CHAPTER 10

Correlations Between Two Variables

Up to this point we have discussed situations involving one or two measured variables at fixed operating conditions. We now turn our attention to describing in quantitative terms the nature of changes in measured variables as operating conditions change. In this section a single numerical measure of the relationship between two random variables is developed using measurements of those two variables over a range of operating conditions. In subsequent sections more complete functional relationships will be described between one random variable (the *output* or *response* variable) and one or more operating variables whose values are altered in a controlled manner.

The following example will be used in this section and following sections. Measurements were made on a filtration system at a number of different operating conditions and some of the results are shown in table 10.1.

Table 10.1 - Filtration Data

Feed flow rate (litre/hour)	Waste solids removed (%)
0.13	24.3
0.23	19.6
0.60	12.9
0.13	25.2
0.88	4.2
0.39	18.4
0.23	20.3
0.75	6.0
0.50	13.3
0.88	5.9

An informative first step in any data analysis operation is to plot the data.

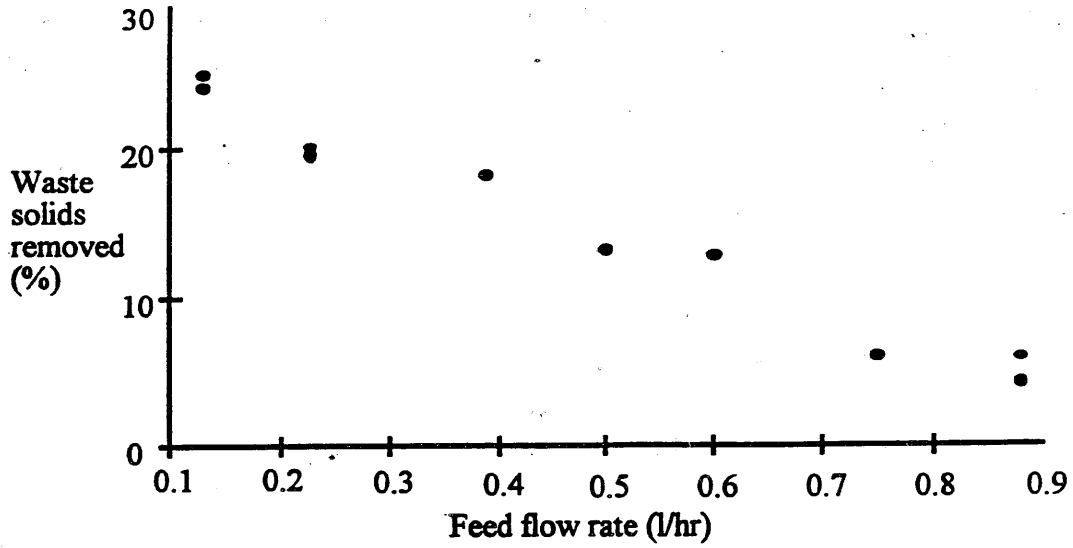

Figure 10.1 - The filtration data.

From figure 10.1 it appears that a straight line would describe the relationship between these two variables very well over the range of conditions spanned by these data.

If feed flow rate and per cent waste solids removed are regarded as random variables, X and Y, one measure of the systematic relationship between them is the covariance $\text{cov}(X,Y)$ defined as

$$\text{cov}(X,Y) = E\{(X-\mu_X)(Y-\mu_Y)\} \tag{10.1}$$

where μ_X and μ_Y are the population means of X and Y respectively. Although the expectation operator E over two random variables simultaneously has not been defined in these notes it can be regarded in (10.1) as the population mean of all possible products of departures of X from its mean μ_X, and Y from its mean μ_Y.

$\text{cov}(X,Y)$ is not a particularly useful measure of systematic association between X and Y because it depends upon the units of the two random variables. A more useful measure that expresses systematic association between two random variables in dimensionless form is the correlation $\rho(X,Y)$ defined as

$$\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{10.2}$$

where σ_X and σ_Y are the population standard deviations of X and Y respectively. The value of $\rho(X,Y)$ can range from -1, indicating an exact linear relationship with negative

Why would this value not equal 1 all the time ? Why does it only apply to lines ?

slope between X and Y, to +1, indicating an exact linear relationship with positive slope between X and Y. A zero value for $\rho(X,Y)$ indicates that no systematic linear relationship exists between X and Y.

An estimator of $\rho(X,Y)$ from n pairs of measurements $(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$ is

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\left(\sum_{i=1}^{n}(x_i-\bar{x})^2\right)\left(\sum_{i=1}^{n}(y_i-\bar{y})^2\right)}} \tag{10.3}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of $(x_1,\cdots,x_n)$ and $(y_1,\cdots,y_n)$ respectively. An alternative expression for r that is algebraically identical to (10.3) and simpler to calculate is

$$r=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)\left(\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right)}}, \tag{10.4}$$

but it suffers from numerical inaccuracy due to roundoff error when the x data and/or the y data are all large values. Values for r can range from -1 to 1.

For the data in table 10.1,

$$n=10 \qquad \sum_{i=1}^{10}x_iy_i=50.57$$

$$\bar{x}=0.472 \qquad \bar{y}=15.01$$

$$\sum_{i=1}^{10}x_i^2=3.01 \qquad \sum_{i=1}^{10}y_i^2=2792$$

so that

$$r=-0.988.$$

The closeness of this value of r to -1 suggests a very strong linear relationship with negative slope between the feed flow rate and waste solids removed. This result was evident from the plot in figure 10.1.

Care must be exercised in interpreting a zero or nearly zero value for r. This statistic is a measure of systematic *linear* association between two variables. Data such as those shown in figure 10.2 would produce a value of r very close to zero indicating essentially no linear relationship between the two variables. In this example, there is clearly a very strong quadratic relationship between Y and X even though $\rho(X,Y)$ may be zero. A plot of the data would guard against misinterpretation of the correlation statistic r. In general, increased scatter of points (x_i, y_i) about the best straight line relating X and Y reduces the value of r (and ρ).

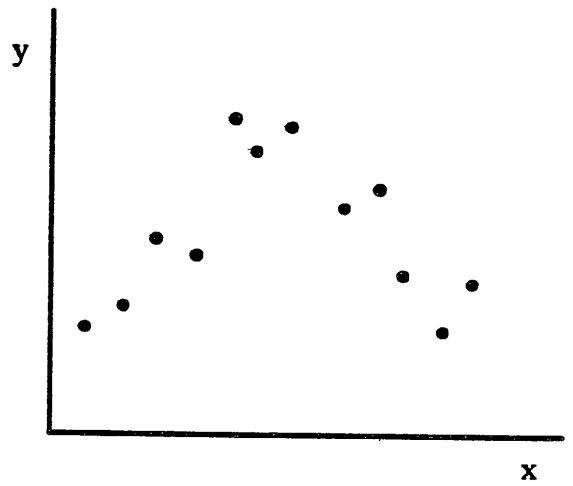

Figure 10.2 - One case where $r = 0$.

The probability density function for r is an awkward algebraic form involving $\rho(X,Y)$ and consequently it cannot be used easily to infer plausible values of ρ from the statistic r. However, R.A. Fisher has shown that the quantity

$$\tanh^{-1} r = \tfrac{1}{2} \ln\left(\frac{1+r}{1-r}\right) \qquad (10.5)$$

is approximately normally distributed with a mean value

$$\tfrac{1}{2} \ln\left(\frac{1+\rho}{1-\rho}\right)$$

and variance

$$\frac{1}{n-3}.$$

Thus by using a normal p.d.f. table such as table 1, an approximate $100(1-\alpha)$ per cent confidence interval for $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ can be calculated as

$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \pm z_{\alpha/2}\left(\frac{1}{\sqrt{n-3}}\right) \tag{10.6}$$

where $z_{\alpha/2}$ is the abscissa value of the unit normal p.d.f. that leaves an upper tail area of $\alpha/2$. If the limits of the interval (10.6) are denoted as L and U, then the corresponding interval for ρ can be calculated directly as

$$\left[\frac{e^{2L}-1}{e^{2L}+1} \text{ to } \frac{e^{2U}-1}{e^{2U}+1}\right] \tag{10.7}$$

For the filtration example ($r=-0.987$, $n=10$) the approximate 95 per cent confidence interval for $\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ obtained using (10.6) is

$$-2.56 \pm 1.96(0.378) = -3.30 \text{ to } -1.82.$$

The resulting 95 per cent confidence interval for ρ using (10.7) is then [-997 to -0.949].

In some applications it may be of interest to determine whether $\rho = 0$ is a plausible value, that is, whether X and Y have any systematic linear relationship. A confidence interval for ρ can provide the answer. Alternatively, a more exact test can be made because if $\rho = 0$ then

$$\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

has a t_{n-2} p.d.f. Thus the null hypothesis $\rho = 0$ can be tested at a level of significance α by referring the calculated value $\frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}}$ to $t_{n-2,\alpha/2}$ for a one-tailed test.

One final word of caution is necessary in interpreting values of the statistic r. Even if a strong systematic association between X and Y is indicated, it should not necessarily be concluded that X and Y have a cause and effect relationship. Huff [1] mentions several examples of "nonsense correlations" (cases where r is very close to 1 or to -1, yet there is no basis for a cause and effect relationship between the two variables in question). One of the best known of these examples is the strong positive correlation that was found between the birth rate in Munich in the early years of this century and the stork population in that city.

REFERENCES

[1] Huff, D., 1954, How to Lie with Statistics. Norton, New York.

objective: model behaviour in mathematical terms (quantify an assumed causal relationship)

CHAPTER 11 *We want our model to be linear with respect to the parameters.

Fitting a Straight Line Relationship Between Two Variables

When two variables are believed to be causally related, it is often of interest to describe their relationship by a mathematical model. An appropriate mechanistic choice of model form may be suggested by physical theory. Sometimes the complexity of the mechanistic model form is such that an empirical model, perhaps a polynomial of low degree, is preferred. This is often the case when a relationship between variables is required only over a relatively restricted operating region. In other cases an empirical model may be the only choice because no relevant theoretical relationship has been developed. An appropriate form of empirical model can often be identified from a plot of data for one variable against the other.

The simplest model form relating two variables is a straight line,

$$E(Y) = \beta_0 + \beta_1 x \tag{11.1}$$

where $E(Y)$ denotes the expected value of a random variable Y.

The assumptions associated with the form (11.1) are as follows. First the "dependent" or "response" variable, Y, is a continuous random variable with probability density function $p(Y)$, so that

$$E(Y) = \int_{-\infty}^{\infty} Y\, p(Y)\, dY$$

The "independent" variable x is NOT a random variable. It's values are regarded as exact, i.e. error free. In practice, the interpretation of this assumption is that any fluctuation associated with a value of x has a much smaller influence on Y than the fluctuation associated with Y itself. The postulated straight line relationship between $E(Y)$ and x is completely characterized by the two *parameters* β_0 and β_1 whose values are unknown. Estimates of β_0 and β_1 can be obtained by fitting model form (11.1) to data for Y and x.

A measured value of Y is denoted by y and (11.1) can be expressed equivalently as

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{11.2}$$

where ε is a random error term having probability density function $p(\varepsilon)$. If the straight line model (11.1) does describe the relationship between $E(Y)$ and x adequately, then the mean value of $p(\varepsilon)$ will be zero. Additional assumptions concerning ε will be discussed later. For now it is sufficient to regard ε as representing the departure of a measured response value y from its expected value which is given by the straight line form (11.1).

