Natural Resources Data Analysis – Lecture Notes
Brian R. Mitchell

VI. Week 6:

A. As I've been reading over assignments, it has occurred to me that I may have glossed over some pretty basic but critical issues. Make sure that you can answer the following questions, since they are the basic foundation on which the rest of this course is built!

1. What is the I-T approach, and why would you want to use it?

   a) Given a predefined model set, the I-T approach allows us to estimate the relative strength of different models, given the data.

   b) It allows us to account for uncertainty in model selection via model averaging.

   c) The I-T approach allows us to estimate the probability of a model in the set being the best model, given the data; this is in contrast to frequentist methods that tell us the probability of the data, given a null model.

2. What is K-L distance?

   a) K-L distance is the distance between a model and truth.

3. What does AIC measure?

   a) Estimated expected relative K-L distance.

   b) Because it is estimated and relative, AIC does not measure distance from truth; it is only useful for determining whether one model or another is closer to truth.

4. What is a model weight (w_i), and why is it important?

   a) This is the weight of evidence in favor of the model being the best model of the set.

   b) Model weights are crucial; they are the basis of model selection and multimodel averaging.

B. Anderson, D. R. and K. R. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. *Journal of Wildlife Management* 66:912-918.

1. Anderson and Burnham list a number of ***scientific issues*** that can affect analyses:

a) ***Poor question***: Use several good alternative hypotheses and models to represent them.

b) ***Too many models***: Think before you analyze data. Do not let the number of models exceed the number of data points; choose a few reasonable models to test.

(1) Note that they are not talking about keeping the number of parameters less than the number of data points (which is a methodological issue that will produce overfit models); they are arguing that the overall model set should also be smaller than the number of data points.

c) The ***true model*** is not in the set: There is no “true” model. You are not trying to find “truth”, you are trying to find the best model to fit the data.

d) Information-theoretic methods are ***not a statistical test***: Don't mix these methods with null hypothesis testing.

2. They also list a number of ***methodological issues***:

a) ***Poor modeling***: Carefully represent each hypothesis with a mathematical model relevant to the data at hand. Consider the data and sampling issues.

b) Failure to consider ***model selection uncertainty***: use multimodel inference to incorporate this uncertainty.

c) Failure to consider ***overdispersion*** in count data: Counts are often overdispersed, so you need to estimate a variance inflation factor.

d) ***Data mining***: Clearly separate inferences that arise from a priori considerations from hunches developed after examining the data in detail.

e) ***Significance versus evidence***: Think about strength of evidence, rather than correct or incorrect decisions.

f) ***Goodness of fit***: Assess goodness of fit with the most highly parameterized model. If the global model fits, so will the most parsimonious one.

g) Failure to ***provide information***: present maximized log-likelihood, number of parameters, information criterion used (i.e. AICc or QAICc), plus differences and weights for each model. A single AIC value is useless; utility is in comparing values across models.

3. Finally, they warn of a number of ***outright mistakes***:

a) Incorrect number of ***estimable parameters***: variance is a parameter in models assuming a normal distribution.

b) No ***sample size correction***: Don't use AIC or QAIC when sample size or number of parameters is large; instead, use AIC_c or $QAIC_c$.

c) Using AIC in ***all subsets selection***: Avoid this "just-the-numbers" approach, since it violates the spirit of the information-theoretic approach.

(1) But note that in some cases an exploratory approach is acceptable, provided it is clearly billed and interpreted as such.

d) ***Data set*** not constant: All models must use the same data and the same response variable.

e) ***Failure of numerical methods*** to converge: if maximum likelihood could not be computed, AIC values are useless.

