Stat 5102 Lecture Slides
Deck 7

Charles J. Geyer
School of Statistics
University of Minnesota

Model Selection

When we have two nested models, we know how to compare them: the likelihood ratio test.

When we have a short sequence of nested models, we can also use the likelihood ratio test to compare each consecutive pair of models. This violates the "do only one test" dogma, but is mostly harmless when there are only a few models being compared.

But what if the models are not nested or if there are thousands or millions of models being compared?

Model Selection (cont.)

This subject has received much theoretical attention in recent years. It is still an area of active research. But some things seem unlikely to change.

Rudimentary efforts at model selection, so-called forward and backward selection procedures, although undeniably things to do (TTD), have no theoretical justification. They are not guaranteed to do anything sensible.

Procedures that are justified theoretically evaluate a criterion function for all models in the class of models under consideration. They "select" the model with the smallest value of the criterion.

Model Selection (cont.)

We will look at two such procedures involving the Akaike information criterion (AIC) and the Bayes information criterion (BIC).

Suppose the log likelihood for model m is denoted l_m, the MLE for model m is denoted $\hat{\boldsymbol{\theta}}_m$, the dimension of $\hat{\boldsymbol{\theta}}_m$ is p_m, and the sample size is n

$$\begin{aligned} \text{AIC}(m) &= -2l_m(\hat{\boldsymbol{\theta}}_m) + 2p_m \\ \text{BIC}(m) &= -2l_m(\hat{\boldsymbol{\theta}}_m) + \log(n)p_m \end{aligned}$$

It is important to understand that both m and $\boldsymbol{\theta}$ are parameters, so $l_m(\boldsymbol{\theta})$ retains all terms in $\log f_{m,\boldsymbol{\theta}}(\mathbf{y})$ that contain m or $\boldsymbol{\theta}$.

Model Selection (cont.)

Suppose we want to select the best model (in some sense) from a class $\mathcal{M}$ which contains a model m_{sup} that contains all models in the class. For example, suppose we have a linear model with q predictors and the class $\mathcal{M}$ consists of all linear models in which the mean vector $\boldsymbol{\mu}$ is a linear function of some subset of these q predictors

$$\boldsymbol{\mu} = \alpha + \sum_{s \in S} \beta_s \mathbf{x}_s$$

where S is a subset, possibly empty, of these predictors. Since there are 2^q subsets, there are 2^q models in the class $\mathcal{M}$. The model m_{sup} is the one containing all q of the predictors.

Model Selection (cont.)

Each model contains an intercept α, so m_{sup} has $q + 1$ parameters.

A model with k predictors has $k + 1$ parameters, including the intercept.

The p_m in AIC or BIC is the number of parameters (including the intercept).

Model Selection (cont.)

There is so much discussion of this situation — the class $\mathcal{M}$ consists of 2^q models, each of which sets some of the coefficients in the model m_{sup} to zero — in the literature that one might think it is the only situation in which model selection arises.

This is not so. We know from our other examples, that even if one starts with only one predictor x_i it is easy to make up other predictors, such as x_i^2, x_i^3, ... in polynomials and $\sin(x_i)$, $\cos(x_i)$, $\sin(2x_i)$, $\cos(2x_i)$, ... in Fourier series.

So there are always infinitely many predictor variables that can be considered. Moreover, it often makes no sense to consider all possible subsets when these "made up" predictors are related.

Model Selection (cont.)

Nevertheless, special software exists only for this 2^q models case, and it is the only case we will do examples for.

The R function `regsubsets` in the `leaps` package does this.

It uses the *branch and bound* algorithm to find the best model of each size p (number of parameters) in a specified range. (With optional arguments, it can find the best k models of each size, for any k.)

Model Selection (cont.)

Having found the best model of each size, what is the best of all of them?

Maximum likelihood cannot be used for that, since it will always pick the supermodel m_{sup}. (The maximum over a superset is always larger.)

Minimum AIC and minimum BIC are two reasonable criteria that have been developed. Each of these procedures selects the set with the smallest value of the criterion.

