

APPENDIX D

Variance components and random effects models in MARK ...

Kenneth P. Burnham, *USGS Colorado Cooperative Fish & Wildlife Research Unit*

The objectives of this appendix are

- to introduce to biologists the concept and nature of what are called (alternative names for the same essential idea) ‘variance components’, ‘random effects’, ‘random coefficient models’, or ‘empirical Bayes estimates’
- present the basic theory and methodology for fitting simple random effects models, including shrinkage estimators, to capture-recapture data (i.e., Cormack-Jolly-Seber and band or tag recovery models)
- extend AIC to simple random effects models embedded into the otherwise fixed-effects capture-recapture likelihood.
- develop some proficiency in executing a variance components analysis and fitting random effects model in program **MARK**

Much of the conceptual material presented in this appendix is derived from a paper authored by Kenneth Burnham and Gary White (2002) – hereafter, we will refer to this paper as ‘B&W’. It is assumed that the reader already has a basic knowledge of some standard encounter-mark-reencounter models as described in detail in this book (e.g., dead recovery and live recapture models – referred to here generically as ‘capture-recapture’).

We introduce the subject of – and some of the motivation for – this appendix by example. In the following we consider two relatively common scenarios (out of a much larger set of possibilities) where a ‘different analytical approach’ might be helpful.

Scenario 1 – parameters as random samples

Consider a Cormack-Jolly-Seber (CJS) time-specific model $\{S_t p_t\}$ wherein survival (S) and capture probabilities (p) are allowed to be time varying for $(k + 2)$ capture occasions, equally spaced in time. If $k \geq 20$ we are adding many survival parameters into our model as if they were unrelated; however, more parsimonious models are often needed. Consider a reduced parameter model – at the extreme, we have the model $\{S, p_t\}$ wherein $S_1 = S_2 = \dots = S_k = S$. However, this model may not fit well even if the general (time-dependent) CJS model fits well and there is no evidence of any explainable

structural time variation, such as a linear time trend, in this set of survival rates, or variation as a function of an environmental covariate. Instead, there may be unstructured time variation in the S_i that is not easily modeled by any simple smooth parametric form, yet which cannot be wisely ignored. In this case it is both realistic and desirable to conceptualize the actual unknown S_i as varying, over these equal-length time intervals, about a conceptual population mean $E(S) = \mu$, with some population variation, σ^2 (Fig. D.1).

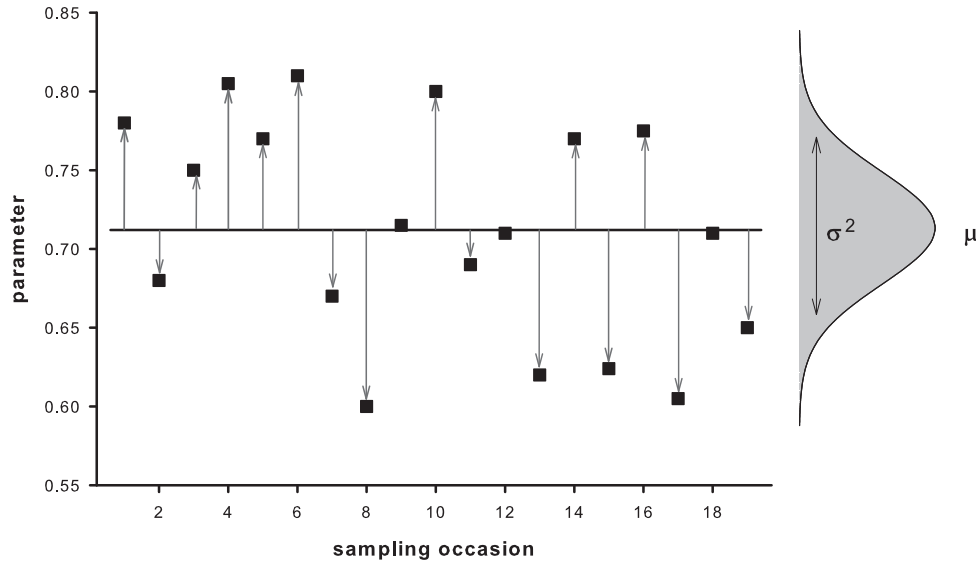


Figure D.1: Schematic representation of variation in occasion-specific parameters θ_i , as if the parameters were drawn randomly from some underlying distribution with mean μ and variance σ^2 .

Here, by population, we will mean a conceptual statistical distribution of survival probabilities, such that the S_i may be considered as a sample from this distribution. Hence, we proceed as if S_i are a *random* sample from a distribution with mean μ and variance σ^2 . Doing so can lead to improved inferences on the S_i regardless of the truth of this conceptualization if the S_i do in fact vary in what seems like a random, or exchangeable, manner. The parameter σ^2 is now the conventional measure of the unstructured variation in the S_i , and we can usefully summarize $S_1 \dots S_k$ by two parameters: μ and σ^2 . The complication is that we do not know the S_i ; we have only estimates \hat{S}_i , subject to non-ignorable sampling variances and covariances, from a capture-recapture model wherein we traditionally consider the S_i as fixed, unrelated parameters. We would like to estimate μ and σ^2 , and adjust our estimates to account for the different contributions to the overall variation in our estimates due to sampling, and the environment. For this, we consider a random effects model.

Scenario 2 – separating sampling + environmental (process) variation

Precise and unbiased estimation of parameter uncertainty (say, the SE of the parameter estimate) is critical to analysis of stochastic demographic models. Consider for example, the estimation of the risk of extinction. It is well known (and entirely intuitive) that any simply stochastic process (say, growth of an age- or size-structured population through time) is more likely to go extinct the more variable a particular ‘vital rate’ is (say, survival or fertility). Thus, if an estimate of the variance of a parameter is

biased high, then this tends to bias high the probability of extinction. We wish to use only estimates of environmental (or process) variation alone, excluding sampling variation, since it is only the magnitude of the former that we want to include in our viability models (White 2000).

Precise estimation of process variation is also critical for analysis of the relationship of the variation of a particular demographic parameter to the projected growth of a population. The process variation in projected growth, λ , is a function of the process variance of a particular demographic parameter. To first order, and assuming no covariances between the a_{ij} elements, this can be expressed as

$$\widehat{\text{var}}(\lambda) \approx \sum_{ij} \left(\frac{\partial \lambda}{\partial a_{ij}} \right) \widehat{\text{var}}(a_{ij}).$$

From this expression, we anticipate that natural selection will select against high process variation in a parameter that λ is most ‘sensitive’ to (i.e., for which $\partial \lambda / \partial a_{ij}$ is greatest) (Pfister 1998; Schmutz 2009). Thus, robust estimation of the process variance of fitness components is critical for life history analysis.

In this appendix, we will consider estimation of variance components, and fitting of random effects models, using program **MARK**. We begin with the development of some of the underlying theory, followed by illustration of the ‘mechanics’ of using program **MARK**, by means of a series of ‘worked examples’.

D.1. Variance components – some basic background theory

The basic idea is relatively simple. We imagine that the S_i are distributed randomly about $E(S) = \mu$ (Fig. D.1). The variation in S_i is σ^2 , as if S_1, \dots, S_k are a sample from a population. It is not required that the sampling be random – merely that the S_1, \dots, S_k are exchangeable (or, more formally, that the conceptual residuals $(\mu - S_i)$ should appear like an iid sample, with no remaining structural information). There are no required distributional assumptions, such as normality.

If we knew the S_i then it follows that

$$\hat{E}(S) = \bar{S} \quad \hat{\sigma}^2 = \frac{\sum^k (S_i - \bar{S})^2}{k - 1}.$$

Of course, except in a computer simulation, we rarely if ever know the S_i . What we might have are ML estimates \hat{S}_i , and estimates of conditional variation $\widehat{\text{var}}(\hat{S}_i | S_i)$.

We can express our estimate of \hat{S}_i in standard linear form as the sum of the mean μ , the deviation of the S_i from the mean, δ_i , and the error term ϵ_i

$$\hat{S}_i = \mu + \delta_i + \epsilon_i,$$

where $\delta_i = (S_i - \mu)$ and $\epsilon_i = (\hat{S}_i - S_i)$. Substituting into our expression for \hat{S}_i ,

$$\begin{aligned} \hat{S}_i &= \mu + \delta_i + \epsilon_i \\ &= \mu + \underbrace{(S_i - \mu)}_{\substack{\uparrow \sigma^2 \\ \text{(process} \\ \text{variance)}}} + \underbrace{(\hat{S}_i - S_i)}_{\substack{\uparrow \text{var}(\hat{S}_i | S_i) \\ \text{(sampling} \\ \text{variance)}}} \end{aligned}$$

Here (and hereafter) we distinguish between ‘process’ (or, environmental) variation, σ^2 , and ‘sampling’

variation, $\text{var}(\hat{S}_i | S_i)$. We refer to the sum of process and sampling variation as total variation, σ_{total}^2 .

$$\begin{aligned} \text{total variation} &= \sigma_{total}^2 = \text{process variation} + \text{sampling variation} \\ &= \sigma^2 + \text{var}(\hat{S}_i | S_i). \end{aligned}$$

It is important to note that *sampling* variation, $\text{var}(\hat{S}_i | S_i)$, depends on the sample size of animals captured, whereas *process* variance σ^2 does not. It is also important to note that if there is sampling covariation, then this should be included in our expression for total variance, σ_{total}^2 :

$$\sigma_{total}^2 = \sigma^2 + \left[E\left(\text{var}(\hat{S}_i | S_i)\right) + E\left(\text{cov}(\hat{S}_i, \hat{S}_j | S_i, S_j)\right) \right].$$

For fully time-dependent models, the sampling covariances of S_i and S_j are often very small for many of the data types we work with in **MARK**, especially relative to process and sampling variance, and the covariance term can often be ignored. We will do so now, for purposes of simplifying the presentation somewhat, but will return to the issue of sampling covariances later on.

If we assume for the moment that all the sampling variances are equal, then the estimate of the overall mean survival is just the mean of the k estimates:

$$\bar{\hat{S}} = \frac{\sum^k \hat{S}_i}{k},$$

with the theoretical variance being the sum of process and sampling variance divided by k :

$$\widehat{\text{var}}(\bar{\hat{S}}) = \frac{\sigma^2 + E\left[\text{var}(\hat{S}_i | S_i)\right]}{k}.$$

Our interest generally lies in estimation of the process variation. By algebra, we see that process variance can be estimated by, in effect, subtracting the sampling variation from the total variation.

$$\begin{aligned} \sigma_{total}^2 &= \text{process variation} + \text{sampling variation} \\ &= \sigma^2 + \widehat{\text{var}}(\hat{S}_i | S_i) \\ \therefore \sigma^2 &= \sigma_{total}^2 - \widehat{\text{var}}(\hat{S}_i | S_i). \end{aligned}$$

Hence, we need an estimate for σ_{total}^2 and $\text{var}(\hat{S}_i | S_i)$.

If we assume that S_1, \dots, S_k are a random sample, with $\bar{S} = \hat{E}(S)$, and population variance σ^2 , then from general least-squares theory

$$\bar{\hat{S}} = \frac{\sum^k w_i \hat{S}_i}{\sum^k w_i},$$

where

$$w_i = \frac{1}{\sigma^2 + \widehat{\text{var}}(\hat{S}_i | S_i)}.$$

Given w_i , the theoretical variance of \hat{S} is

$$\text{var}(\hat{S}) = \frac{1}{\sum^k w_i}.$$

However, although $\text{var}(\hat{S}_i | S_i)$ is estimable, we would still need to know σ^2 .

An alternative approach which leads to an empirical (data-based) estimator is

$$\widehat{\text{var}}(\hat{S}) = \frac{\sum^k w_i (\hat{S}_i - \bar{\hat{S}})^2}{(\sum^k w_i)(k-1)}.$$

We note that if there is no process variation ($\sigma^2 = 0$), then the above reduces to the familiar case of k replicates.

More generally, if we assume that the weights, w_i are equal (or nearly so), then we can re-write the empirical estimator as

$$\widehat{\text{var}}(\hat{S}) = \frac{\sum^k (\hat{S}_i - \bar{\hat{S}})^2}{(k-1)}, \quad \text{if } w_i = w, \forall_i$$

where

$$\bar{\hat{S}} = \frac{\sum^k \hat{S}_i}{k}.$$

The assumption that the weights, w_i are equal is generally reasonable if (i) the $\text{var}(\hat{S}_i | S_i)$ are all nearly equal, or if (ii) they are all small, relative to σ^2 . In theory, when the S_i vary, then the $\text{var}(\hat{S}_i | S_i)$ will also vary. In contrast, with low sampling effort, such that $\text{var}(\hat{S}_i | S_i)$ is much larger than process variance σ^2 , it might be sufficient to use the approximation

$$w_i = \frac{1}{\text{var}(\hat{S}_i | S_i)}.$$

In this case, only relative weights w_i would be needed (since an estimate of σ^2 is not needed).

Now, assume for the moment we are interested in estimating the process variation around the mean \bar{S} . If we also assume that there is no sampling covariance, and that $w_i = w$, and $\sum^k w_i = 1$, then we estimate the total variance as

$$\sigma_{total}^2 = \widehat{\text{var}}(\hat{S}) = \frac{\sum^k (\hat{S}_i - \bar{\hat{S}})^2}{(k-1)},$$

and the sampling variance as the mean of the estimated sample variances

$$E[\widehat{\text{var}}(\hat{S}_i | S_i)] = \frac{\sum^k \text{var}(\hat{S}_i | S_i)}{k}.$$

Hence, our estimate of process variance would be

$$\hat{\sigma}^2 = \frac{\sum^{k-1} (\hat{S}_i - \bar{\hat{S}})^2}{k-1} - \frac{\sum^k \text{var}(\hat{S}_i | S_i)}{k}.$$

However, this estimator (which is essentially the estimator described by Gould & Nichols 1998) is not entirely correct (or efficient). It was derived under the strong assumption that the sampling variances are all equal (i.e., that $\text{SE}(\hat{S}_i)$ are all identical). In practice, this is usually not the case, and thus we refer to the preceding as a ‘naïve’ estimator of process variance.

Instead, we need to weight them to obtain an unbiased estimate of σ^2 . As noted earlier, general least-squares theory suggests using a weight w_i

$$w_i = \frac{1}{\sigma^2 + \text{var}(\hat{S}_i \mid S_i)}.$$

Hence, the estimator of the *weighted* mean survival is

$$\bar{\hat{S}} = \frac{\sum^k w_i \hat{S}_i}{\sum^k w_i},$$

with theoretical variance

$$\text{var}(\bar{\hat{S}}) = \frac{1}{\sum^k w_i},$$

and empirical variance estimator

$$\widehat{\text{var}}(\bar{\hat{S}}) = \frac{\sum^k w_i (\hat{S}_i - \bar{\hat{S}})^2}{\left[\sum^k w_i\right](k-1)}.$$

When the w_i are the true (but unknown) weights, we have

$$\frac{1}{\sum^k w_i} = \frac{\sum^k w_i (\hat{S}_i - \bar{\hat{S}})^2}{\left[\sum^k w_i\right](k-1)},$$

which if we normalize the weights (such that they sum to 1), gives

$$1 = \frac{\sum^k w_i (\hat{S}_i - \bar{\hat{S}})^2}{(k-1)}.$$

Since

$$w_i = \frac{1}{\sigma^2 + \text{var}(\hat{S}_i \mid S_i)} \quad \text{and} \quad \bar{\hat{S}} = \frac{\sum^k w_i \hat{S}_i}{\sum^k w_i},$$

then

$$1 = \frac{\sum^k w_i (\hat{S}_i - \bar{\hat{S}})^2}{(k-1)} = \frac{\sum^k \left[\frac{1}{\sigma^2 + \text{var}(\hat{S}_i \mid S_i)} \left(\hat{S}_i - \sum^k \frac{1}{\sigma^2 + \text{var}(\hat{S}_i \mid S_i)} \hat{S}_i \right)^2 \right]}{(k-1)}$$

We then solve (numerically) for $\hat{\sigma}^2$ (which is the only unknown in the expression) – it is convenient to use the naïve estimate for σ^2 calculated earlier as a starting point in the numerical optimization.

A confidence interval can be constructed for σ^2 by solving two modified versions of this equation. We assume we want a $(1 - \alpha)\%$ CI, where $\alpha = \alpha_U + \alpha_L$ (where U and L stand for upper and lower, respectively). For the upper limit, we solve for σ^2 in the following

$$\frac{\sum_{i=1}^k \left[\frac{1}{\sigma^2 + \text{var}(\hat{S}_i | S_i)} \left(\hat{S}_i - \sum_{i=1}^k \frac{1}{\sigma^2 + \text{var}(\hat{S}_i | S_i)} \hat{S}_i \right)^2 \right]}{(k-1)} = \frac{\chi_{k-1, \alpha_L}^2}{k-1},$$

where χ_{k-1, α_L}^2 is the critical value for the central χ^2 distribution corresponding to $(k-1)$ df and α_L percentile.

To find the lower limit, we substitute $\chi_{k-1, 1-\alpha_U}^2$ in the RHS of the preceding, and solve for σ^2 . If the lower limit does not have a positive solution for σ^2 (since $\sigma \geq 0$), then we set the lower CI to 0 and adjust to a one-sided CI by redefining $\alpha_U = \alpha$.

Burnham *et al.* (1987) describe simplified versions of these estimators for the CI if all the $\text{var}(\hat{S}_i | S_i)$ are the same, or nearly so.

D.2. Variance components estimation – worked examples

Here, we introduce the ‘mechanics’ of the variance decomposition, using a series of progressively more complex examples. We begin with a simple example loosely based on a ‘known fate’ analysis, where survival is estimated as a simple binomial probability, and where there is no covariance among samples.

D.2.1. Binomial survival – simple mean (no sampling covariance)

Imagine a simulated scenario where we are conducting a simple ‘known fate’ analysis (Chapter 17). In each of 10 years ($k = 10$), we mark and release $n = 25$ individuals, and determine the number alive, y , after 1 year (since this is a known-fate analysis, we assume there is no error in determining whether an animal is ‘alive’ or ‘not alive’ on the second sampling occasion). Here, though, we’ll assume that the survival probability in each year, S_i , is drawn from $N(0.5, 0.05)$ (i.e., distributed as an independent normal random variable with mean $\mu = 0.5$ and process variance $\sigma^2 = 0.05^2$).

Conditional on each S_i , we generated y_i (number alive after one year in year i) as an independent binomial random variable $B(n, S_i)$. Thus, our ML estimate of survival for each year is $\hat{S}_i = y_i/n$, with a conditional sampling variance of $\widehat{\text{var}}(\hat{S}_i | S_i) = [\hat{S}_i(1 - \hat{S}_i)]/n$, which given $\mu = 0.5$, and $\sigma^2 = (0.05)^2$ is approximately 0.01.

Table D.1 (top of the next page) gives the values of S_i , y_i and \hat{S}_i for our ‘example data’. Clearly, for a ‘real analysis’, we would not know the true values for S_i – we would have only \hat{S}_i , and generally only have $\widehat{E}_S(\text{var}(\hat{S}_i | S_i))$ as $\widehat{\text{var}}(\hat{S}_i | S_i)$.

In Table D.1 we see that the *empirical* standard deviation of the 10 estimated survival rates (i.e., the \hat{S}_i) is 0.106. However, we should not take $(0.106)^2$ as an estimate of σ^2 because such an estimate includes *both* process and sampling variation. Clearly, we want to subtract the estimated sampling variance from the total variation to get an estimate of the overall process variation.

Table D.1: Single realization from simple binomial survival example, $k = 10$, $E(S) = 0.5$, $\sigma = 0.05$, where $\hat{S}_i = y_i/n$ are $B(25, S_i)$, hence expected $SE(\hat{S}_i|S) \approx 0.1$

year (i)	S_i	\hat{S}_i	$\widehat{SE}(\hat{S}_i S_i)$
1	0.603	0.640	0.096
2	0.467	0.360	0.096
3	0.553	0.480	0.100
4	0.458	0.440	0.100
5	0.506	0.480	0.100
6	0.498	0.320	0.093
7	0.545	0.600	0.098
8	0.439	0.400	0.098
9	0.488	0.560	0.099
10	0.480	0.560	0.099
mean	0.504	0.484	0.100
SD	0.050	0.106	

Using the manual approach...

