Lecture 7 – Models for open populations: Tag recovery and CJS models, Goodness-of-fit

Resources

Chapter 5 *in* Goodness of fit in E. Cooch and G.C. White (eds.), Program MARK: A gentle introduction. http://www.phidot.org/software/mark/docs/book/ 26 September 2007.

Pollock, K.H, Nichols, J.D., Brownie, C. & Hines, J.E. (1990) Statistical inference for capture-recapture experiments. Wildlife Monographs, 107, 1-97

Pollock, K. H., J. D. Nichols, C. Brownie, and J. E. Hines. 1990. Statistical inference for capture-recapture experiments. Wildlife Monographs 107.

Seber, G. A. F. 1982. The estimation of animal abundance and related parameters, 2nd ed. Macmillan, New York, NY.

Chapter 8 *in* White, G. C., Anderson, D.R., Burnham, K.P., and Otis. D.L. 1982. Capture-recapture and removal methods for sampling closed populations. Los Alamos National Laboratory, Rep. LA-8787-NERP. 235pp.

m-arrays

CMR data are often displayed in the form of an ***m*-array**. The m-array is a very concise way to display the data as it relates to the capture histories and the multinomial likelihoods. However capture histories can no be constructed from an ***m***-array.

Each row represents a cohort of marked individuals released on occasion i. This number includes those initially marked at other times and released again at time i. The columns indicate the sampling occasions 2 through m. The values in the array indicate the number of individuals released at i, R(i), and next encountered at time j. The total on the right is the number of individuals released at R(i) that were seen again.

		Occasion									
Occasion	R(i)	j= 2	3	4	5	6	7	8	9	10	$\sum_{i=1}^{m} j_i$
1	51	12	2	8	7	4	3	2	1	4	43
2	49		7	9	1	4	2	1	2	1	27
3	50			5	7	4	5	7	1	2	31
4	51				5	10	5	3	7	3	33
5	49					10	7	3	2	5	27
6	50						6	4	5	7	22
7	50							8	6	1	15
8	49								11	4	15
9	52									4	4

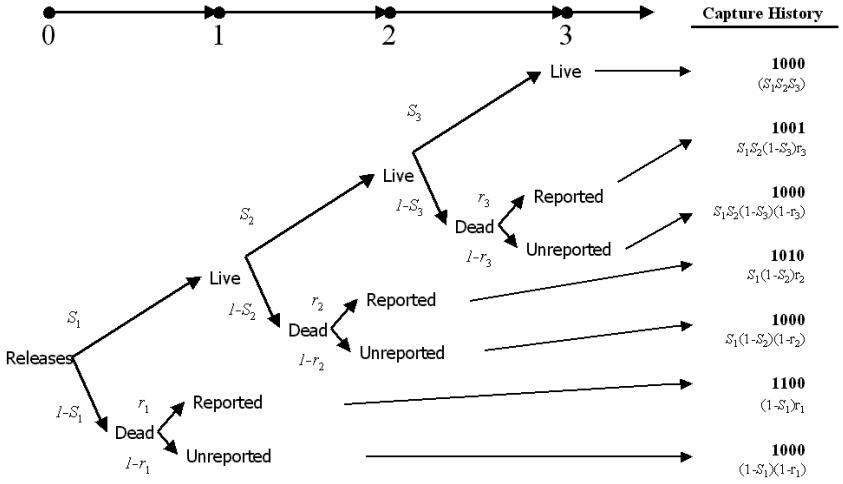
Both band recoveries (dead animals) and recapture/re-sighting data are frequently summarized in this format. Live releases are treated as newly marked animals in year i + 1.

Remember that **each cell** in the m-array has an associated probability, and **each row** has its own multinomial distribution. It is the simultaneous solution of all of the likelihoods that makes this analysis so useful.

Band recovery models

Models are systems of postulates, data, and inferences presented as a mathematical description of an entity or state of affairs (Merriam-Webster). In the context of survival analysis we can think of models are the linkage (via likelihood theory) between the data and an equation or set of assumptions.

There is a large class of models for the analysis of bird banding and fish tagging data. These are



open population models and are generally employed to estimate survival rates based on markers returned from harvested animals. Although there is an assumption that recoveries occurring during a short window of time, it has been demonstrated that these models are relatively robust to violations of this assumption. The figure above illustrates the relationships among the estimated parameters, the timeline of studies and the construction of capture histories for a single release of individuals in a tag recovery study.

1. Constants

   R_i - number of animals marked in year i.

2. Random Variables

   mij - number of bands or tags reported in year j from releases in year i. The m-array of recoveries.

3. Parameters of interest (pij)

   Sj - conditional probability of surviving in period j (the interval between recovery periods), given the animal is alive at the beginning of year j.

   rj - conditional probability of being reported in year j, given the animal died in year j. This parameter is referred to as the recovery probability fi in some publications.

4. Cell probabilities

   It is useful to think of the random variables, m_{ij}, in terms of the cell probabilities and expected number of recoveries. This information will be applied later in assessing goodness of fit.

