Models Where the Fate of Every Individual is Known

This class of models is important because they provide a theory for estimation of survival probability and other parameters from radio-tagged animals. The focus of known fate models is the estimation of survival probability *S*, the probability of surviving an interval between sampling occasions. These are models where it can be assumed that the sampling probabilities are 1. That is, the status (dead or alive) of all tagged animal is known at each sampling occasion. For this reason, precision is typically quite high, even in cases where sample size is often fairly small. The only disadvantages might be the cost of radios and possible effects of the radio on the animal or its behavior. The model is a product of simple binomial likelihoods. Data on egg morality in nests and studies of sessile organisms, such as mollusks, have also been modeled as known fate data.

The Kaplan-Meier Method

The Kaplan-Meier (1958) method has been used commonly in the past and we will mention it as a starting point. This estimator is based on observed data at a series of occasions, where animals are marked and released only at occasion 1. The K-M estimator of the survival function is

$$\hat{S}_t = \prod_{i=1}^{t} \left(\frac{n_i - d_i}{n_i} \right),$$

where n_i is the number of animals alive and at risk at occasion i, d_i is the number known dead at occasion i, and the product is over i up to the t^{th} occasion. Critical here is that n_i is the number known alive at occasion i minus those individuals known dead or censored during the interval. It is rare that a survival study will observe the occasion of death of every individual in the study. Animals are "lost" (i.e., censored) due to radio failure or other reasons. The treatment of such censored animals is often important, but often somewhat subjective. These K-M estimates produce a survival function (see White and Garrott 1990); the cumulative survival up to time t. This is a step function and is useful in comparing, for example, the survival functions for males vs. females.

If there are no animals that are censored, then the survival function (empirical survival function or ESF) is merely,

$$\hat{S}_t = \frac{\text{Number longer than } t}{n} \qquad \text{for } t \geq 0.$$

This is the same as the intuitive estimator where not censoring is occurring;

$$\hat{S}_t = n_{t+1}/n_t \qquad \text{(e.g., } \hat{S}_2 = n_3/n_2\text{)}.$$

The K-M method is an estimate of this survival function in the presence of censoring. Expressions for the variance of these estimates can be found in White and Garrott (1990).

A simple example of this method can be illustrated using the data from Conroy et al. (1989) on 48 radio-tagged black ducks. The data are

	Survived to Occasion							
Week	1	2	3	4	5	6	7	8
Number alive at start	48	47	45	39	34	28	25	24
Number dying	1	2	2	5	4	3	1	0
Number alive at end	47	45	39	34	28	25	24	24
Number censored	0	0	4	0	2	0	0	0

Here, the number alive at the start of an interval are *known* to be alive at the start of sampling occasion *j*). This is equivalent to being alive at the start of interval *j*. For example, 47 animals are known to be alive at the beginning of occasion 2. A further example is that 34 ducks survived to the start of occasion 5. Thus, the MLEs are

$\hat{S}_1 = 47/48 = 0.979$

$\hat{S}_2 = 45/47 = 0.957$

$\hat{S}_3 = 39/41 = 0.951$ (note, only 41 because 4 were censored)

$\hat{S}_4 = 34/39 = 0.872$

$\hat{S}_5 = 28/32 = 0.875$ (note, only 32 because 2 were censored)

$\hat{S}_6 = 25/28 = 0.893$

$\hat{S}_7 = 24/25 = 0.960$

$\hat{S}_8 = 24/24 = 1.000.$

Here one estimates 8 parameters (call this model $S(t)$); one could seek a more parsimonious model in several ways. First, perhaps all the parameters were nearly constant; thus a model with a single survival probability might suffice (i.e., $S(.)$) If something was known about the intervals (similar to the flood years for the European dipper data) one could model these with one parameter and denote the other periods with a second survival parameter. Finally, one

might consider fitting some smooth function across the occasions and, thus, have perhaps only one intercept and one slope parameters (instead of 8 parameters). Still other possibilities exist for both parsimonious modeling and probable heterogeneity of survival probability across animals. These extensions are not possible with the K-M method and K-L-based model selection is not possible.

