

171:290 Model Selection

Lecture VIII: Criteria for Regression Model Selection

Joseph E. Cavanaugh

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

October 13, 2009

Introduction

- The framework of the linear regression model serves as the foundation for statistical modeling.
- Many procedures exist for model selection in linear regression.
- However, some of the most widely known and used methods should probably be avoided.
- In this lecture, we review procedures for model selection as well as model validation in the framework of linear regression.

Introduction

Outline:

- Regression Model Selection Framework
- MSE and Adjusted R^2 , R^2_{adj}
- Procedures for Regression Model Selection
- A Generalized Information Criterion, GIC
- Complexity Penalization
 - Underfitting versus Overfitting
 - Consistency versus Asymptotic Efficiency
- Simulation Study
- Procedures for Model Validation

Regression Model Selection Framework

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

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

- Assume $f(y|\theta_k)$ corresponds to the regression model

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

- Here, y is an $n \times 1$ observation vector, e is an $n \times 1$ error vector, β is a $p \times 1$ parameter vector, and X is an $n \times p$ design matrix of full-column rank. Note that $k = (p + 1)$.
- **Fitted model**: $f(y|\hat{\theta}_k)$.

Popular Criteria for Regression Model Selection

- The Akaike (1973) information criterion:

$$\text{AIC} = -2 \ln f(y | \hat{\theta}_k) + 2(p + 1).$$

- The corrected Akaike (1973) information criterion (Sugiura, 1978; Hurvich and Tsai, 1989):

$$\text{AICc} = -2 \ln f(y | \hat{\theta}_k) + \frac{2(p + 1)n}{n - p - 2}.$$

- The Bayesian information criterion (Schwarz, 1978):

$$\text{BIC} = -2 \ln f(y | \hat{\theta}_k) + (p + 1) \ln n.$$

Popular Criteria for Regression Model Selection

- Note: In the present setting,

$$\begin{aligned} -2 \ln f(y | \hat{\theta}_k) &= n \ln \hat{\sigma}^2 + n(\ln 2\pi + 1) \\ &= n \ln(SS_{Res}/n) + n(\ln 2\pi + 1), \end{aligned}$$

where SS_{Res} denotes the residual sum of squares for the fitted model of interest.

- Thus, for selection criteria of the form

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

the goodness-of-fit term depends solely on the statistic SS_{Res} .

Popular Criteria for Regression Model Selection

- Mallows' (1973) conceptual predictive statistic:

$$C_p = \frac{SS_{Res}}{\tilde{\sigma}_*^2} - n + 2p.$$

Again, SS_{Res} denotes the residual sum of squares for the fitted model of interest; $\tilde{\sigma}_*^2$ denotes the mean square error for the largest fitted model, which typically includes all regressors under consideration.

- Other popular criteria for regression model selection: MSE; the adjusted coefficient of determination, R^2_{adj} .

MSE and Adjusted R^2 , R^2_{adj}

- The addition of regressors to a fitted model will generally decrease (and will never increase) SS_{Res} .
- Thus, one cannot choose the fitted model corresponding to the smallest SS_{Res} .
- However, the addition of regressors to a fitted model can increase $MSE = SS_{Res}/(n - p)$, since this addition decreases both SS_{Res} and $(n - p)$.
- Thus, practitioners often choose the fitted model corresponding to the minimum value of MSE.

MSE and Adjusted R^2 , R^2_{adj}

- The *coefficient of determination*, R^2 , is defined as

$$R^2 = 1 - \frac{SS_{Res}}{SS_{Total}} = \frac{SS_{Reg}}{SS_{Total}},$$

where SS_{Reg} represents the regression sum of squares for the fitted model of interest, and SS_{Total} represents the (corrected) total sum of squares.

- R^2 represents the proportion of variation in the response (SS_{Total}) which is explained by the fitted model (SS_{Reg}).
- The addition of regressors to a fitted model will generally increase (and will never decrease) R^2 .
- Thus, one cannot choose the fitted model corresponding to the largest R^2 .

MSE and Adjusted R^2 , R^2_{adj}

- The *adjusted coefficient of determination*, R^2_{adj} , is a variant of R^2 which imposes a penalization based on the number of regressors included in the fitted model.
- R^2_{adj} is defined as

$$\begin{aligned}R^2_{adj} &= 1 - \frac{(n-1) SS_{Res}}{(n-p) SS_{Total}} \\ &= 1 - \frac{SS_{Res}/(n-p)}{SS_{Total}/(n-1)} \\ &= 1 - \frac{MSE}{MS_{Total}}.\end{aligned}$$

- Note that choosing the fitted model corresponding to the maximum value of R^2_{adj} is equivalent to choosing the fitted model corresponding to the minimum value of MSE.

