10

Regression and model building

Statisticians, like artists, have the bad habit of falling in love with their models. (Attributed to George Box)

Overview

Regression has already been encountered in earlier chapters. In this chapter regression modelling is examined in more detail. In terms of the process model shown earlier in Figure 1.3, regression methods enable the models to be built in terms of linking process inputs (Xs) to process performance measures (Ys) via functional relationships of the form $\mathbf{Y} = f(\mathbf{X})$. The links between regression models and design of experiments will be established. Scenarios in which the response variable is categorical will be dealt with under the heading of logistic regression. The Minitab facilities for the creation, analysis and checking of regression models will be exemplified.

10.1 Regression with a single predictor variable

In Section 3.2.1 reference was made to data on the diameter (Y, mm) of machined automotive components and the temperature (X, °C) of the coolant supplied to the machine at the time of production. Given that the target diameter is 100 mm, a scatterplot indicated the possibility of improving the process through controlling the coolant temperature, thereby leading to less variability in the diameter of the components. Use of **Graph > Scatterplot...** and the **With Regression** option yielded the scatterplot in Figure 10.1, with the addition of the least squares regression line modelling the linear relationship between diameter and temperature. (Use of **Data View...** and the **Regression** tab indicates that the default is to fit a linear model as displayed, but quadratic and cubic models may also be fitted.) The data are available in Diameters.MTW.

Six Sigma Quality Improvement with Minitab, Second Edition. G. Robin Henderson.

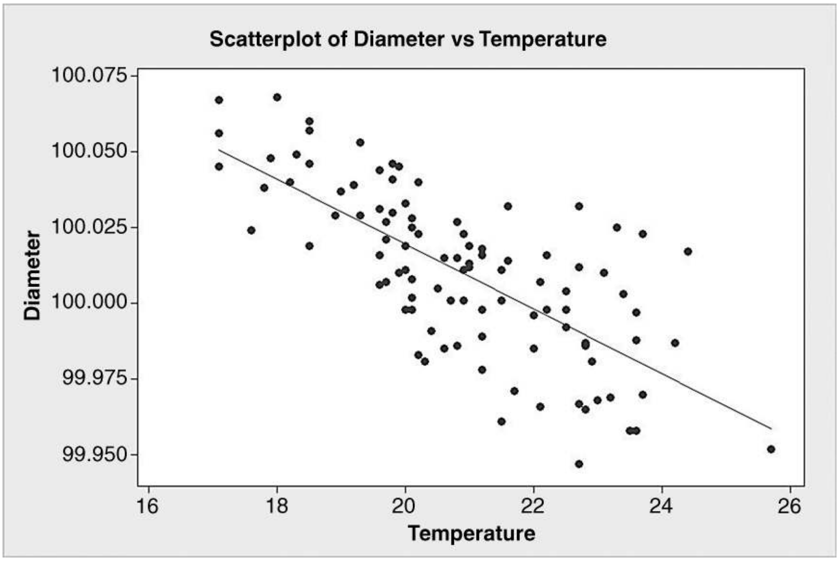


Figure 10.1 Scatterplot of diameter against temperature of coolant.

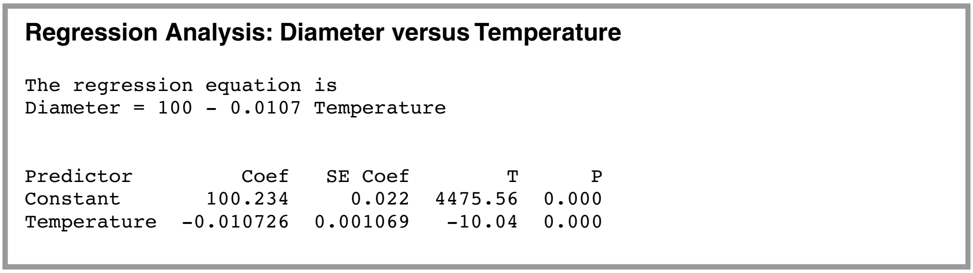
The equation of the least squares regression line may be found using **Stat** > **Regression** > **Regression. . . .** In the dialog **Response:** Diameter and **Predictors:** Temperature must be specified. Thus for a functional relationship $Y = f(X)$, Y is a response and X is a predictor in the terminology used by Minitab. Given that the data are recorded in time order, it is appropriate to check **Four in one** under **Graphs. . . .** Defaults were chosen elsewhere. Various aspects of what is known as *simple linear regression* will now be discussed with reference to the output.

The first portion of the Session Window output is shown Panel 10.1. The equation of the least squares regression line fitted to the data is

$$\text{Diameter} = 100.234 - 0.010\,726 \times \text{Temperature}$$

or, alternatively,

$$y = 100.234 - 0.010\,726x.$$



Regression Analysis: Diameter versus Temperature

```
The regression equation is
Diameter = 100 - 0.0107 Temperature


Predictor        Coef    SE Coef        T      P
Constant      100.234      0.022  4475.56  0.000
Temperature  -0.010726  0.001069   -10.04  0.000
```

Panel 10.1 Regression analysis for diameter versus temperature.

In dealing with regression situations statisticians often make use of the *linear model*

$$Y_i = \alpha + \beta x_i + \varepsilon_i,$$

where α is the intercept parameter, β is the slope parameter and ε_i is the random error, a random variable with mean 0 and standard deviation σ. It follows that

$$\begin{aligned} E(Y_i) &= E(\alpha) + E(\beta x_i) + E(\varepsilon_i) = \alpha + \beta x_i, \\ \mathrm{var}(Y_i) &= \mathrm{var}(\varepsilon_i) = \sigma^2, \end{aligned}$$

where σ^2 is a constant. This means that if Y is observed repeatedly for a particular value of x, then the resulting population of Y-values will have a statistical distribution with mean $\alpha + \beta x$ and variance σ^2.

The further assumptions that the random errors are *independent* and *normally distributed* are also frequently made. One consequence of these assumptions is that the population of Y-values that could be observed for a given x has the normal distribution with mean $\alpha + \beta x$ and variance σ^2.

The fitted least squares regression line may be written as

$$y = a + bx$$

where $a = 100.234$ and $b = -0.010\,726$ respectively estimate the model parameters α and β.

The P-values shown in Panel 10.1 are for two t-tests of hypotheses for the model in which $Y_i \sim N(\alpha + \beta x_i, \sigma^2)$. The first test concerns the intercept parameter, α, in the model:

$$H_0 : \alpha = 0, \quad H_1 : \alpha \neq 0.$$

With P-value 0.000, to three decimal places, we have very strong evidence that the intercept parameter is nonzero.

The second test concerns the slope parameter, β, in the model:

$$H_0 : \beta = 0, \quad H_1 : \beta \neq 0.$$

With P-value 0.000, to three decimal places, we have very strong evidence that the slope parameter is nonzero. This second test may be considered as a test of whether or not there is a linear relationship between Y and x, i.e. between diameter and temperature. (The t-test performed here is equivalent to the test the that the correlation coefficient is zero.)

Whenever a least squares line has been fitted to a set of bivariate data, residuals may be calculated. For any statistical model fitted to data,

$$\text{Data} = \text{Fit} + \text{Residual},$$

as the reader will recall from Chapter 7. The residual may be thought of as that 'component' of the data that remains when the model has performed its task of 'explaining' the data:

$$\text{Residual} = \text{Data} - \text{Fit} = y_i - \hat{Y}_i = y_i - (a + bx_i).$$

The symbol $\hat{Y}_i$ is used for the fitted value of y. For example, for the first data point we have $x_1 = 20.9$ so the fitted value is $\hat{Y}_1 = 100.234 - 0.010\,726 \times 20.9 = 100.234 - 0.224 = 100.010$, to three decimal places. Hence, the first residual is calculated to be

$y_1 - \hat{Y}_1 = 100.001 - 100.010 = -0.009$. The reader is invited to check that the residual for the last data point is 0.026.

Checking **Residuals** and **Fits** under **Storage** during the **Regression** dialog enables columns of the residuals and fitted values to be created. The software gives the first residual as $-0.008\,975\,5$ which rounds to -0.009. Having fitted a linear model to data using least squares, it is standard practice to plot the residuals in various ways. The **Four in one** facility in Minitab yields the following plots:

- histogram of residuals;
- normal probability plot of residuals;
- residuals versus fits;
- residuals versus order.

The four plots for the regression of diameter on temperature are displayed in Figure 10.2.

The shape of the histogram and the linear normal probability plot support the assumption of normality in the model. The fact that the plot of residuals against fitted values has the appearance of a horizontal band of randomly distributed points supports the assumption of constant variability in the model. Finally, the absence of any patterns or trends in the plot of residuals against order suggests that there have been no time-related factors influencing the process. Thus we may consider the residual plots to be satisfactory in this case.

The next portion of the Session window output is shown in Panel 10.2. The value of s provides an estimate of the standard deviation, σ, in the model. The value of R^2 (R-Sq), in this case where there is a single predictor variable, is r^2, the coefficient of determination between diameter and temperature expressed as a percentage – see Section 3.2.2. It indicates the proportion of the variation in diameter that can be attributed to its linear dependence on

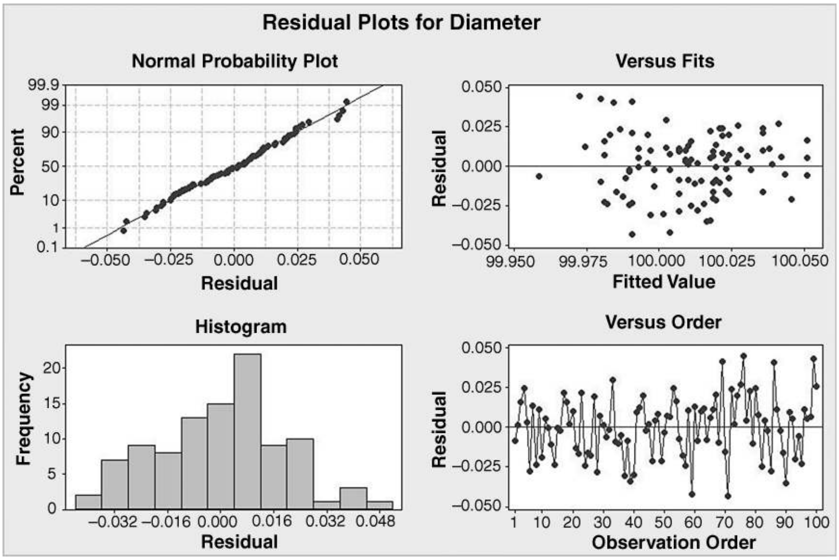


Figure 10.2 Residual plots for regression of diameter on temperature.

```
S = 0.0192354   R-Sq = 50.7%   R-Sq(adj) = 50.2%
```

Panel 10.2 Further output from regression analysis.

temperature. R^2 may also be calculated as the square of the correlation between the observed and fitted values of diameter, the response.

We may summarize the model as follows. The expected value of diameter is linearly related to temperature by the equation

$$\text{Diameter} = 100.234 - 0.010\,726 \times \text{Temperature}$$

For any specific temperature, diameter is normally distributed, with mean estimated by

$$100.234 - 0.010\,726 \times \text{Temperature}$$

and with standard deviation estimated to be 0.0192. Just over half the variation in diameter may be explained through its linear dependence on temperature.

An enhanced version of the plot in Figure 10.1 may be obtained using **Stat > Regression > Fitted Line Plot...**, specifying Diameter as the response and Temperature as the predictor and accepting defaults otherwise. The plot is shown in Figure 10.3. This plot is annotated by Minitab with the equation of the regression line, the estimated standard deviation of the random errors and the value of R^2. (The adjusted value of R^2, denoted by R-Sq(adj) will be discussed later in the chapter.) This plot would clearly complement the written summary of the model in, for example, a project report. (The first and last data points have been indicated by square symbols and vertical line segments drawn from the regression line to these data points. The magnitudes of these segments are the magnitudes of the residuals. The first data point lies below the regression line corresponding to its associated negative residual; the last data point lies above the line

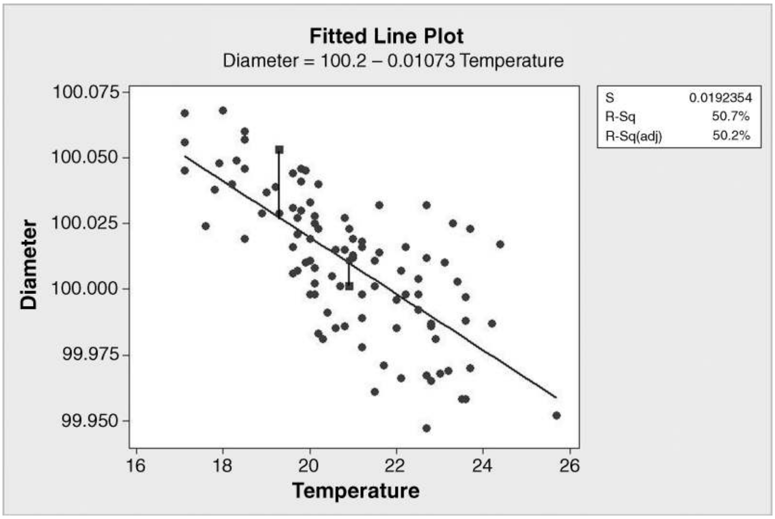


Figure 10.3 Fitted line plot for diameter on temperature.

```
Analysis of Variance

Source      DF         SS         MS       F      P
Regression   1  0.0372752  0.0372752  100.74  0.000
Error       98  0.0362602  0.0003700
Total       99  0.0735354
```

Panel 10.3 ANOVA for regression analysis.

corresponding to its associated positive residual. The fitted least squares regression line is such that the sum of the squares of all 100 residuals is a minimum.)

The Session window output also includes the ANOVA output displayed in Panel 10.3. The ANOVA provides an alternative method of testing the null hypothesis that the slope parameter, β, is zero. In fact the F-statistic of 100.74 quoted is the square, allowing for rounding, of the t-statistic of -10.04 quoted in Panel 10.1. The P-value of 0.000, to three decimal places, provides strong evidence that null hypothesis $H_0: \beta = 0$ should be rejected in favour of the alternative hypothesis $H_1: \beta \neq 0$.

The calculation of residuals has already been detailed. Minitab also computes standardized residuals by dividing the residuals by their standard deviation. The final part of the Session window output alerts the user to any observations in the data set that yield a standardized residual with an absolute value in excess of 2. Details of such observations are listed under the heading Unusual Observations and identified by a letter R at the end of the line – see Panel 10.4. With a standard normal distribution the probability of obtaining values exceeding 2, in absolute value, is approximately 5%. Thus being alerted to six observations in this category, when five would be expected from a sample of 100 observations, need not be a major concern.

The user is also alerted to any observations in the data set that have x-values that give them large influence. An X at the end of the line of output indicates such observations. Observation number 31 is adjudged to have a temperature value which gives it large influence in the analysis. This observation corresponds to the point in the bottom right-hand corner of the scatterplots in Figures 10.1 and 10.3. One would wish to check that the data for this point had been correctly entered or that there were no unusual circumstances affecting the process when the observation was made. One might also investigate the regression analysis with all points flagged as having large influence deleted from the data set.

```
Unusual Observations

Obs  Temperature  Diameter      Fit  SE Fit  Residual  St Resid
 31         25.7    99.952   99.958   0.005    -0.006     -0.35 X
 59         21.5    99.961  100.004   0.002    -0.043     -2.22R
 69         22.7   100.032   99.991   0.003     0.041      2.17R
 71         22.7    99.947   99.991   0.003    -0.044     -2.29R
 76         24.4   100.017   99.972   0.004     0.045      2.37R
 86         23.3   100.025   99.984   0.003     0.041      2.15R
 99         23.7   100.023   99.980   0.004     0.043      2.28R

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large leverage.
```

Panel 10.4 List of unusual observations from regression analysis.

The target diameter was 100 mm. Substitution of this value for diameter into the fitted linear equation yields the following:

$$\begin{aligned}\text{Diameter} &= 100.234 - 0.010\,726 \times \text{Temperature}\\ 100 &= 100.234 - 0.010\,726 \times \text{Temperature}\\ 0.010\,726 \times \text{Temperature} &= 0.234\\ \text{Temperature} &= \frac{0.234}{0.010\,726} = 21.8.\end{aligned}$$

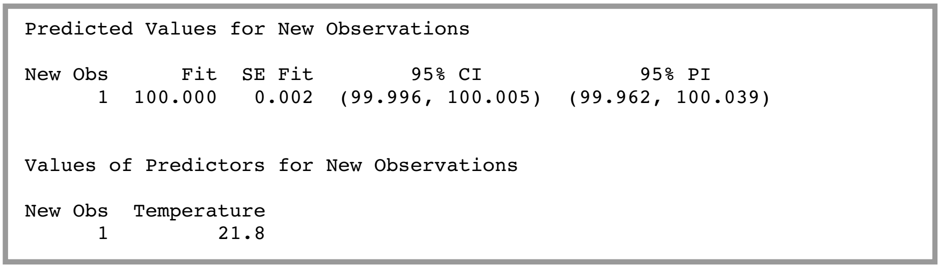
Thus the model suggests that the temperature of the coolant should be controlled at 21.8 °C.

