (see Section 4.2).

If one of the variable selection methods outlined in Section 2.3 has been used, then it would be wise to consider whether a sensible set of independent variables has been included in the final fitted regression model. Again, it is possible that strong interrelationships in the independent variables may have led to the surprise omission of a variable, but this omission should have been balanced by the inclusion of certain variables with which it is highly correlated.

However, it is as well to approach the assessment of the fitted model with scepticism. As well as the possibility that certain of the assumptions outlined in Subsections 1.1.2 or 2.1.2 may have been violated, the actual recorded data might be nonsensical. For instance, the variables measured, or the

observations taken of those variables, may not accurately reflect the phenomenon that they were intended to record. This may have been caused by instrument inaccuracy, error in observation or simply the fact that the phenomenon was not accurately monitored by the equipment. Assessing the actual physical meaning or implications of the fitted model is one way of assessing the credibility of the data and, hence, the model.

One final possibility is that the calculations required to establish the fitted model may have been performed incorrectly. Even when an apparently trustworthy computer program has been used, certain hidden restrictions (such as array space) may have been violated, certain operating instructions misinterpreted or a hardware (or software) malfunction may have gone undetected.

4.1.2 Examining the behaviour of the regression model

Having carefully examined the fitted regression model, the next step is to consider how it behaves when in use.

If we put observed values of the independent variables into the fitted model (44), then we may calculate predicted values of y using

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1(x_{1i} - \bar{x}_1) + \hat{\beta}_2(x_{2i} - \bar{x}_2) + \qquad + \hat{\beta}_k(x_{ki} - \bar{x}_k)$$
$$(\text{for } i = 1, 2, \ldots, n)$$

which correspond to the observed values of y (denoted by $y_1, y_2, \ldots, y_n$). A graph of $\hat{y}_i$ plotted against y_i will then give an immediate visual impression of the performance of the model. At the same time, it will give the opportunity to detect gross errors in the fitted model. Ideally, the graph should be exactly a straight line which passes through the origin and has a slope of 45°, but, usually, the plotted points will scatter about this line. If the plotted points scatter about some other line (i.e. one which does not pass through the origin or which has a slope other than 45°), then some error in calculation must be suspected. Calculating the quantity $\sum_{i=1}^{n}(\hat{y}_i - y_i)$ gives a direct check on previous calculations since this should always be zero (except for rounding error). However, this is clearly not a particularly powerful check.

Having considered the behaviour of the model for the observed values of the independent variables, it is advisable to examine its performance over the range of values for which it might be used. For this purpose, it is useful to have available other data which have not been used in estimating the model (possibly deliberately left out), but nevertheless consist of values of the dependent variable together with associated values of the independent variables. A graph of predicted y plotted against observed y for these new data will again give a rapid visual impression of the performance of the model.

When examining the predictive ability of the model, consideration should be given to any natural constraints that there might be on the dependent variable. For instance, many hydrological variables, such as river flow, by their nature must be positive. Thus, a model which predicted negative values for riverflow under unexceptional conditions of the independent variables must be treated with extreme caution, if not completely discarded.

4.1.3 Stability of the model

Having derived a fitted model for a particular set of data, the investigator may be left with the uncomfortable feeling that, with another set of data from a similar situation, quite a different fitted model might be generated. The concern is that, for a particular set of data, the model only provides an approximation to the data rather than an explanation of the data. In such a situation where the model has no 'physical meaning', it is helpful to see the possible fluctuations that might occur in the fitted model in other ways than just in terms of the variances of estimated coefficients.

Again, the technique of dividing the data (or 'data splitting') is of value. In Subsection 3.1.2, it was suggested that the stability of the model may be investigated by splitting the data into groups according to the value of one of the independent variables. In Subsection 4.1.2, it was suggested that it may be informative to split off some data and not use them in estimating the model, but use them instead to compare values of y predicted by the model with observed values of y. Another possibility is to split the data randomly into groups and fit the model separately to the different groups of data. From the variations in the different fitted models, an immediate idea may be gained of the stability of the model and, in particular, of which are the most stable factors.

4.2 Problems of Numerical Stability

4.2.1 Numerical methods used in regression

In Subsection 2.2.1, the basic problem of estimating $\beta_1, \beta_2, \ldots, \beta_k$ was presented as the problem of solving the equation

$$\mathbf{S}_{xy} = \mathbf{S}_{xx}\hat{\boldsymbol{\beta}}$$

The solution that was suggested involved finding $\mathbf{S}_{xx}^{-1}$, the inverse of $\mathbf{S}_{xx}$. There are a variety of numerical procedures available for achieving this, some notably more successful than others. A popular method used in several regression computer packages is the Gauss–Jordan elimination method.

However, there are alternative ways of determining $\hat{\boldsymbol{\beta}}$ which have gained in popularity in recent years. If we define the column vector $\mathbf{Y}$ to be

$$\mathbf{Y} = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$$

where $y_1, y_2, \ldots, y_n$ and $\bar{y}$ are as defined in Subsection 2.2.1 and, furthermore, we define the matrices $\mathbf{X}$ and $\mathbf{e}$ to be

$$\mathbf{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{21} - \bar{x}_2 & & x_{k1} - \bar{x}_k \\ x_{12} - \bar{x}_1 & x_{22} - \bar{x}_2 & & x_{k2} - \bar{x}_k \\ \vdots & & & \vdots \\ x_{1n} - \bar{x}_1 & x_{2n} - \bar{x}_2 & & x_{kn} - \bar{x}_k \end{bmatrix} \quad \text{and} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

then the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

is equivalent to model (34). The matrix $\mathbf{S}_{xy}$ is equal to $\mathbf{X}'\mathbf{Y}$ and the matrix $\mathbf{S}_{xx}$ is equal to $\mathbf{X}'\mathbf{X}$. Hence, we can see from equation (36) that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. However, if we define the matrix $\mathbf{Z}$ to be $\mathbf{Z} = \mathbf{X}\mathbf{U}^{-1}$ where $\mathbf{U}$ is an upper triangular matrix which is chosen so that the matrix $\mathbf{Z}$ satisfies $\mathbf{Z}'\mathbf{Z} = \boldsymbol{\Lambda}$ where $\boldsymbol{\Lambda}$ is a diagonal matrix, then $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{U}'\mathbf{Z}'\mathbf{Z}\mathbf{U})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{U}^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{Z}'\mathbf{Y}$$

The matrices $\mathbf{U}$ and $\mathbf{Z}$ may be determined by a numerical procedure called 'modified Gram Schmidt orthogonalisation'. Also, we see from equation (37) that $\mathbf{V}_{\hat{\beta}} = \sigma^2\mathbf{S}_{xx}^{-1}$ and, thus, $\mathbf{V}_{\hat{\beta}}$, the variance covariance matrix of the estimates $\hat{\boldsymbol{\beta}}$, is given by

$$\mathbf{V}_{\hat{\beta}} = \sigma^2(\mathbf{U}^{-1}\boldsymbol{\Lambda}^{-1}(\mathbf{U}')^{-1})$$

Hence, we have $\mathbf{U}\mathbf{V}_{\hat{\beta}}\mathbf{U}' = \sigma^2\boldsymbol{\Lambda}^{-1}$ and, since $\boldsymbol{\Lambda}$ is a diagonal matrix and $\mathbf{U}$ and $\mathbf{U}'$ are upper and lower triangular matrices respectively, this allows the elements of $\mathbf{V}_{\hat{\beta}}$ to be found by back substitution.

An alternative approach makes use of a series of orthogonal transformations on $\mathbf{X}$ to obtain the decomposition $\mathbf{X} = \mathbf{Q}\mathbf{R}$ where the first k rows of $\mathbf{R}_{n\times k}$ are upper triangular and the last $n-k$ rows contain all zeros, and $\mathbf{Q}'\mathbf{Q} = \mathbf{I}_{n\times n}$, the $n \times n$ identity matrix. From the equation $\mathbf{X} = \mathbf{Q}\mathbf{R}$, we have $\mathbf{R} = \mathbf{Q}'\mathbf{X} = \begin{bmatrix}\mathbf{S}\\ \mathbf{O}\end{bmatrix}$ where $\mathbf{S}$ is a $k \times k$ matrix and is upper triangular.

Now, we have

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\
&= (\mathbf{R}'\mathbf{Q}'\mathbf{Q}\mathbf{R})^{-1}\mathbf{R}'\mathbf{Q}'\mathbf{Y} \\
&= (\mathbf{R}'\mathbf{R})^{-1}\mathbf{R}'\mathbf{Q}'\mathbf{Y} \\
&= (\mathbf{S}'\mathbf{S})^{-1}[\mathbf{S}'\mathbf{O}']\mathbf{Q}'\mathbf{Y} \\
&= \mathbf{S}^{-1}(\mathbf{S}')^{-1}[\mathbf{S}'\mathbf{O}']\mathbf{Q}'\mathbf{Y} \\
&= \mathbf{S}^{-1}[\mathbf{I}_{k\times k}\mathbf{O}']\mathbf{Q}'\mathbf{Y} \\
&= \mathbf{S}^{-1}\mathbf{V}_1
\end{aligned}$$

where

$$\mathbf{Q}'\mathbf{Y} = \begin{bmatrix}\mathbf{V}_1\\ \mathbf{V}_2\end{bmatrix}$$

and $\mathbf{V}_1$ is a $k \times 1$ matrix and $\mathbf{V}_2$ is a $(n-k) \times 1$ matrix.

For further details of computational techniques associated with multiple regression, the reader is referred to Seber (1977).

4.2.2 The relative merits of the various numerical methods

There have been several large scale numerical investigations which compare the performance of the wide selection of computer packages that implement the methods discussed in Subsection 4.2.1, as well as many other methods.

Important papers in this area are Wampler (1970), Chambers (1973) and Beaton, Rubin and Barone (1976).

Most investigations use 'worst cases' data so as to test the program to its limit. The usual abuse is to make $\mathbf{X}'\mathbf{X}$ nearly singular and this immediately leads to problems for elimination methods which are attempting to find its inverse. A useful measure of the 'near singularity' of a matrix $\mathbf{A}$ is the P condition which is defined by $P(\mathbf{A}) = \lambda/\mu$ where λ is the numerically largest eigenvalue of $\mathbf{A}$ and μ the smallest. Typical 'worst cases' data have $P(\mathbf{A}) \sim 10^{14}$. In these adverse conditions, the methods using Gram Schmidt orthogonalisation or Householder transformations, that is, the last two methods of Subsection 4.2.1, consistently turn out to be the best. One conclusion which seems to be common to all methods is that scaling the x variables so as to arrange for the diagonal terms of $\mathbf{X}'\mathbf{X}$ to be unity does not appear to lead to any improvement in performance. However, no one seems to question the undoubted wisdom of 'subtracting the means' and, thus, of working with the model in the form of model (34) (as shown in Subsection 2.2.1).

A necessary condition for the satisfactory performance of the two successful methods mentioned above is that all inner products are accumulated using double precision arithmetic.

4.2.3 Detecting the failure of the numerical methods

Having performed a regression analysis, the points mentioned in Section 4.1 will help in detecting whether the 'correct' estimates of the regression coefficients have been determined, i.e. whether the correct solution of equation (35) has been found. There are a few additional precautions which might be taken.

Assuming the calculations are to be performed using a computer, a first step might be to run some test problems where the regression coefficients are known, so as to ensure that the computer program is working correctly. At the same time, it may be arranged for the condition of $\mathbf{X}'\mathbf{X}$ to be poor. Alternatively, a set of data published by Longley (1967) appears to give trouble to many of the weaker programs and this data may be tried instead.

A second step might be to monitor the condition of the matrix $\mathbf{X}'\mathbf{X}$ and, hence, to anticipate trouble with a particular regression analysis when $P(\mathbf{X}'\mathbf{X})$ is large.

A third step might be to perform the regression analysis several times with the scale of the x variables altered each time. Interchanging the identity of the x variables (for example, interchanging x_1 and x_3) is a useful device for detecting both elementary mistakes and obscure ones which might otherwise go undetected.

Another useful precautionary measure might be to calculate the vector of residuals, $\hat{\mathbf{e}} = [\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n]' = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ (where $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ are defined by equation (38)), and then to check that $\mathbf{X}'\hat{\mathbf{e}} = \mathbf{O}$, as should obviously be the case. In non-matrix terms, this is equivalent to checking that $\sum_{i=1}^{n} (x_{ri} - \bar{x}_r)\hat{e}_i = 0$ (for $r = 1, 2, 3, \ldots, k$).

Finally, the numerical stability, as well as the 'logical' stability of the calculations might be examined by perturbing the data and reperforming the

regression analysis. Thus, for example, suppose that the variable x_1 can only be recorded to one decimal accuracy giving readings of 12.1, 10.8, 6.9, etc. Then, these readings may be altered to 12.05, 10.75, 6.85, etc. or 12.15, 10.85, 6.95, etc. and the data re-analysed. If the resulting analysis is markedly different from the original analysis, then this may suggest numerical instability. At the same time, this will give the analyst a good idea of the imprecision inherent in his conclusions which has resulted from the imprecision of his original data.

4.3 Analysis of Residuals

4.3.1 Plotting the residuals

Use of the n residuals $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ has already been described in Subsection 4.2.3. Their definition, which is given by equation (38), shows that they are the differences between the observed values of the dependent variable y and the predicted values of y which have been obtained by using model (34) with the least squares estimates of the unknown parameters inserted in place of the corresponding parameters in the model.

The main benefit to be gained from studying these residuals is a knowledge of the adequacy, or more likely the inadequacy, of the assumptions made in multiple linear regression. The terms $e_1, e_2, \ldots, e_n$ are assumed to have zero mean and constant variance σ^2, to be normally distributed and to be mutually independent. Consequently, we will expect similar, if not identical, properties for $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$.

Thus, as a first step in examining residuals, it will be helpful to form a histogram of $\hat{e}_1, \ldots, \hat{e}_n$. Let us now consider the features that we would expect to see, or not to see, in this histogram. Unless there has been an error in calculation, the mean of $\hat{e}_1, \ldots, \hat{e}_n$ will always be zero. The histogram should be reasonably symmetric about zero; in other words, there should be no marked skewness and no strong evidence of bimodality. Also, there should be few points noticeably detached from the rest of the histogram. Existence of detached points may suggest the presence of outliers.

To investigate the shape of this histogram further, various statistics may be calculated from $\hat{e}_1, \ldots, \hat{e}_n$ and these will be discussed in Subsection 4.3.2. However, a further graphical aid is to use normal probability paper which was mentioned in Subsection 3.2.3 to plot the cumulative distribution of $\hat{e}_1, \ldots, \hat{e}_n$. The resulting plot should give a nearly straight line when the assumption of normality of $e_1, \ldots, e_n$ is valid.

The next step might be to plot the graph of $\hat{e}_1, \ldots, \hat{e}_n$ against $\hat{y}_1, \ldots, \hat{y}_n$ (as defined in equation (44)). Ideally, this should produce a graph which is just a horizontal band of points with possibly a slight bulging towards the middle of the graph. Variations on this pattern might be: (1) a band of points of more or less uniform width but which is not horizontal, i.e. it may be rising, falling or curved, or (2) a band of points of non-uniform width.

The first pattern variation usually suggests either an error in calculation or that the model is not adequately representing changes in y. This problem might be overcome by either transforming y or including some polynomial terms of $x_1, \ldots, x_k$ in the model.

