October 7, 2001

BASIC RANDOM EFFECTS MODELS FOR CAPTURE-RECAPTURE DATA

Kenneth P. Burnham
Colorado Cooperative Fish and Wildlife Research Unit (USGS-BRD)
CSU, Fort Collins, Colorado 80523, USA

ABSTRACT

Tag-recovery and live-recapture data analyses focus on survival probabilities between release occasions; time intervals between release occasions are assumed here to be of equal duration, such as one year. We should expect temporal variation in these survival probabilities (S) to be partly random; however, existing models treat these temporal variations as fixed effects. Often survival rates $S_1, \ldots, S_k$ are unrestricted in the general model with the only alternative model assuming $S_1 = \cdots = S_k \equiv S$. Having only these two alternatives is too restrictive when we have 10, 20, or more years of data. If no smooth time-trend or covariate-explained variation occurs in the S_i then there will be no suitable intermediate structural model for these S_i to use as a parsimonious restriction on the general model. In reality, these annual survival rates surely vary over years. Hence, a useful model would be $S_i = E(S) + \epsilon_i$, with the ϵ_i treated as independent random variables, mean 0 variance σ^2. This is a 2-parameter random effects model for the S_i which is intermediate between the models with all S_i different (k parameters) and all S_i the same (1 parameter). This intermediate model allows inference about σ^2 and unconditional inference about $E(S)$, i.e., inference about $E(S)$ allowing for $\sigma^2 > 0$ rather than assuming $\sigma^2 = 0$. While the random effects model would seem to eliminate the individual S_i, in fact it leads to shrinkage estimates ($\widetilde{S}_i$) as an improved (in mean square error relative to unrestricted MLEs, $\hat{S}_i$) conditional inference about the the set of individual annual survival rates. This paper presents new results for random effects models embedded into classical capture-recapture models. Extended results are given wherein a general linear-structural model is imposed on the survival rates as $S_i = \underset{\sim}{x}_i' \underset{\sim}{\beta} + \epsilon_i$, or using a general link function. Such a structural model can represent time-trends or informative covariates, such as weather-related, or both types of structural fixed-effects, yet also allows for residual unexplained variation (i.e., the ϵ_i, hence σ^2) in the survival parameters, S_i. Analysis starts with unrestricted maximum likelihood estimates, $\hat{S}_i$, and their estimated conditional sampling variance-covariance matrix. The solutions, basically from variance components and shrinkage theory, are not closed-form; however, they require only matrix methods and a one-dimensional search over a function of the data and σ^2 to construct point and interval estimates. Finally, AIC is extended to apply to these random effects models, hence allowing unified AIC comparison of both fixed and simple random effects models for capture-recapture data.

Key Words: AIC, band recovery, bird banding, Cormack-Jolly-Seber models, recapture data, shrinkage estimates, variance components.

1. INTRODUCTION

The objectives of this paper are 1) 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 (Longford 1993; see also Carroll et al. 1995, Carlin and Louis 1996); 2) develop 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); and 3) extend AIC to simple random effects models embedded into the otherwise fixed-effects capture-recapture likelihood. It is assumed that the reader already has a basic knowledge of dead-recovery and live-recapture models, referred to here generically as just capture-recapture. The random effects idea is a simple and fundamental one in statistics. It basically admits that the parameters we estimate need sometimes to be considered as random variables.

Consider the 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 (see, e.g., Lebreton et al. 1992). If $k = 20$ or 30 we are adding many survival parameters into our model as if they were unrelated; however, more parsimonious models are often needed (Burnham and Anderson 1998, Link 1999). We can consider reduced explanatory forms for the between-occasion survival rates, $S_1, \ldots, S_k$. At one extreme we have the model $\{S,\ p_t\}$ wherein $S_1 = \cdots = S_k = S$. However, this model may not fit well even if the general 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. Instead, there may be unstructured time variation in the S_i that is not easily modeled by any classical 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 $E(S) = \mu$, and with some population variation, 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 $S_1, \ldots, S_k$ 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 $S_1, \ldots, S_k$ 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-ingnorable sampling variances and covariances, from a capture-recapture model wherein we traditionally consider the S_i as fixed, unrelated parameters.

This is the essence of a random effects model super-imposed on the survival rates in the time-specific CJS model: the individual S_i are acknowledged to vary about some mean value but not in a way we can further explain by any trends or covariates. Hence, the estimation and inference problem now extends to both $E(S)$ and σ^2, as conceptual population parameters applicable to other time periods or other locations. However, we continue to face the usual conditional inference about each S_i, conditional on the actual time periods and study location. Inference about survival rates under traditional CJS models are all conditional on the S_i, i.e., are based on $var(\hat{S}_i \mid S_i)$ which does not involve, or acknowledge, σ^2. If we knew each S_i

exactly (so $\hat{S}_i \equiv S_i$ and $var(\hat{S}_i \mid S_i) \equiv 0$) then we would only be faced with unconditional population-level inferences and we would surely use just $\hat{E}(S) = \overline{S}$ and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}(S_i - \overline{S})^2}{k-1}$$

to summarize the data that would be exactly $S_1, \ldots, S_k$.

Under the simplest capture-recapture model with $\mu = E(S)$ as an unconditional population-level structural parameter, then $\epsilon = S - E(S)$. Rather than use only such a simple random effects model we would like to further explain (i.e., understand) the variation in the S_i, such as by finding covariates, or time trends. Thus we would have models of the form $S_i = a + b\,i + \epsilon_i$ for a linear trend, or $S_i = a + bx_i + cy_i + \epsilon_i$ for some environmental covariates x and y, recorded annually. Whatever the structural model, the unexplained, unconditional (population-level) component of variation is measured by the expected variation in the theoretical residuals, hence by $E(\epsilon^2) = \sigma^2$ (thus, process variation is model-dependent). We can already fit such model structures to capture-recapture data as fixed effects models, i.e., assuming $\sigma^2 = 0$. Methodology is given here for fitting them as random effects models, hence estimating σ^2.

Presumably σ^2 applies to other times or locations. Inference to time intervals other than in the study at hand is always both problematic yet of great interest, as for example in population viability studies where results can depend strongly on the value of σ^2 (see e.g., White 2000). To fully understand population dynamics from recapture or banding data we must allow for process variation in our analysis of these data types (Anderson and Burnham 1976, Link and Nichols 1994, Catchpole et al. 1995). Such inferences, essentially to other time intervals, should be unconditional, hence should include an uncertainty component for σ^2 and must recognize that sample size for this inference is the number of years of the study, not the number of animals marked.

There are practical and philosophic issues that we will not consider here, such as covariates needing to be be measured without error. Also, as noted above, the value of σ^2 depends partly on the adequacy of the unconditional structural model used; σ^2 would be zero given a perfect structural model to "explain" the S_i. It is a deep philosophic issue as to whether there exists some set of fully-informative covariates such that any variations in the S_i are deterministic. "True" σ^2 will then be exactly zero. It suffices to know that σ^2 measures the average unexplained residual variation beyond the expected fit of any *unconditional* deterministic structural model used to explain the variation in the set of S_i.

The notation is chosen here for convenience. All that matters is that $S_1, \ldots, S_k$ represent a set of k estimable parameters of the same type, usually sequential in time. These parameters may be survival rates (S), apparent survival rates (usually denoted as ϕ in CJS models), sampling rates (p, f, or r), fidelity rates (F) in a joint recovery-recapture model (Burnham 1993, Barker 1997), abundance parameters (N, B) in the Jolly-Seber models (Schwarz and Arnason 1996), population finite rates of change (λ) from modified JS models (Pradel 1996), or they may be any of the aforementioned parameters transformed to some other scale based on a link function other than the identity link (cf. Lebreton et al. 1992, White and Burnham 1999). The concepts remain the same as regards separating average conditional sampling variance from process variance.

It is important to know that $var(\hat{S}_i \mid S_i)$ depends on sample size of marked animals in the study. This sampling uncertainty in the estimator $\hat{S}_i$, conditional on S_i, can in principle be driven to 0 by having a very large sample size of marked animals. In contrast, the process variation, σ^2, exists independent of the marking study and hence is independent of the sample size of animals. Precision and bias of $\hat{\sigma}^2$ does depend on sample size of marked animals. Inferences from capture-recapture data about population-level parameters (i.e., to other time periods or areas) that are based only on conditional sampling variation are assuming $\sigma^2 = 0$. Inferences may then be too liberal because relevant total variation is under-estimated. Treating σ^2 as being 0 is acceptable when sampling variances are much larger than σ^2. However, when this is not true, i.e., average $var(\hat{S}_i \mid S_i)$ does not dominate σ^2, we should estimate both sources of variation and make population-level inferences based on total (i.e., unconditional) variation. For $\hat{S}_i$ this would be, pragmatically, $\sigma^2 + var(\hat{S}_i \mid S_i)$. Theoretically, unconditional total variation for $\hat{S}_i$ is $\sigma^2 + E_S[var(\hat{S}_i \mid S_i)]$. This expectation, over S, of the sampling variance is not tractable; to first order we can use $\hat{E}_S[var(\hat{S}_i \mid S_i)] = \hat{var}(\hat{S}_i \mid S_i))$. However, the actual S_i extant during the study are not population-level parameters (as we here use the term), so inference about particular S_i should remain conditional.

We are lead therefore into a duality of thinking. We make population-level inferences (that include the uncertainty represented by σ^2) about population-level parameters under which thinking the S_i are random variables, and individually are not of interest. However, we also want to make the best inferences we can about the actual applicable S_i as conditional parameters irrespective of collective (i.e., population-level) properties of this set of survival rates. Interestingly enough we can have it both ways.

There is yet another concept that arises in these random effects models. The set of maximum likelihood estimators, $\hat{S}_1, \ldots, \hat{S}_k$, of the parameters $S_1, \ldots, S_k$ can be improved upon, in the sense of having smaller expected mean square error, by what are called shrinkage estimators, denoted here as $\tilde{S}_i$. For the simple model with only one population-level parameter, $E(S)$, $\tilde{S}_i$ lies between $\hat{E}(S)$ and $\hat{S}_i$ and the extent of this shrinkage towards $\hat{E}(S)$ depends upon the variance components proportion $\sigma^2/[\sigma^2 + E_S\{var(\hat{S}_i \mid S_i)\}]$ (this formula requires zero covariances for the $\hat{S}_i$). An individual $\tilde{S}_i$ may not improve upon the corresponding MLE $\hat{S}_i$ in the sense of being nearer to S_i in a given case, but overall the shrinkage estimators as a set are to be preferred as being closer to the true S_i if the random effects model applies with $\sigma^2 > 0$ (Efron and Morris 1975, Casella 1985).

Longford (1993) provides a general introduction to ideas about random effects in the context of structural models, including estimation of $E(S)$, or more generally $\underline{\beta}$ in $\mathrm{E}(S) = \underline{x}'\underline{\beta}$, and σ^2 and shrinkage estimators. Some other relevant statistical and ecological literature on these ideas is in Efron and Morris (1975), Johnson (1981, 1889), Morris (1983), Burnham et al. (1987), Robinson (1991), Carlin and Louis (1996), Ver Hoef (1996), Link (1999), and Kubokawa 1999).