We now proceed to determine the particular straight line of form (11.1), i.e. the particular values of β_0 and β_1 that provides closest agreement with the data. It is important to recognize that departures of data points from the best fitting line can arise from two sources. One is experimental error, the lack of exact reproducibility in virtually all scientific measurements. The other is inadequacy of the selected model form. For example, if two variables were related by an exponential decay model form then, even for very small experimental error, departures of data points from a best fitting straight line would occur. In a later section we shall describe procedures for distinguishing between deviation which can be attributed solely to experimental error and deviation which is caused by an inappropriate model in addition to experimental error.

In model form (11.2) the values of the parameters β_0 and β_1 and the random error ε are unknown. However, for any selected values, $\tilde{\beta}_0$ and $\tilde{\beta}_1$, the error, ε, at any data point u can be estimated by the residual e_u defined as

$$e_u = y_u - (\tilde{\beta}_0 + \tilde{\beta}_1 x_u) \tag{11.3}$$

where y_u = the measured response value for data point u,
x_u = the independent variable value for data point u,
$\tilde{\beta}_0$ and $\tilde{\beta}_1$ = selected values for the parameters.

Several criteria can be used to determine the best fitting straight line, but the most common, and the most appropriate under the assumptions to be specified shortly, is the least squares criterion. If n data points are used to fit model (11.2) then the values of β_0 and β_1 for which $\sum_{u=1}^{n} e_u^2$ is minimized are called the least squares estimates of the parameters and are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$.

The distinction between ε and e is that ε involves the true but unknown values of the parameters β_0 and β_1 whereas e involves estimates of those parameters. Even when the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are used, the residuals e are not the random errors ε

since the residuals in this case refer only to the data being used for the fit and not to other sets of data which could be obtained.

As stated above, least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the parameters in (11.2) using data $(x_1, y_1), \cdots, (x_n, y_n)$ are those values of β_0 and β_1 for which

$$\sum_{u=1}^{n} e_u^2 = \sum_{u=1}^{n} \{y_u - (\beta_0 + \beta_1 x_u)\}^2 \tag{11.4}$$

is minimized. If no constraints are imposed on β_0 and β_1, then the minimum occurs at those values of β_0 and β_1 for which

$$\frac{\partial \left\{ \sum_{u=1}^{n} e_u^2 \right\}}{\partial \beta_0} = 0 \tag{11.5}$$

and

$$\frac{\partial \left\{ \sum_{u=1}^{n} e_u^2 \right\}}{\partial \beta_1} = 0 \tag{11.6}$$

Equations (11. 5) and (11. 6) can be expressed as

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{u=1}^{n} x_u = \sum_{u=1}^{n} y_u \tag{11.7}$$

$$\hat{\beta}_0 \sum_{u=1}^{n} x_u + \hat{\beta}_1 \sum_{u=1}^{n} x_u^2 = \sum_{u=1}^{n} x_u y_u \tag{11.8}$$

and their unique solution is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11.9}$$

$$\hat{\beta}_1 = \frac{\sum_{u=1}^{n} x_u y_u - n\bar{x}\bar{y}}{\sum_{u=1}^{n} x_u^2 - n\bar{x}^2} \tag{11.10}$$

where $\bar{x} = \frac{\sum_{u=1}^{n} x_u}{n}$ and $\bar{y} = \frac{\sum_{u=1}^{n} y_u}{n}$. It can be confirmed [1, pp. 215-219] that this solution is in fact a minimum.

From (11.9) and (11.10) it can be seen that the value $\hat{\beta}_0$ depends upon the value $\hat{\beta}_1$. When this dependence occurs the two parameter estimates are said to be *correlated* with one another. More will be said about this a little later in this section.

Using the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ from (11.9) and (11.10), the least squares straight line can be expressed as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{11.11}$$

It is easily verified that this line passes through the point $(\bar{x}, \bar{y})$, the "centroid" of the data. If, however, the proposed straight line model is constrained to pass through a specified point $(x, y) = (a, b)$, then the fitted line need no longer pass through $(\bar{x}, \bar{y})$. The most common example of a constrained straight line model is one that is forced to pass through the origin $(x, y) = (0, 0)$.

To avoid potential problems arising from possible ill-conditioning of the equations (11.7) and (11.8), most computer programs for least squares estimation fit a straight line model of the form

$$y = \gamma_0 + \gamma_1 (x - \bar{x}) + \varepsilon \tag{11.12}$$

where $\bar{x} = \frac{\sum_{u=1}^{n} x_u}{n}$ is the sample mean of the x values to be used. Form (11.12) is algebraically equivalent to form (11.2) with $\beta_0 = \gamma_0 - \gamma_1 \bar{x}$ and $\beta_1 = \gamma_1$. The advantage of form (11.12) can be seen when the partial derivatives of $\sum_{u=1}^{n} e_u^2$ with respect to γ_0 and γ_1 are equated to zero. The resulting least squares estimates are

$$\hat{\gamma}_0 = \bar{y} \tag{11.13}$$

$$\hat{\gamma}_1 = \frac{\sum_{u=1}^{n} x_u y_u - n\bar{x}\bar{y}}{\sum_{u=1}^{n} x_u^2 - n\bar{x}^2} \tag{11.14}$$

and they are not correlated with one another. That is, the least squares estimate $\hat{\gamma}_0$ is independent of the least squares estimate of $\hat{\gamma}_1$. The least squares straight line

$$\hat{y} = \hat{\gamma}_0 + \hat{\gamma}_1(x - \bar{x}) \tag{11.15}$$

is of course identical to that expressed in (11.11).

Example 11.1

Let us consider the filtration data shown in Table 10.1, assuming feed flow rate to be the "error free" independent variable and waste solids removed to be the "error corrupted" response variable. The plot shown in figure 10.1 indicates a straight line to be a most reasonable empirical model form for these data.

From (11.9) and (11.10), least squares estimates of the parameters are

$$\hat{\beta}_1 = \frac{50.566 - 10(0.472)(15.01)}{3.013 - 10(0.472)^2}$$
$$= -25.8$$

and
$$\hat{\beta}_0 = 15.01 - (-25.83)(0.472)$$
$$= 27.2$$

As shown in figure 11.1, the least squares straight line

$$\hat{y} = 27.2 - 25.8x \tag{11.16}$$

is that for which the sum of squares of the vertical distances from each data point to the line is minimized. The use of vertical distances rather than perpendicular or horizontal distances is a consequence of the assumption that all of the "error" in the data is associated with the response variable, waste solids removed.

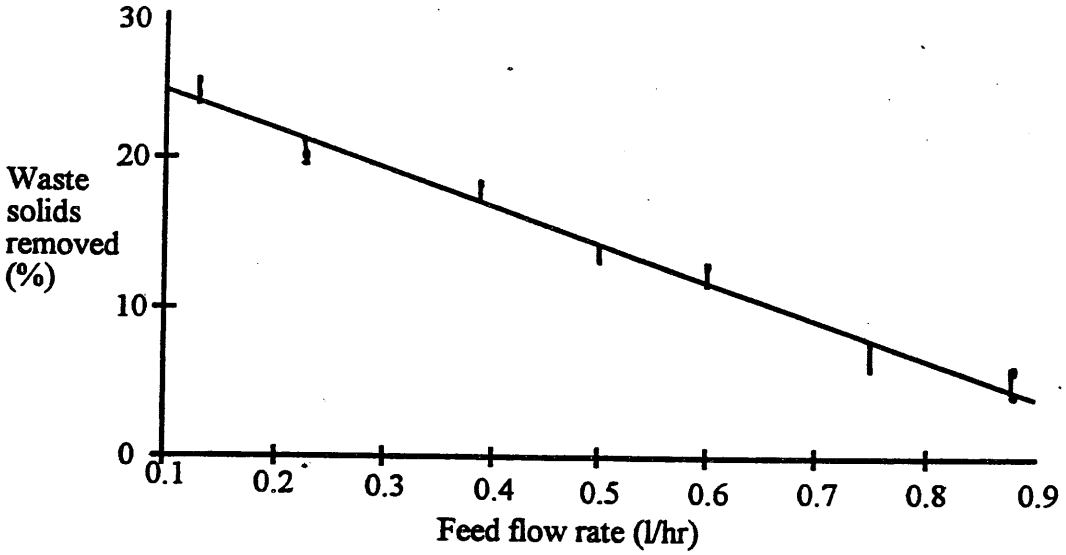

Figure 11.1 - A least squares straight line fit to the filtration data

If, for illustration, the roles of the two variables in this example are reversed so that feed flow rate is the response variable and waste solids removed is the independent variable then the model to be fitted is

$$x = \alpha_0 + \alpha_1 y + \delta \tag{11.17}$$

where α_0 and α_1 are parameters to be estimated and δ is the random error term associated with x. Least squares estimates of α_0 and α_1 are those values that minimize the sum of squares of residuals

$$\sum_{u=1}^{n} \{x_u - (\alpha_0 + \alpha_1 y)\}^2 .$$

Equating partial derivatives with respect to α_0 and α_1 to zero, the least squares estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are given by equations (11.9) and (11.10) with x and y interchanged,

$$\hat{\alpha}_0 = \bar{x} - \hat{\alpha}_1 \bar{y} \tag{11.18}$$

$$\hat{\alpha}_1 = \frac{\sum_{u=1}^{n} y_u x_u - n\bar{y}\bar{x}}{\sum_{u=1}^{n} y_u^2 - n\bar{y}^2} \tag{11.19}$$

For the filtration data, $\hat{\alpha}_0 = 1.036$ and $\hat{\alpha}_1 = -0.0376$ so that this least squares straight line is

$$\hat{x} = 1.04 - 0.038y \tag{11.20}$$

which is *not* the same line as that given by equation (11.16). Since the two fitted lines (11.16) and (11.19) have been obtained by minimizing different functions, they will not in general be the same. Although the two lines in this example lie close to one another because the data are so closely approximated by a straight line, the divergence between two least squares straight lines tends to increase as the scatter of the data points about either of the lines increases. This illustrates the importance of selecting as the response variable that variable to which the error properly belongs.

Least squares analysis of data should not end with the fitting of a proposed model. Tests should be made to determine whether the departures of the data from the fitted model can be attributed to experimental error alone. If not, then the original model form must be modified in light of its inadequacies and the revised model fitted to the data. When a fitted model is judged to be an adequate representation of the data then the precision of the parameter estimates and predicted response values can be evaluated. Before discussing these topics we shall extend the fitting procedure we have described for a straight line model to more general forms of models.

REFERENCES

[1] Beveridge, G.S.G. and Schechter, R.S., 1970: Optimization: Theory and Practice, McGraw-Hill, New York.

