

Lecture 2: Principles of model fitting and model selection

Basics of Parameter Index Matrices

Consider an example with five occasions and four transitions for a time-dependent model (ϕ_t, p_t).

t_1	ϕ_1	t_2	ϕ_2	t_3	ϕ_3	t_4	ϕ_4	t_5
		p_2		p_3		p_4		p_5

In reality, you are not just going to follow one cohort of animals through time, you are going to capture and mark animals on each occasion. In that case, the above table should be expanded as:

	t_1		t_2		t_3		t_4		t_5
1	ϕ_1		ϕ_2		ϕ_3		ϕ_4		
		p_2		p_3		p_4		p_5	
2			ϕ_2		ϕ_3		ϕ_4		
				p_3		p_4		p_5	
3					ϕ_3		ϕ_4		
						p_4		p_5	
4							ϕ_4		
								p_5	

The subscripts on these parameters can be simplified to a triangular *Parameter Index Matrix*:

There are four estimates of local survival ($\phi_1 - \phi_4$), hence:

1	2	3	4
	2	3	4
		3	4
			4

There are also four estimates of recapture rate ($p_2 - p_5$) but to avoid confusion, the recapture rates are sequentially renumbered to indicate that they are different:

5	6	7	8
	6	7	8
		7	8
			8

Subscripts and model notation

Models that are fit to mark-recapture data are denoted with subscripts for the effects within the parameters of interest. The model above is known as the Cormack-Jolly-Seber model (CJS model) and is a basic model that has been widely used. This model has time-dependence in both local survival and recapture rates, and is usually described as $\phi_{\text{time}}, p_{\text{time}}$, or ϕ_t, p_t . If sexes are modeled separately, this could be presented as: $\phi_{\text{sex}*t}, p_{\text{sex}*t}$. The range of available models was limited in early mark-recapture programs. For example, Program JOLLY had four canned models which it would calculate:

Model A: ϕ_t, p_t ; Model B: ϕ_c, p_t ; Model C: ϕ_t, p_c ; Model D: ϕ_c, p_c

Recent software like SURGE and MARK greatly extend these limited options.

Other commonly used subscripts include: 'g' = group, 's' = sex, 't' = time, '2ac' = two age-classes, 'c' or '.' = constant. The notation for mark-recapture models is *not* standardized, so it is always worth defining the terms that you have used. Models that include interactions between main effects are indicated by a "*" (e.g., sex*t, 2ac*t), main effects models with a '+' (e.g., sex+t).

Time-dependent models

This is a time-dependent model, characterized by columns of numbers that are different. If we collapse one rate to a constant, the subscripts are the same, it indicates that all values for that parameter are the same.

Thus a model of (ϕ_t, p_c) would look like:

Local survival				Recapture rate			
1	2	3	4	5	5	5	5
	2	3	4		5	5	5
		3	4			5	5
			4				5

Group models

Often we want to be able to estimate and compare survivorship of animals in different groups. Common categorical analyses include sex (female, male), different sites (poor, good), and size class (small, medium, large). In this case, we create a Parameter Index Matrix for each group

$\phi_{\text{sex}*t}, p_{\text{sex}*t}$

Local survival of females				Local survival of males			
1	2	3	4	5	6	7	8
	2	3	4		6	7	8
		3	4			7	8
			4				8
Capture of females				Capture of males			
9	10	11	12	13	14	15	16
	10	11	12		14	15	16
		11	12			15	16
			12				16

Perhaps we expect that capture rates vary with time but are the same in both sexes. We can build a model with no sex differences in p , denoted:

$\phi_{\text{sex}*t}, p_t$

Local survival of females				Local survival of males			
1	2	3	4	5	6	7	8
	2	3	4		6	7	8
		3	4			7	8
			4				8
Capture of females				Capture of males			
9	10	11	12	9	10	11	12
	10	11	12		10	11	12
		11	12			11	12
			12				12

Perhaps we expect the opposite pattern, that capture rates differ between the sexes but are not changing across time. We can build a model with no time differences in p , denoted:

$\phi_{\text{sex}^*t}, p_{\text{sex}}$							
Local survival of females				Local survival of males			
1	2	3	4	5	6	7	8
	2	3	4		6	7	8
		3	4			7	8
			4				8
Capture of females				Capture of males			
9	9	9		10	10	10	10
	9	9			10	10	10
		9				10	10
			9				10

Note that you are retaining a block of PIMs for each group and that pooling across groups would entail sharing of parameter reference numbers. Also note that you can model effects on apparent survival separately from effects on the encounter rate

Age and Time-Since-Marking (TSM) models

In many cases you may expect age-structure in survival rates. Age models are distinguished by the difference between the diagonal and off diagonal elements. For example, in the model ϕ_{2ac}, p_t , the four parameters labelled 1 along the diagonal separate the transition immediately after first capture from subsequent capture intervals. The 6 parameters labelled 2 would measure apparent survival among individuals that were captured at least once.

