Journal of Advances in Mathematics and Computer Science

34(2): 1-12, 2019; Article no.JAMCS.52709

ISSN: 2456-9968

(Past name: British Journal of Mathematics & Computer Science, Past ISSN: 2231-0851)

Testing the Fairness of a Coin by Akaike's Information Criterion

Kunio Takezawa[1*]

[1] *Division of Informatics and Inventory, Institute for Agro-Environmental Sciences, National Agriculture and Food Research Organization, Kannondai 3-1-3, Tsukuba, Ibaraki 305-8604, Japan.*

Author's contribution

The sole author designed, analysed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/JAMCS/2019/v34i230212

Editor(s):

(1) Dr. Dragos-Patru Covei, Professor, Department of Applied Mathematics, The Bucharest University of Economic Studies, Romania.

(2) Dr. Tian-Xiao He, Professor, Department of Mathematics, Illinois Wesleyan University, USA.

Reviewers:

(1) Xingting Wang, Howard University College of Arts and Sciences, USA.

(2) Peter Stallinga, University of Algarve, Portugal.

(3) Janilson Pinheiro de Assis, Federal Rural University of the Semi-arid Region, Brazil.

(4) Myron Hlynka, University of Windsor, Canada.

(5) A. Ayeshamariam, Khadir Mohideen College, India.

Complete Peer review History: http://www.sdiarticle4.com/review-history/52709

Original Research Article

Received: 22 August 2019

Accepted: 24 October 2019

Published: 26 October 2019

Abstract

In this paper, *AIC* (Akaike's Information Criterion) is used to judge whether a coin is biased or not using the sequence of heads and tails produced by tossing the coin several times. It is well known that $AIC \cdot (-0.5)$ is an efficient estimator of the expected log-likelihood when the true distribution is contained in a specified parametric model. In the coin tossing problem, however, $AIC \cdot (-0.5)$ works as an efficient estimator even if the true distribution is not contained in a specified parametric model. Moreover, the judgement of fairness of coin using *AIC* is equivalent to a statistical test using the Bernoulli distribution with a significance level ranging from 11% to 18%. This indicates that the judgement of the fairness of coin based on *AIC* leads to a higher probability of type I errors than that given by a statistical test with a significance level of 5%. These findings show that we judge the fairness of a coin based on *AIC* when we do not have any prior knowledge about its fairness and we want to judge it from the standpoint of prediction.

**Corresponding author: E-mail: nonpara@gmail.com, takezawa@affrc.go.jp;*

In contrast, a statistical test with a significance level of 5% is adopted when we have prior knowledge that the coin is probably unbiased. Moreover, a statistical test with a 5% significance level allows us to conclude that the coin is biased if we obtain sufficient evidence that permits us to disbelieve the prior knowledge.

Keywords: Akaike's Information Criterion; coin tossing; log-likelihood; future data; predictive estimator; maximum likelihood estimator.

2010 Mathematics Subject Classification: 60G25, 62F10, 62M20.

1 Introduction

AIC (Akaike's Information Criterion) is widely used as a statistic for statistical tests and model selection. This is because $AIC\cdot(-0.5)$ is thought to be an approximation of the expected log-likelihood for universal purposes (e.g. [1]). However, the general derivation of AIC (Section 3 of [2]) assumes that the data satisfy highly restrictive conditions so that $AIC\cdot(-0.5)$ can be regarded as an approximation of the expected log-likelihood. Hence, when we do not know whether the available data satisfy such conditions, $AIC\cdot(-0.5)$ might not be a good approximation of the expected log-likelihood. In fact, when we handle a simple problem that chooses between an exponential distribution and Weibull distribution, $AIC\cdot(-0.5)$ cannot be used as an approximation of the expected log-likelihood in most situations ([3]). Nevertheless, we rarely ascertain that the data at hand satisfy such conditions when we use AIC in practice. One of the reasons for this tendency seems to be that even if $AIC\cdot(-0.5)$ is not an approximation of the expected log-likelihood, AIC works well for the purpose of model selection because it behaves like GCV (Generalized Cross-Validation, [4]), as shown on page 242 of [5]. However, when we use AIC as a model selection criterion on the grounds that $AIC\cdot(-0.5)$ is a good approximation of the expected log-likelihood, we need to confirm that the available data satisfy the conditions for using $AIC\cdot(-0.5)$ as an approximation of the expected log-likelihood. If we cannot confirm that these conditions hold, we can hardly argue the value of AIC as a tool of model selection. Moreover, if $AIC\cdot(-0.5)$ is not an approximation of the expected log-likelihood, Akaike Weights ([6]; Section 7.2 of [7]; Section 2.9 of [1]) do not make sense.

Among such problems, we take the simple example of coin tossing, in which a coin is tossed several times to see whether it lands heads up or tails up. The resultant data are used to judge whether the coin is biased or not on the basis of the Bernoulli distribution. For this problem, we investigate whether or not $AIC\cdot(-0.5)$ works as an approximation of the expected log-likelihood from the perspective of analytical methods and numerical simulations. Then, the results of such analysis are compared with a statistical test with a significance level of 5%.

