**Model selection and inference book (Burnham and Anderson)**

**Post hoc models/ following final model selection – p124**

Can new models reflecting these hypotheses be added to the set of candidate models and more AIC values computed? This question brings up several points. First, if this suggestion was made after examining

the estimates of *pt*2, *pt*3, *pt*4, *. . .* , *pp*7 vs. *pc*2 , *pc*3, *pc*4, *. . .* , *pc*7 and noting

that there seemed to be little difference between successive week-dependent

pairs, then this is a form of data dredging, and any subsequent results should

clearly detail the process by which the additional models were considered. **We**

**encourage full investigation of the data to gain all possible insights; we**

**only want investigators to reveal the extent of any data dredging that took**

**place**. Second, if that suggestion was made on conceptual grounds rather than

by studying the intermediate results, then the new class of models can be added

to the list, AIC computed, the *\_i* , *wi* , and evidence ratio values recomputed,

and inferences made. However, in this second case, the team could be somewhat

faulted for not considering the set of models more carefully in the first

place.

**p147**

Researchers often resort to using a computer program that will examine

all possible models and variables automatically. Here, the hope is that the

computer will discover the important variables and relationships (a “just the

numbers approach” void of any thinking or science). Cook et al. (2001:977)

conducted stepwise linear regression analyses using AIC and Mallows’s *Cp*

(SAS Institute 1988:786) in a study of elk (*Cervus canadensis*) condition.

They found that this approach “*. . .* provided results that often were biologically

unrealistic, unstable due to multicolinearity, and overparameterized (≥ 5

variables).” The literature is full of such failed studies; just because AIC was

used as a selection criterion does not mean that valid inference can be expected.

The primary mistake here is a common one: the failure to posit a small set of

a priori models, each representing a *plausible* research hypothesis.

The presentation of results in scientific publications should detail the logic

used in arriving at a set of candidate models. The model set should have strong

ties to study design and the alternative research hypotheses of interest. Presentation

and discussion of the log(*L*) values, *K*, the appropriate information

criterion for each model, *\_i* , and *wi* is recommended (see Anderson et al.

2001d). Evidence ratios should be presented with other relevant values to

allow a comprehensive assessment of the alternative hypotheses. Such information

allows the merits of each model to be contrasted. If some exploratory

data dredging was done following the formal analysis, this activity should be

clearly noted and the tentative insights from these activities provided

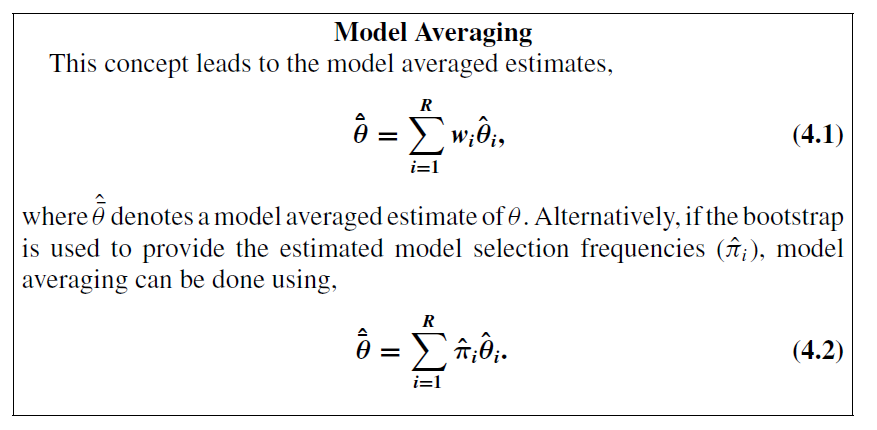
Chapter 4 – Multimodel inference

Part of multimodel inference includes ranking the fitted models from best to worst, based on the Δ*i* values, and then scaling to obtain the relative plausibility of each fitted model (*gi* )by a weight of evidence (*wi* ) relative to the selected best model. Using theconditional sampling variance (var( ˆ *θ*|*x, gi* )) from each model and the Akaikeweights (*wi* ), unconditional inferences about precision can be made over theentire set of models. Model-averaged parameter estimates and estimates ofunconditional sampling variances can be easily computed.

If one of the models was clearly the K-L best (e.g., if its *w* ≥ 0*.*90), then inference could probably be made, conditionally,

on the selected best model. However, it is often the case that no

single model is clearly superior to some of the others in the set.



