7

All Possible Regressions/Data Mining

This primer focuses on the method of all possible regressions. It is central to the approach described in Chapter 11 for variable selection under conditions of high dimensionality in the context of data mining. We illustrate the data mining approach in the worked example for this primer. Here we focus on the core logic of the all possible regressions method. We assume you have read the section on data mining in Chapter 11, but we repeat parts of it here to set context. We also assume you are familiar with multiple regression.

THE LOGIC OF ALL POSSIBLE REGRESSIONS

Given four predictors, X1, X2, X3, and X4, we might want to identify the regression equation that best predicts an outcome, Y. An easy way of doing so is to conduct a regression analysis for every possible combination of predictors and then choose the equation that performs best. In this case, the different regression equations have the following predictors (a) X1, (b) X2, (c) X3, (d) X4, (e) X1 + X2, (f) X1 + X3, (g) X1 + X4, (h) X2 + X3, (i) X2 + X4, (j) X3 + X4, (k) X1 + X2 + X3, (l) X1 + X2 + X4, (m) X1 + X3 + X4, (n) X2 + X3 + X4 and (n) X1 + X2 + X3 + X4. In general, there are $2^k - 1$ equations, where k is the number of predictors. Obviously, the best prediction is going to occur for the equation that contains all of the predictors; but the key question is by how much relative to the other equations and does one really need all of the predictors to achieve adequate prediction. Perhaps we can achieve functionally the same level of prediction but with a more parsimonious model. The all possible regressions approach provides perspectives on this.

In the social sciences, we often put a premium on theoretical parsimony. If one theory can adequately account for a phenomenon with fewer variables or fewer theoretical propositions than another theory, then we prefer the more parsimonious theory. If we are using the all possible regressions approach to help us determine variables to focus our theorizing on, then achieving functionally the same level of prediction with fewer predictors may be desirable.

As stated in Chapter 11, occasionally you may find yourself in situations where you have one or more outcomes you want to understand, a large number of possible explanatory variables (perhaps from a secondary data set), and little theory to guide you on the choice of explanatory variables to focus your theory on. You would like to conduct

exploratory analyses to identify empirically the most promising constructs. Our position is that it is best to avoid this scenario to the extent possible. Instead, you should read relevant literatures about the topic, conduct qualitative research on the topic (if the topic is amenable to such research), think long and hard about the variables in your data set, and then try to generate conceptual logic models for purposes of choosing a subset of explanatory variables. However, if you want the additional vantage point of data-driven suggestions for promising explanatory variables, the data mining approaches discussed in Chapter 11 may be of help. All possible regressions is one tool used in these approaches.

Sometimes you will be in situations where you have data with a large number of variables coupled with a large sample size. In other cases, you will have data with a large number of variables but a small sample size, even to the point where the number of variables may exceed the number of cases. The scenario of many predictors with small sample sizes is called *high dimensionality* or the *curse of dimensionality*. This scenario is our primary focus. The general idea is to divide the predictor pool into workable blocks of 7 to 10 variables, either based on theory or randomly. Then one performs all possible regressions analysis on each block. Based on these analyses, a subset of predictors from each block is retained and the retained predictors are combined into a common pool. These predictors are then blocked into smaller subsets of predictors and the process repeats itself through multiple iterations until a satisfactory subset of predictors is isolated for the outcome. Theory construction surrounding these variables can then commence.

Generally speaking, data mining is more concerned with prediction than explanation. However, one makes use of predictive power as a rough, imperfect tool that might provide one with theoretical ideas in the theory construction process.

CHOOSING MODELS

In this section, we briefly review indices of prediction accuracy in multiple regression.

Root Mean Squared Error

Suppose we wish to characterize the annual income of assistant professors at universities in the United States. If we obtain data on annual income for the entire population of such individuals, we might construct a model that "accounts for" scores on this variable. A simple, but obviously incorrect, model is one that predicts that every individual has a Y score equal to the mean of all the scores. This model can be written as

$$\hat{Y}_i = \mu$$

where μ is the mean calculated across all individuals, and $\hat{Y}$ is the predicted annual income for individual "i." For example, if the mean is $48,348, the model states that everyone has an annual income of $48,348. In multiple regression frameworks, this is called an intercept only model.

