APPLIED REGRESSION ANALYSIS AND EXPERIMENTAL DESIGN

RICHARD J. BROOK
GREGORY C. ARNOLD

APPLIED REGRESSION ANALYSIS AND EXPERIMENTAL DESIGN

RICHARD J. BROOK
GREGORY C. ARNOLD
Department of Mathematics and Statistics
Massey University
Palmerston North, New Zealand

MARCEL DEKKER, INC. New York and Basel

Library of Congress Cataloging in Publication Data

Brook, Richard J.
Applied regression analysis and experimental design.

(Statistics, textbooks and monographs ; vol. 62)
Includes index.
1. Regression analysis. 2. Experimental design.
I. Arnold, G. C. (Gregory C.), [date] . II. Title.
III. Series: Statistics, textbooks and monographs ; v. 62.
QA278.2.B76 1985 519.5'36 85-4361
ISBN 0-8247-7252-0

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

Current printing (last digit):
10 9 8 7 6 5 4 3 2 1

PRINTED IN THE UNITED STATES OF AMERICA

PREFACE

This textbook was written to provide a clear and concise discussion of regression and experimental design models. Equal weighting is given to both of these important topics which are applicable, respectively, to observational data and data collected in a controlled manner. The unifying concepts for these topics are those of linear models so that the principles and applications of such models are considered in some detail.

We have assumed that the reader will have had some exposure to the basic ideas of statistical theory and practice as well as some grounding in linear algebra. Consequently, this text will be found useful in undergraduate/graduate courses as well as being of interest to a wider audience, including numerate practitioners.

We felt that it was important to consider variables, which can be written as columns of data, as geometric vectors. Behind the vector notation is always a geometric picture which we believe helps to make the results intuitively plausible without requiring an excess of theory. In this way we have tried to give readers an understanding of the value and purpose of the methods described, so that the book is not about the theory of linear models, but their applications. To this end, we have included an appendix containing seven data sets. These are referred to frequently throughout the book and they form the basis for many of the problems given at the end of each chapter.

We assume that the reader will have computer packages available. We have not considered in any detail the problems of numerical analysis or the methods of computation. Instead we have discussed the strengths, weaknesses and ambiguities of computer output. For the reader, this means that space-consuming descriptions of computations are kept to a minimum.

We have concentrated on the traditional least squares method but we point out its possible weaknesses and indicate why more recent sophisticated techniques are being explored.

We have included such topics as subset selection procedures, randomization, and blocking. It is our hope that students, having been introduced to these ideas in the general context of the linear model, will be well equipped to pick up the details they need for their future work from more specialised texts.

In the first four chapters, we cover the linear model in the regression context. We consider topics of how to fit a line, how to test whether it is a good fit, variable selection, and how to identify and cope with peculiar values. In the remaining four chapters we turn to experimental design, and consider the problem of constructing and estimating meaningful functions of treatment parameters, of utilising structure in the experimental units as blocks, and of fitting the two together to give a useful experiment.

This book represents the final version of course notes which have evolved over several years. We would like to thank our students for their patience as the course notes were corrected and improved. We acknowledge the value of their comments and less tangible reactions. Our data sets and examples, with varying degrees of modification, have many sources, but we particularly thank John Baker, Selwyn Jebson, David Johns, Mike O'Callaghan and Ken Ryba of Massey University, Dr R. M. Gous of the University of Natal, and Julie Anderson of the New Zealand Dairy Research Institute for giving us access to a wide range of data.

Richard J. Brook
Gregory C. Arnold

CONTENTS

1

FITTING A MODEL TO DATA

1.1 INTRODUCTION

The title of this chapter could well be the title of this book. In the first four chapters, we consider problems associated with fitting a regression model and in the last four we consider experimental designs. Mathematically, the two topics use the same model. The term regression is used when the model is fitted to observational data, and experimental design is used when the data is carefully organized to give the model special properties. For some data, the distinction may not be at all clear or, indeed, relevant. We shall consider sets of data consisting of observations of a variable of interest which we shall call y, and we shall assume that these observations are a random sample from a population, usually infinite, of possible values. It is this population which is of primary interest, and not the sample, for in trying to fit models to the data we are really trying to fit models to the population from which the sample is drawn. For each observation, y, the model will be of the form

$$\text{observed } y = \text{population mean} + \text{deviation} \qquad (1.1.1)$$

The population mean may depend on the corresponding values of a predictor variable which we often label as x. For this reason, y is

called the dependent variable. The deviation term indicates the individual peculiarity of the observation, y, which makes it differ from the population mean.