   The expected values can be computed by multiplying the cell probabilities in each row (cohort) by the number of releases (R_i). Here we examine the 15 cell probabilities for the **saturated model** in a study with 5 releases and 5 sampling (encounter) occasions.

5. Saturated, Global, or General Model.

Cell Probabilities for tag recovery(p_{ij})					
Occasion					
1	2	3	4	5	Not recovered
$(1-S_1)\ r_1$	$S_1(1-S_2)r_2$	$S_1S_2(1-S_3)r_3$	$S_1S_2S_3(1-S_4)r_4$	$S_1S_2S_3S_4(1-S_5)r_5$	$1-\Sigma((p_{1j})$
	$(1-S_6)r_6$	$S_6(1-S_7)r_7$	$S_6S_7(1-S_8)r_8$	$S_6S_7S_8(1-S_9)r_9$	$1-\Sigma((p_{2j})$
		$(1-S_{10})r_{10}$	$S_{10}(1-S_{11})r_{11}$	$S_{10}S_{11}(1-S_{12})r_{12}$	$1-\Sigma((p_{3j})$
			$(1-S_{13})r_{13}$	$S_{13}(1-S_{14})r_{14}$	$1-\Sigma((p_{4j})$
				$(1-S_{15})r_{15}$	$1-\Sigma((p_{5j})$

Expected tag recoveries(E(m_{ij}))						
	Occasion					
Released	1	2	3	4	5	Not recovered
R_1	$R_1(_1\text{-}S_1)r_1$	$R_1S_1(_1\text{-}S_2)r_2$	$R_1S_1S_2(_1\text{-}S_3)r_3$	$R_1S_1S_2S_3(_1\text{-}S_4)r_4$	$R_1S_1S_2S_3S_4(_1\text{-}S_5)r_5$	$R_1\text{-}\Sigma(E(m_{1J}))$
R_2		$R_2(_1\text{-}S_6)r_6$	$R_2S_6(_1\text{-}S_7)r_7$	$R_2S_6S_7(_1\text{-}S_8)r_8$	$R_2S_6S_7S_8(_1\text{-}S9)r9$	$R_2\text{-}\Sigma(E(m_{2J}))$
R_3			$R_3(_1\text{-}S_10)r_10$	$R_3S_10(_1\text{-}S_{11})r_{11}$	$R_3S_10S_{11}(_1\text{-}S_{12})r_{12}$	$R_3\text{-}\Sigma(E(m_{3J}))$
R_4				$R_4(_1\text{-}S_{13})r_{13}$	$R_4S_{13}(_1\text{-}S_{14})r_{14}$	$R_4\text{-}\Sigma(E(m_{4J}))$
R_5					$R_5(_1\text{-}S_{15})r_{15}$	$R_5\text{-}\Sigma(E(m_{5J}))$

NOTE: This model has as many parameters as cells, and while it is important conceptually and for testing goodness-of-fit it is generally of little biological interest. It has deviance = 0 and is the model which best fits the data, but it also has the maximum number of parameters.

6. Model $\{S_t, r_t\}$ the subscript t indicates that parameters vary over time.

Cell Probabilities for tag recovery(p_{ij})					
Occasion					
1	2	3	4	5	Not recovered
$(1-S_1)\ r_1$	$S_1(1-S_2)r_2$	$S_1S_2(1-S_3)r_3$	$S_1S_2S_3(1-S_4)r_4$	$S_1S_2S_3S_4(1-S_5)r_5$	$1-\Sigma(p_{1j})$
	$(1-S_2)r_2$	$S_2(1-S_3)r_3$	$S_2S_3(1-S_4)r_4$	$S_2S_3S_4(1-S_5)r_5$	$1-\Sigma(p_{2j})$
		$(1-S_3)r_3$	$S_3(1-S_4)r_4$	$S_2S_3S_4(1-S_5)r_5$	$1-\Sigma(p_{3j})$

		$(1-S_4)r_4$	$S_4(1-S_5)r_5$	$1-\Sigma(p_{4j})$
			$(1-S_5)r_5$	$1-\Sigma(p_{5j})$

Expected tag recoveries($E(m_{ij})$)						
	Occasion					
Released	1	2	3	4	5	Not recovered
R_1	$R_1(1-S_1)r_1$	$R_1S_1(1-S_2)r_2$	$R_1S_1S_2(1-S_3)r_3$	$R_1S_1S_2S_3(1-S_4)r_4$	$R_1S_1S_2S_3S_4(1-S_5)r_5$	$R_1-\Sigma(E(m_{1j}))$
R_2		$R_2(1-S_2)r_2$	$R_2S_2(1-S_3)r_3$	$R_2S_2S_3(1-S_4)r_4$	$R_2S_2S_3S_4(1-S_5)r_5$	$R_2-\Sigma(E(m_{2j}))$
R_3			$R_3(1-S_3)r_3$	$R_3S_3(1-S_4)r_4$	$R_3S_3S_4(1-S_5)r_5$	$R_3-\Sigma(E(m_{3j}))$
R_4				$R_4(1-S_4)r_4$	$R_4S_4(1-S_5)r_5$	$R_4-\Sigma(E(m_{4j}))$
R_5					$R_5(1-S_5)r_5$	$R_5-\Sigma(E(m_{5j}))$

