AIC MYTHS AND MISUNDERSTANDINGS

Produced and posted by David Anderson and Kenneth Burnham. This site will be updated occasionally. The site is a commentary; we have not spent a great deal of time and effort to refine the wording or be comprehensive in any respect. It is informal and we hope people will benefit from our quick thoughts on various matters.

The most recent changes and additions were on April 12, 2006

Some issues gain an acceptance and have a life of their own, being passed from person to person as fact, when they are actually incorrect. The model selection literature has a number of such issues – myths and misunderstandings of a sort. In addition, there might be honest differences of opinion or philosophy. Here we try to note several of these in reference to our 2002 Springer-Verlag book or concerning K-L based model selection in general. The issues noted below are in no particular order.

- AIC is only for comparing 2 models (Harrell, F. E. 2001. Regression modeling strategies. Springer Verlag, New York, NY.). This statement is simply incorrect and seems to stand-alone, as we have not seen other comments such as this claim.
- Information-theoretic approaches can only be applied when there is one data set. This incorrect notion comes from Peery et al. (2004. Applying the declining population paradigm: diagnosing causes of poor reproduction in the marbled murrelet. *Conservation Biology* 18:1008-1098); they state, "... in MCH (multiple competing hypotheses) multiple data sets are evaluated against predictions from multiple limiting factors, whereas the method of Burnham and Anderson fits multiple models to one data set." Two points can be made here.

First, the authors are confused as to what is meant by a "data set" in the statistical sciences. The literature is full of examples where more "than one data set" is analyzed using information-theoretic methods: e.g., distance sampling data over multiple years or multiple areas or capture-recapture data over two genders or 3 age classes. The most recent analysis of the data on the northern spotted owl [Anthony et al. *Wildlife Monographs* No. 163.] where data from 16 large study areas were analyzed using 2 sex classes and 3-4 age classes. The notion that only "one data set" can be used in the information-theoretic (or Bayesian approaches) is incorrect. Stephens et al (2005) make the same mistake when they state that AIC cannot be used for treatment vs. control data as "the same data must be used." Clearly, the treatment and control data constitute one data set (not 2).

Second, the "...method of Burnham and Anderson..." is also not correct phrasing. The book by Burnham and Anderson is a synthesis of several world-class scientific achievements (e.g., Kullback and Leibler 1951 and Akaike 1973); our contributions include only a careful synthesis of the vast literature on the subject of model selection and inference. We cannot take credit for the methods now available.

• Two individuals have expressed **concern over model averaging to estimate regression coefficients in linear regression** (the βs). There does not seem to be disagreement over model averaging from prediction; the question relates to averaging the regression coefficients. Both individuals are aware that the estimate of β_i depends on what other variables are in the model (see Burnham and Anderson pages 180-181 for examples).

Thus, if one has a particular interest in, say, β_2 then one might want to use the simple 3-parameter model with a slope, regression coefficient on X_2 and the residual variance. The inclusion of other predictor variables influences the estimates of β_2 unless all the predictor variables are orthogonal (which almost never happens unless the data are from an experiment). Then the estimate of β_2 is conditional on the other variables in the model. The alternative approach is to make an "unconditional" estimate of β_2 . This involves model averaging; the resulting estimate is still conditional on the model set. This approach tries to provide a valid estimate of the effect of X_2 on the response variable without making it conditional on other variables in the model. This is often a useful approach. Of course, the parameter β_2 must mean the same thing in every model – the slope on the predictor variable X_2 .

Another point on this same issue is that if a person accepts model-averaged prediction they are implicitly accepting model-averaged structural regression parameters in linear models. This is because any model-averaged prediction is identical to the prediction from the single linear model produced by the model-averaged β 's. (see e.g., Burnham and Anderson 2002, pages 252-253). If model averaged prediction is "good," then (certainly in linear models) model-averaged parameters must also be "good."

