*Natural Resources Data Analysis – Lecture Notes*
**Brian R. Mitchell**

## III. Week 3:

A. Discuss model sets

B. Discuss results from explorations with model selection spreadsheet

1. ***Sample size*** and $AIC_c$ or $QAIC_c$: With low sample sizes, simpler models (lower K) are chosen; as sample size increases the results approach those obtained with AIC or QAIC.

2. Effect of ***different ĉ values***: as ĉ increases, the value of the selection statistic decreases. This means that the relative importance of parameters increases (i.e. models that differ in K are proportionally further apart than with AIC).

3. Effect of incorrect K in all models: If K is off by one for all models, the effect is generally small, unless the sample size is small or the number of parameters in the models is small. For QAIC, as ĉ increases models are less likely to change rank if K is changed by 1 in all models.

C. Main points from B&A, Chapter 2

1. AIC is the ***expected estimated relative Kullback-Leibler (K-L) distance***, where the ***K-L distance*** is the minimum distance between a model and full reality.

a) Because we can never know how close a model is to truth, but we can determine if one model is closer to truth than another, AIC is ***relative***.

b) Because we are not certain about parameter values, we are dealing with an ***estimate***.

c) Because the estimated relative K-L distance is still not calculable, AIC is the ***expected*** estimated relative K-L distance.

d) AIC will pick the best model in the set, even if it is a lousy model (which is why goodness-of-fit testing is important).

2. ***AIC***

a) $AIC = -2\ln(L(\text{model})) + 2K$

b) ***K*** = number of estimated parameters

c) ***L(model)*** is the likelihood of the model

d) This formulation is only valid for ***large sample sizes***!

e) AIC formulas are usually written with "log" instead of "ln"; make sure you use the ***natural log*** (most packages will do this for you).

3. ***Least Squares AIC***

a) $\ln(L(\text{model})) = -\frac{n}{2}\ln(2\pi e\sigma^2)$

b) $AIC = n\ln(2\pi e\hat{\sigma}^2) + 2K$

c) Where $\hat{\sigma}^2 = \frac{\sum residuals^2}{n}$

d) The AIC formula above differs from the formula in the book; this formula includes the constant $2\pi e$ that Burnham and Anderson didn't think was worth printing. Either formula will produce the same AIC differences, but this formula will match likelihoods calculated by statistics packages (e.g. if you happen to you use a likelihood method such as PROC GENMOD in SAS).

e) The ***maximum likelihood estimate of $\sigma^2$*** that you need for AIC may differ from the estimate of $\sigma^2$ from a stats package; you should calculate this on a spreadsheet to make sure. The difference is that stats packages usually divide by the degrees of freedom instead of n.

f) ***Adjust K***: If you are using a statistical technique that assumes a normal error distribution (e.g., GLM models including regression, ANOVA, and ANCOVA), you need to add one (1) to your value for K to accommodate the estimation of $\sigma^2$. Note that this is not limited to least squares estimation; ***this is a requirement of any model assuming a normal error distribution***!

(1) Just to be confusing, the ***nomenclature for $\sigma$*** is not standardized; for example, if you use PROC GENMOD (with normal error distribution and identity link function) in SAS to estimate a model, SAS will report a scale parameter, $\phi$, also called the dispersion parameter. This parameter equals $(\sigma^2)^2$!
(2) The nice thing about using GENMOD is it actually prints the scale parameter as an estimated parameter; so K is just the number of parameters reported in the output.

g) Do not confuse the dispersion parameter with the overdispersion parameter used in QAIC, even though both are often represented as $\phi$ (and don't confuse this $\phi$ with the $\phi$ in mark-recapture, which is survival probability). The documentation for SAS has a decent discussion of the scale/dispersion parameter.

h) Be aware that variance and standard error are biased in maximum likelihood, especially at low sample sizes (Quinn and Keough 2002, p. 25). The relationship between the two quantities is:

(1) $OLS\ SE = \sqrt{\frac{n(MLE\ SE)^2}{df}}$

(2) In other words, the variances differ by n/df.

4. ***AIC$_c$***

a) $AIC_c = -2\log(L(\text{model})) + 2K\left(\frac{n}{n-K-1}\right)$

b) Use $AIC_c$ whenever $n/K < 40$

c) No reason why $AIC_c$ shouldn't be used ALL the time

5. ***QAIC*** and ***QAIC$_c$***

a) $QAIC = \frac{-2\log(L(\text{model}))}{\hat{c}} + 2K$

b) $QAIC_c = \frac{-2\log(L(\text{model}))}{\hat{c}} + 2K\left(\frac{n}{n-K-1}\right)$

c) Identical to equations for AIC, but with an extra parameter $\hat{c}$ to adjust the log-likelihood for overdispersion

d) $\hat{c} = \frac{GOF\ \chi^2}{df}$

e) ***Overdispersion*** is any situation where sampling variance exceeds the theoretical variance based on the distribution assumed by the model. Overdispersion typically comes from two sources:

