WARNING

CONCERNING COPYRIGHT RESTRICTIONS

The Copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or reproduction.

One of three specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research.

If electronic transmission of reserve material is used for purposes in excess of what constitutes "fair use", that user may be liable for copyright infringement.

This policy is in effect for the following document:

TITLE: Examining data (Chapter 3) / from Applied regression analysis and generalized linear models
AUTHOR: Fox, John
SOURCE: Los Angeles: Sage, 2008 pp. 26-49

NO FURTHER TRANSMISSION OR DISTRIBUTION OF THIS MATERIAL IS PERMITTED

SECOND EDITION

APPLIED REGRESSION ANALYSIS and GENERALIZED LINEAR MODELS

John Fox
McMaster University, Hamilton, Ontario, Canada

SAGE
Los Angeles • London • New Delhi • Singapore

For information:

SAGE Publications, Inc.
2455 Teller Road
Thousand Oaks,
California 91320
E-mail: order@sagepub.com

SAGE Publications India Pvt. Ltd.
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road, New Delhi 110 044
India

SAGE Publications Ltd.
1 Oliver's Yard
55 City Road
London EC1Y 1SP
United Kingdom

SAGE Publications Asia-Pacific Pte. Ltd.
33 Pekin Street #02-01
Far East Square
Singapore 048763

Printed in the United States of America

Library of Congress Cataloging-in-Publication Data

Fox, John, 1947-
Applied regression analysis and generalized linear models/John Fox. —2nd ed.
p. cm.
Rev. ed. of: Applied regression analysis, linear models, and related methods. c1997.
Includes bibliographical references and index.
ISBN 978-0-7619-3042-6 (cloth)
1. Regression analysis. 2. Linear models (Statistics) 3. Social sciences—Statistical methods. I. Fox, John, 1947- Applied regression analysis and generalized linear models. II. Title.

HA31.3.F69 2008
300.1′519536—dc22 2007047617

Printed on acid-free paper

08 09 10 11 12 10 9 8 7 6 5 4 3 2 1

Acquisitions Editor:	Vicki Knight
Associate Editor:	Sean Connelly
Editorial Assistant:	Lauren Habib
Production Editor:	Cassandra Margaret Seibel
Copy Editor:	QuADS Prepress (P) Ltd.
Typesetter:	C&M Digitals (P) Ltd.
Proofreader:	Kevin Gleason
Cover Designer:	Candice Harman
Marketing Manager:	Stephanie Adams

3 Examining Data

This chapter, on graphical methods for examining data, and the next, on transformations, represent a digression from the principal focus of the book. Nevertheless, the material here is important to us for two reasons: First, careful data analysis should begin with inspection of the data.[1] You will find in this chapter simple methods for graphing univariate, bivariate, and multivariate data. Second, the techniques for examining and transforming data that are discussed in Chapters 3 and 4 will find direct application to the analysis of data using linear models.[2] Feel free, of course, to pass lightly over topics that are familiar.

To motivate the material in the chapter, and to demonstrate its relevance to the study of linear models, consider the four scatterplots shown in Figure 3.1.[3] The data for these plots, given in Table 3.1, were cleverly contrived by Anscombe (1973) to illustrate the central role of graphical methods in data analysis: Anticipating the material in Chapters 5 and 6, the least-squares regression line and all other common regression "outputs"—such as the correlation coefficient, standard deviation of the residuals, and standard errors of the regression coefficients—are identical in the four data sets.

It is clear, however, that each graph tells a different story about the data. Of course, the data are simply made up, so we have to allow our imagination some latitude:

- In Figure 3.1(a), the least-squares line is a reasonable descriptive summary of the tendency of Y to increase with X.
- In Figure 3.1(b), the linear regression fails to capture the clearly curvilinear relationship between the two variables; we would do much better to fit a quadratic function here,[4] that is, $Y = a + bX + cX^2$.
- In Figure 3.1(c), there is a perfect linear relationship between Y and X for all but one outlying data point. The least-squares line is pulled strongly toward the outlier, distorting the relationship between the two variables for the rest of the data. Perhaps the outlier represents an error in data entry or an observation that differs in some fundamental respect from the others. When we encounter an outlier in real data, we should look for an explanation.[5]
- Finally, in Figure 3.1(d), the values of X are invariant (all are equal to 8), with the exception of one point (which has an X value of 19); the least-squares line would be undefined but for this point—the line necessarily goes through the mean of the 10 Ys that share the value $X = 8$ and through the point for which $X = 19$. Furthermore, if this point were moved,

[1] An eminent statistican who has engaged in frequent consulting (and who will remain nameless for fear of embarrassing him) recently told me that his clients routinely extract only about 30% of the information in their data relevant to their research. He attributed this inefficiency largely to failure to examine the data carefully at an early stage in statistical analysis.

[2] See, for example, the treatments of graphical regression "diagnostics" and transformations in Chapters 11 and 12.

[3] See Section 3.2 for a general discussion of scatterplots.

[4] Quadratic and other polynomial regression models are discussed in Section 17.1.

[5] Outlier detection in linear models is taken up in Chapter 11.

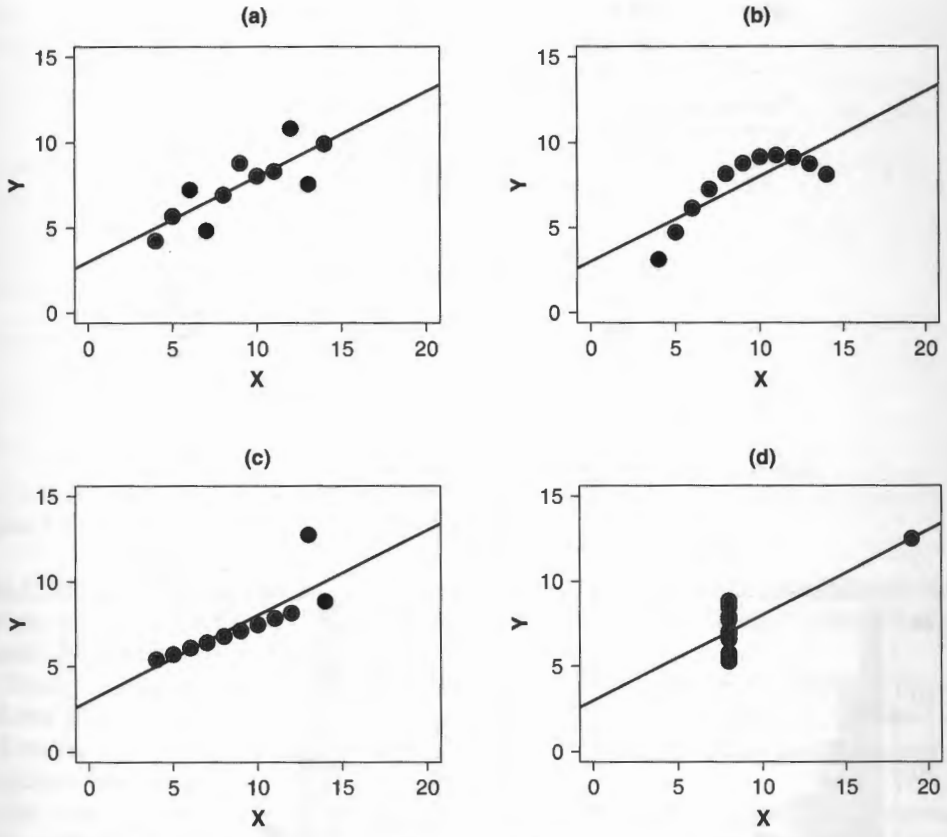

Figure 3.1 Four data sets, due to Anscombe (1973), with identical linear least-squares regressions. In (a), the linear regression is an accurate summary; in (b), the linear regression distorts the curvilinear relationship between Y and X; in (c), the linear regression is drawn toward an outlier; in (d), the linear regression "chases" the influential observation at the right. The least-squares line is shown on each plot.

then the regression line would chase it. We are usually uncomfortable having the result of a data analysis depend so centrally on a single influential observation.[6]

The essential point to be derived from Anscombe's "quartet" (so dubbed by Tufte, 1983) is that it is frequently helpful to examine data graphically. Important characteristics of data are often disguised by numerical summaries and—worse—the summaries can be fundamentally misleading. Moreover, directly examining the numerical data is often uninformative: Only in the fourth data set is the problem immediately apparent upon inspection of the numbers.

> Statistical graphs are central to effective data analysis, both in the early stages of an investigation and in statistical modeling.

[6]Influential data are discussed in Chapter 11.