The model may be used to make predictions of diameter for any specific temperature of interest. Predictions may be obtained by using **Stat > Regression > Regression...** again, with **Response:** Diameter and **Predictors:** Temperature, specified as previously. In order to predict for temperature 21.8, for example, the value 21.8 must be entered under **Options...** by specifying **Prediction intervals for new observations:** 21.8. (More than one temperature value may be entered if required.) The resulting output in the Session window is shown in Panel 10.5.

The order of presentation is arguably not the best. The second section informs the user that diameter has been predicted for temperature 21.8 °C. The first section indicates that the predicted diameter is 100.000 (under the heading Fit), thus confirming that the calculation performed earlier was correct! Two intervals are given. The 95% CI of (99.996,100.005) is a 95% confidence interval for the *mean* diameter obtained when the process is operated with temperature 21.8 °C. (The SE Fit value of 0.002 is the estimated standard deviation required in the computation of the confidence interval.) The 95% PI of (99.962,100.039) is a prediction interval in which we can have 95% confidence that an *individual* diameter, obtained when the process is operated with temperature 21.8 °C, will fall. Note that the prediction interval is wider than the confidence interval.

Instead of listing one or more temperature values under **Options...** in **Prediction intervals for new observations:**, one can create a column of temperature values of interest and insert the column name instead of a list. This was done in order to create Table 10.1. Temperatures from 17 °C to 26 °C, at intervals of 1 °C, were selected in order to cover the range of temperatures encountered in the investigation.

Note that the width of the intervals varies, with the narrowest intervals occurring for temperature 21 °C. In fact the narrowest possible intervals occur when temperature equals the mean of all 100 temperatures recorded in the given data set, i.e. 20.9 °C. In order to display



```
Predicted Values for New Observations

New Obs      Fit  SE Fit        95% CI             95% PI
      1  100.000   0.002  (99.996, 100.005)  (99.962, 100.039)


Values of Predictors for New Observations

New Obs  Temperature
      1         21.8
```

Panel 10.5 Predictions of diameter from model.

Table 10.1 Confidence and prediction intervals for diameter.

Temperature	Predicted diameter	95% confidence limits for diameter		95% prediction limits for diameter	
		Lower	Upper	Lower	Upper
17	100.052	100.043	100.061	100.013	100.091
18	100.041	100.034	100.048	100.002	100.080
19	100.030	100.025	100.036	99.992	100.069
20	100.020	100.015	100.024	99.981	100.058
21	100.009	100.005	100.013	99.971	100.047
22	99.998	99.994	100.003	99.960	100.037
23	99.987	99.982	99.993	99.949	100.026
24	99.977	99.969	99.984	99.938	100.016
25	99.966	99.956	99.976	99.927	100.005
26	99.955	99.944	99.967	99.915	99.995

these intervals use was made of **Stat > Regression > Fitted Line Plot...**, specifying diameter as the response and temperature as the predictor, and under **Options...** checking both **Display confidence interval** and **Display prediction interval** as **Display Options**. The resultant plot is shown in Figure 10.4. The arrowed vertical line segment that has been added to the plot indicates the 95% prediction interval corresponding to temperature 21.8 °C – the temperature predicted by the model to yield the desired mean diameter of 100 mm. From a process improvement point of view, the modelling has demonstrated that there is the potential to reduce the variability of diameter and hence to increase process capability through control of the coolant temperature.

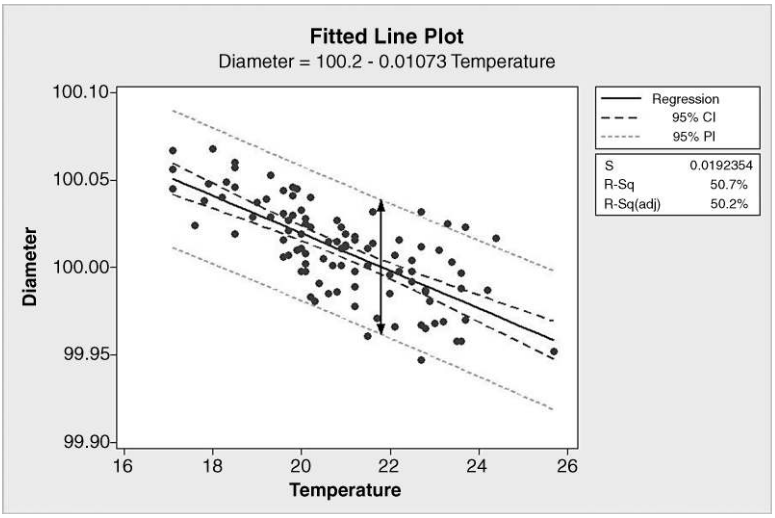


Figure 10.4 Fitted line plot with confidence and prediction intervals.

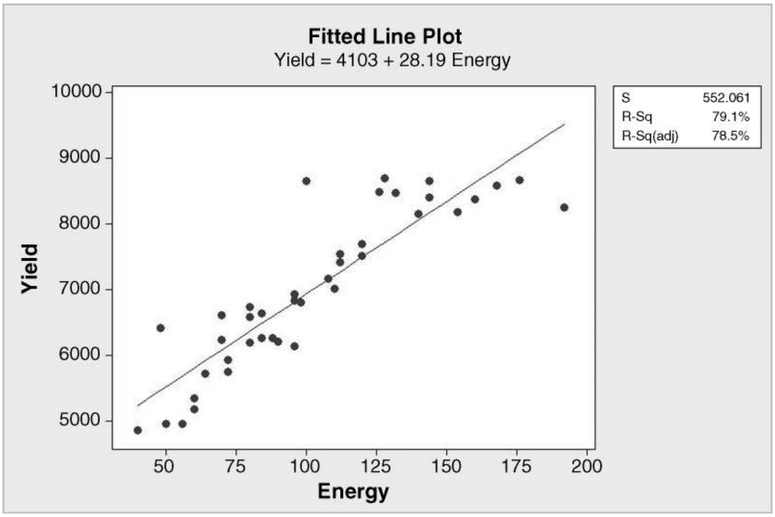


Figure 10.5 Fitted line plot for lactation data.

As a second example, consider the data available in Lactation.MTW. It gives milk yield (kg) and energy intake (MJ/d) for a sample of 40 cows. The fitted line plot for the linear regression of yield on energy is displayed in Figure 10.5. The R^2 value indicates that the linear relationship between yield on energy explains around 80% of the variation in yield. (The reader is invited to verify that the P-value for the slope parameter is 0.000, to three decimal places. Thus there is strong evidence of a linear relationship.)

The plot of residuals against fitted values is shown in Figure 10.6. The fact that this plot has an arched appearance (indicated by the broad arc superimposed on the plot) rather than the

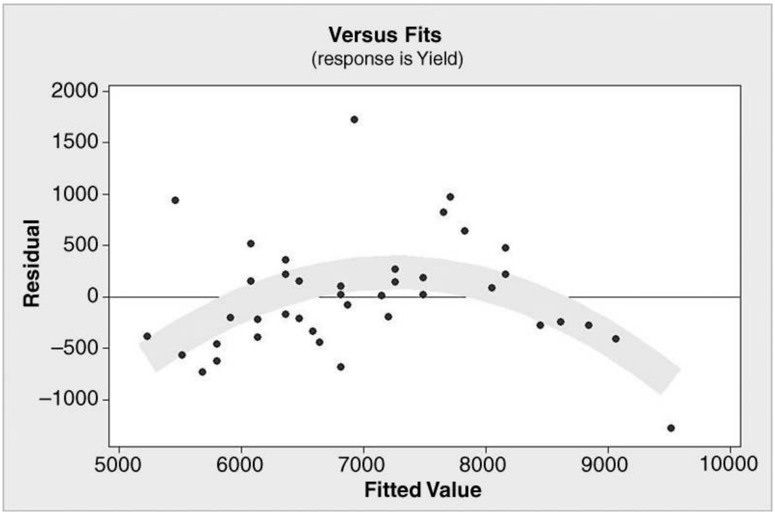


Figure 10.6 Residuals versus fitted values for linear model.

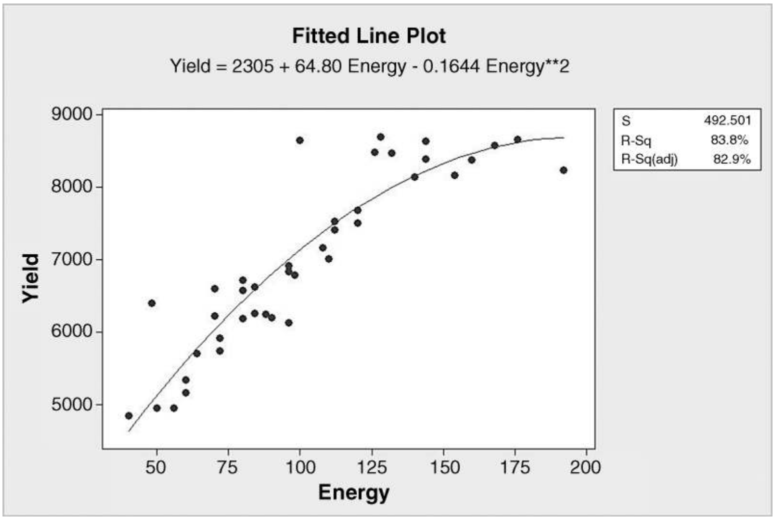


Figure 10.7 Quadratic model fitted to lactation data.

appearance of a horizontal band of randomly distributed points suggests that the linear model is inadequate. The plot indicates that there is still structure of a curvilinear nature in the data that might be utilized in order to improve the model. The simplest way to introduce curvature into the model is to make use of the *quadratic model*

$$Y_i = \alpha + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i,$$

where α is the intercept parameter, β_1 is the slope parameter, β_{11} is the quadratic parameter and ε_i is the random error, with mean 0 and standard deviation σ. This model may be fitted using **Stat > Regression > Fitted Line Plot...** and checking **Quadratic** as the **Type of Regression Model**. Using **Graphs...** the option to plot **Residuals versus fits** was checked under **Residual Plots**.

The fitted line plot is shown in Figure 10.7. The quadratic relationship between yield and energy explains around 84% of the variation in yield – the linear model explained around 80% of the variation, so the addition of the quadratic term has yielded a modest increase in explanatory power.

The plot of residuals against fitted values for the quadratic model is shown in Figure 10.8 and is a more satisfactory residual plot than the one for the previous model. The Session window output is displayed in Panel 10.6. The heading Polynomial Regression Analysis indicates that the software has fitted a polynomial curve, in this case a quadratic curve or parabola, to the data.

The equation of the quadratic curve is given together with the estimate, s, of the standard deviation of the random errors. The R^2 value is given together with an analysis of variance for the overall quadratic regression model. It provides a test of the hypotheses

$$H_0 : \beta_1 = \beta_{11} = 0, \quad H_1 : \text{Not all } \beta\text{s are zero.}$$

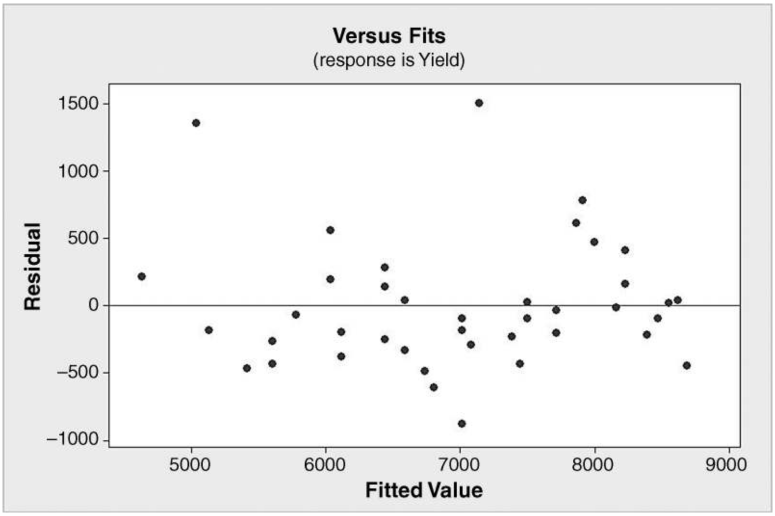
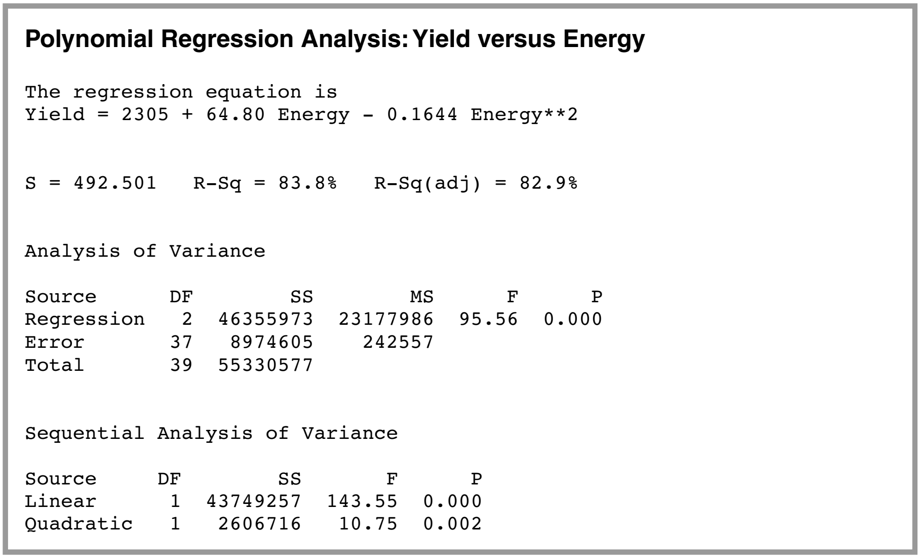


Figure 10.8 Residuals versus fitted values for quadratic model.

With P-value 0.000, to three decimal places, the null hypothesis cannot be rejected. The sequential analysis of variance, with P-values less than 0.01 for both the linear and quadratic terms, provides evidence that both terms are worth including in the model.

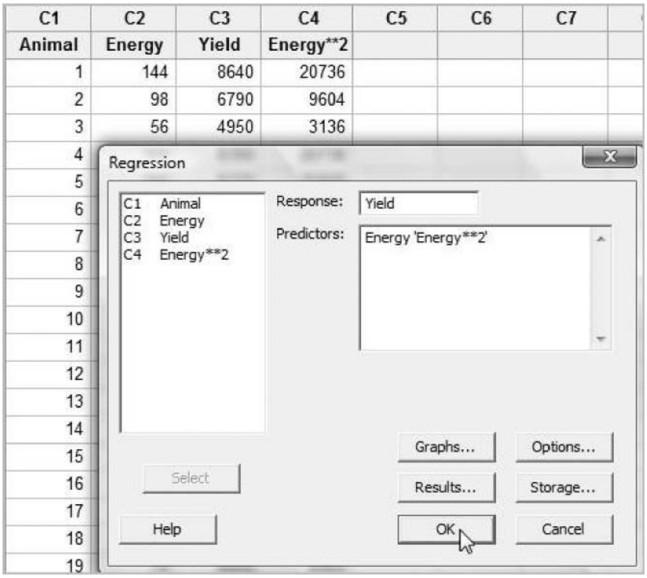
In order to carry out an alternative analysis it is necessary to use **Calc** > **Calculator** to calculate a new column, named Energy**2, say, containing the squares of the energy values. The quadratic regression may be found using **Stat** > **Regression** > **Regression...**. In the dialog **Response:** Yield and **Predictors:** Energy and 'Energy**2' must be specified as indicated in Figure 10.9. Residual plots may be created using **Graphs...** as before.



Polynomial Regression Analysis: Yield versus Energy

```
The regression equation is
Yield = 2305 + 64.80 Energy - 0.1644 Energy**2


S = 492.501   R-Sq = 83.8%   R-Sq(adj) = 82.9%


Analysis of Variance

Source       DF        SS        MS      F      P
Regression    2  46355973  23177986  95.56  0.000
Error        37   8974605    242557
Total        39  55330577


Sequential Analysis of Variance

Source      DF        SS       F      P
Linear       1  43749257  143.55  0.000
Quadratic    1   2606716   10.75  0.002
```

Panel 10.6 Session window output for the quadratic regression model.



C1	C2	C3	C4	C5	C6	C7
Animal	Energy	Yield	Energy**2			
1	144	8640	20736			
2	98	6790	9604			
3	56	4950	3136			

Regression

C1 Animal
C2 Energy
C3 Yield
C4 Energy**2

Response: Yield

Predictors: Energy 'Energy**2'

Figure 10.9 Dialog for fitting quadratic regression model.

Part of the Session window output is shown in Panel 10.7. Here the *P*-values provide evidence that the constant term in the model is nonzero and that the coefficients of the energy term and the energy squared term are both nonzero. The latter two *P*-values are identical to those given in the sequential analysis of variance in Panel 10.6.

From the point of view of quality improvement the creation of a quadratic model can assist with the determination of optimum conditions under which to run a process. The fitted quadratic curve, extrapolated for energy values up to 250, is shown in Figure 10.10. The model suggests to animal scientists that yield of milk could potentially be maximized by targeting energy intake levels of around 200 for the cows. Further investigation would be needed to confirm that yield could be expected to decrease for energy levels beyond 200.

