4.8 Building models

In many situations, a scientist may have measured a large number of predictor variables that could be potentially important, and interest lies in determining which ones are best at explaining variation in the response data cloud and also whether particular combinations of variables, working together, do a better job than other combinations in this respect. More specifically, one may wish to build a model for the response data cloud, using the best possible combination of predictor variables available. There are two primary issues one is faced with when trying to build models in this way: first, what *criterion* should be used to identify a “good” model and second, what *procedure* should one use to select the variables on the basis of said criterion?

In addition to providing tests of hypotheses for specific regression-style problems, DISTLM also provides the user with a flexible model-building tool. A suite of selection procedures and selection criteria are available, as seen in the DISTLM dialog box and described below.

Selection procedures

- *All specified* will simply fit all of the variables in the predictor variables worksheet, either in the order in which they appear in the worksheet (by default) or in the order given explicitly under the ‘Available’ column in the ‘Selection’ dialog. The ‘Selection’ dialog can also be used to force the exclusion or inclusion of certain variables from this or any of the other selection procedures as well.
- *Forward selection* begins with a null model, containing no predictor variables. The predictor variable with the best value for the selection criterion is chosen first, followed by the variable that, together with the first, improves the selection criterion the most, and so on. Forward selection therefore adds one variable at a time to the model, choosing the variable at each step which results in the greatest improvement in the value of the selection criterion. At each step, the conditional test associated with adding that variable to the model is also done. The procedure stops when there is no further possible improvement in the selection criterion.
- *Backward elimination* begins with a full model, containing all of the predictor variables. The variable which, when removed, results in the greatest improvement in the selection criterion is eliminated first. The conditional test associated with removing each variable is also done at each step. Variables are eliminated from the model sequentially, one at a time, until no further improvement in the criterion can be achieved.
- *Step-wise* begins with a null model, like forward selection. First, it seeks to add a variable that will improve the selection criterion. It continues in this fashion, but what distinguishes it from forward selection is that, after every step, it attempts to improve the criterion by removing a term. This approach is therefore like doing a forward selection, followed by a possible backward elimination at every step. The conditional test associated with either the addition or removal of a given variable is done at each step. The procedure stops when no improvements in the selection criterion can be made by either adding or deleting a term. Forward selection is often criticised because it does not allow removal of a term, once it is in the model. The rationale of the step-wise approach responds directly to this

criticism.

- *Best* is a procedure which examines the value of the selection criterion for *all possible combinations* of predictor variables. One can choose the level of detail provided in the output file as ‘Normal’, ‘Brief’ or ‘Detailed’, which mirrors similar choices to be made when running the BEST procedure in PRIMER (see chapter 11 in Clarke & Gorley (2006)). The default output from the *Best* selection procedure in DISTLM is to provide the best 1-variable model, the best 2-variable model, and so on, on the basis of the chosen selection criterion. The overall 10 best models are also provided (by default) in the output, but this number can be increased if desired. Be aware that for large numbers of predictor variables, the time required to fit all possible models can be prohibitive.

Selection criteria

- R^2 is simply the proportion of explained variation for the model, shown in equation (4.1). Clearly, we should wish for models that have good explanatory power and so, arguably, the larger the value of R^2, the better the model. The main drawback to using this as a selection criterion is that, as already noted, its value simply increases with increases in the number of predictor variables. Thus, the model containing all *q* variables will always be chosen as the best one. This ignores the concept of *parsimony*, where we wish to obtain a model having good explanatory power that is, nevertheless, as simple as possible (i.e. having as few predictor variables as are really useful).
- Adjusted R^2 provides a more useful criterion than R^2 for model selection. We may not wish to include predictor variables in the model if they add no more to the explained sum of squares than would be expected by adding some random variable. Adjusted R^2 takes into account the number of parameters (variables) in the model and is defined as: $$ R ^ 2 _ {\text{Adjusted}} = 1 - \frac{ SS _ {\text{Residual}} / \left(N - \nu \right) } { SS _ {\text{Total}} / \left(N - 1 \right) } \tag{4.4} $$ where ν is the number of parameters in the model (e.g., for the full model with all *q* variables, we would have $\nu = q+1$, as we are also fitting an intercept as a separate parameter). Adjusted R^2 will only increase with decreases in the residual mean square, as the total sum of squares is constant. If adding a variable increases the value of ν without sufficiently reducing the value of $SS _ {\text{Residual}}$, then adjusted R^2 will go down and the variable is not worth including in the model.
- *AIC* is an acronym for “Akaike Information Criterion”, and was first described by Akaike (1973) . The criterion comes from likelihood theory and is defined as: $$ AIC = -2 l + 2 \nu \tag{4.5} $$ where l is the log-likelihood associated with a model having ν parameters. Unlike R^2 and adjusted R^2, smaller values of *AIC* indicate a better model. The formulation of *AIC* from normal theory in the univariate case (e.g., see Seber & Lee (2003)) can also be written as: $$ AIC = N \log \left(SS _ { \text{Residual}} / N \right) + 2 \nu \tag{4.6} $$ DISTLM uses a distance-based multivariate analogue to this univariate criterion, by simply inserting the $SS _ {\text{Residual}}$ from the partitioning (as is used in the construction of pseudo-*F*) directly into equation (4.6). Although no explicit qualities of statistical likelihood, *per se*, are necessarily associated with the use of *AIC* in this form, we see no reason why this directly analogous function should not provide