We define the errors in prediction as the difference between the observed and predicted scores:

$$\varepsilon_i = Y_i - \hat{Y}_i$$

where ε_i is the error score for individual "i." When working with models, we usually want to index how far off model predictions are, i.e., the magnitude of the errors. The average of the error scores across individuals is not a useful index, because the positive errors cancel the negative errors during summation and will always produce an average error of zero. A better index is a positive square root average of the error scores. For this index, we first calculate the average squared error:

$$\Sigma \varepsilon_i^2 / N$$

and then take the square root of this result to return the index to its original Y metric:

$$\sqrt{\Sigma \varepsilon_i^2 / N}$$

This index of average error is called the *standard error of estimate* or, somewhat more descriptively, the *root mean squared error*. It is often symbolized by σ_ε or by $\sigma_{Y-\hat{Y}}$. In our income example, if root mean square error (RMSE) is $3,012, this means that the predicted scores for the model are off, on average, by $3,012. The smaller the RMSE, the better the model's predictions. Note that one must take into account the metric of Y when interpreting the RMSE. If Y is the number of children married couples have in their family, then a RMSE that equals 3.0 for a model is considerable, indicating that predictions are off, on average, by 3 children. However, if Y is annual income in units of dollars, then a RMSE that equals 3.0 for a model is small, since predicted scores are off, on average, by only $3. The RMSE is an informative statistic, but it is rarely reported in social science research.

It turns out, the standard deviation of the outcome, Y, is the RMSE for an intercept only regression model. It is the average error in predictions when you predict everyone has a Y value equal to the mean. For this reason, some researchers compare the standard deviation of Y to the RMSE for a model to evaluate how much the predictors have reduced the errors in prediction relative to an intercept only model.

When we calculate the RMSE in a set of sample data, it will underestimate, on average, the RMSE in the population. Statisticians have developed bias corrections to adjust for this. The bias is larger the smaller the sample size, the greater the number of predictors, and the smaller the overall squared R for the prediction equation.

One way of choosing the "best" model from a set of all possible regressions is to focus on the RMSEs for the different regression equations. We first specify *a priori* a threshold value that we believe represents the minimum meaningful amount of error as reflected by the RMSE. For example, if we are predicting annual income, an RMSE of over $1,000 might be considered unacceptable. We then identify the most parsimonious model that has this amount of error or less. Alternatively, we identify the model with the best RMSE. We then locate more parsimonious models that are acceptably close to it in terms of their RMSE. They key to this approach is specifying what is "acceptably close."

Squared Multiple Correlation

Probably the most popular index of prediction errors in the research literature is the squared multiple correlation between the observed (Y) and predicted ($\hat{Y}$) scores. It is based on the concept of explained variance in the outcome measure. Specifically, variability on Y can be decomposed into two parts: (1) that which can be accounted for by the predictors and (2) that which represents errors in prediction (or other variables). The squared multiple correlation is the proportion of variability of Y that can be explained by the model and 1 minus the squared multiple correlation is the proportion of variability in Y that is due to errors in prediction.

It is well known that the sample squared multiple correlation is a biased estimator of the population squared multiple correlation. On average, the sample squared multiple correlation overestimates the population squared multiple correlation. The bias becomes less as the sample size increases, as the number of predictors decrease, and as the magnitude of the population squared multiple correlation increases. A correction factor has been proposed to produce an unbiased estimate of the population squared multiple correlation. This index is referred to as the "adjusted" squared multiple correlation. Although the correction removes bias, it does have shortcomings. For example, under some circumstances, the adjusted squared multiple correlation is negative, which is theoretically nonsensical. The usual practice is to set negative estimates to zero. But by setting negative values to zero, the adjusted squared multiple correlation becomes a biased estimator, much like the unadjusted estimate. To retain an unbiased statistic, one must live with a negative estimate. In addition, the adjusted squared R can exhibit larger sample to sample fluctuations than the unadjusted R square (i.e., it is less efficient). When the true proportion of explained variance in the population is 0.10 or greater and

the total sample size is > 80, there tends to be only minor differences between the adjusted and unadjusted indices.