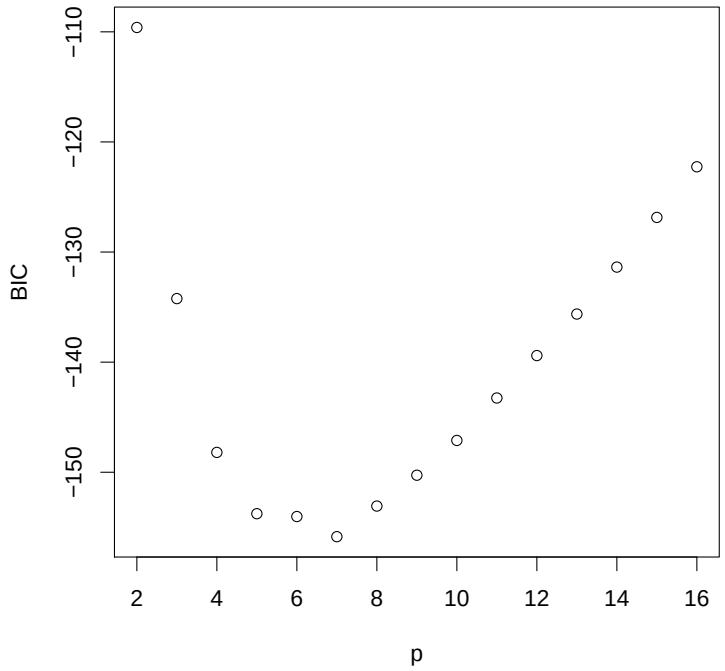
Model Selection (cont.)

Roughly speaking, AIC and BIC each "penalize" larger models. AIC has the smaller penalty $2p_m$; BIC has the larger penalty $\log(n)p_m$. AIC penalizes less and selects larger models; BIC penalizes more and selects smaller models.

The logic for the penalization is different in the two cases. More on that later.

Model Selection (cont.)

Example "when BIC is best" from the computer examples web pages.

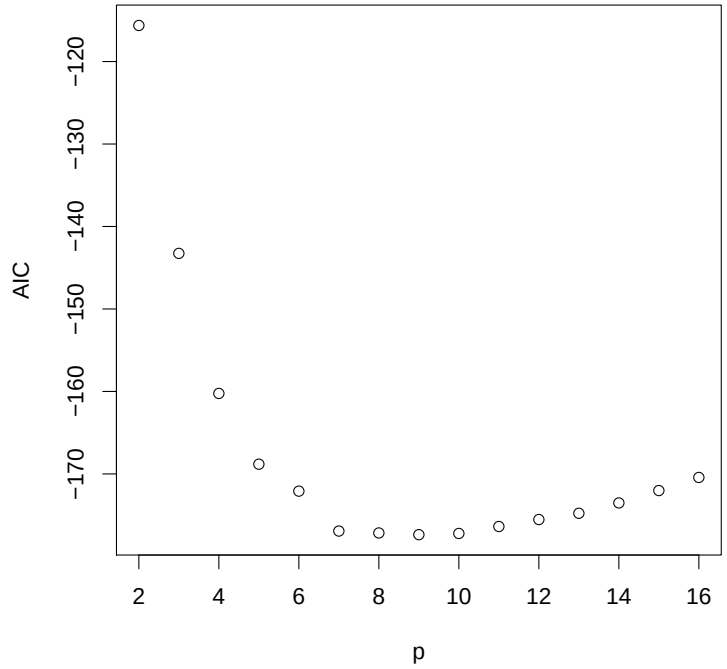


Model Selection (cont.)

An intercept is included in all models so each model has at least one parameter. Possible numbers of parameters range from 1 to 26 (there are 25 predictor variables). The best model according to the BIC criterion has $p = 7$ parameters (six predictors plus intercept).

Model Selection (cont.)

Example "when BIC is best" from the computer examples web pages.



Model Selection (cont.)

An intercept is included in all models so each model has at least one parameter. Possible numbers of parameters range from 1 to 26 (there are 25 predictor variables). The best model according to the AIC criterion has $p = 9$ parameters (eight predictors plus intercept).

These data were simulated, and the simulation truth model ($p = 6$) was closer to the one selected by BIC ($p = 7$). AIC selected a model that was too large ($p = 9$).

Model Selection (cont.)

BIC has a consistency property. When the true unknown model is one of the models under consideration and the sample size n goes to infinity, BIC selects the correct model with probability converging to one as $n \to \infty$.

In practice this means for this story to be approximately realistic, the true unknown model must be one of the models under consideration and must have p much smaller than n, hence only a few nonzero parameters.

In contrast AIC does not provide consistent model selection.

