

2

Looking at Multivariate Data

2.1 Introduction

Most of the chapters in this book are concerned with methods for the analysis of multivariate data, which are based on relatively complex mathematics. This chapter, however, is not. Here we look at some relatively simple graphical procedures and there is no better software for producing graphs than R and S-PLUS[®].

According to Chambers et al. (1983) “there is no statistical tool that is as powerful as a well-chosen graph.” Certainly graphical presentation has a number of advantages over tabular displays of numerical results, not the least of which is creating interest and attracting the attention of the viewer. Graphs are very popular. It has been estimated that between 900 billion (9×10^{11}) and 2 trillion (2×10^{12}) images of statistical graphics are printed each year. Perhaps one of the main reasons for such popularity is that graphical presentation of data often provides the vehicle for discovering the unexpected; the human visual system is very powerful in detecting patterns, although the following caveat from the late Carl Sagan should be kept in mind.

Humans are good at discerning subtle patterns that are really there, but equally so at imagining them when they are altogether absent.

During the last two decades a wide variety of new methods for displaying data graphically have been developed. These will hunt for special effects in data, indicate outliers, identify patterns, diagnose models and generally search for novel and perhaps unexpected phenomena. Large numbers of graphs may be required, and computers are generally needed to generate them for the same reasons they are used for numerical analyses, namely, they are fast and they are accurate.

So, because the machine is doing the work, the question is no longer “Shall we plot?” but rather “What shall we plot?” There are many exciting possibilities including, dynamic graphics (see Cleveland and McGill, 1987), but graphical exploration of data usually begins with some simpler, well-known methods. Univariate marginal views of multivariate data might, for example, be obtained using *histograms*, *stem-and-leaf plots*, or *box plots*. More important for exploring multivariate data are plots that allow the relationships between variables to be assessed. Consequently we begin our discussion of graphics with the ubiquitous scatterplot.

2.2 Scatterplots and Beyond

The simple xy scatterplot has been in used since at least the eighteenth century and has many virtues. Indeed, according to Tufte (1983):

The relational graphic—in its barest form the scatterplot and its variants—is the greatest of all graphical designs. It links at least two variables, encouraging and even imploring the viewer to assess the possible causal relationship between the plotted variables. It confronts causal theories that x causes y with empirical evidence as to the actual relationship between x and y .

To illustrate the use of the scatterplot and the other techniques to be discussed in subsequent sections we shall use the data shown in Table 2.1. These data give

Table 2.1 Air Pollution Data for Regions in the United States

Region	Rainfall	Educ	Popden	Nonwhite	NOX	SO2	Mortality
AkronOH	36	11.4	3243	8.8	15	59	921.9
AlbanyNY	35	11.0	4281	3.5	10	39	997.9
AllenPA	44	9.8	4260	0.8	6	33	962.4
AtlantGA	47	11.1	3125	27.1	8	24	982.3
BaltimMD	43	9.6	6441	24.4	38	206	1071.0
BirmhmAL	53	10.2	3325	38.5	32	72	1030.0
BostonMA	43	12.1	4679	3.5	32	62	934.7
BridgeCT	45	10.6	2140	5.3	4	4	899.5
BufaloNY	36	10.5	6582	8.1	12	37	1002.0
CantonOH	36	10.7	4213	6.7	7	20	912.3
ChatagTN	52	9.6	2302	22.2	8	27	1018.0
ChicagIL	33	10.9	6122	16.3	63	278	1025.0
CinnciOH	40	10.2	4101	13.0	26	146	970.5
ClevehOH	35	11.1	3042	14.7	21	64	986.0
ColombOH	37	11.9	4259	13.1	9	15	958.8
DallasTX	35	11.8	1441	14.8	1	1	860.1
DaytonOH	36	11.4	4029	12.4	4	16	936.2
DenverCO	15	12.2	4824	4.7	8	28	871.8
DetroitMI	31	10.8	4834	15.8	35	124	959.2
FlintMI	30	10.8	3694	13.1	4	11	941.2
FtwortTX	31	11.4	1844	11.5	1	1	891.7
GrndraMI	31	10.9	3226	5.1	3	10	871.3
GrnborNC	42	10.4	2269	22.7	3	5	971.1
HartfdCT	43	11.5	2909	7.2	3	10	887.5

(Continued)

Table 2.1 (*Continued*)

Region	Rainfall	Educ	Popden	Nonwhite	NOX	SO2	Mortality
HoustonTX	46	11.4	2647	21.0	5	1	952.5
IndianIN	39	11.4	4412	15.6	7	33	968.7
KansasMO	35	12.0	3262	12.6	4	4	919.7
LancasPA	43	9.5	3214	2.9	7	32	844.1
LosangCA	11	12.1	4700	7.8	319	130	861.8
LouisvKY	30	9.9	4474	13.1	37	193	989.3
MemphsTN	50	10.4	3497	36.7	18	34	1006.0
MiamiFL	60	11.5	4657	13.5	1	1	861.4
MilwauWI	30	11.1	2934	5.8	23	125	929.2
MinnplMN	25	12.1	2095	2.0	11	26	857.6
NashvlTN	45	10.1	2082	21.0	14	78	961.0
NewhvnCT	46	11.3	3327	8.8	3	8	923.2
NeworIL	54	9.7	3172	31.4	17	1	1113.0
NewyrkNY	42	10.7	7462	11.3	26	108	994.6
PhiladPA	42	10.5	6092	17.5	32	161	1015.0
PittsbPA	36	10.6	3437	8.1	59	263	991.3
PortldOR	37	12.0	3387	3.6	21	44	894.0
ProvdRI	42	10.1	3508	2.2	4	18	938.5
ReadngPA	41	9.6	4843	2.7	11	89	946.2
RichmdVA	44	11.0	3768	28.6	9	48	1026.0
RochtrNY	32	11.1	4355	5.0	4	18	874.3
StLousMO	34	9.7	5160	17.2	15	68	953.6
SandigCA	10	12.1	3033	5.9	66	20	839.7
SanFranCA	18	12.2	4253	13.7	171	86	911.7
SanJosCA	13	12.2	2702	3.0	32	3	790.7
SeatlWA	35	12.2	3626	5.7	7	20	899.3
SpringMA	45	11.1	1883	3.4	4	20	904.2
SyracuNY	38	11.4	4923	3.8	5	25	950.7
ToledoOH	31	10.7	3249	9.5	7	25	972.5
UticaNY	40	10.3	1671	2.5	2	11	912.2
WashDC	41	12.3	5308	25.9	28	102	968.8
WichtaKS	28	12.1	3665	7.5	2	1	823.8
WilmtnDE	45	11.3	3152	12.1	11	42	1004.0
WorctrMA	45	11.1	3678	1.0	3	8	895.7
YorkPA	42	9.0	9699	4.8	8	49	911.8
YoungsOH	38	10.7	3451	11.7	13	39	954.4

Data assumed available as dataframe `airpoll` with variable names as indicated.

information on 60 U.S. metropolitan areas (McDonald and Schwing, 1973; Henderson and Velleman, 1981). For each area the following variables have been recorded:

1. *Rainfall*: mean annual precipitation in inches
2. *Education*: median school years completed for those over 25 in 1960
3. *Popden*: population/mile² in urbanized area in 1960
4. *Nonwhite*: percentage of urban area population that is nonwhite
5. *NOX*: relative pollution potential of oxides of nitrogen
6. *SO2*: relative pollution potential of sulphur dioxide
7. *Mortality*: total age-adjusted mortality rate, expressed as deaths per 100,000

One of the questions about these data might be “How is sulphur dioxide pollution related to mortality?” A first step in answering the question would be to examine a scatterplot of the two variables. Here, in fact, we will produce four versions of the basic scatterplot using the following R and S-PLUS code (we assume that the data are available as the data frame `airpoll` with variable names as above):

```
attach(airpoll)
#set up plotting area to take four graphs
par(mfrow=c(2,2,))
par(pty="s")
plot(SO2,Mortality,pch=1,lwd=2)
title(" (a) ",lwd=2)
plot(SO2,Mortality,pch=1,lwd=2)
#add regression line
abline(lm(Mortality~SO2),lwd=2)
title(" (b) ",lwd=2)
#jitter data
airpoll1<-jitter(cbind(SO2,Mortality))
plot(airpoll1[,1],airpoll1[,2],xlab="SO2",ylab="Mortality",
     pch=1,lwd=2)
title(" (c) ",lwd=2)
plot(SO2,Mortality,pch=1,lwd=2)
#add rug plots
rug(jitter(SO2),side=1)
rug(jitter(Mortality),side=2)
title(" (d) ",lwd=2)
```

Figure 2.1(a) shows the scatterplot of *Mortality* against *SO2*. Figure 2.1(b) shows the same scatterplot with the addition of the simple linear regression fit of *Mortality* on *SO2*. Both plots suggest a possible link between increasing sulphur dioxide level and increasing mortality.