CHAPTER 12

Linear and Nonlinear Models

In statistical analysis, a mathematical model is said to be *linear* if it is a *linear function of the parameters to be estimated.* Any linear model can then be expressed in the form

$$E(Y)=\beta_1 x_1+\cdots+\beta_p x_p \tag{12.1}$$

where the x's are functions of the error free operating variables and the β's are parameters to be estimated. Three examples of linear models are

$$E(Y)=\beta_0+\beta_1 x+\beta_2 x^2 \tag{12.2}$$

$$E(Y)=\beta_1\left(\frac{1}{x_1}\right)+\beta_2 x_2^2 \tag{12.3}$$

$$E(Y)=\beta_0+\beta_1 x_1+\beta_2 \ln\left(\frac{x_1}{x_2}\right)+\beta_3 \exp\left(\frac{-x_1 x_2}{x_3}\right). \tag{12.4}$$

Models (12.2 and 12.4) conform to the general linear model form (12 1) if the parameter β_0 is considered as being multiplied by a variable x_0 whose value is always 1. Polynomial models, whether they involve only one operating variable as in (12.2) or several operating variables, are linear models. As illustrated in all three of the above models, the independent variables in a linear model can be nonlinear functions of the operating variables but the parameters must enter the model in linear fashion. That is, each of the partial derivatives,

$$\frac{\partial E(Y)}{\partial \beta_1},\cdots,\frac{\partial E(Y)}{\partial \beta_p}$$

must *not* be functions of *any* of the parameters $\beta_1,\cdots,\beta_p$.

A nonlinear model is one in which one or more of the parameters to be estimated enters the model in nonlinear fashion. An example is

$$E(Y)=\theta_1\left(1-e^{\theta_2 x}\right) \tag{12.5}$$

where x is the error free operating variable and θ_1 and θ_2 are parameters to be estimated. In this case, both $\partial E(Y)/\partial\theta_1$ and $\partial E(Y)/\partial\theta_2$ are functions of the parameters. Notice in model (12.5) that if θ_2 were assigned a "known" numerical value, so that only θ_1 remained to be estimated, the model would be linear.

Another example of a nonlinear model is

$$E(Y) = \theta_1 \exp(\theta_2 x) \tag{12.6}$$

This model can be transformed to linear form by taking logarithms of both sides,

$$\ln E(Y) = \ln \theta_1 + \theta_2 x = \beta_1 + \beta_2 x \tag{12.7}$$

where $\beta_1 = \ln \theta_1$ and $\beta_2 = \theta_2$. However as will be demonstrated in a later section, linearizing transformations can lead to invalid estimation procedures and misleading parameter estimates unless careful consideration is given to the effects of such transformations on the error variance.

Least squares estimates of the parameters in both linear and nonlinear models are found by minimizing the sum of squares of residuals. For nonlinear models this minimization requires some type of search procedure. For linear models the minimum is located directly by solving a set of linear equations. We now consider least squares fitting of linear models to data.

CHAPTER 13

Fitting Linear Models to Data by Least Squares

As indicated in the preceding section, a general linear model form can be written as

$$E(Y) = \beta_1 x_1 + \cdots + \beta_p x_p \tag{13.1}$$

and an intercept parameter β_0 can be accommodated in the manner described previously. It is obvious that at least p data points must be available if unique estimates of p parameters are to be obtained. Suppose that there are n data points $(n > p)$ which can be tabulated as indicated in table 13.1.

Table 13.1 - Collected data

Data Point Number u	Values of Independent Variables x_1	x_2	$\cdots$	x_p	Measured Response Value y
1	x_{11}	x_{12}	$\cdots$	x_{1p}	y_1
2	x_{21}	x_{22}	$\cdots$	x_{2p}	y_2
3	x_{31}	x_{32}	$\cdots$	x_{3p}	y_3
.	.	.		.	.
.	.	.		.	.
.	.	.		.	.
n	x_{n1}	x_{n2}	$\cdots$	x_{np}	y_n

Let us now define the following matrices for this general case. The dimensions of each matrix are typed in parentheses next to the matrix notation.

$$\mathbf{Y}_{(n\times 1)} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \qquad \beta_{(p\times 1)} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\mathbf{X}_{(nxp)} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}$$

For any selected values $\tilde{\beta} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_p)^T$ where the superscript T denotes the transpose of a matrix, a matrix of residuals $\mathbf{e}^T$ can be constructed where

$$\mathbf{e}_{(nx1)} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} y_1 - (\tilde{\beta}_1 x_{11} + \cdots + \tilde{\beta}_p x_{1p}) \\ \vdots \\ y_n - (\tilde{\beta}_1 x_{n1} + \cdots + \tilde{\beta}_p x_{np}) \end{bmatrix}$$

or $\mathbf{e} = \mathbf{Y} - \mathbf{X}\tilde{\beta}$.

Least squares estimates of the parameters β are obtained by minimizing $\mathbf{e}^T\mathbf{e}$, the sum of squares of residuals with respect to β. The set of equations

$$\frac{\partial(\mathbf{e}^T\mathbf{e})}{\partial\beta_1} = 0$$
$$\vdots$$
$$\frac{\partial(\mathbf{e}^T\mathbf{e})}{\partial\beta_p} = 0$$

can be written in matrix form as

$$\mathbf{X}^T\mathbf{X}\beta = \mathbf{X}^T\mathbf{Y} \qquad (13.2)$$

where $\beta = (\hat{\beta}_1, \cdots, \hat{\beta}_p)^T$ are the least squares estimates of the parameters. The solution of equation (13.2) is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \qquad (13.3)$$

where the superscript -1 denotes the inverse of a matrix. The fitted response values for the *n* data points can be expressed as

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{y}_1 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}} \tag{13.4}$$

and the residuals from the least squares fit as

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}. \tag{13.5}$$

Before illustrating this general linear least squares fitting procedure, two observations should be made. First, it should be noticed that the matrix $\mathbf{X}^T\mathbf{X}$ is a symmetric p x p matrix whose elements are sums of squares and cross products of the individual x values. Second, it must be recognized that solution of the matrix equation (13.2) will not be possible if the matrix $\mathbf{X}^T\mathbf{X}$ is singular for then $\mathbf{X}^T\mathbf{X}$ will have no inverse. Singularity in $\mathbf{X}^T\mathbf{X}$ occurs if all of the elements of any column of the matrix $\mathbf{X}$ are linear combinations of corresponding elements in other columns. Most linear least squares programs print out an appropriate message if this situation exists. More dangerous are "near singularities" in $\mathbf{X}^T\mathbf{X}$ arising from "nearly exact" linear relationships between columns of $\mathbf{X}$. Some computer programs provide warning messages of this situation as well and interpretation of the estimation results should be made with caution. In such a situation, $\mathbf{X}^T\mathbf{X}$ is said to be "ill-conditioned", indicating numerical uncertainty (as opposed to statistical uncertainty) in the results. In essence an ill-conditioned situation is one in which very small changes in the x values produce very large changes in the values of the least squares parameter estimates.

Even without warning messages, an ill-conditioned matrix $\mathbf{X}^T\mathbf{X}$ can be detected from either the correlation matrix of the x data, which is really a transformed version of $\mathbf{X}^T\mathbf{X}$, or the correlation matrix of the parameter estimates. More will be said about this during the discussion of the precision of the parameter estimates.

Example 13.1

To illustrate the matrix notation introduced above two examples are discussed. The first is the straight line fit described in section 11 using waste solids removed as the response variable Y and feed flow rate as the error free independent variable x. The proposed model is

$$E(Y) = \beta_0 + \beta_1 x$$

which can be written in the general linear form (12.1) as

$$E(Y) = \beta_0 x_0 + \beta_1 x_1$$

where $x_0 = 1$ for every data point and x_1 is the feed flow rate. The matrices $\mathbf{Y}$, $\boldsymbol{\beta}$ and $\mathbf{X}$ are then

$$\mathbf{Y} = \begin{bmatrix} 24.3 \\ 19.6 \\ 12.9 \\ 25.2 \\ 4.2 \\ 18.4 \\ 20.3 \\ 6.0 \\ 13.3 \\ 5.9 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 0.13 \\ 1 & 0.23 \\ 1 & 0.60 \\ 1 & 0.13 \\ 1 & 0.88 \\ 1 & 0.39 \\ 1 & 0.23 \\ 1 & 0.75 \\ 1 & 0.50 \\ 1 & 0.88 \end{bmatrix}$$

The matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{Y}$ are

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 10 & 4.72 \\ 4.72 & 3.013 \end{bmatrix} \qquad \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 150.1 \\ 50.566 \end{bmatrix}$$

so that

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.385 & -0.603 \\ -0.603 & 1.279 \end{bmatrix}.$$

Least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can then be obtained from equation (13.3),

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 0.385 & -0.603 \\ -0.603 & 1.279 \end{bmatrix} \begin{bmatrix} 150.1 \\ 50.566 \end{bmatrix} = \begin{bmatrix} 27.2 \\ -25.8 \end{bmatrix}.$$

These values are of course the same as those obtained in section 11.

Example 13.2

The following data were obtained from an experiment designed to test the effects of temperature and roll speed on the tear strength of polyethylene sheeting.

Test Number	Temperature (°C)	Roll Speed (rpm)	Measured Tear Strength (pounds)
1	95	3000	10.29
2	100	2500	8.89
3	95	2000	15.44
4	90	2500	12.57
5	95	3000	11.41
6	90	3500	11.92
7	95	4000	14.11
8	85	3000	7.64
9	95	3000	12.77
10	100	3500	12.83
11	105	3000	6.86

The model form proposed for fitting these data was a full second degree polynomial in the two operating variables, temperature and roll speed.

Improved numerical accuracy in least squares calculations can be achieved by scaling the values of the original variables. In this example the following scaling is employed,

$$y = \text{measured tear strength} - 11.0$$

$$x_1 = \frac{\text{temperature} - 95}{5}$$

$$x_2 = \frac{\text{roll speed} - 3000}{500}$$

With this scaling the data may be expressed as follows:

x_1	x_2	y
0	0	-0.71
1	-1	-2.11
0	-2	4.44
-1	-1	1.57
0	0	0.41
-1	1	0.92
0	2	3.11
-2	0	-3.36
0	0	1.77
1	1	1.83
2	0	-4.14

A full second degree polynomial model is then

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

where the subscripts on each parameter identify the particular function of x_1 and x_2 with which it is associated. The matrices $\mathbf{Y}$, β and $\mathbf{X}$ are then

$$\mathbf{Y} = \begin{bmatrix} -0.71 \\ -2.11 \\ 4.44 \\ 1.57 \\ 0.41 \\ 0.92 \\ 3.11 \\ -3.36 \\ 1.77 \\ 1.83 \\ -4.14 \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{11} \\ \beta_{22} \\ \beta_{12} \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & -1 & 1 & 1 & -1 \\ 1 & 0 & -2 & 0 & 4 & 0 \\ 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 1 & 1 & 1 & -1 \\ 1 & 0 & 2 & 0 & 4 & 0 \\ 1 & -2 & 0 & 4 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 0 & 4 & 0 & 0 \end{bmatrix}$$

The matrices $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}^T\mathbf{Y}$ are

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 11 & 0 & 0 & 12 & 12 & 0 \\ 0 & 12 & 0 & 0 & 0 & 0 \\ 0 & 0 & 12 & 0 & 0 & 0 \\ 12 & 0 & 0 & 36 & 4 & 0 \\ 12 & 0 & 0 & 4 & 36 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix} \qquad \mathbf{X}^T\mathbf{Y} = \begin{bmatrix} 3.73 \\ -4.33 \\ 0.63 \\ -27.79 \\ 32.41 \\ 4.59 \end{bmatrix}$$

so that

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{bmatrix} 0.263 & 0 & 0 & -0.0789 & -0.0789 & 0 \\ 0 & 0.0833 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.0833 & 0 & 0 & 0 \\ -0.0789 & 0 & 0 & 0.0518 & 0.0206 & 0 \\ -0.0789 & 0 & 0 & 0.0206 & 0.0518 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.250 \end{bmatrix}.$$