The second pattern variation usually suggests that the variance of $e_1, \ldots, e_n$ is not remaining constant. This may be caused by the variability in y increasing (or decreasing) as y increases (although the problem of non-uniform density of points which was mentioned in Subsection 3.2.2 should be considered in these circumstances). Also, this pattern may be caused by increasing errors of measurement, inclusion of certain 'bogus', 'contaminated' or spurious results, or the physical process studied being different for different values of y. Rectification may be achieved by transforming the dependent variable y, eliminating certain results or fitting separate models to different portions of the data.

When there is some time sequence associated with $y_1, \ldots, y_n$, it is often helpful to plot $\hat{e}_1, \ldots, \hat{e}_n$ against time or sequence number or, at least, to plot them in the same order as the ys were measured or recorded. This is particularly helpful when time is involved and it has not been included as an independent variable. Again, a horizontal band of points should be expected from this graph but deviations of the type described for the previous graph may occur.

The first pattern variation might indicate that time should have been included as an explanatory variable in the regression equation. The second pattern variation might suggest that the variability of the dependent variable is associated with time; for example, it might suggest that results taken 50 years ago are less precise than results taken nowadays.

Similarly, plotting $\hat{e}_1, \ldots, \hat{e}_n$ against values of each of the independent variables in turn should give a horizontal band of points. A non-horizontal band may suggest that the effect of the independent variable on y has not been fully explained in the model and that, possibly, a polynomial term needs to be included in the model. A widening band may suggest that the variance of the dependent variable is not constant, but possibly related to the independent variable plotted.

Finally, it might be helpful to plot a graph of $\hat{e}_1, \ldots, \hat{e}_n$ against either an independent variable which has so far been omitted from the regression analysis, or a variable that has been thought to be not worth including, or a variable which has been omitted as a result of a stepwise regression calculation. This will help to check for constant variance and that there is no association between the variable plotted and the dependent variable y.

4.3.2 Some tests on the residuals

Exact tests of significance on the residuals tend to be cumbersome because the distribution of the residuals is not simple to deal with. Provided that the model used (i.e. model (34)) is correct, the residuals $\hat{e}_i$ will follow a normal distribution with zero mean. However, the residuals do not have constant variance and they are not independent. In fact, using the notation of Section 4.2, the variance covariance matrix of $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ is $\sigma^2(\mathbf{I}_{n\times n} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$.

Anscombe (1961) describes slightly modified versions of the coefficients of skewness and kurtosis which enable an assessment to be made on the 'normality' of the residuals. However, it has been shown that the test of Shapiro and Wilk mentioned in Subsection 3.2.3 may be applied to the residuals $\hat{e}_1, \ldots, \hat{e}_n$ for a test of normality of $e_1, \ldots, e_n$. Some of the other tests mentioned

in Subsection 3.2.3 (for example, tests for outliers) may be applied to $\hat{e}_1, \ldots, \hat{e}_n$, although they are not strictly valid and must be used with caution since $\hat{e}_1, \ldots, \hat{e}_n$ are not independent and do not have constant variance. However, with a plot of $\hat{e}_1, \ldots, \hat{e}_n$ showing a markedly detached point, a rough idea of its significance may be all that is required.

For testing whether the variances of $e_1, \ldots, e_n$ are constant, a suggested procedure is to divide the data into two groups with an equal number of observations in each. The residual sum of squares is then calculated for each group and their ratio is formed. When the assumption of equal error variance of observations in the two groups is valid, this ratio will follow the distribution $F_{n/2-k-1,n/2-k-1}$. Thus, for example, to test for an increase in variance with increasing y, the smallest $n/2$ values of y would form group 1 and the remaining $n/2$ values would form group 2. (When n is odd, the middle observation would be discarded.)

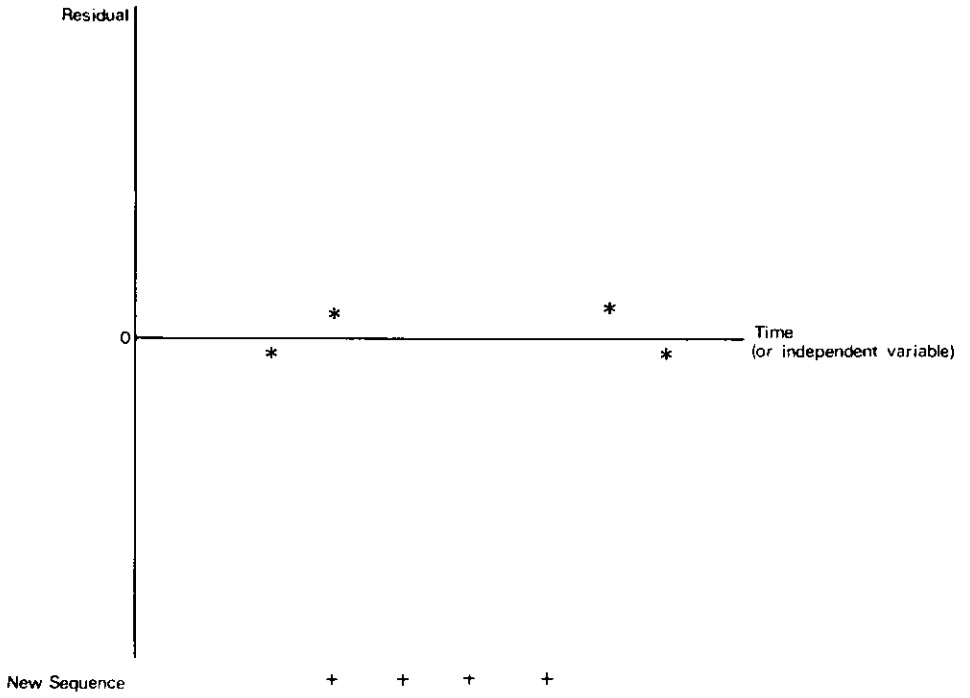


Fig. 20. Plot of residuals and associated signs.

There are several non-parametric tests for detecting changes from a horizontal band in a plotted sequence of readings. The test which will probably be most useful replaces the plotted sequence of points by their sign, as shown in Figure 20.

When there is no trend in the residuals, there should be a random jumble of +s and −s. However, when a trend is evident, there will be a series of runs of +s and −s. The test involves counting the total number of these runs, r. (In the

example above, r is equal to 3.) If there are n_1 +s and n_2 −s in the whole sequence, then the test statistic U is given by

$$U = \frac{r - 1 - 2n_1n_2/(n_1 + n_2)}{\sqrt{\dfrac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}}$$

and $U \sim N(0, 1)$, approximately, whenever the sequence is a random jumble of +s and −s.

Thus, for a $100\alpha\%$ significance test, the hypothesis of a random jumble of +s and −s would be rejected whenever $|U| > Z(\alpha/2)$. Rejection of the hypothesis would suggest that the plot of points did not form a horizontal band.

4.3.3 Other residuals

The residuals that have previously been used are correlated and have differing variances. Consequently, several attempts have been made to derive pseudo residuals which have better properties and this has led to the use of standardised residuals, BLUS residuals, recursive residuals, etc.

If V_i is the ith diagonal term of the variance covariance matrix of the residuals, then $\hat{e}_i/\sqrt{V_i}$ will have variance equal to 1. Using the residual mean square to estimate σ^2, the residuals $\hat{e}_i/\sqrt{V_i}$ are called standardised residuals.

BLUS residuals require the matrix $\mathbf{X}$ (as defined in Subsection 4.2.1) to be partitioned into

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_0 \\ \cdot \\ \mathbf{X}_1 \end{bmatrix}$$

where $\mathbf{X}_0$ is a $k \times k$ matrix and, hence, $\mathbf{X}_1$ is a $(n-k) \times k$ matrix. The h non-zero eigenvalues of $\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'$ are denoted by $\lambda_1^2, \lambda_2^2, \ldots, \lambda_h^2$ and the corresponding normalised eigenvectors are denoted by $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_h$. Thus, $\mathbf{X}_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0'\mathbf{z}_r = \lambda_r^2$ (for $r = 1, 2, \ldots, h$). If the vector of residuals

$$\begin{bmatrix} \hat{e}_1 \\ \cdot \\ \hat{e}_n \end{bmatrix} \text{ is partitioned into two parts } \begin{bmatrix} \hat{\mathbf{e}}_0 \\ \cdot\ \cdot \\ \hat{\mathbf{e}}_1 \end{bmatrix}$$

where $\hat{\mathbf{e}}_0$ contains k rows and, hence, $\hat{\mathbf{e}}_1$ contains $n-k$ rows, then the BLUS residuals are defined to be

$$\hat{\mathbf{u}}_1 = \hat{\mathbf{e}}_1 - \mathbf{X}_1\mathbf{X}_0^{-1}\left[\sum_{r=1}^{h} \frac{\lambda_r}{1 + \lambda_r}\mathbf{z}_r\mathbf{z}_r'\right]\hat{\mathbf{e}}_0$$

These residuals are intended to display the discrepancy between the last $(n-k)$ observed y values and the fitted model; in other words, they have the same purpose as $\hat{\mathbf{e}}_1$. The sum of squares of the $(n-k)$ BLUS residuals, $\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1$, is equal to

the sum of squares of the original residuals, $\sum_{i=1}^{n} \hat{e}_i^2$. However, the $(n-k)$ BLUS residuals are uncorrelated and have constant variance σ^2.

The problem in using these residuals is to decide on a partition of $\mathbf{X}$. By reordering the data, any k of the n readings may be arranged to correspond to $\mathbf{X}_0$. As the readings associated with $\mathbf{X}_0$ are effectively 'lost', i.e. no residuals corresponding to these values are produced, some care must be exercised in the selection of $\mathbf{X}_0$.

When the intention is to use the BLUS residuals to investigate an increase in variance with increasing y, it might be sensible to omit the middle observations. If the middle observations are omitted, leaving m readings at the beginning and m at the end of the sequence of observed y values, then the ratio of the sum of squares of BLUS residuals associated with the first m readings and the sum of squares of those associated with the last m readings should follow an F distribution with m and m degrees of freedom whenever the assumptions stated in Subsection 2.1.2 are valid.

Alternatively, a plot of the BLUS residuals may be informative provided that the partition of $\mathbf{X}$ is performed sensibly.

4.3.4 Autocorrelation

Possible causes of $e_1, \ldots, e_n$ being serially dependent were discussed in Subsection 3.4.1 and some ways of overcoming the problem were suggested in Subsection 3.4.2. However, first of all, it will be necessary to detect whether such a phenomenon exists and, not surprisingly, most procedures to achieve this make use of the residuals, $\hat{e}_1, \ldots, \hat{e}_n$.

The non-parametric test described in Subsection 4.3.2 will help to examine the serial dependence of the residuals. Obviously, it will only be sensible to investigate the possibility of autocorrelation when $y_1, y_2, \ldots, y_n$ represent a series of readings, in some sense. For example, they may represent a sequence of readings in times (such an annual rainfalls) or y_1 may have been recorded first of all, then y_2, then y_3, etc. Suppose that the corresponding residuals $\hat{e}_1, \hat{e}_2, \ldots, \hat{e}_n$ are replaced by their signs (e.g. $++--+--++--$ etc.). Then, autocorrelation between $y_1, \ldots, y_n$ might lead to a non-random sequence of $+$s and $-$s from the residuals which would be detected by the 'runs' test mentioned in Subsection 4.3.2. However, it should be remembered that there may be other causes of a non-random sequence of $+$s and $-$s from the residuals, for example, by an incorrect model having been fitted.

The most frequently used autocorrelation test is probably the Durbin–Watson test. The test statistic d is given by

$$d = \sum_{i=2}^{n} (\hat{e}_i - \hat{e}_{i-1})^2 \Big/ \sum_{i=1}^{n} \hat{e}_i^2$$

This statistic exploits the fact that $E(\hat{e}_i - \hat{e}_{i-1})^2 = E(\hat{e}_i^2) + E(\hat{e}_{i-1}^2) - 2E(\hat{e}_i \hat{e}_{i-1})$.

The right hand term reflects the correlation between successive residuals. Consequently, d will be small when residuals are consistently positively autocorrelated, intermediate when there is no autocorrelation and high when the residuals are negatively autocorrelated. However, the warning given in the

previous paragraph for the runs test also applies to the Durbin–Watson test. If all the assumptions made in fitting a regression model are valid, then the residuals should not display a notable autocorrelation. Thus, detecting autocorrelation in the residuals indicates that something is amiss. It does not necessarily follow that the cause is autocorrelation in the y readings as an incorrect or incomplete model may have been used.

The variance covariance matrix of the residuals given in Subsection 4.3.2 depends on **X**, the matrix of values of the independent variables. It is therefore not surprising to learn that the distribution of d also depends on **X**. To overcome this problem and provide a test which may be easily applied, Durbin and Watson evaluated bounds (d_l, d_u) between which, for a given significance level, the appropriate significance point must lie regardless of **X**. The suggested test procedure is to reject the assumption of independence in favour of positive autocorrelation when $d < d_l$, draw no conclusion when $d_l < d < d_u$ and accept the assumption of independence when $d > d_u$.

This provides a one tailed test of positive autocorrelation versus independence. To investigate the existence of negative autocorrelation, the procedure is to replace d by $4 - d$ in the above and for 'positive autocorrelation' read 'negative autocorrelation'. Tables for d_l and d_u are given in Durbin and Watson's original paper (1951) and are reprinted in several regression books, for example, Theil (1971).

One of the difficulties of applying this test is that the value of d frequently falls in the region (d_l, d_u) and thus the outcome of the test is inconclusive. Durbin and Watson give some guidance for obtaining conclusive results in this situation. However, other workers have shown that when the independent variables are 'smooth', the upper bound d_u provides an approximation to the true significance point. (A 'smooth' variable is defined to be one whose consecutive values show small changes compared with the total range of the variable). A summary of various approximations to the distribution of d which may help when the test is inconclusive is given by Durbin and Watson (1971).

An alternative test of autocorrelation may be derived using the BLUS residuals, $\hat{\mathbf{u}}_1 = [\hat{u}_{1,1}, u_{1,2}, \ldots, \hat{u}_{1,n-k}]'$. The test statistic Q is given by

$$Q = \left(\frac{1}{n-k-1} \sum_{i=1}^{n-k-1} (\hat{u}_{1,i+1} - \hat{u}_{1,i})^2 \right) \Big/ \hat{\sigma}^2$$

where $\hat{\sigma}^2$ is the usual residual mean square given by equation (39). The distribution of Q is tabulated in Theil (1971). Alternatively, for $n - k > 60$, a satisfactory approximation is to assume that Q follows a normal distribution with mean 2 and variance $4/(n-k)$.

Phillips and Harvey (1974) derive an autocorrelation test which is simpler to calculate than the BLUS residuals test but both of these tests have poorer power than the Durbin–Watson test in which an approximate significance point is calculated when $d_l < d < d_u$.

References

Anscombe, F. J. (1961). Proc. 4th Berkeley Symposium, **1**, 1–36.
Beaton, A. E., Rubin, D. B. and Barone, J. L. (1976). *J. Am. Stat. Ass.*, **71**, 158–68.

Chambers, J. L. (1973). Proc. 39th Session of I.S.I., Pt 4, 245–54.
Durbin, J. and Watson, G. S. (1951). *Biometrika*, **38**, 159–78.
Durbin, J. and Watson, G. S. (1971). *Biometrika*, **58**, 1–19.
Langley, J. W. (1967). *J. Am. Stat. Ass.*, **62**, 819–41.
Phillips, G. D. A. and Harvey, A. C. (1974). *J. Am. Stat. Ass.*, **69**, 935–9.
Seber, G. A. F. (1977). *Linear regression analysis*. John Wiley, New York.
Theil, H. (1971). *Principles of Econometrics*. North Holland, Amsterdam.
Wampler, R. H. (1970). *J. Am. Stat. Ass.*, **65**, 549–65.