The data for the final example in this section are reproduced from Gorman and Toman (1966) by permission of the American Society for Quality and are discussed by Hogg and

Regression Analysis: Yield versus Energy, Energy2**

```
The regression equation is
Yield = 2305 + 64.8 Energy - 0.164 Energy**2


Predictor       Coef  SE Coef      T      P
Constant      2305.3    593.7   3.88  0.000
Energy         64.80    11.36   5.70  0.000
Energy**2   -0.16438  0.05014  -3.28  0.002


S = 492.501   R-Sq = 83.8%   R-Sq(adj) = 82.9%
```

Panel 10.7 Session window output for the quadratic regression model.

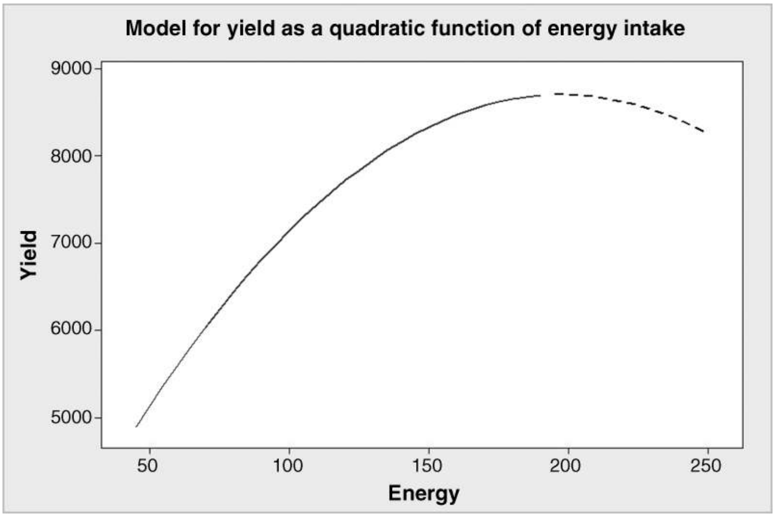


Figure 10.10 Dialog for quadratic regression model.

Ledolter (1992, pp 393–398). They are available in Rut.MTW and give change in rut depth (y) and viscosity of the asphalt (x) for experimental sections of pavement. Scrutiny of the scatterplot of the data in Figure 10.11 suggests that the relationship is nonlinear.

Rather than fit a polynomial model, such as a quadratic or cubic, Hogg and Ledolter applied a logarithmic transformation to both x and y. This transformation may be carried out directly using **Calc** > **Calculator** to calculate new columns containing the logarithms to base 10 of both x and y. The linear regression of $\log_{10}y$ on $\log_{10}x$ may then be obtained using **Stat** > **Regression** > **Regression. . . .** Alternatively, use may be made of **Stat** > **Regression** > **Fitted**

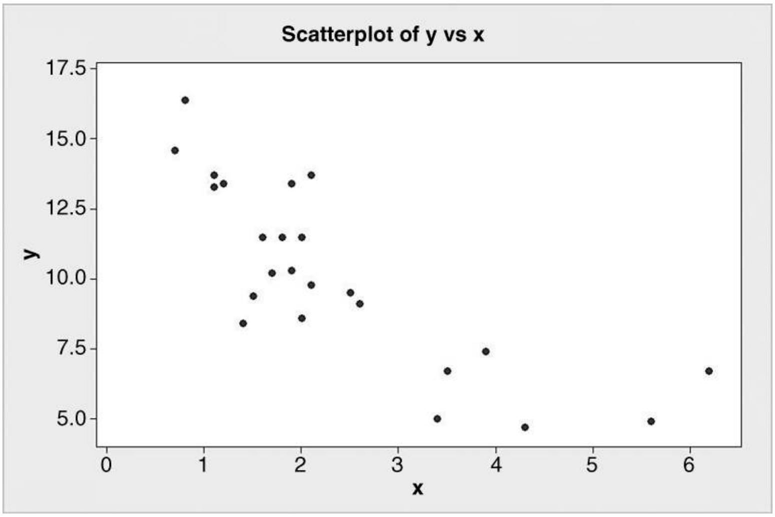


Figure 10.11 Plot of change in rut depth against asphalt viscosity.

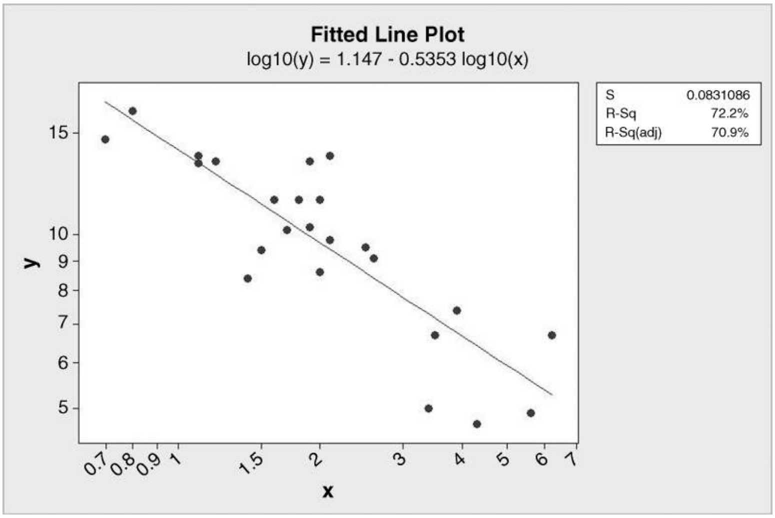


Figure 10.12 Regression of $\log_{10}y$ on $\log_{10}x$.

Line Plot..., specifying y as the response and x as the predictor. Under **Options...**, **Transformations** it is necessary to check **Logten of Y** and **Display logscale for Y variable** together with **Logten of X** and **Display logscale for X variable**. The output is shown in Figure 10.12. This model has an R^2 of 72%, whereas a simple linear regression model fitted to the data yields an R^2 of 64%. The residual plots are satisfactory, so the logarithmic transformations have yielded an adequate model of the situation.

So far we have considered situations where we have created models of the form $Y=f(X)$ for a single response, Y, in terms of a single factor or predictor, X. In the next section we consider models of the form $Y=f(\mathbf{X})$ or $Y=f(X_1,X_2,\ldots)$ for a single response, Y, in terms of a series of factors or predictors, X_1, X_2, ... under the heading of multiple regression.

10.2 Multiple regression

The simplest multiple regression model is the *linear model*

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i,$$

where β_0 is the intercept parameter, β_1 and β_2 are the (partial) regression coefficients, and ε_i is the random error, with mean 0 and standard deviation σ. It follows that

$$\begin{aligned} E(Y_i) &= E(\beta_0) + E(\beta_1 x_{1i}) + E(\beta_2 x_{2i}) + E(\varepsilon_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \\ \mathrm{var}(Y_i) &= \mathrm{var}(\varepsilon_i) = \sigma^2, \end{aligned}$$

where σ^2 is a constant. This means that if Y is measured or observed repeatedly for a particular pair of values of x_1 and x_2, then the resulting values will have a statistical distribution with mean $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ and variance σ^2.

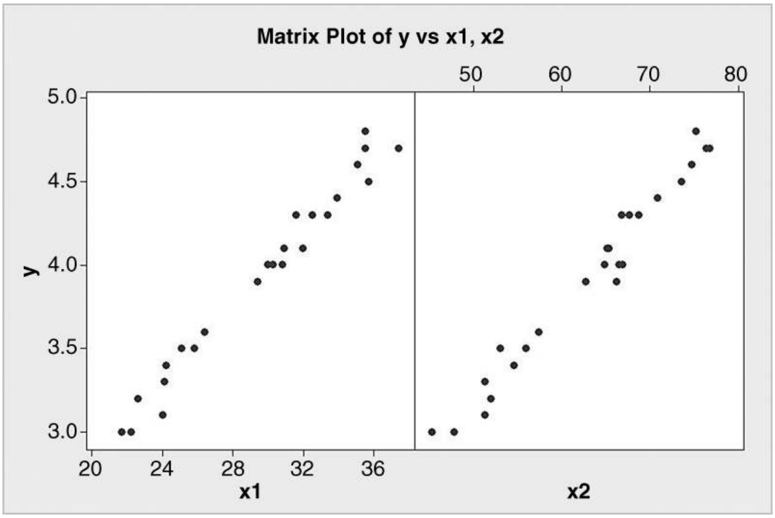


Figure 10.13 Scatterplots of Y versus x_1 and x_2.

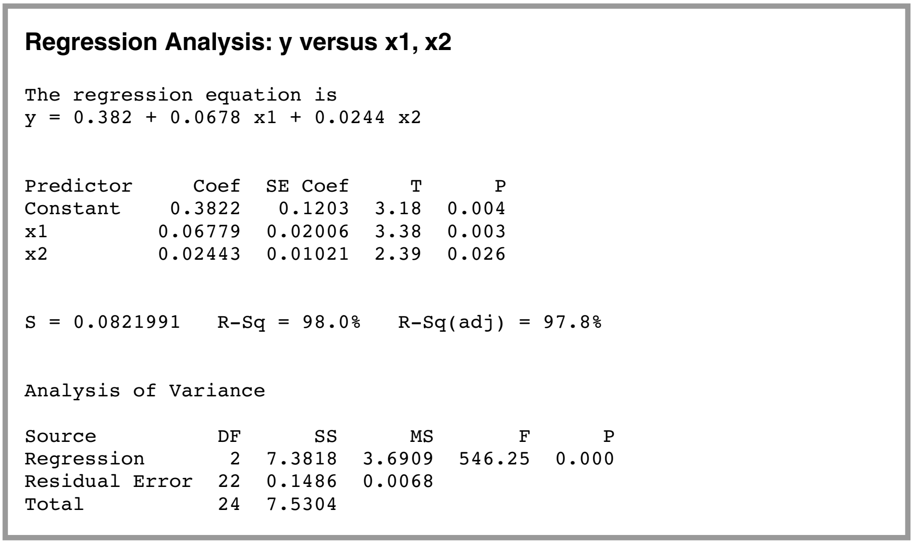
The further assumptions that the random errors are *independent* and *normally distributed* are also frequently made. A consequence of these assumptions is that the population of Y values that could be observed for a given pair of values of x_1 and x_2 has the normal distribution with mean $\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}$ and variance σ^2.

The next example concerns a market research company wishing to predict weekend circulation of daily newspapers in market areas. It ascertained circulation (Y, in thousands) in a sample of 25 market areas, together with total retail sales (x_1, in millions of dollars) and population density (x_2, adults per square mile). The data are available in Circulation.MTW and are from W. Daniel and J. Terrell, *Business Statistics for Management and Economics*, 5th edition, p. 531, © 1989 by Houghton Mifflin Company and used with permission. As a first step the data may be displayed in the form of scatterplots of y versus each of the two x variables as displayed in Figure 10.13. These may be obtained using **Graph > Matrix Plot...** and selecting **Each Y versus each X** and **Simple**.

These plots indicate linear relationships between both Y and x_1 and Y and x_2. A multiple regression model may be fitted using **Stat > Regression > Regression...**, specifying Y as the response and both x_1 and x_2 as predictors. As always, diagnostic plots of the residuals should be selected under **Graphs...**.

Part of the Session window output is shown in Panel 10.8. The multiple regression equation is stated initially. The ANOVA table provides a test of the null hypothesis that both β_1 and β_2 are zero. The P-value of 0.000, to three decimal places, provides strong evidence that the null hypothesis should be rejected in favour of the alternative hypothesis that not both not both of the partial regression coefficients are zero.

In the earlier section of the output the results of individual t-tests on the βs are given. With P-values of 0.004, 0.003 and 0.026 respectively, the null hypothesis that $\beta_0 = 0$, the null hypothesis that $\beta_1 = 0$ and the null hypothesis that $\beta_2 = 0$ would all be rejected at the 5% level of significance. In addition, the standard deviation of the random error is estimated to be 0.0822 and the R^2 value is 98%. Thus 98% of the variation in circulation (Y) can be explained by its



Regression Analysis: y versus x1, x2

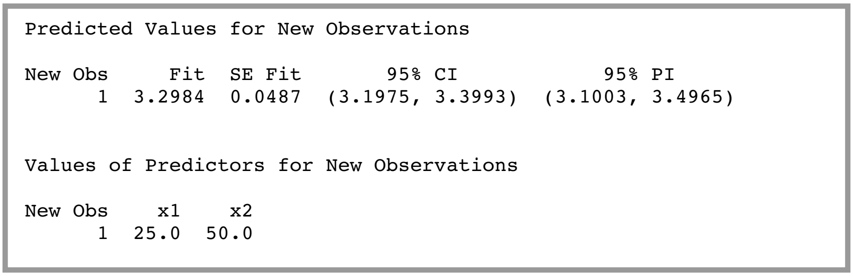
```
The regression equation is
y = 0.382 + 0.0678 x1 + 0.0244 x2


Predictor      Coef  SE Coef     T      P
Constant     0.3822   0.1203  3.18  0.004
x1          0.06779  0.02006  3.38  0.003
x2          0.02443  0.01021  2.39  0.026


S = 0.0821991   R-Sq = 98.0%   R-Sq(adj) = 97.8%


Analysis of Variance

Source          DF      SS      MS       F      P
Regression       2  7.3818  3.6909  546.25  0.000
Residual Error  22  0.1486  0.0068
Total           24  7.5304
```

Panel 10.8 Session window output for multiple regression.

linear dependence on both total retail sales (represented by x_1) and population density (x_2). Both a normal probability plot of the residuals and a plot of the residuals against fitted values appear to be satisfactory. Thus we can be confident that we have an adequate model of the situation.

Suppose that the company wished to predict circulation for a market area with total retail sales of $25 million and population density 50 adults per square mile. Predictions may be obtained by using **Stat > Regression > Regression...** again, with **Response:** *y* and **Predictors:** $x_1 x_2$, specified as previously. In order to predict for $x_1 = 25$ and $x_2 = 50$, under **Options...** specify **Prediction intervals for new observations:** 25 50. Note that order of entry is important here – the values of the predictor variables must be entered in the same order as the variables were entered under predictors. More than one pair of values for x_1 and x_2 may be entered if required, or alternatively, the names of two columns containing matching pairs of values for x_1 and x_2 may be entered.

The resulting Session window output is shown in Panel 10.9. Thus for the market area of interest the model predicts circulation of 3.2984 thousands, i.e. of 3298 to the nearest whole number. The 95% confidence interval converts to (3198, 3399). We can be 95% confident that



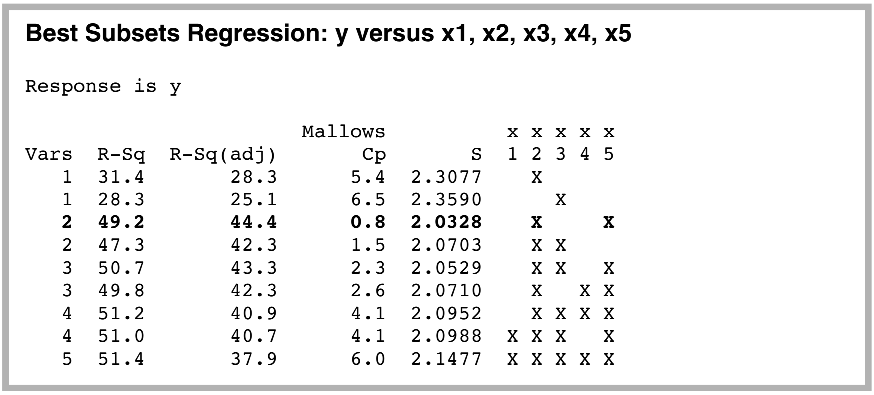
```
Predicted Values for New Observations

New Obs     Fit  SE Fit       95% CI              95% PI
      1  3.2984  0.0487  (3.1975, 3.3993)  (3.1003, 3.4965)


Values of Predictors for New Observations

New Obs    x1    x2
      1  25.0  50.0
```

Panel 10.9 Prediction using the multiple regression model.

the *mean* circulation, for market areas with total retail sales of \$25 million and population density of 50, will lie in this interval. The 95% prediction interval converts to (3100, 3496). We can be 95% confident that the *individual* circulation, for an individual market area with total retail sales of \$25 million and population density of 50, will lie in this interval. Such predictions could be of value to the company in terms of making improvements to production scheduling and distribution.

As a second example, data on percentage elongation of 24 specimens of a steel alloy will be investigated. The data are available in Elongation.MTW and are reproduced by permission of Oxford University Press Inc., New York. The data, available in Elongation.MTW are from p.426 of *Fundamental Concepts in the Design of Experiments*, 5th edition by Charles R. Hicks and Kenneth V. Turner, Jr, copyright © 1964, 1973, 1982, 1993, 1999 and used by permission of Oxford University Press, Inc., New York. Elongation (Y) and the percentages of five specific chemical elements (x_1, x_2, x_3, x_4 and x_5) were determined for each specimen. In terms of creating a multiple regression model here, which is linear in the predictor variables x_1, x_2, x_3, x_4 and x_5, we have two choices for each predictor – either omit or include. Thus there are $2^5 = 32$ possible models, 31 if we discount the trivial model that involves none of the predictors. Initial exploration of potential models may be made using **Stat > Regression > Best Subsets...** with y specified as **Response:**, x_1, x_2, x_3, x_4 and x_5 as **Free Predictors** and defaults accepted otherwise.