The concept of inference being tied to all the models can be used to reduce

model selection bias effects on linear regression coefficient estimates in all

subsets selection. For the linear regression coefficient *βj* associated with predictor

variable *xj* there are two versions of model averaging.

Two different approachs:

* **Option 1 –** Average across models where the parameter is present

First, we have the estimate

ˆ¯ *βj* where *βj* is averaged over all models in which *xj* appears (i.e., when *j* is not zero):

*equation ….*

Thus,

ˆ¯ *βj* is a “natural” average to consider, as it only averages ˆ *βj* over models

where an unknown *βj* parameter appears. Note, however, that the estimator

ˆ¯ *βj* ignores evidence about models *gi* wherein *βj,i* ≡ 0.

* **Option 2** – Average across all model (if parameter not present set to 0)

An alternative way to average over linear regression models is to consider

that variable *xj* is “in” every model, it is just that in some models the corresponding

*βj* is set to zero, rather than being considered unknown. Conditional

on model *gi* being selected, model selection has the effect of biasing ˆ *βj,I* away

from zero (Section 1.6).

**Uncertainty associated with model selection and measures of unconditional precision**

There are three general approaches to assessing model selection uncertainty:

(1) theoretical studies, mostly using Monte Carlo simulation methods; (2) the

bootstrap applied to a given set of data; and (3) utilizing the set of AIC differences

(i.e., *\_i* ) and model weights *wi* from the set of models fit to data.

Useful

insights can be obtained about model selection and associated uncertainties

by extensive Monte Carlo simulations of model selection (e.g., McQuarrie

and Tsai 1998). Use of the bootstrap and ˆ*πi* values applies directly to a single

data set; hence they represent our focus here. The bootstrap may require

10,000 samples for reliable results, and it could take many hours of computer time to apply the bootstrap to complex data-analysis cases. In contrast, the

third method (i.e., use of *wi* values) is easily computable and merits more

development and understanding.

**Bootstrapping**

The fundamental idea of the model-based sampling theory approach to

statistical inference is that the data arise as a sample from some conceptual

probability distribution, *f* , and hence the uncertainties of our inferences can

be measured if we can estimate *f* . There are ways to construct a nonparametric

estimator of (in essence) *f* from the sample data. The fundamental idea of the

bootstrap method (Section 2.13) is that we compute measures of our inference

uncertainty from that estimated sampling distribution of *f* .

4.4 Estimating the Relative Importance of Variables (p167)

Data analysis is sometimes focused on the variables to include versus exclude in

the selected model (e.g., important vs. unimportant).Variable selection is often

the focus of model selection for linear or logistic regression models. Often, an

investigator uses stepwise analysis to arrive at a final model, and from this a

conclusion is drawn that the variables in this model are important, whereas the

other variables are not important. While common, this is poor practice and,

among other issues, fails to fully consider model selection uncertainty. Here,

we provide simple methods to quantify the evidence for the importance of each

variable in the set.

Consider 10 models based on combinations of a number of regressor variables.

Assume that the selected best model includes *x*1 and has an Akaike

weight of only 0.3. There is considerable model selection uncertainty here,

and hence there would seem to be only weak evidence for the importance of

variable *x*1 based on the selected best model. But one must consider the Akaike

168 4. Formal Inference From More Than One Model: Multimodel Inference (MMI)

weights of all other models that include *x*1 in order to quantify the importance

of *x*1. It might be that all models that exclude *x*1 have very low Akaike weights;

that situation would suggest that *x*1 is a very important predictor. The measure

of this importance is to sum the Akaike weights (or the bootstrap ˆ*π i* ) over the

subset of models that include variable *x*1. This idea is applicable in general

to model selection whenever it is equated to variable selection, for linear or

nonlinear models of any type.