From section D.1, if we make the strong assumption that all the sampling variances are equal, then the estimate of the overall mean is the mean of the k estimates:

$$\bar{\hat{S}} = \frac{\sum^k \hat{S}_i}{k},$$

with the theoretical variance being

$$\widehat{\text{var}}(\bar{\hat{S}}) = \frac{\sigma^2 + E[\text{var}(\hat{S}_i | S_i)]}{k}.$$

In other words the total variance is the sum of the process (environmental) variance, σ^2 , plus the expected sampling variance, $E[\text{var}(\hat{S}_i | S_i)]$.

From section D.1, and assuming equal weights w_i , where $\sum^k = 1$, we estimate the *total* variance as

$$\widehat{\text{var}}(\bar{\hat{S}}) = \frac{\sum^k (\hat{S}_i - \bar{\hat{S}})^2}{(k-1)},$$

and the expected *sampling* variance as the mean of the sampling variances

$$E[\widehat{\text{var}}(\hat{S}_i | S_i)] = \frac{\sum^k \text{var}(\hat{S}_i | S_i)}{k}.$$

Thus, we can derive an estimate of the *process* (environmental) variance σ^2 by algebra

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{10} (\hat{S}_i - \bar{\hat{S}})^2}{(10 - 1)} - \frac{\sum_{i=1}^{10} \text{var}(\hat{S}_i | S_i)}{10}.$$

From Table (D.1), the process variance for our $k = 10$ samples is estimated as

$$\begin{aligned} \sigma^2 &= \frac{\sum_{i=1}^{10} (\hat{S}_i - \bar{\hat{S}})^2}{(10 - 1)} - \frac{\sum_{i=1}^{10} \text{var}(\hat{S}_i | S_i)}{10} \\ &= \left(\frac{0.10064}{9} \right) - 0.00959 \\ &= 0.0016 \\ \therefore \sigma &= \sqrt{0.0016} = 0.040. \end{aligned}$$

While our estimate of process variance is not much different from the true underlying value (for this example, true $\sigma^2 = 0.0025$), we noted that this naïve estimator is not entirely correct, since it assumes equal sampling variances. To obtain an unbiased estimate of σ^2 , we weight the sampling variance by

$$w_i = \frac{1}{\sigma^2 + \text{var}(\hat{S}_i | S_i)}.$$

From section D.1, we derive an estimate of process variance over the $k = 10$ samples by solving (numerically) the following for σ^2

$$\begin{aligned} 1 &= \frac{\sum_{i=1}^{10} w_i (\hat{S}_i - \bar{\hat{S}})^2}{(10 - 1)} \\ &= \frac{\sum_{i=1}^{10} \frac{1}{\sigma^2 + \text{var}(\hat{S}_i | S_i)} \left(\hat{S}_i - \frac{\sum_{i=1}^{10} w_i \hat{S}_i}{\sum_{i=1}^{10} w_i} \right)^2}{(10 - 1)} \\ &= \frac{\sum_{i=1}^{10} \left[\frac{1}{\sigma^2 + \text{var}(\hat{S}_i | S_i)} \left(\hat{S}_i - \sum_{i=1}^{10} \frac{1}{\sigma^2 + \text{var}(\hat{S}_i | S_i)} \hat{S}_i \right)^2 \right]}{(10 - 1)}. \end{aligned}$$

For the present example, our estimated process variance is $\hat{\sigma}^2 = 0.00195$.

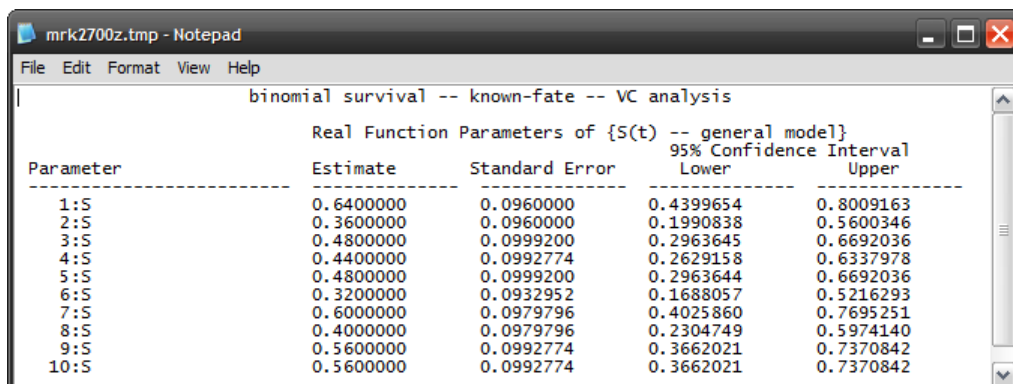
Now, using MARK...

While estimating process variance ‘by hand’ is relatively straightforward for this example, we are clearly interested in using the capabilities of program **MARK** to handle the ‘heavy lifting’ – especially for more

complex problems. Here we will introduce some of the ‘mechanics’ in using program **MARK** to estimate process variance for our simulated ‘known fate’ data. The data (number of marked and released animals that survive the one-year interval; see Table D.1) are formatted for the ‘known fate’ data type, and are contained in **binomial-example.inp**.

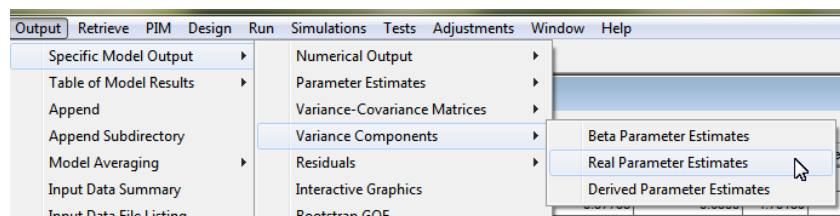
Since our purpose here is to demonstrate the mechanics of ‘variance components analysis’, and not ‘known fate analysis’, we’ve gone ahead and built the basic general model $\{S(t)\}$ for you. Start **MARK**, and open up **binomial-example.dbf** (note: you’ll need to have **binomial-example.fpt** installed in the same directory where you have **binomial-example.dbf**). There is only one model in the browser (for now) – model ‘ $S(t)$ -- general model’. [At some point, you should look at the underlying PIM structure, to see how we are using the known fate data type to model survival using a simple binomial estimator.]

Retrieve model ‘ $S(t)$ -- general model’, and look at the real parameter estimates (shown below). If you compare these survival estimates with those ‘done by hand’ in Table D.1, we see they are identical.

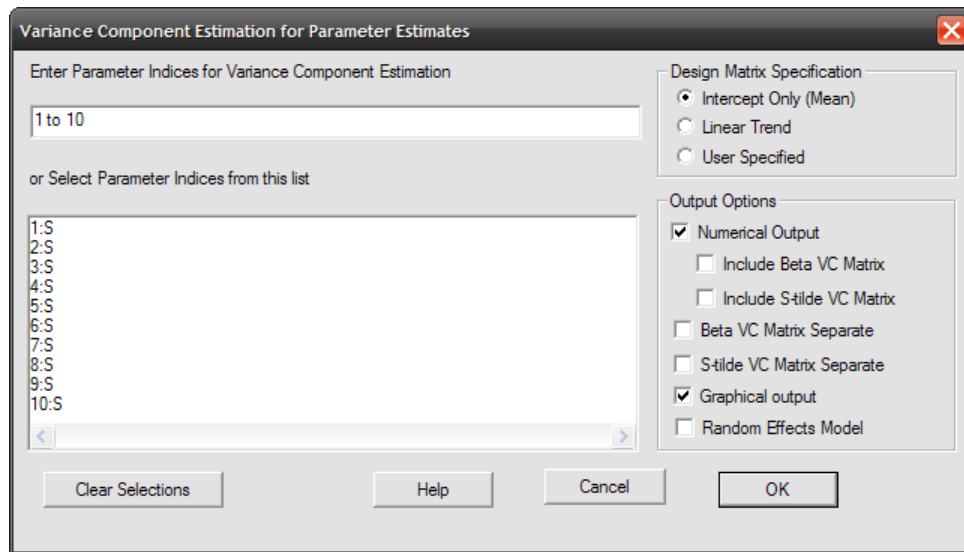


Parameter	Estimate	Standard Error	95% Confidence Interval	
			Lower	Upper
1:S	0.6400000	0.0960000	0.4399654	0.8009163
2:S	0.3600000	0.0960000	0.1990838	0.5600346
3:S	0.4800000	0.0999200	0.2963645	0.6692036
4:S	0.4400000	0.0992774	0.2629158	0.6337978
5:S	0.4800000	0.0999200	0.2963644	0.6692036
6:S	0.3200000	0.0932952	0.1688057	0.5216293
7:S	0.6000000	0.0979796	0.4025860	0.7695251
8:S	0.4000000	0.0979796	0.2304749	0.5974140
9:S	0.5600000	0.0992774	0.3662021	0.7370842
10:S	0.5600000	0.0992774	0.3662021	0.7370842

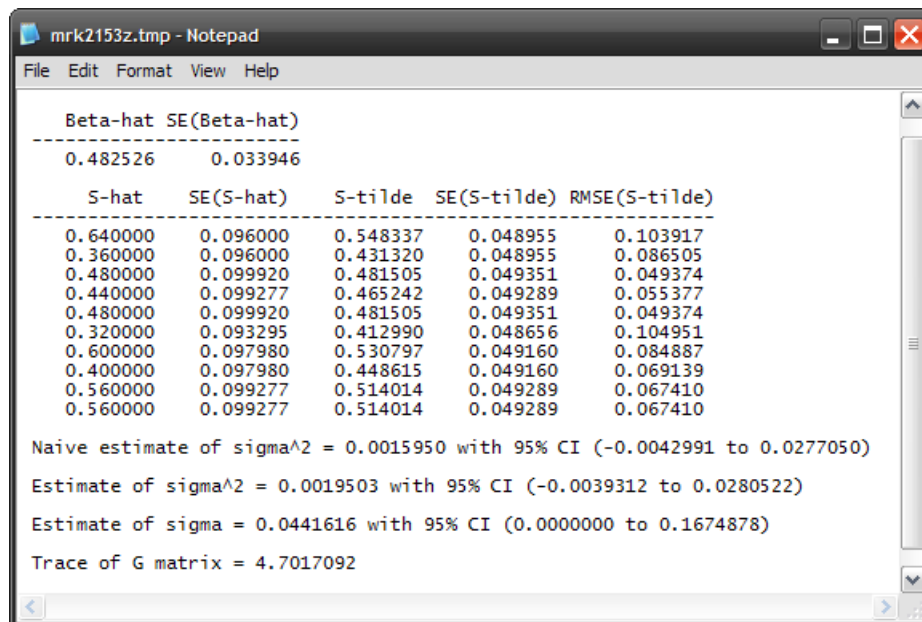
Next, we will use **MARK** to handle the estimation of process variance for us. Start by having a look at the ‘Output | Specific Model Output | Variance Components’ menu. You’ll see that we can execute a variance components analysis on either the β estimates, the real parameter estimates, or on any derived parameter(s). Here, we are interested in the real parameters, so, select that option.



You will then be presented with a window (shown at the top of the next page) asking you to specify which parameters you want to include in the variance component estimation, plus some options concerning the specification of the design matrix, followed by various output options. Note that we specify parameters ‘1 to 10’ to be included in the estimation. Since the ‘data’ were generated based on a sample of survival values drawn from a normal distribution with unknown parametric mean μ and variance σ^2 , it should make intuitive sense that we are going to fit a ‘mean model (i.e., intercept only)’. This is checked by default (we will consider other DM specifications elsewhere). For output options, we will accept the default options (selected) – note that at this point, we are not considering the fitting of a ‘random effects model’ (at least, not directly), so we leave that box unchecked.



Once you have completed entering the parameters and specifying the DM and the output options, click 'OK'. MARK will respond (generally very quickly) by outputting the estimates from the VC analysis into an editor window, shown below (MARK will also generate a plot of various estimates – ignore and close the plot for now):



Starting from the top – the first line reports a 'Beta-hat' of 0.482526. As you might recall from Chapter 6, this is in fact our most robust estimate of the mean survival probability. Note that it is close, but not identical to the simple arithmetic mean $\hat{S}_i = 0.484$. We will outline the reasons for the differences later – for now, we'll accept with deferred proof the statement that 'Beta-hat' represents our best estimate for mean survival, since it is the estimate of the expected value of S as a random variable. This estimate is followed by the estimate of 'SE(Beta-hat)'. We'll defer discussing this for a moment.

Next, a table of various parameter estimates. The first two columns should be self-explanatory – the ML estimates of survival \hat{S}_i ('S-hat'), followed by the binomial standard error for the estimate ('SE(S-hat)'). Next, the 'shrinkage' estimates \tilde{S}_i ('S-tilde') and their corresponding SE and RMSE. The derivation, use and interpretation of the shrinkage estimates is developed in section D.3.

Finally, the estimates for process variation (the **G** matrix we'll get to later). First, **MARK** reports the 'naive estimate of σ^2 ' = 0.001595. This is *exactly* the same value as the 'first approximation' we derived 'by hand' in the preceding section. This is followed by the 'Estimate of σ^2 ' = 0.0019503 (and the 'Estimate of σ ' = 0.044162). Both estimates are *identical* to those we derived 'by hand' using the 'weighted means' approach in the preceding section.

Now, about the 'SE(Beta-hat)' noted above. For this example, in the absence of sampling covariance, it is estimated as the square-root of the sum of estimated process variation, $\hat{\sigma}^2$, and sampling variation, $E[\widehat{\text{var}}(\hat{S}_i | S_i)]$, divided by k , where k is the number of parameter estimates. For our present example, with $k = 10$,

$$\begin{aligned} \text{'SE(Beta-hat)'} &= \sqrt{\frac{\hat{\sigma}^2 + E[\widehat{\text{var}}(\hat{S}_i | S_i)]}{k}} \\ &= \sqrt{\frac{(0.0019503 + 0.0095872)}{10}} = 0.03396, \end{aligned}$$

which is what is reported by **MARK** (to within rounding error). Thus, our estimate of total variance (i.e., the value of the numerator inside the square-root) would be $(\widehat{\text{SE}}^2 \times k) = (0.03396^2 \times 10) = 0.01153$.

The 'SE(Beta-hat)' is useful for calculating 95% CI for 'Beta-hat'. For this example, we can construct a reasonable CI for 'Beta-hat' as $0.482526 \pm (1.96 \times 0.033946) \rightarrow [0.4160, 0.5491]$.

D.2.2. Binomial example extended – simple trend

Here we consider a similar scenario, again involving a simple 'known fate' analysis with no sampling covariance among samples. In each of 15 years ($k = 15$), we mark and release $n = 25$ individuals, and determine the number alive, y , after 1 year. Here, though, we assume that the true mean survival probability in each year, \bar{S}_i , is declining over time (from 0.60 in the first year, to 0.46 in the final year). We'll assume that the survival probability in each year, S_i , is drawn from $\mathcal{N}(\bar{S}_i, 0.05)$. Conditional on each S_i , we generated y_i (number alive after one year in year i) as an independent binomial random variable $B(n, S_i)$. Table D.2 (top of the next page) gives the values of S_i , y_i and \hat{S}_i for our 'example data'.

Now, in the preceding example, the survival probability in each year, S_i , was drawn from $\mathcal{N}(0.5, 0.05)$ (i.e., distributed as an independent normal random variable with mean $\mu = 0.5$ and *process* variance $\sigma^2 = 0.05^2$). Here, though, not only is there random variation around the mean, but the mean itself declines over time. In this example there are 2 sources of process variation: the random variation around the mean survival in any given year, and the decline over time in the value of the mean. As such, we anticipate that the actual process variance will be $> (0.05)^2$.^{*} We also anticipate that if we estimate process variance using a model based on a simple mean (i.e., where we assume that process variation is due only to variation around a mean survival which is the same in all years) the estimate will be biased high (since it will conflate within- and among-year sources of variation). What about sampling variation? The imposition of a trend does not influence sampling variation – in each year, sampling is based on a binomial with the same number of individuals 'released' each year.

^{*} The actual value of the process variance, accounting for both within and among season variation, is $\sigma^2 = 0.0045$.

Table D.2: Single realization from simple binomial survival example, $k = 15$, S declining linearly from $0.6 \rightarrow 0.46$ ($\bar{S} = 0.53$), $\sigma = 0.05$, where $\hat{S}_i = y_i/n$ are $B(25, S_i)$ – hence expected $SE(\hat{S}_i | S) \approx 0.1$.

year (i)	S_i	\hat{S}_i	$\widehat{SE}(\hat{S}_i S_i)$
1	0.647	0.560	0.0099
2	0.595	0.440	0.0099
3	0.667	0.440	0.0099
4	0.580	0.640	0.0092
5	0.532	0.640	0.0092
6	0.475	0.720	0.0081
7	0.624	0.360	0.0092
8	0.516	0.400	0.0096
9	0.640	0.520	0.0100
10	0.430	0.480	0.0100
11	0.503	0.400	0.0096
12	0.509	0.520	0.0100
13	0.533	0.360	0.0092
14	0.394	0.360	0.0092
15	0.490	0.240	0.0073
mean	0.542	0.472	0.0093
SD	0.081	0.129	

We'll test both expectations, using data contained in **binomial-example-trend.inp**. Again, we've provided you with some 'pre-built' models to start with (contained in **binomial-example-trend.dbf** and **binomial-example-trend.fpt**). We'll avoid doing the same 'hand calculations' we worked through in the preceding example (same basic idea, but a fair bit messier because of having to account for both within and among year variation), and simply use **MARK**.

Start **MARK**, and open up **binomial-survival-trend.dbf**. You'll see that there are 2 'pre-built' models in the browser: 'S(t) - DM' (simple time variation) and 'S(T) - DM' (a trend model, where annual estimates are constrained to follow a linear trend). The '-DM' simply indicates that both were constructed using a design matrix. Based on AIC_c weights, there is clearly far more support for the trend model (which is the true generating model) than the model with simple time variation.

Our purpose here is not to do 'model selection' (we'll get there). Our present interest is on estimating the variance components. So, first question. Which model do we want to estimate variance components from? This is a more subtle question than it might seem. On the one hand, if we didn't know there was a trend, it might seem that we should select the time-dependent model since it is more general. On the other hand, you might have prior information suggesting a trend, and might think that it is a better model. Or, you might build both models, see that the trend model has the most support in the data, and use that model as the basis for estimating variance components.

You need to think this through carefully. We are trying to estimate process variance – we want to estimate the magnitude of the joint within- and among-year variation in our data. Thus, we want to use the most general model possible. In this case, model 'S(t) - DM'. We don't use model 'S(T) - DM', since the estimates from that model are constrained to fall on a straight line. Meaning, the only remaining variation would be the annual variation in mean survival (as estimated by the regression equation). Meaning, that any estimate of process variation from such a model would massively underestimate

true process variation in the data.

Start by retrieving model 'S(t) - DM'. Then, select 'Output | Specific model output | Variance components | Real parameter estimates'. With 15 samples we specify '1 to 15'.

What about the 'design matrix specification'? Recall from the preceding example that we used the default 'Intercept Only (mean)' specification. However, there are 2 other options available to you: 'linear trend', and 'user specified'. In effect, the first 2 options ('intercept only' and 'linear trend') are there simply for your convenience, since both models are very commonly used. You could, however, build either model by selecting the 'user specified' option (which essentially is the option you select if you want to build a specific model directly, using the design matrix). We'll defer using the 'user specified' option for now, and simply compare the 'intercept only' and 'linear trend' models. We'll start with the default 'intercept only' option.

Once you click the 'OK' button, MARK will respond with the estimates of year-specific survival probabilities, and the estimates of total and process variance (shown below). Again, the first line is the estimate of the overall mean, $\hat{S} = 0.4711$, and $SE = 0.0342$ (representing total variance). Note that the reported mean is very close to the mean of the true S_i (Table D.2), $\bar{S}_i = 0.472$.