The Session window output is shown in Panel 10.10. By default, summary information is provided in the output for the 'best' two models involving 1, 2, 3 and 4 predictors and the single model involving all 5 predictors. In addition to R^2 and the adjusted R^2, Mallow's C_p statistic is listed for each of the nine models. (The C_p statistic here should not be confused with the capability index C_{p}.) For example, the fourth row of information (highlighted in bold) indicates that the multiple regression involving the two predictors x_2 and x_5 (indicated by the crosses vertically below the 2 and 5 in the headings) has $R^2 = 49.2\%$, adjusted $R^2 = 44.4\%$ and Mallow's C_p statistic of 0.8. Hicks and Turner (1999, p. 425) comment: 'Even though many independent variables may be used, simpler models are more appealing and easier to interpret. Thus, we try to identify the smallest subset of the independent variables that will provide an adequate model.' Many statisticians compare the adjusted R^2 values for candidate models.



Best Subsets Regression: y versus x1, x2, x3, x4, x5

```
Response is y

                      Mallows           x x x x x
Vars  R-Sq  R-Sq(adj)      Cp        S  1 2 3 4 5
   1  31.4       28.3     5.4   2.3077    X
   1  28.3       25.1     6.5   2.3590      X
   2  49.2       44.4     0.8   2.0328    X       X
   2  47.3       42.3     1.5   2.0703    X X
   3  50.7       43.3     2.3   2.0529    X X     X
   3  49.8       42.3     2.6   2.0710    X   X X
   4  51.2       40.9     4.1   2.0952    X X X X
   4  51.0       40.7     4.1   2.0988  X X X   X
   5  51.4       37.9     6.0   2.1477  X X X X X
```

Panel 10.10 Session window output from best subsets regression.

With a simple linear regression model, i.e. one involving a single predictor variable x, and a response y, R^2 is the coefficient of determination given by the square of the correlation coefficient, r, between x and y. With two or more predictor variables, R^2 may be considered as the square of the correlation coefficient between the observed data value, y, and the fitted value $\hat{Y}$. Every new predictor added to a multiple regression model will lead to an increase in the value of R^2. The adjusted R^2, denoted by R-Sq (adj) in Minitab, is given by

$$R^2_{\text{adj}} = 1 - \frac{n-1}{n-p-1}(1-R^2),$$

where n is the number of observations and p is the number of predictor variables in the model. The statistic R^2 is computed from sample data and may therefore be considered as an estimate of a population value. The value of adjusted R^2 provides a better estimate of this population value than does R^2.

A plot of the 'best' adjusted R^2 values for each number of predictors is displayed in Figure 10.14. This points to the model involving the two predictors x_2 and x_5 as being worthy of further investigation. This may be done using **Stat > Regression > Regression...**. The usual residual plots should be examined together with plots of residual against predictor (x) variables not currently in the model. The residual for the 23rd specimen is unusually large. In situations such as this, assuming no data input error has been made, analysis of the data with that observation excluded can be informative. Pattern or structure in residual plots can indicate the need to introduce predictors that are the squares of x variables (curvature terms) or the products of pairs of x variables (interaction terms).

Some statisticians use Mallow's C_p statistic as an aid to model selection. 'The C_p statistic appeals nicely to common sense and is developed from considerations of the proper compromise between excessive bias incurred when one underfits (chooses too few model terms) and excessive prediction variance produced when one overfits (has redundancies in the

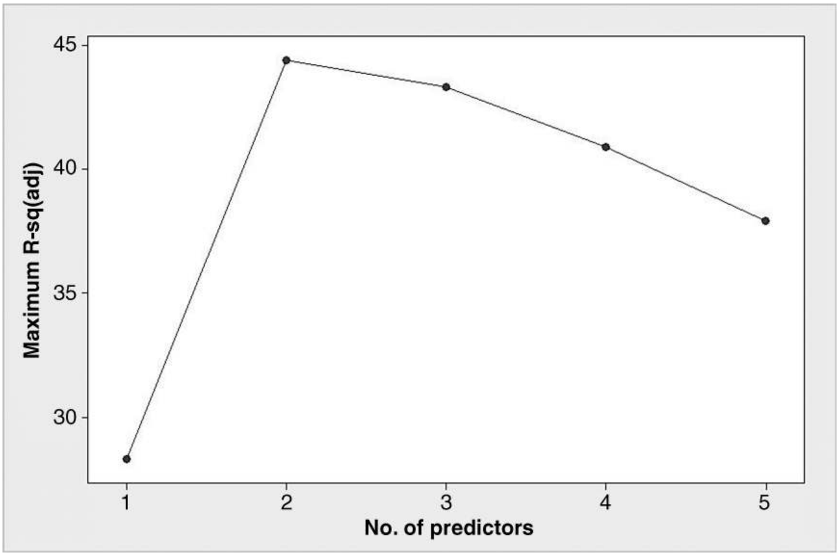


Figure 10.14 Plot of maximum adjusted R^2 versus number of predictors.

model)' (Walpole and Myers, 1989, p. 447). If p denotes the number of model parameters then '$C_p > p$ indicates a model that is biased due to being an underfitted model, while $C_p \approx p$ indicates a reasonable model' (Walpole and Myers, 1989, p. 448).

Minitab provides another major tool for the development of regression models – stepwise regression methods. It will not be considered in this book. Hicks and Turner (1999, p. 428) comment that 'these procedures may miss some models considered by the all-subsets procedure' and 'use of all-subsets regression is recommended when adequate computing facilities are available'. Readers who wish to learn more about variable selection in regression model building might find it beneficial to consult the book by Montgomery *et al.* (2006).

10.3 Response surface methods

The creation of response surface models essentially involves fitting multiple regression models to experimental data. To introduce response surface experimental designs, we consider an experiment carried out to investigate the tensile strength (Y, g/cm) of film used in the food industry. Customers of the manufacturer of the film had been experiencing problems due to the film tearing during food packaging operations. The manufacturer set up a Six Sigma project in order to determine if changes to manufacturing process settings would yield stronger film. The project team decided that the factors seal temperature (°C) and the amount of a plastic additive (%) should be investigated. The settings currently used in production were 140 °C and 4% respectively, and mean tensile strength was stable and predictable with a mean around 63 g/cm. Phase 1 of the experimentation involved a 2^2 factorial design (replicated twice) with low and high levels of 120 °C and 160 °C for temperature and 2% and 6% for amount of additive, supplemented by four runs carried out with the current settings of 140 °C and 4%. The five factor–level combinations involved are displayed in Figure 10.15. There is no simple rule for

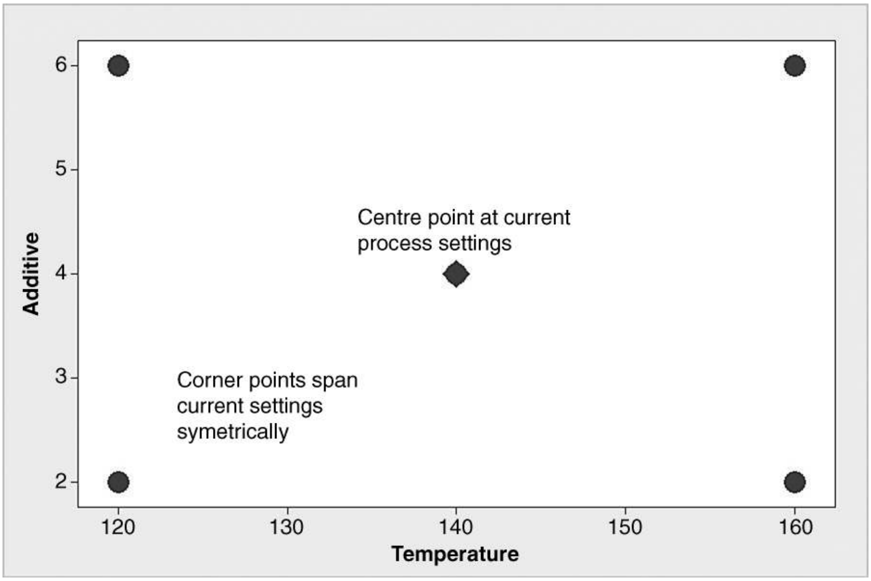


Figure 10.15 Factor–level combinations for 2^2 factorial experiment with centre points.

Table 10.2 Phase 1 data for 2^2 factorial experiment with centre points.

Temperature	Additive	*y*
160	2	64.1
140	4	61.6
160	2	60.8
140	4	62.4
120	2	52.1
120	6	60.6
160	6	69.6
120	2	53.3
120	6	61.7
140	4	63.1
140	4	62.5
160	6	70.5

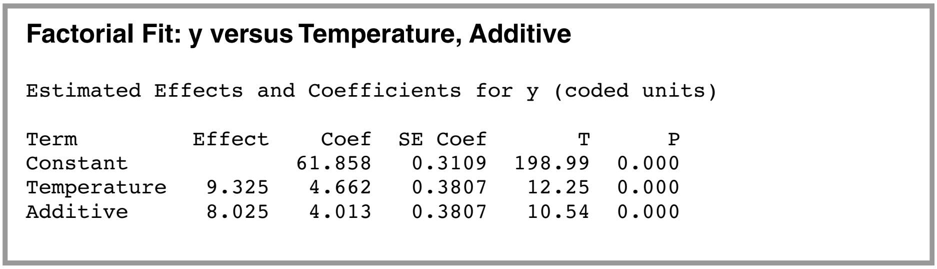
selection of the low and high levels in such scenarios – the project team would need to consider the selection carefully, taking into account the views of people with knowledge and experience of running the process.

Use was made of **Stat > DOE > Factorial > Create Factorial Design...** to carry out the design and create a pro forma for recording results. With **Number of factors:** 2, under **Designs...** the specifications made were **Number of center points per block:** 4, **Number of replicates for corner points:** 2 and **Number of blocks:** 1. Randomization was used. The key data are shown in Table 10.2, with the factor levels in the random order obtained from the software, and the full data set is available in Phase1.MTW. (If the reader wishes to re-create the design and perform the analysis that follows then the response data in the final column of Table 10.2 will have to be entered into his/her worksheet in the appropriate order.)

The initial ANOVA obtained using **Stat > DOE > Factorial > Analyze Factorial Design...** is shown in Panel 10.11. It provides no evidence of any interaction or of any curvature (*P*-values 0.598 and 0.263, respectively). Before creating a contour plot of the response surface one may therefore remove the interaction term from the model. This is achieved by using **Stat > DOE > Factorial > Analyze Factorial Design...** again using **Terms...** to remove the interaction (*AB*) term from the **Selected Terms:** window. One must

```
Analysis of Variance for y (coded units)

Source                  DF   Seq SS   Adj SS   Adj MS       F      P
Main Effects             2  302.712  302.712  151.356  127.42  0.000
  Temperature            1  173.911  173.911  173.911  146.41  0.000
  Additive               1  128.801  128.801  128.801  108.43  0.000
2-Way Interactions       1    0.361    0.361    0.361    0.30  0.598
  Temperature*Additive   1    0.361    0.361    0.361    0.30  0.598
  Curvature              1    1.760    1.760    1.760    1.48  0.263
Residual Error           7    8.315    8.315    1.188
  Pure Error             7    8.315    8.315    1.188
```

Panel 10.11 ANOVA for Phase 1 experiment.



Factorial Fit: y versus Temperature, Additive

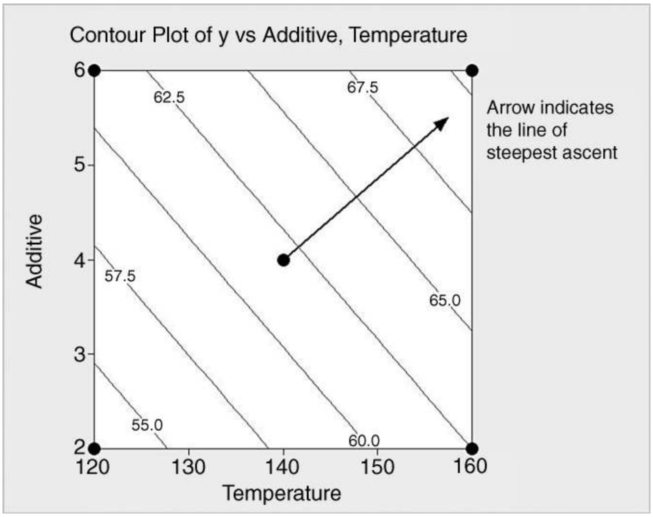
```
Estimated Effects and Coefficients for y (coded units)

Term          Effect     Coef  SE Coef        T      P
Constant               61.858   0.3109   198.99  0.000
Temperature    9.325    4.662   0.3807    12.25  0.000
Additive       8.025    4.013   0.3807    10.54  0.000
```

Panel 10.12 Revised model for Phase 1.

also uncheck **Include center points in the model** in order to create a contour plot as a follow-up to the model revision. With no evidence of interaction or curvature a plane surface will provide an adequate model in the region of the design space explored in Phase 1, so omitting the centre points is justified. Of particular interest for this revised model is the output displayed in Panel 10.12. The *P*-values for temperature and amount of additive provide very strong evidence of important main effects of both.

Having fitted the revised model, one may proceed to use **Stat > DOE > Factorial > Contour/Surface Plots...** to create the required display. **Contour plot** was checked and **Setup...** involved specifying **Response:** C7 y. In addition, under **Contours...** the option to **Use defaults** was accepted for **Contour Levels** and, under **Data Display**, both **Contour Lines** and **Symbols at design points** were checked, but not **Area**. The plot is displayed in Figure 10.16.

Note the solid circle symbols indicating the five FLCs, or design points, used in the experimentation so far and previously displayed in Figure 10.15. A line from the centre point of the design, in the direction of increasing tensile strength, *y*, at right angles to the contour lines, has been added. This is the line of steepest ascent, and Phase 2 of the experimentation involved duplicate process runs at a series of FLCs along the line of steepest ascent. (A method



Contour Plot of y vs Additive, Temperature

Arrow indicates the line of steepest ascent

Figure 10.16 Contour plot for Phase 1.

Table 10.3 Phase 2 data from exploration along line of steepest ascent.

Temperature	Additive	y_1	y_2	Mean
170	6.6	74.8	74.3	74.55
190	8.3	78.8	77.5	78.15
210	10.0	80.1	79.2	79.65
230	11.7	76.3	77.5	76.90
250	13.5	74.4	74.6	74.50

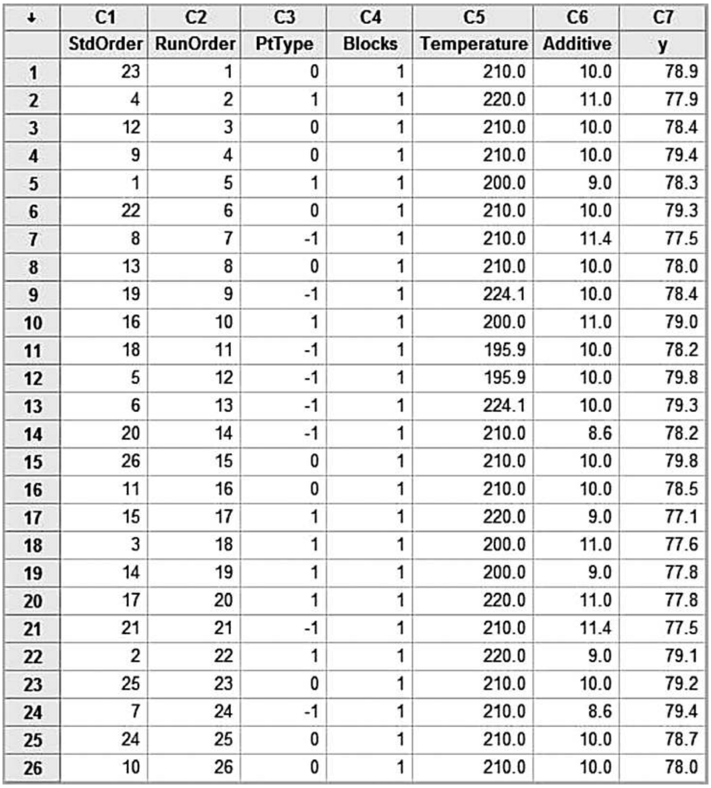
for determining FLCs along a line of steepest ascent is detailed in one of the follow-up exercises.) The data are displayed in Table 10.3 and available in Phase2.MTW.

Note that, as the investigation moved further away from the original centre point, tensile strength began to increase and then decreased again. It peaked with temperature 210 °C and 10% additive, so Phase 3 of the experimentation involved a response surface design, with two replications, centred at this point.