There are other issues but this is the crux of the matter: fit models wherein we can estimate both the population-level process variation and population-level structural parameters that parsimoniously explain sets of related parameters such as $S_1, \ldots, S_k$. This is accomplished by the use of structural plus random effects models that are effectively intermediate between

assuming all $S_i \equiv S$ or having no constraints at all on $S_1, \ldots, S_k$. Simultaneously this framework allows improved conditional inferences, via shrinkage methods, about the S_i considered as year-specific conditional parameters.

Below we elaborate these concepts with an example in which there is independent sampling (measurement) variation on the S_i. Hence, each $\hat{S}_i$ is conditionally independent of all other $\hat{S}_j$. In capture-recapture models there are pairwise sampling correlations which complicate estimation of $E(S)$, σ^2, and shrinkage estimators, at least in the sense that the methods require expression in matrix algebra. In a subsequent section, using matrix methods, we derive theory for the general case of S_i being a linear structural model plus random error, ϵ, applicable to capture-recapture models. Those mathematics may seem difficult; we hope the concepts here are understandable.

2. EXAMPLE OF CONCEPTS AND METHODS

A simple example of random effects variance components is used here that illustrates the key ideas without the complications of capture-recapture. For $k = 10$ (think of as 10 years) we drew one sample of S_i distributed as independent normal random variables with mean $E(S) = 0.5$ and process variance $\sigma^2 = (0.05)^2$. For each year we simulated marking $n = 25$ birds and determining the number alive, y, after 1 year. Conditional on each S_i we generated y_i as an independent binomial(n, S_i) random variable. Hence $\hat{S}_i = y_i/n$ has conditional sampling variance $var(\hat{S}_i \mid S_i) = S_i(1 - S_i)/n$, about $(0.1)^2$. As an estimator we used $\hat{var}(\hat{S}_i \mid S_i) = \hat{S}_i(1 - \hat{S}_i)/(n - 1)$ because it is unbiased. Table 1 gives the results of the sample (by chance we got 0.0499 for the standard deviation of the 10 generated S_i). With real data we would not know S_i; we would only have $\hat{S}_i$, and generally only have $\hat{E}_S(var(\hat{S}_i \mid S_i))$ as $\hat{var}(\hat{S}_i \mid S_i)$. In this simple example we can determine $E_S(var(\hat{S}_i \mid S_i))$ but we have choose to keep the example mimicking what occurs in complex capture-recapture.

From Table 1 we see that the empirical standard deviation of the 10 estimated survival rates (i.e., the $\hat{S}_i$) is 0.106. We should not take $(0.106)^2$ as an estimate of σ^2 because such an estimate includes both process and sampling variation (i.e., includes the conditional binomial variation of the $\hat{S}_i$). Rather, we must subtract the estimated average sampling variance, $\overline{\hat{v}}\ ar(\hat{S}_i \mid S_i)$, from the total variation to get $\hat{\sigma}^2 = 0.011182 - 0.009987 = 0.001195$, or $\hat{\sigma} = 0.0346$. This estimator,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{k}(\hat{S}_i - \overline{\hat{S}})^2}{k-1} - \frac{\sum_{i=1}^{k}\hat{v}ar(\hat{S}_i|S_i)}{k},$$

is theoretically appropriate when the sampling errors, $\hat{S}_i - S_i$, are independent, conditional on the S_i. However, this estimator is neither the generally correct nor the most efficient one. A general method of moments approach is given below (for this case of independent estimators formulae are given in Burnham et al. 1987). Applying that more general theory we get for this example, by numerical methods, $\hat{\sigma} = 0.0394$ with a 95% confidence interval of 0 to 0.1663. As

is the case here, estimated variance components can be imprecise, mostly because of the often small sample size (k) for σ^2. Here that sample size is $k = 10$, not the $k \times n = 250$ marked birds.

In addition to $\hat{\sigma}$, we find here $\hat{E}(S) = 0.4825$ with an estimated unconditional standard error of 0.0339. If we just take the mean of the MLEs and base it's standard error on the set of $\hat{v}ar(\hat{S}_i \mid S_i)$ we get a conditional standard error estimate of 0.0316; this is equivalent to assuming $\sigma^2 = 0$. While the difference in the two standard error estimates is numerically trivial here, the conceptual difference is quite important, especially in making a proper unconditional inference about $E(S)$. The distinction does not matter much in practice when the conditional sampling variances are a lot larger than the process variation – an all too common occurrence.

The most dramatic difference in point estimates, and precision, occurs with the shrinkage estimates of the yearly survival rates. Table 1 shows true S_i, the usual $\hat{S}_i$ (which are MLEs), $\hat{s}e(\hat{S}_i \mid S_i)$, and shrinkage results, $\tilde{S}_i$, $\hat{s}e(\tilde{S}_i \mid \hat{S}_i)$ and $\hat{r}mse(\tilde{S}_i \mid \hat{S}_i)$ (the latter we will explain below). The shrinkage estimator used in this example is

$$\tilde{S}_i = \hat{E}(S) + \sqrt{\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + \hat{v}ar(\hat{S}_i|S_i)}} \times \left[\hat{S}_i - \hat{E}(S)\right] . \qquad (1)$$

Figure 1 provides a graphical display of the MLEs, the shrinkage estimates and $\hat{E}(S)$ from the simple random effects model.

The shrinkage method given in (1) is so called, as one reason, because each residual, $\hat{S}_i - \hat{E}(S)$, arising from the fitted reduced-parameter model, is "shrunken" 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 MLEs by removal of the sampling variation. A shrinkage coefficient different from $\sqrt{\hat{\sigma}^2/(\hat{\sigma}^2 + \hat{v}ar(\hat{S}_i \mid S_i))}$ could be used. The chosen shrinkage coefficient has a very desirable property: if we treat the $\tilde{S}_i$ as if they were a simple random sample, then their sample variance almost exactly equals $\hat{\sigma}^2$. This also means that a plot of the shrinkage residuals, such as is implicit in Fig. 1, gives a correct visual image of estimated process variation in the S_i.

As shown in Table 1, the improvement gained by the shrinkage estimators ($\tilde{S}_i$) appears substantial, they have here about 50% better precision: compare $\hat{s}e(\tilde{S}_i \mid S_i)$ to $\hat{s}e(\hat{S}_i \mid S_i)$. However, because the MLEs here are unbiased and the shrinkage estimators are biased, 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 the example in Table 1, for the MLEs, the SSE is

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

for the shrinkage estimates the SSE is

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

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

To evaluate the two MSEs here we did 10,000 independent simulation trials of this example situation. In each trial a different random set of S_i was generated as noted above. The MSE results, precise to two significant digits, are

$$\hat{E}\left[\sum_{i=1}^{10}(\hat{S}_i - S_i)^2\right] = 0.0990,$$

$$\hat{E}\left[\sum_{i=1}^{10}(\widetilde{S}_i - S_i)^2\right] = 0.0469.$$

Moreover, in 98% of the 10,000 trials the shrinkage estimators were closer, in the SSE measure, to the set of true S_i than were the MLEs.

For the MLEs an approximate 95% confidence interval on S_i is given by $\hat{S}_i \pm 2\hat{s}e(\hat{S}_i \mid S_i)$; this procedure will have good coverage in this example. However, for the shrinkage estimator if we use $\widetilde{S}_i \pm 2\hat{s}e(\widetilde{S}_i \mid S_i)$ coverage will be negatively affected by the bias of $\widetilde{S}_i$. In fact, coverage was about 83% here for any i, based on the 10,000 simulation trials. This is in accord with theory (see Särndal et al. 1992, page 165, Table 5.1) as here the expected $|\text{bias}|/\text{se}$ ratio for the shrinkage estimators is about 1. Essentially correct expected coverage occurs for the interval $\widetilde{S}_i \pm 2\hat{r}mse(\widetilde{S}_i \mid S_i)$ where

$$\hat{r}mse(\widetilde{S}_i \mid S_i) = \sqrt{\hat{v}ar(\widetilde{S}_i \mid S_i) + (\widetilde{S}_i - \hat{S}_i)^2}\,. \tag{2}$$

The expectation over S_i of $\hat{m}se(\widetilde{S}_i \mid S_i) = [\hat{r}mse(\widetilde{S}_i \mid S_i)]^2$ is approximately the mean square error for $\widetilde{S}_i$, MSE_i.

For the MLE, $\hat{r}mse(\hat{S}_i \mid S_i) = \hat{s}e(\hat{S}_i \mid S_i)$ because $\hat{S}_i$ is unbiased. The unbiasedness of the MLE in the general model, together with a high correlation between $\hat{S}_i$ and $\widetilde{S}_i$, and the framework that S_i is random, allows an argument that $\hat{r}mse(\widetilde{S}_i \mid S_i)$ is an estimator of the unconditional sampling standard error of $\widetilde{S}_i$ over conceptual repetitions of the data. It follows that this *rmse* can be a correct basis for a reliable confidence interval. It is rare to have a reliable estimator of MSE for a biased estimator, but when this occurs it makes sense to use, as a basis for a 95% confidence interval, $\pm 2\sqrt{\hat{\text{MSE}}}$ rather than $\pm 2\hat{\text{se}}$. This situation has been noted in Särndal et al. (1992, pages 165-166); they assert there is good large-sample coverage.

Next we provide the general theory, on which this example and a second example below are based.

3. SOME GENERAL THEORY FOR RANDOM EFFECTS

We present some general motivating results and then equations for random effects estimation. A method of moments approach is used here to avoid distributional assumptions.

3.1 Informative Heuristic Considerations

First, we consider some heuristics motivated by the above simple framework of exchangeable random variables in k dimensions. We simplify the problem to one dimension as follows. For given $\mu = E(S)$, σ^2 and $v = E_S[var(\hat{S} \mid S)]$, consider an estimator of S as $\tilde{S} = c(\hat{S} - \mu) + \mu$ where nature generates a random S and then we observe $\hat{S}$ conditional on S; $\hat{S}$ is, conditionally on S, unbiased for all values of S. The value of the constant c is to be determined. The appropriate MSE of $\tilde{S}$ to consider is

$$\text{MSE} = E_S\left[E_{\tilde{S}}\left[(\tilde{S} - S)^2 \mid S\right]\right];$$

$\tilde{S}$ is a function of $\hat{S}$ so expectations can be equally well denoted as being with respect to $\tilde{S}$ or $\hat{S}$. This leads us to write $\tilde{S} - S = c(\hat{S} - S + S - \mu) + \mu - S = c(\hat{S} - S) + (c-1)(S - \mu)$ and hence

$$(\tilde{S} - S)^2 = c^2(\hat{S} - S)^2 + (c-1)^2(S-\mu)^2 + 2c(c-1)(\hat{S} - S)(S - \mu).$$

The cross product term above has conditional expectation, of $\hat{S}$ given S, of 0. Therefore,

$$E_{\tilde{S}}\left[(\tilde{S} - S)^2 \mid S\right] = c^2 v + (c-1)^2(S-\mu)^2,$$

hence

$$E_S\left[E_{\tilde{S}}\left[(\tilde{S} - S)^2 \mid S\right]\right] = c^2 v + (c-1)^2\sigma^2 = \text{MSE} \tag{3}$$

(note that $E_S\left[var(\tilde{S} \mid S)\right] = c^2 v$, $E_S\left[[bias(\tilde{S} \mid S)]^2\right] = (c-1)^2\sigma^2$). Now we can determine c that minimizes MSE.

