Biol 206/306 – Advanced Biostatistics
Lab 9 – Model Selection
Fall 2016

By Philip J. Bergmann

0. Laboratory Objectives

1. Learn why it is advantageous to consider multiple models *a priori*
2. Learn how to calculate the Likelihood of a linear model from its residual sum of squares
3. Learn how to calculate AIC, correct for small sample sizes, delta values, and weights
4. Learn how to compare models and select the best one(s) using the above values

1. Model Selection Background

Model selection implies that you start off with a number of different models and select which fits your data the best. This is a powerful approach because it doesn't simply assume which model fits the data best, but actually treats this part of modeling as an important component of analysis. Note that up to this point, most of the statistics that you have learned assume a linear model to the data and then simply fit that single model to the data, often testing some sort of null hypothesis. Models that are compared can be structurally very different from one another, with different numbers of parameters, and different properties. For example, one model may be linear, while another is logarithmic, and a third is exponential. Models can be very simple or complex, and can make different distributional assumptions about your data.

If you have two continuous variables and fit an OLS regression to them, what sorts of assumptions do you make about your data?

The comparison of models using the framework described here (Likelihood and AIC) is founded in information theory, where **entropy** is a measure of the unpredictability of the data, and **information** is a measure of how much a given model lowers entropy. Essentially, the model that maximizes predictive power has the highest information and is the best one. Models are generally fitted using a maximum likelihood framework. **Likelihood** is proportional to the probability that you would get your data if the model is true, or P(Data | Model). Model fitting algorithms work to find the **parameter** estimates that maximize the likelihood for a set of data. When likelihood is maximized, the parameter values are optimized, and when this is done for multiple models, you are ready to compare the models themselves (of course, in the context of your data).

In OLS regression we learned in a previous lab, what are the parameters of the model? What quantity is used to optimize these parameter estimates? How does this quantity tell you when the model is optimized?

2. Deciding on a Set of Models *a priori*

The first problem that comes up when using a model selection approach, is deciding which models to consider and how to build them. Some models are built from first principles and require considerable effort, training, and knowledge of mathematics and the underlying biology. In a growing number of cases and fields, researchers with this knowledge have implemented a range of relevant models that others can use with their datasets. Examples of this include models of DNA sequence evolution in phylogenetics, and models survivorship in population ecology for mark-recapture data. This increases the accessibility of a model selection approach. A third way to obtaining models is to simply use the linear models that you have been learning about and calculate their likelihoods. This is the approach we will be taking today because you may find it most useful and applicable to your own research.

We will be using the bird abundance dataset that you used in lab 5, and that comes from Loyn (*1987. Effects of patch area and habitat on bird abundances, species numbers and tree health in fragmented Victorian forests. In: Nature Conservation: The Role of Remnants of Native Vegetation (Saunders et al. eds.), pp. 65-77, Surrey Beatty & Sons, Chipping Norton, NSW, Australia*). ***Take a look at the dataset and refresh your memory of what it includes. What is the response variable, and what are the explanatory variables (use the log-transformed versions where available)?***

With five explanatory variables, the fully crossed model has 31 terms. You can create MANY linear models by including some or all of these 31 possible terms (there are literally thousands of possible models, or 5!=120 models if you were to ignore all interactions). Suffice it to say that there are more models than you can fit in a reasonable amount of time, and many of the models are not very interesting. By considering *a priori* which models may be most useful, you not only arrive at a more reasonable amount of work, but come away with a better understanding of the system you are studying. This approach may also limit the number of interactions included in a model, helping to exclude many complex models with many parameters that would likely result in imprecise parameter estimates. Not having collected the data ourselves, we will assume that all five explanatory variables were collected for a reason, and so will consider each in turn and in various combinations. ***Consider the five explanatory variables that are in the dataset. For***

each one, how would you expect it to relate to the response variable from a biological perspective, and why?

The following is the set of models that we will consider:

year	graze + larea	graze + year + alt + larea + ldist
graze	larea + ldist	graze * year * larea
alt	graze + larea + ldist	graze * year * alt * larea * ldist
larea	alt + year + larea	
ldist		

Looking back to lab 5, add one more model to the list provided here. Provide the model that you arrived at using backward stepwise elimination. Include only terms that are either significant or are the main effects of a significant interaction.

You should now have a list of 13 linear models. This is a lot, but far less than if you included all possible models without interactions, or all possible models in general. The next step will be to fit the models, record their fit to the data, using adjusted multiple R^2, and calculate their likelihoods. Likelihoods are generally reported as ln(likelihood), and fortunately, you can calculate ln(likelihood) from the residual sums of squares of each model. The following equation can be used to do this:

$$\ln(L(\hat{\theta})) = \frac{-n}{2} ln\left(\frac{SSe}{n}\right)$$

In this equation, the parameters of the model are θ, the ^ refers to estimates of the parameters, L is the likelihood for those parameter estimates, n is the sample size, and $SS\varepsilon$ is the residual sum of squares for the model. ***What is the sample size, n, for the bird abundance dataset?***

3. Fitting Models

The final piece of information that you need for each model is the number of parameters being estimated. A simple bivariate regression has two parameters: slope and intercept. For each additional term, there is an extra parameter. There is also one extra parameter in each model for estimating the variance, so a regression with one explanatory variable would have 3 parameters in total.

