The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 589-601

Overfitting Probabilities using Dependent F-tests in Regression[1)]

Chan Keun Park[2)]

Abstract

Probabilities of overfitting for model selection criteria 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 show that these comparisons are not independent and discuss overfitting probabilities. Correlation between two F-tests is derived. Finally, probabilities are computed using the dependent F distributions and F distributions based on order statistics of independent Chi-squares.

Keywords : Order statistics, F distribution, AIC, AICc, SIC, SICc, HQ

1. Introduction

The orthogonal regression model is examined, and properties of model selection criteria in orthogonal regression are discussed. 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 overfits by one variable can be written as an F-test. We show that these forward(add one variable) F-tests are not independent and derive their correlation. By assuming independence in the F-tests, upper bounds for probabilities of overfitting can be computed. The distributions for comparing a reduced model of k to nested full models of order $k+1$ are not independent—the within order case. Probabilities of overfitting are dependent 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.

We found probabilities of overfitting for some model selction criteria which are introduced

1) This work is supported by Professor Research Fund of Korea Maritime University in 2001
2) Full-time Instructor, Division of Applied Science, Korea Maritime University, Pusan, 606-791, Korea
E-mail : chapark@hanara.kmaritime.ac.kr

in section 3. McQuarrie(1999) compared probabilities of overfitting for SICc and SIC using independent F-tests. McQuarrie, Shumway and Tsai(1997) also found overfitting probabilities for AICu and some model selection criteria using the same idea. However, we find the probabilities of model selction overfitting in orthogonal using dependent F-tests.

2. Orthogonal regression model

Let the true orthogonal regression model be

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

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

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

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 $X_*=(X_0 \vdots X_1)$ 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. AIC is designed to be an asymptotically unbiased estimator of the Kullback-Leibler information (Kullback and Leibler, 1951) of a fitted model. Kullback-Leibler discrepancy (K-L) is a measure of closeness between two density functions, g and f.

$$K-L = E_*\left[\log\left(\frac{g}{f}\right)\right],$$

where g is the density function of the true model, f is the density function of the candidate model, and E_* denotes expectation under the true model. In regression, let g denote the density of the true model (1) and let f denote the density of the candidate model (2). Under the normality assumption and scaling by $2/n$, we have

$$K-L = \log\left(\frac{\hat{\sigma}_k^2}{\sigma_*^2}\right)+\frac{\sigma_*^2}{\hat{\sigma}_k^2}+\frac{L_2}{\hat{\sigma}_k^2}-1, \tag{4}$$

where

$$L_2 = \frac{1}{n}||X_*\beta_*-X\hat{\beta}||^2=\frac{1}{n}\sum_{i=1}^{n}(\mu_{*i}-x_i'\hat{\beta})^2. \tag{5}$$

One measure of the difference between the true model (1) and the candidate model (2) is the L_2 distance in (5). The expected L_2 distance assuming $X'X=nI_n$ is therefore

$$E_*[L_2(k)] = \frac{\sigma_*^2}{n} k + \sum_{j=k}^{\infty} \beta_{*j}^2 ,$$

where E_* denotes expectation under the true model (1). Now let k' be the model for which $E_*[L_2(k')]$ attains the minimum, and let k_c be the model chosen by a model selection criterion. Then the asymptotic efficiency of the choice is defined to be

$$\lim_{n\to\infty} \frac{E_*[L_2(k')]}{E_*[L_2(k_c)]} .$$

Note that this definition, where the highest efficiency is 1, is the inverse of Shibata's (1980, 1981) definition. A model selection criterion is said to be efficient if P(asymptotic efficiency =1) = 1. AIC(Akaike, 1973) and AICc(Hurvich and Tsai, 1989) are asymptotically efficient criteria.

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

$$AIC = \log(\hat{\sigma}_k^2) + \frac{2(k+1)}{n} . \tag{6}$$

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

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

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

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

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

$$HQ = \log(\hat{\sigma}_k^2) + \frac{2\log\log(n)k}{n} \tag{9}$$

HQ behaves more like the efficient model selection 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

$$SICc = \log(\hat{\sigma}_k^2) + \frac{\log(n)k}{n-k-2}. \tag{10}$$

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 $AIC_{k+1} < AIC_k$.

$$P(AIC_{k+1} < AIC_k)$$
$$= P\left\{F_{1,n-k-1} > (n-k-1)\left(\exp\left(\frac{2}{n}\right)-1\right)\right\} \tag{11}$$

