23. Comparing models using Akaike's Information Criterion (AIC)

The previous chapter explained how to compare nested models using an F test and choose a model using statistical hypothesis testing. This chapter presents an alternative approach that can be applied to both nested and non-nested models, and which does not rely on P values or the concept of statistical significance.

Introducing Akaike's Information Criterion (AIC)

Akaike developed an alternative method for comparing models, based on information theory. The method is called Akaike's Information Criterion, abbreviated AIC.

The logic is not one of hypothesis testing, so you don't state a null hypothesis, don't compute a P value, and don't need to decide on a threshold P value that you deem statistically significant. Rather, the method lets you determine which model is more likely to be correct and quantify how much more likely.

Unlike the F test, which can only be used to compare nested models, Akaike's method can be used to compare either nested or nonnested models.

The theoretical basis of Akaike's method is difficult to follow. It combines maximum likelihood theory, information theory, and the concept of the entropy of information (really!). If you wish to learn more, read *Model Selection and Multimodel Inference -- A practical Information-theoretic approach* by KP Burnham and DR Anderson, second edition, Springer, 2002. It presents the principles of model selection in a way that can be understood by scientists. While it has some mathematical proofs, these are segregated in special chapters and you can follow most of the book without much mathematical background.

How AIC compares models

While the theoretical basis of Akaike's method is difficult to follow, it is easy to do the computations and make sense of the results.

The fit of any model to a data set can be summarized by an information criterion developed by Akaike.

If you accept the usual assumptions of nonlinear regression (that the scatter of points around the curve follows a Gaussian distribution), the AIC is defined by the equation below, where N is the number of data points, K is the number of parameters fit by the regression plus one (because regression is "estimating" the sum-of-squares as well as the values of the parameters), and SS is the sum of the square of the vertical distances of the points from the curve.

$$\text{AIC}=\text{N}\cdot\ln\left(\frac{\text{SS}}{\text{N}}\right)+2\text{K}$$

When you see a new equation, it is often helpful to think about units. N and K are unitless, but SS is in the square of the units you choose to express your data in. This means that you can't really interpret a single AIC value. An AIC value can be positive or negative, and the sign of the AIC doesn't really tell you anything (it can be changed by using different units to express your data). The value of the AIC is in comparing models, so it is only the difference between AIC values that you care about.

Define A to be a simpler model and B to be a more complicated model (with more parameters). The difference in AIC is defined by:

$$\begin{aligned}\Delta\text{AIC}&=\text{AIC}_B\text{-AIC}_A\\&=N\left[\ln\left(\frac{SS_B}{N}\right)-\ln\left(\frac{SS_A}{N}\right)\right]+2(K_B\text{-}K_A)\\&=N\cdot\ln\left(\frac{SS_B}{SS_A}\right)+2\cdot\left(K_B\text{-}K_A\right)\end{aligned}$$

The units of the data no longer matter, because the units cancel when you compute the ratio of the sum-of-squares.

The equation now makes intuitive sense. Like the F test, it balances the change in goodness-of-fit as assessed by sum-of-squares with the change in the number of parameters to be fit. Since model A is the simpler model, it will almost always fit worse, so SS_A will be greater than SS_B. Since the logarithm of a fraction is always negative, the first term will be negative. Model B has more parameters, so K_B is larger than K_A, making the last term positive. If the net result is negative, that means that the difference in sum-of-squares is more than expected based on the difference in number of parameters, so you conclude Model B (the more complicated model) is more likely. If the difference in AIC is positive, then the change in sum-of-squares is not as large as expected from the change in number of parameters, so the data are more likely to have come from Model A (the simpler model).

The equation above helps you get a sense of how AIC works – balancing change in goodness-of-fit vs. the difference in number of parameters. But you don't have to use that equation. Just look at the individual AIC values, and choose the model with the smallest AIC value. That model is most likely to be correct.

A second-order (corrected) AIC

When N is small compared to K, mathematicians have shown that AIC is too small. The corrected AIC value (AIC_C) is more accurate. AIC_C is calculated with the equation below (where N is number of data points, and K is the number of parameters plus 1):

$$\text{AIC}_C=\text{AIC}+\frac{2K(K+1)}{N\text{-}K\text{-}1}$$

If your samples are large, with at least a few dozen times more data points than parameters, this correction will be trivial. N will be much larger than K, so the numerator is small compared to the denominator, so the correction is tiny. With smaller samples, the correction will matter and help you choose the best model.

Note: The AIC_c can only be computed when the number of data points is at least two greater than the number of parameters.