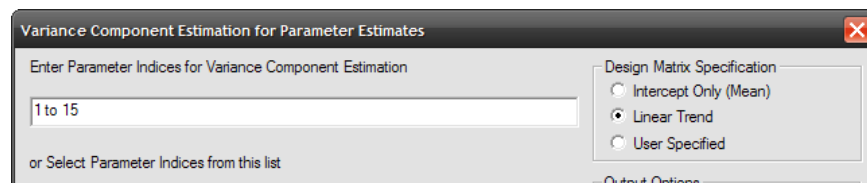
```
binomial survival -- long-term trend -- intercept model|
Beta-hat SE(Beta-hat)
-----
0.471110 0.034211

S-hat SE(S-hat) S-tilde SE(S-tilde) RMSE(S-tilde)
-----
0.560000 0.099277 0.531103 0.069531 0.075297
0.440000 0.099277 0.450114 0.069531 0.070263
0.440000 0.099277 0.450114 0.069531 0.070263
0.640000 0.096000 0.587166 0.068396 0.086426
0.640000 0.096000 0.587166 0.068396 0.086426
0.720000 0.089800 0.648076 0.066078 0.097670
0.360000 0.096000 0.394759 0.068396 0.076722
0.400000 0.097980 0.422774 0.069089 0.072746
0.520000 0.099920 0.503990 0.069747 0.071561
0.480000 0.099920 0.477089 0.069747 0.069808
0.400000 0.097980 0.422774 0.069089 0.072746
0.520000 0.099920 0.503990 0.069747 0.071561
0.360000 0.096000 0.394759 0.068396 0.076722
0.360000 0.096000 0.394759 0.068396 0.076722
0.240000 0.085417 0.302774 0.064296 0.089859

Naive estimate of sigma^2 = 0.0073874 with 95% CI (-0.0003801 to 0.0323057)
Estimate of sigma^2 = 0.0082451 with 95% CI (0.0005168 to 0.0331364)
Estimate of sigma = 0.0908028 with 95% CI (0.0227342 to 0.1820339)
```

What about the estimated variance? The estimate of process variance $\hat{\sigma}^2 = 0.00825$, which as we anticipated is almost twice as large as the true process variance in the data ($\sigma^2 = 0.0045$). Thus, the estimated SE for total variance will also be biased high.

Now, let's do a variance components analysis on the time-dependent model, by checking the 'linear trend' DM option, as shown below:



Here are the parameter estimates.

```

binomial survival -- long-term trend -- trend DM|
-----
Beta-hat SE(Beta-hat)
-----
0.613530 0.060965
-0.017643 0.006631

S-hat SE(S-hat) S-tilde SE(S-tilde) RMSE(S-tilde)
-----
0.560000 0.099277 0.578122 0.064386 0.066887
0.440000 0.099277 0.509807 0.061779 0.093218
0.440000 0.099277 0.500899 0.059476 0.085124
0.640000 0.096000 0.592221 0.056946 0.074335
0.640000 0.096000 0.583535 0.055391 0.079098
0.720000 0.089800 0.620838 0.053140 0.112504
0.360000 0.096000 0.424021 0.053516 0.083443
0.400000 0.097980 0.436196 0.053546 0.064632
0.520000 0.099920 0.486892 0.054014 0.063353
0.480000 0.099920 0.458235 0.054653 0.058827
0.400000 0.097980 0.409730 0.055428 0.056276
0.520000 0.099920 0.460038 0.057235 0.082894
0.360000 0.096000 0.371902 0.058415 0.059616
0.360000 0.096000 0.363215 0.060606 0.060691
0.240000 0.085417 0.288768 0.060512 0.077717

Naive estimate of sigma^2 = 0.0073874 with 95% CI (-0.0003801 to 0.0323057)
Estimate of sigma^2 = 0.0031995 with 95% CI (-0.0025058 to 0.0225038)
Estimate of sigma = 0.0565642 with 95% CI (0.0000000 to 0.1500125)

```

Note that we no longer have an estimate of a single ‘Beta-hat’ (i.e., we no longer have an estimate of just the mean). What we do have, though, is an estimate of the intercept ($\hat{\beta}_1 = 0.61353$), and the slope of the decline over time ($\hat{\beta}_2 = -0.017643$). Both are quite close to the values of $\hat{\beta}_1 = 0.64$ and $\hat{\beta}_2 = -0.0102$, estimated from a regression of the true S_i (Table D.2) on year. This is not surprising. The estimated process variance, $\hat{\sigma}_2 = 0.0031995$, is much more consistent with the true process variance, $\sigma^2 = 0.0045$, than was our estimate under the ‘intercept only’ model.

D.2.3. What about sampling covariance?

In the preceding examples, we considered situations where there was no sampling covariance. While this is a useful place to start, it is not particularly realistic in many situations, where sampling covariances are potentially not small. Recall that our simple (naïve) estimator for process variance, $\hat{\sigma}^2$, for some parameter θ was given as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k-1} (\hat{\theta}_i - \bar{\hat{\theta}})^2}{k-1} - E[\text{var}(\hat{\theta}_i | S_i)].$$

This is a suitable *first* approximation when sampling covariances are 0, or nearly so.

However, when sampling covariances are significant, then we need to modify the estimator for σ^2 to explicitly account for the sampling covariance.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k-1} (\hat{\theta}_i - \bar{\hat{\theta}})^2}{k-1} - E[\text{var}(\hat{\theta}_i | \theta_i)] + E[\text{cov}(\hat{\theta}_i, \hat{\theta}_j | \theta_i, \theta_j)].$$

To illustrate how sampling covariance is handled, we simulated a data set of live-encounter (CJS) data, 21 occasions, 350 newly marked individuals released at each occasion, under true model $\{\varphi_t p_t\}$.

Apparent survival, φ_i , over a given interval i was generated by selecting a random beta deviate drawn from $\mathcal{B}(0.7, 0.005)$. The encounter probability p_i at a sampling occasion i was generated by selecting a random beta deviate drawn from $\mathcal{B}(0.35, 0.005)$. The simulated live-encounter data are contained in **normsim-VC.inp**. The estimates $\hat{\varphi}_i$ from model $\{\varphi_i p_i\}$ are shown at the top of the next page. For a time-dependent model, the terminal φ and p parameter estimates are confounded (reflected in the estimated $\widehat{\text{SE}}(\hat{\varphi}_{20}) = 0.000$). This becomes important when we specify the parameters in the variance-components analysis.

live encounter data -- sampling covariance -- VC/RE example

Real Function Parameters of {phi(t)p(t) -- sin link}

Parameter	Estimate	Standard Error	95% Confidence Lower	Upper
1:Phi	0.7906864	0.0533752	0.6675126	0.8766605
2:Phi	0.6591159	0.0411183	0.5746048	0.7345925
3:Phi	0.7492055	0.0399619	0.6631864	0.8192438
4:Phi	0.8066294	0.0458168	0.7010975	0.8812146
5:Phi	0.7515650	0.0417704	0.6611675	0.8242559
6:Phi	0.8080092	0.0418407	0.7126907	0.8771553
7:Phi	0.6795555	0.0440437	0.5879031	0.7591739
8:Phi	0.5833971	0.0382249	0.5071184	0.6558814
9:Phi	0.7323945	0.0410031	0.6449145	0.8048444
10:Phi	0.8055003	0.0409223	0.7128116	0.8735797
11:Phi	0.6951929	0.0377786	0.6165833	0.7638582
12:Phi	0.7288343	0.0451865	0.6319494	0.8079654
13:Phi	0.6030040	0.0383181	0.5260439	0.6751838
14:Phi	0.7797401	0.0458234	0.6772593	0.8565714
15:Phi	0.6030161	0.0323311	0.5382590	0.6643546
16:Phi	0.7964490	0.0409396	0.7045976	0.8652042
17:Phi	0.5968084	0.0347508	0.5272547	0.6626758
18:Phi	0.7340089	0.0485502	0.6289387	0.8179397
19:Phi	0.6189786	0.0447323	0.5283357	0.7020264
20:Phi	0.5628481	0.0000000	0.5628481	0.5628481

Estimation of σ^2 under the naïve model is straightforward. The only additional complications are that (i) the terms in the estimator are calculated over $\varphi_1 \rightarrow \varphi_{19}$, and (ii) we have to calculate the mean of the sample covariances to estimate $E[\text{cov}(\hat{\varphi}_i, \hat{\varphi}_j \mid \varphi_i, \varphi_j)]$. In practice, this second step isn't too difficult, depending on your facility with computers. You simply need to find a way to calculate the mean of the off-diagonal elements of the V-C matrix (keeping in mind you're calculating the mean over $\varphi_1 \rightarrow \varphi_{19}$). For the present example, $\overline{\text{cov}}(\hat{\varphi}_i, \hat{\varphi}_j \mid \varphi_i, \varphi_j) = -0.00008$. Thus,

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\sum_{k=1}^{k-1} (\hat{\varphi}_i - \bar{\varphi})^2}{k-1} - E[\text{var}(\hat{\varphi}_i \mid \varphi_i)] + E[\text{cov}(\hat{\varphi}_i, \hat{\varphi}_j \mid \varphi_i, \varphi_j)] \\ &= \left(\frac{0.11531}{18} \right) - \left(\frac{0.0338}{19} \right) - 0.00008 = 0.004552.\end{aligned}$$

Clearly, the proportional contribution of the covariance term is very small (2%). This will often be the case, especially for time-dependent models.

If we analyze these live-encounter data using the variance components routines in **MARK**, using the 'intercept only' mean model, the reported value for the 'naïve' estimate (shown at the top of the next page) matches the value we derived by hand on the preceding page. The estimate based on the 'weighted' estimator is almost identical – and both are not too far off the true value of $\sigma^2 = 0.005$.

The near equivalence of the 'naïve' and 'weighted' estimates reflects the fact that sampling variation is small, relative to process variance, in this example (small sampling variance is not surprising, given that the data were generated under a model with $p = 0.35$ and 350 individuals marked and released on

each sampling occasion). Recall from section D.1 that from least-squares theory, we should weight our estimates of total and sampling variance to obtain an unbiased estimate of process variance, σ^2 , using a weight w_i :

$$w_i = \frac{1}{\sigma^2 + \text{var}(\hat{\phi}_i \mid \varphi_i)}.$$

For this example, $\text{var}(\hat{\phi}_i \mid \varphi_i) \ll \sigma^2, \forall i$, and so $w_i \approx 1/\sigma^2$, which is a constant (since σ^2 is a constant). Thus, for this example, the weighting does not change the naïve estimate.

```
normal CJS example -- VC analysis

Beta-hat SE(Beta-hat)
-----
0.709537 0.016171

S-hat SE(S-hat) S-tilde SE(S-tilde) RMSE(S-tilde)
-----
0.790686 0.053375 0.771108 0.041532 0.045915
0.659116 0.041118 0.672438 0.033828 0.036357
0.749206 0.039962 0.746560 0.033366 0.033471
<snipped to save space>
0.596808 0.034750 0.612847 0.029762 0.033809
0.734009 0.048549 0.719898 0.038302 0.040819
0.618979 0.044732 0.634432 0.036430 0.039572

Naive estimate of sigma^2 = 0.0045517 with 95% CI (0.0018018 to 0.0121649)
Estimate of sigma^2 = 0.0045660 with 95% CI (0.0019542 to 0.0120774)
Estimate of sigma = 0.0675725 with 95% CI (0.0442060 to 0.1098971)
```

You may have noticed that the ML estimates ('S-hat') are very close to what we identified earlier as 'shrinkage' estimates ('S-tilde'). Is the near-equivalence of the 'naïve' and 'weighted' estimates for σ^2 related to the 'closeness' of the ML and 'shrinkage' estimates?

D.3. Random effects models and shrinkage estimates

In this section, we introduce what we will refer to as 'random effects' models. We'll begin by having another look at the results from the simple binomial example (section D.2.1):

```
Beta-hat SE(Beta-hat)
-----
0.482526 0.033946

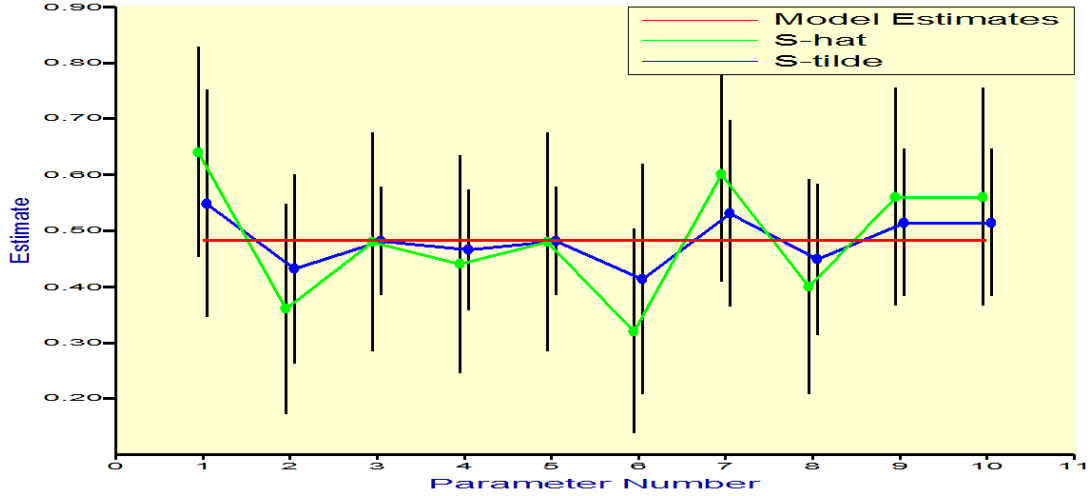
S-hat SE(S-hat) S-tilde SE(S-tilde) RMSE(S-tilde)
-----
0.640000 0.096000 0.548337 0.048955 0.103917
0.360000 0.096000 0.431320 0.048955 0.086505
0.480000 0.099920 0.481505 0.049351 0.049374
0.440000 0.099277 0.465242 0.049289 0.055377
0.480000 0.099920 0.481505 0.049351 0.049374
0.320000 0.093295 0.412990 0.048656 0.104951
0.600000 0.097980 0.530797 0.049160 0.084887
0.400000 0.097980 0.448615 0.049160 0.069139
0.560000 0.099277 0.514014 0.049289 0.067410
0.560000 0.099277 0.514014 0.049289 0.067410

Naive estimate of sigma^2 = 0.0015950 with 95% CI (-0.0042991 to 0.0277050)
Estimate of sigma^2 = 0.0019503 with 95% CI (-0.0039312 to 0.0280522)
Estimate of sigma = 0.0441616 with 95% CI (0.0000000 to 0.1674878)
Trace of G matrix = 4.7017092
```

From left to right are the ML estimates, \hat{S}_i ('S-hat'), the estimated standard error for the ML estimate, $\widehat{SE}(\hat{S}_i \mid S)$ ('SE(S-hat)'), the corresponding 'shrinkage' estimate, \tilde{S}_i ('S-tilde'), the estimated standard

error for the shrinkage estimate, $\widehat{SE}(\tilde{S}_i | \hat{S}_i)$, and the estimated residual mean-squared error (RMSE) for the shrinkage estimate, $\widehat{RMSE}(\tilde{S}_i | \hat{S}_i)$ ('RMSE(S-tilde)').

Here is a plot of the ML estimates, \hat{S}_i (green line), the 'shrinkage' estimates, \tilde{S}_i (blue line), and the model estimates (for the mean model, corresponding to the estimated mean $\hat{\beta} = 0.4825$; red line).



We are familiar with the 'ideas' behind the ML estimates, \hat{S}_i , and the idea of an overall estimate of the mean survival, $\hat{E}(S)$, hopefully makes some intuitive sense. But, what are 'shrinkage' estimates? We'll start with a short-form explanation, focussing on the basic ideas, then jump down into the weeds a bit for a deeper (more technical) discussion. The concept of a 'shrinkage' estimate is perhaps not the easiest thing to understand.* We will follow this by illustrating the mechanics of building and fitting these models in **MARK**, through a series of 'worked examples'.

D.3.1. The basic ideas...

Looking at the tabular output and the plot, there are notable differences in point estimates, and precision, between the ML estimates and the shrinkage estimates. If you look carefully, you'll notice that for most years, the shrinkage estimate falls somewhere between the ML estimate, and the mean. The shrinkage method is so called because each residual arising from the fitted reduced-parameter model ($\hat{S}_i - \hat{E}(S)$) is 'shrunk', then added back to the estimated model structure for observation i under that reduced model. In a heuristic sense, the \tilde{S}_i are derived from the ML estimates by removal of the sampling variation.

When sampling covariances are zero[†], the shrinkage estimator used in **MARK** for the mean-only model (although the structure applies generally) is

$$\tilde{S}_i = \hat{E}(S) + \sqrt{\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{E}_S[\widehat{\text{var}}(\hat{S}_i | S_i)]}} \times [\hat{S}_i - \hat{E}(S)].$$

* In his definitive text on matrix population models, Hal Caswell comments that understanding eigenvalues and eigenvectors (which feature prominently in demographic analysis) requires 'not only a mechanical understanding, but a real intuitive grasp of the slippery little suckers...' (p. 662, 2nd edition). We submit the same sentiment applies to 'shrinkage' estimates.

[†] The full shrinkage estimator, accounting for non-zero sampling covariances, is presented in section D.3.2.

The first term on the right-hand side (RHS) of the expression is the estimate of the mean survival, $\hat{E}(S)$. The last term, $[\hat{S}_i - \hat{E}(S)]$, is simply the residual of the ML estimate from the model. But what about the middle term on the RHS?

We will generally refer to

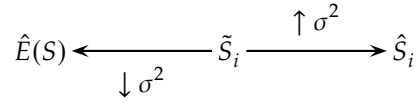
$$\sqrt{\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{E}_S[\widehat{\text{var}}(\hat{S}_i | S_i)]}},$$

as the ‘shrinkage coefficient’, and it clearly is a function of the proportion of total variance – i.e., the sum $\sigma^2 + \text{var}(\hat{S}_i | S_i)$ – due to process (environmental) variation, σ^2 . The square-root is used because then

$$\hat{\sigma}^2 \doteq \frac{\sum^k (\tilde{S}_i - \bar{\tilde{S}})^2}{k-1} \quad \text{and} \quad \hat{E}(S) \doteq \bar{\tilde{S}}.$$

If there is no process variation (i.e., $\sigma^2 = 0$), then the shrinkage coefficient is evaluated at 0, and thus the shrinkage estimate would be the mean, $\hat{E}(S)$. In other words, if you have no environmental variation, then the only variation in the system is sampling. If you remove the sampling variation (which we noted earlier is what shrinkage is doing, at least heuristically), then this makes sense – without process or sampling variation, every shrinkage estimate would simply be the mean, i.e., $\tilde{S}_i = \hat{E}(S)$. In contrast, with increasing process variance (i.e., increasing σ^2), we see that as $\sigma^2 > \text{var}(\hat{S}_i | S_i)$ (i.e., as the proportion of total variance due to process variance increases), then the shrinkage coefficient approaches 1.0, and the shrinkage estimate would approach the ML estimate, \hat{S}_i .

This relationship is depicted in the following diagram where the arrows indicate the direction that decreasing or increasing process variance σ^2 has on the value of the shrinkage estimate, \tilde{S}_i , relative to the arithmetic average of the ML estimate \hat{S}_i and the mean $\hat{E}(S)$.



Thus, another way of looking at it is to view the shrinkage estimate is analogous to an ‘average’ between the two estimates. This is important when we consider model averaging – we defer discussion of that important topic until section D.5.

We should note that a shrinkage coefficient different than

$$\sqrt{\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{E}_S[\widehat{\text{var}}(\hat{S}_i | S_i)]}}$$

could be used. However, this particular shrinkage coefficient has a very desirable property: if we treat the \tilde{S}_i as if they were a random sample, then their sample variance almost exactly equals $\hat{\sigma}^2$. This also means that a plot of the shrinkage residuals (as implicit in the plot at the top of the preceding page) gives a correct visual image of the process variation in the S_i .

This starts to point us in the direction of answering the basic question ‘why derive shrinkage estimators for the S_i ?’. The answer in part comes from the observation we just made that the sample variance of the shrinkage estimates is very close to the estimate of process variance, $\hat{\sigma}^2$. So, the shrinkage estimates are ‘improved’, relative to the ML estimates, since they should have better precision. If we look at the estimates for the binomial survival example, we see that the improvements gained by the

shrinkage estimators, \tilde{S}_i , appears substantial – they have about 50% better precision (simply compare $\widehat{SE}(\tilde{S}_i | S_i)$ to $\widehat{SE}(\hat{S}_i | S_i)$).