MSE and Adjusted R^2 , R^2_{adj}

- In regression analyses, MSE and R^2_{adj} are often used as model selection criteria; however, this practice should be avoided since it frequently leads to choosing overfit models.
- Suppose that the true or generating model $g(y)$ corresponds to the regression model

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

MSE and Adjusted R^2 , R^2_{adj}

- Consider the expected value of MSE under this true model:

$$E\{\text{MSE}\} = \sigma_o^2 + (X_o\beta_o)'(I - H)(X_o\beta_o)/(n - p),$$

where H denotes the projection matrix onto $C(X)$,

$$H = X(X'X)^{-1}X'.$$

- If the fitted model is underspecified, $E\{\text{MSE}\} > \sigma_o^2$.
- If the fitted model is correctly specified or overspecified, $E\{\text{MSE}\} = \sigma_o^2$.
- Hence, choosing the fitted model that corresponds to the minimum value of MSE (or the maximum value of R^2_{adj}) offers no discernable protection from overfitting.

MSE and Adjusted R^2 , R^2_{adj}

- Question: How does the penalization imposed by MSE / R^2_{adj} compare to that imposed by AIC, AICc, BIC, and C_p ?
- We will show that choosing the fitted model corresponding to the minimum value of MSE (or the maximum value of R^2_{adj}) is asymptotically equivalent to choosing the fitted model corresponding to the minimum value of

$$-2 \ln f(y | \hat{\theta}_k) + (p + 1).$$

Procedures for Regression Model Selection

- Since linear regression models are computationally inexpensive to fit, many statistical software packages support *best subsets regression*.
- In best subsets regression, for each regressor subset size (ranging from 1 up to the total number of candidate regressors), the b “best” models are reported.
 - The value of b is specified by the user.
 - The “best” models are the models corresponding to the smallest value of SS_{Res} .
- For many regression model selection criteria, the goodness-of-fit term depends solely on SS_{Res} , and the penalty term depends on p and possibly n .
- Thus, for fitted models of the same size p , an ordering based on SS_{Res} is the same as an ordering based on such selection criteria.

Procedures for Regression Model Selection

- In reporting the best fitting regression models for each regressor subset size, values of model selection criteria can also be output.
- An analyst can then choose a final fitted model, or at least determine a set of viable candidate models for final consideration.
- In SAS, PROC REG supports best subsets regression. Upon request, PROC REG will provide values of AIC, BIC (as SBC), C_p , and R^2_{adj} . Several other criteria are also available.

Procedures for Regression Model Selection

- Automatic variable selection procedures are also popular for regression model selection.
- PROC REG supports forward selection, backwards elimination, and stepwise selection.
- Automatic variable selection procedures are often favored because they are computationally efficient. However, in the regression framework, model fitting is computationally inexpensive; thus, this advantage is somewhat moot.

Procedures for Regression Model Selection

- Recall (from Lecture I) that automatic variable selection algorithms exclude consideration of many candidate models based on many different possible subsets of explanatory variables, and may lead one to a final fitted model based on an inferior subset.
- Example: Suppose that 10 candidate variables are considered for inclusion in a multivariable model.
 - In this setting, there are $2^{10} = 1028$ possible candidate models.
 - Backwards elimination or forward selection will consider at most $10 + 9 + 8 + \dots + 1 = 55$ of these models.

A Generalized Information Criterion, GIC

- We define a *generalized information criterion*, GIC, as

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

- In general, a_n represents a sequence that depends on the sample size n (and possibly the dimension k).
 - a_n may also be a constant.
 - Alternatively, a_n may converge to a constant as $n \rightarrow \infty$.
- In the regression setting, since $k = (p + 1)$, we will write GIC as

$$\text{GIC} = -2 \ln f(y | \hat{\theta}_k) + a_n (p + 1).$$

A Generalized Information Criterion, GIC

- With AIC, $a_n = 2$.
- With AICc, $a_n = 2n/(n - p - 2)$. Note that $a_n \rightarrow 2$ as $n \rightarrow \infty$.
- With C_p , we have argued (Lecture VII) that the selections are asymptotically equivalent to a GIC where $a_n \rightarrow 2$ as $n \rightarrow \infty$.
- With BIC, $a_n = \ln n$.
- With MSE / R^2_{adj} , we will show that the selections are asymptotically equivalent to a GIC where $a_n \rightarrow 1$ as $n \rightarrow \infty$.