The derivative of (3) with respect to c, set to zero, gives $0 = 2c\,v + 2(c-1)\sigma^2$ which has unique solution $c = \sigma^2/(\sigma^2 + v)$. This motivates a minimum MSE shrinkage estimator of the form

$$\tilde{S} = \mu + \left[\tfrac{\sigma^2}{\sigma^2+v}\right](\hat{S} - \mu), \tag{4}$$

whereas (1) is of the form

$$\tilde{S} = \mu + \sqrt{\tfrac{\sigma^2}{\sigma^2+v}}\,(\hat{S} - \mu). \tag{5}$$

Heuristically then, (1) is not likely to give the minimum MSE $\tilde{S}$ (in simulations of the above normal-binomial type, form (4) did have somewhat smaller MSE than form (5)).

Other considerations might apply, rather than just to use a minimum MSE estimator. The process variation is $\sigma^2 = E(S-\mu)^2$. If our shrinkage estimator, $\tilde{S}$, is to be our best estimator of S, we could ask that it satisfy $\sigma^2 = E(\tilde{S} - \mu)^2$. By the same methods as above, $E(\tilde{S} - \mu)^2 = c^2(\sigma^2 + v)$, hence if we want $\sigma^2 = E(\tilde{S} - \mu)^2$, we should use (5), that is, use

$c = \sigma/\sqrt{\sigma^2 + v}$ so that the average variation of the shrinkage estimate is σ^2. By way of contrast, under form (4),

$$E(\widetilde{S} - \mu)^2 = \sigma^2 \left[\frac{\sigma^2}{\sigma^2+v}\right]$$

so this shrinkage estimator is, in a sense, over-shrunk and hence does not reliably related directly to σ^2. This over-shrinkage of the standard shrinkage estimator (4) has been criticized, in a more general setting, by Louis (1984), who also then gave and noted the advantage of (5).

The actual MSEs of the two cases are, for (4),

$$\text{MSE} = v \left[\frac{\sigma^2}{\sigma^2+v}\right]$$

(hence the minimum MSE), and for (5)

$$\text{MSE} = 2\sigma^2 \left[1 - \sqrt{\frac{\sigma^2}{\sigma^2+v}}\right].$$

The above two MSE formulae allow one to compute here exact MSE ratios; these MSE ratios depend only upon σ^2/v. For example, MSE ratios of (4) or (5) relative to $\text{MSE}(\hat{S}) = \sigma^2 + v$ are below:

σ^2/v	(4)	(5)
0.2	0.139	0.197
0.5	0.222	0.282
1	0.250	0.293
2	0.222	0.245
5	0.139	0.145

The take-away message from the above numbers is that use of (5), rather than (4), does not entail a drastically increased MSE; a version of formula (5) is what we are herein recommending. Next we consider a confidence interval on S.

We can expect a standard confidence interval on S based on the MLE, such as $\hat{\text{S}} \pm 2\sqrt{v}$ to have good large-sample coverage. However, from the shrinkage estimator if we use $\widetilde{S} \pm 2\,c\,\sqrt{v}$ coverage will suffer because the conditional bias of $\widetilde{S}$ is $bias(\widetilde{S} \mid S) = (c-1)(S-\mu)$. We cannot just subtract this bias from $\widetilde{S}$ because $bias(\widetilde{S} \mid S)$ is unknown. We could subtract the estimated bias from $\widetilde{S}$ but then the confidence interval should allow that the bias was estimated (and in general c and μ are also estimated), and the sign of the estimated bias might even be wrong.

An alternative approach is to accept a type of unconditional-conditional confidence interval on S, which we sort of must do anyway because our confidence interval must have good coverage over all possible values of S. Hence we should accept as a pivotal quantity $[\widetilde{S} - bias(\widetilde{S} \mid S)] - S$. Now we treat all of $[\widetilde{S} - bias(\widetilde{S} \mid S)]$ as random and we compute its expected (over S) sampling variance, or what is analogous to a sampling variance here: $c^2\,v + E_S[(bias(\widetilde{S} \mid S))^2] = c^2\,v + (c-1)^2\sigma^2 = \text{MSE}$. Thus, we are motivated to use a confidence interval as $\widetilde{\text{S}} \pm 2\,\sqrt{\hat{\text{MSE}}}$.

Now the problem is to get a good estimate of $E_S[(bias(\widetilde{S} \mid S))^2]$ or, just as well, of $(bias(\widetilde{S} \mid S))^2$. The obvious $(\hat{c}-1)^2\hat{\sigma}^2$ is too variable (e.g., $\hat{\sigma}^2$ can be 0). Also, we may be better served by something that better estimates conditional squared bias. In this regard note the obvious: $\hat{S} = \widetilde{S} + (\hat{S} - \widetilde{S})$; hence, $\hat{S} - S = (\widetilde{S} - S) + (\hat{S} - \widetilde{S})$ and therefore

$$0 = E_{\hat{S}}[(\hat{S} - S) \mid S] = E_{\hat{S}}[(\widetilde{S} - S) \mid S] + E_{\hat{S}}[(\hat{S} - \widetilde{S}) \mid S].$$

The above means that $bias(\widetilde{S} \mid S) = -E_{\hat{S}}[(\hat{S} - \widetilde{S}) \mid S]$. Hence, an estimator of $(bias(\widetilde{S} \mid S))^2$ is simply $(\widetilde{S} - \hat{S})^2$ and we may take as a sort of effective sampling variance on $\widetilde{S}$, for purposes of a confidence interval on S,

$$\hat{mse}(\widetilde{S} \mid S) = \hat{var}(\widetilde{S} \mid S) + (\widetilde{S} - \hat{S})^2.$$

This is leads to (2) and thus $\widetilde{S}_i \pm 2\hat{rmse}(\widetilde{S}_i \mid S_i)$ for an approximate 95% confidence interval. Using such a confidence interval based on estimated mean square error, for a biased estimator, is noted to have good coverage properties by Särndal et al. (1992, page 165).

3.2 General Random Effects Inference Theory

The estimators $\hat{S}_1, \ldots, \hat{S}_k$ from capture-recapture are pairwise conditionally correlated, whereas most random effects theory assumes conditional independence of the basic estimators. Thus, extended theory is needed. It is necessary to use matrix methods. Vectors are underlined; all are column vectors. A matrix, X, may be a vector if it has only one column, but we do not then underline X.

We assume $\underline{\hat{S}} = \underline{S} + \underline{\delta}$. Conditional on $\underline{S}$, $\underline{\delta}$ has a zero expectation, hence $E(\underline{\hat{S}} \mid \underline{S}) = \underline{S}$. Also, $\underline{\delta}$ has conditional sampling variance-covariance matrix W, which will generally be a complicated function of $\underline{S}$ and other parameters (such as p or f). The optimal theory requires the unconditional W, $E_{\underline{S}}(W)$, which we do not expect to know.

Unconditionally $\underline{S}$ is a random vector with population-level expectation $X\underline{\beta}$ and variance-covariance matrix $\sigma^2 I$ (generalizing this assumption is not attempted here, but one might worry that the temporal nature of the S_i induces serial correlations). In the simplest case, X is a $k \times 1$ vector of ones and $\underline{\beta}$ is just $E(S)$. By assumption, the process residuals, $\epsilon_i = S_i - E(S_i)$, are independent with homogeneous variance σ^2. Also, we assume mutual independence of sampling errors $\underline{\delta}$ and process errors $\underline{\epsilon}$; this is not a restrictive assumption. We envision fitting a capture-recapture model that does not constrain $\underline{\hat{S}}$ (probably model $\{S_t\}$ for these parameters, but other models also apply), and hence get the MLE $\underline{\hat{S}}$ and the usual likelihood-based estimator of W as an estimator of $E_{\underline{S}}(W)$.

Let $\underline{S}$ be a vector with k elements; let $\underline{\beta}$ have r elements. Let VC denote a variance-covariance matrix. Unconditionally,

$$\underline{\hat{S}} = X\underline{\beta} + \underline{\delta} + \underline{\epsilon}, \quad \mathrm{VC}(\underline{\delta} + \underline{\epsilon}) = D = \sigma^2 I + E_{\underline{S}}(W).$$

We need to estimate $\underline{\beta}$ and σ^2, an unconditional variance-covariance matrix for $\hat{\underline{\beta}}$, compute a confidence interval on σ^2 (on $k-r$ *df*), and compute the shrinkage estimator of $\underline{S}$, $\widetilde{\underline{S}}$, and its conditional sampling variance-covariance matrix. Without any further information or context (like random effects), the MLE is the best conditional (on $\underline{S}$) estimator of $\underline{S}$. However, once we add the random effects structure we can consider an improved (smaller MSE) estimator of $\underline{S}$ (i.e., $\widetilde{\underline{S}}$) for finite sample sizes.

From generalized (weighted) least squares theory (see, e.g., Seber 1977, 1984) for σ^2 given, the best linear unbiased estimator of $\underline{\beta}$ is

$$\hat{\underline{\beta}} = (X'D^{-1}X)^{-1}X'D^{-1}\hat{\underline{S}} . \tag{6}$$

Note that here D ($= \sigma^2 I + E_{\underline{S}}(W)$), hence $\hat{\underline{\beta}}$, is actually a function of σ^2. Assuming normality of $\hat{\underline{S}}$ (approximate normality suffices) then from the same generalized least squares theory the weighted residual sum of squares $(\hat{\underline{S}} - X\hat{\underline{\beta}})'D^{-1}(\hat{\underline{S}} - X\hat{\underline{\beta}})$ has a central chi-squared distribution on $k-r$ degrees of freedom. Therefore, a method of moments estimator of σ^2 is obtained by solving the equation

$$k-r = (\hat{\underline{S}} - X\hat{\underline{\beta}})'D^{-1}(\hat{\underline{S}} - X\hat{\underline{\beta}}) . \tag{7}$$

Equation (7) defines a 1-dimensional numerical solution search problem. Pick a value of σ^2, compute D, then compute $\hat{\underline{\beta}}$ from equation (6), then compute the right hand side of (7). Repeat the process over values of σ^2 until the solution of (7), as $\hat{\sigma}^2$, is found. This also gives $\hat{\underline{\beta}}$. The theoretical unconditional sampling variance-covariance of $\hat{\underline{\beta}}$ is

$$\mathrm{VC}(\hat{\underline{\beta}}) = (X'D^{-1}X)^{-1} . \tag{8}$$

Formula (7) can be simplified by eliminating $\hat{\underline{\beta}}$. First, define

$$A = X(X'D^{-1}X)^{-1}X'.$$

This is the “hat" matrix in linear regression, but we choose to not call it H. Now (7) is

$$k-r = \hat{\underline{S}}'(D^{-1} - D^{-1}AD^{-1})\hat{\underline{S}} .$$

To get a confidence interval on σ^2 we use

$$RSS(\sigma^2) = (\hat{\underline{S}} - X\hat{\underline{\beta}})'D^{-1}(\hat{\underline{S}} - X\hat{\underline{\beta}})$$

as a pivotal quantity. As a function of σ^2 *RSS* is monotonic decreasing. For true σ^2, *RSS* is distributed as central chi-squared on $df = k-r$. A 95% confidence interval on σ^2 is found as the solution points in σ^2 (lower and upper bounds on σ^2, respectively) of $RSS =$ upper 97.5 percentile point of a central χ^2_{df} and $RSS =$ lower 2.5 percentile point of a central χ^2_{df}. As σ^2

goes to ∞, $RSS(\sigma^2)$ goes to 0, hence a finite upper confidence interval always exists. The lower bound on σ^2 is the negative of the smallest eigenvalue of $\hat{E}_{\underline{S}}(W)$. The lower confidence limit can be negative; in fact, even the point estimate and upper confidence limit can be negative. In practice, truncate negative solutions to 0.