Thus the evidence for the importance of variable

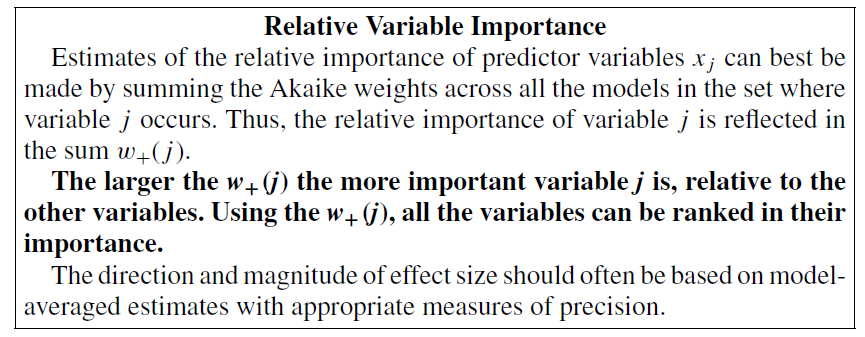
*x*3 is substantially more that just the weight of evidence for the best model.

We can order the three predictor variables in this example by their estimated

importance: *x*3, *x*1, *x*2 with importance weights of 0.85, 0.79, and 0.35. As

with other methods recommended here, we see that we are able to use model

selection to go well beyond just noting the best model from a set of models.



This idea extends to subsets of variables. For example, we can judge the

importance of a pair of variables, as a pair, by the sum of the Akaike weights

of all models that include the pair of variables. For the pair *x*1 & *x*2, the weight

of evidence for the importance of this pair is 0.19. For pair *x*2 & *x*3, the weight

of evidence for importance is 0.23, while for the pair *x*1 & *x*3, the weight of

evidence is 0.65 (compared to 0.5 for the selected model as such). Similar

procedures apply when assessing the relative importance of interaction terms.

When assessing the relative importance of variables using sums of the *wi* ,

it is important to achieve a balance in the number of models that contain each

variable *j* . For example, in the numerical example above, each of the three

variables appeared in four models. This balancing puts each variable on equal

footing.

To summarize, in many contexts the AIC selected best model will include

some variables and exclude others. Yet this inclusion or exclusion by itself

does not distinguish differential evidence for the importance of a variable in

the model. The model weights, *wi* or ˆ*π i* , summed over all models that include

a given variable provide a better weight of evidence for the importance of that

variable in the context of the set of models considered.

4.5 Confidence Set for the K-L Best Model (p169)

There exists a concept of a confidence set for the K-L best model based on the

data, just as there is a confidence interval for a parameter based on a model and

data. For a 95% confidence set on the actual K-L best model, a rational (but

not unique) approach is to sum the Akaike weights from largest to smallest

until that sum is just ≥ 0*.*95; the corresponding subset of models is a type of

confidence set on the K-L best model. In this example (assuming that we have

indexed the models as 1 to 7 in order of decreasing weights), the confidence

set is models {1, 2, 3, 4, 5}, which has sum of weights \_ 0*.*966. In using this

approach to a confidence set of models we are interpreting the Akaike weight

as a posterior probability (i.e., given the data and the set of a priori models)

that model *i* is the K-L best model (see Section 6.4.5). This is not the best

approach but it is easy to understand.

There is another approach to developing a confidence set of models based

on the idea of a *\_i* being a random variable with a sampling distribution.

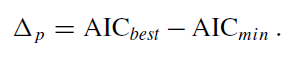
In particular, let index value *best* correspond to the actual expected K-L best

model in the set. There is always a K-L best model in the set of models (ignoring

that ties might occur). It is thus model *gbest* that we should use for the data

analysis; we just do not happen to know a priori the value of *best*. Then the *\_i*

of conceptual interest is



This unobservable random variable (*\_p*) is analogous to *θ*− ˆ *θ*, which can often

be used (after normalization by

\_

se( ˆ *θ*)) as a pivotal value for construction of a

confidence interval on *θ*. A pivotal quantity is one whose sampling distribution

is independent of any unknown parameters, a *t*-distributed pivotal, for example.

The “*p*” in the *\_* defined by (4.12) denotes that this *\_* is a conceptual pivotal

value rather than an actual *\_i* that we can compute from real data.