Although not a real problem here, scatterplots in which there are many points often suffer from overplotting. The problem can be overcome, partially at least, by “jittering” the data, that is, adding a small amount of noise to each observation

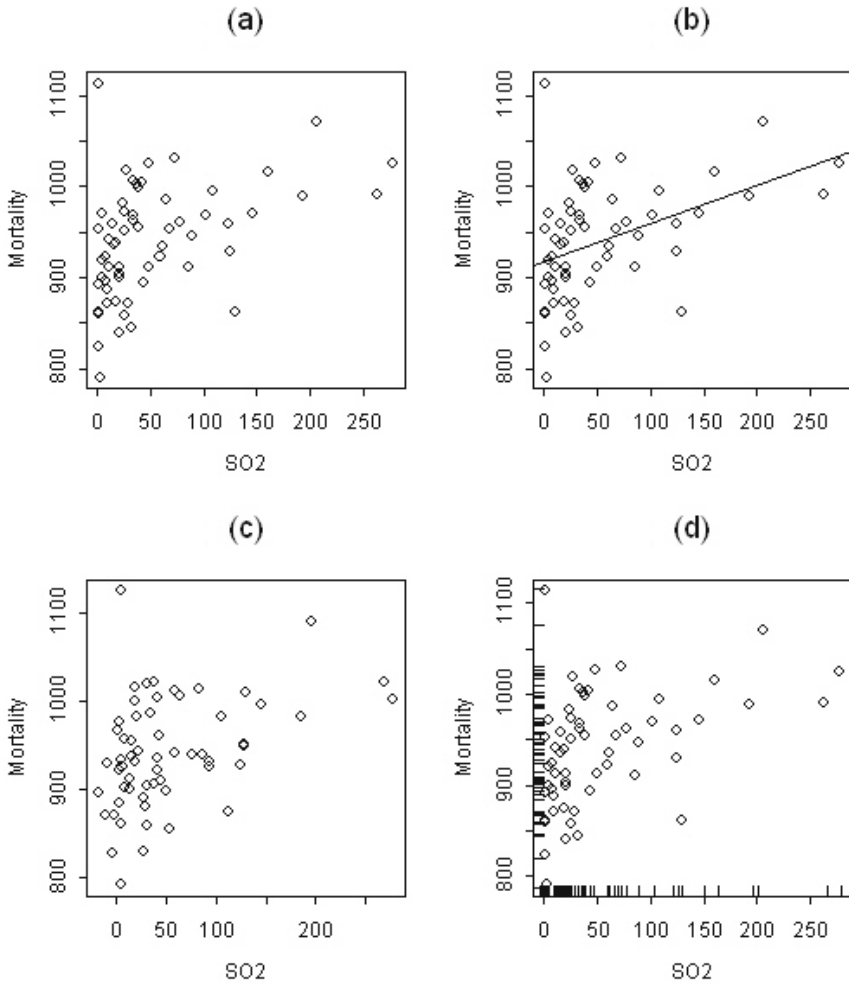


Figure 2.1 (a) Scatterplot for Mortality against SO₂; (b) scatterplot of Mortality against SO₂ with added linear regression fit; (c) jittered scatterplot of Mortality against SO₂; (d) scatterplot of Mortality against SO₂ with information about marginal distributions of the two variables added.

before plotting (see Chambers et al., 1983, for details). Figure 2.1(c) shows the scatterplot in Figure 2.1(a) after jittering. Finally, in Figure 2.1(d), the bivariate scatter of the two variables is framed with a display of the marginal distribution of each variable. Plotting marginal and joint distributions together is usually good data analysis practice.

With these data it might be useful to label the scatterplot with the names of the regions involved. These names are rather long, and if used as they are would

lead to a rather “messy” plot; consequently we shall use the R and S-PLUS function `abbreviate` to shorten them before plotting using the code:

```
names<-abbreviate(row.names(airpoll))
plot(SO2,Mortality,lwd=2,type="n")
text(SO2,Mortality,labels=names,lwd=2)
```

Figure 2.2 highlights some regions with odd combinations of pollution and mortality values. For example, *nwLA* has almost zero SO_2 value, but very high mortality. Perhaps this is a garden suburb where people go to retire?

In Figure 2.1(b) a simple linear regression fit was added to the *Mortality/SO₂* scatterplot. This addition is often very useful for assessing the relationship between the two variables more accurately. Even more useful is to add both the linear regression fit and a *locally weighted regression* or *lowest fit* to the scatterplot. Such fits are described in detail in Cleveland (1979), but essentially they are designed to use the data themselves to suggest the type of fit needed. The model assumed is that

$$y_i = g(x_i) + \epsilon_i,$$

where g is a “smooth” function and the ϵ_i are random variables with zero mean and constant variance. Fitted values, \hat{y}_i , are used to estimate $g(x_i)$ at each x_i by fitting polynomials using weighted least squares, with large weights for points close to

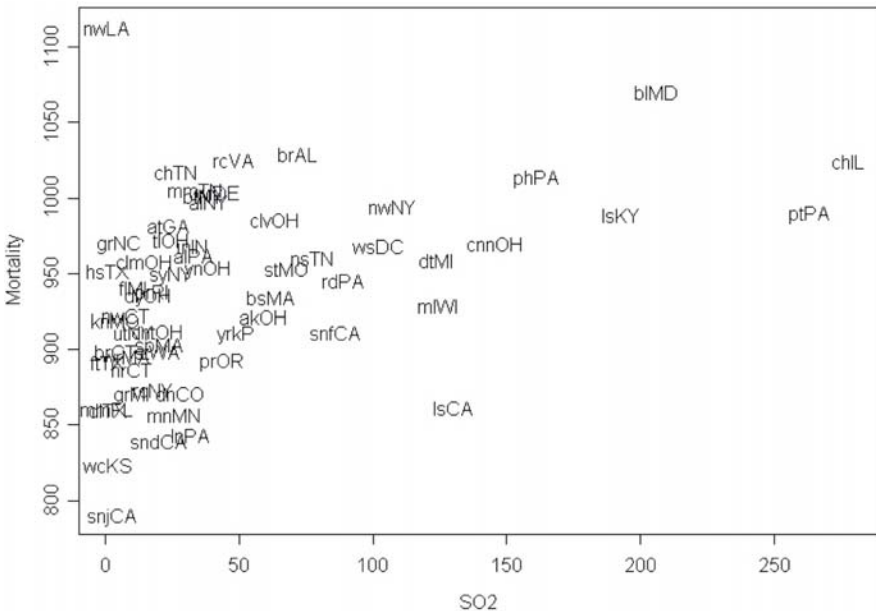


Figure 2.2 Scatterplot of mortality against SO_2 with points labeled by region name.

x_i and small weights otherwise. The degree of “smoothness” of the fitted curve can be controlled by a particular parameter during the fitting process. Examining a scatterplot that includes a locally weighted regression fit can often be a useful antidote to the thoughtless fitting of straight lines with least squares.

To illustrate the use of lowest fits we return to the air pollution data and again concentrate on the two variables, *SO2* and *Mortality*. The following R and S-PLUS code produces a scatterplot with some information about marginal distributions that also includes both a linear regression and a locally weighted regression fit:

```
#set up plotting area for scatterplot
par(fig=c(0,0.7,0,0.7))
plot(SO2,Mortality,lwd=2)
#add regression line
abline(lm(Mortality~SO2),lwd=2)
#add locally weighted regression fit
lines(lowess(SO2,Mortality),lwd=2)
#set up plotting area for histogram
par(fig=c(0,0.7,0.65,1),new=TRUE)
hist(SO2,lwd=2)
#set up plotting area for boxplot
par(fig=c(0.65,1,0,0.7),new=TRUE)
boxplot(Mortality,lwd=2)
```

The resulting diagram is shown in Figure 2.3. Here, apart from a small “wobble” for sulphur dioxide values 0 to 100, the linear fit and the locally weighted fit are very similar.