Define another matrix as

$$H = \sigma D^{-1/2} = \sigma\,(\sigma^2 I + \hat{E}_{\underline{S}}(W))^{-1/2} = (I + \tfrac{1}{\sigma^2}\hat{E}_{\underline{S}}(W)\,)^{-1/2};$$

we will only need, and use, H evaluated at $\hat{\sigma}$. The shrinkage estimate of $\underline{S}$ that we recommend is

$$\begin{aligned} \tilde{\underline{S}} &= H(\hat{\underline{S}} - X\hat{\underline{\beta}}\,) + X\hat{\underline{\beta}}\,, \\ &= H\hat{\underline{S}} + (I - H\,)X\hat{\underline{\beta}}\,. \end{aligned} \tag{9}$$

To get an estimator of the conditional variance of these shrinkage estimators (not exact as the estimation of σ^2 is ignored here, as it is in VC($\hat{\underline{\beta}}$) in (8)) we can define, and compute, a projection matrix G as below:

$$G = H + (I - H\,)AD^{-1}\,; \tag{10}$$

then

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

The theoretical variance-covariance matrix of the shrinkage estimator is

$$\mathrm{VC}(\tilde{\underline{S}} \mid \underline{S}\,) = G\hat{E}_{\underline{S}}(W)G'\,.$$

We propose that comparison to the MLEs, and confidence intervals, should be based on the matrix

$$\hat{\mathrm{M}}\mathrm{SCP}(\tilde{\underline{S}} \mid \underline{S}\,) = \hat{\mathrm{V}}\mathrm{C}(\tilde{\underline{S}} \mid \underline{S}\,) + (\tilde{\underline{S}} - \hat{\underline{S}}\,)(\tilde{\underline{S}} - \hat{\underline{S}}\,)'\,. \tag{11}$$

In particular, from the diagonal elements of (11) we get

$$\hat{rmse}(\tilde{S}_i \mid \underline{S}\,) = \sqrt{\hat{var}(\tilde{S}_i \mid \underline{S}\,) + (\tilde{S}_i - \hat{S}_i)^2}\,. \tag{12}$$

For this method of moments shrinkage estimator we can show, fairly directly from (7) and (9), that

$$\hat{\sigma}^2 = \frac{(\tilde{\underline{S}} - X\hat{\underline{\beta}})'(\tilde{\underline{S}} - X\hat{\underline{\beta}})}{k - r}\,.$$

Hence, the average sum of squares of the shrunk residuals (i.e., $\tilde{\underline{S}} - X\hat{\underline{\beta}}$) produces $\hat{\sigma}^2$. This coherent relationship does not hold for the usual shrinkage estimate found in the statistical

literature (Morris 1983, Louis 1984): $\underline{\tilde{\tilde{S}}} = H^2(\underline{\hat{S}} - X\underline{\hat{\beta}}) + X\underline{\hat{\beta}}$. Moreover, in a bias-precision trade-off there is no compelling reason to think minimum MSE must be the appropriate solution. In fact, the shrinkage estimator defined here is central to being able to obtain a useful, simple extension of AIC to this random effects model. If we used ordinary least squares to fit the model $\underline{\tilde{S}} = X\underline{\beta} + \underline{\epsilon}$ and from this fit computed the residual sum of squares divided by its *df* we would get, essentially, $\hat{\sigma}^2$ as the result. Therefore, in this sense we can say that the shrinkage estimators from (9) are a complete summary of the fitted random effects model and thus provide a basis to evaluate the likelihood, and hence AIC, of the fitted random effects model.

4. 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 $\mathcal{L}(\underline{S}, \underline{\theta}) = \mathcal{L}(S_1, \ldots, S_k, \theta_1, \ldots, \theta_\ell)$. Under this time-specific model, $\{S_t, \theta_t\}$, we have the MLEs, $\underline{\hat{S}}$ and $\underline{\hat{\theta}}$, and the maximized log-likelihood, $\log\mathcal{L}(\underline{\hat{S}}, \underline{\hat{\theta}})$ based on $K = k + \ell$ parameters. Thus (for large sample size, n), AIC for the time-specific model is $-2\log\mathcal{L}(\underline{\hat{S}}, \underline{\hat{\theta}}) + 2K$. The likelihood value for the fitted random effects model is

$$\mathcal{L}(\underline{\tilde{S}}, \underline{\tilde{\hat{\theta}}}) \equiv \mathcal{L}(\underline{\tilde{S}}, \underline{\hat{\theta}}(\underline{\tilde{S}})) = \max_{\underline{\theta}} \mathcal{L}(\underline{\tilde{S}}, \underline{\theta}), \tag{13}$$

where $\underline{\tilde{S}}$ is fixed. A first approximation to this maximized likelihood is just $\mathcal{L}(\underline{\tilde{S}}, \underline{\hat{\theta}})$, where $\underline{\hat{\theta}}$ is the MLE under model $\{S_t, \theta_t\}$. Experience has shown that $\log\mathcal{L}(\underline{\tilde{S}}, \underline{\tilde{\hat{\theta}}})$ and $\log\mathcal{L}(\underline{\tilde{S}}, \underline{\hat{\theta}})$ are often quite close; however, the re-optimization on $\underline{\theta}$ is theoretically appropriate and empirically worth doing (based on Monte Carlo simulation evaluation of these methods using program MARK, White and Burnham 1999).

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

$$K_{re} = \mathrm{tr}(G) + \ell, \tag{14}$$

where G is the projection matrix (formula 10) mapping $\underline{\hat{S}}$ into $\underline{\tilde{S}}$ and $\mathrm{tr}(\cdot)$ is the matrix trace function: $\mathrm{tr}(G) =$ sum of the diagonal elements of G. The mapping $G\underline{\hat{S}} = \underline{\tilde{S}}$ is a type of generalized smoothing. It is know that the effective number of parameters to associate with such smoothing is the trace of the smoother matrix (see e.g., Hastie and Tibshinari 1990, section 3.5).

From (13) and (14), the large-sample AIC for the random effects model is

$$-2\log\mathcal{L}(\underline{\tilde{S}}, \underline{\tilde{\hat{\theta}}}) + 2K_{re} \tag{15}$$

(a proof is sketched below). The more exact version, $\mathrm{AIC_c}$, for the random effects model may, by analogy, be taken as

$$-2\log\mathcal{L}(\underline{\widetilde{S}}\,,\underline{\widetilde{\hat{\theta}}}\,)+2K_{re}+2\tfrac{K_{re}(K_{re}+1)}{n+K_{re}-1)}\,. \tag{16}$$

Results like these are in the literature for AIC generalized to semi-parametric smoothing, see e.g., Hurvich and Simonoff (1998) and Shi and Tsai (1998); these papers are not about random effects models. Instead, these papers note a generalized AIC where the effective number of parameters is the trace of a smoothing matrix.

A detailed derivation of AIC is given in Burnham and Anderson (1998), section 6.2. Using the same notation as in that section, a sketch of part of the derivation of (15) is given here. The MLE under the full time-effects model is

$$\underline{\hat{\tau}} \;=\; \begin{bmatrix} \underline{\hat{S}} \\ \underline{\hat{\theta}} \end{bmatrix}\,.$$

The estimator of τ under the random effects model is

$$\underline{\widetilde{\tau}} \;=\; \begin{bmatrix} \underline{\widetilde{S}} \\ \underline{\hat{\theta}} \end{bmatrix} = \begin{bmatrix} G & O \\ O & I \end{bmatrix} \underline{\hat{\tau}} \;= P\underline{\hat{\tau}}\,.$$

Under Kullback-Leibler based model selection (the foundation of AIC) we need to find an estimator of a target quantity T, defined in Burnham and Anderson (1998); $I(\underline{\tau}_o)$ defined there is also needed here. The data are represented here by vector $\underline{x}$. The key result as regards AIC for the random effects model is then

$$T = E[\log\mathcal{L}(\underline{\widetilde{\tau}}\,(\underline{x}\,))] - \mathrm{tr}[I(\underline{\tau}_o)E\,[(\underline{\widetilde{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)(\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)']]\;; \tag{16}$$

the MLE of τ under the time-effects model is consistent for τ_o . The derivation of (16) is left out here as it is long and involved, it is just a variation on the AIC derivation in Burnham and Anderson (1998). AIC is $-2T$, hence for the random effects model

$$\mathrm{AIC} = -2\log\mathcal{L}(\underline{\widetilde{\tau}}\,(\underline{x}\,)) + 2\hat{\mathrm{tr}}\,[I(\underline{\tau}_o)E\,[(\underline{\widetilde{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)(\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)']]\,.$$

The expectation under the trace operator is basically a covariance matrix. Expanding that term we have

$$(\underline{\widetilde{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)(\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)' = (P\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)(\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)' =$$

$$(P\underline{\hat{\tau}}\,(\underline{x}\,)-P\underline{\tau}_o+P\underline{\tau}_o-\underline{\tau}_o)(\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)' =$$

$$P(\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)(\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)' + (P\underline{\tau}_o-\underline{\tau}_o)(\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)'\,.$$

Taking the the asymptotic (large-sample) expectation over $\underline{x}$ with respect to truth, wherein $E(\underline{\hat{\tau}}\,(\underline{x}\,)) = \underline{\tau}_o$,

$$E[P(\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)(\underline{\hat{\tau}}\,(\underline{x}\,)-\underline{\tau}_o)'] = P\Sigma\,.$$

