171:290 Model Selection
Lecture II: The Akaike Information Criterion

Joseph E. Cavanaugh

Department of Biostatistics
Department of Statistics and Actuarial Science
The University of Iowa

August 28, 2012

Introduction

- AIC, the *Akaike Information Criterion*, is generally regarded as the first model selection criterion.
- Today, AIC continues to be the most widely known and used model selection tool among practitioners.
- AIC was introduced by Hirotugu Akaike in his seminal 1973 paper "Information Theory and an Extension of the Maximum Likelihood Principle" (in: B. N. Petrov and F. Csaki, eds., *2nd International Symposium on Information Theory*, Akademia Kiado, Budapest, pp. 267–281).

Introduction

- The traditional maximum likelihood paradigm, as applied to statistical modeling, provides a mechanism for estimating the unknown parameters of a model having a specified dimension and structure.
- Akaike extended this paradigm by considering a framework in which the model dimension is also unknown, and must therefore be determined from the data.
- Thus, Akaike proposed a framework wherein both model estimation and selection could be simultaneously accomplished.

Introduction

- For a parametric candidate model of interest, the likelihood function reflects the conformity of the model to the observed data.
- As the complexity of the model is increased, the model becomes more capable of adapting to the characteristics of the data.
- Thus, selecting the fitted model that maximizes the empirical likelihood will invariably lead one to choose the most complex model in the candidate collection.
- Model selection based on the likelihood principle, therefore, requires an extension of the traditional likelihood paradigm.

Introduction

Outline:

- Model Selection Framework
- Kullback-Leibler Information
- Derivation of AIC
- Properties and Limitations of AIC
- Use of AIC
- Application

Model Selection Framework

- Suppose a collection of data y has been generated according to an unknown model or density $g(y)$.
- We endeavor to find a fitted parametric model which provides a suitable approximation to $g(y)$.
- Let $\mathcal{F}(k) = \{f(y|\,\theta_k) \mid \theta_k \in \Theta(k)\}$ denote a k–dimensional parametric class, i.e., a class of densities in which the parameter space $\Theta(k)$ consists of k–dimensional vectors whose components are functionally independent.
- Let $\hat{\theta}_k$ denote a vector of estimates obtained by maximizing the likelihood function $f(y|\,\theta_k)$ over $\Theta(k)$.
- Let $f(y|\,\hat{\theta}_k)$ denote the corresponding fitted model.

Model Selection Framework

- Suppose our goal is to search among a collection of classes $\mathcal{F} = \{\mathcal{F}(k_1), \mathcal{F}(k_2), \ldots, \mathcal{F}(k_L)\}$ for the fitted model $f(y|\,\hat{\theta}_k)$, $k \in \{k_1, k_2, \ldots, k_L\}$, which serves as the "best" approximation to $g(y)$.
- *Note:* AIC may be used to delineate between different fitted models having the same dimension. (For instance, fitted regression models based on design matrices having the same size yet different column spaces.)
- For simplicity, our notation and framework assume that each candidate model class $\mathcal{F}(k)$ and the corresponding fitted model $f(y|\,\hat{\theta}_k)$ are distinguished by the dimension k.
- Our model selection problem can therefore be viewed as a problem of dimension determination.

Model Selection Framework

- **True** or **generating model**: $g(y)$.
- **Candidate** or **approximating model**: $f(y|\,\theta_k)$.
- **Fitted model:** $f(y|\,\hat{\theta}_k)$.
- **Candidate family:** $\mathcal{F}$.

Model Selection Framework

- To determine which of the fitted models $\{f(y|\hat{\theta}_{k_1}), f(y|\hat{\theta}_{k_2}), \ldots, f(y|\hat{\theta}_{k_L})\}$ best approximates $g(y)$, we require a measure which provides a suitable reflection of the disparity between the true model $g(y)$ and an approximating model $f(y|\theta_k)$.
- The *Kullback-Leibler information* is one such measure.

Kullback-Leibler Information

- For two arbitrary densities $g(y)$ and $f(y)$ with the same support, the **Kullback-Leibler information** or **Kullback's directed divergence** between $g(y)$ and $f(y)$ with respect to $g(y)$ is defined as

$$I(g, f) = \mathrm{E}\left\{\ln \frac{g(y)}{f(y)}\right\},$$

  where E denotes the expectation under $g(y)$.
- $I(g, f) \geq 0$ with equality if and only if g and f are the same density.
- $I(g, f)$ reflects the separation between g and f.
- $I(g, f)$ is not a formal distance measure.

