Suggestions:

Keep working to understand the concept of a design matrix

Study the figures and model naming conventions of various models in lecture6; this will help understand what the (sub)model is achieving (a picture is worth a 1000 words-type of thing).

Start to review old mid-term exams to understand the type of questions to expect

Keep reading and re-reading materials; as your understanding of the subjects increases, you can better understand points that were unclear on first reading.

Some Rules for Making Design Matrices

The purpose of the design matrix is to allow models that further constrain parameter sets. These constraints provide additional flexibility in modeling and allows researchers to build models that cannot be derived using the simple PIMs in *MARK*. For example, the “+" models (no interaction terms) allow a parallelism in time-specific parameters and this is often very useful in parsimonious modeling. The ability to model parameters as nonlinear functions of covariates is also very important (e.g., modeling survival probability as a logistic function of precipitation).

The whole issue of model selection takes on increased importance as biologists are able to build complicated models (and submodels) of band recovery data and open capture-recapture data (e.g., what link function might be best? Are various interaction terms supported by the data? Is a single model clearly the best? Or are several models somewhat tied for “best"?).

Some simple “rules" that might be helpful:

Use an intercept; a column of 1s in the first column.

If you have k categories (say, 4 study sites), create k-1 columns (then 3 columns). Let the first row of the k categories start with a 1, followed by 0s, and let the last row of the k categories be a row of 0s.

Realize that the columns of the design matrix correspond to the β_i and that the rows correspond to the “real" parameters (i.e., S_j and r_j).

The overall design matrix can be thought of as a large matrix, separated into 4 “quadrants." The top left quadrant is for coding the survival probabilities, while the bottom right quadrant is for coding the recovery/reporting probabilities. The top right and bottom left quadrants are typically all 0s. [Occasionally one might want to model the S_j as a function of the r_j (or, the f_j); then these quadrants can be useful.]

Back-transformation and Link Functions

The link function (e.g., logit(S)) allows one to consider the modeling in a linear fashion. We are all familiar with model building in linear regression; the link function provides this same convenience. Then, one must back-transform to get the derived parameters on the scale of [0, 1], as they are probabilities.

The fundamental model parameters are the β_i and their sampling variance-covariance matrix. These are the MLEs and gotten by maximizing the log-likelihood function, etc. The back-transformation to get the “real" parameters S_j and r_j is one to one; thus these parameters are also MLEs and their sampling variance-covariance can also be estimated (this procedure represents an advanced topic and we will study a little about it when we cover the so-called “delta method"). So, the estimators of the real parameters are *asymptotically* normal, minimum variance and unbiased – all good properties, if samples are large.

The “standard" link function is the logit-link, representing the logistic as a submodel. This is very useful and our typical choice.

Numerically, the sine function is often superior (note, it is the default option in *MARK*). This model is periodic; increasing and decreasing (symmetric) models are “sliced" from parts of the sine wave. In many cases, maximization can best be done using this model, particularly if the MLE lies on or very near a boundary (usually 1).

The log-log and complementary log-log functions are sometimes used because they are slightly asymmetric.

The log function can be used, but does not constrain the estimates to be within [0, 1].

Note, use of link functions always lets one consider the modeling in a *linear* fashion (i.e.,

$$\beta_0 + \beta_1(\mathrm{c}_1) + \beta_2(\mathrm{c}_2) + \cdot\cdot\cdot + \beta_k(\mathrm{c}_k),$$

where the c_i are observed covariates (either continuous or discrete or categorical). Interactions (c_i*c_j) and transformations ($\log_e(\mathrm{c}_k)$) can be included in the linear (sub)model. Such covariates are assumed to be known (i.e., without error, just as in “regression" analysis, one assumed the X_j are known without error).

Interpreting Δ-QAIC$_c$ Values

The various information theoretic methods (AIC, AIC$_c$ and QAIC$_c$) can be used to rank the candidate models from best to worst. Often data do not support only one model as clearly best for data analysis. Instead, suppose three models are essentially tied for best, while another, larger, set of models is clearly not appropriate (either under- or over-fit). Such virtual “ties" for the best approximating model must be carefully considered and admitted. Poskitt and Tremayne (1987) discuss a “portfolio of models" that deserve final consideration. Chatfield (1995b) notes that there may be more than one model that is to be regarded as “useful." The inability to ferret out a single best model is not a defect of AIC or any other selection criterion, rather, it is an indication that the data are simply inadequate to reach such a strong inference. That is, the data are ambivalent concerning some effect or parameterization or structure.

It is perfectly reasonable that several models would serve nearly equally well in approximating a set of data. Inference must admit that there are sometimes competing models and the data do not support selecting only one. Using the Principle of Parsimony, if several models fit the data equally well, the one with the fewest parameters might be preferred; however, some consideration should be given to the other (few) competing models that are essentially tied as the best approximating model. Here the science of the matter should be fully considered. The issue of competing models is especially relevant in including model selection uncertainty into estimators of precision and model averaging. We will only touch on some of these more advanced issues but the reader should understand that formal inference must sometime be based on more than just the (single) “best" model.