However, because the ML estimates are unbiased, and the shrinkage estimators are biased (as we will explain), a necessary basis for a fair comparison is the sum of squared errors (SSE). The SSE is a natural measure of the closeness of a set of estimates to the set of S_i . For example, for the binomial survival example, the SSE for the ML estimates is

$$SSE_{MLE} = \sum_{i=1}^{10} (\hat{S}_i - S_i)^2 = 0.067$$

while for the shrinkage estimates, the SSE is

$$SSE_{shrinkage} = \sum_{i=1}^{10} (\tilde{S}_i - S_i)^2 = 0.019$$

Clearly, in this sample the shrinkage estimates, as a set, are closer to truth. The expected SSE is the mean square error, MSE ($= E[SSE]$), which is a measure of average estimator performance.

Those of you with some background in statistical theory might see the connections between the preceding and James-Stein estimation, wherein (in highly simplified form) when 3 or more parameters are estimated simultaneously, there exist combined estimators more accurate on average (that is, having lower expected MSE) than any method that considers the parameters separately. For example, let θ is a vector consisting of $n \geq 3$ unknown parameters. To estimate these parameters, we take a single measurement X_i for each parameter θ_i , resulting in a vector \mathbf{X} of length n . Suppose the measurements are independent, Gaussian random variables, such that $\mathbf{X} \sim \mathcal{N}(\mu, 1)$. The most obvious approach to parameter estimation would be to use each measurement as an estimate of its corresponding parameter: $\hat{\theta} = \mathbf{X}$. James-Stein demonstrated that this standard (LS) estimator is suboptimal in terms of mean squared error, $E(\theta - \hat{\theta})$. In other words, there exist alternative estimators which always achieve lower mean squared error, no matter what the value of θ is. For example, it can be shown that a combined estimator of the sample and global mean is a better predictor of the future than is the individual sample mean, since the total MSE of the combined estimator is lower than if using the sample means themselves. This clearly points to part of the theory underlying the use of shrinkage estimators – James-Stein says that a combined estimator (say, of the ML estimate and the random mean, which is of course our shrinkage estimate) will have a lower MSE than will the ML estimates themselves (with the degree of improvement increasing with increasing number of sample means being combined).

It is important to note, however, that the combined estimator will be closer to optimal *overall*, since it minimizes the MSE of the estimates overall. However, it is possible that any one individual estimate could be ‘incorrectly shrunk’ (relative to the true value of the parameter), even in the wrong direction. So, shrinkage is conceptually optimal for the set of parameters, but not necessarily for any individual parameter).

If these ideas are new to you, papers by Efron & Morris (1975, 1977) are quite accessible, and provide excellent introductions to the subject.

begin sidebar

Shrinkage estimators and 95% confidence limits

For the ML estimates, an approximate 95% confidence interval on S_i is given by $\hat{S}_i \pm 2\widehat{SE}(\hat{S}_i | S_i)$. This procedure will have good coverage in this example. However, for the shrinkage estimator if we use $\tilde{S}_i \pm 2\widehat{SE}(\tilde{S}_i | S_i)$, coverage will be negatively affected by the bias of \tilde{S} . Theory (discussed in B&W) shows that the correct expected coverage occurs for the interval $\tilde{S}_i \pm 2\widehat{RSME}(\tilde{S}_i | S_i)$, where

$$\widehat{\text{RMSE}}(\tilde{S}_i | S_i) = \sqrt{\widehat{\text{var}}(\tilde{S}_i | S_i) + (\tilde{S}_i - \hat{S}_i)^2}.$$

The expectation over S_i of $\widehat{\text{MSE}}(\tilde{S}_i | S_i) = [\widehat{\text{RMSE}}(\tilde{S}_i | S_i)]^2$ is approximately the mean square error for \tilde{S}_i , $\text{MSE}_{\tilde{S}_i}$. For the ML estimates,

$$\widehat{\text{RMSE}}(\hat{S}_i | S_i) = \widehat{\text{SE}}(\hat{S}_i | S_i),$$

because \hat{S}_i is unbiased. The unbiasedness of the ML estimates in the general model, together with a high correlation between \hat{S}_i and \tilde{S}_i , and the assumption that S_i are random, allows an argument that $\widehat{\text{RMSE}}(\tilde{S}_i | S_i)$ is an estimator of the unconditional sampling standard error of \tilde{S}_i (over conceptual repetitions of the data). It then follows that this RMSE can be a correct basis for a reliable confidence interval. It is rare to have a reliable estimator of the MSE for a biased estimator, but when this occurs it makes sense to use $\pm 2 \sqrt{\widehat{\text{MSE}}}$ rather than $\pm 2 \widehat{\text{SE}}$, as the basis for a 95% CI.

end sidebar

D.3.2. Some technical background...

Most random effects theory assumes conditional independence of the estimators. In the introduction to this section, we started by having another look at the simple binomial survival example introduced earlier in section D.2.1. In that example, there was no sampling covariance – the estimates were all independent of each other.

However, more generally in capture-recapture studies, the estimators $\hat{S}_1, \dots, \hat{S}_k$ are pairwise conditionally correlated. Thus, a more general, extended theory is required, which we develop here in summary form – complete details are found in B&W. While some of the math can get a bit ugly, some familiarity with the ideas (at least) is helpful in more fully understanding ‘what **MARK** is doing’. In the following, vectors (all column) are underlined. Matrices are in bold font. A matrix, \mathbf{X} , may be a vector if it has only a single column. In that case, we do not underline \mathbf{X} .

First, we assume $\hat{\underline{S}} = \underline{S} + \underline{\delta}$, given \underline{S} . Conditional on $\underline{S}, \underline{\delta}$ (which has a zero expectation) has a variance-covariance matrix \mathbf{W} , and $E(\hat{\underline{S}} | \underline{S}) = \underline{S}$ for large samples. Second, unconditionally \underline{S} is a random vector with expectation $\mathbf{X}\underline{\beta}$ and variance-covariance matrix $\sigma^2 \mathbf{I}$, where \mathbf{I} is the identity matrix. (Note: the vector $\underline{\beta}$ is different than the beta parameters of the **MARK** link function). Thus, the process errors $\epsilon_i = S_i - E(S_i)$ are independent with homogeneous variance σ^2 . Also, we assume mutual independence of sampling errors δ and process errors ϵ . We fit a model that does not constrain S , e.g., $\{S_i\}$, and hence get the maximum likelihood estimates $\hat{\underline{S}}$ and an estimate of \mathbf{W} .

Let \underline{S} be a vector with n elements, and $\underline{\beta}$ have k elements. Unconditionally,

$$\begin{aligned} \hat{\underline{S}} &= \mathbf{X}\underline{\beta} + \underline{\delta} + \underline{\epsilon}, \\ \text{VC}(\underline{\delta} + \underline{\epsilon}) &= \mathbf{D} = \sigma^2 \mathbf{I} + \mathbf{W}. \end{aligned}$$

We want to estimate $\underline{\beta}$ and σ^2 , an unconditional variance-covariance matrix for $\hat{\underline{\beta}}$, a confidence interval on σ^2 , and to compute a shrinkage estimator of S (i.e., \tilde{S}) and its conditional sampling variance-covariance matrix. In this random effects context the maximum likelihood estimator is the best conditional estimator of \underline{S} .

However, once we add the random effects structure we can consider an unconditional estimator of \underline{S} (\tilde{S}) and a corresponding unconditional variance-covariance for \tilde{S} , which incorporates σ^2 as well as \mathbf{W} and has $(n - k)$ degrees of freedom (if we are assuming large df for \mathbf{W} and ‘large’ σ^2).

For a given σ^2 we have

$$\underline{\hat{\beta}} = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}^{-1}\underline{\hat{S}}.$$

Note that here \mathbf{D} , hence $\underline{\hat{\beta}}$, is a function of σ^2 . Now we need a criterion that allows us to find an estimator of σ^2 . Assuming normality of $\underline{\hat{S}}$ (approximate normality usually suffices), then the weighted residual sum of squares $(\underline{\hat{S}} - \mathbf{X}\underline{\hat{\beta}})' \mathbf{D}^{-1} (\underline{\hat{S}} - \mathbf{X}\underline{\hat{\beta}})$ is central χ^2 -distributed on $(n - k)$ degrees of freedom. Hence, by the method of moments,*

$$n - k = (\underline{\hat{S}} - \mathbf{X}\underline{\hat{\beta}})' \mathbf{D}^{-1} (\underline{\hat{S}} - \mathbf{X}\underline{\hat{\beta}}).$$

This equation defines a 1-dimensional numerical-solution search problem. Pick an initial (starting) value of σ^2 , compute \mathbf{D} , then compute $\underline{\hat{\beta}}$, then compute the right-hand side of the preceding expression. This process is repeated over values of σ^2 until the solution, as $\underline{\hat{S}}$, is found. This process also gives $\underline{\hat{\beta}}$. The unconditional variance-covariance matrix of $\underline{\hat{\beta}}$ is $\text{VC}(\underline{\hat{\beta}}) = (\mathbf{X}'\mathbf{D}^{-1}\mathbf{X})^{-1}$.

Now we define another matrix as

$$\begin{aligned} \mathbf{H} &= \sigma \sqrt{\mathbf{D}} \\ &= \sigma \sqrt{\sigma^2 \mathbf{I} + E(\mathbf{W})} \\ &= \sqrt{\mathbf{I} + \frac{1}{\sigma^2} E(\mathbf{W})}. \end{aligned}$$

(Here we only need \mathbf{H} at $\hat{\sigma}$.)

The recommended shrinkage estimate (which is what is used in **MARK**) is

$$\begin{aligned} \underline{\tilde{S}} &= \mathbf{H}(\underline{\hat{S}} - \mathbf{X}\underline{\hat{\beta}}) + \mathbf{X}\underline{\hat{\beta}} \\ &= \mathbf{H}\underline{\hat{S}} + (\mathbf{I} - \mathbf{H})\mathbf{X}\underline{\hat{\beta}}. \end{aligned}$$

To get an estimator of the conditional variance of these shrinkage estimators (which is not exact as the estimation of σ^2 is ignored here, as it is for the variance-covariance matrix of $\underline{\hat{\beta}}$), we define and compute a projection matrix \mathbf{G} as follows:

$$\mathbf{G} = \mathbf{H} + (\mathbf{I} - \mathbf{H})\mathbf{A}\mathbf{D}^{-1}.$$

Hence,

$$\underline{\tilde{S}} = \mathbf{G}\underline{\hat{S}}.$$

In other words, \mathbf{G} is the projection matrix which ‘maps’ the vector of ML estimates to the vector of shrinkage estimates.

The conditional variance-covariance matrix of the shrinkage estimator is then $\text{VC}(\underline{\tilde{S}} \mid \underline{\hat{S}}) = \mathbf{G}\mathbf{W}\mathbf{G}'$, whereas $\mathbf{W} = \text{VC}(\underline{\hat{S}} \mid \underline{\hat{S}})$. Because $\underline{\tilde{S}}$ is known to be biased, and because the direction of the bias is known, an improved basis for inference is $\text{VC}(\underline{\tilde{S}} \mid \underline{\hat{S}}) = \mathbf{G}\mathbf{W}\mathbf{G}' + (\underline{\tilde{S}} - \underline{\hat{S}})(\underline{\tilde{S}} - \underline{\hat{S}})'$. The square-roots of the

* Which is why the random effects estimation procedure in **MARK** is sometimes referred to as ‘the moments estimator’.

diagonal elements of this matrix are

$$\widehat{\text{RMSE}}(\tilde{S}_i \mid \underline{S}) = \sqrt{\widehat{\text{var}}(\tilde{S}_i \mid \underline{S}) + (\tilde{S}_i - \hat{S}_i)^2}.$$

As discussed earlier, confidence intervals should be based on this RMSE. The RMSE can exceed the $\text{SE}(\hat{S}_i \mid S_i)$, but on average, the RMSE is smaller.

D.3.3. Deriving an AIC for the random effects model

We will have started with a likelihood for a model at least as general as full time variation on all the parameters, say $\mathcal{L}(\underline{S}, \underline{\theta}) = \mathcal{L}(S_1, \dots, S_k, \theta_1, \dots, \theta_\ell)$. Under this time-specific model, $\{S_t, \theta_t\}$, we have the MLEs, $\hat{\underline{S}}$ and $\hat{\underline{\theta}}$, and the maximized log-likelihood, $\log \mathcal{L}(\hat{\underline{S}}, \hat{\underline{\theta}})$ based on $K = k + \ell$. Thus (for large sample size, n), AIC for the time-specific model is the (now) familiar $-2 \log \mathcal{L}(\hat{\underline{S}}, \hat{\underline{\theta}}) + 2K$.

The dimension of the parameter space to associate with this random effects model is K_{re} ,

$$K_{re} = \text{tr}(\mathbf{G}) + \ell,$$

where \mathbf{G} is the projection matrix mapping $\hat{\underline{S}}$ onto $\tilde{\underline{S}}$ (see above), and ℓ is the number of free parameters not being modeled as random effects. $\text{tr}(\mathbf{G})$ is the matrix trace (i.e., the sum of the diagonal elements of \mathbf{G}). The $\text{tr}(\mathbf{G})$ (and thus K_{re}) is generally not integer.*

Note that the mapping of $\tilde{\underline{S}} = \mathbf{G}\hat{\underline{S}}$ is a type of generalized smoothing. It is known that the effective number of parameters to associate with such smoothing is the trace of the smoother matrix.

Finally, the large-sample AIC for the random effects model is

$$\text{AIC} = -2 \log \mathcal{L}(\tilde{\underline{S}}, \tilde{\underline{\theta}}) + 2K_{re}.$$

A more exact version, AIC_c , for the random effects model may, by analogy, be taken as

$$\text{AIC}_c = -2 \log \mathcal{L}(\tilde{\underline{S}}, \tilde{\underline{\theta}}) + 2K_{re} + 2 \left(\frac{K_{re}(K_{re} + 1)}{n + K_{re} - 1} \right).$$

For a full derivation of the AIC, in both a fixed and random effects context, see Burnham & Anderson (2002).

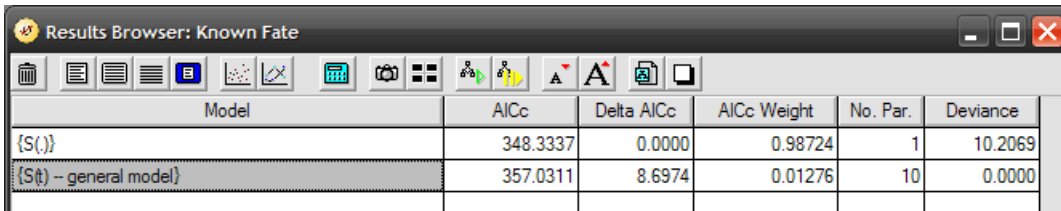
D.4. Random effects models – some worked examples

In the following, we introduce the ‘mechanics’ of fitting random effects models in **MARK**, using a series of progressively more complex examples. Many of the steps were introduced earlier in the context of variance components analysis (section D.2). We begin by revisiting the simple binomial (‘known fate’) analysis we introduced in section D.2.1.

* Which is why the number of parameters reported for random effects models is generally not integer.

D.4.1. Binomial survival revisited – basic mechanics

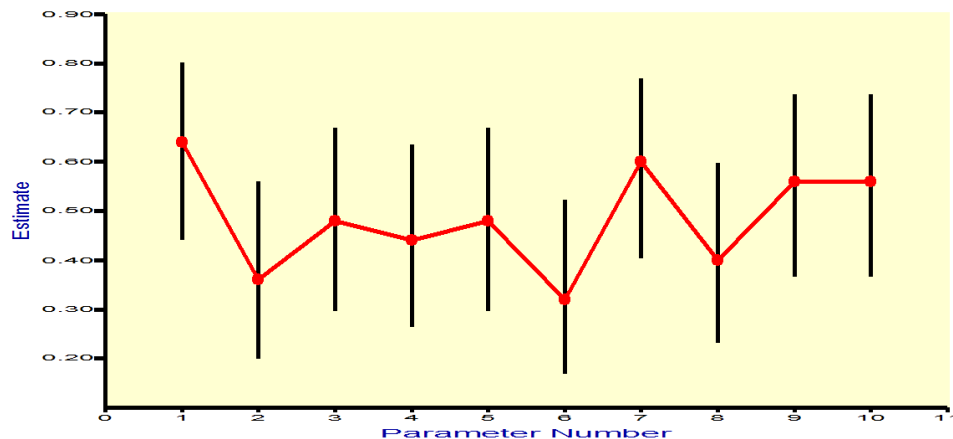
We begin our analysis of the ‘known fate’ data by considering a candidate set of 2 approximating models: model $\{S_t\}$, and model $\{S_{\cdot}\}$. Recall that the former model is our general ‘time-dependent’ model – this is the model we used in our estimation of variance components as detailed in section D.2.1. The second model, $\{S_{\cdot}\}$ is a model where survival S is constrained to be constant over time. Fit both models, and add the results to the browser:



Model	AICc	Delta AICc	AICc Weight	No. Par.	Deviance
$\{S_{\cdot}\}$	348.3337	0.0000	0.98724	1	10.2069
$\{S(t) - \text{general model}\}$	357.0311	8.6974	0.01276	10	0.0000

Clearly, there is overwhelming support in the data for the time-constant model, $\{S_{\cdot}\}$.

Let’s have a look at a plot of the estimates from model $\{S_t\}$:

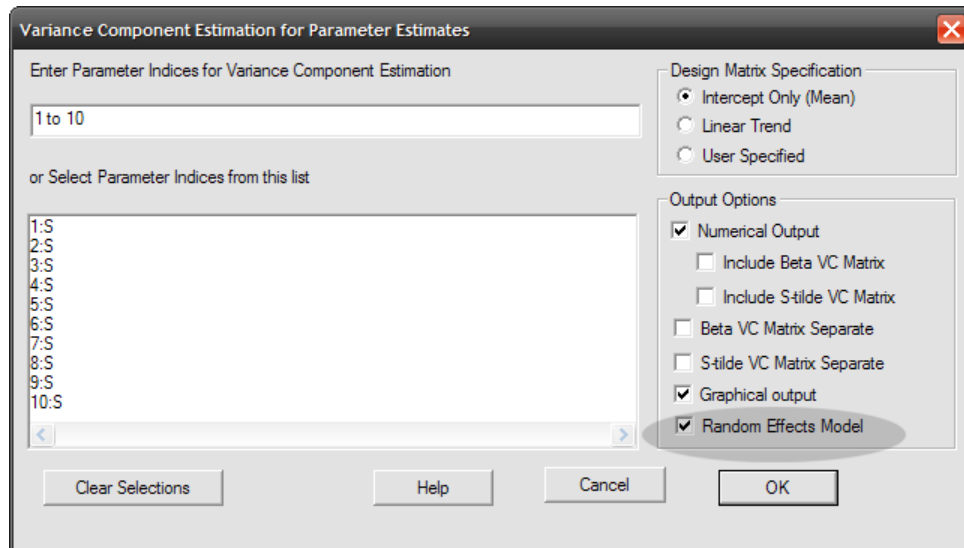


The point estimates \hat{S}_i appear to vary over time, but notice that the confidence bounds on each estimate appear to overlap over all intervals. This is perhaps the underlying explanation for why, despite apparent variation in the point estimates \hat{S}_i , there is essentially no support in the data for the general time-varying model, $\{S_t\}$, relative to a model which constrains the estimates to be constant over time, $\{S_{\cdot}\}$.

