

Lecture 3: Model selection, advanced parameter index matrices and design matrices

Steps in model selection

- 1). Select the variables and a set of candidate models that you want to test *a priori*. Males and females often have different survival rates so 'sex' may be included as an explanatory group variable. Using the results of analyses to guide model selection is termed 'data-dredging'. While it may have a place in exploratory analyses, this approach runs the risk of leading to over-fitted models that detect spurious effects.
- 2). Start with global model that includes all of the terms that you are interested in. Carefully define the model subscript notation. Common subscripts include: s = sex, t = time, g = grp, 2ac = two age-classes, * = main effects and interaction, + = main effects only (additive model).
- 3). A necessary first step is to see whether the global model is a reasonable fit to the data and a variety of goodness-of-fit (GOF) tests are available to do this. This step usually involves calculation of a variation inflation factor (\hat{c} or c-hat) that controls for overdispersion in count data. We will defer this topic to a later lecture.
- 4). Before fitting other models, it is often helpful to examine the unconstrained point estimates generated by the starting global model before proceeding with formal model fitting. If apparent survival is fluctuating wildly and the point estimates have small SE, a constant model is unlikely to improve model fit. You can often reject unproductive paths of model testing if you do this.
- 5). The next step is to fit simpler models with fewer parameters. In Program Mark, models can be constructed with either the parameter index matrices or design matrices. For simple problems, it is possible to calculate all possible models. Consider the Dipper example where we have two sexes and time. For local survival and recapture rates we have at least four possible standard parameterizations (sex*t, t, sex, constant). For each of two rates this gives $4^2=16$ possible combinations. This would be more if you used different structures in the different sexes. It rapidly becomes impossible to calculate all possible models. Using the tools of model selection allows you to eliminate nonproductive pathways if you treat the problem as a descending tree and eliminate branches.
- 6). In more complex problems, the total number of possible models increases rapidly with number of levels within a variable. Drop terms from the global model to create simpler, nested models. Small steps are best. If the AICc value is lower, accept the reduced model. If the AICc value is much greater, retain the model with more terms. Visualize the procedure as pouring gravel through a set of sieves to capture the best possible model.
- 7). Other modeling tactics. One challenge in model selection is that the model fit for ϕ affects estimation of p and vice versa. It is often usually desirable to start with the nuisance parameter like resighting rate to get the best fit there before modeling survival. Some schools of thought suggest leaving the nuisance parameter alone and instead use the full set of unconstrained values for p . If the difference in AICc values is small, a good rule of thumb is to retain the model with more terms as you proceed with model fitting for other parameters, and then retest once you have a better model fit for the other rate. Alternatively, you can proceed with a reduced model and add terms back in once you have a good fit for a reduced model.
- 8) Model fitting is an art not a science.

Selecting link functions

One problem with building linear models for mark-recapture data is that the response variable (e.g., survival or resighting rates) is bounded between 0 and 1. To cope with this problem, MARK offers a variety of ‘link functions’ that transform the data so that they are bounded from negative to positive infinity, which is the assumption for regular linear models.

The two most useful link functions are the *logit* and *sine* links.

The logit (or logistic) link function is defined as:

$$\text{logit}(p) = \ln(p/(1-p))$$

where p is any probability. To back calculate p , the following transformation is used by MARK:

$$p = (1 + e^{-\text{logit}(p)})^{-1}$$

The logit link function allows the probability of an event to be expressed as a linear function of explanatory variables. It is used in MARK to calculate estimates of the intercept (β_0) and slopes (β_1, β_2 etc.), the slope coefficients are then used to calculate estimates of ϕ and p .

Which link function to use? The logit link is most flexible and can be used with both identity and design matrices. The sine link function has better properties if parameters are close to the boundaries of 0 and 1 but cannot be used with design matrices. Both the logit and sine link functions yield parameter estimates bounded from 0-1, some of the other link functions in MARK are not subject to this constraint. The different link functions may yield different numbers of parameters for the same model, depending whether they have trouble estimating rates near the boundaries of 0 and 1. In practice, it is best to use the same link function for all models in a candidate set. If the set of candidate models includes additive models or models with covariates, then the logit link is probably the best link function to use.

Testing user-defined models, additive models and environmental covariates

Last week we learned how to use Parameter Index Matrices (PIMs) to implement basic model structures of MARK:

Group effects: multiple blocks of PIMs

Time-dependence: PIMs with columns

Age-dependence: PIMs with diagonals

Cohort effects: PIMs with rows

Constant models: PIMs with only one value

The basic menu options of MARK offer considerable flexibility in building models, but you may want to extend these to test questions of greater biological relevance. In this class, we will focus on three ways of testing more advanced models.

1. User-defined models using PIMs

1. Pooling subsets of variables in different groups. A common problem is to have two groups, one banded as young, one banded as adults. Perhaps you want to model survival after first capture separately for these two groups but pool them in later transitions. And you want to examine annual variation in both cases. The PIMs of apparent survival of the two groups would be:

adults		young
1 5 6 7		8 5 6 7
2 6 7		9 6 7
3 7		10 7
4		11

This model could be denoted as: $\phi_{age*time}^1, \phi_{time}^2, p_{constant}$ if the PIM for p was a block of 12s.