In order to create such a design use may be made of **Stat > DOE > Response Surface > Create Response Surface Design. . . .** The default **Central Composite** type of design was accepted, with **Number of factors:** 2. Clicking on **Designs. . .** reveals the two available designs for the case of two factors. The default **Full** design involves 13 runs (or FLCs) and one block. This design was adopted with **Number of replicates:** 2 specified and defaults accepted otherwise. Under **Factors:** the default that **Levels Define Cube points** (corners of the square defining the 2^2 factorial that forms the basis of the design in terms of coded units) was accepted and then the factors temperature and amount of additive with the levels selected by the project team (200 °C and 220 °C for temperature and 9% and 11% for amount of additive) were specified in the usual way. The reader should note that these FLCs for corner points are symmetrically placed relative to the centre point levels of 210 °C for temperature and 10% for amount of additive identified in Phase 2 as being a region of the design space worth exploring. The data are displayed in Figure 10.17 and available in Phase3.MTW.

In terms of coordinates in the temperature–additive space, the centre point for the design was (210, 10), with the four corner points (200, 9), (220, 9), (200, 11), (220, 11). In addition, the central composite response surface design involved the FLCs corresponding to the axial points (195.9, 10), (210, 11.4), (224.1, 10), (210, 8.6). Note that Minitab uses codes in the PtType column of the worksheet to indicate centre points (code 0), corner (or cube) points (code 1) and axial points (code − 1).

Use of **Stat > DOE > Response Surface > Analyze Response Surface Design. . .** is required to analyse the data. Once this had been done the contour plot displayed in Figure 10.18 was created using **Stat > DOE > Response Surface > Contour/Surface Plots. . . .** The contour plot indicates that maximum strength, y, of around 79 g/cm can be expected with the FLC of temperature 206 °C and additive 9.8%. Thus, given that with current operating conditions mean strength was around 63 g/cm, the Six Sigma project indicates that a 25% increase to a mean of around 79 g/cm appears feasible. The peak on the response surface has been indicated by the triangular symbol and the nine FLCs for the response surface design are indicated by the solid circles. Hogg and Ledolter (1992, pp. 409–410) provide a calculus-based procedure for calculating the factor levels corresponding to an optimum point and for determining the nature of the optimum point. The solid circles indicate the FLCs used in the response surface design.



↓	C1	C2	C3	C4	C5	C6	C7
	StdOrder	RunOrder	PtType	Blocks	Temperature	Additive	y
1	23	1	0	1	210.0	10.0	78.9
2	4	2	1	1	220.0	11.0	77.9
3	12	3	0	1	210.0	10.0	78.4
4	9	4	0	1	210.0	10.0	79.4
5	1	5	1	1	200.0	9.0	78.3
6	22	6	0	1	210.0	10.0	79.3
7	8	7	-1	1	210.0	11.4	77.5
8	13	8	0	1	210.0	10.0	78.0
9	19	9	-1	1	224.1	10.0	78.4
10	16	10	1	1	200.0	11.0	79.0
11	18	11	-1	1	195.9	10.0	78.2
12	5	12	-1	1	195.9	10.0	79.8
13	6	13	-1	1	224.1	10.0	79.3
14	20	14	-1	1	210.0	8.6	78.2
15	26	15	0	1	210.0	10.0	79.8
16	11	16	0	1	210.0	10.0	78.5
17	15	17	1	1	220.0	9.0	77.1
18	3	18	1	1	200.0	11.0	77.6
19	14	19	1	1	200.0	9.0	77.8
20	17	20	1	1	220.0	11.0	77.8
21	21	21	-1	1	210.0	11.4	77.5
22	2	22	1	1	220.0	9.0	79.1
23	25	23	0	1	210.0	10.0	79.2
24	7	24	-1	1	210.0	8.6	79.4
25	24	25	0	1	210.0	10.0	78.7
26	10	26	0	1	210.0	10.0	78.0

Figure 10.17 Response surface design experiment worksheet.

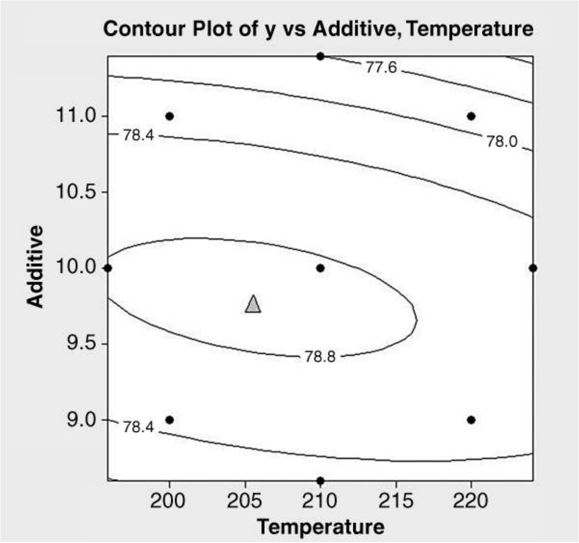


Figure 10.18 Contour plot for Phase 3.

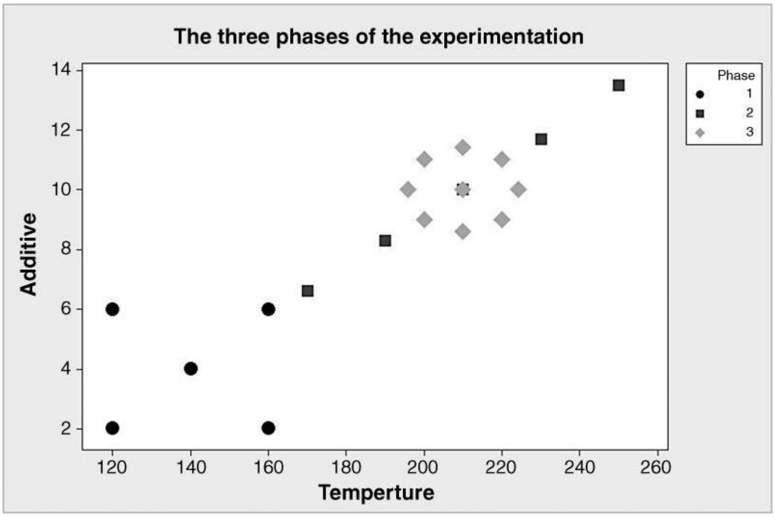


Figure 10.19 Display of all FLCs used in the three phases of the experimentation.

The FLCs used in all three phases of the experimentation are displayed in Figure 10.19. The display highlights the iterative nature of the experimentation. The factorial experiment performed in Phase 1 indicated a promising direction in which to carry out further investigation. Having found an FLC at which yield reached a peak from Phase 2, a response surface design centred in the region of this combination provided Phase 3 and indicated a likely optimal FLC.

Minitab enables central composite response surface designs to be created and analysed for up to 10 factors. In many situations process teams have to consider a number of responses, and it may not be possible to optimize all of them simultaneously. Montgomery (2005a, 2009) gives comprehensive details and examples.

10.4 Categorical data and logistic regression

The use of correlation to investigate relationships or association between continuous random variables and the use of least squares regression to model relationships in which the response was continuous were introduced in Chapter 3 and developed further earlier in this chapter. The rest of this chapter is devoted to the introduction of methods for investigation of association between categorical variables and for modelling relationships in which the response is categorical. There are two main types of measurement scales for categorical variables – *ordinal* and *nominal*. An example of an ordinal scale is the assessment of the condition of a road surface as bad, poor, fair, good or excellent. An example of a nominal scale is the plant where a model of automobile was assembled, e.g. Linwood or Ryton.

10.4.1 Tests of association using the chi-square distribution

A company that manufactures marine radar systems builds scanners using main bearings from two suppliers, A and B. One year after installation of systems, engineers from the company

Table 10.4 Contingency table of scanner data.

Classification of sample of scanners by bearing supplier and bearing state		Bearing state		
		Sound	Worn	
Bearing supplier	A	32	3	35
	B	48	17	65
		80	20	100

service the scanners and check the main bearings for wear. From a random sample of 100 service reports the scanners were categorized according to the bearing supplier and the state of the main bearing, yielding Table 10.4.

The null hypothesis of interest here is that there is no association between bearing supplier and bearing state, with the alternative hypothesis being that there is an association, i.e. that the bearing state is *contingent* on the supplier. Tables such as Table 10.4 are known as *contingency tables*. If the null hypothesis is true then the events that a bearing was supplied by A and is sound are independent. From the table we may then carry out the calculations displayed in Box 10.1. The reader is invited to calculate, using the method outlined in Box 10.1, the expected counts for the other three cells – all four expected counts are shown in Table 10.5. In the case of a 2 × 2 contingency table as we have here, once one expected frequency has been obtained the others may be obtained by ensuring that the correct marginal totals are obtained. Thus a 2 × 2 contingency table has one degree of freedom – an $r \times c$ contingency table, with r rows and c columns, has degrees of freedom given by $(r - 1) \times (c - 1)$.

The test statistic for the hypothesis test is:

$$
\begin{aligned}
\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} &= \frac{(32-28)^2}{28} + \frac{(3-7)^2}{7} + \frac{(48-52)^2}{52} + \frac{(17-13)^2}{13} \\
&= 0.571 + 2.286 + 0.308 + 1.231 \\
&= 4.396, \text{ with 1 degree of freedom.}
\end{aligned}
$$

An estimate for the probability that a bearing was supplied by A is $P(A) = 35/100$. An estimate for the probability that a bearing was sound is $P(S) = 80/100$. If the null hypothesis is true then an estimate of $P(A \cap S)$ is

$$P(A \cap S) = P(A) \times P(S) = \frac{35}{100} \times \frac{80}{100}.$$

Hence, the expected frequency of scanners supplied by A with sound bearings is

$$100 \times \frac{35}{100} \times \frac{80}{100} = \frac{35 \times 80}{100} = \frac{\text{Row total} \times \text{Column total}}{\text{Sample size}}.$$

Box 10.1 Calculation of an expected frequency.

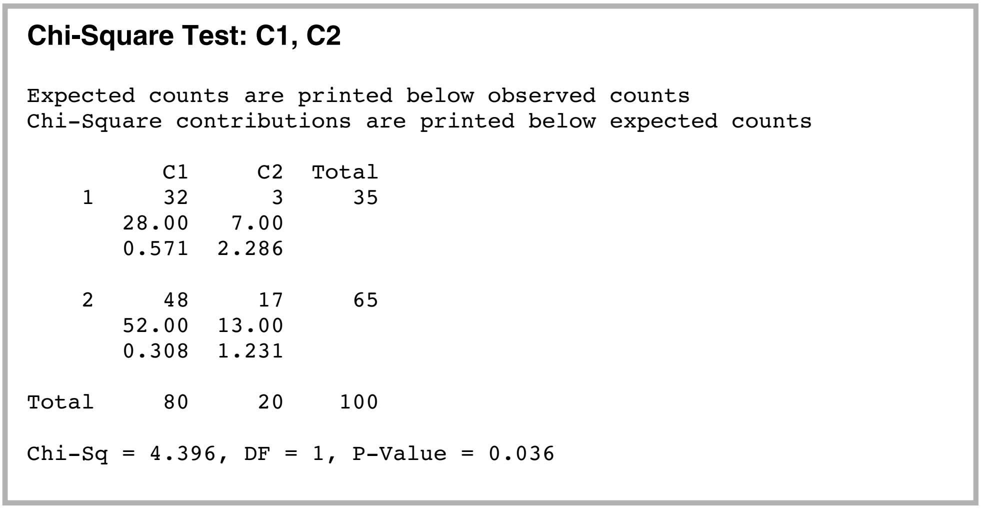
Table 10.5 Contingency table showing observed (O_i) and expected (E_i) counts.

Classification of sample of scanners by bearing supplier and bearing state		Bearing state		
		Sound	Worn	
Bearing supplier	A	32 (28)	3 (7)	35
	B	48 (52)	17 (13)	65
		80	20	100

Use of **Graph** > **Probability Distribution Plot...** > **View Probability** yields the P-value. Under the **Distribution** tab, **Distribution** Chi-Square is specified from the drop-down menu with **Degrees of freedom:** 1. Under the **Shaded Area** tab **X Value** is checked, **X Value:** 4.396 specified and **Right Tail** selected. On clicking **OK** the display reveals the P-value to be 0.036, correct to three decimal places. Since this is less than 0.05 the null hypothesis that there is no association between bearing supplier and bearing state would be rejected in favour of the alternative there is an association, at the 5% level of significance. This test is often referred to as *Pearson's chi-square test* in honour of the statistician Karl Pearson.

This result may be obtained directly from Minitab using **Stat** > **Tables** > **Chi-Square Test (Two-Way Table in Worksheet)...** having set up the four counts from the contingency table in, say, columns C1 and C2 of a worksheet. Having specified **Columns containing the table:** C1 C2, clicking **OK** yields the Session window output in Panel 10.13 which confirms all the calculations performed earlier.

From the contingency table we have an estimate of the probability of wear for a bearing from supplier A, $p_A = 3/35 = 0.0857$, and from supplier B, $p_B = 17/65 = 0.2615$. Thus we could say that we estimate that a bearing from supplier A is approximately one third as likely to