2.2.1 The Convex Hull of Bivariate Data

Scatterplots are often used in association with the calculation of the correlation coefficient of two variables. Outliers, for example, can often considerably distort the value of a correlation coefficient, and a scatterplot may help to identify the offending observations, which might then be excluded from the calculation. Another approach that allows *robust estimation* of the correlation is *convex hull trimming*. The convex hull of a set of bivariate observations consists of the vertices of the smallest convex polyhedron in variable space within which, or on which, all data points lie. Removal of the points lying on the convex hull can eliminate isolated outliers without disturbing the general shape of the bivariate distribution. A robust estimate of the correlation coefficient results from using the remaining observations.

Let’s see how the convex hull approach works with our *Mortality/SO2* scatterplot. We can calculate the correlation coefficient of the two variables using all the observations from the R and S-PLUS instruction:

```
cor(SO2, Mortality)
```

giving a value of 0.426.

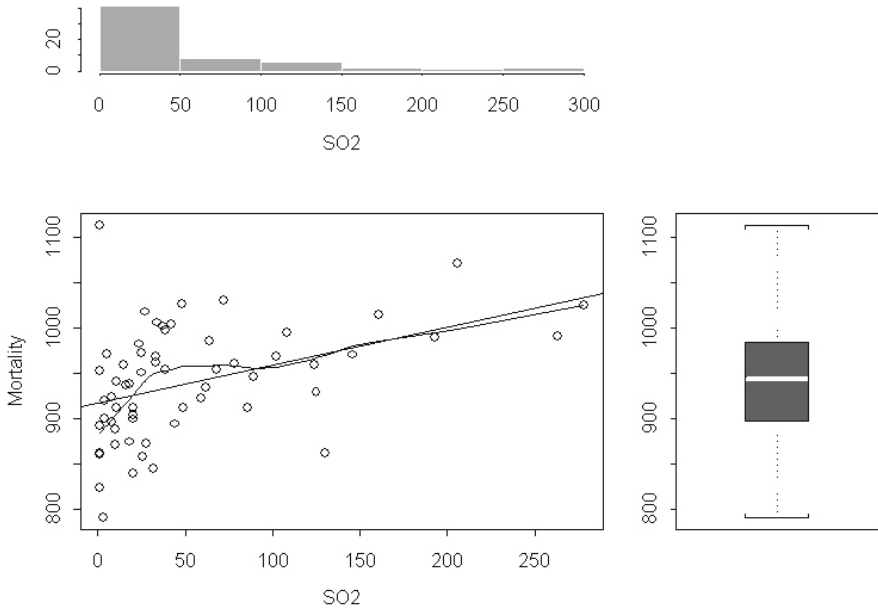


Figure 2.3 Scatterplot of mortality against SO_2 with added linear regression and locally weighted regression fits and marginal distribution information.

Now we can find the convex hull of the data and, for interest, show it on a scatterplot of the two variables using the following R and S-PLUS code:

```
#find points defining convex hull
hull<-chull(SO2,Mortality)
plot(SO2,Mortality,pch=1)
#plot and shade convex hull
polygon(SO2[hull],Mortality[hull],density=15,angle=30)
```

The result is shown in Figure 2.4.

To calculate the correlation coefficient after removal of the points defining the convex hull requires the instruction

```
cor(SO2[-hull],Mortality[-hull])
```

The resulting value of the correlation is now 0.438. In this case the change in the correlation after removal of the points defining the convex hull is very small, surprisingly small, given that some of the defining observations are relatively remote from the body of the data.

2.2.2 The Chiplot

Although the scatterplot is a primary data-analytic tool for assessing the relationship between a pair of continuous variables, it is often difficult to judge whether or not

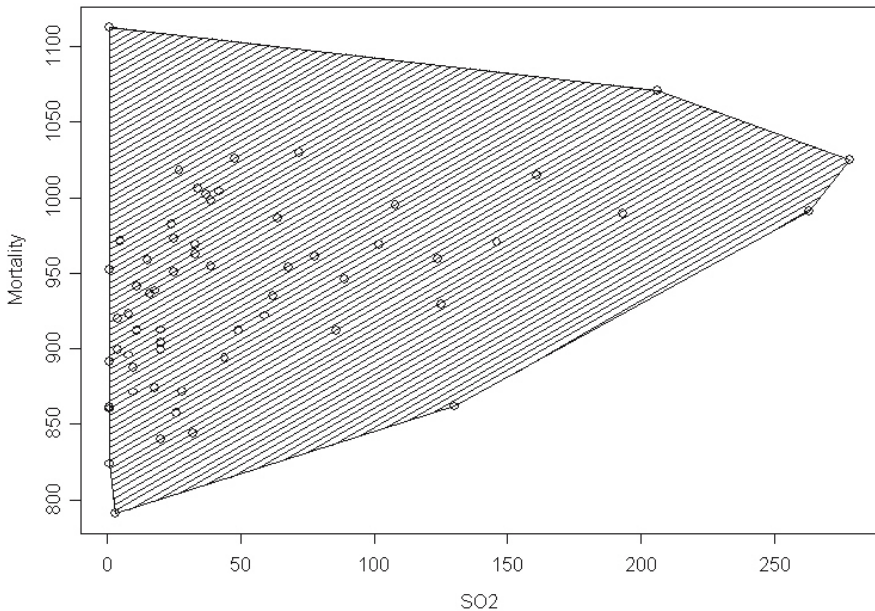


Figure 2.4 Scatterplot of mortality against SO_2 showing convex hull of the data.

the variables are independent. A random scatter of points may be hard for the human eye to judge. Consequently, it is often helpful to augment the scatterplot with an auxiliary display in which independence is itself manifested in a characteristic manner. The *chi-plot* suggested by Fisher and Switzer (1985, 2001) is designed to address the problem. The essentials of this type of plot are described in Display 2.1.

Display 2.1 The Chi-Plot

- A chi-plot is a scatterplot of the pairs

$$(\chi_i, \lambda_i), |\lambda_i| < 4 \left\{ \frac{1}{n-1} - \frac{1}{2} \right\}^2,$$

where

$$\chi_i = (H_i - F_i G_i) / \{F_i(1 - F_i)G_i(1 - G_i)\}^{1/2}$$

$$\lambda_i = 4S_i \max \left\{ \left(F_i - \frac{1}{2} \right)^2, \left(G_i - \frac{1}{2} \right)^2 \right\}$$

and

$$H_i = \sum_{j \neq i} I(x_j \leq x_i, y_j \leq y_i) / (n - 1)$$

$$F_i = \sum_{j \neq i} I(x_j \leq x_i) / (n - 1)$$

$$G_i = \sum_{j \neq i} I(y_j \leq y_i) / (n - 1)$$

$$S_i = \text{sign} \left\{ \left(F_i - \frac{1}{2} \right) \left(G_i - \frac{1}{2} \right) \right\}$$

where $\text{sign}(x)$ is $+1$ if x is positive, 0 if x is zero, and -1 if x is negative; $I(A)$ is the indicator function for the event A , that is, if A is true $I(A) = 1$, if A is not true, $I(A) = 0$.

- When the two variables are independent, the points in a chi-plot will be scattered about a central region. When they are related, the points will tend to lie outside this central region. See the example in the text.

An R and S-PLUS function for producing chi-plots, the `chiplot` is given on the website mentioned in the Preface. To illustrate the chi-plot we shall apply it to the *Mortality* and *SO2* variables of the air pollution data using the code

```
chiplot(SO2, Mortality, vlabs=c("SO2", "Mortality"))
```

The result is Figure 2.5 which shows the scatterplot of *Mortality* plotted against *SO2* alongside the corresponding chi-plot. Departure from independence is indicated in the latter by a lack of points in the horizontal band indicated on the plot. Here there is a clear departure since there are very few of the observations in this region.