However, we know logically that S_i cannot truly be constant over time. There must be *some* true variation in survival, but our data are insufficient (apparently) to support a time-varying model where ‘time’ is modeled as an unconstrained fixed effect. Rather than concluding our analysis of these data here (especially when such a conclusion is based on data insufficiency, rather than biological plausibility), we continue by fitting a random effects model, which we will propose as ‘intermediate’ between constant models, and fully time-dependent models. We will submit at this point that a random effects model is, in fact, such an intermediate model (support for this statement will be developed in a later section). Here, we will try to fit a model where survival varies around some unknown mean μ , with unknown process variance σ^2 – clearly, this corresponds to the ‘intercept only (mean)’ model we first saw in section D.2.1 when we introduced variance components estimation in **MARK**. Such a model also makes

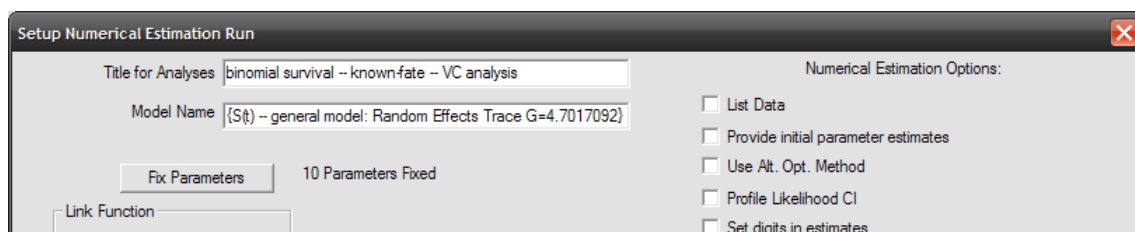
some intuitive sense, if our intuition is guided by the time-series plot of the estimate \hat{S}_i shown on the previous page, where it might be reasonable to ‘imagine’ each S_i as ‘bouncing randomly’ around some mean survival probability, μ (with the magnitude of the ‘bouncing’ around the mean being determined by the process variance, σ^2).

Formally, we will refer to this model as $\{S_{\mu, \sigma^2}\}$. How do we build such a model in **MARK**? Easy enough – we simply run through the steps we need for a variance components analysis for model $\{S_i\}$ (which should be familiar to you by now, if you’ve reached this point in this appendix – if not, go back and review section D.2.1): retrieve our general model $\{S_t\}$, specify parameter indices ‘1 to 10’ in the variance component estimation window, and now, before hitting the ‘OK’ button, we ‘check’ the box to build the random effects model in the lower right-hand corner of the ‘**Variance Component Estimation**’ window:



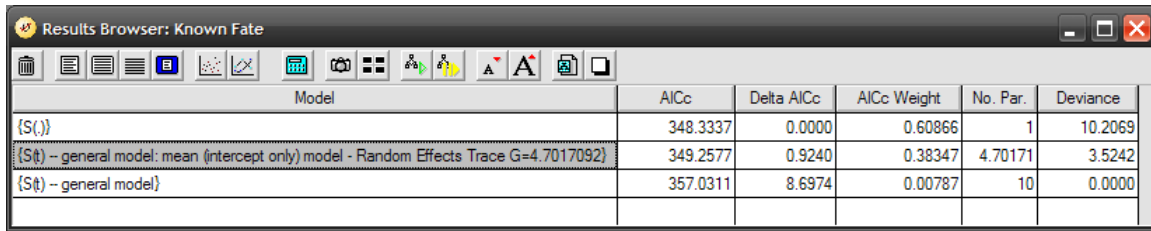
Now, click the ‘OK’ button. As before, **MARK** will respond by outputting various parameters estimates ($\hat{S}_i, \tilde{S}_i, \hat{\sigma}^2$) to the editor, and generating a plot of the time-series for the estimates of survival, and the mean model. The derivation and analysis of both was presented in some detail at the beginning of section D.3 and in section D.3.1, so we won’t repeat them here.

Here, we introduce the next step, which is the actual ‘fitting’ of the estimated random effects model to the data. You may have noticed that in addition to the editor window (containing the parameter estimates) and the plot, **MARK** has also brought up a ‘**Setup Numerical Estimation Run**’ window, the general contents of which will be familiar to you from other analysis you’ve done with **MARK**.



The biggest ‘visual’ difference is that **MARK** has modified the model name. Now, the model is called

'{S(t)f(t) -- sine link: Random Effects Trace G=4.7017092}'. The part of the model name to the left of the colon is what we originally used to name the model. The part to the right (which **MARK** has added) indicates that we're now running a random effects model, and that the 'trace' of the **G** matrix is 4.7017092. Recall from section D.3.2 that the trace of the **G** matrix is related to the number of estimated parameters used in the derivation of the AIC (and that because $\text{tr}(\mathbf{G})$ is generally non-integer, that the number of estimated parameters for random effects models is also usually non-integer). We'll modify the title by adding the words 'intercept only (mean)' somewhere in the title box, to indicate that the model we're fitting is the 'intercept only (mean)' model. Once done, click the 'OK to run' button and add the results to the browser (if you get a warning about **MARK** not being able to import the variance-covariance matrix, ignore it).



Model	AICc	Delta AICc	AICc Weight	No. Par.	Deviance
{S()}	348.3337	0.0000	0.60866	1	10.2069
{S(t) -- general model: mean (intercept only) model - Random Effects Trace G=4.7017092}	349.2577	0.9240	0.38347	4.70171	3.5242
{S(t) -- general model}	357.0311	8.6974	0.00787	10	0.0000

Several things to note here. First, our random effects model now has some significant support in the data (AIC_c weight is 0.383). While not the most parsimonious model in the candidate set, it is clearly better supported than the fixed effect time-dependent model. However, given that a time-invariant model is not logically plausible, then we should select a model where survival varies over time. If we make such a logical choice, then (based on the ideas present in section D.3) our best estimate for annual survival S_i would be the shrinkage estimates \tilde{S}_i from our random effects model, $\{S_{\mu, \sigma^2}\}$.

Second, look at the number of parameters that **MARK** reports as having been estimated for this model (4.70171). We see that this number is identical to $\text{tr}(\mathbf{G})$. Recall from section D.3.3 that the dimension of the parameter space (analogous to the number of estimated parameters in the usual sense) to associate with a random effects model is K_{re} ,

$$K_{re} = \text{tr}(\mathbf{G}) + \ell,$$

where $\text{tr}(\mathbf{G})$ is the trace of the **G** matrix, and ℓ is the number of free parameters not being modeled as a random effect. In this case, all 10 parameters in the model, S_1, \dots, S_{10} are being modeled as a random effect, so $\ell = 0$, and thus $K_{re} = \text{tr}(\mathbf{G}) = 4.70171$.

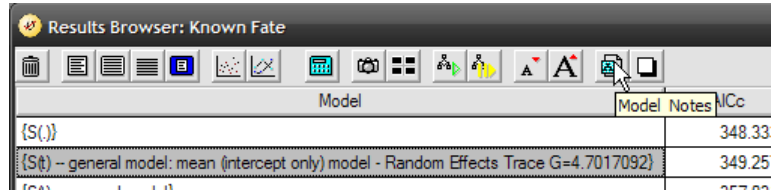
If we next look at the estimates of S_i (below)

binomial survival -- known-fate -- VC analysis					
Real Function Parameters of {S(g) -- general model: (intercept only) Random Effects Trace G=4.7017092}					
Parameter	Estimate	Standard Error	95% Confidence Interval		
			Lower	Upper	
1:S	0.5483372	0.0000000	0.5483372	0.5483372	Fixed
2:S	0.4313200	0.0000000	0.4313200	0.4313200	Fixed
3:S	0.4815048	0.0000000	0.4815048	0.4815048	Fixed
4:S	0.4652419	0.0000000	0.4652419	0.4652419	Fixed
5:S	0.4815048	0.0000000	0.4815048	0.4815048	Fixed
6:S	0.4129904	0.0000000	0.4129904	0.4129904	Fixed
7:S	0.5307974	0.0000000	0.5307974	0.5307974	Fixed
8:S	0.4486149	0.0000000	0.4486149	0.4486149	Fixed
9:S	0.5140139	0.0000000	0.5140139	0.5140139	Fixed
10:S	0.5140139	0.0000000	0.5140139	0.5140139	Fixed

we see that all of the estimates are 'fixed' at the value of \tilde{S}_i . As fixed parameters in the fitted model,

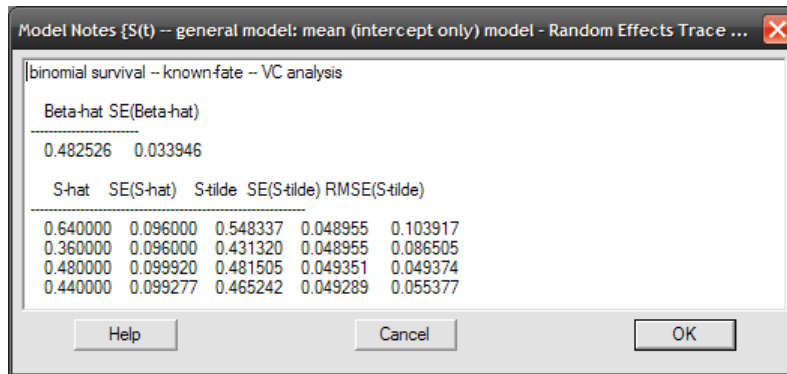
there is no standard error (or CI) estimated (since there is nothing to be estimated for a fixed parameter, obviously).

Finally, if you click on the ‘model notes’ button in the browser toolbar



Model	Model Notes	NCc
{S(.)}		348.33
{S(t) -- general model: mean (intercept only) model - Random Effects Trace G=4.7017092}		349.25

you will be presented with a ‘copy’ of the variance components analysis which was first output to the editor.



binomial survival -- known-fate -- VC analysis

Beta-hat	SE(Beta-hat)
0.482526	0.033946

S-hat	SE(S-hat)	S-tilde	SE(S-tilde)	RMSE(S-tilde)
0.640000	0.096000	0.548337	0.048955	0.103917
0.360000	0.096000	0.431320	0.048955	0.086505
0.480000	0.099920	0.481505	0.049351	0.049374
0.440000	0.099277	0.465242	0.049289	0.055377

This is convenient, since it allows you to ‘store’ the variance components analysis for any particular random effects model you fit to the data (note that the variance components analysis is output to the ‘model notes’ only if you run the random effects model).

D.4.2. A more complex example – California mallard recovery data

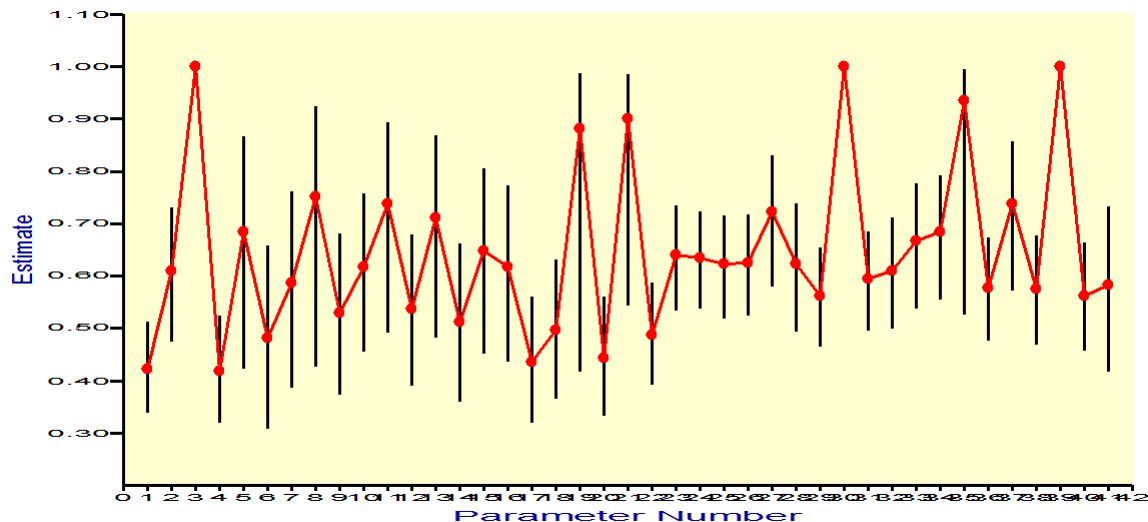
Here we introduce the mechanics, and some of the challenges, of fitting of ‘random effects’ models in **MARK**. We will use a long-term dead recovery data set based on late summer banding of adult male mallards (*Anas platyrhynchos*), banded in California every year from 1955 to 1996 ($k = 41$) (Franklin *et al.*, 2002). The total number of birds banded (marked and released alive) was 42,015, with a total of 7,647 dead recoveries. In a preliminary analysis, the variance inflation factor was estimated as $\hat{c} = 1.1952$. The recovery data are contained in **california-male-mallard.inp**. For your convenience, we’ve also generated 3 candidate models: $\{S_t f_t\}$, $\{S_T f_t\}$, and $\{S_t f_t\}$, where the capital ‘T’ subscript is used to indicate linear trend. These models are contained in the associated .dbf and .fpt files.

Note that we use time-structure for the recovery parameter f . We do so not simply because such a model often makes more ‘biological sense’ than a model where f is constrained (say, f_t), but because any constraint applied to f will impart (or ‘transfer’) more of the variation in the data to the survival parameter S , such that the estimated process variance $\hat{\sigma}^2$ will be ‘inflated’, relative to the true process variance. In general, you want to estimate variance components using a fully time-dependent model, for all parameters, even if such a model is not the most parsimonious given the data.

Here are the results of fitting these 3 models to the data:

Model	QAICc	Delta QAICc	QAICc Weight	Model Likelihood	No. Par.	QDeviance
$\{S(t)f(t) - \sin \text{ link}\}$	54716.5936	0.0000	0.98132	1.0000	83	368.4517
$\{S(T)f(t) - \text{logit link}\}$	54724.5163	7.9227	0.01868	0.0190	44	454.6161
$\{S(f(t)) - \sin \text{ link}\}$	54803.8298	87.2362	0.00000	0.0000	43	535.9342

Based on the relative degree of support in the data it would seem that there is essentially no support for a model where survival is constrained to follow a linear trend, or for a model where survival is constrained to be constant over time. All of the support in the data (among these 3 models) is for model $\{S_t f_t\}$. If this was all we did, we'd come to the relatively uninteresting and uninformative conclusion that there is temporal variation in survival. A plot of the estimates from this model seems to be consistent with this conclusion:



However, rather than concluding our analysis of these data here, or perhaps add some models where annual variation is modeled using a fixed effects approach where annual estimates are constrained to be linear functions of one or more covariates, we might consider models which are 'intermediate' between constant models, and fully time-dependent models. We will submit that a random effects model is, in fact, such an intermediate model.

Let's try to fit a model where survival varies around some unknown mean μ , with unknown process variance σ^2 – clearly, this corresponds to the '**intercept only (mean)**' model. As was the case in our first example involving binomial 'known fate' survival, we will refer to this model as $\{S_{\mu, \sigma^2} f_t\}$. Go ahead and set up this model, first making sure that the fully time-dependent model is the 'active' model (by retrieving it). Specify parameter indices '1 to 41' for the variance component estimation, make sure '**intercept only (mean)**' is selected, and that the '**random effects model**' button is checked (as shown at the top of the next page).

Variance Component Estimation for Parameter Estimates

Enter Parameter Indices for Variance Component Estimation

1 to 41

or Select Parameter Indices from this list

1:S	11:S	21:S	31:S
2:S	12:S	22:S	32:S
3:S	13:S	23:S	33:S
4:S	14:S	24:S	34:S
5:S	15:S	25:S	35:S
6:S	16:S	26:S	36:S
7:S	17:S	27:S	37:S
8:S	18:S	28:S	38:S
9:S	19:S	29:S	39:S
10:S	20:S	30:S	40:S

Design Matrix Specification

☒ Intercept Only (Mean)

☐ Linear Trend

☐ User Specified

Output Options

☒ Numerical Output

☐ Include Beta VC Matrix

☐ Include S-tilde VC Matrix

☐ Beta VC Matrix Separate

☐ S-tilde VC Matrix Separate

☒ Graphical output

☒ Random Effects Model

Clear Selections Help Cancel OK

When you click the 'OK' button, you'll be presented with the estimates of the mean, the ML and 'shrinkage' estimates, and the various estimates of the process variance (to save some space, we've snipped out a number of the estimates for S_i).

California mallard data (males) -- random effects c-hat = 1.1952000

Beta-hat SE(Beta-hat) Label					

	0.629440	0.015657	Intercept		
Par. Num	S-hat	SE(S-hat)	S-tilde	SE(S-tilde)	RMSE(S-tilde)

1	0.423020	0.046257	0.447682	0.040257	0.047211
2	0.689787	0.087061	0.666101	0.060185	0.064678
3	0.666269	0.105860	0.647494	0.064989	0.067647
4	0.526082	0.079841	0.553727	0.055754	0.062231
5	0.685033	0.123080	0.638352	0.068445	0.082849
6	0.481012	0.095575	0.523707	0.058176	0.072161
7	0.586389	0.102964	0.600340	0.062774	0.064306
8	0.751006	0.136897	0.684031	0.071333	0.097847
9	0.529065	0.083367	0.562724	0.056508	0.065774
<<snipped to save space>>					
32	0.610498	0.056343	0.621005	0.045188	0.046393
33	0.667301	0.063895	0.666712	0.049460	0.049464
34	0.685156	0.063149	0.699888	0.048406	0.050598
35	0.935051	0.081332	0.862595	0.058412	0.093069
36	0.577746	0.051772	0.611583	0.041696	0.053699
37	0.738200	0.075641	0.732753	0.054859	0.055129
38	0.846765	0.108590	0.793773	0.066163	0.084768
39	0.535913	0.064761	0.525179	0.047116	0.048323
40	0.675469	0.067890	0.663663	0.052836	0.054139
41	0.583104	0.085443	0.593429	0.061669	0.062527

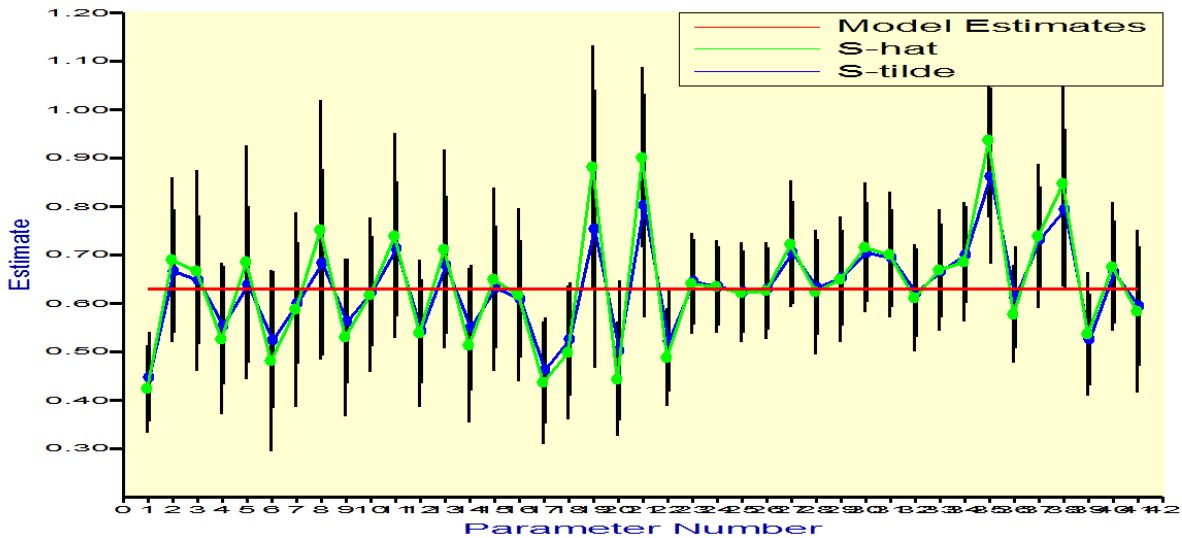
Naive estimate of $\sigma^2 = 0.0079384$ with 95% CI (0.0031332 to 0.0173404)

Estimate of $\sigma^2 = 0.0081075$ with 95% CI (0.0041907 to 0.0166408)

Estimate of $\sigma = 0.0900415$ with 95% CI (0.0647353 to 0.1289993)

Trace of G matrix = 32.0229017

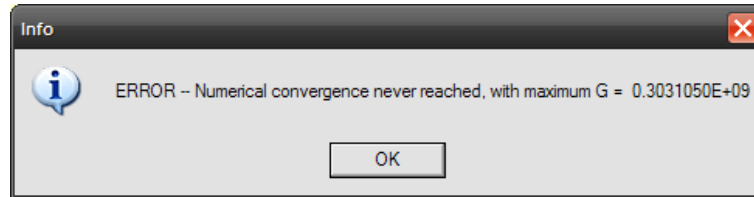
A plot of the ML and shrinkage estimates, and the model from which the shrinkage estimates were derived (in this case, the intercept only mean model), is shown below:



Finally, we come to the estimation run window.