Pollock's Staggered Entry Design

The Kaplan-Meier method assumes that all animals are released at occasion 1 and they are followed during the study until they die or are censored. Often new animals are released at each occasion period (say, weekly); we say this entry is “staggered" (Pollock et al. 1989). Assume, as before, that animals are fitted with radios and that these do not affect the animal's survival probability. This staggered entry fits easily into the
K-M framework by merely redefining the n_i to include the number of new animals released at occasion *i*. Therefore, conceptually, the addition of new animals into the marked population causes no difficulties in data analysis.

The Binomial Model

We focus on the so-called binomial model as this allows standard likelihood inference and is therefore similar to other models in program MARK. There are 3 possible scenarios under the known fate model. Each tagged animal either:

1. survives to end of study (detected at each sampling occasion after its release $\rightarrow$ the fate is known on every occasion).

2. dies sometime during the study (its carcass is found on the first sampling occasion after its death $\rightarrow$ the fate is known).

3. survives up to the point at which time it is censored.

Note, for purposes of estimating survival probabilities, there is no difference between an animal seen alive and then removed from the population at occasion *k* vs. an animal alive at occasion *k* and then censored due to radio failure or whatever.

The binomial model assumes the capture histories are mutually exclusive and exhaustive, that animals are independent, and all animals have the same underlying parameters during interval *j* (homogeneity across individuals).

Known fate data can be modeled by a product of binomials. Let us modify the black duck data slightly, $n_1 = 48$, $n_2 = 44$, and $n_3 = 41$; the first likelihood is

$$\mathcal{L}(S_1 \mid n_1, n_2) = \binom{n_1}{n_2} S_1^{n_2} (1-S_1)^{n_1-n_2} .$$

Clearly, one could find the MLE, $\hat{S}_1$, for this expression (e.g., $\hat{S}_1 = 44/48 = 0.917$). Of course, the other binomial terms are multiplicative, assuming independence. The survival during the second interval is based on $n_2 = 44$ and $n_3 = 41$,

$$\mathcal{L}(S_2 \mid n_2, n_3) = \binom{n_2}{n_3} S_2^{n_3} (1-S_2)^{n_2-n_3} .$$

The likelihood function for the entire set of black duck data (modified to better make some technical points below) is the product of these individual likelihoods. The log-likelihood is the sum of terms such as

$$\log(\mathcal{L}(S_i \mid \underline{n})) = \sum_i n_i \log(\text{Prob.}).$$

This expression is in “standard form" and should now be familiar.

Encounter Histories

Parameterization of encounter histories is critical. Each entry is paired, where the first position is a 1 if the animal is known to be alive at occasion *j*; that is, at the start of the interval. A 0 in this first position indicates the animal was not yet tagged at the start of the interval *j*.

The second position in the pair is 0 if the animal survived to the end of the interval. It is a 1 if it died sometime during the interval. As the fate of every animal is assumed known at every occasion, the sampling probabilities (*p*) and reporting probabilities (*r*) are 1. The examples below will help clarify the coding,

History	Probability	Number Observed
10 10 10 10	$S_1S_2S_3S_4$	17

Tagged at occasion 1 and survived until the end of the study

10 10 11 00	$S_1S_2(1-S_3)$	21

Tagged at occasion 1 and died during the third interval

10 11 00 00	$S_1(1-S_2)$	24

Tagged at occasion 1 and died during the second interval

11 00 00 00 $(1–S_1)$ 43
Tagged at occasion 1 and died during the first interval

10 00 00 11 S_1 $(1–S_4)$ 13
Tagged at occasion 1, censored during intervals 2 and 3, and died during the fourth interval.