The matrix Σ is the asymptotic variance-covariance matrix of $\hat{\underline{\tau}}$. Thus, we get

$$\text{AIC} = -2\log\mathcal{L}(\underline{\tilde{\tau}}) + 2\hat{\text{tr}}[I(\underline{\tau}_o)P\Sigma]$$

$$= -2\log\mathcal{L}(\underline{\tilde{\tau}}) + 2\hat{\text{tr}}[P\Sigma I(\underline{\tau}_o)] .$$

If the model is "good" (close to truth), then $\Sigma \doteq (I(\underline{\tau}_o))^{-1}$ so $\Sigma I(\underline{\tau}_o)$ is an identity matrix and

$$\text{AIC} = -2\log\mathcal{L}(\underline{\tilde{\tau}}) + 2\text{tr}(P) .$$

(The issue of AIC and estimation of this type of trace for model selection is discussed in Burnham and Anderson 1998; the pragmatic assumption is made that the set of models considered contains at least one good model, but not the "true" model). Finally,

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

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

5. TWO EXAMPLES

5.1 A Sage Grouse Band Recovery Example

Band recovery data are used here from a 15-year study (hence $k = 14$) of a non-migratory population of sage grouse (*Centrocercus urophasianus*) banded in North Park, Colorado (a large mountain valley in north-central Colorado). The study, conducted by Clait Braun (Colorado Division of Wildlife), was from 1973 to 1988. These data were analyzed in the M.S. thesis of M. Zablan (1993) before program MARK existed; the analyses reported here were done with MARK. We use just the recovery data from subadult males, of which 1,777 were banded, in Spring, on the leks. There were 312 band recoveries obtained from hunters; hunting was in Fall. These data are well-fit by model $\{S_t , r_t\}$ ($\{S_t , f_t\}$ in the notation of Brownie et al. 1985; see White and Burnham 1999 for the notation and parameterizations used by MARK). The goodness-of-fit $\chi^2 = 13.43$ on 14 df, from program ESTIMATE (Brownie et al. 1985), which is callable from program MARK. Thus we use this time-specific model as our global model for further illustrative random-effects analyses of the survival rates (full time variation was kept on the band recovery rates).

With many years of data the AIC-based selection of model $\{S, r_t\}$, rather than $\{S_t , r_t\}$, can occur because the addition of many parameters beyond just one constant S is not warranted: the "cost" in terms of lack of parsimony is too great. The random-effects model is a legitimate intermediary model. Even if model $\{S_t , r_t\}$ is selected one should also fit the random-effects model and consider using the shrinkage estimates rather than the MLEs. For these subadult male grouse data, the ΔAIC_c values for five models are below:

Model	ΔAIC$_c$	K	Comments re model for survival rates, S_i
$\{S_T\, ,\, r_t\}$	0.00	17	fixed effects, linear time trend
$\{S_{T,\sigma}\, ,\, r_t\}$	0.86	17.00	random effects, linear time trend
$\{S\, ,\, r_t\}$	4.98	16	constant
$\{S_{\mu,\sigma}\, ,\, r_t\}$	6.16	19.96	random time effects
$\{S_t\, ,\, r_t\}$	18.21	29	fixed time effects

The notation $\{S_T\, ,\, r_t\}$ denotes a model with a linear trend imposed on the survival rates. Hence, $S_i = a + bi$ is enforced by substituting $a + bi$ for S_i in the likelihood.

If one looked only at models $\{S, r_t\}$ and $\{S_t\, ,\, r_t\}$ the AIC choice would be strongly in favor of no time effects on survival: an AIC$_c$ difference of 13.23. Selection of that simpler model does not mean we believe there are no time effects. Rather, it means the data do not support having 13 more survival parameters estimated in the fully time-specific model.

The random survival-effects model is denoted as $\{S_{\mu,\sigma}\, ,\, r_t\}$ because in its pure (i.e., marginal likelihood) form it has only two parameters, μ and σ^2, relating to the annual survival rates. All the time variation in the 14 S_i is swept into one parameter (legitimate if the S_i behave like a random sample). Here, model $\{S_{\mu,\sigma}\, ,\, r_t\}$ produces $\hat{\bar{E}}(S) = 0.430$ ($\hat{se} = 0.027$) and $\hat{\sigma} = 0.046$ (95% confidence interval 0 to 0.185). Table 2 gives the MLEs $\hat{S}_i$ under model $\{S_t\, ,\, r_t\}$ and the shrinkage estimates $\tilde{S}_i$ under the random-effects version of model $\{S_t\, ,\, r_t\}$. Figure 2 provides a plot of these estimates. The $\tilde{S}_i$ show temporal variation that corresponds to the magnitude of $\hat{\sigma}$. Moreover, the general temporal pattern we see in the MLEs is mirrored (but muted) in the shrinkage estimates.

Given that we want parsimonious estimates of the S_i, if we otherwise accept model $\{S_t\, ,\, r_t\}$, we should use not the MLEs but rather the shrinkage estimates in Table 2. However, given that $\hat{\sigma}^2 > 0$, the correct assessment of the uncertainty of $\tilde{S}_i$ (and hence confidence intervals) should be based on $\hat{rmse}(\tilde{S}_i \mid S_i)$ in Table 2, not $\hat{se}(\tilde{S}_i \mid S_i)$, as would classically be done. Assuming we use simple confidence intervals of the form $\hat{S} \pm 2\hat{se}$ and $\tilde{S} \pm 2\hat{rmse}$, then the ratio of average confidence interval length based on $\tilde{S}_i$ vs based on $\hat{S}_i$ is $0.111/0.163 = 0.68$; this is a substantial improvement. If we do not need to make specific estimates of each S_i it suffices to report only the two population-level parameter estimates, $\hat{\bar{E}}(S)$ and $\hat{\sigma}$ and their confidence intervals.

We go further with this example for illustrative purposes. Say we wanted, *a-priori* to seeing the data, to examine a model for a simple linear trend in the S_i but with random residual effects about the fitted trend line. To denote this model we use the notation $\{S_{T,\sigma}\, ,\, r_t\}$. Here we are saying that even if there is a linear trend we expect the true S_i do not fall exactly on the trend line. Rather, the situation would be dealt with using standard linear regression if we knew the S_i without a measurement (sampling) variance component. We would then also estimate average residual variation about the fitted line based on $\hat{\sigma}^2$. Inference about $\hat{\sigma}^2$ would use a sample size of k (= 14), not infinity.

However, when we approach such a model by embedding its structure directly into the likelihood as $S_i = a + bi$ (possible via a nonlinear link function) we are forcing the model to accept the condition $\sigma^2 = 0$ rather than including that additional variance component parameter in the inference problem and we proceed as if the sample size of survival rates is infinite, not k.

With the random effects approach we can estimate σ^2 and if we get $\hat{\sigma}^2 > 0$ (we may not if sampling variance is large and the model structure is a good fit) we can get more suitable unconditional variances and also shrinkage estimators about the fitted structural model.

Table 3 gives the MLEs $\hat{S}_i$ under model $\{S_t, r_t\}$, these are the inputs to the random effects model $\{S_{T,\sigma}, r_t\}$ that produced the shrinkage estimates $\tilde{S}_i$ in Table 3. For the random-effects linear time trend model, $\hat{a} = 0.577$ ($\hat{se} = 0.056$), $\hat{b} = -0.021$ ($\hat{se} = 0.007$), and $\hat{\sigma} = 0$ (95% confidence interval 0 to 0.10). Because here $\hat{\sigma} = 0$, the shrinkage estimates exactly satisfy $\tilde{S}_i = \hat{a} + \hat{b}i$. Figure 3 shows this more clearly than Table 3. We do not interpret the result to mean that σ is zero in reality. Rather, we have learned that here the fitted linear time trend accounts for all the explainable variation in the $\hat{S}_i$, the MLEs from model $\{S_t, r_t\}$. Sampling variation is not "explainable." Hence, relative to sampling variation and covariation there is no discernible lack of fit, of the S_i to the fitted line, as would be measured by having $\hat{\sigma}^2 > 0$. This result here is probably due to large sampling variation. Consequently, as a matter of pragmatism we can feel comfortable here making conditional inferences from the fixed-effects trend model, $\{S_T, r_t\}$, which is the AIC best model. Also, model $\{S_{T,\sigma}, r_t\}$ would now be dropped from consideration to avoid model redundancy (Burnham and Anderson 1998, sections 4.2.9 and 4.2.10).

Because the random effects theory applied here lead to $\hat{\sigma} = 0$ we should expect the results of fitting that model indirectly, based on starting with model $\{S_t, r_t\}$, to correspond well to results of direct maximum likelihood estimation under model $\{S_T, r_t\}$. In the latter case we find MLEs as $\hat{a} = 0.595$ ($\hat{se} = 0.058$), and $\hat{b} = -0.020$ ($\hat{se} = 0.007$) compared to the indirect random effects fitting which give $\hat{a} = 0.577$ ($\hat{se} = 0.056$), and $\hat{b} = -0.021$ ($\hat{se} = 0.007$). However, because $\hat{a}$ and $\hat{b}$ under fixed and random effects models are not identical, even though $\hat{\sigma}^2 = 0$ (which results in K_{re} being the same as K for the fixed-effects trend model), the likelihoods of the two models are not identical. This results in AIC for the two models being slightly different; logically, they should be the same. An adjustment is possible (see discussion).

5.2 Mallards Banded for 42 Consecutive Years

This example is from late Summer bandings, in California, of adult male mallards (*Anas platyrhynchos*) banded every year from 1955 to 1996 ($k = 41$) (Franklin et al., in press). The total number of birds banded was 42,015, with a yearly minimum of 268 and a yearly maximum of 2,279. The total number of recoveries was 7,647. The goodness-of-fit test (Brownie et al. 1985, Burnham et al. 1987) to the time-specific model $\{S_t, r_t\}$ gave $\chi^2 = 280.86$ on 235 df ($P = 0.0216$), for a variance inflation factor $\hat{c} = 1.1952$ (see Lebreton et al. 1992 and Burnham and Anderson 1998 for discussions of variance inflation in capture-recapture). To accommodate $c > 0$ the sampling variance-covariance matrix W was adjusted upwards to be $\hat{c}W$ and QAIC (Burnham and Anderson 1998) was used:

$$\text{QAIC} = \frac{-2\log\mathcal{L}}{\hat{c}} + 2K.$$

We also used Akaike weights in this example (see Burnham and Anderson 1998).

