Natural Resources Data Analysis – Lecture Notes
Brian R. Mitchell

V. Week 5:

A. Multivariate AIC

1. For multivariate techniques such as MANOVA or discriminant analysis, there is no single error sum of squares, and there is generally no likelihood listed in statistical output. So how do you calculate AIC?

2. Get the ***pooled error SSCP*** matrix (this is sometimes called "within groups") from your stats package. ***Divide*** this matrix by n (sample size) to get the multivariate equivalent of the maximum likelihood estimate of σ^2 ("divide by n" means you should divide each matrix element). Then use a software package that can calculate the ***determinant*** of a matrix (the Excel formula MDETERM and Mathematica can both do this). The determinant of the matrix is the value you would plug in as σ^2 in the model selection spreadsheet we've been using in class (on the OLS page).

3. ***K*** in this situation includes each unique element in the matrix (i.e. the diagonal and the elements above OR below the diagonal. So K for a 5x5 matrix (i.e. 4 predictor variables and an intercept) is 5 for the betas (predictor and intercept estimates) + 5 for the sums of squares (matrix diagonal) + 10 for the cross-products (off-diagonal) = 20. For a p x p matrix, the contribution of the matrix elements is p(p +1)/2.

4. It turns out the B&A discusses multivariate AIC on pp. 424-426. David Anderson clarified K in a personal communication.

B. Discuss any analysis issues that have come up on individual projects

C. Model Averaging: Parameter Variance Revisited

1. There are 2 formulas for ***unconditional variance*** floating around.

a) The formula presented in Burnham and Anderson (2002):

$$\hat{\text{var}}\left(\bar{\hat{\theta}}\right) = \left[\sum_{i=1}^{R} w_i \sqrt{\hat{\text{var}}\left(\hat{\theta}_i \mid g_i\right) + \left(\hat{\theta}_i - \bar{\hat{\theta}}_i\right)^2}\right]^2$$

b) And the formula presented in Burnham and Anderson (2004)… as well as page 345 of Burnham and Anderson (2002):

$$\hat{\text{var}}\left(\bar{\hat{\theta}}\right) = \sum_{i=1}^{R} w_i \left[\hat{\text{var}}\left(\hat{\theta}_i \mid g_i\right) + \left(\hat{\theta}_i - \bar{\hat{\theta}}_i\right)^2\right]$$

c) The original formula uses the square root of the variance, and then squares the overall sum.

d) In pages 344-345 of Burnham and Anderson (2002), they say that:

(1) The ***first formula assumes a perfect correlation*** of estimates of θ from different models.
(2) The ***first formula introduces inconsistencies*** because the weights are nonlinear (i.e. the weights are squared in the first formula, and not in the second).
(3) The second formula is used in a Bayesian context when the model-averaged posterior is a mixture distribution, and can also be derived in the Kullback-Liebler framework.
(4) The first formula produces a model-averaged variance that is less than or equal to the second.

e) Another important difference between the formulas is that there is no rigorously derived ***covariance formula*** associated with the first formula, while there is for the second (it is presented in Burnham and Anderson (2004) and below in the section on model averaging an outcome).

2. How big is the difference between the two formulas?

a) Based on a quick look at some sample data, the ***difference is generally quite small*** (i.e. the change is in the third decimal place of the variance) if the parameter being averaged is in all models. The difference ***can be quite large*** (up to an order of magnitude difference in just one set of sample data!) if the parameter is not in all models; this is the situation where the assumption of a perfect correlation in estimates of θ across the model set is grossly violated.

b) My personal opinion is that the second formula is more defensible, and should be used instead of the first formula. However, the first formula is in wider use (and seems to still be emphasized in Anderson's workshops), so be sure to cite your use of this formula appropriately.

D. Model Averaging: $\bar{\hat{\theta}}$ versus $\bar{\tilde{\theta}}$

1. The material in B&A (2002) is very confusing regarding $\bar{\hat{\theta}}$ and $\bar{\tilde{\theta}}$. The difference between these two estimators hinges on what to do when a parameter is not in the model.

2. Burnham and Anderson initially assert that $\bar{\hat{\theta}}$ and its variance are calculated by only using models where the parameter occurs.

3. In contrast, $\tilde{\bar{\theta}}$ is calculated by using a zero when a parameter is not present in the model; Burnham and Anderson (2002) indicates that it is not possible to easily estimate the variance in this situation.

4. However, in their examples, workshops, and in Burnham and Anderson (2004), they describe $\tilde{\bar{\theta}}$ while calling it $\hat{\bar{\theta}}$, and in Burnham and Anderson (2004) they imply that the variance is computed by using a zero for the estimated parameter and its variance in the usual variance formula.

