Lecture 5

The Maximum Likelihood method is essentially the only estimation method used in tag recovery or capture recapture models.

Theory goes back to R. A. Fisher in 1912 (when he was a third year undergraduate!) and a flurry of later papers. Actually, Edwards (1974) traces the history of such thinking back to Daniel Bernoulli and Johann Lambert in the 18th century. Maximum likelihood is the primary estimation method in statistics.

Likelihood Theory is the Foundation of Statistics!

The probability of mortality \((1-S)\) includes harvest mortality (including fish killed, but not retrieved) and “natural” mortality.

\(S\) is the probability of survival from one tagging period to the next. \(1-S\) is the probability of mortality and includes both harvest and “natural” mortality. This seems odds, at first, because the data relate to only harvest mortality that is reported.

Generally, tag recovery data do not allow estimates of the 2 components (i.e., harvest vs. natural mortality).

Note that the estimates of survival \((S)\) or mortality \((1-S)\) are based on the data from the harvest!

Haunting Question: If one had only \(R_1\) and \(m_{11},\ m_{12},\ m_{13},\ m_{14},\ \text{and}\ m_{15}\), could they compute MLEs of \(S\) and \(r\)? Why not? Under what model?

Can you see where the assumptions in these models and estimation methods come from? Why are the assumptions explicit?

In the analysis of real data, one must be very concerned about the validity of the assumptions made. If an assumption fails, the method may be very dependent upon the truth of the assumption and large bias in the MLEs may result.

For example, what if 15% of tagged animals lose their tags within a month of being tagged. What effect on \(\hat{S}\) might this have? Why? How could you find out?

Identifiability (a nasty issue)
There are PIMs that specify models to *MARK* that cannot be "identified." That is, the data do not permit the estimation of some parameters. This is an inherent lack of information that keeps some parameters from being estimated.

The issue is like the estimation of the regression model

\[ E(y) = \beta_0 + \beta_1(x) \]

when the sample size \( n = 1 \) (only one sample of \( y \) and \( x \)). Here, the model parameters cannot be identified because there is an infinite number of lines that can be drawn through a point. Of course, if one had a sample size of 2, then a unique line would be defined and the 2 parameters could be estimated (or "identified").

In the fish tagging or bird band recovery models, identifiability of survival would not be an issue if fish were tagged in only a single year AND reporting probability was a constant across years. Consider the bass data, for example,

\[
\begin{array}{cccccc}
2000 & 30 & 70 & 114 & 43 & 15 & 1728,
\end{array}
\]

corresponding to the cell probabilities under model \( \{S, r\} \);

\[
R_1 \quad (1-S)r \quad S(1-S)r \quad SS(1-S)r \quad SSS(1-S)r \quad SSSS(1-S)r^+ .
\]

One can see that \( S \) can be estimated (not an MLE) as

\[
\hat{S} = \frac{m_{14}}{m_{13}} = \frac{SSS(1-S)r}{SS(1-S)r} = S .
\]

Note, the \( r \) in the numerator and denominator cancel, as do the terms \((1-S)\) and 2 of the \( S \), leaving \( \hat{S} = S \). Numerically \( \hat{S} = 43/114 = 0.377 \). So, under this simple model, the constant survival probability can be estimated, if \( r \) is also constant across time.

Under models where \( r \) is allowed to vary by year or age, then identifiability is lost, unless more than one cohort is tagged and released.

A common headache in model \( \{S_t, r_t\} \) is the lack of identifiability of the terms shown below in bold:
Number Tagged

Matrix of Probabilities for $m_{ij} \ E(m_{ij}/R_i)$

<table>
<thead>
<tr>
<th></th>
<th>$S_1(1-S_2)r_2$</th>
<th>$S_1S_2(1-S_3)r_3$</th>
<th>$S_1S_2S_3S_4(1-S_4)r_4S_1S_2S_3S_4(1-S_5)r_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_1$</td>
<td>(1-$S_1$)$r_1$</td>
<td>$S_1S_2(1-S_3)r_3$</td>
<td>$S_1S_2S_3S_4(1-S_4)r_4S_1S_2S_3S_4(1-S_5)r_5$</td>
</tr>
<tr>
<td>$R_2$</td>
<td>(1-$S_2$)$r_2$</td>
<td>$S_2(1-S_3)r_3$</td>
<td>$S_2S_3S_4(1-S_5)r_5$</td>
</tr>
<tr>
<td>$R_3$</td>
<td>(1-$S_3$)$r_3$</td>
<td>$S_3(1-S_4)r_4$</td>
<td>$S_3S_4(1-S_5)r_5$</td>
</tr>
<tr>
<td>$R_4$</td>
<td>(1-$S_4$)$r_4$</td>
<td>$S_4(1-S_5)r_5$</td>
<td></td>
</tr>
<tr>
<td>$R_5$</td>
<td></td>
<td></td>
<td>$(1-S_5)r_5$</td>
</tr>
</tbody>
</table>

In this case, only the product $(1-S_5)r_5$ is identifiable, but not the separate terms. Thus, this model has 4 survival probabilities, 4 reporting probabilities and one product term that can be identified under model $\{S_i, r_i\}$. Total, $K = 9$ (not 10, as you might think/want).

This subject will haunt us continually and more insights will be provided (the concepts of sufficient and minimal sufficient statistics and their dimensionality). Program MARK has clever ways to help understand this issue, but is not perfect for complicated or ill-conditioned models.

**Stupid Assumptions**

Problems occur when the estimation of parameters is attempted when animals are not marked. Assumptions must be made that are without biological basis or support. The classic life table methods are a prime example of a method lacking validity when faced with real-world assumptions (recall the $\ell_x, d_x, m_x$ material you were fed in beginning biology classes?).

Examples of invalid assumptions:
- population stability
- population stationarity
- sampling probabilities constant over years
- age or size classes
- gender.

