Analyzing Environmental Data

Walter W. Piegorsch
University of South Carolina
Columbia, South Carolina

A. John Bailer
Miami University
Oxford, Ohio

John Wiley & Sons, Ltd

Analyzing Environmental Data

Analyzing Environmental Data

Walter W. Piegorsch
University of South Carolina
Columbia, South Carolina

A. John Bailer
Miami University
Oxford, Ohio

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquires): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons, Inc. 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland, 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada, M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging in Publication Data

Piegorsch, Walter W.
Analyzing environmental data / Walter W. Piegorsch, A. John Bailer.
p. cm.
Includes bibliographical references and indexes.
ISBN 0-470-84836-7 (acid-free paper)
1. Environmental sampling. 2. Regression analysis. 3. Correlation (Statistics) I. Bailer, A. John. II. Title.

GE45.S25P54 2005
363.7′0072′7—dc22

2004057769

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-470-84836-7 (HB)

Typeset in 10/12pt Times by Integra Software Services Pvt. Ltd, Pondicherry, India
Printed and bound in Great Britain by TJ International Ltd, Padstow, Cornwall
This book is printed on acid-free paper responsibly manufactured from sustainable forestry
in which at least two trees are planted for each one used for paper production.

To Karen and Jenny for their patience and
encouragement; to Sara, Jacob, Chris, and
Emily for sharing time with research;
and to our readers with thanks for their interest.

Contents

Preface

Data collected by environmental scientists cover a highly diverse set of application areas, ranging from public health studies of toxic environmental exposures to meteorological investigations of chaotic atmospheric phenomena. As a result, analysis of environmental data has itself become a highly diverse effort. In this text we provide a selection of methods for undertaking such analyses, keying on the motivating environmetric features of the observations. We emphasize primarily regression settings where some form of predictor variable is used to make inferences on an outcome variable of environmental interest. (Where possible, however, we also include allied topics, such as uncertainty/sensitivity analysis in Chapter 4 and environmental sampling in Chapter 8.)

This effort proved challenging: the broader field of *environmetrics* has experienced rapid growth in the past few decades (El-Shaarawi and Hunter, 2002; Guttorp, 2003), and it became clear to us that no single text could possibly survey all the modern, intricate statistical methods available for analyzing environmental data. We do not attempt to do so here. In fact, the environmetric paradigm under study in some chapters often leads to a basic, introductory-level presentation, while in other chapters it forces us to describe rather advanced data-analytic approaches. For the latter cases we try where possible to emphasize the simpler models and methods, and also guide readers to more advanced material via citations to the literature. Indeed, to keep the final product manageable, some advanced environmetric topics have been given little or no mention. These include issues in survival analysis (Kalbfleisch and Prentice, 2002), extreme-value analysis (Coles, 2001), experimental design (Mason *et al.*, 2003), Bayesian methods (Carlin and Louis, 2000), geographic information systems (Longley *et al.*, 2001), and applications such as ordination or other multivariate methods popular in quantitative ecology (McGarigal *et al.*, 2000). Readers interested in these important topic areas may benefit from the many books that discuss them in detail, including those cited above.

For an even larger perspective, we recommend the collection of articles given in Wiley's *Encyclopedia of Environmetrics* (El-Shaarawi and Piegorsch, 2002), a project in which we had the pleasure of participating. The *Encyclopedia* was envisioned and produced to give the sort of broad coverage to this diverse field that a single book cannot; in the text below we often refer readers to more in-depth material from the *Encyclopedia* when the limits of our own scope and intent are reached. Alongside and in addition to these references, we also give sourcebook references to the many fine texts that delve into greater detail on topics allied with our own presentation. (As one reviewer quite perceptively remarked, for essentially every topic we present there exists a recent, single sourcebook devoted to that material, although perhaps not with an environmental motivation attached. Our goal was to bring these various topics together under a single cover and with purposeful environmental focus; however, we also try where appropriate to make the reader aware of these other, dedicated products.) We hope the result will be a coherent collection of topics that we have found fundamental for the analysis of environmental data.

Individuals who will benefit most from our presentation are students and researchers who have a sound grounding in statistical methods; we recommend a minimum of two semesters of graduate study in statistical methodology. Even with this background, however, many portions of the book will require more advanced quantitative skills; typically a familiarity with integral and differential calculus, vector and matrix notation/manipulation, and often also knowledge of a few advanced concepts in statistical theory such as probability models and likelihood analysis. We give brief reminders throughout the text on some of these topics; for readers who require a 'refresher' in the more advanced statistical concepts, however, we recommend a detailed study of the review of probability and statistical inference in Appendix A and the references therein. We have also tried to separate and sequester the calculus/linear algebra-based material to the best extent possible, so that adept instructors who wish to use the text for students without a background in calculus and linear algebra may do so with only marginal additional effort.

An integral component of our presentation is appeal to computer implementation for the more intricate analyses. A wealth of computer packages and programming languages are available for this purpose and we give in selected instances Internet URLs that guide users to potentially useful computer applications. (All URLs listed herein are current as of the time of this writing.) For 'hands-on' use, we highlight the SAS® system (SAS Institute Inc., 2000). SAS's ubiquity and extent make it a natural choice, and we assume a majority of readers will already be familiar with at least basic SAS mechanics or can acquire such skills separately. (Users versed in the S-Plus® computer package will find the text by Millard and Neerchal, 2001, to be of complementary use.) Figures containing sample SAS computer code and output are displayed throughout the text. Although these are not intended to be the most efficient way to program the desired operations, they will help illustrate use of the system and (perhaps more importantly) interpretation of the outputs. Outputs from SAS procedures (versions 6.12 and 8.2) are copyright ©2002–2003, SAS Institute Inc., Cary, NC, USA. All Rights Reserved. Reproduced with permission of SAS Institute Inc., Cary, NC. We also appreciate the kind permission of Chapman & Hall/CRC Press to adapt selected material from our earlier text on *Statistics for Environmental Biology and Toxicology* (Piegorsch and Bailer, 1997).

