PARAMETER ESTIMATION

Parameter Estimation

Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured empirical data that has a random component.

The parameters describe an underlying physical setting in such a way that their value affects the distribution of the measured data.

An estimator attempts to approximate the unknown parameters using the measurements.

Parameter Estimation

The *method of least squares* (*least squares estimation, LSE*) is a standard approach to approximate the solution of overdetermined systems, i.e. sets of equations in which there are more equations than unknowns.

"Least squares" means that the overall solution minimizes the sum of the squares of the *residuals* made in the results of every single equation.

Parameter Estimation

Least-squares problems fall into two categories:

- linear or ordinary least squares;
- nonlinear least squares,

depending on whether or not the residuals are linear in all unknowns.

The linear least-squares problem has a closed-form solution. The nonlinear problem is usually solved by iterative refinement; at each iteration the system is approximated by a linear one, and thus the core calculation is similar in both cases.

Parameter Estimation

The objective consists of adjusting the parameters of a model function to best fit a data set.

A simple data set consists of n data pairs $(x_i, y_i), i = 1, \ldots, n$.

The model function has the form $F(x, \boldsymbol{\Theta})$, where m adjustable parameters are held in the vector $\boldsymbol{\Theta}$.

The goal is to find the parameter values $\hat{\theta}_j,\ j = 1, \ldots, m$ for the model that "best" fits the data.

Parameter Estimation

The fit of a model to a data point is measured by its residual, defined as the difference between the actual value of the dependent variable and the value predicted by the model:

$$r(\hat{\boldsymbol{\Theta}})_i = y_i - F(x_i, \hat{\boldsymbol{\Theta}})$$

The LSE method finds the optimal parameter values by minimizing the sum of squared residuals:

$$S(\hat{\boldsymbol{\Theta}}) = \sum_{i=1}^{n} r(\hat{\boldsymbol{\Theta}})_i^2$$

Parameter Estimation

Ex.: Assume that five identical units are being reliability tested. The units fail during the test after operating the following number of hours: 20, 275, 365, 415, and 1020.

Assuming that the data follow exponential distribution, estimate the value of the parameter λ.

Here, the vector $\widehat{\boldsymbol{\Theta}}$ contains single element - $\hat{\lambda}$, and

$$F(t, \boldsymbol{\Theta}) = 1 - e^{-\theta t}.$$

Parameter Estimation

The naïve approach is to minimize the sum of squared residuals as previously specified:

$$S(\widehat{\boldsymbol{\Theta}}) = \sum_{i=1}^{5} r(\widehat{\boldsymbol{\Theta}})_i^2 = \sum_{i=1}^{5} \left(Y_i - \left(1 - e^{-\widehat{\theta} t_i}\right)\right)^2$$

where Y_i are obtained as median ranks.

t_i	Y_i
20	0,129
275	0,314
365	0,5
415	0,686
1020	0,871

However, this would lead to a nonlinear equation with respect to $\widehat{\boldsymbol{\Theta}}$.

We can avoid it by linearizing the *cdf:*

$$F(t,\boldsymbol{\Theta}) = 1 - e^{-\theta t} \qquad 1 - F(t,\boldsymbol{\Theta}) = e^{-\theta t}$$

$$\ln\big(1 - F(t,\boldsymbol{\Theta})\big) = -\theta t \;\Rightarrow\; y = kx + b$$

$$y \equiv \ln\big(1 - F(t,\boldsymbol{\Theta})\big) = \ln(1 - Y)$$

$$k \equiv -\theta$$

$$x \equiv t$$

$$b \equiv 0$$

Then the sum of squared residuals:

$$S(\hat{k}) = \sum_{i=1}^{5} (y_i - \hat{k}x_i)^2 = \sum_{i=1}^{5} (y_i^2 - 2\hat{k}x_i y_i + \hat{k}^2 x_i^2) =$$

$$= \hat{k}^2 \sum_{i=1}^{5} x_i^2 - 2\hat{k} \sum_{i=1}^{5} x_i y_i + \sum_{i=1}^{5} y_i^2$$

To find the minimum of $S(\hat{k})$ we should set $\frac{dS}{d\hat{k}} = 0$.

Parameter Estimation

$$\frac{dS}{d\hat{k}} = 2\hat{k}\sum_{i=1}^{5} x_i^2 - 2\sum_{i=1}^{5} x_i y_i = 0$$

$$\hat{k} = \frac{\sum_{i=1}^{5} x_i y_i}{\sum_{i=1}^{5} x_i^2}$$

Parameter Estimation

	x_i	x_i^2	Y_i	y_i	$x_i y_i$
	20	400	0,129	-0,138	-2,762
	275	75625	0,314	-0,377	-103,641
	365	133225	0,5	-0,693	-252,999
	415	172225	0,686	-1,158	-480,72
	1020	1040400	0,871	-2,048	-2088,902
Σ		1421875			-2929,024

$$\hat{k} = \frac{-2929.024}{1421875} = -2.06 \times 10^{-3}$$

$$\hat{\lambda} = 2.06 \times 10^{-3}$$

Parameter Estimation

The LSE method is quite good for functions that can be linearized. For these distributions, the calculations are relatively easy and straightforward, having closed-form solutions that can readily yield an answer without having to resort to numerical techniques or tables.

