4

Describing Relationships Between Variables

The methods of Chapter 3 are really quite simple. They require little in the way of calculations and are most obviously relevant to the analysis of a single engineering variable. This chapter provides methods that address the more complicated problem of describing relationships between variables and are computationally more demanding.

The chapter begins with least squares fitting of a line to bivariate quantitative data and the assessment of the goodness of that fit. Then the line-fitting ideas are generalized to the fitting of curves to bivariate data and surfaces to multivariate quantitative data. The next topic is the summarization of data from full factorial studies in terms of so-called factorial effects. Next, the notion of data transformations is discussed. Finally, the chapter closes with a short transitional section that argues that further progress in statistics requires some familiarity with the subject of probability.

4.1 Fitting a Line by Least Squares

Bivariate data often arise because a quantitative experimental variable x has been varied between several different settings, producing a number of samples of a response variable y. For purposes of summarization, interpolation, limited extrapolation, and/or process optimization/adjustment, it is extremely helpful to have an equation relating y to x. A linear (or straight line) equation

$$y \approx \beta_0 + \beta_1 x \tag{4.1}$$

relating y to x is about the simplest potentially useful equation to consider after making a simple (x, y) scatterplot.

In this section, the principle of least squares is used to fit a line to (x, y) data. The appropriateness of that fit is assessed using the sample correlation and the coefficient of determination. Plotting of residuals is introduced as an important method for further investigation of possible problems with the fitted equation. A discussion of some practical cautions and the use of statistical software in fitting equations to data follows.

4.1.1 Applying the Least Squares Principle

Example 1

Pressing Pressures and Specimen Densities for a Ceramic Compound

Benson, Locher, and Watkins studied the effects of varying pressing pressures on the density of cylindrical specimens made by dry pressing a ceramic compound. A mixture of Al_2O_3, polyvinyl alcohol, and water was prepared, dried overnight, crushed, and sieved to obtain 100 mesh size grains. These were pressed into cylinders at pressures from 2,000 psi to 10,000 psi, and cylinder densities were calculated. Table 4.1 gives the data that were obtained, and a simple scatterplot of these data is given in Figure 4.1.

Table 4.1
Pressing Pressures and Resultant Specimen Densities

x, Pressure (psi)	y, Density (g/cc)
2,000	2.486
2,000	2.479
2,000	2.472
4,000	2.558
4,000	2.570
4,000	2.580
6,000	2.646
6,000	2.657
6,000	2.653
8,000	2.724
8,000	2.774
8,000	2.808
10,000	2.861
10,000	2.879
10,000	2.858

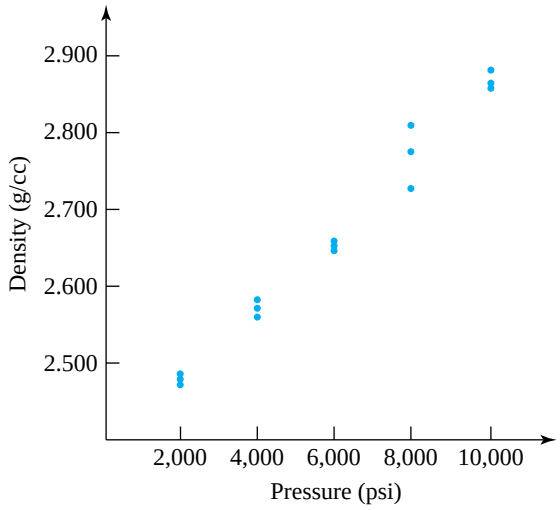

Figure 4.1 Scatterplot of density vs. pressing pressure

It is very easy to imagine sketching a straight line through the plotted points in Figure 4.1. Such a line could then be used to summarize how density depends upon pressing pressure. The principle of **least squares** provides a method of choosing a "best" line to describe the data.

Definition 1

To apply the **principle of least squares** in the fitting of an equation for y to an n-point data set, values of the equation parameters are chosen to minimize

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad \textbf{(4.2)}$$

where $y_1, y_2, \ldots, y_n$ are the observed responses and $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ are corresponding responses predicted or fitted by the equation.

In the context of fitting a line to (x, y) data, the prescription offered by Definition 1 amounts to choosing a slope and intercept so as to minimize the sum of squared vertical distances from (x, y) data points to the line in question. This notion is shown in generic fashion in Figure 4.2 for a fictitious five-point data set. (It is the *squares* of the five indicated differences that must be added and minimized.)

Looking at the form of display (4.1), for the fitting of a line,

$$\hat{y} = \beta_0 + \beta_1 x$$

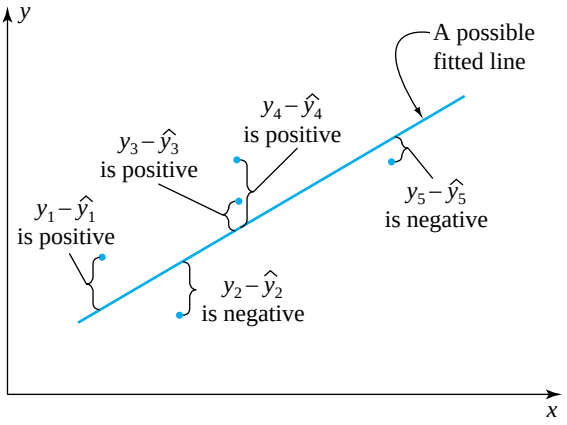

Figure 4.2 Five data points (x, y) and a possible fitted line

Therefore, the expression to be minimized by choice of slope (β_1) and intercept (β_0) is

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2 \tag{4.3}$$

The minimization of the function of two variables $S(\beta_0, \beta_1)$ is an exercise in calculus. The partial derivatives of S with respect to β_0 and β_1 may be set equal to zero, and the two resulting equations may be solved simultaneously for β_0 and β_1. The equations produced in this way are

$$n\beta_0 + \left(\sum_{i=1}^{n} x_i\right)\beta_1 = \sum_{i=1}^{n} y_i \tag{4.4}$$

and

$$\left(\sum_{i=1}^{n} x_i\right)\beta_0 + \left(\sum_{i=1}^{n} x_i^2\right)\beta_1 = \sum_{i=1}^{n} x_i y_i \tag{4.5}$$

For reasons that are not obvious, equations (4.4) and (4.5) are sometimes called the **normal** (as in perpendicular) **equations** for fitting a line. They are two linear equations in two unknowns and can be fairly easily solved for β_0 and β_1 (provided

there are at least two different x_i's in the data set). Simultaneous solution of equations (4.4) and (4.5) produces values of β_1 and β_0 given by

Slope of the least squares line, b_1

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \tag{4.6}$$

and

Intercept of the least squares line, b_0

$$b_0 = \bar{y} - b_1 \bar{x} \tag{4.7}$$

Notice the notational convention here. *The particular numerical slope and intercept minimizing $S(\beta_0, \beta_1)$ are denoted (not as β's but) as b_1 and b_0.*

In display (4.6), somewhat standard practice has been followed (and the summation notation abused) by not indicating the variable or range of summation (i, from 1 to n).

Example 1 (*continued*)

It is possible to verify that the data in Table 4.1 yield the following summary statistics:

$$\sum x_i = 2{,}000 + 2{,}000 + \cdots + 10{,}000 = 90{,}000,$$

$$\text{so } \bar{x} = \frac{90{,}000}{15} = 6{,}000$$

$$\sum (x_i - \bar{x})^2 = (2{,}000 - 6{,}000)^2 + (2{,}000 - 6{,}000)^2 + \cdots + (10{,}000 - 6{,}000)^2 = 120{,}000{,}000$$

$$\sum y_i = 2.486 + 2.479 + \cdots + 2.858 = 40.005,$$

$$\text{so } \bar{y} = \frac{40.005}{15} = 2.667$$

$$\sum (y_i - \bar{y})^2 = (2.486 - 2.667)^2 + (2.479 - 2.667)^2 + \cdots + (2.858 - 2.667)^2 = .289366$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (2{,}000 - 6{,}000)(2.486 - 2.667) + \cdots + (10{,}000 - 6{,}000)(2.858 - 2.667) = 5{,}840$$

Example 1 (*continued*)

Then the least squares slope and intercept, b_1 and b_0, are given via equations (4.6) and (4.7) as

$$b_1 = \frac{5{,}840}{120{,}000{,}000} = .0000486\overline{6} \text{ (g/cc)/psi}$$

and

$$b_0 = 2.667 - (.0000486\overline{6})(6{,}000) = 2.375 \text{ g/cc}$$

Figure 4.3 shows the least squares line

$$\hat{y} = 2.375 + .0000487x$$

Interpretation of the slope of the least squares line

sketched on a scatterplot of the (x, y) points from Table 4.1. Note that the slope on this plot, $b_1 \approx .0000487$ (g/cc)/psi, has physical meaning as the (approximate) increase in y (density) that accompanies a unit (1 psi) increase in x (pressure). The intercept on the plot, $b_0 = 2.375$ g/cc, positions the line vertically and is the value at which the line cuts the y axis. But it should probably not be interpreted as the density that would accompany a pressing pressure of $x = 0$ psi. The point is that the reasonably linear-looking relation that the students found for pressures between 2,000 psi and 10,000 psi could well break down at larger or smaller pressures. Thinking of b_0 as a 0 pressure density amounts to an extrapolation outside the range of data used to fit the equation, something that ought always to be approached with extreme caution.

Extrapolation

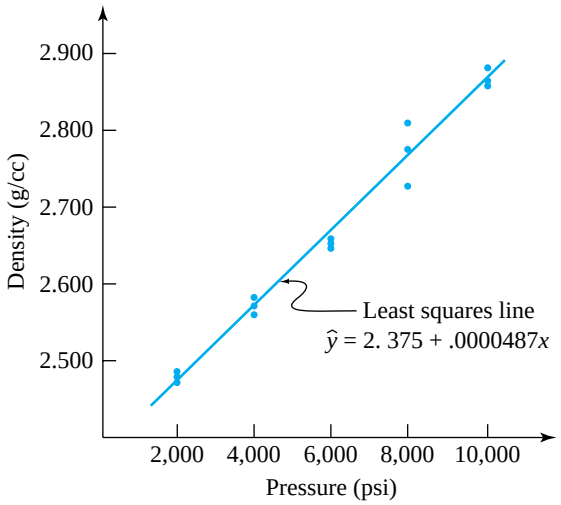

Figure 4.3 Scatterplot of the pressure/density data and the least squares line

As indicated in Definition 1, the value of y on the least squares line corresponding to a given x can be termed a **fitted** or **predicted value**. It can be used to represent likely y behavior at that x.

Example 1 (*continued*)

Consider the problem of determining a typical density corresponding to a pressure of 4,000 psi and one corresponding to 5,000 psi.

First, looking at $x = 4{,}000$, a simple way of representing a typical y is to note that for the three data points having $x = 4{,}000$,

$$\bar{y} = \frac{1}{3}(2.558 + 2.570 + 2.580) = 2.5693 \text{ g/cc}$$

and so to use this as a representative value. But assuming that y is indeed approximately linearly related to x, the fitted value

$$\hat{y} = 2.375 + .0000486(4{,}000) = 2.5697 \text{ g/cc}$$

might be even better for representing average density for 4,000 psi pressure.

Interpolation

Looking then at the situation for $x = 5{,}000$ psi, there are no data with this x value. The only thing one can do to represent density at that pressure is to ask whether interpolation is sensible from a physical viewpoint. If so, the fitted value

$$\hat{y} = 2.375 + .0000486(5{,}000) = 2.6183 \text{ g/cc}$$

can be used to represent density for 5,000 psi pressure.

4.1.2 The Sample Correlation and Coefficient of Determination

Visually, the least squares line in Figure 4.3 seems to do a good job of fitting the plotted points. However, it would be helpful to have methods of quantifying the quality of that fit. One such measure is the **sample correlation**.

Definition 2

The **sample (linear) correlation** between x and y in a sample of n data pairs (x_i, y_i) is

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}} \tag{4.8}$$

Interpreting the sample correlation

The sample correlation always lies in the interval from -1 to 1. Further, it is -1 or 1 only when all (x, y) data points fall on a single straight line. Comparison of

formulas (4.6) and (4.8) shows that $r = b_1 \left(\sum (x_i - \bar{x})^2 / \sum (y_i - \bar{y})^2 \right)^{1/2}$ so that *b_1 and r have the same sign*. So a sample correlation of -1 means that y decreases linearly in increasing x, while a sample correlation of $+1$ means that y increases linearly in increasing x.

Real data sets do not often exhibit perfect ($+1$ or -1) correlation. Instead r is typically between -1 and 1. But drawing on the facts about how it behaves, people take r as a *measure of the strength of an apparent linear relationship*: r near $+1$ or -1 is interpreted as indicating a relatively strong linear relationship; r near 0 is taken as indicating a lack of linear relationship. The sign of r is thought of as indicating whether y tends to increase or decrease with increased x.

Example 1 (*continued*)

For the pressure/density data, the summary statistics in the example following display (4.7) (page 127) produces

$$r = \frac{5{,}840}{\sqrt{(120{,}000{,}000)(.289366)}} = .9911$$

This value of r is near $+1$ and indicates clearly the strong positive linear relationship evident in Figures 4.1 and 4.3.

The **coefficient of determination** is another measure of the quality of a fitted equation. It can be applied not only in the present case of the simple fitting of a line to (x, y) data but more widely as well.

Definition 3

The **coefficient of determination** for an equation fitted to an n-point data set via least squares and producing fitted y values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ is

$$R^2 = \frac{\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} \qquad \textbf{(4.9)}$$

Interpretation of R^2

R^2 may be interpreted as the *fraction of the raw variation in y accounted for using the fitted equation*. That is, provided the fitted equation includes a constant term, $\sum (y_i - \bar{y})^2 \geq \sum (y_i - \hat{y}_i)^2$. Further, $\sum (y_i - \bar{y})^2$ is a measure of raw variability in y, while $\sum (y_i - \hat{y}_i)^2$ is a measure of variation in y remaining after fitting the equation. So the nonnegative difference $\sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2$ is a measure of the variability in y accounted for in the equation-fitting process. R^2 then expresses this difference as a fraction (of the total raw variation).

Example 1 (*continued*)

Using the fitted line, one can find $\hat{y}$ values for all $n = 15$ data points in the original data set. These are given in Table 4.2.

Table 4.2
Fitted Density Values

x, Pressure	$\hat{y}$, Fitted Density
2,000	2.4723
4,000	2.5697
6,000	2.6670
8,000	2.7643
10,000	2.8617

Then, referring again to Table 4.1,

$$\begin{aligned}\sum(y_i - \hat{y}_i)^2 &= (2.486 - 2.4723)^2 + (2.479 - 2.4723)^2 + (2.472 - 2.4723)^2 \\ &\quad + (2.558 - 2.5697)^2 + \cdots + (2.879 - 2.8617)^2 \\ &\quad + (2.858 - 2.8617)^2 \\ &= .005153\end{aligned}$$

Further, since $\sum(y_i - \bar{y})^2 = .289366$, from equation (4.9)

$$R^2 = \frac{.289366 - .005153}{.289366} = .9822$$

and the fitted line accounts for over 98% of the raw variability in density, reducing the "unexplained" variation from .289366 to .005153.

R^2 as a squared correlation

The coefficient of determination has a second useful interpretation. For equations that are linear in the parameters (which are the only ones considered in this text), R^2 turns out to be a squared correlation. It is the squared correlation between the observed values y_i and the fitted values $\hat{y}_i$. (Since in the present situation of fitting a line, the $\hat{y}_i$ values are perfectly correlated with the x_i values, R^2 also turns out to be the squared correlation between the y_i and x_i values.)

Example 1 (*continued*)

For the pressure/density data, the correlation between x and y is

$$r = .9911$$

Example 1 (*continued*)

Since $\hat{y}$ is perfectly correlated with x, this is also the correlation between $\hat{y}$ and y. But notice as well that

$$r^2 = (.9911)^2 = .9822 = R^2$$

so R^2 is indeed the squared sample correlation between y and $\hat{y}$.

4.1.3 Computing and Using Residuals

When fitting an equation to a set of data, the hope is that the equation extracts the main message of the data, leaving behind (unpredicted by the fitted equation) only the variation in y that is uninterpretable. That is, one hopes that the y_i's will look like the $\hat{y}_i$'s except for small fluctuations explainable only as random variation. A way of assessing whether this view is sensible is through the computation and plotting of **residuals**.

Definition 4

If the fitting of an equation or model to a data set with responses $y_1, y_2, \ldots, y_n$ produces fitted values $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$, then the corresponding **residuals** are the values

$$e_i = y_i - \hat{y}_i$$

If a fitted equation is telling the whole story contained in a data set, then its residuals ought to be patternless. So when they're plotted against time order of observation, values of experimental variables, fitted values, or any other sensible quantities, the plots should look randomly scattered. When they don't, the patterns can themselves suggest what has gone unaccounted for in the fitting and/or how the data summary might be improved.

Example 2

Compressive Strength of Fly Ash Cylinders as a Function of Amount of Ammonium Phosphate Additive

As an exaggerated example of the previous point, consider the naive fitting of a line to some data of B. Roth. Roth studied the compressive strength of concrete-like fly ash cylinders. These were made using varying amounts of ammonium phosphate as an additive. Part of Roth's data are given in Table 4.3. The ammonium phosphate values are expressed as a percentage by weight of the amount of fly ash used.

Table 4.3
Additive Concentrations and Compressive Strengths for Fly Ash Cylinders

x, Ammonium Phosphate (%)	y, Compressive Strength (psi)	x, Ammonium Phosphate (%)	y, Compressive Strength (psi)
0	1221	3	1609
0	1207	3	1627
0	1187	3	1642
1	1555	4	1451
1	1562	4	1472
1	1575	4	1465
2	1827	5	1321
2	1839	5	1289
2	1802	5	1292

Using formulas (4.6) and (4.7), it is possible to show that the least squares line through the (x, y) data in Table 4.3 is

$$\hat{y} = 1498.4 - .6381x \qquad \textbf{(4.10)}$$

Then straightforward substitution into equation (4.10) produces fitted values $\hat{y}_i$ and residuals $e_i = y_i - \hat{y}_i$, as given in Table 4.4. The residuals for this straight-line fit are plotted against x in Figure 4.4.

The distinctly "up-then-back-down-again" curvilinear pattern of the plot in Figure 4.4 is not typical of random scatter. Something has been missed in

Table 4.4
Residuals from a Straight-Line Fit to the Fly Ash Data

x	y	$\hat{y}$	$e = y - \hat{y}$	x	y	$\hat{y}$	$e = y - \hat{y}$
0	1221	1498.4	−277.4	3	1609	1496.5	112.5
0	1207	1498.4	−291.4	3	1627	1496.5	130.5
0	1187	1498.4	−311.4	3	1642	1496.5	145.5
1	1555	1497.8	57.2	4	1451	1495.8	−44.8
1	1562	1497.8	64.2	4	1472	1495.8	−23.8
1	1575	1497.8	77.2	4	1465	1495.8	−30.8
2	1827	1497.2	329.8	5	1321	1495.2	−174.2
2	1839	1497.2	341.8	5	1289	1495.2	−206.2
2	1802	1497.2	304.8	5	1292	1495.2	−203.2

Example 2
(*continued*)

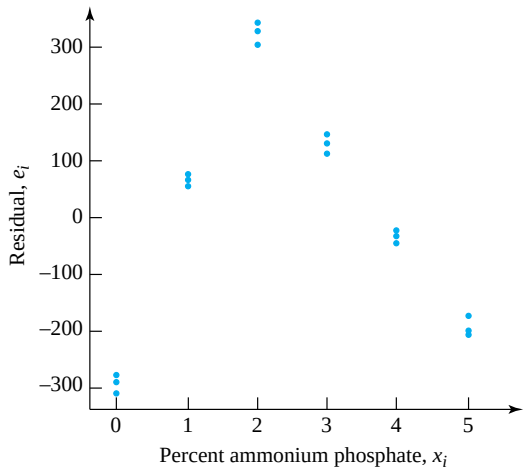

Figure 4.4 Plot of residuals vs. x for a linear fit to the fly ash data

the fitting of a line to Roth's data. Figure 4.5 is a simple scatterplot of Roth's data (which in practice should be made before fitting any curve to such data). It is obvious from the scatterplot that the relationship between the amount of ammonium phosphate and compressive strength is decidedly nonlinear. In fact, a quadratic function would come much closer to fitting the data in Table 4.3.

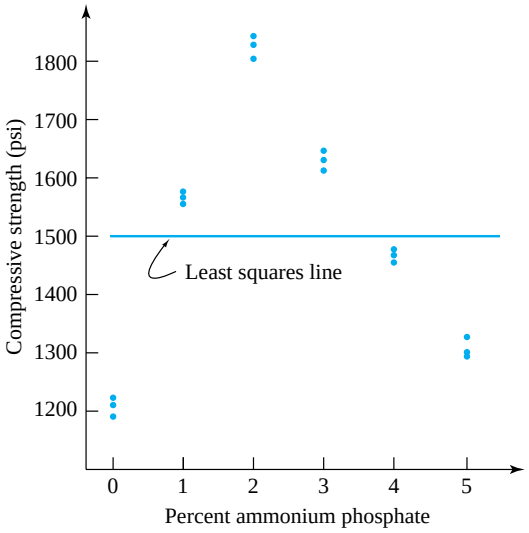

Figure 4.5 Scatterplot of the fly ash data

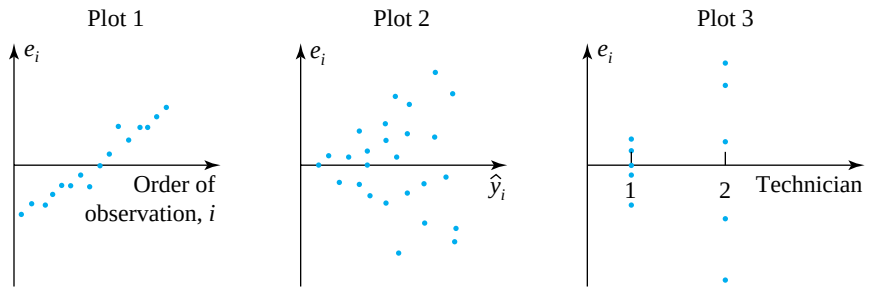

Figure 4.6 Patterns in residual plots

Interpreting patterns on residual plots

Figure 4.6 shows several patterns that can occur in plots of residuals against various variables. Plot 1 of Figure 4.6 shows a trend on a plot of residuals versus time order of observation. The pattern suggests that some variable changing in time is acting on y and has not been accounted for in fitting $\hat{y}$ values. For example, instrument drift (where an instrument reads higher late in a study than it did early on) could produce a pattern like that in Plot 1. Plot 2 shows a fan-shaped pattern on a plot of residuals versus fitted values. Such a pattern indicates that large responses are fitted (and quite possibly produced and/or measured) less consistently than small responses. Plot 3 shows residuals corresponding to observations made by Technician 1 that are on the whole smaller than those made by Technician 2. The suggestion is that Technician 1's work is more precise than that of Technician 2.

Normal-plotting residuals

Another useful way of plotting residuals is to normal-plot them. The idea is that the normal distribution shape is typical of random variation and that normal-plotting of residuals is a way to investigate whether such a distributional shape applies to what is left in the data after fitting an equation or model.

Example 1 (*continued*)

Table 4.5 gives residuals for the fitting of a line to the pressure/density data. The residuals e_i were treated as a sample of 15 numbers and normal-plotted (using the methods of Section 3.2) to produce Figure 4.7.

The central portion of the plot in Figure 4.7 is fairly linear, indicating a generally bell-shaped distribution of residuals. But the plotted point corresponding to the largest residual, and probably the one corresponding to the smallest residual, fail to conform to the linear pattern established by the others. Those residuals seem big in absolute value compared to the others.

From Table 4.5 and the scatterplot in Figure 4.3, one sees that these large residuals both arise from the 8,000 psi condition. And the spread for the three densities at that pressure value does indeed look considerably larger than those at the other pressure values. The normal plot suggests that the pattern of variation at 8,000 psi is genuinely different from those at other pressures. It may be that a different physical compaction mechanism was acting at 8,000 psi than at the other pressures. But it is more likely that there was a problem with laboratory technique, or recording, or the test equipment when the 8,000 psi tests were made.

Example 1 (*continued*)

In any case, the normal plot of residuals helps draw attention to an idiosyncrasy in the data of Table 4.1 that merits further investigation, and perhaps some further data collection.

Table 4.5
Residuals from the Linear Fit to the Pressure/Density Data

x, Pressure	y, Density	$\hat{y}$	$e = y - \hat{y}$
2,000	2.486	2.4723	.0137
2,000	2.479	2.4723	.0067
2,000	2.472	2.4723	−.0003
4,000	2.558	2.5697	−.0117
4,000	2.570	2.5697	.0003
4,000	2.580	2.5697	.0103
6,000	2.646	2.6670	−.0210
6,000	2.657	2.6670	−.0100
6,000	2.653	2.6670	−.0140
8,000	2.724	2.7643	−.0403
8,000	2.774	2.7643	.0097
8,000	2.808	2.7643	.0437
10,000	2.861	2.8617	−.0007
10,000	2.879	2.8617	.0173
10,000	2.858	2.8617	−.0037

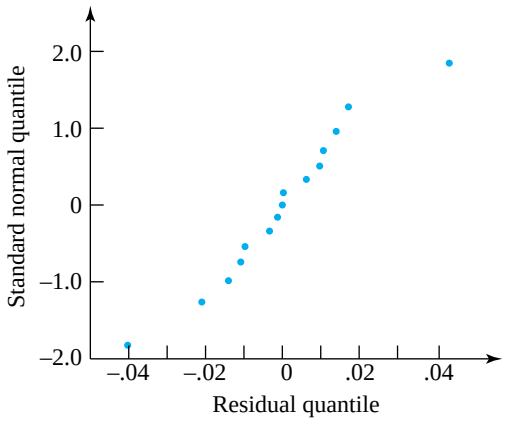

Figure 4.7 Normal plot of residuals from a linear fit to the pressure/density data

4.1.4 Some Cautions

The methods of this section are extremely useful engineering tools when thoughtfully applied. But a few additional comments are in order, warning against some errors in logic that often accompany their use.

r Measures only linear association

The first warning regards the correlation. It must be remembered that r measures only the **linear relationship** between x and y. It is perfectly possible to have a strong *nonlinear* relationship between x and y and yet have a value of r near 0. In fact, Example 2 is an excellent example of this. Compressive strength is strongly related to the ammonium phosphate content. But $r = -.005$, very nearly 0, for the data set in Table 4.3.

Correlation and causation

The second warning is essentially a restatement of one implicit in the early part of Section 1.2: Correlation is not necessarily causation. One may observe a large correlation between x and y in an observational study without it being true that x drives y or vice versa. It may be the case that another variable (say, z) drives the system under study and causes simultaneous changes in both x and y.

The influence of extreme observations

The last warning is that both $R^2(r)$ and least squares fitting can be drastically affected by a few unusual data points. As an example of this, consider the ages and heights of 36 students from an elementary statistics course plotted in Figure 4.8. By the time people reach college age, there is little useful relationship between age and height, but the correlation between ages and heights is .73. This fairly large value is produced by essentially a single data point. If the data point corresponding to the 30-year-old student who happened to be 6 feet 8 inches tall is removed from the data set, the correlation drops to .03.

An engineer's primary insurance against being misled by this kind of phenomenon is the habit of **plotting** data in as many different ways as are necessary to get a feel for how they are structured. Even a simple boxplot of the age data or height

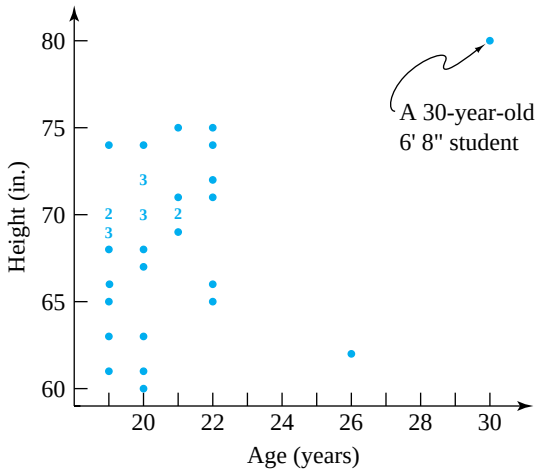

Figure 4.8 Scatterplot of ages and heights of 36 students

data alone would have identified the 30-year-old student in Figure 4.8 as unusual. That would have raised the possibility of that data point strongly influencing both r and any curve that might be fitted via least squares.