Chi-Square Test: C1, C2

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

         C1     C2  Total
    1    32      3     35
      28.00   7.00
      0.571  2.286

    2    48     17     65
      52.00  13.00
      0.308  1.231

Total    80     20    100

Chi-Sq = 4.396, DF = 1, P-Value = 0.036
```

Panel 10.13 Output for chi-square test for association.

show signs of wear as a bearing from supplier B. One measure of association for a two-way contingency table is the *relative risk* defined as the ratio of the two probabilities: here the relative risk (A to B) is $p_A/p_B = 0.0857/0.2615 = 0.338$; similarly, the relative risk (B to A) is the reciprocal of this, 3.051. For an event with probability p the odds are defined as $p/(1 - p)$. Hence, we have that the odds for wear in a bearing from supplier A are

$$\frac{p_A}{1-p_A} = \frac{0.0857}{0.9143} = 0.094,$$

and in a bearing from supplier B are

$$\frac{p_B}{1-p_B} = \frac{0.2615}{0.7385} = 0.354.$$

Another measure of association for a two-way contingency table is the *odds ratio*, defined as the ratio of the two odds. Thus the odds ratio (A to B) is

$$\frac{\text{Odds}_A}{\text{Odds}_B} = \frac{0.0937}{0.3541} = 0.26,$$

and similarly the odds ratio (B to A) is the reciprocal of that, 3.78.

Intuitively, the *relative risk* is an easier measure of association to interpret than the odds ratio. The two are related by the equation

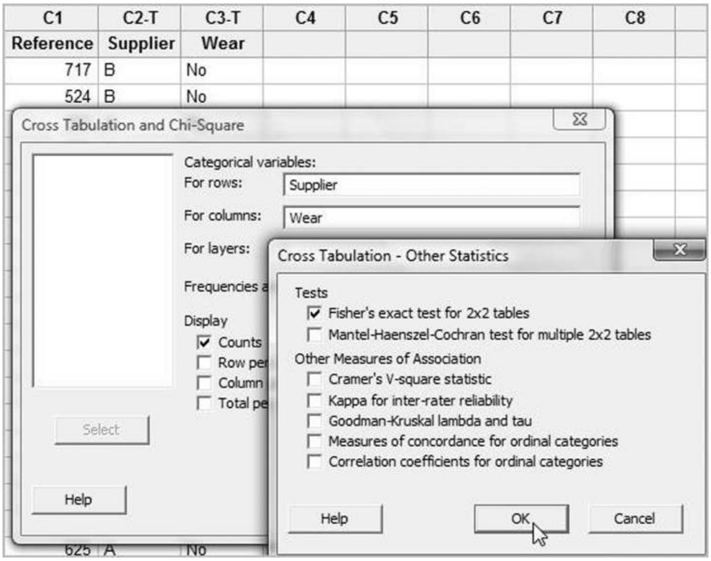
$$\text{Relative risk(A to B)} = \text{Odds ratio(A to B)} \times \frac{1-p_A}{1-p_B}.$$

Odds ratios occur in binary logistic regression analysis and will be referred to again in the next section.

The raw data extracted from the scanner service reports is provided in the worksheet Bearings.MTW. The first column contains scanner service reference numbers, the second the supplier of the main bearing, and the third indicates whether or not signs of wear were found. In order to cross-tabulate the data to form the contingency table and to carry out the chi-square test of association one may use **Stat > Tables > Cross Tabulation and Chi-Square...** as indicated in Figure 10.20. Supplier was allocated to rows and Wear to columns, counts checked under **Display**, **Fisher's exact test for 2 × 2 tables** checked under **Other Stats...** and Chi-Square analysis checked under **Chi-Square...**.

The reader is invited to check that the resultant Session window output includes the key output in Panel 10.13 together with a *P*-value of 0.027 for an alternative chi-square test and a *P*-value of 0.039 for Fisher's exact test for association. The latter test is named in honour of the statistician Ronald Fisher. All three tests lead to the same conclusion, i.e. that the null hypothesis of no association would be rejected at the 5% level of significance.

The Pearson chi-square test involves a degree of approximation. Minitab displays the number of cells that have expected counts less than 5. Minitab Help (**Stat > Tables > Cross Tabulation and Chi-Square... > Help > see also > Methods and formulas > Chi-Square test**) states: 'Some statisticians hesitate to use the χ^2 test if more than 20% of the cells have expected counts below five, especially if the p-value is small and these cells give



C1	C2-T	C3-T	C4	C5	C6	C7	C8
Reference	Supplier	Wear					
717	B	No					
524	B	No					
625	A	No					

Cross Tabulation and Chi-Square

Categorical variables:
For rows: Supplier
For columns: Wear
For layers:
Frequencies a

Display
☑ Counts
☐ Row pe
☐ Column
☐ Total pe

Cross Tabulation - Other Statistics

Tests
☑ Fisher's exact test for 2x2 tables
☐ Mantel-Haenszel-Cochran test for multiple 2x2 tables

Other Measures of Association
☐ Cramer's V-square statistic
☐ Kappa for inter-rater reliability
☐ Goodman-Kruskal lambda and tau
☐ Measures of concordance for ordinal categories
☐ Correlation coefficients for ordinal categories

Figure 10.20 Dialog for cross-tabulation and chi-square test of association.

a large contribution to the total χ^2 value'. If the expected counts for some are small it may be possible to carry out an analysis by combining or omitting row and/or column categories. On the other hand, as the name implies, the Fisher test is exact and may be used with confidence when expected counts in a 2 × 2 contingency table are low.

Duncan (1959) gives the data in Table 10.6 on causes of rejection of metal casting by week of manufacture. The data, from Hunt (1948), reproduced by permission of the American Society for Quality, are available in Castings.MTW. The reader is invited to use **Stat > Tables > Chi-Square Test (Two-Way Table in Worksheet)...** to carry out the test and verify that the Session window output includes the statement 'Chi-Sq = 45.597, DF = 12, P-Value = 0.000'. No expected counts less than 5 were obtained so the analysis provides very strong evidence (P-value < 0.001) that there is a difference in the distribution of rejects from week to week.

Table 10.6 Causes of rejection of metal castings.

Cause of rejection	Week 1	Week 2	Week 3
Sand	97	120	82
Misrun	8	15	4
Shift	18	12	0
Drop	8	13	12
Corebreak	23	21	38
Broken	21	17	25
Other	5	15	19

Table 10.7 Space Shuttle launch O-ring data.

Temperature x	Launches	O-ring failure-free launches	$P(\text{Success}) = p$
53	1	0	0.0
57	1	0	0.0
58	1	0	0.0
63	1	0	0.0
66	1	1	1.0
67	3	3	1.0
68	1	1	1.0
69	1	1	1.0
70	4	2	0.5
72	1	1	1.0
73	1	1	1.0
75	2	1	0.5
76	2	2	1.0
78	1	1	1.0
79	1	1	1.0
81	1	1	1.0

10.4.2 Binary logistic regression

In binary logistic regression the response variable has two categories. In order to introduce the topic, we consider the data in Table 10.7 and Space_Shuttle.MTW from Dalal *et al.* (1989) and reprinted with permission from the *Journal of the American Statistical Association*, all rights reserved. It gives temperature (°F) at the time of launch for 23 Space Shuttle missions and a classification of each mission as either a success or failure in terms of O-ring performance. The classification was based on examination of O-rings that became available for inspection following launches. The temperature column gives the air temperature at the time of launch. For example, for the two launches that took place with air temperature 75 °F, there was one where O-ring failure is known to have occurred. Thus at air temperature 75 °F the estimated probability of a launch free of O-ring failures is $P(\text{Success}) = p = 1/2 = 0.5$.

The question of interest is whether or not there is any relationship between p and x. If the answer is in the affirmative then the question arises of whether or not the relationship can be modelled. As a first step a display was created in the form of a plot of p against x with a Lowess smoother applied. Locally weighted scatterplot smoothing, LOWESS, may be used to explore the relationship between two variables without fitting a specific model and may be added to a scatterplot by right-clicking the graph and selecting **Add > Lowess. . .**. The author used **Degree of smoothing:** 0.45 and **Number of steps:** 2 in creating Figure 10.21.

The fit from the smoother has the approximate appearance of an S-shaped or sigmoid curve. The logistic function is one mathematical function that may be used to model sigmoid curves and is described in Box 10.2.

The logit of p is a linear function of x so we could attempt to estimate the parameters α and β of the logistic model by fitting a straight line to a scatterplot of the logit of p versus x. We encounter an immediate problem with this approach for the current data set in that the logit

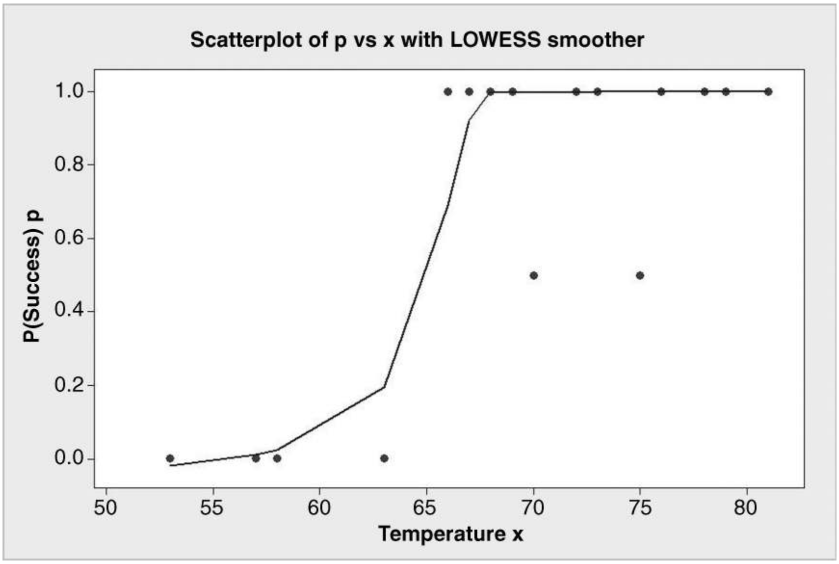


Figure 10.21 Plot of P(Success), p versus temperature, x.

function of p is undefined for $p = 0$ and for $p = 1$. In practice, the method of maximum likelihood is used to estimate the parameters α and β of the logistic model. It may be implemented in Minitab using **Stat > Regression > Binary Logistic Regression...**. The dialog required is shown in Figure 10.22. **Response in event/trial format** was checked as the response data appear in the worksheet as two columns – one containing the number of O-ring failure-free launches and the other containing the number of launches. Thus **Number of events:** O-ring failure free launches and **Number of trials:** Launches were specified. Under **Model:** Temperature x was entered. No entries were required under **Factor:** in this case. Under **Storage...**, **Event probability** was checked under **Characteristics of Estimated Equation** in order to obtain fitted values of the probability of an O-ring failure-free launch for subsequent plotting.

A logistic function linking p to x may be written in the form

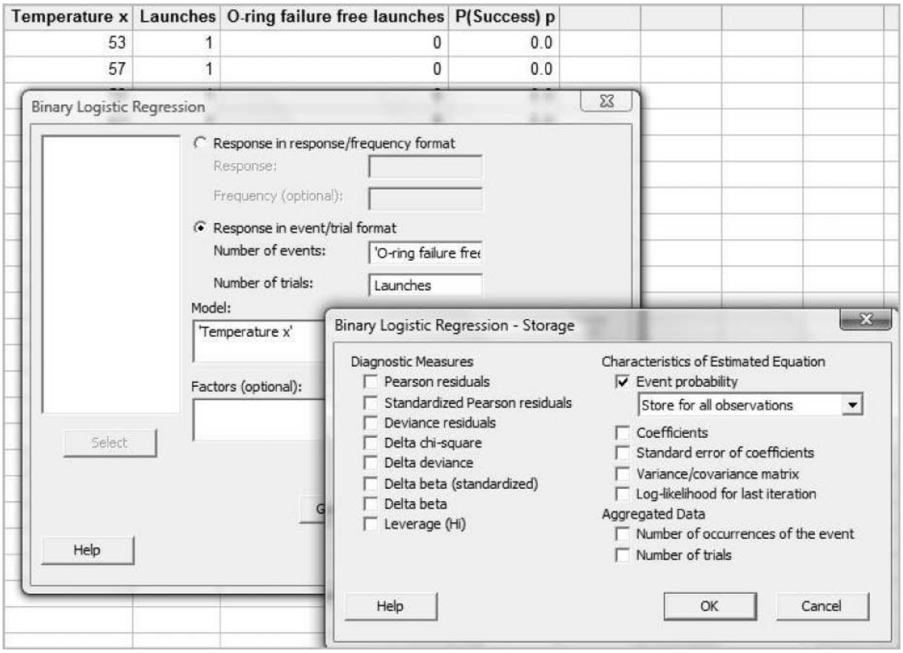
$$p = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)},$$

where α and β are the function parameters. The formula may be conveniently rearranged in the form

$$\ln \frac{p}{1 - p} = \alpha + \beta x.$$

The function $\ln(p/(1 - p))$ is the logit of p.

Box 10.2 The logistic function.



Temperature x	Launches	O-ring failure free launches	P(Success) p
53	1	0	0.0
57	1	0	0.0

Binary Logistic Regression

○ Response in response/frequency format

Response:

Frequency (optional):

● Response in event/trial format

Number of events: 'O-ring failure free

Number of trials: Launches

Model: 'Temperature x'

Factors (optional):

Binary Logistic Regression - Storage

Diagnostic Measures

☐ Pearson residuals
☐ Standardized Pearson residuals
☐ Deviance residuals
☐ Delta chi-square
☐ Delta deviance
☐ Delta beta (standardized)
☐ Delta beta
☐ Leverage (Hi)

Characteristics of Estimated Equation

☑ Event probability
Store for all observations
☐ Coefficients
☐ Standard error of coefficients
☐ Variance/covariance matrix
☐ Log-likelihood for last iteration

Aggregated Data

☐ Number of occurrences of the event
☐ Number of trials

Figure 10.22 Dialog for binary logistic regression.

Summary information given in the first section of the Session window output is shown in Panel 10.14. In simple linear regression, the expected value of the response, Y, is given by the linear function $\alpha + \beta x$, i.e. the expected value of the response is related directly to the linear function. In the form of logistic regression applied here the logit function of the expected probability of success is a linear function $\alpha + \beta x$. Thus in this case the logit function provides the link between the expected response of interest, probability of success, and the linear function of x. Hence the statement 'Link Function: Logit' in Panel 10.14. (Other link functions may be used but will not be considered in this book.) The Response Information summary displayed in Panel 10.14 may be readily checked from Table 10.7.

The next portion of the output is displayed in Panel 10.15. The method of maximum likelihood has provided the estimates of -15.04 and 0.2322 respectively for the logistic

Binary Logistic Regression: O-ring failu, Launches versus Temperature

```
Link Function: Logit


Response Information

Variable                      Value      Count
O-ring failure free launches  Event         16
                              Non-event      7
Launches                      Total         23
```

Panel 10.14 Summary information from logistic regression analysis.

```
Logistic Regression Table

                                                      Odds     95% CI
Predictor            Coef   SE Coef       Z       P  Ratio  Lower  Upper
Constant       -15.0429   7.37862  -2.04  0.041
Temperature x  0.232163  0.108236   2.14  0.032   1.26   1.02   1.56


Log-Likelihood = -10.158
Test that all slopes are zero: G = 7.952, DF = 1, P-Value = 0.005
```

Panel 10.15 The logistic regression table.

parameters α (the constant) and β (the coefficient of temperature x or slope parameter). With P-values of 0.041 and 0.032 we can conclude that both α and β differ significantly from zero.

The logistic model fitted to the data may be written as

$$\ln \frac{p}{1-p} = \alpha + \beta x.$$

The ratio $p/(1 - p)$ is the odds. The odds ratio in this context is the value of $\exp(\beta)$ and is estimated by $\exp(0.2322) = 1.26$, with 95% confidence interval (1.02, 1.56). The odds ratio may be interpreted as the factor by which the odds increase for a unit increase in x, i.e. for an increase of 1 °F in temperature,

The results from goodness-of-fit tests of the logistic model are provided in the next portion of Session window output in Panel 10.16, the null hypothesis being that a logistic model provides a good fit to the data. Since none of the three P-values are small the null hypothesis cannot be rejected, so we conclude that the logistic model provides a potential model of the situation.

The remaining Session window output comprises a table of observed and expected frequencies and a table of measures of association. The observed and fitted probabilities, computed using **Storage...**, from the model are plotted against temperature in Figure 10.23. The default name for the fitted event probability column, EPR01, was changed to pfit. Initially a scatterplot of p versus x was created. On right-clicking the plot **Add** > **Calculated Line...** was selected in order to plot the fitted logistic regression curve. Under **Coordinates** the selections **Y column:** pfit and **X column:** Temperature x were made.

Some statisticians quote the value of x that corresponds to probability 0.5. The dotted lines added to the plot indicate that, for the fitted model, the value of x for which p is 0.5 is

