Journal of the Korean Statistical Society (1998), 27: 4, pp 397-405

A Note on Parametric Bootstrap Model Selection†

Kee-Won Lee[1] and Songyong Sim[2]

ABSTRACT

We develop parametric bootstrap model selection criteria in an example to fit a random sample to either a general normal distribution or a normal distribution with prespecified mean. We apply the bootstrap methods in two ways; one considers the direct substitution of estimated parameter for the unknown parameter, and the other focuses on the bias correction. These bootstrap model selection criteria are compared with AIC. We illustrate that all the selection rules reduce to the one sample t-test, where the cutoff points converge to some certain points as the sample size increases.

Keywords: AIC, Polygamma Function, T-test

1. INTRODUCTION

1.1. Problem Description

Suppose that we have a random sample $X_1, \ldots, X_n$ from a normal distribution with unknown mean and variance μ and σ^2 respectively. We wish to check whether the proposed model can be simplified by specifying the mean as $\mu = \mu_0$; a prespecified value.

Denote the probability density function of a normal distribution with parameter $\theta \in \Theta = \{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$ by $g(\cdot, \theta)$, and the distribution itself by $N(\mu, \sigma^2)$. Then, the problem reduces to the model selection between $\theta \in \Theta$ and $\theta \in \Theta_0$, where $\Theta_0 = \{(\mu, \sigma^2) : \mu = \mu_0, \sigma^2 > 0\}$.

---

†Part of the work was done while the first author visited Centre for Mathematics and its Applications, Australian National University, Canberra, ACT 0200, Australia.

[1]Department of Statistics, Chunchon, 200-702 S. Korea

This research was supported by Korea Science & Engineering Foundation 951-0103-032-2

[2]Department of Statistics, Hallym University, Chunchon, 200-702 S. Korea

1.2. Selection Criterion

We use the following measure of discrepancy, which is based upon Kullback-Leibler information quantity, as our model selection criterion to choose one distribution over another:

$$E\Big[-2nE\Big\{ \log g(Z, \hat{\theta}) \Big\}\Big], \tag{1.1}$$

where Z is independent and identically distributed with X_i's. Z is usually termed a future observation. $\hat{\theta}$ is the maximum likelihood estimator (MLE), which minimizes $-2\sum_{i=1}^{n} \log g(X_i, \theta)$. We select a model with the smallest value of (1.1). The inner expectation in (1.1) is taken with respect to the future observation Z, and the outer expectation with respect to $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$ to average out possible sampling variation. The employment of a future observation reflects the idea that the model selection should be based on the performance of the model over a new observation rather than what we currently have.

1.3. Asymptotic Approach

Akaike (1973) developed AIC to estimate (1.1) from an asymptotic approach. AIC is given by

$$-2\sum_{i=1}^{n} \log g(X_i, \hat{\theta}) + 2p, \tag{1.2}$$

where p is the number of parameters in the model.

An introductory review of AIC, including a motivation of AIC through Kullback Leibler information quantity and an explanation as to the historical reason why the mysterious number 2 appears in (1.1), can be found in Sakamoto, Ishiguro, and Kitagawa (1986, chap. 4).

1.4. Bootstrap Approach

A key idea in using the bootstrap method is to substitute a consistent estimator in place of the unknown parameter. The expectations in (1.1) are taken with respect to the true distribution which we believe generated the random sample. Therefore, the bootstrap sample should also be drawn from the estimated true distribution, in principle. See Linhart and Zucchini (1986, chap. 2) for a nonparametric bootstrap approach to model selection problems. Since we have two competing parametric distributions to choose from, the bootstrap samples should

be drawn from the two estimated distributions in a parametric way to assess the expected values.

The bootstrap method can be used to estimate (1.1) in two ways. In a classical plug-in bootstrap approach, we estimate (1.1) directly by plugging-in the estimated parameter in place of the unknown parameter, while in a refined one we focus on the bias correction of $-2\sum_{i=1}^{n}\log g(X_i,\hat{\theta})$ as an estimator of (1.1).

Efron and Tibshirani (1993, chap. 21) gives an overview of parametric bootstrap.