5. I think that the best match to the apparent intent of Burnham and Anderson is to: 1) use parameter estimates of zero and parameter variances of zero for parameters that are not present in a given model, and 2) call the estimators $\hat{\bar{\theta}}$ and $\hat{\text{var}}\,\hat{\bar{\theta}}$.

E. Model Averaging: Unconditional confidence interval

1. Once you have the unconditional parameter estimate and its variance, just calculate ***confidence intervals*** as you normally would. One typical approach is based on the z distribution:

2. $\hat{\bar{\theta}} \pm z_{1-\alpha/2}\sqrt{\hat{\text{var}}\left(\hat{\bar{\theta}}\right)}$

3. Demonstrate formula on spreadsheet.

F. Model Averaging: Estimating an Outcome (y) and Its Variance

1. Some background

a) In the following discussion, it is important to distinguish between ***x values*** (e.g. an actual measurement of temperature or some other variable) and ***β values*** (e.g. the model estimate of the slope that is applied to the x values in the model equation).

b) The ***good news*** is that you can always estimate your outcome parameter and its variance, even when your model set is not nested and you cannot model average your predictors.

c) The ***bad news***:

(1) A model-averaged outcome and variance is only valid for a specific combination of x values. For continuous variables, the mean is typically chosen, although you should consider calculating outcomes and variances for other important values. If you are using categorical predictors, you will need to pick some representative scenarios to calculate outcomes for.

(2) Calculating model averaged outcome variances can be a headache.

2. ***Estimating the outcome***

   a) When the model set is ***nested***

      (1) Write out your global model equation
      (2) Substitute your model averaged beta estimates and your x values
      (3) Add them up!
      (4) Example:
         (a) Given the regression equation: $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$
         (b) Given the model averaged estimates: $b_0 = 1.45$, $b_1 = 0.43$, $b_2 = 2.33$, and $b_3 = -0.68$.
         (c) For $x_1 = 1$, $x_2 = 0$, and $x_3 = 3.4$,
         (d) $y = 1.45 + 0.43*1 + 2.33*0 – 0.68*3.4 = -0.432$

   b) When the model set is ***not nested***

      (1) For each model, write out the model equation, substitute in the model-specific beta values and your chosen x values, and calculate your model-specific point estimate.
      (2) Once you have a point estimate of the outcome for each model, apply the formula for a model averaged parameter to calculate the model averaged outcome estimate:
      $\hat{\bar{y}} = \sum_{i=1}^{R} w_i \hat{y}_i$
      (3) Note that this method will also work for a nested set.

   c) Let's program these calculations on a spreadsheet.

      (1) For nested models: 1) copy model averaged estimates to a new sheet, 2) enter values for x, 3) use SUMPRODUCT to calculate the outcome.
      (2) For non-nested models: 1) copy model estimates to a new sheet, 2) copy weights to the new sheet, 3) enter values for x, 4) use SUMPRODUCT for each model to generate a model estimate, and then use SUMPRODUCT with those estimates and the model weights to get the model-averaged estimate.

3. Estimating the ***model averaged variance of the outcome***

   a) How do you calculate the ***variance of an outcome***?

      (1) This is a deceptively simple question. What you are actually trying to do is considered "***error propagation***"; you are summing a series of terms, each of which has a certain amount of uncertainty.

(2) The usual approach to dealing with this problem is the "***Method of Moments***", which can get pretty messy and includes lots of partial derivatives.
(3) Luckily, the problem is simplified in our case, since we are dealing with a linear equation, albeit one with correlated variables. The formula (from Hosmer and Lemeshow 2000, p. 41) to use in this situation is:

$$\hat{Var}(\hat{y}) = \sum_{j=0}^{p} x_j^2 \, \hat{\text{var}}(\hat{\beta}_j) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2 x_j x_k \, \hat{\text{cov}}(\hat{\beta}_j, \hat{\beta}_k)$$

(4) What do we need to know in order to use this formula? [x values, variances, and covariances]
(5) Let's program this formula on a spreadsheet: 1) Decide on x-values, 2) bring in variances and use SUMPRODUCT to multiply them with the squared x terms, 3) put x-values above and to the left of the matrix, 4) calculate a new matrix with the covariance term; clever use of relative vs. absolute formulas can speed this step up, 5) sum the formulas in the new matrix, 6) add with the SUMPRODUCT of the variances and the squared x values.

b) ***Method 1***: Model averaged outcome variance based on individual model outcomes

(1) For each model, determine the point estimate and variance of the outcome, using the error propagation formula and the same set of x values for each model.
(2) Use the formulas for averaging predictor estimates and variances:

$$\hat{\bar{y}} = \sum_{i=1}^{R} w_i \hat{y}_i$$

$$\hat{\text{var}}\left(\hat{\bar{y}}\right) = \sum_{i=1}^{R} w_i \left[\hat{\text{var}}(\hat{y}_i \mid g_i) + \left(\hat{y}_i - \hat{\bar{y}}_i\right)^2 \right]$$

c) ***Method 2***: Model averaged outcome variance based on model averaged predictor variances and covariances

