Properties of the residuals from two tag-recovery models*

Robert J. Latour
John M. Hoenig

Department of Fisheries Science
Virginia Institute of Marine Science
College of William and Mary
Gloucester Point, Virginia 23062
E-mail address (for R. J. Latour): latour@vims.edu

Kenneth H. Pollock

Biomathematics Graduate Program
Department of Statistics
North Carolina State University
Raleigh, North Carolina 27695

Researchers often use multiyear tag-recovery studies to assess fish populations; yet deriving useful stock assessment parameter estimates from the resulting data can be difficult. The reliability of those parameter estimates generally depends on data quality and meeting the assumptions inherent to the models used for analysis. As a result, practical application of multiyear tag-recovery models generally requires that a large portion of the data analysis involve investigation and evaluation of biases due to potential assumption violation.

Brownie et al. (1985) developed a class of models that has become widely used for the analysis of multiyear tag-recovery data. These models constitute a generalization of the class of models developed by Seber (1970), which have recently been resurrected as an important tool for the analysis of multiyear tag-recovery data by the development of the software program MARK (White and Burnham, 1999). Although these models are fairly simple and robust, in practical situations at least one of the assumptions is often not supported by the data.

Approaches that are commonly used to assess the fit of multiyear tag-recovery models include the formal goodness-of-fit test, Akaike's information criterion (AIC) (Akaike, 1973; Burnham and Anderson, 1992; Burnham et al., 1995) and other related measures such as quasilikelihood AIC (Akaike, 1985). Although these measures are informative about overall model fit, they do not provide any information about why a model fit is poor or which assumption(s) is (are) possibly in violation. To remedy this problem, Latour et al. (2001a) conducted a series of simulations and demonstrated that distinct patterns in model residuals will be evident if particular assumptions are violated. They discussed in detail the residuals associated with the time-specific parameterizations of the Seber (1970) and Brownie et al. (1985) models, as well as the time-specific instantaneous rates model developed by Hoenig et al. (1998).

The genesis of the work by Latour et al. (2001a) can be traced to two particular applications of multiyear tag-recovery models. Specifically, Latour et al. (2001b) analyzed tag-recovery data of red drum (*Sciaenops ocellatus*) in South Carolina and found systematic patterns along the diagonals in the upper right corner of the residuals matrix. Frusher and Hoenig (2001) applied a series of tag-recovery models to Australian rock lobster (*Jasus edwardsii*) data and found consistent patterns in the columns of the residuals matrix. In both instances, the researchers could only speculate as to the cause of these patterns in residuals. Although the simulations conducted by Latour et al. (2001a) have since provided reasonable explanations for the observed patterns, the development of those diagnostic procedures led to the discovery that the residuals associated with the time-specific Seber (1970) and Brownie et al. (1985) models are subject to several constraints.

This note contains a series of simple mathematical arguments that verify the assertions made by Latour et al. (2001a) about the residuals of the time-specific parameterizations of the Seber (1970) and Brownie et al. (1985) models. Unfortunately, the constraints inherent to the residuals of those models partially cloud a researcher's ability to assess the existence of a pattern. As such, knowledge of the inherent properties of the residuals of these models is of particular importance, especially because the time-specific parameterizations are commonly used for the analysis of tag-recovery data.

Materials and methods

Multiyear tag-recovery models

Multiyear tagging data are generally represented by an upper triangular matrix of tag recoveries. For example, the matrix for a study with I years of tagging and J years of tag-recovery would be, when $I = J,$

$$r = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1J} \\ - & r_{22} & \cdots & r_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ - & - & - & r_{IJ} \end{bmatrix}, \qquad (1)$$

where r_{ij} = the number of tags recovered in year j that were released in year i (note, $i = 1, \ldots , I; j = i, \ldots , J$).