2. Annual covariates. Perhaps you wanted to model survival as a function of flood conditions. This is a question that is explored in the Dipper example in the Lebreton et al. monograph. In this case the PIMs might be coded as 1 = non flood year vs. 2 = flood year (in occasions 2, 3 and 6).

1 2 2 1 1 2 1
2 2 1 1 2 1
2 1 1 2 1
1 1 2 1
1 2 1
2 1
1

In this case, you need to consider carefully how covariates an environmental conditions line up. The first column of twos will apply to the transition between t_2 and t_3 and so forth.

2. The additive model

In the modeling we have done so far, we have skipped one important test. The problem with PIMs is that they can only be used to construct models that are fully factorial (*) or only one of the main effects. Consider a test for an effect of time where we compare model ϕ_{sex*t, p_c} with model ϕ_t, p_c . Or more formally:

$$\phi = \beta_0 + \beta_1SEX + \beta_2TIME + \beta_3SEX*TIME$$

$$\phi = \beta_0 + \beta_1SEX$$

$$\beta_2TIME + \beta_3SEX*TIME$$

The difference between these models is *two* terms. Thus, we have taken two steps where it would be preferable to take one smaller step. Effectively, we have not investigated the significance of the interaction term.

$$\phi = \beta_0 + \beta_1SEX + \beta_2TIME + \beta_3SEX*TIME$$

$$\phi = \beta_0 + \beta_1SEX + \beta_2TIME$$

$$\beta_3SEX*TIME$$

To understand what testing whether an interaction terms is significant, consider the following 2-way ANCOVA (sex is categorical, time is continuous).

model mass = sex time sex*time

sex	ns	sig	ns	sig	ns	sig
time	ns	ns	sig	sig	ns	ns
interaction	ns	ns	ns	ns	sig	sig

Before starting a mark-recapture analysis, it always useful to look at parameter values in your starting global model (e.g., ϕ_{sex*t} , p_{sex*t}). If there is a constant difference, an additive model might be the best fit. Additive models are tested in MARK by using dummy variables.

3. Models with annual covariates

Perhaps we want to model survival or recapture as a function of annual conditions. Why is it not acceptable to generate point estimates and plunk them in to a post hoc regression? Two reasons from Lebreton et al. 1992: pp 76-77:

1. Scaling/range problems. A change in covariate x is not proportional to an equivalent change in local survival ϕ . This problem arises because local survival is bounded 0-1. Mark-recapture analyses cope with this problem by using the logit transformation as a link function. By using the logit transformation, $\text{logit}(\text{local survival})$ is not bounded and can range from $-\infty$ to $+\infty$.

2. Autocorrelation among survival estimates. Survival estimates are not independent because different samples of animals span each survival transition with overlap among them. Local survival will covary with covariate x if x varies systematically. If you want to calculate the percentage of variation explained, and this can be done for a series of nested models:

- 1) $\phi_{\text{time}, p}$
 - a) χ^2
- 2) $\phi_{\text{covariate } x, p}$
 - b) χ^2
 - c) χ^2
- 3) $\phi_{\text{constant}, p}$

The percentage of variation explained by covariate $x = [\text{b } \chi^2] / [\text{c } \chi^2] \approx r^2\text{-value}$.

Covariates can be modeled in MARK with design matrices based on dummy variables.

4. Design matrices based on dummy variables

General linear models are based on dummy variable coding, which you are usually shielded from by the interface of most statistical programs. Basically, the model structure is rewritten in a binary form using 0 and 1 to code levels of each variables. You may have been exposed to this if you have used multiple regression. Regression assumes that the independent variables are continuous, but you can include categorical variables if you recode them as dummy variables. For mark-recapture analyses, look at the following examples.

a) The additive model

Consider the model: ϕ_{sex*t} , p_c . You could fit this model in MARK two ways: using the PIMs and

an identity matrix or by using dummy variables in a design matrix.

The PIMs for this model would be as follows:

Local survival females	Local survival males	Recapture females	Recapture males
1 2 3 4	5 6 7 8	9 9 9 9	9 9 9 9
2 3 4	6 7 8	9 9 9	9 9 9
3 4	7 8	9 9	9 9
4	8	9	9

If you ran $\phi_{\text{sex}^*t, p_c}$ by using the PIM menu options, Program Mark would prompt you with a question about using an ‘identity matrix’, to which you would say ‘yes’.

An identity matrix is basically a block with a diagonal of ones:

```
1 0 0 0 0 0
0 1 0 0 0 0
0 0 1 0 0 0
0 0 0 1 0 0
0 0 0 0 1 0
0 0 0 0 0 1
0 0 0 0 0 1
```

The same model for phi can also be coded using dummy variables, where you would need n-1 columns where n = levels of variable x

	Accurate but redundant	Better
Two sexes	1 0 0 1	1 0
Four intervals	1 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0	1 0 0 0 1 0 0 0 1 0 0 0

The full model for ϕ_{sex^*t} is below.