(1) Calculate the predictor variable model averaged variances using the usual formula:

$$\hat{\text{var}}\left(\hat{\bar{\theta}}\right) = \sum_{i=1}^{R} w_i \left[\hat{\text{var}}\left(\hat{\theta}_i \mid g_i\right) + \left(\hat{\theta}_i - \hat{\bar{\theta}}_i\right)^2 \right]$$

(2) Calculate the covariance for each combination of predictor variables using the formula:

$$\hat{\text{cov}}\left(\hat{\bar{\theta}}, \hat{\bar{\tau}}\right) = \sum_{i=1}^{R} w_i \left[\hat{\text{cov}}\left(\hat{\theta}_i, \hat{\tau}_i \mid g_i\right) + \left(\hat{\theta}_i - \hat{\bar{\theta}}\right)\left(\hat{\tau}_i - \hat{\bar{\tau}}\right) \right]$$

(3) Then use the formula for calculating the variance of an outcome (note that θ and τ are replaced by θ_j and θ_k):

$$\hat{\text{var}}\left(\hat{\bar{y}}\right) = \sum_{j=0}^{p} x_j^2 \, \hat{\text{var}}\left(\hat{\bar{\theta}}_j\right) + \sum_{j=0}^{p} \sum_{k=j+1}^{p} 2 x_j x_k \, \hat{\text{cov}}\left(\hat{\bar{\theta}}_j, \hat{\bar{\theta}}_k\right)$$

d) ***Which method should be used?***

(1) ***Theoretically***, both methods should produce equivalent results with nested model sets. For non-nested model sets, Method 1 is the only option.
(2) Method 1 has been tested in ***simulation studies*** (David Anderson, personal communication), and the results were good. Method 2 has not been rigorously investigated.
(3) For my ***sample data set***, Method 1 yielded a variance estimate that seemed reasonable; the variance was slightly larger than the variances for the most heavily weighted models. Method 2 yielded a variance estimate that I felt was unreasonable; the variance was much lower than the estimates for the most heavily weighted models. My belief is that a variance that accounts for model selection uncertainty should be larger than the heavily weighted models, not smaller.
(4) Until this formula is revised appropriately, ***I suggest sticking with Method 1***.

e) Let's program Method 1 on a spreadsheet: 1) Calculate model specific variances as above, 2) for each model, add the variance to the squared difference between the outcome estimate and the model averaged outcome estimate, 3) multiply the result by the model weight, and 4) sum that result across the models.

f) Look at the difference between Method 1 and Method 2 on Brian's model averaging spreadsheet.

G. ***Examples of model averaging*** from B&A, Chapter 4

1. ***Cement Data***

a) Discusses model averaging the outcome variable without actually going over the procedure (they use SAS to get model-specific outcome estimates and variances).

b) Don't seem too concerned that the unconditional SE (1.9) is 33% smaller than the bootstrapped estimate of the unconditional SE (3.0). Note that if the alternate formula for variance is used, the unconditional SE is 2.3. However, the bootstrapped estimate of the SE would probably also be larger with this formula.

c) In this example, when calculating model averaged parameters, Burnham and Anderson only use the models where the parameter appears. This produces a ***biased*** (large) result, as they discuss in sections 4.2.2 and 1.6. My impression is that they no longer recommend this procedure.

d) They reduce their model set by 4 models (from 15 to 11, NOT 16 to 12 as they mistakenly write) because of colinearity between variable 2 and 4. This is a useful example of when you may be forced to ***alter the model set*** during an early stage of your analysis; Burnham and Anderson do not feel that this hurts a claim to a confirmatory analysis, and I agree.

e) They improperly apply their procedure for estimating ***parameter importance*** by using it on an unbalanced model set. In the original (full) model set, each parameter was represented in 8 models. In the reduced set, variables 1 and 3 were in 6 models, and 2 and 4 were in only 4. It was simply not valid to use their variable importance approach here.

2. ***Durban Storm Data***

a) As usual, I'm not convinced by the ***sample size*** used here. They suggest that sample size is 2,474 (once each week for 47-48 years). However, since the data are subsequently collapsed into counts by week, I think the sample size is 52 (essentially, the data for each week estimates the likelihood of rain in that week; we have the advantage of a good estimate of that probability, but we don't get to inflate the sample size). They seem to be using the correct sample size in table 4.4 for degrees of freedom, but they should clearly have used $QAIC_c$ instead of QAIC in this case.

b) The discussion of calculating the ***confidence interval*** of the probability of a storm for a given week is a little unclear. Essentially, they are calculating a model averaged outcome based on individual model estimates of the outcome and variance. This procedure is complicated by the logistic model; the safest strategy is to calculate your estimates and confidence intervals in logit space, then transform the end points into probabilities. By doing this for each week, they were able to generate the data for figure 4.2.

