6 ERROR MODELS AND ESTIMATION

Model specification usually includes a description of the behaviour of the mean in the distribution and a random effect or error. This error is the difference between the model expected value and the observed value, and can be classified as:

- **Structural error.** The estimation procedure uses structural relations between the populations at different times. These relations will only be approximately correct. For instance, CPUE is unlikely to be precisely proportional to abundance, although this model makes estimation easier.

- **Measurement error.** This error is not only caused by the sampling gear, (e.g. originating from variation in weather conditions during surveys), but also by patchy distributions of the population.

- **Process error.** Random effects may also affect underlying dynamics models. For instance, random weather and oceanographic changes may increase or decrease natural mortality and recruitment in an unpredictable way. The difference between process and measurement error is that process error introduces a real change in the system, whereas measurement errors introduce no underlying change and therefore do not affect future observations.

6.1 LIKELIHOOD

In fitting models to data, the models describing the stock and observations cease to become descriptions of the process in their own right, but instead become descriptions of how parameters in probability models change. For example, in many applications a model is used to describe the μ parameter in the normal distribution, which also happens to be the mean.

These probability models describe how likely the observed data are, given the parameters. The likelihood concept simply turns this on its head. Likelihood is the probability that a set of parameters is correct given the data. This makes no substantive difference to the probability distribution, but conceptually underpins most criteria for fitting models to data. For example, maximum likelihood defines a set of parameters when the likelihood function reaches its maximum point, and a Bayesian estimator uses likelihood, along with prior probabilities and a cost function, to define a set of parameters where the expected cost is minimised.

Although using likelihood (i.e. a full probability model) is theoretically better, least-squares is often used in the analysis of fisheries data. This can be justified by the following points:

- least-squares is maximum likelihood where the probability distribution is the Normal.

- true likelihoods can rarely be specified with any certainty. Least-squares with some appropriate transformation is probably as good as any alternative without

more information on error structure. Therefore, a more complex procedure may not be justified.

- maximum likelihood for many parametric likelihoods (e.g. Poisson etc.) can be reformulated in terms of least-squares.

- least-squares numerical procedures are much simpler and less likely to break down than more general approaches, particularly where there are many parameters to estimate.

The fact that least-squares are relatively easy to use, adaptable, objective and may often be close to the maximum likelihood solution has resulted in its wide use. Nevertheless, there is increasing interest in alternative approaches that are theoretically more appealing and for which robust numerical methods are being developed. However, even if other more complicated methods are used to fit models, in most cases least-squares can still form the starting point of the analysis, so the methods discussed here are always likely to be of use.

6.2 LEAST-SQUARES ESTIMATION

Least-squares estimation is based on the same principle as curve fitting. Given a set of observations, define a model and then establish the set of the parameters that gives the "best fit" of this model to the observed data. The "goodness of fit" is usually the sum of squared differences between observed and calculated dependent variables (i.e. squared residuals). The sum of squares (SSQ) is also sometimes called the Euclidean norm. The "best fit" is the set of parameters that minimises the squared difference between the observations and the model's expected values.

There is no guarantee that the "best fit" model is correct, and the model may be entirely inadequate for reflecting the dynamics. Analyses of a data set therefore should always include comparison between observations and the fitted model for inspection. This is often done graphically by plotting the residuals against the independent variables or by plotting on the same graph both the observations and the fitted curve. The residuals should show a random scatter and should not exhibit any remaining pattern.

Maximum likelihood estimation (e.g. Lehman 1983) includes the same elements as a least-squares estimation, only the goodness of fit measure will often differ based on the explicit assumption of the form of the error model.

The elements in least-squares estimation are:

The model: $y^{obs} = Model(\beta) + \varepsilon$ (47)

where β is the vector of parameters, and ε is the difference (error) between the observations and value calculated from the model. In practice this means all uncertainty is treated as observation error whether created by measurement noise, model mis-specification or otherwise. The mean of the observed quantity can be defined as:

$$E\left\{y^{obs}\right\} = Model(estimated\ \beta) = Model(\overline{\beta}) \tag{48}$$

Therefore, the mean error becomes: $E\{\varepsilon\} = 0$. Where the expected error is not zero, this is often referred to as bias.