As an example, $y could be the price paid for a house in a certain city. The population mean could be thought of as the mean price paid for houses in that city, presumably in a given time period. In this case the deviation term could be very large as house prices would vary greatly depending on a number of factors such as the size and condition of the house as well as its position in the city. In New Zealand, each house is given a government valuation, GV, which is reconsidered on a five year cycle. The price paid for a house will depend to some extent on its GV. The regression model could then be written in terms of $x, the GV, as:

$$\underset{\text{price}}{y} = \underset{\text{population mean}}{\alpha + \beta x} + \underset{\text{deviation}}{\varepsilon} \qquad (1.1.2)$$

As the population mean is now written as a function of the GV, the deviations will tend to be smaller. Figure 1.1.1 indicates possible values of y when x=20,000 and x=50,000. Theoretically, all values of y may be possible for each value of x but, in practice, the y values would be reasonably close to the value representing the population mean.

The model could easily be extended by adding other predictor variables such as the age of the house or its size. Each deviation

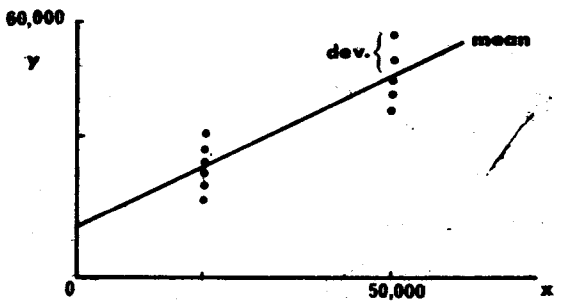

FIGURE 1.1.1 House prices, y, regressed against GV, x.

term would tend to be smaller now as the population mean accounts for the variation in prices due to these additional variables. The deviation term can be thought of as accounting for the variations in prices unexplained by the mean.

Another example, this time from horticulture, would be a model in which y is the yield, in kilograms, of apples per hectare for different orchards. The population mean could be written as a function of the amount of fertilizer added, the amount of insecticide spray used, and the rainfall. In this case, the deviation term would include unexplained physical factors such as varying fertility of the soils as well as possible errors of measurement in weighing the apples.

In each of these examples, a model is postulated and as it relates to the population, of which we know only the small amount of information provided by the sample, then we must use some method of deciding which part of y relates to the population mean and which to the deviation. We shall use the method of least squares to do this.

1.2 HOW TO FIT A LINE

1.2.1 The Method of Least Squares

As the deviation term involves the unexplained variation in y, we try to minimise this in some way. Suppose we postulate that the mean value of y is a function of x. That is

$$E(y) = f(x)$$

Then for a sample of n pairs of y's with their corresponding x's we have

$$\underset{\text{observed } y}{y_i} = \underset{\text{mean of } y}{f(x_i)} + \underset{\text{deviation}}{\varepsilon_i} \qquad 1 \le i \le n \qquad (1.2.1)$$

The above notation assumes that the x's are not random variables but are fixed in advance. If the x's were in fact random variables we should write

$$f(x_i) = E(y_i \mid X_i = x_i)$$

$$= \text{mean of } Y_i \text{ given that } X_i = x_i$$

which gives the same results. We will therefore assume in future that the x's are fixed.

The simplest example of a function f would arise if y was proportional to x. We could imagine a situation where an inspector of weights and measures set out to test the scales used by shopkeepers. In this case, the x's would be the weights of standard measures while y's would be the corresponding weights indicated by the shopkeeper's scales. The model would be

$$\underset{\text{weight shown by scales}}{y_i} = \underset{\text{parameter standard measure}}{\beta x_i} + \underset{\text{deviation}}{\varepsilon_i} \qquad (1.2.2)$$

The mean value of y when $x = x_i$ is given by

$$E(y_i) = \beta x_i = f(x_i) \qquad (1.2.3)$$

This is called a regression curve. In this simple example we would expect the parameter β to be 1, or at least close to 1. We think of the parameters as being fixed numbers which describe some attributes of the population.