Least squares estimates are

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_{11} \\ \hat{\beta}_{22} \\ \hat{\beta}_{12} \end{bmatrix} = \begin{bmatrix} 0.617 \\ -0.361 \\ 0.053 \\ -1.068 \\ 0.813 \\ 1.147 \end{bmatrix}$$

so that the fitted model is

$$\hat{y} = 0.617 - 0.361x_1 + 0.053x_2 - 1.068x_1^2 + 0.813x_2^2 + 1.147x_1x_2.$$

In terms of the original variables, this fitted model is

$$\begin{aligned}(\text{Estimated tear strength} - 11) = 0.617 &- 0.361\left(\frac{\text{temperature} - 95}{5}\right) \\ &+ 0.053\left(\frac{\text{roll speed} - 3000}{500}\right) \\ &- 1.068\left(\frac{\text{temperature} - 95}{5}\right)^2 \\ &+ 0.813\left(\frac{\text{roll speed} - 3000}{500}\right)^2 \\ &+ 1.147\left(\frac{\text{temperature} - 95}{5}\right)\left(\frac{\text{roll speed} - 3000}{500}\right)\end{aligned}$$

that is,

$$\begin{aligned}\text{Estimated tear strength} = &-218.361 + 6.67(\text{temperature}) - 0.063(\text{roll speed}) \\ &- 0.043(\text{temperature})^2 + 3.25\text{x}10^{-6}(\text{roll speed})^2 \\ &+ 4.59\text{x}10^{-4}(\text{temperature})(\text{roll speed})\end{aligned}$$

We now examine the assumptions upon which least squares estimation is based and then consider methods of testing the adequacy of a fitted model.

CHAPTER 14

Assumptions in Least Squares Estimation

All information which can be used to assess the adequacy of a fitted model is contained in the residuals $e_u = (y_u - \hat{y}_u)$, $u = 1,2,\cdots,n$. Therefore if a fitted model is to be judged adequate, then the behaviour of the residuals should match the behaviour assumed for the random error term ε. The assumptions about ε that are inherent in the use of least squares, whether for a linear or a nonlinear model, are as follows.

Assumption 1:

The values of the operating variables are known exactly. Statistically, this is expressed as $\text{var}(x_{ui}) = 0$, $u = 1,2,\cdots,n$, $i = 1,2,\cdots,p$, where var( ) denotes the variance of the variable in parentheses. In practice this is interpreted to mean that any uncertainty associated with a value of an operating variable has much less effect on the response value than the uncertainty associated with a measured value of the response itself.

This assumption is likely to be valid for the two examples described above but can create difficulties in other situations to be discussed later.

Assumption 2:

The form of the model is appropriate. Statistically, this is expressed as $E(\varepsilon) = 0$ for all data.

This assumption is often invalid during the early stages of analysis of a set of data. In the following section, procedures will be described for using residuals to detect deficiencies in a model form.

Assumption 3:

The variance of the random error term is constant over the region of the operating variables used to collect the data. Statistically, this is expressed as $\text{var}(\varepsilon_u) = \sigma^2$, $u = 1,2,\cdots,n$.

When the variance of the random error term varies over the operating region, such as in calibration of a gas chromatograph, for example, then either weighted least squares must be used [1, pp.77-81] or a transformation of the response variable must be made [2]. Arbitrary transformations such as logarithmic or reciprocal transformations to convert a nonlinear model form to a linear form often lead to violation of this assumption and produce misleading parameter estimates. More will be said about this in a later section.

Assumption 4:

There is no systematic association of the random error for any one data point with the random error for any other data point. Statistically, this is expressed as $\text{corr}(\varepsilon_u, \varepsilon_v) = 0, \quad u, v = 1,2,\cdots,n; u \neq v,$ where corr(,) denotes the linear correlation between the two variables in parentheses.

Violation of this assumption can have a much more dramatic effect on the parameter estimates than violation of any of the other three assumptions.

The assumption that the random error term has a normal probability density function is not. required for least squares estimation. However, if this assumption is valid in addition to the four assumptions listed above, then the parameter estimates obtained by least squares are maximum likelihood estimates [l, p.60], that is, the most likely values for the parameters for the postulated model form in the light of the available data. As explained in an earlier section, a normal probability density function very often is appropriate for data from the physical and engineering sciences, because of the Central Limit Theorem.

Other criteria for parameter estimation include (i) minimizing the sum of the absolute values of the residuals, (ii) minimizing the perpendicular distances of the data points to the fitted curve, and (iii) minimizing the maximum absolute residual value. Gauss proved for a linear model under the four assumptions stated above that of all parameter estimates which are linear combinations of the measured response values, least squares estimates have individually the minimum possible variances. For this reason, least squares estimation is advocated unless (i) violation of one or more of the above four assumptions implies the need for an alternative fitting criterion or (ii) the form of probability density function of the random error is known, and this form is not normal, thereby leading to maximum likelihood estimation which does not involve minimization of the sum of squares of residuals.

REFERENCES

[1] Draper, N.R. and Smith, H., 1966: Applied Regression Analysis, John Wiley, New York.

[2] Box, G.E.P. and Hill, W.J., 1974: Correcting inhomogeneity of variance with power transformation weighting, Technometrics, 16 (3), 385-389.

CHAPTER 15

Testing the Adequacy of a Fitted Model

15.1 RESIDUAL PLOTS

It has been wisely observed that the last thing one should do with a fitted model is believe it. The first type of test that should be made on a fitted model is to plot the residuals against the fitted response values. Recalling that residuals from an adequate fitted model should behave randomly, the plot should be examined to detect any unusual behaviour.

As shown in figure 15.1a, one or two residuals may be very large compared with the others, indicating that those data points lie far from the fitted curve. There is an unfortunate tendency to discard such *outliers* automatically on the grounds that "the reading must have been in error". Such action is not only unscientific but can lead to incorrect conclusions about the true relationship between the response variable and the operating variables. As Daniel and Wood [1] point out, outliers may be valid data points which, if investigated, can often reveal surprising information about the relationship being studied.

In figure 15.1(b), the variance of the residuals appears to be increasing systematically, thereby violating assumption 2 for least squares estimation. Unless a changing variance is properly accounted for in fitting a model, the parameter estimates and the fitted model may be misleading. Weighted least squares [2] may be used to estimate parameters in this situation, with

$\sum_{u=1}^{n} w_u e_u^2$ being minimized where w_u is the relative weight assigned to data point u. Each weight should be inversely proportional to the variance of the random error at the associated data point. Because good estimates of the variances for all of the data points are seldom available, transformation of the response variable [3] is a more dependable method for taking account of a changing variance. Indeed, even in cases where variance estimates are available, the transformation approach may be necessary [4].

Other violations of the least squares assumptions can be identified from residual plots. Systematic association between residuals for successive runs (violation of assumption 4) may be detected by plotting residuals against the order in which the data points were obtained. The situation in figure 15.1(c) can be corrected by adding a term βt_u to the original model form to account for a linear trend in the random errors with time of observation t. If the data points are equally spaced in time, the autocorrelation function

[5] of the residuals can reveal more complex forms of systematic association among successive residuals.

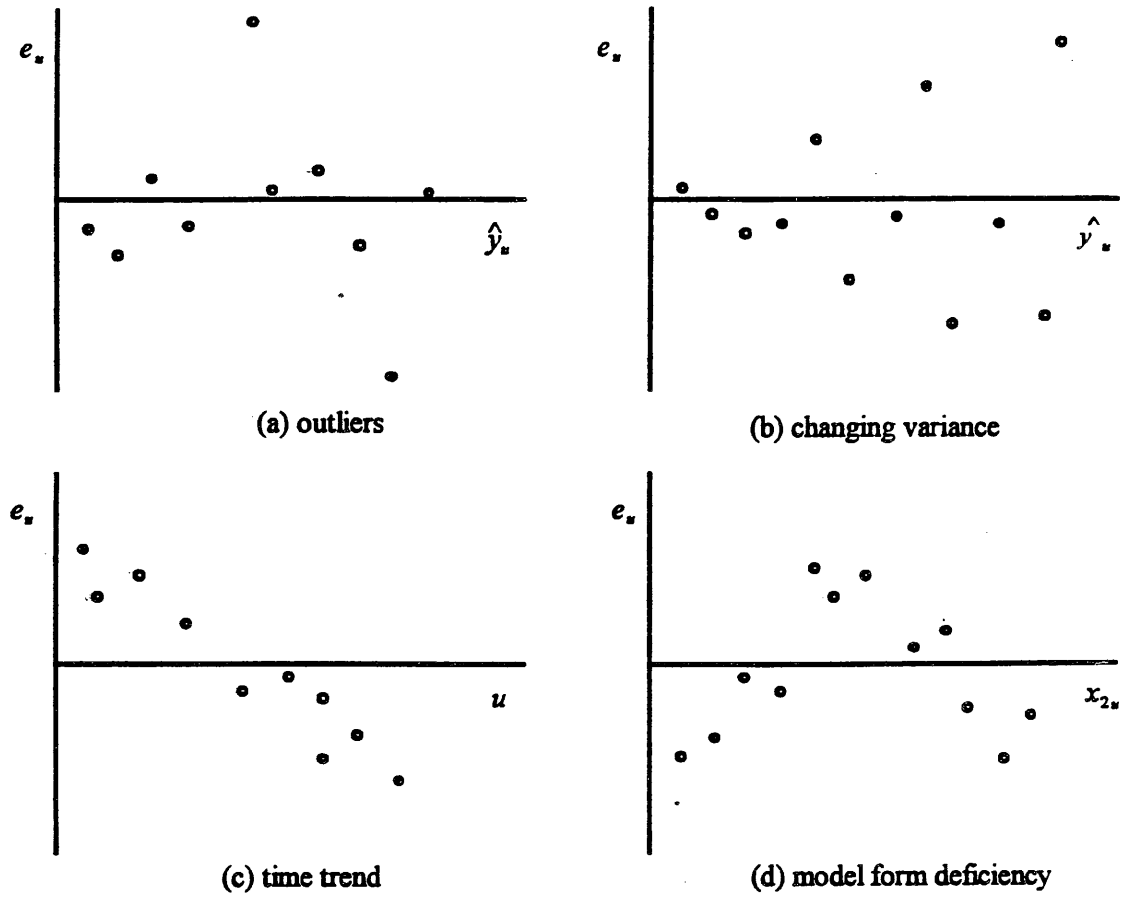

Figure 15.1 - Residual plots from inadequate model.

The need for additional terms in a model form can be revealed by plots of residuals against variables which are not present in the current model. The situation in figure 15.1(d) indicates the need for a term $\beta_{22}x_2^2$ in the model form. Notice that this residual pattern could be obtained even if the fitted model contained a term $\beta_2 x_2$.

Two residual plots are shown in figure 15.2 for the fitted model (11.16) obtained for the filtration data. Neither of these plots shows any evidence of systematic behaviour in the residuals, suggesting that the fitted model provides an adequate representation of the data.