For the goodness of fit measure, the sum-of-squares, is used:

$$L = \sum_{obs}\left[y^{obs} - Model(\beta)\right]^2 \tag{49}$$

The parameters in the model are chosen so that this sum-of-squares is at its minimum. This goodness of fit measure is often extended to account for the observations having different variances. In this case, the goodness-of-fit measure becomes:

$$L = \sum_{i}\sum_{obs}\frac{\left[y_i^{obs} - Model(\beta)\right]^2}{\sigma_i^2} \tag{50}$$

where data subset *i* contains all observations with the same variance σ_i^2.

A least-squares estimator will produce maximum likelihood estimates and confidence intervals if $\varepsilon = \varphi(0,\sigma^2)$, that is normally distributed with mean 0 and variance σ^2. It will even produce maximum likelihood estimates if the variance is constant over the range of explanatory variables or the proportional change in variance is known in a weighted least-squares scheme. However, if the true error distribution is not symmetrical, the variance changes in an unknown manner, or there are process or structural errors (as there almost certainly always are), the estimates will not be maximum likelihood.

It has been found with fisheries data that least-squares by itself provides a poor fit. For this reason, it is a common practice to use transformations to approximate alternative distributions. The transformations form part of the link model and often are used to represent alternative error distributions besides the normal. Elliott (1983) suggests the following transformations for stabilising the variance:

Error distributions	Observation	Transformation
Log-normal	*y*: continuous	$ln\ (y)$
Poisson	*y*: discrete	$1/y$
Binomial	*h*: frequency	$2\ arcsin(\sqrt{h})$
General frequency distribution	*h*: frequency	$ln(\ h/(1-h)\ )$
Taylor expansion	*y*: discrete	y^{α}

The most commonly assumed error is the log-normal, which is dealt with by taking logarithms of the data and then assuming that errors are Normally distributed. The use of the lognormal might be theoretically justified in some instances. For example, consider a cohort being subject to random survival rates between egg release and recruitment:

$$R_{\sum \Delta t} = R_0 e^{-\sum_i M_i \Delta t} \tag{51}$$

If the mortality is made up of a large sum of small random effects (M_i), the final total mortality, by the Central Limit Theorem, will be normally distributed even if the individual random components are not. Hence, this will result in a lognormal distribution.

As well as the practical observation that models fitted to log transformed data fit the data better, the log-normal has several other advantages:

- Taking logs often makes errors symmetrical around the mean. The Normal distribution does not discriminate against negative values, so for example, it allows for negative populations of fish, which are clearly impossible. In effect, this produces a bias towards larger observations in the analysis. The log-normal assumes negative values are impossible and corrects this bias. As a result, note that the exponent of the log-normal parameter, exp(μ), is not the same as the arithmetic mean, but lower as μ is the mean of the log values. The arithmetic mean will depend on both the log-mean (μ) and the variance.

- The log-normal does not assume a constant error variance, but assumes the variance increases with the arithmetic mean. Again, this has generally been observed in fisheries data, where increasing catches and effort tends to produce greater variability. The log-normal corrects estimates for this effect.

Nevertheless, the main argument for the log-normal remains pragmatic. It represents the observed error distribution better than would the normal, producing better estimates. However, you should always check which transformation if any is appropriate for your data by examining model residuals. The fact that a procedure is widely used is not a justification for its use in any particular case.

The model for which we want to estimate the parameters, provides the mean in the unknown probability distribution. Hence, for example, CPUE might be modelled as:

$$\ln CPUE = \ln q + \ln P + \varepsilon \tag{52}$$

where ε = measurement error.

Based on the above model we expect that the mean value, the logarithmic mean CPUE over many stations randomly spread out in the survey area, will be a linear function of the population P and that this mean value can be measured without bias. The estimation equation becomes:

$$\sum_a \sum_y \left[\ln Catch_{ay}^{obs} - \ln Catch_{ay}^{\mathrm{mod}}\right]^2 + \sum_i \sum_a \sum_y \left[\ln Cpue_{iay}^{obs} - \ln Cpue_{iay}^{\mathrm{mod}}\right]^2 + ... = MIN\{\text{model parameters}\} \quad (53)$$

where subscript i = abundance index, a = age and y = year of the observation. Most often there are several "tuning" data series available. A basic feature of the ADAPTIVE framework (Chapter 8) is to sum these individual contributions as in Equation 53. It is the researcher's responsibility to build the estimation equations relevant for each individual stock assessment.

