Regression Analysis by Example

SAMPRIT CHATTERJEE
BERTRAM PRICE

Regression Analysis by Example

SAMPRIT CHATTERJEE
BERTRAM PRICE

New York University
New York, New York

JOHN WILEY & SONS
New York • Chichester • Brisbane • Toronto

Library of Congress Cataloging in Publication Data:

Chatterjee, Samprit, 1938-
Regression analysis by example.

(Wiley series in probability and mathematical statistics)
Includes bibliographies and index.
1. Regression analysis. I. Price, Bertram, 1939- joint author. II. Title.
QA278.2.C5 519.5'36 77-24510

ISBN 0-471-01521-0

Printed in the United States of America

10 9 8 7 6

Preface

Regression analysis has become one of the most widely used statistical tools for analyzing multifactor data. It is appealing because it provides a conceptually simple method for investigating functional relationships among variables. The standard approach in regression analysis is to use a sample of data to compute an estimate of the proposed relationship, and then evaluate the fit using statistics such as t, F, and R^2.

Our approach is much broader. We view regression analysis as a set of data analytic techniques that are used to help understand the interrelationships among a given set of variables. The emphasis is not on formal statistical tests and probability computations. We argue for an informal analysis directed towards uncovering patterns in the data.

We utilize most standard and some not so standard summary statistics on the basis of their intuitive appeal. We are not overly concerned with precise probability evaluations. We rely heavily on graphical representations of the data. In particular, many variations of plots of regression residuals are used. Graphical methods for exploring residuals can suggest model deficiencies or point to "troublesome" observations. Upon further investigation into their origin, the troublesome observations often turn out to be more informative than the well-behaved observations. We feel that more information is obtained from an informal examination of a plot of residuals than from the formal test of statistical significance of some limited null-hypothesis. In short, our presentation in the chapters of this book is guided by the principles and concepts of exploratory data analysis.

The various concepts and techniques of regression analysis are developed with the aid of examples. In each example, we have isolated one or two techniques and discussed them in some detail. The data was chosen to highlight the techniques being presented. Although when analyzing a given set of data it is usually necessary to employ many techniques, we have tried to choose the various data sets so that it would not be necessary to discuss the same technique more than once. Our hope is that after working through the book, the reader will be ready and able to analyze his or her own data methodically, thoroughly, and confidently.

No attempt is made to derive the techniques used. Techniques are described, the required assumptions are given, and finally, the success of the technique in the particular example is assessed. Although derivations of the techniques are not included, we have tried to refer the reader in each case to sources in which such discussion is available. Our hope is that some of these sources will be investigated by the reader who wants a more thorough grounding in theory. The emphasis in this book is not on formulas, tests of hypotheses, or confidence intervals, but on the analysis of data.

We have taken for granted the availability of a computer and a statistical analysis system. We feel that there has been a qualitative change in the analysis of linear models, from model fitting to model building, from overall tests to clinical examinations of data, from macroscopic to the microscopic analysis. To do this kind of analysis a computer is essential and we have assumed its availability. No specific machines or plotters are needed to carry out the different analyses. Almost all of the analyses we use are now available in software packages at most computer centers in universities, business, and government agencies.

The material presented is intended for anyone who is involved in analyzing data. The book should be helpful to anyone who has some knowledge of the basic concepts of statistics. In the university, it could be a text for a course in regression analysis for students whose specialization is not statistics, but nevertheless use regression analysis quite extensively in their work. For students whose major emphasis is statistics, and who take a course on regression analysis from a book at the level of Searle, Plackett, or Rao, this book can be used to balance and complement the theoretical aspects of the subject with practical applications. Outside the university, this book can be profitably used by those people whose present approach to analyzing multifactor data consists of looking at standard computer output (t, F, R^2, standard errors, etc.), but who want to go beyond these summaries for a more thorough analysis.

We have attempted to write a book for a group of readers with diverse backgrounds. We have also tried to put emphasis on the art of data analysis rather than on the development of statistical theory. We are fortunate to have had assistance and encouragement from several friends, colleagues, and associates. We particularly want to mention Professor Martin J. Gardner of the University of Southampton, U. K. who read an early draft and made valuable comments. Some of our colleagues at New York University have used portions of the material in their courses and have shared with us their comments and comments of their students. The students in our classes on regression analysis have all contributed by asking penetrating questions and demanding meaningful and understand-

able answers. Mr. Tak Lo has assisted with some aspects of the computer work. Ms. Roberta Mollot typed the final manuscript with haste and accuracy. To each of these people and to the many others that have provided assitance and encouragement we are grateful and express our thanks.