Such “assumptions" are simply stupid, given what is known about vertebrates and invertebrates. Methods based on such assumptions do not lead to valid inferences nor do they hold up to tests of their assumptions (see Burnham and Anderson 1979 and Anderson et al. 1981; both appear in Appendix C of Brownie et al. 1985).
GOODNESS OF FIT

The deviance can be used as a measure of goodness-of-fit for band recovery models. The deviance for model $j$ is defined as

$$\text{deviance} = -2 \log_e(L_j(\hat{\theta})) + 2\log(L_{sat}(\hat{\theta})),$$

evaluated at the MLEs for model $j$ and the saturated model (the saturated multinomial model is a model with as many parameters as cells). Assuming “large” sample size, the deviance is distributed as a $\chi^2$ random variable with degrees of freedom equal to the number of cells minus the number of estimable parameters.

Two other standard procedures that will be useful in understanding the fit of a model to a data set. The traditional goodness-of-fit test is based on the Pearson statistic, which is asymptotically $\chi^2$ distributed;

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}},$$

where $O_{ij}$ are the observed values and $\hat{E}_{ij}$ are the expected values under a particular model. This approach lacks the additivity property associated with the G-statistic (below). Fully developed GOF tests do exist for tag recovery models; the theory of these GOF tests is well established (Brownie et al. 1985, Burnham et al. 1987). Program MARK does produce basic GOF information in the form,

$$\chi^2 = \sum \frac{(m_{ij} - \hat{E}(m_{ij}))^2}{\hat{E}(m_{ij})}.$$

The estimated cell expectations are based on the specific model form and the MLEs of the parameters in that particular model. The array of tag recovery counts and the array of $\hat{E}(m_{ij})$ are given at the end of the output from MARK for recovery data.

The less well known, but preferred, test is based on likelihood theory, often called the G test, and is also distributed as $\chi^2$ when sample size is suitably large,

$$\chi^2 = 2 \sum_i \sum_j O_{ij} \cdot \log_e \left( \frac{O_{ij}}{\hat{E}_{ij}} \right).$$

This general approach for the banding/tagging models is
\[ \chi^2 = 2 \sum_i \sum_j m_{ij} \cdot \log_e \left( \frac{m_{ij}}{E_{ij}} \right), \]

where \( m_{ij} \) represents the matrix of tag recoveries and the matrix of expected values is model-specific. For example, under model \( \{S_t, r_t\} \),

\[ \hat{E}_{14} = N_1 \hat{S}_1 \hat{S}_2 \hat{S}_3 (1-\hat{S}_4) \hat{r}_4 \]

and under model \( \{S, r\} \),

\[ \hat{E}_{14} = N_1 \hat{SS} (1-\hat{S}) \hat{r}, \]

of course, the parameters estimates here are all MLEs.

In both cases, MARK provides test statistics where some pooling has been done to improve the approximation to \( \chi^2 \) when the expected values for some cells are small (say, below 2). Under either approach, the cell counts \( m_{ij} \) and expectations \( \hat{E}(m_{ij}) \) are pooled. This pooling affects the degrees of freedom and the P-value of the GOF test. These GOF test results from MARK can be improved on some for sparse data by more exact GOF methods. However, the GOF information for recovery data produced by program MARK is quite useful and worth examining if you have concerns about model fit.

Both of these statistics are approximations to the deviance and all three are somewhat asymptotically equivalent. The advantage of the traditional Pearson statistic and the G statistic is that individual cells can be examined for lack of fit and patterns in the lack of fit. This helps understand the data and the model in question.

The degrees of freedom for these tests is always the number of cells – \( K \). The number of cells in these banding/tagging models is \( \left( k(k+1) \right) / 2 \) for data sets where the number of years of banding \( k \) equals the number of years of reporting (“triangular”). The column showing the number of animals never reported is ignored in computing the number of cells. The number of estimable parameters \( K \) must include terms such as \( S_{k-1} r_k \) as one parameter (not two); it is only the product that can be estimated.

The hypotheses for these goodness-of-fit tests are:

\[ H_0: \text{ the model fits} \]

\[ H_A: \text{ the model does not fit. (a general alternative).} \]

Program MARK provides various GOF tests and statistics and these are important to examine in the analysis of real data (see also, program RELEASE). Often, you will want to examine
the matrix of reported tags/bands (the $m_{ij}$), their associated expectations under a given model (the $\hat{E}_{ij}$) and the associated individual chi-square contributions (denoted as $\chi^2_{ij}$). Goodness of fit testing can become a science of its own for the banding data and capture-recapture data for open populations (see Burnham et al. 1987 for such material).

**Multiple Groups**

The full power of current theory and computer software comes into play when there are band recovery data on more than a single group. The simplest case arises when there are parallel data sets for males and females. For example, consider the data sets,

<table>
<thead>
<tr>
<th>MALES</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2583</td>
<td>91</td>
<td>89</td>
<td>24</td>
<td>18</td>
</tr>
<tr>
<td>3075</td>
<td>141</td>
<td>45</td>
<td>52</td>
<td>50</td>
</tr>
<tr>
<td>1195</td>
<td>27</td>
<td>31</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>3418</td>
<td>156</td>
<td>92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FEMALES</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1478</td>
<td>40</td>
<td>31</td>
<td>8</td>
<td>11</td>
</tr>
<tr>
<td>1525</td>
<td>72</td>
<td>20</td>
<td>15</td>
<td>7</td>
</tr>
<tr>
<td>319</td>
<td>8</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>1805</td>
<td>63</td>
<td>27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1400</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These data are from winter banding of mallards in Illinois in the early 1960s (see Brownie et al. (1985: 146) for the full data set and the (now-out-of-date) analysis of sex-specific parameters.

Even with two groups there are many models possible. One could model each group with separate parameters and let both $S$ and $r$ be time-specific. Thus, this is model $\{S_{gt}, r_{gt}\}$, where $K = 18$. This is like two separate C-J-S models. At the other extreme, one could assume survival and reporting probabilities where constant over both years and groups. This is model $\{S, r\}$, with $K = 2$.

Many models of the multi-group case can be done in *MARK* using the PIMs, while more sophisticated models must await the concept of the design matrix. The important issue to begin to emerge here is that of *a priori* knowledge about the questions of interest to be asked of these data. What is known about the science of the matter? Why were the data collected in
the first place? What controversy surrounds these data? What are the issues/questions of interest?