```
Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             11.1303  14  0.676
Deviance            11.9974  14  0.607
Hosmer-Lemeshow     11.0395   8  0.199
```

Panel 10.16 Goodness-of-fit tests for the logistic regression model.

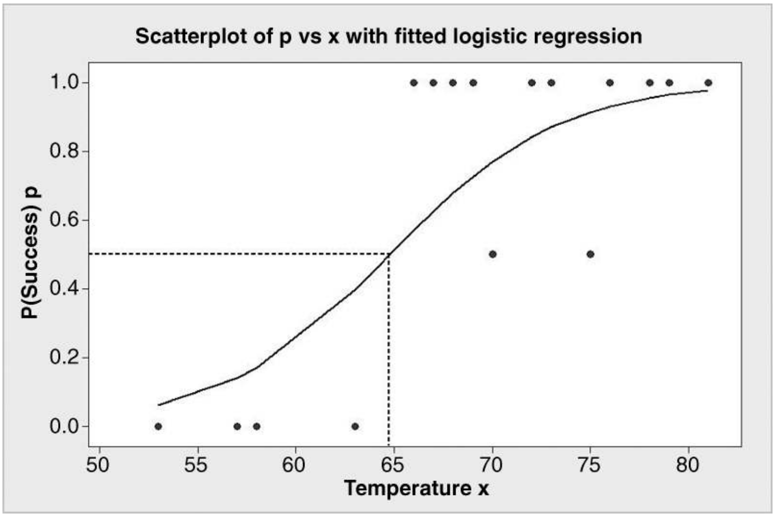


Figure 10.23 Binary logistic regression model fitted to data.

approximately 65. Thus the model predicts that there is probability of 0.5 that a Space Shuttle launch will be O-ring failure-free at temperature 65 °F. The calculation is as follows. We have $p = 0.5$, so

$$\ln\frac{0.5}{1-0.5} = \ln 1 = 0 = \alpha + \beta x$$

so

$$x = -\frac{\alpha}{\beta}.$$

Thus in this case the required temperature is estimated by

$$-\frac{-15.04}{0.2322} = 64.8.$$

The modelling carried out has established a relationship between the probability of a Space Shuttle launch being O-ring failure-free and air temperature. One may think of the random variable Y that takes the value 0 for a Space Shuttle launch with O-ring failures and the value 1 for a launch that was O-ring failure-free. The fitted logistic model enables prediction of the value of Y for a given value of x in the sense that prediction may be made of the conditional probability

$$P(Y = 1|\text{Temperature} = x) = p = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}.$$

For temperature 31 °F the model predicts

$$P(Y=1|x=31)=p=\frac{\exp(-15.04+0.2322\times 31)}{1+\exp(-15.04+0.2322\times 31)}=0.004.$$

Thus for a Space Shuttle launch taking place at 31 °F the model predicts a probability of 0.004 for it to be O-ring failure-free. (Such a prediction should be viewed with some caution as it involves extrapolation, i.e. prediction at a temperature well below any observed value. However, scrutiny of the column of predicted probabilities reveals that the fitted logistic model yields a probability of 0.06 for a launch at 53 °F, the lowest observed temperature, to be O-ring failure free.) At the time of the *Challenger* Space Shuttle launch on 28 January 1986 the air temperature was 31 °F. The catastrophe that occurred was attributed to failure of an O-ring.

When there are two or more predictor variables x_1, x_2, ..., the logistic regression model becomes

$$\ln\frac{p}{1-p}=\beta_0+\beta_1x_1+\beta_2x_2+\ \ldots.$$

As an example, we will consider data discussed by Everitt (1994, pp. 54–66) on patients suffering from acute myeloblastic leukaemia. The patients were given a course of treatment and the binary response recorded was whether or not a patient responded to treatment. The data are from a paper by Hart (1977), available in Leukaemia.MTW and reproduced by permission of John Wiley and Sons Inc., New York.

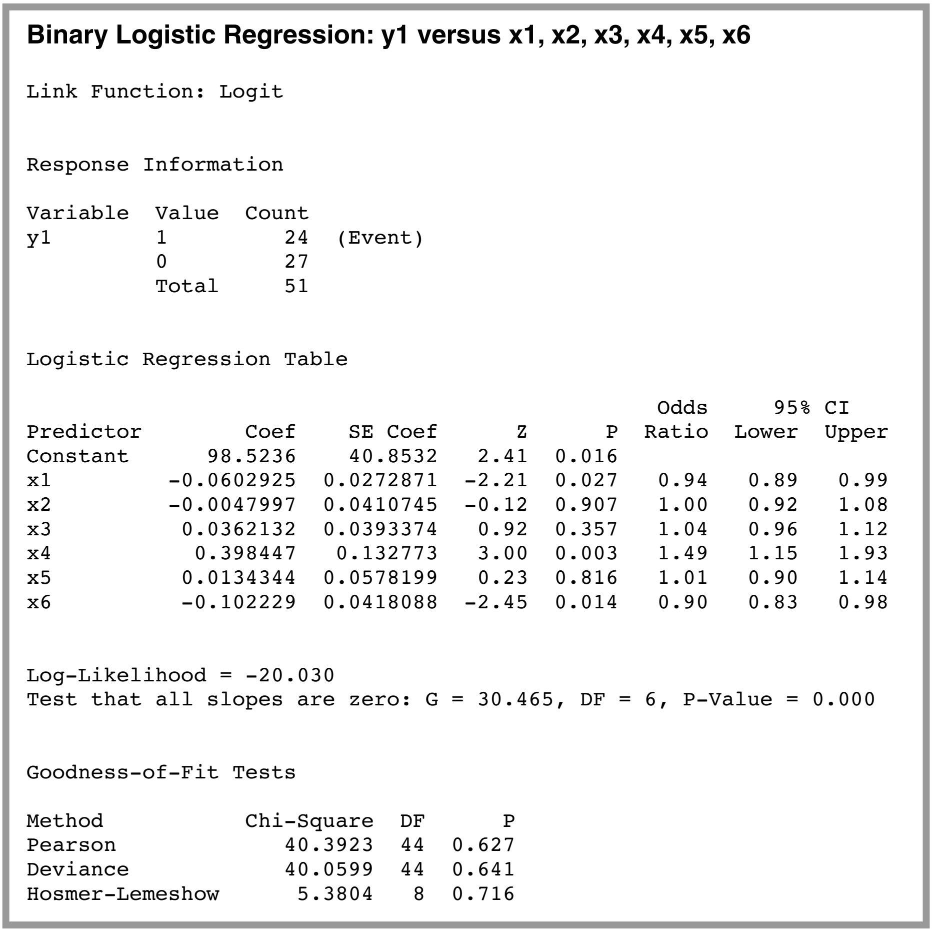
Six variables were recorded for each patient prior to treatment: age at diagnosis, x_1; smear differential percentage of blasts, x_2; percentage of absolute marrow leukaemia infiltrate, x_3; percentage labelling index of the bone marrow leukaemia cells, x_4; absolute blasts, x_5; and highest temperature prior to treatment, x_6. The responses recorded were: the response to treatment, y_1, where 0 = Fails to respond to treatment, 1 = Responds to treatment; survival time from diagnosis (months), y_2; and status, y_3, where 0 = Dead, 1 = Still alive. (No reference will be made to y_2 and y_3 in this book.)

To begin the analysis **Stat > Regression > Binary Logistic Regression...** was used to fit a model involving all six of the candidate predictor variables recorded prior to treatment. Here **Response:** y1 was specified on checking **Response in response/frequency format** as the raw data are in raw binary format, and **Model:** x1 x2 x3 x4 x5 x6 was entered.

Key Session window output is displayed in Panel 10.17. The Response Information indicates that 24 of the 51 patients in the sample responded to the treatment. The test of the null hypothesis that all slopes are zero,

$$H_0:\beta_1=\beta_2=\beta_3=\beta_4=\beta_5=\beta_6=0,$$

yields P-value 0.000, to three decimal places. Thus the null hypothesis would be rejected in favour of the alternative hypothesis that at least one of the βs is nonzero. None of the goodness-of-fit tests has a small P-value, so there is no evidence of lack of fit. However, scrutiny of the P-values in the Logistic Regression Table suggests that x_2, x_3 and x_5 do not contribute to the model. Thus a second model, involving only x_1, x_4 and x_6, was fitted. In addition, under **Graphs...** the two diagnostic plots of **Delta chi-square versus probability** and **Delta chi-square versus leverage** were selected. An explanation of the background to these plots



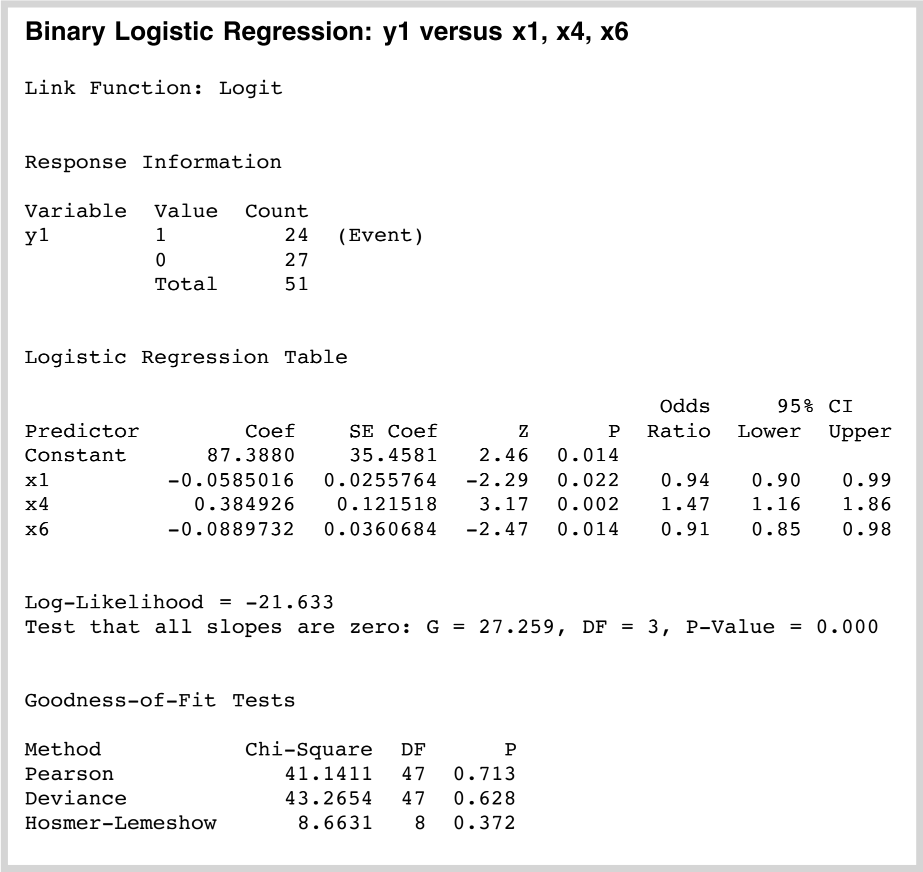
Binary Logistic Regression: y1 versus x1, x2, x3, x4, x5, x6

```
Link Function: Logit


Response Information

Variable  Value  Count
y1        1         24  (Event)
          0         27
          Total     51


Logistic Regression Table

                                                         Odds     95% CI
Predictor        Coef    SE Coef       Z      P  Ratio  Lower  Upper
Constant      98.5236    40.8532    2.41  0.016
x1         -0.0602925  0.0272871   -2.21  0.027   0.94   0.89   0.99
x2         -0.0047997  0.0410745   -0.12  0.907   1.00   0.92   1.08
x3          0.0362132  0.0393374    0.92  0.357   1.04   0.96   1.12
x4           0.398447   0.132773    3.00  0.003   1.49   1.15   1.93
x5          0.0134344  0.0578199    0.23  0.816   1.01   0.90   1.14
x6          -0.102229  0.0418088   -2.45  0.014   0.90   0.83   0.98


Log-Likelihood = -20.030
Test that all slopes are zero: G = 30.465, DF = 6, P-Value = 0.000


Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             40.3923  44  0.627
Deviance            40.0599  44  0.641
Hosmer-Lemeshow      5.3804   8  0.716
```

Panel 10.17 Initial logistic regression analysis of leukaemia data.

is beyond the scope of this book, but values of delta chi-square of around 4 or higher flag the possible presence of unusual observations in the data, observations that merit further scrutiny.

The Logistic Regression Table for the second model is shown in Panel 10.18. The estimated value of the coefficient β_1 of age (x_1) in the model is -0.0585, and the corresponding odds ratio is 0.94. Consider two patients similar in all respects except that one is 1 year older than the other. The model predicts that the odds in favour of the older patient responding to the treatment are 94% of the odds in favour of the younger patient responding to the treatment. The estimated value of the coefficient β_4 of percentage labelling index of the bone marrow leukaemia cells (x_4) in the model is 0.3849, and the corresponding odds ratio is 1.47. Consider two patients similar in all respects except that one has percentage labelling index of the bone marrow leukaemia cells that is 1% higher than the other. The model predicts that the odds in favour of the patient with the higher percentage responding to the treatment are 147% of the odds in favour of the patient with the lower percentage responding to the treatment.



Binary Logistic Regression: y1 versus x1, x4, x6

```
Link Function: Logit


Response Information

Variable  Value  Count
y1        1         24  (Event)
          0         27
          Total     51


Logistic Regression Table

                                                     Odds     95% CI
Predictor         Coef    SE Coef      Z      P  Ratio  Lower  Upper
Constant       87.3880    35.4581   2.46  0.014
x1          -0.0585016  0.0255764  -2.29  0.022   0.94   0.90   0.99
x4            0.384926   0.121518   3.17  0.002   1.47   1.16   1.86
x6          -0.0889732  0.0360684  -2.47  0.014   0.91   0.85   0.98


Log-Likelihood = -21.633
Test that all slopes are zero: G = 27.259, DF = 3, P-Value = 0.000


Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             41.1411  47  0.713
Deviance            43.2654  47  0.628
Hosmer-Lemeshow      8.6631   8  0.372
```

Panel 10.18 Logistic regression table for second model.

The plot of delta chi-square versus leverage is displayed in Figure 10.24. Brushing (introduced in Chapter 3) was used to identify the row numbers, which match the patient numbers, for those observations yielding values of delta chi-square in excess of 4. It certainly suggests that patient number 47 is unusual in some respect, and possibly also patients 48 and 50.

Consider again the data in Table 10.5 in the previous section presented in the form displayed in the worksheet in Figure 10.25. Also shown is the dialog required to analyse the data via binary logistic regression. Note that Supplier is specified both for **Model:** and **Factors (optional):**. The relevant Session window output is shown in Panel 10.19. Note that it gives the odds ratio (B to A) as 3.78, as obtained in the previous section, and that in addition it gives the 95% confidence interval (1.02, 13.95) for this ratio. An odds ratio of 1 corresponds to bearing wear being independent of supplier. The fact that the confidence interval does not include 1 provides evidence, at the 5% level of significance, that bearing wear is dependent on supplier. Since all values in the interval exceed 1 the conclusion from the data is that the odds of wear being present in the bearings are significantly greater for supplier B.

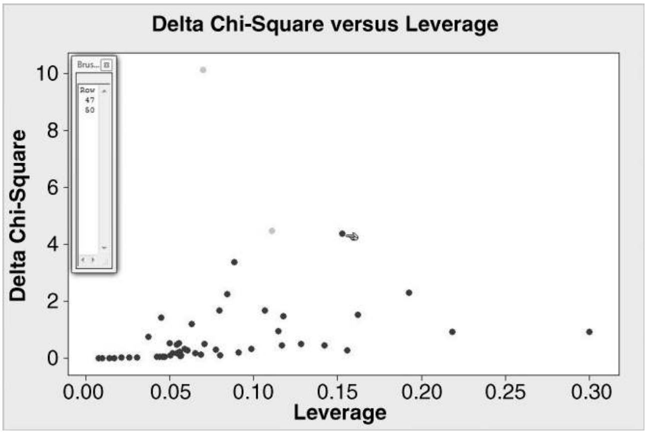
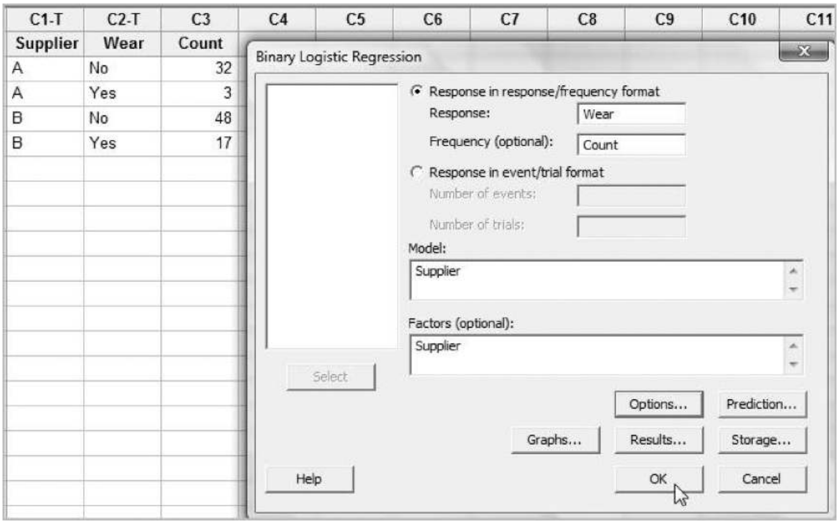


Figure 10.24 Delta chi-square versus leverage plot.



C1-T	C2-T	C3
Supplier	Wear	Count
A	No	32
A	Yes	3
B	No	48
B	Yes	17

Binary Logistic Regression

● Response in response/frequency format
Response: Wear
Frequency (optional): Count
○ Response in event/trial format
Number of events:
Number of trials:
Model: Supplier
Factors (optional): Supplier

Figure 10.25 2 × 2 contingency table set up for analysis using binary logistic regression.