- Brian Ripley, University of Oxford, takes issue on several accounts via various web sites. DRA wrote him asking for clarification (December, 2003) but he chose not to respond. Several of his statements have been picked up and several people have asked us about these, so we will try to note his position and ours.
 - 1. **AIC** assumes a true model. This is not correct; see Chapter 7 of Burnham and Anderson (2000). This error may come from the fact that there are several derivations from K-L information to AIC. One such derivation notes that the deviance (the first term in AIC) has a relationship to the chi-squared distribution (and is therefore the basis for likelihood ratio tests). The chi-squared distribution of the test statistic comes about only if the models are nested. We are guessing, but this might be the source of the confusion. Our clear position on this issue is that nothing need be assumed about a true model when justifying or using AIC (or AICc or QAICc); see details in Sections 7.2 and 7.2 of our 2002 book.
 - 2. **AIC is only for nested models**. This is unfounded. AIC as an estimator of relative, expected Kullback-Leibler information is for both nested and nonnested models. We have not seen this claim in others sources; it is simply incorrect.

3. The so-called penalty term in AIC (i.e., 2K) is not a bias correction term. This is incorrect, see Chapter 7 in Burnham and Anderson (2002). There are certainly dozens of journal papers that clearly show that the maximized $\log(L)$ is a biased estimator of relative, expected K-L information and that to a first order a defensible asymptotic bias correction term is K, the number of estimable parameters in the model. So, $E(K-L) = \log(L) - K$. To obtain his AIC, Akaike multiplied **both** terms by -2. Thus, AIC was $-2\log(L) + 2K$. Note, the 2 is not arbitrary; it is the result of multiplying by -2 such that the first term in the AIC is the (well known) deviance, a measure of lack of fit of the model.

The rigorous derivation of the estimator of expected relative K-L, without assuming the model is true, leads to a bias correction term that is the trace of the product of two matrices: $\operatorname{tr}(J^*\Gamma^1)$. If the model, g, in question is the "true" model, f, then this trace term equals K. If g is a good (in K-L sense) approximation to f then $\operatorname{tr}(J^*\Gamma^1)$ is not very different from K. Moreover, any estimator of this trace (hence, TIC model selection) is so variable (i.e., poor) that its better to take this trace term as K rather than to estimate it.

- 4. Ripley states, "Burnham and Anderson (2002) is a book I would recommend people NOT read until they have read the primary literature. I see no evidence that the authors have actually read Akaike's papers." The first statement is Dr. Ripley's opinion and he is certainly entitled to it. The second statement is simply wrong. We have read all of Akaike's papers in detail and have corresponded with him. Our book is full of specific references to Akaike's papers. It surely must be clear from our 1998 and 2002 books that we read these papers.
- Guthrey et al. (2005. Information theory in wildlife science: critique and viewpoint. *Journal of Wildlife Management* 69:457-465) represents a paper in a special class. This was an invited paper by Mike Morrison, then editor of the *Journal of Wildlife Management*. No *JWM* reviewers are listed or mentioned and we assume that Morrison handled all details with respect to this manuscript. We have written Morrison and asked for an invitation to respond to various points made by Guthrey et al. Morrison offered only a *Letter to the Editor*, which we did not view as appropriate (or fair to us or the science community that *JWM* serves).

The paper represents a near *delirious rant* and perhaps cannot be taken seriously. People are certainly free to critique the various information-theoretic approaches; however, we submit that those offering a critique should have some basic understanding of the philosophical and mathematical-technical issues before publishing in an open science forum (as in the *Journal of Wildlife Management*). As some scientists have read the paper by Guthrey et al. we will provide the following material for consideration. We have noted only a few of the worst aspects here; a full explanation of the problems would require much more time than we have been willing to spend.

- 1. In our philosophical view data contain information (if collected according to some fundamental principles; e.g., a well-founded sampling protocol or experimental design). Here, we use the word *information* in a technical sense (Boltzman, Shannon, and Kullback-Leibler information). Extraction of such information from the data can take a couple of general forms:
- In simple, often 2-3 dimensional problems, one can extract some information in the data by plotting or computing simple summary statistics.
- In more interesting/difficult situations (e.g., the real world) models are useful in extracting the information in the data and allowing an understanding of the issues. This is broadly termed "model based inference" and has been useful in the empirical sciences, medicine, and engineering over the past several decades if not centuries.

Model based inference is a huge subject and well accepted; it is basically unavoidable as well as very powerful. Guthrey et al. seem to almost deny the value (actually the necessity) of quantification in the empirical sciences. Model based inference begs the question "which model should be used" as it is rarely clear *a priori* that one model is somehow *known* to be "best." Further thought begs the further questions about the technical meaning of "best model."