Thus, one might develop 5-8 models for the analysis of these data before any data dredging. This important issue will be compromised a bit in FW663 as we did not collect the data, nor were we closely associated with reasons for data collection in the early 1960s. We are certainly allowed to know that the sample size is very large, especially for males. Hunting seasons for waterfowl in the early 1960s were quite variable; thus one might expect year-specific variation in the reporting probabilities. It would seem that the simple models \( \{S, r\} \) with \( K = 2 \) and \( \{S_g, r_g\} \) with \( K = 4 \) could be ignored, given the amount of data and the dynamics of hunting seasons at the time. Other models can be formulated in class and discussed.

An important issue here is **do not run all possible models or as many as can be thought of and implemented**. This is a senseless analysis strategy and leads to over-fitting and the “finding” of spurious effects.

A class of models of particular interest in the multi-group case are ones that constrain, say, survival parameters to be similar for the two genders. Thus, if survival probabilities for males had a certain pattern across years, then perhaps these parameters for females might “parallel” the pattern for males. A good year for males would be a good year for females, etc., etc. These models can be built and are very effective in many situations. The design matrix will be needed to develop such models.

In considering the multi-group case, one must continue to remember the *Principle of Parsimony*. Adding parameters will tend to decrease bias, but with an increase in variation (uncertainty). Thus, there is a trade-off between bias and variance. Demonstrations in class and laboratory exercises will help with model building in the multi-group case.
The analysis of multi-group data sets is at the current state of the art. The recent monograph on population dynamics of the Northern Spotted Owl is an example of what can be done (in that case some 100 groups). Lebreton et al. (1992) is the primary reference for the modeling of several groups.

More on Identifiability and Related Issues

Any model for tag recovery data is based on interpretable parameters, especially of the type $S$ and $r$ (or $S$ and $f$) explicitly appearing in the model structure for $E(m_{ij} \mid R_i)$. However, just because a parameter appears in the model does not mean that parameter can in fact be estimated from data. Most parameters in the model are estimable, but not all; it depends on the model.

The idea of parameters not being estimable is illustrated by trying to estimate $S_1$ from one year of tag recoveries, $m_{11}$:

$$\frac{E(m_{11})}{R_1} = (1 - S_1)r_1.$$ 

You cannot do it; you only have an estimate of the product, $(1 - S_1)r_1$ and there is no way to separately estimate $S_1$ and/or $r_1$.

For all models we know (or can know) what parameters are estimable and hence we know the number, $K$, needed for $\text{QAIC}_c$ (or likelihood ratio tests). (For some models this information is embedded in the help file of MARK). You do not need to know $K$ to fit models. MARK tries to determine $K$ by numerical methods, it does not always succeed. So there are times when $K$ needs to be input to MARK to get $\text{QAIC}_c$ computed correctly. You also need to beware of interpreting numerical results for non-estimable parameters as meaning anything. You can tell such cases by the estimated standard error: it will be either huge, or paradoxically, trivially small (near zero).

There is another practical problem that arises in model fitting: point estimates that are on a boundary (i.e., and $\hat{S} = 1$, or $r = 1$). This can lead to MARK computing the wrong $K$ for the model. Also, when this occurs, it suggests the model is too general, hence not the best one to use. When a parameter estimate is “pegged” on a boundary its estimated standard error will generally be quite wrong (too small), and this event (estimate on a boundary) can cause the estimated standard errors of other estimates to be wrong. You need to look at fitted models to be sure no such anomalies have occurred and to see if MARK has $K$ correct, so $\text{QAIC}_c$ is correct. The issue of the correct $K$ is a difficult one for us to provide advice about. However, anomalous point estimates and weird standard errors are indicators of either basic parameter non estimability, or may just reflect sparse (poor) data or a bad fitting model.
An example of a problem with sparse data that produced an estimate on a boundary is the below. First, the input data were

```c
/* Release recovery data for RELEASES >= 711 MM (28 INCHES) long, data for Chesapeake Bay, from Cynthia Goshorn via Dave Smith. Years are 1987 to 1996 */
recovery matrix group=1;
   1  0  2  0  1  2  1  0  0  0  0;
   6  8  7 14  6  1  3  0  0  0;
   9 17 17  6  4  3  5  2  0  0;
  23 16 11  5  2  4  0  0  0;
  47 24 20  4  9  3  0  0  0;
  44 28 18 16  7  0  0  0  0;
  52 42 22  0  0  0  0  0  0;
  61 29  0  0  0  0  0  0  0;
  92  0  0  0  0  0  0  0  0;

29 129 221 304 396 438 628 545 529 862;
```

The first year of releases and that cohort of recoveries are too small to be worth including in the analysis. However, dropping that year does not solve the problems with analysis of these data under model \{S(t), r(t)\}. Output under this model follows.
```plaintext
model=\{S(t), r(t)\}

<table>
<thead>
<tr>
<th>I</th>
<th>S(I)</th>
<th>Standard Error</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.6882106</td>
<td>0.2549591</td>
<td>0.1769578</td>
<td>0.9577357</td>
</tr>
<tr>
<td>2</td>
<td>0.9597202</td>
<td>0.0162191</td>
<td>0.9128110</td>
<td>0.9818921</td>
</tr>
<tr>
<td>3</td>
<td>0.9477962</td>
<td>0.0117022</td>
<td>0.9194888</td>
<td>0.9665133</td>
</tr>
<tr>
<td>4</td>
<td>0.5873970</td>
<td>0.0721886</td>
<td>0.4426089</td>
<td>0.7184953</td>
</tr>
<tr>
<td>5</td>
<td>0.5735394</td>
<td>0.0754826</td>
<td>0.4234658</td>
<td>0.7111909</td>
</tr>
<tr>
<td>6</td>
<td>0.6190184</td>
<td>0.0745679</td>
<td>0.4664646</td>
<td>0.7512166</td>
</tr>
<tr>
<td>7</td>
<td>0.6884013</td>
<td>0.0816988</td>
<td>0.5115350</td>
<td>0.8233431</td>
</tr>
<tr>
<td>8</td>
<td>0.7018095</td>
<td>0.0960572</td>
<td>0.4890719</td>
<td>0.8526551</td>
</tr>
<tr>
<td>9</td>
<td>0.4699632</td>
<td>0.0792827</td>
<td>0.3221038</td>
<td>0.6232899</td>
</tr>
<tr>
<td>10</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>I</th>
<th>r(I)</th>
<th>Standard Error</th>
<th>Lower</th>
<th>Upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>0.1105959</td>
<td>0.1346352</td>
<td>0.0084308</td>
<td>0.6452122</td>
</tr>
<tr>
<td>12</td>
<td>1.0000000</td>
<td>0.5302730E-05</td>
<td>0.9999904</td>
<td>1.0000004</td>
</tr>
<tr>
<td>13</td>
<td>1.0000000</td>
<td>0.4885705E-05</td>
<td>0.9999904</td>
<td>1.0000009</td>
</tr>
<tr>
<td>14</td>
<td>0.1755289</td>
<td>0.0378107</td>
<td>0.1131433</td>
<td>0.2621452</td>
</tr>
<tr>
<td>15</td>
<td>0.2866250</td>
<td>0.0637381</td>
<td>0.1790485</td>
<td>0.4253479</td>
</tr>
<tr>
<td>16</td>
<td>0.2762157</td>
<td>0.0667091</td>
<td>0.1655747</td>
<td>0.4232850</td>
</tr>
<tr>
<td>17</td>
<td>0.3217343</td>
<td>0.0938853</td>
<td>0.1695157</td>
<td>0.5243401</td>
</tr>
<tr>
<td>18</td>
<td>0.3133693</td>
<td>0.1132253</td>
<td>0.1399389</td>
<td>0.5614302</td>
</tr>
<tr>
<td>19</td>
<td>0.2263497</td>
<td>0.0466889</td>
<td>0.1478443</td>
<td>0.3303798</td>
</tr>
<tr>
<td>20</td>
<td>0.1067285</td>
<td>0.0105167</td>
<td>0.0877991</td>
<td>0.1291613</td>
</tr>
</tbody>
</table>
```