Another way of choosing the "best" model from a set of all possible regressions is to focus on the squared correlations (or adjusted squared correlations) for the different regression equations. We first specify *a priori* a threshold value that we believe represents the minimum meaningful decrement in the squared R, such as a decrement of 0.05 or 5% explained variance. We specify what an acceptable squared R is and then choose the most parsimonious model that meets or exceeds that value. Alternatively, we identify the model with the best R squared. We then locate more parsimonious models that are acceptably close to it in terms of their R squared. They key, again, is specifying what is "acceptably close."

Information Theory Indices

A third class of statistics for choosing a model uses information theory indices, namely the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). These indices are discussed in the primer on regression mixture modeling. The section from that primer with minor edits is appended to this primer for easier access. These indices are related to the likelihood of observing the sample data given that a population model (in this case, a given regression equation) is true. The indices are such that the lower the value, the greater the likelihood of the model, everything else being equal. The indices include a penalty function for lack of parsimony. For interpretational details, see the appendix. To choose a model, one finds the model with lowest AIC or BIC values and then finds the most parsimonious model that is closest to it that has a comparable AIC or BIC using the guidelines in the appendix. The worked example accompanying this primer uses this approach.

THE ROLE OF SIGNIFICANCE TESTS

Tests of statistical significance usually are not relied upon when selecting models in an all possible regressions approach. The underlying statistical theory for making valid comparisons in such contexts is not well developed.

CATEGORICAL PREDICTORS

For the types of exploratory analyses described in Chapter 11 that use all possible regressions, a simple way to deal with categorical variables is to create a single pseudo-continuous variable with the mean Y value for the group the individual is in as that individual's score on the pseudo-variable. For example, if there are three levels of the

categorical variable with Y means of 1.1, 1.5 and 2.0, respectively, a new variable is created for use in the analysis that assigns to all people in group 1 a score of 1.1., to all people in group 2 a score of 1.5, and to all people in group 3 a score of 2.0. Then use that one "continuous" variable as a predictor in the model in place of the categorical variable.

CONCLUDING COMMENTS

All possible regressions is a statistical method used in data mining to help identify viable predictors of an outcome. It is quite demanding computationally as the number of predictors becomes large. For example, for 10 predictors, there are 1,023 regression equations that result. Most data mining uses of it are pursued in the context of algorithms that dictate 10 or fewer predictors at a time. The method has limitations and can identify some false positives as well as false negatives, but it also has some informational value.

REFERENCES

Burnham, K. & Anderson, D. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, 33, 261-304.

Kuha, J. (2004). AIC and BIC : Comparisons of assumptions and performance. *Sociological Methods and Research*, 33, 188-208.

Raftery, A.E. (1995). Bayesian model selection in social research (with discussion). *Sociological Methodology*, 25, 111-195.

Sclove, L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, 2, 333-343

APPENDIX: INFORMATION INDICES FOR MODEL CHOICE

When choosing between the different models to determine the number of classes, a commonly used set of comparative fit indices is based in a statistical theory known as *information theory*. Two such indices are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). In general, researchers calculate an AIC index and/or a BIC index for the different models and then choose the model that has the best BIC or AIC value. In this appendix, we develop the logic of these indices, taking a few liberties in the interest of pedagogy. We first develop the concept of a log likelihood, a concept that is central to both the AIC and BIC. We then describe the model comparison process for the AIC, followed by consideration of that process for the BIC.

Log Likelihoods

Suppose we have a very large population and half the population is male and half the population is female. The probability of a randomly selected case being a male is 0.50 and this also is true for being a female. Stated more formally:

p(male) = 0.50 p(female) = 0.50

If we randomly select two cases, the probability of a given joint result across the two selections or "trials" is the product of their probabilities. As such, the probability of observing two males is

p(male)*p(male) = (0.50)(0.50) = 0.25

This is known as the multiplication rule for independent trials. Stated more formally, let p(A) = the probability of event A on a trial and p(B) = the probability of event B on a second (independent) trial. The joint probability of both events A and B is the product of the individual probabilities p(A) p(B). To be more concrete, there are four combinations that can result, each with a probability of 0.25:

Probability of a male on the first trial followed by a male on the second trial: 0.25
Probability of a male on the first trial followed by a female on the second trial: 0.25
Probability of a female on the first trial followed by a male on the second trial: 0.25
Probability of a female on the first trial followed by a female on the second trial: 0.25

and if we do not care about the order of appearance in the trials,

Probability of two males: 0.25
Probability of a male and a female: 0.50
Probability of two females: 0.25