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

$$P\{AIC_{C_{k+1}} < AIC_{C_k}\}$$
$$= P\left\{F_{1,n-k-1} > (n-k-1)\left(\exp\left(\frac{2(n-1)}{(n-k-3)(n-k-2)}\right)-1\right)\right\}. \tag{12}$$

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

$$P\{SIC_{k+1} < SIC_k\}$$
$$= P\left\{F_{1,n-k-1} > (n-k-1)\left(\exp\left(\frac{\log(n)}{n}\right)-1\right)\right\}. \tag{13}$$

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

$$P\{SIC_{C_{k+1}} < SIC_{C_k}\}$$
$$= P\left\{F_{1,n-k-1} > (n-k-1)\left(\exp\left(\frac{\log(n)(n-2)}{(n-k-3)(n-k-2)}\right)-1\right)\right\}. \tag{14}$$

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

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

$$= P\left\{F_{1,n-k-1} > (n-k-1)\left(\exp\left(\frac{2\log\log(n)}{n}\right)-1\right)\right\}. \quad (15)$$

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

Now, we consider the SSE of a full model (k+1 variables) and a reduced model (k variables) where the full model and the reduced model are nested. The SSE(reduced)-SSE(full) follows the $\sigma^2\chi_1^2(\lambda)$ distribution since the variables are orthogonal and all these $\sigma^2\chi_1^2(\lambda)$ are independent. We denote the reduced model by k and the full model by k+1. In this case, $\frac{SSE_k - SSE_{k+1}}{\sigma^2}$ follows a χ_1^2 (possibly non-central) distribution since we only consider the overfit case and the reduced model contains every important variables ; a $\frac{SSE_{k+1}}{\sigma^2}$ follows central χ_{n-k-1}^2 since we assume that the k+1 full model includes every important variable. These χ^2 distributions are independent and form the basis of the F-distribution.

Table 1 presents probabilities using equations (11)-(15) 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 any 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. Table 2 presents probabilities assuming *i.i.d.*. F-tests. K denote the maximum possible model order (total number of variables plus the intercept) and k denote the model order. There are $K-k$ overfit candidate models from which to choose. 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 (11)-(15) 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. The probability of overfitting AIC decreases when the sample size is increased. The probability of overfitting AICc increases when the sample size is increased. The probabilities of overfitting SIC and HQ decrease when the sample size increased.

The probabilities of overfitting for most of these model selection criteria are large when the sample size is small. When K increases, probabilities are quite large due to the increase in candidate models. Unlike Table 1, Table 2 display the impact of K. 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.

5. Dependent F-tests overfitting probabilities

Consider the case of a model with m degrees of freedom (the reduced model). Suppose we want to consider adding one-additional variable from the choice of X_1, X_2 such that all variables X_0 in the reduced model, X_1 and X_2 are all orthogonal such that $X_j'X_j = n$. Let

$$F_1 = \frac{(X_1'Y)^2/n}{Y'(I_n - \frac{1}{n}X_0X_0' - \frac{1}{n}X_1X_1')Y/(m-1)} \sim F_{1,m-1}$$

be the F-test statistic for including X_1 and

$$F_2 = \frac{(X_2'Y)^2/n}{Y'(I_n - \frac{1}{n}X_0X_0' - \frac{1}{n}X_2X_2')Y/(m-1)} \sim F_{1,m-1}$$

be the F-test statistic for adding X_2. It can be shown that $E[F_1] = E[F_2] = (m-1)/(m-3)$ with variance,

$$var[F_1] = var[F_2] = 2(m-1)^2(m-2)/((m-5)(m-3)^2).$$

Let $A \sim \chi_1^2$, $B \sim \chi_1^2$ and $C \sim \chi_{m-2}^2$ be independent. Reduced model has m degrees of freedom and the nested full model has $m-1$ degrees of freedom. Then, F_1 has the same distribution as $(m-1)A/(C+B)$ and F_2 has the same distribution as $(m-1)B/(C+A)$. The correlation between F_1 and F_2 is the same as the correlation between $C_1 = A/(C+B)$ and $C_2 = B/(C+A)$. It can be shown that the joint distribution between C_1 and C_2 is

$$f(c_1, c_2)dc_1dc_2 = \frac{(m-2)(1-c_1c_2)^{(m/2)-3}}{2\pi\sqrt{c_1c_2}((1+c_1)(1+c_2))^{(m/2)-2}}dc_1dc_2, \quad (16)$$

$$c_1 > 0,\ c_2 > 0,\ c_1c_2 < 1.$$

Evaluating $E[c_1c_2]$ has no closed form but can be evaluated numerically. For correlation,

$m>5$ due to the full model having $m-1$ degrees of freedom. Table 3 presents correlation between F_1 and F_2. For all reduced model degrees of freedom, the strongest correlation is -0.0958 at $m=8$. As m, $m>8$, increases, the correlation goes to zero as F_1 and F_2 are asymptotically independent. This follow from the denominators of F_1 and $F_2 \rightarrow \sigma^2$ a.s. In general, the dependence is weak.