All examples end with the symbol 🌎. Large data sets used in any examples and exercises in Chapters 5 and 6 have been archived online at the publisher's website, http://www.wiley.com/go/environmental. In the text, these are presented in reduced tabular form to show only a few representative observations. We indicate this wherever it occurs.

By way of acknowledgments, our warmest gratitude goes to our colleague Don Edwards, who reviewed a number of chapters for us and also gave extensive input into the material in Chapters 5 and 6. Extremely helpful suggestions and input came also from Timothy G. Gregoire, Andrew B. Lawson, Mary C. Christman, James Oris, R. Webster West, Philip M. Dixon, Dwayne E. Porter, Oliver Schabenberger, Jay M. Ver Hoef, Rebecca R. Sharitz, John M. Grego, Kerrie P. Nelson, Maureen O. Petkewich, and three anonymous reviewers. We are also indebted to the Wiley editorial group headed by Siân Jones, along with her colleague Helen Ramsey, for their professionalism, support, and encouragement throughout the preparation of the manuscript. Of course, despite the fine efforts of all these individuals, some errors may have slipped into the text, and we recognize these are wholly our own responsibility. We would appreciate hearing from readers who identify any inconsistencies that they may come across. Finally, we hope this book will help our readers gain insights into and develop strategies for analyzing environmental data.

WALTER W. PIEGORSCH AND A. JOHN BAILER
Columbia, SC and Oxford, OH
May 2004

1

Linear regression

Considerable effort in the environmental sciences is directed at predicting an environmental or ecological response from a collection of other variables. That is, an observed *response variable*, Y, is recorded alongside one or more *predictor variables*, and these latter quantities are used to describe the deterministic aspects of Y. If we denote the predictor variables as $x_1, x_2, \ldots, x_p$, it is natural to model the deterministic aspects of the response via some function, say, $g(x_1, x_2, \ldots, x_p; \boldsymbol{\beta})$, where $\boldsymbol{\beta} = [\beta_0\ \beta_1 \ldots \beta_p]^{\mathrm{T}}$ is a column vector of $p+1$ unknown parameters. (A *vector* is an array of numbers arranged as a row or column. The superscript $^{\mathrm{T}}$ indicates transposition of the vector, so that, for example, $[a_1 a_2]^{\mathrm{T}} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$. More generally, a *matrix* is an array of numbers arranged in a square or rectangular fashion; one can view a matrix as a collection of vectors, all of equal length. Background material on matrices and vectors appears in Appendix A. For a more general introduction to the use of matrix algebra in regression, see Neter *et al.*, 1996, Ch. 5.) We use the function g(·) to describe how Y changes as a function of the x_js.

As part of the model, we often include an additive error term to account for any random, or *stochastic*, aspects of the response. Formally, then, an observation Y_i is assumed to take the form

$$Y_i = g(x_{i1}, x_{i2}, \ldots, x_{ip}; \boldsymbol{\beta}) + \varepsilon_i, \tag{1.1}$$

$i = 1, \ldots, n$, where the additive error terms ε_i are assigned some form of probability distribution and the *sample size* n is the number of recorded observations. Unless otherwise specified, we assume the Y_is constitute a random sample of statistically independent observations. If Y represents a continuous measurement, it is common to take $\varepsilon_i \sim$ i.i.d. $N(0, \sigma^2)$, 'i.i.d.' being a shorthand notation for *i*ndependent and *i*dentically *d*istributed (see Appendix A). Coupled with the additivity assumption in (1.1), this is known as a *regression* of Y on the x_js.

Analyzing Environmental Data W. W. Piegorsch and A. J. Bailer
© 2005 John Wiley & Sons, Ltd ISBN: 0-470-84836-7 (HB)

Note also that we require the x_j predictor variables to be fixed values to which no stochastic variability may be ascribed (or, at least, that the analysis be conditioned on the observed pattern of the predictor variables).

We will devote a large portion of this text to environmetric analysis for a variety of regression problems. In this chapter, we give a short review of some elementary regression models, and then move on to a selection of more complex forms. We start with the most basic case: simple linear regression.

1.1 Simple linear regression

The simple linear case involves only one predictor variable ($p = 1$), and sets $g(x_{i1}; \boldsymbol{\beta})$ equal to a linear function of x_{i1}. For simplicity, when $p = 1$ we write x_{i1} as x_i. Equation (1.1) becomes

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

$i = 1, \ldots, n$, and we call $\beta_0 + \beta_1 x_i$ the *linear predictor*. The linear predictor is the deterministic component of the regression model. Since this also models the population mean of Y_i, we often write $\mu(x_i) = \beta_0 + \beta_1 x_i$, and refer to $\mu(x)$ as the *mean response function*.

The simple linear regression model can also be expressed as the matrix equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{Y} = [Y_1 \ldots Y_n]^{\mathrm{T}}, \boldsymbol{\varepsilon} = [\varepsilon_0 \ldots \varepsilon_n]^{\mathrm{T}}$ and $\mathbf{X}$ is a matrix whose columns are the two vectors $\mathbf{J} = [1 \ldots 1]^{\mathrm{T}}$ – i.e., a column vector of ones – and $[x_1 \ldots x_n]^{\mathrm{T}}$.