“←” indicates problem estimates

What went wrong? Insight can be gained by looking at a parameterization from Brownie et al. (1985) where

\[ f_j = (1 - S_j)r_j . \]

Under this model parameterization, the unrestricted MLEs of \( S_2 \) and \( S_3 \) are \( > 1 \).
model={$S(t), f(t)$}

<table>
<thead>
<tr>
<th></th>
<th>$S(I)$</th>
<th>Standard Error</th>
<th>95% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5931043</td>
<td>0.2271318</td>
<td>0.1872679 0.9021626</td>
</tr>
<tr>
<td>2</td>
<td>1.0797344</td>
<td>0.1820650</td>
<td>0.7228870 1.4365818</td>
</tr>
<tr>
<td>3</td>
<td>1.1707343</td>
<td>0.1903254</td>
<td>0.7976965 1.5437720</td>
</tr>
<tr>
<td>4</td>
<td>0.5099335</td>
<td>0.0772790</td>
<td>0.3620754 0.6560738</td>
</tr>
<tr>
<td>5</td>
<td>0.5735387</td>
<td>0.075834</td>
<td>0.4234656 0.7111900</td>
</tr>
<tr>
<td>6</td>
<td>0.6270057</td>
<td>0.0755834</td>
<td>0.4715178 0.7600288</td>
</tr>
<tr>
<td>7</td>
<td>0.6796323</td>
<td>0.0807166</td>
<td>0.5063700 0.8143737</td>
</tr>
<tr>
<td>8</td>
<td>0.7018086</td>
<td>0.0960566</td>
<td>0.4890726 0.8526536</td>
</tr>
<tr>
<td>9</td>
<td>0.4699636</td>
<td>0.0792828</td>
<td>0.3221041 0.6232904</td>
</tr>
<tr>
<td>10</td>
<td>0.0344828</td>
<td>0.0338830</td>
<td>0.0048356 0.2079196</td>
</tr>
<tr>
<td>11</td>
<td>0.0410397</td>
<td>0.0164941</td>
<td>0.0184737 0.0886799</td>
</tr>
<tr>
<td>12</td>
<td>0.0501508</td>
<td>0.0117315</td>
<td>0.0315548 0.0788141</td>
</tr>
<tr>
<td>13</td>
<td>0.0628728</td>
<td>0.0104667</td>
<td>0.0452223 0.0867861</td>
</tr>
<tr>
<td>14</td>
<td>0.1222342</td>
<td>0.0137126</td>
<td>0.0977977 0.1517497</td>
</tr>
<tr>
<td>15</td>
<td>0.1052331</td>
<td>0.0119675</td>
<td>0.0839769 0.1310996</td>
</tr>
<tr>
<td>16</td>
<td>0.0989749</td>
<td>0.0099066</td>
<td>0.0811819 0.1201578</td>
</tr>
<tr>
<td>17</td>
<td>0.0934437</td>
<td>0.0099059</td>
<td>0.0757538 0.1147516</td>
</tr>
<tr>
<td>18</td>
<td>0.1199738</td>
<td>0.0125178</td>
<td>0.0975225 0.1467532</td>
</tr>
<tr>
<td>19</td>
<td>0.1067285</td>
<td>0.0105167</td>
<td>0.0877991 0.1291613</td>
</tr>
</tbody>
</table>

This makes it clear that two estimates are out of range and this creates problems. We prefer to treat such cases as a diagnostic that a model with too many parameters have been used for the analysis of the data.

**Unequal Time Intervals Between Banding Times**

We have been treating the striped bass data as if the time intervals between tagging sessions are exactly one year apart. This is probably the best thing to do. However, capture-recapture and recovery studies may utilize unequal time intervals. Then you must know the time interval lengths and input this information to program MARK.

Let the nominal time instant of capture session $i$ be at $\tau_i$, $i = 1, 2, \ldots, t$. Then the length of the $i$th time interval is $\Delta_i = \tau_{i+1} - \tau_i$. (I would take $\tau_i$ as the weighted average Julian date at which fish were caught, tagged and released; weights = number of fish released on the date). Survival rate $S_i$ is for this length of time, i.e., $\Delta_i$. If the time intervals are not equal, interpretations of $\hat{S}_i$ are hindered under any $\{S(t)\}$ model. Moreover, models like $\{S(.), T\}$, or $\{S(T)\}$, are not meaningful unless they properly allow for unequal time intervals.