Kullback-Leibler Information

- For our purposes, we will consider the Kullback-Leibler information between the true model $g(y)$ and the approximating model $f(y|\,\theta_k)$ with respect to $g(y)$, which we will denote as $I(\theta_k)$:

$$I(\theta_k) = \mathrm{E}\left\{\ln \frac{g(y)}{f(y|\,\theta_k)}\right\}.$$

- In the preceding expression and henceforth, E denotes the expectation under the true density or model.

Kullback-Leibler Information

- Let
$$d(\theta_k) = \mathrm{E}\{-2 \ln f(y|\,\theta_k)\}.$$
- $d(\theta_k)$ is often called the **Kullback discrepancy**.
- Note that we can write
$$2I(\theta_k) = d(\theta_k) - \mathrm{E}\{-2 \ln g(y)\}.$$
- Since $\mathrm{E}\{-2 \ln g(y)\}$ does not depend on θ_k, any ranking of a set of candidate models corresponding to values of $I(\theta_k)$ would be identical to a ranking corresponding to values of $d(\theta_k)$.
- Hence, for the purpose of discriminating among various candidate models, $d(\theta_k)$ serves as a valid substitute for $I(\theta_k)$.

Derivation of AIC

- The measure

$$d(\hat{\theta}_k) = \mathrm{E}\{-2 \ln f(y|\, \theta_k)\}|_{\theta_k = \hat{\theta}_k}$$

reflects the separation between the generating model $g(y)$ and a fitted model $f(y|\, \hat{\theta}_k)$.

- Evaluating $d(\hat{\theta}_k)$ is not possible, since doing so requires knowledge of $g(y)$.
- The work of Akaike (1973), however, suggests that $-2 \ln f(y|\, \hat{\theta}_k)$ serves as a biased estimator of $d(\hat{\theta}_k)$, and that the bias adjustment

$$\mathrm{E}\{d(\hat{\theta}_k)\} - \mathrm{E}\{-2 \ln f(y|\, \hat{\theta}_k)\}$$

can often be asymptotically estimated by twice the dimension of θ_k.

Derivation of AIC

- Since k denotes the dimension of θ_k, under appropriate conditions, the expected value of

$$\text{AIC} = -2 \ln f(y|\hat{\theta}_k) + 2k$$

should asymptotically approach the expected value of $d(\hat{\theta}_k)$, say

$$\Delta(k) = \text{E}\{d(\hat{\theta}_k)\}.$$

- Specifically, one can establish that

$$\text{E}\{\text{AIC}\} + o(1) = \Delta(k).$$

- AIC therefore provides an asymptotically unbiased estimator of $\Delta(k)$.

Derivation of AIC

- $\Delta(k)$ is often called the **expected Kullback discrepancy**.
- $\Delta(k)$ reflects the average separation between the generating model $g(y)$ and fitted models having the same structure as $f(y|\hat{\theta}_k)$.

Derivation of AIC

- The asymptotic unbiasedness property of AIC requires the assumption that $g(y) \in \mathcal{F}(k)$.
- This assumption implies that the true model or density is a member of the parametric class $\mathcal{F}(k)$, and can therefore be written as $f(y|\,\theta_o)$, where $\theta_o \in \Theta(k)$.
- From a practical perspective, the assumption that $f(y|\,\theta_o) \in \mathcal{F}(k)$ implies that the fitted model $f(y|\,\hat{\theta}_k)$ is either *correctly specified* or *overfitted.*
- Is this a problematic assumption?

Derivation of AIC

- To justify the asymptotic unbiasedness of AIC, consider writing $\Delta(k)$ as follows:

$$\begin{aligned}
&\Delta(k) \\
&= \mathrm{E}\{d(\hat{\theta}_k)\} \\
&= \mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\} \\
&\quad + \Big[\mathrm{E}\{-2\ln f(y|\,\theta_o)\} - \mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\}\Big] \qquad (1) \\
&\quad + \Big[\mathrm{E}\{d(\hat{\theta}_k)\} - \mathrm{E}\{-2\ln f(y|\,\theta_o)\}\Big]. \qquad (2)
\end{aligned}$$

- The following lemma asserts that (1) and (2) are both within $o(1)$ of k.