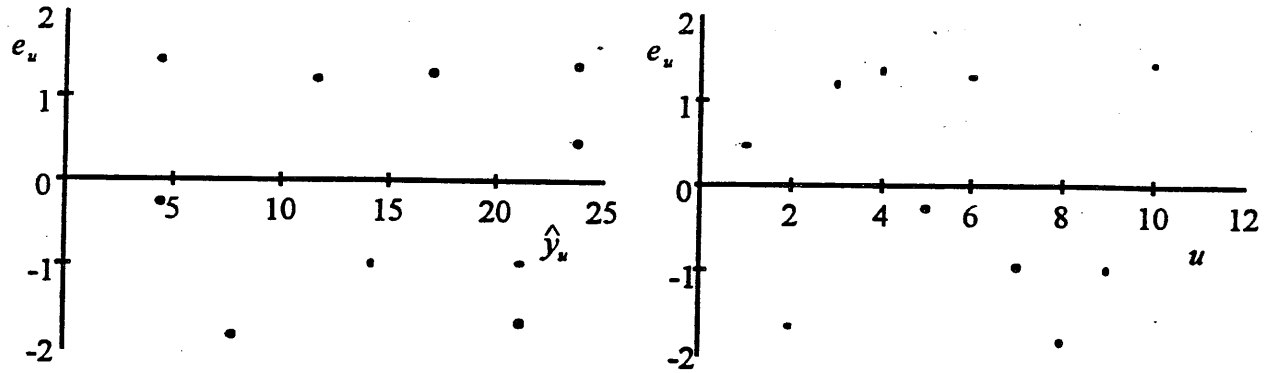

Figure 15.2 - Residual plots for the filtration data.

15.2 SIGNIFICANCE TESTS FOR MODEL INADEQUACY

Interpretations of residual plots are subjective and success in detecting inadequacies in a fitted model will depend upon the skill and experience of the data analyst. More objective tests can also be made. As stated earlier, lack of agreement between data points and a fitted curve arises from two sources: (i) experimental error and (ii) inadequacy in the form of the model being fitted. The tests of model inadequacy to be described now are significance tests for determining whether the residual component from source (ii) is appreciably larger than that from source (i).

The variance of experimental or "pure" error can be estimated from *replicate* data points. Replicate data are measured values obtained from independent tests carried out at the same operating conditions. In the filtration data shown in table 10.1 there are three pairs of replicates. To be genuine replicates, the differences among the measured response values of the replicate set must reflect the influences of all sources of variation. The replicates in the filtration data are genuine replicates. It is a common error to regard multiple readings taken while operating conditions are fixed as a set of genuine replicates. They are not genuine replicates because sources of variation arising from the setting of individual operating conditions has not been allowed to contribute to the difference among the readings.

For one set of m replicates, σ^2, the pure error variance, can be estimated using the well known expression

$$\hat{\sigma}^2 = \frac{\sum_{u=1}^{m} (y_u - \bar{y})^2}{m-1} \qquad (15.1)$$

where $\bar{y} = \sum_{u=1}^{m} y_u \Big/ m$ is the sample mean response for that replicate set. For l sets of replicates, each set at different operating conditions, the individual variance estimates, $\hat{\sigma}_1^2, \cdots, \hat{\sigma}_l^2$, may be pooled to obtain a superior estimate $\hat{\sigma}_p^2$ using the expression

$$\hat{\sigma}_p^2 = \frac{\sum_{i=1}^{l} (m_i - 1)\hat{\sigma}_i^2}{\sum_{i=1}^{l} (m_i - 1)} \qquad (15.2)$$

where m_i is the number of data points in the i'th set of replicates. Pooling of variance estimates should only be carried out when it can be safely assumed that all of the estimates $\hat{\sigma}_i^2$, $i = 1, \cdots, l$, are estimates of a common variance. In the above discussion this means that the pure error variance is assumed to be the same at all operating conditions in the data. This assumption can be tested using Bartlett's test [6] or, if there is suspicion that the random error is not normally distributed, Levene's test [7].

If the data contain replicates, the residual sum of squares can be divided as shown below into a pure error component and another component that reflects model inadequacy. The notation used in the derivation [2] is defined as follows :

- y_u = measured response value for data point u,
- $\hat{y}_u$ = value of the fitted response value (the ordinate of the point on the fitted curve) for data point u,
- n = total number of data points used for fitting the model,
- $\bar{y}_u$ = sample mean of all measured response values at the operating conditions for data point u (note that unless there are replicates at point u, $\bar{y}_u = y_u$),
- k = number of distinct sets of operating conditions in the data,
- m_i = number of data points at operating condition i.

The residual sum of squares is

$$\begin{aligned}\sum_{u=1}^{n} e_u^2 &= \sum_{u=1}^{n} (y_u - \hat{y}_u)^2 \\ &= \sum_{u=1}^{n} \{(y_u - \bar{y}_u) + (\bar{y}_u - \hat{y}_u)\}^2 \\ &= \sum_{u=1}^{n} (y_u - \bar{y}_u)^2 + \sum_{u=1}^{n} (\bar{y}_u - \hat{y}_u)^2 + 2\sum_{u=1}^{n} (y_u - \bar{y}_u)(\bar{y}_u - \hat{y}_u)\end{aligned} \tag{15.3}$$

Now the last term in expression (15.3) can be shown to be equal to zero as follows,

$$\begin{aligned}\sum_{u=1}^{n} (y_u - \bar{y}_u)(\bar{y}_u - \hat{y}_u) &= \sum_{u=1}^{n} y_u \bar{y}_u - \sum_{u=1}^{n} \bar{y}_u^2 - \sum_{u=1}^{n} y_u \hat{y}_u + \sum_{u=1}^{n} \bar{y}_u \hat{y}_u \\ &= \sum_{i=1}^{k} m_i \bar{y}_i^2 - \sum_{i=1}^{k} m_i \bar{y}_i^2 - \sum_{i=1}^{k} m_i \bar{y}_i \hat{y}_i + \sum_{i=1}^{k} m_i \bar{y}_i \hat{y}_i \,. \\ &= 0\end{aligned}$$

Therefore from equation (15.3),

$$\sum_{u=1}^{n} e_u^2 = \sum_{u=1}^{n} (y_u - \bar{y}_u)^2 + \sum_{u=1}^{n} (\bar{y}_u - \hat{y}_u)^2 \,. \tag{15.4}$$

The first term on the right hand of (15.4) arises from pure error alone since it is independent of the form of model that may be proposed. It can be confirmed that this term is in fact equal to the numerator of the expression (15.2) for $\hat{\sigma}_p^2$. Since there are $\sum_{i=1}^{l} (m_i - 1)$ independent non-zero components $(y_u - \bar{y}_u)$ in this term, where l is the number of sets of replicates in the data, the number of degrees of freedom is $\sum_{i=1}^{l} (m_i - 1)$.

The second term on the right hand side of (15.4) arises from both pure error and inadequacy in the form of model being fitted. The following argument can be used to obtain its associated degrees of freedom. The number of degrees of freedom for $\sum_{u=1}^{n} e_u^2$ is $(n-p)$, where p is the number of estimated parameters in the fitted model, because the n residuals $e_1, \cdots, e_n$ have been constrained by the p linear equations (13.2) to ensure that $\sum_{u=1}^{n} e_u^2$ is minimized. Since the number of degrees of freedom associated with each side of

equation (15.4) must be the same, the number of degrees of freedom associated with $\sum_{u=1}^{n}(\bar{y}_u - \hat{y}_u)^2$ is $n - p - \sum_{i=1}^{l}(m_i - 1)$.

As indicated by expression (15.2), $\sum_{u=1}^{n}(y_u - \bar{y}_u)^2 \Big/ \sum_{i=1}^{l}(m_i - 1) = \hat{\sigma}_p^2$ is an estimate of σ^2, the pure error variance. Following Draper and Smith [2], it can be shown that $\sum_{u=1}^{n}(\bar{y}_u - \hat{y}_u)^2 \Big/ \left\{n - p - \sum_{i=1}^{l}(m_i - 1)\right\}$ is an estimate of a linear combination of pure error variance and bias due to inadequacy of the model form.

The quantity $\sum_{u=1}^{n}(\bar{y}_u - \hat{y}_u)^2$ is more easily calculated from (15.4) as

$$\sum_{u=1}^{n}(\bar{y}_u - \hat{y}_u)^2 = \sum_{u=1}^{n} e_u^2 - \sum_{u=1}^{n}(y_u - \bar{y}_u)^2 \tag{15.5}$$

where $\sum_{u=1}^{n}(y_u - \bar{y}_u)^2 = \hat{\sigma}_p^2 \sum_{i=1}^{l}(m_i - 1)$.

The significance of any model inadequacy can be tested by comparing the value of the ratio

$$R = \frac{\sum_{u=1}^{n}(\bar{y}_u - \hat{y}_u)^2 \Big/ \left\{n - p - \sum_{i=1}^{l}(m_i - 1)\right\}}{\hat{\sigma}_p^2} \tag{15.6}$$

with the value of F_{ν_1,ν_2} at the desired probability level where

$$\nu_1 = n - p - \sum_{i=1}^{l}(m_i - 1)$$

and

$$\nu_2 = \sum_{i=1}^{l}(m_i - 1).$$

This is a test of the null hypothesis that the bias due to model inadequacy is zero.

If the value of R is larger than the upper α per cent value of F_{ν_1,ν_2}, where α is the selected level of significance, then the fitted model is said to display "lack of fit". Notice that, as in any significance test, this procedure gives no information about the

nature of the inadequacy in the fitted model. Such information may be obtained from residual plots.

Even if the value of R is not "significant", residual plots from the fitted model should be examined. The sensitivity of the test ratio (15.6) to model inadequacy decreases sharply with the number of degrees of freedom ν_2 associated with $\hat{\sigma}_p^2$. Thus, unless the data contain several replicates, this procedure will detect only gross inadequacy in a fitted model.

Example

This test for model inadequacy is now applied to the fitted model (11.6) for the filtration data. From (11.4), the sum of squares of residuals is

$$\sum_{u=1}^{10} e_u^2 = 15.89$$

and has $10-2=8$ degrees of freedom.

There are three pairs of replicate response values, (24.3, 25.2), (19.6, 20.3) and (4.2, 5.9), yielding three estimates of pure error variance, $\hat{\sigma}_1^2 = 0.41$, $\hat{\sigma}_2^2 = 0.25$ and $\hat{\sigma}_3^2 = 1.45$, respectively. Application of both Bartlett's test and Levene's test reveal no reason for doubting that the pure error variance is the same at all three operating conditions. Since these conditions include both extreme feed flow rates, it is reasonable to assume that the pure error variance is constant over the data set.

Using (15.2) for the three variance estimates,

$$\hat{\sigma}_p^2 = \frac{\sum_{i=1}^{3} \hat{\sigma}_i^2}{3}$$
$$= 0.70$$

and has 3 degrees of freedom. From (15.5),

$$\sum_{u=1}^{n} (\bar{y}_u - \hat{y}_u)^2 = 15.89 - 3(0.70)$$
$$= 13.79$$

and has $10-2-3=5$ degrees of freedom.

The test ratio (15.6) has the value

$$R = \frac{13.79/5}{0.70}$$
$$= 3.94$$

Comparing this value with $F_{5,3,0.05} = 9.01$ suggests that any inadequacy in the fitted straight line model (11.16) is not significantly larger (at the 0.05 significance level) than experimental error.

Draper and Smith [3] refer to the foregoing test as one using a "internal" estimate of pure error variance because the estimate $\hat{\sigma}_p^2$ has been obtained from data that are part of the total data set to which the model has been fitted. For situations in which the data contain no replicates, Draper and Smith describe a test for model inadequacy that uses an "external" estimate of pure error variance. The external estimate may be one based upon past information or upon data from another laboratory or from another process similar to the one under study. The data from which the external estimate is derived are not part of the data to which the model is fitted.