As a first step in any regression analysis, we recommend that a graphical display of the data pairs (x_i, Y_i) be produced. Plotted, this is called a *scatterplot*; see Fig. 1.1 in Example 1.1, below. The scatterplot is used to visualize the data and begin the process of assessing the model fit: straight-line relationships suggest a simple linear model, while curvilinear relationships suggest a more complex model. We discuss nonlinear regression modeling in Chapter 2.

Under the common assumptions that $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{Var}[\varepsilon_i] = \sigma^2$ for all $i = 1, \ldots, n$, the model parameters in $\boldsymbol{\beta} = [\beta_0 \; \beta_1]^{\mathrm{T}}$ have interpretations as the Y-intercept (β_0) and slope (β_1) of $\mu(x_i)$. In particular, for any unit increase in x_i, $\mu(x_i)$ increases by β_1 units. To estimate the unknown parameters we appeal to the least squares (LS) method, where the sum of squared errors $\sum_{i=1}^n \{Y_i - \mu(x_i)\}^2$ is minimized (LS estimation is reviewed in §A.4.1). The LS estimators of β_0 and β_1 here are

$$b_0 = \overline{Y} - b_1 \overline{x}$$

and

$$b_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n x_i Y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2}, \qquad (1.2)$$

where $\overline{Y} = \sum_{i=1}^{n} Y_i/n$ and $\overline{x} = \sum_{i=1}^{n} x_i/n$. The algebra here can be simplified using matrix notation: if $\mathbf{b} = [b_0\ b_1]^{\mathrm{T}}$ is the vector of LS estimators, then $\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ being the *inverse* of the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ (see §A.4.3).

If we further assume that $\varepsilon_i \sim$ i.i.d. $N(0, \sigma^2)$, then the LS estimates will correspond to maximum likelihood (ML) estimates for β_0 and β_1. (ML estimation is reviewed in §A.4.3.) The LS/ML estimate of the mean response, $\mu(x) = \beta_0 + \beta_1 x$, for any x, is simply $\hat{\mu}(x) = b_0 + b_1 x$.

We should warn that calculation of b_1 can be adversely affected by a number of factors. For example, if the x_is are spaced unevenly, highly separated values of x_i can exert strong *leverage* on b_1 by pulling the estimated regression line too far up or down. (See the web applet at http://www.stat.sc.edu/~west/javahtml/Regression.html for a visual demonstration. Also see the discussion on regression diagnostics, below.) To avoid this, the predictor variables should be spaced as evenly as possible, or some transformation of the x_is should be applied before performing the regression calculations. The natural logarithm is a typical choice here, since it tends to compress very disparate values. If when applying the logarithm, one of the x_i values is zero, say $x_1 = 0$, one can average the other log-transformed x_is to approximate an equally spaced value associated with $x_1 = 0$. This is *consecutive-dose average spacing* (Margolin *et al.*, 1986): denote the transformed predictor by $u_i = \log(x_i)$, $i = 2, \ldots, n$. Then at $x_1 = 0$, use

$$u_1 = u_2 - \frac{u_n - u_2}{n - 1}. \tag{1.3}$$

A useful tabular device for collecting important statistical information from a linear regression analysis is known as the *analysis of variance (ANOVA) table*. The table lays out *sums of squares* that measure variation in the data attributable to various components of the model. It also gives the *degrees of freedom* (df) for each component. The df represent the amount of information in the data available to estimate that particular source of variation. The ratio of a sum of squares to its corresponding df is called a *mean square*.

For example, to identify the amount of variability explained by the linear regression of Y on x, the sum of squares for regression is $\mathrm{SSR} = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$, where $\hat{Y}_i = b_0 + b_1 x_i$ is the ith *predicted value* (also called a *fitted value*). SSR has degrees of freedom equal to the number of regression parameters estimated minus one; here, $\mathrm{df_r} = 1$. Thus the mean square for regression when $p = 1$ is $\mathrm{MSR} = \mathrm{SSR}/1$.

We can also estimate the unknown variance parameter, σ^2, via ANOVA computations. Find the sum of squared errors $\mathrm{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ and divide this by the error df (the number of observations minus the number of regression parameters estimated), $\mathrm{df_e} = n - 2$. The resulting *mean squared error* is

$$\mathrm{MSE} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n - 2},$$

and this is an unbiased estimator of σ^2. We often call $\sqrt{\mathrm{MSE}}$ the *root mean squared error*.

We do not go into further detail here on the construction of sums of squares and ANOVA tables, although we will mention other aspects of linear modeling and ANOVA below. Readers unfamiliar with ANOVA computations can find useful expositions in texts on linear regression analysis, such as Neter *et al.* (1996) or Christensen (1996).

We use the MSE to calculate the *standard errors* of the LS/ML estimators. (A standard error is the square root or estimated square root of an estimator's variance; see §A.4.3.) Here, these are

$$se[b_0] = \sqrt{\mathrm{MSE}\left\{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right\}}$$

and

$$se[b_1] = \sqrt{\frac{\mathrm{MSE}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}. \tag{1.4}$$

Standard errors (and variances) quantify the variability of the point estimator, helping to gauge how meaningful the magnitude of a given estimate is. They also give insight into the impact of different experimental designs on estimating regression coefficients. For example, notice that $se[b_0]$ is smallest for x_is chosen so that $\overline{x} = 0$, while $se[b_1]$ is minimized when $\sum_{i=1}^{n}(x_i - \overline{x})^2$ is taken to be as large as possible.