10 00 00 00 S_1 9
Tagged at occasion 1, known to be alive at the end of the first interval, but not released at occasion 2 and thus was censored after the first interval.

Estimation of survival probabilities is based on a release (1) at the start of an interval and survival to the end of the interval (0), mortality probabilities are based on a release (1) and death (1) during the interval; if the animal then was censored, it does not provide information about S_i or $1–S_i$).

Some “rules" for encounter history coding:
A. The two-digit pairs each pertain to an interval (the period of time between occasions).
B. There are only 3 possible entries for each interval:
- 10 = an animal survived the interval, given it was alive at the start of the interval
- 11 = an animal died during the interval, given it was alive at the start of the interval
- 00 = an animal was censored for this interval

C. In order to know the fate of an animal during an interval, one must have encountered it BOTH at the beginning AND the end of the interval.

Censoring

Censoring appears “innocent" but it is often not. If a substantial proportion of the animals do not have exactly known fates, it might be better to consider models that allow the sampling parameters to be < 1. In practice, one almost inevitably loose track of some animals. Reasons for uncertainty about an animal's fate include radio transmitters that fail (this may or may not be independent of mortality) or animals that leave the study area. In such cases, the encounter histories must be coded correctly to allow these animals to be censored. Censoring often require some judgment.

When an animal is not detected at the end of an interval (i.e., immediately before occasion j) or at the beginning of the next interval (i.e., immediately after occasion j+1), then its fate is unknown and must be entered as a 00 in the encounter history matrix. Generally, this results in 2 pairs with a 00 history; this is caused by the fact that interval j is a 00 because the ending fate was not known and the fact that the beginning fate for the next interval (j+1) was not known. Censored intervals almost always occur in runs of two or more (e. g., 00 00 or 00 00 00). See the example above where the history was 10 00 00 11.

In this example, the animal was censored but re-encountered at the beginning of interval 4 (alive) and it died during that interval. It might seem intuitive to infer that the animal was alive and, thus, fill in the 2 censored intervals with 10 10 – this is incorrect and results in bias.

Censoring is assumed to be independent of the fate of the animal; this is an important assumption. If, for example, radio failure is due to mortality, bias will result in estimators of $\hat{S}$. Of course, censoring reduces sample size, so there is a trade-off here. If many animals must be censored, then the possible dependence of fates and censoring must be a concern.

Binomial Likelihood Functions Allowing Each Animal To Have Its Own Survival Parameter

Before we move into models for individual covariates, some quick review of the binomial likelihood might be helpful. Consider the usual *n* flips of a coin where,

$p =$ probability the coin lands heads;
$q = 1 - p =$ probability the coin lands tails.

Let *n* = 16 flips (trials). We often write the likelihood in a compact form as

$$\mathcal{L}(p \mid n, y) = (p)^y (1-p)^{n-y},$$

where $y =$ number of heads. If we observe $y = 5$, then

$$\mathcal{L}(p \mid 16, 5) = (p)^5 (1-p)^{16-5}.$$

Alternatively, we could write the likelihood for each individual outcome and take the product of these terms as the likelihood function. One alternative is to merely write the likelihood as (using the convention that $q = (1-p)$),

$$\mathcal{L}(p \mid 16, 5) = ppppp \cdot qqqqqqqqqqq.$$

Alternatively, we could write this as,

$$\mathcal{L}(p \mid 16, 5) = \prod_{i=1}^{5} p_i \cdot \prod_{i=6}^{16} (1-p_i).$$

Finally, we could define an indicator variable to denote head or tail; let $y = 1$ if heads, 0 if tails. Then the likelihood can be written for the i^{th} flip as

$$\mathcal{L}\left(p \mid n, \{y_1, y_2, \ldots, y_{16}\}\right) = \prod_{i=1}^{16} \left[(p)^{y_i} (1-p)^{1-y_i}\right].$$

Note, the subscript *i* is for *individual* coin flips; *i* = 1, 2, ..., 16 flips).