Table 3.1 Four Contrived Regression Data Sets From Anscombe (1973)

$X_{a,b,c}$	Y_a	Y_b	Y_c	X_d	Y_d
10	8.04	9.14	7.46	8	6.58
8	6.95	8.14	6.77	8	5.76
13	7.58	8.74	12.74	8	7.71
9	8.81	8.77	7.11	8	8.84
11	8.33	9.26	7.81	8	8.47
14	9.96	8.10	8.84	8	7.04
6	7.24	6.13	6.08	8	5.25
4	4.26	3.10	5.39	19	12.50
12	10.84	9.13	8.15	8	5.56
7	4.82	7.26	6.42	8	7.91
5	5.68	4.74	5.73	8	6.89

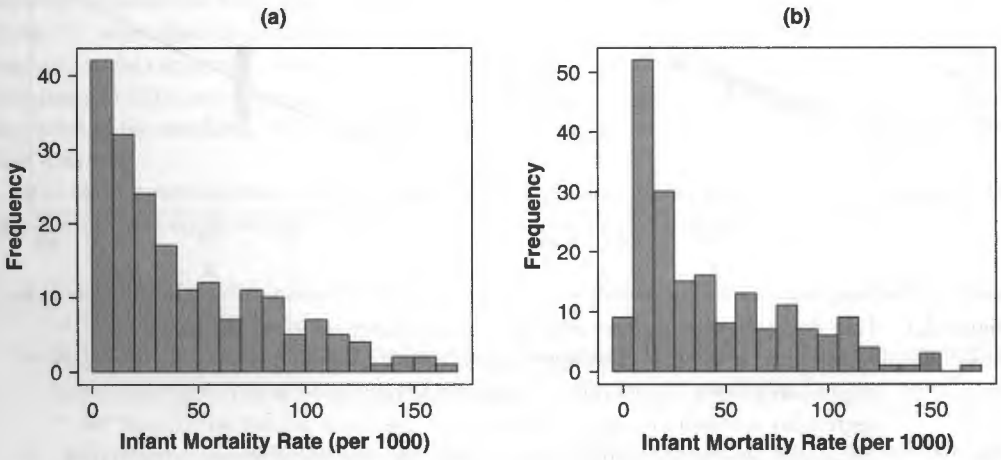

Figure 3.2 Histograms of infant morality for 193 nations. The histograms both use bins of width 10; histogram (a) employs bins that start at 0, while (b) employs bins that start at −5.

SOURCE: United Nations (1998).

3.1 Univariate Displays

3.1.1 Histograms

Figure 3.2 shows two *histograms* for the distribution of infant mortality among 193 countries, as reported in 1998 by the United Nations. The infant mortality rate is expressed as number of deaths of children aged less than 1 year per 1,000 live births. I assume that the histogram is a familiar graphical display, so I will offer only a brief description: To construct a histogram for infant mortality, dissect the range of the variable into equal-width intervals (called "bins"); count the number of observations falling in each bin; and display the frequency counts in a bar graph.

Both histograms in Figure 3.2 use bins of width 10; they differ in that the bins in Figure 3.2(a) start at 0 (i.e., 0 to 10, 10 to 20, etc.), while those in Figure 3.2(b) start at −5 (i.e., −5 to 5,

```
1 | 2: represents 12
 leaf unit: 1
        n: 193

   39     0 | 345555555666666667777777777778888999999
   72     1 | 000122222333344444555566778888999
   95     2 | 00112223333444455556669
  (19)    3 | 0001233445577889999
   79     4 | 012344456889
   67     5 | 11246667888
   56     6 | 01255568
   48     7 | 122347788
   39     8 | 00222456669
   28     9 | 025678
   22    10 | 234677
   16    11 | 023445
   10    12 | 2445
    6    13 | 2
    5    14 | 29

HI: 153 [Liberia], 154 [Afghanistan), 169 [Sierra Leone]
```

Figure 3.3 Stem-and-leaf display for infant mortality.

5 to 15, etc.).[7] The two histograms for infant mortality are more similar than different—both, for example, show that the distribution of infant mortality is positively skewed—but they do give slightly different impressions of the shape of the distribution.

Figure 3.3 shows an alternative form of histogram, called a *stem-and-leaf display*. The stem-and-leaf plot, introduced by John Tukey (1972, 1977), ingeniously employs the numerical data to form the bars of the histogram. As Tukey suggests, it is simple to construct a stem-and-leaf display by hand to "scratch down" a small data set.

You may be familiar with the stem-and-leaf display. Here is a relatively compressed explanation:

- Each data value is broken between two adjacent digits into a "stem" and a "leaf": In Figure 3.3, the break takes place between the tens and units digits. For example, the infant mortality rate in Albania was 32, which translates into the stem 3 and leaf 2.
- Stems (here, 0, 1, ..., 14) are constructed to cover the data, implicitly defining a system of bins, each of width 10. Each leaf is placed to the right of its stem, and the leaves on each stem are then sorted into ascending order. We can produce a finer system of bins by dividing each stem into two parts (taking, respectively, leaves 0–4 and 5–9), or five parts (0–1, 2–3, 4–5, 6–7, 8–9); for the infant mortality data, two-part stems would correspond to bins of width 5 and five-part stems to bins of width 2. We could employ still finer bins by dividing stems from leaves between the ones and tenths digits, but, for infant mortality, that would produce a display with almost as many bins as observations. Similarly, a coarser division between the hundreds and tens digits would yield only two stems—0 and 1.
- Unusually large values—*outliers*—are collected on a special "HI" stem and displayed individually. Here, there are three countries with unusually large infant mortality rates. Were there countries with unusually small infant mortality rates, then these would be collected and displayed individually on a "LO" stem.[8]
- The column of *depths* counts in toward the median from both ends of the distribution. The median is the observation at depth $(n + 1)/2$, where (as usual) n is the number of observations. For the infant mortality data, the median is at depth $(193 + 1)/2 = 97$. In

[7] Because infant mortality cannot be negative, the contrast between Figures 3.2(a) and (b) is somewhat artificial.

[8] The rule for identifying outliers is explained in Section 3.1.4 on boxplots.

Figure 3.3, there are 39 observations at stem 0, 72 at and below stem 1, and so on; there are five observations (including the outliers) at and above stem 14, six at and above stem 13, and so forth. The count at the stem containing the median is shown in parentheses—here, 19 at stem 3. Note that $95 + 19 + 79 = 193$.

In constructing histograms (including stem-and-leaf displays), we want enough bins to preserve some detail, but not so many that the display is too rough and dominated by sampling variation. Let n^* represent the number of nonoutlying observations. Then, for $n^* \leq 100$, it usually works well to use no more than about $2\sqrt{n^*}$ bins; likewise, for $n^* > 100$, we can use a maximum of about $10 \times \log_{10} n^*$ bins. Of course, in constructing a histogram, we also want bins that start and end at "nice" numbers (e.g., 10 to 20 rather than 9.5843 to 21.0457); in a stem-and-leaf display, we are limited to bins that correspond to breaks between digits of the data values. Computer programs that construct histograms incorporate rules such as these.[9]

For the distribution of infant mortality, $n^* = 193 - 3 = 190$, so we should aim for no more than $10 \times \log_{10}(190) \approx 23$ bins. The stem-and-leaf display in Figure 3.3 uses 15 stems (plus the "HI" stem).

Histograms, including stem-and-leaf displays, are very useful graphs, but they suffer from several problems:

- As we have seen, the visual impression of the data conveyed by a histogram can depend on the arbitrary origin of the bin system.
- Because the bin system dissects the range of the variable into class intervals, the histogram is discontinuous (i.e., rough) even if, as in the case of infant mortality, the variable is continuous.[10]
- The form of the histogram depends on the arbitrary width of the bins.
- Moreover, if we use bins that are narrow enough to capture detail where data are plentiful—usually near the center of the distribution—then they may be too narrow to avoid "noise" where data are sparse—usually in the tails of the distribution.

3.1.2 Nonparametric Density Estimation

Nonparametric density estimation addresses the deficiencies of traditional histograms by averaging and smoothing. As the term implies, "density estimation" can be construed formally as an attempt to estimate the probability density function of a variable based on a sample, but it can also be thought of informally as a descriptive technique for smoothing histograms.

In fact, the histogram—suitably rescaled—is a simple density estimator.[11] Imagine that the origin of the bin system is at x_0, and that each of the m bins has width $2h$; the end points of the

[9] More sophisticated rules for the number of bins take into account information beyond n. For example, Freedman and Diaconis (1981) suggest

$$\text{number of bins} \approx \left\lceil \frac{n^{1/3}\left(x_{(n)} - x_{(1)}\right)}{2(Q_3 - Q_1)} \right\rceil$$

where $x_{(n)} - x_{(1)}$ is the range of the data, $Q_3 - Q_1$ is the inter-quartile range, and the "ceiling" brackets indicate rounding *up* to the next integer.

[10] That is, infant mortality rates are continuous for practical purposes in that they can take on many different values. Actually, infant mortality rates are ratios of integers and hence are rational numbers, and the rates in the U.N. data set are rounded to the nearest whole number.

[11] Rescaling is required because a density function encloses a total area of 1. Histograms are typically scaled so that the height of each bar represents frequency (or percent), and thus the heights of the bars sum to the sample size n (or 100). If each bar spans a bin of width $2h$ (anticipating the notation below), then the total area enclosed by the bars is $n \times 2h$. Dividing the height of each bar by $2nh$ therefore produces the requisite density rescaling.

bins are then at $x_0, x_0 + 2h, x_0 + 4h, \ldots, x_0 + 2mh$. An observation X_i falls in the jth bin if (by convention)

$$x_0 + 2(j-1)h \leq X_i < x_0 + 2jh$$

The histogram estimator of the density at any x value located in the jth bin is based on the number of observations that fall in that bin:

$$\widehat{p}(x) = \frac{\#_{i=1}^{n}[x_0 + 2(j-1)h \leq X_i < x_0 + 2jh]}{2nh}$$

where # is the counting operator.