Similarly, the standard error of $\hat{\mu}(x)$ is

$$se[\hat{\mu}(x)] = \sqrt{\mathrm{MSE}\left\{\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right\}}.$$

Notice that, as with $\hat{\mu}(x), se[\hat{\mu}(x)]$ varies with x. It attains its minimum at $x = \overline{x}$ and then increases as x departs from $\overline{x}$ in either direction. One may say, therefore, that precision in $\hat{\mu}(x)$ is greatest near the center of the predictor range – i.e., at $\overline{x}$ – and diminishes as x moves away from it. Indeed, if one drives x too far away from the predictor range, $se[\hat{\mu}(x)]$ can grow so large as to make $\hat{\mu}(x)$ essentially useless. This illustrates the oft-cited concern that *extrapolation* away from the range of the data leads to imprecise, inaccurate, and in some cases even senseless statistical predictions.

The standard errors are used in constructing statistical inferences on the β_js or on $\mu(x)$. For example, notice that if $\beta_1 = 0$ then the predictor variable has no effect on the response and the simple linear model collapses to $Y_i = \beta_0 + \varepsilon_i$, a 'constant + error' model for Y. To assess this, assume that the $N(0, \sigma^2)$ assumption on the ε_is is valid. Then, a $1 - \alpha$ *confidence interval* for β_1 is

$$b_1 \pm t_{\alpha/2}(n-2)se[b_1].$$

(The theory of confidence intervals is reviewed in §A.5.1.) An alternative inference is available by conducting a *hypothesis test* of the null hypothesis $H_0\colon \beta_1 = 0$ vs. the alternative hypothesis $H_a: \beta_1 \neq 0$. (The theory of hypothesis tests is reviewed in §A.5.3.) Here, we find the test statistic

$$|t_{\text{calc}}| = \frac{|b_1|}{se[b_1]}$$

based on Student's t-distribution (§A.2.11), and reject H_0 when $|t_{\text{calc}}| \geq t_{\alpha/2}(n-2)$. (We use the subscript 'calc' to indicate a statistic that is wholly calculable from the data.) Equivalently, we can reject H_0 when the corresponding *P-value*, here

$$P = 2\mathrm{P}\left[t(n-2) \geq \frac{|b_1|}{se(b_1)}\right],$$

drops below the preset *significance level* α (see §A.5.3).

For testing against a one-sided alternative such as $H_a: \beta_1 > 0$, we reject H_0 when $t_{\text{calc}} = b_1/se[b_1] \geq t_\alpha(n-2)$. The P-value is then $\mathrm{P}[t(n-2) \geq b_1/se(b_1)]$. Similar constructions are available for β_0; for example, a $1-\alpha$ confidence interval is $b_0 \pm t_{\alpha/2}(n-2)se[b_0]$.

All these operations can be conducted by computer, and indeed, many statistical computing packages perform simple linear regression. Herein, we highlight the SAS® system (SAS Institute Inc., 2000), which provides LS/ML estimates $\mathbf{b} = [b_0\ b_1]^{\mathrm{T}}$, their standard errors $se[b_j]$, an ANOVA table that includes an unbiased estimator of σ^2 via the MSE, and other summary statistics, via its PROC GLM or PROC REG procedures.

Example 1.1 (Motor vehicle CO$_2$) To illustrate use of the simple linear regression model, consider the following example. In the United Kingdom (and in most other industrialized nations) it has been noted that as motor vehicle use increases, so do emissions of various byproducts of hydrocarbon combustion. Public awareness of this potential polluting effect has bolstered industry aspirations to 'uncouple' detrimental emissions from vehicle use. In many cases, emission controls and other efforts have reduced the levels of hazardous pollutants such as small particulate matter (PM) and nitrogen oxides. One crucial counter-example to this trend, however, is the ongoing increases in the greenhouse gas carbon dioxide (CO_2). For example, Redfern *et al.* (2003) discuss data on $x =$ UK motor vehicle use (in kilometers per year) vs. $Y = CO_2$ emissions (as a relative index; $1970 = 100$). Table 1.1 presents the data.

A plot of the data in Table 1.1 shows a clear, increasing, linear trend (Fig. 1.1). Assuming that the simple linear model with normal errors is appropriate for these data, we find the LS/ML estimates to be $b_0 = 28.3603$ and $b_1 = 0.7442$. The corresponding standard errors are $se[b_0] = 2.1349$ and $se[b_1] = 0.0127$. Since $n = 28$, a 95% confidence interval for β_1 is $0.7742 \pm t_{0.025}(26) \times 0.0127 = 0.7742 \pm 2.056 \times 0.0127 = 0.7742 \pm 0.0261$. (We find $t_{0.025}(26)$ from Table B.2 or via the SAS function `tinv`; see Fig. A.4.) Based on this 95% interval, the CO_2 index increases approximately 0.75 to 0.80 units (relative to 1970 levels) with each additional kilometer.

Table 1.1 Yearly CO_2 emissions (rel. index; 1970 = 100) vs. motor vehicle use (rel. km/yr; 1970 = 100) in the United Kingdom, 1971–1998

Year	1971	1972	1973	1974	1975	1976	1977
x = vehicle use	105.742	110.995	116.742	114.592	115.605	121.467	123.123
$Y = CO_2$	104.619	109.785	117.197	114.404	111.994	116.898	119.915
Year	1978	1979	1980	1981	1982	1983	1984
x = vehicle use	127.953	127.648	135.660	138.139	141.911	143.707	151.205
$Y = CO_2$	126.070	128.759	130.196	126.409	130.136	134.212	140.721
Year	1985	1986	1987	1988	1989	1990	1991
x = vehicle use	154.487	162.285	174.837	187.403	202.985	204.959	205.325
$Y = CO_2$	143.462	153.074	159.999	170.312	177.810	182.686	181.348
Year	1992	1993	1994	1995	1996	1997	1998
x = vehicle use	205.598	205.641	210.826	214.947	220.753	225.742	229.027
$Y = CO_2$	183.757	185.869	186.872	185.100	192.249	194.667	193.438

Source: Redfern *et al.* (2003).