Application of multiyear tag-recovery models generally involves constructing a matrix of expected values and comparing them to the observed

data. The matrix of expected values corresponding to the time-specific parameterization of Brownie et al. (1985), which is referred to as model 1, takes the form

$$E_r = \begin{bmatrix} N_1f_1 & N_1S_1f_2 & \cdots & N_1(S_1\cdots S_{J-1})f_J \\ - & N_2f_2 & \cdots & N_2(S_2\cdots S_{J-1})f_J \\ \vdots & \vdots & \ddots & \vdots \\ - & - & - & N_If_J \end{bmatrix}, \qquad (2)$$

where N_i = the number tagged in year i;
   f_i = the tag recovery rate in year i; and
   S_i = the survival rate in year i.

   As stated above, the Brownie et al. (1985) models constitute a generalization of those developed by Seber (1970). The only difference lies in the definition of the tag recovery rate. Specifically, Seber (1970) modeled the tag recovery rate in year i as $f_i = (1 - S_i)r_i$, where r_i is the rate at which tags are reported from killed fish in year i regardless of the source of mortality. The matrix of expected values associated with time-specific parameterization of the Seber (1970) models, which we will refer to as model 1^*, takes the form (when $I = J$)

$$E_r = \begin{bmatrix} N_1(1-S_1)r_1 & N_1S_1(1-S_2)r_2 & \cdots & N_1(S_1\cdots S_{J-1})(1-S_J)r_J \\ - & N_2(1-S_2)r_2 & \cdots & N_2(S_2\cdots S_{J-1})(1-S_J)r_J \\ \vdots & \vdots & \ddots & \vdots \\ - & - & - & N_1(1-S_J)r_J \end{bmatrix}. (3)$$

The data in each row of Equation 1 follow a multinomial distribution and maximum likelihood estimation can be used to derive parameter estimates from either model 1 or model 1^*. Program MARK has emerged as the leading software package for deriving these estimates (White and Burnham, 1999).

Patterns in residuals

Latour et al. (2001a) manipulated a hypothetical perfect data set (i.e. the observed number of tag recoveries was equal to the expected number of tag recoveries) to simulate four specific forms of assumption violation for multiyear tag-recovery models. For each scenario, they analyzed the modified data with model 1, model 1^*, and a time-specific parameterization of the instantaneous rates (IR) models (Hoenig et al., 1998) and noted any patterns in the residuals matrix that resulted from each particular assumption violation. Specifically, they found the following: 1) the presence of nonmixing (which violates the assumption that the tagged population is representative of the target population) leads to consistent patterns on the main and super diagonals of the residuals matrix (the main diagonal contains the $(1,1),(2,2),\ldots,(I,I)$ cells and the first super diagonal contains the $(1,2),(2,3),\ldots(I-1,I)$ cells in a square matrix); 2) permanent emigration from the study area of individuals within a tagged cohort (which violates the assumption that all tagged fish within a cohort are subject to the same annual survival and tag-recovery rates) leads

to a pattern of negative residuals along the diagonals of the upper right corner of the residuals matrix; 3) tag-induced mortality or immediate loss of tags due to poor tagging (which violates the assumptions that tags are not lost and survival rates are not affected by tagging) leads to row patterns in the residuals matrix (note that these patterns are detectable only in the residuals matrix of the IR model); and 4) a change in the natural mortality rate (which violates the frequently imposed assumption that natural mortality is constant over time) leads to column patterns in the residuals matrix (again, this only applies to the IR model).

Constraints on residuals of model 1 and model 1^*

Latour et al. (2001a) asserted without proof that the residuals associated with model 1 and model 1^* are subject to several constraints. Specifically, they stated that the relationship $E_{II} = r_{II}$ always holds, regardless of the number of years of tagging and tag-recovery (note that E_{ij} is the expected number of tags recovered in year j that were released in year i). This implies that the observed data and the expected value associated with the (1,1) cell are always identical and that the residual for that cell is always equal to zero. They also stated that the residuals associated with the implicit "never seen again" category are also always equal to zero (recall that under a multinomial formulation, one of the possible outcomes is to never recapture a tagged fish). Collectively, these constraints imply that the residuals matrix derived from using model 1 or model 1^* to analyze data from a study with I years of tagging and J years of tag-recovery takes the form