Assignment: (8 points)

Fit each of the above 13 models using the summary() and lm() functions. Save each summary table as an object. In the table below, fill in the 13th model from the previous page, and record the number of parameters (K), adjusted R^2, residual degrees of freedom, and residual sum of squares. You can calculate the residual sums of squares using the residuals() function on your summary objects. Residual sums of squares is simply the residuals squared and all of these values summed. Then calculate the ln(L) for each model using the equation on the previous page. Complete the table below wit this information. We will complete the last three columns shortly. Hint: If you do this table in Excel, you can fill down the equations, reducing risk of error.

Model	**K**	R^2	**df**$_\varepsilon$	**SS**$_\varepsilon$	**ln(*L*)**	**AIC**$_c$	Δ_i	w_i
year	3							
graze								
alt								
larea								
ldist								
graze + larea								
larea + ldist								
graze + larea + ldist								
alt + year + larea								
graze + year + alt + larea + ldist								
graze * year * larea								
graze * year * alt * larea * ldist								

In Excel, plot R^2 against SSe, and K against SSe. What do you notice about the relationships, particularly the second one? Explain why you see these patterns.

4. Comparing Models

We will compare our fitted models by completing the table on the previous page. The next step is using the likelihood to get an indication of the amount of information in each model. The Akaike Information Criterion (AIC) is used to do this. Note that although the best model tends to have the highest likelihood, the best model always has the lowest AIC. Also note that the AIC becomes biased when the sample size is small relative to the number of parameters in a model, and so a corrected version, AICc, is used when the sample size is less than about 40 times the number of parameters (this is just a rule of thumb). The AIC and AICc can be calculated as follows:

$$AIC = -2\ln\left(L(\hat{\theta})\right) + 2K \qquad \text{and} \qquad AICc = AIC + \frac{2K(K+1)}{n-K-1}$$

AIC and AICc also have terms that take into account the number of parameters in a model. ***As the number of parameters in a model increases, what happens to the AIC (and AICc)?***

If lower AIC values indicate better model fit, and more complex models typically fit a data set better, what is the purpose of the term in AIC that includes K?

Calculate AICc for each of the 13 models and fill in the appropriate column in the table. Which model is the best? Why? What is the worst model? Highlight the cells for the best model in yellow and make the text for the worst model red in the table.

Delta values, Δ_i, for each model *i*, are a measure of how much worse each model is than the best model, given the data. Delta values are simply calculated by first finding the lowest AIC or AICc, and then subtracting that value from each of the others. Delta values do not really contain any new information, but make it easier to interpret because the best model will have a value of zero. A typical rule of thumb is that models with delta values within about two of the best model are about equal to the best model. Models with deltas greater than two and up to around 8-10 are certainly worse than the best model, but may still have some support. Models with deltas >8-10 have essentially no support. ***Do this now and complete the appropriate column in your table or spreadsheet. Which models are pretty much of the same information content as the best one?***

Which models have at least some support if you use a $\Delta_i = 8$ as the cut-off?

The next step is to calculate model weights. These are derived from the Δ_i values and are useful because they give something akin to the probability of each model, given the set of models that you are considering. Model weights range from zero to one. The closer to one a model weight is, the more probable that is the right model from the set. Be very careful when using model weights because they are dependent on the set of models. If you add another model or get rid of one, the model weights will change. Model weights are calculated as:

$$w_i = \frac{e^{-\Delta_i/2}}{\sum e^{-\Delta_j/2}}$$

In the equation, *j* is the number of models. Note that *e* Euler's number or constant, and that in Excel *e to the power of x* is denoted as "EXP(x)". If you calculate the numerator for each model, then the denominator is simply the sum of those values. ***Calculate w_i for each model. What are the weights of the two best models?***

Finally, a good way of thinking about how much support there is for a given model relative to another model is using evidence ratios. These are not typically presented in a publication, but are helpful in thinking about how your models compare. An evidence ratio (ER) is simply a measure of how much more evidence for one model there is, relative to another model. An evidence ratio of 5 means that there is five times as much evidence for the better model than the worse model. You can use this to compare any two models in your set. If you consider the weights of two models, then the evidence ratio is:

$$ER = \frac{w_{larger}}{w_{smaller}}$$

<u>Assignment: (2 points)</u>

What is the evidence ratio for the best model relative to the second best?

What is the ER for the second best relative to the third best?

At first glance, model 10 (the one with K = 7) doesn’t seem too bad. What is the ER for the best model relative to model 10? How many times better is the best model?