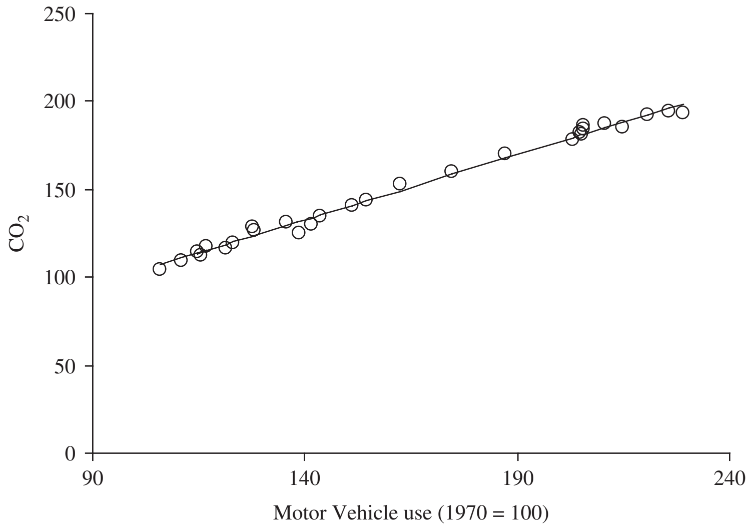

Figure 1.1 Scatterplot and estimated LS line for motor vehicle CO_2 data from Table 1.1

Alternatively, we can test the significance of the slope with these data. Specifically, since one would expect *a priori* that increased motor vehicle use would increase CO_2 emissions, the hypotheses $H_0\colon \beta_1 = 0$ vs. $H_a\colon \beta_1 > 0$ are a natural choice. Suppose we set our significance level to $\alpha = 0.01$. For these data, the test statistic is $t_{\text{calc}} = b_1/se[b_1] = 0.7742/0.0127 = 60.96$, with corresponding P-value $P[t(26) \geq 60.96] < 0.0001$. This is well below α, hence we conclude that a significant, increasing effect exists on CO_2 emissions associated with the observed pattern of motor vehicle use in the UK between 1971 and 1998.

The sample size in Example 1.1, $n = 28$, is not atypical for a simple linear regression data set, but of course analysts can encounter much larger sample sizes in environmental practice. We will study selected examples of this in the chapters on nonlinear regression (Chapter 2), temporal data (Chapter 5), and spatially correlated data (Chapter 6), below.

Once a model has been fitted to data, it is important to assess the quality of the fit in order to gauge the validity of the consequent inferences and predictions. In practice, any statistical analysis of environmental data should include a critical examination of the assumptions made about the statistical model, in order to identify if any unsupported assumptions are being made and to alert the user to possible unanticipated or undesired consequences. At the simplest level, a numerical summary for the quality of a regression fit is the *coefficient of determination* $\{\sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y})\}^2/\{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2\}$, denoted as R^2. This may also be computed from the ANOVA table as $R^2 = \mathrm{SSR}/\{\mathrm{SSR} + \mathrm{SSE}\}$. Under a linear model, R^2 has interpretation as the proportion of variation in Y_i that can be attributed to the variation in x_i. If the predictor variable explains Y precisely (i.e., the x_i, Y_i pairs all coincide on a straight line), R^2 attains its maximum value of 1.0. Alternatively, if there is *no* linear relationship between x_i and Y_i (so $\beta_1 = 0$), $R^2 = 0.0$. As such, higher values of R^2 indicate higher-quality explanatory value in x_i.

More intricate *regression diagnostics* can include a broad variety of procedures for assessing model fit (Davison and Tsai, 1992; Neter *et al.*, 1996, Ch. 3). Most basic among these is study of the *residuals* $r_i = Y_i - \hat{Y}_i$. Almost every analysis of a regression relationship should include a graph of the residuals, r_i, against the predicted values, $\hat{Y}_i$ (or, if $p = 1$, against x_i). Such a *residual plot* can provide information on a number of features. For instance, if there is an underlying curvilinear trend in the data that was not picked up by the original scatterplot, the residual plot may highlight the curvilinear aspects not explained by the simple linear terms. Or, if the assumption of variance homogeneity is inappropriate – i.e., if $\mathrm{Var}[\varepsilon_i]$ is not constant over changing x_i – the residual plot may show a fan-shaped pattern of increasing or decreasing residuals (or both) as $\hat{Y}_i$ increases. Figure 1.2 illustrates both these sorts of patterns. Notice in Fig. 1.2(b) that variability increases with increasing mean response; this sort of pattern is not uncommon with environmental data.

If the residual plot shows a generally uniform or random pattern, then evidence exists for a reasonable model fit.

Example 1.2 (Motor vehicle CO_2, cont'd) Returning to the data on motor vehicle use in the UK, we find $\mathrm{SSR} = 26\,045.2953$ and $\mathrm{SSE} = 196.0457$. This gives $R^2 = 0.9925$, from which it appears that variation in CO_2 emissions is strongly explained by variation in motor vehicle use.