ϕ_{2ac}, p_t	ϕ_{3ac}, p_t	ϕ_{2ac^*t}, p_t	ϕ_{3ac^*t}, p_t
Two age only	Three age only	Two age and time	Three age and time
Local survival			
1 2 2 2	1 2 3 3	1 5 6 7	1 5 8 9
1 2 2	1 2 3	2 6 7	2 6 9
1 2	1 2	3 7	3 7
1	1	4	4
Recapture			
3 4 5 6	4 5 6 7	8 9 10 11	10 11 12 13
4 5 6	5 6 7	9 10 11	11 12 13
5 6	6 7	10 11	12 13
6	7	11	13

Age models

Animals are aged in two main ways: i) newly marked each year as young and ii) by independent characters. The latter method is often used to block animals into age-classes (e.g. birds: first-year or adult, in thrushes by buff-coloured feathers on wing, in grouse by pigmentation and shape of leading primaries, mammals: virgin or adult, in microtine rodents by vaginal perforation).

Why might you expect age-specific variation in apparent survival? First, it could be due to age effects on true survival.

1. Selection: Poor quality individuals die young.
2. A constraint: juveniles are inexperienced at foraging and avoiding predators.

3. A restraint: competition for breeding opportunities is high and leads to delayed maturity.
4. Senescence: declining performance in older individuals, usually only seen in long-lived animals.

Second, it could be due to age effects on site fidelity.

5. Natal dispersal: Young animals often disperse from their natal site and it is not possible to distinguish losses due to permanent emigration from study plots from mortality.

Time since marking models

The same models can also be applied to a sample of animals where age is unknown or all individuals were initially marked as adults. ‘Time since marking’ models effectively control for animals with capture histories where they were seen once and never again. Apparent survival in the transition after first marking can be lower for several nonexclusive reasons:

Again, it could be due to effects on true survival:

1. Relative age: newly marked animals may be indistinguishable from previously marked animals but could have lower true survival or site-fidelity rates because their relative age is younger than returning individuals
2. Handling effects: the action of capturing and marking an animal induces a higher rate of mortality or a behavioural switch where the animal permanently disperses from your study plot.

Or it could be due to effects on site fidelity:

3. Transients: some individuals stop at the study site but then move on. This is often the case with systematic mistnet studies that are capturing migratory birds at a stopover site.
4. Heterogeneity of capture: some animals are encountered less frequently than others. For example, perhaps you are studying a territorial animal with territories that are large relative to the size of your study area (e.g., red squirrels, snow geese feeding areas). Animals with territories that border are your study area are more likely to be encountered only once than animals with territories on the study area. This may lead to permanent emigration.

In practice, these four alternatives are difficult to distinguish without corollary information.

Cohort models

Cohort models define individuals on the basis of when they were first captured.

$\phi_{\text{cohort}, p_t}$	$\phi_{\text{cohort}^*t, p_t}$	$\phi_{\text{age}^*t, p_t}$
Local survival		
1 1 1 1	1 2 3 4	1 5 8 10
2 2 2	5 6 7	2 6 9
3 3	8 9	3 7
4	10	4
Recapture		
5 6 7 8	11 12 13 14	11 12 13 14
6 7 8	12 13 14	12 13 14
7 8	13 14	13 14
8	14	14

In practice, cohort models are rarely used because their assumptions are not usually met. Also, the parameterization of a saturated cohort model is equivalent to that of an age model.