3. ***Flour Beetle Mortality***

a) Between 49 and 63 beetles tested at 8 different dosage levels, 471 beetles total. The models work with the probability of mortality at each dosage level. Is the ***sample size*** of 471 that Burnham and Anderson use really appropriate? My feeling is that the correct sample size is 8 (we get a good estimate of the mortality rate at each dosage level by using lots of beetles; but the number of individuals used to get that estimate is not relevant to the question at hand). I think that they should have used AIC_c for model selection.

b) Dosages ranged from 49 to 77 mg/L. How reasonable is it to use any model to estimate the effect of a 40 mg/L dose? In this type of problem

we can expect the mortality to be lower than at the 49 mg/L dose… but I don't think this sort of ***extrapolation*** is good science.

c) I find it interesting that they use ***different link functions*** for models in their model set. My impression was that this violates the mandate against transforming outcome values… but on further reflection and digging, I discovered that link functions are apparently applied to the predictors in statistical software (e.g. the program MARK help on link functions shows the functions applied to the predictors). So apparently it is OK to compare models that use different link functions (e.g. probit, logit, etc.). It is still not OK to blatantly transform an outcome variable (e.g. y in some models and log(y) in others).

H. Writing about and ***presenting model selection results***

Burnham and Anderson (2002) and Anderson et al. (2001) make numerous suggestions of things to include in papers that use model selection:

1. State your objectives and note whether your analysis is ***confirmatory or exploratory*** (*Introduction*)

a) If your analysis is confirmatory, make sure the details in your Methods section back this up.

2. Describe and justify the working ***hypotheses*** and the ***model set***, and how the models relate to the study objectives (*Methods*).

a) For large model sets, I recommend moving much of the model set description to an appendix. While it is OK to fully enumerate a large model set in a thesis or dissertation, I doubt that many journals will want to see this.

b) It may be simpler to describe your model set based on groups of parameters (e.g. the model set used four functional forms for time and four for season, in all possible combinations, to produce 16 models).

3. ***Justify*** your choice of AIC and other formulas with specific references (*Methods*)

a) You will need to cite your choice of AIC formula, model averaging formulas (i.e. parameter estimate, variance estimate, covariance estimate).

4. ***Document*** your methods and use citations whenever possible, especially when they differ from the "typical" approach (*Methods*).

a) For example, the appropriate reference for the use of zeroes in the point estimate and variance formulas when a parameter is not in the model is the monograph Burnham and Anderson (2004); in Burnham and

Anderson (2002) they use a different notation and do not overtly advocate this approach.

b) Some procedures that I advocate are not necessarily cited clearly anywhere (e.g. using a zero for the variance when the parameter is not in the model). In these cases, you will need to justify your approach.

c) Don't forget to reference the notations you use (e.g. w_i for Akaike weights).

5. Assess the fit of your ***global model*** (*Methods*)

a) Your methods section should state the procedure you use to assess fit (e.g. residual plots, Hosmer-Lemeshow test, etc.).

b) It is probably not necessary to present fit results (although it is worth including for a thesis or dissertation). For a research paper, it will likely suffice to say in your methods that the global model did indeed fit.

6. Include a table of ***model selection statistics***, including ln(L), K, AIC, AIC differences, and Akaike weights (*Results*).

a) Sometimes the raw AIC value or the ln(L) is not included; this can save space and one can be calculated from the other

b) It usually helps to sort by decreasing Akaike weight.

c) For large model sets, include all models with Akaike weights high enough to affect parameter estimates; exclude the rest and use a footnote stating that *n* models were excluded from the table and that the models did not affect inference.

7. Estimate ***important parameters*** and include ***confidence intervals*** (*Results*).

a) If you are model averaging an outcome, make sure that you choose an appropriate and representative sample of different predictors for your table or graph.

8. ***Don't mix*** frequentist and information-theoretic approaches

a) Avoid using the terms "significant" and "rejected", and don't use statistical significance tests.

9. Acknowledge ***data dredging***.

a) Emphasize that it is exploratory and that results need to be confirmed with future studies or a new data set.

10. In many situations the terms "***independent***" and "***dependent***" do not make sense when applied to variables; x variables are often correlated! It is clearer to use the terms "predictor" or "explanatory" and "outcome" to refer to x and y variables, respectively.

11. "Avoid confusing low frequencies with small sample sizes. If one finds only 4 birds on 230 plots, the proportion of plots with birds can be precisely estimated. Alternatively, if the birds are the object of study, the 230 plots are irrelevant, and the sample size (4) is very small" – Anderson et al. (2001), p. 377.

a) I couldn't agree more with this statement! I just wish it were applied more carefully in the examples that Burnham and Anderson (2002) use.