

Single Hauls 2

Parameterisation

In the previous session we looked at a number of different selectivity curves. The formulas presented there were all parameterised by use of α and β (and an additional δ parameter for the Richards curve). These are called generic parameterisations. In the context of selectivity research it is however often more of interest to focus on the so-called selectivity parameters (L_{50} ; SR). L_{50} is defined as the length at which 50% of the fish are retained. Analogously L_{25} and L_{75} are the lengths at 25% and 75% retentions respectively. SR is defined by $L_{75} - L_{25}$.

There is a one-to-one correspondence between the generic parameters and the selectivity parameters for the four 2-parameter curves. (The same holds true for Richards curve given the δ parameter, but it will not be dealt with here). It is therefore possible to express the curves in terms of (L_{50} ; SR) rather than (α ; β). These expressions are the so-called selectivity parameterisations.

Curve	Functional form	α	β
Logit	$\frac{\exp\left(2 \cdot \log(3) \frac{\ell - L_{50}}{SR}\right)}{1 + \exp\left(2 \cdot \log(3) \frac{\ell - L_{50}}{SR}\right)}$	$-\frac{L_{50}}{SR} \cdot 2 \log(3)$	$\frac{2 \cdot \log(3)}{SR}$
Probit	$\Phi\left(2 \cdot \Phi^{-1}(0.75) \cdot \frac{\ell - L_{50}}{SR}\right)^*$	$-\frac{L_{50}}{SR} \cdot 2 \cdot \Phi^{-1}(0.75)$	$\frac{2 \cdot \Phi^{-1}(0.75)}{SR}$
Log-Log	$\exp\left(-\exp\left(-\left(k_1 + k_2 \frac{\ell - L_{50}}{SR}\right)\right)\right)^{**}$	$k_1 - k_2 \cdot \frac{L_{50}}{SR}$	$\frac{k_2}{SR}$
C-Log-Log	$1 - \exp\left(-\exp\left(-k_1 + k_2 \cdot \frac{\ell - L_{50}}{SR}\right)\right)^{**}$	$-k_1 - k_2 \cdot \frac{L_{50}}{SR}$	$\frac{k_2}{SR}$

* Φ is the Probability function for the standard normal distribution

** $k_1 = -\log(\log(2))$ and $k_2 = \log\left(\frac{\log(0.25)}{\log(0.75)}\right)$

Apart from the scientific interest in selectivity parameters rather than generic parameters there is another reason for preferring this parameterisation. Depending on the scale used the selectivity curves are often located well to the right on the length-axis. This results in α and β often being highly correlated. Since L_{50} is typically close to the mid-point of the observed length classes a selectivity parameterisation can be considered as a centralisation of the curve. It does therefore often present a higher numerical stability. In practical work, a selectivity parameterisation may sometimes provide results where the generic parameterisation fails to converge.



Sub-sampling

Catches are often quite large and total measurements may become insurmountable. In such cases only a sub-sample is measured. There are various strategies for taking sub-samples (by weight, by counts or otherwise). The strategies will not be dealt with here. Instead we will assume that the sub-sample to the total catch can be represented by some sub-sampling ratio q . Note that it is implicitly assumed that all length classes have been sub-sampled by the same ratio. Later this constrain will be relaxed and we will consider length-based sub-sampling.

Traditionally data from a sub-sampled catch have been transformed by raising the counts relatively to q . The argument for this is that the raised numbers represent what would have been measured had all fish been measured. By taking a sub-sample an additional uncertainty is however introduced into the data. If say sub-sampling ratio of 5% was applied to a given catch and only 10 fish of a given length class was measured the raised number would be 200. Using a number of 200 in the analysis is pretending that 200 fish were actually measured. Since the standard deviation of the estimate decreases with increasing sample size it is clear that this approach affects the uncertainty with which the parameters are estimated in an erroneous way.

A different approach to handling sub-sampled data is to consider q as an effort parameter. Continuing the example above, we would present the measurement of the 10 fish as a result of applying an effort of 5%.

The SELECT model is modified accordingly. The principle can be illustrated by considering the covered codend catch. Assume that the codend and the cover have been sub-sampled by ratios q_{codend} and q_{cover} respectively. Recall from the previous note that the number of fish in the two compartments were assumed Poisson distributed with rates:

- codend: $r(\ell) \cdot \lambda_\ell$
- cover: $(1 - r(\ell)) \cdot \lambda_\ell$

From SPR Theorem 2 it follows that the number of fish measured from the two compartments are

- codend: $q_{\text{codend}} \cdot r(\ell) \cdot \lambda_\ell$
- cover: $q_{\text{cover}} \cdot (1 - r(\ell)) \cdot \lambda_\ell$

Instead of considering the proportion of the total catch caught in the codend we now consider the proportion of the total number of fish measured taken from the codend:

$$\begin{aligned} \Phi(\ell) &= \frac{q_{\text{codend}} \cdot r(\ell) \cdot \lambda_\ell}{q_{\text{codend}} \cdot r(\ell) \cdot \lambda_\ell + q_{\text{cover}} \cdot (1 - r(\ell)) \cdot \lambda_\ell} \\ &= \frac{q_{\text{codend}} \cdot r(\ell)}{q_{\text{codend}} \cdot r(\ell) + q_{\text{cover}} \cdot (1 - r(\ell))} \\ &= \frac{r(\ell)}{r(\ell) + \gamma \cdot (1 - r(\ell))} \end{aligned}$$

where $\gamma = \frac{q_{\text{cover}}}{q_{\text{codend}}}$.