2.2.3 The Bivariate Boxplot

A further helpful enhancement to the scatterplot is often provided by the two-dimensional analogue of the boxplot for univariate data, known as the *bivariate boxplot* (Goldberg and Iglewicz, 1992). This type of boxplot may be useful in indicating the distributional properties of the data and in identifying possible outliers. The bivariate boxplot is based on calculating “robust” measures of location, scale, and correlation. It consists essentially of a pair of concentric ellipses, one of which (the “hinge”) includes 50% of the data and the other (called the “fence”) which delineates potential troublesome outliers. In addition, resistant regression lines of both y on x and x on y are shown, with their intersection showing the bivariate location estimator. The acute angle between the regression lines will be small

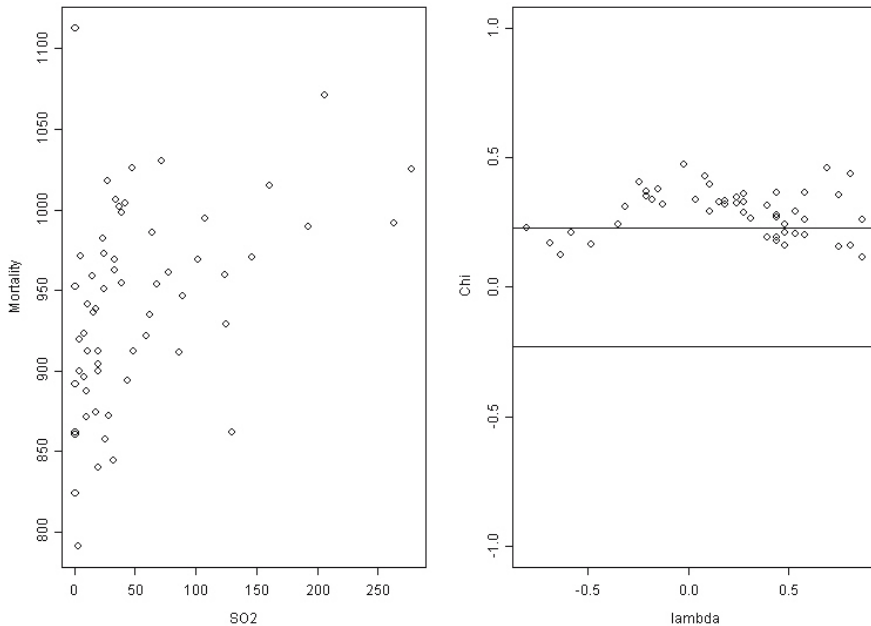


Figure 2.5 Chi-plot of *Mortality* and *SO2*.

for a large absolute value of correlations and large for a small one. Details of the construction of a bivariate boxplot are as given in Display 2.2:

Display 2.2 Constructing a Bivariate Boxplot

- The bivariate boxplot is the two-dimensional analogue of the familiar boxplot for univariate data and consists of a pair of concentric ellipses, the “hinge” and the “fence.”
- To draw the elliptical fence and hinge, location (T_x^*, T_y^*) , scale (S_x^*, S_y^*) , and correlation (R^*) estimators are needed, in addition to a constant D that regulates the distance of the fence from the hinge. In general $D = 7$ is recommended since this corresponds to an approximate 99% confidence bound on a single observation.
- In general, robust estimators of location, scale, and correlation are recommended since they are better at handling data with outliers or with density or shape differing moderately from the elliptical bivariate normal. Goldberg and Iglewicz (1992) discuss a number of possibilities.

- To draw the bivariate boxplot, first calculate the median E_m and the maximum E_{\max} of the standardized errors, E_i , which are essentially the generalized distances of each point from the centre (T_x^*, T_y^*) . Specifically, the E_i are defined by

$$E_i = \sqrt{\frac{X_{si}^2 + Y_{si}^2 - 2R^* X_{si} Y_{si}}{1 - R^{*2}}},$$

where $X_{si} = (X_i - T_x^*)/S_x^*$ is the standardized X_i value and Y_{si} is similarly defined.

- Then

$$E_m = \text{median} \{E_i: i = 1, 2, K, n\}$$

and

$$E_{\max} = \text{maximum} \{E_i: E_i^2 < DE_m^2\}.$$

- To draw the hinge, let

$$R_1 = E_m \sqrt{\frac{1 + R^*}{2}}, \quad R_2 = E_m \sqrt{\frac{1 - R^*}{2}}.$$

- For $\theta = 0$ to 360 in steps of 2, 3, 4, or 5 degrees, let

$$\begin{aligned} \Theta_1 &= R_1 \cos \theta, \\ \Theta_2 &= R_2 \sin \theta, \\ X &= T_x^* + (\Theta_1 + \Theta_2)S_x^*, \\ Y &= T_y^* + (\Theta_1 - \Theta_2)S_y^*. \end{aligned}$$

- Finally, plot X, Y .

To illustrate the use of a bivariate boxplot we shall again use the *SO2* and *Mortality* scatterplot. An R and S-PLUS function, `bivbox`, for constructing and plotting the boxplot is given on the website (see Preface) and can be used as follows,

```
bivbox(cbind(SO2, Mortality), xlab="SO2", ylab="Mortality")
```

to give the diagram shown in Figure 2.6.

In Figure 2.6 robust estimators of scale and location have been used and the diagram suggests that there are five outliers in the data. To use the nonrobust

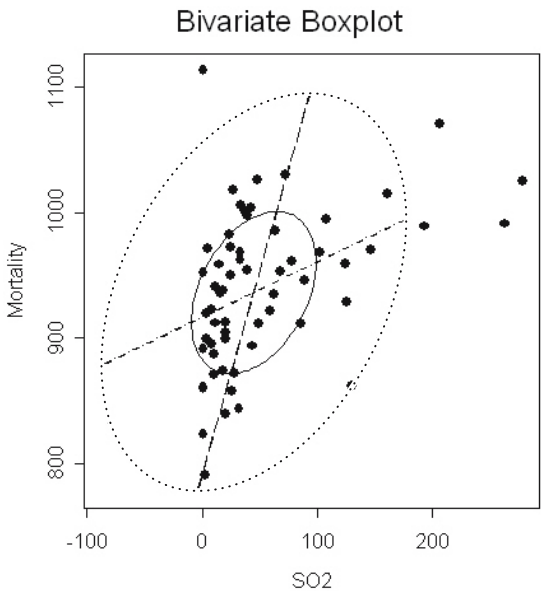


Figure 2.6 Bivariate boxplot of *SO2* and *Mortality* (robust estimators of location, scale, and correlation).

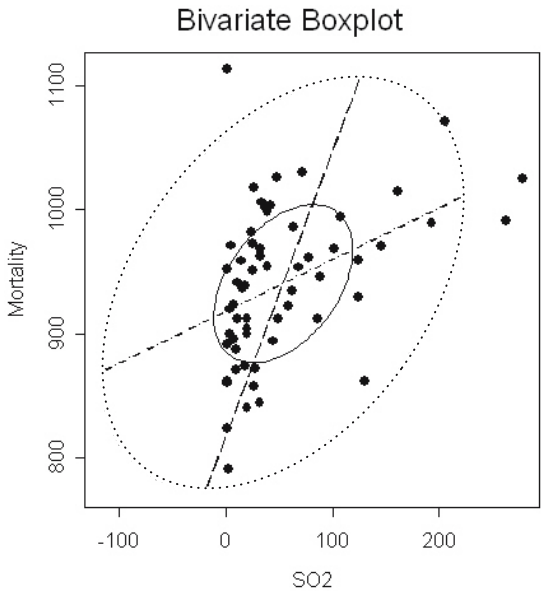


Figure 2.7 Bivariate boxplot of *SO2* and *Mortality* (nonrobust estimators).

estimators, that is, the usual means, variances, and correlation coefficient, the necessary code is

```
bivbox(cbind(SO2,Mortality),xlab="SO2",ylab="Mortality",
      method="O")
```

The resulting diagram is shown in Figure 2.7. Now only three outliers are identified. In general the use of the robust estimator version of the `bivbox` function is recommended.