A Generalized Information Criterion, GIC

- For the candidate model of interest, let $\hat{\sigma}^2$ denote the MLE of σ^2 .
- In the normal linear regression setting,

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

- Since $\hat{\sigma}^2 = SS_{Res}/n$, we also have

$$\text{MSE} = \frac{SS_{Res}}{(n-p)} = \frac{n}{(n-p)} \hat{\sigma}^2.$$

A Generalized Information Criterion, GIC

- Note that choosing the fitted model corresponding to minimum value of MSE is equivalent to choosing the fitted model corresponding to the minimum value of

$$\begin{aligned} & n \ln\{\text{MSE}\} + n(2\pi + 1) + 1 \\ &= n \ln \left\{ \frac{n}{(n-p)} \hat{\sigma}^2 \right\} + n(2\pi + 1) + 1 \\ &= \{n \ln \hat{\sigma}^2 + n(2\pi + 1)\} + n \ln \left\{ \frac{n}{(n-p)} \right\} + 1 \\ &= -2 \ln f(y|\hat{\theta}_k) + n \ln \left\{ \frac{n}{(n-p)} \right\} + 1. \end{aligned}$$

A Generalized Information Criterion, GIC

- Consider a first-order Taylor series expansion of $n \ln \{n/(n-p)\}$ in the argument $\{n/(n-p)\}$ about the point 1.
- We have

$$\begin{aligned} n \ln \left\{ \frac{n}{(n-p)} \right\} &\approx n \ln(1) + \left(\frac{n}{(n-p)} - 1 \right) n \\ &= \frac{np}{(n-p)}. \end{aligned}$$

A Generalized Information Criterion, GIC

- Thus, choosing the fitted model corresponding to the minimum value of MSE is equivalent to choosing the fitted model corresponding to the minimum value of

$$\begin{aligned} & -2 \ln f(y|\hat{\theta}_k) + n \ln \left\{ \frac{n}{(n-p)} \right\} + 1 \\ & \approx -2 \ln f(y|\hat{\theta}_k) + \frac{np}{(n-p)} + 1 \\ & = -2 \ln f(y|\hat{\theta}_k) + a_n(p+1), \end{aligned}$$

where

$$a_n = \frac{n}{(n-p)} - \frac{p}{(n-p)(p+1)}.$$

- Note that $a_n \rightarrow 1$ as $n \rightarrow \infty$.

A Generalized Information Criterion, GIC

Large-Sample Characterization of Popular Criteria

Criterion	Large-sample value of a_n
MSE, R^2_{adj}	1
AIC, AICc, C_p	2
BIC	$\ln n$

Complexity Penalization: Underfitting versus Overfitting

- What is the appropriate degree of complexity penalization for a model selection criterion?
- Consider a criterion the form of GIC.
- The goodness-of-fit term is $-2 \ln f(y | \hat{\theta}_k)$.
- The penalty term of GIC is $a_n(p + 1)$.
- The larger the size of a_n , the lower the probability of GIC choosing an overfit model, and the higher the probability of GIC choosing an underfit model.

Complexity Penalization: Underfitting versus Overfitting

- Consider a setting in which the candidate family \mathcal{F} consists of both underspecified and overspecified models.
- The goodness-of-fit term is $O(n)$.
- For any GIC where the penalty term sequence a_n is such that $(a_n/n) \rightarrow 0$ as $n \rightarrow \infty$, the asymptotic probability of GIC selecting an underfit model is zero.
- Based on likelihood-ratio theory, the asymptotic probability of GIC selecting an overfit model containing L extraneous regressors is given by $P(\chi_L^2 > a_n L)$, where χ_L^2 is a centrally distributed chi-squared random variable based on L degrees of freedom.

Complexity Penalization: Underfitting versus Overfitting

- With BIC, $a_n = \ln n$. The probability $P(\chi_L^2 > a_n L)$ becomes smaller as the sample size grows.
- With MSE, R^2_{adj} , AIC, AICc, and C_p , a_n is either a constant or converges to a constant: specifically, 1 or 2. The probability $P(\chi_L^2 > a_n L)$ is always appreciably greater than zero.

Complexity Penalization: Underfitting versus Overfitting

- Recall the definitions of consistency and asymptotic efficiency.
- 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.

Asymptotic Efficiency and Consistency

- Asymptotic efficiency was introduced by Shibata (1980) in the context of autoregressive time series models.
- It was later extended by Shibata (1981) to linear regression models.
- For the linear regression framework, in the setting outlined by Shibata (1981), AIC, AICc, and C_p are asymptotically efficient; MSE, R^2_{adj} , and BIC are not.
- BIC is consistent; MSE, R^2_{adj} , AIC, AICc, C_p , are not.