Program MARK handles this issue just fine provided you input the $\Delta_i$ when these intervals differ in length (the MARK default for equal time interval lengths is $\Delta_i \equiv 1$). In the case of unequal $\Delta_i$ the basic parameterization used is in terms of survival rates on a per-unit
time basis (whatever units your \( \Delta_i \) are expressed in). Given unequal \( \Delta_i \) \textit{MARK} uses the interval survival rates expressed as

\[
(S_i)^{\Delta_i}
\]

and this form is used in the likelihoods. Using this form the underlying survival rate, \( S_i \), for the \( i \)th interval is on a standardized time unit basis for all intervals. Thus the \{\( S(.) \)\} models are sensible.

It is worth taking a look at the actual times when striped bass are caught, tagged and released to calculate if the \( \Delta_i \) are quite equal to 1 year, or if perhaps there is variation there that we should allow for in fitting the tag recovery models.

A lot of the original developments for tag recovery models (in the 1970s) use the parameterization in \( S \) and \( f \). Hence expected multinomial cells (i.e., \( E(m_{ij})/R_i \)) in the models were like

\[
\begin{array}{cccc}
  f_1 & S_1f_2 & S_1S_2f_3 & S_1S_2S_3f_4 \\
  f_2 & S_2f_3 & S_2S_3f_4 \\
  f_3 & S_3f_4 \\
  f_4
\end{array}
\]

An alternative form has been used and is in \textit{MARK}:

\[
(1 - S_1)r_1 \quad S_1(1 - S_2)r_2 \quad S_1S_2(1 - S_3)r_3 \quad S_1S_2S_3(1 - S_4)r_4 \\
(1 - S_2)r_2 \quad S_2(1 - S_3)r_3 \quad S_2S_3(1 - S_4)r_4 \\
(1 - S_3)r_3 \quad S_3(1 - S_4)r_4 \\
(1 - S_4)r_4
\]

(actually, now both forms are in \textit{MARK}).

The structural relationship between these models is just

\[
f_i = (1 - S_i)r_i
\]

With good data the estimated \( S_i \) are identical under full time effects (models \{\( S(t), f(t) \)\} and \( S(t), r(t) \)). In general, there are no substantive differences in the survival rate estimates you will get from either model form, and in particular the meaning of the survival rates is the same under either form. The model forms appear to be “saying” different things, but that is not the
case in terms of what the survival parameter here means: survival of banded animals from the beginning of one year to the beginning of the next year.

Alas, these tag recovery studies allow estimates of survival probability of tags in the population. If a fish is caught, its tag removed, and the fish released, then the survival probability of fish will be underestimated! The fish and its tag should be one unit; if they are separated, problems occur. This has happened in the national program for striped bass on the east coast of the USA!

The Sampling Variance-Covariance Matrix

This is a good time to reread Burnham et al. (1987) pages 12-14.

The variance-covariance matrix is a square matrix, with sampling variances down the diagonal. These are measures of precision or repeatability. The off-diagonal elements are sampling covariances and measure the dependence between estimators (e.g., $\hat{S}$ and $\hat{p}$). Such dependency is caused by different estimators using common data. This is a statistical dependence, measured as $\text{cov}(\cdot, \cdot)$.

Often easier to interpret is the sampling correlation $\text{corr}(\cdot, \cdot)$. For example,

$$\text{corr}(\hat{S}, \hat{p}) = \frac{\text{cov}(\hat{S}, \hat{p})}{\text{se}(\hat{S}) \text{se}(\hat{p})},$$

where $-1 \leq \text{corr}(\cdot, \cdot) \leq 1$.

These sampling covariances and correlations are not related to population correlations, $\rho$ anymore than $\sigma_S$ is related to $\text{var}(\hat{S})$.

In the fish tagging or bird banding model $\{S_t, r_t\}$ there are several nasty sampling correlations between various estimators. One such issue is the dependence of $\hat{S}_i$ on $\hat{S}_{i+1}$. For example,

$$\text{corr}(\hat{S}_2, \hat{S}_3) = \frac{\text{cov}(\hat{S}_2, \hat{S}_3)}{\text{se}(\hat{S}_2) \text{se}(\hat{S}_3)}$$

is negative.

There is a pair-wise dependence of successive estimators of survival probability. This gives a plot of estimated survival probabilities a “see saw” look that is unrelated to the parameters. The naive on-looker might think such a plot suggested some sort of density dependence!
Curiously, the sampling correlations of estimators such as $S_2$ vs. $S_1$ or $S_2$ vs. $S_7$ are nearly zero. The variance-covariance matrix for the survival probabilities is “band diagonal.”

Another dependency is between the estimators or survival vs. reporting probabilities. Program *MARK* provides estimates of the variance-covariance (often just called the covariance matrix) matrix as an option.

The *shape* of the log-likelihood function is related to the sampling variance as a measure of precision. Can you intuit show the shape of the log-likelihood function might look when the sampling correlation was near either –1 or +1 or 0?

**INFORMATION THEORY AND LOG-LIKELIHOOD MODELS: A BASIS FOR MODEL SELECTION AND INFERENCE**

Full reality cannot be included in a model; thus we seek a good model to *approximate* the effects or factors supported by the empirical data. The selection of an appropriate approximating model is critical to statistical inference from many types of empirical data. Information theory (see Guiasu 1977 and Cover and Thomas 1991) and, in particular the Kullback-Leibler (Kullback and Liebler 1951) “distance," or “information" forms the deep theoretical basis for data-based model selection. Akaike (1973) found a simple relationship between expected Kullback-Leibler information and Fisher's maximized log-likelihood function (see deLeeuw 1992 for a brief review). This relationship leads to a simple, effective, and very general methodology for selecting a parsimonious model for the analysis of empirical data.

**Some Background**

Well over a century ago measures were derived for assessing the “distance" between two models or probability distributions. Most relevant here is Boltzmann's (1877) concept of generalized entropy in physics and thermodynamics (see Akaike 1985 for a brief review). Shannon (1948) employed entropy in his famous treatise on communication theory. Kullback and Leibler (1951) derived an information measure that happened to be the negative of Boltzmann's entropy: now referred to as the Kullback-Leibler (K-L) distance. The motivation for Kullback and Leibler's work was to provide a rigorous definition of “information" in relation to Fisher's “sufficient statistics." The K-L distance has also been called the K-L discrepancy, divergence, information and number – these terms are synonyms, we tend to use *distance* or *information* in the material to follow.