Results are given here for three models: $\{S_t\,,\,r_t\}$, $\{S,\,r_t\}$, and $\{S_{\mu,\sigma}\,,\,r_t\}$:

Model	K	$\Delta QAIC$	Akaike weight	Comment re model for survival
$\{S_{\mu,\sigma}\,,\,r_t\}$	73.26	0.00	0.9984	random time effects
$\{S_t\,,\,r_t\}$	83	12.87	0.0016	fixed time effects
$\{S\,,\,r_t\}$	43	100.11	0.0000	time-constant S

The random-effects model is by far the AIC best model here. From the random effects model $\hat{\sigma} = 0.0843$, 95% confidence interval on σ of 0.0582 to 0.125, and $\hat{\mu} \equiv \hat{E}(S) = 0.630$ with unconditional $\hat{se} = 0.014$. This estimated standard error includes the uncertainty due to σ. If one takes the mean of the 41 MLEs, $\hat{S}_i$, from model $\{S_t\,,\,r_t\}$ and bases the standard error of that mean on just the sampling variance-covariance matrix, $\hat{c}W$, the result is 0.638 with $\hat{se} = 0.0054$ (a result that is conditional on $\underline{S}$, hence excludes σ). This is an incorrect standard error to use when inferences are meant to apply to other mallard populations or time periods. Just as bad, if one computes the standard deviation of the MLEs from model $\{S_t\,,\,r_t\}$ as an estimate of σ, the result is 0.121, a value much inflated by sampling variation, and almost outside the proper 95% confidence interval on σ.

Finally, consider the improvement in precision achieved by the shrinkage estimates, $\tilde{S}_i$, from model $\{S_{\mu,\sigma}\,,\,r_t\}$ compared to the MLEs, $\hat{S}_i$, from model $\{S_t\,,\,r_t\}$. As a basis for this comparison we use the ratio of average $\hat{rmse}$ to $\hat{se}$ below:

$$\frac{\bar{\hat{r}}\,mse(\tilde{S}_i|S_i)}{\bar{\hat{s}}e(\hat{S}_i|S_i)} = \frac{0.06476}{0.07870} = 0.823\ .$$

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.

The simple random effects model is both necessary here for inference about process variation, σ^2, and very useful for improved inferences about time-varying survival rates.

6. DISCUSSION

For capture-recapture data the MLEs and their variance-covariance matrix are conditional on the survival probability parameters as being fixed effects. Thus, traditional statistical inferences for such data (e.g., CJS models) do not in theory extend to conditions other than those extant when and where the data were collected. This consideration also applies in principle if we embed the model restriction $\underline{S} = X\underline{\beta}$ (or with a nonlinear link function) directly into the likelihood, $\mathcal{L}(S_1\,,\,\ldots\,,\,S_k\,,\,\underline{\theta}\,)$ (i.e., the "direct" approach) and hence get a direct MLE of $\underline{\beta}$. Indeed, that MLE value will be very similar to the estimate of $\underline{\beta}$ from the random effects approach based on the MLEs $\hat{S}_1\,,\,\ldots\,,\,\hat{S}_k$ from this same likelihood. However, what the direct MLE approach lacks is an evaluation of the fit of the structural model, such as $\underline{S} = X\underline{\beta}$, to the true S_i in the sense of obtaining a valid estimate of σ^2. Rather, the direct likelihood approach implicitly assumes no lack of fit, hence $\sigma^2 \equiv 0$, and the uncertainty of the MLE of $\underline{\beta}$ will not

include a component for σ^2. Such a direct approach is thus *not* analogous to what occurs in the regression approach we would do if knew the S_i. The random effects approach properly recognizes the two variance components affecting $\underline{\hat{S}}$ and hence produces a correct estimator of the unconditional VC($\underline{\hat{\beta}}$) when $\hat{\sigma}^2 > 0$. Therefore, we can get correct unconditional inferences on a conceptual population-level $\underline{\beta}$ as well as get $\hat{\sigma}^2$. We also get improved conditional inferences on the S_i by use of the shrinkage estimators, $\widetilde{S}_i$, given by (9) and their conditional $\hat{rmse}(\widetilde{S}_i \mid \underline{S})$ given by (12).

The approach to random effects models presented here could, in principle, be replaced by one of two exact approaches, one Bayesian the other frequentist. However, those approaches are both computationally very demanding and both require assuming a probability distribution on S as a random variable, say $f(S)$ (as a pdf). Under the frequentist approach we first compute the marginal distribution of the data by integrating out S. The resulting marginal likelihood has only the fixed-effects parameters $\underline{\beta}$, σ^2, and $\underline{\theta}$, and standard AIC applies. In the simplest case the marginalized likelihood is given by

$$\mathcal{L}(\mu, \sigma^2, \underline{\theta}) = E_{\underline{S}}[\mathcal{L}(\underline{S}, \underline{\theta})] = \int \cdots \int \mathcal{L}\{\underline{S}, \underline{\theta}\} \prod_{i=1}^{k} f(S_i \mid \mu, \sigma^2)\, dS_1 \cdots dS_k . \qquad (17)$$

This is a k-dimensional integral and k can be 10 to 30, or more.

For capture-recapture models we cannot ever do (17) analytically (because the likelihoods are far too complicated as functions of their parameters) and perhaps not well, or easily, by classical numerical integration methods. It should be possible to compute (17) by Monte Carlo methods, including Markov Chain Monte Carlo used in Bayesian marginalization methods (see e.g., Gelfand and Smith 1990, Zeger and Karim 1991, Smith and Gelfand 1992). Those methods are very computer intensive. Bear in mind that the entire optimization of $\mathcal{L}(\mu, \sigma^2, \underline{\theta})$ for the MLEs must be numerical, hence the integration in (17) must be done many times. Some simple investigation suggests that for capture-recapture perhaps 1,000,000 Monte Carlo trials are be needed to get useful precision in the evaluation of (17), and that is for one point in the parameter space of $\underline{\theta}$, μ, σ^2. Full numerical maximization of (17) will require evaluation at many points in the parameter space; in principle this can be done. The upshot is that the exact frequentist approach is not recommended. However, a full Bayesian approach using MCMC is probably feasible and worth exploring for capture-recapture models that are in effect mixed-effects models possibly with nested or multiple levels of random effects. The advantage of the MCMC approach is its generality; the disadvantage is that it is very computer intensive.

As with the grouse example for models $\{S_T, r_t\}$ and $\{S_{T,\sigma}, r_t\}$ the fitted random-effects model may produce $\hat{\sigma}^2 = 0$. In this case tr(G) is exactly the number of structural parameters, $\underline{\beta}$, in the model for $\underline{S}$ (2 for this grouse example). Also, K_{re} is then identical to K (17 in this example) for the fixed effects version of the random-effects model structure. However, the two estimates of the structural parameter $\underline{\beta}$ will not be identical (they will be quite close) and this results in the two AIC values being slightly different. This is a nuisance only, as in this case one should discard the redundant random effects model. However, it is possible to eliminate this discontinuity between the two AICs in such a case.

First consider the simple random effects model with one structural parameter, $E(S)$. The fixed-effects analogy is the model wherein all $S_i = S$. Let the MLE from that model be denoted $\hat{S}$. From the random effects model we have the estimator $\hat{E}(S)$. Now instead of computing the random effects likelihood from (13) as

$$\mathcal{L}(\underline{\widetilde{S}}\,, \underline{\widetilde{\hat{\theta}}}\,) = \max_{\underline{\theta}} \mathcal{L}(\underline{\widetilde{S}}\,, \underline{\theta}\,)$$

based on the $\widetilde{S}_i$, we modify these survival probabilities to be

$$\widetilde{S}_i + (\hat{S} - \hat{E}(S))\left[\left[\frac{1}{\mathrm{tr}(G)}\right]\left[\frac{k-\mathrm{tr}(G)}{k-1}\right]\right]. \tag{18}$$

The likelihood function is then re-optimized for $\underline{\theta}$ at these values of the survival probabilities. This smoothly eliminates the AIC discontinuity because at $\hat{\hat{\sigma}}^2 = 0$, $\mathrm{tr}(G) = 1$, the shrinkage estimates "collapse" to $\hat{E}(S)$ and (18) becomes $\hat{S}$, the MLE under the matching fixed effects model. Upon re-optimization the likelihood for the fitted random effects model is identical to that of its structural matching fixed effects model and the two AIC values will be identical.

The idea is easily generalized to a model structure $\underline{S} = X\underline{\beta}$, where β is r-dimensional. Let $\underline{\hat{\beta}}$ be the MLE under the fixed effects version of the model. From the fitted random effects version of the model we get $\hat{E}(\underline{S}) = X\underline{\hat{\hat{\beta}}}$, the shrinkage estimate $\underline{\widetilde{S}}$ about $\hat{E}(\underline{S})$, and $\hat{\hat{\sigma}}^2$. Now re-optimize (for $\underline{\theta}$) the likelihood not at $\underline{\widetilde{S}}$, but rather at

$$\underline{\widetilde{S}} + (X\underline{\hat{\beta}} - \hat{E}(\underline{S}))\left[\left[\frac{r}{\mathrm{tr}(G)}\right]\left[\frac{k-\mathrm{tr}(G)}{k-r}\right]\right]. \tag{19}$$

If $\hat{\hat{\sigma}}^2 = 0$, then $\mathrm{tr}(G) = r$, the shrinkage estimate collapses to $\hat{E}(\underline{S})$ and (19) becomes $\underline{S} = X\underline{\hat{\beta}}$ so the re-optimized likelihood has the same value as for the fixed effects model, and the two AICs are the same. The main reason to do this is so one sees at a glance from the AICs that the random effects model is totally redundant.

Model redundancy in the context of AIC selection is discussed in Burnham and Anderson (1998, sections 4.2.9 and 4.2.10). The random effects model is substantially redundant of its fixed effects likelihood version. Report results from only one of these models. A general suggestion now is to not model average over the random effects models because their primary purpose is to allow estimation of process variance, σ^2, a parameter not in any of the fixed effects models. Whereas analysis strategies for capture-recapture data do exist (e.g, Lebreton et al. 1992, Anderson and Burnham 1999), they do not consider random effects models. One strategy suggestion is to first fit fixed effects models. Then for the AIC selected model, e.g., of the form $\{S_{X\underline{\beta}}, \underline{\theta}\}$, fit the corresponding random effects model $\underline{S} = X\underline{\beta} + \underline{\epsilon}$. If it has similar or smaller AIC and $\hat{\hat{\sigma}}^2 > 0$ then make inferences from that random effects model. As a general suggestion be sure and check for overdispersion (as by goodness-of-fit) and use QAIC if overdispersion is found.

For fitting a random effects model we suggest using the MLEs from the general time specific model (plus any age or group effects needed). This generality is required not only for $\hat{\underline{S}}$, but also for other parameters in the model (e.g., p or f). Those other parameters should be allowed to be fully time-varying so that no erroneous structure imposed on them can affect (wrongly constrain) $\hat{\underline{S}}$ and thereby bias $\hat{\sigma}^2$.

When generating the MLEs $\hat{S}_1, \ldots, \hat{S}_k$ to input to the random effects model we recommend using an identity link (and use an identity link in the random effects fitting). The reason for this recommendation is that only the identity link is free of problems as to estimation of the sampling variance-covariance matrix, W, that occur when estimates are on a boundary. With a logit link, in particular, when an $\hat{S}_i$ is at or very near 1 a result is that the numerical-determined value of $\hat{var}(\hat{S}_i \mid S_i)$ is often 0 (which is quite wrong). Then in the weighting used for random effects this estimate erroneously gets very high weight. The result is bias in estimates from the random effects model fitting. An estimator that minimizes (but does not eliminate) this problem is

$$\hat{\sigma}^2 = \frac{\Sigma(\hat{S}_i - \bar{S})^2}{k-1} - \bar{v}ar + \bar{c}ov ,$$

where $\bar{v}ar$ and $\bar{c}ov$ are the average sampling variances and covariances, respectively, computed from W. This is a quick, easy estimator (given in program MARK as the naive estimator), it is not as efficient as the weighted estimator.