7. Model $\{S., r_t\}$ survival constant; recoveries vary among years

Cell Probabilities for tag recovery(p_{ij})					
Occasion					
1	2	3	4	5	Not recovered
$(1-S_1)\ r_1$	$S_1(1-S_1)r_2$	$S_1S_1(1-S_1)r_3$	$S_1S_1S_1(1-S_1)r_4$	$S_1S_1S_1S_1(1-S_1)r_5$	$1-\Sigma(p_{1j})$
	$(1-S_1)r_2$	$S_1(1-S_1)r_3$	$S_1S_1(1-S_1)r_4$	$S_1S_1S_1(1-S_1)r_5$	$1-\Sigma(p_{2j})$
		$(1-S_1)r_3$	$S_1(1-S_1)r_4$	$S_1S_1S_1(1-S_1)r_5$	$1-\Sigma(p_{3j})$
			$(1-S_1)r_4$	$S_1(1-S_1)r_5$	$1-\Sigma(p_{4j})$
				$(1-S_1)r_5$	$1-\Sigma(p_{5j})$

Expected tag recoveries($E(m_{ij})$)						
	Occasion					
Released	1	2	3	4	5	Not recovered
R_1	$R_1(1-S_1)r_1$	$R_1S_1(1-S_1)r_2$	$R_1S_1S_1(1-S_1)r_3$	$R_1S_1S_1S_1(1-S_1)r_4$	$R_1S_1S_1S_1S_1(1-S_1)r_5$	$R_1-\Sigma(E(m_{1j}))$
R_2		$R_2(1-S_1)r_2$	$R_2S_1(1-S_1)r_3$	$R_2S_1S_1(1-S_1)r_4$	$R_2S_1S_1S_1(1-S_1)r_5$	$R_2-\Sigma(E(m_{2j}))$
R_3			$R_3(1-S_1)r_3$	$R_3S_1(1-S_1)r_4$	$R_3S_1S_1(1-S_1)r_5$	$R_3-\Sigma(E(m_{3j}))$
R_4				$R_4(1-S_1)r_4$	$R_4S_1(1-S_1)r_5$	$R_4-\Sigma(E(m_{4j}))$
R_5					$R_5(1-S_1)r_5$	$R_5-\Sigma(E(m_{5j}))$

8. Model $\{S., r.\}$ survival and recovery constant

Cell Probabilities for tag recovery(p_{ij})					
Occasion					
1	2	3	4	5	Not recovered
$(1-S_1)\ r_1$	$S_1(1-S_1)r_1$	$S_1S_1(1-S_1)r_1$	$S_1S_1S_1(1-S_1)r_1$	$S_1S_1S_1S_1(1-S_1)r_1$	$1-\Sigma(p_{1j})$

$(1-S_1)r_1$	$S_1(1-S_1)r_1$	$S_1S_1(1-S_1)r_1$	$S_1S_1S_1(1-S_1)r_1$	$1-\Sigma(p_{2j})$
	$(1-S_1)r_1$	$S_1(1-S_1)r_1$	$S_1S_1S_1(1-S_1)r_1$	$1-\Sigma(p_{3j})$
		$(1-S_1)r_1$	$S_1(1-S_1)r_1$	$1-\Sigma(p_{4j})$
			$(1-S_1)r_1$	$1-\Sigma(p_{5j})$

Expected tag recoveries($E(m_{ij})$)						
	Occasion					
Released	1	2	3	4	5	Not recovered
R_1	$R_1(1-S_1)r_1$	$R_1S_1(1-S_1)r_1$	$R_1S_1S_1(1-S_1)r_1$	$R_1S_1S_1S_1(1-S_1)r_1$	$R_1S_1S_1S_1S_1(1-S_1)r_1$	$R_1-\Sigma(E(m_{1j}))$
R_2		$R_1(1-S_1)r_1$	$R_1S_1(1-S_1)r_1$	$R_1S_1S_1(1-S_1)r_1$	$R_1S_1S_1S_1(1-S_1)r_1$	$R_2-\Sigma(E(m_{2j}))$
R_3			$R_1(1-S_1)r_1$	$R_1S_1(1-S_1)r_1$	$R_1S_1S_1(1-S_1)r_1$	$R_3-\Sigma(E(m_{3j}))$
R_4				$R_1(1-S_1)r_1$	$R_1S_1(1-S_1)r_1$	$R_4-\Sigma(E(m_{4j}))$
R_5					$R_1(1-S_1)r_1$	$R_5-\Sigma(E(m_{5j}))$