2.3 Estimating Bivariate Densities

Often the aim in examining scatterplots is to identify regions where there are high or low densities of observations, “clusters,” or to spot outliers. But humans are not particularly good at visually examining point density, and it is often a very helpful aid to add some type of *bivariate density estimate* to the scatterplot. In general a nonparametric estimate is most useful since we are unlikely, in most cases, to want to assume some particular parametric form such as the bivariate normality. There is now a considerable literature on density estimation; see, for example, Silverman (1986) and Wand and Jones (1995). Basically, density estimates are “smoothed” two-dimensional histograms. A brief summary of the mathematics of bivariate density estimation is given in Display 2.3.

Display 2.3 Estimating Bivariate Densities

- The data set whose underlying density is to be estimated is $\mathbf{X}_1, \mathbf{X}_2, L, \mathbf{X}_n$.
- The bivariate kernel density estimator with kernel K and window width h is defined by

$$\hat{f}(\mathbf{x}) = \frac{1}{nh^2} \sum_{i=1}^n K \left\{ \frac{1}{h}(\mathbf{x} - \mathbf{X}_i) \right\}.$$

- The kernel function $K(\mathbf{x})$ is a function, defined for bivariate \mathbf{x} , satisfying

$$\int K(\mathbf{x}) d\mathbf{x} = 1.$$

- Usually $K(\mathbf{x})$ will be a radially symmetric unimodal probability density function, for example, the standard bivariate normal density function:

$$K(\mathbf{x}) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{x}\right).$$

Let us look at a simple two-dimensional histogram of the *Mortality*/*SO2* observations found and then displayed as a perspective plot by using the S-PLUS code

```
h2d<-hist2d(SO2,Mortality)
persp(h2d,xlab="SO2",ylab="Mortality",zlab="Frequency")
```

The result is shown in Figure 2.8. The density estimate given by the histogram is really too rough to be useful. (The function `hist2d` appears to be unavailable in R, but this is of little consequence since, in practice, unsmoothed two-dimensional histograms are of little use.)

Now we can use the R and S-PLUS function `bivden` given on the website to find a smoother estimate of the bivariate density of *Mortality* and *SO2* and to then display the estimated density as both a contour and perspective plot. The necessary code is

```
#get bivariate density estimates using a normal kernel
den1<-bivden(SO2,Mortality)
#construct a perspective plot of the density values
persp(den1$seqx,den1$seqy,den1$den,xlab="SO2",
      ylab="Mortality",
      zlab="Density",lwd=2)
#
plot(SO2,Mortality)
#add a contour plot of the density values to the scatterplot
contour(den1$seqx,den1$seqy,den1$den,lwd=2,nlevels=20,add=T)
```

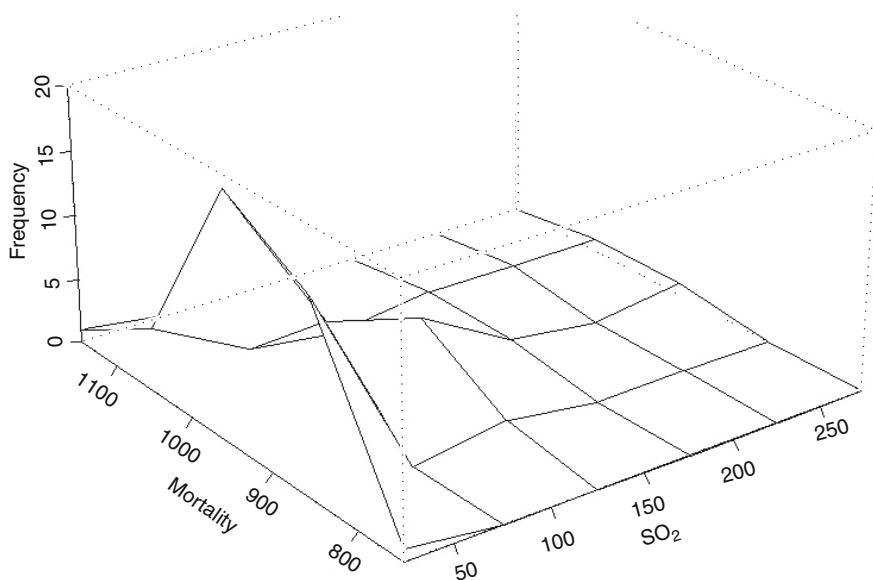


Figure 2.8 Two-dimensional histogram of *Mortality* and *SO2*.

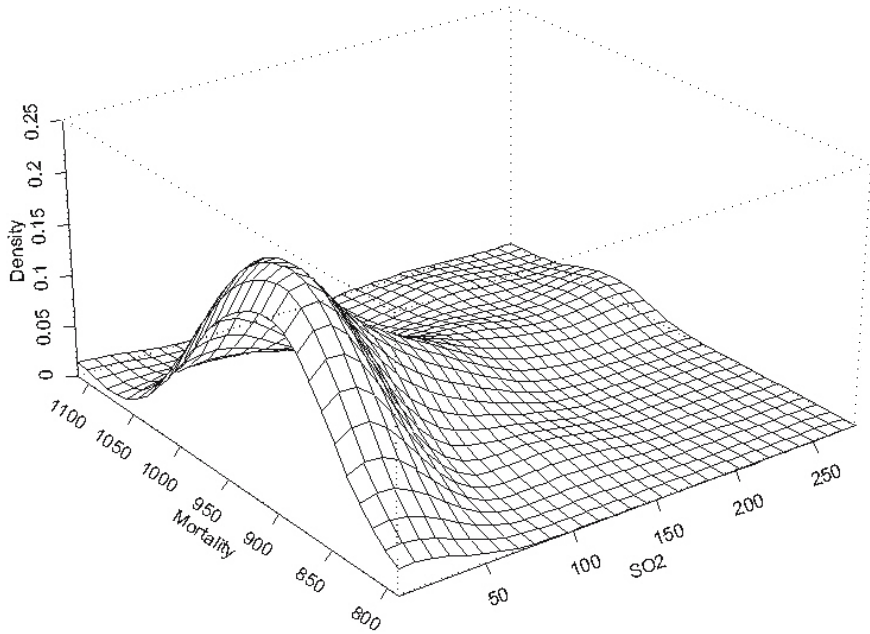


Figure 2.9 Perspective plot of estimated bivariate density of *Mortality* and *SO2*.

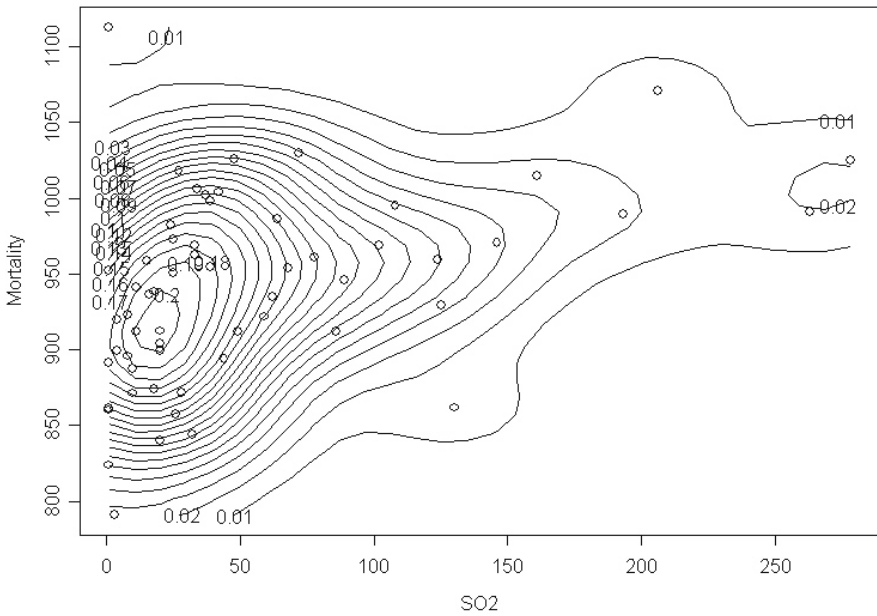


Figure 2.10 Contour plot of estimated bivariate density of *Mortality* and *SO2*.

The results are shown in Figures 2.9 and 2.10. Both plots give a clear indication of the skewness in the bivariate density of the two variables. (The diagrams shown result from using S-PLUS; those from R are a little different.)

In R the `bkde2D` function from the *KernSmooth* library might also be used to provide bivariate density estimates; see Exercise 2.7.

