Akaike Without Tears
Ken Burnham and Elliott Sober

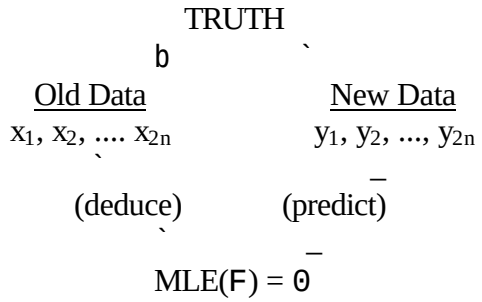
Consider two populations of corn plants -- do they have the same or different mean heights? These two possibilities are expressed in two models:

(SAME) $F_1 = F_2$ (= F)

(DIFF) $F_1 \ldots F_2$

To make the problem simple, we will suppose that height is normally distributed and that both models assign the same known variance (F^2).

We are going to evaluate these models, not by asking which is true, or which is more probably true, but by asking how well each will do in the following prediction problem. We draw n individuals from the first population and measure their heights, thereby obtaining the observations $x_1, x_2, ..., x_n$; we then do the same for the second population, thus obtaining the observations $x_{n+1}, x_{n+2}, ...,x_{2n}$. We then use maximum likelihood estimation (MLE) to estimate the parameters in the models, thus obtaining the two fitted models L(SAME) and L(DIFF). We then draw n new individuals from each of the two populations and measure their heights and thereby obtain the new observations $y_1, y_2, ... , y_{2n}$. We want to know how well each fitted model will do in predicting these new observations. The following diagram depicts this procedure when it is applied to (SAME):



In this simple setting, the MLE of the true mean, based on the 2n old observations, is just the sample mean, 0. A measure of how well SAME does in predicting the new data when it is fitted to old data in this way is given by

$$\Pr(y_1, y_2 ..., y_{2n} \mid F = 0).$$

Taking the log of this probability, we can define the predictive accuracy as follows:

> The predictive accuracy of SAME with respect to these two data sets =
> $\text{Log}[\Pr(y_1, y_2 ..., y_{2n} \mid 0)] = \sum_i \text{Log}[\Pr(y_i \mid 0)]$.

This expression describes how well SAME does with respect to the two specific data sets we have been considering. We now want to define how well SAME will do *on average*, as different

old and new data sets are drawn from the two populations of corn plants. This average performance gets represented as a double expectation:

The expected predictive accuracy of SAME = $E_x E_y \sum_i \text{Log}[\Pr(y_i * \theta)]$.

Although we can simply compute SAME's predictive accuracy with respect to the two specific data sets given, we can't compute SAME's expected predictive accuracy for all the possible data sets that The Truth might generate. The reason is that The Truth is not known to us, and as the above diagram suggests, how well SAME will do in the prediction problem depends on what The Truth is. If the two populations have the same mean height, SAME will do better in predicting new data from old than SAME will do if the two populations have radically different mean heights.

Although we can't calculate the model's expected predictive accuracy, we can try to estimate it. Akaike's theorem, a simple instance of which we'll now derive, shows how to calculate an unbiased estimate of a model's expected predictive accuracy. This estimate can be obtained from a single data set; this is fortunate, since investigators inevitably have just one data set with which to work -- namely, the single data set consisting of all the data they possess.

First of all, we know that

$E_x E_y \sum_i \text{Log}[\Pr(y_i * \theta)] \% E_x E_y (-1/2\sigma^2) \sum_i (y_i - \theta)^2$,

the right-hand side of which expands to

$E_x E_y (-1/2\sigma^2) \sum_i [(y_i - \bar{y} + \bar{y} - \theta)^2]$,

and this to

(*) $E_x E_y (-1/2\sigma^2) \sum_i [(y_i - \bar{y})^2 + 2(y_i - \bar{y})(\bar{y} - \theta) + (\bar{y} - \theta)^2]$.

Here the expected values of θ and $\bar{y}$ are to be understood as the averages obtained by repeatedly making 2n observations.

We need to evaluate the three terms in (*). With respect to the first, we know that

$\sum_i \text{Log}[\Pr(y_i * \bar{y})]$ is an unbiased estimate of $E_y (-1/2\sigma^2) \sum_i (y_i - \bar{y})^2$.

Given the relationship of the old and new data sets, it follows that

(1) $\sum_i \text{Log}[\Pr(x_i * \theta)]$ is an unbiased estimate of $E_y (-1/2\sigma^2) \sum_i (y_i - \bar{y})^2$.