Naive Plug-in Bootstrap

In this approach (1.1) is estimated directly by substituting $\hat{\theta}$ for θ in (1.1) as follows:

Step 1. Draw a bootstrap sample $X_1^*,\ldots,X_n^*$ from the fitted normal distribution with estimated parameter $\hat{\theta}$.

Step 2. Compute the bootstrap version of $\hat{\theta}$ which minimizes $-2\sum_{i=1}^{n}\log g(X_i^*,\theta)$ Denote it by $\hat{\theta}^*$.

Step 3. Evaluate the bootstrap version of (1.1);

$$E^*\big[-2nE\big\{\log g(Z^*,\hat{\theta}^*)\big\}\big], \tag{1.3}$$

where Z^* is a future observation of the bootstrap sample.

Bias Corrected Bootstrap

Shortly, we will check that (1.3) still has a downward bias of amount roughly equal to the number of parameters, one major reason being that the bias of the obvious estimate is still large relative to its standard error.

A refined bootstrap approach focuses on the bias correction. The bias of $-2\sum_{i=1}^{n}\log g(X_i,\hat{\theta})$ as an estimator of (1.1) is given by

$$\text{bias}(\theta)=E\Big\{-2\sum_{i=1}^{n}\log g(X_i,\hat{\theta})\Big\}-E\Big[-2nE\big\{\log g(Z,\hat{\theta})\big\}\Big]. \tag{1.4}$$

In fact, AIC uses $-2p$ as an estimator of (1.4) from an asymptotic approach. We obtain the bias corrected bootstrap estimator of (1.1) by following the steps in the Naive plug-in bootstrap approach with a minor modification at the final stage.

Step3'. Evaluate the bootstrap estimator of (1.4);

$$\text{bias}(\hat{\theta}) = E^*\Big\{ -2\sum_{i=1}^{n} \log g(X_i^*, \hat{\theta}^*) \Big\} - E^*\Big[-2nE\big\{ \log g(Z^*, \hat{\theta}^*) \big\} \Big], \tag{1.5}$$

and add it to $-2\sum_{i=1}^{n} \log g(X_i, \hat{\theta})$.

2. SELECTION CRITERIA

First, we need the followng facts about the polygamma function on which statistical properties of naive bootstrap selection criteria depend heavily.

2.1. The polygamma function

Polygamma functions are defined as the derivatives of $\log\Gamma(x)$. In general, the s-gamma function is defined as $d^{s-1}\log\Gamma(x)/dx^{s-1} = \Psi^{(s-2)}(x)$ for $s \geq 2$, where $\Psi(x) = d\log\Gamma(x)/dx$ is the digamma function. See Abramowitz and Stegun (1968) for more properties.

This special function appears in statistical literature when we express higher order moments and cumulants of a natural logarithm of the gamma distribution. In particular, if we denote a chi-square distribution with ν degrees of freedom by χ^2_ν, then the cumulant generating function of $\log\chi^2_\nu$ is given by $t\log 2 + \log\Gamma(t+\nu/2) - \log\Gamma(\nu/2)$. From this, we can compute the mean and variance of $\log\chi^2_\nu$ as

$$\begin{aligned} E\big(\log\chi^2_\nu\big) &= \log 2 + \Psi(\nu/2), \\ Var\big(\log\chi^2_\nu\big) &= \Psi'(\nu/2). \end{aligned} \tag{2.1}$$

Using the expansion formula in Abramowitz and Stegun (1968), we can approximate the digamma and the trigamma function as

$$\begin{aligned} \Psi(\nu) &= \log\nu - 1/(2\nu) + O(\nu^{-2}), \\ \Psi'(\nu) &= (\nu - 1/2)^{-1} + O(\nu^{-3}), \end{aligned} \tag{2.2}$$

for large ν. See Johnson and Kotz (1970) for further applications of the polygamma function in statistical sciences.