6.2.1 Weights

When there is more than one "tuning" time series available the data of the different series are usually not obtained with the same measurement variance. In this case, it is preferable to introduce a weighting of the data series, by specifying a weight parameter, λ_i, for each data series i. In theory, these weights should be the inverse of the variance of the measurements. Estimates of the true variances are often available from abundance survey data, but they are more difficult to estimate for data from the commercial fisheries. However, it is unnecessary to obtain the absolute weights, only relative weighting with respect to some primary data series.

In least-squares theory, the variances can be estimated from the sum-of-squares as:

$$\frac{\sum_{\mathrm{obs}} \left[X^{\mathrm{obs}} - X^{\mathrm{mod}}\right]^2}{(n-p)} \quad (54)$$

where n is the number of observations and p is the number of parameters in the model (e.g. Lehmann 1983). It is not possible to estimate weights within the estimation procedure, only once the model is fitted. This is illustrated by a simple system with two "tuning" data series $CPUE_1$ and $CPUE_2$:

$$\sum_{\mathrm{obs}} \left[\ln CPUE_1^{\mathrm{obs}} - \ln CPUE_1^{\mathrm{mod}}\right]^2 + \lambda \sum_{\mathrm{obs}} \left[\ln CPUE_2^{\mathrm{obs}} - \ln CPUE_2^{\mathrm{mod}}\right]^2 = MIN \quad (55)$$

This sum-of-squares clearly has a minimum for $\lambda = 0$ ($\lambda \geq 0$), as this eliminates the second contribution to the sum of squares. Therefore, weights need to be treated as external variables estimated through some other means.

Extended Survivor Analysis (Darby and Flatman 1994) includes an internal weighting procedure, treating each age group and each data series separately. These weights are introduced in a double iteration inherent in the method. Details of this procedure are discussed in Section 8.3.

All data types above can be integrated in the combined estimation least-squares expression:

$$\sum_{a\in age}\sum_{y\in year}\left[\ln Catch_{ay}^{obs}-\ln Catch_{ay}^{\mathrm{mod}}\right]^2+$$
$$\sum_{i\in Abundance\ indices}\lambda_i^{CPUE}\sum_{a\in age}\sum_{y\in year}\left[\ln CPUE_{iay}^{obs}-\ln CPUE_{iay}^{\mathrm{mod}}\right]^2+$$
$$\sum_{i\in biomass\ indices}\lambda_i^{biomass}\sum_{y\in year}\left[\ln I_{iy}^{obs}-\ln I_{iy}^{\mathrm{mod}}\right]^2+ \tag{56}$$
$$\sum_{i\in effort\ indices}\lambda_i^{effort\ index}\sum_{y\in year}\left[\ln E_{iy}^{obs}-\ln E_{iy}^{\mathrm{mod}}\right]^2=MIN\{\mathrm{model\ parameters}\}$$

where λ_i are the weights applied to data series.

6.3 FINDING THE LEAST-SQUARES SOLUTION

Finding the least-squares solution is the common problem of finding the minimum for a function.

$$\sum_{obs}\left(y^{obs}-Model(\beta)\right)^2=MIN \tag{57}$$

This problem is converted into an equivalent problem of solving a set of simultaneous equations. In any function a minimum occurs where the partial differentials of the parameters are equal to zero, so:

$f(\beta)=MIN$ with respect to the set of β's is equivalent to

$\frac{\partial f(\beta)}{\partial \beta_i}=0$ *for* $i=1,2,...$

In least-squares, the function f is the χ^2 function (the sum-of-squares), and the parameters are the parameters of the model. Any numerical routine could be used to find a solution, and good robust routines exist. Why not just use the canned, black-box routines available in many software packages? The simple answer is no reason in many cases, and as long as the researcher checks such routines have successfully found the minimum, they are recommended. However, in some cases they are not adequate, particularly when the number of parameters is very large. Faster, more reliable and more accurate methods may be developed for a problem by considering the numerical solution yourself.

