CHAPTER 1. INTRODUCTION

A *regression model* describes how the distribution of a *response* (or *dependent*) *variable*—or some characteristic of that distribution, typically its mean—changes with the values of one or more *explanatory* (or *independent*) *variables*. *Regression diagnostics* are methods for determining whether a regression model that has been fit to data adequately represents the structure of the data. For example, if the model assumes a linear (straight-line) relationship between the response and an explanatory variable, is the assumption of linearity warranted? Regression diagnostics not only reveal deficiencies in a regression model that has been fit to data but in many instances may suggest how the model can be improved.

This monograph considers two important classes of regression models:

- The *normal linear regression model*, in which the response variable is quantitative and assumed to have a normal (or *Gaussian*) distribution conditional on the values of the explanatory variables. The observations on the response are further assumed to be independent of one another, to be a linear function (i.e., a weighted sum) of the parameters of the model, and to have constant conditional variance. The normal linear model fit by the method of least squares is the focus of the monograph both because it is often used in practice and because it provides a basis for diagnostics for the other regression models considered here.

- *Generalized linear models* (*GLMs*), in which the conditional distribution of the response variable is a member of an *exponential family*, such as the families of Gaussian, binomial, and Poisson distributions, and in which the mean of the response is transformed to a linear function of the parameters of the model. The GLMs include the normal linear model, *logistic regression* for a dichotomous response, and *Poisson regression* for count data as important special cases. GLMs can also be extended to nonexponential distributions and to situations in which an explicit conditional response distribution isn't assumed.

As a preliminary example of what can go wrong in linear least-squares regression, consider the four scatterplots from Anscombe (1973) shown in Figure 1.1 and dubbed "Anscombe's quartet" by Edward Tufte in an influential treatise on statistical graphics (Tufte, 1983). One of the goals of statistical analysis is to provide an adequate descriptive summary of the data. All four of Anscombe's data sets were contrived cleverly to produce the same standard linear regression outputs—slope, intercept, correlation, residual standard deviation, coefficient standard errors, and statistical tests—but, importantly, not the same residuals.

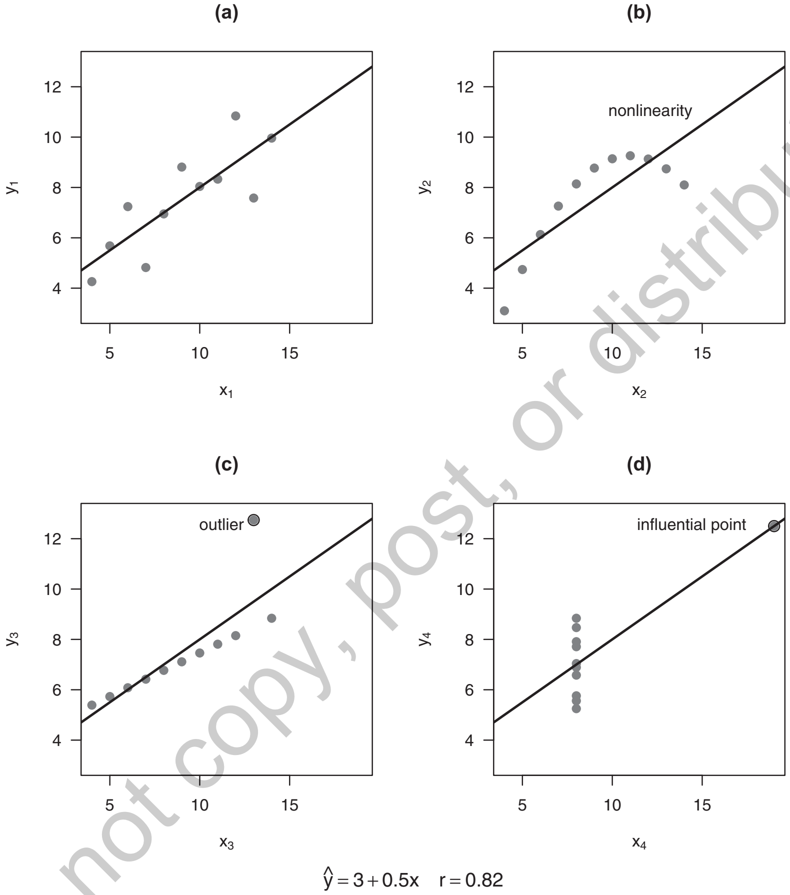
- In Figure 1.1(a), the least-squares line is a reasonable description of the tendency for y to increase with x.
- In Figure 1.1(b), the linear regression fails to capture the obviously curvilinear pattern of the data—the linear model is clearly wrong.
- In Figure 1.1(c), one data point (an *outlier*) is out of line with the others and has an undue *influence* on the fitted least-squares line. A line through the other points fits them perfectly. Ideally in this case, we want to understand why the outlying case differs from the others—possibly it is special in some way (e.g., it is strongly affected by a variable other than x, or represents an error in recording the data). Of course, we are exercising our imaginations here, because Anscombe's data are simply made up, but the essential point is that we should address anomalous data substantively.
- In Figure 1.1(d), in contrast, we are unable to fit a line at all but for the rightmost data point; the least-squares line goes through this influential point and through the mean of the remaining values of y above the common x-value of 8. At the very least, we should be reluctant to trust the estimated regression coefficients because of their dependence on one unusual point.