We can dispense with the arbitrary origin x_0 of the bin system by counting locally within a continuously moving window of half-width h centered at x:

$$\widehat{p}(x) = \frac{\#_{i=1}^{n}(x - h \leq X_i < x + h)}{2nh}$$

In practice, of course, we would use a computer program to evaluate $\widehat{p}(x)$ at a large number of x values covering the range of X. This "naive density estimator" (so named by Silverman, 1986) is equivalent to locally weighted averaging, using a rectangular weight function:

$$\widehat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} W\left(\frac{x - X_i}{h}\right) \tag{3.1}$$

where

$$W(z) = \begin{cases} \frac{1}{2} & \text{for } |z| < 1 \\ 0 & \text{otherwise} \end{cases}$$

a formulation that will be useful below when we consider alternative weight functions to smooth the density. Here z is a "stand-in" for the argument to the $W(\cdot)$ weight function—that is, $z = (x - X_i)/h$. The naive estimator is like a histogram that uses bins of width $2h$ but has no fixed origin, and is similar in spirit to the naive nonparametric-regression estimator introduced in Chapter 2.

An illustration, using the U.N. infant mortality data, appears in Figure 3.4, and reveals the principal problem with the naive estimator: Because the estimated density jumps up and down as observations enter and leave the window, the naive density estimator is intrinsically rough.

The rectangular weight function $W(z)$ in Equation 3.1 is defined to enclose an area of $2 \times \frac{1}{2} = 1$, producing a density estimate that (as required) also encloses an area of 1. Any function that has this property—probability density functions are obvious choices—may be used as a weight function, called a *kernel*. Choosing a kernel that is smooth, symmetric, and unimodal smooths out the rough edges of the naive density estimator. This is the essential insight of *kernel density estimation*.

The general kernel density estimator is, then, given by

$$\widehat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

There are many reasonable choices of the kernel function $K(z)$, including the familiar standard normal density function, $\phi(z)$, which is what I will use here. While the naive density estimator in effect sums suitably scaled rectangles centered at the observations, the more general kernel

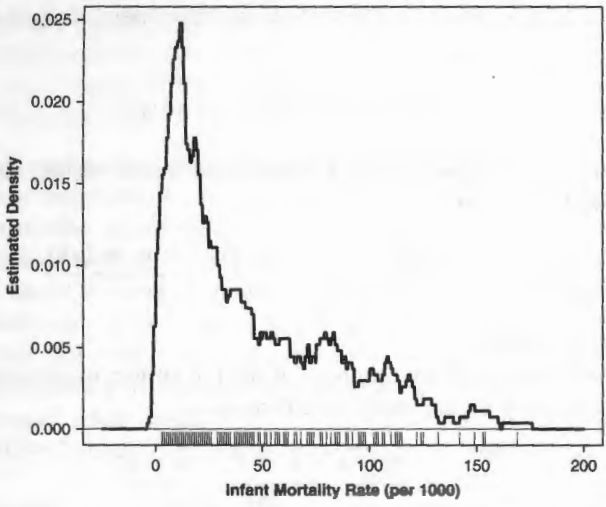

Figure 3.4 Naive density estimator for infant mortality, using a window half-width of $h = 7$. Note the roughness of the estimator. A *rug-plot* (or "one-dimensional scatterplot") appears at the bottom of the graph, showing the location of the data values.

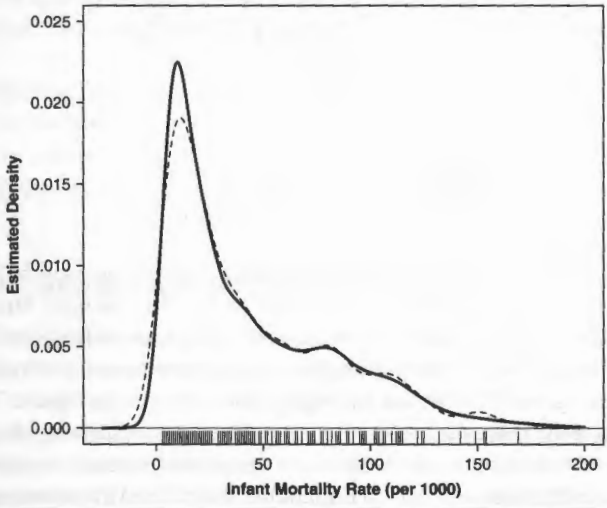

Figure 3.5 Kernel (broken line) and adaptive-kernel (solid line) density estimates for the distribution of infant mortality, using a normal kernel and a window half-width of $h = 7$. Note the relative "lumpiness" of the kernel estimator at the right, where data are sparse.

estimator sums smooth lumps. An example is shown in Figure 3.5, in which the kernel density estimator is given by the broken line.[12]

Selecting the window width for the kernel estimator is primarily a matter of trial and error—we want a value small enough to reveal detail but large enough to suppress random noise. We can,

[12]Notice that there is nonzero estimated density in Figure 3.5 below an infant mortality rate of 0. Of course, this does not make sense, and although I will not pursue it here, it is possible to constrain the lower and upper limits of the kernel estimator.

however, look to statistical theory for rough guidance:[13] If the underlying density that we are trying to estimate is normal with standard deviation σ, then (for the normal kernel) estimation is most efficient with the window half-width

$$h = 0.9\sigma n^{-1/5} \tag{3.2}$$

As is intuitively reasonable, the optimal window grows gradually narrower as the sample size is increased, permitting finer detail in large samples than in small ones.[14]

Although we might, by reflex, be tempted to replace the unknown σ in Equation 3.2 with the sample standard deviation S, it is prudent to be more cautious, for if the underlying density is sufficiently non-normal, then the sample standard deviation may be seriously inflated. A common compromise is to use an "adaptive" estimator of spread:

$$A = \min\left(S, \frac{\text{interquartile range}}{1.349}\right) \tag{3.3}$$

The factor 1.349 is the interquartile range of the standard normal distribution, making (interquartile range)/1.349 a robust estimator of σ in the normal setting.

One further caveat: If the underlying density is substantially non-normal—in particular, if it is skewed or multimodal—then basing h on the adaptive estimator A generally produces a window that is too wide. A good procedure, then, is to start with

$$h = 0.9An^{-1/5}$$

and to adjust this value downwards until the resulting density plot becomes too rough. This is the procedure that was used to find the window width in Figure 3.5, where $S = 38.55$ and (interquartile range)/1.349 $= (68 - 13)/1.349 = 40.77$. Here, the "optimal" window width is $h = 0.9 \times 38.55 \times 197^{-1/5} = 12.061$.

The kernel density estimator usually does a pretty good job, but the window half-width h remains a compromise: We would prefer a narrower window where data are plentiful (to preserve detail) and a wider one where data are sparse (to suppress noise). Because "plentiful" and "sparse" refer implicitly to the underlying density that we are trying to estimate, it is natural to begin with an initial estimate of the density, and to adjust the window half-width on the basis of the initial estimate.[15] The result is the *adaptive-kernel estimator* (not to be confused with the adaptive estimator of spread in Equation 3.3).

1. Calculate an initial density estimate, $\widetilde{p}(x)$—for example, by the kernel method.

2. Using the initial estimate, compute local window factors by evaluating the estimated density at the observations:

$$f_i = \left[\frac{\widetilde{p}(X_i)}{\widetilde{p}}\right]^{-1/2}$$

In this formula, $\widetilde{p}$ is the geometric mean of the initial density estimates at the observations—that is,

$$\widetilde{p} = \left[\prod_{i=1}^{n} \widetilde{p}(X_i)\right]^{1/n}$$

[13] See, for example, Silverman (1986, chap. 3) for a detailed discussion of these issues.

[14] If we really knew that the density were normal, then it would be even more efficient to estimate it parametrically by substituting the sample mean $\overline{X}$ and standard deviation S for μ and σ in the formula for the normal density, $p(x) = (2\pi\sigma^2)^{-1/2}\exp[-(x-\mu)^2/2\sigma^2]$.

[15] An alternative is to use a *nearest-neighbor* approach, as in the nonparametric-regression methods discussed in Chapter 2.

(where the operator $\prod$ indicates continued multiplication). As a consequence of this definition, the f_is have a product of 1, and hence a geometric mean of 1, ensuring that the area under the density estimate remains equal to 1.

3. Calculate the adaptive-kernel density estimator using the local window factors to adjust the width of the kernels centered at the observations:

$$\widehat{p}(x) = \frac{1}{nh}\sum_{i=1}^{n}\frac{1}{f_i}K\left(\frac{x - X_i}{f_i h}\right)$$

Applying the adaptive kernel estimator to the infant mortality distribution produces the solid line in Figure 3.5: For this distribution the kernel and adaptive-kernel estimates are very similar, although the adaptive kernel more sharply defines the principal mode of the distribution near 20, and produces a smoother long right tail.

3.1.3 Quantile-Comparison Plots

Quantile-comparison plots are useful for comparing an empirical sample distribution with a theoretical distribution, such as the normal distribution—something that is more commonly of interest for derived quantities such as test statistics or residuals than for observed variables. A strength of the display is that it does not require the use of arbitrary bins or windows.