9. Model $\{S., r_3\}$ survival constant and recovery different in year 3

Cell Probabilities for tag recovery(p_{ij})					
Occasion					
1	2	3	4	5	Not recovered
$(1-S_1)r_1$	$S_1(1-S_1)r_1$	$S_1S_1(1-S_1)r_3$	$S_1S_1S_1(1-S_1)r_1$	$S_1S_1S_1S_1(1-S_1)r_1$	$1-\Sigma(p_{1j})$
	$(1-S_1)r_1$	$S_1(1-S_1)r_3$	$S_1S_1(1-S_1)r_1$	$S_1S_1S_1(1-S_1)r_1$	$1-\Sigma(p_{2j})$
		$(1-S_1)r_3$	$S_1(1-S_1)r_1$	$S_1S_1S_1(1-S_1)r_1$	$1-\Sigma(p_{3j})$
			$(1-S_1)r_1$	$S_1(1-S_1)r_1$	$1-\Sigma(p_{4j})$
				$(1-S_1)r_1$	$1-\Sigma(p_{5j})$

Specifying tag recovery models using PIMs in MARK

Like m-arrays, in PIMs the columns correspond to occasions and rows correspond to cohorts (releases). In MARK each parameter gets an index number; parameters that are constrained to be equal are given the same index number.

For example the cell probabilities for the Model: $\{S.r.\}$ are:

Cell Probabilities for tag recovery(p_{ij})				
1	2	3	4	5
$(1-S_1)r_1$	$S_1(1-S_1)r_1$	$S_1S_1(1-S_1)r_1$	$S_1S_1S_1(1-S_1)r_1$	$S_1S_1S_1S_1(1-S_1)r_1$
	$(1-S_1)r_1$	$S_1(1-S_1)r_1$	$S_1S_1(1-S_1)r_1$	$S_1S_1S_1(1-S_1)r_1$
		$(1-S_1)r_1$	$S_1(1-S_1)r_1$	$S_1S_1S_1(1-S_1)r_1$
			$(1-S_1)r_1$	$S_1(1-S_1)r_1$
				$(1-S_1)r_1$

Since all of the S's and all of the $r's$ are constrained to be equal, only two parameters are estimated and the corresponding PIMs are:

Survival Probabilities					Reporting Probabilities				
1	1	1	1	1	2	2	2	2	2
	1	1	1	1		2	2	2	2
		1	1	1			2	2	2
			1	1				2	2
				1					2

Similarly the cell probabilities for the Model: $\{S_t r_.\}$ are:

Cell Probabilities for tag recovery(p_{ij})				
Occasion				
1	2	3	4	5
$(1-S_1)\ r_1$	$S_1(1-S_2)r_1$	$S_1S_2(1-S_3)r_1$	$S_1S_2S_3(1-S_4)r_1$	$S_1S_2S_3S_4(1-S_5)r_1$
	$(1-S_2)r_1$	$S_2(1-S_3)r_1$	$S_2S_3(1-S_4)r_1$	$S_2S_3S_4(1-S_5)r_1$
		$(1-S_3)r_1$	$S_3(1-S_4)r_1$	$S_2S_3S_4(1-S_5)r_1$
			$(1-S_4)r_1$	$S_4(1-S_5)r_1$
				$(1-S_5)r_1$

The S's are year specific (n = 5) and all of the r's are constrained to be equal, so a total of six parameters are estimated:

Survival Probabilities					Reporting Probabilities				
1	2	3	4	5	6	6	6	6	6
	2	3	4	5		6	6	6	6
		3	4	5			6	6	6
			4	5				6	6
				5					6

The cell probabilities for an age specific Model: $\{S_j S_s S_a r\}$ where animals are marked as young with different survival probabilities for each of three age classes (juvenile, subadult, and adult) and a single recovery probability are:

Cell Probabilities for tag recovery(p_{ij})
Occasion

$(1-S_j)r$	$S_j(1-S_s)r$	$S_jS_s(1-S_a)r$	$S_jS_2S_a(1-S_a)r$	$S_jS_2S_aS_a(1-S_a)r$
	$(1-S_j)r$	$S_j(1-S_s)r$	$S_jS_s(1-S_a)r$	$S_jS_2S_a(1-S_a)r$
		$(1-S_j)r$	$S_j(1-S_s)r$	$S_jS_s(1-S_a)r$
			$(1-S_j)r$	$S_j(1-S_s)r$
				$(1-S_j)r$

The corresponding PIMs are:

Survival Probabilities					Reporting Probabilities				
1	2	3	3	3	4	4	4	4	4
	1	2	3	3		4	4	4	4
		1	2	3			4	4	4
			1	2				4	4
				1					4

The cell probabilities for $S_t r_t$, the **SATURATED MODEL**, are:

Cell Probabilities for tag recovery(p_{ij})				
Occasion				
$(1-S_1)r_1$	$S_1(1-S_2)r_2$	$S_1S_2(1-S_3)r_3$	$S_1S_2S_3(1-S_4)r_4$	$S_1S_2S_3S_4(1-S_5)r_5$
	$(1-S_6)r_6$	$S_6(1-S_7)r_7$	$S_6S_7(1-S_8)r_8$	$S_6S_7S_8(1-S_9)r_9$
		$(1-S_{10})r_{10}$	$S_{10}(1-S_{11})r_{11}$	$S_{10}S_{11}(1-S_{12})r_{12}$
			$(1-S_{13})r_{13}$	$S_{13}(1-S_{14})r_{14}$
				$(1-S_{15})r_{15}$

The corresponding PIMs are:

Survival Probabilities					Reporting Probabilities				
1	2	3	4	5	16	17	18	19	20
	6	7	8	9		21	22	23	24
		10	11	12			25	26	27
			13	14				28	29
				15					30

Multiple Groups

Multiple groups can represent any number of situations. In the recent monograph on the Northern Spotted Owl, models considering 100 identifiable groups were identified. The most common use of groups occurs when parallel studies are conducted on (1) males and females, (2) different study

areas, or (3) animals experimentally subjected to different treatments. For example White et al. present data taken from Brownie et al.: 146) for male and female mallards in m-array form:

MALES						FEMALES					
	Occasion (j)						Occasion (j)				
R_i	1	2	3	4	5	R_i	1	2	3	4	5
2583	91	89	24	18	16	1478	40	31	8	11	2
3075		141	45	52	50	1525		72	20	15	7
1195			27	31	21	319			8	7	3
3418				156	92	1805				63	27
3100					113	1400					39

With 2 groups a much larger array of models becomes obvious. At the extremes are models $\{S., r.\}$ with constant survival across groups and over time and $\{S_{g*t}, r_{g*t}\}$ where both S and r are time-specific. An important concept to grasp at this point is that it is crucial to develop a suitable family of models *a priori* and not to run all possible models. This strategy is often referred to as *data-dredging* which often leads to "over-fitting" and the "detection" of spurious effects.

Unequal time intervals

When time intervals are unequal, CMR analysis handles data in much the same way we use DSRs to calculate nest success. Thus, you specify the relative length of the intervals t and interval survival rates are calculated based on $S_i^{t_i}$, that is the survival rate (e.g., annual) raised to the power of the length of the interval.

For example if there was no fishery in year three of a fish tagging study, the cell probabilities would become.

Cell probabilities			
1	2	3-4	5
$(1\text{-}S_1)r$	$S_1(1\text{-}S_2)r$	$S_1\mathbf{S_2S_3}(1\text{-}S_4)r$	$S_1\mathbf{S_2S_3}S_4(1\text{-}S_5)r$
	$(1\text{-}S_2)r$	$\mathbf{S_2S_3}(1\text{-}S_4)r$	$\mathbf{S_2S_3}S_4(1\text{-}S_5)r$
		$(1\text{-}S_4)r$	$S_4(1\text{-}S_5)r$
			$(1\text{-}S_5)r$

Identifiability

Some parameters cannot be "identified," i.e., a unique solution cannot be achieved. This is somewhat analogous to attempting to solve for a regression model with only one observation (n = 1).

1. Lack of identifiability is a common problem among models where r varies by year or age unless more than one cohort is marked.
2. Another common problem in the Model: $\{S_t\, r_t\}$ is the lack of identifiability in the last term in each row of the m-array (in red):

$$\begin{array}{ccccc} (1\text{-}S_1)r_1 & S_1(1\text{-}S_2)r_2 & S_1S_2(1\text{-}S_3)r_3 & S_1S_2S_3(1\text{-}S_4)r_4 & S_1S_2S_3S_4(1\text{-}S_5)r_5 \\ & (1\text{-}S_2)r_2 & S_2(1\text{-}S_3)r_3 & S_2S_3(1\text{-}S_4)r_4 & S_2S_3S_4(1\text{-}S_5)r_5 \end{array}$$

$$\begin{matrix} (1\text{-}S_3)r_3 & S_3(1\text{-}S_4)r_4 & S_3S_4(1\text{-}S_5)r_5 \\ & (1\text{-}S_4)r_4 & S_4(1\text{-}S_5)r_5 \\ & & (1\text{-}S_5)r_5 \end{matrix}$$

Because the terms S_5 and r_5 do not appear separately anywhere else in the matrix only $(1\text{-}S_5)r_5$ they can be estimated.

3. Identifiability also becomes an issue in several other situations:
   a. When K can not be determined by numerical methods:
   b. When parameter estimates ($\hat{S}$ or $\hat{r}$) lie on or near 0 or 1 (the boundary).
   c. When data are sparse. This problem arises when the recoveries or encounters are not sufficient to fill all of the cells in the m-array, thus all of the terms in the likelihood are not estimable.