We routinely recommend computing (and presenting in publications) the **AIC differences** (rather than the actual AIC values),

$$\begin{aligned}\Delta_i &= \text{AIC}_i - \text{minAIC},\\ &\doteq \text{E}_{\hat{\theta}}[\hat{I}(f, g_i)] - \text{minE}_{\hat{\theta}}[\hat{I}(f, g_i)],\end{aligned}$$

over all candidate models in the set (*MARK* provides these values). Such differences estimate the relative expected K-L differences between full truth f and the set of R approximating models $g_i(x \mid \theta)$. These differences apply to AIC, AIC$_c$, or QAIC$_c$. These Δ_i values are easy to interpret and allow a quick comparison and ranking of candidate models and are also useful in computing Akaike weights (see below).

The larger Δ_i is, the less plausible is the i^{th} approximating model as being the K-L best model for the data. The following rough rules of thumb are useful:

For models where

$\Delta_i \leq 2$ have substantial support and should receive consideration in making inferences,

Δ_i of about 4 to 7 have considerably less support,

$\Delta_i > 10$ have either essentially no support, and might be omitted from further consideration (or at least those models fail to explain some substantial explainable variation in the data).

If observations are not independent but are assumed to be independent, or if the sample size is very small, then these simple guidelines cannot be expected to hold.

The Likelihood of a Model, Given the Data

We can extend the usual concept of the likelihood of the parameters, given both the data and model, i.e.,

$$\mathcal{L}(\underline{\theta} \mid \underline{x}, M_i),$$

to **a concept of the likelihood of the model, given the data**, hence

$$\mathcal{L}(M_i \mid \underline{x}) \propto \exp(-\tfrac{1}{2}\Delta_i)$$

where $\propto$ means “proportional to." Further, it is useful to normalize the $\mathcal{L}(M_i \mid \underline{x})$ to be a set of positive “**Akaike weights**" w_i , adding to 1:

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

This idea of **the likelihood of the model given the data**, and hence these model weights, has been suggested for many years by Akaike (e.g., Akaike 1978b, 1979, 1980, 1981b and 1983b; also see Bozdogan 1987 and Kishino 1991) and has been researched some by Buckland et al. (1997). These weights are called Akaike weights apply also when using AIC$_c$, QAIC, and QAIC$_c$ *(*and even TIC, not yet discussed). The likelihood of each model, given the data and the Akaike weights are not yet programmed in *MARK*.

The bigger a Δ_i is, the smaller the w_i, and the less plausible is model *i* as being the actual K-L best model for *f* based on the design and sample size used. For example, consider *R*=7 models, defined *a priori*. The 7 Δ_i **values** are (in order)

$$0,\ \ 1.2,\ \ 1.9,\ \ 3.5,\ \ 4.1,\ \ 5.8,\ \ \text{and}\ \ 7.3,$$

the **relative likelihoods** of the models are proportional to the values,

$$1,\ \ 0.55,\ \ 0.39,\ \ 0.17,\ \ 0.13,\ \ 0.05,\ \ \text{and}\ \ 0.03$$

if these are normalized to **Akaike weights, w_i** , we have,

$$0.43,\ \ 0.24,\ \ 0.17,\ \ 0.08,\ \ 0.06,\ \ 0.02,\ \ \text{and}\ \ 0.01.$$

In this example, the evidence suggests that the selected best model is not convincingly best (i.e., 0.43 vs. 0.24). The best model is only about twice as likely as the next best model. This weak support for the best model suggests we should expect to see a lot of variation in the selected best model from sample to sample if we could, in this situation, draw multiple independent samples.

In general, likelihood theory provides a quantitative measure of data-based weight of evidence about parameter values, given a model and data (see e..g, Royall 1997). This concept extends to evidence about the relative likelihood of models, given an *a priori* set of models and the data.

Model Selection Uncertainty

Clearly, there is uncertainty in what model is selected as “best" as a basis for inference. Given several sets of data, independently generated from the same process, we would expect some variation in the selected best model using information theoretic criteria (or any other method, for that matter). This variance component should be reflected in estimates of precision of the estimated model parameters.

Consider a multiple, linear regression situation. If one had 100 data sets, each with *k* *x*-variables and each of sample size *n*, on a particular system, then one expects the Least Square estimates of the regression (slope) parameters to vary from data set to data set. This variation is measured by the standard error of the estimated regression parameters (the β_i). This is a measure of precision (across other data sets, even though such other data sets are not available). This is a conceptual issue – measuring the precision (repeatability) as if other data sets existed on the same process.

The selection of a best model is similar in that the best model would likely vary from data set to data set, if many such data sets were to exist. Thus, the measure of precision about an estimated parameter should, ideally, include the usual standard error, given a model, plus the variation associated with the repeatability of selecting a particular model.

We will not provide details on the incorporation of model selection uncertainty into estimates of precision of parameter estimators, but people should be aware that this is an important issue in the analysis of empirical data. Further details (more than you might think you want to know!) are given in Burnham and Anderson (1998).