Overfitting Probabilities of Model Selection Criteria

C. K. Park

Abstract

Probabilities of overfitting for model selection criteria in regression are derived for several different situations. First, one candidate model with one extra variable is compared to the current model. This is expanded to m candidate models. We assume that these comparisons are independent and discuss upper bounds for overfitting probabilities. We found the overfitting probabilities of AIC, AICc, SIC, SICc, and HQ on one extra variable case and multiple extra variable case.

1. Introduction

We introduce the forms of model selection criteria and find the probabilities of overfitting. We then expand the probabilities of overfitting to the multiple candidate model case where none of the additional variables are important to the model. This is similar to a repeated testing problem where all null hypotheses are true. The distribution of SSE(Sum of Square of Errors) and the distribution for the difference in SSE between two nested models are discussed. The probability that a model selection criterion is overfitted by one variable can be written as an F-test. By assuming independence in the F-tests, upper bounds for probabilities of overfitting can be easily computed. Probabilities of overfitting are independent across orders due to variables entering the model on the basis of their order statistics. We begin with a discussion of the orthogonal regression model.

2. Orthogonal regression model

Let the true model orthogonal regression model be

Department of Applied Mathematics, Korea Maritime University

$$Y = X_*\beta_* + \varepsilon_*, \tag{1}$$

where $\varepsilon_* \sim N(0, \sigma_*^2 I_n)$, X_* is the $n \times k_*$ design matrix with $k_* = rank(X_*)$, Y is an $n \times 1$ vector of observations, , β_* is a $k_* \times 1$ vector of unknown parameters, and ε_* is an $n \times 1$ vector of errors. The candidate orthogonal regression model is

$$Y = X\beta + \varepsilon, \tag{2}$$

where $\varepsilon \sim N(0, \sigma^2 I_n)$ and $k = rank(X)$. Without loss of generality, we assume that the design matrices $X_* = (1, x_1, \cdots, x_{k_*-1})$ and $X = (1, x_1, \cdots, x_{k-1})$satisfy $X_*'X_* = nI_{k*}$, and $X'X = nI_k$, respectively, where $x_j = (x_{j,1}, \cdots, x_{j,n})'$. In addition, we define underfitting as $k < k_*$ $(X \subset X_*)$ and overfitting as $k > k_*$ $(X_* \subset X)$.

Based on candidate model (2), the least estimator of β is $\hat{\beta} = (X'X)^{-1}X'Y = X'Y/n$, where $Y = (y_1, \cdots, y_n)'$, and the resulting sum of squares of errors is

$$SSE_k = (Y - \overline{Y})'(Y - \overline{Y}) - \sum_{j=1}^{j=1} \frac{1}{n}(X_j'Y)^2, \tag{3}$$

where X_j represents the *jth* variable included in the model. The unbiased and maximum likelihood estimates of σ^2 are $s_k^2 = SSE_k/(n-k)$ and $\hat{\sigma}_k^2 = SSE_k/n$, respectively.

One consequence of orthogonality is that to compare all subsets of the available candidate variables for orthogonal regression, one only needs to compute SSE for all one variable models. In this case, when the *jth* variable X_j is added to the candidate model, the variable count increases by one and the SSE decreases by $(X_j'Y)^2/n$. The best one variable model is that for which $(Y - \overline{Y})'(Y - \overline{Y}) - (X_j'Y)^2/n$ is the smallest (or alternatively, that which consists of the variable with the largest $(X_j'Y)^2/n)$. Without loss of generality, in the discussions that follow we assume that candidate variables have been sorted in this way.

The variable with the largest $(X_j'Y)^2/n$ are entered into the model first. Order

k = 1 refers to the intercept only model. k = 2 represents the best 1-variable model in the sense that this model has the smallest SSE for all 1-variable models. The k = 2 model contains the variable with largest $(X_j' Y)^2/n$. In general, the order k model refers to the k-1 variable model with the smallest SSE and containing variables with the largest $(X_j' Y)^2/n$. Order K represents the order of the model with all X_j include.

The $(X_j' Y)^2/n$ are independent (possibly non-central) χ_1^2 random variables. Reduction in SSE has a distribution based on the order statistics of independent random variables. However, they may have a central or non-central distribution. In the simplest case, we have independent identically distributed order statistics. Typically, some of the X_j are important (yielding non-central χ_1^2) and we have independent but not identically distributed.