a reasonable approach. Unlike R^2, the value of *AIC* will not continue to get better with increases in the number of predictor variables in the model. The "$+ 2 \nu$" term effectively adds a "penalty" for increases in the number of predictor variables.

- *AIC*$_c$ is a modification of the *AIC* criterion that was developed to handle situations where the number of samples (*N*) is small relative to the number of predictor variables (*q*). *AIC* was found to perform rather poorly in these situations (Sugiura (1978) , Sakamoto, Ishiguro & Kitigawa (1986) , Hurvich & Tsai (1989)). *AIC*$_c$ is calculated as: $$ AIC _ c = N \log \left(SS _ { \text{Residual}} / N \right) + 2 \nu \left(N / \left(N - \nu - 1 \right) \right) \tag{4.7} $$
  In essence, the usual *AIC* penalty term ($+2 \nu$) has been adjusted by multiplying it by the following correction factor: (*N* / (*N* – ν – 1)). Burnham & Anderson (2002) recommend, in the analysis of a univariate response variable, that *AIC*$_c$ should be used instead of *AIC* whenever the ratio *N* / ν is small. They further suggest that a ratio of (say) *N* / ν < 40 should be considered as "small"! As the use of information criteria such as this in multivariate analysis (including based on resemblance matrices) is still very much in its infancy, we shall make no specific recommendations about this at present; further research and simulation studies are clearly needed.
- *BIC*, an acronym for "Bayesian Information Criterion" (Schwarz (1978)), is much like *AIC* in flavour (it is not actually Bayesian in a strict sense). Smaller values of *BIC* also indicate a better model. The difference is that it includes a more severe penalty for the inclusion of extraneous predictor variables. Namely, it replaces the "$+2 \nu$" in equation (4.6) with "$+ \log(N) \nu$" instead. In the DISTLM context, it is calculated as: $$ BIC = N \log \left(SS _ { \text{Residual}} / N \right) + \log \left(N \right) \nu \tag{4.8} $$ For any data set having a sample size of *N* ≥ 8, then log(*N*) > 2, and the *BIC* penalty for including variables in the model will be larger (so more strict) than the *AIC* penalty.

Depending on the resemblance measure used (and, to perhaps a lesser extent, the scales of the original response variables), it is possible for *AIC* (or *AIC*$_c$ or *BIC*) to be negative for a given model. This is caused, not by the model having a negative residual sum of squares, but rather by $SS _ {\text{Residual}}$ being less than 1.0 in value. When the log is taken of a value less than 1.0, the result is a negative number. However, in these cases (as in all others), smaller values of *AIC* (or *AIC*$_c$ or *BIC*) still correspond to a better model.

Although there are other model selection criteria, we included in DISTLM the ones which presently seem to have the greatest general following in the literature (e.g., Burnham & Anderson (2002)). For example, Godinez-Dominguez & Freire (2003) used a multivariate analogue to *AIC* in order to choose among competing models in a multivariate canonical correspondence analysis (CCA). However, the properties and behaviour of these proposed criteria are still largely unknown in the context of dbRDA, especially with multiple response variables (ρ > 1) and for non-Euclidean resemblance measures. More research in this area is certainly required. In the context of univariate model selection, *AIC* is known to be a rather generous criterion, and will tend to err on the side of including rather too many predictor variables; that is, to "overfit" (e.g., Nishii (1984) , Zhang (1992) , Seber & Lee (2003)). On the other hand, trials using *BIC* suggest it may be a bit too severe, requiring the removal of rather too many potentially useful variables. Thus, we suggest that if the use of *AIC* and *BIC* yield similar results for a given dataset, then you are probably on the

right track! One possibility is to plot a scatter-plot of the *AIC* and *BIC* values for the top 20 or so models obtained for a given dataset and see which models fall in the lower left-hand corner (that is, those which have relatively low values using either of these criteria). These are the ones that should be considered as the best current contenders for a parsimonious model. An example of this type of approach is given in an analysis of the Ekofisk macrofauna below.

Revision #25
Created 12 August 2022 09:27:54 by Arden
Updated 5 September 2022 13:11:02 by Arden