Cormack-Jolly-Seber models

These models are similar to the band/tag recovery models with a few important differences. Animals are captured, given a unique mark, and released. On subsequent occasions marked and unmarked animals are captured. Marked animals are recorded and released. Unmarked animals are marked and released. Accidental deaths on capture are allowed and recorded. The general Jolly-Seber model allows year specific estimates of apparent survival, capture probability, population size, and number of individuals entering the population. While population size and the number of new individuals can be estimated they are subject to substantial bias due to heterogeneity and other issues.

The Cormack-Jolly-Seber model is a restricted model that estimates time-specific apparent survival rates and recapture probabilities. An important issue is that only apparent survival (i.e., 1 – (mortality + emigration)) is estimated, thus apparent survival $\leq$ the true survival rate. Thus studies are restricted to a specific locality.

MARK example data set (dipper.dbf)

1. Constants

   R_i – The number of animals released in year i is known and included the number of previously marked animals recaptured and re-released in year i.

2. Random variables

   m_{ij} – As before, these are the data, the matrix of **first** recaptures of individuals released in year i; and recaptured in year j.

   Recaptures can be encounters of any type including recapture, re-sighting, or otherwise detecting the presence and status of marked individuals.

3. Parameters

   Φ_j – Conditional probability of apparent survival in interval j, given alive at the beginning of the interval j

   p_j – Conditional probability of capture or recapture at time j, given alive at the beginning of the interval j

   K – the number of estimable parameters in the model.

4. Capture histories and m-arrays

The encounter data are summarized in a capture-history or encounter-history matrix. Using 1 to represent and encounter and 0 not encountered on occasion j, typical encounter histories for 5 occasions might look like:

{11111} - released on occasion 1 and encountered on every sampling occasion

{10101} - released on occasion 1 and encountered on occasions 3 and 5.

{10000} - released on occasion 1 and never seen again

{00101} - first released on occasion 3 and seen again on the 5th occasion.

The formulation of the m-array is slightly different since the R_i include the number of re-releases and the m_{ij} include only the first capture of each individual. Thus each individual appears only once in each row of the m-array. For example consider the following m-array for a simulated study with 2000 **new** released on each occasion:

i	R_i	1	2	3	4
1	2000	30	70	114	43
2	2030		80	97	55
3	2150			167	46
4	2378				72

5. Expectations

The cell expectations are likewise different because they must incorporate the probability of survival with no recapture. Take for example the capture history:

{10101},

representing an individual released at time 1 and encountered again at time 3 and time 5, the relationship between this encounter history and the estimable parameters can be represented by:

$$\begin{array}{ccccccccc} 1 & \rightarrow & 0 & \rightarrow & 1 & \rightarrow & 0 & \rightarrow & 1 \\ & & \Phi_1 & & \Phi_2 & & \Phi_3 & & \Phi_4 \\ & & & p_2 & & p_3 & & p_4 & & p_5 \end{array}$$

The probability of observing this encounter history is thus:

$$\Phi_1(1-p_2)\Phi_2 p_3 \Phi_3(1-p_4)\Phi_4 p_5$$

A more complicated example occurs when an animal is not encountered on the last occasion, because the last term must include the probability that the animal died before the last occasion. For example the probability of :

{11110}

is

$$\Phi_1 p_2 \Phi_2 p_3 \Phi_3 p_4 \Phi_4(1-p_5)+(1-\Phi_3).$$

When used to develop the expected values for the m-array, we see additional differences. For example given the$m$-array:

i	R_i	2	3	4	5
1	R_1	m_{12}	m_{13}	m_{14}	m_{15}
2	R_2		m_{23}	m_{24}	m_{25}
3	R_3			m_{34}	m_{35}
4	R_4				m_{45}

For example m_{23} includes animals released (initially) or re-released on occasion 2 and captured on occasion 3 (i.e., capture histories {1**11**...} and {0**11**...}). Thus,

$$E(m_{23}) = R_2\Phi_2 p_3$$

Likewise, the capture histories {1**101**0} and {0**101**00} contribute to m_{24}. Thus

$$E(m_{24}) = R_2\Phi_2(1-p_3)\Phi_3 p_4.$$

Thus, it should be fairly obvious that capture histories can be used to construct the m-array, but the m-array can not be used to reconstruct the capture history matrix.

6. Model specification

The basic CJS model is model $\{\Phi_t p_t\}$ (similar to $\{S_t r_t\}$), but models may be constructed using PIMs and Design Matrix in MARK similar to the methods used for tag-recovery models. Like band-recovery models, the likelihoods are based on the multinomial distribution and constructed similarly.

GOODNESS OF FIT (GOF)

1. What is goodness of fit (GOF) and why do I care?

There is an underlying assumption that the data fit the saturated or most general model, which can be interpreted to mean that the following more specific assumptions are met:

a. Every marked animal present in the population at time i has the same probability of recapture (p_i)

b. Every marked animal in the population immediately after time i has the same probability of surviving to time i+1.

c. Marks are not lost or missed.

d. All samples are instantaneous, relative to the interval between occasion i and i+1, and each release is made immediately after the sample.