Figure 1.3 shows the residual plot from the simple linear model fit. The residual points appear randomly dispersed, with no obvious structure or pattern. This suggests that the variability in CO_2 levels about the regression line is constant and so the homogeneous variance assumption is supported. One could also graph a histogram or normal probability plot of the residuals to assess the adequacy of the normality assumption. If the histogram appears roughly bell-shaped, or if the normal plot produces a roughly straight line, then the assumption of normal errors may be reasonable. For the residuals in Fig. 1.3, a normal probability plot constructed using PROC UNIVARIATE in SAS (via its `plot` option; output suppressed) does plot as

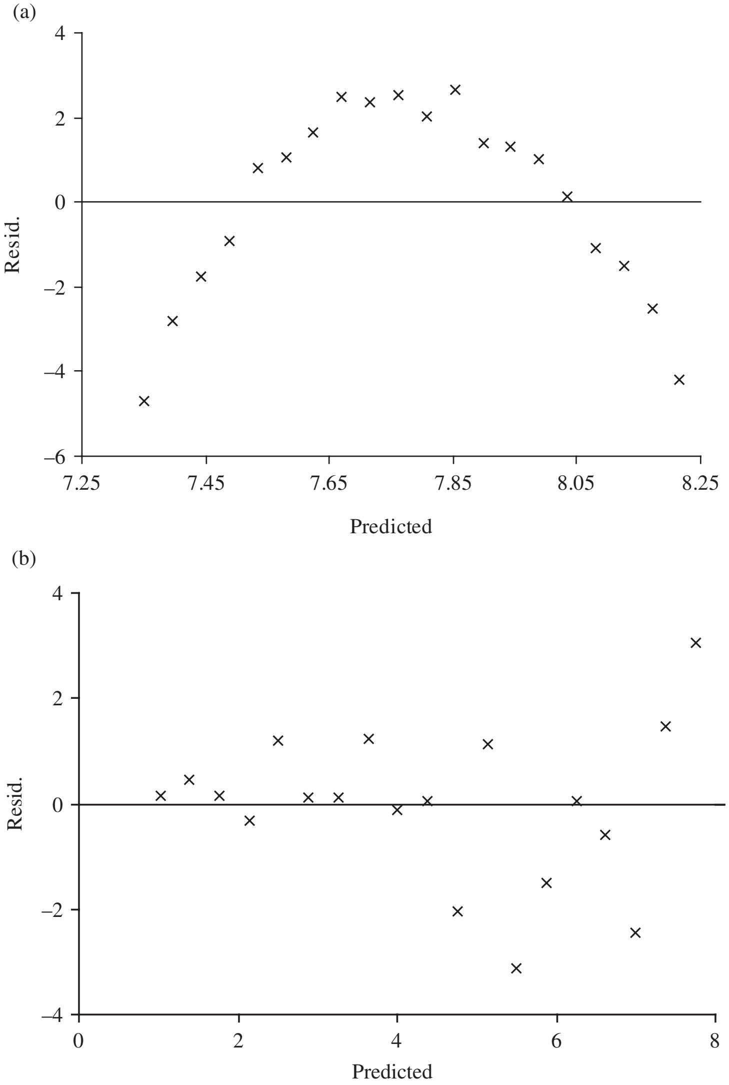

Figure 1.2 Typical residual plots in the presence of model misspecification. (a) Curvilinear residual trend indicates curvilinearity not fit by the model. (b) Widening residual spread indicates possible variance heterogeneity. Horizontal reference lines indicate residual = 0

roughly linear. Or one can call for normal probability plots directly in PROC REG, using the statement

```
plot nqq.*r. npp.*r.;
```

The plot statement in PROC REG can also be used to generate a residual plot, via

```
plot r.*p.;
```

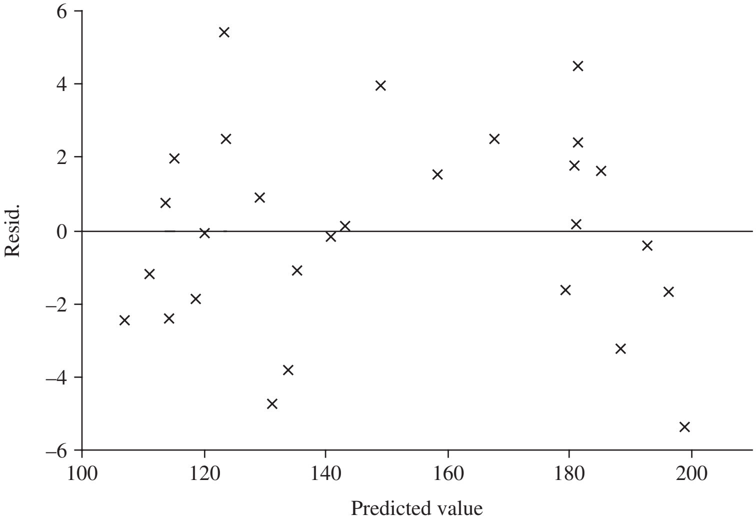

Figure 1.3 Residual plot for motor vehicle CO_2 data from Table 1.1. Horizontal bar indicates residual = 0

or an overlay of the data and the predicted regression line, via

```
plot Y*x p.*x/overlay;
```

When the residual plot identifies a departure from variance homogeneity, inferences on the unknown parameters based on the simple linear fit can be incorrect, and some adjustment is required. If the heterogeneous variation can be modeled or otherwise quantified, it is common to weight each observation in inverse proportion to its variance and apply weighted least squares (WLS; see §A.4.1). For example, suppose it is known or anticipated that the variance changes as a function of x_i, say $\mathrm{Var}[Y_i] \propto h(x_i)$. Then, a common weighting scheme employs $w_i = 1/h(x_i)$.