(1) ***Lack of independence***. Correlated data causes the actual variance to exceed the theoretical expectation.
(2) ***Heterogeneity***. If a model assumes that a parameter is the same for all records, but in reality there are multiple values for different groups of records (or perhaps even individually-specific values), then the true variance will be greater than that calculated according to the theoretical value. A good example of this is survival rate, which can be expected to differ for different groups of individuals (e.g. males and females).

f) Overdispersion can occur when the assumed error distribution is ***binomial***, ***multinomial*** or ***Poisson***. The problem is most common with mark-recapture data.

g) Because a poorly fitting model can appear overdispersed (note that ĉ is estimated using a goodness of fit statistic!), every attempt must be made to model your system accurately.

h) Calculate ĉ based on your most global model; if ĉ is greater than 4 there is probably a structural problem with your data. ***If ĉ is between 1 and 4***, use $QAIC_c$ for model selection and add 1 to your estimate of K. If ĉ is less than one, use $AIC_c$ (i.e. use ĉ = 1). The cut-point at 4 is, of course, essentially arbitrary.

i) ***Important***: If you use an overdispersion parameter in model selection, you MUST multiply all your variances and covariances by ĉ.

j) There is not complete agreement on whether estimating ĉ requires an increase in K; Gary White (developer of MARK) disagrees with this practice. However, I have not seen anything published that contradicts the conservative approach of increasing K when you use an overdispersion parameter (and be aware that MARK will not correctly increment K for this analysis; the option to do so has a bug as of January 28, 2005).

6. ***AIC differences***

a) $\Delta_i = AIC_i - AIC_{min}$

b) For each model, subtract that model's AIC from the minimum AIC value across the model set.

c) The arbitrary cutoff values for levels of empirical support (e.g. 0 – 2 = substantial support, 4 – 7 = some support) are not important if you will be using model averaging.

7. ***AIC weights***

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

b) The weight is the weight of evidence in favor of the model being the best model in the model set; the weights are normalized to sum to 1.

D. Some important points

1. ***Data set consistency***

    a) The data set must be the same for all models.

        (1) If some records are ***missing data*** for some variables, the records should be excluded from the analysis.

        (2) ***Grouping data***: On page 81 of Burnham and Anderson (2002), they write: "AIC cannot be used to compare models where the data are ungrouped in one case (Model U) and grouped (e.g. grouped into histogram classes) in another." They do not provide any more detail on this issue.

            (a) This statement led me to assume that a given predictor variable could not appear in a model set as both a continuous variable and a classified variable. For example, if wind speed is a predictor for our outcome variable, it can not be in km/hr in some models and the Beaufort scale in others.

            (b) However, it appears that this statement refers to grouping a response variable. In other words, it is not valid to collapse data records into a smaller number of groups in some models and not in others. For example, if you measure energy consumption every month for two years, you can't make monthly comparisons in some models and yearly comparisons in others.

            (c) A simple thought experiment should illustrate why the example in (a) is OK. Assume you went out and measured wind speeds on 8 different days (along with some other data that you suspect affects the amount of bird song you hear). On 4 of those days, wind speed happened to be 2.1 km/hr. On the other 4, it happened to be 4.6 km/hr. Regardless of whether you analyze this wind data as a continuous variable or a nominal variable (i.e. "Low" and "High"), you get exactly the same results. If there were more groups or some variability in the wind speeds, the results would not be exactly the same, but I see no reason why comparing AIC for these models should not be allowed.

2. ***Response variable*** must be the same in all models. In other words, you can't use Y as your outcome in some models, and log(Y) in others.

    a) However, it is possible to compare models with ***different error distributions*** (e.g. normal versus lognormal). See section 6.7 in B&A 2002 for a discussion; using models with different error distributions requires careful calculation of model likelihoods… and there is no guarantee that different statistical software packages use the full calculation.

b) Because transforming Y is typically done to normalize the error distribution, modeling different error distributions is ***functionally equivalent*** to comparing models with different transformations of Y.

3. ***Non-nested models***: one of the powerful features of AIC is that it can be used to compare non-nested models. However, you will not be able to model average parameters that take different forms in different models. If your model set is not nested, it is likely that you will not be able to average some or all of your parameters (but you can always average your model outcome).

4. ***Models within 2 AIC units***: If models differ by one parameter and are within 2 AIC units, check the log likelihoods. If the likelihoods are similar, the more parameterized model does not add any new information (the new parameter is not important) (B&A p. 131).

a) This is really only an issue if you are basing inference on the single best model.

b) If you have models within 2 AIC units, you should be using model averaging, and it will be clear from the parameter estimates that the parameter is not important.

5. ***Large increments in K***: Ideally, you should avoid large increments in K between different models (i.e. several models with few parameters, and several with many) (B&A p. 136).

E. AIC versus BIC (Burnham and Anderson 2004)

1. Models, by definition, are only approximations to reality; there are no true models that perfectly reflect full reality. George Box made the famous statement, "***All models are wrong but some are useful***."