Generally assume that 3 & 4 are met even though we know that some marks are lost over time and we know that sampling and releases are not instantaneous. It is 1 & 2 that we are concerned with in testing GOF.

Example from genetics and general biology labs.

2. Calculating $\hat{c}$ - three approaches

Remember the limited discussion of $\hat{c}$ when we covered QAICc? $\hat{c}$ attempts to estimate the degree of lack of fit (aka extra-binomial variation (EBV) or overdispersion) in the data. GOF testing is a form of contingency table analysis, which addresses the question: Do the frequencies of individuals exhibiting particular encounter histories (m_{ij}) match those expected under general model; given the number released on each occasion? The degree of

overdispersion is estimated by the parameter $\hat{c}$, which can also be thought of as a variance inflation factor.

A naïve approximation of $\hat{c}$ is based on contingency tables is:

$$\hat{c} = \frac{\chi^2}{df}$$

Recall from the early discussion AIC and the quasi-likelihood (QAICc):

$$QAIC_C = \frac{-2\ln(\mathcal{L})}{\hat{c}} + 2K + \frac{2K(K+1)}{n-K-1}$$

Thus, the model deviance, is divided by $\hat{c}$. Thus, if $\hat{c}$ >1, then the contribution to the QAIC value from the model deviance will decline, and the relative penalty for parameters will increase. Thus, as $\hat{c}$ increases, $QAIC_c$ increasingly favors models with fewer parameters.

a. Model deviance divided by model degrees of freedom.

$$\hat{c} = \frac{-2\ln(\mathcal{L}(\theta))}{df}$$

Assumes that the deviance is distributed as χ^2 , But the deviance is not always distributed as χ^2. This estimate is always biased **high**.

b. Use Pearson GOF χ^2 statistic divided by the model df (less subject to bias) but unavailable for all models (sparse data sets and models with covariates). Often used to determine where LOF is occurring.

c. Bootstrap approach – theoretically robust, only an estimate of $\hat{c}$.

Uses parametric Monte-Carlo methods to approximate the expected distribution of either the deviance or deviance $\hat{c}$ (as in a.) based only on the multinomial distribution. The observed value of either the deviance or $\hat{c}$ is then compared to the resulting distribution to determine the probability of a greater value than observed. If this probability is low, then it is likely that overdispersion exists in the data. A new value of $\hat{c}$ is then calculated by either dividing either the original deviance or the original $\hat{c}$ by the mean of the simulated values.

This estimate is biased low, with the bias increasing as the number of occasions increases and the apparent survival rate increases for the Cormack-Jolly-Seber data type.

d. Median $\hat{c}$ - **PREFERRED METHOD**

This approach also is based on simulating data, this time with a range deviance $\hat{c}$ values. Logistic regression is then used to estimate the median $\hat{c}$ of the distribution. The range of $\hat{c}$ values to simulate (lower and upper bounds, and the total number of points based on these bounds) and the number of simulations for each of the specified values are specified by the user. A small set of values over a wide range should be used to generate the resulting deviance $\hat{c}$s. The logistic regression analysis is performed by MARK as a known fate model. Output consists of the estimated value of $\hat{c}$ and a SE derived from the logistic regression analysis.

The median chat is biased high. However, it has a much smaller standard deviation than the chat estimated by RELEASE. Thus median $\hat{c}$ is closer to truth than the RELEASE chat.

e. Program RELEASE GOF

RELEASE – also available as stand-alone program for survival analysis, but only runs certain specialized models for experimental projects.

Generates 3 standard tests

1) TEST 1

Omnibus test for the comparison of groups. NOT used for GOF. Possible to do much more sophisticated tests in MARK.

2) TEST 2 - Directly test assumptions 1 (above).

Assumption – all marked animals in the population have the same chances of being captured at any given time.

Examples of violations: animals of a particular age or size are more (or less) likely to be captured than animals of a different age or size, animals captured at occasion $_i$ are more (or less) likely to be captured later, animals temporarily leave study area, animals always exist in pairs or schools. For estimation of survival in open populations, marked animals have the same probability of recapture. For estimation of population size (abundance), both marked and unmarked animals must have the same probability of capture.

3) TEST 3 – Directly tests assumption 2 (above).

Assumption – among the marked individuals in the population, all animals have the same probability of surviving, regardless of when they were marked.

Examples of violations: individuals caught at early dates are more (or less) prone to mortality during a given period than individuals caught later, individual marked as offspring at (i -1) will be older or larger at occasion (i).

To run RELEASE from within MARK, simply pull down the 'Tests' menu, and select 'Program RELEASE GOF' (only available with 'Recaptures' data type). Results will be output into a Notepad window.

4) Output:

a) Information concerning recent updates to RELEASE, program limits
b) Listing of capture histories, summary tabulation as an m-array
c) TEST 3 and TEST 2 results for each group (respectively)
d) Summary statistics.
e) TEST2 – "of those marked animals not seen at (i +1), but known to be alive at (i +1), does when they were next seen (i +2, i +3...) depend on whether or not they were seen at (i)?".
f) TEST2.C – examines capture heterogeneity usually pooling results in a 2x2 matrix.

sensitive to short-term capture effects, or non-random temporary emigration.

highlights failure of the homogeneity assumption (assumption a), among animals and between occasions.

basic assumption of "equal catchability" of marked animals.