In particular if $r(\cdot)$ is the logistic curve we get:

$$\begin{aligned}\Phi(\ell) &= \frac{\exp(\alpha + \beta \cdot \ell)}{\exp(\alpha + \beta \cdot \ell) + \gamma} \\ &= \frac{\exp(\alpha - \log(\gamma) + \beta \cdot \ell)}{\exp(\alpha - \log(\gamma) + \beta \cdot \ell) + 1} \\ &= \frac{\exp(\tilde{\alpha} + \beta \cdot \ell)}{\exp(\tilde{\alpha} + \beta \cdot \ell) + 1}\end{aligned}$$

Proof: Exercise.

The expression in the last line is itself a logistic function. Consequently sub-sampled data from a covered codend experiment can be analysed using the logistic function can be analysed without any changes with a sub-sequent adjustment of α . Note however that this pertains to this particular case only.



The model is easily amended to handle length-based sub-sampling. This is achieved by adding a length index to the γ parameter:

$$\Phi(\ell) = \frac{r(\ell)}{r(\ell) + \gamma_\ell \cdot (1 - r(\ell))}$$



The model is modified similarly for data from experiments with paired gears.



Raising data mainly affects the variances of the estimates, whereas the parameter estimates tend to be about the same.

Model Check

After fitting a curve to the data, the fit should be carefully examined. The most important tools for doing this are the deviance statistic and inspection of the residuals. They are closely related and also closely related to the likelihood.

Residuals

The deviance residual for length class ℓ is defined by

$$r_{\ell} = \text{sign}(\pi_{\text{obs},\ell} - \hat{\pi}_{\ell}) \cdot \sqrt{\left(2 \cdot n_{\ell,+} \left(\hat{\pi}_{\ell} \cdot \log\left(\frac{\pi_{\text{obs},\ell}}{\hat{\pi}_{\ell}}\right) + (1 - \hat{\pi}_{\ell}) \cdot \log\left(\frac{1 - \pi_{\text{obs},\ell}}{1 - \hat{\pi}_{\ell}}\right) \right)\right)}$$

Under the model the deviance residuals are independent and (asymptotically) standard normally distributed (i.e., mean 0 and variance 1). Inspection of the residuals should therefore focus on potential systematic patterns (long runs of positive or negative residuals) and also if too many residuals exceed the ± 2 boundaries.

There are others (and more intuitive) residuals. The deviance residual is however generally preferred for non-normal data due to its asymptotic properties.

Deviance

The deviance is a single statistic that measures the overall departure from the model. It is defined as the sum of squares of the deviance residuals, defined above:

$$D = \sum_{\ell} r_{\ell}^2$$

From SPR Theorem 3 we get $D \sim \chi^2(\text{dof})$, where *dof* is L minus the number of parameters. This property should however not be taken too rigorously. A good rule of thumb is that D should be about the same as *dof*. It is commonly observed that biological data show some degree of over-dispersion; i.e. $D > \text{dof}$. Under-dispersion can occur, but is less common.

Variations

Estimating the variances of the parameters is an essential part of the inference. It is however not an easy task to achieve. In the maximum-likelihood framework the variances are estimated as the

inverse of the observed information $I(\theta) = -\left\{ \frac{\partial^2 l}{\partial \theta^2} \right\}_{\theta = \hat{\theta}}$ (See SPR)

Most often the interest focuses on the selectivity parameters ($L50, SR$) rather than the generic parameters (α, β) . The variances of these estimates can either be estimated by first estimating the variance matrix for (α, β) and then use the delta theorem to convert these into (approximate) variance estimates for $(L50, SR)$. This approach is well described in the ICES manual.

If however a selectivity parameterisation is used in the likelihood function, variance estimates for $(L50, SR)$ can be obtained in a more direct (and more precise) manner.

Here we only consider the covered codend-case with un-sampled data.

First we denominate the entries in the information matrix:

$$I(\theta) = \begin{Bmatrix} A_{11} & A_{12} \\ A_{12} & A_{22} \end{Bmatrix}$$

Note that it is symmetrical, so that only three entries must be calculated

Set $\eta = \frac{2\log(3)}{SR}(\ell - L_{50\%})$ and produce some temporary variables

$$t_1 = \frac{\exp(\eta)}{(1 + \exp(\eta))^2}$$

$$t_2 = \frac{1 - \exp(\eta)}{1 + \exp(\eta)}$$

$$t_3 = \frac{n_{\ell, \text{codend}}}{r(\ell)} - \frac{n_{\ell, \text{cover}}}{1 - r(\ell)}$$

$$t_4 = \frac{n_{\ell, \text{codend}}}{r(\ell)^2} - \frac{n_{\ell, \text{cover}}}{(1 - r(\ell))^2}$$

(The dependence on length-classes has been suppressed)
 The entries in the information matrix is now given by

$$A_{11} = \sum_{\ell} t_1 \cdot \frac{\log(18)}{SR^2} \cdot (t_1 \cdot t_4 - t_2 \cdot t_3)$$

$$A_{12} = \sum_{\ell} t_1 \cdot \frac{\log(9)}{SR^2} \cdot (t_1 \cdot t_4 \cdot \eta - (t_2 \cdot \eta + 1) \cdot t_3)$$

$$A_{22} = \sum_{\ell} t_1 \cdot \frac{\eta}{SR^2} \cdot (t_1 \cdot t_4 \cdot \eta - (t_2 \cdot \eta + 2) \cdot t_3)$$

Now the matrix can be inverted to produce estimates of the covariance-matrix:

$$\begin{pmatrix} \sigma_{L50}^2 & \sigma_{L50,SR} \\ \sigma_{L50,SR} & \sigma_{SR}^2 \end{pmatrix} = \frac{1}{A_{11} \cdot A_{22} - A_{12}^2} \begin{pmatrix} A_{22} & A_{12} \\ A_{12} & A_{11} \end{pmatrix}$$