Summary of Parameter Index Matrices for five types of models:

1. Constant model: the PIM contains one parameter for all values.
2. Group models: the structure of the PIM may be the same but the parameter values are different among *multiple PIMS*
3. Time models: the PIM is blocked into different *columns*
4. Age and Time since marking models: the PIM is blocked into different *diagonals*
5. Cohort models: the PIM is blocked into different *rows*

Explicit probability statements and the log-likelihood expression

Consider the following simple example with only 3 parameters.

t_1	t_2	t_3
ϕ_1	ϕ_2	
	p_2	p_3

There are three parameters in the model that we want to estimate: ϕ_1 , p_2 and β_3 . We cannot estimate p_3 without capture information beyond the last occasion t_3 so the last two parameters cannot be separated and their product forms an inestimable beta term ($\phi_2 * p_3 = \beta_3$). For individuals marked on the first occasion, it is possible to write an explicit probability statement for each possible capture history:

Capture History	Probability	Observed frequency
111	$\phi_1 * p_2 * \beta_3$	52
110	$\phi_1 * p_2 * (1 - \beta_3)$	20
101	$\phi_1 * (1 - p_2) * \beta_3$	6
100	$1 - \phi_1 * p_2 - \phi_1 * (1 - p_2) * \beta_3$	22

The likelihood expression can then be written as:

$$\mathcal{L} = K * (\phi_1 * p_2 * \beta_3)^{52} * (\phi_1 * p_2 * (1 - \beta_3))^{20} * (\phi_1 * (1 - p_2) * \beta_3)^6 * (1 - \phi_1 * p_2 - \phi_1 * (1 - p_2) * \beta_3)^{22}$$

Taking the natural log of this expression puts it into a much better format: that of a general linear model. The log-likelihood is the joint probability of all of these outcomes:

$$\ln \mathcal{L} = \ln(K) + 52 \ln(\phi_1 * p_2 * \beta_3) + 20 \ln(\phi_1 * p_2 * (1 - \beta_3)) + 6 \ln(\phi_1 * (1 - p_2) * \beta_3) + 22 \ln(1 - \phi_1 * p_2 - \phi_1 * (1 - p_2) * \beta_3)$$

where K = a multinomial constant.

Mark-recapture software such as Program Mark solves likelihood expressions like this by iteratively fitting values to the multinomial equation. The ‘maximum-likelihood’ is the set of values for the parameters that maximize the value of the $\ln \mathcal{L}$. This obviously gets quite complex with multiple groups and long-time series.

Model fitting with maximum likelihood estimation yields a number of metrics that are the measure of the model fit.

1. The *deviance* ($\text{Dev} = -2\ln\mathcal{L}$): The ‘deviance’ or $\text{Dev} = -2\ln\mathcal{L}$ can be thought of as the unexplained variation in the model or as an index of model fit (low=good fit, high=bad fit). The more parameters in the model, usually the lower the deviance value that can be obtained.
2. The *number of parameters* (np or K): the number of separate identifiable terms in the model (determined by the number of intervals, groups and age-classes). These are the number of individual survival and resighting rates in a particular model (e.g. 3 for the above model).
3. The model also yields *point estimates for each of the parameters and their variance*.

For the time-dependent model above, Program MARK would generate the following values:

1. A deviance ($-2\ln\mathcal{L}$) = 2,327.7
2. The number of parameters (ϕ_1, p_2, β_3) = 3
3. Point estimates of each of the parameters, $\phi_1=0.803 \pm 0.044\text{SE}$, $p_2=0.897 \pm 0.040$, $\beta_3=0.721 \pm 0.031$

Information theory and model selection

The principle of parsimony. An idea that is commonly used in statistics, that is boiling relationships down to their essential elements. The object is to capture the few independent variables that describe the largest amount of variation in the dependent variable. This is essentially the aim of stepwise multiple regression, PCA and other multivariate techniques. One useful analogy is predicting weather patterns. The more parameters we include in a model, the less biased our model will be and mean predictions will be closer to truth. However, adding parameters is expected to increase the variance in our predictions because each parameter is estimated with some uncertainty. In practical terms, the costs of collecting additional information are also expected to increase with an increasing number of parameters. In mark-recapture methods, AIC values are used to identify parsimonious models.