Derivation of AIC

- We assume the necessary regularity conditions required to ensure the consistency and asymptotic normality of the maximum likelihood vector $\hat{\theta}_k$.

Lemma

$$E\{-2\ln f(y|\theta_o)\} - E\{-2\ln f(y|\hat{\theta}_k)\} = k + o(1), \quad (1)$$

$$E\{d(\hat{\theta}_k)\} - E\{-2\ln f(y|\theta_o)\} = k + o(1). \quad (2)$$

Proof of Lemma

Proof:

- Define

$$\mathsf{I}(\theta_k) = \mathsf{E}\left[-\frac{\partial^2 \ln f(y|\,\theta_k)}{\partial\theta_k \partial\theta_k'}\right] \text{ and } \mathcal{I}(\theta_k, y) = \left[-\frac{\partial^2 \ln f(y|\,\theta_k)}{\partial\theta_k \partial\theta_k'}\right].$$

- $\mathsf{I}(\theta_k)$ is the **expected Fisher information matrix**.
- $\mathcal{I}(\theta_k, y)$ is the **observed Fisher information matrix**.

Proof of Lemma

- First, consider taking a second-order expansion of $-2\ln f(y|\,\theta_o)$ about $\hat{\theta}_k$, and evaluating the expectation of the result.
- We obtain

$$\begin{aligned}\mathrm{E}\{-2\ln f(y|\,\theta_o)\} &= \mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\} \\ &\quad +\mathrm{E}\left\{(\hat{\theta}_k-\theta_o)'\{\mathcal{I}(\hat{\theta}_k,y)\}(\hat{\theta}_k-\theta_o)\right\} \\ &\quad +o(1).\end{aligned}$$

- Thus,

$$\begin{aligned}&\mathrm{E}\{-2\ln f(y|\,\theta_o)\}-\mathrm{E}\{-2\ln f(y|\,\hat{\theta}_k)\} \\ &=\mathrm{E}\left\{(\hat{\theta}_k-\theta_o)'\{\mathcal{I}(\hat{\theta}_k,y)\}(\hat{\theta}_k-\theta_o)\right\}+o(1). \quad (3)\end{aligned}$$

Proof of Lemma

- Next, consider taking a second-order expansion of $d(\hat{\theta}_k)$ about θ_o, again evaluating the expectation of the result.
- We obtain

$$\begin{aligned} \mathrm{E}\{d(\hat{\theta}_k)\} &= \mathrm{E}\{-2\ln f(y|\,\theta_o)\} \\ &\quad +\mathrm{E}\left\{(\hat{\theta}_k-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}_k-\theta_o)\right\} \\ &\quad +o(1). \end{aligned}$$

- Thus,

$$\begin{aligned} &\mathrm{E}\{d(\hat{\theta}_k)\}-\mathrm{E}\{-2\ln f(y|\,\theta_o)\} \\ &=\mathrm{E}\left\{(\hat{\theta}_k-\theta_o)'\{\mathrm{I}(\theta_o)\}(\hat{\theta}_k-\theta_o)\right\}+o(1). \end{aligned} \tag{4}$$

Proof of Lemma

- The quadratic forms

  $$(\hat{\theta}_k - \theta_o)' \{\mathcal{I}(\hat{\theta}_k, y)\}(\hat{\theta}_k - \theta_o) \text{ and } (\hat{\theta}_k - \theta_o)' \{\mathsf{I}(\theta_o)\}(\hat{\theta}_k - \theta_o)$$

  both converge to centrally distributed chi-square random variables with k degrees of freedom.
- Recall again that we are assuming $\theta_o \in \Theta(k)$.
- Thus, the expectations of both quadratic forms are within $o(1)$ of k.
- This fact along with (3) and (4) establishes (1) and (2). $\square$

Bias Correction

- AIC provides us with an approximately unbiased estimator of $\Delta(k)$ in settings where n is large and k is comparatively small.
- In settings where n is small and k is comparatively large (e.g., $k \approx n/2$), $2k$ is often much smaller than the bias adjustment, making AIC substantially negatively biased as an estimator of $\Delta(k)$.
- In small-sample applications, better estimators of the bias adjustment are available than $2k$. (These estimators will be discussed in future lectures.)