We are grateful to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., to Dr. Frank Yates, F.R.S., and to Longman Group Ltd., London, for permission to reprint four columns from Table III (p. 46) from their book, *Statistical Tables for Biological, Agricultural, and Medical Research* (6th edition, 1974).

SAMPRIT CHATTERJEE
BERTRAM PRICE

Eagle Island, Maine
White Plains, New York
March 1977

Contents

CHAPTER 1

Simple Linear Regression

1.1. INTRODUCTION

Regression analysis may be broadly defined as the analysis of relationships among variables. It is one of the most widely used statistical tools because it provides a simple method for establishing a functional relationship among variables. The relationship is expressed in the form of an equation connecting the response or dependent variable y, and one or more independent variables, $x_1, x_2, \ldots, x_p$. The equation, or to be more precise, the regression equation takes the form

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p,$$

where $b_0, b_1, b_2, \ldots, b_p$ are called the regression coefficients and are determined from the data. A regression equation containing only one independent variable is called a simple regression equation. An equation containing more than one independent variable is referred to as a multiple regression equation. An example of simple regression would be an analysis in which the time to repair a machine is studied in relation to the number of components to be repaired. Here we have one dependent variable (time to repair the machine) and one independent variable (number of components to be repaired). An example of a very complex multiple regression situation would be an attempt to explain the age-adjusted mortality rates prevailing in different geographic regions (dependent variable) by a large number of environmental and socioeconomic factors (independent variables). Both types of problems are treated in the text. In fact, these two particular examples are included, one in the first chapter, the other in the last chapter.

The explicit determination of the regression equation is in a sense the

final product of the analysis. It is a summary of the relationship between y (the dependent variable) and the set of independent variables, the x's. The equation may be used for several purposes. It may be used to evaluate the importance of individual x's, to analyze the effects of policy that involves changing values of the x's, or to forecast values of y for a given set of x's. *Although the regression equation is the final product, there are many* important by-products. We view regression analysis as a set of data analytic techniques that are used to help understand the interrelationships among variables in a certain environment. It is assumed that data from the environment is available. Sometimes the data will have been collected in a controlled setting so that factors that are not of primary interest can be held constant. Most often the data will have been collected under nonexperimental conditions where very little can be controlled by the investigator. The task of regression analysis is to learn as much as possible about the environment represented by the data. We emphasize that what is uncovered along the way to the formulation of the equation may often be as valuable and informative as the final equation.

We begin our study by considering the simple linear regression model. In this chapter the model is formulated, assumptions are stated, and the standard theoretical results are recorded. There are no formal derivations. Familiarity with standard results is developed through the examples. Formulas are presented, but only for purposes of reference. It is assumed throughout that the necessary summary statistics will be computer generated from an existing regression package.* The reader familiar with the basic concepts of regression analysis may choose to begin with the section labeled, "Analysis of Residuals" and then proceed to the example on p. 10, referring back to the formulas as necessary. Readers interested in mathematical derivations are referred to the bibliographic notes at the end of this chapter where a number of books that contain a formal development of the regression problem are listed.

1.2. DESCRIPTION OF THE DATA AND MODEL

The data consists of n observations on a dependent or response variable and an independent or explanatory variable x_1. The observations are

*Most computer centers and commercial computer time vendors offer one or more regression analysis packages. We assume that these programs have been thoroughly tested and produce numerically accurate answers. For the most part the assumption is a safe one, but for some data sets, different programs have given dramatically different results. See Beaton, Rubin, and Barone (1976), or Longley (1967), for a discussion of this problem.

usually recorded as follows:

Observation number	y	x_1
1	y_1	x_{11}
2	y_2	x_{12}
3	y_3	x_{13}
.	.	.
.	.	.
.	.	.
n	y_n	x_{1n}

The relationship between y and x_1 is formulated as a linear* model

$$y_i = \beta_0 + \beta_1 x_{1i} + u_i, \qquad i = 1, 2, \ldots, n, \tag{1.1}$$

where β_0 and β_1 are constants and are called the model regression parameters, and u_i is a random disturbance. It is assumed that in the range of the observations studied, the linear equation (1.1) provides an acceptable approximation to the true relation between y and x_1. In other words, y is approximately a linear function of x_1, and u measures the discrepancy in that approximation. It is assumed that for every fixed value of x_1, the u's are random quantities independently distributed with mean zero and a common variance denoted by σ^2. The coefficient β_1 may be interpreted as the increment in y corresponding to a unit increase in x_1.