Again, we notice that **MARK** has modified the model name. Now, the model is called ' $\{S(t)f(t) -- \sin \text{ link: Random Effects Trace } G=32.0229017\}$ '. The part of the model name to the left of the colon is what we originally used to name the model. The part to the right (which **MARK** has added) indicates that we're now running a random effects model, and that the 'trace' of the **G** matrix is 32.0229. Again, we're going to modify the title slightly, to indicate that the model we're going to fit is the '**intercept only (mean)**' model – we'll simply add the words 'intercept only' somewhere in the title box.

We hit the ‘**Ok to run**’ button and...

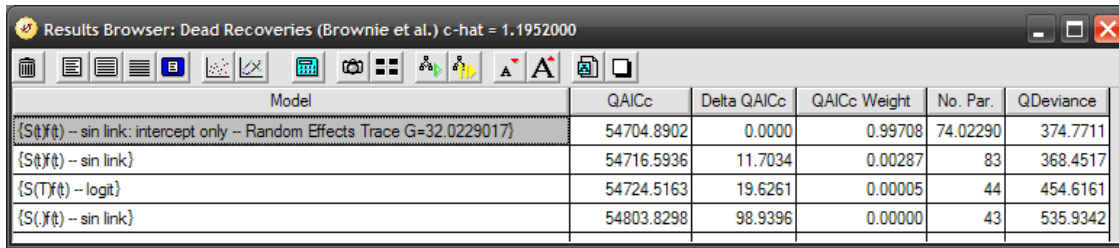


Clearly, something has gone wrong. Generally when you see the phrase ‘numerical convergence never reached’ (or something to that effect) embedded in an error message, your first response should be to consider trying ‘better’ starting values. Often, such convergence issues reflect some underlying ‘problems’ with the data (sparseness, one or more parameters estimated near either 0 or 1), and **MARK** is potentially having difficulty estimating the likelihood – a problem which might be exacerbated (or simply an artifact) of the default starting values used in the numerical optimization.

One straightforward approach is to use different starting values – in this case, the ML estimates from model $\{S_t f_t\}$. To do this, simply check the **‘provide initial parameter estimates’** box in the numerical estimation run window, before running the random effects model. Now, when you click **‘OK to run’**, you will be presented with a window asking you to specify the initial parameter estimates for the numerical estimation. To use the estimates from model $\{S_t f_t\}$, simply click the **‘retrieve’** button, and select the appropriate model (labeled ‘ $S(t)f(t)$ -- sine link’). This will populate the boxes in the **‘initial values’** windows with the ML estimates. Then, once you click the **‘OK’** button, **MARK** will attempt the numerical optimization. For this example, using these different starting values solves the problem – the random effects model converges successfully.

An alternative approach which also generally works (albeit at the expense of some extended computational time in many cases), and which does not require good starting values for the optimization (which you may not always have), is to use simulated annealing for the numerical optimization. You may recall (from Chapter 10) that you can specify using simulated annealing for the optimization by selecting the **‘alternate optimization’** checkbox on the right-hand side of the **‘run numerical estimation’** window. What simulated annealing does during the optimization is to periodically make a random jump to a new parameter value. It is this characteristic is what allows the algorithm more flexibility in finding the global maximum (in cases where there may in fact be local maxima in the likelihood; see Chapter 10 for a discussion of this in the context of multi-state models), and minimizes the chances that the numerical solution is determined by starting values (simulated annealing starts with the defaults, but then makes the random jumps around the parameter space, as described).

To use simulated annealing for our mallard analysis, you simply retrieve model $\{S_t f_t\}$ (our general model), run through the variance components analysis (remembering to check the **‘random effects model’** box), and then try again – this time, before hitting the **‘OK to run’** button for the generated random effects model, make sure the **‘Use Alt. Opt. Method’** button is checked. Change the title (we’ll add **‘intercept only model -- SA’** to indicate both the model, and the optimization method used to maximize the likelihood), then click **‘OK to run’**. Simulated annealing takes significantly longer to converge than does the default optimization routine – how much longer will depend on how fast your computer is. Nonetheless, this approach also works fine, and yields the same model fit as the model fit using different initial values for the optimization. We’ll only keep one of these in the browser (shown at the top of the next page).



Results Browser: Dead Recoveries (Brownie et al.) c-hat = 1.1952000

Model	QAICc	Delta QAICc	QAICc Weight	No. Par.	QDeviance
{S(t)f(t)} -- sin link: intercept only -- Random Effects Trace G=32.0229017	54704.8902	0.0000	0.99708	74.02290	374.7711
{S(t)f(t)} -- sin link	54716.5936	11.7034	0.00287	83	368.4517
{S(t)f(t)} -- logit	54724.5163	19.6261	0.00005	44	454.6161
{S(t)f(t)} -- sin link	54803.8298	98.9396	0.00000	43	535.9342

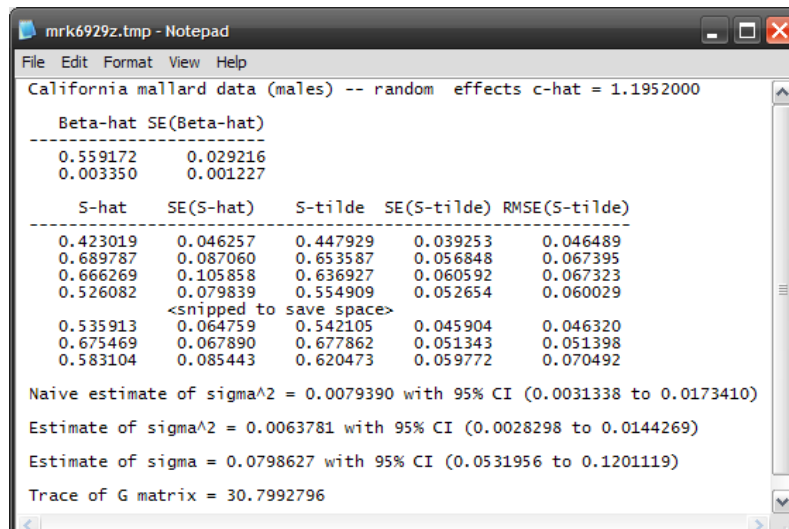
We see that the ‘intercept only’ random effects model has virtually all the support in the data, even relative to our previous ‘best model’ $\{S_t f_t\}$. Again, we note that the number of parameters estimated for the random effects model is non-integer. Note, however, that the number of parameters estimated (74.02363) is not simply $\text{tr}(\mathbf{G})$ ($= 32.02363$). The difference between the two values is $(74.02363 - 32.02363) = 42$. Where does the 42 come from? Recall that the number of parameters estimated for the random effects model, K_{re} is given as $K_{re} = \text{tr}(\mathbf{G}) + \ell$, where ℓ is the number of free parameters not being modeled as a random effect. In our mallard example, we modeled the 41 survival parameters S_1, \dots, S_{41} as a random effect, but we left the recovery parameter f modeled over time as a simple fixed effect. How many f parameters in our model? $\ell = 42$, which of course is why the number of parameters estimated is 42 more than $\text{tr}(\mathbf{G})$.

What more can we about our results so far? Consider the improvement in precision achieved by the shrinkage estimates, \tilde{S}_i , from model $\{S_{\mu,\sigma} f_t\}$ compared to the ML estimates, \hat{S}_i from model $\{S_t f_t\}$. As discussed in section D.3.2, a convenient basis for this comparison is the ratio of average \widehat{RMSE} to \widehat{SE} :

$$\frac{\widehat{RMSE}(\tilde{S}_i | S_i)}{\widehat{SE}(\hat{S}_i | S_i)} = \frac{0.06476}{0.07870} = 0.823.$$

The average precision of the shrinkage estimates is improved, relative to MLEs, by 18%, hence confidence intervals on S_i would be on average 18% shorter.

Let’s continue by fitting a linear trend random effects model. First, retrieve model $\{S_t f_t\}$. Then, start a ‘**variance components**’ analysis – this time, selecting the ‘**linear trend**’ design matrix specification, instead of the default ‘**intercept only (mean)**’. Here is a truncated listing of the numerical estimates.



mrk6929z.tmp - Notepad

California mallard data (males) -- random effects c-hat = 1.1952000

Beta-hat SE(Beta-hat)	
0.559172	0.029216
0.003350	0.001227

S-hat	SE(S-hat)	S-tilde	SE(S-tilde)	RMSE(S-tilde)
0.423019	0.046257	0.447929	0.039253	0.046489
0.689787	0.087060	0.653587	0.056848	0.067395
0.666269	0.105858	0.636927	0.060592	0.067323
0.526082	0.079839	0.554909	0.052654	0.060029
<snipped to save space>				
0.535913	0.064759	0.542105	0.045904	0.046320
0.675469	0.067890	0.677862	0.051343	0.051398
0.583104	0.085443	0.620473	0.059772	0.070492

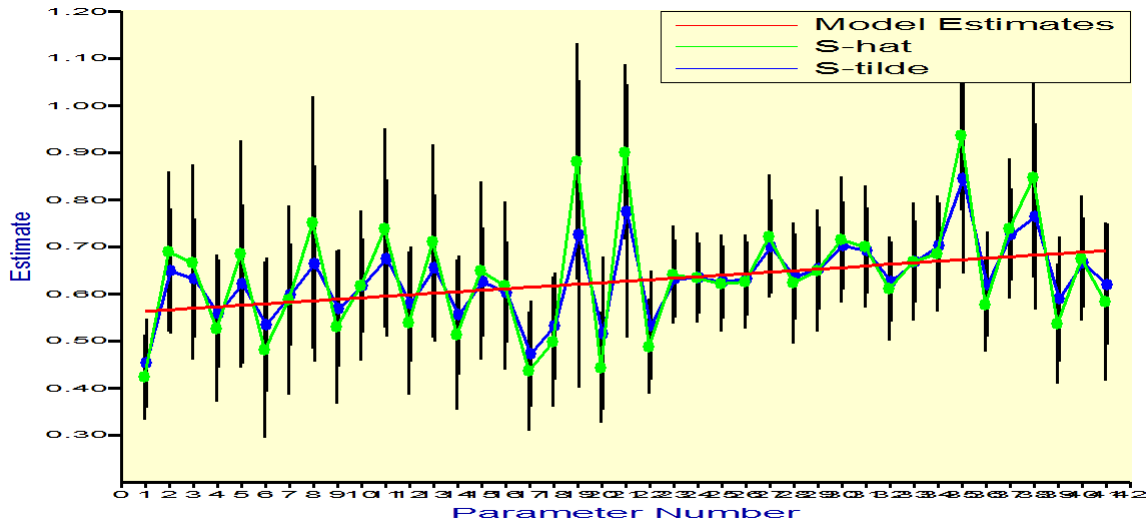
Naive estimate of $\sigma^2 = 0.0079390$ with 95% CI (0.0031338 to 0.0173410)

Estimate of $\sigma^2 = 0.0063781$ with 95% CI (0.0028298 to 0.0144269)

Estimate of $\sigma = 0.0798627$ with 95% CI (0.0531956 to 0.1201119)

Trace of G matrix = 30.7992796

We see that the estimated process variance is nearly half the value estimated from the intercept only model. We also see that the estimate for the slope is positive ($\hat{\beta} = 0.0034$). This is reflected in the plot of the ML and shrinkage estimates against the model, shown below:



Next, we'll go ahead and fit the estimated 'linear trend' RE model to the data, after adding the phrase 'linear trend' to the title.

Here is the results browser with the 'linear trend' random effects model results added:

Model	QAICc	Delta QAICc	QAICc Weight	No. Par.	QDeviance
{S(t)f(t) -- sin link: linear trend -- Random Effects Trace G=30.7983785}	54703.3469	0.0000	0.68324	72.79838	375.6855
{S(t)f(t) -- sin link: Random Effects Trace G=32.0229017}	54704.8902	1.5433	0.31583	74.02290	374.7711
{S(t)f(t) -- sin link}	54716.5936	13.2467	0.00091	83	368.4517
{S(T)f(t) -- logit}	54724.5163	21.1694	0.00002	44	454.6161
{S(.)f(t) -- sin link}	54803.8298	100.4829	0.00000	43	535.9342

We see clear evidence of strong support for random variation in the individual S_i around the trend line – this model has almost twice the support in the data as the next best model (our intercept only model). What is of particular note is that if we hadn't built the random effects models, and had based our inference solely on the 3 starting models, we would have concluded there was no evidence whatsoever of a trend, when in fact, the random effects trend model ended up being the best supported by the data.

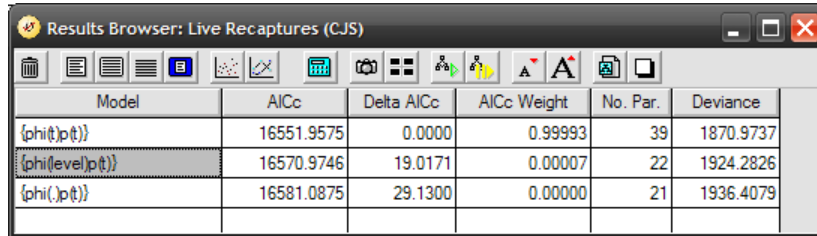
The simple random effects models we used here are both necessary for inference about process variation, σ^2 , and also for improved inferences about time-varying survival rates.

D.4.3. Random effects – environmental covariates

Here, we consider fitting a random effects model when survival differs as a function of some environmental covariate. Suppose we have some live encounter (CJS) data collected on a fish population studied in a river that is subject to differences in water level. You hypothesize that annual fish survival is influenced by variation in water level. We have $k = 21$ occasions of mark-recapture data (contained

in **level-covar.inp**). Over each of the 20 intervals between occasions, water flow was characterized as either ‘average’ (A) or ‘low’ (L) (more specific covariate information was not available). Here is the time series of flow covariates: {AAAAALLAAALALALLLAL}.

We begin our analysis by considering 3 fixed effect models for apparent survival, $\varphi: \{\varphi_t p_t\}, \{\varphi, p_t\}$ and $\{\varphi_{level} p_t\}$. Here are the results from fitting these 3 models to the data:

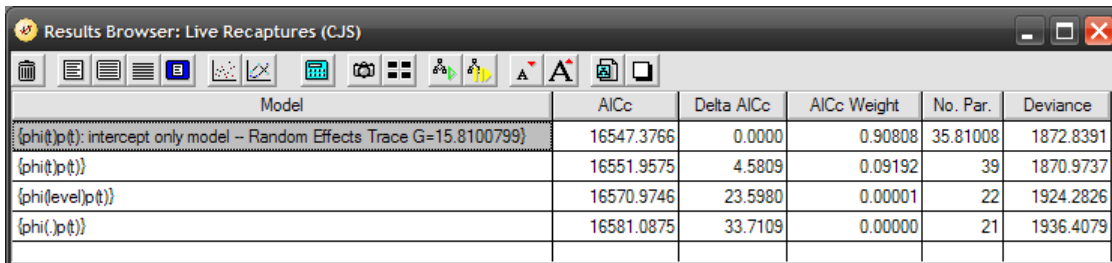


Model	AICc	Delta AICc	AICc Weight	No. Par.	Deviance
{phi(t)p(t)}	16551.9575	0.0000	0.99993	39	1870.9737
{phi(level)p(t)}	16570.9746	19.0171	0.00007	22	1924.2826
{phi(.)p(t)}	16581.0875	29.1300	0.00000	21	1936.4079

We see strong evidence for variation over time in apparent survival, but no support for an effect of water level. If you look at the estimates from model $\{\varphi_{level}\}$ for average ($\hat{\varphi}_{avg} = 0.709$, $SE = 0.0106$) and low ($\hat{\varphi}_{low} = 0.650$, $SE = 0.0100$), the lack of any support for this model may not be surprising. At least, based on considering water level as a fixed effect.

Now let's consider some random effects models. We'll build 2 different models – one a simple intercept only (mean) model, which would seem to be consistent with the strong support for the simple time variation model $\{\varphi_t\}$, and one where survival is thought to vary randomly around a level-specific mean. In other words, we hypothesize $\mu_{low} \neq \mu_{avg}$. We'll assume, however, that $\sigma_{low}^2 = \sigma_{high}^2$. This is not directly testable using the moments-based variance components approach in **MARK**, but is testable using an MCMC approach (see appendix E).

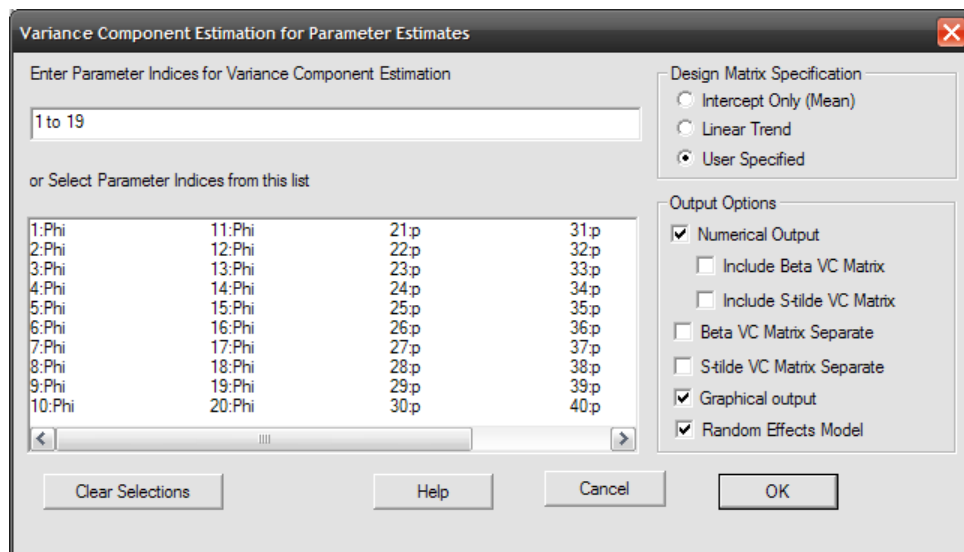
By this point, building and fitting the intercept only model for the survival parameters $\varphi_1 \rightarrow \varphi_{19}$ (remember, we don't include φ_{20} since it is confounded with our estimate of p_{21} for our time-dependent model) should be straightforward, so we'll skip the description of the mechanics, and will simply present the results – we've added the ‘intercept only’ model to the browser (below).



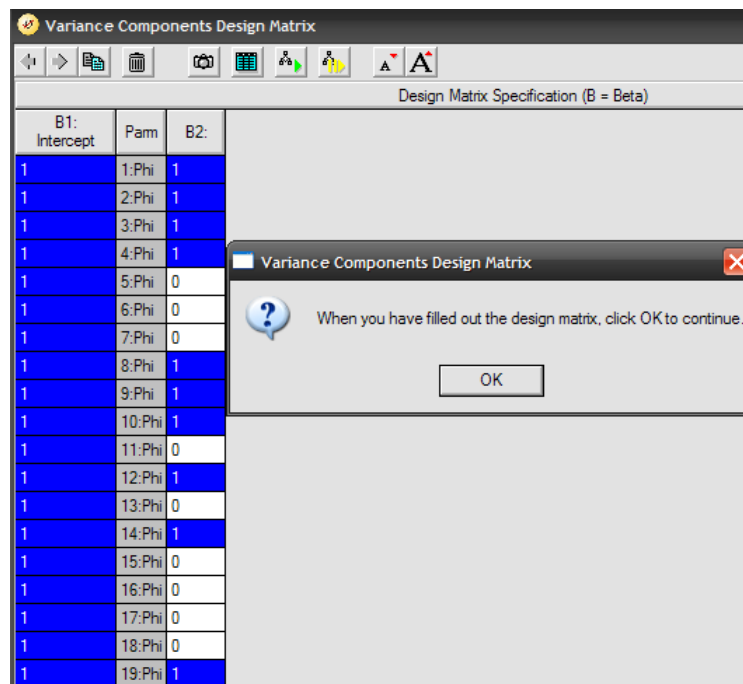
Model	AICc	Delta AICc	AICc Weight	No. Par.	Deviance
{phi(t)p(t): intercept only model – Random Effects Trace G=15.8100799}	16547.3766	0.0000	0.90808	35.81008	1872.8391
{phi(t)p(t)}	16551.9575	4.5809	0.09192	39	1870.9737
{phi(level)p(t)}	16570.9746	23.5980	0.00001	22	1924.2826
{phi(.)p(t)}	16581.0875	33.7109	0.00000	21	1936.4079

We see strong evidence that the intercept only random effects model is more parsimonious given the data than any other model.