It is not exact to consider *\_p* \_ AIC*best* −AIC*min* as a pivotal quantity,

but it seems a useful approximation in some contexts. The context it seems

useful in is one of complex truth; tapering effect sizes; many models, some

being good approximations to truth, with full truth not in the set of models

used; and a lot of nested sequences of models (as in the starling experiment

example in Chapter 3). Monte Carlo studies on the above *\_p* can be done;

we have done many of these and results support the conclusion that in this

context, the sampling distribution of this *\_p* has substantial stability and the

95th percentile of the sampling distribution of *\_p* is generally much less that

10, and in fact generally less than 7 (often closer to 4 in simple situations). This

means that an alternative rule of thumb for an approximate 95% confidence

set on the K-L best model is the subset of all models *gi* having *\_i* ≤ some

value that is roughly in the range 4 to 7. In fact, the *\_* value to use when a

model is not competitive as a candidate for the K-L best model is variable, but

is probably somewhere between 2 and 10 in many situations. Thus, a *\_i* of 2

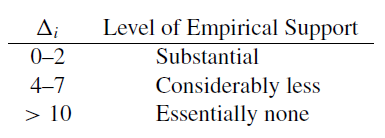
is not large, while a *\_i* \_ 10 is strong evidence against model *gi* being the

K-L best model in the set of models considered, if sample size is not small.

These guidelines, rough as they are, are useful.

We review this interpretation of evidence from the *\_i* when observations

are independent, sample sizes are large, and models are nested:



Models with*\_ >* 10 represent very strong evidence that the model is not the

K-L best model. The reader should not take these guidelines as inviolate since

as there are situations to which they do not apply well (such as when there is a

small sample size or dependent observations). Likewise, if there are thousands

of models, these guidelines may not hold.

4.9 It is clear that we select, based on strong evidence, the model structure

E(*y*) \_ *c* + *dz* as the better model. The estimates of the structural parameters

of this model, and their conditional standard errors, are ˆ *c* \_ −1917*.*6 (

\_

se \_

252*.*9), ˆ *d* \_ 183*.*3 (

\_

se \_ 9*.*3).

One wood data – p184

We do not have away we would consider reliable

to compute unconditional standard errors when a parameter is unique to a single

model (which *d* definitely is). However, when the evidence is strongly in favor

of the selected model, such as here, it is reasonable to act as if we considered

only that model, hence act as if that model would always be the one fit to such

data, in which case conditional standard errors apply. So here we accept use of

the conditional standard errors as a measure of estimator precision. As a rule

of thumb we will hazard the suggestion that if the selected model has Akaike

weight ≥ 0*.*90, it is acceptable to use the conditional standard errors. The

exact value (i.e., 0.90) is not critical; the concept is that if (and only if) the data

support the selected model strongly enough (*wmin* ≥ 0*.*9 seems also be a safe

rule of thumb; also see Royall (1997), where a similar rule is proposed),

*6.8.2 Use AICc, Not AIC, When K Is Large (p325)*

*Grueber et al (2011)*

*Model averaging – which models*

***1)* Cut-off criterion** to use to delineate a ‘top model set’, such as using the top 2AICC of models (Burnham & Anderson, 2002), top 6AICC (Richards, 2008), top 10AICC (Bolker et al., 2009) or 95% confidence (summed weight, Burnham & Anderson, 2002).

2**) Particular factor of interest** – not in top model : An added complication is how to decide what to do if a particular factor of interest (such as an experimental treatment) is not present in a model captured within the top model set (see Appendix: Step 4). Solutions in such cases are to either conclude that there is little evidence that the factor of interest explains variation in the response variable or extend the cut-off criteria to include at least one model that contains the factor of interest (for example, in cases where a parameter estimate is essential to further analysis). The latter solution may result in very large model sets, and ⁄ or inconsistent cut-off criteria for different response variables. High cut-offs are discouraged as they can lead not only to spurious results as described earlier but also to the inclusion of overly complex models (Richards, 2008). Such overly complex models may have similar weight as simpler versions in the set, and model averaging these can potentially result in overweighting the parameters they contain. Simulation studies have shown that removing complex models from the set does not necessarily impact the chance of selecting parsimonious models and also reduces the total number of models selected (Richards et al., 2010). A tentative solution therefore is **to exclude models from the set that are more complex versions of those with lower AICc** (Burnham & Anderson, 2002; Richards, 2008). However, careful scrutiny of these complex models may reveal that they are characterized by the presence of unique predictors of potentially strong biological importance and therefore in such cases should not be removed. Determining how to resolve the issue of nested models is likely to depend on the context of the particular study, but there are currently few clear guidelines on this

**Model averaging method**: After a top model set is defined, the method used to compute the model-averaged parameters should also be chosen carefully. There are two methods by which the estimate and error for each parameter are weighted (detailed in Burnham & Anderson, 2002; Nakagawa & Freckleton, 2010).

