

171:290 Model Selection

Lecture XIV: The Application of Model Selection Criteria

Joseph E. Cavanaugh

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

December 1, 2009

Introduction

- In this lecture, we will discuss and review important principles, concepts, procedures, and guidelines related to model selection and the use of model selection criteria.
- Each point of discussion will be introduced by a question.
- Our focus will be on applications and practice.

Introduction

Topics for Discussion

- The use of likelihood-based criteria to compare models with likelihoods of different distributional forms.
- Nested versus nonnested models.
- Model selection criteria versus hypothesis testing.
- “Data dredging” and selection criteria.
- The consequences of overfitting versus underfitting.
- Relative values of selection criteria.
- Overfitting: AIC versus hypothesis testing.

Question

- **Question:** Is it permissible to use AIC and other likelihood-based criteria (e.g., TIC, BIC, HQ) to compare models that have likelihoods of different distributional forms (e.g., normal, Poisson, binomial)?
- **Answer:** Yes.
 - Model selection criteria can be used to compare models that have likelihoods from different distributional families.
 - When likelihood-based criteria are computed, no constants should be discarded from the goodness-of-fit term $-2 \ln f(y | \hat{\theta}_k)$.
 - Keep in mind that certain statistical software packages routinely discard such constants.

Application

Illustrative AIC Application

- Burnham and Anderson (2002, §6.7.3) feature an example where the response variable is T_4 cell counts per cubic milliliter of blood.
- The response is measured for 20 patients in remission from Hodgkin's disease and for 20 controls.
- Several probability distributions are considered for the response, including normal, log-normal, gamma, Poisson, and negative binomial.
- Two mean structures are entertained: one that allows for two means for the two different groups (i.e., allows for a "treatment effect") and one that does not.

Application

Results:

Distribution	One Mean		Two Means	
	k	AIC	k	AIC
Normal	2	608.8	3	606.4
Log-Normal	2	590.1	3	588.6
Gamma	2	591.3	3	588.0
Poisson	1	11652.0	2	10204.0
Negative Binomial	2	589.2	3	586.0

Application

Conclusions

- There appears to be evidence in favor of a treatment effect: for each distribution, the AIC value for the two-mean model is less than the value for the one-mean model.
- The negative binomial distribution is favored by AIC.
- The Poisson distribution appears entirely inappropriate for this data.
- The Poisson models are the only models that do not include a separate dispersion parameter.
- Among the models that include a separate dispersion parameter, those based on a right-skewed distribution (log-normal, gamma, negative binomial) are favored over those based on the symmetric distribution (normal).

Question

- **Question:** Can a model selection criterion be used to compare nonnested models?
- **Answer:** Yes.
 - Model selection criteria can (and should) be used to compare nonnested models.
 - However, when nonnested models are compared using selection criteria, the distributions of differences in criterion values are less straightforward to characterize.
 - Thus, when comparing nonnested models using model selection criteria, marginal differences should be considered carefully.

Nonnested Models

Useful quotes on the use of AIC-type criteria

- 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.”

Nonnested Models

- Kitagawa (1987) states: “It must be emphasized that conceptually there are no difficulties in the comparison of Kullback-Leibler information numbers of nonnested models and there is a significant difference in the concept of likelihood as an estimate of the Kullback-Leibler number from the one in the statistical testing framework.”
- However, Kitagawa (1987) adds: “. . . differences in AIC values may be more variable in the nonnested situation. But here again . . . although some care is required in comparing marginal differences, there are no conceptual difficulties in comparing nonnested models.”

Question

- **Question:** Although hypothesis tests are developed under a different paradigm than model selection criteria, for the purpose of model selection, do both procedures share the same objective? Will they generally yield the same conclusion?
- **Answer:** No.
 - Consider two models M_{k_1} and M_{k_2} of dimensions k_1 and k_2 .
 - Assume that model M_{k_1} is nested within model M_{k_2} , so that M_{k_2} features additional effects not included in M_{k_1} .
 - A model selection criterion assesses whether the estimation error associated with including the additional effects is offset by the approximation error associated with excluding them.
 - A hypothesis test assumes that the larger model M_{k_2} is true. The test checks whether the additional effects are nonzero; i.e., whether there is any approximation error associated with excluding the effects.

Selection Criteria versus Hypothesis Testing

- The following example is from Sakamoto, Ishiguro, and Kitagawa (1986).
- Suppose that the larger model M_{k_2} features $L = k_2 - k_1$ additional effects not included in the smaller model M_{k_1} .
- Let $AIC(k_1)$ denote AIC for model M_{k_1} , and let $AIC(k_2)$ denote AIC for model M_{k_2} .
- Assume that the sample size is large enough for the penalty term of AIC to provide an adequate bias adjustment.