2.4 Representing Other Variables on a Scatterplot

The scatterplot can only display two variables. But there have been a number of suggestions as to how extra variables may be included. In this section we shall illustrate one of these, the *bubbleplot*, in which three variables are displayed. Two variables are used to form the scatterplot itself, and then the values of the third variable are represented by circles with radii proportional to these values and centered on the appropriate point in the scatterplot. To illustrate the bubbleplot we shall use the three variables, *SO2*, *Rainfall*, and *Mortality* from the air pollution data. The R and S-PLUS code needed to produce the required bubble plot is

```
plot(SO2,Mortality,pch=1,lwd=2,ylim=c(700,1200),
     xlim=c(-5,300))
#add circles to scatterplot
symbols(SO2,Mortality,circles=Rainfall,inches=0.4,add=TRUE,
       lwd=2)
```

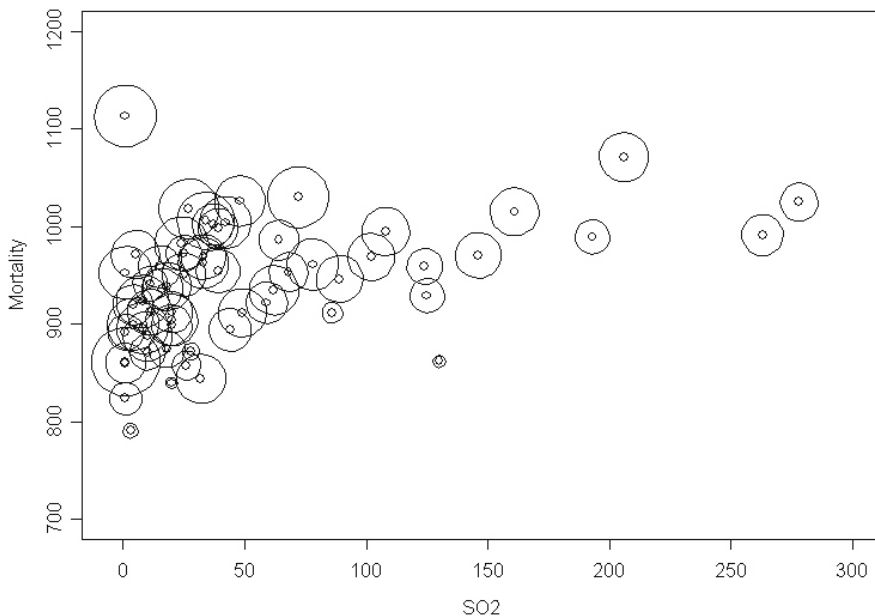


Figure 2.11 Bubbleplot of *Mortality* and *SO2* with *Rainfall* represented by radii of circles.

The resulting diagram is shown in Figure 2.11. Two particular observations to note are the one with high mortality and rainfall but very low sulphur dioxide level (NworILA) and the one with relatively low mortality and low rainfall but moderate sulphur dioxide level (IosangCA).

2.5 The Scatterplot Matrix

There are seven variables in the air pollution data which between them generate 21 possible scatterplots, and it is very important that the separate plots are presented in the best way to an in overall comprehension and understanding of the data. The *scatterplot matrix* is intended to accomplish this objective.

A scatterplot matrix is defined as a square, symmetric grid of bivariate scatterplots. The grid has q rows and columns, each one corresponding to a different variable. Each of the grid's cells shows a scatterplot of two variables. Variable j is plotted against variable i in the ij th cell, and the same variables appear in cell ji with the x - and y -axes of the scatterplots interchanged. The reason for including both the upper and lower triangles of the grid, despite the seeming redundancy, is that it enables a row and a column to be visually scanned to see one variable against all others, with the scales for the one variable lined up along the horizontal or the vertical.

To produce the basic scatterplot matrix of the air pollution variable we can use the `pairs` function in both R and S-PLUS

```
pairs(airpoll)
```

The result is Figure 2.12. The plot highlights that many pairs of variables in the air pollution data appear to be related in a relatively complex fashion, and that there are some potentially troublesome outliers in the data.

Rather than having variable labels on the main diagonal as in Figure 2.10, we may like to have some graphical representation of the marginal distribution of the corresponding variable, for example, a histogram. And here is a convenient point in the discussion to illustrate briefly the “click-and-point” features of the S-PLUS GUI since these can be useful in some situations, although for serious work the command line approach used up to now, and in most of the remainder of the book, is to be recommended. So to construct the required plot:

- Click on **Graph** in the toolbar;
- Select **2D plot**;
- In Axes Type highlight **Matrix**;
- Click **OK**;
- In **Scatterplot Matrix Dialogue** select `airpoll` as data set;
- Highlight all variables names in **x-column slot**;
- Check **Line/Histogram** tab;
- Check **Draw Histogram**;
- Click on **OK**.

The resulting diagram appears in Figure 2.13.

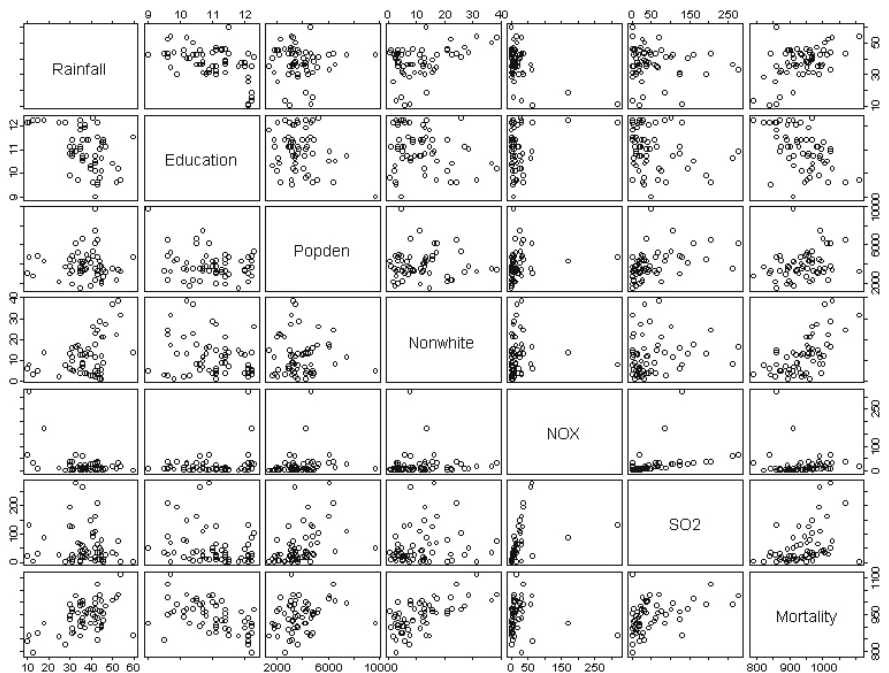


Figure 2.12 Scatterplot matrix of air pollution data.

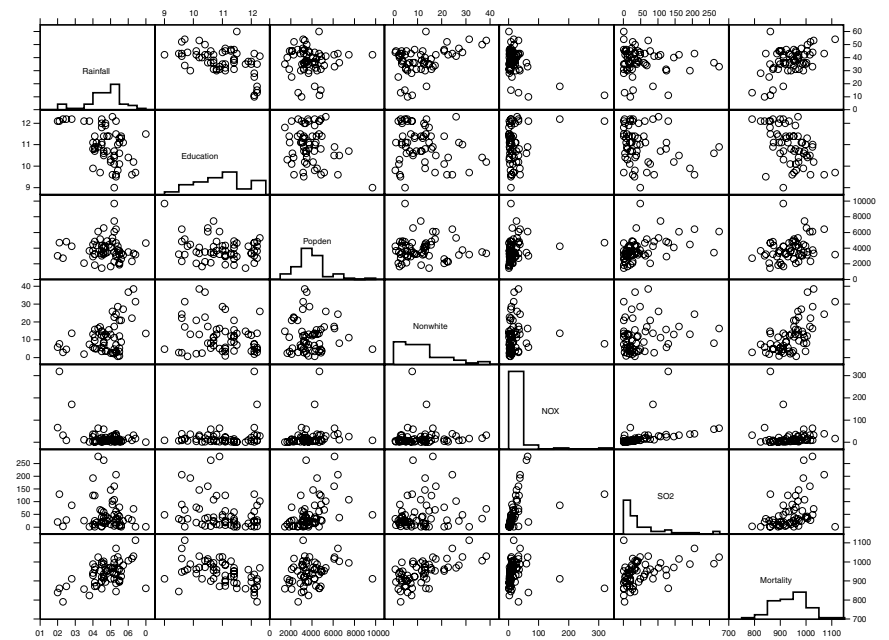


Figure 2.13 Scatterplot matrix of air pollution data showing histograms of each variable on the main diagonal.