With large numbers of parameters, the N-dimensional parameter space can become very complex. Routines written for general functions can make no assumptions about those functions. They therefore tend to crawl around the parameter space very slowly to avoid overstepping the minimum. This may still not avoid missing the minimum and can take inordinate amounts of time. Routines to find the least-squares minimum take advantage of attributes of the χ^2 function, increasing the chance of success and the speed at which the minimum is found.

Canned black box routines are also widely available for finding least-squares, so why is the detail of methodology given here? The reason is largely the same. The researcher can take advantage of their knowledge of the function (i.e. the stock assessment model) to increase the chance of success and speed of the method. Any canned routine would treat the stock assessment model as a single function. However, a researcher will often see how the function could be broken down into simpler components, each amenable to simpler analysis. As will be seen, this approach is used in XSA, where a simple linear regression to estimate parameters of the model linking the population to the CPUE index, so these parameters can be solved by a single calculation. The parameters belonging to the more complex non-linear model still need to be found through iteration, but the number of parameters has been greatly reduced.

6.3.1 Linear Models

On the face of it, linear models would be of little use in stock assessment as most realistic population models are non-linear. However, there are often linear components, which can be estimated separately. The advantages of dealing with linear parts separately is purely pragmatic. Linear parameters can be found by calculation rather than iterative numerical procedures, which speeds up model fitting.

Where the model is linear, the least-squares equations are linear as well and can be solved directly through calculation. The solution is obtained by solving the M simultaneous linear equations, where M is the number of parameters or independent x variables. The solution of linear simultaneous equations is subject to standard linear algebra techniques. Assuming equal variances, we wish to find the solution to a set of M equations:

$$\frac{\partial L}{\partial \beta_i} = 0$$

where

$$L = \sum_{k=1}^{N} \left(y_k^{obs} - y_k^{\mathrm{mod}} \right)^2 \qquad (58)$$

$$y_k^{\mathrm{mod}} = \sum_{i=1}^{M} \beta_i x_{ik}$$

and there are M parameters and N data points. The set of differential equations can be found easily for a linear model:

$$\frac{\partial L}{\partial \beta_j} = -2 \sum_{k=1}^{N} \left(y_k^{obs} - y_k^{\mathrm{mod}} \right) \frac{\partial y_k^{\mathrm{mod}}}{\partial \beta_j}$$

$$\frac{\partial y_k^{\mathrm{mod}}}{\partial \beta_j} = x_{jk} \qquad (59)$$

$$\therefore \quad \frac{\partial L}{\partial \beta_j} = -2 \sum_{k=1}^{N} \left(x_{jk}\, y_k^{obs} - x_{jk} \sum_{i=1}^{M} \beta_i x_{ik} \right) = 0$$

So, now we have M equations in terms of the x and y data variables and the parameters $\{\beta\}$, each equation equal to zero. These are rearranged suitable for a matrix format:

$$\begin{Bmatrix} \sum_k x_{1k}x_{1k} & \sum_k x_{2k}x_{1k} & \cdots & \sum_k x_{Mk}x_{1k} \\ \sum_k x_{1k}x_{2k} & \sum_k x_{2k}x_{2k} & \cdots & \sum_k x_{Mk}x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_k x_{1k}x_{Mk} & \sum_k x_{2k}x_{Mk} & \cdots & \sum_k x_{Mk}x_{Mk} \end{Bmatrix} \begin{Bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_M \end{Bmatrix} = \begin{Bmatrix} \sum_k y_1 x_{1k} \\ \sum_k y_k x_{2k} \\ \vdots \\ \sum_k y_k x_{Mk} \end{Bmatrix} \quad (60)$$

All the terms in Equation 59 now appear in Equation 60, but arranged as matrices. The solution for β_i is conceptually simple. We multiply both sides of Equation 60 by the inverted matrix appearing on the left-hand side, isolating the $\{\beta\}$ vector. Reintroducing the variances for completeness, the solution can be written:

$$\{\beta\}_j = \sum_{i=1}^{M} \left[\left\{ \sum_{k=1}^{N} \frac{x_{jk}x_{ik}}{\sigma_k^2} \right\}^{-1}_{ij} \quad \left\{ \sum_{k=1}^{N} \frac{y_k x_{ik}}{\sigma_i^2} \right\}_i \right] \quad (61)$$

The σ^2_i is the variance associated with each data point, and often is assumed equal among data points. This matrix equation allows the least-squares estimate to be obtained in one iteration.

