- 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 fc
lowess smoo
levels of GDI

- Scatterplots in whic
because the bulk of
ple, the scatterplot
Figure 3.13. It often
between Y and X.[23]
- Scatterplots in whicl
instance of this phe
vocabulary test agai
Surveys conducted
and include in total
discrete—is to focu
for example, can be
discussed below). A
random quantity to
random variable on
Paradoxically, the te
randomly "jittered"

The bivariate scatterp
quantitative variables.
nonparametric regress
Scatterplots of the rel
jittering the data.

[23] See Chapter 4.

[24] The idea of jittering a scatter

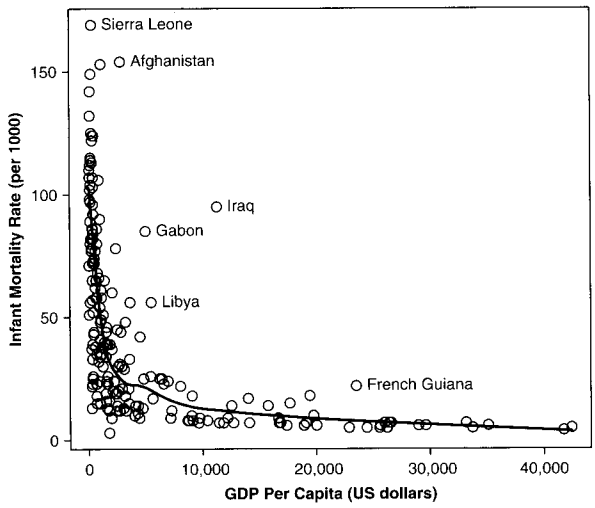


Figure 3.13 Scatterplot for infant mortality and GDP per capita for 193 nations. The line is for a lowess smooth with a span of 1/2. Several nations with high infant mortality for their levels of GDP are identified.

- 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).

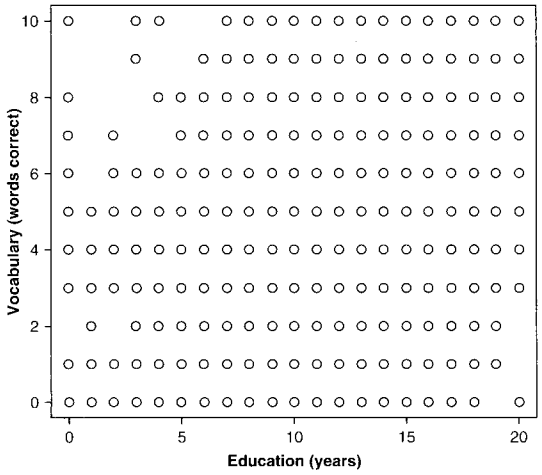


Figure 3.14 Scatterplot of scores on a 10-item vocabulary test versus years of education. Although there are nearly 22,000 observations in the data set, most of the plotted points fall on top of one another.

SOURCE: National Opinion Research Center (2005).

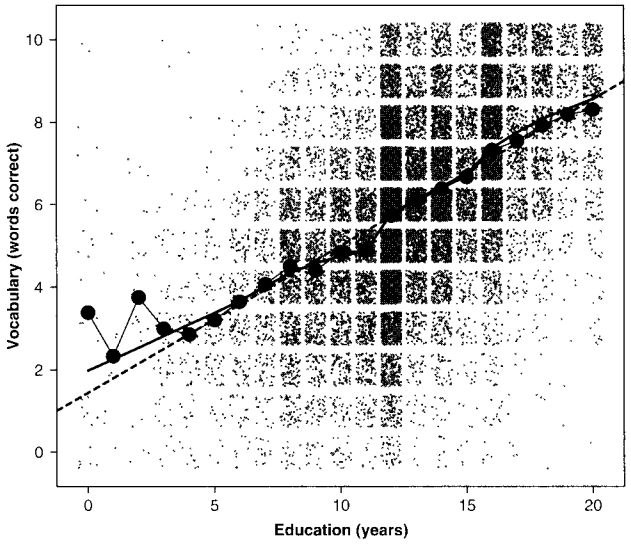


Figure 3.15 Jittered scatterplot for vocabulary score versus years of education. A uniformly distributed random quantity between −0.4 and +0.4 was added to each score for both variables. The heavier solid line is for a lowess fit to the data, with a span of 0.2; the broken line is the linear least-squares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

SOURCE: Personal commu

variable in this graph is tl
by each firm with other
corporation is controlle

It is apparent from t
foreign and Canadian c
and the United States.
the conditional distribut
association between leve
firms than among U.K.