For weights given as w_i, $i = 1, \ldots, n$, the WLS estimators become

$$\tilde{b}_0 = \left(\sum_{i=1}^{n} w_i\right)^{-1}\left(\sum_{i=1}^{n} w_i Y_i - b_1 \sum_{i=1}^{n} w_i x_i\right) \qquad (1.5)$$

and

$$\tilde{b}_1 = \frac{\sum_{i=1}^{n} w_i x_i Y_i - \left(\sum_{i=1}^{n} w_i\right)^{-1}\left(\sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i Y_i\right)}{\sum_{i=1}^{n} w_i x_i^2 - \left(\sum_{i=1}^{n} w_i\right)^{-1}\left(\sum_{i=1}^{n} w_i x_i\right)^2}. \qquad (1.6)$$

The standard errors require similar modification; for example,

$$se[\tilde{b}_1] = \frac{\sqrt{\tilde{\mathrm{MSE}}}}{\sqrt{\sum_{i=1}^n w_i x_i^2 - \left(\sum_{i=1}^n w_i\right)^{-1}\left(\sum_{i=1}^n w_i x_i\right)^2}},$$

where $\tilde{\mathrm{MSE}}$ is the weighted mean square $\sum_{i=1}^n w_i(Y_i - \tilde{b}_0 - \tilde{b}_1 x_i)^2/(n-2)$. Inferences on β_1 then mimic those described above for the simple linear case. In SAS, both PROC GLM and PROC REG can incorporate these (or any other) weighting schemes, using the `weight` statement. Neter *et al.* (1996, §10.1) give further details on the use of WLS methods.

If appropriate weights cannot be identified, it is often possible to stabilize the variances by transforming the original observations. A common transformation in many environmental applications is the (natural) logarithm: $V_i = \log(Y_i)$. This is part of a larger class of transformations, known as the Box–Cox power transformations (Box and Cox, 1964). The general form is $V_i = (Y_i^\lambda - 1)/\lambda$, for some specified transformation parameter λ. The natural logarithm is the limiting case at $\lambda = 0$. Other popular transformations include the square root ($\lambda = 1/2$), the quadratic ($\lambda = 2$), and the reciprocal ($\lambda = -1$). One can also estimate λ from the data, although this can lead to loss of independence among the V_is. Users should proceed with caution when estimating a power transformation parameter; see Carroll and Ruppert (1988) for more on this and other issues regarding data transformation in regression. Another useful transformation, often employed with percentage data, is the *logit transform*: if Y_i is a percentage between 0 and 100, take $V_i = \log\{Y_i/(100 - Y_i)\}$. We employ this in Example 1.5, below.

Many other procedures are available for diagnosing and assessing model fit, correcting for various model perturbations and inadequacies, and analyzing linear relationships. A full description of all these methods for the simple linear model is beyond the scope of this chapter, however. Details can be found in the targeted textbook by Belsley *et al.* (1980), or in general texts on statistics such as Samuels and Witmer (2003, Ch. 12) and Neter *et al.* (1996, Chs. 1–5).

1.2 Multiple linear regression

The simplest statistical model for the case of $p > 1$ predictor variables in (1.1) employs a linear term for each predictor: set $g(x_{i1}, x_{i2}, \ldots, x_{ip}; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$. This is a *multiple linear regression* model. The parameter β_j may be interpreted as the change in $\mathrm{E}[Y_i]$ that occurs for a unit increase in x_{ij} – the 'slope' of the *j*th predictor – assuming all the other *x*-variables are held fixed. (When it is not possible to vary one predictor while holding all others constant, then of course this interpretation may not make sense. An example of such occurs with polynomial regression models; see §1.5.) We require $n > p + 1$.

Assuming, as above, that the errors satisfy $\mathrm{E}[\varepsilon_i] = 0$ and $\mathrm{Var}[\varepsilon_i] = \sigma^2$ for all $i = 1, \ldots, n$, the LS estimators for $\boldsymbol{\beta} = [\beta_0\ \beta_1 \ldots \beta_p]^{\mathrm{T}}$ can be derived using multivariable calculus. When the additional assumption is made that $\varepsilon_i \sim$ i.i.d. $N(0, \sigma^2)$, these LS estimates will correspond to ML estimates.

Unfortunately, the LS/ML estimators for $\boldsymbol{\beta}$ are not easily written in closed form. The effort can be accomplished using vector and matrix notation in similar fashion to that mentioned in §1.1, although actual calculation of the estimates is most efficiently performed by computer. Almost any statistical computing package can fit a multiple linear regression via LS or WLS methods; in Example 1.3, below, we illustrate use of SAS.

Similar to the simple linear case, we can test whether any particular predictor variable, x_{ij}, is important in modeling $\mathrm{E}[Y_i]$ via appeal to a t-test: find $t_{\mathrm{calc}} = b_j/se[b_j]$ and reject H_0: $\beta_j = 0$ in favor of H_a: $\beta_j \neq 0$ when $|t_{\mathrm{calc}}| = |b_j|/se[b_j] \geq t_{\alpha/2}(n-p-1)$. Note that this tests the significance of the jth predictor variable given that all the other predictor variables are present in the model. In this sense, we call it an *adjusted test* or a *partial test* of significance. Confidence intervals are similar; for example, a pointwise $1-\alpha$ confidence interval for β_j is $b_j \pm t_{\alpha/2}(n-p-1)se[b_j]$; $j = 1, \ldots, p$. Notice the change in $\mathrm{df_e}$ from the simple linear case where $p = 1$: estimation of each additional β_j results in a loss of 1 additional df for error, so we have gone from $\mathrm{df_e} = n-2$ to $\mathrm{df_e} = n-(p+1)$.