The Kullback-Leibler distance can be conceptualized as a directed “distance" between two models, say $f$ and $g$ (Kullback 1959). Strictly speaking this is a measure of “discrepancy"; it is not a simple distance because the measure from $f$ to $g$ is not the same as the measure from $g$ to $f$ – it is a directed or oriented distance. The K-L distance is perhaps the most fundamental of all information measures in the sense of being derived from minimal
assumptions and its additivity property. The K-L distance between models is a fundamental quantity in science and information theory (see Akaike 1983) and is the logical basis for model selection as defined by Akaike. In the heuristics given here, we will assume the models in question are continuous probability distributions denoted as $f$ and $g$. Biologists are familiar with the normal, log-normal, gamma and other continuous distributions. We will, of course, not limit ourselves to these common, simple types.

It is useful to think of $f$ as full reality and let it have (conceptually) an infinite number of parameters. This “crutch” of infinite dimensionality at least keeps the concept of reality even though it is in some unattainable perspective.

Let $g$ be the approximating model being compared to (measured against) $f$. We use $x$ to denote the data being modeled and $\theta$ to denote the parameters in the approximating model $g$. We use $g(x)$ as an approximating model, whose parameters must be estimated from these data (in fact, we will make this explicit using the notation $g(x \mid \theta)$, read as “the approximating model $g$ for data $x$ given the parameters $\theta$). If the parameters of the model $g$ have been estimated, using ML or LS methods, we will denote this by $g(x \mid \hat{\theta})$. Generally, in any real world problem, the model $g(x \mid \theta)$ is a function of sample data (often multivariate) and the number of parameters ($\theta$) in $g$ might often be of high dimensionality. Finally, we will want to consider a set of approximating models as candidates for the representation of the data; this set of models is denoted $\{g_i(x \mid \theta), i = 1, ..., R\}$. It is critical that this set of models be defined prior to probing examination of the data (i.e., no data dredging).
The Kullback-Leibler Distance or Information

The K-L distance between the models \( f \) and \( g \) is defined for continuous functions as the (usually multi-dimensional) integral

\[
I(f, g) = \int f(x) \log \left( \frac{f(x)}{g(x|\theta)} \right) dx,
\]

where \( \log \) denotes the natural logarithm. Kullback and Leibler (1951) developed this quantity from “information theory,” thus they used the notation \( I(f, g) \);

\( I(f, g) \) is the “information” lost when \( g \) is used to approximate \( f \).

Of course, we seek an approximating model that loses as little information as possible; this is equivalent to minimizing \( I(f, g) \), over \( g \). \( f \) is considered to be given (fixed) and only \( g \) varies over a space of models indexed by \( \theta \). An equivalent interpretation of minimizing \( I(f, g) \) is finding an approximating model that is the “shortest distance” away from truth. We will use both interpretations as both seem useful.

The expression for the K-L distance in the case of discrete distributions such as the Poisson, binomial or multinomial, is

\[
I(f, g) = \sum_{i=1}^{k} p_i \log \left( \frac{p_i}{\pi_i} \right).
\]

Here, there are \( k \) possible outcomes of the underlying random variable; the true probability of the \( i^{th} \) outcome is given by \( p_i \), while the \( \pi_1, \ldots, \pi_k \) constitute the approximating probability distribution (i.e., the approximating model). In the discrete case, we have \( 0 < p_i < 1, 0 < \pi_i < 1 \) and \( \sum p_i = \sum \pi_i = 1 \). Hence, here \( f \) and \( g \) correspond to the \( p \) and \( \pi \), respectively.

The material above makes it obvious that both \( f \) and \( g \) (and their parameters) must be known to compute the K-L distance between these 2 models. We see that this requirement is diminished as \( I(f, g) \) can be written equivalently as

\[
I(f, g) = \int f(x) \log (f(x)) \, dx - \int f(x) \log (g(x | \theta)) \, dx.
\]

Note, each of the two terms on the right of the above expression is a statistical expectation with respect to \( f \) (truth). Thus, the K-L distance (above) can be expressed as a difference between two expectations,

\[
I(f, g) = E_f[\log(f(x))] - E_f[\log(g(x | \theta))],
\]
each with respect to the true distribution $f$. This last expression provides easy insights into the derivation of AIC. The important point is that the K-L distance $I(f, g)$ is a measure of the directed “distance” between the probability models $f$ and $g$.

The first expectation $E_f[\log(f(x))]$ is a constant that depends only on the unknown true distribution and it is clearly not known (i.e., we do not know $f(x)$ in actual data analysis). Therefore, treating this unknown term as a constant, only a measure of relative, directed distance is possible (Bozdogan 1987, Kapur and Kesavan 1992: 155) in practice. Clearly, if one computed the second expectation, $E_f[\log(g(x | \theta))]$, one could estimate $I(f, g)$, up to a constant (namely $E_f[\log(f(x))]$),

$$I(f, g) = \text{Constant} - E_f[\log(g(x | \theta))],$$

or

$$I(f, g) - \text{Constant} = - E_f[\log(g(x | \theta))].$$

The term $\left(I(f, g) - \text{Constant}\right)$ is a relative, directed distance between the two models $f$ and $g$, if one could compute or estimate $E_f[\log(g(x | \theta))]$. Thus, $E_f[\log(g(x | \theta))]$ becomes the quantity of interest.

In data analysis the model parameters must be estimated and there is usually substantial uncertainty in this estimation. Models based on estimated parameters, hence on $\hat{\theta}$ not $\theta$, represent a major distinction from the case where model parameters would be known. This distinction affects how we must use K-L distance as a basis for model selection. The difference between having $\theta$ (we do not) and having $\hat{\theta}$ (we do) is quite important and basically causes us to change our model selection criterion to that of minimizing expected estimated K-L distance rather than minimizing known K-L distance (over the set of $R$ a priori models considered).

Thus, we use the concept of selecting a model based on minimizing the estimated Kullback-Leibler distance

$$\hat{I}(f, g) = \int f(x) \log \left( \frac{f(x)}{g(x | \theta(y))} \right) \, dx.$$ 

Consequently, we can determine a method to select the model $g_i$ that on average minimizes, over the set of models $g_1, \ldots, g_R$, a very relevant expected K-L distance.