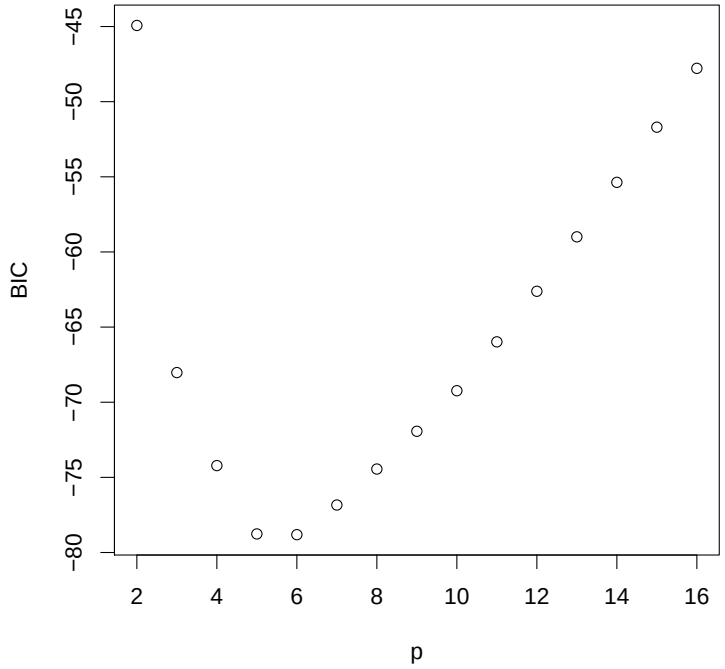
Model Selection (cont.)

This theoretical story, although much woofed about by statisticians, is not realistic in real applications. In scientific data, usually all predictors have some relation to the response, however weak. Moreover, many unmeasured predictors may also have some relation to the response. Thus the true model never has only a few nonzero parameters and never is in the class of models under consideration.

In this situation, the BIC penalty is too strong. It always selects small models which are never correct. AIC was developed to do approximately the right thing in this situation.

Model Selection (cont.)

Example "when AIC is best" from the computer examples web pages.

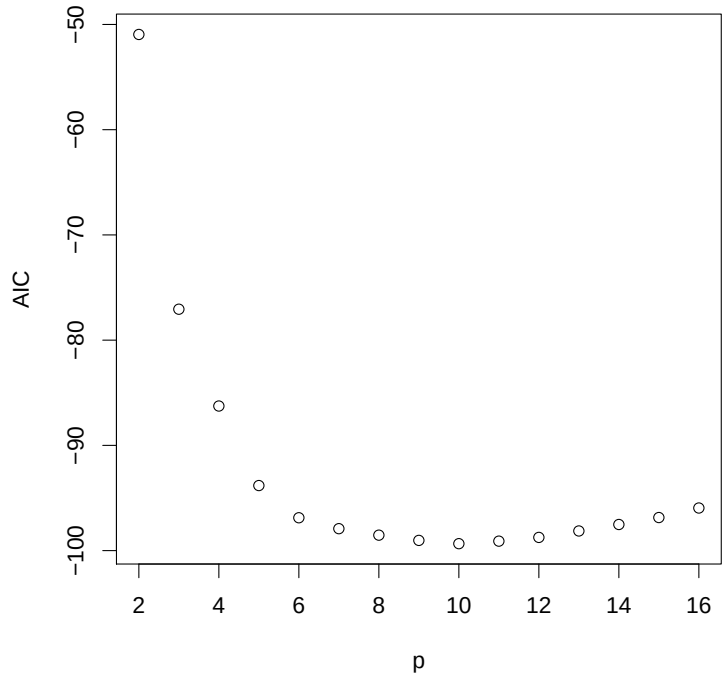


Model Selection (cont.)

An intercept is included in all models so each model has at least one parameter. Possible numbers of parameters range from 1 to 26 (there are 25 predictor variables). The best model according to the BIC criterion has $p = 6$ parameters (five predictors plus intercept).

Model Selection (cont.)

Example "when AIC is best" from the computer examples web pages.



Model Selection (cont.)

An intercept is included in all models so each model has at least one parameter. Possible numbers of parameters range from 1 to 26 (there are 25 predictor variables). The best model according to the AIC criterion has $p = 10$ parameters (nine predictors plus intercept).

These data were simulated, and the simulation truth model had nonzero regression coefficients for all 25 predictor variables. Both BIC and AIC selected a model that was too small, but AIC is always closer to correct in this situation, since it always selects a larger model.

Model Selection (cont.)

A slogan from one of my teachers (Werner Stutzle).