2 AIC for the Coin Tossing Problem

The heads or tails probability as a result of coin tossing obeys the Bernoulli distribution. The probability mass function of the Bernoulli distribution is represented as

$$f(x_i|\theta) = \theta^{x_i}(1-\theta)^{(1-x_i)}, \tag{2.1}$$

where x_i takes a value of either 0 or 1. Here, $x_i = 1$ indicates 'heads', while $x_i = 0$ indicates 'tails'; θ is the parameter of Bernoulli distribution; $f(1|\theta)$ denotes the probability of landing heads; and $f(0|\theta)$ denotes the probability of landing tails. The number of trials is denoted as n. Then, the result of the trials is written as $\mathbf{x} = (x_1, x_2, x_3, \cdots, x_n)$.

A hypothesis test is performed using $\mathbf{x}$, and the true value of θ is represented as θ_0. In this setting, the null hypothesis is

H_0: The coin is fair (i.e. not biased). That is, $\theta_0 = 0.5$.

The alternative hypothesis is

H_1: The coin is biased. That is, $\theta_0 \neq 0.5$.

We assume that we have the data $\mathbf{x}$. The log-likelihood for these data is

$$l(\theta|\mathbf{x}) = \sum_{i=1}^{n} \log\Big(f(x_i|\theta)\Big) = \sum_{i=1}^{n} \Big(x_i \log(\theta) + (1 - x_i)\log(1 - \theta)\Big). \tag{2.2}$$

Differentiation of $l(\theta|\mathbf{x})$ with respect to θ gives

$$\frac{\partial l(\theta|\mathbf{x})}{\partial \theta} = \frac{1}{\theta(1-\theta)} \sum_{i=1}^{n} (x_i - \theta). \tag{2.3}$$

When this equation is set to 0, we obtain

$$\hat{\theta}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{2.4}$$

Here, $\hat{\theta}(\mathbf{x})$ is the maximum likelihood estimator when $\mathbf{x}$ is our data. Then, the log-likelihood of $\hat{\theta}(\mathbf{x})$ in the light of $\mathbf{x}$ is written as

$$l(\hat{\theta}(\mathbf{x})|\mathbf{x}) = \sum_{i=1}^{n} \log\Big(f(x_i|\hat{\theta}(\mathbf{x}))\Big) = \log\Big(f(\mathbf{x}|\hat{\theta}(\mathbf{x}))\Big). \tag{2.5}$$

Next, future data are set as $\mathbf{x}^* = (x_1^*, x_2^*, \ \ldots, x_n^*)$. The log-likelihood of $\hat{\theta}(\mathbf{x})$ in the light of $\mathbf{x}^*$ is depicted as

$$l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*) = \sum_{i=1}^{n} \log\Big(f(x_i^*|\hat{\theta}(\mathbf{x}))\Big) = \log\Big(f(\mathbf{x}^*|\hat{\theta}(\mathbf{x}))\Big). \tag{2.6}$$

When the expectation with respect to the future data ($\mathbf{x}^*$) is represented as $E_{\{\mathbf{x}^*\}}\Big[\cdot\Big]$, a large value of $E_{\{\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big]$ indicates that the estimator $\hat{\theta}(\mathbf{x})$ fits well to future data. Although $\hat{\theta}(\mathbf{x})$ is guaranteed to fit well to the available data, it is not guaranteed to fit closely to future data. However, because fitting well to future data is the most important property when using a model for practical purposes, we should estimate the value of $E_{\{\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big]$.
Then, we assume

$$l(\hat{\theta}(\mathbf{x})|\mathbf{x}) = E_{\{\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big] + b. \tag{2.7}$$

That is, the remainder of subtraction of the log-likelihood of $\hat{\theta}(\mathbf{x})$ in the light of future data from the log-likelihood of $\hat{\theta}(\mathbf{x})$ in the light of available data is 'b'. Because $l(\hat{\theta}(\mathbf{x})|\mathbf{x})$ is calculated using Eq.(2.5), derivation of b yields the value of $E_{\{\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big]$. However, we have to prepare an infinite number of future data ($\mathbf{x}^*$) to obtain the value of b. Then, the expectation of both sides of Eq.(2.8) with respect to available data ($\mathbf{x}$) is taken. This yields

$$E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big] = E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big] + b, \tag{2.8}$$

where $E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[\cdot\Big]$ indicates the expectation with respect to both of $\mathbf{x}$ and $\mathbf{x}^*$. Expectation $E_{\{\mathbf{x}\}}\Big[\cdot\Big]$ denotes the expectation with respect to $\mathbf{x}$. Hence, b is written as

$$b = E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big] - E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big], \tag{2.9}$$

where both $E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big]$ and $E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big]$ are functions of θ_0 and are nonrandom variables. Therefore, b is also a function of θ_0 and is a nonrandom variable. Because the value of θ_0 is usually unknown, the value of b is also unknown if b depends upon θ_0. However, if b is independent of θ_0 (it needs some specific conditions), b is regarded as a constant independent of θ_0. In such a situation, if the value of this constant is derived, the value of $E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big]$ is estimated because $l(\hat{\theta}(\mathbf{x})|\mathbf{x})$ (where $\mathbf{x}$ is the available data) can be used as an approximation of $E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big]$ in Eq.(2.8); this procedure does not need the value of θ_0. The usefulness of AIC is based on this principle.

