

# 171:290 Model Selection

## Lecture IV: Corrected AIC and Modified AIC, AICc and MAIC

Joseph E. Cavanaugh

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

September 15, 2009

## Introduction

- The Akaike information criterion, AIC, is derived as an estimator of the expected Kullback discrepancy between the true model and a fitted candidate model.
- AIC is applicable in a broad array of modeling frameworks, since its justification only requires conventional large-sample properties of maximum likelihood estimators.
- The asymptotic justification of the criterion requires two strong assumptions:
  - (i) that the true model is contained in the candidate class under consideration,
  - (ii) that the vector of maximum likelihood estimators satisfies the conventional large-sample properties of MLE's.

## Introduction

- Can these assumptions be relaxed?
- The corrected Akaike information criterion, AICc, is based on a development where the large-sample requirement in (ii) is relaxed.
- The Takeuchi (1976) information criterion, TIC, is based on a development where the true model assumption (i) is relaxed.
- The modified Akaike information criterion, MAIC, is based on a development where the both the true model assumption (i) and the large-sample requirement in (ii) are relaxed.

# Introduction

- In Lecture III, we introduced AICc.
- In Lecture IV, we introduce MAIC.
- In Lecture V, we introduce TIC.

# Introduction

## Outline:

- Review of AIC and AICc (Lectures II and III)
- Proof of AICc Lemma
- The Modified Akaike Information Criterion, MAIC
- Discussion
- Application

## Review of AIC and AICc

Key Constructs:

- **True or generating model:**  $g(y)$ .
- **Candidate or approximating model:**  $f(y|\theta_k)$ .
- **Candidate class:**

$$\mathcal{F}(k) = \{f(y|\theta_k) \mid \theta_k \in \Theta(k)\}.$$

- **Fitted model:**  $f(y|\hat{\theta}_k)$ .

## Review of AIC and AICc

- **Kullback discrepancy** between  $g(y)$  and  $f(y|\hat{\theta}_k)$  with respect to  $g(y)$ :

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

- **Expected Kullback discrepancy:**

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

## Review of AIC and AICc

- We impose 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  $\theta_o \in \Theta(k)$ .

## Review of AIC and AICc

Consider writing  $\Delta(k)$  as follows:

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

$$+ \left[ E\{d(\hat{\theta}_k)\} - E\{-2 \ln f(y|\theta_o)\} \right]. \quad (2)$$

## Review of AIC and AICc

- The derivation of AIC is based on the following lemma.

## Lemma

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

- Conditions under which the lemma holds:
  - $g(y) \in \mathcal{F}(k)$ ,
  - the maximum likelihood vector  $\hat{\theta}_k$  satisfies the conventional large-sample properties of MLE's.

## Review of AIC and AICc

- The derivation of AICc is based on the following lemma.
- We assume that the candidate class  $\mathcal{F}(k)$  consists of normal linear regression models, and that  $g(y) \in \mathcal{F}(k)$ .
- As before, we write  $g(y)$  as  $f(y|\theta_o)$ .
- We use  $p$  to denote the rank of the design matrix, meaning  $k = (p + 1)$ .

## Review of AIC and AICc

## Lemma

$$\begin{aligned} & E\{-2 \ln f(y|\theta_o)\} - E\{-2 \ln f(y|\hat{\theta}_k)\} \\ &= n \ln(n/2) - n\psi\left(\frac{n-p}{2}\right), \\ & E\{d(\hat{\theta}_k)\} - E\{-2 \ln f(y|\theta_o)\} \\ &= -n \ln(n/2) + n\psi\left(\frac{n-p}{2}\right) + \frac{2(p+1)n}{n-p-2}. \end{aligned}$$

## Proof of AICc Lemma

- Suppose that the generating model for the data is given by

$$y = X\beta_o + e, \quad e \sim N_n(0, \sigma_o^2 I),$$

and that the candidate model postulated for the data is of the form

$$y = X\beta + e, \quad e \sim N_n(0, \sigma^2 I).$$

- Here,  $y$  is an  $n \times 1$  observation vector,  $e$  is an  $n \times 1$  error vector,  $\beta_o$  and  $\beta$  are  $p \times 1$  parameter vectors, and  $X$  is an  $n \times p$  design matrix of full-column rank.