Many approaches have been put forward over the past 50-80 years to address the central issues of "model selection." It is a historical fact that poor model selection tools happened to be developed first (e.g., stepwise regression, likelihood ratio tests). Such *ad hoc* tools have been used extensively because much better tools were late in coming. Making matters worse is that these more powerful methods have not been widely taught by statistics departments, at least to non-majors.

By "much better tools" we are referring to various Bayesian approaches that have become computationally feasible only in the past 15-20 years, information-theoretic methods that have been introduced over the past 20-30 years, and various computer intensive approaches (e.g., cross validation, bootstrapping) in the past 20 or so years.

- 2. Lack of a fundamental understanding of what Guthery et al. call the "algorithm" is illustrated by the following (page 459),
 - "... if the global model is presumably valid, why should it be pared? If a pared model is presumably valid, why would one advance a presumably invalid global model?"

The central point is that one does not know *a priori* which of the *R* models might be "presumably valid." How is one supposed to know that a simple linear model with 8 unknown parameters is "valid" while other models might be nonlinear and have, say, 3-11 or perhaps 30 unknown parameters? Even if one could determine "presumably valid" would there not be some uncertainty about this matter? What if two co-investigators disagreed on the "presumably best model"? What if there is disagreement on the

"presumably best model" in a courtroom situation? Futhermore, it is certainly unclear as to what Guthery et al. might mean by a "valid" model.

How does sample size play into these technical issues? Should a product multinomial model with 8 survival probabilities, 8 sampling probabilities and no interaction effects serve for sample sizes ranging from 120 tagged animals to perhaps 54,000 or 388,000 tagged animals? Clearly one cannot often judge that one model out of many is "presumably valid" nor would we expect others to agree when there is some level of controversy involved.

A key issue people often fail to properly comprehend is that our context is that of fitting models to data to learn what the data "have to tell us" via estimated parameters and the strength of evidence about the different models themselves. The parameters are not a priori known. The size of model (number of parameters) that can reliably be fit to given data depends in part on the sample size, and is not a prior known. In this context "validity" of a model is not a useful concept. What is important is that fitting the models to the data should lead to useful and reliable knowledge. Fitting too general a model risks spurious results; fitting too simple a model risks failing to identify interesting real effects. The "balance" point is not known a prior and depends on sample size.

These errors in both logic and understanding apply to many simple situations. Consider a simple control vs. treatment experiment (completely randomized design) which results in two conceptual models – one with a treatment effect and the other without a treatment effect. How would a person just say one of these models was "presumably valid"? What would be the scientific basis for saying "the model with no treatment effect is valid" with no analysis? Is not science supposed to be about objectivity?

- 3. Several bold statements by Guthery et al. seem to defy any printable response. These must include:
 - "The point is that statistical assumptions serve the artificial world of statistical theory; real world ecological processes operate largely independently of statistical assumptions and theory." (Page 460)
 - "Contrary to popular opinion, the statistical principle of parsimony borrows no legitimacy from Ockahm's Razor, ..." (Page 460)
- 4. Guthery et al. remark,
 - "How should wildlife scientists address this dizzying array of information criteria and other model selection approaches?"

These authors are just now starting to raise questions that we hit upon in 1989-90 and other investigators hit upon much earlier. During 1989-90 we were working with Drs. Jean-Dominique Lebreton and Jean Clobert on the open population capture-recapture models whereby we had practical problems involving easily 10-30 models with the

number of unknown parameters ranging from perhaps 2 to 60. These models were rarely nested; hence likelihood ratio tests were uninterpretable. Goodness-of-fit tests were of little help in understanding and were critically dependent on the arbitrary α -level. Some models had high precision of estimated model parameters while other models made more realistic assumptions but precision was sacrificed.

This lead us to the extensive literature on model selection theory and methods. Necessity lead us to Akaike's information criterion (AIC). We read and tried to understand why such a simple equation could be so potentially useful. We completed an *Ecological Monograph* (Lebreton et al. 1992) using a mix of null hypothesis testing and Akaike's information criterion.