Variable b (Eq.(2.10)) is estimated below along the lines of Section 3 of [2]. First, we decompose b as follows.

$$\begin{aligned} b &= E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big] - E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big] \\ &= E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big] - E_{\{\mathbf{x}\}}\Big[l(\theta_0|\mathbf{x})\Big] \\ &\quad + E_{\{\mathbf{x}\}}\Big[l(\theta_0|\mathbf{x})\Big] - E_{\{\mathbf{x}^*\}}\Big[l(\theta_0|\mathbf{x}^*)\Big] \\ &\quad + E_{\{\mathbf{x}^*\}}\Big[l(\theta_0|\mathbf{x}^*)\Big] - E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big], \end{aligned} \tag{2.10}$$

where $E_{\{\mathbf{x}^*\}}\Big[\cdot\Big]$ indicates the expectation with respect to $\mathbf{x}^*$.

Next, we set

$$D_1 = E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big] - E_{\{\mathbf{x}\}}\Big[l(\theta_0|\mathbf{x})\Big], \tag{2.11}$$

$$D_2 = E_{\{\mathbf{x}\}}\Big[l(\theta_0|\mathbf{x})\Big] - E_{\{\mathbf{x}^*\}}\Big[l(\theta_0|\mathbf{x}^*)\Big], \tag{2.12}$$

$$D_3 = E_{\{\mathbf{x}^*\}}\Big[l(\theta_0|\mathbf{x}^*)\Big] - E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big]. \tag{2.13}$$

Then, Eq.(2.10) leads to

$$b = D_1 + D_2 + D_3. \tag{2.14}$$

A Taylor expansion of D_1 around $\hat{\theta}(\mathbf{x})$ yields

$$\begin{aligned} D_1 &= E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big] - E_{\{\mathbf{x}\}}\Big[l(\theta_0|\mathbf{x})\Big] \\ &\approx E_{\{\mathbf{x}\}}\Big[-(\theta_0 - \hat{\theta}(\mathbf{x}))\frac{\partial l(\hat{\theta}(\mathbf{x})|\mathbf{x})}{\partial \hat{\theta}} - \frac{1}{2}(\theta_0 - \hat{\theta}(\mathbf{x}))^2\frac{\partial^2 l(\hat{\theta}(\mathbf{x})|\mathbf{x})}{\partial \hat{\theta}^2}\Big] \\ &\approx E_{\{\mathbf{x}\}}\Big[-\frac{1}{2}(\theta_0 - \hat{\theta}(\mathbf{x}))^2\frac{\partial^2 l(\hat{\theta}(\mathbf{x})|\mathbf{x})}{\partial \hat{\theta}^2}\Big], \end{aligned} \tag{2.15}$$

where the following equation is used on the basis that $\hat{\theta}(\mathbf{x})$ is the maximum likelihood estimator.

$$\frac{\partial l(\hat{\theta}(\mathbf{x})|\mathbf{x})}{\partial \hat{\theta}} = 0. \tag{2.16}$$

Differentiation of Eq.(2.3) with respect to θ gives

$$\frac{\partial^2 l(\theta|\mathbf{x})}{\partial \theta^2} = \Big(-\frac{1}{(1-\theta)\theta^2} + \frac{1}{(1-\theta)^2\theta}\Big)\sum_{i=1}^{n}(x_i - \theta) - \frac{n}{\theta(1-\theta)}. \tag{2.17}$$

Then, if we set $\theta = \hat{\theta}(\mathbf{x})$, the equation below is obtained.

$$\begin{aligned} \frac{\partial^2 l(\hat{\theta}(\mathbf{x})|\mathbf{x})}{\partial \hat{\theta}^2} &= \Big(-\frac{1}{(1-\hat{\theta}(\mathbf{x}))\hat{\theta}(\mathbf{x})^2} + \frac{1}{(1-\hat{\theta}(\mathbf{x}))^2\hat{\theta}(\mathbf{x})}\Big)\sum_{i=1}^{n}(x_i - \hat{\theta}(\mathbf{x})) - \frac{n}{\hat{\theta}(\mathbf{x})(1-\hat{\theta}(\mathbf{x}))} \\ &= -\frac{n}{\hat{\theta}(\mathbf{x})(1-\hat{\theta}(\mathbf{x}))}, \end{aligned} \tag{2.18}$$

where Eq.(2.4) is used.

Substitution of Eq.(2.18) into Eq.(2.15) leads to

$$E_{\{\mathbf{x}\}}\Big[-\frac{1}{2}(\theta_0-\hat{\theta}(\mathbf{x}))^2\frac{\partial^2 l(\hat{\theta}(\mathbf{x})|\mathbf{x})}{\partial\hat{\theta}^2}\Big]=\frac{n}{2}E_{\{\mathbf{x}\}}\Big[(\theta_0-\hat{\theta}(\mathbf{x}))^2\frac{1}{\hat{\theta}(\mathbf{x})(1-\hat{\theta}(\mathbf{x}))}\Big]. \tag{2.19}$$

When $\hat{\theta}(\mathbf{x})$ is close to the true value, that is, $\hat{\theta}(\mathbf{x})\approx\theta_0$ holds, the following equation is derived:

$$\frac{1}{\hat{\theta}(\mathbf{x})(1-\hat{\theta}(\mathbf{x}))}\approx\frac{1}{\theta_0(1-\theta_0)}. \tag{2.20}$$