4.1.5 Computing

The examples in this section have no doubt left the impression that computations were done "by hand." In practice, such computations are almost always done with a statistical analysis package. The fitting of a line by least squares is done using a **regression program**. Such programs usually also compute R^2 and have an option that allows the computing and plotting of residuals.

It is not the purpose of this text to teach or recommend the use of any particular statistical package, but annotated printouts will occasionally be included to show how MINITAB formats its output. Printout 1 is such a printout for an analysis of the pressure/density data in Table 4.1, paralleling the discussion in this section. (MINITAB's regression routine is found under its "Stat/Regression/Regression" menu.) MINITAB gives its user much more in the way of analysis for least squares curve fitting than has been discussed to this point, so your understanding of Printout 1 will be incomplete. But it should be possible to locate values of the major summary statistics discussed here. The printout shown doesn't include plots, but it's worth noting that the program has options for saving fitted values and residuals for later plotting.

Printout 1 Fitting the Least Squares Line to the Pressure/Density Data

```
Regression Analysis


The regression equation is
density = 2.38 +0.000049 pressure

Predictor        Coef       StDev          T        P
Constant      2.37500     0.01206     197.01    0.000
pressure   0.00004867  0.00000182      26.78    0.000

S = 0.01991     R-Sq = 98.2%     R-Sq(adj) = 98.1%


Analysis of Variance

Source            DF          SS          MS         F        P
Regression         1     0.28421     0.28421    717.06    0.000
Residual Error    13     0.00515     0.00040
Total             14     0.28937

Obs   pressure    density         Fit   StDev Fit    Residual    St Resid
  1       2000    2.48600     2.47233     0.00890     0.01367        0.77
  2       2000    2.47900     2.47233     0.00890     0.00667        0.37
  3       2000    2.47200     2.47233     0.00890    -0.00033       -0.02
  4       4000    2.55800     2.56967     0.00630    -0.01167       -0.62
  5       4000    2.57000     2.56967     0.00630     0.00033        0.02
  6       4000    2.58000     2.56967     0.00630     0.01033        0.55
  7       6000    2.64600     2.66700     0.00514    -0.02100       -1.09
  8       6000    2.65700     2.66700     0.00514    -0.01000       -0.52
```

9	6000	2.65300	2.66700	0.00514	-0.01400	-0.73
10	8000	2.72400	2.76433	0.00630	-0.04033	-2.14R
11	8000	2.77400	2.76433	0.00630	0.00967	0.51
12	8000	2.80800	2.76433	0.00630	0.04367	2.31R
13	10000	2.86100	2.86167	0.00890	-0.00067	-0.04
14	10000	2.87900	2.86167	0.00890	0.01733	0.97
15	10000	2.85800	2.86167	0.00890	-0.00367	-0.21

R denotes an observation with a large standardized residual

At the end of Section 3.3 we warned that using spreadsheet software in place of high-quality statistical software can, without warning, produce spectacularly wrong answers. The example provided at the end of Section 3.3 concerns a badly wrong sample variance of only three numbers. It is important to note that the potential for numerical inaccuracy shown in that example carries over to the rest of the statistical methods discussed in this book, including those of the present section. For example, consider the $n = 6$ hypothetical (x, y) pairs listed in Table 4.6. For fitting a line to these data via least squares, MINITAB correctly produces $R^2 = .997$. But as recently as late 1999, the current version of the leading spreadsheet program returned the ridiculously wrong value, $R^2 = -.81648$. (This data set comes from a posting by Mark Eakin on the "edstat" electronic bulletin board that can be found at http://jse.stat.ncsu.edu/archives/.)

Table 4.6
6 Hypothetical Data Pairs

x	y	x	y
10,000,000.1	1.1	10,000,000.4	3.9
10,000,000.2	1.9	10,000,000.5	4.9
10,000,000.3	3.1	10,000,000.6	6.1

Section 1 Exercises

1. The following is a small set of artificial data. Show the hand calculations necessary to do the indicated tasks.

x	1	2	3	4	5
y	8	8	6	6	4

(a) Obtain the least squares line through these data. Make a scatterplot of the data and sketch this line on that scatterplot.

(b) Obtain the sample correlation between x and y for these data.

(c) Obtain the sample correlation between y and $\hat{y}$ for these data and compare it to your answer to part (b).

(d) Use the formula in Definition 3 and compute R^2 for these data. Compare it to the square of your answers to parts (b) and (c).

(e) Find the five residuals from your fit in part (a). How are they portrayed geometrically on the scatterplot for (a)?

2. Use a computer package and redo the computations and plotting required in Exercise 1. Annotate your output, indicating where on the printout you can

find the equation of the least squares line, the value of r, the value of R^2, and the residuals.

3. The article "Polyglycol Modified Poly (Ethylene Ether Carbonate) Polyols by Molecular Weight Advancement" by R. Harris (*Journal of Applied Polymer Science*, 1990) contains some data on the effect of reaction temperature on the molecular weight of resulting poly polyols. The data for eight experimental runs at temperatures 165°C and above are as follows:

Pot Temperature, x (°C)	Average Molecular Weight, y
165	808
176	940
188	1183
205	1545
220	2012
235	2362
250	2742
260	2935

Use a statistical package to help you complete the following (both the plotting and computations):

(a) What fraction of the observed raw variation in y is accounted for by a linear equation in x?

(b) Fit a linear relationship $y \approx \beta_0 + \beta_1 x$ to these data via least squares. About what change in average molecular weight seems to accompany a 1°C increase in pot temperature (at least over the experimental range of temperatures)?

(c) Compute and plot residuals from the linear relationship fit in (b). Discuss what they suggest about the appropriateness of that fitted equation. (Plot residuals versus x, residuals versus $\hat{y}$, and make a normal plot of them.)

(d) These data came from an experiment where the investigator managed the value of x. There is a fairly glaring weakness in the experimenter's data collection efforts. What is it?

(e) Based on your analysis of these data, what average molecular weight would you predict for an additional reaction run at 188°C? At 200°C? Why would or wouldn't you be willing to make a similar prediction of average molecular weight if the reaction is run at 70°C?

4. Upon changing measurement scales, nonlinear relationships between two variables can sometimes be made linear. The article "The Effect of Experimental Error on the Determination of the Optimum Metal-Cutting Conditions" by Ermer and Wu (*The Journal of Engineering for Industry*, 1967) contains a data set gathered in a study of tool life in a turning operation. The data here are part of that data set.

Cutting Speed, x (sfpm)	Tool Life, y (min)
800	1.00, 0.90, 0.74, 0.66
700	1.00, 1.20, 1.50, 1.60
600	2.35, 2.65, 3.00, 3.60
500	6.40, 7.80, 9.80, 16.50
400	21.50, 24.50, 26.00, 33.00

(a) Plot y versus x and calculate R^2 for fitting a linear function of x to y. Does the relationship $y \approx \beta_0 + \beta_1 x$ look like a reasonable explanation of tool life in terms of cutting speed?

(b) Take natural logs of both x and y and repeat part (a) with these log cutting speeds and log tool lives.

(c) Using the logged variables as in (b), fit a linear relationship between the two variables using least squares. Based on this fitted equation, what tool life would you predict for a cutting speed of 550? What approximate relationship between x and y is implied by a linear approximate relationship between $\ln(x)$ and $\ln(y)$? (Give an equation for this relationship.) By the way, Taylor's equation for tool life is $yx^{\alpha} = C$.

4.2 Fitting Curves and Surfaces by Least Squares

The basic ideas introduced in Section 4.1 generalize to produce a powerful engineering tool: **multiple linear regression**, which is introduced in this section. (Since the term *regression* may seem obscure, the more descriptive terms **curve fitting** and **surface fitting** will be used here, at least initially.)

This section first covers fitting curves defined by polynomials and other functions that are linear in their parameters to (x, y) data. Next comes the fitting of surfaces to data where a response y depends upon the values of several variables $x_1, x_2, \ldots, x_k$. In both cases, the discussion will stress how useful R^2 and residual plotting are and will consider the question of choosing between possible fitted equations. Lastly, we include some additional practical cautions.

4.2.1 Curve Fitting by Least Squares

In the previous section, a straight line did a reasonable job of describing the pressure/density data. But in the fly ash study, the ammonium phosphate/compressive strength data were very poorly described by a straight line. This section first investigates the possibility of fitting curves more complicated than a straight line to (x, y) data. As an example, an attempt will be made to find a better equation for describing the fly ash data.

A natural generalization of the linear equation

$$y \approx \beta_0 + \beta_1 x \tag{4.11}$$

is the **polynomial equation**

$$y \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k \tag{4.12}$$

The least squares fitting of equation (4.12) to a set of n pairs (x_i, y_i) is conceptually only slightly more difficult than the task of fitting equation (4.11). The function of $k + 1$ variables

$$S(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k)\right)^2$$

must be minimized. Upon setting the partial derivatives of $S(\beta_0, \beta_1, \ldots, \beta_k)$ equal to 0, the set of **normal equations** is obtained for this least squares problem, generalizing the pair of equations (4.4) and (4.5). There are $k + 1$ linear equations in the $k + 1$ unknowns $\beta_0, \beta_1, \ldots, \beta_k$. And typically, they can be solved simultaneously for a single set of values, $b_0, b_1, \ldots, b_k$, minimizing $S(\beta_0, \beta_1, \ldots, \beta_k)$. The mechanics of that solution are carried out using a **multiple linear regression program**.

Example 3
(*Example 2 continued*)

More on the Fly Ash Data of Table 4.3

Return to the fly ash study of B. Roth. A quadratic equation might fit the data better than the linear one. So consider fitting the $k = 2$ version of equation (4.12)

$$y \approx \beta_0 + \beta_1 x + \beta_2 x^2 \qquad \textbf{(4.13)}$$

to the data of Table 4.3. Printout 2 shows the MINITAB run. (After entering x and y values from Table 4.3 into two columns of the worksheet, an additional column was created by squaring the x values.)

Printout 2 Quadratic Fit to the Fly Ash Data

```
Regression Analysis


The regression equation is
y = 1243 + 383 x - 76.7 x**2

Predictor        Coef       StDev          T        P
Constant      1242.89       42.98      28.92    0.000
x              382.67       40.43       9.46    0.000
x**2          -76.661       7.762      -9.88    0.000

S = 82.14       R-Sq = 86.7%     R-Sq(adj) = 84.9%


Analysis of Variance


Source            DF          SS          MS         F        P
Regression         2      658230      329115     48.78    0.000
Residual Error    15      101206        6747
Total             17      759437

Source       DF      Seq SS
x             1          21
x**2          1      658209
```

The fitted quadratic equation is

$$\hat{y} = 1242.9 + 382.7x - 76.7x^2$$

Figure 4.9 shows the fitted curve sketched on a scatterplot of the (x, y) data. Although the quadratic curve is not an altogether satisfactory summary of Roth's data, it does a much better job of following the trend of the data than the line sketched in Figure 4.5.

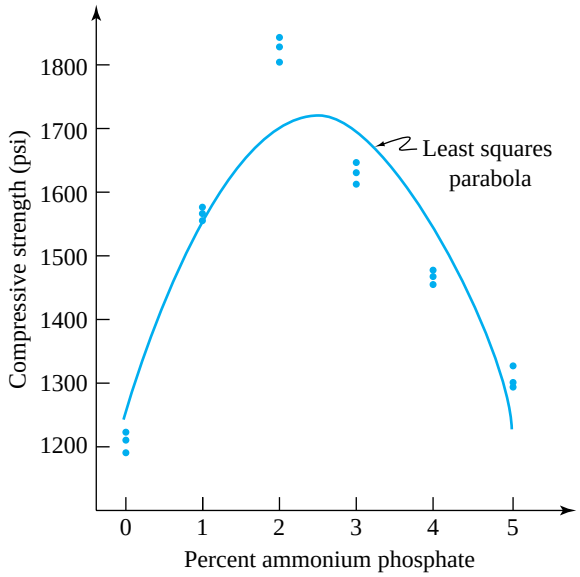

Figure 4.9 Scatterplot and fitted quadratic for the fly ash data

The previous section showed that when fitting a line to (x, y) data, it is helpful to quantify the goodness of that fit using R^2. The coefficient of determination can also be used when fitting a polynomial of form (4.12). Recall once more from Definition 3 that

Coefficient of determination

$$R^2 = \frac{\sum(y_i - \bar{y})^2 - \sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \qquad \textbf{(4.14)}$$

is the fraction of the raw variability in y accounted for by the fitted equation. Calculation by hand from formula (4.14) is possible, but of course the easiest way to obtain R^2 is to use a computer package.

Example 3 (*continued*)

Consulting Printout 2, it can be seen that the equation $\hat{y} = 1242.9 + 382.7x - 76.7x^2$ produces $R^2 = .867$. So 86.7% of the raw variability in compressive strength is accounted for using the fitted quadratic. The sample correlation between the observed strengths y_i and fitted strengths $\hat{y}_i$ is $+\sqrt{.867} = .93$.

Comparing what has been done in the present section to what was done in Section 4.1, it is interesting that for the fitting of a line to the fly ash data, R^2 obtained there was only .000 (to three decimal places). The present quadratic is a remarkable improvement over a linear equation for summarizing these data.

A natural question to raise is "What about a cubic version of equation (4.12)?" Printout 3 shows some results of a MINITAB run made to investigate this possibility, and Figure 4.10 shows a scatterplot of the data and a plot of the fitted cubic

Example 3 (*continued*)

equation. (x values were squared and cubed to provide x, x^2, and x^3 for each y value to use in the fitting.)

Printout 3 Cubic Fit to the Fly Ash Data

```
Regression Analysis


The regression equation is
y = 1188 + 633 x - 214 x**2 + 18.3 x**3

Predictor         Coef       StDev          T        P
Constant      1188.05       28.79      41.27    0.000
x              633.11       55.91      11.32    0.000
x**2          -213.77       27.79      -7.69    0.000
x**3           18.281       3.649       5.01    0.000

S = 50.88       R-Sq = 95.2%     R-Sq(adj) = 94.2%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      723197      241066     93.13    0.000
Residual Error    14       36240        2589
Total             17      759437
```

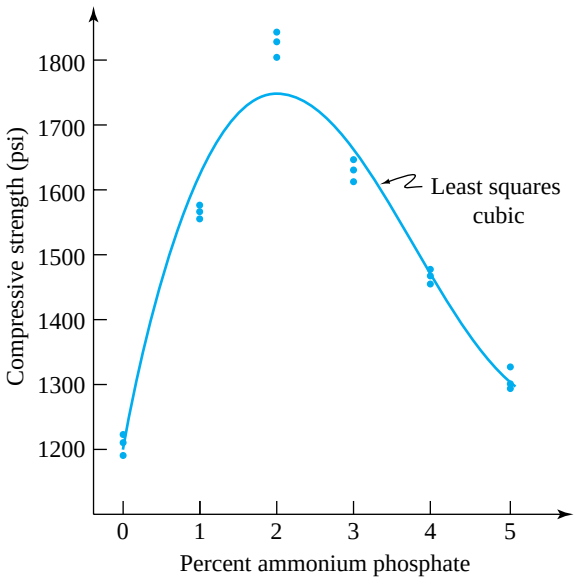

Figure 4.10 Scatterplot and fitted cubic for the fly ash data

R^2 for the cubic equation is .952, somewhat larger than for the quadratic. But it is fairly clear from Figure 4.10 that even a cubic polynomial is not totally satisfactory as a summary of these data. In particular, both the fitted quadratic in Figure 4.9 and the fitted cubic in Figure 4.10 fail to fit the data adequately near an ammonium phosphate level of 2%. Unfortunately, this is where compressive strength is greatest—precisely the area of greatest practical interest.

The example illustrates that R^2 is not the only consideration when it comes to judging the appropriateness of a fitted polynomial. The examination of plots is also important. Not only scatterplots of y versus x with superimposed fitted curves but plots of residuals can be helpful. This can be illustrated on a data set where y is expected to be nearly perfectly quadratic in x.

Example 4

Analysis of the Bob Drop Data of Section 1.4

Consider again the experimental determination of the acceleration due to gravity (through the dropping of the steel bob) data given in Table 1.4 and reproduced here in the first two columns of Table 4.7. Recall that the positions y were recorded at $\frac{1}{60}$ sec intervals beginning at some unknown time t_0 (less than $\frac{1}{60}$ sec) after the bob was released. Since Newtonian mechanics predicts the bob displacement to be

$$\text{displacement} = \frac{gt^2}{2}$$

one expects

$$\begin{aligned} y &\approx \frac{1}{2}g\left(t_0 + \frac{1}{60}(x-1)\right)^2 \\ &= \frac{g}{2}\left(\frac{x}{60}\right)^2 + g\left(t_0 - \frac{1}{60}\right)\left(\frac{x}{60}\right) + \frac{g}{2}\left(t_0 - \frac{1}{60}\right)^2 \\ &= \frac{g}{7200}x^2 + \frac{g}{60}\left(t_0 - \frac{1}{60}\right)x + \frac{g}{2}\left(t_0 - \frac{1}{60}\right)^2 \end{aligned} \qquad \textbf{(4.15)}$$

That is, y is expected to be approximately quadratic in x and, indeed, the plot of (x, y) points in Figure 1.8 (p. 22) appears to have that character.

As a slight digression, note that expression () shows that if a quadratic is fitted to the data in Table 4.7 via least squares,

$$\hat{y} = b_0 + b_1 x + b_2 x^2 \qquad \textbf{(4.16)}$$

is obtained and an experimentally determined value of g (in mm/sec^2) will be

Example 4
(*continued*)

Table 4.7
Data, Fitted Values, and Residuals for a Quadratic Fit to the Bob Displacement

x, Point Number	y, Displacement	$\hat{y}$, Fitted Displacement	e, Residual
1	.8	.95	−.15
2	4.8	4.56	.24
3	10.8	10.89	−.09
4	20.1	19.93	.17
5	31.9	31.70	.20
6	45.9	46.19	−.29
7	63.3	63.39	−.09
8	83.1	83.31	−.21
9	105.8	105.96	−.16
10	131.3	131.32	−.02
11	159.5	159.40	.10
12	190.5	190.21	.29
13	223.8	223.73	.07
14	260.0	259.97	.03
15	299.2	298.93	.27
16	340.5	340.61	−.11
17	385.0	385.01	−.01
18	432.2	432.13	.07
19	481.8	481.97	−.17
20	534.2	534.53	−.33
21	589.8	589.80	.00
22	647.7	647.80	−.10
23	708.8	708.52	.28

$7200b_2$. This is in fact how the value 9.79 m/sec^2, quoted in Section 1.4, was obtained.

A multiple linear regression program fits equation (4.16) to the bob drop data giving

$$\hat{y} = .0645 - .4716x + 1.3597x^2$$

(from which $g \approx 9790$ mm/sec^2) with R^2 that is 1.0 to 6 decimal places. Residuals for this fit can be calculated using Definition 4 and are also given in Table 4.7. Figure 4.11 is a normal plot of the residuals. It is reasonably linear and thus not remarkable (except for some small suggestion that the largest residual or two may not be as extreme as might be expected, a circumstance that suggests no obvious physical explanation).

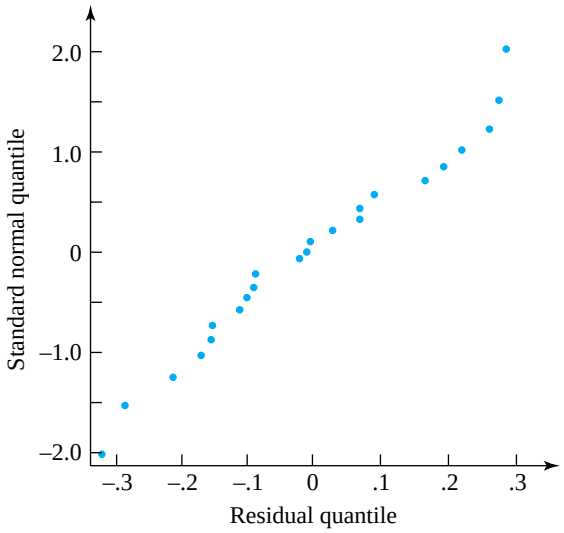

Figure 4.11 Normal plot of the residuals from a quadratic fit to the bob drop data

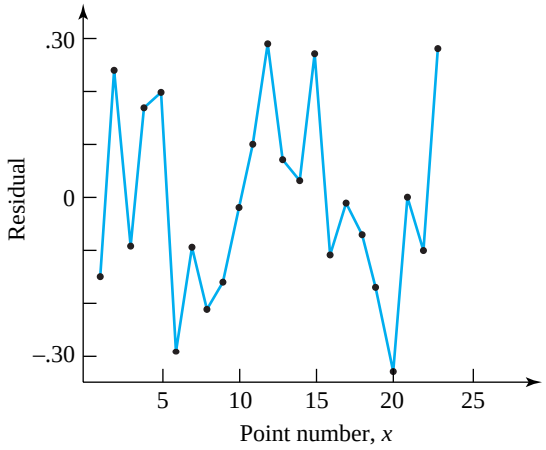

Figure 4.12 Plot of the residuals from the bob drop quadratic fit vs. x

However, a plot of residuals versus x (the time variable) is interesting. Figure 4.12 is such a plot, where successive plotted points have been connected with line segments. There is at least a hint in Figure 4.12 of a **cyclical pattern** in the residuals. Observed displacements are alternately too big, too small, too big, etc. It would be a good idea to look at several more tapes, to see if a cyclical pattern appears consistently, before seriously thinking about its origin. But should the

Example 4 (*continued*)

pattern suggested by Figure 4.12 reappear consistently, it would indicate that something in the mechanism generating the 60 cycle current may cause cycles to be alternately slightly shorter then slightly longer than $\frac{1}{60}$ sec. The practical implication of this would be that if a better determination of g were desired, the regularity of the AC current waveform is one matter to be addressed.

What if a polynomial doesn't fit (x, y) data?

Examples 3 and 4 (respectively) illustrate only partial success and then great success in describing an (x, y) data set by means of a polynomial equation. Situations like Example 3 obviously do sometimes occur, and it is reasonable to wonder what to do when they happen. There are two simple things to keep in mind.

For one, although a polynomial may be unsatisfactory as a global description of a relationship between x and y, it may be quite adequate **locally**—i.e., for a relatively restricted range of x values. For example, in the fly ash study, the quadratic representation of compressive strength as a function of percent ammonium phosphate is not appropriate over the range 0 to 5%. But having identified the region around 2% as being of practical interest, it would make good sense to conduct a follow-up study concentrating on (say) 1.5 to 2.5% ammonium phosphate. It is quite possible that a quadratic fit only to data with $1.5 \leq x \leq 2.5$ would be both adequate and helpful as a summarization of the follow-up data.

The second observation is that the terms $x, x^2, x^3, \ldots, x^k$ in equation (4.12) can be replaced by any (known) functions of x and what we have said here will remain essentially unchanged. The normal equations will still be $k + 1$ linear equations in $\beta_0, \beta_1, \ldots, \beta_k$, and a multiple linear regression program will still produce least squares values $b_0, b_1, \ldots, b_k$. This can be quite useful when there are theoretical reasons to expect a particular (nonlinear but) simple functional relationship between x and y. For example, Taylor's equation for tool life is of the form

$$y \approx \alpha x^{\beta}$$

for y tool life (e.g., in minutes) and x the cutting speed used (e.g., in sfpm). Taking logarithms,

$$\ln(y) \approx \ln(\alpha) + \beta \ln(x)$$

This is an equation for $\ln(y)$ that is linear in the parameters $\ln(\alpha)$ and β involving the variable $\ln(x)$. So, presented with a set of (x, y) data, empirical values for α and β could be determined by

1. taking logs of both x's and y's,
2. fitting the linear version of (4.12), and
3. identifying $\ln(\alpha)$ with β_0 (and thus α with $\exp(\beta_0)$) and β with β_1.

4.2.2 Surface Fitting by Least Squares

It is a small step from the idea of fitting a line or a polynomial curve to realizing that essentially the same methods can be used to summarize the effects of several different quantitative variables $x_1, x_2, \ldots, x_k$ on some response y. Geometrically the problem is fitting a surface described by an equation

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \qquad \textbf{(4.17)}$$

to the data using the least squares principle. This is pictured for a $k = 2$ case in Figure 4.13, where six (x_1, x_2, y) data points are pictured in three dimensions, along with a possible fitted surface of the form (4.17). To fit a surface defined by equation (4.17) to a set of n data points $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ via least squares, the function of $k + 1$ variables

$$S(\beta_0, \beta_1, \beta_2, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})\right)^2$$

must be minimized by choice of the coefficients $\beta_0, \beta_1, \ldots, \beta_k$. Setting partial derivatives with respect to the β's equal to 0 gives normal equations generalizing equations (4.4) and (4.5). The solution of these $k + 1$ linear equations in the $k + 1$ unknowns $\beta_0, \beta_1, \ldots, \beta_k$ is the first task of a multiple linear regression program. The fitted coefficients $b_0, b_1, \ldots, b_k$ that it produces minimize $S(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$.

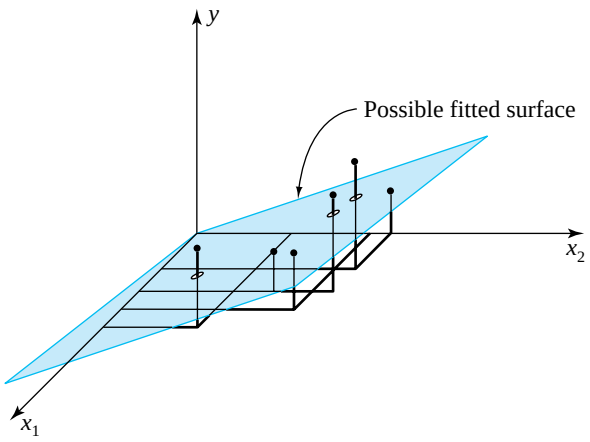

Figure 4.13 Six data points (x_1, x_2, y) and a possible fitted plane

Example 5

Surface Fitting and Brownlee's Stack Loss Data

Table 4.8 contains part of a set of data on the operation of a plant for the oxidation of ammonia to nitric acid that appeared first in Brownlee's *Statistical Theory and Methodology in Science and Engineering*. In plant operation, the nitric oxides produced are absorbed in a countercurrent absorption tower.

The air flow variable, x_1, represents the rate of operation of the plant. The acid concentration variable, x_3, is the percent circulating minus 50 times 10. The response variable, y, is ten times the percentage of ingoing ammonia that escapes from the absorption column unabsorbed (i.e., an inverse measure of overall plant efficiency). For purposes of understanding, predicting, and possibly ultimately optimizing plant performance, it would be useful to have an equation describing how y depends on x_1, x_2, and x_3. Surface fitting via least squares is a method of developing such an empirical equation.

Printout 4 shows results from a MINITAB run made to obtain a fitted equation of the form

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

Table 4.8
Brownlee's Stack Loss Data

i, Observation Number	x_{1i}, Air Flow	x_{2i}, Cooling Water Inlet Temperature	x_{3i}, Acid Concentration	y_i, Stack Loss
1	80	27	88	37
2	62	22	87	18
3	62	23	87	18
4	62	24	93	19
5	62	24	93	20
6	58	23	87	15
7	58	18	80	14
8	58	18	89	14
9	58	17	88	13
10	58	18	82	11
11	58	19	93	12
12	50	18	89	8
13	50	18	86	7
14	50	19	72	8
15	50	19	79	8
16	50	20	80	9
17	56	20	82	15

The equation produced by the program is

$$\hat{y} = -37.65 + .80x_1 + .58x_2 - .07x_3 \qquad \textbf{(4.18)}$$

Interpreting fitted coefficients from a multiple regression

with $R^2 = .975$. The coefficients in this equation can be thought of as rates of change of stack loss with respect to the individual variables x_1, x_2, and x_3, holding the others fixed. For example, $b_1 = .80$ can be interpreted as the increase in stack loss y that accompanies a one-unit increase in air flow x_1 if inlet temperature x_2 and acid concentration x_3 are held fixed. The signs on the coefficients indicate whether y tends to increase or decrease with increases in the corresponding x. For example, the fact that b_1 is positive indicates that the higher the rate at which the plant is run, the larger y tends to be (i.e., the less efficiently the plant operates). The large value of R^2 is a preliminary indicator that the equation (4.18) is an effective summarization of the data.