$$resid = \begin{bmatrix} 0.00 & (r_{12} - E_{12}) & \cdots & (r_{1J} - E_{1J}) & 0.00 \\ - & (r_{22} - E_{22}) & \cdots & (r_{2J} - E_{2J}) & 0.00 \\ \vdots & \vdots & \ddots & \vdots & 0.00 \\ - & - & \cdots & (r_{IJ} - E_{IJ}) & 0.00 \end{bmatrix}, \qquad (4)$$

where r_{ij} and E_{ij} are as defined previously and the last column of the matrix represents the residuals associated with the "never seen again" category.

   In addition to the aforementioned zero residuals, Latour et al. (2001a) stated that the sum of each row and each column of the residuals matrix must equal zero and that for the case when $I = J$ (i.e. the recovery matrix is square), the constraint that $E_{II} = r_{II}$ is also present (i.e. the residual associated with the (I, I) cell is always equal to zero).

   In the context of searching for patterns in residuals, these constraints have the following implications. First, the presence of residuals that are constrained to be zero essentially reduces the total number of values that are available for inspection and ultimately forces conclusions about the existence of a pattern to be based on the signs of fewer residuals. For short-term tagging studies (e.g. 3–4 years), the loss of residuals for inspection makes it extremely difficult to evaluate model performance because each row, column, and diagonal of the residuals matrix already contains only a few values. Second, because the sum of each row of the residuals matrix must total zero, and

because the residuals corresponding to the "never seen again" cells are zero, it is not possible for a row pattern to be expressed in the residuals matrices of either model 1 or model 1.[*] This constraint renders it very difficult to detect assumption violations that are cohort-specific—the most common being tag-induced and handling mortality and short-term tag loss.

Results

To verify that the aforementioned constraints about the structure of the residuals matrices associated with model 1 and model 1[*] are true, we offer the following mathematical arguments. The proofs simply involve algebraic manipulation of equations involving the analytical formulae for the maximum likelihood parameter estimates (MLE) of model 1. The formulae for the MLEs were originally developed by Seber (1970) and can be applied to both to model 1 and model 1[*] (because r_i can be expressed as a function of S_i and f_i, the invariance property of MLEs implies that an MLE of r_i can be obtained by the transform). Hence, the proofs are developed for model 1, and we note that similar arguments could be constructed for the residuals of model 1[*].

Recall that the analytical solutions for the maximum likelihood estimates of f_i and S_i from Seber (1970) and Brownie et al. (1985) are given by

$$\hat{f}_i = \frac{R_i C_i}{N_i T_i} \quad \text{and} \quad \hat{S}_i = \frac{R_i}{N_i}\frac{(T_i - C_i)}{T_i}\frac{N_{i+1}}{R_{i+1}},$$

where R_i and C_i are the row and column totals of the observed data in year i; and

$$\begin{aligned} T_1 &= R_1 \\ T_i &= R_i + T_{i-1} - C_{i-1} & i = 2, \dots, I \\ T_{I+j} &= T_{I+j-1} - C_{I+j-1} & j = 1, \dots, J-I \text{ if } J > I. \end{aligned}$$

The (1,1) cell

To show that the residual associated with the (1,1) cell is always zero, we must demonstrate that the difference between the observed and estimated expected value in the first cell is always zero. Hence, we have

$$\begin{aligned} r_{11} - \hat{E}_{11} &= r_{11} - N_1\hat{f}_1 \\ &= r_{11} - N_1\left(\frac{R_1 C_1}{N_1 T_1}\right) \text{(substituting for } \hat{f}_1) \\ &= r_{11} - C_1 \text{ (since } T_1 = R_1) \\ &= r_{11} - r_{11} = 0 \text{ (because the column total} \\ &\qquad\qquad \text{in the first recovery year is } r_{11}). \end{aligned}$$