We can also make statements on subsets or groupings of the β-parameters. For example, consider a test of the null hypothesis that a group of $k > 1$ of the β_js is equal to zero, say, H_0: $\beta_{j+1} = \cdots = \beta_{j+k} = 0$. Rejection of H_0 suggests that the corresponding group of k predictor variables has a significant impact on the regression relationship. A general approach for such a test involves construction of *discrepancy measures* that quantify the fit of the general (or *full*) model with all $p+1$ of the β-parameters, and the *reduced model* with $p-k+1$ (non-zero) β-parameters. For the multiple regression model with normally distributed errors, a useful discrepancy measure is the sum of squared errors $\mathrm{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i = b_0 + b_1x_{i1} + \cdots + b_px_{ip}$ is the ith predicted value under the full model. For clarity, we augment the SSE notation by indicating if it is calculated under the full model (FM) or under the reduced model (RM): SSE(FM) or SSE(RM). The SSEs are used to quantify the relative quality of each model's fit to the data: if H_0 is false, we expect SSE(RM) to be larger than SSE(FM), since the model under which it is fitted fails to include important predictor variables. Corresponding to these terms, we also write the degrees of freedom associated with each error terms as $\mathrm{df_e(FM)}$ and $\mathrm{df_e(RM)}$, respectively. The difference between the two is $\Delta_\mathrm{e} = \mathrm{df_e(RM)} - \mathrm{df_e(FM)}$. Here, $\mathrm{df_e(FM)} = n-p-1$, while $\mathrm{df_e(RM)} = n+k-p-1$, so that $\Delta_\mathrm{e} = k$ is the number of parameters constrained by the null hypothesis.

To use this discrepancy approach for testing H_0, calculate the test statistic

$$F_{\mathrm{calc}} = \frac{\{\mathrm{SSE(RM)} - \mathrm{SSE(FM)}\}/\Delta_\mathrm{e}}{\mathrm{SSE(FM)}/\mathrm{df_e(FM)}}, \tag{1.7}$$

which under H_0 is distributed as per an F-distribution with Δ_e and $\mathrm{df_e(FM)}$ degrees of freedom (§A.2.11). We denote this as $F_{\mathrm{calc}} \sim F[\Delta_\mathrm{e}, \mathrm{df_e(FM)}]$. Reject H_0 in favor of an alternative that allows at least one of the β_js in H_0 to be non-zero when F_{calc} exceeds the appropriate upper-α F-critical point, $F_\alpha(\Delta_\mathrm{e}, \mathrm{df_e[FM]})$. For the multiple regression setting, this is $F_{\mathrm{calc}} \geq F_\alpha(k, n-p-1)$. The P-value is $P = \mathrm{P}[\mathrm{F}(k, n-p-1) \geq F_{\mathrm{calc}}]$. This testing strategy corresponds to a form of generalized likelihood ratio test (§A.5).

In many cases, the various measures in (1.7) can be read directly from an ANOVA table for the full model (Neter *et al.*, 1996, Ch. 16). For example, if SSR(FM) is the full model's sum of squares for regression and the reduced model contains only the intercept β_0 (so $k = p$), $F_{\text{calc}} = \{\text{SSR(FM)}/p\}/\text{MSE}$. Also, an extension of the coefficient of determination from the simple linear setting is the *coefficient of multiple determination*: $R^2 = \text{SSR(FM)}/\{\text{SSR(FM)} + \text{SSE(FM)}\}$. As in the simple linear case, R^2 measures the proportion of variation in Y_i that can be accounted for by variation in the collection of x_{ij}s.

For this approach to be valid, the parameters represented under the RM must be a true subset of those under the FM. We say then that the models are *nested*. If the relationship between the RM and FM does not satisfy a nested hierarchy, F_{calc} under H_0 may not follow (or even approximate) an F-distribution. The family of models are then said to be *separate* (Cox, 1961, 1962); inferences for testing separate families are still an area of developing environmetric research (Hinde, 1992; Schork, 1993).

Example 1.3 (Soil pH) Edenharder *et al.* (2000) report on soil acidity in west-central Germany, as a function of various soil composition measures. For Y = soil pH, three predictor variables (all percentages) were employed: x_{i1} = soil texture (as clay), x_{i2} = organic matter, and x_{i3} = carbonate composition ($CaCO_3$ by weight). The $n = 17$ data points are given in Table 1.2.

Table 1.2 Soil pH vs. soil composition variables in west-central Germany

x_1 = % Clay	x_2 = % Organics	x_3 = Carbonate	Y = pH
51.1	4.3	6.1	7.1
22.0	2.6	0.0	5.4
17.0	3.0	2.0	7.0
16.8	3.0	0.0	6.1
5.5	4.0	0.0	3.7
21.2	3.3	0.1	7.0
14.1	3.7	16.8	7.4
16.6	0.7	17.3	7.4
35.9	3.7	15.6	7.3
29.9	3.3	11.9	7.5
2.4	3.1	2.8	7.4
1.6	2.8	6.2	7.4
17.0	1.8	0.3	7.5
32.6	2.3	9.1	7.3
10.5	4.0	0.0	4.0
33.0	5.1	26.0	7.1
26.0	1.9	0.0	5.6

Source: Edenharder *et al.* (2000).