Consider the underfit model $Y = X_0\beta_{*0} + \varepsilon_*$, where X_1 has been omitted from the model. Underfit models tend to be too simplistic and make poor predictions. $\hat{\beta}$ is unbiased for β but s^2 is biased high for σ^2. The overfit candidate model is $Y = X_0\beta_{*0} + X_1\beta_{*1} + X_2\beta_{*2} + \varepsilon_*$ where $X_* = (X_0 : X_1)$ and this model contains the extra variables in X_2. The model is needlessly complex. Both $\hat{\beta}$ and s^2 are unbiased. However, when k, the number of parameters including the intercept, is close to the sample size n, we can get biased estimates. The overfit model can also make poor predictions, which is unnecessarily complex. The controlling of underfitting and overfitting is an important rule for finding the best model in regression.

3. Review of model selection criteria

Now, we review some common efficient criteria. Akaike(1973) showed that AIC is asymptotically unbiased for the Kullback-Leibler information (Kullback and Leibler, 1951) up to a constant. $AIC = nlog(\hat{\sigma}_k^2) + 2(k+1) + nlog(2\pi) + n$; the last two terms are not important for model selection, so we can ignore them. Simplifying and scaling by n, we get

$$AIC = \log(\hat{\sigma}_k^2) + \frac{2(k+1)}{n}. \tag{4}$$

The model which minimizes AIC is considered to be closest to the true model. However, AIC tends to be overfitted in small samples (Nishii, 1984 ; Hurvich and Tsai, 1989). Hurvich and Tsai (1989) attained the bias-corrected, in terms of selected order, version of AIC. AICc estimates the expectation of K-L and performs better than AIC in small samples.

AICc is a better criterion than AIC to find the true model in small samples. However, AICc is asymptotically equivalent to AIC in large samples. Hurvich and Tsai modified AIC to provide an exactly unbiased estimator for the expected K-L information, assuming that the errors have a normal

$$AICc = \log(\hat{\sigma}_k^2) + \frac{n+k}{n-k-2} \tag{5}$$

It can be shown that

$$AICc = AIC + \frac{2(k+1)(k+2)}{n-k-2} + n.$$

When k increases to n-2, the second term of above equation goes to a plus infinity. AICc is AIC plus an additional penalty term.

SIC(BIC) (Schwarz, 1978 ; Akaike, 1978) can be overfitted in small samples due to the linear (in k) penalty function. The equation of SIC is

$$SIC = \log(\hat{\sigma}_k^2) + \frac{\log(n)k}{n}. \tag{6}$$

In large samples, the penalty term $\frac{\log(n)k}{n}$ is much larger than the 2(k+1) penalty term in AIC. This large penalty function prevents overfitting in large samples.

HQ (Hannan and Quinn, 1979) is a strongly consistent estimation procedure based on the law of the iterated logarithm. The equation of HQ is

$$HQ = \log(\hat{\sigma}_k^2) + \frac{2\log\log(n)k}{n} \tag{7}$$

HQ behaves more like the efficient model selection AIC. When the sample size is small, the penalty function of HQ is similar to that of AIC. For example, *loglog*(100) = 1.527, *loglog*(1000) = 1.933, and *loglog*(10000) = 2.220. The *loglog*(n) term represents the ratio of the HQ penalty function to the AIC penalty function. Indeed for n = 200000, loglog(200000) is 2.502, and the penalty function of HQ is only approximately 2.5 times larger than that of AIC.

The last criterion we consider is SICc (McQuarrie, 1999). SICc can be derived by using the relationship between AIC and AICc. The penalty function of SICc is the penalty function of SIC scaled by $\frac{n}{n-k-2}$. SICc is defined as

$$SICc = \log(\hat{\sigma}_k^2) + \frac{\log(n)k}{n-k-2}. \tag{8}$$

4. Probabilities of overfitting

We are now examining probabilities of overfitting for these criteria. We denote the reduced model k and the full model by k+1. We begin with the one candidate model case only and compare the true model to this one candidate model. Suppose that the true model is k, and add only one variable to the true model. We will find the probability of overfitting of this situation (add one variable).