Parallel boxplots dis
discrete (categorical

3.3 Plotting M

Because paper and cor
data is intrinsically diffi
dimensional "point clou
to view a higher-dimens
case below). The essenti
reveal important charact
on the basis of a statistic

[25] We will revisit this example
outliers in the plot.

[26] We will apply these powerfu

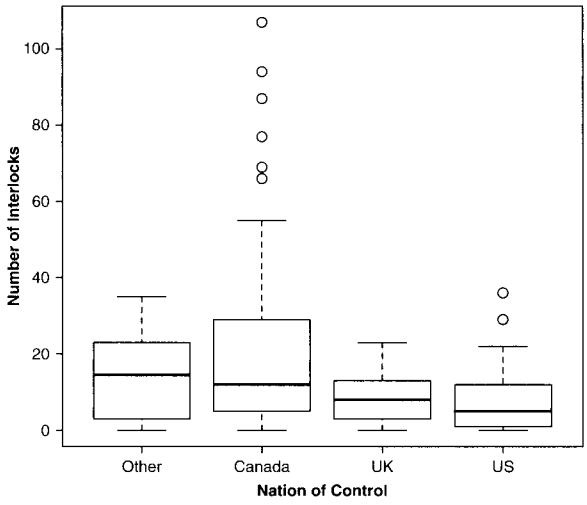


Figure 3.16 Number of interlocking directorate and executive positions by nation of control, for 248 dominant Canadian firms.

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.