Printout 4 Multiple Regression for the Stack Loss Data

```
Regression Analysis


The regression equation is
stack = - 37.7 + 0.798 air + 0.577 water - 0.0671 acid

Predictor        Coef       StDev          T        P
Constant      -37.652       4.732      -7.96    0.000
air           0.79769     0.06744      11.83    0.000
water          0.5773      0.1660       3.48    0.004
acid         -0.06706     0.06160      -1.09    0.296

S = 1.253       R-Sq = 97.5%     R-Sq(adj) = 96.9%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      795.83      265.28    169.04    0.000
Residual Error    13       20.40        1.57
Total             16      816.24

Source       DF      Seq SS
air           1      775.48
water         1       18.49
acid          1        1.86

Unusual Observations
Obs        air      stack         Fit   StDev Fit    Residual    St Resid
 10       58.0     11.000      13.506       0.552      -2.506       -2.23R

R denotes an observation with a large standardized residual
```

Although the mechanics of fitting equations of the form (4.17) to multivariate data are relatively straightforward, the **choice and interpretation of appropriate equations** are not so clear-cut. Where many x variables are involved, the number

The goal of multiple regression

of potential equations of form (4.17) is huge. To make matters worse, there is no completely satisfactory way to plot multivariate $(x_1, x_2, \ldots, x_k, y)$ data to "see" how an equation is fitting. About all that we can do at this point is to (1) offer the broad advice that what is wanted is *the simplest equation that adequately fits the data* and then (2) provide examples of how R^2 and residual plotting can be helpful tools in clearing up the difficulties that arise.

Example 5 (*continued*)

In the context of the nitrogen plant, it is sensible to ask whether all three variables, x_1, x_2, and x_3, are required to adequately account for the observed variation in y. For example, the behavior of stack loss might be adequately explained using only one or two of the three x variables. There would be several consequences of practical engineering importance if this were so. For one, in such a case, a simple or **parsimonious** version of equation (4.17) could be used in describing the oxidation process. And if a variable is not needed to predict y, then it is possible that the expense of measuring it might be saved. Or, if a variable doesn't seem to have much impact on y (because it doesn't seem to be essential to include it when writing an equation for y), it may be possible to choose its level on purely economic grounds, without fear of degrading process performance.

As a means of investigating whether indeed some subset of x_1, x_2, and x_3 is adequate to explain stack loss behavior, R^2 values for equations based on all possible subsets of x_1, x_2, and x_3 were obtained and placed in Table 4.9. This shows, for example, that 95% of the raw variability in y can be accounted for using a linear equation in only the air flow variable x_1. Use of both x_1 and the water temperature variable x_2 can account for 97.3% of the raw variability in stack loss. Inclusion of x_3, the acid concentration variable, in an equation already involving x_1 and x_2, increases R^2 only from .973 to .975.

If identifying a simple equation for stack loss that seems to fit the data well is the goal, the message in Table 4.9 would seem to be "Consider an x_1 term first, and then possibly an x_2 term." On the basis of R^2, including an x_3 term in an equation for y seems unnecessary. And in retrospect, this is entirely consistent with the character of the fitted equation (4.18): x_3 varies from 72 to 93 in the original data set, and this means that $\hat{y}$ changes only a total amount

$$.07(93 - 72) \approx 1.5$$

based on changes in x_3. (Remember that $.07 = b_3 =$ the fitted rate of change in y with respect to x_3.) 1.5 is relatively small in comparison to the range in the observed y values.

Once R^2 values have been used to identify potential simplifications of the equation

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$$

these can and should go through thorough residual analyses before they are adopted as data summaries. As an example, consider a fitted equation involving

Table 4.9
R^2's for Equations Predicting Stack Loss

Equation Fit	R^2
$y \approx \beta_0 + \beta_1 x_1$	.950
$y \approx \beta_0 + \beta_2 x_2$	.695
$y \approx \beta_0 + \beta_3 x_3$	.165
$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$	.973
$y \approx \beta_0 + \beta_1 x_1 + \beta_3 x_3$	.952
$y \approx \beta_0 + \beta_2 x_2 + \beta_3 x_3$	.706
$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$	.975

x_1 and x_2. A multiple linear regression program can be used to produce the fitted equation

$$\hat{y} = -42.00 - .78x_1 + .57x_2 \qquad \textbf{(4.19)}$$

Dropping variables from a fitted equation typically changes coefficients

(Notice that b_0, b_1, and b_2 in equation (4.19) differ somewhat from the corresponding values in equation (4.18). That is, equation (4.19) was not obtained from equation (4.18) by simply dropping the last term in the equation. In general, the values of the coefficients b will change depending on which x variables are and are not included in the fitting.)

Residuals for equation (4.19) can be computed and plotted in any number of potentially useful ways. Figure 4.14 shows a normal plot of the residuals and three other plots of the residuals against, respectively, x_1, x_2, and $\hat{y}$. There are no really strong messages carried by the plots in Figure 4.14 except that the data set contains one unusually large x_1 value and one unusually large $\hat{y}$ (which corresponds to the large x_1). But there is enough of a curvilinear "up-then-down-then-back-up-again" pattern in the plot of residuals against x_1 to suggest the possibility of adding an x_1^2 term to the fitted equation (4.19).

You might want to verify that fitting the equation

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2$$

to the data of Table 4.8 yields approximately

$$\hat{y} = -15.409 - .069x_1 + .528x_2 + .007x_1^2 \qquad \textbf{(4.20)}$$

with corresponding $R^2 = .980$ and residuals that show even less of a pattern than those for the fitted equation (4.19). In particular, the hint of curvature on the plot of residuals versus x_1 for equation (4.19) is not present in the corresponding plot for equation (4.20). Interestingly, looking back over this example, one sees that fitted equation (4.20) has a better R^2 value than even fitted equation (4.18), in

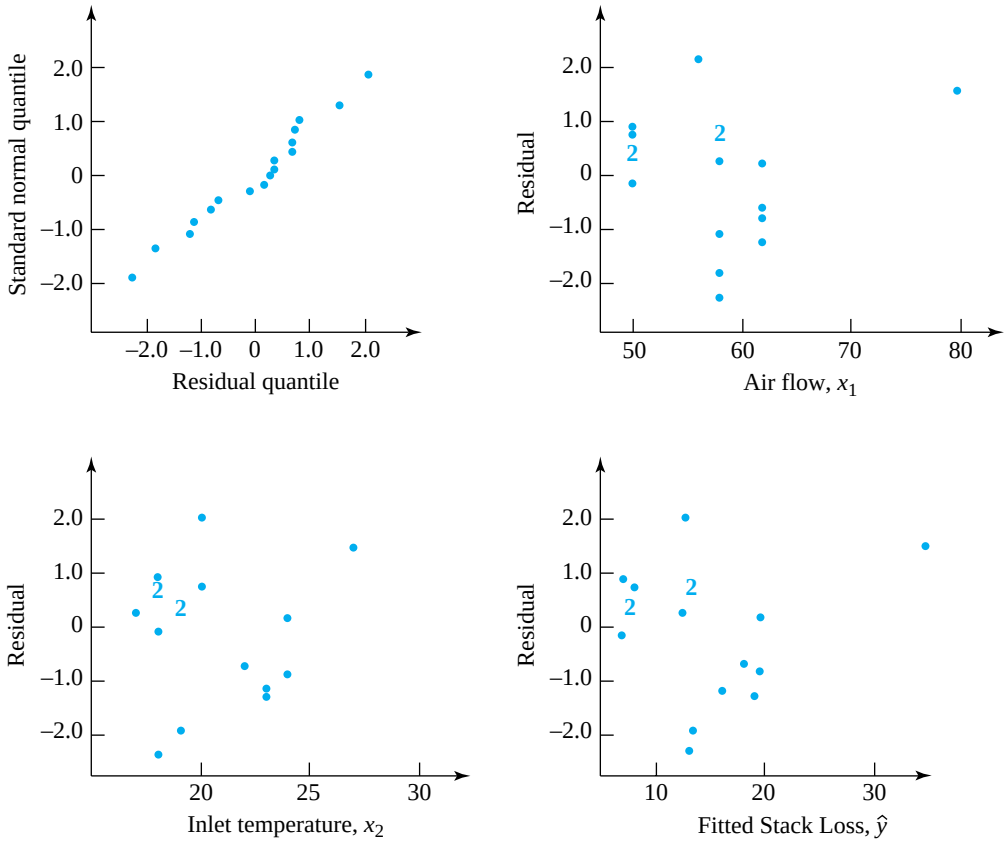

Figure 4.14 Plots of residuals from a two-variable equation fit to the stack loss data ($\hat{y} = -42.00 - .78x_1 + .57x_2$)

Example 5 *(continued)*

spite of the fact that equation (4.18) involves the process variable x_3 and equation (4.20) does not.

Equation (4.20) is somewhat more complicated than equation (4.19). But because it still really only involves two different input x's and also eliminates the slight pattern seen on the plot of residuals for equation (4.19) versus x_1, it seems an attractive choice for summarizing the stack loss data. A two-dimensional representation of the fitted surface defined by equation (4.20) is given in Figure 4.15. The slight curvature on the plotted curves is a result of the x_1^2 term appearing in equation (4.20). Since most of the data have x_1 from 50 to 62 and x_2 from 17 to 24, the curves carry the message that over these ranges, changes in x_1 seem to produce larger changes in stack loss than do changes in x_2. This conclusion is consistent with the discussion centered around Table 4.9.

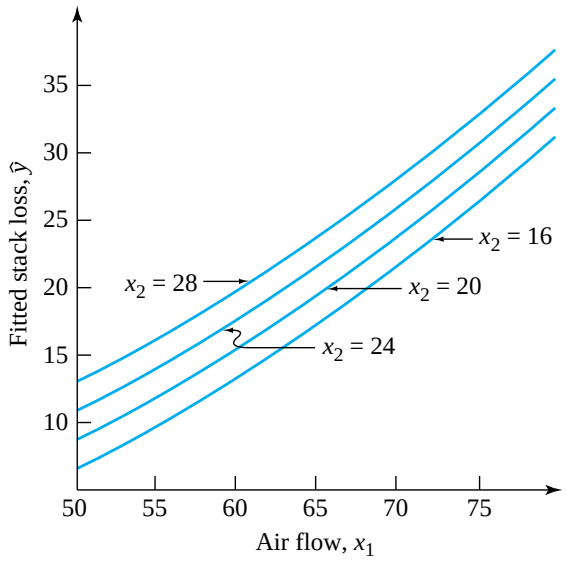

Figure 4.15 Plots of fitted stack loss from equation (4.20)

Common residual plots in multiple regression

The plots of residuals used in Example 5 are typical. They are

1. normal plots of residuals,
2. plots of residuals against all x variables,
3. plots of residuals against $\hat{y}$,
4. plots of residuals against time order of observation, and
5. plots of residuals against variables (like machine number or operator) not used in the fitted equation but potentially of importance.

All of these can be used to help assess the appropriateness of surfaces fit to multivariate data, and they all have the potential to tell an engineer something not previously discovered about a set of data and the process that generated them.

Earlier in this section, there was a discussion of the fact that an "x term" in the equations fitted via least squares can be a known function (e.g., a logarithm) of a basic process variable. In fact, it is frequently helpful to allow an "x term" in equation (4.17) (page 149) to be a known function of *several* basic process variables. The next example illustrates this point.

Example 6

Lift/Drag Ratio for a Three-Surface Configuration

P. Burris studied the effects of the positions relative to the wing of a canard (a forward lifting surface) and tail on the lift/drag ratio for a three-surface configuration. Part of his data are given in Table 4.10, where

x_1 = canard placement in inches above the plane defined by the main wing
x_2 = tail placement in inches above the plane defined by the main wing

(The front-to-rear positions of the three surfaces were constant throughout the study.)

A straightforward least squares fitting of the equation

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

to these data produces R^2 of only .394. Even the addition of squared terms in both x_1 and x_2, i.e., the fitting of

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2$$

produces an increase in R^2 to only .513. However, Printout 5 shows that fitting the equation

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

yields $R^2 = .641$ and the fitted relationship

$$\hat{y} = 3.4284 + .5361 x_1 + .3201 x_2 - .5042 x_1 x_2 \qquad \textbf{(4.21)}$$

Table 4.10
Lift/Drag Ratios for 9 Canard/Tail Position Combinations

x_1, Canard Position	x_2, Tail Position	y, Lift/Drag Ratio
−1.2	−1.2	.858
−1.2	0.0	3.156
−1.2	1.2	3.644
0.0	−1.2	4.281
0.0	0.0	3.481
0.0	1.2	3.918
1.2	−1.2	4.136
1.2	0.0	3.364
1.2	1.2	4.018

Printout 5 Multiple Regression for the Lift/Drag Ratio Data

```
Regression Analysis


The regression equation is
y = 3.43 + 0.536 x1 + 0.320 x2 - 0.504 x1*x2

Predictor         Coef       StDev          T        P
Constant        3.4284      0.2613      13.12    0.000
x1              0.5361      0.2667       2.01    0.101
x2              0.3201      0.2667       1.20    0.284
x1*x2          -0.5042      0.2722      -1.85    0.123

S = 0.7839      R-Sq = 64.1%     R-Sq(adj) = 42.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      5.4771      1.8257      2.97    0.136
Residual Error     5      3.0724      0.6145
Total              8      8.5495
```

(After reading x_1, x_2, and y values from Table 4.10 into columns of MINITAB's worksheet, x_1x_2 products were created and y fitted to the three predictor variables x_1, x_2, and x_1x_2 in order to create this printout.)

Figure 4.16 shows the nature of the fitted surface (4.21). Raising the canard (increasing x_1) has noticeably different predicted impacts on y, depending on the value of x_2 (the tail position). (It appears that the canard and tail should not be lined up—i.e., x_1 should not be near x_2. For large predicted response, one wants small x_1 for large x_2 and large x_1 for small x_2.) It is the cross-product term x_1x_2 in relationship (4.21) that allows the response curves to have different characters for different x_2 values. Without it, the slices of the fitted $(x_1, x_2, \hat{y})$ surface would be parallel for various x_2, much like the situation in Figure 4.15.

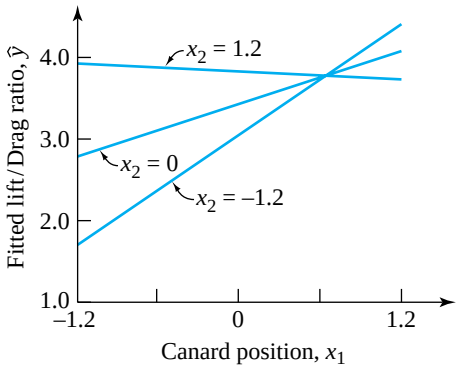

Figure 4.16 Plots of fitted lift/drag from equation (4.21)

Example 6 (*continued*)

Although the main new point of this example has by now been made, it probably should be mentioned that equation (4.21) is not the last word for fitting the data of Table 4.10. Figure 4.17 gives a plot of the residuals for relationship (4.21) versus canard position x_1, and it shows a strong curvilinear pattern. In fact, the fitted equation

$$\hat{y} = 3.9833 + .5361x_1 + .3201x_2 - .4843x_1^2 - .5042x_1x_2 \qquad \textbf{(4.22)}$$

provides $R^2 = .754$ and generally random-looking residuals. It can be verified by plotting $\hat{y}$ versus x_1 curves for several x_2 values that the fitted relationship (4.22) yields nonparallel parabolic slices of the fitted $(x_1, x_2, \hat{y})$ surface, instead of the nonparallel linear slices seen in Figure 4.16.

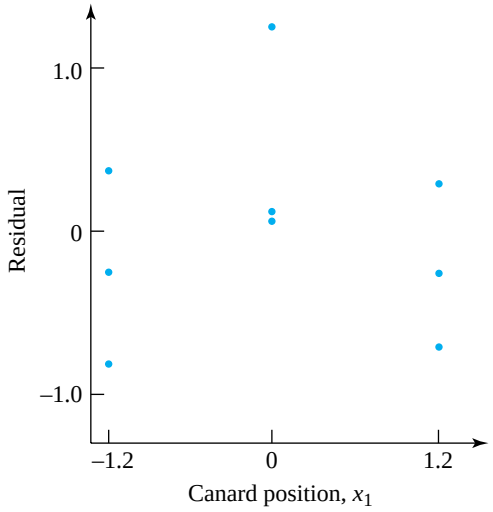

Figure 4.17 Plot of residuals from equation (4.21) vs. x_1

4.2.3 Some Additional Cautions

Least squares fitting of curves and surfaces is of substantial engineering importance—but it must be handled with care and thought. Before leaving the subject until Chapter 9, which explains methods of formal inference associated with it, a few more warnings must be given.

Extrapolation

First, it is necessary to warn of the dangers of extrapolation substantially outside the "range" of the $(x_1, x_2, \ldots, x_k, y)$ data. It is sensible to count on a fitted equation to describe the relation of y to a particular set of inputs $x_1, x_2, \ldots, x_k$ only if they are like the sets used to create the equation. The challenge surface fitting affords is

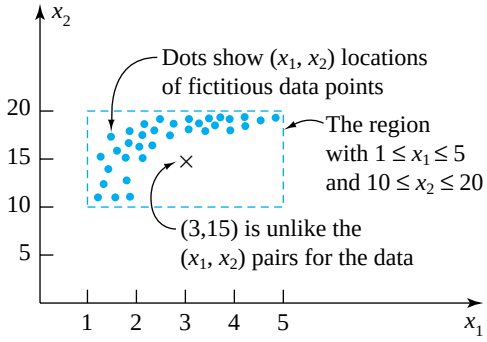

Figure 4.18 Hypothetical plot of (x_1, x_2) pairs

that when several different x variables are involved, it is difficult to tell whether a particular $(x_1, x_2, \ldots, x_k)$ vector is a large extrapolation. About all one can do is check to see that it comes close to matching some *single data point* in the set *on each coordinate* $x_1, x_2, \ldots, x_k$. It is not sufficient that there be some point with x_1 value near the one of interest, another point with x_2 value near the one of interest, etc. For example, having data with $1 \le x_1 \le 5$ and $10 \le x_2 \le 20$ doesn't mean that the (x_1, x_2) pair (3, 15) is necessarily like any of the pairs in the data set. This fact is illustrated in Figure 4.18 for a fictitious set of (x_1, x_2) values.

The influence of outlying data vectors

Another potential pitfall is that the fitting of curves and surfaces via least squares can be strongly affected by a few outlying or extreme data points. One can try to identify such points by examining plots and comparing fits made with and without the suspicious point(s).

Example 5 (*continued*)

Figure 4.14 earlier called attention to the fact that the nitrogen plant data set contains one point with an extreme x_1 value. Figure 4.19 is a scatterplot of (x_1, x_2) pairs for the data in Table 4.8 (page 150). It shows that by most qualitative standards, observation 1 in Table 4.8 is unusual or outlying.

If the fitting of equation (4.20) is redone using only the last 16 data points in Table 4.8, the equation

$$\hat{y} = -56.797 + 1.404x_1 + .601x_2 - .007x_1^2 \qquad \textbf{(4.23)}$$

and $R^2 = .942$ are obtained. Using equation (4.23) as a description of stack loss and limiting attention to x_1 in the range 50 to 62 could be considered. But it is possible to verify that though some of the coefficients (the b's) in equations (4.20) and (4.23) differ substantially, the two equations produce comparable $\hat{y}$ values for the 16 data points with x_1 between 50 and 62. In fact, the largest difference in fitted values is about .4. So, since point 1 in Table 4.8 doesn't

Example 5
(*continued*)

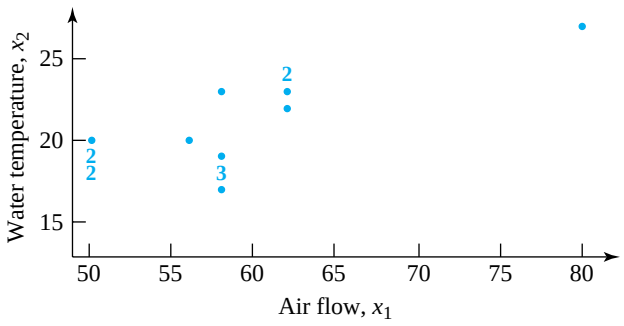

Figure 4.19 Plot of (x_1, x_2) pairs for the stack loss data

radically change predictions made using the fitted equation, it makes sense to leave it in consideration, adopt equation (4.20), and use it to describe stack loss for (x_1, x_2) pairs interior to the pattern of scatter in Figure 4.19.

Replication and surface fitting

A third warning has to do with the notion of replication (first discussed in Section 2.3). It is the fact that the fly ash data of Example 3 has several y's for each x that makes it so clear that even the quadratic and cubic curves sketched in Figures 4.9 and 4.10 are inadequate descriptions of the relationship between phosphate and strength. The fitted curves pass clearly outside the range of what look like believable values of y for some values of x. Without such replication, what is permissible variation about a fitted curve or surface can't be known with confidence. For example, the structure of the lift/drag data set in Example 6 is weak from this viewpoint. There is no replication represented in Table 4.10, so an external value for typical experimental precision would be needed in order to identify a fitted value as obviously incompatible with an observed one.

The nitrogen plant data set of Example 5 was presumably derived from a primarily observational study, where no conscious attempt was made to replicate (x_1, x_2, x_3) settings. However, points number 4 and 5 in Table 4.8 (page 150) do represent the replication of a single (x_1, x_2, x_3) combination and show a difference in observed stack loss of 1. And this makes the residuals for equation (4.20) (which range from -2.0 to 2.3) seem at least not obviously out of line.

Section 9.2 discusses more formal and precise ways of using data from studies with some replication to judge whether or not a fitted curve or surface misses some observed y's too badly. For now, simply note that among replication's many virtues is the fact that it allows more reliable judgments about the appropriateness of a fitted equation than are otherwise possible.

The possibility of overfitting

The fourth caution is that the notion of equation simplicity (*parsimony*) is important for reasons in addition to simplicity of interpretation and reduced expense involved in using the equation. It is also important from the point of view of typically giving smooth interpolation and not **overfitting** a data set. As a hypothetical example,

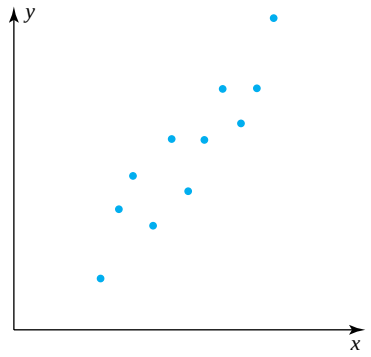

Figure 4.20 Scatterplot of 11 pairs (x, y)

consider the artificial, generally linear (x, y) data plotted in Figure 4.20. It would be possible to run a (wiggly) $k = 10$ version of the polynomial (4.12) through each of these points. But in most physical problems, such a curve would do a much worse job of predicting y at values of x not represented by a data point than would a simple fitted line. A tenth-order polynomial would overfit the data in hand.

Empirical models and engineering

As a final point in this section, consider how the methods discussed here fit into the broad picture of using models for attacking engineering problems. It must be said that physical theories of physics, chemistry, materials, etc. rarely produce equations of the forms (4.12) or (4.17). Sometimes pertinent equations from those theories can be rewritten in such forms, as was possible with Taylor's equation for tool life earlier in this section. But the majority of engineering applications of the methods in this section are to the large number of problems where no commonly known and simple physical theory is available, and a simple **empirical** description of the situation would be helpful. In such cases, the tool of least squares fitting of curves and surfaces can function as a kind of "mathematical French curve," allowing an engineer to develop approximate empirical descriptions of how a response y is related to system inputs $x_1, x_2, \ldots, x_k$.

Section 2 Exercises

1. Return to Exercise 3 of Section 4.1. Fit a quadratic relationship $y \approx \beta_0 + \beta_1 x + \beta_2 x^2$ to the data via least squares. By appropriately plotting residuals and examining R^2 values, determine the advisability of using a quadratic rather than a linear equation to describe the relationship between x and y. If a quadratic fitted equation is used, how does the predicted mean molecular weight at 200°C compare to that obtained in part (e) of the earlier exercise?

2. Here are some data taken from the article "Chemithermomechanical Pulp from Mixed High Density Hardwoods" by Miller, Shankar, and Peterson (*Tappi Journal*, 1988). Given are the percent NaOH used as a pretreatment chemical, x_1, the pretreatment time in minutes, x_2, and the resulting value of a specific surface area variable, y (with units of cm^3/g), for nine batches of pulp produced from a mixture of hardwoods at a treatment temperature of 75°C in mechanical pulping.

% NaOH, x_1	Time, x_2	Specific Surface Area, y
3.0	30	5.95
3.0	60	5.60
3.0	90	5.44
9.0	30	6.22
9.0	60	5.85
9.0	90	5.61
15.0	30	8.36
15.0	60	7.30
15.0	90	6.43

(a) Fit the approximate relationship $y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$ to these data via least squares. Interpret the coefficients b_1 and b_2 in the fitted equation. What fraction of the observed raw variation in y is accounted for using this equation?

(b) Compute and plot residuals for your fitted equation from (a). Discuss what these plots indicate about the adequacy of your fitted equation. (At a minimum, you should plot residuals against all of x_1, x_2, and $\hat{y}$ and normal-plot the residuals.)

(c) Make a plot of y versus x_1 for the nine data points and sketch on that plot the three different linear functions of x_1 produced by setting x_2 first at 30, then 60, and then 90 in your fitted equation from (a). How well do fitted responses appear to match observed responses?

(d) What specific surface area would you predict for an additional batch of pulp of this type produced using a 10% NaOH treatment for a time of 70 minutes? Would you be willing to make a similar prediction for 10% NaOH used for 120 minutes based on your fitted equation? Why or why not?

(e) There are many other possible approximate relationships that might be fitted to these data via least squares, one of which is $y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$. Fit this equation to the preceding data and compare the resulting coefficient of determination to the one found in (a). On the basis of these alone, does the use of the more complicated equation seem necessary?

(f) For the equation fit in part (e), repeat the steps of part (c) and compare the plot made here to the one made earlier.

(g) What is an intrinsic weakness of this real published data set?

(h) What terminology (for data structures) introduced in Section 1.2 describes this data set? It turns out that since the data set has this special structure and all nine sample sizes are the same (i.e., are all 1), some special relationships hold between the equation fit in (a) and what you get by separately fitting linear equations in x_1 and then in x_2 to the y data. Fit such one-variable linear equations and compare coefficients and R^2 values to what you obtained in (a). What relationships exist between these?

4.3 Fitted Effects for Factorial Data

The previous two sections have centered on the least squares fitting of equations to data sets where a quantitative response y is presumed to depend on the levels $x_1, x_2, \ldots, x_k$ of *quantitative factors*. In many engineering applications, at least some of the system "knobs" whose effects must be assessed are basically *qualitative* rather than quantitative. When a data set has complete factorial structure (review the meaning of this terminology in Section 1.2), it is still possible to describe it in terms of an equation. This equation involves so-called fitted factorial effects. Sometimes, when a few of these fitted effects dominate the rest, a parsimonious version of this

equation can adequately describe the data and have intuitively appealing and understandable interpretations. The use of simple plots and residuals will be discussed, as tools helpful in assessing whether such a simple structure holds.

The discussion begins with the 2-factor case, then considers three (or, by analogy, more) factors. Finally, the special case where each factor has only two levels is discussed.