The probabilities of overfitting assuming independence tend to underestimate the probabilities of overfitting when the F-test dependence is included. Table 4 summarizes the multiple candidate model overfitting case using the dependence of the F-tests. Table 4 was constructed assuming that in the true model, the y_1 are $i.i.d.$ $N(\mu, \sigma^2)$ and all χ^2 distributions are central. The true model is the intercept only. For a randomly selected reduced model with k regressors, we have $K-k$ possible regressors left over. This yields $n-k$ degrees of freedom for the reduced model. A key assumption in Table 4 is that we have a random sample of remaining possible regressors. Thus, the $K-k$ χ_1^2 are independent.

Probabilities in Table 4 are produced when the F-tests are not indenpent. Table 5 presents probabilities of overfitting using order statistics from a model where none of the X_j are important to the model.

We found that the weak correlation has little impact on Table 4, and we use the probability of overfitting AIC decreases when the sample size is increased. The probability of overfitting of AICc increases when the sample size is increased. The probabilities of overfitting of SIC and HQ decrease when the sample size is increased. The probabilities of overfitting for most of these model selection criteria are large when the sample size is a small. SIC has a smaller probability when n is large and should have less overfitting than other criteria when the sample is large. SICc has smaller probabilities than SIC. When K increases, probabilities are quite large due to the increase in candidate models.

Selecting one of the 1-variable overfit models is equivalent to considering the maximum of the F-tests. Suppose there are $v=K-k$ extra variables. If $P\{F_{(v)}>f\}$ then an overfit model is selected. These F-test for adding one variable to the current model are not independent. As seen in comparing Tables 2 and 4, assuming independent F-tests leads to underestimating probabilities of overfitting. For criteria with small probability of overfitting, as useful upper bound is $P\{\text{overfit}\} \leq v\alpha$ where v is the number of extra variables and α is the probability from (11)-(15). For model selection criteria with weak penalty functions (α large), this bound is not useful.

Table 5 below considers the usual regression situation where the remaining regressors are dependent on the included regressors. Table 4 assumes independent χ_1^2 for the . remaining $K-k$ variables. These probabilities compare the current model with a candidate model that includes one additional variable. From selection 2, we showed that the variables are entered

into the model according to the size of their $(X_j'X)^2/n$, which have chi-square distributions. Consider the case where none of the variables are important to the model.. Here, all chi-square distributions are central and independent due to the orthogonality of the model. Now, probabilities of including one additional variable depend on the order statistic from $i.i.d.$ χ_1^2 random variables. The variable with corresponding to the largest order statistic is entered first. Once a variable has been entered into the model, the remaining variables are no longer independent χ_1^2 due to the ordering of the $(X_j'X)^2/n$. The distribution of the reduction in SSE for adding the best second variable does not have the same distribution as $K-2$ independent χ_1^2. Since the independence probabilities underestimate the probability of the maximum, the independence probabilities can bound the smaller order statistics. This can be seen in Table 5. Treating the orders k individually inflates the probabilities of overfitting. In Table 5, we can see that the probabilities of overfitttting decreases quickly for $k>1$. This is due to the ordering of the $(X_j'Y)^2/n$.

In general, adding additional variables follows from the order statistics of χ_1^2 random variables. Overfitting in orthogonal regression involves order statistics from $i.i.d.$ central χ_1^2. Assuming independence provides an upper bound for overfitting by more than 1 variable.

6. Conclusion

Usual comparisons of one reduced vs. one full model describe the basic behavior of a model selection criterion. Criteria with stronger penalty functions have smaller probabilities of overfitting. According to comparing of Table 2 and Table 4, the pattern of overfitting probabilities of model selection criteria is a little same. Also, the weak correlation has little impact on Table 4.

Assuming independence for these comparisons can be lead to overfitting the probability of overfitting due to the variables entering into the model according to their order statistics. Assuming independent comparisons makes model selection criteria with weak penalty functions appear to overfit much worse than they do in practice.