Parameter (ϕ)	Intercept	Sex effect	Time effect	Interaction
1	1	1	1 0 0	1 0 0
2	1	1	0 1 0	0 1 0
3	1	1	0 0 1	0 0 1
4	1	1	0 0 0	0 0 0
5	1	0	1 0 0	0 0 0
6	1	0	0 1 0	0 0 0
7	1	0	0 0 1	0 0 0
8	1	0	0 0 0	0 0 0

If you ran the same model with both PIMs and also the design matrix, you should get the same DEV, np, and AICc both times if you use the same link function (logit) in both cases. Why are you bothering with all this?

To be able to fit the additive model $\phi_{\text{sex}+t, p_c}$ and models with external constraints. These two types of models cannot be fit from the PIMs and can only be developed using design matrices. The dummy variable structure for the additive model is:

Intercept	Sex effect	Time effect
1	1	1 0 0
1	1	0 1 0
1	1	0 0 1
1	1	0 0 0
1	0	1 0 0
1	0	0 1 0
1	0	0 0 1
1	0	0 0 0

What are the number of parameters in this model? Four time periods plus one constant difference plus one parameter in recapture rate is 6 parameters. Additive models generally have one more parameter than the next more reduced model.

b) Linear covariates

We can also use dummy variable structures to examine linear change in phi or p. Perhaps we want to test whether survival has declined over the length of the study period. In the coding above time is essentially coded as class variable and each interval is treated differently. Here, time is coded as a continuous variable and the intervals are numbered in sequence.

Full model for $\phi_{\text{sex}*\text{linear}}$

Intercept	Sex effect	Time effect	Interaction
1	1	1	1
1	1	2	2
1	1	3	3
1	1	4	4
1	0	1	0
1	0	2	0
1	0	3	0
1	0	4	0

Additive model for $\phi_{\text{sex}+\text{linear}}$

Intercept	Sex effect	Time effect
1	1	1
1	1	2
1	1	3
1	1	4
1	0	1
1	0	2
1	0	3
1	0	4

What are the number of parameters in these models? It takes two parameters to describe a linear regression: a slope and an intercept. Therefore, for $\phi_{\text{sex}*\text{linear}}, p_c, np$ would be 4 for phi and 1 for $p = 5$ total. For the additive linear model $\phi_{\text{sex}+\text{linear}}, p_c, np$ would be 3 for phi and 1 for $p = 4$ total.

3. Annual covariates: environmental factors, sampling effort

We can use the same approach to model local survival and recapture rates as a function of annual covariates (rainfall, rates of nest predation, effort in hours afield). For example, maybe we work at a site that is influenced by flooding and variation in annual rainfall:

Year	Flooding	Rainfall
1996	+	201
1997	-	75
1998	-	52
1999	+	155
2000	-	178

The corresponding dummy variable constraints for flooding and rainfall would be:

Full model for $\phi_{\text{sex}*flood}$

Intercept	Sex effect	Flood effect	Interaction
1	1	1	1
1	1	0	0
1	1	0	0
1	1	1	1
1	0	1	0
1	0	0	0
1	0	0	0
1	0	1	0

Full model for $\phi_{\text{sex}*rain}$

Intercept	Sex effect	Rain effect	Interaction
1	1	201	201
1	1	75	75
1	1	52	52
1	1	155	155
1	0	201	0
1	0	75	0
1	0	52	0
1	0	155	0

The information for 2000 is not used because we are looking at the subsequent effects of annual conditions on survival in the following transitions.

5. What are the dummy variables doing?

Remember, MARK is fitting models to the likelihood equation subject to the logit transformation:

$$\phi = \beta_0 + \beta_1 \text{SEX} + \beta_2 \text{TIME} + \beta_3 \text{SEX} * \text{TIME}$$

If you fit a constraint, MARK will generate estimates of the intercept and slopes $\beta_0 - \beta_3$:

INTERCEPT	0.577	(β_0)
SLOPE N1	-0.184	(β_1)
SLOPE N2	-0.763	(β_2)
SLOPE N3	0.259	(β_3)

Thus, to obtain a value of phi for females at, e.g., time 1-2, MARK substitutes in the dummy variables for this parameter, in this case a structure of “1 0 0”:

$$\text{logit}(\phi) = (0.577)(1) + (-0.184)(1) + (-0.763)(0) + (0.259)(0)$$

$$\text{logit}(\phi) = 0.393$$

$$\text{The back transformation is then: } \hat{\phi}_{1-2} = (1 + e^{-0.393})^{-1} = e^{0.393} / (1 + e^{0.393}) = 0.597$$

6. Parameters reflect model structure

A useful feature of MARK is that it generates estimates of local survival and recapture rates under the constraints set by the model structure. If you fit a model with constant apparent survival and constant encounter rate, only two parameters will be estimated. Inspection of these reconstituted parameters in the MARK output can be a useful test of whether the constraint is doing what you think it should be doing. For example, if you are fitting an additive model, the parameter estimates should display some degree of parallelism. Note that the reconstituted parameters will not be exactly parallel when they are plotted because they have been back transformed from being exactly parallel on the logit scale.