We now shift gears review another facet of statistical theory that we will make use of. If we know that a very large set of scores is normally distributed with a certain mean and standard deviation, then we can use knowledge of the probability density function for a normal distribution to compute the probability of obtaining any given value when we randomly select a case from that distribution. The density formula is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{\frac{-.5(x-\mu)^2}{\sigma^2}}$$

where x is the score value in question, μ is the mean of the distribution, σ is the standard deviation of the distribution, π is the mathematical constant pi, *e* is the constant associated with the Naperian logarithm, and the density describes the height of the normal curve at the value of x. We can use this density in conjunction with calculus to calculate the probability of observing the score in question. As an example, if scores are normally distributed with a mean of 100 and a standard deviation of 13.77, then, using the above formula, we find that the likelihood of a score of 99 is 0.0289. For a score of 87, it is 0.0186.[1]

Suppose we randomly select two scores from an extremely large population where scores are normally distributed with a mean of 100 and a standard deviation of 13.77. The probability that the scores will be 87 and 99, using the joint probability theorem described above, is (0.0289)(0.0186) = 0.00053754. Stated another way, the probability of observing these two data points given that the mean is 100 and the standard deviation is 13.77 (and assuming a normal distribution) is 0.00053754, with further adjustments to account for disinterest in the order of selection.

Suppose we randomly sample 100 data points from the population and calculate the likelihood of those 100 data points occurring using a strategy similar to the above method. The strategy would involve multiplying each probability by one another, with the result being a very, very small number. To make things more manageable and so as not to work with such small numbers, statisticians transform the final result by calculating the log of it, yielding what is called a *log likelihood*. The log likelihood is indicative of (but not equal to) the probability of obtaining the sample data given a "model" that states (a) the scores are normally distributed, (b) the mean is 100, and (c) the standard deviation is 13.77.

Log likelihoods are negative because the log of numbers less than 1.00 is always negative. For example, the natural log of 1.00 is zero, the natural log of 0.50 is -0.69, the natural log of 0.25 is -1.39, and the natural log of .01 is -4.61.[2]

Now, let's turn the above situation on its head. Suppose we have a set of 100 data points but we do not know the mean and standard deviation of the (assumed normal) distribution from which they come. We might, based on theory or logic, decide to "test" a model that states the mean is 95 and the standard deviation is 15. Using the probability density function from above and the strategies described, we can calculate the log likelihood for this model. The closer the log likelihood value is to zero (i.e., the less

[1] Technically, the probability of observing an exact value for a continuous variable is zero. We compute the likelihoods here by focusing on the interval defined by the real limits of the number (e.g., 98.5 to 99.5) in conjunction with the integral that scales the area under the curve to 1.00.

[2] Actually, some operationalizations of log likelihoods can yield positive numbers, but discussion of this point is beyond the scope of this primer.

negative it is), the more likely the data came from the postulated model. We might formulate a second (competing) model that the mean is 100 and the standard deviation is 13.75 and calculate the log likelihood for it. Again, the closer the value of the log likelihood for this model is to zero, the more likely it is the data came from the model positing a mean of 100 and a standard deviation of 13.75.

We can compare the log likelihood values for the two models and we might find that one model results in a log likelihood closer to 0 than the other model. The model with the log likelihood closer to zero is more likely to have produced the data, hence we would prefer it to the model with the more negative log likelihood. Such is the fundamental logic of choosing between models based on their relative log likelihoods: We calculate the log likelihood of competing models and then choose the model with the log likelihood that is closest to zero. To be sure, the above explanation is simplistic and glosses over technicalities, but hopefully it conveys the general idea of comparing log likelihoods for two models.

As an aside, the above logic also is central to the well-known method of estimation called *maximum likelihood estimation*. In this approach, to estimate the mean of a distribution, one conceptually posits different models each representing a possible population mean value, calculates the likelihood of observing the data given the "model," and then selects the value/model that has the maximum likelihood.