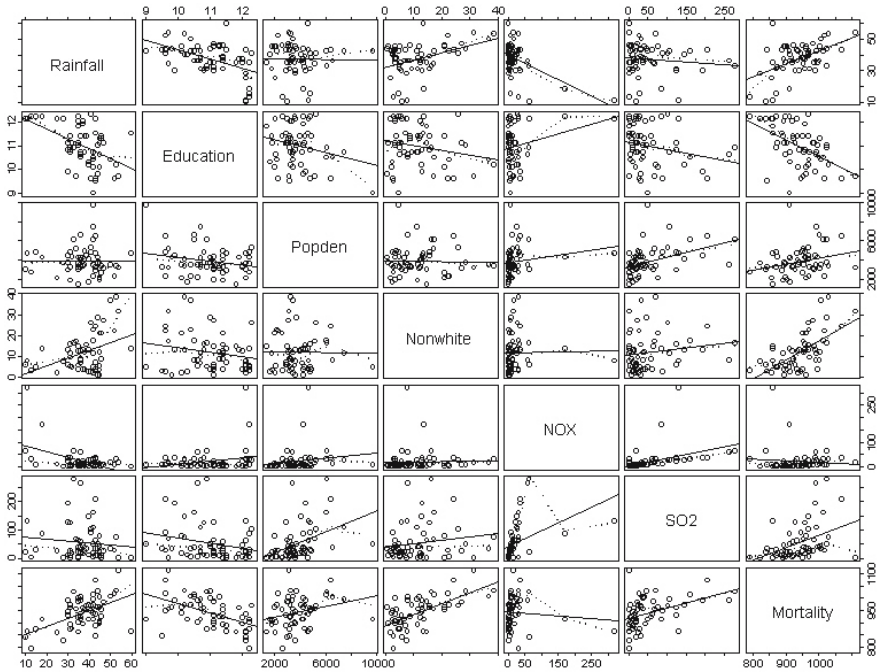


Figure 2.14 Scatterplot matrix of air pollution data showing linear and locally weighted regression fits on each panel.

Previously in this chapter we looked at a variety of ways in which individual scatterplots can be enhanced to make them more useful. These enhancements can, of course, also be used on each panel of a scatterplot matrix. For example, we can add linear and locally weighted regression fits to the air pollution diagram using the following code in either R or S-PLUS

```
pairs(airpoll, panel=function(x,y) {abline(lsfrit(x,y)$coef,
                                           lwd=2)
                                   lines(lowess(x,y), lty=2,
                                           lwd=2)
                                   points(x,y) })
```

to give Figure 2.14. Other possibilities for enhancing the panels of a scatterplot matrix are considered in the exercises.

2.6 Three-Dimensional Plots

In S-PLUS there are a variety of three-dimensional plots that can often be usefully applied to multivariate data. We will illustrate some of the possibilities using once

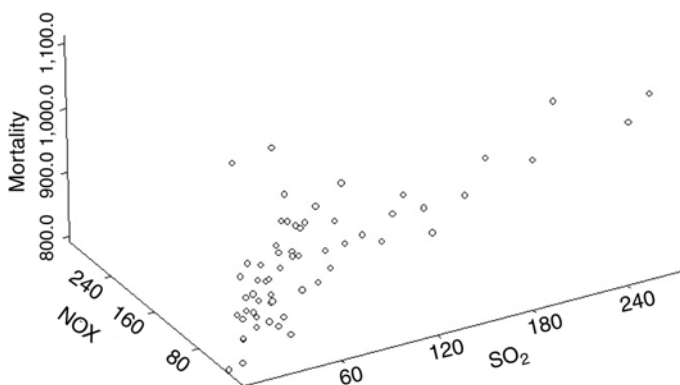


Figure 2.15 Three-dimensional plot of SO_2 , NOX , and $Mortality$.

again the air pollution data. To begin we will construct a simple three-dimensional plot of SO_2 , NOX , and $Mortality$ again using the S-PLUS GUI:

- Click **Graph** on the tool bar;
- Select **3D**;
- In **Insert Graph Dialogue**, choose **3D Scatter**, and click **OK**;
- In the **3D Line/Scatterplot [1] dialogue** select **Data Set** *airpoll*;
- Select SO_2 for **x Column**, NOX for **y Column**, and $Mortality$ for **z**;
- Click **OK**.

$Mortality$ appears to increase rapidly with increasing NOX values but more modestly with increasing levels of SO_2 . (A similar diagram can be found using the `cloud` function in S-PLUS and in R where it is available in the *lattice* library.)

Often it is easier to see what is happening in such a plot if lines are used to join drop-line plot. Such a plot is obtained using the instructions above but in **Insert Graph dialogue**, choose **3D Scatter with Drop Line**. The result is shown in Figure 2.16

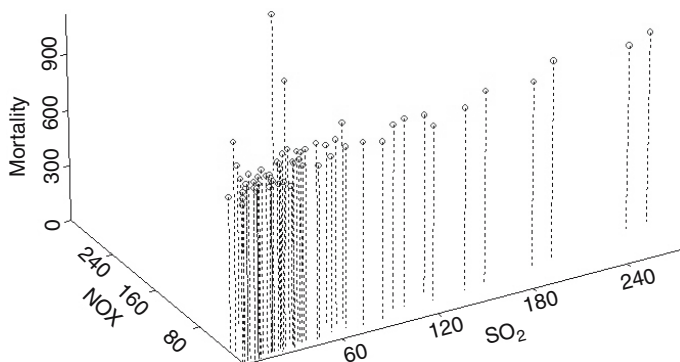


Figure 2.16 Three-dimensional drop line plot of SO_2 , NOX , and $Mortality$.

2.7 Conditioning Plots and Trellis Graphics

The conditioning plot or *coplot* is a potentially powerful visualization tool for studying the bivariate relationship of a pair of variables conditional on the values of one or more other variables. Such plots can often highlight the presence of interactions between the variables where the degree and/or direction of the bivariate relationship differs in the different levels of the third variable.

To illustrate we will construct a coplot of *Mortality* against *SO2* conditioned on population density (*Popden*) for the air pollution data. We need the R and S-PLUS function `coplot`

```
coplot(Mortality~SO2 | Popden)
```

The resulting plot is shown in Figure 2.17. In this diagram, the panel at the top is known as the given panel; the panels below are dependence panels. Each rectangle in the given panel specifies a range of values of population density. On a corresponding dependence panel, *Mortality* is plotted against *SO2* for those regions with population densities within one of the intervals in the given panel. To match the latter to the dependence panels, these panels need to be examined from left to right in the bottom row and then again left to right in subsequent rows.

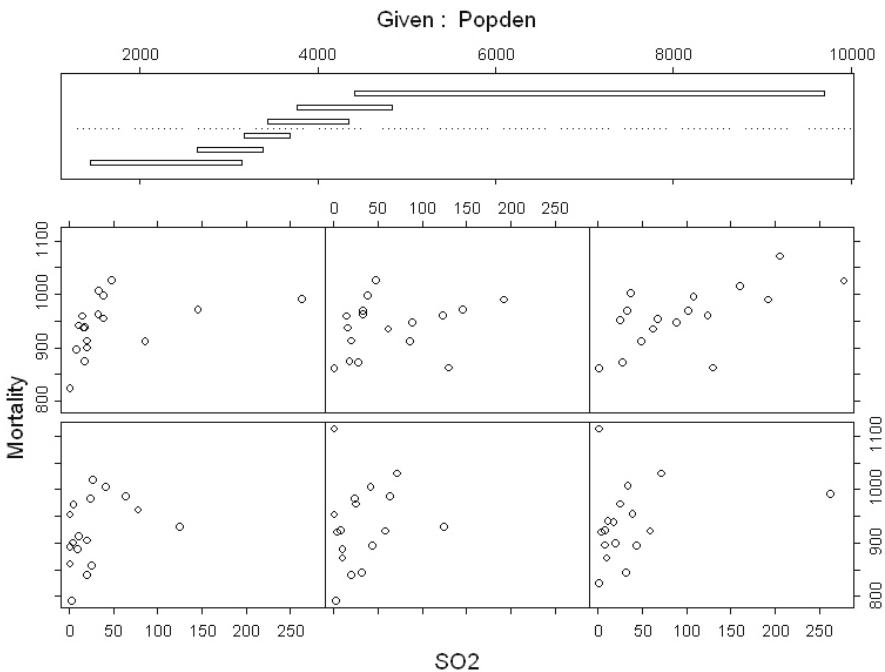


Figure 2.17 Coplot of *SO2* and *Mortality* conditional on population density.