The (I,I) cell when I = J

To show that the residual associated with the (I,I) cell is always zero when $I = J$, again we must demonstrate

that the difference between the observed and estimated expected values in that cell is always zero. Hence, we have

$$\begin{aligned} r_{II} - \hat{E}_{II} &= r_{II} - N_I\hat{f}_I \\ &= r_{II} - N_I\left(\frac{R_I C_I}{N_I T_I}\right) \text{(substituting for } \hat{f}_I) \\ &= r_{II} - R_I \text{ (because } T_I = C_I) \\ &= r_{II} - r_{II} = 0 \text{ (because the row total in the} \\ &\qquad\qquad \text{final recovery year is } r_{II}). \end{aligned}$$

Column sums when I = J

To show that the column sums of the residuals matrix equal zero, we must demonstrate the column sum of the observed data equals that of the expected values. Consider the sum of the expected values associated with the Ith column of the recovery matrix, that is

$$Q = E_{1I} + E_{2I} + \cdots + E_{II} = N_1(\hat{S}_1\cdots\hat{S}_{I-1})\hat{f}_I + N_2(\hat{S}_2\cdots\hat{S}_{I-1})\hat{f}_I + \cdots + N_{I-1}\hat{S}_{I-1}\hat{f}_I + N_I\hat{f}_I.$$

Now substitute for $\hat{f}_i$ and $\hat{S}_i$ on the right hand side:

$$\begin{aligned} Q = N_1&\left(\left(\frac{R_1}{N_1}\frac{(T_1-C_1)}{T_1}\frac{N_2}{R_2}\right)\cdots\right. \\ &\left.\left(\frac{R_{I-1}}{N_{I-1}}\frac{(T_{I-1}-C_{I-1})}{T_{I-1}}\frac{N_I}{R_I}\right)\right)\left(\frac{R_I C_I}{N_I T_I}\right)+ \\ N_2&\left(\left(\frac{R_2}{N_2}\frac{(T_2-C_2)}{T_2}\frac{N_3}{R_3}\right)\cdots\left(\frac{R_{I-1}}{N_{I-1}}\frac{(T_{I-1}-C_{I-1})}{T_{I-1}}\frac{N_I}{R_I}\right)\right)\left(\frac{R_I C_I}{N_I T_I}\right)+ \\ \cdots &+ N_{I-1}\left(\frac{R_{I-1}}{N_{I-1}}\frac{(T_{I-1}-C_{I-1})}{T_{I-1}}\frac{R_I}{N_I}\right)\left(\frac{R_I C_I}{N_I T_I}\right) + N_I\left(\frac{R_I C_I}{N_I T_I}\right). \end{aligned}$$

Cancel terms and factor out the term $\frac{C_I}{T_I}$:

$$\begin{aligned} Q = &\left(\frac{C_I}{T_I}\right)\left[\left(\frac{R_1(T_1-C_1)}{T_1}\right)\left(\frac{T_2-C_2}{T_2}\right)\cdots\left(\frac{T_{I-1}-C_{I-1}}{T_{I-1}}\right)+ \right. \\ &\left.\left(\frac{R_2(T_2-C_2)}{T_2}\right)\left(\frac{T_3-C_3}{T_3}\right)\cdots\left(\frac{T_{I-1}-C_{I-1}}{T_{I-1}}\right)\right]+ \\ &+\cdots+\left(\frac{C_I}{T_I}\right)\left[\left(\frac{R_{I-1}(T_{I-1}-C_{I-1})}{T_{I-1}}\right)+R_I\right]. \end{aligned}$$

Systematically factor out terms of the form $\frac{(T_i - C_i)}{T_i}$ (recall that $T_1 = R_1$):

$$Q = \left(\frac{C_I}{T_I}\right)\left[\left(\frac{T_{I-1}-C_{I-1}}{T_{I-1}}\right)\left(\frac{T_{I-2}-C_{I-2}}{T_{I-2}}\right)\right]$$

$$\left[\cdots\left[\left(\frac{T_2-C_2}{T_2}\right)(R_1-C_1+R_2)+R_3\right]+R_4\right]+\cdots+R_{I-1}\right]+R_I\right].$$