Let $P(x)$ represent the theoretical *cumulative distribution function* (CDF) with which we want to compare the data; that is, $P(x) = \Pr(X \leq x)$. A simple (but not terribly useful) procedure is to graph the *empirical cumulative distribution function* (ECDF) for the observed data, which is simply the proportion of data below each value of x, as x moves continuously from left to right:

$$\widehat{P}(x) = \frac{\#_{i=1}^{n}(X_i \leq x)}{n}$$

As illustrated in Figure 3.6, however, the ECDF is a "stair-step" function (where each step occurs at an observation, and is of height $1/n$), while the CDF is typically smooth, making the comparison difficult.

The quantile-comparison plot avoids this problem by never constructing the ECDF explicitly:

1. Order the data values from smallest to largest, $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. The $X_{(i)}$ are called the *order statistics* of the sample.

2. By convention, the cumulative proportion of the data "below" $X_{(i)}$ is given by[16]

$$P_i = \frac{i - \frac{1}{2}}{n}$$

3. Use the inverse of the CDF (that is, the *quantile function*) to find the value z_i corresponding to the cumulative probability P_i; that is,[17]

$$z_i = P^{-1}\left(\frac{i - \frac{1}{2}}{n}\right)$$

[16]This definition avoids cumulative proportions of 0 or 1, which would be an embarrassment in step 3 for distributions, like the normal, that never quite reach cumulative probabilities of 0 or 1. In effect, we count half of each observation below its exact value and half above. Another common convention is to use $P_i = \left(i - \frac{1}{3}\right)/\left(n + \frac{1}{3}\right)$.

[17]This operation assumes that the CDF has an inverse—that is, that P is a strictly increasing function (one that never quite levels off). The common continuous probability distributions in statistics—for example, the normal, t-, F-, and χ^2 distributions—all have this property. These and other distributions are reviewed in Appendix D on probability and estimation.

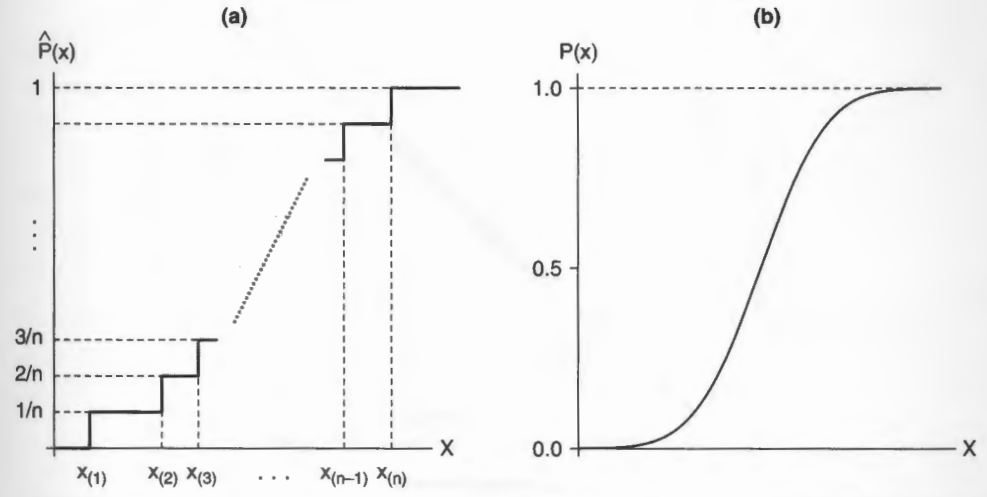

Figure 3.6 A "typical" empirical cumulative distribution function (ECDF) is shown in (a), a "typical" theoretical cumulative distribution function (CDF) in (b). $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ represent the data values ordered from smallest to largest. Note that the ordered data values are not, in general, equally spaced.

4. Plot the z_i as horizontal coordinates against the $X_{(i)}$ as vertical coordinates. If X is sampled from the distribution P, then $X_{(i)} \approx z_i$. That is, the plot should be approximately linear, with an intercept of 0 and slope of 1. This relationship is only approximate because of sampling error (see point 6). If the distributions are identical except for location, then the plot is approximately linear with a nonzero intercept, $X_{(i)} \approx \mu + z_i$; if the distributions are identical except for scale, then the plot is approximately linear with a slope different from 1, $X_{(i)} \approx \sigma z_i$; finally, if the distributions differ both in location and scale but have the same shape, then $X_{(i)} \approx \mu + \sigma z_i$.

5. It is often helpful to place a comparison line on the plot to facilitate the perception of departures from linearity. The line can be plotted by eye, attending to the central part of the data, or we can draw a line connecting the quartiles. For a normal quantile-comparison plot—comparing the distribution of the data with the standard normal distribution—we can alternatively use the median as a robust estimator of μ and the interquartile range/1.349 as a robust estimator of σ. (The more conventional estimates $\widehat{\mu} = \overline{X}$ and $\widehat{\sigma} = S$ will not work well when the data are substantially non-normal.)

6. We expect some departure from linearity because of sampling variation; it therefore assists interpretation to display the expected degree of sampling error in the plot. The standard error of the order statistic $X_{(i)}$ is

$$\mathrm{SE}(X_{(i)}) = \frac{\widehat{\sigma}}{p(z_i)}\sqrt{\frac{P_i(1-P_i)}{n}}$$

where $p(z)$ is the probability density function corresponding to the CDF $P(z)$. The values along the fitted line are given by $\widehat{X}_{(i)} = \widehat{\mu} + \widehat{\sigma} z_i$. An approximate 95% confidence "envelope" around the fitted line is, therefore,[18]

[18]By the method of construction, the 95% confidence level applies (point-wise) to each $\widehat{X}_{(i)}$, not to the whole envelope: There is a greater probability that *at least one* point strays outside the envelope even if the data are sampled from the

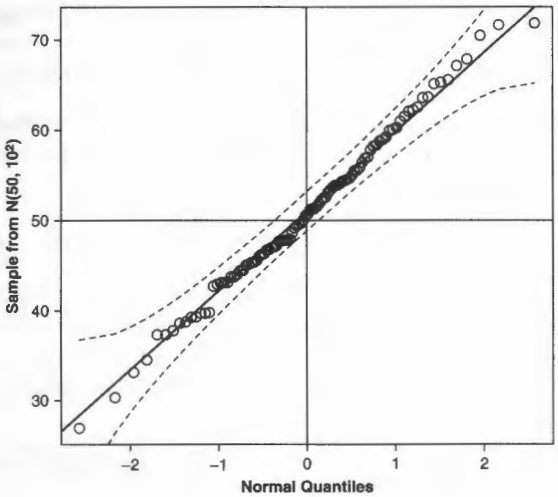

Figure 3.7 Normal quantile-comparison plot for a sample of 100 observations drawn from a normal distribution with mean 50 and standard deviation 10. The fitted line is through the quartiles of the distribution, and the broken lines give a point-wise 95% confidence interval around the fit.

$$\widehat{X}_{(i)} \pm 2 \times \mathrm{SE}(X_{(i)})$$

Figures 3.7 to 3.10 display normal quantile-comparison plots for several illustrative distributions:

- Figure 3.7 plots a sample of $n = 100$ observations from a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 10$. The plotted points are reasonably linear and stay within the rough 95% confidence envelope.
- Figure 3.8 plots a sample of $n = 100$ observations from the positively skewed chi-square distribution with 2 degrees of freedom. The positive skew of the data is reflected in points that lie *above* the comparison line in both tails of the distribution. (In contrast, the tails of negatively skewed data would lie *below* the comparison line.)
- Figure 3.9 plots a sample of $n = 100$ observations from the heavy-tailed t-distribution with 2 degrees of freedom. In this case, values in the upper tail lie above the corresponding normal quantiles, and values in the lower tail below the corresponding normal quantiles.
- Figure 3.10 shows the normal quantile-comparison plot for the distribution of infant mortality. The positive skew of the distribution is readily apparent. The possibly bimodal character of the data, however, is not easily discerned in this display.

Quantile-comparison plots highlight the tails of distributions. This is important, because the behavior of the tails is often problematic for standard estimation methods like least squares, but it is useful to supplement quantile-comparison plots with other displays—such as histograms or kernel-density estimates—that provide more intuitive representations of distributions. A key point is that there is no reason to limit ourselves to a single picture of a distribution when different pictures bring different aspects of the distribution into relief.

---

comparison distribution. Determining a *simultaneous* 95% confidence envelope would be a formidable task, because the order statistics are not independent.

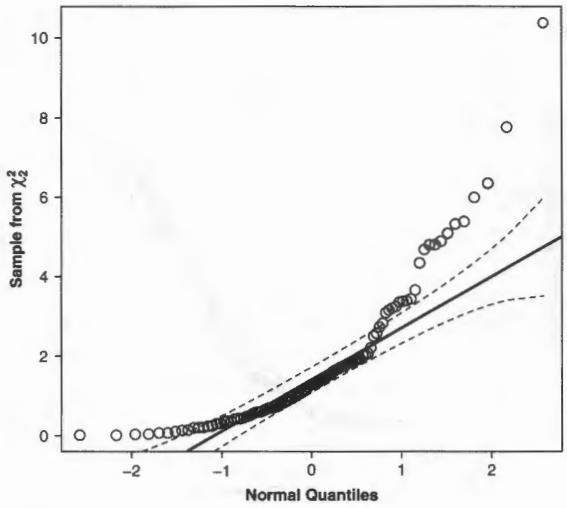

Figure 3.8 Normal quantile-comparison plot for a sample of 100 observations from the positively skewed chi-square distribution with 2 degrees of freedom.