C. Discuss and critique published examples of model selection and averaging

1. Gibson, L. A., B. A. Wilson, D. M. Cahill, and J. Hill. 2004. Spatial prediction of rufous bristlebird habitat in a coastal heathland: a GIS-based approach. *Journal of Applied Ecology* 41:213-223.

a) Did their modeling methods take spatial autocorrelation into account? For example, birds may not be present in some areas because historical conditions drove them extinct, and they simply haven't recolonized yet. So if this is a metapopulation or a recovering population, suitable areas that are vacant will be modeled as unsuitable. (p. 214) They partially address this on page 220, stating that the importance of distance to coast may be a recolonization effect.

b) Surveys were all done on roads; this introduces a large bias that will affect the applicability of the models. (p. 215)

c) There is no way to estimate detection probability for these surveys; there is no way to know if the "absent" sites are truly absent. They probably should have used some sort of "presence only" modeling approach, or attempted to estimate detectability. (p. 215)

d) Figure 2 lists "survey points". There are about 83 of them, presumably the 30 detections and 53 non-detections. This figure would have been much more useful if the two types of points were represented with different symbols. (p. 215)

e) What led to the choice of 53 "absent" sites? Seems like a strange choice. (p. 216)

f) They note that they are taking an exploratory approach, but this is hidden in the “GIS coverage” section (p. 216). They do pick this up again in the Discussion (p. 221).

g) “based on the second-order Akaike’s information criterion corrected for sample size” should be the second-order AIC *or* AIC corrected for sample size, not both. (p. 216)

h) Note that this is not truly an all-subsets model set, since the null model is missing. (p. 216/217)

i) States that bootstrap selection probabilities are robust to the effects of sampling error. That may be true for a large data set and fairly low sampling error, but I don’t know if it is true for a small data set. (p. 216)

j) Examining the model set results (Table 1) reveals that models 3 and 4 differ from models 1 and 2 by about 2 AIC units, and the addition of a single parameter (sun index). Sun index is therefore not an important parameter (as evidenced by a SE three times larger than the parameter estimate (Table 2). Similarly, habitat complexity does not seem to add much. Sun index (and perhaps habitat complexity) should probably have been left out of the predictive equation, especially since their predictive model does not include uncertainty in the estimates (so the sun index affects the outcome estimate when it really does not have an effect). Alternatively, could have re-run the models lacking sun index (and habitat complexity) for inferential purposes. (p. 217/218)

k) The conditional SE estimates for the best model are irrelevant and should not have been included. (p. 218)

l) What are the units for the distance and elevation variables? The values for coast should have either been calculated at different units, or shown to more significant digits. For all we know, the coefficient is 0.00050, and the SE is 0.00049. (Table 2, p. 218)

m) “The small differences in unconditional and conditional standard errors observed in Table 2 were likely to be a reflection of the relatively strong support for the best model”. Not true (especially since the best model was not highly supported). More a function of similarity of parameter estimates among the highly weighted models. (p. 218)

n) Regarding the HP analysis, a significant z-score does NOT mean that the variable should necessarily be retained. Let’s use some common sense, and consider the effect size and precision of the estimated effect size. The HP analysis adds nothing to this study, mixes the use of the I-T approach and significance testing, and leads to the erroneous conclusion that all variables should be used for inference. (p. 218)

o) I would have liked to have seen the incorporation of outcome uncertainty into the map, or at least a calculation of a confidence interval on the outcome at mean values of the predictors. As it is, there is a pretty map that I don't think is very meaningful. (p. 219)

p) They persist in interpreting sun index as an important variable, as well as habitat structure measured by their GIS approach, when the evidence for the importance of these two predictors is weak. (p. 220)

q) The ROC area of 0.97 is extraordinarily high; such a high number is usually not possible in logistic regression without nearly complete separation in the contingency table. This may be due to the small size of the data set.