The advantage of linear models should now be apparent, and occurs because differentiation eliminates the parameter from the equation, enabling the linearity of the equations to be maintained. This allows an 'easy' solution. With non-linear models, the set of Equations 59 will not be linear, and therefore no simple solution exists.

In the simplest case with only one parameter, Equation 61 becomes:

$$\beta = \frac{\sum_{k=1}^{N} \frac{y_k x_k}{\sigma_k^2}}{\sum_{k=1}^{N} \frac{x_k^2}{\sigma_k^2}} \quad (62)$$

If the σ^2_i are constant among data points this equation becomes the sum of the product of the y and x variables divided by the sum of squares of the x variable. This result is often very useful in estimating parameters in models linking observed variables to underlying population dynamics variables. For example, consider the case where we have generated a population time series from a model, and we wish to relate it to a CPUE index which requires estimating a single parameter q, as:

$$CPUE_t = qP_t + \varepsilon \quad (63)$$

Assuming least-squares, constant variance and time independent errors, q can be calculated as:

$$q = \frac{\sum_{t=1}^{T} CPUE_t P_t}{\sum_{t=1}^{T} P_t^2} \tag{64}$$

This avoids the need to estimate q as part of the minimisation process. Instead q is calculated each cycle, and the numerical routine concentrates on solving the non-linear parameters associated with the population model.

A similar simple procedure can be undertaken with two parameters. In this case, we try to fit the logarithm of the CPUE to the log population size, assuming a non-linear relationship:

$$\begin{aligned} CPUE_t &= qP_t^v \\ \ln CPUE_t &= \ln q + v \ln P_t \end{aligned} \tag{65}$$

In this case, the parameter v is the slope, but we also have an intercept parameter ($\ln q$). In the linear framework, constant parameters (non-covariates are often called factors) are estimated using dummy variables. Dummy variables take on values of 1 or 0 depending on whether the parameter applies to any particular observation or not. In this case, the constant applies to all observations, so the first x variable is always 1:

$$\begin{aligned} &y_t^{obs} = \beta_1 x_{1t} + \beta_2 x_{2t} \\ &\text{where} \\ &y_t^{obs} = \ln CPUE_t \\ &x_{1t} = 1, \quad x_{2t} = \ln P_t \\ &\beta_1 = \ln q, \quad \beta_2 = v \end{aligned} \tag{66}$$

Solving for β_1 and β_2 using the general Equation 61 gives:

$$\begin{aligned} \beta_1 &= \frac{S_{22}S_{y1} - S_{12}S_{y2}}{S_{11}S_{22} - S_{12}S_{21}} \\ \beta_2 &= \frac{S_{11}S_{y2} - S_{21}S_{y1}}{S_{11}S_{22} - S_{12}S_{21}} \end{aligned} \tag{67}$$

where

$$
\begin{aligned}
S_{11} &= \sum_{t=1}^{T} (x_{1t})^2 = T \\
S_{22} &= \sum_{t=1}^{T} (x_{2t})^2 \\
S_{12} &= S_{21} = \sum_{t=1}^{T} x_{1t} x_{2t} = \sum_{t=1}^{T} x_{2t} \\
S_{y1} &= \sum_{t=1}^{T} x_{1t} y_t = \sum_{t=1}^{T} y_t \\
S_{y2} &= \sum_{t=1}^{T} x_{2t} y_t
\end{aligned}
\qquad (68)
$$

Notice that the subscripts refer to the elements of the matrices and vectors in Equation 60. S_{11} is the element in the first row and first column of the matrix on the left-hand side, and S_{y1} is the element in the first row of the vector on the right-hand side. For two parameters, the matrix inversion is very simple and it is possible to write out the result in a simple equation as above. This simplicity rapidly disappears with larger matrices. Inversion is closely related to calculating determinants, which is a sum involving all row-column combinations of elements. For large matrices these calculations are not trivial and may take a considerable amount of time, although the method remains faster (or at least more exact) than its non-linear cousin.

A second problem with inverting the matrix, is it may be singular. This may occur through aliasing, or high correlations pushing the inversion calculations beyond the computer's precision. An alternative solution to removing the offending parts of the model is to use matrix transformation techniques, notably Singular Value Decomposition (SVD). These techniques do not produce different results, merely skirt around singularity problems (Press *et al.* 1989). However, a detailed description falls beyond the scope of this manual.