AIC : AIC overfits if $AIC_{k+1} < AIC_k$.

$$P(AIC_{k+1} < AIC_k) =$$

$$P\left\{F_{1,n-k-1} > (n-k-1)\left(\exp\left(\frac{2}{n}\right)-1\right)\right\} \tag{9}$$

AICc : AICc overfits if $AIC_{C_{k+1}} < AIC_{C_k}$

$$P\{AIC_{C_{k+1}} < AIC_{C_k}\} =$$

$$P\left\{F_{1,n-k-1} > (n-k-1)\left(\exp\left(\frac{2(n-1)}{(n-k-3)(n-k-2)}\right)-1\right)\right\}. \tag{10}$$

SIC : SIC overfits if $SIC_{k+1} < SIC_k$.

$$P\{SIC_{k+1} < SIC_k\}$$

$$P\left\{F_{1,n-k-1} > (n-k-1)\left(\exp\left(\frac{\log(n)}{n}\right)-1\right)\right\}. \tag{11}$$

SICc : SICc overfits if $SIC_{C_{k+1}} < SIC_{C_k}$.

$$P\{SIC_{C_{k+1}} < SIC_{C_k}\} =$$

$$P\left\{F_{1,n-k-1} > (n-k-1)\left(\exp\left(\frac{\log(n)(n-2)}{(n-k-3)(n-k-2)}\right)-1\right)\right\}. \tag{12}$$

HQ : HQ overfits if $HQ_{k+1} < HQ_k$.

$$P\{HQ_{k+1} < HQ_k\} =$$

$$P\left\{F_{1,n-k-1} > (n-k-1)\left(\exp\left(\frac{2\log\log(n)}{n}\right)-1\right)\right\}. \tag{13}$$

We see that these probabilities all follow the F distribution and will be referred to as F-tests.

Table 1 presents probabilities using equations (9)-(13) for preferring order *k*+1 over the current order *k*. In table 1, *n* is the sample size, $k = Rank(X)$, and *K* is the number of total variables including the intercept. When the sample size increases, the probabilities of overfitting for AIC, SIC, SICc, and HQ tend to decrease, probabilities of overfitting for AICc increase. When *K* increases, there is no change in the probabilities in Table 1. When *k* increases, probabilities of overfitting of AIC, SIC, and HQ increase due to linear penalty functions of their equations. When *k* increases, probabilities of overfitting of AICc and SICc decrease due to dividing by *n-k-2* in their penalty functions. Probabilities of overfitting for SICc are smaller than those of the other model selection criteria. We say SICc has the strongest penalty function.

Consider the case where more than one candidate model is considered, which is a multiple testing situation. Orthogonal regression yields independent chi-squares, and we overfit if any of the overfit candidate models are selected. However, the F-tests are not independent as shown below. Table 2 presents probabilities assuming *i.i.d.* F-tests. We will compare these probabilities to those where we include the dependence of the F-tests. Note that the probabilities in Table 2 are much easier to compute. *K* denote the maximum possible model order (total number of variables plus the intercept) and *k* denote the model order. There are *K-k* 1-additional-variable models to compare with the current model. Let α be the probabilities of selecting one additional variable when only the current model is compared to one candidate model containing one additional variable. Equations (9)-(13) represent α probabilities. Assuming independence for illustration purpose, the probability of favoring an order *k*+1 model over the current order *k* model is

$1-(1-\alpha)^{K-k}$. Table 2 presents these probabilities.

In Table 2, we can see that the probability of overfitting increases as K increases. With more candidate models to choose from, the higher the chance of overfitting. As in Table 1, model selection criteria with stronger penalty functions have smaller probability of overfitting. Probabilities for K = 6 and k = 5 are the same as in Table 1 since there is only one candidate model to compare to the current model. Although the variables are orthogonal, the F-tests are not independent as we show below. However, the patterns in overfitting probabilities are the same as including the dependence. Model selection criteria with weaker penalty functions overfits with higher probability.

6. Conclusion and Further Research

Usual comparisons of one reduced vs. one full model describe the basic behavior of model selection criterion. Criteria with stronger penalty functions have smaller probabilities of overfitting. Assuming independence for these comparisons can lead to overestimating the probability of overfitting due to the variables entering into the model according to their order statistics.

We now think about comparisons of one candidate model with one extra variable and expansion of m candidate models when these comparisons are not independent. Probabilities will be computed using the dependence of F distributions and F distributions based on order statistics of independent Chi-squares.

Table 1. Single candidate model case.