Model Comparisons using the AIC

The AIC is an index of model likelihood or "model fit" based on a log likelihood. A common representation of it is

$$AIC = (-2)\ (LL) + 2k \qquad [1]$$

where LL is the log likelihood associated with the model in question and k is the number of estimable parameters in the model (such as when we estimate an intercept and the various regression coefficients). By multiplying the log likelihood by -2, the AIC essentially becomes a positive number, with larger numbers indicating lower likelihoods of the model. The AIC also includes what is often referred to as a penalty function for lack of parsimony, namely 2k. If the model has many parameters in it that must be estimated, then the AIC will be larger, everything else being equal. With the AIC, model parsimony is rewarded.[3] In general, the smaller the value of AIC, the better the "fit" of the model to the data. To make this intuitive, if the probability of the data given the

[3] Technically, the 2k term is part of the mathematical theory underlying the derivation of AIC. Also, choosing the value of -2 to multiply the LL by is not arbitrary. This value has a clear rationale. See Burnham and Anderson (2004).

model is 0.25, the log likelihood will be -1.39 and multiplying this by -2 yields 2.78. If the probability of the data given the model is much higher, say 0.50, the log likelihood is -0.69 and multiplying this by -2 yields 1.38. So, the smaller the value, the better the model. To this term, a penalty function is added that inflates the value of AIC for models that estimate more parameters

There are many variations of the AIC. For example, some researchers use the above formula but with a small sample bias correction incorporated into it. This is sometimes referred to as AIC_c. The nuances of the different versions of the AIC are described in Burnham and Anderson (2004). Do not be surprised if for some software you observe AIC indices that are quite different in magnitude from other software. The important idea for all them is that we can compare different models using their respective AICs and then choose models that have "better" AICs when compared to other models.

Sometimes we compare more than two models, i.e., we might compare three, four or five models. When comparing more than two models, it is common to first identify the model with the lowest AIC value (which is the best fitting model of all the models being considered). One then calculates the difference in AIC values between each of the models and this best fitting model (subtracting the latter from the former). For the best fitting model, the difference will be zero and for all other models, it will be positive in value, with the larger the disparity, the worse the fit of the target model relative to the best fitting model.

General rules of thumb have been proposed to contextualize the magnitude of the difference in AICs between models (see Burnham & Anderson, 2004). The most common rules of thumb are as follows:

1. If the disparity in AICs is < 2, then the two models have about the same support

2. If the disparity in AICs is > 2 and < 4, then the better fitting model has positive support relative to the model it is compared with

3. If the disparity in AICs is > 4 and < 10, then the better fitting model has strong support relative to the model it is compared with

4. If the disparity in AICs is > 10, then the better fitting model has very strong support relative to the model it is compared with.

Of course, one must be careful when applying rules of thumb like this because they may not apply in all contexts. Indeed, some analysts object to their specification, arguing that

they can result in the same rigid and counterproductive use of a criterion like "$p < 0.05$" that plagues null hypothesis testing frameworks.

Another standard for comparing two models vis-a-vis the AIC is to examine what is called the *evidence ratio*. Let D = the AIC for the worse fitting model of the two models minus the AIC for the better fitting model of the two models (and let *e* be the traditional Naperian constant). The evidence ratio is defined as

$$ER = 1 / e^{(-D / 2)}$$

where ER stands for "evidence ratio." It indicates how much more likely the better fitting model is (given the data) than the worse fitting model (given the data). For example, if the AIC for the better fitting model is 100 and for the worse fitting model it is 102, then the evidence ratio is

$$1 / e^{-(102-100) / 2)} = 2.63$$

The better fitting model is 2.63 times more likely to have yielded the data than the model it is being compared with.

Finally, some researchers normalize AIC differences relative to all models being compared so that they sum to 1. These are called *Akaike weights* and indicate the "weight of evidence" in favor of a model relative to *all* models in the comparison set. Akaike weights are distinct from evidence ratios because Akaike weights are impacted by the particular set of models being compared when the number of models is greater than two. Let us first describe how Akaike weights are calculated and then we will make them more concrete with an example.

To calculate the Akaike weight, each model is assigned an index of its likelihood relative to that of the best fitting model using the value from the denominator of the evidence ratio, $e^{(-D/2)}$, as the index. Let T = the sum of the $e^{(-D/2)}$ values across all the models being considered. Then the Akaike weight for a given model is defined as

$$e^{(-D/2)} / T$$

The weight ranges from 0 to 1.00, with higher values favoring the model in question.