4.3.1 Fitted Effects for 2-Factor Studies

Example 9 of Chapter 3 (page 101) illustrated how informative a plot of sample means versus levels of one of the factors can be in a 2-factor study. Such plotting is always the place to begin in understanding the story carried by two-way factorial data. In addition, it is helpful to calculate the factor level (marginal) averages of the sample means and the grand average of the sample means. For factor A having I levels and factor B having J levels, the following notation will be used:

Notation for sample means and their averages

$$\bar{y}_{ij} = \text{the sample mean response when factor A is at level } i \text{ and factor B is at level } j$$

$$\bar{y}_{i.} = \frac{1}{J}\sum_{j=1}^{J} \bar{y}_{ij}$$

$$= \text{the average sample mean when factor A is at level } i$$

$$\bar{y}_{.j} = \frac{1}{I}\sum_{i=1}^{I} \bar{y}_{ij}$$

$$= \text{the average sample mean when factor B is at level } j$$

$$\bar{y}_{..} = \frac{1}{IJ}\sum_{i,j} \bar{y}_{ij}$$

$$= \text{the grand average sample mean}$$

The $\bar{y}_{i.}$ and $\bar{y}_{.j}$ are row and column averages when one thinks of the $\bar{y}_{ij}$ laid out in a two-dimensional format, as shown in Figure 4.21.

Example 7

Joint Strengths for Three Different Joint Types in Three Different Woods

Kotlers, MacFarland, and Tomlinson studied the tensile strength of three different types of joints made on three different types of wood. Butt, lap, and beveled joints were made in nominal $1'' \times 4'' \times 12''$ pine, oak, and walnut specimens using a resin glue. The original intention was to test two specimens of each Joint Type/Wood Type combination. But one operator error and one specimen failure not related to its joint removed two of the original data points from consideration and gave the data in Table 4.11. These data have complete 3×3 factorial struc-

		Factor B			
		Level 1	Level 2		Level J
Factor A	Level 1	$\bar{y}_{11}$	$\bar{y}_{12}$	• • •	$\bar{y}_{1J}$	$\bar{y}_{1.}$
	Level 2	$\bar{y}_{21}$	$\bar{y}_{22}$	• • •	$\bar{y}_{2J}$	$\bar{y}_{2.}$
	⋮	⋮	⋮		⋮	⋮
	Level I	$\bar{y}_{I1}$	$\bar{y}_{I2}$	• • •	$\bar{y}_{IJ}$	$\bar{y}_{I.}$
		$\bar{y}_{.1}$	$\bar{y}_{.2}$	• • •	$\bar{y}_{.J}$	$\bar{y}_{..}$

Figure 4.21 Cell sample means and row, column, and grand average sample means for a two-way factorial

Example 7 (*continued*)

Table 4.11
Measured Strengths of 16 Wood Joints

Specimen	Joint	Wood	y, Stress at Failure (psi)
1	beveled	oak	1518
2	butt	pine	829
3	beveled	walnut	2571
4	butt	oak	1169
5	beveled	oak	1927
6	beveled	pine	1348
7	lap	walnut	1489
8	beveled	walnut	2443
9	butt	walnut	1263
10	lap	oak	1295
11	lap	oak	1561
12	lap	pine	1000
13	butt	pine	596
14	lap	pine	859
15	butt	walnut	1029
16	beveled	pine	1207

Table 4.12
Sample Means for Nine Wood/Joint Combinations

		Wood			
		1 (Pine)	*2 (Oak)*	*3 (Walnut)*	
Joint	*1 (Butt)*	$\bar{y}_{11} = 712.5$	$\bar{y}_{12} = 1169.0$	$\bar{y}_{13} = 1146.0$	$\bar{y}_{1.} = 1009.17$
	2 (Beveled)	$\bar{y}_{21} = 1277.5$	$\bar{y}_{22} = 1722.5$	$\bar{y}_{23} = 2507.0$	$\bar{y}_{2.} = 1835.67$
	3 (Lap)	$\bar{y}_{31} = 929.5$	$\bar{y}_{32} = 1428.0$	$\bar{y}_{33} = 1489.0$	$\bar{y}_{3.} = 1282.17$
		$\bar{y}_{.1} = 973.17$	$\bar{y}_{.2} = 1439.83$	$\bar{y}_{.3} = 1714.00$	$\bar{y}_{..} = 1375.67$

Interaction Plot

ture. Collecting y's for the nine different combinations into separate samples and calculating means, the $\bar{y}_{ij}$'s are as presented in tabular form in Table 4.12 and plotted in Figure 4.22. This figure is a so-called **interaction plot** of these means.

The qualitative messages given by the plot are as follows:

1. Joint types ordered by strength are "beveled is stronger than lap, which in turn is stronger than butt."

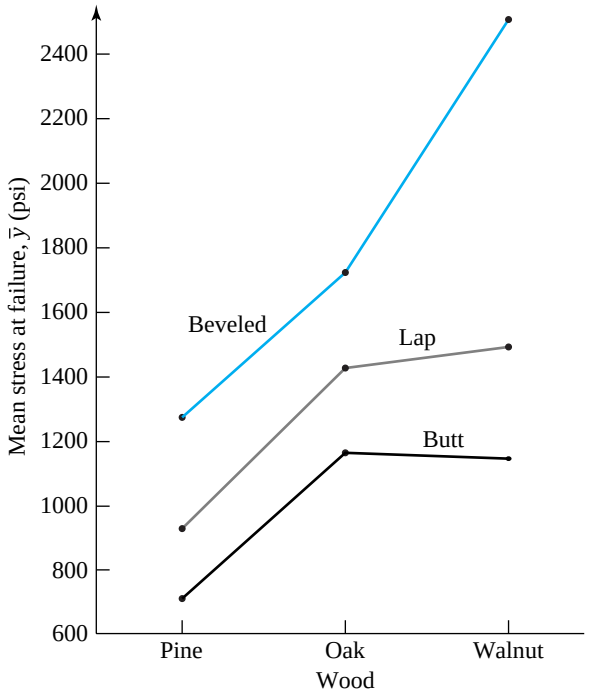

Figure 4.22 Interaction plot of joint strength sample means

Example 7 (*continued*)

2. Woods ordered by overall strength seem to be "walnut is stronger than oak, which in turn is stronger than pine."
3. The strength pattern across woods is not consistent from joint type to joint type (or equivalently, the strength pattern across joints is not consistent from wood type to wood type).

The idea of fitted effects is to invent a way of quantifying such qualitative summaries.

The row and column average means ($\bar{y}_{i\cdot}$'s and $\bar{y}_{\cdot j}$'s, respectively) might be taken as measures of average response behavior at different levels of the factors in question. If so, it then makes sense to use the differences between these and the grand average mean $\bar{y}_{\cdot\cdot}$ as measures of the effects of those levels on mean response. This leads to Definition 5.

Definition 5

In a two-way complete factorial study with factors A and B, the **fitted main effect of factor A at its *i*th level** is

$$a_i = \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}$$

Similarly, the **fitted main effect of factor B at its *j*th level** is

$$b_j = \bar{y}_{\cdot j} - \bar{y}_{\cdot\cdot}$$

Example 7 (*continued*)

Simple arithmetic and the $\bar{y}$'s in Table 4.12 yield the fitted main effects for the joint strength study of Kotlers, MacFarland, and Tomlinson. First for factor A (the Joint Type),

$$\begin{aligned} a_1 &= \text{the Joint Type fitted main effect for butt joints} \\ &= 1009.17 - 1375.67 \\ &= -366.5 \text{ psi} \\ a_2 &= \text{the Joint Type fitted main effect for beveled joints} \\ &= 1835.67 - 1375.67 \\ &= 460.0 \text{ psi} \\ a_3 &= \text{the Joint Type fitted main effect for lap joints} \\ &= 1282.17 - 1375.67 \\ &= -93.5 \text{ psi} \end{aligned}$$

Similarly for factor B (the Wood Type),

$$\begin{aligned} b_1 &= \text{the Wood Type fitted main effect for pine} \\ &= 973.17 - 1375.67 \\ &= -402.5 \text{ psi} \\ b_2 &= \text{the Wood Type fitted main effect for oak} \\ &= 1439.83 - 1375.67 \\ &= 64.17 \text{ psi} \\ b_3 &= \text{the Wood Type fitted main effect for walnut} \\ &= 1714.00 - 1375.67 \\ &= 338.33 \text{ psi} \end{aligned}$$

These fitted main effects quantify the first two qualitative messages carried by the data and listed as (1) and (2) before Definition 5. For example,

$$a_2 > a_3 > a_1$$

says that beveled joints are strongest and butt joints the weakest. Further, the fact that the a_i's and b_j's are of roughly the same order of magnitude says that the Joint Type and Wood Type factors are of comparable importance in determining tensile strength.

A difference between fitted main effects for a factor amounts to a difference between corresponding row or column averages and quantifies how different response behavior is for those two levels.

Example 7 (*continued*)

For example, comparing pine and oak wood types,

$$\begin{aligned} b_1 - b_2 &= (\bar{y}_{.1} - \bar{y}_{..}) - (\bar{y}_{.2} - \bar{y}_{..}) \\ &= \bar{y}_{.1} - \bar{y}_{.2} \\ &= 973.17 - 1439.83 \\ &= -466.67 \text{ psi} \end{aligned}$$

which indicates that pine joint average strength is about 467 psi less than oak joint average strength.

In *some* two-factor factorial studies, the fitted main effects as defined in Definition 5 pretty much summarize the story told by the means $\bar{y}_{ij}$, in the sense that

$$\bar{y}_{ij} \approx \bar{y}_{..} + a_i + b_j \qquad \text{for every } i \text{ and } j \tag{4.24}$$

Display (4.24) implies, for example, that the pattern of mean responses for level 1 of factor A is the same as for level 2 of A. That is, changing levels of factor B (from say j to j') produces the same change in mean response for level 2 as for level 1 (namely, $b_{j'} - b_j$). In fact, if relation (4.24) holds, there are **parallel traces** on an interaction plot of means.

Example 7 (*continued*)

To illustrate the meaning of expression (4.24), the fitted effects for the Joint Type/Wood Type data have been used to calculate $3 \times 3 = 9$ values of $\bar{y}_{..} + a_i + b_j$ corresponding to the nine experimental combinations. These are given in Table 4.13.

For comparison purposes, the $\bar{y}_{ij}$ from Table 4.12 and the $\bar{y}_{..} + a_i + b_j$ from Table 4.13 are plotted on the same sets of axes in Figure 4.23. Notice the parallel traces for the $\bar{y}_{..} + a_i + b_j$ values for the three different joint types. The traces for

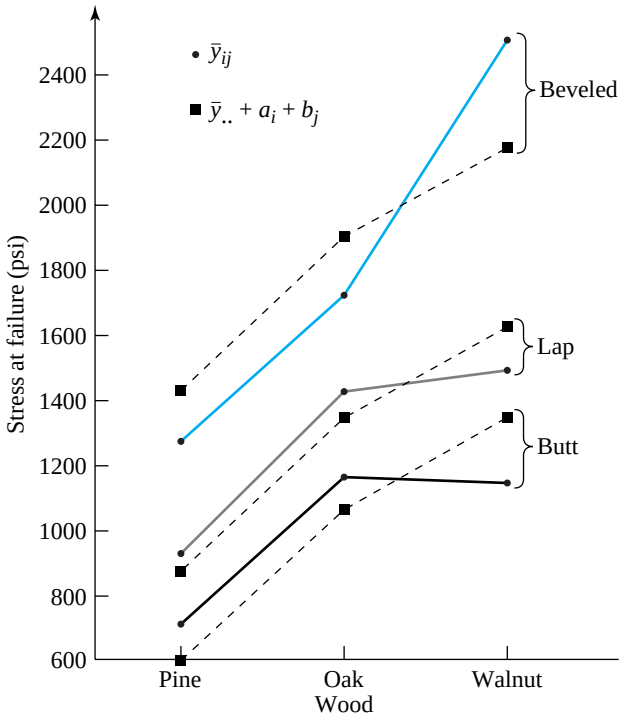

Figure 4.23 Plots of $\bar{y}_{ij}$ and $\bar{y}_{..} + a_i + b_j$ vs. wood type for three joint types

Table 4.13
Values of $\bar{y}_{..} + a_i + b_j$ for the Joint Strength Study

		Wood		
		1 (Pine)	*2 (Oak)*	*3 (Walnut)*
Joint	*1 (Butt)*	$\bar{y}_{..} + a_1 + b_1 =$ 606.67	$\bar{y}_{..} + a_1 + b_2 =$ 1073.33	$\bar{y}_{..} + a_1 + b_3 =$ 1347.50
	2 (Beveled)	$\bar{y}_{..} + a_2 + b_1 =$ 1433.17	$\bar{y}_{..} + a_2 + b_2 =$ 1899.83	$\bar{y}_{..} + a_2 + b_3 =$ 2174.00
	3 (Lap)	$\bar{y}_{..} + a_3 + b_1 =$ 879.67	$\bar{y}_{..} + a_3 + b_2 =$ 1346.33	$\bar{y}_{..} + a_3 + b_3 =$ 1620.50

the $\bar{y}_{ij}$ values for the three different joint types are not parallel (particularly when walnut is considered), so there are apparently substantial differences between the $\bar{y}_{ij}$'s and the $\bar{y}_{..} + a_i + b_j$'s.

When relationship (4.24) fails to hold, the patterns in mean response across levels of one factor depend on the levels of the second factor. In such cases, the differences between the combination means $\bar{y}_{ij}$ and the values $\bar{y}_{..} + a_i + b_j$ can serve as useful *measures of lack of parallelism* on the plots of means, and this leads to another definition.

Definition 6

In a two-way complete factorial study with factors A and B, the **fitted interaction of factor A at its *i*th level and factor B at its *j*th level** is

$$ab_{ij} = \bar{y}_{ij} - (\bar{y}_{..} + a_i + b_j)$$

Interpretation of interactions in a two-way factorial study

The fitted interactions in some sense measure how much pattern the combination means $\bar{y}_{ij}$ carry that is not explainable in terms of the factors A and B acting separately. Clearly, when relationship (4.24) holds, the fitted interactions ab_{ij} are all small (nearly 0), and system behavior can be thought of as depending separately on level of A and level of B. In such cases, an important practical consequence is that it is possible to develop recommendations for levels of the two factors independently of each other. For example, one need not recommend one level of A if B is at its level 1 and another if B is at its level 2.

Consider a study of the effects of factors Tool Type and Turning Speed on the metal removal rate for a lathe. If the fitted interactions are small, turning speed recommendations that remain valid for all tool types can be made. However, if the fitted interactions are important, turning speed recommendations might vary according to tool type.

Example 7 (*continued*)

Again using the Joint Type/Wood Type data, consider calculating the fitted interactions. The raw material for these calculations already exists in Tables 4.12 and 4.13. Simply taking differences between entries in these tables cell-by-cell yields the fitted interactions given in Table 4.14.

It is interesting to compare these fitted interactions to themselves and to the fitted main effects. The largest (in absolute value) fitted interaction (ab_{23}) corresponds to beveled walnut joints. This is consistent with one visual message in Figures 4.22 and 4.23: This Joint Type/Wood Type combination is in some sense most responsible for destroying any nearly parallel structure that might otherwise appear. The fact that (on the whole) the ab_{ij}'s are not as large as the a_i's or b_j's is consistent with a second visual message in Figures 4.22 and 4.23: The lack of parallelism, while important, is not as important as differences in Joint Types or Wood Types.

Table 4.14
Fitted Interactions for the Joint Strength Study

		Wood		
		1 (Pine)	*2 (Oak)*	*3 (Walnut)*
Joint	*1 (Butt)*	$ab_{11} = 105.83$	$ab_{12} = 95.67$	$ab_{13} = -201.5$
	2 (Beveled)	$ab_{21} = -155.66$	$ab_{22} = -177.33$	$ab_{23} = 333.0$
	3 (Lap)	$ab_{31} = 49.83$	$ab_{32} = 81.67$	$ab_{33} = -131.5$

Example 7 has proceeded "by hand." But using a statistical package can make the calculations painless. For example, Printout 6 illustrates that most of the results of Example 7 are readily available in MINITAB's "General Linear Model" routine (found under the "Stat/ANOVA/General Linear Model" menu). Comparing this printout to the example does bring up one point regarding the fitted effects defined in Definitions 5 and 6. Note that the printout provides values of only two (of three) Joint main effects, two (of three) Wood main effects, and four (of nine) Joint × Wood interactions. These are all that are needed, since it is a consequence of Definition 5 that *fitted main effects for a given factor must total to 0*, and it is a consequence of Definition 6 that *fitted interactions must sum to zero across any row or down any column* of the two-way table of factor combinations. The fitted effects not provided by the printout are easily deduced from the ones that are given.

Fitted effects sum to zero

Printout 6 Computations for the Joint Strength Data

```
General Linear Model

Factor     Type Levels  Values
joint      fixed      3 beveled butt    lap
wood       fixed      3     oak pine walnut
```

```
Analysis of Variance for strength, using Adjusted SS for Tests

Source       DF     Seq SS     Adj SS     Adj MS       F      P
joint         2    2153879    1881650     940825   32.67  0.000
wood          2    1641095    1481377     740689   25.72  0.001
joint*wood    4     468408     468408     117102    4.07  0.052
Error         7     201614     201614      28802
Total        15    4464996

Term                  Coef     StDev        T      P
Constant           1375.67     44.22    31.11  0.000
  joint
beveled             460.00     59.63     7.71  0.000
butt               -366.50     63.95    -5.73  0.001
  wood
oak                  64.17     63.95     1.00  0.349
pine               -402.50     59.63    -6.75  0.000
  joint*  wood
beveled oak        -177.33     85.38    -2.08  0.076
beveled pine       -155.67     82.20    -1.89  0.100
butt    oak          95.67     97.07     0.99  0.357
butt    pine        105.83     85.38     1.24  0.255

Unusual Observations for strength

Obs  strength       Fit   StDev Fit  Residual   St Resid
  4   1169.00   1169.00      169.71      0.00         * X
  7   1489.00   1489.00      169.71      0.00         * X

X denotes an observation whose X value gives it large influence.

Least Squares Means for strength

  joint              Mean     StDev
beveled            1835.7     69.28
butt               1009.2     80.00
lap                1282.2     80.00
  wood
oak                1439.8     80.00
pine                973.2     69.28
walnut             1714.0     80.00
  joint*  wood
beveled oak        1722.5    120.00
beveled pine       1277.5    120.00
beveled walnut     2507.0    120.00
butt    oak        1169.0    169.71
butt    pine        712.5    120.00
butt    walnut     1146.0    120.00
lap     oak        1428.0    120.00
lap     pine        929.5    120.00
lap     walnut     1489.0    169.71
```

4.3.2 Simpler Descriptions for Some Two-Way Data Sets

Rewriting the equation for ab_{ij} from Definition 6,

$$\bar{y}_{ij} = \bar{y}_{..} + a_i + b_j + ab_{ij} \tag{4.25}$$

That is, $\bar{y}_{\cdot\cdot}$, the fitted main effects, and the fitted interactions provide a decomposition or breakdown of the combination sample means into interpretable pieces. These pieces correspond to an overall effect, the effects of factors acting separately, and the effects of factors acting jointly.

Taking a hint from the equation fitting done in the previous two sections, it makes sense to think of (4.25) as a fitted version of an approximate relationship,

$$y \approx \mu + \alpha_i + \beta_j + \alpha\beta_{ij} \qquad \textbf{(4.26)}$$

where $\mu, \alpha_1, \alpha_2, \ldots, \alpha_I, \beta_1, \beta_2, \ldots, \beta_J, \alpha\beta_{11}, \ldots, \alpha\beta_{1J}, \alpha\beta_{21}, \ldots, \alpha\beta_{IJ}$ are some constants and the levels of factors A and B associated with a particular response y pick out which of the α_i's, β_j's, and $\alpha\beta_{ij}$'s are appropriate in equation (4.26). By analogy with the previous two sections, the possibility should be considered that a relationship even simpler than equation (4.26) might hold, perhaps not involving $\alpha\beta_{ij}$'s or even α_i's or perhaps β_j's.

It has already been said that when relationship (4.24) is in force, or equivalently

$$ab_{ij} \approx 0 \qquad \text{for every } i \text{ and } j$$

it is possible to understand an observed set of $\bar{y}_{ij}$'s in simplified terms of the factors acting separately. This possibility corresponds to the simplified version of equation (4.26),

$$y \approx \mu + \alpha_i + \beta_j$$

and there are other simplified versions of equation (4.26) that also have appealing interpretations. For example, the simplified version of equation (4.26),

$$y \approx \mu + \alpha_i$$

says that only factor A (not factor B) is important in determining response y. ($\alpha_1, \alpha_2, \ldots, \alpha_I$ still allow for different response behavior for different levels of A.)

Two questions naturally follow on this kind of reasoning: "How is a *reduced* or *simplified* version of equation (4.26) fitted to a data set? And after fitting such an equation, how is the appropriateness of the result determined?" General answers to these questions are subtle. But there is one circumstance in which it is possible to give fairly straightforward answers. That is the case where the data are **balanced**—in the sense that all of the samples (leading to the $\bar{y}_{ij}$'s) have the same size. With balanced data, the fitted effects from Definitions 5 and 6 and simple addition produce fitted responses. And based on such fitted values, the R^2 and residual plotting ideas from the last two sections can be applied here as well. That is, when working with balanced data, least squares fitting of a simplified version of equation (4.26) can be accomplished by

1. calculating fitted effects according to Definitions 5 and 6 and then

2. adding those corresponding to terms in the reduced equation to compute fitted responses, $\hat{y}$.

Residuals are then (as always)

Residuals

$$e = y - \hat{y}$$

(and should look like noise if the simplified equation is an adequate description of the data set). Further, the fraction of raw variation in y accounted for in the fitting process is (as always)

Coefficient of determination

$$R^2 = \frac{\sum(y - \bar{y})^2 - \sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} \qquad \textbf{(4.27)}$$

where the sums are over all observed y's. (Summation notation is being abused even further than usual, by not even subscripting the y's and $\hat{y}$'s.)

Example 8
(*Example 12, Chapter 2, revisited—p. 49*)

Simplified Description of Two-Way Factorial Golf Ball Flight Data

G. Gronberg tested drive flight distances for golf balls of several different compressions on several different evenings. Table 4.15 gives a small part of the data that he collected, representing 80 and 100 compression flight distances (in yards) from two different evenings. Notice that these data are balanced, all four sample sizes being 10.

Table 4.15
Golf Ball Flight Distances for Four Compression/Evening Combinations

		Evening (B)			
		1		*2*	
Compression (A)	*80*	180	192	196	180
		193	190	192	195
		197	182	191	197
		189	192	194	192
		187	179	186	193
	100	180	175	190	185
		185	190	195	167
		167	185	180	180
		162	180	170	180
		170	185	180	165

Example 8 (*continued*)

These data have complete two-way factorial structure. The factor Evening is not really of primary interest. Rather, it is a blocking factor, its levels creating homogeneous environments in which to compare 80 and 100 compression flight distances. Figure 4.24 is a graphic using boxplots to represent the four samples and emphasizing the factorial structure.

Calculating sample means corresponding to the four cells in Table 4.15 and then finding fitted effects is straightforward. Table 4.16 displays cell, row, column, and grand average means. And based on those values,

$$
\begin{aligned}
a_1 &= 189.85 - 184.20 = 5.65 \text{ yards} \\
a_2 &= 178.55 - 184.20 = -5.65 \text{ yards} \\
b_1 &= 183.00 - 184.20 = -1.20 \text{ yards} \\
b_2 &= 185.40 - 184.20 = 1.20 \text{ yards} \\
ab_{11} &= 188.1 - (184.20 + 5.65 + (-1.20)) = -.55 \text{ yards} \\
ab_{12} &= 191.6 - (184.20 + 5.65 + 1.20) = .55 \text{ yards} \\
ab_{21} &= 177.9 - (184.20 + (-5.65) + (-1.20)) = .55 \text{ yards} \\
ab_{22} &= 179.2 - (184.20 + (-5.65) + 1.20) = -.55 \text{ yards}
\end{aligned}
$$

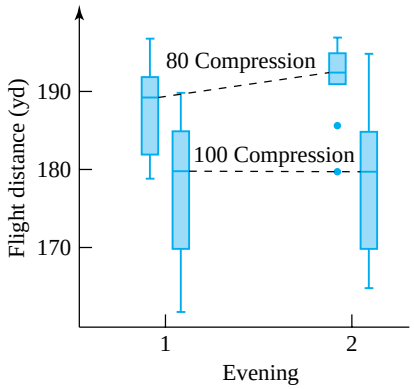

Figure 4.24 Golf ball flight distance boxplots for four combinations of Compression and Evening

Table 4.16
Cell, Row, Column, and Grand Average Means for the Golf Ball Flight Data

		Evening (B)		
		1	*2*	
Compression (A)	*80*	$\bar{y}_{11} = 188.1$	$\bar{y}_{12} = 191.6$	189.85
	100	$\bar{y}_{21} = 177.9$	$\bar{y}_{22} = 179.2$	178.55
		183.00	185.40	184.20

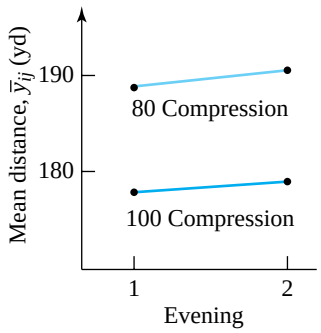

Figure 4.25 Interaction plot for the golf ball flight data

The fitted effects indicate that most of the differences in the cell means in Table 4.16 are understandable in terms of differences between 80 and 100 compression balls. The effect of differences between evenings appears to be on the order of one-fourth the size of the effect of differences between ball compressions. Further, the pattern of flight distances across the two compressions changed relatively little from evening to evening. These facts are portrayed graphically in the interaction plot of Figure 4.25.

The story told by the fitted effects in this example probably agrees with most readers' intuition. There is little reason a priori to expect the relative behaviors of 80 and 100 compression flight distances to change much from evening to evening. But there is slightly more reason to expect the distances to be longer overall on some nights than on others.

It is worth investigating whether the data in Table 4.15 allow the simplest

"Compression effects only"

description, or require the somewhat more complicated

"Compression effects and Evening effects but no interactions"

description, or really demand to be described in terms of

"Compression, Evening, and interaction effects"

To do so, fitted responses are first calculated corresponding to the three different possible corresponding relationships

$$y \approx \mu + \alpha_i \tag{4.28}$$

$$y \approx \mu + \alpha_i + \beta_j \tag{4.29}$$

$$y \approx \mu + \alpha_i + \beta_j + \alpha\beta_{ij} \tag{4.30}$$

Example 8 (*continued*)

Table 4.17
Fitted Responses Corresponding to Equations (4.28), (4.29), and (4.30)

Compression	Evening	For (4.28) $\bar{y}_{..} + a_i = \bar{y}_{i.}$	For (4.29) $\bar{y}_{..} + a_i + b_j$	For (4.30) $\bar{y}_{..} + a_i + b_j + ab_{ij} = \bar{y}_{ij}$
80	1	189.85	188.65	188.10
100	1	178.55	177.35	177.90
80	2	189.85	191.05	191.60
100	2	178.55	179.75	179.20

These are generated using the fitted effects. They are collected in Table 4.17 (not surprisingly, the first and third sets of fitted responses are, respectively, row average and cell means).

Residuals $e = y - \hat{y}$ for fitting the three equations (4.28), (4.29), and (4.30) are obtained by subtracting the appropriate entries in, respectively, the third, fourth, or fifth column of Table 4.17 from each of the data values listed in Table 4.15. For example, 40 residuals for the fitting of the "A main effects only" equation (4.28) would be obtained by subtracting 189.85 from every entry in the upper left cell of Table 4.15, subtracting 178.55 from every entry in the lower left cell, 189.85 from every entry in the upper right cell, and 178.55 from every entry in the lower right cell.

Figure 4.26 provides normal plots of the residuals from the fitting of the three equations (4.28), (4.29), and (4.30). None of the normal plots is especially linear, but at the same time, none of them is grossly nonlinear either. In particular, the first two, corresponding to simplified versions of relationship 4.26, are not significantly worse than the last one, which corresponds to the use of all fitted effects (both main effects and interactions). From the limited viewpoint of producing residuals with an approximately bell-shaped distribution, the fitting of any of the three equations (4.28), (4.29), and (4.30) would appear approximately equally effective.