Utilize the definition of $T_i = R_i + T_{i-1} - C_{i-1}$ to systematically simplify and cancel starting with the innermost parenthetic expression:

$$Q = \left(\frac{C_I}{T_I}\right)\left[\left(\frac{T_{I-1}-C_{I-1}}{T_{I-1}}\right)T_{I-1}+R_I\right]$$

$$= \left(\frac{C_I}{T_I}\right)[T_{I-1}-C_{I-1}+R_I]$$

$$= \left(\frac{C_I}{T_I}\right)T_I = C_I = r_{1I}+r_{2I}+\cdots+r_{IJ},$$

which demonstrates that the column sum of expected values equals the column sum of observed recoveries, as desired for the Ith column. Similar arguments hold for the other columns.

Column sums when $I > J$

The proof that the column sums of the residuals matrix equal zero when the recovery matrix is nonsquare is similar to the proof above except for making use of the definition $T_{I+j} = T_{I+j-1} - C_{I+j-1}$.

Row sums when $I = J$

To show that the row sums (excluding the "never seen again" cell) of the residuals matrix equal zero, we must demonstrate the sum of the observed data equals that of the expected values. Consider the sum of the expected values associated with the first row of the recovery matrix:

$$Q = E_{11} + E_{12} + \cdots + E_{1I} = N_1\hat{f}_1 + N_1\hat{S}_1\hat{f}_2 + \cdots + N_1(\hat{S}_1\cdots\hat{S}_{I-1})\hat{f}_I.$$

Now substitute for $\hat{f}_i$ and $\hat{S}_i$ on the right hand side:

$$Q = N_1\left(\frac{R_1C_1}{N_1T_1}\right) + N_1\left(\frac{R_1}{N_1}\frac{(T_1-C_1)}{T_1}\frac{N_2}{R_2}\right)\left(\frac{R_2C_2}{N_2T_2}\right) +$$

$$\cdots + N_1\left(\left(\frac{R_1}{N_1}\frac{(T_1-C_1)}{T_1}\frac{N_2}{R_2}\right)\cdots\right.$$

$$\left.\left(\frac{R_{I-1}}{N_{I-1}}\frac{(T_{I-1}-C_{I-1})}{T_{I-1}}\frac{N_I}{R_I}\right)\right)\left(\frac{R_IC_I}{N_IT_I}\right).$$

Cancel and factor out the term $T_1 - C_1$ (recall that $T_1 - R_1$):

$$Q = \left(\frac{R_1C_1}{T_1}\right) + (T_1-C_1)\left[\frac{C_2}{T_2} + \left(\frac{T_2-C_2}{T_2}\right)+\right]$$

$$\cdots + \left(\frac{T_2-C_2}{T_2}\right)\left(\frac{T_3C_3}{T_3}\right)\cdots\left(\frac{T_{I-1}-C_{I-1}}{T_{I-1}}\right)\left(\frac{C_I}{T_I}\right)\right].$$

Systematically factor out terms of the form $\frac{(T_i - C_i)}{T_i}$:

$$Q = \left(\frac{R_1C_1}{T_1}\right) + (T_1-C_1)$$

$$\left[\frac{C_2}{T_2}\left(\frac{T_2-C_2}{T_2}\right)\left[\frac{C_3}{T_3}+\cdots+\left[\frac{C_{I-1}}{T_{I-1}}+\left(\frac{T_{I-1}-C_{I-1}}{T_{I-1}}\right)\right]\left[\frac{C_I}{T_I}\right]\cdots\right]\right].$$

The expression inside the innermost square brackets is equal to 1 (recall $T_I = C_I$). Hence, we have

$$Q = \left(\frac{R_1C_1}{T_1}\right) + (T_1-C_1) = \frac{R_1C_1 + R_1T_1 - R_1C_1}{T_1} = R_1$$

$$= r_{11} + r_{12} + \cdots + r_{1I},$$

which shows the sum of the expected values in row 1 is equal to the sum of the observed data. Similar arguments hold for the other rows.