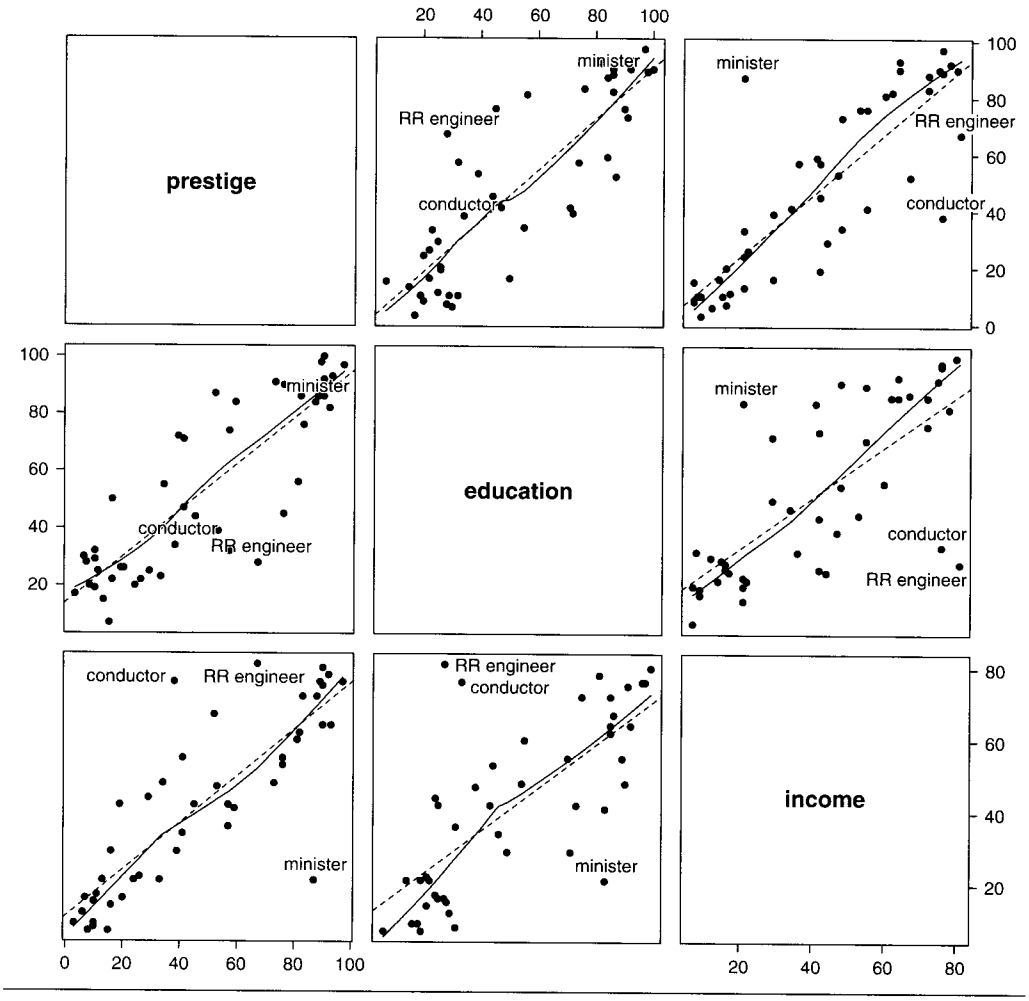


Figure 3.17 Scatterplot matrix for occupational prestige, level of education, and level of income, for 45 U.S. occupations in 1950. The least-squares regression line (broken line) and lowess smooth (for a span of 0.6, solid line) are shown on each plot. Three unusual observations are identified.

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 whic
were no direct prestige ratir

The variable names on th
columns of the display: Fc
display is "prestige"; the hc
Thus, the scatterplot in the
education (on the horizonta

It is important to unders
analyzing multivariate data:
the plot focuses on the *marg*
object of data analysis for s
ships (between pairs of var
associations. For example, i
ship of prestige to educatior
prestige and education igno

The response variable Y
partial relationship between
to be a partial association b
Xs themselves are nonlinear
can be nonlinear even when

Despite this intrinsic lim
data, and this is indeed the
vations: *Ministers* have rel
relatively high prestige for
neers have relatively high i
conductors also have relativ
ill for the least-squares linea

3.3.2 Coded Scatte

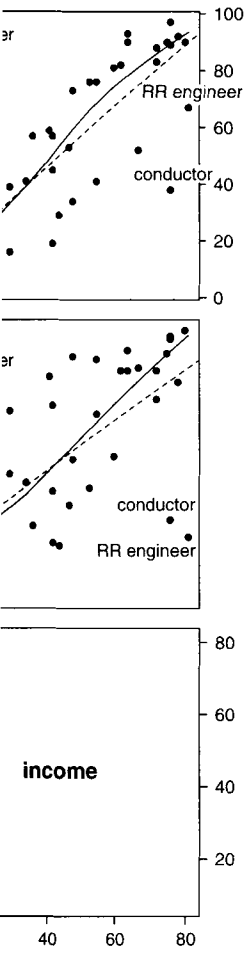
Information about a categ
the plotting symbols. The n
degrees of fill, distinguishab

Figure 3.18 shows a scat
vations for men are display
12—which, recall, represent
near the line $Y = X$; it is als
as one would expect, and tha
heavier than everyone else.

3.3.3 Three-Dimen

Another useful multivaria
dimensional scatterplot. Mc

[27] We will return to this regression
[28] These ideas are explored in Chap
[29] See the discussion of Duncan's o
[30] See Spence and Lewandowsky (1
on coded scatterplots.
[31] Davis's data were introduced in C



ion, and level of income,
on line (broken line) and
each plot. Three unusual

stical model, is to examine
catterplot matrix produces

on, and income levels of
lyzed by Duncan (1961),
d an occupation as "good"
mbents in the occupation
represents the percentage
Duncan's purpose was to
dict 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.

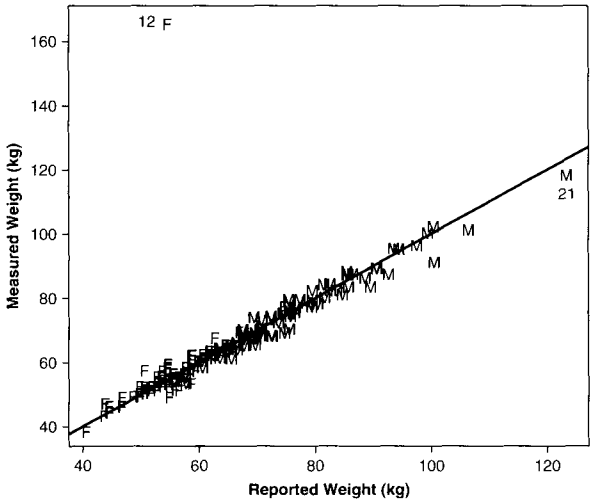


Figure 3.18 Davis's data on measured and reported weight, by gender. Data points for men are represented by Ms, for women by Fs, and are jittered slightly to reduce overplotting. The line on the graph is $Y = X$. In the combined data set for men and women, the outlying observation is number 12.

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).