Substitution of Eq.(2.20) into Eq.(2.19) turns out to be

$$E_{\{\mathbf{x}\}}\Big[-\frac{1}{2}(\theta_0-\hat{\theta}(\mathbf{x}))^2\frac{\partial^2 l(\hat{\theta}(\mathbf{x})|\mathbf{x})}{\partial\hat{\theta}^2}\Big]\approx\frac{n}{2\theta_0(1-\theta_0)}E_{\{\mathbf{x}\}}\Big[(\theta_0-\hat{\theta}(\mathbf{x}))^2\Big]. \tag{2.21}$$

Note that the equation below holds.

$$\begin{aligned}
E_{\{\mathbf{x}\}}\Big[(\theta_0-\hat{\theta}(\mathbf{x}))^2\Big] &= E_{\{\mathbf{x}\}}\Big[\Big(\theta_0-\frac{\sum_{i=1}^n x_i}{n}\Big)^2\Big]\\
&= \theta_0^2-\frac{2\theta_0}{n}E_{\{\mathbf{x}\}}\Big[\sum_{i=1}^n x_i\Big]+\frac{1}{n^2}E_{\{\mathbf{x}\}}\Big[\Big(\sum_{i=1}^n x_i\Big)^2\Big]\\
&= \theta_0^2-\frac{2\theta_0}{n}E_{\{\mathbf{x}\}}\Big[\sum_{i=1}^n x_i\Big]+\frac{1}{n^2}E_{\{\mathbf{x}\}}\Big[\sum_{i=1}^n x_i^2\Big]+\frac{1}{n^2}E_{\{\mathbf{x}\}}\Big[\sum_{i,j=1(i\neq j)}^n x_i x_j\Big]\\
&= \theta_0^2-2\theta_0^2+\frac{\theta_0}{n}+\frac{\theta_0^2(n^2-n)}{n^2}\\
&= \frac{\theta_0(1-\theta_0)}{n},
\end{aligned} \tag{2.22}$$

where $\sum_{i,j=1(i\neq j)}^n$ indicates the summation with respect to both i and j except for the cases when $i=j$. Substitution of Eq.(2.22) into Eq.(2.21) leads to

$$D_1\approx E_{\{\mathbf{x}\}}\Big[-\frac{1}{2}(\theta_0-\hat{\theta})^2\frac{\partial^2 l(\hat{\theta}(\mathbf{x})|\mathbf{x})}{\partial\hat{\theta}^2}\Big]\approx\frac{1}{2}. \tag{2.23}$$

Because $\mathbf{x}$ and $\mathbf{x}^*$ are samples from the same population, Eq.(2.12) becomes

$$D_2=E_{\{\mathbf{x}\}}\Big[l(\theta_0|\mathbf{x})\Big]-E_{\{\mathbf{x}^*\}}\Big[l(\theta_0|\mathbf{x}^*)\Big]=0. \tag{2.24}$$

The Taylor expansion of D_3(Eq.(2.13)) around θ_0 provides

$$\begin{aligned}
D_3 &= E_{\{\mathbf{x}^*\}}\Big[l(\theta_0|\mathbf{x}^*)\Big]-E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big]\\
&\approx E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[-(\hat{\theta}(\mathbf{x})-\theta_0)\frac{\partial l(\theta_0|\mathbf{x}^*)}{\partial\theta_0}-\frac{1}{2}(\hat{\theta}(\mathbf{x})-\theta_0)^2\frac{\partial^2 l(\theta_0|\mathbf{x}^*)}{\partial\theta_0^2}\Big]\\
&\approx E_{\{\mathbf{x}\}}\Big[-(\hat{\theta}(\mathbf{x})-\theta_0)\Big]E_{\{\mathbf{x}^*\}}\Big[\frac{\partial l(\theta_0|\mathbf{x}^*)}{\partial\theta_0}\Big]-E_{\{\mathbf{x}\}}\Big[\frac{1}{2}(\hat{\theta}(\mathbf{x})-\theta_0)^2\Big]E_{\{\mathbf{x}^*\}}\Big[\frac{\partial^2 l(\theta_0|\mathbf{x}^*)}{\partial\theta_0^2}\Big].
\end{aligned} \tag{2.25}$$

Equation (2.3) leads to

$$E_{\{\mathbf{x}^*\}}\Big[\frac{\partial l(\theta_0|\mathbf{x}^*)}{\partial\theta_0}\Big]=\frac{1}{\theta_0(1-\theta_0)}E_{\{\mathbf{x}^*\}}\Big[\sum_{i=1}^n(x_i^*-\theta_0)\Big]=0. \tag{2.26}$$

Hence, Eq.(2.25) becomes

$$D_3 \approx -E_{\{\mathbf{x}\}}\Big[\frac{1}{2}(\hat{\theta}(\mathbf{x})-\theta_0)^2\Big]E_{\{\mathbf{x}^*\}}\Big[\frac{\partial^2 l(\theta_0|\mathbf{x}^*)}{\partial\theta_0^2}\Big]. \tag{2.27}$$

Equation (2.17) yields

$$\begin{aligned} E_{\{\mathbf{x}^*\}}\Big[\frac{\partial^2 l(\theta_0|\mathbf{x}^*)}{\partial\theta_0^2}\Big] &= \Big(-\frac{1}{(1-\theta_0)\theta_0^2}+\frac{1}{(1-\theta_0)^2\theta_0}\Big)E_{\{\mathbf{x}^*\}}\Big[\sum_{i=1}^{n}(x_i^*-\theta_0)\Big]-\frac{n}{\theta_0(1-\theta_0)} \\ &= -\frac{n}{\theta_0(1-\theta_0)}. \end{aligned} \tag{2.28}$$