## Proof of AICc Lemma

- Let  $\theta_o$  and  $\theta_k$  respectively denote the  $k = (p + 1)$  dimensional vectors  $(\beta'_o, \sigma_o^2)'$  and  $(\beta', \sigma^2)'$ .
- Assume  $\beta_o$  is such that for some  $0 < p_o \leq p$ , the last  $(p - p_o)$  components of  $\beta_o$  are zero.
- Thus, the true model is nested within the candidate model.
- Note that the nesting ensures that  $f(y|\theta_o) \in \mathcal{F}(k)$ .
- Let  $\hat{\beta}$  denote the least-squares estimator of  $\beta$ , and let  $\hat{\sigma}^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/n$ .
- Let  $\hat{\theta}_k = (\hat{\beta}', \hat{\sigma}^2)'$  denote the MLE of  $\theta_k$ .

## Proof of AICc Lemma

## Preliminary Results:

- Let  $\chi^2$  be a random variable having a central chi-square distribution with  $d$  degrees of freedom.

- $$E \left\{ \frac{1}{\chi^2} \right\} = \frac{1}{d-2}.$$

- $$E \{ \ln \chi^2 \} = \ln 2 + \psi \left( \frac{d}{2} \right),$$

where  $\psi$  is the *digamma* or *psi* function.

## Proof of AICc Lemma

## Proof:

- The log-likelihood for the candidate model is given by

$$\ln f(y|\theta_k) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (y - X\beta)' (y - X\beta).$$

- The following relations can be established:

$$E\{-2 \ln f(y|\theta_o)\} = n \ln \sigma_o^2 + n(1 + \ln 2\pi),$$

$$E\{-2 \ln f(y|\hat{\theta}_k)\} = E\{n \ln \hat{\sigma}^2\} + n(1 + \ln 2\pi),$$

$$\begin{aligned} d(\hat{\theta}_k) &= n \ln \hat{\sigma}^2 + \frac{n\sigma_o^2}{\hat{\sigma}^2} \\ &\quad + \frac{1}{\hat{\sigma}^2} (\hat{\beta} - \beta_o)' (X'X) (\hat{\beta} - \beta_o) + n \ln 2\pi. \end{aligned}$$

## Proof of AICc Lemma

To evaluate the expected value of  $d(\hat{\theta}_k)$ , note the following:

- $(n\hat{\sigma}^2/\sigma_o^2)$  has a chi-square distribution with  $(n - p)$  degrees of freedom,
- the quadratic form  $\{(\hat{\beta} - \beta_o)' \{(1/\sigma_o^2)(X'X)\}(\hat{\beta} - \beta_o)\}$  has a chi-square distribution with  $p$  degrees of freedom,
- $\hat{\sigma}^2$  and  $\hat{\beta}$  are independent.

$$\begin{aligned}
& E\{d(\hat{\theta}_k)\} \\
&= E\{n \ln \hat{\sigma}^2\} + E\left\{\frac{n\sigma_o^2}{\hat{\sigma}^2}\right\} \\
&\quad + E\left\{\frac{1}{\hat{\sigma}^2}\right\} E\left\{(\hat{\beta} - \beta_o)'(X'X)(\hat{\beta} - \beta_o)\right\} + n \ln 2\pi \\
&= E\{n \ln \hat{\sigma}^2\} + n \ln 2\pi + n^2 E\left\{\frac{\sigma_o^2}{n\hat{\sigma}^2}\right\} \\
&\quad + n E\left\{\frac{\sigma_o^2}{n\hat{\sigma}^2}\right\} E\left\{(\hat{\beta} - \beta_o)' \{(1/\sigma_o^2)(X'X)\}(\hat{\beta} - \beta_o)\right\} \\
&= [E\{n \ln \hat{\sigma}^2\} + n(1 + \ln 2\pi)] - n + n^2 \{1/(n - p - 2)\} \\
&\quad + n \{1/(n - p - 2)\} (p) \\
&= [E\{n \ln \hat{\sigma}^2\} + n(1 + \ln 2\pi)] + \frac{2n(p + 1)}{n - p - 2}.
\end{aligned}$$

We now simplify the first and second terms of the bias adjustment.