2.2. Key Notations and Formulae

For $\theta \in \Theta_0$, $\hat{\theta} = (\mu_0, \tilde{\sigma}^2)$ with $\tilde{\sigma}^2 = \sum_{i=1}^{n}(X_i - \mu_0)^2/n$. Draw a bootstrap sample $X_1^*, \ldots, X_n^*$ from $N(\mu_0, \tilde{\sigma}^2)$. The bootstrap MLE $\hat{\theta}^* = (\mu_0, \tilde{\sigma}^{*2})$ with

$\tilde{\sigma}^{*2} = \sum_{i=1}^{n}(X_i^* - \mu_0)^2/n$. We have

$$\begin{array}{rcl} -2\sum_{i=1}^{n} \log g(X_i, \hat{\theta}) & = & n \log 2\pi\tilde{\sigma}^2 + n, \\ -2\sum_{i=1}^{n} \log g(X_i^*, \hat{\theta}^*) & = & n \log 2\pi\tilde{\sigma}^{*2} + n. \end{array} \tag{2.3}$$

For $\theta \in \Theta$, $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$ with $\hat{\mu} = \sum_{i=1}^{n} X_i/n$ and $\hat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \hat{\mu})^2/n$. Draw a bootstrap sample $X_1^*, \ldots, X_n^*$ from $N(\hat{\mu}, \hat{\sigma}^2)$. The bootstrap MLE $\hat{\theta}^* = (\hat{\mu}^*, \hat{\sigma}^{*2})$ with $\hat{\mu}^* = \sum_{i=1}^{n} X_i^*/n$ and $\hat{\sigma}^{*2} = \sum_{i=1}^{n}(X_i^* - \hat{\mu}^*)^2/n$. We have

$$\begin{array}{rcl} -2\sum_{i=1}^{n} \log g(X_i, \hat{\theta}) & = & n \log 2\pi\hat{\sigma}^2 + n, \\ -2\sum_{i=1}^{n} \log g(X_i^*, \hat{\theta}^*) & = & n \log 2\pi\hat{\sigma}^{*2} + n. \end{array} \tag{2.4}$$

Note that $n\tilde{\sigma}^{*2}/\tilde{\sigma}^2$ and $n\hat{\sigma}^{*2}/\hat{\sigma}^2$ are chi-square random variables with n and $(n-1)$ degrees of freedom respectivey.

From (2.3) and (2.4), AIC's are given by

$$\begin{array}{ll} n \log 2\pi\tilde{\sigma}^2 + n + 2 & \text{for } \theta \in \Theta_0, \\ n \log 2\pi\hat{\sigma}^2 + n + 4 & \text{for } \theta \in \Theta. \end{array} \tag{2.5}$$

2.3. Bootstrap Approach

Naive Plug-in Bootstrap

For $\theta = (\mu_0, \sigma^2)$, we have

$$\begin{array}{rcl} -2nE\{\log g(Z^*, \hat{\theta}^*)\} & = & n\big[\log 2\pi\tilde{\sigma}^{*2} + E\{(Z^* - \mu_0)^2\}/\tilde{\sigma}^{*2}\big] \\ & = & n(\log 2\pi\tilde{\sigma}^{*2} + \tilde{\sigma}^2/\tilde{\sigma}^{*2}). \end{array} \tag{2.6}$$

Taking expectation with respect to $N(\mu_0, \tilde{\sigma}^2)$, (1.3) reduces to

$$n\{\log 2\pi\tilde{\sigma}^2 + \Psi(n/2) - \log(n/2) + n/(n-2)\}, \tag{2.7}$$

from (2.1). For $\theta = (\mu, \sigma^2)$, we have

$$\begin{array}{rcl} -2nE\{\log g(Z^*, \hat{\theta}^*)\} & = & n\big[\log 2\pi\hat{\sigma}^{*2} + E\{(Z^* - \hat{\mu}^*)^2\}/\hat{\sigma}^{*2}\big] \\ & = & n\{\log 2\pi\hat{\sigma}^{*2} + (1 + 1/n)\hat{\sigma}^2/\hat{\sigma}^{*2}\}. \end{array} \tag{2.8}$$

Taking expectation with respect to $N(\hat{\mu}, \hat{\sigma}^2)$, (1.3) reduces to

$$n\{\log 2\pi\hat{\sigma}^2 + \Psi\big((n-1)/2\big) - \log(n/2) + (n+1)/(n-3)\}, \tag{2.9}$$

from (2.1). Note that the final forms of the naive plug-in bootstrap selection criteria do not depend on any particular realization of a bootstrap sample. From

the approximation formula (2.2), we can check that the naive plug-in bootstrap selection criteria are roughly equivalent to AIC $- p$. That is, this approach leads to a seriously biased estimated selection criterion. Refer Chung, et. al. (1996) for a more general dicussion of this phenomenon.