```
Logistic Regression Table

                                                       Odds     95% CI
Predictor       Coef   SE Coef       Z       P  Ratio  Lower  Upper
Constant   -2.36712  0.603807   -3.92   0.000
Supplier
 B          1.32914  0.666513    1.99   0.046    3.78   1.02  13.95
```

Panel 10.19 Logistic regression table for 2 × 2 contingency table.

Minitab also provides ordinal logistic regression and nominal logistic regression. If patients in a hospital classified the level of pain experienced during recovery from a knee replacement operation as being mild, moderate or severe then ordinal logistic regression would be appropriate in an investigation of the relationship between pain experienced and predictors such as gender and age. This is the case as there is a natural ordering in the response – moderate represents a greater level of pain than mild and, in turn, severe represents a greater level of pain than moderate. In addition to predictors, factors such as type of replacement joint may be introduced into the modelling. Nominal logistic regression would be appropriate where there is no natural ordering in possible values for the response. An investigation of preferred methods of payment for supermarket customers, with possible values credit card, bank card, cheque and cash, in relation to gender, age and disposable income could be undertaken using nominal logistic regression. Minitab Help provides examples.

Logistic regression is one example of a generalized linear model. Montgomery (2005a, p. 563) provides further details and examples and makes the following comments. 'Generalized linear models have found extensive application in biomedical and pharmaceutical research and development. As more software packages incorporate this capability, it will find widespread application in the general industrial research and development environment.'

10.5 Exercises and follow-up activities

1. Table 10.8 gives the mass (tonnes) and fuel usage (kilometres/litre) for a sample of 10 vehicles.

   (a) Obtain the least squares regression of y on x and store the residuals and fitted values.

   (b) Give an interpretation of the slope of the regression line.

   (c) Give an interpretation of the value of R^2.

   (d) Check the values of some fitted values and residuals.

   (e) Perform diagnostic checks of the model using residual plots.

2. In an investigation of the shelf life of a cereal, data were obtained on shelf time, x (days), and percentage moisture content, y. The data are given in Table 10.9.

   (a) Investigate the regression of y on x.

Table 10.8 Mass and fuel usage for a sample of vehicles.

Mass x	1.27	1.68	1.63	1.45	1.86	1.18	1.63	1.54	1.72	1.22
Fuel Usage y	6.1	5.3	5.5	5.8	5.2	6.3	5.6	5.5	5.5	6.0

Table 10.9 Moisture content and shelf time.

x	0	3	6	8	10	13	16	20	24	27	30	34	37	41
y	2.4	2.6	2.7	2.8	3.0	3.0	3.1	2.7	3.4	3.6	3.7	3.9	4.0	4.5

(b) Obtain a 95% prediction interval for the moisture content of an individual box of the cereal that has been stored on a shelf for 30 days.

Consumer testing indicated that the cereal is unacceptably soggy when the moisture content is greater than 4.0. On the basis of the prediction interval you have calculated, would you recommend to a supermarket manager that he continue to stipulate a shelf life of 30 days for this brand of cereal? Does analysis of the data following removal of the unusual data value alter your conclusion?

3. In the manufacture of glass bottles gobs of molten glass are poured from the furnace into the moulds in which the containers are formed by the action of compressed air. The gob temperature is of major importance and the manufacturer was interested in being able to predict gob temperature from the temperature obtained from a sensor located in the fore-hearth of the furnace. An experiment was conducted from which a series of 35 values of gob temperature (y) and fore-hearth temperature (x) were obtained. The data are available in the worksheet Gob.MTW and reproduced by permission of Ardagh Glass Ltd., Barnsley.

   (a) Carry out a regression analysis of y on x with diagnostic checks of the residuals.

   (b) Display the data, with fitted line and 95% prediction interval curves.

   (c) Obtain a 99% prediction interval for gob temperature when fore-hearth temperature is 1165 and state, with justification, whether or not you would advise the furnace supervisor to pour glass for a container requiring a target gob temperature of 1150 when fore-hearth temperature is displayed as 1165 on the furnace control panel.

   (d) State the value of fore-hearth temperature that would yield the narrowest prediction interval for gob temperature.

4. Table 10.10 gives systolic blood pressure (SBP, y) measured in millimetres of mercury and the age (x) measured in years for a sample of women considered to be in good health.

   Obtain the regression line of y on x and show that the slope differs significantly from zero.

   (a) Plot residuals and comment on the adequacy of the model.

   (b) Obtain 95% prediction intervals for the systolic blood pressure of women aged 50 years and 20 years, respectively.

5. During the development of a biocide for use in hospitals, a microbiologist carried out an experiment using a trial solution. Twelve beakers were set up, each containing 50 ml of a nutrient broth with microbial spores in suspension. At the start of the experiment a

Table 10.10 SBP and age for a sample of women.

Age (x)	71	53	40	42	47	49	74	43	50	49	67	37
SBP (y)	158	139	126	131	128	141	167	116	128	126	148	121

Table 10.11 Data from biocide experiment.

Time (minutes)	Spore count (organisms per ml)
10	16 518
20	15 177
30	10 084
40	8 533
50	8 823
60	6 690
70	6 042
80	4 890
90	4 065
100	3 166
110	3 012
120	1 852

fixed amount of the biocide was added simultaneously to each beaker. At 10-minute intervals thereafter, a beaker was selected at random and the spore count determined by a method that meant that the beaker could no longer be part of the experiment. The data in Table 10.11 were obtained and are provided in Biocide.MTW.

(a) Fit a least squares regression line of spore count on time.

(b) Explain how the plot of residuals confirms the poor fit of the model.
In order to improve the model she considered the equation

$$N = N_0 e^{-kt},$$

where N represents the spore count and t the time.

(c) Express $\log_e N$ as a linear function of t and obtain the least squares regression of $\log_e N$ on t.

(d) Perform diagnostic checks of the revised model.

(e) Use the revised model to estimate the decimal reduction time D, i.e. the time predicted by the model for the biocide to reduce the number of spores to one-tenth of its initial value.

6. The Minitab worksheet EXH_Regr.MTW, available in the Minitab Sample Data folder supplied with the software, contains in columns C9 and C10 values of energy consumption (y) and setting (x) for a type of machine. Use **Stat > Regression > Fitted Line Plot...** to fit a quadratic model to the data. Is the model improved by making a logarithmic transformation of the response? Create a column containing the values of x^2, fit the models using **Stat > Regression > Regression...** and carry out diagnostic checks of the residuals.

7. The Minitab worksheet Trees.MTW, available in the Minitab Sample Data folder supplied with the software, gives diameter, height and volume for a sample of 31 black

cherry trees from Allegheny National Forest in the USA. Diameter (feet) was measured 4.5 feet above ground. Height was measured in feet and volume in cubic feet.

Suppose that the forest management team wish to develop a model to enable prediction of timber production, i.e. a model that may be used to predict volume, which is difficult to measure, from diameter and height measurements. Explore models involving the three variables and also the three variables after logarithmic transformation of all three. Which model would you recommend?

8. The Minitab worksheet EXH_Regr.MTW (see Exercise 6 above) contains in columns C3 to C8 data, from a project on solar thermal energy, on total heat flux from homes. It is desirable to ascertain whether total heat flux can be predicted from insolation, the position of the focal points in the east, south, and north directions and from time.

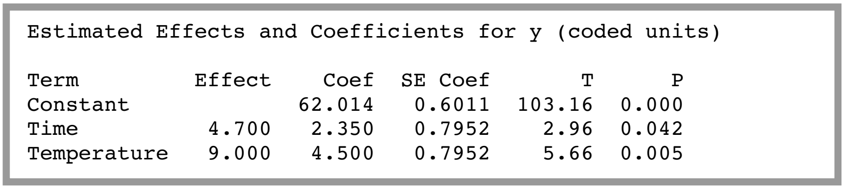
   (i) Use best subsets regression to select a potential regression model to predict heat flux from a subset of the five candidate predictor variables.

   (ii) Fit your selected model and carry out diagnostic checks of the residuals.

9. The worksheet BHH1.MTW contains data from Phase 1 of an illustration of response surface methodology given by Box *et al.* (1978, pp. 514–525). (All the data in this exercise are reproduced by permission of John Wiley & Sons, Inc., New York.) The design used was a single replication of a 2^2 factorial with the addition of three centre points. The response was yield (g) from a laboratory-scale chemical production process and the factors were time (low 70 minutes, high 80 minutes) and temperature (low 127.5 °C, high 132.5 °C). The centre point (75, 130) represented current operating factor levels prior to the experimentation.

   (i) Verify, from scrutiny of the ANOVA from **Stat > DOE > Factorial > Analyze Factorial Design...**, that there is no evidence of curvature and no evidence of interaction.

   (ii) In order to create a contour plot, repeat the analysis having, under **Terms...**, removed the interaction term from the model and unchecked **Include center points in the model**. You should obtain the Session window output displayed in Panel 10.20 – providing evidence that both time and temperature appear to influence yield and indicating that the equation of the fitted model, in terms of the **coded** units, is $y = 62.0 + 2.35x_1 + 4.50x_2$ (with the constant and the coefficients quoted to three significant figures).



```
Estimated Effects and Coefficients for y (coded units)

Term          Effect     Coef  SE Coef       T      P
Constant               62.014   0.6011  103.16  0.000
Time           4.700    2.350   0.7952    2.96  0.042
Temperature    9.000    4.500   0.7952    5.66  0.005
```

Panel 10.20 Part of Session window output.

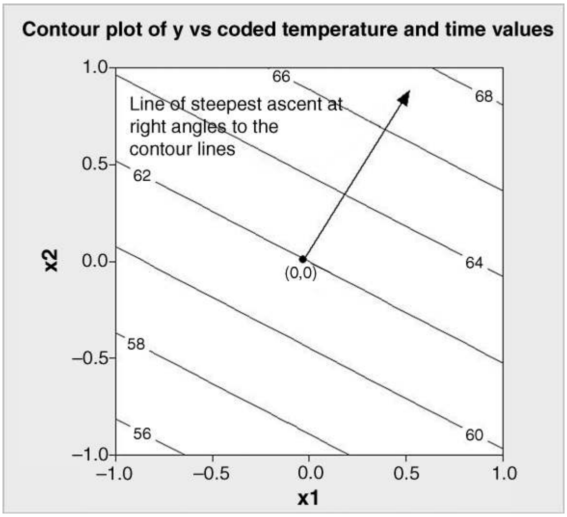


Figure 10.26 Line of steepest ascent.

(iii) Use **Stat > DOE > Factorial > Contour/Surface Plots...** to create a contour plot. Choose **Contour Plot** and under **Setup...** check **Display plots using: Coded units**. Under **Contours...** insert **Values:** 56 58 60 62 64 66 68 as **Contour Levels**. Under **Data Display** select **Contour lines** and **Symbols at design points**. This will create the plot in Figure 10.26, but without the annotation.The point (0, 0) in coded units represents the centre point of the design.

The line of steepest ascent is the line through (0,0) in the above plot at right angles to the contours in the direction in which yield increases. Some mathematics is presented in Box 10.3 that enables FLCs on the line of steepest ascent to be calculated.

(iv) Set up, in a Minitab worksheet, columns named x1, x2, Time and Temperature and assign values 0, 1, 2, 3, 4 and 5 to x1. Use **Calc > Calculator** to compute x2 from the equation of the line of steepest ascent and to decode the values in x1 and x2 to give the time and temperature values. The values you should obtain are shown in Table 10.12.

The first row corresponds to the centre point (a useful cross-check on the computations) and the average yield for the three runs at this FLC is readily verified to be 62.3. The process team decided to carry out a further run with the FLC (80, 135). This gave the encouraging yield of 73.3, so the next run was carried out with the combination (100, 154). The result was a disappointing yield of 58.2. Then use of (90, 144) gave yield 86.8, representing substantial improvement.

(v) Add the yield column to your worksheet and display the Phase 2 data. (The data are available in BHH2.MTW)

As a result of the knowledge gained from Phase 2, the Phase 3 experimentation involved a central composite design centred on time 90 minutes and temperature 145 °C. The data are available in worksheet BHH3.MTW.

In terms of the coded units, the contour lines are specified by the model equation

$$y = 62.0 + 2.35x_1 + 4.50x_2.$$

In particular, the contour for yield 62.0 is specified by

$$62.0 = 62.0 + 2.35x_1 + 4.50x_2.$$

This equation may be written as

$$x_2 = -\frac{2.35}{4.50}x_1.$$

To obtain the equation of the line of steepest ascent it is necessary to change the sign on the right-hand side and to invert the ratio which forms the coefficient of x_1. Thus the equation of the line of steepest ascent is given by

$$x_2 = +\frac{4.50}{2.35}x_1 = 1.91x_1.$$

The coding equations are

$$x_1 = \frac{\text{Time} - 75}{5} \quad \text{and} \quad x_2 = \frac{\text{Temperature} - 130}{2.5}.$$

They may be rearranged to give

$$\text{Time} = 5x_1 + 75 \quad \text{and} \quad \text{Temperature} = 2.5x_2 + 130.$$

Box 10.3 Determination of the line of steepest ascent.

Table 10.12 Data from Phase 2 exploration along line of steepest ascent.

Phase 2 Experimentation on the line of steepest ascent				
x_1	x_2	Time	Temperature	Yield
0	0.00	75	130.000	62.3
1	1.91	80	134.775	**73.3**
2	3.82	85	139.550	
3	5.73	90	144.325	**86.8**
4	7.64	95	149.100	
5	9.55	100	153.875	**58.2**

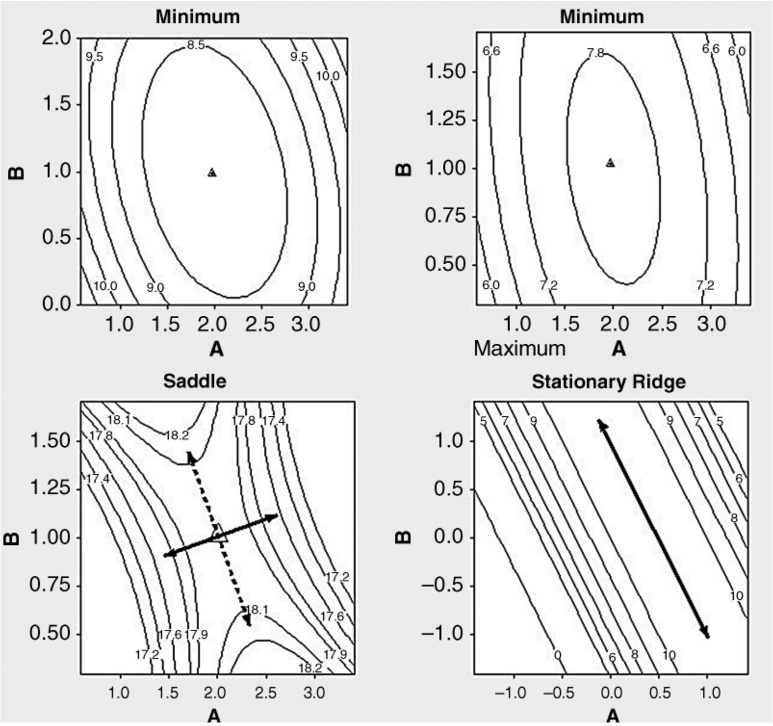


Figure 10.27 Examples of response surfaces.

(vi) Analyse the data and display the response surface as a contour plot, with time on the horizontal axis and temperature on the vertical axis.

Box *et al.* conclude the illustration by stating that the surface ‘represents an oblique rising ridge with yields increasing from the lower right to the top left corner of the diagram, that is, yield increases as we progressively *increase* temperature and simultaneously *reduce* reaction time’. Were it desirable to increase yield still further then ‘subsequent experimentation would have followed and further explored this rising ridge’.

10. The contour plots in Figure 10.27 display four types of response surface encountered in modelling process behaviour. In the cases of the saddle and the stationary ridge, what recommendations would you make, given that the goal was to maximize the responses?

11. Open the Pulse.MTW worksheet provided in the Minitab Sample Data folder. Use **Help** > **? Help** and the **Search** tab to perform a search for Pulse. Double clicking on PULSE.MTW then reveals a description of and key to the data set.

   (i) Form a 2 × 2 contingency table that classifies the sample by smoking habit and gender and test for association using both the chi-square and Fisher’s exact test.

   (ii) Form a contingency table that classifies the sample by smoking habit and level of activity and test for association. Note that a 2 × 4 contingency table results for which the chi-square analysis involves three expected frequencies less than 5, so

Table 10.13 Contingency table classifying defects in wafers.

	Type of defect			
Shift	A	B	C	D
1	15	21	45	13
2	26	31	34	5
3	33	17	49	20

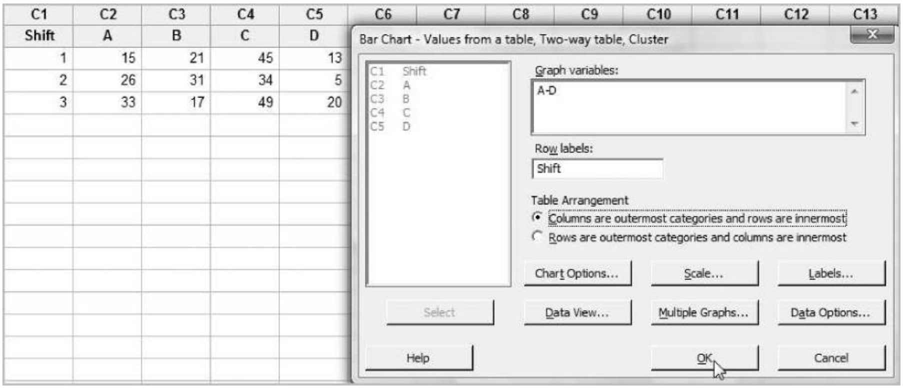
Minitab issues a warning in the Session window output that the chi-square analysis is probably invalid.

Code activity levels 0 and 1 as low, 2 as moderate and 3 as high, and form and analyse the corresponding contingency table.

12. NIST/SEMATECH (2005) gives an industrial example in which 309 wafer defects were recorded and the defects were classified according to type (A,B,C, or D) and according to the production shift at time of manufacture of the wafer (1, 2, or 3). The data are given in Table 10.13.

 Emphasis has been placed on the display of data in this book, so prior to formal analysis the reader is invited to set up the contingency table as shown in Figure 10.28 and use **Graph** > **Bar Chart**. . . with **Bars represent:** Values from a table, clicking on **Cluster** under **Two-way table**, clicking on **OK**, then selecting **Graph variables:** A-D, **Row labels:** Shift and accepting defaults otherwise to display the data. Repeat with the second option for Table Arrangement. Do you consider one display to be more informative than the other? Carry out a formal test for association.

13. Cox (1970, p. 86) provides data on the duration of heating, *T*, for ingots and on the numbers not ready for rolling, *R*, summarized in Table 10.14 and reproduced by permission of Taylor & Francis Group. Model the relationship between readiness for rolling and duration of heating. On a scatterplot of the observed probability of readiness for rolling versus duration of heating superimpose the curve giving the fitted probability of readiness for rolling as a function of duration of heating.



C1	C2	C3	C4	C5
Shift	A	B	C	D
1	15	21	45	13
2	26	31	34	5
3	33	17	49	20

Bar Chart - Values from a table, Two-way table, Cluster

C1 Shift
C2 A
C3 B
C4 C
C5 D

Graph variables: A-D

Row labels: Shift

Table Arrangement
● Columns are outermost categories and rows are innermost
○ Rows are outermost categories and columns are innermost

Figure 10.28 Contingency classifying defects in wafers.

Table 10.14 Data on ingot readiness for rolling.

Duration of heating T	No. not ready for rolling R	No. tested N
7	0	55
14	2	157
27	7	159
51	3	16

14. The worksheet ChemProc.MTW gives data on the reaction time and catalyst used in a chemical production process and an indication of whether or not the final product was of prime quality. Use binary logistic regression to fit a model involving both reaction time and catalyst. Catalyst has to be specified as a factor as well as being included in the model. Store the event probabilities (for all observations) and display them plotted against reaction time using **Graph** > **Scatterplot...**, selecting **With Connect and Groups** and specifying Catalyst under **Categorical variables for grouping**. State recommendations that can be made with regard to running the process.

15. Obtain the odds ratio with a 95% confidence interval for each of the scenarios Tables 10.15 and 10.16. Comment.

Table 10.15 Scenario 1.

Classification of sample of scanners by bearing supplier and bearing state		Bearing state		
		Sound	Worn	
Bearing supplier	P	5	5	10
	Q	7	3	10
		12	8	20

Table 10.16 Scenario 2.

Classification of sample of scanners by bearing supplier and bearing state		Bearing state		
		Sound	Worn	
Bearing supplier	P	50	50	100
	Q	70	30	100
		120	80	200