We recommend that you always use the AIC$_c$ rather than the AIC. With small sample sizes commonly encountered in biological data analysis, the AIC$_c$ is more accurate. With large sample sizes, the two values are very similar.

The change in AICc tells you the likelihood that a model is correct

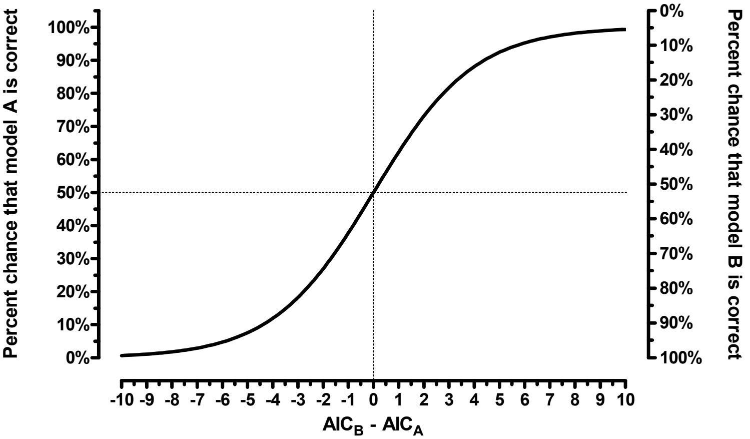
The model with the lower AIC$_c$ score is the model more likely to be correct. But how much more likely?

If the AIC$_c$ scores are very close, there isn't much evidence to choose one model over the other. If the AICc scores are far apart, then the evidence is overwhelming. The probability that you have chosen the correct model is computed by the following equation, where Δ is the difference between AIC$_c$ scores.

$$\text{probability}=\frac{e^{-0.5\Delta}}{1+e^{-0.5\Delta}}$$

Note: The probabilities are based on the difference between AIC$_c$ scores. The probabilities are the same if the AIC$_c$ scores are 620,000 and 620,010 as they are if the AIC$_c$ scores are 1 and 11. Only the absolute difference matters, not the relative difference.

This graph shows the relationship between the difference in AIC (or AIC$_c$) scores and the probability that each model is true.



Note: These probabilities are also called Akaike's weights.

If the two AIC$_c$ values are the same, then the difference is zero, and each model is equally likely to be correct. The graph shows that there is a 50% chance that model A is correct, and a 50% chance that model B is correct. If the difference is 2.0, with model A having the lower score, there is a 73% probability that model A is correct, and a 27% chance that model B is correct. Another way to look at this is that model A is 73/27 or 2.7 times more likely to be correct than model B. If the difference between AIC$_c$ scores is 6.0, then model

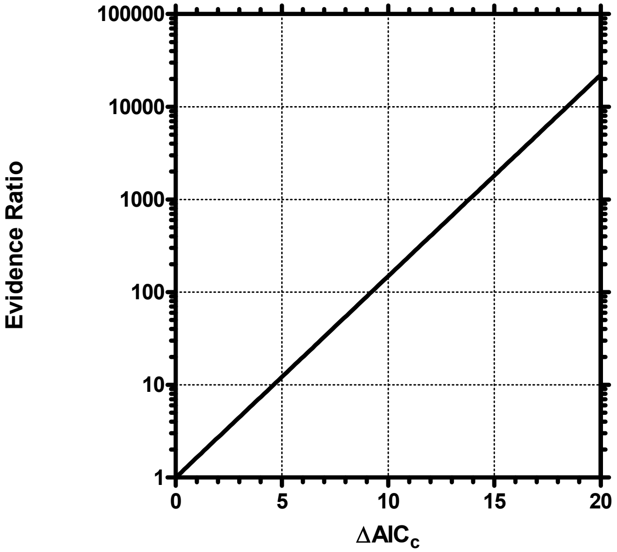
A has a 95% chance of being correct so is 20 (95/5) times more likely to be correct than model B.

Note: Akaike's weights compute the relative probability of two models. It can be extended to compute the relative probabilities of a family of three or more models. But it is always possible that another model is even more likely. The AIC$_c$ method only compares the models you choose, and can't tell you if a different model is more likely still.

The relative likelihood or evidence ratio

When comparing two models, you can divide the probability that one model is correct by the probability the other model is correct to obtain the *evidence ratio*, which is defined by this equation.

$$\text{Evidence Ratio}=\frac{\text{Probability that model 1 is correct}}{\text{Probability that model 2 is correct}}=\frac{1}{e^{-0.5\cdot\Delta\text{AIC}_c}}$$



Note: The evidence ratio is based on the absolute difference between AIC$_c$ scores, not the relative difference. You'll get the same evidence ratio with AIC$_c$ scores of 4567 and 4577 and AIC$_c$ scores of 1 and 11. In both cases, the difference is 10.