To make this concrete, suppose we fit five different models to a set of data. Here is a table with the AICs, the differences between the model AIC versus the model with the lowest AIC, and the Akaike weights (w):

Model	AIC	D	$e^{(-D/2)}$	$w = e^{(-D/2)}/T$
1	204	2	0.3678	0.2242
2	202	0	1.0000	0.6094
3	206	4	0.1353	0.0824
4	206	4	0.1353	0.0824
5	214	12	0.0024	0.0015
Sum			T = 1.6408	1.0000

The sum of the weights across all five models is 1.00. The weights represent a continuous measure of relative strength of evidence for each model. Each weight can be crudely interpreted as the probability that the model is the best model among the set. In the present case, the data support Model 2.

The basic idea when evaluating models is to examine multiple criteria, including the magnitude of the difference in AICs, the evidence ratios, the Akaike weights, and the substantive meaning/logical coherence of the models, in order to choose the best one.

Model Comparisons using the BIC

We describe the logic of the BIC using the Schwartz BIC, which is formally defined as

$$BIC = -2\ LL + \ln(N)\ k \qquad [2]$$

where k = the number of estimable parameters in the model, N = the sample size, and LL = the model log likelihood. Like the AIC, the smaller the BIC, the better the model fit, everything else being equal. Like the AIC, there is a penalty function for lack of parsimony, but the penalty is different than the AIC. The penalty is somewhat harsher for the BIC as opposed to the AIC. There are other instantiations of the BIC, and we discuss these below. For current purposes, we use the Schwartz formulation.

Like the AIC, it is not uncommon for the model with the smallest BIC to be used as a reference point for comparing models, with a common practice being to calculate the difference between each model in the model set and the model with the best BIC, like we did for the AIC. For the best fitting model, this difference will be zero.

To evaluate models in terms of BIC differences, general rules of thumb are (see Raftery, 1995):

1. If the BIC disparity < 2.2, then the better fitting model and the model it is compared with have about the same support

2. If the BIC disparity > 2.2 and < 6, then the better fitting model has positive support relative to the model it is compared with

3. If the BIC disparity > 6 and < 10, then the better fitting model has strong support relative to the model it is compared with

4. If the BIC disparity > 10 then the better fitting model has very strong support relative to the model it is compared with

For similar but slightly different standards, see Wasserman (1997).

One also can calculate what is called a *Bayes Factor* (BF) for each model relative to the best fitting model. It is defined as

$BF = e^{(D'/2)}$

where D' is the BIC difference between the target model and the best fitting model. The Bayes factor is the probability that the model with the lower BIC produced the data divided by the probability the model in question produced the data. For example, a BF = 10 means it is 10 times more likely the model with the minimum BIC produced the data than the model in question.

Finally, a relative model weight, analogous to the Akaike weight, can be computed by normalizing model likelihoods relative to *all* models in the comparison set so that they sum to 1. Let D = the difference in the BIC for the model in question minus the value of the BIC for the best fitting model, T = the sum of the index $e^{(-D/2)}$ across each model. The relative weight for a model is

$e^{(-D/2)} / T$

The weight ranges from 0 to 1.00, with higher values favoring the model. Again, the sum of the weights across models is 1.00.

As with the AIC, the basic idea when evaluating models is to examine multiple criteria, including the magnitude of the difference in BICs, the Bayes factors, the relative weights, and the substantive meaning/logical coherence of the models, in order to choose the best one.

You will encounter variants of the BIC, but the basic logic in applying them is the same. For example, like the AIC_c, there is a sample size adjusted BIC that is similar to

Schwartz' BIC, but it applies a somewhat milder penalty function (Sclove, 1987). There also are variants of both the AIC and BIC to deal with dispersion issues in count regression models (called QAIC and QBIC).

Which Method is Better, AIC or BIC?

A debated topic in statistics is which approach to model comparison is better, one based on AICs or one based on BICs. There are advocates on both sides of the matter and we dare not venture into this controversy here. The BIC tends to favor simpler models more so than the AIC. This can be both a strength and a weakness. Interested readers are referred to Burnham and Anderson (2004), Yang (2005), and Kuha (2004). Kuha argues for the use of both indices.

An issue with both approaches is that researchers can be lulled into thinking that the best fitting model within a set of models is the true model. This is not necessarily the case. Researchers can choose the best of a set of wrong models, which is not our goal.