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

Table 3. Correlation (ρ) of two F-tests. By degrees of freedom for reduced model (df).

df	ρ	df	ρ	df	ρ
6	-0.0679	17	-0.0588	28	-0.0363
7	-0.0902	18	-0.0557	29	-0.0355
8	-0.0958	19	-0.0529	30	-0.0337
9	-0.0934	20	-0.0505	50	-0.0206
10	-0.0892	21	-0.0482	100	-0.0096
11	-0.0845	22	-0.0460	500	-0.0019
12	-0.0794	23	-0.0441	1000	-0.0011
13	-0.0746	24	-0.0421	10000	0.0005
14	-0.0710	25	-0.0410	100000	-0.0002
15	-0.0658	26	-0.0390	∞	0
16	-0.0622	27	-0.0379		

Table 4. Multiple candidate models case, no independence.

		K=6					K=11				
n	k	AIC	AICc	SIC	SICc	HQ	AIC	AICc	SIC	SICc	HQ
10	1	0.792	0.354	0.732	0.339	0.854	-	-	-	-	-
10	3	0.698	0.076	0.646	0.071	0.757	-	-	-	-	-
10	5	0.400	0.001	0.366	0.001	0.442	-	-	-	-	-
20	1	0.674	0.494	0.453	0.292	0.628	0.924	0.788	0.746	0.533	0.897
20	3	0.529	0.266	0.348	0.136	0.489	0.902	0.607	0.731	0.345	0.874
20	5	0.245	0.070	0.155	0.031	0.223	0.859	0.385	0.690	0.181	0.829
50	1	0.613	0.547	0.247	0.198	0.443	0.862	0.808	0.443	0.365	0.704
50	3	0.448	0.354	0.169	0.110	0.311	0.809	0.703	0.400	0.274	0.645
50	5	0.187	0.124	0.065	0.033	0.123	0.726	0.564	0.342	0.189	0.561
100	1	0.593	0.561	0.161	0.141	0.360	0.840	0.814	0.299	0.265	0.597
100	3	0.424	0.379	0.105	0.082	0.242	0.777	0.726	0.258	0.207	0.529
100	5	0.172	0.141	0.038	0.026	0.091	0.683	0.606	0.210	0.150	0.442
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

Table 5. Order statistics, conditional probabilities.

		K=6					K=11				
n	k	AIC	AICc	SIC	SICc	HQ	AIC	AICc	SIC	SICc	HQ
10	1	0.792	0.354	0.732	0.339	0.854	-	-	-	-	-
10	3	0.303	0.008	0.253	0.007	0.371	-	-	-	-	-
10	5	0.034	0.000	0.027	0.000	0.045	-	-	-	-	-
20	1	0.674	0.494	0.453	0.292	0.628	0.924	0.788	0.746	0.533	0.897
20	3	0.106	0.019	0.036	0.004	0.085	0.576	0.178	0.292	0.050	0.511
20	5	0.003	0.000	0.001	0.000	0.002	0.230	0.013	0.083	0.002	0.189
50	1	0.613	0.547	0.247	0.198	0.443	0.862	0.808	0.443	0.365	0.704
50	3	0.051	0.025	0.003	0.001	0.018	0.318	0.188	0.033	0.012	0.140
50	5	0.000	0.000	0.000	0.000	0.000	0.046	0.011	0.001	0.000	0.011
100	1	0.593	0.561	0.161	0.141	0.360	0.840	0.814	0.299	0.265	0.597
100	3	0.040	0.028	0.001	0.000	0.007	0.253	0.193	0.007	0.004	0.064
100	5	0.000	0.000	0.000	0.000	0.000	0.025	0.012	0.000	0.000	0.002
10000	1	0.575	0.575	0.012	0.012	0.164	0.786	0.786	0.022	0.021	0.275
10000	3	0.030	0.030	0.000	0.000	0.000	0.200	0.199	0.000	0.000	0.004
10000	5	0.000	0.000	0.000	0.000	0.000	0.012	0.012	0.000	0.000	0.000

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] David, H.A. (1981). *Order Statistics*. Wiley, New York.

[4] 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.

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

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

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

[8] McQuarrie, A.D., Shumway, R.H., and Tsai, C.L. (1997). The model selection criterion AICu. *Statistics & Probability Letters* 34, 285-292

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

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

[11] Shibata, R. (1980). Asymptotically efficient selection of the order of the model for estimating parameters of linear process. *The Annals of Statistics* 8, 147-164.

[12] Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* 68, 45-54.

[13] Shibata, R. (1986). Consistency of model selection and parameter estimation. In Essays in time series and applied processes. *Journal of Applied Probability* 23A, 127-141.