The test procedure using an external estimate of the error variance is similar in form to the one described above. The test ratio is

$$T = \frac{\sum_{u=1}^{n} e_u^2 \Big/ (n-p)}{\hat{\sigma}_E^2} \tag{15.7}$$

where $\hat{\sigma}_E^2$ is the external estimate of pure error variance having ν_E degrees of freedom. The significance of any inadequacy in the fitted model is tested by comparing the value of T with the value of $F_{(n-p),\nu_E}$ at the desired probability level.

Because the data from which $\hat{\sigma}_E^2$ is obtained may not have exactly the same error structure as the data used to fit the model, $\hat{\sigma}_E^2$ is not in general as reliable an estimate of σ^2 as one obtained from replicates within the data set being used for fitting the model.

REFERENCES

[1] Daniel, C. and Wood, F.S. (1971), *Fitting Equations to Data*, John Wiley, New York.

[2] Draper, N.R. and Smith, H. (1966), *Applied Regression Analysis*, John Wiley, New York.

[3] Box, G.E.P. and Hill, W.J. (1974), Technometrics, 16, pp. 385-389.

[4] Pritchard, D.J., Bacon, D.W. and Downie, J. (1977), Technometrics, 19, pp. 227-236.

[5] Box, G.E.P. and Jenkins, G.M. (1970), *Time Series Analysis, Forecasting and Control*, Holden Day, San Francisco.

[6] Chatfield, C. (1969), *Statistics for Technology*, Penguin.

[7] Levene, H. (1960), *Contributions to Probability and Statistics, edited by I. Olkin*, Stanford University Press, Stanford, California, pp. 278-292.

CHAPTER 16

Computing the Sum of Squares of Residuals

The sum of squares of residuals $\sum_{u=1}^{n} e_u^2$ can obviously be computed by squaring each residual and adding the squares. Although this method of calculation produces an answer whose numerical accuracy is generally good, most linear least squares computer programs use a different calculation procedure which, although less accurate, is faster.

The sum of squares of residuals can be written in matrix notation as $\mathbf{e}^T\mathbf{e}$ where in the notation introduced in section 13,

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}. \tag{16.1}$$

From (16.1),

$$\begin{aligned} \mathbf{e}^T\mathbf{e} &= \left(\mathbf{Y} - \hat{\mathbf{Y}}\right)^T\left(\mathbf{Y} - \hat{\mathbf{Y}}\right) \\ &= \mathbf{Y}^T\mathbf{Y} - 2\hat{\mathbf{Y}}^T\mathbf{Y} + \hat{\mathbf{Y}}^T\hat{\mathbf{Y}} \end{aligned} \tag{16.2}$$

where for a linear model,

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}. \tag{16.3}$$

Therefore in (16.2),

$$\hat{\mathbf{Y}}^T\mathbf{Y} = (\mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{Y} = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y} \tag{16.4}$$

and

$$\hat{\mathbf{Y}}^T\hat{\mathbf{Y}} = (\mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}}. \tag{16.5}$$

Now from (13.3),

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}.$$

Substituting for $\hat{\beta}$ in (16.5),

$$\hat{\mathbf{Y}}^{\mathrm{T}}\hat{\mathbf{Y}} = \hat{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} = \hat{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{16.6}$$

Substituting (16.4) and (16.6) in (16.2),

$$\mathbf{e}^{\mathrm{T}}\mathbf{e} = \mathbf{Y}^{\mathrm{T}}\mathbf{Y} - \hat{\beta}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{Y}. \tag{16.7}$$

Expression (16.7) is customarily used in linear least squares programs to compute the sum of squares of residuals since the vectors $\mathbf{X}^{\mathrm{T}}\mathbf{Y}$ and $\hat{\beta}$ have already been calculated in obtaining estimates of the parameters.

CHAPTER 17

Testing the Need for Particular Terms in a Fitted Model

If neither an internal nor an external estimate of pure error variance is available, a test for specific types of inadequacy in a fitted model can still be carried out. The specified potential inadequacy must be expressed in the form of an extension of the current model. Since the number of possible extensions to any model form is infinite, this test procedure cannot be comprehensive. Nevertheless, many situations arise in which a suspected deficiency in a model form can be expressed as a particular type of extension. For example, deficiency in a straight line model may be accounted for by the addition of a squared term in the operating variable. Deficiency in a model form involving two operating variables may be accounted for by one or more additional terms involving a third operating variable. In cases like these the need for specific additional terms in a model can be assessed using the following procedure.

The original model form will be denoted as model A. The extended model form, denoted as model B, is model A with additional terms. Thus model B contains all of the parameters of model A plus some additional parameters. The ratio

$$Q = \frac{1}{\hat{\sigma}^2} \frac{\begin{pmatrix}\text{sum of squares of}\\ \text{residuals from model A}\end{pmatrix} - \begin{pmatrix}\text{sum of squares of}\\ \text{residuals from model B}\end{pmatrix}}{\begin{pmatrix}\text{number of parameters}\\ \text{estimated in model B}\end{pmatrix} - \begin{pmatrix}\text{number of parameters}\\ \text{estimated in model A}\end{pmatrix}}, \qquad (17.1)$$

where $\hat{\sigma}^2$ is an estimate of the pure error variance σ^2, is a measure of the improvement achieved by model B in fitting the data over the fit of model A. If Q is large, then the improvement is worthwhile, indicating that model A is deficient and that the additional parameters in model B are warranted. If Q is small (it clearly cannot be less than zero), then the addition of the extra parameters in model B is not worthwhile. In the latter situation, model A cannot be judged adequate, but it has at least passed a "test" against the specific alternative model B.

Assessment of how large the ratio Q must be, in order that model B can be judged to be an improvement on model A, is accomplished by referring the value of Q to an F_{ν_1,ν_2} distribution. The number of degrees of freedom, ν_1, is the difference in the number of estimated parameters in models A and B; ν_2 is the number of degrees of freedom

associated with $\hat{\sigma}^2$. The probability level for F_{ν_1,ν_2} will reflect the data analyst's definition of what constitutes a significant improvement over the fit of model A. This test is a test of significance of the null hypothesis that the true values of all of the additional parameters in model B are zero.

The estimate $\hat{\sigma}^2$ required in Q may be obtained from replicates within the data set, if they exist, or from an external estimate as described previously. If neither an internal nor an external estimate of σ^2 is available, then as a last resort the residual mean square from model B, $\sum_{u=1}^{n} e_u^2 \Big/ (n-p)$, where p is the number of parameters estimated in model B, can be used as a rough estimate of $\hat{\sigma}^2$. Problems arise from using this estimate because it is a valid estimate of σ^2 only if model B provides an adequate fit of the data. However, no conclusive test of the adequacy of a fitted model can be made without a pure error variance estimate (either internal or external) obtained from replicates If model B is inadequate, then it's residual mean square will over estimate σ^2 and the sensitivity of the test based on ratio Q will decrease.

Example 17.1

The polyethylene data described in example 13.2 will be used to illustrate the test described above for assessing the need for a particular extension of a fitted model.

Using the coding suggested in example 13.2, a first degree polynomial can be fitted to these data with the following result,

$$\hat{y} = 0.339 - 0.361x_1 + 0.052x_2 . \qquad (17.2)$$

The sum of squares of residuals from this fit is 69.872 with $11-3=8$ degrees of freedom.

From the three replicates $\hat{\sigma}_p^2 = 1.542$ with 2 degrees of freedom. The value of the ratio R defined in (15.6) is 7.219 and since $F_{6,2,0.05} = 19.3$ and $F_{6,2,0.10} = 9.3$, it might be concluded that any lack of fit in model (17.2) is not significant. However, this test is not very sensitive in this case since $\hat{\sigma}_p^2$ has only 2 degrees of freedom. Tests of extensions of this model would be of interest.

In example 13.2, the second degree polynomial that provided a least squares fit to these data was found to be

$$\hat{y} = 0.617 - 0.361x_1 + 0.053x_2 - 1.068x_1^2 + 0.813x_2^2 + 1.147x_1x_2 . \qquad (17.3)$$

It's residual sum of squares is 7.541 with $11-6=5$ degrees of freedom.

The value of the ratio R for model (17.3) is 0.963, much more reassuring than the R ratio for model (17.2) since if no model inadequacy exists, the expected value of R is 1. However, the same criticism concerning lack of sensitivity of this test due to the small number of replicates applies.

Comparing the two fitted models (17.2) and (17.3), the value of ratio Q defined by (17.1) using $\hat{\sigma}^2 = \hat{\sigma}_p^2 = 1.542$ with 2 degrees of freedom is

$$Q = \frac{1}{1.542}\left(\frac{69.872 - 7.541}{6-3}\right)$$
$$= 13.48$$

Referring this value to the $F_{3,2}$ distribution, it is found to be significant at the 0.10 probability level, but not significant at the 0.05 probability level. The occurrence of this sort of borderline case is a reminder of the danger of attempting to use a significance test to label all results as "significant" or "non-significant". Ideally, more data should be obtained to increase the sensitivity of the test but if this is not possible, then other comparisons must be used.

Because the value of ratio R for model (17.3) gave no hint of inadequacy for that model, it is reasonable to recalculate ratio Q using the residual mean square from the fitted model (17.3) as an estimate of $\hat{\sigma}^2$. Further support for using this estimate comes from the close agreement between the value of the residual mean square, $7.54/5 = 1.51$ and the internal estimate of pure error variance, $\hat{\sigma}_p^2 = 1.542$. The advantage of using the residual mean square is its larger number of degrees of freedom, resulting in a more sensitive comparison of models (17.2) and (17.3).

Using $\hat{\sigma}^2 = 1.51$ with 5 degrees of freedom,

$$Q = \frac{1}{1.51}\left(\frac{69.872 - 7.541}{6-3}\right)$$
$$= 13.76$$

and, when referred to the $F_{3,5}$ distribution, this value is found to be significant at the 0.01 probability level.

From these comparisons and individual tests of model inadequacy, the second degree polynomial model (17.3) appears to provide a distinctly superior fit to the polyethylene data to that provided by the first degree polynomial model (17.2).

The ratio Q can also be used to test whether simplification of an adequate model B is possible by deleting terms to form a model A that is still an adequate representation of the data. Reduction of an adequate fitted model to its simplest form is a useful exercise since, as explained in section 19, redundant parameters in a fitted model inflate the variances of predicted response values unnecessarily.

One linear least squares estimation procedure that has achieved considerable popularity is stepwise regression. From a number of specified potential independent variables, the variable is selected that is most highly correlated with the response variable. Then a sequence of fitted models is developed, each model containing all of the terms in the preceding model plus the one additional test for which the ratio Q in (17.1) is largest. For this stepwise procedure, Q will always be a comparison of two models that differ by only one term. The variance σ^2 is estimated by the residual mean square from model B, the model having the larger number of parameters, *whether or not model B is adequate*. The decision at each step about whether to add a term to the existing model is made within the routine by comparing ratio Q to an F value specified by the user. This F value remains constant throughout the sequence of fits.