Substitution of Eq.(2.22) and Eq.(2.28) into Eq.(2.27) gives

$$D_3 \approx \frac{1}{2}. \tag{2.29}$$

Substitution of Eq.(2.23), Eq.(2.24), and Eq.(2.29) into Eq.(2.14) yields

$$b = D_1 + D_2 + D_3 \approx 1. \tag{2.30}$$

By substituting this result into Eq.(2.8), we obtain

$$E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big] \approx E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big]-1. \tag{2.31}$$

The result of only n trials (that is, one set of data) is available in ordinary situations. Therefore, $l(\hat{\theta}(\mathbf{x})|\mathbf{x})$ is used as an approximation of $E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big]$. Using this approximation and multiplying by (-2), we obtain the statistic defined as AIC. Hence, when $\hat{\theta}(\mathbf{x})$ is used as an estimate of the parameter (θ_0) of the Bernoulli distribution, AIC is obtained as

$$AIC = -2\cdot l(\hat{\theta}(\mathbf{x})|\mathbf{x})+2. \tag{2.32}$$

That is, b (Eq.(2.8)) becomes a constant independent of θ_0.

Equation (2.32) is derived when the following regression equation is assumed:

$$f(x_i|\hat{\theta}(\mathbf{x})) = \hat{\theta}(\mathbf{x})^{x_i}(1-\hat{\theta}(\mathbf{x}))^{(1-x_i)}. \tag{2.33}$$

The use of Eq.(2.32) allows us to approximate the expected log-likelihood as $AIC\cdot(-0.5)$ when $\hat{\theta}(\mathbf{x})$, given by the maximum likelihood method, is adopted as the parameter.

We also consider the regression equation given below.

$$f(x_i|\theta=0.5) = 0.5^{x_i}\cdot 0.5^{(1-x_i)} = 0.5 \tag{2.34}$$

This equation is based on the assumption that a coin is not biased. This assumption turns Eq.(2.10) into

$$\begin{aligned} b &= E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big]-E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big] \\ &= E_{\{\mathbf{x}\}}\Big[\sum_{i=1}^{n}\Big(x_i\log(0.5)+(1-x_i)\log(1-0.5)\Big)\Big] \\ &\quad -E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[\sum_{i=1}^{n}\Big(x_i^*\log(0.5)+(1-x_i^*)\log(1-0.5)\Big)\Big] \\ &= n\log(0.5)-n\log(0.5) \\ &= 0. \end{aligned} \tag{2.35}$$

Hence, we have

$$E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big] = E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big]. \tag{2.36}$$

By substituting $\theta = 0.5$ into Eq.(2.2), AIC in this setting becomes

$$AIC = -2 \cdot l(0.5|\mathbf{x}) = -2n\log(0.5). \tag{2.37}$$

Note that when Eq.(2.34) is assumed, Eq.(2.35), Eq. (2.36), and Eq.(2.37) hold even if the population (that is, the true distribution) that generates $\mathbf{x}^*$ is described as $\theta_0 = \theta_1$ ($\theta_1 \neq 0.5$).

That is, when we assume the model of $\theta_0 = 0.5$, Eq.(2.36) gives an approximation of the expected log-likelihood in both situations: i) the true distribution satisfies $\theta_0 = \theta_1$ ($\theta_1 \neq 0.5$) and ii) the true distribution satisfies $\theta_0 = 0.5$.

The general derivation of AIC uses the condition that the specified model contains the true distribution as a special case. This is because Eq.(3.105) on page 61 in [2] is derived using this condition.

However, the model of $\theta_0 = 0.5$ does not contain that of $\theta_0 = \theta_1$ ($\theta_1 \neq 0.5$) as a special case. Nevertheless, when the model of $\theta_0 = 0.5$ is assumed for a Bernoulli distribution, $AIC \cdot (-0.5)$, which is given by Eq.(2.37), works as an approximation of the expected log-likelihood if the true distribution satisfies $\theta_0 = 0.5$ and if the true distribution satisfies $\theta_0 = \theta_1$ ($\theta_1 \neq 0.5$).

Therefore, in the coin tossing problem, if the model of $\theta_0 = 0.5$ is assumed (that is, the coin is not biased), we can use Eq.(2.37) regardless of whether or not the coin is actually biased. Moreover, when we assume that the coin may be biased (i.e. $\theta_0 = \hat{\theta}$), the specified model contains the true distribution as a special case. Hence, we can use Eq.(2.32).

The findings above show that in the coin tossing problem, a comparison of the AIC given by Eq.(2.32) with that given by Eq.(2.37) enables us to estimate the fairness of a coin using the expected log-likelihood.