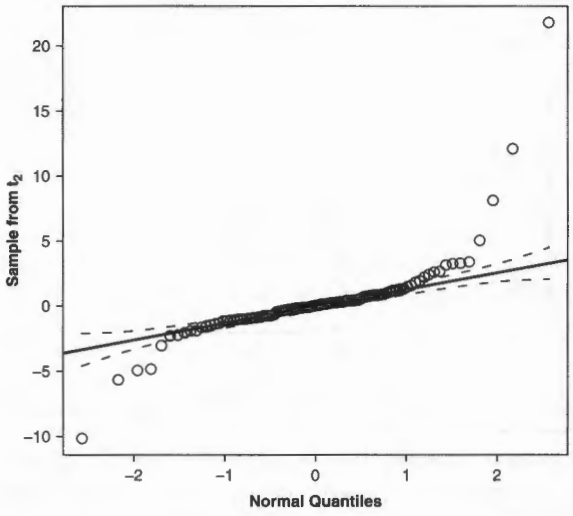

Figure 3.9 Normal quantile-comparison plot for a sample of 100 observations from the heavy-tailed t-distribution with 2 degrees of freedom.

3.1.4 Boxplots

Unlike histograms, density plots, and quantile-comparison plots, *boxplots* (due to Tukey, 1977) present only summary information on center, spread, and skewness, along with individual outlying observations. Boxplots are constructed from the *five-number summary* of a distribution—the minimum, first quartile, median, third quartile, and maximum—and outliers, if they are present. Boxplots, therefore, are useful when we require a compact representation of a distribution (as, for example, in the margins of a scatterplot), when we wish to compare the principal characteristics

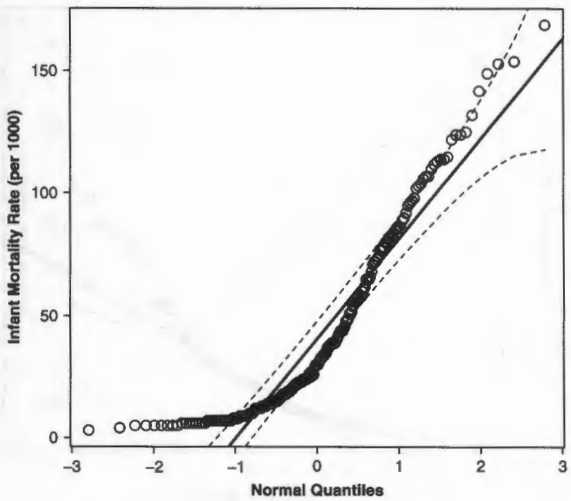

Figure 3.10 Normal quantile-comparison plot for the distribution of infant mortality. Note the positive skew.

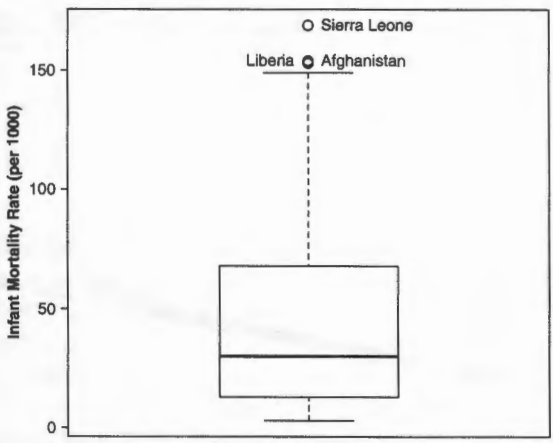

Figure 3.11 Boxplot for infant mortality. The central box is drawn between the hinges; the position of the median is marked in the box; and outlying observations are displayed individually.

of several distributions,[19] or when we want to select a transformation that makes a distribution more symmetric.[20]

An illustrative boxplot for infant mortality appears in Figure 3.11. This plot is constructed according to the following conventions (illustrated in the schematic horizontal boxplot in Figure 3.12):

1. A scale is laid off to accommodate the extremes of the data. The infant mortality data, for example, range between 3 and 169.

[19] See Section 3.2.

[20] Transformations to symmetry are discussed in Chapter 4.

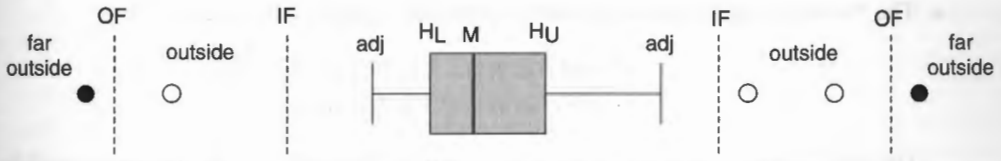

Figure 3.12 Schematic boxplot, showing the median (M), hinges (H_L and H_U), adjacent values (adj), inner and outer fences (IF and OF), and outside and far-outside observations.

2. The central box is drawn between the *hinges*, which are simply defined first and third quartiles, and therefore encompasses the middle half of the data. The line in the central box represents the median. Recall that the depth of the median is

$$\text{depth}(M) = \frac{n+1}{2}$$

giving the position of the middle observation after the data are ordered from smallest to largest: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$. When n is even, the depth of the median has a fractional part; using "floor" brackets to represent truncation to an integer, we count in from either end to average the two observations at depth $\lfloor (n+1)/2 \rfloor$. For the infant mortality data, $\text{depth}(M) = (193+1)/2 = 97$, and $M = X_{(97)} = 30$.
Likewise, the depth of the hinges is

$$\text{depth}(H) = \frac{\lfloor \text{depth}(M) \rfloor + 1}{2}$$

If $\text{depth}(H)$ has a fractional part, then, for each hinge, we average the two observations at the adjacent positions, that is, at $\lfloor \text{depth}(H) \rfloor$ and $\lfloor \text{depth}(H) \rfloor + 1$. For the infant mortality distribution, $\text{depth}(H) = (97+1)/2 = 49$. The lower hinge is, therefore, $H_L = X_{(49)} = 13$, and the upper hinge is $H_U = X_{(149)} = 73$. (Counting down 97 observations from the top yields the subscript $197 - 49 + 1 = 149$.)

3. The following rules are used to identify outliers, which are shown individually in the boxplot:

- The *hinge-spread* (roughly the interquartile range) is the difference between the hinges:

$$H\text{-spread} = H_U - H_L$$

- The lower and upper "inner fences" are located 1.5 hinge-spreads beyond the hinges:

$$\text{IF}_L = H_L - 1.5 \times H\text{-spread}$$
$$\text{IF}_U = H_U + 1.5 \times H\text{-spread}$$

Observations beyond the inner fences (but within the outer fences, defined below) are termed "outside" and are represented by open circles. The fences themselves are not shown in the display.

- The "outer fences" are located three hinge-spreads beyond the hinges:[21]

$$\text{OF}_L = H_L - 3 \times H\text{-spread}$$
$$\text{OF}_U = H_U + 3 \times H\text{-spread}$$

  Observations beyond the outer fences are termed "far outside" and are represented by filled circles. There are no far-outside observations in the infant mortality data.
- The "whisker" growing from each end of the central box extends either to the extreme observation on its side of the distribution (as at the low end of the infant mortality data) or to the most extreme nonoutlying observation, called the "adjacent value" (as at the high end of the infant mortality distribution).[22]

The boxplot of infant mortality in Figure 3.11 clearly reveals the skewness of the distribution: The lower whisker is much shorter than the upper whisker; the median is closer to the lower hinge than to the upper hinge; and there are several outside observations at the upper end of the infant mortality distribution, but none at the lower end. The apparent bimodality of the infant mortality data is not captured by the boxplot, however.

There are many useful univariate displays, including the traditional histogram. The stem-and-leaf plot is a modern variant of the histogram for small data sets, constructed directly from numerical data. Nonparametric density estimation may be employed to smooth a histogram. Quantile comparison plots are useful for comparing data with a theoretical probability distribution. Boxplots summarize some of the most important characteristics of a distribution, including center, spread, skewness, and outliers.

3.2 Plotting Bivariate Data

The *scatterplot*—a direct geometric representation of observations on two quantitative variables (generically, Y and X)—is the most useful of all statistical graphs. The scatterplot is a natural representation of data partly because the media on which we draw plots—paper, computer screens—are intrinsically two dimensional. Scatterplots are as familiar and essentially simple as they are useful; I will therefore limit this presentation to a few points. There are many examples of bivariate scatterplots in this book, including in the preceding chapter.

- In analyzing data, it is convenient to work in a computing environment that permits the interactive identification of observations in a scatterplot.
- Because relationships between variables in the social sciences are often weak, scatterplots can be dominated visually by "noise." It often helps, therefore, to plot a nonparametric regression of Y on X.

[21] Here is a rough justification for the fences: In a normal population, the hinge-spread is 1.349 standard deviations, and so $1.5 \times H\text{-spread} = 1.5 \times 1.349 \times \sigma \approx 2\sigma$. The hinges are located $1.349/2 \approx 0.7$ standard deviations above and below the mean. The inner fences are, therefore, approximately at $\mu \pm 2.7\sigma$, and the outer fences at $\mu \pm 4.7\sigma$. From the standard normal table, $\Pr(Z > 2.7) \approx .003$, so we expect slightly less than 1% of the observations beyond the inner fences ($2 \times .003 = .006$); likewise, because $\Pr(Z > 4.7) \approx 1.3 \times 10^{-6}$, we expect less than one observation in 100,000 beyond the outer fences.