Another decision is also made at each step: whether to delete any one of the terms in the current fitted model. Because of correlations among the independent variables, a particular variable that was a worthwhile addition at one step may become redundant after certain other variables are added to the model. This situation arises because of the nature of the test based on ratio Q. Judgment of the merit of adding or deleting any individual term is contingent on the particular collection of other terms that are present in the model at that step.

For the reasons described above, stepwise regression should be employed with care. It does not necessarily produce the fitted model with fewest terms nor even that combination of a specified number of terms that provides the best fit to the data. Efficient procedures do exist [2] for examining all possible subsets of variables.

If the column of matrix $\mathbf{X}$ corresponding to a particular independent variable is orthogonal to all of the other columns in $\mathbf{X}$, then this variable is uncorrelated with all of the other independent variables and a test for its inclusion in the model can be conducted without regard to the terms already in the model. It is sometimes possible to achieve the situation where all columns of $\mathbf{X}$ are mutually orthogonal so that every term in the model can be tested independently of the other terms. This condition will arise only if operating variables for the individual tests have been deliberately set according to a planned pattern. Planning the collection of experimental data is a topic that will be discussed in considerable detail in subsequent sections.

Once a fitted model has been judged adequate its precision should be assessed. This will reveal the value of the model (i) as a description of the data to which it has been fitted and (ii) as a predictor of performance at other operating conditions. In the next section, the precision of the parameter estimates in a fitted model is discussed. Then follows a section dealing with the precision of predictions using the fitted model.

REFERENCES

[1] Efroymson, M.A. (1960), *Mathematical Methods for Digital Computers*, (Vol 1), edited by A. Ralston and H.S. Wilf, John Wiley, New York.

[2] Hocking, R.R. (1976), Biometrics, 32, pp 1-49.

CHAPTER 18

Precision of the Parameter Estimates

Even in current scientific literature it is common to find values of parameter estimates quoted without any accompanying measure of their precision. Absence of information about precision can lead to misinterpretation of the quoted values.

Consider a simple example from reaction kinetics. If the estimated value of a rate constant is reported as $\hat{k} = 5.61\text{x}10^{-3}$, what is to be inferred about the true value k? If the precision of the estimate $\hat{k}$ were described by a 95 per cent confidence interval, then of course $5.61\text{x}10^{-3} \pm 2\text{x}10^{-4}$ would suggest that the true value k was very close to $5.6\text{x}10^{-3}$. On the other hand, $5.61\text{x}10^{-3} \pm 2\text{x}10^{-2}$ would indicate a very broad range of plausible values for k including the value zero.

Uncertainty about the true parameter values results from the random errors associated with the measured responses. All basic information concerning the precision of parameter estimates from a fitted model is contained in the matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2$, where σ^2 is the variance of the random errors. This matrix is called the *covariance matrix of the parameter estimates*. Obviously, an estimate $\hat{\sigma}^2$ must be used for σ^2 and the choice of an appropriate estimate has been thoroughly discussed in the preceding section.

For the general linear model,

$$E(Y) = \beta_1 x_1 + \cdots + \beta_p x_p \tag{18.1}$$

the covariance matrix of the parameter estimates is a $\hat{\beta}_i$, $i = 1, \cdots, p$ p x p symmetric matrix whose i'th row (or i'th column) is associated with the parameter estimate $\hat{\beta}_i, i = 1, \cdots, p$.

Individual elements of the estimated covariance matrix of the parameter estimates,

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\hat{\sigma}^2 = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1p} \\ & c_{22} & \cdots & c_{2p} \\ & & \ddots & \vdots \\ & & & c_{pp} \end{bmatrix} \tag{18.2}$$

have the following interpretations. The diagonal element $c_{ii}, i = 1, \cdots, p$, is the estimated variance of $\hat{\beta}_i$. The off-diagonal element $c_{ij}, i, j = 1, \cdots, p; i \neq j$, is the estimated covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$. Covariance was discussed briefly in section 10. In this case,

$$\operatorname{cov}\left(\hat{\beta}_i, \hat{\beta}_j\right) = E\left[\left\{\hat{\beta}_i - E\left(\hat{\beta}_i\right)\right\}\left\{\hat{\beta}_j - E\left(\hat{\beta}_j\right)\right\}\right]$$

As explained in section 10, a more useful measure of association between two random variables is their correlation ρ. For parameter estimates $\hat{\beta}_i$ and $\hat{\beta}_j$,

$$\rho\left(\hat{\beta}_i, \hat{\beta}_j\right) = \frac{\operatorname{cov}\left(\hat{\beta}_i, \hat{\beta}_j\right)}{\sqrt{\operatorname{var}\left(\hat{\beta}_i\right)\operatorname{var}\left(\hat{\beta}_j\right)}}.$$

Using (18.2) the estimated correlation $\hat{\rho}\left(\hat{\beta}_i, \hat{\beta}_j\right)$ between $\hat{\beta}_i$ and $\hat{\beta}_j$ is

$$\hat{\rho}\left(\hat{\beta}_i, \hat{\beta}_j\right) = \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}. \tag{18.3}$$

Example 18.1

To illustrate the meaning of correlation between parameter estimates, the straight line fit to the filtration data will again be used. In section 11 the fitted model was found to be

$$\hat{y} = 27.2 - 25.8x$$

where y and x denoted waste solids removed and feed flow rate, respectively. As found in section 15, an internal estimate of pure error variance, $\hat{\sigma}_p^2 = 0.698$, was available from the data. The covariance matrix of the parameter estimates is then

$$(\mathbf{X}^T\mathbf{X})^{-1}\hat{\sigma}^2 = \begin{bmatrix} 0.384 & -0.601 \\ -0.601 & 1.274 \end{bmatrix}(0.698)$$
$$= \begin{bmatrix} 0.268 & -0.419 \\ -0.419 & 0.889 \end{bmatrix} \quad (18.4)$$

From (18.4) it can be seen that the estimated variances of $\hat{\beta}_0$ and $\hat{\beta}_1$. are 0.268 and 0.899, respectively. The estimated correlation between the two parameter estimates is

$$\hat{\rho}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-0.419}{\sqrt{0.268(0.889)}}$$
$$= -0.858$$

Since correlation values must lie between -1 and 1 this value represents a reasonably large negative linear association between $\hat{\beta}_0$ and $\hat{\beta}_1$.

A graphical demonstration of the precision of the parameter estimates is given in figure 18.1. Here the relationship between the residual sum of squares $\sum_{u=1}^{10} e_u^2$ and the two parameters β_0 and β_1 is shown in graphical form as a set of concentric ellipses. The number shown beside each of the ellipses is the residual sum of squares obtained by using any pair of parameter values on the that ellipse. The centre of the elliptical set: is located at the least squares value $\hat{\beta}_0 = 27.2$ and $\hat{\beta}_1 = -25.8$. The intervals produced by projecting the end points of any ellipse on the β_0 and β_1 axes reflect the relative uncertainties of their estimates. Because the variance of $\hat{\beta}_1$ is larger than the variance of $\hat{\beta}_0$, the interval on the β_1 axis is larger than the corresponding interval on the β_0 axis.

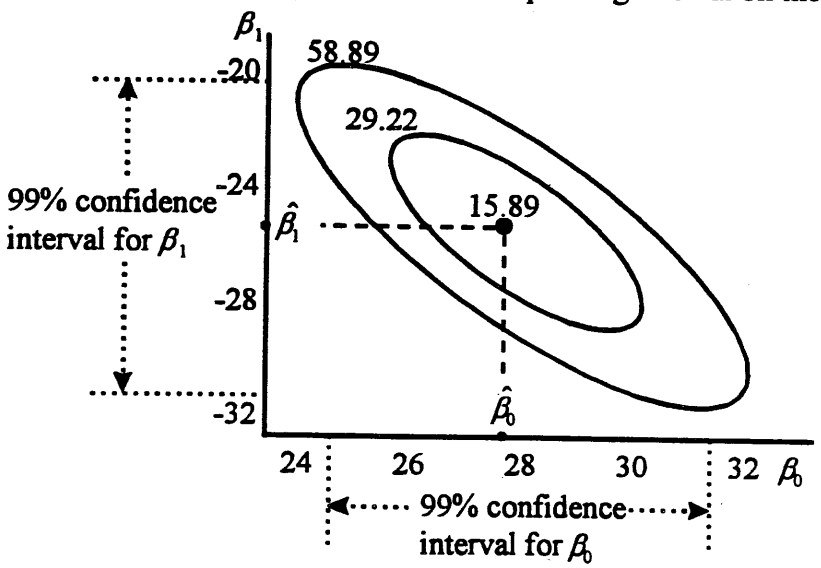

Figure 18.1 - Residual sum of squares surface for a straight line fit to the filtration data.

The correlation between $\hat{\beta}_0$ and $\hat{\beta}_1$ represents the degree to which the least squares estimate of one of the parameters depends upon the value of the other parameter. From figure 18.1, for example, it can be seen that the least squares estimate of β_1 decreases as the value of β_0 increases. Had the estimated correlation $\hat{\rho}\left(\hat{\beta}_0, \hat{\beta}_1\right)$ been positive instead of negative, the major axes of the ellipses would have had a positive slope, indicating an increase in the least squares value of β_1 as β_0 increased. The closer a correlation is to 1 or -1, the "thinner" each ellipse becomes about it's major axis, i.e. the closer the association between $\hat{\beta}_0$ and $\hat{\beta}_1$ approaches an exact one to one relationship.

For a fitted model of the general linear form (18.1), the precision of each parameter estimate may be expressed by a confidence interval for the true value of that parameter. A $100(1-\alpha)$ per cent confidence interval for $\beta_i, i = 1, \cdots, p$, is given by the expression

$$\hat{\beta}_i \pm t_{\nu,\alpha/2}\sqrt{\left(\text{estimated variance of } \hat{\beta}_i\right)} \tag{18.5}$$

where $\hat{\beta}_i$ are the least squares estimate of β_i,

$t_{\nu,\alpha/2}$ is the upper $\alpha/2$ percent value of the t_ν distribution,

ν is the number of degrees of freedom associated with the pure error variance estimate, $\hat{\sigma}^2$.

The estimated variance of c_{ii} is the diagonal element c_{ii} in expression (18.2).

Unfortunately individual confidence intervals for the parameter estimates do not provide information about correlation that may exist between those estimates. Without correlation values the inferences drawn about plausible values of the parameters $\beta_1, \cdots, \beta_p$ may be very wrong [1, pp. 64-67].

When parameter estimates are correlated, their precision can be described completely by a *joint confidence region*. A $100(1-\alpha)$ per cent joint confidence region for a set of parameters $\beta_1, \cdots, \beta_p$ is a region in the parameter space that has probability $(1-\alpha)$ of containing the true values of $\beta_1, \cdots, \beta_p$. For the general linear model (18.1), a $100(1-\alpha)$ per cent confidence region for the p parameters β is defined as those values of the parameters that satisfy the following inequality,

$$\left(\beta-\hat{\beta}\right)^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\left(\beta-\hat{\beta}\right) \le p\hat{\sigma}^2 F_{p,\nu,\alpha} \tag{18.6}$$

where $\hat{\beta}$ are the least squares estimate,

$\hat{\sigma}^2$ is an estimate of pure error variance with ν degrees of freedom,

$F_{p,\nu,\alpha}$ is the upper α per cent point of the F distribution having p and ν degrees of freedom.