Akaike (1973) showed firstly that the maximized log-likelihood is biased upward as an estimator of the model selection criterion. Second, he showed that under certain conditions (these conditions are important, but quite technical), that this bias is approximately equal to $K$, the number of estimable parameters in the approximating model. Thus, an approximately unbiased estimator of the relative, expected K-L information is

$$\log(\mathcal{L}(\hat{\theta} | y)) - K,$$
This result is equivalent to

\[ \log(\mathcal{L}(\hat{\theta} \mid y)) - K = \text{Constant} - \hat{E}_\theta[\hat{I}(f,g)] \]

or

\[ -\log(\mathcal{L}(\hat{\theta} \mid y)) + K = \text{estimated relative expected K-L distance.} \]


Akaike (1973) then defined “an information criterion" (AIC) by multiplying by \( -2 \) (“taking historical reasons into account”) to get

\[ \text{AIC} = -2\log(\mathcal{L}(\hat{\theta} \mid y)) + 2K. \]

This has became known as “Akaike's Information Criterion" or AIC. Here it is important to note that AIC has a strong theoretical underpinning, based on information theory and Kullback-Leibler information within a realistic data analysis philosophy that no model is true, rather truth as \( f \) is far more complex than any model used. Akaike's inferential breakthrough was realizing that a predictive expectation version of the log-likelihood could (as one approach) be used to estimate the relative expected K-L distance between the approximating model and the true generating mechanism. Thus, rather than having a simple measure of the directed distance between two models (i.e., the K-L distance), one has instead an \emph{estimate} of the expected relative, directed distance between the fitted model and the unknown true mechanism (perhaps of infinite dimension) which actually generated the observed data. Because the expectation of the logarithm of \( f(x) \) drops out as a constant, independent of the data, AIC is defined without specific reference to a "true model" (Akaike 1985:13).

Thus, one should select the model that yields the smallest value of AIC because this model is estimated to be “closest" to the unknown reality that generated the data, from among the candidate models considered. This seems a very natural, simple concept; select the fitted approximating model that is estimated, on average, to be closest to the unknown \( f \).

Perhaps none of the models in the set are good, but AIC attempts to select the best approximating model of those in the candidate set. Thus, every effort must be made to assure that the set of models is well founded.
AIC Differences

Because AIC is on an relative scale, we routinely recommend computing (and presenting in publications) the AIC differences (rather than the actual AIC values),

\[ \Delta_i = \text{AIC}_i - \text{minAIC}, \]

over all candidate models in the set. Such differences estimate the relative expected K-L differences between \( f \) and \( g_i(x | \theta) \). These \( \Delta_i \) values are easy to interpret and allow a quick comparison and ranking of candidate models and are also useful in computing Akaike weights and other quantities of interest.

The larger \( \Delta_i \) is, the less plausible is the fitted model \( g_i(x | \hat{\theta}) \) as being the K-L best model for samples such as the data one has. As a rough rule of thumb, models for which \( \Delta_i \leq 2 \) have substantial support and should receive consideration in making inferences. Models having \( \Delta_i \) of about 4 to 7 have considerably less support, while models with \( \Delta_i > 10 \) have either essentially no support, and might be omitted from further consideration, or at least those models fail to explain some substantial explainable variation in the data. If observations are not independent but are assumed to be independent then these simple guidelines cannot be expected to hold.

Important Refinements to AIC

A Second Order AIC

Akaike derived an estimator of the K-L information quantity, however, AIC may perform poorly if there are too many parameters in relation to the size of the sample (Sugiura 1978, Sakamoto et al. 1986). Sugiura (1978) derived a second order variant of AIC that he called c-AIC. Hurvich and Tsai (1989) further studied this small-sample (second order) bias adjustment which led to a criterion that is called AIC\(_c\),

\[ \text{AIC}_c = -2 \log(\hat{L}(\hat{\theta})) + 2K \left( \frac{n}{n-K-1} \right), \]

where the penalty term is multiplied by the correction factor \( n/(n-K-1) \). This can be rewritten as

\[ \text{AIC}_c = -2 \log(\hat{L}(\hat{\theta})) + 2K + \frac{2K(K+1)}{n-K-1}, \]

or, equivalently,
\[ \text{AIC}_c = \text{AIC} + \frac{2K(K+1)}{n-K-1}, \]

where \( n \) is sample size (also see Sugiura 1978). AIC\(_c\) merely has an additional bias correction term. If \( n \) is large with respect to \( K \), then the second order correction is negligible and AIC should perform well. Generally, we advocate the use of AIC\(_c\) when the ratio \( n/K \) is small (say < 40). In reaching a decision about the use of AIC vs. AIC\(_c\), one must use the value of \( K \) for the highest dimensioned model in the set of candidates. If the ratio \( n/K \) is sufficiently large, then AIC and AIC\(_c\) are similar and will tend to select the same model. One should use either AIC or AIC\(_c\) consistently in a given analysis; rather than mixing the two criteria. **Unless the sample size is large with respect to the number of estimated parameters, use of AIC\(_c\) is recommended.**

**Modification to AIC for Overdispersed Count Data**

Count data have been known not to conform to simple variance assumptions based on binomial or multinomial distributions. If the sampling variance exceeds the theoretical (model based) variance, the situation is called “overdispersion.” Our focus here is on a lack of independence in the data leading to overdispersion or “extra-binomial variation.” Eberhardt (1978) provides a clear review of these issues in the biological sciences. For example, Canada geese (\textit{Branta spp.}) frequently mate for life and the pair behaves almost as an individual, rather than as two independent “trials.” The young of some species continue to live with the parents for a period of time, which can also cause a lack of independence of individual responses. Further reasons for overdispersion in biological systems include species whose members exist in schools or flocks. Members of such populations can be expected to have positive correlations among individuals within the group; such dependence causes overdispersion. A different type of overdispersion stems from parameter heterogeneity; that is individuals having unique parameters rather than the same parameter (such as survival probability) applying to all individuals.