Bias Correction

- If AIC severely underestimates $\Delta(k)$ for higher dimensional fitted models in the candidate set, the criterion may favor the higher dimensional models even when the expected discrepancy between these models and the generating model is rather large.
- Examples illustrating this phenomenon appear in Linhart and Zucchini (1986, pages 86–88), who comment (page 78) that "in some cases the criterion simply continues to decrease as the number of parameters in the approximating model is increased."

Asymptotic Efficiency

- AIC is *asymptotically efficient* in the sense of Shibata (1980, 1981), yet it is not *consistent*.
- Suppose that the generating model is of a finite dimension, and that this model is represented in the candidate collection under consideration. A **consistent** criterion will asymptotically select the fitted candidate model having the correct structure with probability one.
- On the other hand, suppose that the generating model is of an infinite dimension, and therefore lies outside of the candidate collection under consideration. An **asymptotically efficient** criterion will asymptotically select the fitted candidate model which minimizes the mean squared error of prediction.

Use of AIC

- AIC is applicable in a broad array of modeling frameworks, since its justification only requires conventional large-sample properties of maximum likelihood estimators.
- The application of the criterion does not require the assumption that one of the candidate models is the "true" or "correct" model, although the derivation implies otherwise.

Use of AIC

- AIC can be used to compare non-nested models.
- Burnham and Anderson (2002, p. 88): "A substantial advantage in using information-theoretic criteria is that they are valid for nonnested models. Of course, traditional likelihood ratio tests are defined only for nested models, and this represents another substantial limitation in the use of hypothesis testing in model selection."

Use of AIC

- AIC can be used to compare models based on different probability distributions.
- However, when the criterion values are computed, no constants should be discarded from the goodness-of-fit term $-2 \ln f(y|\hat{\theta}_k)$ (such as $n \ln 2\pi$).
- Keep in mind that certain statistical software packages routinely discard constants in the evaluation of likelihood-based selection criteria: e.g., in normal linear regression, $n \ln \hat{\sigma}^2$ is often used as the goodness-of-fit term as opposed to $n \ln \hat{\sigma}^2 + n(\ln 2\pi + 1)$.

Use of AIC

- In a model selection application, the optimal fitted model is identified by the minimum value of AIC.
- However, the criterion values are important; models with similar values should receive the same "ranking" in assessing criterion preferences.

Use of AIC

- Question: What constitutes a substantial difference in criterion values?
- For AIC, Burnham and Anderson (2002, p. 70) feature the following table.

$AIC_i - AIC_{min}$	Level of Empirical Support for Model i
0 - 2	Substantial
4 - 7	Considerably Less
> 10	Essentially None

Application

- The Iowa Fluoride Study is a longitudinal study based on several hundred children throughout the state of Iowa.
- Brofitt, Levy, Warren, and Cavanaugh (2007) analyzed data from the Iowa Fluoride Study to investigate the association between bottled water use and caries prevalence.
- Since fluoride is typically not added to bottled water, children who drink primarily bottled water may be more prone to caries than those who drink primarily fluoridated tap water.

Application

- Our modeling analyses were based on 413 children.
- The subjects were classified as "bottled water users" or "non bottled water users" based on data collected from questionnaires completed by the parents every six months.

Application

- The response variable was defined as a count of the number of affected tooth surfaces among the permanent incisors and first molars at the time of the mixed dentition exam.
- The covariate of primary interest was the dichotomous bottled water usage variable (BW).
- Other covariates considered were the age of the child at the time of the examination and toothbrushing frequency (coded as a five-level score).
- Generalized linear models (GLMs) were employed.
  - The log link was used.
  - For the random component, two probability distributions were considered: Poisson and negative binomial.

Application

- Model selections were made using AIC.
- We will consider 4 models:
  - Poisson GLM with BW,
  - Poisson GLM without BW,
  - Negative binomial GLM with BW,
  - Negative binomial GLM without BW.

Application

AIC Results

Distribution	With BW	Without BW
Poisson	753.0	751.0
Neg. Binomial	650.2	648.3

References

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: B. N. Petrov and F. Csáki, eds., *2nd International Symposium on Information Theory*, Akadémia Kiadó, Budapest, pp. 267–281.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **AC-19**, 716–723.
- Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45–54.

H. Akaike

Hirotugu Akaike (1927–2009)