The readings of the scales, the y's, will fluctuate, some being above the mean, f(x), in which case the deviation, ε, will be positive while others will be below the mean and the corresponding ε will be negative.

The method of least squares uses the sample of n values of x and y to estimate population parameters by minimizing the deviations ε. More specifically, we seek a value of β which we will label b to minimize the sum of squares of the ε_i, that is

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2 \qquad (1.2.4)$$

If the mean, f(x), has the simple structure of the model (1.2.2)

$$S = \sum_{i=1}^{n} [y_i - \beta x_i]^2 \qquad (1.2.5)$$

Methods of algebra or calculus can be employed to yield

$$\sum_{i=1}^{n} [y_i - b x_i] x_i = 0 \qquad (1.2.6)$$

Rearranging (1.2.6), the least squares estimate of β is the value b which solves the equation

$$\sum_{i=1}^{n} b x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$\text{or} \quad b = \Sigma x_i y_i / \Sigma x_i^2 \qquad (1.2.7)$$

This equation is called the normal equation. For those who appreciate calculus, it could be noted that this equation (1.2.7) can also be written as

$$\sum [y_i - f(x_i)] \frac{\partial f}{\partial \beta} = 0 \qquad (1.2.8)$$

where $\frac{\partial f}{\partial \beta}$ is the partial derivative of $f(x;\beta)$ with respect to β. For this simple model without a constant, we have:

the regression curve is $E(y_i) = f(x_i) = \beta x_i$

and the estimate of it is $\hat{y}_i = \hat{f}(x_i) = bx_i \qquad (1.2.9)$

Equation 1.2.9 is called the prediction curve. Notice that:

(i) $\hat{y}_i$ estimates the mean value of y when $x = x_i$.
(ii) The difference $y_i - \hat{y}_i = e_i$, which is called the residual.

(iii) Parameters are written as Greek letters.
(iv) Estimates of the parameters are written in Roman letters.

Even with the simple problem of calibration of scales it may be sensible to add an intercept term into the model for it may be conceivable that all the scales weigh consistently on the high side by an amount α. The model is then

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{1.2.10}$$

The normal equations become

$$\sum [y_i - f(x_i)] \frac{\partial f}{\partial \alpha} = 0$$

$$\sum [y_i - f(x_i)] \frac{\partial f}{\partial \beta} = 0 \tag{1.2.11}$$

From (1.2.11), or using algebra, and noting that $\Sigma a_i = na$, we obtain

$$a\,n + b \Sigma x_i = \Sigma y_i$$

$$a \Sigma x_i + b \Sigma x_i^2 = \Sigma x_i y_i \tag{1.2.12}$$

Elementary texts give the solution of these normal equations as

$$b = [\Sigma (x_i - \bar{x})(y_i - \bar{y})] / [\Sigma (x_i - \bar{x})^2] \tag{1.2.13}$$

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

Here, $\bar{x}$ and $\bar{y}$ are the sample means.

It is easy to extend (1.2.12) to many variables. For a model with k variables we need to use double subscripts as follows

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$$

where $x_{i0} = 1$ if an intercept term is included. The normal equations are

$$
\begin{array}{lccccc}
 & C0 & C1 & & Ck & Cy \\
R0 & b_0 \sum x_{i0}^2 & + b_1 \sum x_{i0}x_{i1} & +\cdots+ & b_k \sum x_{i0}x_{ik} & = \sum x_{i0}y_i \\
R1 & b_0 \sum x_{i1}x_{i0} & + b_1 \sum x_{i1}^2 & +\cdots+ & b_k \sum x_{i1}x_{ik} & = \sum x_{i1}y_i \\
 & \vdots & \vdots & & \vdots & \vdots \\
 & \vdots & \vdots & & \vdots & \vdots \\
Rk & b_0 \sum x_{ik}x_{i0} & + b1 \sum x_{ik}x_{i1} & +\cdots+ & b_k \sum x_{ik}^2 & = \sum x_{ik}y_i
\end{array}
\qquad (1.2.14)
$$

Notice that R0 (Row 0) involves x_0 in every term and in general Rj involves x_j, which is analagous to (1.2.11) with the derivative taken with respect to β_j. Similarly C0 (Col 0) involves x_0 in every term, and in general Cj involves x_j and Cy involves y in every term.