Chapter 5

SOME EXAMPLES

5.1 An Example of Fitting and Comparing Several Regression Lines

The source of the data used in this example is Report No. 73 of the Institute of Hydrology by M. Robinson (1980). This report examined the effect of pre-afforestation drainage on the streamflow and water quality of a small upland catchment, the Coalburn catchment, located approximately 40 km north-east of Carlisle. Table 10 of the report gives rainfall (R) and run-off (Q) in millimetres, for winter (October–March) and summer (April–September) over a five-year period prior to drainage work and over a similar period after drainage.

Table 2 Rainfall and runoff on the Coalburn catchment

Winter				*Summer*			
Period	*R*	*Q*	*% Runoff*	*Period*	*R*	*Q*	*% Runoff*
Pre-draining							
1967–8	926	729	78.7	1967	669	343	51.2
1968–9	494	455	92.1	1968	632	310	49.1
1969–70	577	446	77.3	1969	579	259	44.7
1970–1	652	599	91.9	1970	575	305	53.0
1971–2	542	465	85.8	1971	457	196	42.8
Mean	638	539	84.5	Mean	582	283	48.6
Post-draining							
1973–4	542	480	88.6	1974	497	235	47.3
1974–5	794	636	80.1	1975	642	370	57.6
1975–6	546	478	87.5	1976	449	199	44.3
1976–7	622	593	95.3	1977	584	315	53.9
1977–8	763	704	92.3				
Mean	653	578	88.5	Mean	543	279	51.4

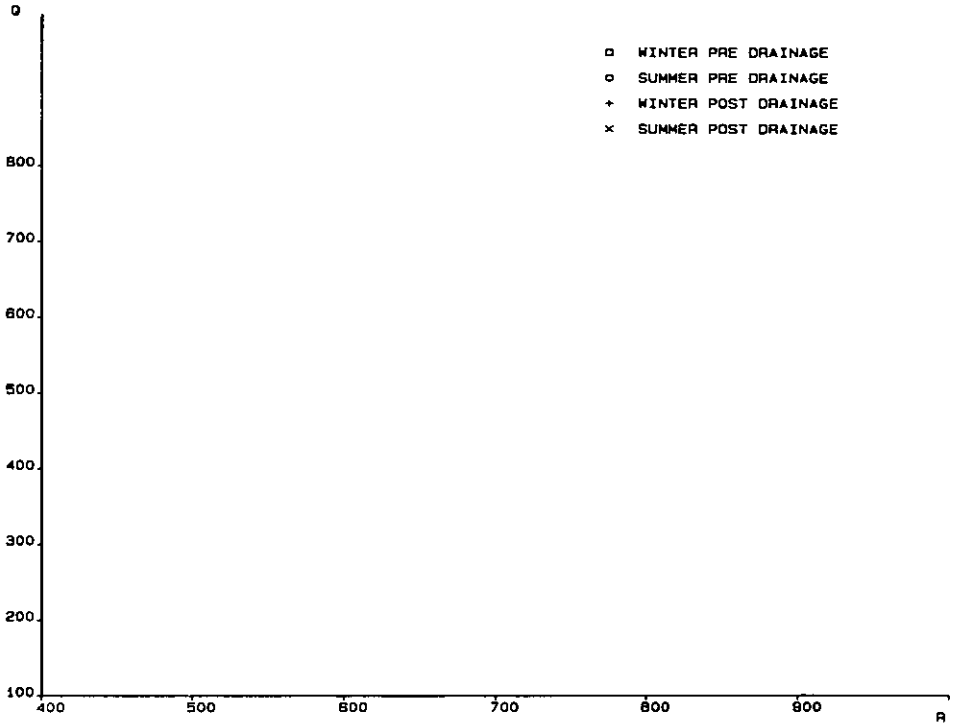


Fig. 21. Coalburn catchment data.

Figure 21 gives a plot of the whole data set. From this plot of the data we might tentatively conclude that:

A linear regression model for relating Q and R seems reasonable for each of the four groups of data.
There is a difference in position (but not slope) between the winter and summer regression lines.
There is less scatter about the line for the summer data.

We would probably also observe that:

The winter data cover a larger span of values of R.
There is a shorter series of values in the summer post-drainage category than in the other three categories.

Although not strictly necessary as a step in the analysis of this data set, let us start by fitting a straight line to just the winter pre-drainage data. Using the notation of Section 1.2, we take rainfall to be the independent variable, x, and runoff to be the dependent variable y for reasons similar to those advanced in Subsection 1.1.3. The table shown on page 108 gives the various necessary calculations.

Column 1 gives the values of the initial calculations made on the winter pre-draining data. From these statistics, the useful intermediate statistics S_{xx}, etc. are calculated and their values are given in column 2. The formulae for S_{xx}, etc. are given at the end of Subsection 1.2.2. Finally, column 3 gives the most pertinent statistics in fitting a straight line, namely the intercept and slope

$n = 5$		
$\sum_{i=1}^{n} x_i = 3191$	$\bar{x} = 638.2$	$\hat{a} = 99.34$
$\sum_{i=1}^{n} y_i = 2694$	$\bar{y} = 538.8$	$\hat{b} = 0.6886$
$\sum_{i=1}^{n} x_i^2 = 2153309$	$S_{xx} = 116812.8$	S.E.$(\hat{a}) = 82.19$
$\sum_{i=1}^{n} x_i y_i = 1799744$	$S_{xy} = 80433.2$	S.E.$(\hat{b}) = 0.1252$
$\sum_{i=1}^{n} y_i^2 = 1512408$	$S_{yy} = 60880.8$	$\hat{\sigma}^2 = 1832.4$

together with their standard errors. These are calculated from equations (5), (6), (7) and (8) respectively with the estimate of σ^2 used in equations (7) and (8) being calculated by equation (10). S.E.$(\hat{a})$ denotes the estimated standard error of $\hat{a}$ and is equal to the square root of equation (7) when the estimate for σ^2 given by equation (10) has been inserted. Thus we conclude that our fitted model is

$$\text{Runoff} = 99.34 + 0.6886 \times \text{Rainfall}$$

Testing the hypothesis $a = 0$ assesses the evidence to support a 'straight line through the origin' model to relate runoff and rainfall. Referring to the first test

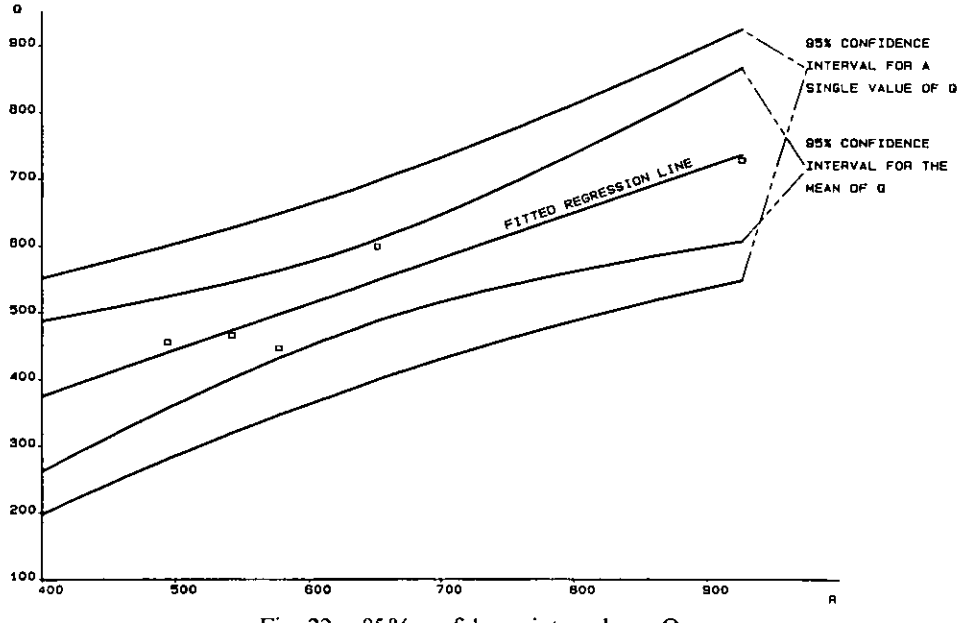


Fig. 22. 95% confidence intervals on Q.

in Subsection 1.2.3 the test statistic given may, in fact, be written as $\hat{a}/[\sqrt{\text{S.E.}(\hat{a})}]$, the value of this expression is 1.21. Using a 5% significance test gives $t(n-2, 1-\alpha/2) = 3.18$ and hence we cannot reject the hypothesis $a = 0$. Testing the hypothesis $b = 0$ assesses whether rainfall has any affect on runoff. The second test statistic in Subsection 1.2.3 may also be written in the simpler form $\hat{b}/[\sqrt{\text{S.E.}(\hat{b})}]$. Its value is 5.5 and hence, using a 5% significance test, we reject the hypothesis $b = 0$, implying that there is some linear association between rainfall and runoff values.

Figure 22 gives the 95% confidence intervals stated in equations (13) and (14). Thus, at a particular rainfall value, the outer curves give a 95% confidence interval for a single reading of runoff and the inner curves give a 95% confidence interval for the mean runoff at that rainfall value. All observations lie comfortably within the outer curves which suggest no obvious outliers. Indeed, with so few observations, it would have to be a quite exceptional observation to show up as an outlier.

Fitting the linear regression model to each of the four sets of data separately gives the following set of summary statistics:

Table 3 Regression lines for the Coalburn catchment data

	Pre-draining		*Post-draining*	
	Winter	*Summer*	*Winter*	*Summer*
$\hat{a}$	99.34	−110.11	81.03	−205.04
$\hat{b}$	0.688 6	0.674 3	0.760 9	0.892 8
S.E.($\hat{a}$)	82.19	74.61	126.75	13.89
S.E.($\hat{b}$)	0.125 2	0.127 1	0.191 5	0.025 4
$\hat{\sigma}^2$	1 832.4	416.3	2 078.7	14.46

Examining the speculations we made earlier we see from these statistics that it is apparent that the intercept parameter, a, differs between the winter and summer data. Furthermore, there is less scatter about the line ($\hat{\sigma}^2$) in the summer months. In addition, there is a slight suggestion of a higher slope parameter (b) in the post-draining data. However, the most striking difference of all, and the most inconvenient, is the very low value of $\hat{\sigma}^2$ for the summer post-draining data.

Formally, a test of significance on the four $\hat{\sigma}^2$ values, using the test of equality of variance outlined in Subsection 1.3.2, gives $M = 8.94$. With $\chi^2(3, 0.95) = 7.81$, this suggests that, using a 5% significance test, we should reject the hypothesis of equal variance about the line for the four categories of data. Reference to Figure 21 will confirm just how unusually linear the summer post-draining data are. If this effect were real then it might well be the most interesting finding of the analysis, namely that draining has led to runoff being closely related to rainfall. However, with only four observations one must treat any conclusions with caution and perhaps even scepticism.

Computing the analysis of variance table given in Subsection 1.3.2 gives the following table:

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*	*Mean square ratio*
Overall regression	281 725	1	281 725	238
Difference in positions	202 948	3	67 649	57.2
Difference in slopes	924	3	308	0.26
Residual	13 010	11	1 182.7	
Total	498 607	18		

Formally, we would accept a hypothesis of equal slopes and reject a hypothesis of equal intercepts, the latter being significant at the 0.1 % level. Thus our suspicion about the intercepts is confirmed but the difference in slopes is not significant. Referring back to the table of slopes and intercepts we see that the difference in intercepts relative to their standard errors is much greater than the difference in slopes relative to their standard errors. Any conclusions from the analysis of variance table are made on the assumption of equal variance about the line in the four sets of data but it would seem unlikely that the degree by which this assumption is violated would greatly alter the conclusions.

Further comparisons would be of interest, in particular to establish whether there is a change in slope between pre- and post-draining data. To take the analysis further, it helps to see that the model for these data can be written as a multiple regression model similar to equation (34). The model used so far is equation (23) in Subsection 1.3.2 and to keep the complexity to a minimum, let us suppose that the data had only consisted of two groups with two pairs of values in each group. Equation (23) becomes

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_{11} & 0 \\ 1 & 0 & x_{12} & 0 \\ 0 & 1 & 0 & x_{21} \\ 0 & 1 & 0 & x_{22} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{21} \\ e_{22} \end{bmatrix}$$

which is of the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$ and, as demonstrated in Subsection 4.2.1, this is equivalent to model (34). Thus significance tests on a_1, a_2, b_1, b_2 may be conducted using the general linear hypothesis results of Subsection 2.3.1. In general, provided that the model is linear in the unknown parameters and that the error term is an additive component, it will usually be possible by using dummy variables in conjunction with 'real' variables to write the model in multiple regression form. However, because of the general nature of the test statistic (equation (48)), it is often unnecessary to actually derive the multiple regression version of the model in question. This will be apparent from the analysis in the remainder of this section. Let us now use this approach to compare the values of a and b for pre-draining data with those for post-draining but, at the same time retain the winter/summer division in the data. This will

allow us to focus on the pre/post draining division in its effect on a and b but, at the same time, acknowledge that there might be a difference (probably in a) between the winter and summer regression lines. We need to compute two residual sums of squares, the residual sum of squares R, fitting a separate regression line to each of the four groups of data ($=13\,010$ from the previous analysis of variance table) and the residual sum of squares R_H computed by assuming the hypothesis to be true. The hypothesis, H, is that a and b are the same for pre- and post-draining data (but possibly different for winter and summer data). Merging the pre- and post-draining data produces a set of data with just two categories, winter and summer. Applying the same type of analysis as we did with four categories gives an analysis of variance table with residual sum of squares, $R_H = 17\,416.68$ (R_H must be $\geq R$). The test statistic for H, given by equation (48), is $[(17\,416.68 - 13\,010)/4]/[13\,010/(19-8)] = 0.93$ which should follow $F_{4,11}$ if H is true. The 5% point of $F_{4,11}$ is 3.36 and hence we must accept H.

It is enlightening to approach the same test from a slightly different angle. Let us analyse the winter and summer data separately but with the same objective as previously, to decide whether a and b are the same for pre- and post-draining data. From Table 3, we know that the residual sum of squares, fitting separate regression lines to winter pre- and post-draining data, is $1832.4 \times 3 + 2078.7 \times 3 = 11\,733.3 = R$. Combining pre- and post-draining data and fitting a single regression line to the winter data give a residual sum of squares of $13\,972.3 = R_H$.

Our test statistic for H is now $[(13\,972.3 - 11\,733.3)/2]/[11\,733.3/(10-4)] = 0.57$ which should follow $F_{2,6}$. Again we accept H. Repeating the calculations for the summer data gives $R = 416.3 \times 3 + 14.46 \times 2 = 1277.82$ and $R_H = 3444.38$. Our test statistic is now $[(3444.38 - 1277.82)/2]/[1277.82/(9-4)] = 4.24$ which should follow $F_{2,5}$. The 5% point is 5.79 and hence using a 5% significance test we must accept H, but the 10% point is 3.78 at which level we would reject H. Consequently we cannot be very sure about the validity of H. There might be some grounds for supposing that in the summer months, the pre- and post-draining regression lines differ. This difference in conclusions for summer and winter is due to a large extent to the much smaller variation about the line in summer data ($R = 1277.82$ for the summer, $R = 11\,733.3$ for the winter).