Bias Corrected Bootstrap

For $\theta \in \Theta_0$, the bootstrap bias estimation (1.5) reduces to

$$E^*\{n(1 - \tilde{\sigma}^2/\tilde{\sigma}^{*2})\} = -2n/(n-2)$$

by (2.3) and (2.6). Similarly for $\theta \in \Theta$, (1.5) reduces to

$$E^*\{n - (n+1)\hat{\sigma}^2/\hat{\sigma}^{*2}\} = -4n/(n-3)$$

by (2.4) and (2.8). Therefore, the bias corrected bootstrap selection criteria are given by

$$\begin{array}{ll} n\log 2\pi\tilde{\sigma}^2 + n + 2n/(n-2) & \text{for } \theta \in \Theta_0, \\ n\log 2\pi\hat{\sigma}^2 + n + 4n/(n-3) & \text{for } \theta \in \Theta. \end{array} \tag{2.10}$$

Refer Lee and Sim (1996) for a more general discussion.

2.4. Comparison with T-test

Let $T_{n-1} = (n-1)^{1/2}(\hat{\mu} - \mu_0)/\hat{\sigma}$ be the Student t-test statistic. We can check that AIC selects $\theta \in \Theta$ if $n\log\tilde{\sigma}^2/\hat{\sigma}^2 > 2$, which simplifies to

$$|T_{n-1}| > (n-1)^{1/2}\{\exp(2/n) - 1\}^{1/2} = \sqrt{2} + O(n^{-2}).$$

Naive plug-in bootstrap approach selects $\theta \in \Theta$ if $|T_{n-1}|$ is greater than

$$(n-1)^{1/2}\big[\exp\big\{-\Psi(n/2) + \Psi\big((n-1)/2\big) - 2/(n-2) + 4/(n-3)\big\} - 1\big]^{1/2},$$

which can be approximated as $1 + 2/(n-3) + O(n^{-2})$ by the formula (2.2). Bias corrected bootstrap approach selects $\theta \in \Theta$ if

$$\begin{array}{rcl} |T_{n-1}| & > & (n-1)^{1/2}\big[\exp\big\{2(n-1)/\big((n-2)(n-3)\big) - 1\big\}\big]^{1/2} \\ & = & \sqrt{2} + O(n^{-2}). \end{array}$$

It is interesting to note that the cutoff points for AIC and the bias corrected bootstrap converge to $\sqrt{2}$ from below and from above respectively, while the cutoff points for the naive plug-in bootstrap converge to 1 from above. Table 2.1 presents cutoff points and corresponding levels of the two-sided t-test.

Table 2.1: Cutoff points and corresponding levels of two-sided t-test for AIC, Naive plug-in bootstrap, and bias corrected bootstrap. Bias corrected bootstrap approach is denoted by Bootstrap-a, and naive plug-in bootstrap by Bootstrap-b. Cutoff points are denoted by c_n and the levels are denoted by α.

	AIC		Bootstrap-a		Bootstrap-b	
n	c_n	α	c_n	α	c_n	α
5	1.4026	0.2334	3.3429	0.0288	2.7321	0.0523
10	1.4116	0.1916	1.8471	0.0978	1.4278	0.1871
20	1.4136	0.1737	1.5850	0.1295	1.1752	0.2544
100	1.4142	0.1604	1.4435	0.1520	1.0309	0.3051

The selection procedures based on these criteria lead to a very conservative decision and therefore should be used with caution.