Cox and Snell (1989) discuss modeling of count data and note that the first useful approximation is based on a single variance inflation factor (\( \hat{c} \)) which can be estimated from the goodness-of-fit chi-square statistic (\( \chi^2 \)) of the global model and its degrees of freedom,

\[ \hat{c} = \chi^2/df. \]

The variance inflation factor should be estimated from the global model.

Given \( \hat{c} \), empirical estimates of sampling variances (\( \text{var}_e(\hat{\theta}_i) \)) and covariances (\( \text{cov}_e(\hat{\theta}_i, \hat{\theta}_j) \)) can be computed by multiplying the estimates of the theoretical (model-based) variances and covariances by \( \hat{c} \) (a technique that has long been used, see e.g., Finney 1971). These empirical measures of variation (i.e., \( \hat{c} \cdot \text{var}_e(\hat{\theta}_i) \)) must be treated as having the degrees of freedom used to compute \( \hat{c} \) for purposes of setting confidence limits (or testing hypotheses).
Generally, quasi-likelihood adjustments (i.e., use of \( \hat{c} > 1 \)) are made only if some reasonable lack of fit has been found (for example if the observed significance level \( P \leq 0.15 \) or \( 0.25 \)) and the degrees of freedom \( \geq 10 \), as rough guidelines.

Patterns in the goodness-of-fit statistics (Pearson \( \chi^2 \) or G-statistics) might be an indication of structural problems with the model. Of course, the biology of the organism in question should provide clues as to the existence of overdispersion; one should not rely only on statistical considerations in this matter.

Principles of quasi-likelihood suggest simple modifications to AIC and \( \text{AIC}_c \); we denote these modifications as (Lebreton et al. 1992),

\[
\text{QAIC} = - \left[ 2 \log(L(\hat{\theta}))/\hat{\theta} \right] + 2K, \]

and

\[
\text{QAIC}_c = - \left[ 2 \log(L(\hat{\theta}))/\hat{\theta} \right] + 2K + \frac{2K(K+1)}{n-K-1},
\]

\[
= \text{QAIC} + \frac{2K(K+1)}{n-K-1}.
\]

Of course, when no overdispersion exists, \( c = 1 \), the formulae for QAIC and QAIC\(_c\) reduce to AIC and \( \text{AIC}_c \), respectively.

Some History

Akaike (1973) considered AIC and its information theoretic foundations “... a natural extension of the classical maximum likelihood principle.” Interestingly, Fisher (1936) anticipated such an advance over 60 years ago when he wrote,

“... an even wider type of inductive argument may some day be developed, which shall discuss methods of assigning from the data the functional form of the population.”

This comment was quite insightful; of course, we might expect this from R. A. Fisher! Akaike was perhaps kind to consider AIC an extension of classical ML theory; he might just as well have said that classical likelihood theory was a special application of the more general information theory. In fact, Kullback believed in the importance of information theory as a unifying principle in statistics.

Interpreting Differences Among AIC Values
Akaike's Information Criterion (AIC) and other information theoretic methods can be used to rank the candidate models from best to worst. Often data do not support only one model as clearly best for data analysis. Instead, suppose three models are essentially tied for best, while another, larger, set of models is clearly not appropriate (either under- or over-fit). Such virtual “ties” for the best approximating model must be carefully considered and admitted. Poskitt and Tremayne (1987) discuss a “portfolio of models” that deserve final consideration. Chatfield (1995b) notes that there may be more than one model that is to be regarded as “useful.” The inability to ferret out a single best model is not a defect of AIC or any other selection criterion, rather, it is an indication that the data are simply inadequate to reach such a strong inference. That is, the data are ambivalent concerning some effect or parameterization or structure.

It is perfectly reasonable that several models would serve nearly equally well in approximating a set of data. Inference must admit that there are sometimes competing models and the data do not support selecting only one. Using the Principle of Parsimony, if several models fit the data equally well, the one with the fewest parameters might be preferred; however, some consideration should be given to the other (few) competing models that are essentially tied as the best approximating model. Here the science of the matter should be fully considered. The issue of competing models is especially relevant in including model selection uncertainty into estimators of precision and model averaging.

A well thought out global model (where applicable) is important and substantial prior knowledge is required during the entire survey or experiment, including the clear statement of the question to be addressed and the collection of the data. This prior knowledge is then carefully input into the development of the set of candidate models. Without this background science, the entire investigation should probably be considered only very preliminary.

Model Selection Uncertainty

One must keep in mind that there is often considerable uncertainty in the selection of a particular model as the “best” approximating model. The observed data are conceptualized as random variables; their values would be different if another, independent set were available. It is this “sampling variability” that results in uncertain statistical inference from the particular data set being analyzed. While we would like to make inferences that would be robust to other (hypothetical) data sets, our ability to do so is still quite limited, even with procedures such as AIC, with its cross validation properties, and with independent and identically distributed sample data. Various computer intensive, resampling methods will further improve our assessment of the uncertainty of our inferences, but it remains important to understand that proper model selection is accompanied by a substantial amount of uncertainty. Quantification of many of these issues is beyond the scope of the material here (see Burnham and Anderson 1998 for advanced methods).

AIC When Different Data Sets are to be Compared
Models can only be compared using AIC when they have been fitted to exactly the same set of data (this applies also to likelihood ratio tests). For example, if nonlinear regression model A is fitted to a data set with \( n = 140 \) observations, one cannot validly compare it with Model B when 7 outliers have been deleted, leaving only \( n = 133 \). Furthermore, AIC cannot be used to compare models where the data are ungrouped in one case (Model U) and grouped (e.g., grouped into histograms classes) in another (Model G).

Summary

The Principle of Parsimony provides a conceptual guide to model selection, while expected K-L information provides an objective criterion, based on a deep theoretical justification. AIC, AIC\(_c\) and QAIC\(_c\) provide a practical method for model selection and associated data analysis and are estimates of expected, relative K-L information. AIC, AIC\(_c\) and QAIC represent an extensions of classical likelihood theory, are applicable across a very wide range of scientific questions, and are quite simple to use in practice.

Some References On Model Selection


Anderson, D. R., and Burnham, K. P. (199_). General strategies for the collection and analysis of ringing data. Bird Study __, __-__.

Anderson, D. R., and Burnham, K. P. (199_). Understanding information criteria for selection among capture-recapture or ring recovery models. Bird Study __, __-__.