1.3. ESTIMATION AND TESTS OF HYPOTHESES

The parameters β_0 and β_1 are estimated by the method of least squares which involves minimizing the sum of squares of the residuals $S(\beta_0, \beta_1)$, where

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i})^2.$$

The values of β_0 and β_1 that minimize $S(\beta_0, \beta_1)$, b_0 and b_1 are given by

$$b_1 = \frac{\sum (y_i - \bar{y})(x_{1i} - \bar{x}_1)}{\sum (x_{1i} - \bar{x}_1)^2} \tag{1.2}$$

*The adjective linear has a dual role here. It may be taken to describe the fact that the relationship between y and x_1 is linear. More generally, the word linear refers to the fact that the regression parameters enter Equation (1.1) in a linear fashion. As we shall encounter later, $y = \beta_0 + \beta_1 x_1^2 + u$ is also a linear model even though the relationship between y and x_1 is quadratic.

and

$$b_0 = \bar{y} - b_1 \bar{x}_1, \tag{1.3}$$

where

$$\bar{y} = \frac{\sum y_i}{n} \quad \text{and} \quad \bar{x}_1 = \frac{\sum x_{1i}}{n}.$$

Based on the assumptions described previously concerning the u's, it follows that the quantities, b_0 and b_1, are unbiased estimates of β_0 and β_1. Their variances are

$$\text{Var}(b_1) = \frac{\sigma^2}{\sum (x_{1i} - \bar{x}_1)^2}, \tag{1.4}$$

$$\text{Var}(b_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}_1^2}{\sum (x_{1i} - \bar{x}_1)^2} \right]. \tag{1.5}$$

An unbiased estimate of σ^2 is s^2 given as

$$s^2 = \frac{\sum (y_i - b_0 - b_1 x_{1i})^2}{n-2} \tag{1.6}$$

Replacing σ^2 by s^2 in (1.4) and (1.5) we get unbiased estimates of the variances of b_0 and b_1.

Corresponding to the ith observation, the response value predicted by the model is given as

$$\hat{y}_i = b_0 + b_1 x_{1i}. \tag{1.7}$$

The residual corresponding to the ith observation is

$$e_i = y_i - \hat{y}_i \tag{1.8}$$

and the standardized value of the ith residual is defined by

$$e_{is} = \frac{e_i}{s}, \tag{1.9}$$

where s is obtained from (1.6).

In order to construct confidence intervals and to perform tests of hypotheses about the parameters in the regression model, we have to make

an additional assumption about the probability law of the u's. The u's are assumed to have a normal distribution.

With this assumption of normality, the least squares estimate, b_1, of β_1 is normally distributed with mean β_1 and variance as given in (1.4). To test the null hypothesis $H_0(\beta_1 = \beta_1^0)$, where β_1^0 is a constant chosen by the investigator, the appropriate test statistic is

$$t = \frac{(b_1 - \beta_1^0)}{\text{s.e.}(b_1)}, \tag{1.10}$$

where s.e.(b_1) is the standard error of b_1 and is given by

$$\text{s.e.}(b_1) = \frac{s}{\left\{\sum (x_{1i} - \bar{x}_1)^2\right\}^{1/2}}. \tag{1.11}$$

The statistic t in (1.10) is distributed as a Student's t with $(n-2)$ degrees of freedom. The test is carried out by comparing the observed value with the appropriate tabulated critical t value. The usual test is for $\beta_1^0 = 0$ in which case t reduces to the ratio of b_1 to its standard error.

The confidence limits for β_1 with confidence coefficient $(1-\alpha)$ are given by

$$b_1 \pm t\left(n-2, \frac{\alpha}{2}\right)\left[\text{s.e.}(b_1)\right], \tag{1.12}$$

where $t(n-2, \alpha)$ is the $(1-\alpha)$ percentile of a t distribution with $(n-2)$ degrees of freedom. The intercept of the regression line, b_0, is normally distributed with mean β_0 and variance given in (1.5). The statistic for testing $H_0(\beta_0 = \beta_0')$, where β_0' is a value specified by the investigator, is

$$t = \frac{b_0 - \beta_0'}{\text{s.e.}(b_0)}, \tag{1.13}$$

where

$$\text{s.e.}(b_0) = s\left[\frac{1}{n} + \frac{\bar{x}_1^2}{\sum (x_{1i} - \bar{x}_1)^2}\right]^{1/2} \tag{1.14}$$

and is distributed as a Student's t with $(n-2)$ degrees of freedom. The confidence limits for β_0 with confidence coefficient $(1-\alpha)$ are

$$\left\{b_0 \pm t\left(n-2, \frac{\alpha}{2}\right)\left[\text{s.e.}(b_0)\right]\right\}. \tag{1.15}$$