The calculation of R^2 values for equations (4.28), (4.29), and (4.30) proceeds as follows. First, since the grand average of all 40 flight distances is $\bar{y} = 184.2$ yards (which in this case also turns out to be $\bar{y}_{..}$) ,

$$\begin{aligned}\sum(y-\bar{y})^2 &= (180-184.2)^2 + \cdots + (179-184.2)^2 \\ &\quad + (180-184.2)^2 + \cdots + (185-184.2)^2 \\ &\quad + (196-184.2)^2 + \cdots + (193-184.2)^2 \\ &\quad + (190-184.2)^2 + \cdots + (165-184.2)^2 \\ &= 3{,}492.4\end{aligned}$$

(This value can easily be obtained on a pocket calculator by using 39 ($= 40 - 1 = n - 1$) times the sample variance of all 40 flight distances.) Then $\sum(y-\hat{y})^2$

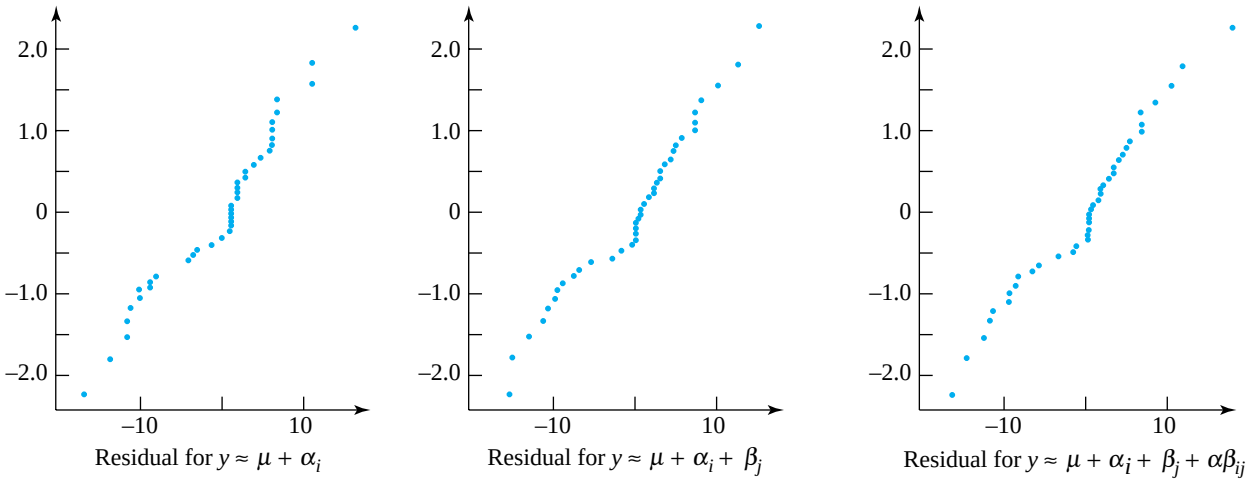

Figure 4.26 Normal plots of residuals from three different equations fitted to the golf data

values for the three equations are obtained as the sums of the squared residuals. For example, using Tables 4.15 and 4.17, for equation (4.29),

$$\begin{aligned}\sum(y-\hat{y})^2 &= (180-188.65)^2+\cdots+(179-188.65)^2\\ &\quad+(180-177.35)^2+\cdots+(185-177.35)^2\\ &\quad+(196-191.05)^2+\cdots+(193-191.05)^2\\ &\quad+(190-179.75)^2+\cdots+(165-179.75)^2\\ &= 2{,}157.90\end{aligned}$$

Finally, equation (4.27) is used. Table 4.18 gives the three values of R^2.

The story told by the R^2 values is consistent with everything else that's been said in this example. None of the values is terribly big, which is consistent with the large within-sample variation in flight distances evident in Figure 4.24. But

Table 4.18
R^2 Values for Fitting Equations (4.28), (4.29), and (4.30) to Gronberg's Data

Equation	R^2
$y \approx \mu + \alpha_i$	.366
$y \approx \mu + \alpha_i + \beta_j$	.382
$y \approx \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$	.386

Example 8 (*continued*)

considering A (Compression) main effects does account for some of the observed variation in flight distance, and the addition of B (Evening) main effects adds slightly to the variation accounted for. Introducing interactions into consideration adds little additional accounting power.

The computations in Example 8 are straightforward but tedious. The kind of software used to produce Printout 6 typically allows for the painless fitting of simplified relationships like (4.28), (4.29), and (4.30) and computation (and later plotting) of the associated residuals.

4.3.3 Fitted Effects for Three-Way (and Higher) Factorials

The reasoning that has been applied to two-way factorial data is naturally generalized to complete factorial data structures that are three-way and higher. First, fitted main effects and various kinds of interactions are computed. Then one hopes to discover that a data set can be adequately described in terms of a few of these that are interpretable when taken as a group. This subsection shows how this is carried out for 3-factor situations. Once the pattern has been made clear, the reader can carry it out for situations involving more than three factors, working by analogy.

In order to deal with three-way factorial data, yet more notation is needed. Unfortunately, this involves triple subscripts. For factor A having I levels, factor B having J levels, and factor C having K levels, the following notation will be used:

Notation for sample means and their averages (for three-way factorial data)

$\bar{y}_{ijk}$ = the sample mean response when factor A is at level i, factor B is at level j, and factor C is at level k

$$\bar{y}_{...} = \frac{1}{IJK}\sum_{i,j,k}\bar{y}_{ijk}$$

= the grand average sample mean

$$\bar{y}_{i..} = \frac{1}{JK}\sum_{j,k}\bar{y}_{ijk}$$

= the average sample mean when factor A is at level i

$$\bar{y}_{.j.} = \frac{1}{IK}\sum_{i,k}\bar{y}_{ijk}$$

= the average sample mean when factor B is at level j

$$\bar{y}_{..k} = \frac{1}{IJ}\sum_{i,j}\bar{y}_{ijk}$$

= the average sample mean when factor C is at level k

$$\bar{y}_{ij.} = \frac{1}{K}\sum_{k}\bar{y}_{ijk}$$

$$= \text{the average sample mean when factor A is at level } i \text{ and factor B is at level } j$$

$$\bar{y}_{i.k} = \frac{1}{J}\sum_{j}\bar{y}_{ijk}$$

$$= \text{the average sample mean when factor A is at level } i \text{ and factor C is at level } k$$

$$\bar{y}_{.jk} = \frac{1}{I}\sum_{i}\bar{y}_{ijk}$$

$$= \text{the average sample mean when factor B is at level } j \text{ and factor C is at level } k$$

In these expressions, where a subscript is used as an index of summation, the summation is assumed to extend over all of its I, J, or K possible values.

It is most natural to think of the means from a 3-factor study laid out in three dimensions. Figure 4.27 illustrates this general situation, and the next example employs another common three-dimensional display in a 2^3 context.

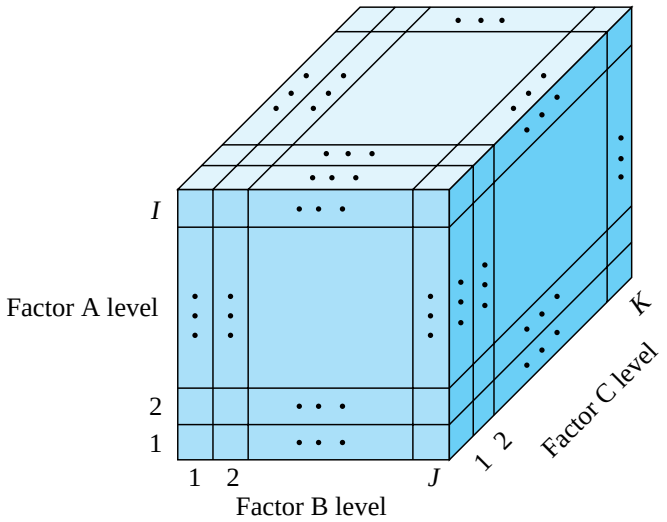

Figure 4.27 *IJK* cells in a three-dimensional table

Example 9

A 2^3 Factorial Experiment on the Strength of a Composite Material

In his article "Application of Two-Cubed Factorial Designs to Process Studies" (*ASQC Technical Supplement Experiments in Industry*, 1985), G. Kinzer discusses a successful 3-factor industrial experiment.

The strength of a proprietary composite material was thought to be related to three process variables, as indicated in Table 4.19. Five specimens were produced under each of the $2^3 = 8$ combinations of factor levels, and their moduli of rupture were measured (in psi) and averaged to produce the means in Table 4.20. (There were also apparently 10 specimens made with an autoclave temperature of 315°F, an autoclave time of 8 hr, and a time span of 8 hr, but this will be ignored for present purposes.)

Cube plot for displaying 2^3 means

A helpful display of these means can be made using the corners of a cube, as in Figure 4.28. Using this three-dimensional picture, one can think of average sample means as averages of $\bar{y}_{ijk}$'s sharing a face or edge of the cube.

Table 4.19
Levels of Three Process Variables in a 2^3 Study of Material Strength

Factor	Process Variable	Level 1	Level 2
A	Autoclave temperature	300°F	330°F
B	Autoclave time	4 hr	12 hr
C	Time span (between product formation and autoclaving)	4 hr	12 hr

Table 4.20
Sample Mean Strengths for 2^3 Treatment Combinations

i, Factor A Level	j, Factor B Level	k, Factor C Level	$\bar{y}_{ijk}$, Sample Mean Strength (psi)
1	1	1	1520
2	1	1	2450
1	2	1	2340
2	2	1	2900
1	1	2	1670
2	1	2	2540
1	2	2	2230
2	2	2	3230

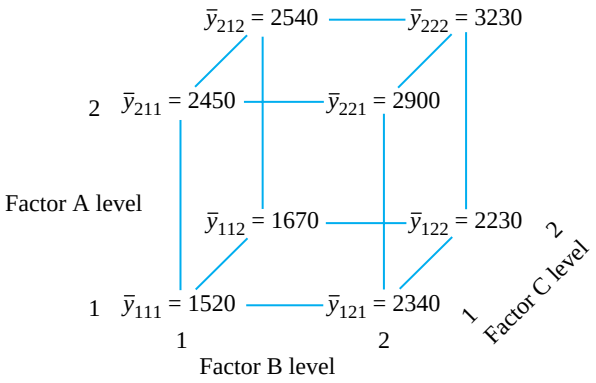

Figure 4.28 2^3 sample mean strengths displayed on a cube plot

For example,

$$\bar{y}_{1..} = \frac{1}{2 \cdot 2}(1520 + 2340 + 1670 + 2230) = 1940 \text{ psi}$$

is the average mean on the bottom face, while

$$\bar{y}_{11.} = \frac{1}{2}(1520 + 1670) = 1595 \text{ psi}$$

is the average mean on the lower left edge. For future reference, all of the average sample means are collected here:

$$\begin{aligned}
\bar{y}_{...} &= 2360 \text{ psi} \\
\bar{y}_{1..} &= 1940 \text{ psi} \qquad \bar{y}_{2..} = 2780 \text{ psi} \\
\bar{y}_{.1.} &= 2045 \text{ psi} \qquad \bar{y}_{.2.} = 2675 \text{ psi} \\
\bar{y}_{..1} &= 2302.5 \text{ psi} \qquad \bar{y}_{..2} = 2417.5 \text{ psi} \\
\bar{y}_{11.} &= 1595 \text{ psi} \qquad \bar{y}_{12.} = 2285 \text{ psi} \\
\bar{y}_{21.} &= 2495 \text{ psi} \qquad \bar{y}_{22.} = 3065 \text{ psi} \\
\bar{y}_{1.1} &= 1930 \text{ psi} \qquad \bar{y}_{1.2} = 1950 \text{ psi} \\
\bar{y}_{2.1} &= 2675 \text{ psi} \qquad \bar{y}_{2.2} = 2885 \text{ psi} \\
\bar{y}_{.11} &= 1985 \text{ psi} \qquad \bar{y}_{.12} = 2105 \text{ psi} \\
\bar{y}_{.21} &= 2620 \text{ psi} \qquad \bar{y}_{.22} = 2730 \text{ psi}
\end{aligned}$$

Analogy with Definition 5 provides definitions of fitted main effects in a 3-factor study as the differences between factor-level average means and the grand average mean.

Definition 7

In a three-way complete factorial study with factors A, B, and C, the **fitted main effect of factor A at its *i*th level** is

$$a_i = \bar{y}_{i..} - \bar{y}_{...}$$

The **fitted main effect of factor B at its *j*th level** is

$$b_j = \bar{y}_{.j.} - \bar{y}_{...}$$

And the **fitted main effect of factor C at its *k*th level** is

$$c_k = \bar{y}_{..k} - \bar{y}_{...}$$

Using the geometrical representation of factor-level combinations given in Figure 4.28, these fitted effects are averages of $\bar{y}_{ijk}$'s along planes (parallel to one set of faces of the rectangular solid) minus the grand average sample mean.

Next, analogy with Definition 6 produces definitions of fitted two-way interactions in a 3-factor study.

Definition 8

In a three-way complete factorial study with factors A, B, and C, the **fitted 2-factor interaction of factor A at its *i*th level** and **factor B at its *j*th level** is

$$ab_{ij} = \bar{y}_{ij.} - (\bar{y}_{...} + a_i + b_j)$$

the **fitted 2-factor interaction of factor A at its *i*th level** and **factor C at its *k*th level** is

$$ac_{ik} = \bar{y}_{i.k} - (\bar{y}_{...} + a_i + c_k)$$

and the **fitted 2-factor interaction of factor B at its *j*th level** and **factor C at its *k*th level** is

$$bc_{jk} = \bar{y}_{.jk} - (\bar{y}_{...} + b_j + c_k)$$

Interpreting two-way interactions in a three-way study

These fitted 2-factor interactions can be thought of in two equivalent ways:

1. as what one gets as fitted interactions upon averaging across all levels of the factor that is not under consideration to obtain a single two-way table of (average) means and then calculating as per Definition 6 (page 169);
2. as what one gets as averages, across all levels of the factor not under consideration, of the fitted two-factor interactions calculated as per Definition 6, one level of the excluded factor at a time.

Example 9 (*continued*)

To illustrate the meaning of Definitions 7 and 8, return to the composite material strength study. For example, the fitted A main effects are

$$a_1 = \bar{y}_{1..} - \bar{y}_{...} = 1940 - 2360 = -420 \text{ psi}$$
$$a_2 = \bar{y}_{2..} - \bar{y}_{...} = 2780 - 2360 = 420 \text{ psi}$$

And the fitted AB 2-factor interaction for levels 1 of A and 1 of B is

$$ab_{11} = \bar{y}_{11.} - (\bar{y}_{...} + a_1 + b_1) = 1595 - (2360 + (-420) + (2045 - 2360))$$
$$= -30 \text{ psi}$$

The entire set of fitted effects for the means of Table 4.20 is as follows.

$a_1 = -420$ psi	$b_1 = -315$ psi	$c_1 = -57.5$ psi
$a_2 = 420$ psi	$b_2 = 315$ psi	$c_2 = 57.5$ psi
$ab_{11} = -30$ psi	$ac_{11} = 47.5$ psi	$bc_{11} = -2.5$ psi
$ab_{12} = 30$ psi	$ac_{12} = -47.5$ psi	$bc_{12} = 2.5$ psi
$ab_{21} = 30$ psi	$ac_{21} = -47.5$ psi	$bc_{21} = 2.5$ psi
$ab_{22} = -30$ psi	$ac_{22} = 47.5$ psi	$bc_{22} = -2.5$ psi

Interpretation of three-way interactions

Remember equation (4.25) (page 171). It says that in 2-factor studies, the fitted grand mean, main effects, and two-factor interactions completely describe a factorial set of sample means. Such is not the case in three-factor studies. Instead, a new possibility arises: *3-factor interaction*. Roughly speaking, the fitted three-factor interactions in a 3-factor study measure how much pattern the combination means carry that is not explainable in terms of the factors A, B, and C acting separately and in pairs.

Definition 9

In a three-way complete factorial study with factors A, B, and C, **the fitted 3-factor interaction of A at its *i*th level, B at its *j*th level**, and **C at its *k*th level** is

$$abc_{ijk} = \bar{y}_{ijk} - (\bar{y}_{...} + a_i + b_j + c_k + ab_{ij} + ac_{ik} + bc_{jk})$$

Example 9 (*continued*)

To illustrate the meaning of Definition 9, consider again the composite material study. Using the previously calculated fitted main effects and 2-factor interactions,

$$abc_{111} = 1520 - (2360 + (-420) + (-315) + (-57.5) + (-30) + 47.5 + (-2.5)) = -62.5\text{psi}$$

Similar calculations can be made to verify that the entire set of 3-factor interactions for the means of Table 4.20 is as follows:

$$\begin{aligned} abc_{111} &= -62.5 \text{ psi} & abc_{211} &= 62.5 \text{ psi} \\ abc_{121} &= 62.5 \text{ psi} & abc_{221} &= -62.5 \text{ psi} \\ abc_{112} &= 62.5 \text{ psi} & abc_{212} &= -62.5 \text{ psi} \\ abc_{122} &= -62.5 \text{ psi} & abc_{222} &= 62.5 \text{ psi} \end{aligned}$$

A second interpretation of three-way interactions

Main effects and 2-factor interactions are more easily interpreted than 3-factor interactions. One insight into their meaning was given immediately before Definition 9. Another is the following. If at the different levels of (say) factor C, the fitted AB interactions are calculated and the fitted AB interactions (the pattern of parallelism or nonparallelism) are essentially the same on all levels of C, then the 3-factor interactions are small (near 0). Otherwise, large 3-factor interactions allow the pattern of AB interaction to change, from one level of C to another.

4.3.4 Simpler Descriptions of Some Three-Way Data Sets

Rewriting the equation in Definition 9,

$$\bar{y}_{ijk} = \bar{y}_{\cdots} + a_i + b_j + c_k + ab_{ij} + ac_{ik} + bc_{jk} + abc_{ijk} \qquad \textbf{(4.31)}$$

This is a breakdown of the combination sample means into somewhat interpretable pieces, corresponding to an overall effect, the factors acting separately, the factors acting in pairs, and the factors acting jointly. Display (4.31) may be thought of as a fitted version of an approximate relationship

$$y \approx \mu + \alpha_i + \beta_j + \gamma_k + \alpha\beta_{ij} + \alpha\gamma_{ik} + \beta\gamma_{jk} + \alpha\beta\gamma_{ijk} \qquad \textbf{(4.32)}$$

When beginning the analysis of three-way factorial data, one hopes to discover a simplified version of equation (4.32) that is both interpretable and an adequate description of the data. (Indeed, if it is not possible to do so, little is gained by using the factorial breakdown rather than simply treating the data in question as *IJK unstructured* samples.)

As was the case earlier with two-way factorial data, the process of fitting a simplified version of display (4.32) via least squares is, in general, unfortunately somewhat complicated. But *when all sample sizes are equal* (i.e., the data are

balanced), the fitting process can be accomplished by simply adding appropriate fitted effects defined in Definitions 7, 8, and 9. Then the fitted responses lead to residuals that can be used in residual plotting and the calculation of R^2.

Example 9 (*continued*)

Looking over the magnitudes of the fitted effects for Kinzer's composite material strength study, the A and B main effects clearly dwarf the others, suggesting the possibility that the relationship

$$y \approx \mu + \alpha_i + \beta_j \qquad \textbf{(4.33)}$$

could be used as a description of the physical system. This relationship doesn't involve factor C at all (either by itself or in combination with A or B) and indicates that responses for a particular AB combination will be comparable for both time spans studied. Further, the fact that display (4.33) doesn't include the $\alpha\beta_{ij}$ term says that factors A and B act on product strength separately, so that their levels can be chosen independently. In geometrical terms corresponding to the cube plot in Figure 4.28, display (4.33) means that observations from the cube's back face will be comparable to corresponding ones on the front face and that parallelism will prevail on both the front and back faces.

Kinzer's article gives only $\bar{y}_{ijk}$ values, not raw data, so a residual analysis and calculation of R^2 are not possible. But because of the balanced nature of the original data set, fitted values are easily obtained. For example, with factor A at level 1 and B at level 1, using the simplified relationship (4.33) and the fitted main effects found earlier produces the fitted value

$$\hat{y} = \bar{y}_{\cdots} + a_1 + b_1 = 2360 + (-420) + (-315) = 1625 \text{ psi}$$

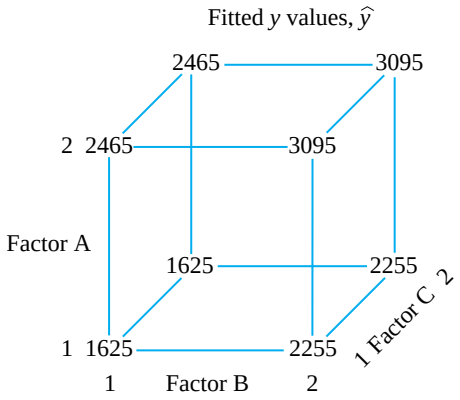

Figure 4.29 Eight fitted responses for relationship (4.33) and the composite strength study

Example 9 (*continued*)

All eight fitted values corresponding to equation (4.33) are shown geometrically in Figure 4.29. The fitted values given in the figure might be combined with product requirements and cost information to allow a process engineer to make sound decisions about autoclave temperature, autoclave time, and time span.

In Example 9, the simplified version of display (4.32) was especially interpretable because it involved only main effects. But sometimes even versions of relation (4.32) involving interactions can draw attention to what is going on in a data set.

Example 10

Interactions in a 3-Factor Paper Airplane Experiment

Schmittenberg and Riesterer studied the effects of three factors, each at two levels, on flight distance of paper airplanes. The factors were Plane Design (A) (design 1 versus design 2), Plane Size (B) (large versus small), and Paper Type (C) (heavy versus light). The means of flight distances they obtained for 15 flights of each of the $8 = 2 \times 2 \times 2$ types of planes are given in Figure 4.30.

Calculate the fitted effects corresponding to the $\bar{y}_{ijk}$'s given in Figure 4.30 "by hand." (Printout 7 also gives the fitted effects.) By far the biggest fitted effects (more than three times the size of any others) are the AC interactions. This makes perfect sense. The strongest message in Figure 4.30 is that plane design 1 should be made with light paper and plane design 2 with heavy paper. This is a perfect example of a strong 2-factor interaction in a 3-factor study (where, incidentally, the fitted 3-factor interactions are roughly $\frac{1}{4}$ the size of any other fitted effects). Any simplified version of display (4.32) used to represent this situation would certainly have to include the $\alpha\gamma_{ik}$ term.

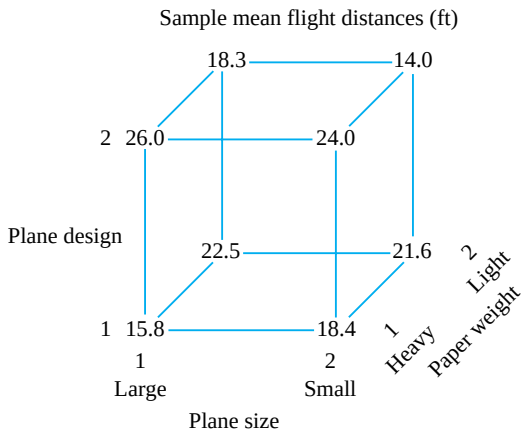

Figure 4.30 2^3 sample mean flight distances displayed on the corners of a cube

Printout 7 Calculation of Fitted Effects for the Airplane Experiment

```
General Linear Model


Factor     Type Levels Values
design     fixed      2    1 2
size       fixed      2    1 2
paper      fixed      2    1 2

Analysis of Variance for mean dis, using Adjusted SS for Tests

Source               DF     Seq SS     Adj SS     Adj MS       F      P
design                1      2.000      2.000      2.000      **
size                  1      2.645      2.645      2.645      **
paper                 1      7.605      7.605      7.605      **
design*size           1      8.000      8.000      8.000      **
design*paper          1     95.220     95.220     95.220      **
size*paper            1      4.205      4.205      4.205      **
design*size*paper     1      0.180      0.180      0.180      **
Error                 0      0.000      0.000      0.000
Total                 7    119.855

** Denominator of F-test is zero.

Term                     Coef     StDev        T      P
Constant              20.0750    0.0000        *      *
design
1                    -0.500000  0.000000        *      *
size
1                     0.575000  0.000000        *      *
paper
1                     0.975000  0.000000        *      *
design*size
1      1              -1.00000   0.00000        *      *
design*paper
1      1              -3.45000   0.00000        *      *
size*paper
1    1              -0.725000  0.000000        *      *
design*size*paper
1      1    1        -0.150000  0.000000        *      *
```

4.3.5 Special Devices for 2^p Studies

All of the discussion in this section has been general, in the sense that any value has been permissible for the number of levels for a factor. In particular, all of the definitions of fitted effects in the section work as well for $3 \times 5 \times 7$ studies as they do for $2 \times 2 \times 2$ studies. But from here on in the section, attention will be restricted to 2^p data structures.

Special 2^p factorial notation

Restricting attention to two-level factors affords several conveniences. One is notational. It is possible to reduce the clutter caused by the multiple subscript "ijk" notation, as follows. One level of each factor is designated as a "high" (or "+") level and the other as a "low" (or "−") level. Then the 2^p factorial combinations are labeled with letters corresponding to those factors appearing in the combination at

Table 4.21
Shorthand Names for the 2^3 Factorial Treatment Combinations

Level of Factor A	Level of Factor B	Level of Factor C	Combination Name
1	1	1	(1)
2	1	1	a
1	2	1	b
2	2	1	ab
1	1	2	c
2	1	2	ac
1	2	2	bc
2	2	2	abc

their high levels. For example, if level 2 of each of factors A, B, and C is designated the high level, shorthand names for the $2^3 = 8$ different ABC combinations are as given in Table 4.21. Using these names, for example, $\bar{y}_a$ can stand for a sample mean where factor A is at its high (or second) level and all other factors are at their low (or first) levels.

Special relationship between 2^p effects of a given type

A second convenience special to two-level factorial data structures is the fact that all effects of a given type have the same absolute value. This has already been illustrated in Example 9. For example, looking back, for the data of Table 4.20,

$$a_2 = 420 = -(-420) = -a_1$$

and

$$bc_{22} = -2.5 = bc_{11} = -bc_{12} = -bc_{21}$$

This is always the case for fitted effects in 2^p factorials. In fact, if two fitted effects of the same type are such that an even number of $1 \to 2$ or $2 \to 1$ subscript changes are required to get the second from the first, the fitted effects are equal (e.g., $bc_{22} = bc_{11}$). If an odd number are required, then the second fitted effect is -1 times the first (e.g., $bc_{12} = -bc_{22}$). This fact is so useful because one needs only to do the arithmetic necessary to find one fitted effect of each type and then choose appropriate signs to get all others of that type.

A statistician named Frank Yates is credited with discovering an efficient, mechanical way of generating one fitted effect of each type for a 2^p study. His method is easy to implement "by hand" and produces fitted effects with all "2" subscripts (i.e., corresponding to the "all factors at their high level" combination).

The Yates algorithm for computing fitted 2^p factorial effects

The **Yates algorithm** consists of the following steps.

Step 1 Write down the 2^p sample means in a column in what is called **Yates standard order**. Standard order is easily remembered by beginning

with (1) and a, then multiplying these two names (algebraically) by b to get b and ab, then multiplying these four names by c to get c, ac, bc, abc, etc.

Step 2 Make up another column of numbers by first adding and then subtracting (first from second) the entries in the previous column in pairs.

Step 3 Follow step 2 a total of p times, and then make up a final column by dividing the entries in the last column by the value 2^p.

The last column (made via step 3) gives fitted effects (all factors at level 2), again in standard order.

Example 9 (*continued*)

Table 4.22 shows the use of the Yates algorithm to calculate fitted effects for the 2^3 composite material study. The entries in the final column of this table are, of course, exactly as listed earlier, and the rest of the fitted effects are easily obtained via appropriate sign changes. This final column is an extremely concise summary of the fitted effects, which quickly reveals which types of fitted effects are larger than others.