[22] All of the folksy terminology—"hinges," "fences," "whiskers," and so on—originates with Tukey (1977).

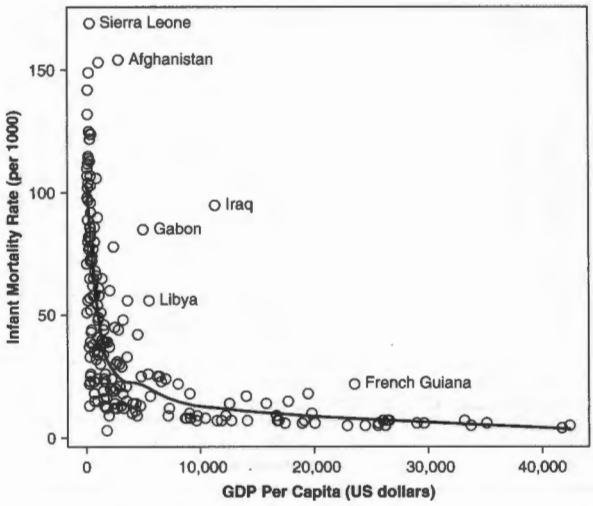

Figure 3.13 Scatterplot for infant mortality and GDP per capita for 193 nations. The line is for a lowess smooth with a span of 1/2. Several nations with high infant mortality for their levels of GDP are identified.

- Scatterplots in which one or both variables are highly skewed are difficult to examine, because the bulk of the data congregate in a small part of the display. Consider, for example, the scatterplot for infant mortality and gross domestic product (GDP) per capita in Figure 3.13. It often helps to "correct" substantial skews prior to examining the relationship between Y and X.[23]
- Scatterplots in which the variables are discrete can also be difficult to examine. An extreme instance of this phenomenon is shown in Figure 3.14, which plots scores on a 10-item vocabulary test against years of education. The data are from 16 of the U.S. General Social Surveys conducted by the National Opinion Research Center between 1974 and 2004, and include in total 21,638 observations. One solution—especially useful when only X is discrete—is to focus on the conditional distribution of Y for each value of X. Boxplots, for example, can be employed to represent the conditional distributions (see Figure 3.16, discussed below). Another solution is to separate overlapping points by adding a small random quantity to the discrete scores. In Figure 3.15, for example, I have added a uniform random variable on the interval $[-0.4, +0.4]$ to each value of vocabulary and education. Paradoxically, the tendency for vocabulary to increase with education is much clearer in the randomly "jittered" display.[24]

> The bivariate scatterplot is a natural graphical display of the relationship between two quantitative variables. Interpretation of a scatterplot can often be assisted by graphing a nonparametric regression, which summarizes the relationship between the two variables. Scatterplots of the relationship between discrete variables can be enhanced by randomly jittering the data.

[23] See Chapter 4.

[24] The idea of jittering a scatterplot, as well as the terminology, is due to Cleveland (1994).

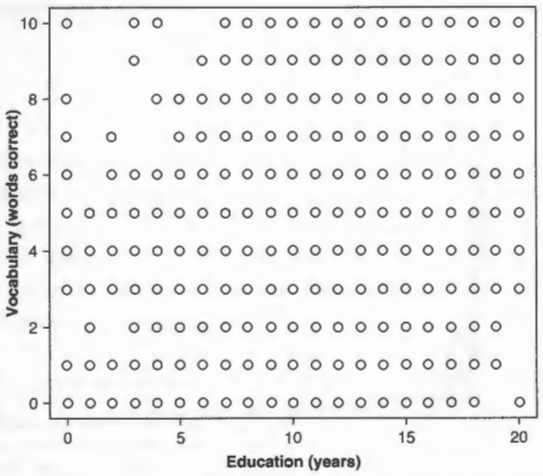

Figure 3.14 Scatterplot of scores on a 10-item vocabulary test versus years of education. Although there are nearly 22,000 observations in the data set, most of the plotted points fall on top of one another.

SOURCE: National Opinion Research Center (2005).

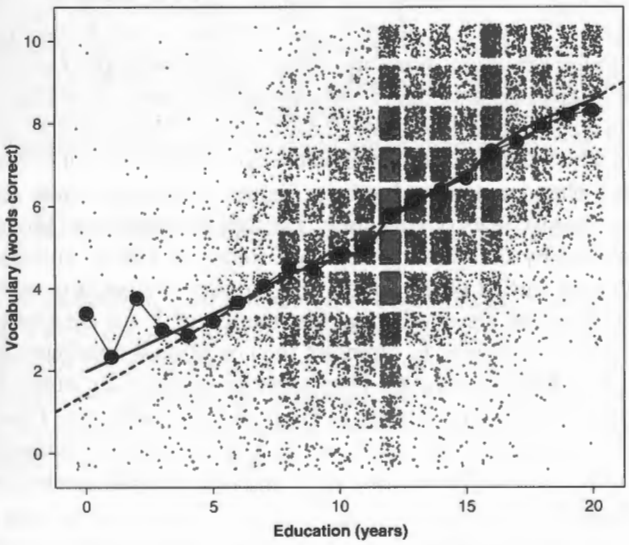

Figure 3.15 Jittered scatterplot for vocabulary score versus years of education. A uniformly distributed random quantity between −0.4 and +0.4 was added to each score for both variables. The heavier solid line is for a lowess fit to the data, with a span of 0.2; the broken line is the linear least-squares fit; the conditional means for vocabulary given education are represented by the dots, connected by the lighter solid line.

As mentioned, when the explanatory variable is discrete, parallel boxplots can be used to display the conditional distributions of Y. One common case occurs when the explanatory variable is a qualitative/categorical variable. An example is shown in Figure 3.16, using data collected by Michael Ornstein (1976) on interlocking directorates among the 248 largest Canadian firms. The response

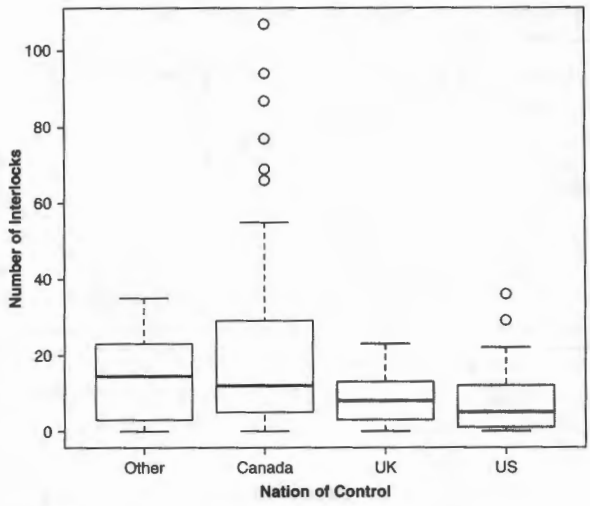

Figure 3.16 Number of interlocking directorate and executive positions by nation of control, for 248 dominant Canadian firms.

SOURCE: Personal communication from Michael Ornstein.

variable in this graph is the number of interlocking directorships and executive positions maintained by each firm with others in the group of 248. The explanatory variable is the nation in which the corporation is controlled, coded as Canada, United Kingdom, United States, and other foreign.

It is apparent from the graph that the average level of interlocking is greater among other-foreign and Canadian corporations than among corporations controlled in the United Kingdom and the United States. It is relatively difficult to discern detail in this display: first, because the conditional distributions of interlocks are positively skewed; and; second, because there is an association between level and spread—variation is also greater among other-foreign and Canadian firms than among U.K. and U.S. firms.[25]

Parallel boxplots display the relationship between a quantitative response variable and a discrete (categorical or quantitative) explanatory variable.

3.3 Plotting Multivariate Data

Because paper and computer screens are two dimensional, graphical display of multivariate data is intrinsically difficult. Multivariate displays for quantitative data often project the higher-dimensional "point cloud" of the data onto a two-dimensional space. It is, of course, impossible to view a higher-dimensional scatterplot directly (but see the discussion of the three-dimensional case below). The essential trick of effective multidimensional display is to select projections that reveal important characteristics of the data. In certain circumstances, projections can be selected on the basis of a statistical model fit to the data or on the basis of explicitly stated criteria.[26]

[25]We will revisit this example in Section 4.4. Because the names of the firms are unavailable, I have not identified the outliers in the plot.

[26]We will apply these powerful ideas in Chapters 11 and 12.