Regression is for prediction, not explanation.

When the true model is not even in the class of models under consideration, it is clear that the model "selected" cannot be true and cannot "explain" correctly. It can nevertheless predict well.

This slogan correctly summarizes the statistical properties of regression (LM and GLM). Most scientists are unhappy with it, because they want explanation. The slogan is a reminder of the unattainability of this desire.

Kullback-Leibler Information

The *Kullback-Leibler Information* (KLI) of a distribution with PDF/PMF f with respect to a distribution with PDF/PMF g is

$$\lambda(f) = -E_g\left\{\log\left(\frac{f(Y)}{g(Y)}\right)\right\}$$

Since $\exp(x) \geq 1 + x$, we have $\log(1 + x) \leq x$ and $\log(y) \leq y - 1$. Thus

$$\begin{aligned}\lambda(f) &\geq -E_g\left\{\frac{f(Y)}{g(Y)} - 1\right\} \\ &= -\int f(y)\,dy + \int g(y)\,dy \\ &= 0\end{aligned}$$

Clearly $\lambda(g) = 0$.

Kullback-Leibler Information (cont.)

Up to constants, KLI is negative expected log likelihood

$$E_g\{l(\boldsymbol{\theta})\} = E_g\{\log f_{\boldsymbol{\theta}}(Y) + h(Y)\} = E_g\{\log f_{\boldsymbol{\theta}}(Y)\} + E_g\{h(Y)\}$$

and

$$-\lambda(\boldsymbol{\theta}) = E_g\left\{\log\left(\frac{f_{\boldsymbol{\theta}}(Y)}{g(Y)}\right)\right\} = E_g\{\log f_{\boldsymbol{\theta}}(Y)\} - E_g\{\log g(Y)\}$$

and the terms that contain $\boldsymbol{\theta}$ agree.

Thus KLI measures how far f is from g in the same sense that log likelihood approximates.

Misspecified Maximum Likelihood

We know that when the model is correct, maximum likelihood is consistent, asymptotically normal, and efficient

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}_0$$

and

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, I(\boldsymbol{\theta}_0)^{-1}\Big)$$

When the model is not correct, maximum likelihood is not consistent. It cannot be since there is no $\boldsymbol{\theta}$ that corresponds to the true distribution of the data. In this case

$$\hat{\boldsymbol{\theta}}_n \xrightarrow{P} \boldsymbol{\theta}^*$$

where $\boldsymbol{\theta}^*$ minimizes KLI with respect to the true distribution of the data.

Misspecified Maximum Likelihood

When the model is misspecified the log likelihood derivative identities no longer hold. Because $\boldsymbol{\theta}^*$ minimizes KLI, we do have

$$E\{\nabla l_n(\boldsymbol{\theta}^*)\} = 0$$

which plays the role of the usual first log likelihood derivative identity in asymptotic theory. Fisher information can no longer be defined two ways.

$$I_n(\theta) = \text{var}\{\nabla l_n(\boldsymbol{\theta})\}$$
$$J_n(\theta) = -E\{\nabla^2 l_n(\boldsymbol{\theta})\}$$

are no longer equal. Each plays part of the role Fisher information plays in asymptotic theory. The resulting asymptotics are

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, J_1(\boldsymbol{\theta}^*)^{-1} I_1(\boldsymbol{\theta}^*) J_1(\boldsymbol{\theta}^*)^{-1}\Big)$$

AIC revisited

The *Akaike information criterion* (AIC) was developed as an unbiased estimate of twice KLI plus a constant (which does not matter). The idea is that it gives the best estimate possible of KLI that only depends on the log likelihood and p_m.

Better estimates have been developed but they are much more complicated. For example, the Takeuchi Information Criterion (TIC)

$$\mathrm{TIC}(m) = -2l_m(\hat{\boldsymbol{\theta}}_m) + 2\,\mathrm{tr}\Big[I_1(\boldsymbol{\theta}^*)J_1(\boldsymbol{\theta}^*)^{-1}\Big]$$

AIC revisited (cont.)

TIC does not assume any of the models under consideration are actually correct. It merely tries to find which of the models under consideration is closest to correct in the sense of KLI.

TIC(m) reduces to AIC(m) when model m is correct.

BIC revisited

BIC was developed to approximate unnormalized Bayes factors. Under the asymptotics of Bayesian estimation the within model priors do not affect the asymptotics (deck 4, slides 88–90). That is why BIC does not involve priors.