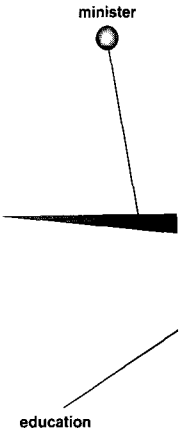


Figure 3.19 Three-dimension
an orientation th
least-squares reg

gender and the year of the surv
panels of the coplot; that is, g
between vocabulary score and
In a few panels, the lowess lin
this region are quite sparse.

Although they can be effec
two, or perhaps three, condit
relationship between the respo
variables. Second, because cop
for large data sets, an issue that

Visualizing multivariate d
higher-dimensional scatte
cloud onto two or three
dynamic three-dimensiona

Summary

- Statistical graphs are ce
tigation and in statistica

Data points for men are
tly to reduce overplotting.
r men and women, the

l onto a three-dimensional
rplots.[32]
al scatterplot is an illusion
tion of a three-dimensional
e.g., rotation) and the ability
of perspective, color, depth
directly examining objects

the static, two-dimensional
l interactively into a reveal-
the data, the three unusual

another graphical device for
to focus on the relationship
ividing the data into groups
ariables. If the conditioning
l. If a conditioning variable
ping bins, which are called

ata, is shown in Figure 3.20.
education "controlling for"

onent-plus-residual plots discussed

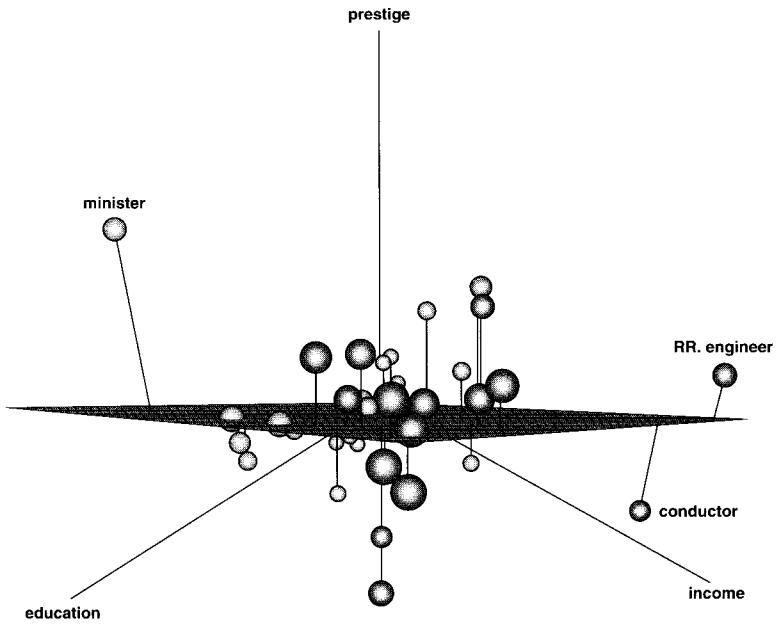


Figure 3.19 Three-dimensional scatterplot for Duncan's occupational prestige data, rotated into an orientation that reveals three unusual observations. From this orientation, the least-squares regression plane, also shown in the plot, is viewed nearly edge on.

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.