Note that adding the two R values $1277.82 + 11\,733.3 = 13\,011.12$ gives (except for rounding error) the value of R ($=13\,010$) for the whole set of data and adding the values of R_H, $13\,972.3 + 3444.38 = 17\,416.68$ gives the value of R_H for the whole data set. Thus we can imagine combining the results of the summer and winter tests into a single significance test. However, intuition should not be relied upon too heavily in this area as this additive property only occurs if certain conditions about the matrix $\mathbf{X}$ are satisfied, namely that certain columns of $\mathbf{X}$ are mutually orthogonal.

So far, we have only simultaneously tested whether both a and b differ between the pre- and post-drainage data. To emphasise the extent to which differences in regression lines may be examined, suppose we pursued the hint of a difference in pre- and post-draining regression line in the summer months. Is the difference primarily in the intercept a, or the slope, b or both? There are

four hypotheses which we could investigate with the summer data and they are listed below together with the residual sum of squares derived by assuming that hypothesis to be true.

Hypothesis	*Intercept*	*Slope*	*Residual sum of squares*	*Degrees of freedom*
H_1	different	different	1 277.88	5
H_2	same	different	1 647.01	6
H_3	different	same	1 850.30	6
H_4	same	same	3 444.38	7

To test the hypothesis (H_3) that the pre- and post-drainage regression lines have the same slope (but not necessarily the same intercept) we would compute $[(1850.30 - 1277.88)/1]/(1277.82/5) = 2.24$ which should follow $F_{1,5}$. The 5% point of $F_{1,5}$ is 6.61 and hence we accept this hypothesis using a 5% significance test. A similar calculation would lead us to accept hypothesis H_2 that the regression lines have the same intercept (but not necessarily the same slope). However the hypothesis H_4, that the regression lines have both the same slope and the same intercept is the hypothesis we considered earlier which gave a test statistic of 4.24 which should follow $F_{2,5}$.

For this particular set of data, the conclusion to be drawn from examining the three hypotheses H_2, H_3 and H_4 is not clear cut, although in other situations it might be quite informative. From examining H_4 there is some suggestion that the lines differ but by examining H_2 and H_3 it is not clear that this difference is confined to either the slope alone or the intercept alone. It is more that there is a marginal difference between the two sets of data which can be accommodated by having one parameter different in the two regression lines and it does not matter greatly whether that parameter is the slope or the intercept. There is a marginal preference for it being the slope.

A similar set of calculations on the winter data produces the following table:

Hypothesis	*Intercept*	*Slope*	*Residual sum of squares*	*Degrees of freedom*
H_1	different	different	11 733.5	6
H_2	same	different	11 742.0	7
H_3	different	same	11 933.2	7
H_4	same	same	13 972.3	8

As we saw previously, the test statistic for H_4 is 0.57 which should follow $F_{2,6}$. Consequently there is very little reason to question the validity of H_4. If there were, then it would again be the slope parameter that differed 'more' than the intercept parameter.

A similar approach could be made to substantiate the difference in the intercept parameter, a, between winter and summer months. This time the

division would be winter/summer rather than pre/post draining and the parameter of interest would be a, the intercept, rather than b, the slope. Alternatively, a rough assessment of the significance of the difference between two estimates of a (and similarly of b) may be gained by computing

$$\frac{\hat{a}_1 - \hat{a}_2}{\sqrt{[\text{S.E.}(\hat{a}_1)]^2 + [\text{S.E.}(\hat{a}_2)]^2}}$$

where the suffices 1 and 2 refer to the two groups of data, winter/summer. On the null hypothesis of equal intercepts, the distribution of this quantity will tend to a Normal distribution with increasing sample sizes. For an approximate $100\alpha\%$ significance test we would accept the null hypothesis if

$$\left| \frac{\hat{a}_1 - \hat{a}_2}{\sqrt{[\text{S.E.}(\hat{a}_1)]^2 + [\text{S.E.}(\hat{a}_2)]^2}} \right| < Z(\alpha/2)$$

Thus, comparing the two post-drainage intercepts gives a test statistic of $[81.03 - (-205.04)]/[\sqrt{(126.75)^2 + (13.89)^2}] = 2.24$ and consequently we should reject the null hypothesis of equal intercepts using a 5% significance test. However, for this example, sample sizes are very small and it would be foolish to place much weight on this conclusion. An exact test of significance of the same hypothesis would just fail to reject the hypothesis of equal intercepts, but one advantage of the approximate test presented here is that it allows for the possibility of different variances in the different groups which the exact test does not.

In summary, to produce an overall model for the four groups of data it would seem that we should allow for different intercepts for winter and summer

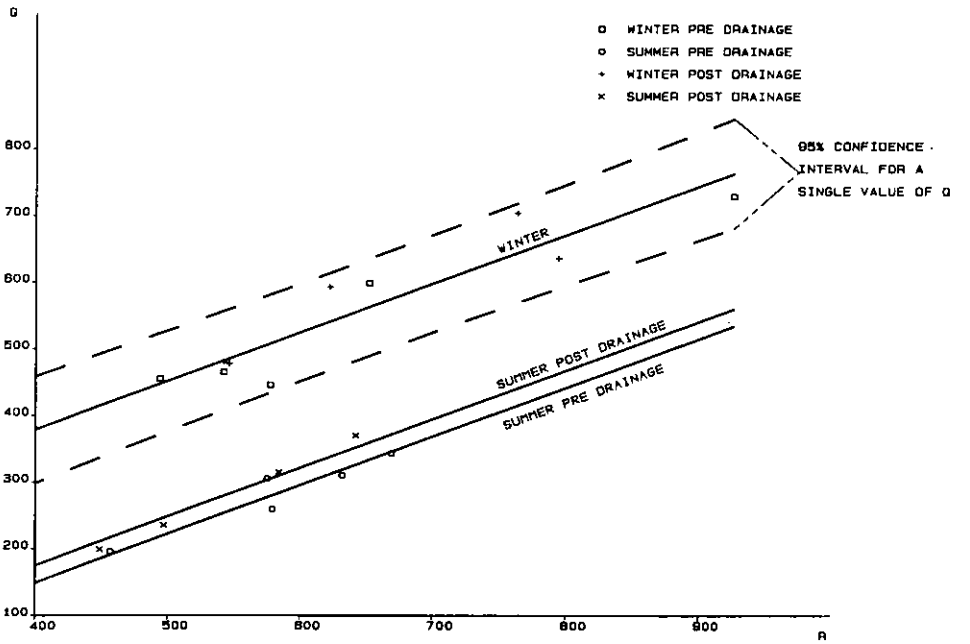


Fig. 23. Parallel regression lines for the Coalburn data.

data and either different intercepts or different slopes for pre- and post-drainage in the summer. Opting for three different intercepts but a common slope, Figure 23 shows the resulting three parallel regression lines. $Q = (86.48, -143.08, -117.13) + 0.7309R$ for winter, summer pre- and summer post-drainage respectively. The 95% confidence interval for a single value of Q is also given; for clarity the intervals for the other two lines have been omitted. The calculations may either be carried out by using dummy variables to introduce the different intercepts, as already explained in this section, or by making use of the general results usually referred to as analysis of covariance. These lead to some economy of calculation for regression models which consist of some dummy variables and some genuinely quantitative independent variables.

The residual sum of squares from fitting this model to the whole data set is 15943 with 15 degrees of freedom. A test of the assumptions made about the slope and intercept parameters (common slope for all four groups and common intercept parameter for both winter groups) could be performed using the general linear hypothesis approach of Subsection 2.3.1. The value of R_H will be 15943 with 15 degrees of freedom and the value of R will be 13010 with 11 degrees of freedom (the residual sum of squares from fitting a model with different slope and intercept parameters for each group). Our test statistic will be $[(15943 - 13010)/(15 - 11)]/(13010/11) = 0.62$ which should follow $F_{4,11}$. We could clearly not reject the hypothesis of common slope and common winter intercept parameter.

5.2 Multiple Regression on Mean Annual Flood

5.2.1 Introduction

Data were provided by the Institute of Hydrology on the annual maximum flood and various catchment characteristics for 83 catchments distributed over the whole of England, Wales and Scotland. This data set will be referred to as data set X. The catchment characteristics were area (AREA), stream frequency (STMFRQ), stream slope (S1085), mainstream length (MSL), standard annual average rainfall (SAAR), one day rainfall of 5 year return period less effective mean soil moisture deficit (RSMD), urban index (URBAN), lake index (LAKE) and soil index (SOIL). Detailed explanation of these characteristics is given in Report No. 49 of the Institute of Hydrology by J. V. Sutcliffe (1978). The length of record varied considerably with as many as 43 annual maxima available in one catchment and 2 in another, giving a total of 905 readings of annual maxima in all. Individual readings of annual maxima will be denoted by Q_{ij} and the mean of all values of Q_{ij} for a particular catchment by Q_i.

The Flood Studies Report (1975) gave methods for assessing the statistical distribution of floods suitable for a range of cases depending on the amount of information available and relevant to any given site. As part of this overall scheme there was a requirement to be able to predict the mean annual maximum flood ($\bar{Q}$) for sites at which no flow-gaugings at all are available. This predicted mean value would then be used as an input by other procedures. Objectively determined catchment characteristics such as those listed above

were found for existing gauge sites together with values of $\bar{Q}$, which was essentially just the average of the individual yearly maximum instantaneous flow rates for each site: however, for gauges with only short records, improved estimates of mean annual flood were sometimes used. For the study here the individual yearly maxima are available and so a somewhat more general analysis can be made, looking at the relationship of the distribution of yearly maxima with catchment characteristics. Data set X consists of the same catchments as available for the Flood Studies Report except that only the 83 out of 643 catchments with areas of less than 72 km^2 are included: the period of data is also the same.

The Flood Studies Report had suggested a countrywide equation relating catchment characteristics and mean annual flood $\bar{Q}$ which was linear in the logs of the variables. It was suggested that the equation could be improved if regional multipliers were used instead of a single countrywide constant. In regression terms the countrywide equation with a single constant could be derived by applying the multiple regression technique to the dependent variable $\log \bar{Q}_i$ and independent variables log AREA, log STMFRQ etc. or to the dependent variable $\log Q_{ij}$ and the same set of independent variables. The difference between these two approaches will be discussed at the start of Subsection 5.2.2.

Table 4 gives the regression coefficients quoted in the Flood Studies Report and those derived with equation (36) using data set X with $\log Q_{ij}$ as the dependent variable.

Table 4 Regression coefficients for predicting mean annual flood, $\bar{Q}$

	log AREA	*log STMFRQ*	*log S1085*	*log SAAR*	*log RSMD*	*log (1 + URBAN)*	*log (1 + LAKE)*	*log SOIL*
Flood Studies Report	0.94	0.27	0.16	—	1.03	—	−0.85	1.23
Data set X	0.79	0.23	0.13	0.85	—	2.44	—	1.29

The reason for transforming URBAN and LAKE is that both indices can be zero and hence will give values of $-\infty$ when logged. Many other transformations could of course be applied but 1 + URBAN, 1 + LAKE have the advantage, in interpretation, that they have no effect on the resultant prediction if the catchment in question includes no urban or lake area.

In general there seems to be reasonable agreement between the two sets of coefficients. The data sets differ in that data set X is of a limited number of smaller catchments. Variables SAAR and RSMD have a correlation of 0.93 and consequently it is no surprise to find the one substituting for the other in the equations. Data set X contains very few catchments with lakes and hence LAKE is not useful for predicting $\bar{Q}$ for this data set. This small point illustrates the dependency of any derived equation on the scope, quality and context of the data set used to derive it.

To illustrate the theory given in Sections 2.2 and 2.3, we will give some further details of the calculations involved in the regression using data set X.

Table 5 Regression coefficients and their standard errors

Variable	*Identity*	*Regression coefficient*	*Standard error*	*Test statistic*
y	log annual maximum flood			
x_1	log AREA	$\hat{\beta}_1 = 0.79$	0.042	18.50
x_2	log STMFRQ	$\hat{\beta}_2 = 0.23$	0.027	8.65
x_3	log S1085	$\hat{\beta}_3 = 0.13$	0.042	3.09
x_4	log MSL	$\hat{\beta}_4 = 0.10$	0.056	1.73
x_5	log SAAR	$\hat{\beta}_5 = 0.85$	0.114	7.47
x_6	log RSMD	$\hat{\beta}_6 = 0.02$	0.109	0.20
x_7	log (1 + URBAN)	$\hat{\beta}_7 = 2.44$	0.214	11.42
x_8	log (1 + LAKE)	$\hat{\beta}_8 = 0.16$	0.200	0.82
x_9	log SOIL	$\hat{\beta}_9 = 1.29$	0.084	15.34
		$\hat{\alpha} = -5.83$	0.53	−10.95

The regression coefficients have been derived from equation (36) and their standard errors from equation (37) using equation (39) to estimate σ^2. In Subsection 2.3.2 we suggest that a test of significance of $\beta_i = 0$ $(i = 1, 2, \ldots, 9)$ involves computing the ratio of the regression coefficient and its standard error (last column) and comparing this with the t distribution with $n - k - 1$ $(= 895)$ degrees of freedom. The value of $t(895, 0.975)$ is 1.96 and thus we only accept $\beta_i = 0$ for $i = 4$, 6 and 8 using a 5% significance test.

If we had wanted to test the hypothesis $\beta_1 = \beta_2 = \cdots = \beta_9 = 0$ (y is not linearly related to $x_1, x_2, \ldots, x_9$) we could have formed the analysis of variance table given below as described in Subsection 2.3.2.

Source	*Sum of squares*	*Degrees of freedom*	*Mean square*
Regression	691.97	9	76.89
Residual	297.96	895	0.33
Total	989.93		

The ratio of the two mean squares is ~233 which is certainly greater than $F(9, 895, 0.95)$ $(= 2.21)$ and hence we would reject the hypothesis of no linear association between y and $x_1, x_2, \ldots, x_9$.

Having observed that the two data sets seem quite similar and that a straightforward application of the multiple regression technique of Chapter 2 can establish a relationship between Q and catchment characteristic, let us just stand back for a moment and consider what we have done. We chose to work with the logs of all variables. On what grounds could this be justified and were the assumptions of multiple regression satisfied? Is there a better transformation than log? Is it reasonable to assume a countrywide equation, or to allow a different constant term for different areas, or should there really be a different equation for different areas?

5.2.2 Transformations and weights on annual maximum flood

Confining our attention from now on to data set X, the regression equation given in Subsection 5.2.1 was calculated by computing the logs of all readings of annual maximum flood (Q_{ij}) and regressing them on log catchment characteristics. Because, within a given catchment the values of the independent variable will be constant, this is in some ways practically equivalent to regressing the mean of the logs of annual maximum flood on log catchment characteristics. However the latter regression would need to be a weighted regression (see Subsection 2.4.2) as there are more readings in some catchments than others, and hence the means will have different variances. The weights should be equal to the number of readings in each catchment. Either method would then be 'valid' provided that the variance of $\log Q_{ij}$ remained constant under all catchment conditions, although the degrees of freedom on the residual sum of squares will be larger, and hence the tests of significance will be more sensitive, if the original 905 values of Q_{ij}, rather than the 83 means, are used. They will correspond to a regression on $\log \bar{Q}$ if $\bar{Q}$ is interpreted as being the geometric mean of annual maximum floods. However they will differ slightly from a regression in which $\bar{Q}$ is taken to be the more usual estimate of mean annual flood, the arithmetic mean of annual maxima.

For examining the assumption of constant variance of $\log \bar{Q}_{ij}$ it is particularly useful that the data set has repeated readings of the dependent variable at fixed values of the independent variables (several readings of annual maximum flood from each catchment). The test of homogeneity of variance used in Subsection 1.3.2 may be slightly modified to compare the 83 estimates of variance obtained from the different catchments. Unfortunately this test shows a significant difference ($p < 0.001$) in the variance of $\log \bar{Q}_{ij}$ from catchment to catchment.