3 Numerical Simulations

Using Eq.(2.2), Eq.(2.4), Eq.(2.5), and Eq.(2.6), b (Eq.(2.10)) is approximated as

$$\begin{aligned} b &= E_{\{\mathbf{x}\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x})\Big] - E_{\{\mathbf{x},\mathbf{x}^*\}}\Big[l(\hat{\theta}(\mathbf{x})|\mathbf{x}^*)\Big] \\ &\approx \frac{1}{K}\sum_{k=1}^{K}\sum_{j=1}^{n}\Big(x_{jk}\log\Big(\frac{1}{n}\sum_{i=1}^{n}x_{ik}\Big) + (1-x_{jk})\log\Big(1-\frac{1}{n}\sum_{i=1}^{n}x_{ik}\Big)\Big) \\ &\quad -\frac{1}{KQ}\sum_{q=1}^{Q}\sum_{k=1}^{K}\sum_{j=1}^{n}\Big(x^*_{jq}\log\Big(\frac{1}{n}\sum_{i=1}^{n}x_{ik}\Big) + (1-x^*_{jq})\log\Big(1-\frac{1}{n}\sum_{i=1}^{n}x_{ik}\Big)\Big), \end{aligned} \tag{3.1}$$

where $\{x_{ik}\}$ denotes the available data and $\{x^*_{iq}\}$ denotes future data. In addition, $\{x_{ik}\}$ and $\{x^*_{iq}\}$ are realizations of the Bernoulli distribution with parameter $\theta = \theta_0$. In this numerical simulation, θ_0 is one of $\{0.3, 0.4, 0.5\}$ and n is one of $\{50, 200\}$. Furthermore, $K = 100$ and $Q = 20$ are set. By varying the initial number of pseudo-random values, the values of b (Eq.(3.2)) are calculated $2,000$ times. Note that we exclude the case when $\sum_{i=1}^{n} x_{ik} = 0$ because log-likelihood is unavailable when $\hat{\theta} = 0$. The resultant histograms of the value of b are illustrated in Fig. 1. The average of the value of b is approximately 1 and the values of b are distributed close to 1. These graphs indicate that $AIC \cdot (-0.5)$ is regarded as an approximation of the expected log-likelihood. They also show that when the value of θ_0 is 0.3, that is, a coin is considerably biased, the variance of b is large.

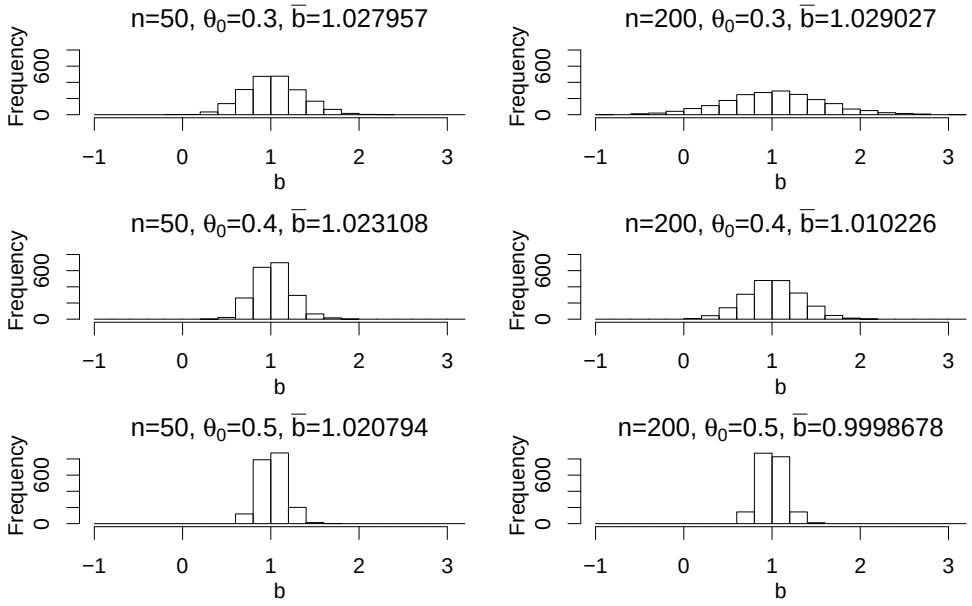

Fig. 1. Distribution of the value of b derived by Eq.(3.2). The three graphs on the left-hand side are obtained when $n = 50$ and θ_0 is one of $\{0.3, 0.4, 0.5\}$. The three graphs on the right-hand side are obtained when $n = 200$ and θ_0 is one of $\{0.3, 0.4, 0.5\}$. Here, $\bar{b}$ stands for the average of $2,000$ values of b

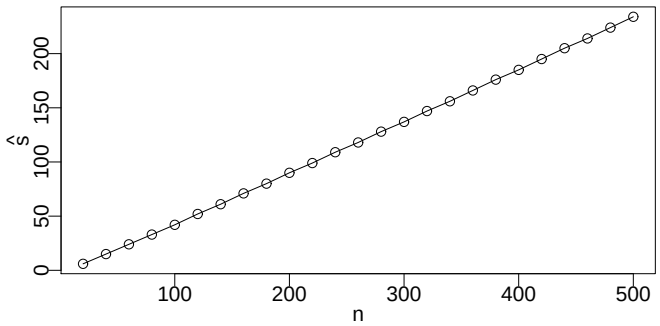

Fig. 2. Relationship between the number of trials (n) and $\hat{s}$ (the maximal s in the critical region where $s < 0.5n$ holds)