This technique, of estimating linear parameters separate from the iterative fit, is widely used, in XSA for example. The benefits should not be underestimated. Searches for the minimum of non-linear functions is not trivial where there are large numbers of parameters and any method that reduces this number should be used.

6.3.2 Non-linear Models

A usual approach to finding the minimum is the Newton iteration scheme:

$$
\beta_i^{new} = \beta_i^{old} - \{H\}_{ij}^{-1} \left\{ \frac{\partial L(\beta^{old})}{\partial \beta_i} \right\}_i
$$

where (69)

$$
\{H\}_{ij} = \left\{ \frac{\partial^2 L}{\partial \beta_i \, \partial \beta_j} \right\}_{ij}
$$

$\{H\}_{ij}^{-1}$ is the inverse matrix of second partial derivatives of the sum of squares with respect to parameter pairs, often called the Hessian matrix.

As in the linear case, the aim is to find the point where the simultaneous partial derivative equations are zero. This approach to minimisation works on the principle that the step length moving parameters towards the zero point should be ratio of the differential to the slope of the differential (i.e. the second derivative) at the current position, which will produce a the correct step length where the model is linear. On each iteration, $\{\beta\}_i^{new}$ is generated and becomes the $\{\beta\}_i^{old}$ for the next cycle, so eventually $\{\beta\}_i^{new}$ converges to $\{\beta\}_i^{old}$ and the iterations can stop. The starting point for $\{\beta\}_i$ is important, but reasonable estimates are often available in VPA applications (e.g. F=0.5 year^{-1}).

It is only for a few problems when the second order derivatives are actually evaluated analytically. Instead computer-oriented methods are based on numerical approximations to the first and the second order derivatives, which are based on calculations of the function at small departures (h), e.g.

$$\frac{\partial L(\beta)}{\partial \beta_i} \cong \frac{L(\beta_i + h_i) - L(\beta)}{h_i} = \varphi_i \tag{70}$$

defines the first derivative and

$$\frac{\partial^2 L}{\partial \beta_i \partial \beta_j} \cong \frac{L(\beta_i + h_i, \beta_j + h_j) - L(\beta_i + h_i, \beta_j) - L(\beta_i, \beta_j + h_j) + L(\beta_i, \beta_j)}{h_i h_j} = H_{ij} \tag{71}$$

defines the second derivative and could therefore be used to calculate the Hessian matrix, $\{H\}_{ij}$. More sophisticated methods are usually used as these calculations may not be accurate (see Abramowich and Stegun 1966). An actual application will very often use standard implementations (see Press *et al.* 1989), which work with all but the most ill-behaved functions. However, we can take particular advantage of what is known about the least-squares function to improve both the speed and chance of success in finding the minimum.

The minimisation problem is first converted into the normal equations. Because the sum of squares is at a minimum point, we know that:

$$\frac{\partial L(\beta)}{\partial \beta_i} = -2 \sum_{obs} \left[y^{obs} - Model(\beta) \right] \frac{\partial Model(\beta)}{\partial \beta_i} = 0 \quad \text{for i} = 1,2,...,\text{p} \tag{72}$$

Likewise, explicit differentiation to produce the Hessian terms gives:

$$\begin{aligned} \frac{\partial^2 L(\beta)}{\partial \beta_i \, \partial \beta_j} = \quad & -2 \quad \sum_{obs} \left(y^{obs} - Model(\beta) \right) \frac{\partial^2 Model(\beta)}{\partial \beta_i \, \partial \beta_j} \\ & +2 \quad \sum_{obs} \frac{\partial Model(\beta)}{\partial \beta_i} \frac{\partial Model(\beta)}{\partial \beta_j} \end{aligned} \tag{73}$$