$$\begin{aligned}
 & E\{-2 \ln f(y|\theta_o)\} - E\{-2 \ln f(y|\hat{\theta}_k)\} \\
 &= \{n \ln \sigma_o^2 + n(1 + \ln 2\pi)\} - \{E\{n \ln \hat{\sigma}^2\} + n(1 + \ln 2\pi)\} \\
 &= E\left\{n \ln \left(\frac{\sigma_o^2}{\hat{\sigma}^2}\right)\right\} \\
 &= E\left\{n \ln \left\{n \left(\frac{\sigma_o^2}{n\hat{\sigma}^2}\right)\right\}\right\} \\
 &= n \ln n - nE\left\{\ln \left(\frac{n\hat{\sigma}^2}{\sigma_o^2}\right)\right\} \\
 &= n \ln n - n\left\{\ln 2 + \psi\left(\frac{n-p}{2}\right)\right\} \\
 &= n \ln(n/2) - n\psi\left(\frac{n-p}{2}\right).
 \end{aligned}$$

$$\begin{aligned}
& E\{d(\hat{\theta}_k)\} - E\{-2 \ln f(y|\theta_o)\} \\
&= \left\{ [E\{n \ln \hat{\sigma}^2\} + n(1 + \ln 2\pi)] + \frac{2n(p+1)}{n-p-2} \right\} \\
&\quad - \{n \ln \sigma_o^2 + n(1 + \ln 2\pi)\} \\
&= E \left\{ n \ln \left( \frac{\hat{\sigma}^2}{\sigma_o^2} \right) \right\} + \frac{2n(p+1)}{n-p-2} \\
&= E \left\{ n \ln \left\{ \frac{1}{n} \left( \frac{n\hat{\sigma}^2}{\sigma_o^2} \right) \right\} \right\} + \frac{2n(p+1)}{n-p-2} \\
&= -n \ln n + n E \left\{ \ln \left( \frac{n\hat{\sigma}^2}{\sigma_o^2} \right) \right\} + \frac{2n(p+1)}{n-p-2} \\
&= -n \ln n + n \left\{ \ln 2 + \psi \left( \frac{n-p}{2} \right) \right\} + \frac{2n(p+1)}{n-p-2} \\
&= -n \ln(n/2) + n\psi \left( \frac{n-p}{2} \right) + \frac{2n(p+1)}{n-p-2}. \quad \square
\end{aligned}$$

## Proof of AICc Lemma

- AICc is obtained by adding the bias adjustment terms derived in the previous lemma to the baseline estimator of  $\Delta(k)$ ,  $-2 \ln f(y|\hat{\theta}_k)$ .
- We have:

$$\begin{aligned} \text{AICc} &= -2 \ln f(y|\hat{\theta}_k) \\ &\quad + \left\{ n \ln(n/2) - n\psi\left(\frac{n-p}{2}\right) \right\} \\ &\quad + \left\{ -n \ln(n/2) + n\psi\left(\frac{n-p}{2}\right) + \frac{2(p+1)n}{n-p-2} \right\} \\ &= -2 \ln f(y|\hat{\theta}_k) + \frac{2(p+1)n}{n-p-2}. \end{aligned}$$

## The Modified Akaike Information Criterion, MAIC

- The asymptotic unbiasedness of AIC and exact unbiasedness of AICc (in the normal linear regression framework) require the assumption that  $g(y) \in \mathcal{F}(k)$ .
- This assumption implies that the candidate model of interest,  $f(y|\theta_k)$ , is either correctly specified or overspecified.
- The modified Akaike information criterion, MAIC, is one of several AIC variants based on a development that relaxes this assumption.
- MAIC was introduced for the framework of normal multivariate linear regression models by Fujikoshi and Satoh (1997).
- We will introduce MAIC in the framework of normal univariate linear regression models.

## The Modified Akaike Information Criterion, MAIC

- Let  $\mathcal{F} = \{\mathcal{F}(k_1), \mathcal{F}(k_2), \dots, \mathcal{F}(k_L)\}$  represent the candidate family.
- Assume that the largest candidate class in  $\mathcal{F}$  is  $\mathcal{F}(K)$  (i.e.,  $K = \max\{k_1, k_2, \dots, k_L\}$ ).
- MAIC is based on the assumption that  $g(y) \in \mathcal{F}(K)$ .
- Under this assumption, the largest candidate model,  $f(y|\theta_K)$ , is either correctly specified or overspecified.
- The candidate model of interest,  $f(y|\theta_k)$ , may be correctly specified, underspecified, or overspecified.