Now, we consider the final model where survival is thought to vary randomly around a level-specific mean. We'll refer to this as model $\{\varphi_{\mu_{level}\sigma_{level}^2}\}$. How do we construct this model in **MARK**? Here we finally make use of the ‘**User Specified**’ design matrix option in the variance components setup window (shown at the top of the next page). Simply retrieve the time-specific fixed effect model $\{\varphi_t p_t\}$, start the variance components analysis, specify the parameters (1 to 20), and check the ‘**User Specified**’ design matrix option. Make sure you've also checked the ‘**Random Effects Model**’ option as well.



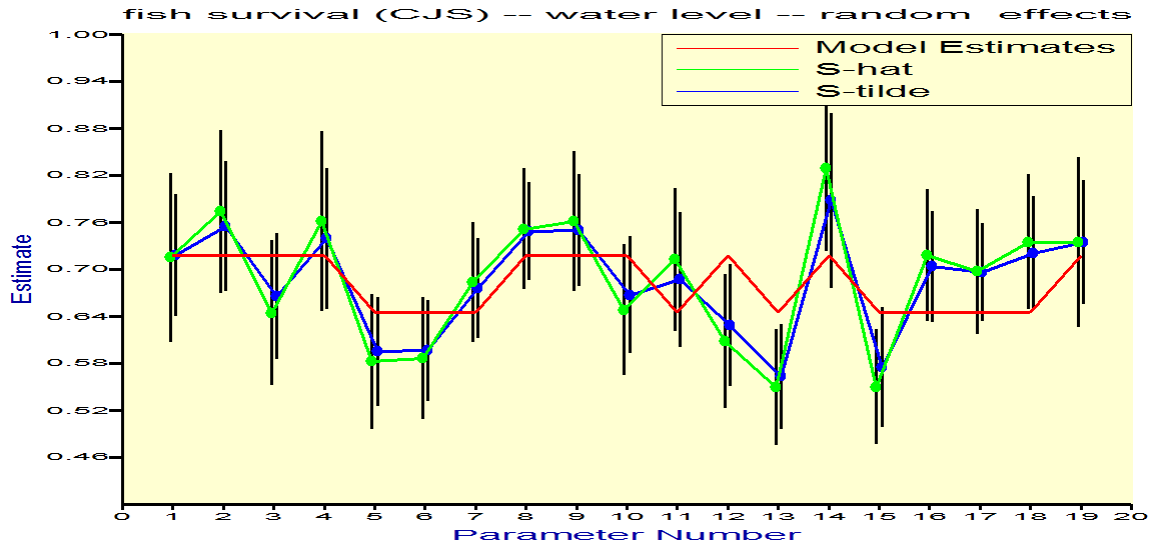
Now, all we need to do is click the 'OK' button. Once you do so, MARK will present you with a small pop-up window asking you to specify how many covariate columns you want in the user-specified design matrix. The default is 3, but here, for our model, we need only 2: one to code for the intercept, and one to code for the water level (a single column since there are only two levels of the water covariate). The design matrix entry window is shown below:



Note that there are only 19 rows in the DM, since we're applying a random effect to $\varphi_1 \rightarrow \varphi_{19}$ only. Also note the small 'pop-up' window indicating that 'When you have filled out the design matrix, click OK to continue'. Meaning you should do as it says – specify your design matrix for the survival

parameters, and then when you're sure it's correct, click the 'OK' button.

First we're presented with a plot (below), showing the ML and shrinkage estimates, and (importantly here) the underlying model (the red line).



The red line clearly indicates that there are 2 separate means being modeled, for the low and average water flow years, respectively. The estimated process variance is $\hat{\sigma}^2 = 0.00313$, and the estimate for $\hat{\beta}_1 = 0.0717$ in the linear model indicates that survival is higher in 'average' water level years (since we used 'low' level years as the reference level in our design matrix, above). What is also very important here, is that the shrinkage estimates are clearly not constrained to fall exactly 'on the red line' – they represent shrunk estimates of apparent survival as if each estimate was drawn randomly from a sample with a water level-specific mean.

OK, what about the results of fitting this random effects model to the data?

Results Browser: Live Recaptures (CJS)					
Model	AICc	Delta AICc	AICc Weight	No. Par.	Deviance
{phi(t)p(t): water level model -- Random Effects Trace G=15.2401619}	16546.4694	0.0000	0.58836	35.24016	1873.0840
{phi(t)p(t): intercept only model -- Random Effects Trace G=15.8100799}	16547.3766	0.9072	0.37380	35.81008	1872.8391
{phi(t)p(t)}	16551.9575	5.4881	0.03784	39	1870.9737
{phi(level)p(t)}	16570.9746	24.5052	0.00000	22	1924.2826
{phi(.)p(t)}	16581.0875	34.6181	0.00000	21	1936.4079

What is especially noteworthy here is that the random effects model with water level-specific means, $\{\varphi_{\mu_{level}\sigma_{level}^2}\}$, is now the most parsimonious model in the model set, despite 'water level' having no support whatsoever when considered in a fixed effects design. The near equivalence of the AICc weights between this model and the simpler 'intercept only' random effects model suggest that we can't differentiate between the two, but whereas our initial conclusion strongly rejected the hypothesis that there was an influence of water level on apparent survival, our random effects modeling would suggest that perhaps we shouldn't be quite so sure.

D.4.4. Worked example – λ -Pradel model

We conclude with analysis of a famous set of data, the moth (*Gonodontis bidentata*) data reported on by Bishop *et al.* (1978) and compulsively analyzed by many others (e.g., Link & Barker 2005). The data consist of records for 689 male moths that were captured, marked, and released daily over 17 days in northwest England. These moths were nonmelanic; demographic parameters were estimated as part of a larger study looking at comparative fitness of distinct color morphs.*

Here we will use random effect Pradel models (Chapter 13), focussing on estimation of process variance, and possible trend, in realized growth rate λ . The data we'll work with are contained in **moth-example.inp**. Our focus here is on variation in λ . Recall that in a Pradel model, λ can be estimated as either a structural (real) parameter (for data type '**survival and lambda**'), or as a derived parameter (for any of the other Pradel model data types). For the purposes of estimating process variance on λ , it doesn't particularly matter which data type we use, since we can estimate process variance for either real or derived parameters. We have already seen variance components analysis of real parameters in earlier examples, so here, we'll demonstrate the process of estimation for λ as a derived parameter.

To start, select the data type '**Pradel survival and seniority**'. 17 occasions. Now, you may recall that there are some challenges with parameter confounding for fully time-dependent Pradel models. As is often the case, some of these problems can be handled by applying constraints to one or more parameters. Since our purpose here is simply to demonstrate some mechanics, and not conduct an exhaustive analysis of these data, we'll avoid some of these issues by simply setting the encounter probability p to be constant over time. So, our general model will be $\{\varphi_t p, \gamma_t\}$. Go ahead and modify the PIM chart to construct this model, and add the results to the browser. Model deviance is 236.708. Here are the derived estimates

mothe example -- Pradel models -- RE

Estimates of Derived Parameters
Lambda Estimates of $\{\phi(t)p(.)\gamma(t)\}$

Grp.	Occ.	Lambda-hat	Standard Error	95% Confidence Interval	
				Lower	Upper
1	1	3.2172358	0.8774335	1.4974662	4.9370054
1	2	1.0801263	0.1782850	0.7306877	1.4295648
1	3	1.1301282	0.1904042	0.7569360	1.5033205
1	4	0.4113538	0.0848096	0.2451269	0.5775807
<snipped to save space>					
1	13	0.6402288	0.0938706	0.4562424	0.8242152
1	14	1.0036080	0.1579969	0.6939340	1.3132820
1	15	0.5873034	0.1132084	0.3654149	0.8091919
1	16	0.2268143	0.0836395	0.0628808	0.3907478

log(Lambda) Estimates of $\{\phi(t)p(.)\gamma(t)\}$

Grp.	Occ.	log(Lambda-hat)	Standard Error	95% Confidence Interval	
				Lower	Upper
1	1	1.1685225	0.2727290	0.6339737	1.7030714
1	2	0.0770780	0.1650594	-0.2464384	0.4005943
1	3	0.1223311	0.1684802	-0.2078901	0.4525523
1	4	-0.8883016	0.2061720	-1.2923987	-0.4842045
<snipped to save space>					
1	13	-0.4459297	0.1466204	-0.7333057	-0.1585537
1	14	0.0036015	0.1574289	-0.3049592	0.3121622
1	15	-0.5322138	0.1927597	-0.9100228	-0.1544047
1	16	-1.4836237	0.3687578	-2.2063889	-0.7608585

Note that **MARK** generates derived estimates of λ , and $\log \lambda$. While there are important considerations as to which is more appropriate for analysis (see discussion in Chapter 13), our purpose here is

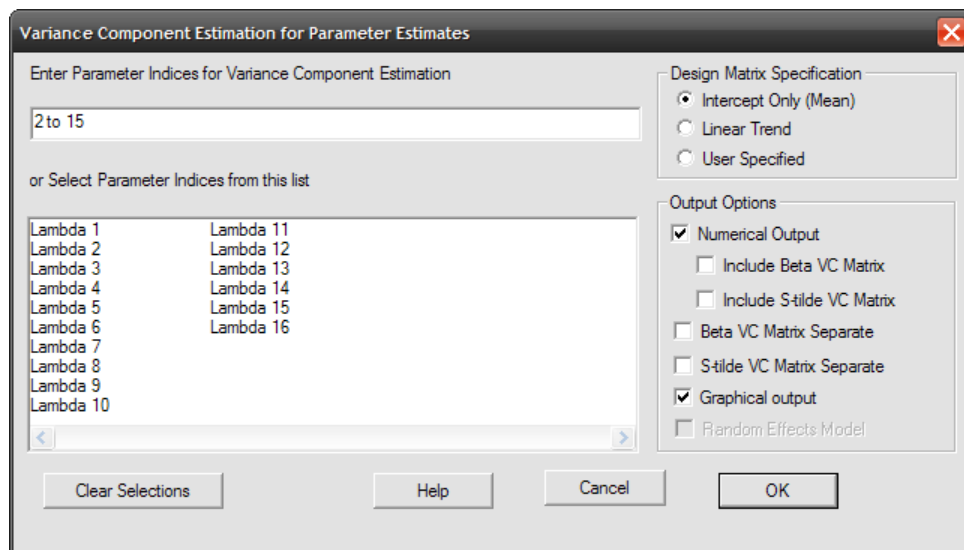
* The underlying motivation for this study should be very familiar to any of you with some background in evolutionary biology.

simply to demonstrate some of the ‘mechanics’, so we’ll focus on estimates of λ on the linear scale.

Let’s construct a random effects model for λ , using $\{\varphi_t p, \gamma_t\}$ as our general model. The steps are the same, except that when you access the ‘**variance components**’ sub-menu, you need to specify ‘**derived parameter estimates**’ for the parameter type. You’ll be asked to select either ‘**Lambda Population Change**’ or ‘**log(Lambda) Population Change**’. We’ll select the former.

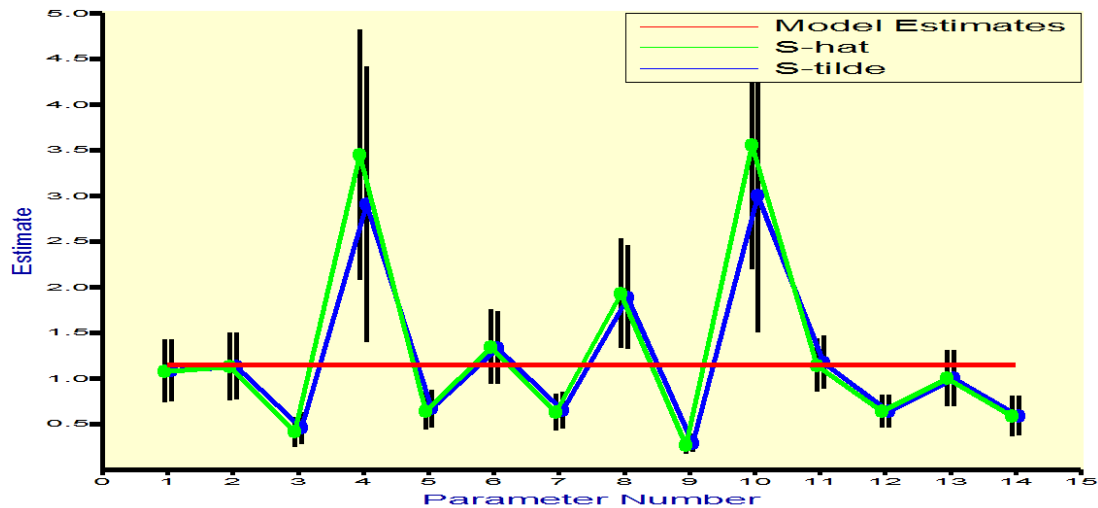
Now, for the first ‘challenge’ – specifying the parameter index values for λ . You might simply assume that you would select all of them: 1 to 16. However, here (and generally), we need to be a bit careful. When specifying the parameters to include in the random effect, you should not include parameters that are confounded, or that are otherwise known to have ‘issues’. For Pradel models, there is potential for confounding, but constraining encounter probability p to be constant over time should solve some of that problem. However, experience has shown that the first and last estimates of λ in Pradel models are often biased. If you look at the estimates for $\hat{\lambda}_1$ and $\hat{\lambda}_{16}$ shown on the previous page, you’ll see some evidence that this might be the case for these data. For example, $\hat{\lambda}_1 = 3.217$. This is often a judgement call (for example, you might say that $\hat{\lambda}_{11} = 3.55$ is even larger, so using scale as evidence for a ‘problem’ with $\hat{\lambda}_1$ is perhaps not a great criterion). For now, we’ll be ‘conservative’, and include only $\lambda_2 \rightarrow \lambda_{15}$ in specifying the random effect.

Now, for one important difference in fitting random effects models to derived parameters – you can’t. If you look in the lower-right-hand corner, you’ll notice that the ‘**Random Effects Model**’ check-box has been ‘greyed-out’.



If you think about it, this should make some sense. The random effects ‘constraint’ can be applied only to real (structural) parameters in the model. Here, we are estimating λ as a derived parameter (i.e., by algebra), and thus we can’t build and fit a random effects model using the Pradel ‘**Survival and seniority only**’ data type. We’ll revisit this point in a moment. Go ahead and click the ‘**OK**’ button. The plot of the real estimates $\hat{\lambda}_i$ and shrinkage estimates $\tilde{\lambda}_i$ is shown at the top of the next page. The estimated mean $\hat{\lambda} = 1.147$, with an estimated process variance of $\hat{\sigma}^2 = 0.722$. The plot shows clear evidence of fairly large swings in realized growth – this is not overly surprising for an insect, where large changes in reproduction and survival are relatively commonplace (there aren’t enough data available to test for ‘boom-bust’ cycles, another common occurrence with insect growth dynamics).

Now, what if we wanted to fit a random effects model for λ , and not simply estimate the mean and process variance? To do this, we need to change the data type, to one where λ is included as a structural



parameter in the model. Simply select '**PIM | Change data type**'. You will be presented with a box contained the different data types you can apply to these data. You'll see that one of those data types presented is '**Pradel survival and lambda**'. Since that is the only model containing λ as a real structural parameter, select that data type, and click '**OK**'.

You'll find that you've been returned back to the browser view – with no indication that anything has changed. But, if you look at the PIM chart, you'll see that the current 'active' model structure is $\{\varphi_t p_t \lambda_t\}$. In other words, the parameter structure of the model has changed (as expected), but the underlying model has reverted back to the fully time-dependent model. **MARK** has basically assumed that if you're switching data types, you are (in effect) starting over. So, we want to make the encounter probability p constant, so that our general model under this data type, $\{\varphi_t p, \lambda_t\}$ is consistent with the general model we used under the data type where λ was estimated as a derived parameter.

Go ahead and run model $\{\varphi_t p, \lambda_t\}$, and add the results to the browser.

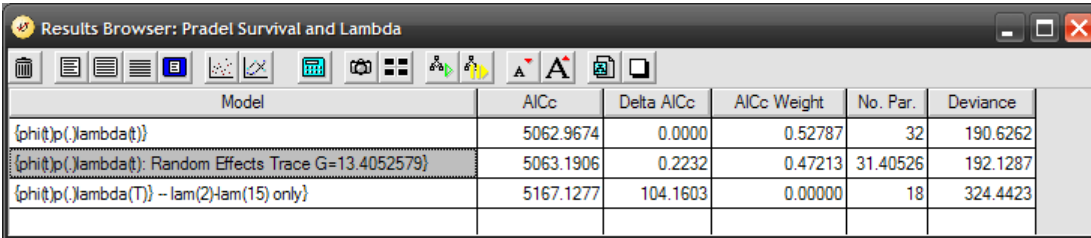
Results Browser: Pradel Survival and Lambda					
Model	AICc	Delta AICc	AICc Weight	No. Par.	Deviance
$\{\varphi_t p(\lambda_{t\hat{d}})\}$	5062.9674	0.0000	1.00000	32	190.6262
$\{\varphi_t p(\gamma_{t\hat{d}})\}$ – general model	5111.2040	48.2366	0.00000	33	236.7079

Unfortunately we see that these are not the same models, even though in theory they should be – the model deviances and the number of estimated parameters are both different. There might be a couple of issues here. First, if you look at the estimates, you'll see that the first and final estimates of apparent survival φ and the final estimate of λ are both poorly estimated, and are probably confounded (despite setting p constant). Second, and perhaps more likely, the logical constraint that $\lambda_i > \varphi_i$ is not enforced for 1 or more parameters. This is not uncommon in fitting the '**Pradel survival and lambda**' model (for discussion of this issue, see the -sidebar- beginning on p. 9 of Chapter 13). As such, comparing model $\{\varphi_t p, \lambda_t\}$ with $\{\varphi_t p, \gamma_t\}$ makes little sense, although the estimates of $\hat{\lambda} = 1.170$ and $\hat{\sigma}^2 = 0.697$ from model $\{\varphi_t p, \lambda_t\}$ (using $\lambda_2 \rightarrow \lambda_{15}$) are fairly close to the estimates derived for model $\{\varphi_t p, \gamma_t\}$.

For the moment, we'll skip over this issue, and focus on some additional 'mechanics'. Delete model

$\{\varphi_t p, \gamma_t\}$ from the browser – we’ll focus on model $\{\varphi_t p, \lambda_t\}$. We are interested in analysis of trend in λ . Trend in realized λ is potentially of great significance for conservation and management, and robust estimation of trend would seem to be something worth pursuing. Now, based on the plot of $\hat{\lambda}_i$ and $\tilde{\lambda}_i$ shown at the top of the preceding page, we don’t expect a ‘trend’ model of any flavor to have much support in the data. But, to demonstrate the mechanics, we’ll proceed anyway.

With model $\{\varphi_t p, \lambda_t\}$, we can see there are at least 2 ways we could proceed. We could build an ultrastructural model where λ is constrained to fall on a straight trend line, or we could build a random effects trend model. We’ll do both, for purposes of comparison. To make the comparison ‘fair’, we’ll build the ‘T’ (trend) ultrastructural model applied to $\lambda_2 \rightarrow \lambda_{15}$ only, since these are the parameters we would use in the random effects model. Then, we’ll build the random effects ‘trend’ model, also using $\lambda_2 \rightarrow \lambda_{15}$ only, and add the results of both to the browser.



Model	AICc	Delta AICc	AICc Weight	No. Par.	Deviance
$\{\phi(t)p(\cdot)\lambda(t)\}$	5062.9674	0.0000	0.52787	32	190.6262
$\{\phi(t)p(\cdot)\lambda(t): \text{Random Effects Trace } G=13.4052579\}$	5063.1906	0.2232	0.47213	31.40526	192.1287
$\{\phi(t)p(\cdot)\lambda(t): \text{--lam(2)-lam(15) only}\}$	5167.1277	104.1603	0.00000	18	324.4423

We notice immediately that the random effects model with trend actually gets a fair amount of support in the data, especially relative to the ultrastructural fixed effects trend model. So, despite appearances, there might be some evidence of trend in the data. Enough that the overall support for this model is fairly compelling.