Row sums when $I > J$

As with the proof of the column sums when $I > J$, the definition $T_{I+j} = T_{I+j-1} - C_{I+j-1}$ is needed to show the row sum of a nonsquare matrix (excluding the "never seen again" cells) are zero.

"Never seen again" cells

The likelihood function for the Brownie-type model is a product multinomial and the parameters for each row are constrained to sum to one. Therefore, the expected values in a row are simply an apportionment of the number tagged to the years of recovery and the "never seen again" category. Hence, the sum of the estimated expected values has to equal the row sum (including the "never seen again" cell), which implies the residuals of the "never seen again" cells are always equal to zero.

Discussion

The residuals of multiyear tag-recovery models can be very helpful for evaluating model performance. Unfortunately, examining the residuals matrix for patterns is not a commonly employed procedure for assessing model fit in practical situations. The work by Latour et al. (2001a) was intended

to demonstrate the insight a researcher can acquire by using residuals as a diagnostic probe to gauge the possibility of assumption violation. Similarly, the work presented here is intended to further guide researchers by explicitly delineating the properties of the residuals associated with two commonly applied multiyear tag-recovery models.

Model 1 and model 1^* represent parameterizations of only two classes of tag-recovery models. Properties of the residuals associated with other classes of models (e.g. movement models, age-structured tag-recovery models, capture-recapture models) have not been studied. We feel strongly that similar types of insight about model performance and model fit can be acquired by examining the residuals for patterns. As such, we recommend that the residuals from other classes of models be more thoroughly investigated.

Literature cited

Akaike, H.
    1973.   Information theory as an extension of the maximum likelihood principle.   *In* Second international symposium on information theory (B. N. Petrov and F. Csaki. eds.), p. 267–281.   Akademiai Kiado, Budapest.
Akaike, H.
    1985.   Prediction and entropy.   *In* A celebration of statistics (A. C. Atkinson and S. E. Fienberg, eds.), p. 1–24.   Springer, New York, NY.
Brownie, C., D. R. Anderson, K. P. Burhnam, and D. S. Robson.
    1985.   Statistical inference from band recovery data: a hand-book, 2nd ed., 305 p.   U.S. Fish and Wildl. Serv. Resour. Publ. 156.
Burnham, K. P., and D. R. Anderson.
    1992.   Data-based selection of an appropriate biological model: the key to modern data analysis.   *In* Wildlife 2001: populations (D. R. McCullough and R. H. Barrett, eds.), p. 16–30.   Elsevier Science Publishers, London.
Burnham, K. P., G. C. White, and D. R. Anderson.
    1995.   Model selection strategy in the analysis of capture-recapture data.   Biometrics 51:888–898.
Frusher, S. D., and J. M. Hoenig.
    2001.   Estimating natural and fishing mortality and tag reporting rate of rock lobster from a multiyear tagging model.   Can. J. Fish. Aquat. Sci. 58:2490–2501.
Hoenig, J. M., N. J. Barrowman, W. S. Hearn, and K. H. Pollock.
    1998.   Multiyear tagging studies incorporating fishing effort data.   Can. J. Fish. Aquat. Sci. 55:1466–1476.
Latour, R. J., J. M. Hoenig, J. E. Olney, and K. H. Pollock.
    2001a.   Diagnostics for multiyear tagging models with application to Atlantic striped bass (*Morone saxatilis*).   Can. J. Fish. Aquat. Sci. 58:1717–1726.
Latour, R. J., K. H. Pollock, C. A. Wenner, and J. M. Hoenig.
    2001b.   Estimates of fishing and natural mortality for red drum *Sciaenops ocellatus* in South Carolina waters.   N. Am. J. Fish. Manage. 21:733–744.
Seber, G. A. F.
    1970.   Estimating time-specific survival and reporting rates for adult birds from band returns.   Biometrika 57:313–318.
White, G. C., and K. P. Burnham.
    1999.   Program MARK–survival estimation from populations of marked animals.   Bird Study 46:120–138.