An argument for using the logarithm of Q_{ij} might have been that the variance of Q_{ij} was not constant from catchment to catchment but varied with the square of the mean (see Subsection 3.3.1). As mentioned there, by plotting estimated mean and standard deviation of annual maximum flood against each other, such a possibility could easily be investigated. A straight line should ensue from such a plot.

For this particular data set a straight line goes a long way to explaining the relationship between mean and standard deviation but by no means completely. This is confirmed by the outcome of applying the Box–Cox transformation (see Subsection 3.3.3) with Q_{ij} as dependent variable and log of catchment characteristics as the independent variables. A value of $\lambda = 0.115$ gives the maximum value of $L_b(\lambda)$ and the hypothesis $\lambda = 0$ (implying a log transformation) is rejected using a 0.1% significance test ($2(L_b(\lambda_{\max}) - L_b(\lambda_0)) = 35.06 > \chi^2(1, 0.999) = 10.83$).

Thus it would unfortunately appear to be invalid to apply multiple regression as described in Subsection 5.2.1 to $\log Q_{ij}$ and log of catchment characteristics. The appropriate transformation for this data set would be to take $(Q_{ij})^{0.115}$ when using the logs of the catchment variables. This would still lead to a multiplicative model in the catchment variable but that model would no longer be predicting yearly annual maximum flood but the function

antilog (annual maxim flood)$^{0.115}$. If it were felt, on physical grounds, or for mathematical simplicity, to be sensible to use log annual maximum flood and log catchment characteristics then a weighted regression could be performed with the mean values of $\log Q_{ij}$ for each catchment as the dependent variable and with weights = (no. of readings in the catchment)/(sample variance of $\log Q_{ij}$ for that catchment). A further variant might be to relieve the parameter λ in the Box–Cox transformation of the burden of equalising all the variances and instead, incorporate the Box–Cox transformation into a weighted regression of $(\bar{Q}_i)^\lambda$ on catchment characteristics (logged or otherwise).

5.2.3 Regression of the standard deviation

There are several reasons for trying to predict the standard deviation of the annual maximum flood. It would be interesting to see which catchment characteristics influence variability, it would enable us to 'smooth' our weights prior to a regression analysis of the type described above and it would allow statements about precision of prediction of annual maximum flood to be given in terms of value of certain catchment characteristics.

However, although with this data set, we can readily estimate the standard deviation of annual maximum flood for each catchment, the variance of that statistic will not be constant over all catchments. Thus, it would be invalid to use that statistic in an unweighted multiple regression. A weighted multiple regression analysis could be performed making use of the fact that, for a particular catchment, the estimate of standard deviation

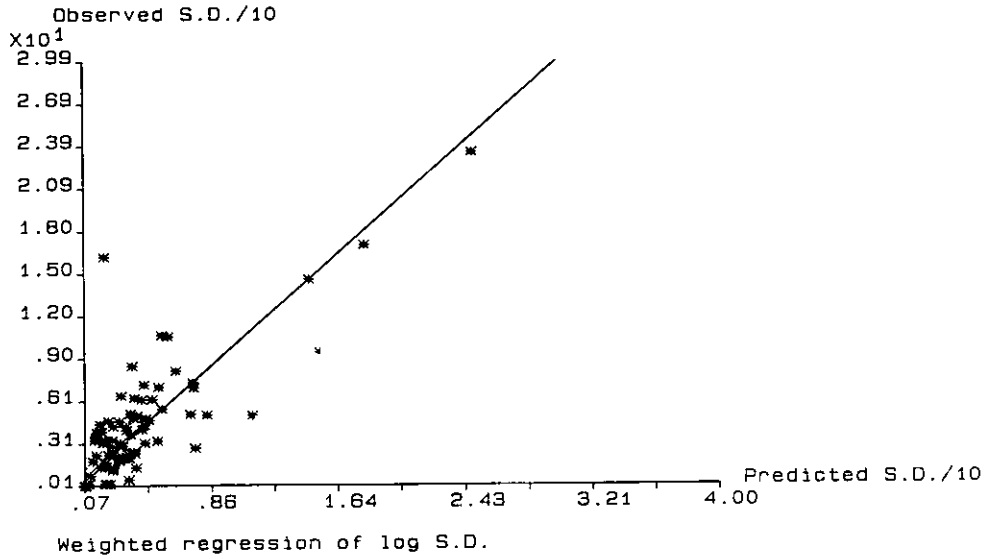
$$\sqrt{\frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Q_{ij} - \bar{Q}_i)^2}$$

has variance approximately equal to σ_i^2/n_i for a normal population where the catchment standard deviation is σ_i. Other possibilities might also be considered such as a transformation to stabilise the variance using the technique of Subsection 3.3.1 or a Box–Cox transformation. Because of the relationship between the estimated standard deviation and its variance, the appropriate variance stabilising transformation is log (estimated standard deviation).

A weighted multiple regression analysis of log (estimated standard deviation) on catchment characteristics and mean annual flood with weights equal to the number of readings in the catchment is only moderately successful. Two independent variables have regression coefficients which are significantly different from zero, $\bar{Q}$ and SOIL. Applying the Box–Cox transformation suggests a transformation (estimated standard deviation)$^{0.46}$ and the hypotheses $\lambda = 0$ (log transformation) and $\lambda = 1$ (no transformation) are both rejected using 0.1% level significance tests. A combination of weighted multiple regression and a Box–Cox transformation suggests the transformation (estimated standard deviation)$^{0.42}$.

Using either of these transformations produces a much more successful regression. Several regression coefficients are significantly different from zero, S1085, SAAR LAKE and $\bar{Q}$, suggesting that each of these factors is associated with variability in annual maximum flood, and there are very few outliers evident

when observed and predicted values are compared (see Figure 24). There are, of course, other more objective ways of comparing the relative success of different regressions. Chapter 12 of Seber (1977) gives details and references to a selection of methods. There are only two 'significantly' large residuals and in both cases the regression model has underestimated the standard deviation. It would be interesting to examine the particular features of these catchments



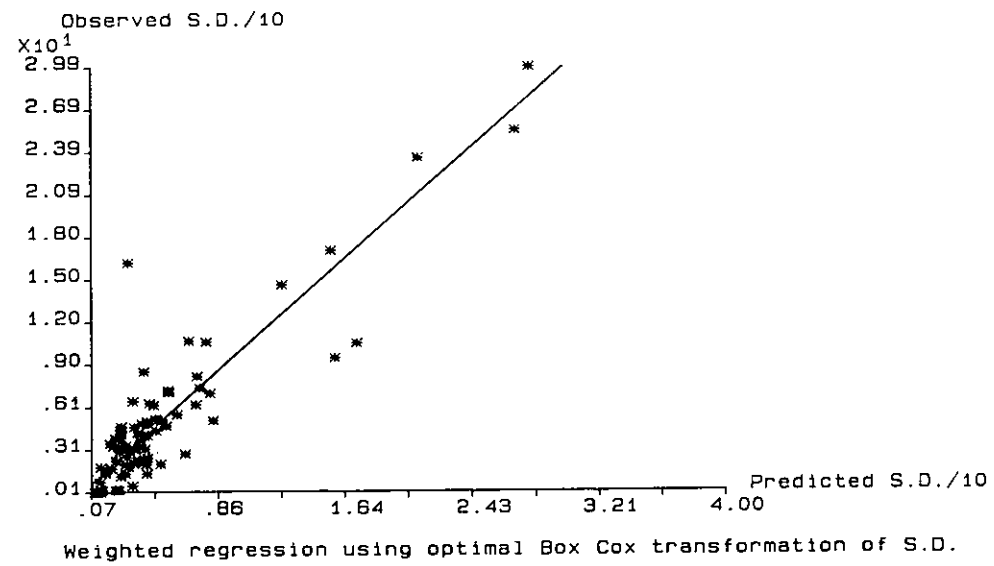


Fig. 24. Observed and predicted values of S.D.

which apparently have more variability in annual maximum flood than one would expect; the gauge numbers are 40 006 and 55 008. The feature is that they produce mostly relatively low annual maximum flood values except for a few exceptionally high ones, thus the mean is low but the standard deviation is relatively high.

Returning now to predicting $\bar{Q}$ from catchment characteristics a weighted regression with the catchment standard deviations predicted from the multiple regression on estimated standard deviation described above, proves to be much more satisfactory than one using the sample standard deviation for each catchment. More of the regression coefficients are significantly different from zero and there are very few outliers amongst the residuals, in fact just two again. These correspond to catchments 67 003 and 87 801 both of which are predicted to have a higher value of $\bar{Q}$ than they in fact achieve. The significant variables are AREA, S1085, SAAR, URBAN and SOIL. A Box–Cox transformation of $\bar{Q}$ could be considered within the weighted regression but the weights should be altered to take account of the transformation of $\bar{Q}$. The results of such regressions should not be taken too seriously. Weighted linear regression as described in Subsection 2.4.2 assumes the weights to be known exactly whereas here they are estimated and furthermore that estimation involved using variables that are then used as dependent and independent variables in the weighted regression.

5.2.4 Comparisons between regions

Let us now examine the wisdom of using a single countrywide equation to predict annual maximum flood. As was mentioned in Subsection 5.2.1, the Flood Studies Report suggested using a single equation but with regional multipliers, but a further alternative (also considered in the Flood Studies Report) would be to use a completely different equation for each region.

Although intended to give some coherence to the data analysis, the main purpose of this subsection is to illustrate the use of techniques described in Subsection 2.4.1, the comparison of several regression lines. Consequently, we shall not concern ourselves further with the problems of transformation or weighting but simply assume that a log transformation of all variables produces a set of data which satisfy the basic assumptions for a multiple regression analysis. It would, of course, be more correct to pursue the transformations examined in the previous subsections.

Data set X was drawn from nine regions (regions 1–9 in Sutcliffe (1978)) but there were insufficient catchments in each region to be able to treat them separately in a multiple regression analysis. Consequently four composite regions were formed, as follows:

Composite region	Region
A (West of England)	4, 8, 9
B (East of England)	5, 6, 7
C (North England, South Scotland)	2, 3
D (Northern Scotland)	1

Forming the analysis of variance table as described in Subsection 2.4.1 gives the following results:

Source	*Sum of squares*	*Degrees of freedom*
Overall regression	691.97	9
Difference in positions	10.57	3
Difference in regressions	76.13	27
Residual	211.26	865
Total	989.93	904

(Because for this data set we have repeated observations of the dependent variable for each set of values of the independent variable, it would have been possible to split the residual sum of squares into two components, a systematic departure from the regression lines component and a new residual sum of squares which would measure 'within catchment variation'. The form of the sums of squares would be similar to those given in Subsection 1.3.1).

Testing firstly for parallelism of the regression lines for the four composite regions ($\beta_{j1} = \beta_{j2} = \beta_{j3} = \beta_{j4}$ for $j = 1, 2, \ldots, 9$) gives a test statistic of $(76.13 \times 865)/(211.26 \times 27) = 11.54$ which should follow an F distribution with 27 and 865 degrees of freedom if the hypothesis is true. As $F(27, 865, 0.999) = 2.1$ we strongly reject the hypothesis of parallelism. Thus it would appear from the 83 catchments studied, that using separate regression equations for each region will provide a much more accurate description of the data than a single regression equation with regional multipliers.

If we insisted on a single regression equation but were unsure about the merits of regional multipliers as opposed to a single constant, then the second test given at the end of Subsection 2.4.2 ($\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$) would give some insight into this question. The test statistic is $(10.57 \times 865)/(211.26 \times 3) = 14.43$ and $F(3, 865, 0.999) = 5.4$ and consequently we must also reject this hypothesis. Thus, if we insist on using a single regression equation, it is much better to use regional multipliers than a single constant term.

Multiple regression computer programs are readily available these days but many only cope with a single set of data as described in Subsection 2.2.1 and do not have the immediate facility for handling several groups of data as described in Subsection 2.4.1. However, by using dummy variables (see Subsection 2.5.4), noting the generality of the general linear hypothesis method of testing (see Subsection 2.3.1) and running the program several times, the analysis just described can be performed. Three separate groups of runs are required:

(a) a multiple regression analysis separately on each composite region's set of data;
(b) a multiple regression analysis on the whole data set but with further independent variables which are dummy variables defining the particular composite region from which the data came;
(c) a multiple regression analysis on the whole data set ignoring the existence of composite regions.

For this data the set of runs under (a) produced the following residual sums of squares:

Table 6 Individual regressions on each region

Composite region	*Residual sum of squares*	*Degrees of freedom*
A	33.35	197
B	133.17	281
C	27.29	207
D	17.45	180
(Total)	211.26	865

Run (b) used three dummy variables which took the following values:

Composite region	x_{10}	x_{11}	x_{12}
A	0	0	0
B	1	0	0
C	0	1	0
D	0	0	1

and produced a residual sum of squares of 287.39 with 892 degrees of freedom. Run (c) produced a residual sum of squares of 297.96 with 895 degrees of freedom (see Subsection 5.2.1) and a total sum of squares of 989.93 with 904 degrees of freedom.

The analysis of variance table given previously in this subsection is now formed as follows.

Source	*Sum of squares*	*Degrees of freedom*
Overall regression	$989.93 - 297.96 = 691.97$	$904 - 895 = 9$
Difference in positions	$297.96 - 287.39 = 10.57$	$895 - 892 = 3$
Difference in regressions	$287.39 - 211.26 = 76.13$	$892 - 865 = 27$
Residual	211.26	865
Total	989.93	904

Returning to the interpretation of the analysis, Table 7 gives the significant regression coefficients for the four composite regions arising from run (a) mentioned above.

As the analysis of variance has suggested, there is considerable variation in the regression coefficients between the four composite regions. Although AREA

and STMFRQ have a roughly similar role in regions A, B and C they are not particularly relevant in region D. This probably reflects the nature of data set X, in that only catchments with small area were included, rather than that area is not, in general, a significant factor in predicting annual maximum flood. Similar comparisons may be made with the coefficients of other variables although one must always bear in mind that intercorrelation between these variables can lead to one regression coefficient being 'traded off' against another and even to one variable being omitted because of the inclusion of another (see Subsection 5.2.1).

Table 7 Regression of log annual maximum flood on log catchment characteristics

Composite region	*Constant*	*log AREA*	*log STMFRO*	*log S1085*	*log MSL*	*log SAAR*	*log RSMD*	*log (1+URBAN)*	*log (1+LAKE)*	*log SOIL*
A	−9.13	1.28	0.25	−0.31	−1.13	1.14		−29.87	−4.49	−1.88
B	−4.53	0.78	0.16	0.97	0.41	—	—	2.75	—	1.17
C	−4.39	0.82	0.38	−0.16	—	0.66	0.41	—	—	2.26
D	−15.8	—	—	0.74	2.67	1.05	—	—	11.18	−3.53

One final difference to point out is the much larger residual sum of squares associated with composite region B, evident in Table 6 and in the plot of observed and predicted values given in Figure 25. This appears to suggest that for the Eastern region of England, annual maximum flood is much more difficult to predict from catchment characteristics than for other parts of the country, but such a hypothesis would not be supported by physical considerations.