There are relatively few observations in each panel on which to draw conclusions about possible differences in the relationship of *SO2* and *Mortality* at different levels of population density although there do appear to be some differences. In such cases it is often helpful to enhance the coplot dependence panels in some way. Here we add a locally weighted regression fit using the R and S-PLUS code:

```
coplot (Mortality~SO2 | Popden, panel=function(x,y,col,pch)
  panel.smooth(x,y,span=1) )
```

The result is shown in Figure 2.18. This plot suggests that the relationship between mortality and sulphur dioxide for lower levels of population density is more complex than at higher levels, although the number of points on which this claim is made is rather small.

Conditional graphical displays are simple examples of a more general scheme known as *trellis graphics* (Becker and Cleveland, 1994). This is an approach to examining high-dimensional structure in data by means of one-, two-, and three-dimensional graphs. The problem addressed is how observations of one or more variables depend on the observations of the other variables. The essential feature of this approach is the multiple conditioning that allows some type of plot to be displayed for different values of a given variable (or variables). The aim is to help in

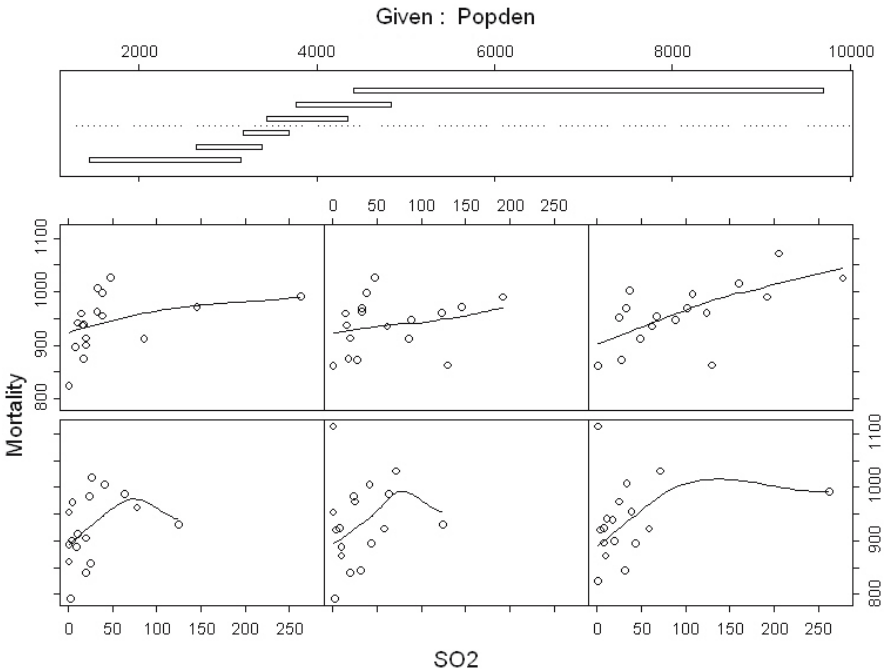




Figure 2.18 Coplot of *SO2* and *Mortality* conditional on population density with added locally weighted regression fit.

understanding both the structure of the data and how well proposed models describe the structure. An excellent recent example of the application of trellis graphics is given in Verbyla et al. (1999). To illustrate the possibilities we shall construct a three-dimensional plot of *SO₂*, *NOX*, and *Mortality* conditional on *Popden*. The necessary “click-and-point” steps are:

- Click on **Data** in the tool bar;
- In **Select Data** box choose `airpoll1`;
- Click **OK**;
- Click on 3D plots button, , to get 3D plot palette;
- Highlight *NOX* in spreadsheet and then ctrl click on *SO₂*, *Mortality*, and *Popden*;
- Turn conditioning button  on;
- Choose **Drop line scatter** from 3D palette.

The resulting diagram is shown in Figure 2.19.

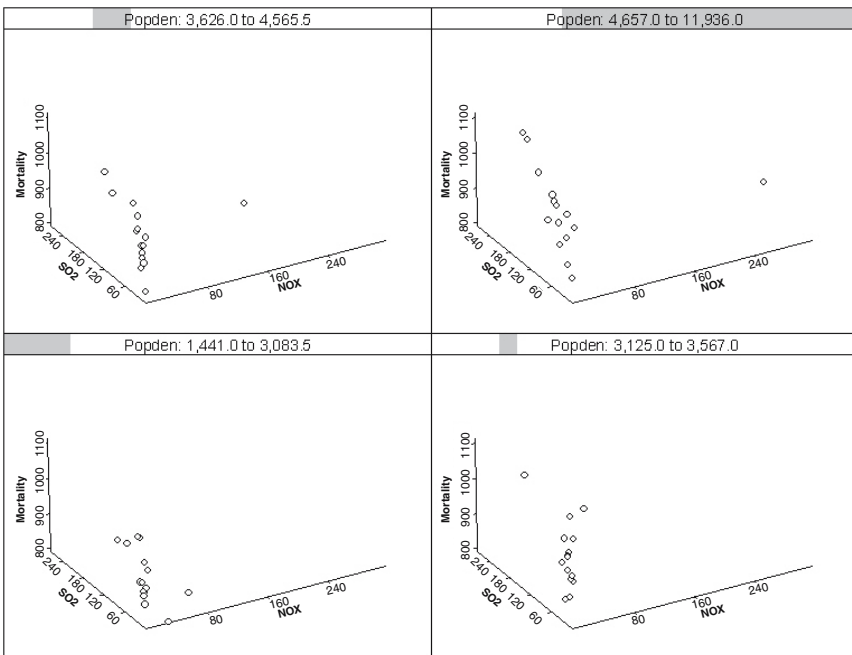


Figure 2.19 Three-dimensional plot for *NOX*, *SO₂*, and *Mortality* conditional on *Popden*.

2.8 Summary

Plotting multivariate data is an essential first step in trying to understand their message. The possibilities are almost limitless with software such as R and S-PLUS and readers are encouraged to explore more fully what is available. The methods covered in this chapter provide just some basic ideas for taking an initial look at multivariate data.

Exercises

- 2.1 The bubbleplot makes it possible to accommodate three variable values on a scatterplot. More than three variables can be accommodated by using what might be termed a *star plot* in which the extra variables are represented by the lengths of the sides of a “star.” Construct such a plot for all seven variables in the air pollution data using say *Rainfall* and *SO2* to form the basic scatterplot. (Use the `symbols` function.)
- 2.2 Construct a scatterplot matrix of the air pollution data in which each panel shows a bivariate density estimate of the pair of variables.
- 2.3 Construct a trellis graphic showing a scatterplot of *SO2* and *Mortality* conditioned on both rainfall and population density.
- 2.4 Construct a three-dimensional plot of *Rainfall*, *SO2*, and *Mortality* showing the estimated regression surface of *Mortality* on the other two variables.
- 2.5 Construct a three-dimensional plot of *SO2*, *NOX*, and *Rainfall* in which the observations are labelled by an abbreviated form of the region name.
- 2.6 Investigate the use of the `chiplot` function on all pairs of variables in the air pollution data.
- 2.7 Investigate the use of the `bkde2D` function in the *KernSmooth* library of R to calculate the bivariate density of *SO2* and *Mortality* in the air pollution data. Use the `wireframe` function available in the *lattice* library in R to construct a perspective plot of the estimated density.
- 2.8 Produce a similar diagram to that given in Figure 2.19 using the `cloud` function.