Next, we assume that the number of trials is n and the number of 1's in $\{x_i\}$ is s. Then, the number of 0's in $\{x_i\}$ is $(n - s)$. Using this setting, we compare the AIC (Eq.(2.32)) given by the maximum likelihood estimator (Eq.(2.4)) with the AIC (Eq.(2.37)) given by $\theta_0 = 0.5$. Here, the AIC defined by Eq.(2.37) is called AIC_0 and the AIC defined by Eq.(2.2), Eq.(2.4), and Eq.(2.32) is called AIC_1. That is, AIC_0 and AIC_1 are respectively defined as

$$AIC_0 = -2n\log(0.5), \tag{3.2}$$

and

$$\begin{aligned} AIC_1 &= -2\Big(\sum_{i=1}^{n}\Big(x_i\log(\hat{\theta}(\mathbf{x})) + (1-x_i)\log(1-\hat{\theta}(\mathbf{x}))\Big)\Big) + 2 \\ &= -2\Big(\sum_{i=1}^{n}\Big(x_i\log\big(\frac{1}{n}\sum_{j=1}^{n}x_j\big) + (1-x_i)\log\big(1-\frac{1}{n}\sum_{j=1}^{n}x_j\big)\Big)\Big) + 2 \\ &= -2\Big(\sum_{i=1}^{n}\Big(x_i\log\big(\frac{s}{n}\big) + (1-x_i)\log\big(1-\frac{s}{n}\big)\Big)\Big) + 2, \end{aligned} \tag{3.3}$$

where s $(0 \leq s \leq s_{max})$ stand for the number of times of the coin lands heads up when the number of trials is n, s_{max} is the maximal number of integers that are less than $0.5n$. In contrast, when a coin is not biased, that is, $\theta_0 = 0.5$ holds, $p(s)$, which is the probability that the number of times the coin lands heads up is less than or equal to s, is

$$p(s) = \sum_{i=0}^{s}\Big({}_nC_i \cdot 0.5^i \cdot (1-0.5)^{(n-i)}\Big) = 0.5^n \sum_{i=0}^{s} {}_nC_i. \tag{3.4}$$

Because AIC_1 cannot be defined when we assume $s = 0$, s is set to one of $\{1, 2, 3, \ldots, s_{max} - 1\}$ to find the minimal value of s that satisfies

$$(AIC_1(s) - AIC_0) \cdot (AIC_1(s+1) - AIC_0) < 0. \tag{3.5}$$

The resultant s is termed as $\hat{s}$. Here, $\hat{s}$ is the maximal value of s that is located in the critical region where $s < 0.5n$ holds. Substitution of $\hat{s}$ into s in Eq.(3.4) leads to the value of $p(\hat{s})$. Since this test is two-sided, $2 \cdot p(\hat{s})$ is the significance level of the hypothesis test regarding the assumption of a Bernoulli distribution. Fig. 2. illustrates the relationship between n and $\hat{s}$ when the number of trials (n) is one of $\{20, 40, 60, \cdots, 500\}$, and Fig. 3. illustrates the relationship between n and $2 \cdot p(\hat{s})$.

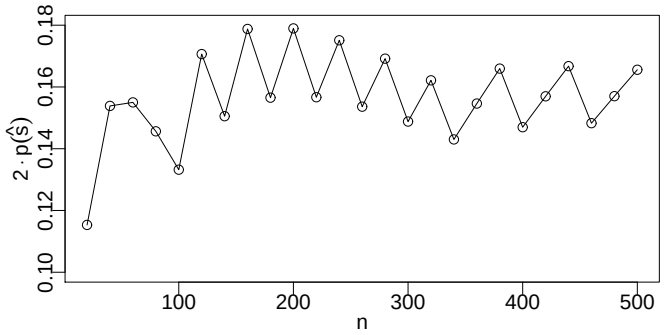

Fig. 3. Relationship between the number of trials (n) and $2 \cdot p(\hat{s})$

Next, we carried out a numerical simulation in which the experiment of tossing a coin n times was repeated $10,000$ times to derive AIC_0(Eq.(3.2)) and AIC_1(Eq.(3.3)) and counted the number of times that $AIC_0 > AIC_1$ holds; the result of the counting is denoted by m. The resultant values of m when the number of trials (n) in each experiment is $\{20, 40, 60, \cdots, 500\}$ are given in Fig. 4. This graph looks similar to the one in Fig. 3.

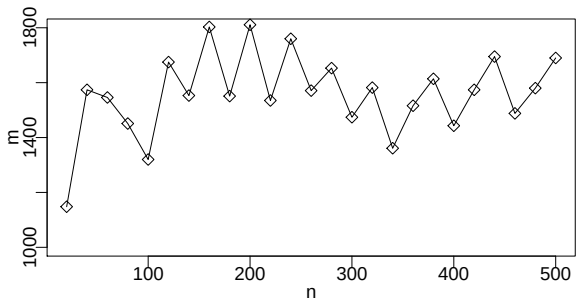

Fig. 4. Relationship between m and n when the experiment of tossing a coin n times was repeated $10,000$ times

Both Fig. 3 and Fig. 4 show that the probability of wrongly identifying a fair coin as a biased coin does not tend to decline as the number of trials increases. This tendency corresponds to the fact that AIC does not provide a consistent estimator of orders ([8]; [9]; [10]; page 74 in [2]). This is not a defect in AIC. Because the significance level of a significance test is usually set at 5% regardless of the number of data, AIC has a similar characteristic to that of a significance test in the sense that the significance level of a significance test that is equivalent to AIC is hardly affected by the number of data.