5.2.5 Examination of assumptions

The various models outlined above for relating annual maximum floods to catchments characteristics can be thought of as attempts to find a regression-like structure in which the residual errors have a constant variance—so that at least this one of the basic assumptions of standard regression theory would hold. However, the extent of departure from the other assumptions also needs consideration, together with the possible effect on any conclusions. For this example, it is reasonable, on an intuitive basis at least, that flood events on adjacent or neighbouring catchments will be correlated, so that residuals of annual maximum flood in the same year will also be correlated. While it is possible to account for correlated residuals within the overall analysis by using generalised least squares, in this case the estimation of these correlations is problematic in view of the small number of observations available to estimate each correlation. It would probably be enough to note that the correlation would generally be positive and that only a relatively small number of pairs of residuals are correlated, since residuals in different years are assumed independent. Thus the overall effect would be that the variances of the estimated parameters would be underestimated by a small amount if the analysis ignored the inter-correlation.

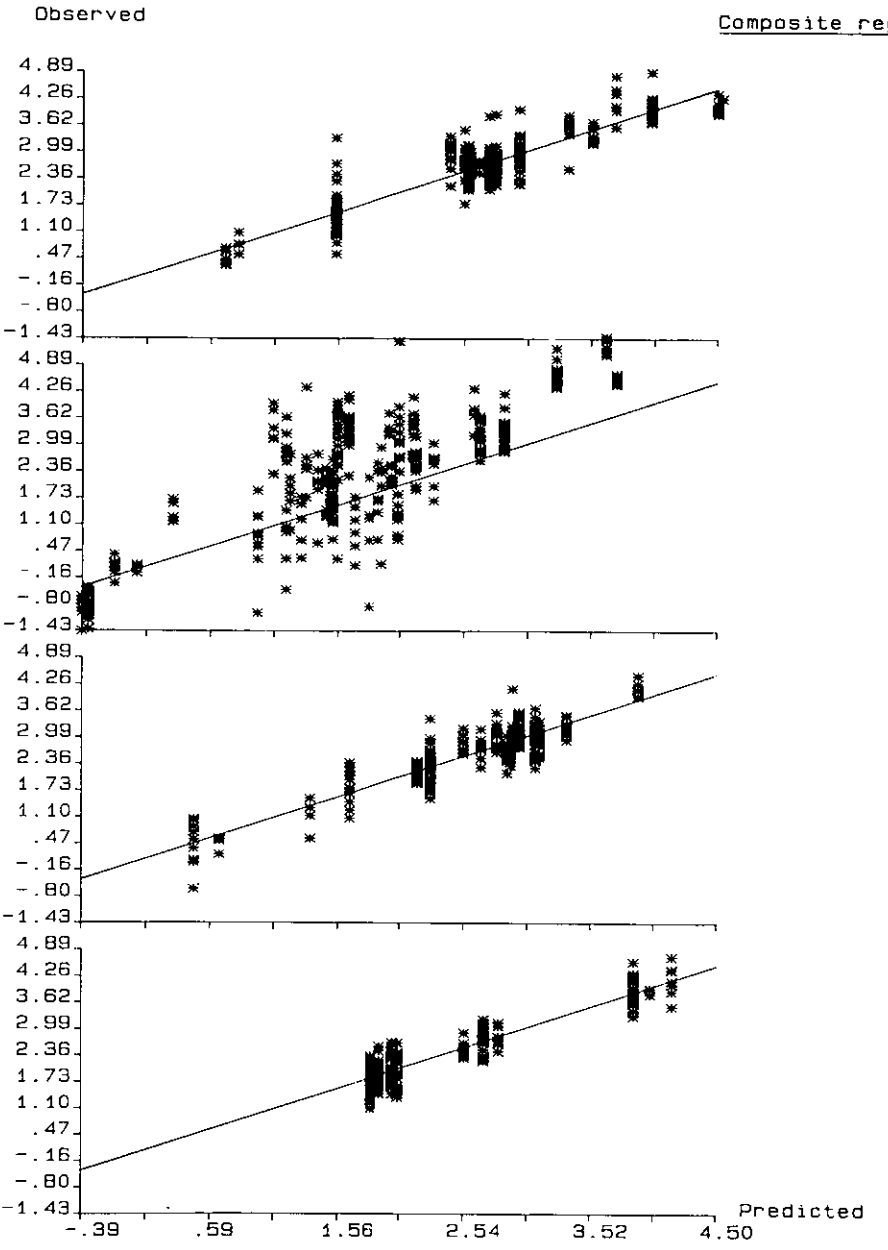


Fig. 25. Regression of $\log Q_{ij}$ on log catchment characteristics for composite regions.

The third assumption that would need to be checked is the Normality of the residual errors, for example, using the probability plots discussed in Subsection 3.2.3. To some extent a choice between competing models might be based on the closeness to Normality of the residuals, but this is possibly less important than other considerations such as the simplicity of structure of the final form of model.

5.3 Stepwise Regression—Choosing the Best Predictors

5.3.1 Introduction

Forecasts of monthly streamflow can be of use in a range of circumstances: in the efficient operation of dams for agriculture or hydropower, for example, or simply as background information providing some advance warning of whether flows will be high or low. The way in which such forecasts can best be obtained obviously depends to a great extent on the type of data that would be routinely available for making the forecasts, and on other general considerations such as the size of the catchment concerned. The problem examined in this section is the prediction of monthly streamflow in the Mekong at Pakse in Laos. Since the catchment here is extremely large (545 000 km^2) and is not all within one country, values for rainfall from an appropriate widely-spread set of sites would be difficult to obtain. Values for flow in major tributaries and for points upstream on the main channel would be more easily obtainable and more useful, but it might still not be possible to arrange to receive these values within sufficient time on a routine basis. Ideally such data might be used in some form of flow-routing model, perhaps operating on a daily time scale. The problem considered here will be that of predicting, in some simple way, future monthly total flows on the basis of routinely-made daily readings of flow at the same site. For this exploratory analysis the data used consisted of 48 years of records of monthly total flow, flow on the last day of the month and flow on the second last day.

Monthly flow varies considerably from month to month and within a given month, as illustrated by the means and estimated standard deviations given in Table 8 and the histograms given in Figure 26.

Table 8 Monthly flow (million cubic metres) in the Mekong at Pakse (April 1934–March 1982)

Month	*Mean*	*S.D.*	*Month*	*Mean*	*S.D.*
January	7652	1245	July	46596	11328
February	5333	824	August	73085	14111
March	4793	663	September	74437	13673
April	4440	785	October	45475	10431
May	7621	2107	November	21821	4617
June	23715	7392	December	11766	1971

The effect of the monsoon rains is beginning to be evident in May, but notice that the onset of the monsoon produces much more variability in flow (May and June) than a similar level of flow at the end of the rainy season (November and December). This fact and the suspicion that it will be unlikely that the flow in March and April will predict the time of arrival or magnitude of the monsoon rains suggests, even at the early stage, that it will be difficult to predict monthly flow in May or June.

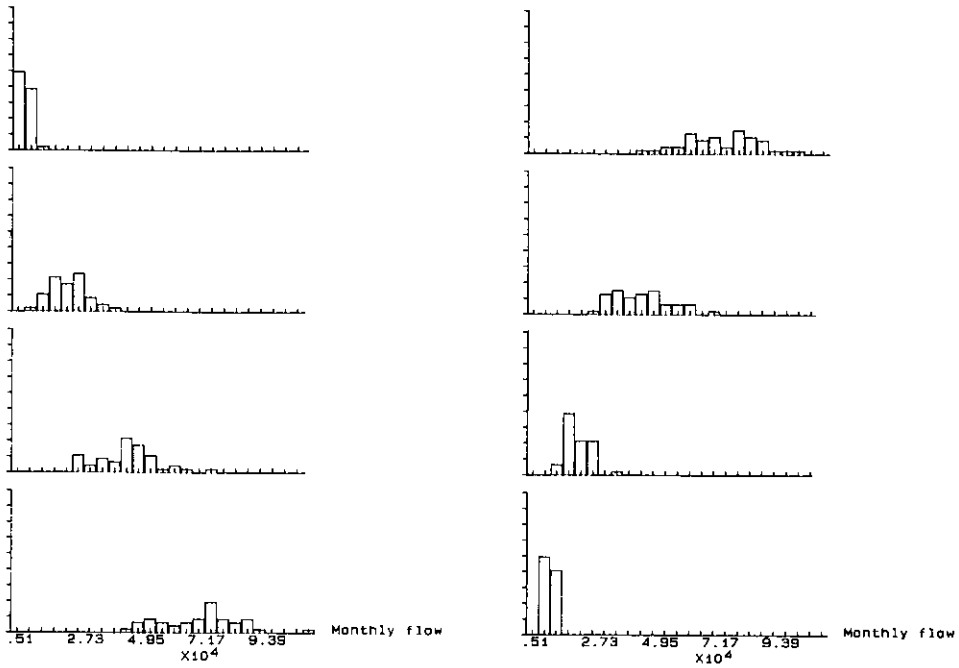


Fig. 26. Histograms of monthly flows in the Mekong at Pakse.

5.3.2 An example of stepwise regression

For a particular month, we wish to predict the flow, at a particular time, from any information available before that time. Thus, for August 1942, we could potentially draw on any values of monthly flow for months preceding that date (100 values in this data set), any values of flow in the last day of the month for months preceding that date, etc. In a regression context, if we take the dependent variable y to be monthly flow in August and the independent variables, flows in previous months, we may state the objective to be to predict monthly flow in August, but other features of the problem do not immediately fit into the regression mould.

First of all the independent variables may well be previous values of the dependent variable. Furthermore, the number of independent variables will increase, as values of the dependent variable are taken from more and more recent years. Problems arising from using previous (lagged) values of the dependent variable as independent variables were discussed at the end of Subsection 3.4.2. Clearly we should proceed with some caution. As far as the problem of an increasing number of independent variables is concerned, one approach, and it is not the only approach, would be to decide beforehand how many years' previous information to use. Studying a correlation matrix of flow for the month in question versus flows in previous months might suggest that there is little point in using information from more than two years ago.

For this particular set of data, it would appear that there is little value in information more than three years before the month in question. Adopting that suggestion, the number of available sets of observations will be 45 but the

potential number of independent variables is still very large. The number of independent variables must be at least one less than the number of observations for a solution to equation (35) to exist. Consequently, we must either exclude further independent variables at this stage or use forward selection (Subsection 2.3.5) or stepwise regression (Subsection 2.3.7) to build up gradually a regression equation from the pool of available independent variables. In that stepwise regression is a forward selection with additional safeguards, the former is usually preferable.

To illustrate the technique let us apply stepwise regression to monthly flow in January, with independent variables Mr, LDr, SDr ($r = 1, 2, 3, \ldots, 12, 24, 36$) where Mr, LDr and SDr indicate monthly flow, flow on the last day of the month and flow on the second last day of the month for that month which is r months before that of the dependent variable. Following the steps as outlined in Subsection 2.3.5, Step 1 is to perform simple linear regression of y on each of the independent variables. Testing the significance of the hypothesis $b = 0$ (which is equivalent to testing the significance of the correlation coefficient between y and the relevant independent variable) gives the set of independent variables which are significantly associated (using, in this case, a 5% significance level) with monthly flow in January. Starting with the most highly significant, they are: $LD1$, $SD1$, $M1$, $LD11$, $SD11$, $M11$, $M10$, $SD12$, $M12$, $LD12$, $M9$, $LD10$, $SD10$, $M7$, $SD7$, $LD7$, $LD8$, $LD2$, $SD9$, $SD8$, $SD2$ and $LD9$. Consequently, we select $LD1$ for inclusion in the regression equation.

Step 2 requires computing the partial correlation coefficients between monthly flow in January and each independent variable conditional on $LD1$. As an example, let us compute the partial correlation of monthly flow in January (M) with $LD12$. The correlations we need for equation (49) are as follows:

Variables	*Correlation*
$LD12$ and $LD1$	0.304
$LD12$ and M	0.452
$LD1$ and M	0.887

The partial correlation between M and $LD12$ is, from equation (49), 0.415. Repeating this calculation for each independent variable gives the set of independent variables which might be added to the regression equation at this stage. They are, in order of absolute value of partial correlation coefficient, $M9$, $M10$, $SD12$, $LD12$, $SD10$, $LD10$, $M12$, $SD9$, $LD11$, $SD11$, $M11$ and $LD9$.

Step 3 will mean including $M9$ in the regression equation. Testing the joint significance of $LD1$ and $M9$ using the analysis of variance table given in Subsection 2.3.2 gives a test statistic of 96.1 which means that we must strongly reject the hypothesis of no linear association between M and the independent variables, $LD1$ and $M9$. As mentioned in Subsection 2.3.2, the merit of $M9$ in addition to the variables already included in the model ($LD1$) may be judged by comparing the ratio of the regression coefficient of $M9$ and its standard error to the appropriate t distribution. The ratio is 3.2 and consequently we strongly

reject the hypothesis that $M9$ should be omitted from the model when $LD1$ is included.

Stepwise regression differs from forward selection in that it also examines the possibility, at each stage, of excluding a variable already included in the model. Thus, at this stage, we should also examine the value of $LD1$ once $M9$ has been included in the model. Using the same test as for $M9$ gives a ratio of 12.3 which, again, strongly supports the inclusion of the variable ($LD1$) in the model.

Step 4 examines the partial correlation between M and the other independent variables, conditional on $LD1$ and $M9$. The largest partial correlation is between M and $SD12$ ($=0.30$) and the next largest is between M and $LD12$ ($=0.28$). Including $SD12$ in the model and repeating step 3 leads to a significant overall model, but the individual contribution of $SD12$ is not significant, giving a ratio of 1.8. Thus the procedure stops with a fitted model

$$M = 189.48 + 19.54 \times LD1 + 0.38 \times M9$$

Table 9 summarises the outcome of applying the stepwise procedure to each month individually:

Table 9 Stepwise regression for each monthly flow

Dependent variable: Monthly flow in—	*Independent variables (in order of selection)*	*Multiple correlation coefficient*
January	*LD1*, *M9*	0.91
February	*M1*, *LD2*, *SD7*, *LD6*	0.95
March	*M1*, *LD12*, *M7*	0.91
April	*LD1*	0.83
May	*SD3*, *SD6*	0.53
June	*LD1*, *LD8*	0.51
July	*LD1*, *M12*	0.63
August	*LD1*, *LD3*, *SD1*, *SD11*	0.73
September	*LD1*, *LD24*, *M4*, *LD12*, *LD6*, *SD1*	0.77
October	*LD1*, *SD1*, *LD2*	0.86
November	*LD1*, *LD3*, *SD3*	0.82
December	*LD1*, *SD10*, *SD4*	0.84

There are several aspects of Table 9 on which to comment but the one most frequently taken for granted is the tremendous 'data reduction' which stepwise regression achieves. As can be seen from the detailed description of the analysis of the January data, the collection of variables correlated with monthly flow is reduced to a small subset which contains most of the pertinent information. It does not follow that the variables listed above are the 'correct' set or that they describe the 'true' relationship. In interpreting such a table it is always worth considering those variables that just missed being included, perhaps because of high correlation with an already included variable. The best that can be thought of the relationships set out in Table 9 is that within a certain framework they best represent the data provided. They are only as good as the data reflects the nature and extent of the phenomenon being studied.

Turning now to the independent variables selected by stepwise regression, it is immediately apparent that, as would be expected for predicting monthly flow, some feature of flow in the previous month is the most important variable, amongst the variables considered. The one exception is the month of May when, almost certainly, it is the timing of the onset of the monsoon which has most influence on the monthly flow. Notice, however, the strong preference for the flow on the last or second last day, rather than monthly flow, as a predictor of the following month's flow. The appearance of $LD12$, $M12$ or even $LD24$ in certain months suggests a pattern from year to year for these months.