Example 1.2.1

Consider the simple example of the calibrating of scales where x kg is the "true" weight and y kg the weight indicated by a certain scale. The values of x and y are given in Table 1.2.1. For the model without an intercept term

$$\hat{y} = bx = 0.97\,x \quad \text{from (1.2.7)}$$

If an intercept term is included, the normal equations of (1.2.12) become

$$5.0\,a + 7.5\,b = 7.55$$
$$7.5\,a + 13.75\,b = 13.375$$

TABLE 1.2.1 Scale Calibration Data

y	x
0.70	0.5
1.15	1.0
1.35	1.5
2.05	2.0
2.30	2.5
$\Sigma y = 7.55$	$\Sigma x = 7.5$
$\Sigma xy = 13.375$	$\Sigma x^2 = 13.75$

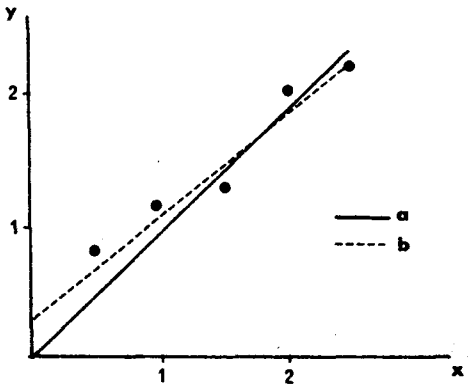

FIGURE 1.2.1 Prediction curves. a: with intercept, b: no intercept.

The solution to these equations is a = 0.28, b = 0.82 giving the prediction curve

$$\hat{y} = 0.28 + 0.82\ x$$

The prediction curves are shown in Figure 1.2.1.

1.2.2 The Assumptions of Least Squares

We have used the method of least squares without considering assumptions on the model. It is usual, however, to make certain assumptions which justify the use of the least squares approach. In particular, the estimates and predicted values we obtain will be optimal in the sense of being unbiased and having the smallest variance among all unbiased linear estimates provided that the following four assumptions hold:

(i) The x values are fixed and not random variables
(ii) The deviations are independent
(iii) The deviations have a mean of zero and
(iv) The variance of the deviations is constant and does not depend on (say) the x values.

If we add a fifth assumption, namely,

(v) The deviations are normally distributed,

then the estimates of the parameters are the same as would be obtained from maximum likelihood, which gives us further theoretical assurances. For the development followed in this book, we are more concerned that this property ensures that estimates of parameters and predicted values of y are also distributed normally leading to F-tests and confidence intervals based on the t-statistics. In fact, means, normality and the method of least squares go hand in hand. It is not very surprising that least squares is an optimal approach if the above assumptions are true.

1.2.3 Other Ways of Fitting a Curve

The main problem with the approach of least squares is that a large deviation will have an even larger square and this deviation may have an unduly large influence on the fitted curve. To guard against such distortions we could try to isolate large deviations. We consider this in more detail in Chapter 4 under outliers and sensitive points. Alternatively, we could seek estimates which minimize a different function of the deviations.

If the model is expressed in terms of the population median of y, rather than its mean, another method of fitting a curve would be by minimizing T, the sum of the absolute values of deviations, that is

$$T = \sum_{i=1}^{n} |\varepsilon_i|$$

Although this is a sensible approach which works well, the actual mathematics is difficult when the distributions of estimates are sought. Hogg (1974) suggests minimizing

$$T = \sum |\varepsilon_i|^p \quad \text{with} \quad 1<p<2$$

and $p = 1.5$, in particular, may be a reasonable compromise. Again it is difficult to determine the exact distributions of the resulting estimates. If we are not so much interested in testing hypothesis as