For example, if the AIC$_c$ scores differ by 5.0, then the evidence ratio equals 12.18. The model with the lower AIC$_c$ score is a bit more than twelve times more likely to be correct than the other model. Most people don't consider this to be completely persuasive. If the difference in AIC$_c$ scores equals 10, then the evidence ratio is 148, so the evidence is overwhelmingly in favor of the model with the lower AIC$_c$.

Don't overinterpret the evidence ratio. The ratio tells you the relative likelihood, *given your experimental design*, of the two models being correct. If you have few data points, the simpler model might fit best with a very high evidence ratio. This tells you that you can be quite sure that the simple model is adequate to explain your data. With so few points, you don't have any real evidence that the more complicated model is right. This doesn't mean that a more complicated model might not explain your system better. If you had lots of data, you might find that a more complicated model is more likely to be correct.

Terminology to avoid when using AIC_c

The AIC is derived from information theory, which is different from statistical hypothesis testing. You can use the AIC_c method to determine the relative likelihood of two (or more) models. You can't use AIC_c to decide whether the data fit "significantly" better to one model so you should "reject" the other. Therefore you should never use the terms "significant" or "reject" when presenting the conclusions of model comparison by AIC_c. These terms have very definite meanings in statistics that only apply when you compute a P value and use the construct of statistical hypothesis testing. Those terms carry too much baggage, so should only be used in the context of statistical hypothesis testing.

How to compare models with AIC_C by hand

Even if your nonlinear regression program does not compare models with the Akaike's Information Criteria, you can do so fairly simply. These steps review the equations presented in this chapter, so you can compute AIC_c by hand.

1. Fit the first model using nonlinear regression.

2. Look at the results of nonlinear regression and write down the sum-of-squares; call this value SS. If you used any weighting factors, then record the weighted sum-of-squares.

3. Define N to be the number of data points. Be sure to account for replicates properly. If you have 13 X values with duplicate determinations of Y, and you asked the program to treat each replicate as a separate value, then N is 26.

4. Define K to be the number of parameters fit by nonlinear regression plus 1. Don't count parameters that are constrained to constant values. If in doubt, count the number of distinct SE values reported by nonlinear regression, then add 1 to get the value of K. (Why add one? Because nonlinear regression is also "estimating" the value of the sum-of-squares.)

5. Compute AIC_C:

$$AIC_c = N \cdot \ln\left(\frac{SS}{N}\right) + 2K + \frac{2K(K+1)}{N-K-1}$$

6. Repeat steps 1-5 with the other model.

7. The model with the lower AIC_c score is more likely to be correct.

8. Calculate the evidence ratio from the difference in AIC_c scores:

$$\text{Evidence Ratio}=\frac{1}{e^{-0.5\cdot\Delta AIC_c}}$$

One-way ANOVA by AICc

The previous chapter showed how to turn one-way ANOVA into a model comparison problem, and did the comparison by the extra sum-of-squares F test. Here are the same data, comparing models by AIC$_c$.

Model	SS	N	Pars	AIC$_c$	Probability	Evidence ratio
Null hypothesis	40.84	18	1	19.55	37.53%	1.66
Alternative hypothesis	27.23	18	3	18.53	62.47%	

The alternative hypothesis (that the group means are not all identical) has the lower AIC$_c$. Thus, it is more likely to be correct than the null hypothesis that all the data come from populations with the same mean. But the evidence ratio is only 1.66. So the alternative hypothesis is more likely to be correct than the null hypothesis, but only 1.66 times more likely. With so few data and so much scatter, you really don't have enough evidence to decide between the two models.

24. How should you compare models -- AIC$_c$ or F test?

A review of the approaches to comparing models

Chapter 21 discussed the general approach to use when comparing models. Most important:

Before comparing models statistically, use common sense. Only use statistical comparisons to compare two models that make scientific sense, and where each parameter has a sensible best-fit value and a reasonably narrow confidence interval. Many scientists rush to look at statistical comparisons of models too soon. Take the time to ask whether each fit is sensible, and only rely on statistical comparisons when both models fit the data well.

If you wish to compare two models that are not related ("nested"), then your only choice is to compare the fits with the AIC$_c$ method. The F test should not be used to compare nonnested models. Since you'll rarely find it helpful to compare fits of biological models that are not nested, you'll almost always compare nested models.

If the two models are related, and both fit the data with sensible parameter values, you can choose between the F test and AIC method. The rest of this chapter explains the advantages and disadvantages of the two approaches.