3. BIAS AND VARIANCE UNDER $\theta \in \Theta_0$

What if, in fact, $N(\mu_0, \sigma^2)$ is the true model? Then, we can compute (1.1) exactly, and use it to assess the relative performance of AIC and its bootstrap competitors. All the selection criteria for $\theta \in \Theta_0$ have $n \log 2\pi\tilde{\sigma}^2$ in common, and those criteria for $\theta \in \Theta$ have $n \log 2\pi\hat{\sigma}^2$ in common. Under $\theta \in \Theta_0$, the expected values and variances of these common terms are given by

$$\begin{aligned} &E(n \log 2\pi\tilde{\sigma}^2) = n\{ \log 2\pi\sigma^2 + \Psi(n/2) - \log(n/2)\}, \\ &E(n \log 2\pi\tilde{\sigma}^2) = n\{ \log 2\pi\sigma^2 + \Psi\big((n-1)/2\big) - \log(n/2)\}, \\ &Var(n \log 2\pi\tilde{\sigma}^2) = n^2\Psi'(n/2), \\ &Var(n \log 2\pi\tilde{\sigma}^2) = n^2\Psi'\big((n-1)/2\big), \end{aligned} \tag{3.1}$$

by the fact that $n\tilde{\sigma}^2/\sigma^2$ and $n\hat{\sigma}^2/\sigma^2$ are chi-square random variables with n and $(n-1)$ degrees of freedom respectively under $\theta \in \Theta_0$ and the formula (2.1). From (3.1), (1.1) can be evaluated as

$$\begin{aligned} &n\{ \log 2\pi\sigma^2 + \Psi(n/2) - \log(n/2) + n/(n-2)\} && \text{for } \theta \in \Theta_0, \\ &n\{ \log 2\pi\sigma^2 + \Psi\big((n-1)/2\big) - \log(n/2) + (n+1)/(n-3)\} && \text{for } \theta \in \Theta. \end{aligned}$$

Expected values of AIC are

$$\begin{array}{ll} n\{\log 2\pi\sigma^2 + \Psi(n/2) - \log(n/2) + (n+2)/n\} & \text{for } \theta \in \Theta_0, \\ n\{\log 2\pi\sigma^2 + \Psi\big((n-1)/2\big) - \log(n/2) + (n+4)/n\} & \text{for } \theta \in \Theta. \end{array}$$

Hence the biases of AIC are $-4/(n-2)$ for $\theta \in \Theta_0$ and $-12/(n-4)$ for $\theta \in \Theta$. The biases of naive plug-in bootstrap selection criteria reduce to

$$\begin{array}{ll} E\big(n\log\tilde{\sigma}^2/\sigma^2\big) = n\big\{\Psi(n/2) - \log(n/2)\big\} = -1 + O(n^{-1}) & \text{for } \theta \in \Theta_0, \\ E\big(n\log\hat{\sigma}^2/\sigma^2\big) = n\big\{\Psi\big((n-1)/2\big) - \log(n/2)\big\} = -2 + O(n^{-1}) & \text{for } \theta \in \Theta, \end{array}$$

which means that the naive plug-in bootstrap approach still has a downward bias of amount roughly equal to the number of parameters in the model.

Most notably the *bias corrected bootstrap selection criteria are unbiased*. Since the variances are all the same for each selection critria, we may conclude that the bias corrected boostrap approach gives the best overall performance.

References

Abramowitz, M., and Stegun, I. A. (eds)(1968). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Washington, U.S. Goverment Printing Office.

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle", in *Proceedings of the 2nd International Symposium on Information Theory*, eds. Petrov, B. N., and Csáki, F., Akademiai Kiádo, 267–281.

Chung, H-Y, Lee, K-W, and Koo, J-Y (1996). "A note on bootstrap model selection criterion", *Statistics and Probability Letters*, **26**, 35–41.

Efron, B., and R. J. Tibshirani (1993). *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Johnson, N. L., and Kotz, S. (1970). *Distributions in Statistics: Continuous Univariate Distributions—2*, New York: John Wiley.

Lee, K-W and Sim, S. (1996). "On the bias of bootstrap model selection criteria", *Journal of the Korean Statistical Society*, **25-2**, 195–204.

Linhart, H., and Zucchini, W. (1986). *Model Selection*, New York: John Wiley.

Sakamoto, Y., Ishiguro, M., and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*, KTK Scientific Publishers: Tokyo.