Table 4.22
The Yates Algorithm Applied to the Means of Table 4.20

Combination	$\bar{y}$	Cycle 1	Cycle 2	Cycle 3	Cycle 3 ÷ 8
(1)	1520	3970	9210	18,880	$2360 = \bar{y}_{\cdots}$
a	2450	5240	9670	3,360	$420 = a_2$
b	2340	4210	1490	2,520	$315 = b_2$
ab	2900	5460	1870	−240	$-30 = ab_{22}$
c	1670	930	1270	460	$57.5 = c_2$
ac	2540	560	1250	380	$47.5 = ac_{22}$
bc	2230	870	−370	−20	$-2.5 = bc_{22}$
abc	3230	1000	130	500	$62.5 = abc_{222}$

The reverse Yates algorithm and easy computation of fitted responses

The Yates algorithm is useful beyond finding fitted effects. For balanced data sets, it is also possible to modify it slightly to find fitted responses, $\hat{y}$, corresponding to a simplified version of a relation like display (4.32). First, the desired (all factors at their high level) fitted effects (using 0's for those types not considered) are written down in reverse standard order. Then, by applying p cycles of the Yates additions and subtractions, the fitted values, $\hat{y}$, are obtained, listed in reverse standard order. (Note that no final division is required in this **reverse Yates algorithm**.)

Example 9 (*continued*)

Consider fitting the relationship (4.33) to the balanced data set that led to the means of Table 4.20 via the reverse Yates algorithm. Table 4.23 gives the details. The fitted values in the final column are exactly as shown earlier in Figure 4.29.

Table 4.23
The Reverse Yates Algorithm Applied to Fitting the "A and B Main Effects Only" Equation (4.33) to the Data of Table 4.20

Fitted Effect	Value	Cycle 1	Cycle 2	Cycle 3 ($\hat{y}$)
abc_{222}	0	0	0	$3095 = \hat{y}_{abc}$
bc_{22}	0	0	3095	$2255 = \hat{y}_{bc}$
ac_{22}	0	315	0	$2465 = \hat{y}_{ac}$
c_2	0	2780	2255	$1625 = \hat{y}_{c}$
ab_{22}	0	0	0	$3095 = \hat{y}_{ab}$
b_2	315	0	2465	$2255 = \hat{y}_{b}$
a_2	420	315	0	$2465 = \hat{y}_{a}$
$\bar{y}_{...}$	2360	1940	1625	$1625 = \hat{y}_{(1)}$

The importance of two-level factorials

The restriction to two-level factors that makes these notational and computational devices possible is not as specialized as it may at first seem. When an engineer wishes to study the effects of a large number of factors, even 2^p will be a large number of conditions to investigate. If more than two levels of factors are considered, the sheer size of a complete factorial study quickly becomes unmanageable. Recognizing this, two-level studies are often used for screening to identify a few (from many) process variables for subsequent study at more levels on the basis of their large perceived effects in the screening study. So this 2^p material is in fact quite important to the practice of engineering statistics.

Section 3 Exercises

1. Since the data of Exercise 2 of Section 4.2 have complete factorial structure, it is possible (at least temporarily) to ignore the fact that the two experimental factors are basically quantitative and make a factorial analysis of the data.

(a) Compute all fitted factorial main effects and interactions for the data of Exercise 2 of Section 4.2. Interpret the relative sizes of these fitted effects, using a interaction plot like Figure 4.22 to facilitate your discussion.

(b) Compute nine fitted responses for the "main effects only" explanation of y, $y \approx \mu + \alpha_i + \beta_j$. Plot these versus level of the NaOH variable, connecting fitted values having the same level of the Time variable with line segments, as in Figure 4.23. Discuss how this plot compares to the two plots of fitted y versus x_1 made in Exercise 2 of Section 4.2.

(c) Use the fitted values computed in (b) and find a value of R^2 appropriate to the "main effects only" representation of y. How does it compare to the R^2 values from multiple regressions? Also use the fitted values to compute

residuals for this "main effects only" representation. Plot these (versus level of NaOH, level of Time, and $\hat{y}$, and in normal plot form). What do they indicate about the present "no interaction" explanation of specific area?

2. Bachman, Herzberg, and Rich conducted a 2^3 factorial study of fluid flow through thin tubes. They measured the time required for the liquid level in a fluid holding tank to drop from 4 in. to 2 in. for two drain tube diameters and two fluid types. Two different technicians did the measuring. Their data are as follows:

Technician	Diameter (in.)	Fluid	Time (sec)
1	.188	water	21.12, 21.11, 20.80
2	.188	water	21.82, 21.87, 21.78
1	.314	water	6.06, 6.04, 5.92
2	.314	water	6.09, 5.91, 6.01
1	.188	ethylene glycol	51.25, 46.03, 46.09
2	.188	ethylene glycol	45.61, 47.00, 50.71
1	.314	ethylene glycol	7.85, 7.91, 7.97
2	.314	ethylene glycol	7.73, 8.01, 8.32

(a) Compute (using the Yates algorithm or otherwise) the values of all the fitted main effects, two-way interactions, and three-way interactions for these data. Do any simple interpretations of these suggest themselves?

(b) The students actually had some physical theory suggesting that the log of the drain time might be a more convenient response variable than the raw time. Take the logs of the y's and recompute the factorial effects. Does an interpretation of this system in terms of only main effects seem more plausible on the log scale than on the original scale?

(c) Considering the logged drain times as the responses, find fitted values and residuals for a "Diameter and Fluid main effects only" explanation of these data. Compute R^2 appropriate to such a view and compare it to R^2 that results from using all factorial effects to describe log drain time. Make and interpret appropriate residual plots.

(d) Based on the analysis from (c), what change in log drain time seems to accompany a change from .188 in. diameter to .314 in. diameter? What does this translate to in terms of raw drain time? Physical theory suggests that raw time is inversely proportional to the fourth power of drain tube radius. Does your answer here seem compatible with that theory? Why or why not?

3. When analyzing a full factorial data set where the factors involved are quantitative, either the surface-fitting technology of Section 4.2 or the factorial analysis material of Section 4.3 can be applied. What practical engineering advantage does the first offer over the second in such cases?

4.4 Transformations and Choice of Measurement Scale (*Optional*)

Sections 4.2 and 4.3 are an introduction to one of the main themes of engineering statistical analysis: the discovery and use of simple structure in complicated situations. Sometimes this can be done by reexpressing variables on some other (nonlinear) scales of measurement besides the ones that first come to mind. That is, sometimes simple structure may not be obvious on initial scales of measurement, but may emerge after some or all variables have been transformed. This section presents several examples where transformations are helpful. In the process, some comments about commonly used types of transformations, and more specific reasons for using them, are offered.

4.4.1 Transformations and Single Samples

In Chapter 5, there are a number of standard theoretical distributions. When one of these standard models can be used to describe a response y, all that is known about the model can be brought to bear in making predictions and inferences regarding y. However, when no standard distributional shape can be found to describe y, it may nevertheless be possible to so describe $g(y)$ for some function $g(\cdot)$.

Example 11

Discovery Times at an Auto Shop

Elliot, Kibby, and Meyer studied operations at an auto repair shop. They collected some data on what they called the "discovery time" associated with diagnosing what repairs the mechanics were going to recommend to the car owners. Thirty such discovery times (in minutes) are given in Figure 4.31, in the form of a stem-and-leaf plot.

The stem-and-leaf plot shows these data to be somewhat skewed to the right. Many of the most common methods of statistical inference are based on an assumption that a data-generating mechanism will in the long run produce not skewed, but rather symmetrical and bell-shaped data. Therefore, using these methods to draw inferences and make predictions about discovery times at this shop is highly questionable. However, suppose that some transformation could be applied to produce a bell-shaped distribution of transformed discovery times. The standard methods could be used to draw inferences about transformed discovery times, which could then be translated (by undoing the transformation) to inferences about raw discovery times.

One common transformation that has the effect of shortening the right tail of a distribution is the logarithmic transformation, $g(y) = \ln(y)$. To illustrate its use in the present context, normal plots of both discovery times and log discovery times are given in Figure 4.32. These plots indicate that Elliot, Kibby, and Meyer could not have reasonably applied standard methods of inference to the discovery times, but they could have used the methods with log discovery times. The second normal plot is far more linear than the first.

0	4 3
0	6 5 5 5 6 6 8 9 8 8
1	4 0 3 0
1	7 5 9 5 6 9 5
2	0
2	9 5 7 9
3	2
3	6

Figure 4.31 Stem-and-leaf plot of discovery times

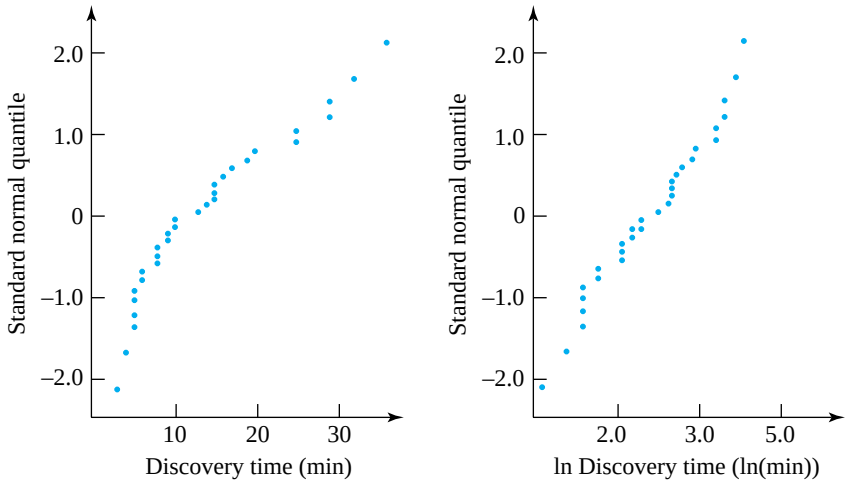

Figure 4.32 Normal plots for discovery times and log discovery times

The logarithmic transformation was useful in the preceding example in reducing the skewness of a response distribution. Some other transformations commonly employed to change the shape of a response distribution in statistical engineering studies are the **power transformations**,

Power transformations

$$g(y) = (y - \gamma)^{\alpha} \tag{4.34}$$

In transformation (4.34), the number γ is often taken as a threshold value, corresponding to a minimum possible response. The number α governs the basic shape of a plot of $g(y)$ versus y. For $\alpha > 1$, transformation (4.34) tends to lengthen the right tail of a distribution for y. For $0 < \alpha < 1$, the transformation tends to shorten the right tail of a distribution for y, the shortening becoming more drastic as α approaches 0 but not as pronounced as that caused by the **logarithmic transformation**

Logarithmic transformation

$$g(y) = \ln(y - \gamma)$$

4.4.2 Transformations and Multiple Samples

Comparing several sets of process conditions is one of the fundamental problems of statistical engineering analysis. It is advantageous to do the comparison on a scale where the samples have comparable variabilities, for at least two reasons. The first is the obvious fact that comparisons then reduce simply to comparisons between response means. Second, standard methods of statistical inference often have well-understood properties only when response variability is comparable for the different sets of conditions.

When response variability is not comparable under different sets of conditions, a transformation can sometimes be applied to all observations to remedy this. This possibility of **transforming to stabilize variance** exists when response variance is roughly a function of response mean. Some theoretical calculations suggest the following guidelines as a place to begin looking for an appropriate variance-stabilizing transformation:

Transformations to stabilize response variance

1. If response standard deviation is approximately proportional to response mean, try a logarithmic transformation.
2. If response standard deviation is approximately proportional to the δ power of the response mean, try transformation (4.34) with $\alpha = 1 - \delta$.

Where several samples (and corresponding $\bar{y}$ and s values) are involved, an empirical way of investigating whether (1) or (2) above might be useful is to plot $\ln(s)$ versus $\ln(\bar{y})$ and see if there is approximate linearity. If so, a slope of roughly 1 makes (1) appropriate, while a slope of $\delta \neq 1$ signals what version of (2) might be helpful.

In addition to this empirical way of identifying a potentially variance-stabilizing transformation, theoretical considerations can sometimes provide guidance. Standard theoretical distributions (like those introduced in Chapter 5) have their own relationships between their (theoretical) means and variances, which can help pick out an appropriate version of (1) or (2) above.

4.4.3 Transformations and Simple Structure in Multifactor Studies

In Section 4.2, Taylor's equation for tool life y in terms of cutting speed x was advantageously reexpressed as a linear equation for $\ln(y)$ in terms of $\ln(x)$. This is just one manifestation of the general fact that many approximate laws of physical science and engineering are **power laws**, expressing one quantity as a product of a constant and powers (some possibly negative) of other quantities. That is, they are of the form

A power law

$$y \approx \alpha x_1^{\beta_1} x_2^{\beta_2} \cdots x_k^{\beta_k} \tag{4.35}$$

Of course, upon taking logarithms in equation (4.35),

$$\ln(y) \approx \ln(\alpha) + \beta_1 \ln(x_1) + \beta_2 \ln(x_2) + \cdots + \beta_k \ln(x_k) \tag{4.36}$$

which immediately suggests the wide usefulness of the logarithmic transformation for both y and x variables in surface-fitting applications involving power laws.

But there is something else in display (4.36) that bears examination: The k functions of the fundamental x variables enter the equation **additively**. In the language of the previous section, there are *no interactions* between the factors whose levels are specified by the variables $x_1, x_2, \ldots, x_k$. This suggests that even in studies involving only seemingly qualitative factors, if a power law for y is at work and the factors

act on different fundamental variables x, a logarithmic transformation will tend to create a simple structure. It will do so by eliminating the need for interactions in describing the response.

Example 12

Daniel's Drill Advance Rate Study

In his book *Applications of Statistics to Industrial Experimentation*, Cuthbert Daniel gives an extensive discussion of an unreplicated 2^4 factorial study of the behavior of a new piece of drilling equipment. The response y is a rate of advance of the drill (no units are given), and the experimental factors are Load on the small stone drill (A), Flow Rate through the drill (B), Rotational Speed (C), and Type of Mud used in drilling (D). Daniel's data are given in Table 4.24.

Application of the Yates algorithm to the data in Table 4.24 ($p = 4$ cycles are required, as is division of the results of the last cycle by 2^4) gives the fitted effects:

$$\bar{y}_{\cdots} = 6.1550$$

$a_2 = .4563$	$b_2 = 1.6488$	$c_2 = 3.2163$	$d_2 = 1.1425$
$ab_{22} = .0750$	$ac_{22} = .2975$	$ad_{22} = .4213$	
$bc_{22} = .7525$	$bd_{22} = .2213$	$cd_{22} = .7987$	
$abc_{222} = .0838$	$abd_{222} = .2950$	$acd_{222} = .3775$	$bcd_{222} = .0900$
$abcd_{2222} = .2688$			

Looking at the magnitudes of these fitted effects, the candidate relationships

$$y \approx \mu + \beta_j + \gamma_k + \delta_l \tag{4.37}$$

and

$$y \approx \mu + \beta_j + \gamma_k + \delta_l + \beta\gamma_{jk} + \gamma\delta_{kl} \tag{4.38}$$

Table 4.24
Daniel's 2^4 Drill Advance Rate Data

Combination	y	Combination	y
(1)	1.68	d	2.07
a	1.98	ad	2.44
b	3.28	bd	4.09
ab	3.44	abd	4.53
c	4.98	cd	7.77
ac	5.70	acd	9.43
bc	9.97	bcd	11.75
abc	9.07	abcd	16.30

Example 12 (*continued*)

are suggested. (The five largest fitted effects are, in order of decreasing magnitude, the main effects of C, B, and D, and then the two-factor interactions of C with D and B with C.) Fitting equation (4.37) to the balanced data of Table 4.24 produces $R^2 = .875$, and fitting relationship (4.38) produces $R^2 = .948$. But upon closer examination, neither fitted equation turns out to be a very good description of these data.

Figure 4.33 shows a normal plot and a plot against $\hat{y}$ for residuals from a fitted version of equation (4.37). It shows that the fitted version of equation (4.37) produces several disturbingly large residuals and fitted values that are systematically too small for responses that are small and large, but too large for moderate responses. Such a curved plot of residuals versus $\hat{y}$ in general suggests that a nonlinear transformation of y may potentially be effective.

The reader is invited to verify that residual plots for equation (4.38) look even worse than those in Figure 4.33. In particular, it is the bigger responses that are

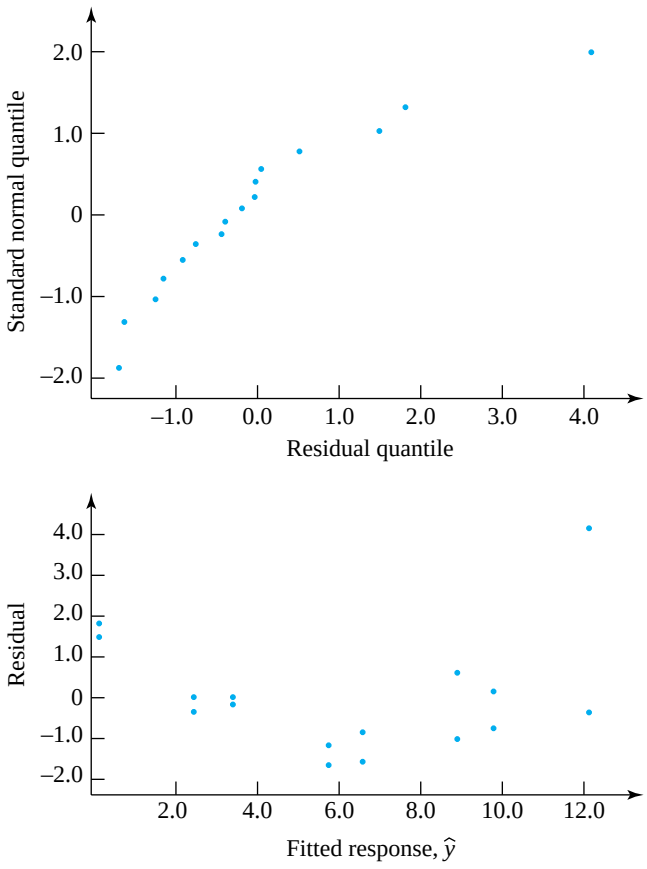

Figure 4.33 Residual plots from fitting equation (4.37) to Daniel's data

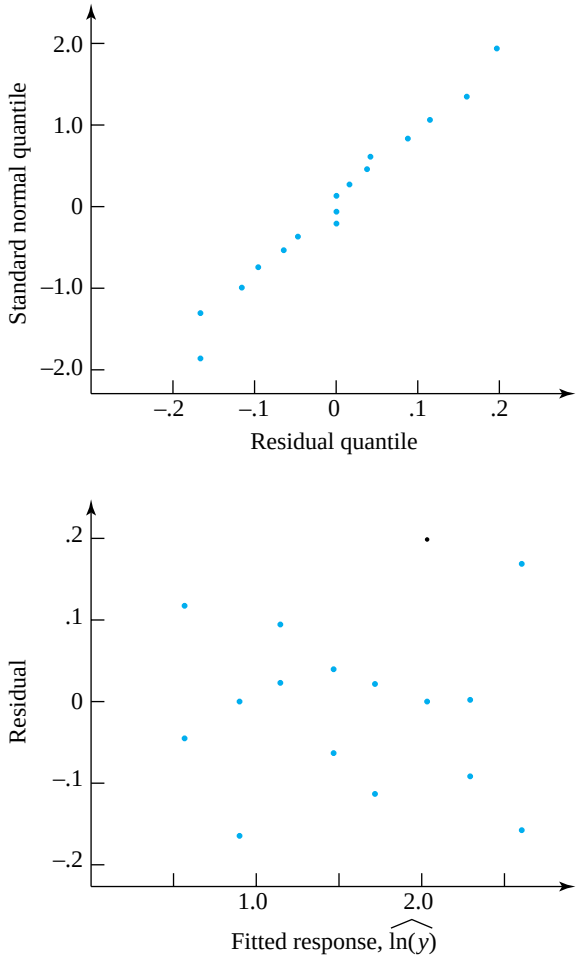

Figure 4.34 Residual plots from fitting equation (4.39) to Daniel's data

fitted relatively badly by relationship (4.38). This is an unfortunate circumstance, since presumably one study goal is the optimization of response.

But using $y' = \ln(y)$ as a response variable, the situation is much different. The Yates algorithm produces the following fitted effects.

$$\overline{y'}_{....} = 1.5977$$

$a_2 = .0650$	$b_2 = .2900$	$c_2 = .5772$	$d_2 = .1633$
$ab_{22} = -.0172$	$ac_{22} = .0052$	$ad_{22} = .0334$	
$bc_{22} = -.0251$	$bd_{22} = -.0075$	$cd_{22} = .0491$	
$abc_{222} = .0052$	$abd_{222} = .0261$	$acd_{222} = .0266$	$bcd_{222} = -.0173$
$abcd_{2222} = .0193$			

Example 12 (*continued*)

For the logged drill advance rates, the simple relationship

$$\ln(y) \approx \mu + \beta_j + \gamma_k + \delta_l \qquad \textbf{(4.39)}$$

yields $R^2 = .976$ and absolutely unremarkable residuals. Figure 4.34 shows a normal plot of these and a plot of them against $\widehat{\ln(y)}$.

The point here is that the logarithmic scale appears to be the natural one on which to study drill advance rate. The data can be better described on the log scale without using interaction terms than is possible with interactions on the original scale.

There are sometimes other reasons to consider a logarithmic transformation of a response variable in a multifactor study, besides its potential to produce simple structure. In cases where responses vary over several orders of magnitude, simple curves and surfaces typically don't fit raw y values very well, but they can do a much better job of fitting $\ln(y)$ values (which will usually vary over less than a single order of magnitude). Another potentially helpful property of a log-transformed analysis is that it will never yield physically impossible negative fitted values for a positive variable y. In contrast, an analysis on an original scale of measurement can, rather embarrassingly, do so.

Example 13

A 3^2 Factorial Chemical Process Experiment

The data in Table 4.25 are from an article by Hill and Demler ("More on Planning Experiments to Increase Research Efficiency," *Industrial and Engineering Chemistry*, 1970). The data concern the running of a chemical process where the objective is to achieve high yield y_1 and low filtration time y_2 by choosing settings for Condensation Temperature, x_1, and the Amount of B employed, x_2.

For purposes of this example, consider the second response, filtration time. Fitting the approximate (quadratic) relationship

$$y_2 \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

to these data produces the equation

$$\hat{y}_2 = 5179.8 - 56.90x_1 - 146.0x_2 + .1733x_1^2 + 1.222x_2^2 + .6837x_1x_2 \qquad \textbf{(4.40)}$$

and $R^2 = .866$. Equation (4.40) defines a bowl-shaped surface in three dimensions, which has a minimum at about the set of conditions $x_1 = 103.2^\circ$C and $x_2 = 30.88$ cc. At first glance, it might seem that the development of equation

Table 4.25
Yields and Filtration Times in a 3^2 Factorial Chemical Process Study

x_1, Condensation Temperature (°C)	x_2, Amount of B (cc)	y_1, Yield (g)	y_2, Filtration Time (sec)
90	24.4	21.1	150
90	29.3	23.7	10
90	34.2	20.7	8
100	24.4	21.1	35
100	29.3	24.1	8
100	34.2	22.2	7
110	24.4	18.4	18
110	29.3	23.4	8
110	34.2	21.9	10

(4.40) rates as a statistical engineering success story. But there is the embarrassing fact that upon substituting $x_1 = 103.2$ and $x_2 = 30.88$ into equation (4.40), one gets $\hat{y}_2 = -11$ sec, hardly a possible filtration time.

Looking again at the data, it is not hard to see what has gone wrong. The largest response is more than 20 times the smallest. So in order to come close to fitting both the extremely large and more moderate responses, the fitted quadratic surface needs to be very steep—so steep that it is forced to dip below the (x_1, x_2)-plane and produce negative $\hat{y}_2$ values before it can "get turned around" and start to climb again as it moves away from the point of minimum $\hat{y}_2$ toward larger x_1 and x_2.

One cure for the problem of negative predicted filtration times is to use $\ln(y_2)$ as a response variable. Values of $\ln(y_2)$ are given in Table 4.26 to illustrate the moderating effect the logarithm has on the factor of 20 disparity between the largest and smallest filtration times.

Fitting the approximate quadratic relationship

$$\ln(y_2) \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

to the $\ln(y_2)$ values produces the equation

$$\widehat{\ln(y_2)} = 99.69 - .8869x_1 - 3.348x_2 + .002506x_1^2 + .03375x_2^2 + .01196x_1x_2 \quad \textbf{(4.41)}$$

and $R^2 = .975$. Equation (4.41) also represents a bowl-shaped surface in three dimensions and has a minimum approximately at the set of conditions $x_1 = 101.5$°C and $x_2 = 31.6$ cc. The minimum fitted log filtration time is $\widehat{\ln(y_2)} = 1.7582$ ln (sec), which translates to a filtration time of 5.8 sec, a far more sensible value than the negative one given earlier.

Example 13 (*continued*)

Table 4.26
Raw Filtration Times and Corresponding Logged Filtration Times

y_2, Filtration Time (sec)	$\ln(y_2)$, Log Filtration Time (ln(sec))
150	5.0106
10	2.3026
8	2.0794
35	3.5553
8	2.0794
7	1.9459
18	2.8904
8	2.0794
10	2.3026

The taking of logs in this example had two beneficial effects. The first was to cut the ratio of largest response to smallest down to about 2.5 (from over 20), allowing a good fit (as measured by R^2) for a fitted quadratic in two variables, x_1 and x_2. The second was to ensure that minimum predicted filtration time was positive.

Of course, other transformations besides the logarithmic one are also useful in describing the structure of multifactor data sets. Sometimes they are applied to the responses and sometimes to other system variables. As an example of a situation where a power transformation like that specified by equation (4.34) is useful in understanding the structure of a sample of bivariate data, consider the following.

Example 14

Yield Strengths of Copper Deposits and Hall-Petch Theory

In their article "Mechanical Property Testing of Copper Deposits for Printed Circuit Boards" (*Plating and Surface Finishing*, 1988), Lin, Kim, and Weil present some data relating the yield strength of electroless copper deposits to the average grain diameters measured for these deposits. The values in Table 4.27 were deduced from a scatterplot in their paper. These values are plotted in Figure 4.35. They don't seem to promise a simple relationship between grain diameter and yield strength. But in fact, the so called Hall-Petch relationship says that yield strengths of most crystalline materials are proportional to the reciprocal square root of grain diameter. That is, Hall-Petch theory predicts a linear relationship between y and $x^{-.5}$ or between x and y^{-2}. Thus, before trying to further detail the relationship between the two variables, application of transformation (4.34) with $\alpha = -.5$ to x or transformation (4.34) with $\alpha = -2$ to y seems in order. Figure 4.36 shows the partial effectiveness of the reciprocal square root transformation (applied to x) in producing a linear relationship in this context.

Table 4.27
Average Grain Diameters and Yield Strengths for Copper Deposits

x, Average Grain Diameter (μm)	y, Yield Strength (MPa)	x, Average Grain Diameter (μm)	y, Yield Strength (MPa)
.22	330	.48	236
.27	370	.49	224
.33	266	.51	236
.41	270	.90	210

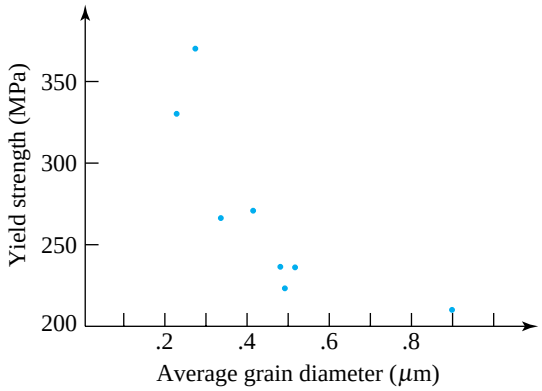

Figure 4.35 Scatterplot of yield strength versus average grain diameter

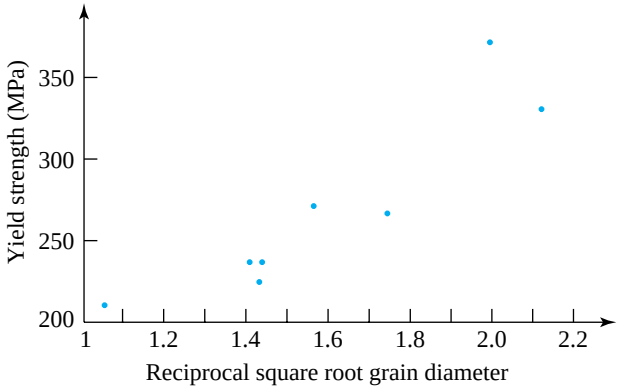

Figure 4.36 Scatterplot of yield strength versus the reciprocal square root average grain diameter

In the preceding example, a directly applicable and well-known physical theory suggests a natural transformation. Sometimes physical or mathematical considerations that are related to a problem, but do not directly address it, may also suggest some things to try in looking for transformations to produce simple structure. For example, suppose some other property besides yield strength were of interest and thought to be related to grain size. If a relationship with diameter is not obvious, quantifying grain size in terms of cross-sectional area or volume might be considered, and this might lead to squaring or cubing a measured diameter. To take a different example, if some handling characteristic of a car is thought to be related to its speed and a relationship with velocity is not obvious, you might remember that kinetic energy is related to velocity squared, thus being led to square the velocity.