At least 2 points to make here. First, there are some general concerns with the ultrastructure approach, where constraints are applied to λ . Since growth is a function of *per capita* survival plus *per capita* recruitment, $\lambda = \varphi + f$, then any constraint applied to λ enforces a strict negative covariance between $\hat{\varphi}$ and \hat{f} (see chapter 13, and discussions in Franklin 2001). Such a covariance is clearly artificial, and may make little to no biological sense (since it would imply that any increase in survival is perfectly balanced by an equal and opposite change in recruitment). Second, the random effects approach to trend analysis can’t be applied to λ when λ is estimated as a *derived* parameter – you can estimate the intercept and slope of the trend, and the process variance, but you can’t fit a random effects model for a derived parameter (i.e., you cannot add it to the browser).

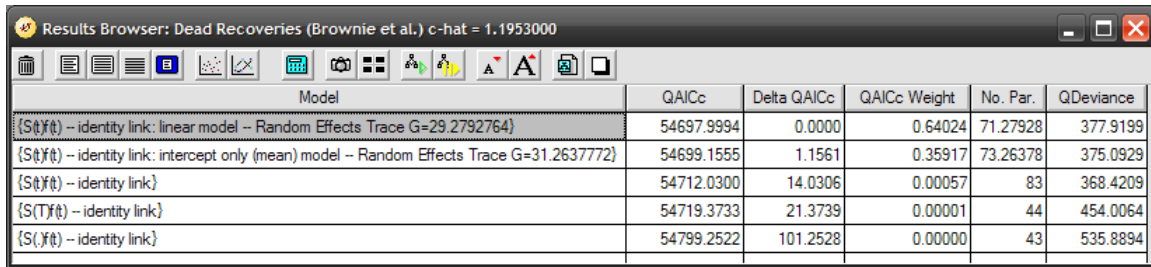
While there are some technical challenges with Pradel models in general (especially for time-dependent models, which are exacerbated for random effects models which are generally applied to time-dependent models), it seems clear that considering variance components and random effects analysis for λ in Pradel models has potential to be extremely useful.

D.5. Model averaging?

At this point, you might be asking yourself, ‘what about model averaging?’. This important topic is introduced in some detail in Chapter 4 (see the Burnham & Anderson ‘model selection’ book for a comprehensive treatment).

There are at least two issues to consider here. First, if our candidate model set contains both fixed effect and random effect models, should we model average over the entire set? In fact, **MARK** does nothing ‘mechanically’ to prevent you from doing so. For example, take the mallard analysis introduced in section D.4.2. Assume that we have the following 5 models in the browser – 3 fixed effects

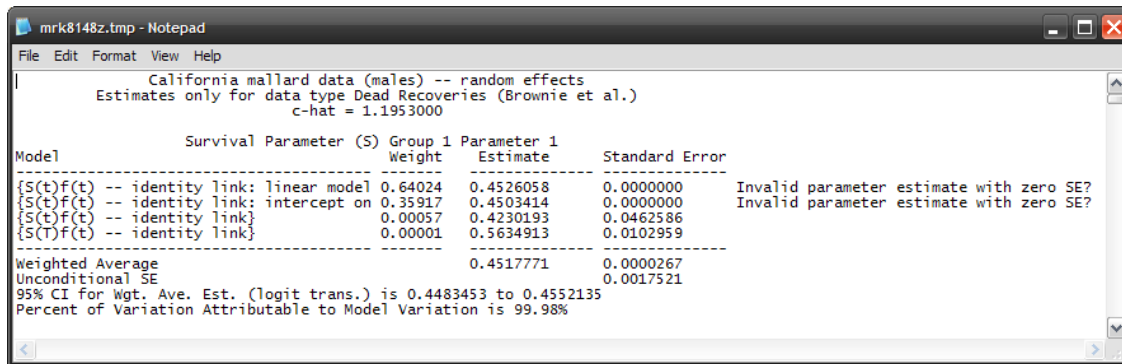
models ($\{S_t f_t\}$, $\{S_t f_t\}$, $\{S_T f_t\}$), and 2 random effect models ('linear trend model', and 'intercept only (mean) model':



Results Browser: Dead Recoveries (Brownie et al.) c-hat = 1.1953000

Model	QAICc	Delta QAICc	QAICc Weight	No. Par.	QDeviance
{S(t)f(t) -- identity link: linear model -- Random Effects Trace G=29.2792764}	54697.9994	0.0000	0.64024	71.27928	377.9199
{S(t)f(t) -- identity link: intercept only (mean) model -- Random Effects Trace G=31.2637772}	54699.1555	1.1561	0.35917	73.26378	375.0929
{S(t)f(t) -- identity link}	54712.0300	14.0306	0.00057	83	368.4209
{S(T)f(t) -- identity link}	54719.3733	21.3739	0.00001	44	454.0064
{S(.)f(t) -- identity link}	54799.2522	101.2528	0.00000	43	535.8894

You can go ahead and run the model averaging routines on the survival estimates – **MARK** simply assumes you want to average over the entire model set. Your only hint that you might need to be a bit careful comes when you look at the model averaging output. Here is a snippet of the output for the mallard analysis, for the parameter corresponding to S_1 :



```

mrk8148z.tmp - Notepad
File Edit Format View Help
California mallard data (males) -- random effects
Estimates only for data type Dead Recoveries (Brownie et al.)
c-hat = 1.1953000

Model      Survival Parameter (S) Group 1 Parameter 1
              Weight      Estimate      Standard Error
-----
{S(t)f(t) -- identity link: linear model 0.64024 0.4526058 0.0000000 Invalid parameter estimate with zero SE?
{S(t)f(t) -- identity link: intercept on 0.35917 0.4503414 0.0000000 Invalid parameter estimate with zero SE?
{S(t)f(t) -- identity link} 0.00057 0.4230193 0.0462586
{S(T)f(t) -- identity link} 0.00001 0.5634913 0.0102959

Weighted Average 0.4517771 0.0000267
Unconditional SE 0.0017521
95% CI for Wgt. Ave. Est. (logit trans.) is 0.4483453 to 0.4552135
Percent of Variation Attributable to Model Variation is 99.98%

```

Pay particular attention to the right-hand side. Notice that for the 2 random effects models, you get a 'warning' about 'invalid parameter estimates' with 'zero SE?'. **MARK** has 'noticed' that the parameter estimates for parameters modeled as random effects (the survival parameters, in this case) have a zero SE. **MARK** 'suspects' there might be a problem, but in fact this is exactly what you should see. As discussed earlier, the random effects model is fit to the data after fixing the random parameters to their shrinkage estimates, and thus, there is no SE for those parameters. And, since model averaging intends to generate an unconditional parameter estimate, by accounting for conditional uncertainty in the parameter for each model, then there are clear problems when one or more of the parameters 'appear' to be estimated without any uncertainty (i.e., have zero SE's). So, if **MARK** included the random effect parameter estimates in the model averaging, then the estimate of the unconditional SE for that parameter would be very negatively biased (i.e., much too small).

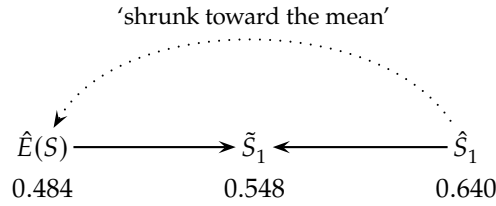
The second issue is 'conceptual', and relates broadly to the issue of 'model redundancy'. In the extreme, having 2 structurally identical models in the same candidate model set would clearly render model averaged estimates invalid. Model redundancy in the context of AIC selection is discussed in Burnham and Anderson (sections 4.2.9 and 4.2.10). With some thought, you might see that the random effects model is substantially redundant to its fixed effects likelihood version. For example, model $\{S_{\mu,\sigma} f_t\}$ would be redundant to some degree to model $\{S_t f_t\}$.

In one sense, a random effects model is a model which is intermediate between a time-invariant ('dot', say $S_.$) model, and a fully time-varying fixed effects model (say, S_t). Recall from section D.3 that

shrinkage estimates \tilde{S}_i generally lie between $\hat{E}(S)$ and \hat{S}_i . Recall from section D.3.1 that in the absence of sampling covariance, the shrinkage estimator used in **MARK** is

$$\tilde{S}_i = \hat{E}(S) + \sqrt{\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{E}_S[\text{var}(\hat{S}_i | S_i)]}} \times [\hat{S}_i - \hat{E}(S)].$$

Consider the estimates of survival for the first interval from the binomial example presented in section D.2.1: the ML estimate $\hat{S}_1 = 0.640$, the mean survival $\hat{E}(S) = 0.484$, and the shrunk estimate $\tilde{S}_1 = 0.548$. As shown in the following



the ML estimate is ‘shrunk’ toward the mean, with the resulting shrinkage estimate, \tilde{S}_1 being ‘intermediate’ between the mean, $\hat{E}(S)$ (which is equivalent to a ‘dot’ model, in some respects) and the ML estimate \hat{S}_1 (which comes from the ‘time-dependent’ model). So, the shrinkage estimate is analogous to an ‘average’ between the two estimates. As noted in section D.3, the degree of shrinkage is determined by the magnitude of the process variance σ^2 , relative to the sampling variance, $\hat{E}_S[\text{var}(\hat{S}_i | S_i)]$. If the process variance is relatively small, then the shrunk estimates will approach the estimate for the mean $\hat{E}(S)$, whereas if the process variance is large relative to the sampling variance, then the shrunk estimates will be closer to the ML estimates, \hat{S}_i . So, the shrinkage model $\{S_{\mu,\sigma}, \underline{\theta}\}$, is *intermediate* between model $\{S_t, \underline{\theta}\}$ and model $\{S_., \underline{\theta}\}$, and is thus strongly analogous to an ‘average model’. As a result, when the process and sampling variances are similar, the estimates from model $\{S_{\mu,\sigma}, \underline{\theta}\}$ will tend to be quite similar to the model-averaged estimates estimated over model $\{S_t, \underline{\theta}\}$ and model $\{S_., \underline{\theta}\}$.

D.6. Caveats, warnings, and general recommendations

Using random effects as a basis for modeling collections of related parameters is a long-standing approach in statistics and one that can be very effective. Use of the random effects approach in capture-recapture is relatively new – in the nearly 10 years since the publication of B&W, there have been relatively few applications of these models to real data, despite what we believe are several interesting opportunities made available by these methods.

However, we also believe that the methodology needs to be better understood as to any potential pitfalls and as to its operating characteristics. The following is a summary of our experience to date with random effects models, particularly as implemented in **MARK**. This material is largely abstracted from B&W, and accumulated experience with such models since the time of that publication.

1. the ‘method of moments’ described in the appendix, and as implemented in **MARK**, has been shown to perform well, especially when $\sigma^2 > 0.025$. The method(s) may not do so well if $\sigma^2 \rightarrow 0$. However, we think it reasonable to believe that for a worthwhile study yielding good data, process variation, σ^2 , will generally not be too small, relative to average sampling variation and it is for these conditions (of ‘good data’) that we need effective random effects inference methods.

2. Another issue to be aware of, as regards estimation of the parameter σ^2 , is the matter of unequal, rather than equal length, time intervals. Let the time interval i have length Δ_i . Then we should parameterize the model as $S_i = (\psi_i)^{1/\Delta_i}$ where now each survival probability ψ_i is on the same unit time basis. It may then make biological sense to consider parameters that are a mean and variation for ψ_1, \dots, ψ_k . But this may just as well not make sense, because the time intervals are intrinsically not comparable as they may be in very different times of the annual cycle. It becomes a subject matter judgement as to whether random effects analysis will be meaningful with unequal time intervals. For the moment, don't apply random effects models or variance components analysis to situations where the intervals between sampling occasions are unequal (even specifying unequal interval length will generally yield negatively biased estimates of process variance, σ^2).
3. In practice if there is over-dispersion in the data, as measured by a scalar often denoted by c (see Chapter 4 and Chapter 5), the estimated sampling variance-covariance must be adjusted by a reliable \hat{c} (see discussion in B&W, and Franklin *et al.* 2002 for an example with real data).
4. A key design feature to focus on to meet the criterion of 'having good data' when applying random effects is k , the number of estimable random effects parameters (time intervals, locations, etc.). The sample size for estimating σ^2 is k . Therefore, one must not have k too small; < 10 is too small. Even if we knew all the underlying S_i a sample of size $k < 10$ is too small for reliable inference about the variation of these parameters (even given a random sample of them, which is not required here). Inference performance has been shown to be acceptable when $k > 15$. The benefits (includes shrinkage estimates) of random effects models become greater as the number of underlying parameters, k , increases.
5. The other influential design feature is number of individuals marked and released. Both numbers initially released and numbers recaptured are important to the performance of inferences from random effects models. While there are no 'hard and fast' rules, we can make some general recommendation. B&W showed that low numbers of marked and released animals, especially for low survival and encounter probabilities, generally led to point estimates of $\sigma^2 \rightarrow 0$. This is because the sampling variation was much larger than process variance in these cases.
6. The situation where inferences from a random effects model are most advantageous seems to be for when σ^2 is about the same as average sampling variance, $(\hat{S}_i | S_i)$ (recall that sampling variance is strongly influenced by sample size of animals capture and reencountered, whereas process variance is not). If one or the other variance component dominates the total variation in the MLE's \hat{S}_i then the data strongly favor either the simple model $\{S, p_t\}$ (sample variance dominates), or the general model $\{S_t p_t\}$ (process variation dominates), rather than the random effects model.

However, it is not a problem, as regards inference about σ^2 , to have large sample sizes of animals, hence small sampling variances, so that should be one's design goal. If it then turns out that sampling variance is similar to process variance, the random effects model will be markedly superior to model $\{S_t p_t\}$. Thus, in a sense the random effects model is optimal at the 'intermediate' sample size case. As sample size of animals increases, the random effects model converges to model $\{S_t p_t\}$.

7. A potential technical issue is the 'boundary effect' (at least under what is basically a likelihood approach). As discussed in B&W, if one enforces the constraint $S < 1$ when the unbounded MLE $\hat{S} \geq 1$, then standard numerical methods used in **MARK** to get the observed information matrix fails. As a result, the estimated information matrix is incorrect for any terms concerning the \hat{S} that is at the bound of 1 (and the inverse information

matrix is likely wrong in all elements). Experience shows that, in this case, the resultant point estimate of σ^2 can be very different from what one gets when the survival parameter MLE's are allowed to be unbounded. The difference can be substantial. Using an identity link, B&W found $\hat{\sigma}^2$ to be unbiased in many cases.* With good data we rarely observe an unbounded MLE of S that exceeds 1. This might be explored in a Bayesian context, where it is easy (in a MCMC analysis) to allow S to have its distribution over an interval such as 0 to 2 (rather than 0 to 1). B&W considered this, and found a strong effect of the upper-bound on the point estimate (and entire posterior distribution) for σ^2 , and for that particular S . (Note: MCMC applications in **MARK** are discussed in Appendix E).

D.7. Summary

This appendix has considered random effects models. The name 'random effects' can be misleading in that a person may think it means that underlying years or areas (when spatial variation is considered, rather than temporal) must be selected at random. This is neither true, nor possible, for a set of contiguous years. Variance components is a better name, in that at the heart of the method is the separation of process and sampling variance components. The issue of what inferential meaning we can ascribe to σ^2 is indeed tied to design and subject matter considerations. However, the shrinkage estimators do not depend on any inferential interpretation of σ^2 ; rather, they can always be considered as improvements, in a MSE sense, over the MLEs based on full time-varying S_i . The random effects model only requires that the residuals ($S_i - \mathbf{XB}$) are exchangeable.

When are we interested in these sorts of models? Often, if a data set is sparse, but with many (≥ 10) occasions, model $\{S(\cdot)\}$ will be selected. Clearly, this is a *model*, as we know that conditional survival probability cannot remain exactly the same over any significant length of time. While model $\{S(\cdot)\}$ might be 'best' in the sense of a bias-variance trade-off (i.e., it is identified by AIC as most parsimonious among the candidate models), it might leave the investigator wondering about the variation in the parameters. Thus, the estimation of σ^2 has relevance. At the other extreme, assume that model $\{S_i\}$ is selected; here the investigator might have (say) 25 estimates of the survival parameters, each perhaps with substantial sampling variation. This makes it difficult to see patterns (e.g., time trends or associations) or understand the variation in the parameters. Further, analysis of stochastic population models is often complicated by uncertainty concerning the relative variation in estimates of one or more demographic parameters.

Random effects models are a very interesting class of models which can address both issues (amongst many more), but even a partial understanding is somewhat difficult to achieve. The intent of this appendix was to try to convey the general notion of random effects models, and the idea of 'variance components', as implemented in program **MARK**, in a reasonably accessible fashion. The subject of random effects models and variance components as applied to data from marked individuals is treated in considerably more depth in several of the following papers. An alternative approach to estimating variance components, based on Bayesian inference (*sensu* Royle and Link, 2002) and Markov Chain Monte Carlo (MCMC), is presented in Appendix E.

* Note that we do not suggest routinely accepting final inferences that include survival estimates exceeding 1. In fact, the shrinkage estimates will generally not exceed 1, so using \tilde{S}_i and not \hat{S}_i will be the needed improved inference. However, to get to this final inference it may be desirable to pass through an imaginary space ($S > 1$), just as imaginary numbers can facilitate real solutions to real problems. Models only need to possess utility, not full reality.

D.8. References

- Burnham, K. P., Anderson, D. R., White, G. C., Brownie, C., and Pollock, K. H. (1987) *Design and Analysis Methods for Fish Survival Experiments Based on Release-Recapture*. American Fisheries Society Monograph No. 5. Bethesda, Maryland, USA. 437 pp.
- Burnham, K. P., and White, G. C. (2002) Evaluation of some random effects methodology applicable to bird ringing data. *Journal of Applied Statistics*, **29**, 245-264.
- Efron, B., and Morris, D. (1975) Data analysis using Stein's Estimator and its generalizations. *Journal of the American Statistical Association*, **70**, 311-319.
- Efron, B., and Morris, D. (1977) Stein's paradox in statistics. *Scientific American*, **238**, 119-127.
- Franklin, A. B., Anderson, D. R., and Burnham, K. P. (2002) Estimation of long-term trends and variation in avian survival probabilities using random effects models. *Journal of Applied Statistics*, **29**, 267-287.
- Gould, W. R., and Nichols, J. D. (1998) Estimation of temporal variability of survival in animal populations. *Ecology*, **79**, 2531-2538.
- Pfister, C. A. (1998) Patterns of variance in stage-structured populations: evolutionary predictions and ecological implications. *Proceedings of National Academy of Science*, **95**, 213-218.
- Royle, J. A., and Link, W. A. (2002) Random effects and shrinkage estimation in capture-recapture models. *Journal of Applied Statistics*, **29**, 329-351.
- Schmutz, J. A. (2009) Stochastic variation in avian survival rates: life-history predictions, population consequences, and the potential responses to human perturbations and climate change. Pages 441-461 in D. L. Thomson, E. G. Cooch, and M. J. Conroy, editors. *Modeling Demographic Processes in Marked Populations*. Springer, Berlin.
- White, G. C. (2000) Population viability analysis: data requirements and essential analyses. Pages 288-331 in L. Boitani and T. K. Fuller, editors. *Research Techniques in Animal Ecology: Controversies and Consequences*. Columbia University Press, New York, New York, USA.
- White, G. C., Burnham, K. P., and Anderson, D. R. (2001) Advanced features of Program MARK. Pages 368-377 in R. Field, R. J. Warren, H. Okarma, and P. R. Sievert, editors. *Wildlife, Land, and People: Priorities for the 21st Century*. Proceedings of the Second International Wildlife Management Congress. The Wildlife Society, Bethesda, Maryland, USA.
- White, G. C. (2000) Population viability analysis: data requirements and essential analyses. Pages 288-331 in L. Boitani and T. K. Fuller, editors. *Research Techniques in Animal Ecology: Controversies and Consequences*. Columbia University Press, New York, New York, USA.