Akaike’s Information Criterion (AIC). In model testing, you want to minimize both the NP and the DEV, and indices that combine these terms are used in several branches of statistics (e.g., Mallows’s C_p in multiple regression). Here, we combine the deviance and the number of parameters to calculate Akaike’s Information Criterion (AIC):

$$\text{AIC} = \text{Dev} + 2K$$

where $\text{Dev} = -2\ln\mathcal{L}$ and K = the number of parameters estimated. Why $2 \times K$? Simulation models with dummy data sets have established that 2 is the best value. The actual value of AIC is meaningless, in one set of analyses it could be 630-680, another 2150-2275. What is important are the absolute differences between models. Program MARK uses a modified version of this formula which includes a correction term for small sample size. This improved version of AIC is termed corrected Akaike’s Information Criterion (AICc) and is given by:

$$\text{AICc} = \text{Dev} + 2K + \frac{2K(K+1)}{(N-K-1)}$$

where N = the effective or finite sample size (determined from the capture histories). Program MARK will also allow for corrections due to overdispersion if a variance inflation factor or \hat{c} can

be estimated (a topic for a future class). Quasi-Akaike's Information Criterion is calculated as:

$$\text{QAICc} = \frac{\text{Dev}}{\hat{c}} + 2K + \frac{2K(K+1)}{(N-K-1)}$$

Model Ranking

In a set of candidate models, the model with all of the terms present is the 'starting' or 'global' model. Further models are calculated by sequentially dropping terms from that model. The actual value of AIC has no meaning, what is relevant is the difference in AIC between the models (Δ_i). The 'best fit model' is the model with the lowest AIC value (minAIC) which is then set to a value of zero. All other models are then ranked vs. this model by the difference in AIC values:

$$\Delta_i = \text{AIC}_i - \text{min AIC}$$

Two models that differ in AIC values by less than a value of two ($\Delta_i \leq 2$) are usually considered to be equally parsimonious.

Another useful metric is the Akaike weights (w_i) which are automatically calculated by Program MARK. Akaike weights are calculated as:

$$w_i = \frac{\exp\left(-\frac{1}{2}\Delta_i\right)}{\sum_{r=1}^R \exp\left(-\frac{1}{2}\Delta_i\right)}$$

and sum to one across all of the candidate models that are fit in a particular data set.

The ratio of Akaike weights is an indication of the relative support between two models. For example, if the w_i for ϕ_t, p_t and ϕ_t, p_c is 0.507 and 0.343, respectively, then ϕ_t, p_t has $0.507/0.343 = 1.5$ times the support of model ϕ_t, p_c .

Frequentist methods

Hypothesis testing with Likelihood ratio tests (LRT) are an early method that was used for model selection but use of this approach is now discouraged. They appear in the older literature, including the Lebreton et al. (1992) monograph. LRT allow you to exact an exact probability-value to a comparison of two models. A limitation of LRT tests is that they allow comparison of models that are *nested*, models that only differ by the presence/absence of a factor.

Example analysis of Dipper data set from Lebreton et al. (1992)

$\chi^2_4 = 2.78$ $P = 0.595$	$\phi_t p_t$ DEV = 656.95 np = 11 AIC = 678.95	$\chi^2_4 = 7.53$ $P = 0.110$
$\phi_t p_c$ DEV = 659.73 np = 7 AIC = 673.73	$\chi^2_9 = 9.89$ $P = 0.359$	$\phi_c p_t$ DEV = 664.48 np = 7 AIC = 678.48
$\chi^2_5 = 7.11$ $P = 0.212$	$\phi_c p_c$ DEV = 666.84 np = 2 AIC = 670.84	$\chi^2_5 = 2.36$ $P = 0.797$

Calculation of the likelihood ratio test: Models can be compared with LRT because the distribution of the deviance approximates a χ^2 -distribution. The difference between the two nested models is taken as the χ^2 -value with the difference in number of parameters as the degrees of freedom:

$$\chi^2\text{-value} = | (\text{Deviance of model 1}) - (\text{Deviance of model 2}) |$$

$$\text{df} = (\text{np of model 1}) - (\text{np of model 2})$$

MARK will do the calculations among all pairs of models in a model set.

Problems with hypothesis testing and LRT

LRT can only be applied to nested models.

The theoretical foundation is weak for LRT but is good for AIC.

A α -level of 0.05 is arbitrary, and is unreasonable to apply across a range of N. Better to calculate effect sizes?

In fitting models to CMR data, one is fitting multiple models to the same set of data so hypothesis testing is inappropriate.

Hypothesis testing seeks to identify one single model that fits the data but a group of models may be equally good fit.

Final thoughts

Two final points to keep in mind. When modeling mark-recapture or other data the ‘true’ model is unknown. Statistical models are only approximations of reality. Box (1976) suggested: ‘all models are wrong, but some are useful’. The tools of model selection are used to identify the best model that explains the most variation with the fewest number of parameters. Note that as the sample size of data increases, the complexity of the models that can be supported also increases.