The goal of data transformation

To repeat the main point of this section, the search for appropriate transformations is a quest for measurement scales on which structure is transparent and simple. If the original/untransformed scales are the most natural ones on which to report the findings of a study, the data analysis should be done on the transformed scales but then "untransformed" to state the final results.

Section 4 Exercises

1. What are benefits that can sometimes be derived from transforming data before applying standard statistical techniques?
2. Suppose that a response variable, y, obeys an approximate power law in at least two quantitative variables (say, x_1 and x_2). Will there be important interactions? If the log of y is analyzed instead, will there be important interactions? (In order to make this concrete, you may if you wish consider the relationship $y \approx kx_1^2x_2^{-3}$. Plot, for at least two different values of x_2, y as a function of x_1. Then plot, for at least two different values of x_2, $\ln(y)$ as a function of x_1. What do these plots show in the way of parallelism?)

4.5 Beyond Descriptive Statistics

We hope that these first four chapters have made you genuinely ready to accept the need for methods of formal statistical inference. Many real data sets have been examined, and many instances of useful structure have been discovered—this in spite of the fact that the structure is often obscured by what might be termed *background noise*. Recognizing the existence of such variation, one realizes that the data in hand are probably not a perfect representation of the population or process from which they were taken. Thus, generalizing from the sample to a broader sphere will have to be somehow hedged. To this point, the hedging has been largely verbal, specific to the case, and qualitative. There is a need for ways to quantitatively express the precision and reliability of any generalizations about a population or process that are made from data in hand. For example, the chemical filtration time problem of Example 13 produced the conclusion that with the temperature set at 101.5°C and using 31.6 cc of B, a predicted filtration time is 5.8 sec. But will the next time be 5.8 sec $\pm$ 3 sec or $\pm$ .05 sec? If you decide on $\pm$ *somevalue*, how sure can you be of those tolerances?

In order to quantify precision and reliability for inferences based on samples, the mathematics of probability must be employed. Mathematical descriptions of data generation that are applicable to the original data collection (and any future collection) are necessary. Those mathematical models must explicitly allow for the kind of variation that has been faced in the last two chapters.

The models that are most familiar to engineers do not explicitly account for variation. Rather, they are **deterministic**. For example, Newtonian physics predicts that the displacement of a body in free fall in a time t is exactly $\frac{1}{2}gt^2$. In this statement, there is no explicit allowance for variability. Any observed deviation from the Newtonian predictions is completely unaccounted for. Thus, there is really no logical framework in which to extrapolate from data that don't fit Newtonian predictions exactly.

Stochastic (or probabilistic) models do explicitly incorporate the feature that even measurements generated under the same set of conditions will exhibit variation. Therefore, they can function as descriptions of real-world data collection processes, where many small, unidentifiable causes act to produce the background noise seen in real data sets. Variation is predicted by stochastic or probabilistic models. So they provide a logical framework in which to quantify precision and reliability and to extrapolate from noisy data to contexts larger than the data set in hand.

In the next chapter, some fundamental concepts of probability will be introduced. Then Chapter 6 begins to use probability as a tool in statistical inference.

Section 5 Exercises

1. Read again Section 1.4 and the present one. Then describe in your own words the difference between deterministic and stochastic/probabilistic models. Give an example of a deterministic model that is useful in your field.

Chapter 4 Exercises

1. Nicholson and Bartle studied the effect of the water/cement ratio on 14-day compressive strength for Portland cement concrete. The water/cement ratios (by volume) and compressive strengths of nine concrete specimens are given next.

Water/Cement Ratio, x	14-Day Compressive Strength, y (psi)
.45	2954, 2913, 2923
.50	2743, 2779, 2739
.55	2652, 2607, 2583

   (a) Fit a line to the data here via least squares, showing the hand calculations.
   (b) Compute the sample correlation between x and y by hand. Interpret this value.
   (c) What fraction of the raw variability in y is accounted for in the fitting of a line to the data?
   (d) Compute the residuals from your fitted line and make a normal plot of them. Interpret this plot.
   (e) What compressive strength would you predict, based on your calculations from (a), for specimens made using a .48 water/cement ratio?
   (f) Use a statistical package to find the least squares line, the sample correlation, R^2, and the residuals for this data set.

2. Griffith and Tesdall studied the elapsed time in $\frac{1}{4}$ mile runs of a Camaro Z-28 fitted with different

sizes of carburetor jetting. Their data from six runs of the car follow:

Jetting Size, x	Elapsed Time, y (sec)
66	14.90
68	14.67
70	14.50
72	14.53
74	14.79
76	15.02

(a) What is an obvious weakness in the students' data collection plan?
(b) Fit both a line and a quadratic equation ($y \approx \beta_0 + \beta_1 x + \beta_2 x^2$) to these data via least squares. Plot both of these equations on a scatterplot of the data.
(c) What fractions of the raw variation in elapsed time are accounted for by the two different fitted equations?
(d) Use your fitted quadratic equation to predict an optimal jetting size (allowing fractional sizes).

3. The following are some data taken from "Kinetics of Grain Growth in Powder-formed IN-792: A Nickel-Base Super-alloy" by Huda and Ralph (*Materials Characterization*, September 1990). Three different Temperatures, x_1 (°K), and three different Times, x_2 (min), were used in the heat treating of specimens of a material, and the response

$$y = \text{mean grain diameter } (\mu\text{m})$$

was measured.

Temperature, x_1	Time, x_2	Grain Size, y
1443	20	5
1443	120	6
1443	1320	9
1493	20	14
1493	120	17
1493	1320	25
1543	20	29
1543	120	38
1543	1320	60

(a) What type of data structure did the researchers employ? (Use the terminology of Section 1.2.) What was an obvious weakness in their data collection plan?
(b) Use a regression program to fit the following equations to these data:

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$$
$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 \ln(x_2)$$
$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 \ln(x_2) + \beta_3 x_1 \ln(x_2)$$

What are the R^2 values for the three different fitted equations? Compare the three fitted equations in terms of complexity and apparent ability to predict y.
(c) Compute the residuals for the third fitted equation in (b). Plot them against x_1, x_2, and $\hat{y}$. Also normal-plot them. Do any of these plots suggest that the third fitted equation is inadequate as summary of these data? What, if any, possible improvement over the third equation is suggested by these plots?
(d) As a means of understanding the nature of the third fitted equation in (b), make a scatterplot of y vs. x_2 using a logarithmic scale for x_2. On this plot, plot three lines representing $\hat{y}$ as a function of x_2 for the three different values of x_1. Qualitatively, how would a similar plot for the second equation differ from this one?
(e) Using the third equation in (b), what mean grain diameter would you predict for $x_1 = 1500$ and $x_2 = 500$?
(f) It is possible to ignore the fact that the Temperature and Time factors are quantitative and make a factorial analysis of these data. Do so. Begin by making an interaction plot similar to Figure 4.22 for these data. Based on that plot, discuss the apparent relative sizes of the Time and Temperature main effects and the Time × Temperature interactions. Then compute the fitted factorial effects (the fitted main effects and interactions).

4. The article "Cyanoacetamide Accelerators for the Epoxide/Isocyanate Reaction" by Eldin and Renner (*Journal of Applied Polymer Science*, 1990)

reports the results of a 2^3 factorial experiment. Using cyanoacetamides as catalysts for an epoxy/isocyanate reaction, various mechanical properties of a resulting polymer were studied. One of these was

$$y = \text{impact strength (kJ/mm}^2\text{)}$$

The three experimental factors employed and their corresponding experimental levels were as follows:

Factor A	Initial Epoxy/Isocyanate Ratio 0.4 (−) vs. 1.2 (+)
Factor B	Flexibilizer Concentration 10 mol % (−) vs. 40 mol % (+)
Factor C	Accelerator Concentration 1/240 mol % (−) vs. 1/30 mol% (+)

(The flexibilizer and accelerator concentrations are relative to the amount of epoxy present initially.) The impact strength data obtained (one observation per combination of levels of the three factors) were as follows:

Combination	y	Combination	y
(1)	6.7	c	6.3
a	11.9	ac	15.1
b	8.5	bc	6.7
ab	16.5	abc	16.4

(a) What is an obvious weakness in the researchers' data collection plan?

(b) Use the Yates algorithm and compute fitted factorial effects corresponding to the "all high" treatment combination (i.e., compute $\bar{y}_{\ldots}$, a_2, b_2, etc.). Interpret these in the context of the original study. (Describe in words which factors and/or combinations of factors appear to have the largest effect(s) on impact strength and interpret the sign or signs.)

(c) Suppose only factor A is judged to be of importance in determining impact strength. What predicted/fitted impact strengths correspond to this judgment? (Find $\hat{y}$ values using the reverse Yates algorithm or otherwise.) Use these eight values of $\hat{y}$ and compute R^2 for the "A main effects only" description of impact strength. (The formula in Definition 3 works in this context as well as in regression.)

(d) Now recognize that the experimental factors here are quantitative, so methods of curve and surface fitting may be applicable. Fit the equation $y \approx \beta_0 + \beta_1$(epoxy/isocyanate ratio) to the data. What eight values of $\hat{y}$ and value of R^2 accompany this fit?

5. Timp and M-Sidek studied the strength of mechanical pencil lead. They taped pieces of lead to a desk, with various lengths protruding over the edge of the desk. After fitting a small piece of tape on the free end of a lead piece to act as a stop, they loaded it with paper clips until failure. In one part of their study, they tested leads of two different Diameters, used two different Lengths protruding over the edge of the desk, and tested two different lead Hardnesses. That is, they ran a 2^3 factorial study. Their factors and levels were as follows:

Factor A	Diameter	.3 mm (−) vs. .7 mm (+)
Factor B	Length Protruding	3 cm (−) vs. 4.5 cm (+)
Factor C	Hardness	B (−) vs. 2H (+)

and $m = 2$ trials were made at each of the $2^3 = 8$ different sets of conditions. The data the students obtained are given here.

Combination	Number of Clips
(1)	13, 13
a	74, 76
b	9, 10
ab	43, 42
c	16, 15
ac	89, 88
bc	10, 12
abc	54, 55

(a) It appears that analysis of these data in terms of the natural logarithms of the numbers of

clips first causing failure is more straightforward than the analysis of the raw numbers of clips. So take natural logs and compute the fitted 2^3 factorial effects. Interpret these. In particular, what (in quantitative terms) does the size of the fitted A main effect say about lead strength? Does lead hardness appear to play a dominant role in determining this kind of breaking strength?

(b) Suppose only the main effects of Diameter are judged to be of importance in determining lead strength. Find a predicted log breaking strength for .7 mm, 2H lead when the length protruding is 4.5 cm. Use this to predict the number of clips required to break such a piece of lead.

(c) What, if any, engineering reasons do you have for expecting the analysis of breaking strength to be more straightforward on the log scale than on the original scale?

6. Ceramic engineering researchers Leigh and Taylor, in their paper "Computer Generated Experimental Designs" (*Ceramic Bulletin*, 1990), studied the packing properties of crushed T-61 tabular alumina powder. The densities of batches of the material were measured under a total of eight different sets of conditions having a 2^3 factorial structure. The following factors and levels were employed in the study:

Factor A	Mesh Size of Powder Particles 6 mesh (−) vs. 60 mesh (+)
Factor B	Volume of Graduated Cylinder 100 cc (−) vs. 500 cc (+)
Factor C	Vibration of Cylinder no (−) vs. yes (+)

The mean densities (in g/cc) obtained in $m = 5$ determinations for each set of conditions were as follows:

$\bar{y}_{(1)} = 2.348$ $\quad \bar{y}_{a} = 2.080$
$\bar{y}_{b} = 2.298$ $\quad \bar{y}_{ab} = 1.980$
$\bar{y}_{c} = 2.354$ $\quad \bar{y}_{ac} = 2.314$
$\bar{y}_{bc} = 2.404$ $\quad \bar{y}_{abc} = 2.374$

(a) Compute the fitted 2^3 factorial effects (main effects, 2-factor interactions and 3-factor interactions) corresponding to the following set of conditions: 60 mesh, 500 cc, vibrated cylinder.

(b) If your arithmetic for part (a) is correct, you should have found that the largest of the fitted effects (in absolute value) are (respectively) the C main effect, the A main effect, and then the AC 2-factor interaction. (The next largest fitted effect is only about half of the smallest of these, the AC interaction.) Now, suppose you judge these three fitted effects to summarize the main features of the data set. Interpret this data summary (A and C main effects and AC interactions) in the context of this 3-factor study.

(c) Using your fitted effects from (a) and the data summary from (b) (A and C main effects and AC interactions), what fitted response would you have for these conditions: 60 mesh, 500 cc, vibrated cylinder?

(d) Using your fitted effects from (a), what average change in density would you say accompanies the vibration of the graduated cylinder before density determination?

7. The article "An Analysis of Transformations" by Box and Cox (*Journal of the Royal Statistical Society, Series B*, 1964) contains a classical unreplicated 3^3 factorial data set originally taken from an unpublished technical report of Barella and Sust. These researchers studied the behavior of worsted yarns under repeated loading. The response variable was

$$y = \text{the numbers of cycles till failure}$$

for specimens tested with various values of

$x_1 =$ length (mm)

$x_2 =$ amplitude of the loading cycle (mm)

$x_3 =$ load (g)

The researchers' data are given in the accompanying table.

x_1	x_2	x_3	y	x_1	x_2	x_3	y
250	8	40	674	300	9	50	438
250	8	45	370	300	10	40	442
250	8	50	292	300	10	45	332
250	9	40	338	300	10	50	220
250	9	45	266	350	8	40	3,636
250	9	50	210	350	8	45	3,184
250	10	40	170	350	8	50	2,000
250	10	45	118	350	9	40	1,568
250	10	50	90	350	9	45	1,070
300	8	40	1,414	350	9	50	566
300	8	45	1,198	350	10	40	1,140
300	8	50	634	350	10	45	884
300	9	40	1,022	350	10	50	360
300	9	45	620				

(a) To find an equation to represent these data, you might first try to fit multivariable polynomials. Use a regression program and fit a full quadratic equation to these data. That is, fit

$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1^2 + \beta_5 x_2^2 + \beta_6 x_3^2 + \beta_7 x_1 x_2 + \beta_8 x_1 x_3 + \beta_9 x_2 x_3$$

to the data. What fraction of the observed variation in y does it account for? In terms of parsimony (or providing a simple data summary), how does this quadratic equation do as a data summary?

(b) Notice the huge range of values of the response variable. In cases like this, where the response varies over an order of magnitude, taking logarithms of the response often helps produce a simple fitted equation. Here, take (natural) logarithms of all of x_1, x_2, x_3, and y, producing (say) x_1', x_2', x_3', and y', and fit the equation

$$y' \approx \beta_0 + \beta_1 x_1' + \beta_2 x_2' + \beta_3 x_3'$$

to the data. What fraction of the observed variability in $y = \ln(y)$ does this equation account for? What change in y' seems to accompany a unit (a 1 $\ln(g)$) increase in x_3'?

(c) To carry the analysis one step further, note that your fitted coefficients for x_1' and x_2' are nearly the negatives of each other. That suggests that y' depends only on the difference between x_1' and x_2'. To see how this works, fit the equation

$$y' \approx \beta_0 + \beta_1 (x_1' - x_2') + \beta_2 x_3'$$

to the data. Compute and plot residuals from this relationship (still on the log scale). How does this relationship appear to do as a data summary? What power law for y (on the original scale) in terms of x_1, x_2, and x_3 (on their original scales) is implied by this last fitted equation? How does this equation compare to the one from (a) in terms of parsimony?

(d) Use your equation from (c) to predict the life of an additional specimen of length 300 mm, at an amplitude of 9 mm, under a load of 45 g. Do the same for an additional specimen of length 325 mm, at an amplitude of 9.5 mm, under a load of 47 g. Why would or wouldn't you be willing to make a similar projection for an additional specimen of length 375 mm, at an amplitude of 10.5 mm, under a load of 51 g?

8. Bauer, Dirks, Palkovic, and Wittmer fired tennis balls out of a "Polish cannon" inclined at an angle of 45°, using three different Propellants and two different Charge Sizes of propellant. They observed the distances traveled in the air by the tennis balls. Their data are given in the accompanying table. (Five trials were made for each Propellant/Charge Size combination and the values given are in feet.)

		Propellant		
		Lighter Fluid	*Gasoline*	*Carburetor Fluid*
Charge Size	*2.5 ml*	58	76	90
		50	79	86
		53	84	79
		49	73	82
		59	71	86
	5.0 ml	65	96	107
		59	101	102
		61	94	91
		68	91	95
		67	87	97

Complete a factorial analysis of these data, including a plot of sample means useful for judging the size of Charge Size × Propellant interactions and the computing of fitted main effects and interactions. Write a paragraph summarizing what these data seem to say about how these two variables affect flight distance.

9. The following data are taken from the article "An Analysis of Means for Attribute Data Applied to a 2^4 Factorial Design" by R. Zwickl (*ASQC Electronics Division Technical Supplement*, Fall 1985). They represent numbers of bonds (out of 96) showing evidence of ceramic pull-out on an electronic device called a dual in-line package. (Low numbers are good.) Experimental factors and their levels were:

Factor A Ceramic Surface
unglazed (−) vs. glazed (+)

Factor B Metal Film Thickness
normal (−) vs. 1.5 times normal (+)

Factor C Annealing Time
normal (−) vs. 4 times normal (+)

Factor D Prebond Clean
normal clean (−) vs. no clean (+)

The resultant numbers of pull-outs for the 2^4 treatment combinations are given next.

Combination	Pull-Outs	Combination	Pull-Outs
(1)	9	c	13
a	70	ac	55
b	8	bc	7
ab	42	abc	19
d	3	cd	5
ad	6	acd	28
bd	1	bcd	3
abd	7	abcd	6

(a) Use the Yates algorithm and identify dominant effects here.

(b) Based on your analysis from (a), postulate a possible "few effects" explanation for these data. Use the reverse Yates algorithm to find fitted responses for such a simplified description of the system. Use the fitted values to compute residuals. Normal-plot these and plot them against levels of each of the four factors, looking for obvious problems with your representation of system behavior.

(c) Based on your "few effects" description of bond strength, make a recommendation for the future making of these devices. (All else being equal, you should choose what appear to be the least expensive levels of factors.)

10. Exercise 5 of Chapter 3 concerns a replicated 3^3 factorial data set (weighings of three different masses on three different scales by three different students). Use a full-featured statistical package that will compute fitted effects for such data and write a short summary report stating what those fitted effects reveal about the structure of the weighings data.

11. When it is an appropriate description of a two-way factorial data set, what practical engineering advantages does a "main effects only" description offer over a "main effects plus interactions" description?

12. The article referred to in Exercise 4 of Section 4.1 actually considers the effects of both cutting speed and feed rate on tool life. The whole data

set from the article follows. (The data in Section 4.1 are the $x_2 = .01725$ data only.)

Cutting Speed, x_1 (sfpm)	Feed, x_2 (ipr)	Tool Life, y (min)
800	.01725	1.00, 0.90, 0.74, 0.66
700	.01725	1.00, 1.20, 1.50, 1.60
700	.01570	1.75, 1.85, 2.00, 2.20
600	.02200	1.20, 1.50, 1.60, 1.60
600	.01725	2.35, 2.65, 3.00, 3.60
500	.01725	6.40, 7.80, 9.80, 16.50
500	.01570	8.80, 11.00, 11.75, 19.00
450	.02200	4.00, 4.70, 5.30, 6.00
400	.01725	21.50, 24.50, 26.00, 33.00

(a) Taylor's expanded tool life equation is $yx_1^{\alpha_1}x_2^{\alpha_2} = C$. This relationship suggests that $\ln(y)$ may well be approximately linear in both $\ln(x_1)$ and $\ln(x_2)$. Use a multiple linear regression program to fit the relationship

$$\ln(y) \approx \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2)$$

to these data. What fraction of the raw variability in $\ln(y)$ is accounted for in the fitting process? What estimates of the parameters α_1, α_2, and C follow from your fitted equation?

(b) Compute and plot residuals (continuing to work on log scales) for the equation you fit in part (a). Make at least plots of residuals versus fitted $\ln(y)$ and both $\ln(x_1)$ and $\ln(x_2)$, and make a normal plot of these residuals. Do these plots reveal any particular problems with the fitted equation?

(c) Use your fitted equation to predict first a log tool life and then a tool life, if in this machining application a cutting speed of 550 and a feed of .01650 is used.

(d) Plot the ordered pairs appearing in the data set in the (x_1, x_2)-plane. Outline a region in the plane where you would feel reasonably safe using the equation you fit in part (a) to predict tool life.

13. K. Casali conducted a gas mileage study on his well-used four-year-old economy car. He drove a 107-mile course a total of eight different times (in comparable weather conditions) at four different speeds, using two different types of gasoline, and ended up with an unreplicated 4×2 factorial study. His data are given in the table below.

Test	Speed (mph)	Gasoline Octane	Gallons Used	Mileage (mpg)
1	65	87	3.2	33.4
2	60	87	3.1	34.5
3	70	87	3.4	31.5
4	55	87	3.0	35.7
5	65	90	3.2	33.4
6	55	90	2.9	36.9
7	70	90	3.3	32.4
8	60	90	3.0	35.7

(a) Make a plot of the mileages that is useful for judging the size of Speed $\times$ Octane interactions. Does it look as if the interactions are large in comparison to the main effects?

(b) Compute the fitted main effects and interactions for the mileages, using the formulas of Section 4.3. Make a plot like Figure 4.23 for comparing the observed mileages to fitted mileages computed supposing that there are no Speed $\times$ Octane interactions.

(c) Now fit the equation

$$\text{Mileage} \approx \beta_0 + \beta_1(\text{Speed}) + \beta_2(\text{Octane})$$

to the data and plot lines representing the predicted mileages versus Speed for both the 87 octane and the 90 octane gasolines on the same set of axes.

(d) Now fit the equation Mileage $\approx \beta_0 + \beta_1$ (Speed) separately, first to the 87 octane data and then to the 90 octane data. Plot the two different lines on the same set of axes.

(e) Discuss the different appearances of the plots you made in parts (a) through (d) of this exercise in terms of how well they fit the original

data and the different natures of the assumptions involved in producing them.

(f) What was the fundamental weakness in Casali's data collection scheme? A weakness of secondary importance has to do with the fact that tests 1–4 were made ten days earlier than tests 5–8. Why is this a potential problem?

14. The article "Accelerated Testing of Solid Film Lubricants" by Hopkins and Lavik (*Lubrication Engineering*, 1972) contains a nice example of the engineering use of multiple regression. In the study, $m = 3$ sets of journal bearing tests were made on a Mil-L-8937 type film at each combination of three different Loads and three different Speeds. The wear lives of journal bearings, y, in hours, are given next for the tests run by the authors.

Speed, x_1 (rpm)	Load, x_2 (psi)	Wear Life, y (hs)
20	3,000	300.2, 310.8, 333.0
20	6,000	99.6, 136.2, 142.4
20	10,000	20.2, 28.2, 102.7
60	3,000	67.3, 77.9, 93.9
60	6,000	43.0, 44.5, 65.9
60	10,000	10.7, 34.1, 39.1
100	3,000	26.5, 22.3, 34.8
100	6,000	32.8, 25.6, 32.7
100	10,000	2.3, 4.4, 5.8

(a) The authors expected to be able to describe wear life as roughly following the relationship $yx_1x_2 = C$, but they did not find this relationship to be a completely satisfactory model. So instead, they tried using the more general relationship $yx_1^{\alpha_1}x_2^{\alpha_2} = C$. Use a multiple linear regression program to fit the relationship

$$\ln(y) \approx \beta_0 + \beta_1 \ln(x_1) + \beta_2 \ln(x_2)$$

to these data. What fraction of the raw variability in $\ln(y)$ is accounted for in the fitting process? What estimates of the parameters α_1, α_2, and C follow from your fitted equation? Using your estimates of α_1, α_2, and C, plot on the same set of (x_1, y) axes the functional relationships between x_1 and y implied by your fitted equation for x_2 equal to 3,000, 6,000, and then 10,000 psi, respectively.

(b) Compute and plot residuals (continuing to work on log scales) for the equation you fit in part (a). Make at least plots of residuals versus fitted $\ln(y)$ and both $\ln(x_1)$ and $\ln(x_2)$, and make a normal plot of these residuals. Do these plots reveal any particular problems with the fitted equation?

(c) Use your fitted equation to predict first a log wear life and then a wear life, if in this application a speed of 20 rpm and a load of 10,000 psi are used.

(d) **(Accelerated life testing)** As a means of trying to make intelligent data-based predictions of wear life at low stress levels (and correspondingly large lifetimes that would be impractical to observe directly), you might (fully recognizing the inherent dangers of the practice) try to extrapolate using the fitted equation. Use your fitted equation to predict first a log wear life and then a wear life if a speed of 15 rpm and load of 1,500 psi are used in this application.

15. The article "Statistical Methods for Controlling the Brown Oxide Process in Multilayer Board Processing" by S. Imadi (*Plating and Surface Finishing*, 1988) discusses an experiment conducted to help a circuit board manufacturer measure the concentration of important components in a chemical bath. Various combinations of levels of

x_1 = % by volume of component A (a proprietary formulation, the major component of which is sodium chlorite)

and

x_2 = % by volume of component B (a proprietary formulation, the major component of which is sodium hydroxide)

were set in the chemical bath, and the variables

$y_1 =$ ml of 1N H_2SO_4 used in the first phase of a titration

and

$y_2 =$ ml of 1N H_2SO_4 used in the second phase of a titration

were measured. Part of the original data collected (corresponding to bath conditions free of Na_2CO_3) follow:

x_1	x_2	y_1	y_2
15	25	3.3	.4
20	25	3.4	.4
20	30	4.1	.4
25	30	4.3	.3
25	35	5.0	.5
30	35	5.0	.3
30	40	5.7	.5
35	40	5.8	.4

(a) Fit equations for both y_1 and y_2 linear in both of the variables x_1 and x_2. Does it appear that the variables y_1 and y_2 are adequately described as linear functions of x_1 and x_2?

(b) Solve your two fitted equations from (a) for x_2 (the concentration of primary interest here) in terms of y_1 and y_2. (Eliminate x_1 by solving the first for x_1 in terms of the other three variables and plugging that expression for x_1 into the second equation.) How does this equation seem to do in terms of, so to speak, predicting x_2 from y_1 and y_2 for the original data? Chemical theory in this situation indicated that $x_2 \approx 8(y_1 - y_2)$. Does your equation seem to do better than the one from chemical theory?

(c) A possible alternative to the calculations in (b) is to simply fit an equation for x_2 in terms of y_1 and y_2 directly via least squares. Fit $x_2 \approx \beta_0 + \beta_1 y_1 + \beta_2 y_2$ to the data, using a regression program. Is this equation the same one you found in part (b)?

(d) If you were to compare the equations for x_2 derived in (b) and (c) in terms of the sum of squared differences between the predicted and observed values of x_2, which is guaranteed to be the winner? Why?

16. The article "Nonbloated Burned Clay Aggregate Concrete" by Martin, Ledbetter, Ahmad, and Britton (*Journal of Materials*, 1972) contains data on both composition and resulting physical property test results for a number of different batches of concrete made using burned clay aggregates. The accompanying data are compressive strength measurements, y (made according to ASTM C 39 and recorded in psi), and splitting tensile strength measurements, x (made according to ASTM C 496 and recorded in psi), for ten of the batches used in the study.