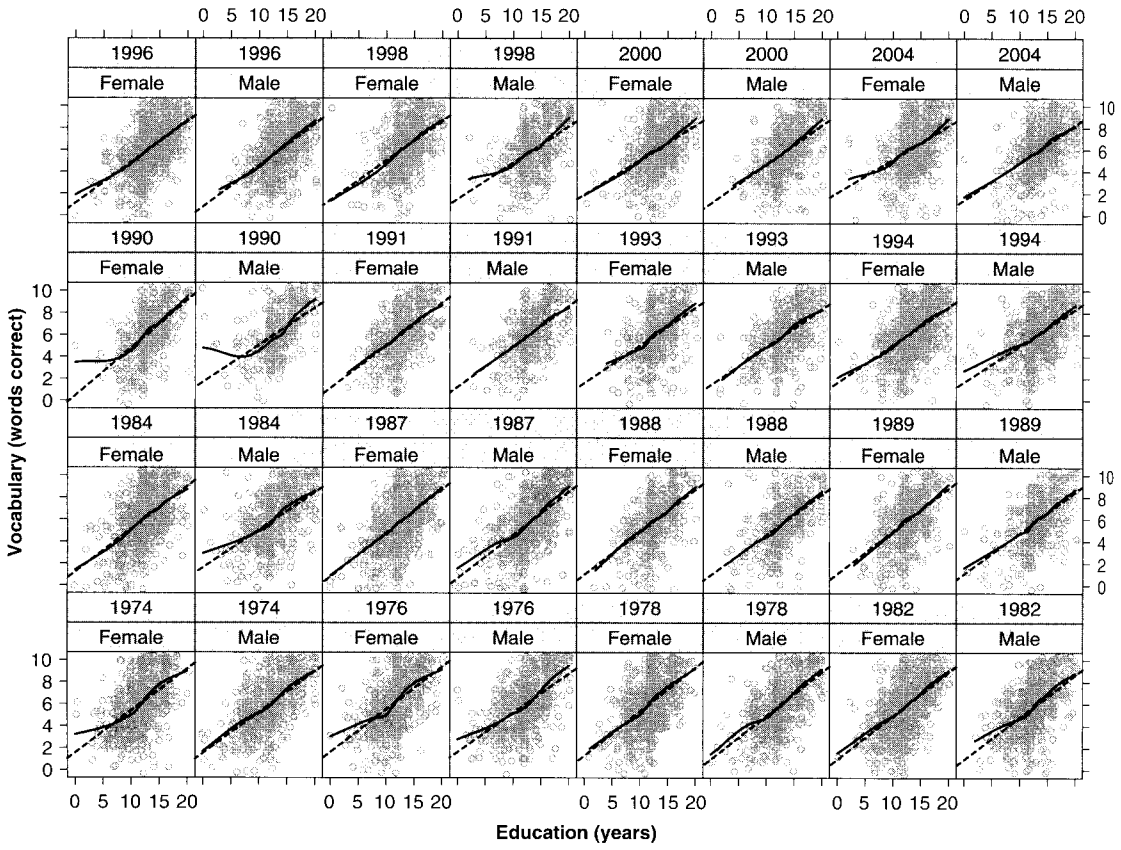


Figure 3.20 Coplot showing the relationship between vocabulary score and education controlling for year and gender. The points in each panel are jittered to reduce over-plotting. The broken line shows the linear least-square fit, while the solid line gives the lowess fit for a span of 0.6.

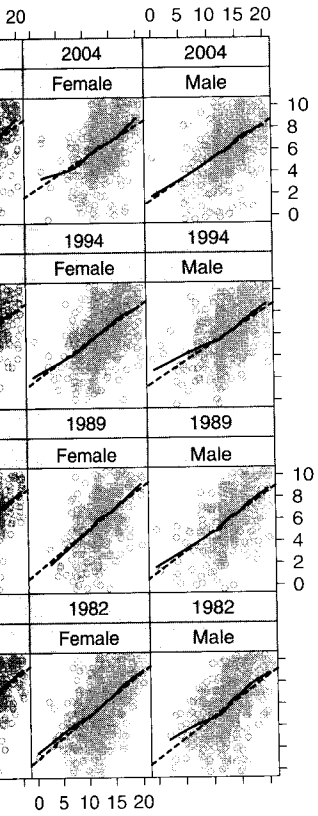
- There are many useful univariate displays, including the traditional histogram. The stem-and-leaf plot is a modern variant of the histogram for small data sets, constructed directly from numerical data. Nonparametric density estimation may be employed to smooth a histogram. Quantile-comparison plots 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 thre
dynamic three-dimensio

Recommended Re

The literature—especially the
furnish only the briefest of bit

- Fox (2000c) presents a l
history of the subject. J
social scientists.
- Tufte's (1983) influenti
opinionated but well w
graphics, broadly constr
- Modern interest in statis
data analysis; unfortun
(Tukey, 1977) difficult
introduction to the topic
exploratory data analys
1985).
- Tukey's influence made
described in two access
in Chambers, Clevelan
start.
- Modern statistical grap
The S statistical compu
1998; Chambers & Ha
graphical capabilities. F
ace. Cook and Weisber
(Tierney, 1990) to prod
a variety of statistical g
of the methods describ
ern statistical graphs u
statistical computing e
- Atkinson (1985) prese
as do Cook (1998) and



e and education
are jittered to reduce
e fit, while the solid line

onal histogram. The stem-
a sets, constructed directly
employed to smooth a his-
ata with a theoretical prob-
portant characteristics of a

relationship between two
be assisted by graphing a
between the two variables.
be enhanced by randomly

ve response variable and a

ve cannot directly examine
e 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).