The first term in Equation 73 containing the second partial derivative is generally ignored in estimating the Hessian matrix for two reasons. Firstly, the second derivatives are often small compared to the first derivatives (they are zero in linear models for example), so their inclusion may not improve the efficiency of the fitting. Secondly, in practice the first term will sum to a small value when $Model(\beta)$ estimates are close to the expected value of the y^{obs} (i.e. the mean). Therefore, the procedure may be most efficient when the initial estimates are reasonably close to the best-fit estimates. The fitting process now becomes:

$$\{\beta\}_i^{new} = \{\beta\}_i^{old} - \left\{ \sum_{k=1}^{N} \frac{\partial Model(\beta)}{\partial \beta_i} \frac{\partial Model(\beta)}{\partial \beta_j} \right\}_{ij}^{-1} \left\{ \frac{\partial Model(\beta)}{\partial \beta_i} \right\}_i \tag{74}$$

In some cases, the Hessian in this form may be easy to obtain analytically. For example, notice that where the model is linear, the Hessian matrix is the same as that in Equation 60. Using the true differentials should improve the efficiency of the fit. In other cases, Equation 74 will not help as the first derivative is just too complicated to derive and simple numerical methods (e.g. Equation 70) are instead used to generate both the Hessian matrix and vector of first derivatives.

The scheme proposed in Press *et al.* (1989) is the Levenberg-Marquardt method, which uses either the Hessian matrix or a simple step routine where the Hessian is a poor approximation to the shape of the χ^2 function. Although this approach should be used in many cases, it may well still be worthwhile exploring the simultaneous partial differential equations and Hessian matrix. While it may not be worthwhile pursuing the analytical approach, some analysis may help in understanding the behaviour of the model and potential pitfalls in attempting to find the least-squares solution numerically.

6.4 ESTIMABLE PARAMETERS

While it is possible to formulate a least-squares function for any model it does not follow that all parameters can be estimated. This can be inherent in the model formulation or it can be because of a lack of sufficient information.

An example of a model that cannot be fully identified is where parameters multiply or add together in a way that cannot be separated by the data collected, such as $Model(\beta)=\beta_1 \beta_2 x$, where only the product of the two parameters can be estimated. In fishery biology, an example is the population equation:

$$N_{a+1,y+1} = N_{ay} e^{-M_a - F_{ay}} \tag{75}$$

The equation contains such a problem in parameters F_{ay} and M_a unless data can be brought to bear to estimate these parameters separately. It is this basic problem that explains the minimum data requirement for an analytical assessment. To separate the two, the catch in numbers by age and by year combined with observations on either the fishing mortality or the stock in numbers are required. Usually M_a is just fixed as an external parameter.

It is not only the model structure that makes certain parameters inestimable. The data structure can also have features that prevent the estimation of all parameters. This is the collinearity problem, indicated by high parameter correlation estimates. In extreme cases, parameters may be “aliased” which implies the data are inadequate to provide separate parameter estimates.

A simple example where such correlation occurs is in estimating fishing power based on vessel characteristics. Most characteristics are dictated by vessel size. So the size of net, vessel speed, hold size, number of crew, sophistication of gear all relate back to the size of vessel. In essence, because we do not have observations on catch rates of large vessels with small engines or small vessels with large engines, it is not possible to separate the effects of engine size and vessel length. What appears to be a large amount of data, all the different characteristics of the fleet, boils down to very little real information to separate vessels. Methods such as principle components analysis should be used to reduce a large number of correlated variables into a few representative uncorrelated components for this type of analysis.

A more worrying example for stock assessment is the possible relationship between stock size and catchability. Vessels aggregate in areas where catchability is highest. Fish aggregate to improve spawning success and minimise their natural mortality. There are several cases where it is suspected that as the population decreases, fish density on the fishing grounds remains constant, so effectively catchability is increasing as the population falls.

While correlations in linear models are relatively straightforward, it is much more complicated in non-linear models such as those used in fish stock assessment. It is not clear how terminal cohort sizes might be correlated with catchability estimates for CPUE indices before doing a full analysis.

Statistical experimental design ensures that such collinearity does not occur in experimental data. However in fisheries or oceanographic surveys, the researcher does not have the same degree of control over the system under investigation and such data, because of the oceanographic or biological links occurring in nature, often show some degree of correlation between the independent variables.