2. The "***best model***", for analysis of data, depends on sample size; smaller effects can often only be revealed as sample size increases. The idea is not to model the data; instead, it is to model the information in the data.

3. What is BIC?

a) $BIC = -2\ln(L(\text{model})) + K\ln(n)$

b) Approximates the natural log of the Bayes factor.

c) The model weight (calculated the same as the Akaike weight, but using BIC) is the inferred probability that the model is the quasi-true model in the model set.

d) For small or moderate sample sizes obtained in practice, the model selected by BIC may be much simpler than the most parsimonious model (i.e. ***BIC may underfit***, leading to biased parameter estimates).

4. Comparison

a) ***Effects in the data***: BIC performs well when there are a few large effects and no small effects, while AIC performs well when there are tapering effects.

b) ***Target model***: BIC is "***dimension consistent***" or "***order consistent***"; this means that regardless of sample size, BIC aims to find the most parsimonious model with the lowest K-L distance. AIC is a "***minimum total discrepancy***" criteria (Taper 2004); this means that the target model may change with sample size, and the goal is to minimize predictive errors. AIC will select models with more parameters as sample size increases.

c) ***Performance***: BIC performs better with nested models and large sample sizes; at small sample sizes the BIC-selected model can be quite biased (underfit), especially if there are tapering effects.

d) ***Fit of selected model***: Based on simulations, the model selected by AIC always fits if the global model fits; the model selected by BIC does not always fit.

5. ***AIC as a Bayesian result***: BIC model selection arises in the context of a large-sample approximation to the Bayes factor, conjoined with assuming equal priors on models. In reality, it makes more sense to use a savvy prior based on the sample size and the number of parameters in a given model (this acknowledges that a small data set has less ability to resolve parameters).

a) The "***savvy prior***" that is presented is the prior that mathematically converts BIC to AIC.

b) There is ***no justification given*** for why this particular prior was chosen, other than that the math happens to work out and it includes sample size and the number of parameters.

c) I am personally wary of claiming that AIC is a Bayesian result without seeing a realistic derivation of the savvy prior from basic principles (as opposed to mathematical sleight of hand)!

6. Based on simulations, ***model averaging*** is always better than prediction based on the best model only (for AIC and BIC).

F. B&A Chapter 3

1. ***Important reminders***: 1) Use model averaging, even though this chapter focuses on model selection. 2) Assess goodness of fit of the global model early in the analysis. Examine outliers, leveraged points, symmetry, trends, and autocorrelations in the residuals. Numerous standard diagnostic procedures should help with this.

2. ***Example 1: Cement Hardening Data***: A simple example of multiple regression analysis. 13 cases and the 4 predictor variables, analysis looks at all possible non-interactive models. Note that there is significant model selection uncertainty.

    a) Note that the number of candidate models exceeds the sample size. This is OK for an exploratory analysis (when you could easily have MANY more models than samples), but should be avoided for confirmatory analyses since it increases the chance that the chosen model will fit the data best simply by chance.

    b) The key is not so much the number of models, but ensuring you have a much larger sample size than your number of parameters ($n > K$). If $n/K < 40$, make sure to use $AIC_c$.

3. ***Example 2: Time Distribution of an Insecticide***: What are the important chemical, physical, biological phenomena governing pesticide distribution in a simulated ecosystem? Take 3 measurements at 12 different time periods; measure percent of radioactivity in fish, soil/plants, and water. Modeling technique is complicated, but basically used a good set of 7 candidate models. $AIC_c$ led to the same results obtained by the original authors.

    a) What is the ***sample size***? 36 is used in the example, but I suspect that the true sample size is 12 (each set of measurements is a single sample).

    b) The ***results change*** with a reduction in sample size; using the smaller sample size results in selection of model 3a instead of 4a; 3a has a weight of 0.684. The results are highly dependent on the sample size!

4. ***Example 3: Nestling Starlings***: Generated Monte Carlo data with 34 parameters and many tapering treatment effects. Global model has 4 fewer parameters than the generating model. What approximating model can be used to analyze the data that leads to valid inferences about system structure, parameters, and treatment effects?

    a) Effective ***sample size*** = number of initial captures + all resightings. I think that the correct sample size is the number of initial captures.

    b) Part of the success was an approach that allowed for modeling the treatment effects (as a sine wave) rather than estimating re-sight and survival probabilities for every week.

5. ***Example 4: Sage Grouse Survival***: Another mark-recapture example. Resightings are mortalities (hunter band returns) in this case.

a) Effective ***sample size*** = number of banded birds. In this case, I agree.

b) Note that parameters that are not included in the best model may still have an effect... it is just that the effect is probably small.

G. References for this class

1. Taper, M. L. 2004. Model identification from many candidates. *In:* The Nature of Scientific Evidence. M. L. Taper and S. R. Lele, editors. The University of Chicago Press, Chicago, IL, USA. pp. 3-16.