$$\Psi(x_i) = (2i - 1)/2n, \qquad (8.5.2.3)$$

where $\Psi(x)$ is the cumulative distribution function for the normal p.d.f. A plot of R_i against x_i should be a straight line with zero intercept and unit slope. A straight line with a slope greater than one suggests that the model is satisfactory, but that the variances of the data points have been systematically underestimated. Lack of fit is suggested if the curve has a higher slope near the ends, indicating that large residuals occur with greater frequency than would be predicted by the normal p.d.f.

The sorted residuals tend to be strongly correlated. A positive displacement from a smooth curve tends to be followed by another positive displacement, and a negative one by another negative one, which gives the *Q–Q* plot a wavy appearance, and it may be difficult to decide whether it is a straight line or not. Because of this, a useful alternative to the *Q–Q* plot is the conditional *Q–Q* plot (Kafadar & Spiegelman, 1986), so called because the abscissa for plotting the *i*th sorted residual is the mean of a conditional p.d.f. for that residual given the observed values of all the others. To construct a conditional *Q–Q* plot, first transform the distribution to a uniform p.d.f. by

$$U_i = \Psi(R_i, \mu, \sigma), \qquad (8.5.2.4)$$

where μ and σ are resistant estimates (Section 8.2.2) of the mean and standard deviation of the p.d.f., such as the median and 0.75 times the interquartile range, and Ψ represents the cumulative distribution function. Letting $U_0 = 0$ and $U_{n+1} = 1$, the expected value of U_i, given all the others, is

$$\langle U_i \rangle = (U_{i-1} + U_{i+1})/2. \qquad (8.5.2.5)$$

The *i*th abscissa for the *Q–Q* plot is then

$$x_i = \Psi^{-1}(\langle U_i \rangle, \mu, \sigma), \qquad (8.5.2.6)$$

where $\Psi^{-1}(y, \mu, \sigma)$ is a *per cent point function*, or p.p.f., the value of x for which $\Psi(x, \mu, \sigma) = y$.

Q–Q plots for subsets of the data can reveal, by nonzero intercepts, that those subsets are subject to a systematic bias. Because of its property of removing short-range kinks in the curve, the conditional *Q–Q* plot can be particularly useful in this application. The values of μ and σ used for the transformation to a uniform distribution, as in (8.5.2.4), should be those determined from the entire data set.

A *Q–Q* plot will reveal data points that are in poor agreement with the model, but that do not belong to any easily identifiable subset. Because of the central limit theorem (Section 8.4.1), however, the least-squares method tends to force the distribution of the residuals toward a normal distribution, and the discrepant points may not be clearly evident. A robust/resistant procedure (see Section 8.2.2), because it reduces the influence of strongly discrepant data points, helps to separate them from the body of the data. Therefore, if a data set contains discrepant points, a *Q–Q* plot of the residuals from a robust/resistant fit will tend to have greater curvature at the extremes than one from a corresponding least-squares fit. If the discrepant data points that are thus identified have a pattern, this information may enable a systematic error to be characterized.

8.5.3. Influential data points

Section 8.4.4 discusses the influence of individual data points on the estimation of parameters and how to identify the data points that should be measured with particular care in order to make the most precise estimates of particular parameters. The same properties that cause these influential data points to be most effective in reducing the uncertainty of a parameter estimate when the model is a correct predictor for the observations also cause them to have the greatest potential for introducing bias if there is a flaw in the model or, correspondingly, if they are subject to systematic error. Reviews of procedures for studying the effects of influential data points and outliers have been given by Beckman & Cook (1983), by Chatterjee & Hadi (1986), and by Belsley (1991).