6.5 ROBUST REGRESSION

An alternative approach to least-squares is to apply “robust regression” (e.g. see Chen and Paloheimo 1995). The least-squares fit is based on minimising the squared sum of residuals and this sum can be strongly dependent on a few outliers (cf. the example above). Robust regression exists in different forms, but is based on either replacing the sum of squares of the residuals by some other measure of “goodness of fit”, (e.g. the median) or ignoring a certain percentage of the largest residuals in the fitting procedure (trimmed LSQ). Using the median, the least-squares problem is reformulated to finding the best curve where 50% of the observations have positive and 50% negative residuals. Obviously the magnitude of the residuals is of no importance and therefore outliers have less influence on the final result than when normal least-squares is applied. The approach can be formulated based on fitting a model as:

$$Median\left\{\left(y_i - Model(\beta_i)\right)^2, i = 1,2,....\right\} \quad (76)$$

6.6 CATCH

In many applications catch errors are either ignored (e.g. in most ADAPT and XSA methods) or the errors are assumed to be log-normal (e.g. in the ICA or in the CAGEAN methods). The reason for ignoring these errors in the catch data is that the stochastic error in the catch data is often insignificant compared to the noise in the survey data. This is probably correct in many fish stock assessments, but only for the more abundant age groups. The number of old fish caught, if constituting only a few percent of the total catch, is unlikely to be precisely estimated.

Methot (1990) suggested as part of his "Synthetic Model" that the error structure of the catch data be decomposed into two contributions:

- Estimate of the overall catch in weight
- Estimate of the age composition

The first contribution can be assumed to have lognormal errors. For the second contribution, Methot (1990) suggests that a multinomial distribution is more appropriate. The estimation of the catch in numbers, C, is often obtained through a fisheries statistics programme that provides total landings by species and by time period supplemented by a biological sampling programme that takes a length sample (n_l) and an age-length key (m_{la}) (ALK). The estimation of the catch-at-age for the population model is:

$$C_a = C_\bullet \sum_l \frac{n_l}{n_\bullet} \frac{m_{la}}{m_{l\bullet}} \quad (77)$$

where the dot subscript indicates summation over that subscript (Lewy and Lassen 1997).

However, where the observed and expected catch is included as part of the sum-of-squares, we can use the age composition observations directly. In the simple case, the age distribution is a random sample of the catches with observed frequencies in numbers of fish, h_{ay}. The catch composition in the model is given by:

$$\Theta_{ay} = \frac{C_{ay}^{\text{mod}}}{C_{\bullet y}^{\text{mod}}} \quad (78)$$

$$\text{where} \quad \sum_a C_{ay}^{\text{mod}} = C_{\bullet y}^{\text{mod}} \qquad \therefore \sum_a \Theta_{ay} = 1$$

Therefore Θ_{ay} is essentially independent of the total catch. In this case, the catch term contribution:

$$\sum_a \sum_y \left[\ln C_{ay}^{obs} - \ln C_{ay}^{\mathrm{mod}}\right]^2 \tag{79}$$

with the multinomial likelihood becomes:

$$\sum_y \left[\ln C_{\bullet y}^{obs} - \ln C_{\bullet y}^{\mathrm{mod}}\right]^2 + \sum_{a,y} h_{ay} \ln \Theta_{ay} + ... = MIN\{\text{model parameters}\} \tag{80}$$

Fortunately, because only the age sample, but not the catches, appear with the Θ_{ay} parameters, they can be estimated independently of the catch data and population model by finding the maximum of the multinomial likelihood function. Combining length sampling with ALK gives a similar result, but the formulae are more complicated (Lewy and Lassen 1997).

The variance can be found as:

$$\frac{Var(C_a)}{C_a^2} = \frac{Var(C_\bullet)}{C_\bullet^2} + \frac{Var(h_a)}{h_a^2} \tag{81}$$

for the simple situation when the age sample is a random sample of the catch. If it can be assumed that the variance contribution from the total landings can be neglected compared to the error due to ageing, and if the ageing error can be approximated by a multinomial distribution, then this can be simplified to:

$$Var(C_{ay}) \cong \frac{1}{n}\frac{C_{ay}}{C_{\bullet y}}\left(1 - \frac{C_{ay}}{C_{\bullet y}}\right) \Bigg/ \left(\frac{C_{ay}}{C_{\bullet y}}\right)^2 = \frac{1}{n}\frac{1 - \frac{C_{ay}}{C_{\bullet y}}}{\frac{C_{ay}}{C_{\bullet y}}} \tag{82}$$

where n is the number of fish in the sample.