In connection with the second addend in (*), we know that

(2) $E_x E_y (-1/2\sigma^2) \sum_i [2(y_i - \bar{y})(\bar{y} - \theta)] = 0$,

since

$$\Sigma_i (y_i - \bar{y}) = 0.$$

Finally, we need to evaluate $E_x E_y (-1/2\sigma^2) 2n(\bar{y} - \theta)^2$, which is the third addend in (*). This expands to $E_x E_y (-1/2\sigma^2) 2n(\theta^2 - 2\theta\bar{y} + \bar{y}^2)$. At this point, we need the following facts:

$$[E(\theta)]^2 = [E(\bar{y})]^2 = \sigma^2.$$

The variance of the mean (θ or $\bar{y}$) $= \sigma^2/2n = E(\theta)^2 - [E(\theta)]^2$. Hence $E(\theta^2) = \sigma^2/2n + \sigma^2$.

It follows that the third term in (*) is equal to

(3) $\quad (-2n/2\sigma^2)(\sigma^2/2n + \sigma^2 - 2\sigma^2 + \sigma^2/2n + \sigma^2) = -1.$

Combining (1), (2), and (3), we conclude that an unbiased estimate of $E_x E_y \Sigma_i \text{Log}[\Pr(y_i * \theta)]$ is given by $\Sigma_i \text{Log}[\Pr(x_i * \theta)] - 1$. In other words,

(**) An unbiased estimate of the expected predictive accuracy of SAME =
log[Pr(Data * L(SAME)] - 1.

What is the estimate for DIFF? Here we can exploit the fact that DIFF is, in effect, just the conjunction of two models like SAME, one for each population. By carrying through the line of reasoning just followed for SAME, we can derive the result that an unbiased estimate of the expected predictive accuracy of DIFF is given by

$$\log \Sigma_i \Pr(x_1, x_2, \ldots x_n * \theta_f)] - 1 + \log \Sigma_i \Pr(x_{n+1}, x_{n+2}, \ldots x_{2n} * \theta_s)] - 1,$$

where θ_f is the MLE of the first population's mean and θ_s is the MLE of the second population's mean, both based on the old data. In other words,

(***) An unbiased estimate of the expected predictive accuracy of DIFF =
log[Pr(Data * L(DIFF)] - 2.

Notice that when we use (**) and (***) to compare estimates of the expected predictive accuracies of SAME and DIFF, we are attending to the log-likelihoods of the two fitted models, but also to the number of adjustable parameters that the two models contain.

Comment #1: In addition to the assumption that height is normally distributed and that the variance is known, there is an additional special feature of this derivation that is worth noticing. When it came to estimating the expected predictive accuracy of DIFF, we did this by using the estimate already obtained for SAME "twice." DIFF is, so speak, just two SAME models, one for each population. However, competing models often fail to be related in this way. Consider, for example, LIN, which says that $y = a + bx + \text{Error}$ and PAR, which says that $y = a + bx + cx^2 + \text{Error}$. A derivation of Akaike's theorem for this pair of models will have to

proceed differently. PAR can't be conceptualized as two applications of LIN, one for each part of the data.

Comment #2: When we described how SAME's predictive accuracy is to be measured, we used the expression $\Pr(y_1, y_2 \ldots, y_{2n} \mid F = 0)$. In the context of the normal model assumed here, this is proportional to $(-1/2F^2) \sum_i (y_i - 0)^2$; thus, likelihood in this case bears a special connection to the sum of *squared* distances. However, there are other situations in which expected log-likelihood is proportional to the absolute value of the distance (a measure proposed by Kolmogorov). In this case, a different mathematical derivation is needed if one wants an unbiased estimate of a model's expected predictive accuracy. Perhaps it is not unreasonable to conjecture that parsimony (as measured by number of adjustable parameters) is relevant to estimating expected predictive accuracy in this setting as well.

Comment #3: Akaike's theorem provides a method for obtaining unbiased estimates of a model's expected predictive accuracy. However, it is not plausible to claim that an estimator's being unbiased is sufficient to justify us in using it. This is obvious when you consider the fact that we can construct infinitely many unbiased estimators of a quantity, once we have identified one of them. So what kind of argument can be produced that begins with Akaike's theorem and ends with an endorsement of AIC? It is important to understand this problem comparatively – the goal is to compare AIC with other estimators that have been described, not to show that AIC is the best of all possible estimators (including ones that no one has thought of as yet). What goes for scientific theories also goes for scientific methods – we can and should evaluate them comparatively, not absolutely.