2. Maestas, J. D., R. L. Knight, and W. C. Gilgert. Biodiversity across a rural land-use gradient. *Conservation Biology* 17(5):1425-1434.

a) It's a bit difficult to interpret their randomization. It appears to me that the points were randomly selected, but forced to fall within the 2 reserves, 3 ranches, and 2 developments (at appropriate elevations, etc…). So the assumption is that these are "typical" reserves, ranches, and developments within the study area. (p. 1427)

b) The methods are pretty sparse for the bird density estimation via AIC. What is a "reliable detection function"? I'm assuming this means that an adequate number of birds were detected. What detection function models were used? If these are not described in the references for this section (Thomas et al. 1998 and Buckland et al. 1993) then they need to be described here. What is "seemed plausible"? Need to be explicit about when averaging was used. (p. 1428)

c) Note that since model selection is being used only for the specific task of identifying the best detection function to use for estimating bird density (based on calculations conducted by program "Distance"), I think it is OK to take the resulting density estimates and analyze them with inferential statistics. Similarly, if MARK were used to estimate population size of a number of different populations, and different models were fit and averaged within MARK, it would probably be OK to take the results and compare populations in different habitats via inferential statistics. (p. 1428)

d) How were final density estimates calculated? Did they incorporate the uncertainty in component density estimates? (p. 1428)

e) It is a little difficult to judge the statistical results without more detail. But based on the confidence intervals in the figures, it appears that these were calculated correctly. (p. 1429).

f) Figure 2 is not complete without the data for Lark Sparrow, Western Meadowlark, and Mourning Dove. (p. 1429)

g) Table 1 is not particularly useful; would be much more meaningful if the number of detections of the species were used instead of the symbol. (p. 1429)

h) There are an awful lot of p-values in this paper. Although this gets tedious to read, I don't feel that the use is inappropriate, especially since the authors are good about providing confidence intervals for all estimates.

i) Do any of the statistical tests seem gratuitous? Perhaps the results for mesopredators are a tad obvious (clearly dogs and cats will be more prevalent near homes, and coyotes are known to prefer open rangelands, while bobcats are known to prefer woody and brushy habitats).

j) Since the goal of this study was comparing densities of species in 3 habitat types, I do not think there is much of a role for a model selection approach. What sort of question would use similar methods and be amenable to the information-theoretic approach? Perhaps an approach that is more interested in modeling density across the landscape, and that includes a variety of factors (rather than reducing the data to a mean and SE density per habitat per species).

3. Miyakoshi, Y., M. Nagata, and S. Kitada. Effect of smolt size on postrelease survival of hatchery-reared masu salmon Oncorhynchus masou. Fisheries Science 67:134-137.

a) Might be nice to know just how many marked salmon were released.

b) Fig 1: I don't think the confidence intervals shown in the graph are the ones predicted from the model. The intervals should all increase with increasing weight at release (I think). Also, the 15 g CI seems unusually small. So this is probably information that comes from the raw data, but it is not labeled as such.

c) Perhaps the information on Fig 1 is all of the raw data. If that is the case, then I don't know that the sample space has been covered well enough to determine a functional form.

d) The functional form is poorly chosen for this model set. Since estimated recoveries are low (particularly at low smolt weights), a functional form should have been chosen that prevents y from dropping below 0. A negative recovery rate is predicted at smolt weights at or below 14 g; this can not be correct.

e) Table: Note that the betas can not be compared; this is a non-nested model set.

f) Table: Model 4 uses a transformed y. This model is invalid, unless re-cast as $y = e^{\alpha+\beta \ln(x)}$.

g) Table: no model weights are reported. The chosen model actually only has a weight of 0.31; model 2 has a weight of 0.22, and 3 other models have weights above 0.10.

h) This is clearly a situation where multimodel averaging is needed. Note that since the models are not linear, they can not be averaged on the spreadsheet we have been using in class, since the spreadsheet cannot handle some of the model terms for calculating point estimates, and the full "method of moments" is needed to calculate outcome variances. With the full data set in hand, this could be done using something like PROC NLIN in SAS.

i) I'm not sure if there is a biological rationale for some of these models (e.g. models 5 and 8).

j) Model 5 does not produce meaningful results unless the coefficient for the gamma term is negative.

D. Goodness of fit testing – Logistic Regression

1. See notes from Week 4