Random effects models can be fit to parameters on a transformed scale, in particular such as logit(S) $= \beta$ (but estimates of both β and $var(\hat{\beta} \mid \beta)$ can be unreliable when $\hat{S}$ is too near a boundary). However, $\hat{\sigma}$ is now on a logit scale; if we want process variation for S a back transform is required. In general, if $S = t(\beta)$, then to a first order $\sigma_S = \mid t'(E(\beta)) \mid \sigma_\beta$; this may not be a good approximation. If β has a Gausian distribution then a second order result is $\sigma_S^2 = [t'(E(\beta))]^2\sigma_\beta^2 + \frac{1}{2}[t''(E(\beta))\sigma_\beta^2]^2$. A seemingly better procedure (not explored yet) is to obtain the $\tilde{\beta}_i$, compute $\tilde{S}_i = t(\tilde{\beta}_i)$, $i = 1, \ldots, k$, and then estimate σ_S as the sample standard deviation of $\tilde{S}_1, \ldots, \tilde{S}_k$.

The methods here are for equal-length time intervals. It is not clear how to relate process variation to quite unequal time intervals. Surely the biology of the animals would have to then be a consideration.

There may be a need to consider time series issues for $S_1, \ldots, S_k$ since the time intervals are sequential. The sequence of survival rates, S_i, may not be independent. However, given their conceptual nature it is difficult to come to intellectual grips with the issue of their independence or possible lag-correlations. To generalize the method here the key step is to generalize matrix D to be $D = \sigma^2 C + E_{\underline{S}}(W)$ for some correlation matrix on $\underline{S}$. For example, assuming the S_i are AR(1) adds one correlation parameter, ρ, to the problem. The extension is worth developing just to see if the estimators then perform well or not.

The estimation theory presented here (for exchangeable S_i) is optimal only when the weight matrix is $D = \sigma^2 I + E_{\underline{S}}(W)$. However we will generally have only $\hat{E}_{\underline{S}}(W) = \hat{W}$ from numerical second partial derivative methods (i.e., from an empirical information matrix). The effect of this difference between theory and application is not well known yet, but based on preliminary Monte Carlo simulations it does not appear to be problematic. Also, note that some

positive bias in $\hat{\sigma}^2$ may occur because negative estimates (which occur) must be truncation to 0, yet the estimation method is nearly unbiased only when one allows negative estimates of σ^2.

There is a large literature on capture-recapture models (see, e.g., Schwarz and Seber 1999), yet very little has been done on random effects models for capture-recapture. This paper is a start. Much remains to be learned about random effects in capture-recapture, but here is a class of models we have not been using, but need to, that are intermediate between completely unstructured time variation and a model like $S_i \equiv S$ that eliminates all the process variation. The latter modelling forces $\sigma^2 = 0$, which is unrealistic, although as a model it may often be useful. However, we need to start using random effects models with capture-recapture data both to estimate and study process variation, and because of the efficiency advantages of shrinkage estimators.

ACKNOWLEDGEMENTS

The author thanks Alan Franklin for obtaining the mallard data from the U.S.G.S. Bird Banding Lab and formatting the data for MARK. Clait Braun is thanked for the grouse data. Helpful reviews were provided by Doug Johnson, David Anderson and an anonymous reviewer. A huge thank you to Gary White for creating program MARK.

LITERATURE CITED

Anderson, D. R. and Burnham, K. P. (1976) *Population Ecology of the Mallard VI. The effects of exploitation on survival*. U. S. Fish and Wildlife Service, Resource Publication **128**.

Anderson, D. R., and Burnham, K. P. (1999) General strategies for the analysis of ringing data. *Bird Study* **46**Supplement, 261-270.

Barker, R. J. (1997) Joint modeling of live-recapture, tag-resighting and tag-recovery data. *Biometrics* **53**, 666-677.

Brownie, C., Anderson, D. R., Burnham, K. P., and Robson D. S. (1985) *Statistical inference from band recovery data--a handbook (2nd Ed.)* U. S. Fish and Wildlife Service Resource Publication **156**.

Burnham, K. P. (1993) A theory for combined analysis of ring recovery and recapture data. In *Marked Individuals in the Study of Bird Population*, J. D. Lebreton and P. M. North (eds), Birkhäuser Verlag, Basel, Switzerland. pp. 199-213

Burnham, K. P. and Anderson, D. R. (1998) *Model Selection and Inference: A Practical Information-Theoretical Approach*, Springer-Verlag, New York, NY.

Burnham, K. P., Anderson, D. R., White, G. C., Brownie, C., and K. H. Pollock (1987) *Design and Analysis of Fish Survival Experiments Based on Release-recapture Data*. American Fisheries Society, Monograph 5. Bethesda, Maryland.

Carlin, B. P., and Louis, T. A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman and Hall, London.

Carroll, R. J., Ruppert, D., and Stefanski, L. A. (1995) *Measurement Error in Nonlinear Models*, Chapman and Hall, London.

Casella, G. (1985) An introduction to empirical Bayes data analysis. *The American Statistician* **39**, 83-87.

Catchpole, E. A., Freeman, S. N., and Morgan, B. J. T. (1995) Modelling age variation in survival and reporting rates for recovery models. *Journal of Applied Statistics* **22**, 597-609.

Efron, B., and Morris, C. (1975) Data analysis using Stein's Estimator and its generalizations. *Journal of the American Statistical Association* **70**, 311-319.

Franklin, A. B., Anderson, D. R., and Burnham, K. P. (in press) Estimation of long-term trends and variation in avian survival probabilities using random effects models. *Journal of Applied Statistics* (for first half of 2002).

Gelfand, A. E., and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Hastie, T, and Tibshirani, R. (1990) *Generalized Additive Models*, Chapman and Hall, London.

Hurvich, C. M., and Simonoff, J. S. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60**, 271-293.

Johnson, D. H. (1981) Improved population estimates through the use of auxiliary information. *Studies in Avian Biology No* **6**, 436-440.

Johnson, D. H. 1989. An empirical Bayes approach to analyzing recurrent animal surveys. *Ecology* **70**, 945-952.

Kubokawa, T. 1999. Shrinkage and modification techniques in estimation of variance and related problems: a review. *Communications in Statistics – Theory and Methods* **28**, 613-650.

Lebreton, J. D., Burnham, K. P., Clobert, J., and Anderson, D. R. (1992) Modeling survival and testing biological hypotheses using marked animals: case studies and recent advances. *Ecological Monographs* **62**, 67-118.

Link, W. A. (1999) Modeling pattern in collections of parameters. *Journal of Wildlife Management* **63**, 1017-1027.

Link, W. A., and Nichols, J. D. (1994) On the importance of sampling variation to investigations of temporal variation in animal population size. *Oikos* **69**, 539-544.

Longford, N. T. (1993) *Random Coefficient Models*, Oxford University Press, Inc., New York, NY.

Louis, T. A. (1984) Estimating a population of parmeter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association* **79**, 393-398.

Morris, C. N. (1983) Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* **78**, 47-65.

Pradel, R. (1996) Utilization of capture-mark-recapture for the study of recruitment and population growth rate. *Biometrics* **52**, 703-709.

Robinson, G. K. (1991) That BLUP is a good thing: the estimation of random effects. *Statistical Science* **6**, 15-51.

Särndal, C., Swensson, B., and Wretman, J. (1992) *Model Assisted Sampling*, Spring-Verlag, New York, NY.

Schwarz, C. J., and Arnason, A. N. (1996) A general methodology for the analysis of capture-recapture experiments in open populations. *Biometrics* **52**, 860-873.

Schwarz, C. J., and Seber, G. A. F. (1999) Estimating Animal Abundance: Review III. *Statistical Science* **14**, 427-456.

Seber, G. A. F. (1977) *Linear Regression Analysis*, John Wiley and Sons, New York, NY.

Seber, G. A. F. (1984) *Multivariate Observations*, John Wiley and Sons, New York, NY.

Shi, P., and Tsai, C-L (1998) A note on the unification of the Akaike information criterion. *Journal of the Royal Statistical Society, Series B* **60**, 551-558.

Smith, A. F. M., and Gelfand, A. E. (1992) Bayesian statistics without tears: a sampling-resampling perspective. *The American Statistician* **46**, 84-88.

Ver Hoef, J. M. (1996) Parametric empirical Bayes methods for ecological applications. *Ecological Applications* **6**, 1047-1055.

White, G. C. (2000) Population viability analysis: data requirements and essential analysis. In *Research Techniques in Animal Ecology: Controversies and Consequences*, L. Boitani and T. K. Fuller (eds), Columbia University Press, New York, NY. pp 288-331.

White, G. C., and Burnham, K. P. (1999) Program MARK – survival estimation from populations of marked animals. *Bird Study* **46**Supplement, 120-138.

Zablan, M. A. (1993) Evaluation of sage grouse banding program in North Park, Colorado. M.S. Thesis, Colorado State University, Fort Collins, CO, 80523, USA.

Zeger, S. L., and Karim, M. R. (1991) Generalized linear models with random effects: a Gibbs sampling approach. *Journal of the American Statistical Association* **86**, 79-86.