## The Modified Akaike Information Criterion, MAIC

- In the regression framework, the model  $f(y|\theta_K)$  would typically represent all covariates under consideration.
- Let  $P$  denote the rank of the design matrix for this model.
- Let  $\sigma_*^2$  denote the error variance for this model.
- Let  $\hat{\sigma}_*^2$  denote the maximum likelihood estimator of  $\sigma_*^2$ .
- Define MAIC as

$$\begin{aligned} \text{MAIC} = & -2 \ln f(y|\hat{\theta}_k) + \frac{2n(p+1)}{(n-p-2)} \\ & + \left[ 2p \left\{ \frac{(n-p)\hat{\sigma}_*^2}{(n-P)\hat{\sigma}^2} - 1 \right\} - 2 \left\{ \frac{(n-p)\hat{\sigma}_*^2}{(n-P)\hat{\sigma}^2} - 1 \right\}^2 \right]. \end{aligned}$$

## Discussion

- MAIC can be written as

$$\text{MAIC} = \text{AICc} + \left[ 2p \left\{ \frac{(n-p)\hat{\sigma}_*^2}{(n-P)\hat{\sigma}^2} - 1 \right\} - 2 \left\{ \frac{(n-p)\hat{\sigma}_*^2}{(n-P)\hat{\sigma}^2} - 1 \right\}^2 \right].$$

- The additional data-dependent penalization added to AICc is designed to be approximately zero for correctly specified and overfitted models.
- For underfitted models, this penalization is designed to reduce the bias of AICc.

## Discussion

Bias Properties of AIC, AICc, MAIC  
(Normal Linear Regression Models)

Fitted Model	AIC	AICc	MAIC
Underfitted	$O(1)$	$O(1)$	$O(1/n)$
Correctly Specified or Overfitted	$O(1/n)$	0	$O(1/n^2)$

## Discussion

- Does the additional stochastic penalization added to AICc to produce MAIC yield a practical improvement?
- In simulation results featured in Fujikoshi and Satoh (1997), MAIC marginally outperforms AICc in terms of overfitted selections, yet performs the same as AICc in terms of underfitted selections.
- In the next lecture, we will examine the performance of AIC, AICc, and MAIC in a simulation study.
- We will also introduce a general criterion that provides an asymptotically unbiased estimator for  $\Delta(k)$  without requiring  $g(y) \in \mathcal{F}(k)$ , the Takeuchi (1976) information criterion.

## Application

- We consider a linear regression analysis based on data collected from 30 trauma patients in the state of Missouri during the 1990's.
- The outcome variable is the hospital charges incurred by the patient (in dollars), log transformed (LN\_HCHRG).
- The main explanatory variable is the length of the patient's hospital stay (in days), log transformed (LN\_LOS).

## Application

- An additional explanatory variable of interest is the *Injury Severity Score* (ISS), which quantifies the degree of anatomic derangement suffered by the trauma patient.
- There are two versions of this score.
- One version is computed by a hospital staff member based on a record of the patient's injury at the time of hospital admission (ISS\_HS).
- The other version is computed by a program that converts ICD-9 codes (International Classification of Diseases, Version 9) to an ISS (ISS\_ICD9).

## Application

- The primary question of interest is whether the hospital-staff coded ISS (ISS\_HS) or the ICD-9 ISS (ISS\_ICD9) is the better explanatory variable for use in conjunction with length of stay to predict hospital charges.
- Clinically, it would seem superfluous to use both ISS variables in conjunction with length of stay to predict hospital charges.

## Application

## AIC, AICc, MAIC Results

Model	AIC	AICc	MAIC
LOG_LOS + ISS_HS	-41.4	-39.8	-39.6
LOG_LOS + ISS_ICD9	-37.8	-36.2	-36.8
LOG_LOS + ISS_HS + ISS_ICD9	-39.7	-37.2	-37.2

# Application

## Conclusions

- All criteria indicate that the hospital-staff coded ISS (ISS\_HS) is superior to the ICD-9 ISS (ISS\_ICD9) for use in conjunction with length of stay to predict hospital charges.
- For all three models, the AICc and MAIC values are very similar.
- The model that includes both ISS variables yields higher criterion values than the model that only includes ISS\_HS. However, the difference in criterion values is *greater* for AICc and MAIC than for AIC.
  - Difference in AIC values:  $-1.7$ .
  - Difference in AICc values:  $-2.6$ .
  - Difference in MAIC values:  $-2.4$ .

## References

- Cavanaugh, J. E. (1997). Unifying the derivations of the Akaike and corrected Akaike information criteria. *Statistics & Probability Letters*, **33**, 201–208.
- Fujikoshi, Y., and Satoh, K. (1997). Modified AIC and  $C_p$  in multivariate linear regression. *Biometrika* **84**, 707–716.