Selection Criteria versus Hypothesis Testing

- Suppose that $AIC(k_1) = AIC(k_2)$.
 - Interpretation: For the given sample size, the estimation error associated with including the L additional effects is equivalent to the approximation error associated with excluding them.
 - Neither model is preferred by AIC.
 - Based on parsimony, one might favor model M_{k_1}
- Assume that a likelihood-ratio test (LRT) is conducted to determine whether to include the L additional effects represented in model M_{k_2} .
- The following table provides the LRT p-values for various numbers of additional effects L .

Selection Criteria versus Hypothesis Testing

Table: P-Values for LRT Tests
Between Two Nested Models with Equivalent AIC Values.
(Larger Model Features L Additional Effects.)

L	P-Value
1	0.1573
2	0.1353
3	0.1116
4	0.0916
5	0.0752
10	0.0293
15	0.0119
20	0.0050

Selection Criteria versus Hypothesis Testing

- The p-values for the LRT test indicate increasingly strong support for the larger alternative model M_{k_2} .
- However, with AIC, both the null model and the alternative model are supported equally.
- The LRT test is only based on assessing approximation error (bias), and does not take the effect of estimation error into account (variability).
- Akaike (1974) noted “The use of a fixed level of significance for the comparison of models with various numbers of parameters is wrong, because it does not take into account the increase of the variability of the estimates when the number of parameters is increased.”

Question

- **Question:** Does the use of model selection criteria permit “data dredging”: post hoc analyses involving the exploration of a large set of conceivable variables and effects?
- **Answer:** No.
 - Post hoc exploratory analyses are inherent to statistical modeling, yet an exhaustive consideration of variables and effects often leads to the selection of a fitted model that is excessively tailored to the data at hand.
 - Reliance on model selection criteria provides protection from “chasing statistical significance,” yet may still lead to fitted models plagued by spurious, unimportant effects.

Data Dredging

- Burnham and Anderson (2002, p. 147) discuss an elk study by Cook et al. (2001) where analyses were conducted using stepwise linear regression, AIC, and Mallows' C_p .
- Cook et al. (2001) found that their approach produced a model that was “biologically unreasonable, unstable due to multicollinearity, and overparameterized.”
- Burnham and Anderson (2002, p. 147) claim that a common mistake in such studies is the “failure to posit a small set of a priori models, each representing a *plausible* research hypothesis.”
- Whenever possible, the candidate collection should be carefully configured based on a thorough consideration of the clinical and scientific principles governing the underlying phenomenon.

Question

- **Question:** From a predictive standpoint, is underfitting is more problematic than overfitting?
- **Answer:** In general, yes.
 - Recall that underfitting induces bias, whereas overfitting induces increased variability.
 - In large-sample settings, the additional variability resulting from overfitting is of less concern than the bias resulting from underfitting.
 - However, in small-sample to moderate-sample settings, the inflated variability resulting from gross overfitting may have a much more pronounced effect on overall predictive accuracy than the bias resulting from underfitting.
 - The preceding principles are reflected in the values of expected discrepancies (e.g., Kullback, Gauss).

Underfitting versus Overfitting

- Burnham and Anderson (2002, p. 17), in reference to Shibata (1989), state “While one must worry about errors due to both underfitting and overfitting, it seems that modest overfitting is less damaging than underfitting.”
- It is important to realize that the bias due to underfitting is not attenuated as the sample size is increased, whereas any inflation in variability due to overfitting may be compensated by an appropriate increase in the sample size.

Question

- **Question:** In a model selection application, the optimal fitted model is identified by the minimum (or maximum) value of the criterion. Are differences between criterion values therefore irrelevant?
- **Answer:** No.
 - In general, a model selection criterion scores every fitted model in a candidate family in accordance with how well each model balances the competing objectives of conformity to the data and parsimony.
 - The sizes of the scores are important; models with similar criterion values should receive the same “ranking” in assessing criterion preferences.

Relative Criterion Values

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

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

Relative Criterion Values

- For BIC, Kass and Raftery (1995, p. 777) feature the following table (slightly revised for presentation). (Lecture VI)

$BIC_i - BIC_{min}$	Evidence Against Model i
0 - 2	Not worth more than a bare mention
2 - 6	Positive
6 - 10	Strong
> 10	Very Strong

Relative Criterion Values

- Both of the preceding sets of guidelines imply that one should treat any AIC/BIC values that are within 2 units of one another as comparable.
- This “within 2” rule-of-thumb can be justified from various perspectives.
- We present an initial justification based on Bayes factors, posterior probabilities, and BIC.