Anscombe's simple illustrations serve to introduce several of the themes of this monograph, including nonlinearity, outlying data, influential data, and the effectiveness of graphical displays. The usual numeric regression outputs clearly do not tell the whole story. Diagnostic methods—many of them graphical—help to fill in the gaps.

The plan of the monograph is as follows:

- Chapter 2 reviews the normal linear regression model estimated by the method of least squares.
- Chapter 3 introduces simple graphical methods for examining regression data and discusses how to transform variables to deal with common data analysis problems.

Figure 1.1 Anscombe's quartet: Four data sets with identical standard regression outputs (e.g., the equation of the common least-squares line and correlation coefficient are shown below the graphs).



Source: Adapted from Anscombe (1973). Reprinted by permission of the American Statistical Association, www.amstat.org.

- Chapter 4 describes methods for detecting unusual data in least-squares regression, distinguishing among high-leverage cases, outliers, and influential cases.

- Chapter 5 takes up the problems of nonnormally distributed errors and nonconstant error variance.

- Chapter 6 discusses methods for detecting and correcting nonlinearity.

- Chapter 7 describes methods for diagnosing collinearity.
- Chapter 8 extends the diagnostics discussed in the preceding chapters to GLMs.
- Chapter 9 makes recommendations for incorporating diagnostics in the work flow of regression analysis and suggests complementary readings.

My aim is to explain clearly the various kinds of problems that regression diagnostics address, to provide effective methods for detecting these problems, and, where appropriate, to suggest possible remedies. All the problems discussed in this monograph vary in degree from trivial to catastrophic, but I view nonlinearity as intrinsically the most serious problem, because it implies that we're fitting the wrong equation to the data.

The first edition of this monograph was published in 1991. This new edition has been thoroughly revised and rewritten, partly reflecting more recent developments in regression diagnostics, partly extending the coverage to GLMs, and partly reflecting my evolving understanding of the subject. I feel that it is only right to mention that I've addressed partially overlapping material in Fox (2016) and (with Sanford Weisberg) in Fox and Weisberg (2019). Although this monograph was written independently of these other sources, I have adapted some of the examples that appear in them and I'm aware that I may express myself similarly when writing about similar subject matter.

I have prepared a website for the monograph, with data and R code for the examples in the text at **https://tinyurl.com/RegDiag**. If you have difficulty finding the website, there is also a link to the supporting materials on the SAGE website at https://www.sagepub.com: After navigating to the SAGE website, search for "John Fox" to locate the SAGE webpage for the monograph.