The effects of possible systematic error can be studied by identifying influential data points and then observing the effects of deleting them one by one from the refinement. The deletion of a data point should affect the standard uncertainty of an estimate, but should not cause a shift in its mean that is more than a small multiple of the resulting standard uncertainty. As in Section 8.4.4, we define the design matrix, $\boldsymbol{A}$, by

$$A_{ij} = \partial M_i(\mathbf{x})/\partial x_j, \qquad (8.5.3.1)$$

where $M_i(\mathbf{x})$ is the model function for the *i*th data point, and $\mathbf{x}$ is a vector of adjustable parameters. Let $\boldsymbol{R}$ be the upper triangular Cholesky factor of the weight matrix, so that $\boldsymbol{W} = \boldsymbol{R}^T\boldsymbol{R}$, and define the weighted design matrix by $\boldsymbol{Z} = \boldsymbol{RA}$ and the weighted vector of observations by $\mathbf{y}' = \boldsymbol{R}\mathbf{y}$. The least-squares estimate of $\mathbf{x}$ is then

$$\widehat{\mathbf{x}} = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{y}', \qquad (8.5.3.2)$$

and the vector of predicted values is

$$\widehat{\mathbf{y}}' = \mathbf{Z}(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\mathbf{y}' = \boldsymbol{P}\mathbf{y}', \qquad (8.5.3.3)$$

where $\boldsymbol{P}$ is the projection, or hat, matrix. A diagonal element, P_{ii}, of $\boldsymbol{P}$ is a measure of the leverage, that is of the relative influence, of the *i*th data point, and therefore of the sensitivity of the estimates of the elements of $\mathbf{x}$ to an error in the measurement of that data point. P_{ii} lies in the range $0 \le P_{ii} \le 1$, and it has average value p/n, so that data points with values of P_{ii} greater than $2p/n$ can be considered particularly influential.

Let $\boldsymbol{H} = \boldsymbol{Z}^T\boldsymbol{Z}$ be the normal-equations matrix, let $\boldsymbol{V} = \boldsymbol{H}^{-1}$ be the estimated variance–covariance matrix, and let $\mathbf{q} = \boldsymbol{Z}^T\mathbf{y}'$, so that $\widehat{\mathbf{x}} = \boldsymbol{V}\mathbf{q}$. Let $\mathbf{z}_i$ be the *i*th row of $\boldsymbol{Z}$, and denote by $\boldsymbol{Z}^{(i)}$, $\boldsymbol{H}^{(i)}$, $\boldsymbol{V}^{(i)}$, $\mathbf{q}^{(i)}$, and $\widehat{\mathbf{x}}^{(i)}$ the respective matrices and vectors computed with the *i*th data point deleted from the data set. We wish to find large values of $|\widehat{x}_j - \widehat{x}_j^{(i)}|/[V_{jj}^{(i)}]^{1/2}$, so we need to compute $\boldsymbol{V}^{(i)}$ and $\mathbf{x}^{(i)}$. With a derivation similar to that for (8.4.4.7), it can be shown (Fedorov, 1972; Prince & Nicholson, 1985) that

$$\boldsymbol{V}^{(i)} = \boldsymbol{V} + \frac{\boldsymbol{V}\mathbf{z}_i^T\mathbf{z}_i\boldsymbol{V}}{(1 - \mathbf{z}_i\boldsymbol{V}\mathbf{z}_i^T)} = \boldsymbol{V} + \frac{\boldsymbol{V}\mathbf{z}_i^T\mathbf{z}_i\boldsymbol{V}}{(1 - P_{ii})}. \qquad (8.5.3.4)$$

Note that, if $P_{ii} = 1$, all elements of $\boldsymbol{V}^{(i)}$ become infinite, implying that $\boldsymbol{H}^{(i)}$ is singular. Thus, if such a data point is deleted, the solution is no longer determinate. Now,

$$\widehat{\mathbf{x}}^{(i)} = \boldsymbol{V}^{(i)}\mathbf{q}^{(i)} \qquad (8.5.3.5)$$

and

$$\mathbf{q}^{(i)} = \mathbf{q} - y_i'\mathbf{z}_i^T, \qquad (8.5.3.6)$$

so that, when $\boldsymbol{V}$ and $\widehat{\mathbf{x}}$ have been computed once, it is a straightforward and inexpensive additional computation to determine whether any parameter has been strongly influenced, and therefore potentially biased, by the inclusion of any data point in the refinement. If there is any reason to be concerned about possible systematic error, the leverage of every data point included in the refinement should be computed, and the effects of deletion of all of those with leverage greater than $2p/n$ should be observed.