In these last three forms, each outcome (head or tail) has a probability term in the likelihood. The likelihood is the product of these individual probabilities. These formulations are useful in understanding the modeling of individual covariates.

INDIVIDUAL COVARIATES

Many of the data types allow modeling parameters as functions of covariates that are unique to each individual *i*. The band recovery models and the Cormack-Jolly-Seber models allow individual covariates. We introduce this important subject here in the context of the know fate models.

A number of people have suggested modeling of the individual animals, allowing covariates that vary by individual (e.g., White and Garrott 1990, Smith et al. 1994). This approach is very useful in the biological sciences. Here, each animal (*i*) has a unique survival probability in the likelihood. In the black duck example (slightly modified), $n_1 = 48$ and $n_2 = 44$ and the binomial likelihood for the survival probability during the first week (i.e., S_1) can be written as

$$\mathcal{L}(S_1 \mid n_1, n_2) = \binom{n_1}{n_2} S_1^{n_2} (1–S_1)^{n_1-n_2}$$

This can be re-expressed (omitting the multinomial coefficient) as

$$\mathcal{L}(S_1 \mid n_1, n_2) = \prod_{i=1}^{n_2} S_i \cdot \prod_{i=n_2+1}^{n_1} (1–S_i),$$

where the subscript *i* is over all the tagged ducks (48 ducks in the study). Thus, the first term in the likelihood is the product of the survival probability over 44 (= n_2) ducks that survival, while the second term in the product of of the mortality probabilities (1–*S*) for the 4 ducks that died during the first week ($4 = n_2–n_1 = 44–48$). So, a final expression of this likelihood is

$$\mathcal{L}(S_1 \mid n_1, n_2) = \prod_{i=1}^{44} S_i \cdot \prod_{i=45}^{48} (1-S_i),$$

Here, the likelihood is expressed as if all the animals that survived had the first 44 tag numbers, whereas the 4 animals that died had the final tag numbers. This is just to allow simple expressions to represent the data. Program MARK handles the fact that animals die or survive independently of the tag numbers and keeps track of which covariate is associated with which individual (from the encounter history matrix).
Thus, this is not an issue the investigator must worry about; other than having the correct information in the encounter history matrix.

Now we consider *modeling* the survival probability of these individuals as a nonlinear function of some covariate that varies for each individual animal *i*. The natural (but arbitrary) choice is the logit relationship

$$S_i = \frac{1}{1+\exp(-[\beta_0+\beta_1(X_i)])}$$

with link function

$$\log_e\left(S_i/(1-S_i)\right) = \beta_0 + \beta_1(X_i)$$

where X_i is the value of the covariate for the i^{th} individual. Of course, other functions could be used (log, log-log, complementary log-log, etc.). More than one covariate can also be measured and used with this general approach. If we substitute the logit submodel and its individual covariate into the likelihood above, the expression *looks* messy, but is conceptually familiar,

$$\mathcal{L}(\beta_0, \beta_1 \mid n_1, n_2, X_i) = \prod_{i=1}^{n_2} \left(\frac{1}{1+\exp(-[\beta_0+\beta_1(X_i)])} \right) \times$$

$$\prod_{i=n_2+1}^{n_1} \left(1- \left(\frac{1}{1+\exp(-[\beta_0+\beta_1(X_i)])} \right)\right)$$

or

$$\mathcal{L}(\beta_0, \beta_1 \mid n_1, n_2, X_i) = \prod_{i=1}^{44} \left(\frac{1}{1+\exp(-[\beta_0+\beta_1(X_i)])} \right) \times$$

$$\prod_{i=45}^{48} \left(1- \left(\frac{1}{1+\exp(-[\beta_0+\beta_1(X_i)])} \right)\right)$$

Thus, the MLEs for β_0 and β_1 (the intercept and slope, respectively) are the focus of the estimation. The individual survival parameters in the likelihood have been replaced with the β_j. Of course, additional binomial terms could be multiplied for the parameters S_2, S_3, ..., S_8 in the black duck example to obtain an overall expression for the likelihood function.