Table 1. One realization from the random effects example, $k = 10$, $E(S) = 0.5$, $\sigma = 0.05$ where $\hat{S}_i = y_i / n$ are binomial(25, S_i), hence expected $se(\hat{S}_i \mid S_i)$ is almost 0.1; also shown are shrinkage estimates, their estimated conditional standard errors and root mean square errors, and the sum of squares between $\hat{S}_i$ and S_i and between $\tilde{S}_i$ and S_i.

i	S_i	$\hat{S}_i$	$\hat{se}(\hat{S}_i \mid S_i)$	$\tilde{S}_i$	$\hat{se}(\tilde{S}_i \mid S_i)$	$\hat{rmse}(\tilde{S}_i \mid S_i)$
1	0.603	0.640	0.098	0.541	0.047	0.109
2	0.467	0.360	0.098	0.437	0.047	0.090
3	0.553	0.480	0.102	0.482	0.047	0.047
4	0.458	0.440	0.101	0.467	0.047	0.054
5	0.506	0.480	0.102	0.482	0.047	0.047
6	0.498	0.320	0.095	0.420	0.047	0.111
7	0.545	0.600	0.100	0.526	0.047	0.088
8	0.439	0.400	0.100	0.452	0.047	0.070
9	0.488	0.560	0.101	0.511	0.047	0.068
10	0.480	0.560	0.101	0.511	0.047	0.068
mean	0.504	0.484	0.100	0.483	0.047	0.075
st.dev.	0.050	0.106		0.039		

$$\sum_{i=1}^{10}(\hat{S}_i - S_i)^2 = 0.067 \qquad 0.019 = \sum_{i=1}^{10}(\tilde{S}_i - S_i)^2$$

Table 2. MLEs and shrinkage estimates, their relevant estimated standard errors, and the shrinkage $\hat{rmse}$, for the male subadult sage grouse ring recovery data in Zablan (1993) (see text for details) fit to the simple random effects model with estimated parameters, $\hat{E}(S) = 0.430$ and $\hat{\sigma} = 0.046$ (95% C.I. 0 to 0.186).

i	$\hat{S}_i$	$\hat{se}(\hat{S}_i \mid S_i)$	$\tilde{S}_i$	$\hat{se}(\tilde{S}_i \mid S_i)$	$\hat{rmse}(\tilde{S}_i \mid S_i)$
1	0.579	0.204	0.479	0.050	0.111
2	0.667	0.211	0.494	0.050	0.180
3	0.366	0.101	0.427	0.045	0.076
4	0.626	0.156	0.493	0.049	0.141
5	0.521	0.139	0.477	0.047	0.065
6	0.535	0.176	0.463	0.049	0.087
7	0.365	0.128	0.411	0.047	0.065
8	0.319	0.110	0.389	0.046	0.083
9	0.705	0.267	0.466	0.051	0.245
10	0.261	0.102	0.367	0.045	0.115
11	0.507	0.172	0.438	0.050	0.085
12	0.295	0.128	0.381	0.048	0.098
13	0.396	0.219	0.411	0.051	0.053
14	0.227	0.162	0.369	0.050	0.150
mean	0.455	0.163	0.433	0.048	0.111
st. dev.	0.157		0.046		

Table 3. MLEs and shrinkage estimates, their relevant estimated standard errors, and the shrinkage $\hat{rmse}$, for the male subadult sage grouse ring recovery data in Zablan (1993) (see text for details) fit to the linear time trend random effects model with estimated parameters, $\hat{a} = 0.577$, $\hat{b} = -0.021$ and $\hat{\sigma} = 0.0$ (95% C.I. 0.0 to 0.10).

i	$\hat{S}_i$	$\hat{se}(\hat{S}_i \mid S_i)$	$\tilde{S}_i$	$\hat{se}(\tilde{S}_i \mid S_i)$	$\hat{rmse}(\tilde{S}_i \mid S_i)$
1	0.579	0.204	0.556	0.049	0.054
2	0.667	0.211	0.535	0.043	0.139
3	0.366	0.101	0.514	0.038	0.153
4	0.626	0.156	0.493	0.032	0.136
5	0.521	0.139	0.473	0.028	0.056
6	0.535	0.176	0.452	0.025	0.087
7	0.365	0.128	0.431	0.023	0.070
8	0.319	0.110	0.410	0.024	0.094
9	0.705	0.267	0.389	0.027	0.317
10	0.261	0.102	0.368	0.031	0.111
11	0.507	0.172	0.347	0.036	0.164
12	0.295	0.128	0.326	0.042	0.052
13	0.396	0.219	0.305	0.048	0.103
14	0.227	0.162	0.284	0.054	0.079
mean	0.455	0.163	0.420	0.036	0.115
st.dev.	0.157				

Fig. 1 Survival estimates from Table 1, MLEs and shrunk estimates

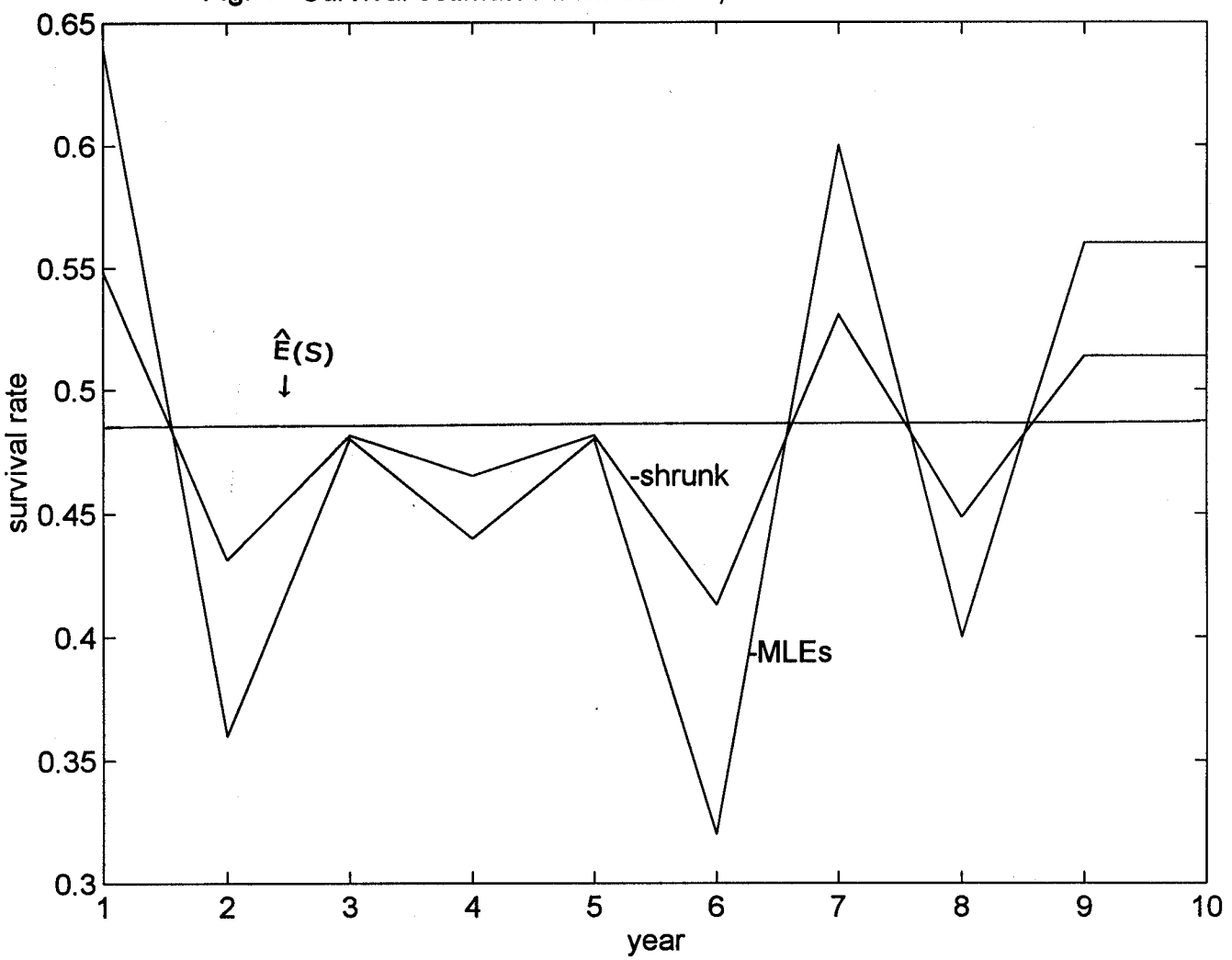

Fig. 2 Male, subadult Sage grouse, Zablan 1993, results from MARK

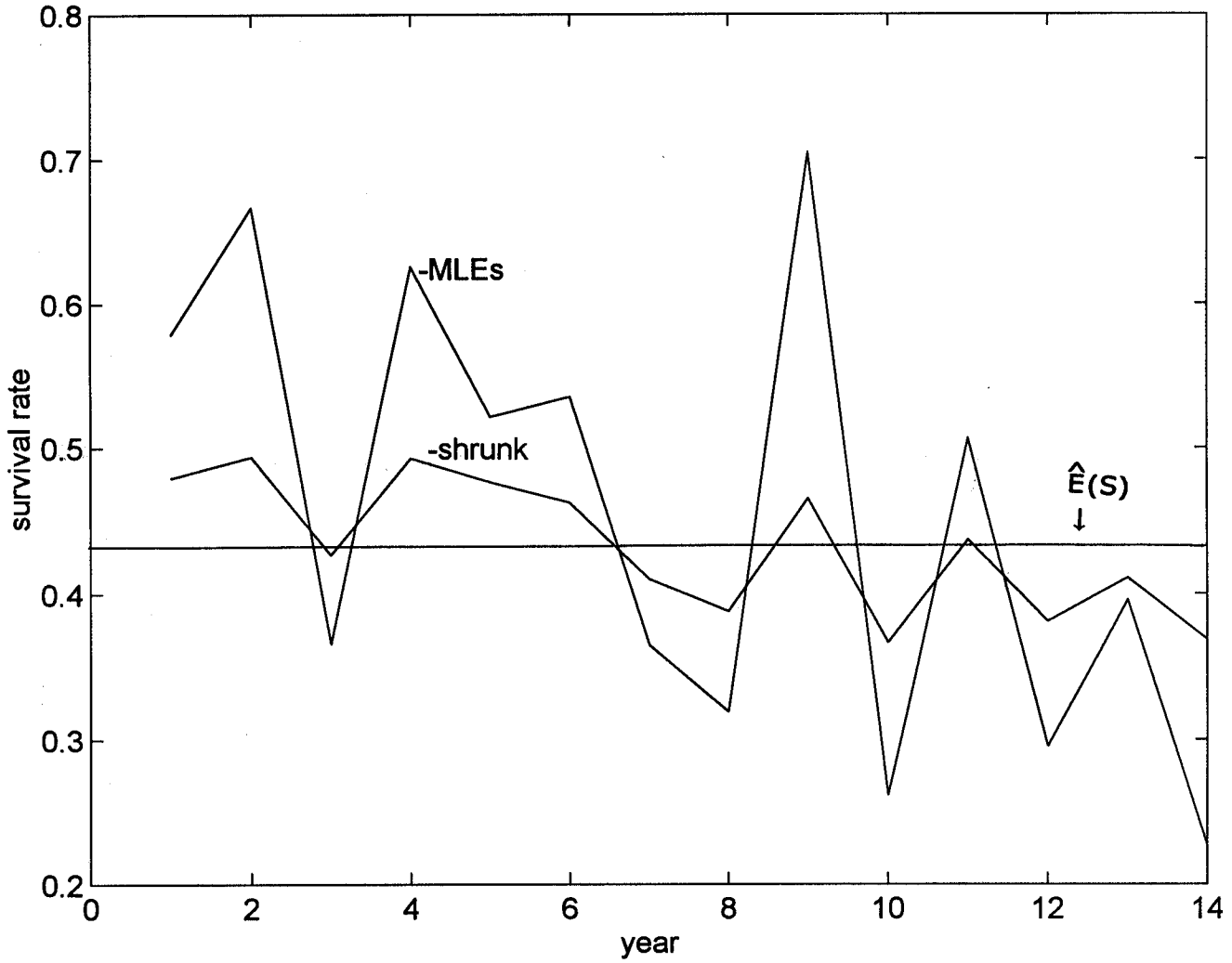

Fig. 3 Male, subadult Sage grouse, Zablan 1993, results from MARK