Asymptotic Efficiency versus Consistency

- Which property, consistency or asymptotic efficiency, is preferential?
- In addressing this question, keep in mind that both properties are *asymptotic*. The finite sample behavior of model selection criteria will often not reflect asymptotic optimality properties.
- McQuarrie and Tsai (1998): “Researchers who believe that the system they study is infinitely complicated, or that there is no way to measure all the important variables, choose models based on efficiency.”
- McQuarrie and Tsai (1998): “[Consistency is preferred when] the researcher believes that all variables can be measured, and furthermore, that enough is known about the physical system being studied to write the list of all important variables.”

Simulation Study

Study Outline:

(See Lecture VII)

- In each of four simulation sets, one thousand samples of size n are generated from a true regression model which has an n by $p_o = 5$ design matrix, a parameter vector of the form $\beta_o = (1, 1, 1, 1, 1)'$, and an error variance of $\sigma_o^2 = 16$.
- For every sample, candidate models with nested design matrices of ranks $p = 2, 3, \dots, P = 8$ are fit to the data.
 - The first column of every design matrix is a vector of ones.
 - The design matrix of rank $p_o = 5$ is correctly specified.
- The covariates are generated as *iid* replicates from a uniform $(0, 10)$ distribution.

Simulation Study

- In the four simulation sets, four different sample sizes n are employed: 30 (“small”), 60 (“moderate”), 200 (“large”), and 1000 (“very large”).
- We examine the effectiveness of MSE / R^2_{adj} , AIC, AICc, C_p , and BIC at selecting p , the order of the model.

Simulation Study

Set I: Order selections with $n = 30$.

p	MSE	AIC	AIC _c	C_p	BIC
2	0	0	0	0	0
3	0	1	5	1	7
4	2	6	20	11	22
5	449	640	846	718	829
6	176	154	88	132	88
7	163	91	28	68	32
8	210	108	13	70	22

Simulation Study

Set II: Order selections with $n = 60$.

p	MSE	AIC	AIC _c	C_p	BIC
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	1	1	2
5	475	712	811	744	923
6	173	132	113	125	57
7	143	83	45	72	13
8	209	73	30	58	5

Simulation Study

Set III: Order selections with $n = 200$.

p	MSE	AIC	AIC _c	C_p	BIC
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	490	783	806	787	977
6	171	111	105	110	20
7	176	70	62	71	2
8	163	36	27	32	1

Simulation Study

Set IV: Order selections with $n = 1000$.

p	MSE	AIC	AIC _c	C_p	BIC
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	448	776	784	781	992
6	193	142	135	138	8
7	167	51	50	50	0
8	172	31	31	31	0

Procedures for Model Validation

- A model selection criterion attempts to find the “best” fitted model among those models in a candidate collection.
- However, there is no guarantee that the selected model will be an adequate model, since all of the models in the candidate collection could be inappropriate.
- Recall (from Lecture I) that an optimal statistical model is characterized by three fundamental attributes.
 - Parsimony: model simplicity.
 - Goodness-of-fit: conformity of the fitted model to the data at hand.
 - Generalizability: applicability of the fitted model to describe or predict new data.

Procedures for Model Validation

- *Model validation* refers to the process of ensuring that the selected fitted model provides an adequate fit to the data used in its own construction, and is capable of adequately describing and predicting new data.

Procedures for Model Validation

- Goodness-of-fit can be investigated by checking whether the residuals mimic the assumed distributional characteristics of the model errors: $e \sim N_n(0, \sigma^2 I)$.
 - Residual plots are useful for checking the mean-zero and constant-variance assumptions, as well as the assumption of independence.
 - Histograms, boxplots, and quantile-quantile plots for residuals are useful for checking the normality assumption.
 - Tests for normality (e.g., Kolmogorov-Smirnov) can also be applied to the residuals, but they should be used with caution.

Procedures for Model Validation

- Lack-of-fit tests allow one to determine whether the mean structure for the candidate model is misspecified.
 - The classical lack-of-fit test requires exact replicates.
 - Near replicate lack-of-fit tests are also available (Shillington, 1979; Christensen, 1989, 1991; Utts, 1982.)
- The predictive effectiveness of a model can be assessed using *cross-validation*.
 - Leave-out cross validation procedures may be used for both model selection and model validation.
 - However, a true assessment of predictive efficacy must be based on “new” data, or on split-sample validation.

Procedures for Model Validation

- In split-sample validation, the sample is randomly split into two parts, a *training sample* and a *test sample*.
- The model is fit based on the training sample.
- The predictive efficacy of the fitted model is then assessed based on the ability of the model to predict the data in the test sample.
- From Barnard (1974): “The simple idea of splitting a sample into two and then developing the hypothesis [model] on the basis of one part and testing it on the remainder may perhaps . . . be one of the most seriously neglected ideas in statistics, if we measure the degree of neglect by the ratio of the number of cases where a method could give help to the number where it is actually used.”