4 Conclusions

The discussion above shows that, to determine whether or not a coin is biased using the results of coin tossing, AIC can be used from the perspective of fitting closely to future data because $AIC \cdot (-0.5)$ is regarded as the approximation of the expected log-likelihood. It also reveals that determination of the fairness of a coin using AIC is almost equivalent to a statistical test with a significance level ranging from 11% to 18% as shown in Fig.3 and Fig.4. We are tempted to feel that if the probability of wrongly identifying a fair coin as a biased coin is located between 11% and 18%, this is too high, especially in comparison to 5%, which is a common significance level for a significance test. However, we conclude that when we do not tentatively assume that the coin is unbiased, a significance level between 11% and 18% is appropriate if fitting well to future data is our only concern. In contrast, if we suspect that the coin is not biased, we determine that the coin is biased only if we obtain evidence that disproves this suspicion with certainty. A significance test with a significance level of 5% is usually used for such a situation. This means that when we suspect that the coin is not biased, the significance level is set at a lower level than it would be if our purpose were prediction. In this sense, however, the significance level can be 8% or 3%, for example. The main reasons a significance level of 5% is currently popular are that 5% is psychologically acceptable and has been used historically (e.g. [11]).

One thing we can say for sure is that for the coin-tossing problem, AIC and a significance test with a significance level of 5% should be used differently. If we think of AIC and the 5% test as belonging to different paradigms ([12]), this obscures the essential notion that the difference is based on that of significance level; the difference should be attributed to the problem setting and the goal. Moreover, about the selection of the predictive variables of multiple regression analysis, Section 8.2.1 of [13] says:

If prediction performance is the goal, then a 15 to 20% cutoff may work best, although methods designed more directly for optimal prediction should be preferred.

Section 7.4 of [14] states that the threshold for the backward selection method should be set between 10% and 15%.

In contrast, [15] says:

AIC optimization corresponds to significance-based selection at a significance level of 0.157.

This is because Wilks' theorem ([16]) shows that the probability of misidentifying a fair coin as a biased coin is 15.73% if the number of trials is very large. Section 5.6 of [5] derives similar results and shows that *GCV* leads to a similar tendency. Therefore, the backward selection method for creating multiple regression equations yields beneficial results if *AIC* is adopted as a selection criterion. However, this does not necessarily mean that *AIC* performs well as an approximation of the expected log-likelihood (Section 5.2 of [5]). We should always keep in mind that the general procedure for deriving *AIC* (Section 3 of [2]) needs the assumption that restrictive conditions are satisfied.

Although *AIC* is widely used for practical purposes (e.g., [17]; [18]), little attention has been paid to how *AIC* performs in each problem. The simple problem of coin tossing treated here clarifies one aspect of *AIC*. We expect that the characteristics of *AIC* will be examined in more various practical activities.

Acknowledgement

The author is very grateful to the referees for carefully reading the paper and for their comments and suggestions which have improved the paper.

Competing Interests

The author declares that no competing interests exist.

References

[1] Burnham KP, Anderson DR. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, Second Edition. New York: Springer; 2011.

[2] Konishi S, Kitagawa G. Information criteria and statistical modelling. New York: Springer; 2008.

[3] Takezawa K. A Simulation Study for the AIC and Likelihood Cross-validation: The Case of Exponential Versus Weibull Distributions. Journal of Advances in Mathematics and Computer Science. 2018;28(3):1-13.

[4] Craven P, Wahba G. Smoothing noisy data with spline functions. Numerische Mathematik. 1979;31(4):377-390.

[5] Takezawa K. Learning Regression Analysis by Simulation. Springer; 2014.

[6] Buckland ST, Burnham KP, Augustin NH. Model selection: An integral part of inference. Biometrics, 1997;53:603-618.

[7] Claeskens G, Hjort NL. Model Selection and Model Averaging. Cambridge University Press; 2008.

[8] Shibata R. Selection of the order of an autoregressive model by Akaikefs information criterion. Biometrika. 1976;63:117-126.

[9] Nishii R. Asymptotic properties of criteria for selection of variables in multiple regression. Annals of Statistics. 1984;12:758-765.

[10] Shao J. An asymptotic theory for linear model selection.(with discussion) Statistica Sinica, 1997;7:221-264.

[11] Cowles M, Davis C. On the origins of the .05 level of statistical significance. American Psychologist. 1982;37(5):553-558.

[12] Burnham KP, Anderson DR. P values are only an index to evidence: 20th- vs. 21st-century statistical science. Ecology, 2014;95(3):627-630.

[13] Faraway JJ. Linear Models with R, second edition. Chapman & Hall/CRC. Boca Raton, FL, U.S.A.; 2014.

[14] Rawlings JO, Pantula SG, Dickey DA. Applied Regression Analysis: A Research Tool, second edition. Springer; 1998.

[15] Heinze G, Wallisch C, Dunkler D. Variable selection - A review and recommendations for the practicing statistician. Biometrical Journal. 2018;60(3):431-449.

[16] Wilks SS. The large-sample distribution of the likelihood ratio for testing composite hypotheses. The Annals of Mathematical Statistics. 1938;9:60-62.

[17] Taylor DC, Snipes M, Barber NA. Indicators of hotel profitability: model selection using Akaike information criteria. Tourism and Hospitality Research. 2018;18(1):61-71.

[18] Pho KH, Ly Sel, Ly Sal, Lukusa TM. Comparison among Akaike Information Criterion, Bayesian information criterion and Vuong's test in model selection: a case study of violated speed regulation in Taiwan. Journal of Advanced Engineering and Computation. 2019;3(1):293-303.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/52709