		K=6					K=11				
n	k	AIC	AICc	SIC	SICc	HQ	AIC	AICc	SIC	SICc	HQ
10	1	0.220	0.072	0.188	0.069	0.263	-	-	-	-	-
10	3	0.293	0.025	0.259	0.024	0.337	-	-	-	-	-
10	5	0.400	0.001	0.366	0.001	0.442	-	-	-	-	-
20	1	0.186	0.118	0.105	0.062	0.166	0.186	0.118	0.105	0.062	0.166
20	3	0.213	0.094	0.127	0.046	0.192	0.213	0.094	0.127	0.046	0.192
20	5	0.245	0.070	0.155	0.031	0.223	0.245	0.070	0.155	0.031	0.223
50	1	0.168	0.142	0.054	0.042	0.107	0.168	0.142	0.054	0.042	0.107
50	3	0.177	0.133	0.059	0.038	0.115	0.177	0.133	0.059	0.038	0.115
50	5	0.187	0.124	0.065	0.033	0.123	0.187	0.124	0.065	0.033	0.123
100	1	0.163	0.150	0.034	0.030	0.084	0.163	0.150	0.034	0.030	0.084
100	3	0.167	0.146	0.036	0.028	0.088	0.167	0.146	0.036	0.028	0.088
100	5	0.172	0.141	0.038	0.026	0.091	0.172	0.141	0.038	0.026	0.091
10000	1	0.157	0.157	0.002	0.002	0.035	0.157	0.157	0.002	0.002	0.035
10000	3	0.157	0.157	0.002	0.002	0.035	0.157	0.157	0.002	0.002	0.035
10000	5	0.157	0.157	0.002	0.002	0.035	0.157	0.157	0.002	0.002	0.035

Table 2. Multiple candidate models case, with independence.

		K=6					K=11				
n	k	AIC	AICc	SIC	SICc	HQ	AIC	AICc	SIC	SICc	HQ
10	1	0.711	0.313	0.647	0.301	0.782	-	-	-	-	-
10	3	0.647	0.074	0.593	0.069	0.708	-	-	-	-	-
10	5	0.400	0.001	0.366	0.001	0.442	-	-	-	-	-
20	1	0.642	0.466	0.427	0.275	0.596	0.843	0.676	0.633	0.440	0.804
20	3	0.513	0.256	0.336	0.132	0.473	0.813	0.498	0.615	0.280	0.775
20	5	0.245	0.070	0.155	0.031	0.223	0.755	0.303	0.569	0.144	0.717
50	1	0.602	0.536	0.242	0.194	0.433	0.809	0.749	0.393	0.322	0.640
50	3	0.443	0.349	0.167	0.109	0.307	0.745	0.633	0.348	0.236	0.575
50	5	0.187	0.124	0.065	0.033	0.123	0.645	0.485	0.286	0.156	0.482
100	1	0.588	0.556	0.159	0.140	0.357	0.798	0.768	0.268	0.238	0.548
100	3	0.422	0.376	0.104	0.082	0.241	0.722	0.668	0.226	0.180	0.474
100	5	0.172	0.141	0.038	0.026	0.091	0.610	0.533	0.176	0.125	0.380
10000	1	0.575	0.575	0.012	0.012	0.164	0.786	0.786	0.022	0.021	0.275
10000	3	0.402	0.401	0.007	0.007	0.102	0.698	0.698	0.017	0.017	0.222
10000	5	0.157	0.157	0.002	0.002	0.035	0.575	0.575	0.012	0.012	0.164

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, *in 2nd International symposium on Information Theory* 267-281. (Eds) B.N. Petrov and F.Csaki, Akademia Kiado, Budapest.

[2] Akaike, H. (1978). A bayesian analysis of the minimum AIC procedure. *Annals of the Institute of Statistical Mathematics* 30, Part A, 9-14.

[3] Hannan, E.J. and Quinn, B.G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society*, B 41, 190-195.

[4] Hurvich, C.M. and Tsai, C.L. (1989). Regression and time series model selection in small samples. *Biometrika* 76, 297-307.

[5] Kullback, S. and Leibler, R.A. (1951) On information and sufficiency. *Annals of Mathematical Statistics* 22, 79-86.

[6] McQuarrie, A.D. (1999). A small-sample correlation for the Schwarz SIC model selection criterion. *Statistics & Probability Letters* 44, 79-86.

[7] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, 12(2), 758-765.

[8] Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461-464.