1.4. INDEX OF FIT

After obtaining estimates of β_0, β_1 and σ^2, it is desirable to evaluate the goodness of fit of the model in Equation (1.1) to the observed data. The index most widely used for this purpose is the sample correlation coefficient computed for y and $\hat{y}$, defined* as

$$R = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\left[\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2\right]^{1/2}}, \tag{1.16}$$

where $\bar{\hat{y}}$ is the average of the y's. The numerical value of R lies between 1 and -1. This goodness of fit index may be viewed as a measure of the strength of the linear relationship between y and x_1. The square of the correlation coefficient, R^2 may be written as

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}. \tag{1.17}$$

The definitions of R given in (1.16) and (1.17) are algebraically equivalent. The definition given in (1.17) provides us with an alternative interpretation. The index R^2 may be interpreted as the proportion of total variability in y that is explained by x_1. If R^2 is near 1, then x_1 explains a large part of variation in y. To examine whether x_1 explains a significant amount of variation in y, the null hypothesis tested is $H_0(\rho = 0)$ against an alternative $H_0(\rho \neq 0)$ where ρ is the population correlation coefficient. The appropriate statistic for testing this hypothesis is

$$t = \frac{|R|\sqrt{n-2}}{\sqrt{1-R^2}}, \tag{1.18}$$

where t is a Student's variable with $(n-2)$ degrees of freedom. The test is carried out by comparing the observed t value with a tabulated t value with appropriate degrees of freedom.

It is clear that if no linear relationship exists between y and x_1, then β_1, the population regression coefficient, is zero. Consequently, the statistical tests for $H_0(\beta_1 = 0)$ and $H_0(\rho = 0)$ should be identical. Although the statistics for testing these hypotheses given in (1.10) and (1.18) look different, it can be demonstrated that they are algebraically equivalent.

*The numerical value of R may also be obtained by computing the correlation coefficient between y and x_1.

1.5. PREDICTED VALUES AND STANDARD ERRORS

The fitted regression equation can be used to predict the value of the dependent variable y which corresponds to any chosen value, x_1^0, of the independent variable. The predicted value $\hat{y}_0$ is

$$\hat{y}_0 = b_0 + b_1 x_1^0 \tag{1.19}$$

and has a variance,

$$\text{Var}(\hat{y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_1^0 - \bar{x}_1)^2}{\sum_i (x_{1i} - \bar{x}_1)^2} \right]. \tag{1.20}$$

An estimate of the variance of $\hat{y}_0$ is obtained by replacing σ^2 by s^2 in (1.20). The confidence limits for the predicted value with confidence coefficient $(1-\alpha)$ are $\hat{y}_0 \pm t(n-2, \alpha/2)\text{s.e.}(\hat{y}_0)$, where

$$\text{s.e.}(\hat{y}_0) = s \left[1 + \frac{1}{n} + \frac{(x_1^0 - \bar{x}_1)^2}{\sum (x_{1i} - \bar{x}_1)^2} \right]^{1/2}. \tag{1.21}$$

The predicted response has a normal distribution with mean, $\mu_0 = \beta_0 + \beta_1 x_1^0$. If our interest is in the mean response, then μ_0 is estimated as

$$\hat{\mu}_0 = b_0 + b_1 x_1^0 \tag{1.22}$$

with variance

$$\text{Var}(\hat{\mu}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(x_1^0 - \bar{x}_1)^2}{\sum (x_{1i} - \bar{x}_1)^2} \right]. \tag{1.23}$$

Note that the point estimate of μ_0 is identical to the predicted response, $\hat{y}_0$. The difference in interpretation is reflected in the variances of the respective quantities.

1.6. EVALUATING THE FIT

We have stated the basic results that are used for making inferences in the context of the simple linear regression model. The results are based on summary statistics that are computed from the data. The results are valid and have meaning only insofar as the assumptions concerning the residual