Because $\mathbf{X}^T\mathbf{X}$ in expression (18.6) is a positive definite matrix, the boundary of the joint confidence region (18.6) is an ellipse. The inner ellipse in figure 18.1 is a 95 percent joint confidence region for β_0 and β_1 for a straight line representation of the filtration data The outer ellipse is a 99 per cent joint confidence region for the same example. For comparison, the individual 99 per cent confidence intervals for β_0 and β_1 are also shown in figure 18.1. In calculating the regions and the intervals, the internal estimate $\hat{\sigma}_p^2 = 0.698$ with 3 degrees of freedom has been used as an estimate of pure error variance.

For fitted models having more than three parameters, interpretation of the joint confidence region (18.6) will be difficult unless all parameters are mutually uncorrelated. In this case $\mathbf{X}^T\mathbf{X}$ will be a diagonal matrix and the principal axes of the hyper ellipsoidal region will be parallel to the co-ordinate axes for the parameters and individual assessment of each parameter can be made without regard to the values of the other parameters. Such a situation can be ensured only by deliberately designing a set of test runs to achieve this result.

When correlations exist among parameter estimates in a multi-parameter model, the inequality (18.6) can be used to assess the plausibility of various combinations of parameter values. In particular the effects of setting selected subsets of parameters to zero can be tested.

REFERENCES

[1] Draper, N.R. and Smith, H., (1966), Applied Regression Analysis, John Wiley, New York.

CHAPTER 19

Precision of Predicted Responses

The effects of the random errors associated with the measured response values on the parameter estimates have been discussed in the preceding section. This section describes how these effects are transmitted to the predicted responses.

The least squares fitted model of general form can be written as

$$\hat{y} = \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p .$$

For any set of operating conditions "k" of interest, whether they be at one of the existing data points or at an untested point within the operating region or at a point outside the operating region, a prediction of the response can be made,

$$\hat{y}_k = \hat{\beta}_1 x_{k1} + \cdots + \hat{\beta}_p x_{kp}$$

or, in matrix notation,

$$\hat{y}_k = \mathbf{x}_k^{\mathrm{T}} \hat{\boldsymbol{\beta}}$$

where $\hat{y}_k$ is the predicted response at operating conditions k and $\mathbf{x}_k^{\mathrm{T}} = \begin{bmatrix} x_{k1} & \cdots & x_{kp} \end{bmatrix}$ is a p element row vector of values of the independent variables at operating conditions k.

If it is assumed that the fitted model is adequate at operating conditions k, then $E(\varepsilon_k) = 0$ and consequently $E(\hat{y}_k) = E(y_k)$. Under this assumption the variance of $\hat{y}_k$ can be expressed as

$$\operatorname{var}(\hat{y}_k) = \sum_{i=1}^{p} x_{ki}^2 \operatorname{var}(\hat{\beta}_i) + 2 \sum_{i=1}^{p} \sum_{j=i+1}^{p} x_{ki} x_{kj} \operatorname{cov}(\hat{\beta}_i, \hat{\beta}_j) . \qquad (19.1)$$

Expression (19.1) can be written in matrix form as

$$\operatorname{var}(\hat{y}_k) = \mathbf{x}_k^{\mathrm{T}} (\mathbf{X}^{\mathrm{T}} \mathbf{X})^{-1} \mathbf{x}_k \sigma^2 . \qquad (19.2)$$

For a straight line model

$$E(Y) = \beta_0 + \beta_1 x$$

fitted to n data it can be confirmed that

$$\text{var}(\hat{y}_k) = \sigma^2 \left\{ \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{u=1}^{n} (x_u - \bar{x})^2} \right\} \qquad (19.3)$$

where $\bar{x} = \sum_{u=1}^{n} x_u \Big/ n$ is the sample mean of the values of the operating variable used in fitting the model.

From expression (19.3) it can be seen that for a fitted straight line model, the variance of the predicted response increases monotonically as the operating conditions at which the prediction is made move away from $\bar{x}$. This effect, which is not necessarily true for other model forms, is shown in figure 19.1. The danger of excessive extrapolation of a fitted model is evident.

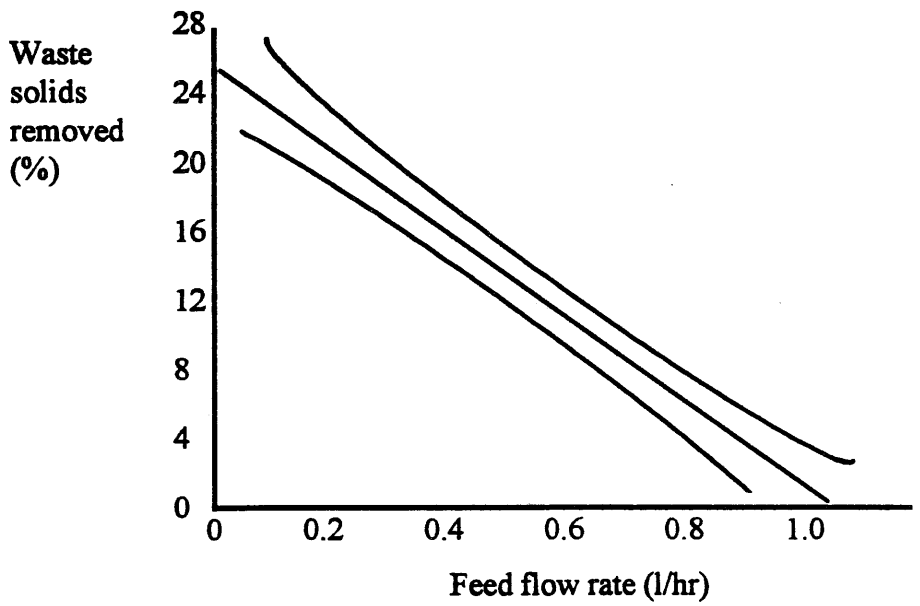

Figure 19.1 - Upper and lower 95% confidence limits for predicted responses.

In practice σ^2 in expression, (19.2) is estimated by $\hat{\sigma}^2$ as described previously. The resulting expression will be an estimate of the variance of $\hat{y}_k$.

Using (19.2), a $100(1-\alpha)$ per cent confidence interval for $E(\hat{y}_k)$ can then be calculated as

$$\hat{y}_k \pm t_{\nu,\alpha/2}\sqrt{\text{estimated variance of } \hat{y}_k} \qquad (19.4)$$

where again ν is the number of degrees of freedom associated with the estimate $\hat{\sigma}^2$.

For illustration, a 95 per cent confidence interval is now calculated for the expected value of waste solids removed at a feed flow rate of 0.3 l/hr. From (11.16) the fitted straight line model is

$$\hat{y} = 27.2 - 25.8x$$

so that the predicted waste solids removed at $x_k = 0.3$ is $\hat{y}_k = 19.46$. Using (19.1),

$$\begin{aligned}\text{est. var}(\hat{y}_k) &= \text{var}(\hat{\beta}_0) + (0.3)^2\,\text{var}(\hat{\beta}_1) + 0.6\,\text{cov}(\hat{\beta}_0, \hat{\beta}_1)\\ &= 0.268 + (0.3)^2(0.889) + 0.6(-0.419)\\ &= 0.0966\end{aligned}$$

Using the available internal estimate of pure error variance $\hat{\sigma}_p^2 = 0.698$ with 3 degrees of freedom a 95 per cent confidence interval for $E(\hat{y}_k)$ is

$$19.46 \pm t_{3,0.025}(0.0966)^2 = 19.46 \pm 0.99.$$

The outer lines in figure 19.1 define the upper and lower limits of 95 per cent confidence intervals for this example.

Because the matrix $(\mathbf{X}^T\mathbf{X})^{-1}$ in expression (19.2) is positive definite, the variance of a predicted response must increase as more terms are added to the model. To avoid unnecessarily poor precision for predicted responses, redundant parameters should be removed from a fitted model using the procedure described in section 17.

For an adequate model expression (19.4) is a confidence interval for both $E(\hat{y}_k)$ and $E(y_k)$. Although the range of plausible values for the population mean value of the response at specified operating conditions will be of interest to the experimenter, he will ordinarily be more interested in the range of plausible values for an actual measurement of

the response at those conditions, particularly if he intends to carry out further tests. If a future individual measured response value at operating conditions k is denoted as y_k^f, then from the fitted model, the least squares estimate of y_k^f is $\hat{y}_k$ and the (unknown) population mean value of y_k^f is $E(y_k)$.

There are two sources of uncertainty associated with y_k^f: (i) the uncertainty about the value of it's mean $E(y_k)$ and (ii) pure error fluctuation of y_k^f about it's (unknown) mean $E(y_k)$. Uncertainty about $E(y_k)$ can be represented by the variance of $\hat{y}_k$ and pure error fluctuation of y_k^f about its men can be represented by σ^2. The variance of y_k^f is then the sum of these two components,

$$\text{var}(y_k^f) = \text{var}(\hat{y}_k) + \sigma^2$$

or, using an estimate $\hat{\sigma}^2$,

$$\text{est. var}(y_k^f) = \text{est. var}(\hat{y}_k) + \hat{\sigma}^2 \qquad (19.5)$$

where the estimated variance of $\hat{y}_k$ is given by expression (19.2) using $\hat{\sigma}^2$ in place of σ^2.

A $100(1-\alpha)$ per cent confidence interval for the future individual measured response value y_k^f can now be expressed as

$$\hat{y}_k \pm t_{\nu,\alpha/2}\sqrt{\text{estimated variance of } y_k^f} \qquad (19.6)$$

This interval will naturally be larger than that for $E(\hat{y}_k)$ obtained from (19.4).

Again using the least squares straight line fit to the filtration data, the estimated variance of y_k^f at $x_k = 0.3$ l/hr feed flow rate is

$$\begin{aligned}\text{est. var}(\hat{y}_k) + \hat{\sigma}^2 &= 0.0966 + 0.698 \\ &= 0.795\end{aligned}$$

A 95 per cent confidence interval for y_k^f is then

$$19.46 \pm t_{3,0.025}\sqrt{0.795} = 19.46 \pm 2.84 .$$

CHAPTER 20

Interpretation of a Fitted Model

A mechanistic model is usually derived on the assumption that changes in the operating variables cause changes in the response variable. An empirical model fitted to process data also provides a description of the relationship between changes in selected operating variables and corresponding changes in the response variable but this relationship may not be a cause and effect relationship. Operating data can display systematic associations that defy logical explanation. Sometimes such relationships can be produced by what Box [1] has called a "lurking variable" which affects two or more variables simultaneously, thereby producing a systematic association between the affected variables. Extreme examples of such meaningless relationships in other fields have been cited by Huff [2].

Other obstacles prevent proper interpretation of routine operating data. In normal process operation, important operating variables are often held in tight control. Under these circumstances the effects of such operating variables on the response may be completely masked by random error. Sometimes routine changes in process variables are deliberately correlated. For example, a reduction in temperature may be regularly accompanied by a reduction pressure. In such cases it will be impossible to extract the individual effects of these operating variables on the response.

For these reasons, as Box [1] has advised, "to find out what happens to a system when you interfere with it, you have to interfere with it (not just passively observe it)".

"Interference" with a system to obtain specific information is accomplished most effectively and efficiently using experimental plans or strategies which are discussed in the following sections.

REFERENCES

[1] Box, G.E.P., (1966), "Use and abuse of regression", Technometrics, 8, pp 625-629.

[2] Huff, D. (1954), How to Lie with Statistics, Norton.