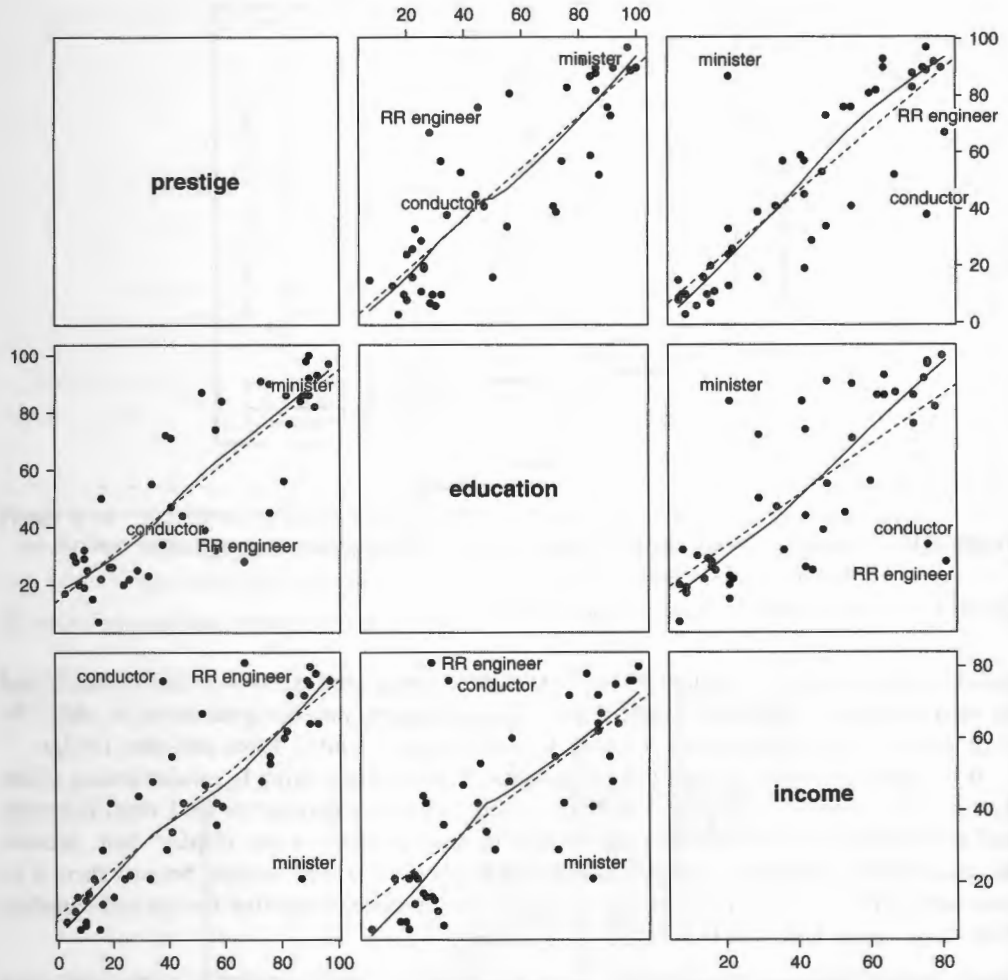

Figure 3.17 Scatterplot matrix for occupational prestige, level of education, and level of income, for 45 U.S. occupations in 1950. The least-squares regression line (broken line) and lowess smooth (for a span of 0.6, solid line) are shown on each plot. Three unusual observations are identified.

SOURCE: Duncan (1961).

3.3.1 Scatterplot Matrices

A simple approach to multivariate data, which does not require a statistical model, is to examine bivariate scatterplots for all pairs of variables. Arraying these plots in a *scatterplot matrix* produces a graphical analog to the correlation matrix.

An illustrative scatterplot matrix, for data on the prestige, education, and income levels of 45 U.S. occupations, appears in Figure 3.17. In this data set, first analyzed by Duncan (1961), "prestige" represents the percentage of respondents in a survey who rated an occupation as "good" or "excellent" in prestige; "education" represents the percentage of incumbents in the occupation in the 1950 U.S. Census who were high-school graduates; and "income" represents the percentage of occupational incumbents who earned incomes in excess of $3500. Duncan's purpose was to use a regression analysis of prestige on income and education to predict the prestige levels of

other occupations, for which data on income and education were available, but for which there were no direct prestige ratings.[27]

The variable names on the diagonal of the scatterplot matrix in Figure 3.17 label the rows and columns of the display: For example, the vertical axis for the two plots in the first row of the display is "prestige"; the horizontal axis for the two plots in the second column is "education." Thus, the scatterplot in the first row, second column is for prestige (on the vertical axis) versus education (on the horizontal axis).

It is important to understand an essential limitation of the scatterplot matrix as a device for analyzing multivariate data: By projecting the multidimensional point cloud onto pairs of axes, the plot focuses on the *marginal* relationships between the corresponding pairs of variables. The object of data analysis for several variables, however, is typically to investigate *partial* relationships (between pairs of variables, "controlling" statistically for other variables), not marginal associations. For example, in the Duncan data set, we are more interested in the partial relationship of prestige to education holding income constant than in the marginal relationship between prestige and education ignoring income.

The response variable Y can be related marginally to a particular X, even when there is no partial relationship between the two variables controlling for other Xs. It is also possible for there to be a partial association between Y and an X but no marginal association. Furthermore, if the Xs themselves are nonlinearly related, then the marginal relationship between Y and a specific X can be nonlinear even when their partial relationship is linear.[28]

Despite this intrinsic limitation, scatterplot matrices often uncover interesting features of the data, and this is indeed the case in Figure 3.17, where the display reveals three unusual observations: *Ministers* have relatively low income for their relatively high level of education, and relatively high prestige for their relatively low income; *railroad conductors* and *railroad engineers* have relatively high incomes for their more-or-less average levels of education; *railroad conductors* also have relatively low prestige for their relatively high incomes. This pattern bodes ill for the least-squares linear regression of prestige on income and education.[29]

3.3.2 Coded Scatterplots

Information about a categorical third variable can be entered on a bivariate scatterplot by coding the plotting symbols. The most effective codes use different colors to represent categories, but degrees of fill, distinguishable shapes, and distinguishable letters can also be effective.[30]

Figure 3.18 shows a scatterplot of Davis's data on measured and reported weight.[31] Observations for men are displayed as Ms, for women as Fs. Except for the outlying point (number 12—which, recall, represents an error in the data), the points both for men and for women cluster near the line $Y = X$; it is also clear from the display that most men are heavier than most women, as one would expect, and that, discounting the bad data point, one man (number 21) is quite a bit heavier than everyone else.

3.3.3 Three-Dimensional Scatterplots

Another useful multivariate display, directly applicable to three variables at a time, is the *three-dimensional scatterplot*. Moreover, just as data can be projected onto a judiciously chosen plane

[27] We will return to this regression problem in Chapter 5.

[28] These ideas are explored in Chapter 12.

[29] See the discussion of Duncan's occupational-prestige regression in Chapter 11.

[30] See Spence and Lewandowsky (1990) for a fine review of the literature on graphical perception, including information on coded scatterplots.

[31] Davis's data were introduced in Chapter 2, where only the data for women were presented.

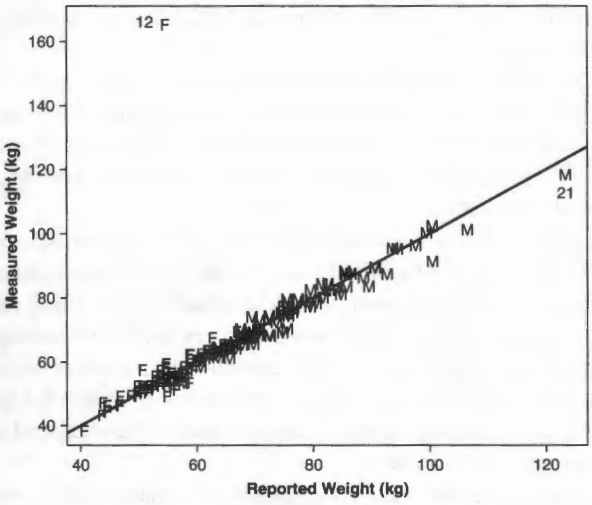

Figure 3.18 Davis's data on measured and reported weight, by gender. Data points for men are represented by Ms, for women by Fs, and are jittered slightly to reduce overplotting. The line on the graph is $Y = X$. In the combined data set for men and women, the outlying observation is number 12.

in a two-dimensional plot, higher-dimensional data can be projected onto a three-dimensional space, expanding the range of application of three-dimensional scatterplots.[32]

Barring the use of a true stereoscopic display, the three-dimensional scatterplot is an illusion produced by modern statistical software: The graph represents a projection of a three-dimensional space onto a two-dimensional computer screen. Nevertheless, motion (e.g., rotation) and the ability to interact with the display—possibly combined with the effective use of perspective, color, depth cueing, and other visual devices—can produce a vivid impression of directly examining objects in three-dimensional space.

It is literally impossible to convey this impression adequately on the static, two-dimensional page of a book, but Figure 3.19 shows Duncan's prestige data rotated interactively into a revealing orientation: Looking down the cigar-shaped scatter of most of the data, the three unusual observations stand out very clearly.

3.3.4 Conditioning Plots

Conditioning plots (or *coplots*), described in Cleveland (1993), are another graphical device for examining multidimensional data. The essential idea of the coplot is to focus on the relationship between the response variable and a particular explanatory variable, dividing the data into groups based on the values of other explanatory variables—the *conditioning variables*. If the conditioning variables are discrete, then this division is straightforward and natural. If a conditioning variable is continuous, it can be binned: Cleveland suggests using overlapping bins, which are called "shingles."

An illustrative coplot, for the General Social Survey vocabulary data, is shown in Figure 3.20. This graph displays the relationship between vocabulary score and education "controlling for"

[32]For example, there are three-dimensional versions of the added-variable and component-plus-residual plots discussed in Chapters 11 and 12. See, for example, Cook and Weisberg (1989).