Lebreton and Clobert went on with their research on the open models while we decided to study the model selection issue more thoroughly. We became aware of the multitude of alternatives available (see Burnham and Anderson 2002:37 for a brief summary) and the large technical literature (including several books) on model selection. We did not write the first edition of our book to blindly showcase AIC over other methods. We began our 7-8 year research by trying to understand the foundation of each of the model selection approaches. What assumptions were required in the derivation of the various approaches? Several papers and books helped to isolate a few good methods (e.g., AIC, BIC, RIC) while indicating that other approaches were almost universally poor (e.g, R² stepwise hypothesis testing). We began to understand that there were few mathematical errors in the literature, but deep differences in philosophy. These philosophical differences lead to differing approaches and it became important to sort these out in our first edition (Burnham and Anderson 1998).

We believe there are compelling reasons to use AICc and QAICc as effective methods for *general* use. For a given situation, a specialized method can potentially be developed and might be superior to Akaike's information-theoretic approach – we do not deny this (however see the 1998 Springer-Verlag book by McQuarrie and Tsai, *Regression and time series model selection*). However, there are, to us, extensive reasons to believe that AICc and QAICc are at the current state of the science for a general approach. The statement by Guthrey et al. might have been appropriate in 1990, but not today. How could these authors have read or understand the material in our 2002 book?

5. Guthery et al. (page 462) state,

"...'Akaike best' models suggest faulty data whenever the 'best' model does not contain the trivially obvious, such as year effects on the annual survival of an *r*-selected species."

This statement seems to be a farrago. Data analysis can show what inferences the data support, not the exact nature of full reality. Often year-specific parameterizations are simply not supported by the data available, particularly when sample size is small (see Section 3.5 of our book for an example using sage grouse). The 3 best models for the sage grouse data lacked a year-specific parameterization; but it is incorrect to claim,

therefore, that these data were "faulty." The use of likelihood ratio tests selected a model for these sage grouse data with 58 parameters ($\Delta = 36.3$), while AICc selected a parsimonious model with only 4 parameters. The evidence ratio (K = 4 vs. 58 parameters) was 76 million. Clearly, too much uncertainty (high variances) exists when trying to estimate 58 unknown year-specific parameters when the data set is fairly sparse. Still, just because the best model does not contain year-specific parameters is not a logical basis for claiming that the data are "faulty." There is an analogy with null hypothesis testing: failure to reject H_0 does not mean the data are faulty nor that the null hypothesis is true.

- 6. Kadane and Lazar (2004, JASA, 279-290 in Guthery's tirade) state that the frequentist criteria *are ad hoc*, having "no guiding principle." Our strong belief is that the deep guiding principle is Kullback-Leibler information loss and that methods stemming from this foundation allow a rigorous theory for model selection as an approximation to full reality. The Bayesian literature (including BIC) so often assumes that (1) a true model exists and (2) that it is in the set! This (true model) approach is not mathematically wrong; but it is philosophically absurd (e.g., once one has found this best model, are they to think that it represents full reality in all respects?). It is important to understand that data arise from reality the complex system of interest. There is no true model (that produced the data); data do not come from models!
- 7. One cannot avoid quantification in the empirical sciences. Rigor in science is lacking when only qualitative approaches are employed. Guthery et al. seem to deny or minimize all quantification and this is a poor approach as a science strategy. More emphasis must be placed on the science of our work, carefully hypothesizing alternatives, developing good models of these hypotheses, and using modern analytical methods to help understand and provide evidence for alternative *j*, given the hypotheses/models in the set. The calculations required under the information-theoretic approach are simple yet very effective. This allows a further focus on the science issues. Finally, the most powerful set of tools relates to making formal inferences from all the hypotheses/models in the set (multimodel inference).
- Stephens et al. (2005, Information theory and hypothesis testing: a call for pluralism, *Journal of Animal Ecology*, 42:4-12) try to suggest benefits from using both null hypothesis testing and information-theoretic approaches in data analysis and inference. The paper contains a number of serious errors and misunderstandings. A manuscript addressing some of these issues has been submitted to the editor of the *Journal of Animal Ecology* (February, 2006). People are certainly entitled to their own opinions and the field is strengthened by debate over technical issues. However, unless the published material is well grounded (and the authors having some basic competence in the fundamentals), then time and effort are generally sacrificed.

Next up: BIC consistent, AIC not