There are two notions to think clearly about:

(1) the survival probabilities are replaced by a logit (sub)model of the individual covariate X_i. Conceptually, every animal *i* has its own survival probability and this may be related to the covariate X_i.

(2) during the analysis, the covariate of the i^{th} animal must correspond to the survival probability of that animal. Program *MARK* handles this detail. Note, in the last expression of the likelihood (above) we assume it is the 48^{th} duck that died and this corresponds to the 48^{th} covariate, X_{48}.

Inference follows usual likelihood methods: K-L information can be estimated using AIC, AIC_c, or $QAIC_c$, models can be ranked using the Δ_i, Akaike weights and evidence ratios can measure strength of evidence for various covariate models. Further inference can be based on the $\hat{\beta}$ and $\widehat{se}(\hat{\beta})$, etc.

An Example of the Application

Assume a biologist has found 88 active nests of the red-cockaded woodpecker in which nest initiation occurred on the same day. She selects a single nestling from each of the 88 nests and measures 3 covariates on each of these 88 nestlings. The covariates measured are the number of ectoparasites found, the number of hatchlings in the nest, and the weight at hatching. The “first" occasion is actually on day 3 following hatching. Each bird is tagged uniquely with a colored leg band to allow it to be identified and its fate determined visually by daily inspection of the nest. Birds are followed for 12 days (while they are still in the nest;

they typically start to leave the nest after 15 days) and their fate is determined daily. Thus, the data follow the known fate scenario, even though animals are not fitted with radios. Overdispersion should not be a factor as only one bird in each nest is the subject of the study. Sample size is 88 (no staggered entry) and there are 12 occasions (days 3, 4, ..., 15).

If the data were modeled without an occasion effect (i.e., model $\{S_{(.)}$, rather than model $\{S_t\}$, one might potentially include the model with all three covariates for each individual woodpecker (i), as

$$\log_e\left(S_i/(1\text{-}S_i)\right) = \beta_0 + \beta_1(E_i) + \beta_2(H_i) + \beta_3(W_i) ,$$

where,

E_i is the number of ectoparasites on day 3 (= occasion 1)
H_i is the number of nest mates on day 3
W_i is the weight of the nesting on day 3.

Of course, other *a priori* models would be considered in making inferences from these data, this is just an example. One might ask if any of the individual covariates are important. If so, which covariate is more important? Should 1 or 2 or all 3 covariates be included in a good model for these data?

The estimation would focus on the β parameters, but the *interpretation* would be interesting. For example, one might look at the mean of the S_i given that all the covariates were held at their average values. Then this mean might be compared with means for low vs. high values of weight and the number of nest mates. One could compute the values of S for a range of ecoparasites, while holding the other covariates at their mean values. Other possibilities exist and could be explored for the selected model.

The estimates of the survival probabilities (by individual) can be plotted in Excel or any other convenient software package. That is, given the logit (or whatever link function is chosen), the individual covariates and the $\hat{\beta}_j$, one can plot the derived S_i,
program MARK does not provide such plots (but you can export the beta estimates to Excel and do the plots yourself). Instead, $\hat{S}_1$ (either the derived survival for the first animal in the encounter history matrix, or for the mean of the individual covariate, or else a value you specify) is provided as a check.

Now, one can see that individual covariates can be used in the band recovery models and the open capture-recapture models. *Too few biologists are taking full advantage of the information contained in individual covariates.*

Note, there are problems if the covariate changes through time in the band recovery and open C-R models. For example, if weight changes throughout the study period, one only has weights for those animals recaptured at various occasions. Thus, when animals are not captured (e.g., the “never recaptured" animals) then the value of their covariate at that time is

not known! This issue is not a problem with the known fate models if animals are actually handled (as opposed to merely being resighted).