Long range dependencies such as seen here are not necessarily unrealistic, since part of the flow of the Mekong derives from glaciers in the Himalayas and glacier-fed rivers often exhibit fluctuations in flow over periods of years related to the extension and recession of the glaciers. However, the extent of any man-made influences would also need to be checked but are likely to be small in comparison with the large natural flows.

Too literal an interpretation of the inclusion of some of the other variables may be misleading. For instance, in November, $LD3$ would appear to be of importance yet $LD2$ is not. However, the correlations of $LD2$ and $LD3$ with M are 0.13 and -0.04 respectively, indicating that when judged on their own, $LD2$ has a stronger association with M than has $LD3$. The reason for $LD3$ being included in preference to $LD2$ is that the partial correlations of $LD2$ and $LD3$ with M, after allowing for the effect of $LD1$, are -0.17 and -0.30 and thus $LD3$ provides the most additional information after $LD1$ has been included.

However, there are some interesting patterns which are worth noting such as the inclusion of $SD1$ as well as $LD1$ for the three consecutive months of August, September and October. In each case $LD1$ has a positive regression coefficient whereas $SD1$ has a negative regression coefficient which suggests that this combination of variables may be detecting whether the peak flow following the monsoon has been passed and that flows are now decreasing, or whether flows are still rising. A similar explanation might be the reason for the inclusion of $LD2$ in addition to $M1$, in the independent variables for February. The regression coefficient for $LD2$ is negative and thus the combination of $M1$ and $LD2$ may reflect whether monthly flow is still falling in the preceding months and, if so, how rapidly. Given these observations and the fact that LDr and SDr are highly positively correlated, one might consider replacing them by their sums and differences in the set of independent variables.

Various other patterns would be worth exploring such as the inclusion of $LD2$, $SD3$ and $SD4$ in October, November and December respectively, but the purpose of this section is to illustrate the use of regression techniques as opposed to a detailed analysis of the data set. However, it is perhaps important to draw attention to one more obvious point, namely that the months around the onset of the monsoon rains prove to be the most difficult to predict. Values of the multiple correlation coefficient are at their lowest in May and June but then climb steadily thereafter to peak in the months of January to March where flow is steadily falling in stable conditions. It is interesting to note that flow in April is less easy to predict than flows in previous months perhaps partly due to the occasional early monsoon and partly because flows have reached their lowest level and it is unexpected events which will alter that flow.

In Subsection 4.3.1 it was suggested that forming a histogram of the residuals and plotting the residuals against various other variables might be informative in assessing the validity of the assumptions made in multiple regression, as well as giving some further insight into the data. Figure 27 gives histograms of the residuals on the far right of the graphs and a plot of $\hat{e}_1, \ldots, \hat{e}_n$ versus $y_1, \ldots, y_n$ on the left. Notice that the months whose histogram of residuals show the least spread do not correspond to months with the highest multiple correlation coefficient in Table 9. The multiple correlation is essentially a ratio of residual variance to total variance. When this ratio is small, and hence when prediction is successful, the multiple correlation will be large. We can assess this ratio by eye from the plots of $\hat{e}_1, \ldots, \hat{e}_n$ against $y_1, \ldots, y_n$ by comparing the spread of points in the 'y axis' direction with that in the 'x axis' direction. Where this ratio is smaller, for example in October, November and December, the multiple correlation will be high. However such a plot should not be used to assess the validity of the model. It can be shown that, under the usual assumptions for multiple regressions, $\hat{e}_1, \ldots, \hat{e}_n$ and $y_1, \ldots, y_n$ will be correlated and hence one might expect a trend of the type evident in these plots even when all the assumptions of multiple regression hold. Indeed in a variety of quite different circumstances, described in Hoaglin *et al.* (1983) one might expect the plot to follow a 45° line exactly.

It may also be demonstrated that $\hat{e}_1, \ldots, \hat{e}_n$ and $\hat{y}_1, \ldots, \hat{y}_n$ are uncorrelated and hence it is more usual to plot these two quantities, as in Figure 28. If we had seen a trend in this graph of the type evident in Figure 27, we might have concluded that the model was quite inadequate, overpredicting low values and underpredicting high values. However, there is no such trend. There is some evidence of unsatisfactory prediction in the July figures, as shown by a skewed histogram of residuals and an uneven spread of points in the vertical direction, indicating that although a few values are noticeably underpredicted the majority are overpredicted. A transformation of the dependent variable may help in this situation.

5.3.3 Some further regressions

While scanning the data in the hope of spotting some patterns hitherto undetected, it became apparent that there was the slight suspicion of a cyclical trend over the years in a given month's values. Although such a trend would be extremely hazardous to employ in any predictor, that is unless its physical mechanism could be understood, it provided a nice opportunity to apply the technique of periodic regression (see Subsection 2.5.3). Analysing each month's results separately, cosine and sine terms of periods ranging from 3 to 12 years were used as the independent variables in a regression on monthly flow. If the reader has access to a multiple regression program or package but which does not contain periodic regression, he may prefer to carry out the calculations by running the regression program with the independent variables as the cosines and sines of the required periods rather than writing separate code for the formulae given in Subsection 2.5.3. This may be extremely inefficient in computer time as it will lead to the unnecessary inversion of a $2k \times 2k$ matrix when terms of period $n, n/2, \ldots, n/k$ are used but it may be a

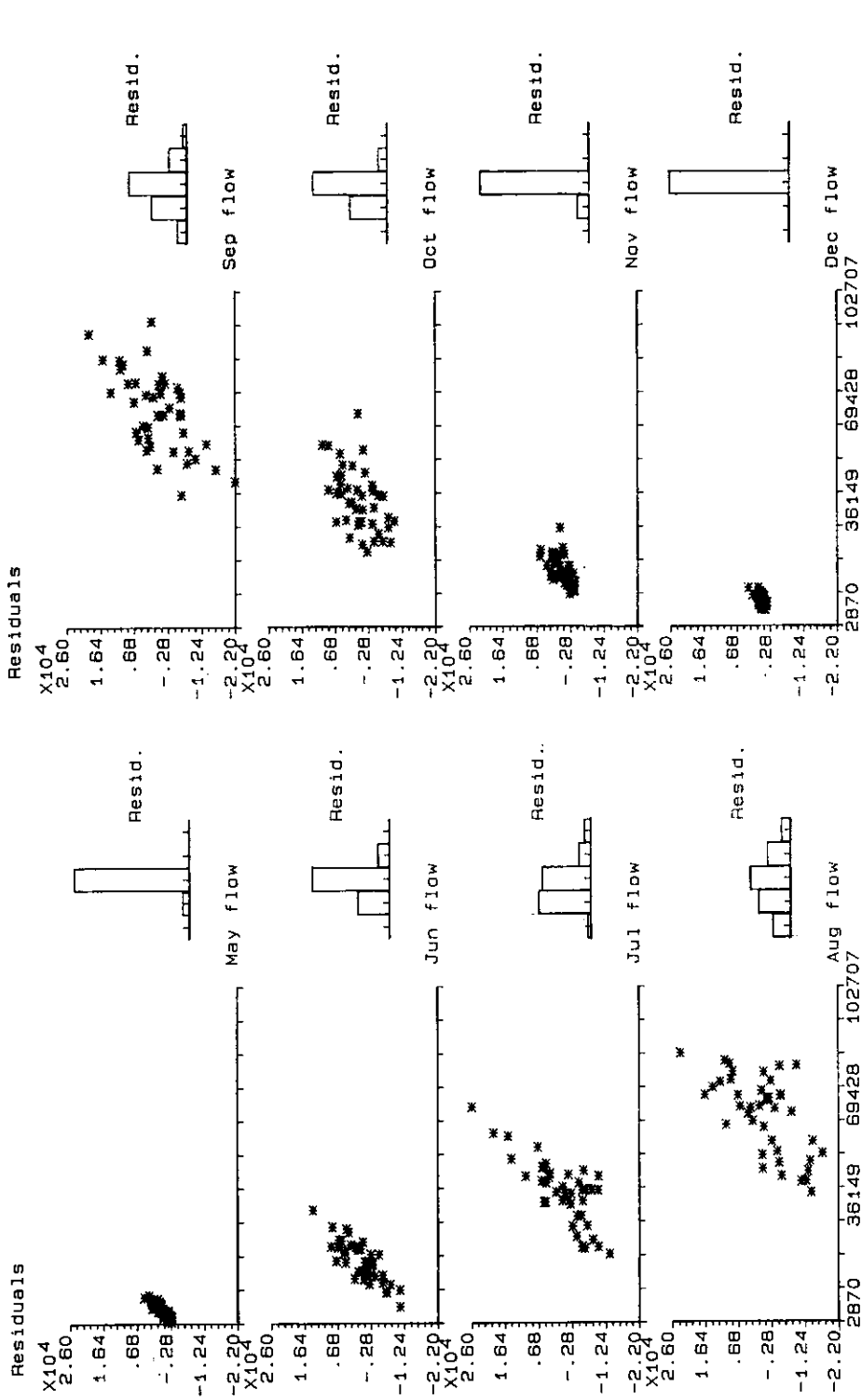


Fig. 27. Regression of monthly flow on the best predictors—residuals versus observed flows.

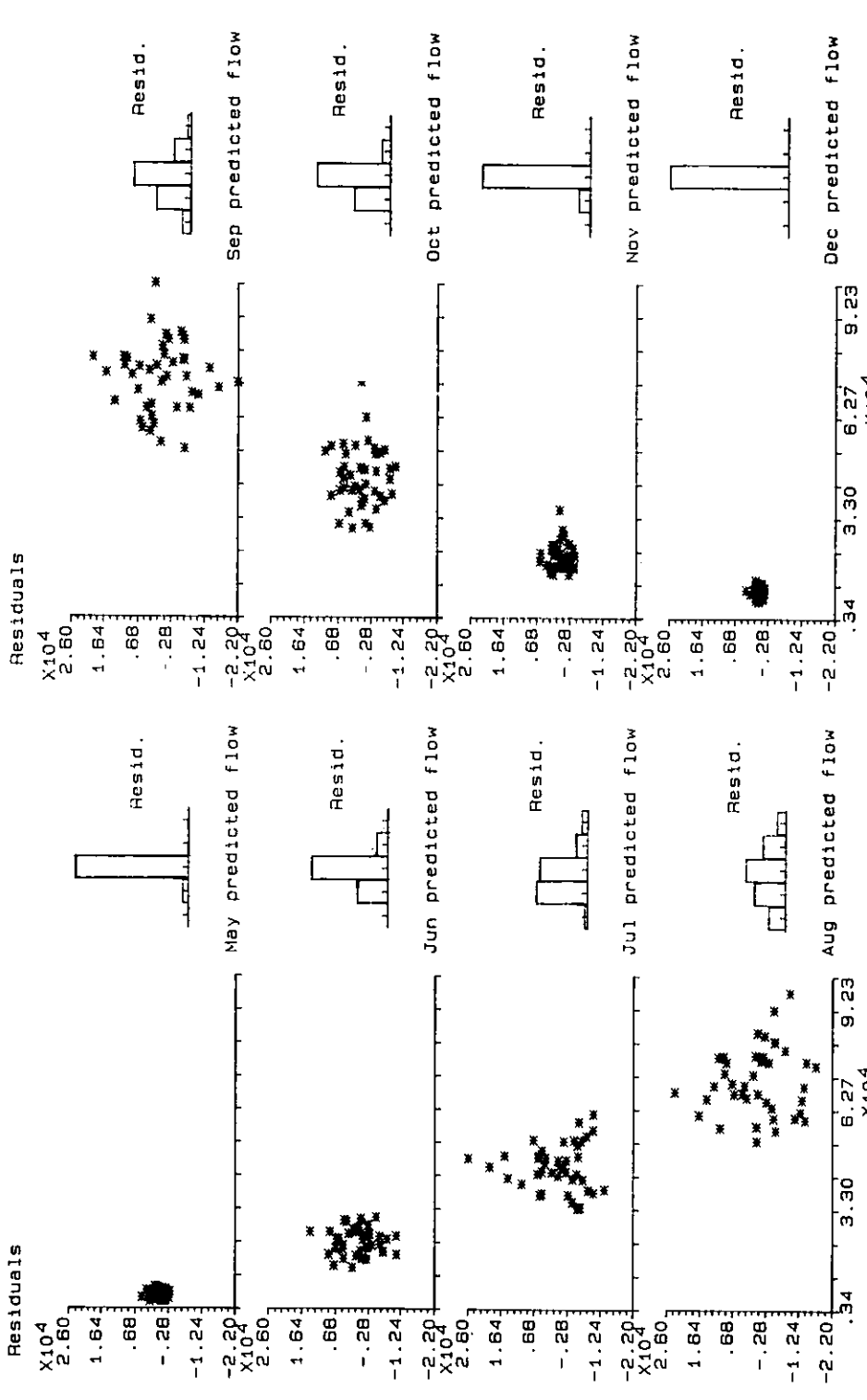


Fig. 28. Regression of monthly flow on the best predictors—residuals versus predicted flows.

more efficient use of the individual's time if it is only intended to perform such regressions occasionally. However, with a stepwise regression program, it would be advisable to ensure that the pair of variables (cosine and sine) of the same period are either included or excluded at each stage.

Table 10 sets out the periodicity of the terms which were significant in this analysis. Thus, for monthly flow in January, there is evidence of two cycles, the stronger being a nine-year cycle and the weaker a five-year cycle. To find any evidence of such a cycle is surprising but what is even more interesting is that a nine-year cycle appears to be evident in consecutive months from December to to May (excluding April). Adjacent months are strongly correlated and so some similarity of year-to-year fluctuations is to be expected, but this degree of consistency does suggest that the nine-year cycle may be more than just an artefact.

Table 10 Periodic regression in monthly flow

Month	*Periodicity of significant terms*	*Multiple correlation coefficient*
January	9 , 5	0.42
February	9 , 8	0.46
March	9	0.28
April	—	—
May	9	0.30
June	—	—
July	8 , 5	0.40
August	11 , 3	0.42
September	—	—
October	—	—
November	12	0.35
December	9	0.26

The suggestion to apply periodic regression to monthly flows came from studying the original data and noticing some pattern. In a similar way, when studying the data and considering the problem of predicting monthly flow in May, the main problem to emerge was to predict when the monsoon would arrive. Monthly flows for months preceding May would seem to be of no great value in predicting the onset of the monsoon except, perhaps, in recording when the previous rainy season finally ended. The highly tenuous hypothesis would then be that when the previous rainy season is abnormally late, this may be associated with a later arrival of the next rainy season. Again, from inspecting the data, there did appear to be some suggestion of this phenomenon, although it is all to easy to see what you want to see in a set of data.

Pursuing this idea partly out of interest and partly for illustrative purposes, we will need some indicator of when the rainy season ends. Such a variable could then be included amongst the independent variables in a regression on monthly flow in May. No such variable was observed directly in this data set

but it may be possible to construct such a variable from those observed. There is no reason why the independent variables used in a linear regression should not be some function of those variables which might have been used as independent variables. The major effect of the rainy season on flow seems to be coming to an end by December and thus an indicator of a late end to the rainy season might be x_1 = (Flow in December)/(Flow in January). A high value for this variable would indicate a later duration of the rainy season provided, that is, that the effect of the rains did not continue unabated through December and January. Perhaps the further inclusion of a variable x_2 = (Flow in January)/(Flow in February) may be a safeguard against this unlikely alternative.

Taking an extremely simplistic view of predicting the flow in May, one might include a variable which simply attempts to make a proportionate change to the monthly flow in the previous May, such as

$$x_3 = \left(\frac{\text{monthly flow in April of year } x}{\text{monthly flow in April of year } (x-1)}\right) \times (\text{monthly flow in May of year } (x-1))$$

The general point to be made is that it might be worth giving some thought to constructing artificial independent variables which, perhaps, more directly reflect the phenomenon being investigated. Not surprisingly there are many pitfalls to such adventurous use of the data, especially in this context where some of the independent variables are lagged values of the dependent variable.