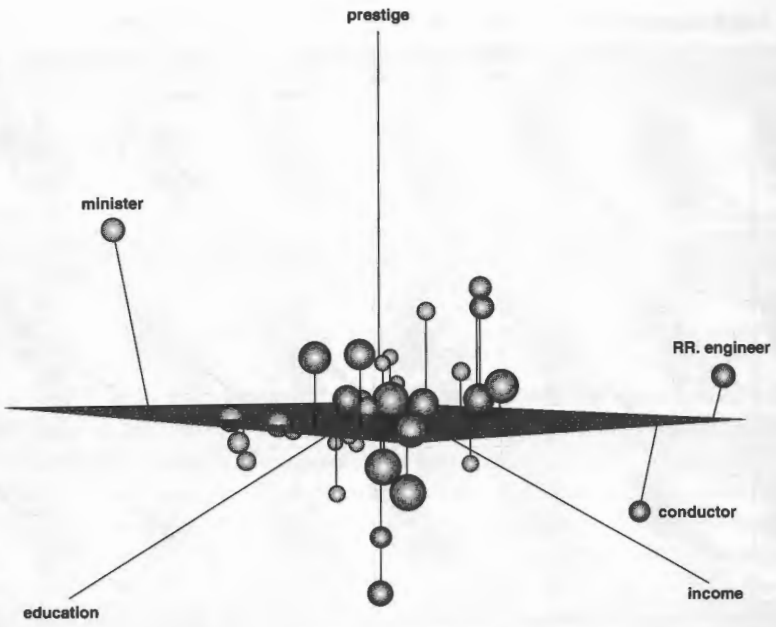

Figure 3.19 Three-dimensional scatterplot for Duncan's occupational prestige data, rotated into an orientation that reveals three unusual observations. From this orientation, the least-squares regression plane, also shown in the plot, is viewed nearly edge on.

gender and the year of the survey. The partial relationships are remarkably similar in the different panels of the coplot; that is, gender and year appear to make little difference to the relationship between vocabulary score and education. The relationships also appear to be very close to linear: In a few panels, the lowess line departs from the linear least-square line at the far left, but data in this region are quite sparse.

Although they can be effective graphs, coplots have limitations: First, if there are more than two, or perhaps three, conditioning variables, it becomes difficult to perceive how the partial relationship between the response and the focal explanatory variable changes with the conditioning variables. Second, because coplots require the division of the data into groups, they are most useful for large data sets, an issue that grows more acute as the number of conditioning variables increases.

Visualizing multivariate data is intrinsically difficult because we cannot directly examine higher-dimensional scatterplots. Effective displays project the higher-dimensional point cloud onto two or three dimensions; these displays include the scatterplot matrix, the dynamic three-dimensional scatterplot, and the conditioning plot.

Summary

- Statistical graphs are central to effective data analysis, both in the early stages of an investigation and in statistical modeling.

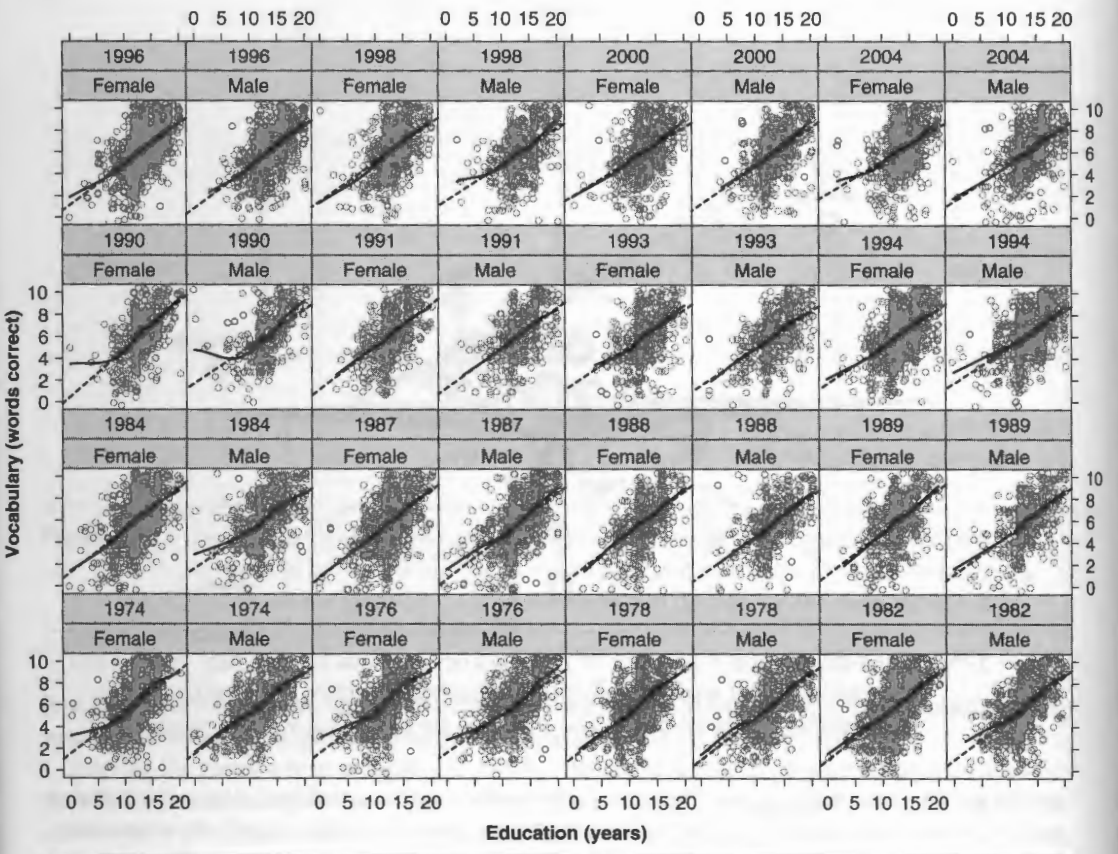

Figure 3.20 Coplot showing the relationship between vocabulary score and education controlling for year and gender. The points in each panel are jittered to reduce over-plotting. The broken line shows the linear least-square fit, while the solid line gives the lowess fit for a span of 0.6.

- There are many useful univariate displays, including the traditional histogram. The stem-and-leaf plot is a modern variant of the histogram for small data sets, constructed directly from numerical data. Nonparametric density estimation may be employed to smooth a histogram. Quantile-comparison plots are useful for comparing data with a theoretical probability distribution. Boxplots summarize some of the most important characteristics of a distribution, including center, spread, skewness, and outliers.
- The bivariate scatterplot is a natural graphical display of the relationship between two quantitative variables. Interpretation of a scatterplot can often be assisted by graphing a nonparametric regression, which summarizes the relationship between the two variables. Scatterplots of the relationship between discrete variables can be enhanced by randomly jittering the data.
- Parallel boxplots display the relationship between a quantitative response variable and a discrete explanatory variable.
- Visualizing multivariate data is intrinsically difficult because we cannot directly examine higher-dimensional scatterplots. Effective displays project the higher-dimensional point

cloud onto two or three dimensions; these displays include the scatterplot matrix, the dynamic three-dimensional scatterplot, and the conditioning plot.

Recommended Reading

The literature—especially the recent literature—on statistical graphics is truly voluminous. I will furnish only the briefest of bibliographies:

- Fox (2000c) presents a brief overview of statistical graphics, including information on the history of the subject. Jacoby (1997, 1998) gives a more extended overview addressed to social scientists.
- Tufte's (1983) influential book on graphical presentation of quantitative information is opinionated but well worth reading. (Tufte has since published several other books on graphics, broadly construed, but I prefer his first book.)
- Modern interest in statistical graphics is the direct result of John Tukey's work on exploratory data analysis; unfortunately, Tukey's idiosyncratic writing style makes his seminal book (Tukey, 1977) difficult to read. Velleman and Hoaglin (1981) provide a more digestible introduction to the topic. There is interesting information on the statistical theory underlying exploratory data analysis in two volumes edited by Hoaglin, Mosteller, and Tukey (1983, 1985).
- Tukey's influence made Bell Labs a center of work on statistical graphics, much of which is described in two accessible and interesting books by William Cleveland (1993, 1994) and in Chambers, Cleveland, Kleiner, and Tukey (1983). Cleveland (1994) is a good place to start.
- Modern statistical graphics is closely associated with advances in statistical computing: The S statistical computing environment (Becker, Chambers, & Wilks, 1988; Chambers, 1998; Chambers & Hastie, 1992), also a product of Bell Labs, is particularly strong in its graphical capabilities. R, a free, open-source implementation of S, was mentioned in the preface. Cook and Weisberg (1994, 1999) use the Lisp-Stat statistical computing environment (Tierney, 1990) to produce an impressive statistical package, called Arc, which incorporates a variety of statistical graphics of particular relevance to regression analysis (including many of the methods described later in this text). Friendly (1991) describes how to construct modern statistical graphs using the SAS/Graph system. Brief presentations of these and other statistical computing environments appear in a book edited by Stine and Fox (1996).
- Atkinson (1985) presents a variety of innovative graphs in support of regression analysis, as do Cook (1998) and Cook and Weisberg (1994, 1999).