Pros and cons of using the F test to compare models

The F test is based on traditional statistical hypothesis testing.

The null hypothesis is that the simpler model (the one with fewer parameters) is correct. The improvement of the more complicated model is quantified as the difference in sum-of-squares. You expect some improvement just by chance, and the amount you expect by chance is determined by the number of degrees of freedom in each model. The F test compares the difference in sum-of-squares with the difference you'd expect by chance. The result is expressed as the F ratio, from which a P value is calculated.

The P value answers this question: If the null hypothesis is really correct, in what fraction of experiments (the size of yours) will the difference in sum-of-squares be as large as you observed or larger? If the P value is small, you'll conclude that the simple model (the null hypothesis) is wrong, and accept the more complicated model. Usually the threshold P value is set at its traditional value of 0.05. If the P value is less than 0.05, then you reject the simpler (null) model and conclude that the more complicated model fits significantly better.

> Reminder: The F test should only be used to compare two related ("nested") models. If the two models are not nested, use the AIC$_c$ method to compare them.

The main advantage of the F test is familiarity. It uses the statistical hypothesis testing paradigm that will be familiar to the people who read your papers or attend your presentations. Many will even be familiar with the use of the F test to compare models.

The F test also has some disadvantages:

- You have to set an arbitrary value of alpha, the threshold P value below which you deem the more complicated model to be "significantly" better so "reject" the simpler model. Traditionally, this value of alpha is set to 0.05, but this is arbitrary.

- The value of alpha is generally set to the same value regardless of sample size. This means that even if you did an experiment with many thousands of data points, there is a 5% chance of rejecting the simpler model even when it is correct. If you think about it, this doesn't make much sense. As you collect more and more data, you should be increasingly sure about which model is correct. It doesn't really make sense to always allow a 5% chance of rejecting the simple model falsely.

- The F test cannot be readily extended to compare three or more models. The problem is one of interpreting multiple P values.

Pros and cons of using AIC$_c$ to compare models

The AIC$_c$ method is based on information theory, and does not use the traditional "hypothesis testing" statistical paradigm. Therefore it does not generate a P value, does not reach conclusions about "statistical significance", and does not "reject" any model.

The AIC$_c$ model determines how well the data supports each model. The model with the lowest AIC$_c$ score is most likely to be correct. The difference in AIC$_c$ score tells you how much more likely. Two models can be compared with a likelihood ratio, which tells you how many times more likely one model is compared to the other.

The main advantage of the AIC approach is that it doesn't tell you just which model is more likely, it tells you how much more likely. This makes it much easier to interpret your results. If one model is hundreds or thousands of times more likely than another, you can make a firm conclusion. If one model is only a few times more likely than the other, you can only make a very tentative conclusion.

Another feature of the AIC approach (which we consider to be an advantage) is that the method doesn't make any decisions for you. It tells you which model is more likely to be correct, and how much more likely. It is up to you to make a decision about whether to conclude that one model is clearly correct, or to conclude that the results are ambiguous and the experiment needs to be repeated. You can simply accept the more likely model. You can accept the simpler model unless the more complicated model is much more likely. Or you can conclude that the data are ambiguous and make no decision about which model is best until you collect more data. This seems like a much more sensible approach to us than the rigid approach of statistical hypothesis testing.

A final advantage of the AIC$_c$ approach is that it is easily extended to compare more than two models. Since you don't set arbitrary values of alpha and declare "statistical significance", you don't get trapped in the logical quandaries that accompany multiple statistical comparisons.

The only real disadvantage of the AIC_c approach is that it is unfamiliar to many scientists. If you use this approach, you may have to explain the method as well as your data.

Which method should you use?

When comparing models, by far the most important step comes before using either the F test or the AIC_c method. You should look at the graph of the fits and at the curve fitting results. You should reject a model if one of the fits has best-fit values that make no scientific sense, if the confidence intervals for the parameters are extremely wide, or if a two-phase (or two component) model fits with one phase having a tiny magnitude. Only when both fits are sensible should you go on to compare the models statistically.

If you compare nonnested models, then the F test is not appropriate. Use the AIC_c approach.

If you compare three or more models, then the whole approach of significance testing gets tricky. Use the AIC_c approach.

In most cases, you'll compare two nested models, so there is no clear advantage of one approach over the other. We prefer the AIC_c method for the reasons listed in the previous section, mostly because AIC_c quantifies the strength of the evidence much better than does a P value. However, this is not a strong preference, and the F test method works well and has the advantage of familiarity.

Tip: Pick a method and stick with it. It is not appropriate to compare with both methods and pick the results that you like the best.