The outcome of additionally using x_1, x_2 and x_3 described above in predicting y = monthly flow in May is quite interesting. The artificial variable x_3 is significantly correlated with y, but there is a larger correlation between $SD3$ and y and hence a stepwise regression includes $SD3$ as a first step. After including $SD3$, the importance of x_3 falls and $SD6$ is, once again, the next most important variable. However, it is at this point that x_1 emerges as the next most important factor with x_3 no longer of much importance at all. As would be expected, the regression coefficient of x_1 is negative thus supporting the hypothesis that a late end to the previous rainy season (high value of x_3) will lead to a later onset of the next rainy season (lower flow in May). At no stage does x_2 appear to be of any importance. Using x_1 in addition to $SD3$ and $SD6$ increases the multiple correlation with monthly flow in May to 0.56.

5.3.4 A simple predictor for monthly flow

As was mentioned in Subsection 5.3.1, the particular problem in mind, which led to this data set being collected, was to predict monthly streamflow in the Mekong from measurements of previous flows. A further objective was that this predictor should be as simple as possible, and it was for this reason that transformations of the dependent and independent variables were not considered. There is, of course, no one interpretation of the term 'simple predictor'. It could be one that is easy to calculate, one whose required measurements are easy to collect or one whose form is similar from one month to another. Once again, our objective will give the opportunity to demonstrate a range of applications of multiple regression but the ensuing analysis should

not, under any circumstances, be regarded as a model analysis of the set of data.

As our first attempt, let us consider a single equation to predict all monthly flows, but just as we used 'regional multipliers' in Section 5.2, so here we will include individual constants for the months by introducing eleven dummy variables (as in Subsection 5.2.4). A stepwise regression produces a very neat solution; the selected subset of independent variables out of the set Mr, LDr, SDr ($r = 1, 2, 3, \ldots, 12, 24, 36$) consists of $LD1$, $LD2$ and $LD3$ together with dummy variables for months May, June, July, August, September and October. Again, the last day flow appears to be more useful than the monthly flow.

Table 11 Correlation between observed and predicted monthly flow

Month	*Stepwise regression for each month (equation A)*	*Single equation with individual monthly constants (equation B)*	*Predictor using periodic equation for LD1 coefficient (equation C)*
January	0.91	0.77	0.89
February	0.95	0.89	0.89
March	0.91	0.84	0.84
April	0.83	0.82	0.83
May	0.53	0.42	0.41
June	0.51	0.43	0.43
July	0.63	0.58	0.58
August	0.73	0.56	0.54
September	0.77	0.56	0.56
October	0.86	0.79	0.78
November	0.82	0.71	0.78
December	0.84	0.70	0.79

Table 11 gives the multiple correlation coefficients for various attempts at predicting monthly flow. The first column simply reproduces those values, given in Table 9, derived from applying stepwise regression to each month separately. To some extent this column will be taken as a reference level, as amongst linear functions of the independent variables it should be close to the best that can be achieved. However, within the context of a simple predictor, it would be cumbersome to implement as it requires quite different collections of variables to predict different monthly flows. We will refer to the predictors given by applying stepwise regression in this manner as equation A. We see from Table 11 that the single equation predictor described above (now referred to as equation B) is generally quite good, giving correlations which are quite close to these achieved by equation A. Except for the difficulty of predicting the May monthly flow, it is August/September and December/January flows which are noticeably less successful with equation B.

For our second attempt at a simple predictor we might try to accommodate the variation from month to month in the relationship of monthly flow with the available independent variables. An approach to this objective might be to

decide on a small collection of independent variables which appear in most of the predictors given in Table 9, compute separate regression equations for each month and then try to establish a fairly simple relationship to describe how the regression coefficients of a particular variable vary over the months. To illustrate this approach, the collection of variables $LD1$, $LD3$, $LD6$, $LD9$ and $LD12$ were selected, partly because they cover the preceding twelve months in a reasonably uniform way and partly because of their popularity (or that of a near neighbour) in the regressions summarised in Table 9. It is of some benefit in this context to scale both the dependent and the independent variables to have zero mean and unit standard deviation across the data set studied. With such a standardisation, the regression coefficient in a regression of y on x would just be the correlation coefficient between y and x. Carrying out that standardisation and computing separate regressions for each month of monthly flow versus $LD1$, $LD3$, $LD6$, $LD9$ and $LD12$ gives the set of regression coefficients for $LD1$ given in the first column of Table 12.

Table 12 Calculated and predicted regression coefficients for $LD1$

Month	*Calculated regression coefficient*	*Predicted regression coefficient*
January	0.84	0.88
February	0.88	0.84
March	0.77	0.74
April	0.89	0.62
May	0.20	0.51
June	0.39	0.44
July	0.55	0.42
August	0.53	0.47
September	0.44	0.56
October	0.79	0.68
November	0.81	0.79
December	0.77	0.87

As might have been expected, there appears to be a cyclical trend in these coefficients and hence a possible model to describe these values might be a periodic regression with one sine and cosine term each of period 12 months. Fitting such a model as described in Subsection 2.5.3 gives the fitted model $LD1$ regression coefficient $= 0.653 + 0.215 \cos(2\pi r/12) + 0.0911 \sin(2\pi r/12)$ where $r = 1, 2, \ldots, 12$ for months January, February, etc., respectively. Using this equation to predict the regression coefficients of the standardised version of $LD1$ gives the set of values given in the second column of Table 12.

Ignoring the variables $LD3$, $LD6$, $LD9$ and $LD12$ and just using the formula

$$\frac{\text{Flow in month } r - \text{mean}}{\text{S.D.}}$$

$$= \left(0.653 + 0.215 \cos\left(\frac{2\pi r}{12}\right) + 0.0911 \sin\left(\frac{2\pi r}{12}\right)\right)\left(\frac{LD1 - \text{mean } LD1}{\text{S.D. of } LD1}\right)$$

(now referred to as equation C) to predict the flow in each month gives the set of correlations given in column 3 of Table 11. Notice that, because each monthly flow has been standardised about its monthly mean, this regression is equivalent to fitting separate constant terms for each month and in that sense is similar to equations A and B.

Comparing columns 2 and 3 of Table 11, there is very little to choose between the two predictors. They give almost equally good prediction for most months but equation C has a marginally better performance in November–January. Which of the predictors to recommend would depend largely on other factors such as the purpose for which the predictor is to be used. Is it primarily for monitoring the months with high flow, the months with low flow or flow in the months at the beginning of the rainy season? Equally, which formula is actually most useful will depend on other factors.

In principle, tables of month-by-month coefficients could readily be supplied for any of the predictors to keep computation to a minimum. While a predictor which uses only the single independent variable $LD1$ is attractive, there would be little extra difficulty in implementing predictors based on more variables, particularly if hand calculators were used. There may be a preference for predictors whose coefficients change smoothly over the year, since then one would feel confident about interpolating the coefficients to produce forecasts of total flow, from the middle of one month to the middle of the next, based on the latest available day's flow and on total flow over the previous mid-month to mid-month, and so on.

Our final attempt at a simple predictor stems from regarding equation C, because it uses standardised variables, as predicting the departure from the long run mean for a given month. This suggests predicting the long run monthly value with one equation which makes use of previous flows for that month and then predicting the departure from the long run value with an equation which makes use of flows in the immediately preceding months. The predictors from such a two-stage predictor cannot be better than that from a single equation incorporating all of the variables (in a linear model context). However, a two-stage predictor may be easier and more flexible to administer and the equations will probably be easier to interpret than the rather haphazard collection of variables which stepwise regression produces.

Let us start, therefore, by considering the prediction of monthly flow from previous flows for that month. Using the previous three years' monthly flows and flow on the last day as independent variables, a separate regression for each month gives a set of predictors referred to as equation D, whose multiple correlation coefficients are given in the first column of Table 13.

Not surprisingly, it is the immediately preceding year's flow which figures most prominently in these predictors and there are an equal number of cases in which monthly flow and flow on the last day of the month are the preferred variables. Also, not surprisingly, it is the more stable months of January to April which are most amenable to a predictor of this type. The months of May and June have proved difficult to predict in all of the approaches adopted so far but we see here that flows in the ensuing months of August to December are not easily predicted from long term historical data.

The second stage of this two-stage predictor is to incorporate flows from the

Table 13 Two stage predictor of monthly flow

Month	*Stage one predictor (equation D)*	*Stage two 'correctors'*	
		Using errors in previous months	*Additionally using flows in previous months*
January	0.62	0.85	0.93
February	0.57	0.81	0.92
March	0.68	0.90	0.89
April	0.53	0.77	0.81
May	0.42	0.38	0.41
June	0.45	0.49	0.49
July	0.39	0.56	0.58
August	0.21	0.40	0.79
September	0.47	0.64	0.74
October	0.39	0.59	0.87
November	0.47	0.62	0.82
December	0.30	0.72	0.82

immediately preceding months. Two examples are given of how this might be accomplished. The first involves using stepwise regression with a dependent variable, the flow for the month in question, and with independent variables, the predicted monthly flow using equation D for the month in question, and the difference between the predicted monthly flow by equation D, and the observed monthly flow for each other month, making twelve independent variables in all. Thus our second stage is very much in the nature of a 'corrector', using the predicted value for the month in question and the errors in prediction for the previous months. Column 2 of Table 13 gives the multiple correlation coefficients for this set of regressions which compare reasonably well with those for equations B and C given in Table 11. However, it is the set of variables

Table 14 Variables selected for the 'corrector', first example

	Jan.	*Feb.*	*Mar.*	*Apr.*	*May*	*Jun.*	*Jul.*	*Aug.*	*Sep.*	*Oct.*	*Nov.*	*Dec.*
January	P					P–O						P–O
February	P–O	P										
March		P–O	P					P–O				
April		P–O	P–O	P								
May		P–O										
June						P						
July						P–O	P					
August							P–O					
September					P–O				P			
October									P–O	P		
November										P–O	P	
December											P–O	P

selected by this stepwise regression which is the more interesting result from this analysis.

Table 14 summarises the variables selected. For a given month (row), a *P* in a particular column indicates that the predicted value for that month (column) using equation D was selected as an independent variable and a *P–O* indicates that the difference between the predicted and observed values for that month (column) was selected. Thus the dominant pattern is, in 8 out of 12 cases, the selection of the predicted value using equation D for the month in question and the difference between observed and predicted for the immediately previous month as independent variables. This very simple structure may make this particular two-stage procedure very easy to operate in practice.

As a second example, the set of independent variables used in the first example was augmented by including the variables *LDr* and *SDr* ($r = 1, 2, \ldots, 12$). Although this leads to a set of multiple correlations (column 3 of Table 13) which are very similar to those achieved with equation A (see Table 11, column 1) the set of variables selected (Table 15) is a more confusing mixture than in the first example.

Table 15 Variables selected for the 'corrector', second example

	Jan.	*Feb.*	*Mar.*	*Apr.*	*May*	*Jun.*	*Jul.*	*Aug.*	*Sep.*	*Oct.*	*Nov.*	*Dec.*
January			S						P–O			L
February							P–O					
March		S	P									
April			L									
May		S										
June						P						
July						L						
August					L		L, S		S, P–O			
September					P–O			L	P			
October		P–O							L, S			
November										I	P	
December		P–O									L	

The notation used in Table 15 is the same as in Table 14 with the addition of symbols L and S to indicate flow on the last day and second to last day respectively. There is again a strong diagonal tendency in the table but the pattern of the (*P–O*) variables is very haphazard. As one might expect, there is a strong similarity in the months of the variables selected and the months of the variables selected for inclusion in equation A (see Table 9). However, although this second example produces the better set of multiple correlations, for simplicity of predictor, the approach in the first example may be preferable. As mentioned earlier, which predictor would actually be the simplest to use in practice depends on many more factors than the mathematical complexity of the equation suggested. However, having the objective of simplicity has been a convenient way of demonstrating a variety of ways of using multiple regression.

References

Flood Studies Report (1975). Natural Environment Research Council, London.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (1983). *Understanding Robust and Exploratory Data Analysis*. John Wiley, New York.

Robinson, M. (1980). The effect of pre-afforestation drainage on the stream-flow and water quality of a small upland catchment. Institute of Hydrology Report No. 73.

Seber, G. A. F. *Linear regression analysis*. John Wiley, New York.

Sutcliffe, J. V. (1978). Methods of flood estimation: a guide to the Flood Studies Report. Institute of Hydrology Report No. 49.

POSTSCRIPT

Since this book was originally drafted, several computer packages of regression programs have been introduced. The availability of these packages removes many of the computational problems associated with the use of multiple regression techniques. However, an understanding of the theoretical basis of multiple regression techniques is as important as ever. A short reference list is given to some computer packages which contain a substantial number of regression programs. The list is by no means exhaustive, nor is it intended to single out the 'best' packages. In addition, a further reference list is given to some currently available regression books which may be used to supplement or extend the material presented in this text.

Finally, some references are given to a few recently published research papers. These may help to give some idea of the direction of current thinking although the reader can gain a more complete picture by referring to Section 6.1 of recent issues of Statistical Theory and Methods abstracts, published by the International Statistical Institute.

References

1. Computer packages

Alvey, N. G. *et al.* (1977). GENSTAT. *A general statistical program.* Rothamsted Experimental Station.

Baker, R. J. and Nelder, J. A. (1978). *The GLIM System, Release 3.* Numerical Algorithms Group, Oxford.

Dixon, W. J. and Brown, M. (1985). *BDMP Biomedical Computer Programs, 1985.* University of California, Berkeley.

Dongarra, J., Bunch, J. R., Moler, C. B. and Stewart, G. W. (1979). *The linpack users guide. SIAM.* Philadelphia.

Nie, N. H., Hull, G. H., Jenkins, J. G., Steinbrenner, K. and Bent, D. H. (1975). *SPSS Statistical package for the social sciences.* McGraw-Hill, New York.

Numerical Algorithms Group (1982). *NAG Fortran Library Manual (Mark 9).* NAG, Oxford.

2. Further reading

Belsey, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression diagnostics: influential data and sources of collinearity*. John Wiley, New York.

Chatterjee, S. and Price, B. (1977). *Regression analysis by example*. John Wiley, New York.

Cook, R. D. and Weisberg, S. (1982). *Residuals and influence in regression*. Chapman and Hall, London.

Draper, M. and Smith, H. (1981). *Applied regression analysis*. John Wiley, New York.

Seber, G. A. F. (1977). *Linear regression analysis*. John Wiley, New York.

Vinod, H. D. and Ullah, A. (1981). *Recent advances in regression methods*. Marcel Dekker, New York.

Weisberg, S. (1980). *Applied linear regression*. John Wiley, New York.

3. Recent research papers

Copas, J. B. (1983). Regression, Prediction and Shrinkage. *J. Roy. Stat. Soc. B*, **45**(2), 311–54.

Friedman, J. H. and Stuetzle, W. (1981). Projection Pursuit Regression. *J. Am. Stat. Ass.*, **76**, 817–23.

Hocking, R. R. and Pendleton, O. J. (1983). The Regression Dilemma. Communications in Statistics—Theory and Methods, Vol. 12, pp. 497–527.

Hocking, R. R. (1983). Developments in linear regression methodology. *Technometrics*, **25**, 219–49.

The complete issue of Communications in Statistics Theory and Method (1984). Vol. 13, No. 2, is devoted to Ridge Regression.

INDEX