Batch	**1**	**2**	**3**	**4**	**5**
y	1420	1950	2230	3070	3060
x	207	233	254	328	325
Batch	**6**	**7**	**8**	**9**	**10**
y	3110	2650	3130	2960	2760
x	302	258	335	315	302

(a) Make a scatterplot of these data and comment on how linear the relation between x and y appears to be for concretes of this type.

(b) Compute the sample correlation between x and y by hand. Interpret this value.

(c) Fit a line to these data using the least squares principle. Show the necessary hand calculations. Sketch this fitted line on your scatterplot from (a).

(d) About what increase in compressive strength appears to accompany an increase of 1 psi in splitting tensile strength?

(e) What fraction of the raw variability in compressive strength is accounted for in the fitting of a line to the data?

(f) Based on your answer to (c), what measured compressive strength would you predict for a

batch of concrete of this type if you were to measure a splitting tensile strength of 245 psi?

(g) Compute the residuals from your fitted line. Plot the residuals against x and against $\hat{y}$. Then make a normal plot of the residuals. What do these plots indicate about the linearity of the relationship between splitting tensile strength and compressive strength?

(h) Use a statistical package to find the least squares line, the sample correlation, R^2, and the residuals for these data.

(i) Fit the quadratic relationship $y \approx \beta_0 + \beta_1 x + \beta_2 x^2$ to the data, using a statistical package. Sketch this fitted parabola on your scatterplot from part (a). Does this fitted quadratic appear to be an important improvement over the line you fit in (c) in terms of describing the relationship of y to x?

(j) How do the R^2 values from parts (h) and (i) compare? Does the increase in R^2 in part (i) speak strongly for the use of the quadratic (as opposed to linear) description of the relationship of y to x for concretes of this type?

(k) If you use the fitted relationship from part (i) to predict y for $x = 245$, how does the prediction compare to your answer for part (f)?

(l) What do the fitted relationships from parts (c) and (i) give for predicted compressive strengths when $x = 400$ psi? Do these compare to each other as well as your answers to parts (f) and (k)? Why would it be unwise to use either of these predictions without further data collection and analysis?

17. In the previous exercise, both x and y were really response variables. As such, they were not subject to direct manipulation by the experimenters. That made it difficult to get several (x, y) pairs with a single x value into the data set. In experimental situations where an engineer gets to choose values of an experimental variable x, why is it useful/important to get several y observations for at least some x's?

18. Chemical engineering graduate student S. Osoka studied the effects of an agitator speed, x_1, and a polymer concentration, x_2, on percent recoveries of pyrite, y_1, and kaolin, y_2, from a step of an ore refining process. (High pyrite recovery and low kaolin recovery rates were desirable.) Data from one set of $n = 9$ experimental runs are given here.

x_1 (rpm)	x_2 (ppm)	y_1 (%)	y_2 (%)
1350	80	77	67
950	80	83	54
600	80	91	70
1350	100	80	52
950	100	87	57
600	100	87	66
1350	120	67	54
950	120	80	52
600	120	81	44

(a) What type of data structure did the researcher use? (Use the terminology of Section 1.2.) What was an obvious weakness in his data collection plan?

(b) Use a regression program to fit the following equations to these data:

$$y_1 \approx \beta_0 + \beta_1 x_1$$
$$y_1 \approx \beta_0 + \beta_2 x_2$$
$$y_1 \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

What are the R^2 values for the three different fitted equations? Compare the three fitted equations in terms of complexity and apparent ability to predict y_1.

(c) Compute the residuals for the third fitted equation in part (b). Plot them against x_1, x_2, and $\hat{y}_1$. Also normal-plot them. Do any of these plots suggest that the third fitted equation is inadequate as a summary of these data?

(d) As a means of understanding the nature of the third fitted equation from part (b), make a scatterplot of y_1 vs. x_2. On this plot, plot three lines representing $\hat{y}_1$ as a function of x_2 for the three different values of x_1 represented in the data set.

(e) Using the third equation from part (b), what pyrite recovery rate would you predict for $x_1 = 1000$ rpm and $x_2 = 110$ ppm?

(f) Consider also a multivariable quadratic description of the dependence of y_1 on x_1 and x_2. That is, fit the equation

$$y_1 \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$$

to the data. How does the R^2 value here compare with the ones in part (b)? As a means of understanding this fitted equation, plot on a single set of axes the three different quadratic functions of x_2 obtained by holding x_1 at one of the values in the data set.

(g) It is possible to ignore the fact that the speed and concentration factors are quantitative and to make a factorial analysis of these y_1 data. Do so. Begin by making an interaction plot similar to Figure 4.22 for these data. Based on that plot, discuss the apparent relative sizes of the Speed and Concentration main effects and the Speed $\times$ Concentration interactions. Then compute the fitted factorial effects (the fitted main effects and interactions).

(h) If the third equation in part (b) governed y_1, would it lead to Speed $\times$ Concentration interactions? What about the equation in part (f)? Explain.

19. The data given in the previous exercise concern both responses y_1 and y_2. The previous analysis dealt with only y_1. Redo all parts of the problem, replacing the response y_1 with y_2 throughout.

20. K. Fellows conducted a 4-factor experiment, with the response variable the flight distance of a paper airplane when propelled from a launcher fabricated specially for the study. This exercise concerns part of the data he collected, constituting a complete 2^4 factorial. The experimental factors involved and levels used were as given here.

Factor A	Plane Design straight wing (−) vs. t wing (+)
Factor B	Nose Weight none (−) vs. paper clip (+)
Factor C	Paper Type notebook (−) vs. construction (+)
Factor D	Wing Tips straight (−) vs. bent up (+)

The mean flight distances, y (ft), recorded by Fellows for two launches of each plane were as shown in the accompanying table.

(a) Use the Yates algorithm and compute the fitted factorial effects corresponding to the "all high" treatment combination.

(b) Interpret the results of your calculations from (a) in the context of the study. (Describe in words which factors and/or combinations of factors appear to have the largest effect(s) on flight distance. What are the practical implications of these effects?)

Combination	y	Combination	y
(1)	6.25	d	7.00
a	15.50	ad	10.00
b	7.00	bd	10.00
ab	16.50	abd	16.00
c	4.75	cd	4.50
ac	5.50	acd	6.00
bc	4.50	bcd	4.50
abc	6.00	abcd	5.75

(c) Suppose factors B and D are judged to be inert as far as determining flight distance is concerned. (The main effects of B and D and all interactions involving them are negligible.) What fitted/predicted values correspond to this description of flight distance (A and C main effects and AC interactions only)? Use these 16 values of $\hat{y}$ to compute residuals, $y - \hat{y}$. Plot these against $\hat{y}$, levels of A, levels of B, levels of C, and levels of D. Also

normal-plot these residuals. Comment on any interpretable patterns in your plots.

(d) Compute R^2 corresponding to the description of flight distance used in part (c). (The formula in Definition 3 works in this context as well as in regression. So does the representation of R^2 as the squared sample correlation between y and $\hat{y}$.) Does it seem that the grand mean, A and C main effects, and AC 2-factor interactions provide an effective summary of flight distance?

21. The data in the accompanying table appear in the text *Quality Control and Industrial Statistics* by Duncan (and were from a paper of L. E. Simon). The data were collected in a study of the effectiveness of armor plate. Armor-piercing bullets were fired at an angle of 40° against armor plate of thickness x_1 (in .001 in.) and Brinell hardness number x_2, and the resulting so-called ballistic limit, y (in ft/sec), was measured.

x_1	x_2	y	x_1	x_2	y
253	317	927	253	407	1393
258	321	978	252	426	1401
259	341	1028	246	432	1436
247	350	906	250	469	1327
256	352	1159	242	257	950
246	363	1055	243	302	998
257	365	1335	239	331	1144
262	375	1392	242	355	1080
255	373	1362	244	385	1276
258	391	1374	234	426	1062

(a) Use a regression program to fit the following equations to these data:

$$y \approx \beta_0 + \beta_1 x_1$$
$$y \approx \beta_0 + \beta_2 x_2$$
$$y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

What are the R^2 values for the three different fitted equations? Compare the three fitted equations in terms of complexity and apparent ability to predict y.

(b) What is the correlation between x_1 and y? The correlation between x_2 and y?

(c) Based on (a) and (b), describe how strongly Thickness and Hardness appear to affect ballistic limit. Review the raw data and speculate as to why the variable with the smaller influence on y seems to be of only minor importance in this data set (although logic says that it must in general have a sizable influence on y).

(d) Compute the residuals for the third fitted equation from (a). Plot them against x_1, x_2, and $\hat{y}$. Also normal-plot them. Do any of these plots suggest that the third fitted equation is seriously deficient as a summary of these data?

(e) Plot the (x_1, x_2) pairs represented in the data set. Why would it be unwise to use any of the fitted equations to predict y for $x_1 = 265$ and $x_2 = 440$?

22. Basgall, Dahl, and Warren experimented with smooth and treaded bicycle tires of different widths. Tires were mounted on the same wheel, placed on a bicycle wind trainer, and accelerated to a velocity of 25 miles per hour. Then pedaling was stopped, and the time required for the wheel to stop rolling was recorded. The sample means, y, of five trials for each of six different tires were as follows:

Tire Width	Tread	Time to Stop, y (sec)
700/19c	smooth	7.30
700/25c	smooth	8.44
700/32c	smooth	9.27
700/19c	treaded	6.63
700/25c	treaded	6.87
700/32c	treaded	7.07

(a) Carefully make an interaction plot of times required to stop, useful for investigating the sizes of Width and Tread main effects and Width × Tread interactions here. Comment briefly on what the plot shows about these effects. Be sure to label the plot very clearly.

(b) Compute the fitted main effects of Width, the fitted main effects of Tread, and the fitted Width × Tread interactions from the y's. Discuss how they quantify features that are evident in your plot from (a).

23. Below are some data read from a graph in the article "Chemical Explosives" by W. B. Sudweeks that appears as Chapter 30 in *Riegel's Handbook of Industrial Chemistry*. The x values are densities (in g/cc) of pentaerythritol tetranitrate (PETN) samples and the y values are corresponding detonation velocities (in km/sec).

x	y	x	y	x	y
.19	2.65	.50	3.95	.91	5.29
.20	2.71	.50	3.87	.91	5.11
.24	2.79	.50	3.57	.95	5.33
.24	3.19	.55	3.84	.95	5.27
.25	2.83	.75	4.70	.97	5.30
.30	3.52	.77	4.19	1.00	5.52
.30	3.41	.80	4.75	1.00	5.46
.32	3.51	.80	4.38	1.00	5.30
.43	3.38	.85	4.83	1.03	5.59
.45	3.13	.85	5.32	1.04	5.71

(a) Make a scatterplot of these data and comment on the apparent linearity (or the lack thereof) of the relationship between y and x.

(b) Compute the sample correlation between y and x. Interpret this value.

(c) Show the "hand" calculations necessary to fit a line to these data by least squares. Then plot your line on the graph from (a).

(d) About what increase in detonation velocity appears to accompany a unit (1 g/cc) increase in PETN density? What increase in detonation velocity would then accompany a .1 g/cc increase in PETN density?

(e) What fraction of the raw variability in detonation velocity is "accounted for" by the fitted line from part (c)?

(f) Based on your analysis, about what detonation velocity would you predict for a PETN density of 0.65 g/cc? If it was your job to produce a PETN explosive charge with a 5.00 km/sec detonation velocity, what PETN density would you employ?

(g) Compute the residuals from your fitted line. Plot them against x and against $\hat{y}$. Then make a normal plot of the residuals. What do these indicate about the linearity of the relationship between y and x?

(h) Use a statistical package and compute the least squares line, the sample correlation, R^2, and the residuals from the least squares line for these data.

24. Some data collected in a study intended to reduce a thread stripping problem in an assembly process follow. Studs screwed into a metal block were stripping out of the block when a nut holding another part on the block was tightened. It was thought that the depth the stud was screwed into the block (the thread engagement) might affect the torque at which the stud stripped out. In the table below, x is the depth (in 10^{-3} inches above .400) and y is the torque at failure (in lbs/in.).

x	y	x	y	x	y	x	y
80	15	40	70	75	70	20	70
76	15	36	65	25	70	40	65
88	25	30	65	30	60	30	75
35	60	0	45	78	25	74	25
75	35	44	50	60	45		

(a) Use a regression program and fit both a linear equation and a quadratic equation to these data. Plot them on a scatterplot of the data. What are the fractions of raw variability in y accounted for by these two equations?

(b) Redo part (a) after dropping the $x = 0$ and $y = 45$ data point from consideration. Do your conclusions about how best to describe the relationship between x and y change appreciably? What does this say about the extent to which a single data point can affect a curve-fitting analysis?

(c) Use your quadratic equation from part (a) and find a thread engagement that provides an optimal predicted failure torque. What would

you probably want to do before recommending this depth for use in this assembly process?

25. The textbook *Introduction to Contemporary Statistical Methods* by L. H. Koopmans contains a data set from the testing of automobile tires. A tire under study is mounted on a test trailer and pulled at a standard velocity. Using a braking mechanism, a standard amount of drag (measured in %) is applied to the tire and the force (in pounds) with which it grips the road is measured. The following data are from tests on 19 different tires of the same design made under the same set of road conditions. $x = 0\%$ indicates no braking and $x = 100\%$ indicates the brake is locked.

Drag, x (%)	Grip Force, y (lb)
10	550, 460, 610
20	510, 410, 580
30	470, 360, 480
50	390, 310, 400
70	300, 280, 340
100	250, 200, 200, 200

(a) Make a scatterplot of these data and comment on "how linear" the relation between y and x appears to be.

In fact, physical theory can be called upon to predict that instead of being linear, the relationship between y and x is of the form $y \approx \alpha \exp(\beta x)$ for suitable α and β. Note that if natural logarithms are taken of both sides of this expression, $\ln(y) \approx \ln(\alpha) + \beta x$. Calling $\ln(\alpha)$ by the name β_0 and β by the name β_1, one then has a linear relationship of the form used in Section 4.1.

(b) Make a scatterplot of $y' = \ln(y)$ versus x. Does this plot look more linear than the one in (a)?

(c) Compute the sample correlation between y' and x "by hand." Interpret this value.

(d) Fit a line to the drags and logged grip forces using the least squares principle. Show the necessary hand calculations. Sketch this line on your scatterplot from (b).

(e) About what increase in log grip force appears to accompany an increase in drag of 10% of the total possible? This corresponds to what kind of change in raw grip force?

(f) What fraction of the raw variability in log grip force is accounted for in the fitting of a line to the data in part (d)?

(g) Based on your answer to (d), what log grip force would you predict for a tire of this type under these conditions using 40% of the possible drag? What raw grip force?

(h) Compute the residuals from your fitted line. Plot the residuals against x and against $\hat{y}$. Then make a normal plot of the residuals. What do these plots indicate about the linearity of the relationship between drag and log grip force?

(i) Use a statistical package to find the least squares line, the sample correlation, R^2, and the residuals for these (x, y') data.

26. The article "Laboratory Testing of Asphalt Concrete for Porous Pavements" by Woelfl, Wei, Faulstich, and Litwack (*Journal of Testing and Evaluation*, 1981) studied the effect of asphalt content on the permeability of open-graded asphalt concrete. Four specimens were tested for each of six different asphalt contents, with the following results:

Asphalt Content, x (% by weight)	Permeability, y (in./hr water loss)
3	1189, 840, 1020, 980
4	1440, 1227, 1022, 1293
5	1227, 1180, 980, 1210
6	707, 927, 1067, 822
7	835, 900, 733, 585
8	395, 270, 310, 208

(a) Make a scatterplot of these data and comment on how linear the relation between y and x appears to be. If you focus on asphalt contents between, say, 5% and 7%, does linearity seem to be an adequate description of the relationship between y and x?

Temporarily restrict your attention to the $x = 5, 6,$ and 7 data.

(b) Compute the sample correlation between y and x "by hand." Interpret this value.

(c) Fit a line to the asphalt contents and permeabilities using the least squares principle. Show the necessary hand calculations. Sketch this fitted line on your scatterplot from (a).

(d) About what increase in permeability appears to accompany a 1% (by weight) increase in asphalt content?

(e) What fraction of the raw variability in permeability is "accounted for" in the fitting of a line to the $x = 5, 6,$ and 7 data in part (c)?

(f) Based on your answer to (c), what measured permeability would you predict for a specimen of this material with an asphalt content of 5.5%?

(g) Compute the residuals from your fitted line. Plot the residuals against x and against $\hat{y}$. Then make a normal plot of the residuals. What do these plots indicate about the linearity of the relationship between asphalt content and permeability?

(h) Use a statistical package and values for $x = 5, 6,$ and 7 to find the least squares line, the sample correlation, R^2, and the residuals for these data.

Now consider again the entire data set.

(i) Fit the quadratic relationship $y \approx \beta_0 + \beta_1 x + \beta_2 x^2$ to the data using a statistical package. Sketch this fitted parabola on your second scatterplot from part (a). Does this fitted quadratic appear to be an important improvement over the line you fit in (c) in terms of describing the relationship over the range $3 \le x \le 8$?

(j) Fit the linear relation $y \approx \beta_0 + \beta_1 x$ to the entire data set. How do the R^2 values for this fit and the one in (i) compare? Does the larger R^2 in (i) speak strongly for the use of a quadratic (as opposed to a linear) description of the relationship of y to x in this situation?

(k) If one uses the fitted relationship from (i) to predict y for $x = 5.5$, how does the prediction compare to your answer for (f)?

(l) What do the fitted relationships from (c), (i) and (j) give for predicted permeabilities when $x = 2\%$? Compare these to each other as well as your answers to (f) and (k). Why would it be unwise to use any of these predictions without further data collection?

27. Some data collected by Koh, Morden, and Ogbourne in a study of axial breaking strengths (y) for wooden dowel rods follow. The students tested $m = 4$ different dowels for each of nine combinations of three different diameters (x_1) and three different lengths (x_2).

x_1 (in.)	x_2 (in.)	y (lb)
.125	4	51.5, 37.4, 59.3, 58.5
.125	8	5.2, 6.4, 9.0, 6.3
.125	12	2.5, 3.3, 2.6, 1.9
.1875	4	225.3, 233.9, 211.2, 212.8
.1875	8	47.0, 79.2, 88.7, 70.2
.1875	12	18.4, 22.4, 18.9, 16.6
.250	4	358.8, 309.6, 343.5, 357.8
.250	8	127.1, 158.0, 194.0, 133.0
.250	12	68.9, 40.5, 50.3, 65.6

(a) Make a plot of the 3×3 means, $\bar{y}$, corresponding to the different combinations of diameter and length used in the study, plotting $\bar{y}$ vs. x_2 and connecting the three means for a given diameter with line segments. What does this plot suggest about how successful an equation for y that is linear in x_2 for each fixed x_1 might be in explaining these data?

(b) Replace the strength values with their natural logarithms, $y' = \ln(y)$, and redo the plotting of part (a). Does this second plot suggest that the logarithm of strength might be a linear function of length for fixed diameter?

(c) Fit the following three equations to the data via least squares:

$$y' \approx \beta_0 + \beta_1 x_1,$$
$$y' \approx \beta_0 + \beta_2 x_2,$$
$$y' \approx \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

What are the coefficients of determination for the three fitted equations? Compare the equations in terms of their complexity and their apparent ability to predict y'.

(d) Add three lines to your plot from part (b), showing predicted log strength (from your third fitted equation) as a function of x_2 for the three different values of x_1 included in the study. Use your third fitted equation to predict first a log strength and then a strength for a dowel of diameter .20 in. and length 10 in. Why shouldn't you be willing to use your equation to predict the strength of a rod with diameter .50 in. and length 24 in.?

(e) Compute and plot residuals for the third equation you fit in part (c). Make plots of residuals vs. fitted response and both x_1 and x_2, and normal-plot the residuals. Do these plots suggest any potential inadequacies of the third fitted equation? How might these be remedied?

(f) The students who did this study were strongly suspicious that the ratio $x_3 = x_1^2/x_2$ is the principal determiner of dowel strength. In fact, it is possible to empirically discover the importance of this quantity as follows. Try fitting the equation

$$y' \approx \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2$$

to these data and notice that the fitted coefficients of $\ln x_1$ and $\ln x_2$ are roughly in the ratio of 4 to -2, i.e., 2 to -1. (What does this fitted equation for $\ln(y)$ say about y?) Then plot y vs. x_3 and fit the linear equation

$$y \approx \beta_0 + \beta_3 x_3$$

to these data. Finally, add three curves to your plot from part (a) based on this fitted equation linear in x_3, showing predicted strength as a function of x_2. Make one for each of the three different values of x_1 included in the study.

(g) Since the students' data have a (replicated) 3×3 factorial structure, you can do a factorial analysis as an alternative to the preceding analysis. Looking again at your plot from (a), does it seem that the interactions of Diameter and Length will be important in describing the raw strengths, y? Compute the fitted factorial effects and comment on the relative sizes of the main effects and interactions.

(h) Redo part (g), referring to the graph from part (b) and working with the logarithms of dowel strength.

28. The paper "Design of a Metal-Cutting Drilling Experiment—A Discrete Two-Variable Problem" by E. Mielnik (*Quality Engineering*, 1993–1994) reports a drilling study run on an aluminum alloy (7075-T6). The thrust (or axial force), y_1, and torque, y_2, required to rotate drills of various diameters x_1 at various feeds (rates of drill penetration into the workpiece) x_2, were measured with the following results:

Diameter, x_1 (in.)	Feed Rate, x_2 (in. rev)	Thrust, y_1 (lb)	Torque, y_2 (ft-lb)
.250	.006	230	1.0
.406	.006	375	2.1
.406	.013	570	3.8
.250	.013	375	2.1
.225	.009	280	1.0
.318	.005	225	1.1
.450	.009	580	3.8
.318	.017	565	3.4
.318	.009	400	2.2
.318	.009	400	2.1
.318	.009	380	2.1
.318	.009	380	1.9

Drilling theory suggests that $y_1 \approx \kappa_1 x_1^a x_2^b$ and $y_2 \approx \kappa_2 x_1^c x_2^d$ for appropriate constants $\kappa_1, \kappa_2, a, b, c$, and d. (Note that upon taking natural logarithms, there are linear relationships between $y_1' = \ln(y_1)$ or $y_2' = \ln(y_2)$ and $x_1' = \ln(x_1)$ and $x_2' = \ln(x_2)$.)

(a) Use a regression program to fit the following equations to these data:

$$y_1' \approx \beta_0 + \beta_1 x_1',$$
$$y_1' \approx \beta_0 + \beta_2 x_2',$$
$$y_1' \approx \beta_0 + \beta_1 x_1' + \beta_2 x_2'$$

What are the R^2 values for the three different fitted equations? Compare the three fitted equations in terms of complexity and apparent ability to predict y_1'.

(b) Compute and plot residuals (continuing to work on log scales) for the third equation you fit in part (a). Make plots of residuals vs. fitted y_1' and both x_1' and x_2', and normal-plot these residuals. Do these plots reveal any particular problems with the fitted equation?

(c) Use your third equation from (a) to predict first a log thrust and then a thrust if a drill of diameter .360 in. and a feed of .011 in./rev are used. Why would it be unwise to make a similar prediction for $x_1 = .450$ and $x_2 = .017$? (*Hint:* Make a plot of the (x_1, x_2) pairs in the data set and locate this second set of conditions on that plot.)

(d) If the third equation fit in part (a) governed y_1, would it lead to Diameter × Feed interactions for y_1 measured on the log scale? To help you answer this question, plot $\widehat{y_1'}$ vs. x_2 (or x_2') for each of $x_1 = .250, .318$, and $.406$. Does this equation lead to Diameter × Feed interactions for raw y_1?

(e) The first four data points listed in the table constitute a very small complete factorial study (an unreplicated 2×2 factorial in the factors Diameter and Feed). Considering only these data points, do a "factorial" analysis of this part of the y_1 data. Begin by making an interaction plot similar to Figure 4.22 for these data. Based on that plot, discuss the apparent relative sizes of the Diameter and Feed main effects on thrust. Then carry out the arithmetic necessary to compute the fitted factorial effects (the main effects and interactions).

(f) Redo part (e), using y_1' as the response variable.

(g) Do your answers to parts (e) and (f) complement those of part (d)? Explain.

29. The article "A Simple Method to Study Dispersion Effects From Non-Necessarily Replicated Data in Industrial Contexts" by Ferrer and Romero (*Quality Engineering*, 1995) describes an unreplicated 2^4 experiment done to improve the adhesive force obtained when gluing on polyurethane sheets as the inner lining of some hollow metal parts. The factors studied were the amount of glue used (A), the predrying temperature (B), the tunnel temperature (C), and the pressure applied (D). The exact levels of the variables employed were not given in the article (presumably for reasons of corporate security). The response variable was the adhesive force, y, in Newtons, and the data reported in the article follow:

Combination	y	Combination	y
(1)	3.80	d	3.29
a	4.34	ad	2.82
b	3.54	bd	4.59
ab	4.59	abd	4.68
c	3.95	cd	2.73
ac	4.83	acd	4.31
bc	4.86	bcd	5.16
abc	5.28	abcd	6.06

(a) Compute the fitted factorial effects corresponding to the "all high" treatment combination.

(b) Interpret the results of your calculations in the context of the study. Which factors and/or combinations of factors appear to have the largest effects on the adhesive force? Suppose that only the A, B, and C main effects and the B × D interactions were judged to be of importance here. Make a corresponding statement to your engineering manager about how the factors impact adhesive force.

(c) Using the reverse Yates algorithm or otherwise, compute fitted/predicted values corresponding to an "A, B, and C main effects and BD interactions" description of adhesive force. Then use these 16 values to compute residuals, $e = y - \hat{y}$. Plot these against $\hat{y}$, and against levels of A, B, C, and D. Also normal-plot them. Comment on any interpretable patterns you see. Particularly in reference to the plot of residuals vs. level of D, what does this graph suggest if one is interested not only in high mean adhesive force but in consistent adhesive force as well?

(d) Find and interpret a value of R^2 corresponding to the description of y used in part (c).

30. The article "Chemical Vapor Deposition of Tungsten Step Coverage and Thickness Uniformity Experiments" by J. Chang (*Thin Solid Films*, 1992) describes an unreplicated 2^4 factorial experiment aimed at understanding the effects of the factors

Factor A Chamber Pressure (Torr)
8 (−) vs. 9 (+)

Factor B H_2 Flow (standard cm^3/min)
500 (−) vs. 1000 (+)

Factor C SiH_4 Flow (standard cm^3/min)
15 (−) vs. 25 (+)

Factor D WF_6 Flow (standard cm^3/min)
50 (−) vs. 60 (+)

on a number of response variables in the chemical vapor deposition tungsten films. One response variable reported was the average sheet resistivity, y (mΩ/cm) of the resultant film, and the values reported in the paper follow.

Combination	y	Combination	y
(1)	646	d	666
a	623	ad	597
b	714	bd	718
ab	643	abd	661
c	360	cd	304
ac	359	acd	309
bc	325	bcd	360
abc	318	abcd	318

(a) Use the Yates algorithm and compute the fitted factorial effects corresponding to the "all high" treatment combination. (You will need to employ four cycles in the calculations.)

(b) Interpret the results of your calculations from (a) in the context of the study. (Describe in words which factors and/or combinations of factors appear to have the largest effect(s) on average sheet resistivity. What are the practical implications of these effects?)

(c) Suppose that you judge all factors except C to be "inert" as far as determining sheet resistivity is concerned (the main effects of A, B, and D and all interactions involving them are negligible). What fitted/predicted values correspond to this "C main effects only" description of average sheet resistivity? Use these 16 values to compute residuals, $e = y - \hat{y}$. Plot these against $\hat{y}$, level of A, level of B, level of C, and level of D. Also normal-plot these residuals. Comment on any interpretable patterns in your plots.

(d) Compute an R^2 value corresponding to the description of average sheet resistivity used in part (c). Does it seem that the grand mean and C main effects provide an effective summary of average sheet resistivity? Why?