The Corrected AIC ($\text{AIC}_C$)

Although AIC is supposed to result in the selection of a parsimonious model, this will not always be the case. In fact, if given the chance, AIC will prefer a model with $n$ parameters to any other model! Thus, AIC can severely violate the principle of parsimony, in extreme circumstances. The failure of AIC to select an adequately parsimonious model can be a problem whenever the number of parameters in the model under consideration is more than (roughly) 30% of the sample size. Therefore, if we keep $p+q$ small in comparison to $n$, AIC will work well, but in small samples, we might want to consider some models where $p+q$ is a substantial fraction of the sample size, and AIC may perform poorly.

It could be argued that a good model selection criterion should work even if the user tries a "bad" (e.g., over-parametrized) model: If the model is bad, the criterion should be able to detect this. In this regard, AIC fails. In order to remove this deficiency, Hurvich and Tsai (1989)\(^1\) introduced a corrected version, $\text{AIC}_C$, defined by

$$AIC_C(p, q) = -2\text{log}[\text{likelihood}(p, q)] + 2(p + q + 1) \frac{n}{n - p - q - 2}.$$  

In $AIC_C$, we take the penalty term for AIC, which we can consider to be $2(p+q+1)$ (see the note below), and multiply it by the correction factor, $\frac{n}{n - p - q - 2}$. When $p+q$ is small compared to $n$, $\text{AIC}$ and $\text{AIC}_C$ will behave similarly, since the correction factor will be close to 1 in this case. But as $p+q$ gets larger, the penalty term of $\text{AIC}_C$ becomes much stronger than that of $\text{AIC}$. Therefore, in small samples, or whenever the largest value of $p+q$ tried is a substantial fraction of $n$, $\text{AIC}_C$ will tend to select a more parsimonious model than $\text{AIC}$. Otherwise (i.e., in the situations where $\text{AIC}$ works well), the two criteria will typically make the same selection. Since $\text{AIC}_C$ will give a reasonable answer even in extreme circumstances, and since it is not much harder to compute than $\text{AIC}$, we recommend that $\text{AIC}_C$ be used. We have found that $\text{AIC}_C$ retains all of the advantages of $\text{AIC}$, while removing many of the disadvantages.

(Incidentally, Akaike originally said that the penalty term of AIC should be twice the number of parameters in the model. This is $2(p+q+1)$, since we have $p+q$ ARMA parameters, as well as the variance of the innovations, $\epsilon_i$. In the case of $AIC$ it doesn’t matter whether we use $2(p+q)$ or $2(p+q+1)$, since the two penalties differ by a constant. In the case of $AIC_C$, it does matter.)

To compute $AIC_C$ from the MINITAB output, we proceed as follows. Note that MINITAB uses least-squares instead of maximum likelihood estimation. Thus, the likelihood is not available. Instead, we will use $SS$, which is the residual sum of squares (backforecasts excluded), for each model under consideration. Define $N = n - d$ (the effective sample size, after differencing). Then using log to denote the natural logarithm, we have

$$AIC = N \log \left( \frac{SS}{N} \right) + 2(p+q+1) \frac{N}{N-p-q-2} \quad \text{if no constant term in model ,}$$

$$AIC = N \log \left( \frac{SS}{N} \right) + 2(p+q+2) \frac{N}{N-p-q-3} \quad \text{if constant term is included .}$$

It must be clearly understood that $AIC_C$ should not be used for the selection of $d$, since it was not designed for this purpose.

The penalty terms used for $AIC_C$ and $AIC$ may seem arbitrary, but they are not. Here is the philosophy underlying $AIC_C$ (and also $AIC$ in large samples). Suppose that there is some true model which generated our time series. This true model is not ARMA. But we do want to consider using ARMA models to describe our data, since they provide a flexible, estimable, and interpretable class of models. Although the ARMA models have only a few parameters, the true model presumably has a huge number of parameters (perhaps infinitely many). By using the ARMA models to describe our data, we are providing a simple description of what is in fact an extremely complicated situation (i.e., "real life"). It is possible to define a "distance" between each (fitted) candidate model and the truth. Fortunately, even though we will not know in practice what the truth is, it is possible to estimate the distance between our candidate model and the truth. The $AIC_C$ provides an estimate of this distance, which can be calculated even though the true model is not known, and may not even have a finite
number of parameters. The choice of the penalty term in $AIC_C$ guarantees that the distance described above can be estimated (essentially) without any bias. In particular, the bias in $AIC_C$ is much less than the bias in $AIC$ (for estimating how far our candidate model is from the truth) in small samples. In large samples, both criteria are nearly unbiased, and therefore provide good model selections. Finally, we note that it has been shown that, in large samples, $AIC$ and $AIC_C$ give the best possible model selections, assuming that the true model is infinitely complex.