a) In the **so-called natural average method (**Burnham & Anderson, 2002; p. 152), the parameter estimate for each predictor is averaged only over models in which that predictor appears and is weighted by the summed weights of these models.

b) Alternatively, in the **so-called zero method** (Burnham & Anderson, 2002), a parameter estimate (and error) of zero is substituted into those models where the given parameter is absent, and the parameter estimate is obtained by averaging over all models in the top model set. Thus, the zero method decreases the effect sizes (and errors) of predictors that only appear in models with small model weights (particularly when the predictors have weak effects), diluting the parameter estimates of these predictors (shrinkage towards zero) (Lukacs et al., 2010).

Which method for which circumstances??? Although no clear distinction has been made as to the circumstances under which either of these two methods is more appropriate, Nakagawa & Freckleton (2010) recommend that the zero method should be used when the aim of the study is to determine which factors have the strongest effect on the response variable. Conversely, when there is a particular factor of interest and it is possible that this factor may have a weak effect compared to other covariates, the natural average method should be used to avoid shrinkage towards zero (

*Interpreting model estimates*

When model-averaged estimates are derived, it is essential to interpret both the direction (positive or negative) of parameter estimates and their magnitudes (effect sizes) in relation to one another (see Appendix: Step 4). Such an assessment can be problematic when input variables are measured on different scales (Gelman, 2008), and interactions are present.

Standardizing Centralizing predictors is essential when model averaging is employed, and standardization facilitates the interpretation of the relative strength of parameter estimates. In addition, it is recommended that input variables (not predictors) are standardized to a mean of 0 and a SD of 0.5 before model analysis (see Appendix: Step 2). The value 0.5 is used, rather than 1 SD, as this allows the standardization of binary predictors [and ⁄ or categorical variables, as ‘dummy variables’ are created (Schielzeth, 2010)] and continuous predictor variables to a common scale (Gelman, 2008; s

When interpreting the model, it is therefore important to remember that parameter estimates are on this scale. Such standardizations have sometimes been criticized (King, 1986; Bring, 1994; Hereford et al., 2004; Schielzeth, 2010) because parameter estimates are on the transformed scales, which are difficult to interpret biologically. However, back-transformations (described below) of these estimates are straightforward and we recommend that where point estimates of the response variable are derived, authors present them in the original scale (see Appendix: Step 5).

Using the model for prediction In many cases, the final model is ultimately used to generate a point estimate for the response variable under a given set of circumstances (i.e at fixed points for each predictor variable). In studies of inbreeding, for example, we are interested in comparing the predicted survival point estimates of highly inbred vs. outbred individuals (e.g. Keller & Waller, 2002). There are nearly unlimited combinations of predictor levels (‘conditions’) that could conceivably be substituted into the model statement to evaluate survival estimates, and the choice of levels made will depend on the question being investigated. For example, one may choose to use a ‘worst-case-scenario’ (by substituting in extreme values for the predictors) to compare the responses at one site to those of another, to compare conservation management strategies or any others. When predictors have been centred and standardized following the approach of Gelman (2008), one can substitute 0 as the mean and ð Þ xi x =ð Þ 2 rx for different levels (xi) of a parameter of interest (with a mean x and standard deviation rx) (see Appendix: Step 5). It is essential to remember to back-transform the result. Effects of a parameter of interest should be computed at the mean of all other parameters as a matter of routine, to allow comparisons across studies

**Missing data** - Model comparisons using IT approaches require data sets with no missing data, as deleting cases containing missing values can severely affect the results of model selection under IT approaches (Nakagawa & Freckleton, 2010). This has been recently covered in detail by other authors (Nakagawa & Freckleton, 2010).

Nonetheless, in the current discussion, we have identified a number of areas for more research:

• Which IT criteria should be used when comparing models, given the difficulties presented by including random factors?

• In determining the cut-off for a top model set when examining a factor of interest – how many models is ‘too many’ for model averaging?

• How should we decide which nested models to remove from the model set?

• How do we quantify model fit in mixed-effects models? In addition, we emphasize the importance of standardizing variables where model averaging is employed, as to fail to do so renders the results of model averaging uninterpretable in the presence of interactions (c.f. Schielzeth, 2010).