	When seen again?

Seen at i	i+1	i+2	i+3
No	n	n	n
Yes	n	n	n

g) TEST 3 –assumption that all marked animals alive at (i) have the same probability of surviving to (i +1).

h) TEST 3 – "of those individuals seen at occasion (i), how many were ever seen again, and when?".

i) TEST3.SR:

seen before i	seen again	not seen again
Yes	n	n
No	n	n

does the probability that an individual, known to be alive at occasion i, is ever seen again depend on whether it was marked at or before occasion i?

If there is only a single release cohort, then "seen before i ?" becomes "seen before i , excluding initial release?".

TEST3.SR: Presented for TEST 3 in the version of RELEASE bundled with MARK.

TEST3.Sm: "among those animals seen again, does WHEN they were seen depend on whether they were marked on or before occasion (i)?".

	When seen again?		
seen before i	i+1	i+2	i+3
No	n	n	n
Yes	n	n	n

TEST 3 deals with "survival heterogeneity"

Although not the preferred method, the combined χ^2 statistics from TEST 2 and TEST 3 divided by the degrees of freedom are an approximation of $\hat{c}$, the variance inflation factor used in $QAIC_c$.

5) When the general model is rejected

Determine whether the mode is appropriate. Often straight forward:

a) examine the detailed TEST 2 and TEST 3 contingency tables

"Systematic" rejection (or bias) in the individual tables?

Particular cell (or cells) in one of the test tables is consistently over- or under-populated (i.e. below or above the predicted value).

b) Example for CJS (recapture) data

(1) TEST2 (recapture rates) is OK

(2) TEST3 (survival rates) is rejected. Next examine each of the TEST3 tables for each cohort.

(3) TEST3.Sm is accepted.

(4) TEST3.SR is rejected

(5) TEST3.SR – of those individuals seen either on or before occasion i, what proportion were ever seen again?

If rejected, then there is a difference in "survival" among individuals, depending on whether or not they were seen for the first time either on or before occasion i. However, TEST3.Sm only looks at individuals who WERE seen again. Among these individuals, when they were seen again does not depend on whether or not they were seen for the first time at occasion i.

(6) Examine each individual TEST3.SR table
(a "+" indicates more individuals than expected)
(a "-" indicates fewer individuals than expected)

	TEST3.SR	
Seen at *i*	Seen again	Not seen again
Seen before	+	-
Not seen before	-	+

In this example the pattern seems to be present in the majority (e.g., 8/10) of the tables. What does this suggest?

Among individuals seen for the first time at occasion *i*, more are never seen again than expected (i.e., newly marked individuals were consistently less likely than previously marked individuals to ever be seen again).

Interpretation?

If individuals were marked as juveniles then this could reflect lower survival rates for newly marked juveniles.

Marking effect

Presence of transients (migratory individuals)

Heterogeneity in capture rates

6) What can you do if you do reject CJS?

If the individual TEST2 or TEST3 results seem to show systematic deviations among occasions the CJS assumptions are not met and a different starting model is indicated (e.g., the age-structured model).

Use analogous GOF tests for the new starting model.

RELEASE only tests the CJS model

RELEASE can be used for other models, under some conditions (see MARKBOOK or AFS monograph)

f. Extra-binomial variation

If the deviations in the contingency tables are not "consistent" among batches

No clear "explanation" (biological or otherwise) for the violation of TEST2 or TEST3.

Remember that the conceptual basis of all models is "data = structure + residual variation".

Suggests that there is more residual variation than expected

Excess variation shows up in the model GOF.

$$\hat{c} = \frac{-2\ln(L(\theta))}{df} > 1$$

Remember though that in general, as $\hat{c}$ increases your model selection is more conservative, and favor reduced parameter models.

g. When is $\hat{c}$ too big?

If the model fits perfectly, then $\hat{c} \leq 1$. What if $\hat{c}$ =2, or $\hat{c}$ =10?

"Rule of thumb", $\hat{c} \leq 3.0$ is ok (see Lebreton et al. 1992 - pp. 84-85).

Considerations

1) Ok to adjust for lack of fit simply by using $\hat{c}$.
2) Large $\hat{c}$ (>2) warrants careful examination of the model structure.
   If the problem is due to structure problems in the general model:
   a) Examine TEST 2 and TEST 3 results.
   b) Consider a different general model.
3) If not a structural problem
   a) Ok to adjust for lack of fit simply by using $\hat{c}$.
   b) Make the parameter estimates as robust and valid as possible.
   c) Blindly doing so may obscure important insights concerning your data.
   d) If model structure is correct and $\hat{c}$ is >>3, data may not be adequate for analysis.