Relative Criterion Values

- Consider two models M_{k_1} and M_{k_2} .
- Let $\text{BIC}(k_1)$ denote BIC for model M_{k_1} , and let $\text{BIC}(k_2)$ denote BIC for model M_{k_2} .
- Let $P(k_1 | y)$ denote the posterior probability for model M_{k_1} , and let $P(k_2 | y)$ denote the posterior probability for model M_{k_2} .
- Let B_{12} denote the Bayes factor for model M_{k_1} relative to model M_{k_2} .
- If M_{k_1} and M_{k_2} are assigned equal prior probabilities, the Bayes factor is equivalent to the ratio of posterior probabilities $P(k_1 | y)/P(k_2 | y)$, often called the posterior odds.

Relative Criterion Values

- As outlined in Lecture VI, the posterior odds can be approximated as follows:

$$P(k_1 | y) / P(k_2 | y) \approx \exp \{ (-1/2) (\text{BIC}(k_1) - \text{BIC}(k_2)) \} .$$

- In comparing two competing hypotheses (or models), Jeffereys' (1961, appendix B) claimed that the evidence provided by a Bayes' factor between 1 and 3 is "not worth a bare mention," and that the evidence provided by a Bayes' factor greater than 3 is "substantial."
- Assuming equal priors on M_{k_1} and M_{k_2} , a Bayes factor of 3 implies that model M_{k_1} is three times as likely as M_{k_2} .
- Based on the BIC approximation, a Bayes factor (or ratio of posterior probabilities) of 3 translates to a difference in BIC values of roughly 2 (2.19).

Relative Criterion Values

- Consider a generalized information criterion, GIC, defined as

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

where a_n represents a sequence that depends on the sample size n (and possibly the dimension k). (Lecture VIII)

- Let $\text{GIC}(k_1)$ denote GIC for model M_{k_1} , and let $\text{GIC}(k_2)$ denote GIC for model M_{k_2} .
- Based on the BIC approximation to $P(k_1 | y) / P(k_2 | y)$, an **evidence ratio** may be defined as

$$\exp \{ (-1/2)(\text{GIC}(k_1) - \text{GIC}(k_2)) \}.$$

- An evidence ratio of 3 translates to a difference in GIC values of roughly 2.

Question

- **Question:** Does the use of AIC (and other asymptotically equivalent criteria) tend to provide less protection against overfitting than the use of conventional hypothesis testing?
- **Answer:** **Not necessarily.**
 - Choosing a fitted candidate model based on the smallest AIC value, without consideration of any alternate fitted models with comparable AIC values, may provide less protection against overfitting.
 - However, choosing the most parsimonious model among all fitted models with AIC values comparable to the minimum AIC value may actually provide better protection against overfitting.
 - As previously discussed, a common guideline is to treat any AIC values that are within 2 units of one another as comparable.

Overfitting

- Consider a setting where AIC is used to choose between two fitted linear regression models, one of which is correctly specified, the other of which is overfit and contains L extraneous regressors.
- Based on likelihood-ratio theory, the asymptotic probability of AIC selecting the overfit model containing L extraneous regressors is given by $P(\chi_L^2 > 2L)$, where χ_L^2 is a centrally distributed chi-squared random variable based on L degrees of freedom. (Lecture VIII)
- The preceding fact can be used to tabulate asymptotic overfitting probabilities for AIC corresponding to $L = 1, 2, \dots$ extraneous regressors.

Overfitting

Table: Asymptotic Probability of Overfitting by L Variables Based on Strict AIC Values.

L	Probability
1	0.1573
2	0.1353
3	0.1116
4	0.0916
5	0.0752
6	0.0620
7	0.0512
8	0.0424

Overfitting

- Now suppose we regard two AIC values as comparable if the absolute difference between the values is less than 2.
- Additionally, if two AIC values are comparable and the models differ in complexity, suppose we choose the more parsimonious model.
- Based on this strategy, the asymptotic probability of selecting the overfit model containing L extraneous regressors is given by $P(\chi_L^2 > 2L + 2)$, where χ_L^2 is a centrally distributed chi-squared random variable based on L degrees of freedom.
- The preceding fact can be used to tabulate asymptotic overfitting probabilities corresponding to $L = 1, 2, \dots$ extraneous regressors.

Overfitting

Table: Asymptotic Probability of Overfitting by L Variables Based on Comparable AIC Values.

L	Probability
1	0.0455
2	0.0498
3	0.0460
4	0.0404
5	0.0348
6	0.0296
7	0.0251
8	0.0212