LSE is generally best used with data sets containing complete data, that is, data consisting only of single times-to-failure with no censored or interval data.

Parameter Estimation

In statistics, *maximum likelihood estimation* (*MLE*) is a method of estimating the parameters of a statistical model so the observed data is most probable.

Specifically, this is done by finding the value of the parameter (or parameter vector) $\widehat{\boldsymbol{\Theta}}$ that maximizes the likelihood function $\mathcal{L}\left(\widehat{\boldsymbol{\Theta}}\right)$, which is the joint probability (or probability density) of the observed data over a parameter space.

Parameter Estimation

The vector $\widehat{\boldsymbol{\Theta}}$ that maximizes the likelihood function is called the *maximum likelihood estimate*.

The logic of maximum likelihood is both intuitive and flexible, and as such the method has become a dominant means of inference within much of the quantitative research of the social and medical sciences and engineering.

Parameter Estimation

Consider X – a continuous random variable with *pdf*:

$$f(x, \boldsymbol{\Theta}) \equiv f(x, \theta_1, \theta_2, \ldots, \theta_k)$$

where $\theta_1, \theta_2, \ldots, \theta_k$ are k unknown parameters which need to be estimated, with N independent observations, $x_1, x_2, \ldots, x_N$, which correspond in the case of life data analysis to failure times.

The *likelihood function* is given by:

$$\mathcal{L}(\widehat{\boldsymbol{\Theta}}) = \prod_{i=1}^{N} f(x_i, \widehat{\boldsymbol{\Theta}})$$

Parameter Estimation

In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the *logarithmic likelihood* (*log-likelihood*) *function*:

$$\Lambda(\widehat{\Theta}) = \ln \mathcal{L}(\widehat{\Theta}) = \sum_{i=1}^{N} \ln f(x_i, \widehat{\Theta})$$

The maximum likelihood estimators (or parameter values) of $\theta_1, \theta_2, \ldots, \theta_k$ are obtained by maximizing $\mathcal{L}(\widehat{\boldsymbol{\Theta}})$ or $\Lambda(\widehat{\boldsymbol{\Theta}})$:

$$\frac{\partial \Lambda}{\partial \hat{\theta}_j} = 0, \qquad j = 1,2, \ldots, k$$

Parameter Estimation

Ex.: Assume that five identical units are being reliability tested. The units fail during the test after operating the following number of hours: 20, 275, 365, 415, and 1020.

Assuming that the data follow exponential distribution, estimate the value of the parameter λ.

Here, the vector $\widehat{\boldsymbol{\Theta}}$ contains single element - $\hat{\lambda}$, and

$$f(t, \boldsymbol{\Theta}) = \theta e^{-\theta t}.$$

Parameter Estimation

The log-likelihood function:

$$\Lambda(\hat{\Theta}) = \sum_{i=1}^{5} \ln\left(\hat{\Theta} e^{-\hat{\Theta} t_i}\right) = 5 \ln \hat{\Theta} - \hat{\Theta} \sum_{i=1}^{5} t_i$$

Substituting failure times for t_i, we get:

$$\Lambda(\hat{\Theta}) = 5 \ln \hat{\Theta} - 2095 \cdot \hat{\Theta}$$

$$\frac{\partial \Lambda}{\partial \hat{\Theta}} = \frac{5}{\hat{\Theta}} - 2095 = 0 \quad \Rightarrow \quad \hat{\Theta} = \frac{5}{2095} = 2.39 \times 10^{-3}$$

$$\hat{\lambda} = 2.39 \times 10^{-3}$$

Parameter Estimation

Analyzing the results of two previous examples, you should notice that parameter estimates differ from one another, and we can't specify which result is better.

We can evaluate *residual sum of squares* (*RSS*) or *mean squared error* (*MSE*):

$$RSS(\hat{\boldsymbol{\Theta}}) = \sum_{i=1}^{N} \left(Y_i - F(X_i, \hat{\boldsymbol{\Theta}})\right)^2 \quad MSE(\hat{\boldsymbol{\Theta}}) = \frac{RSS(\hat{\boldsymbol{\Theta}})}{N}$$

Parameter Estimation

We also can determine the likelihood of either result by calculating $\Lambda(\hat{\boldsymbol{\Theta}})$, or, $-2\Lambda(\hat{\boldsymbol{\Theta}})$ - the metric used in various statistical quality tests.

Let's compare these metrics obtained with the results of LSE and MLE:

	RSS	-2Λ
$\hat{\lambda}_{LSE} = 2.06 \times 10^{-3}$	0.035	70.482
$\hat{\lambda}_{MLE} = 2.39 \times 10^{-3}$	0.047	70.379

Parameter Estimation

	RSS	-2Λ
$\hat{\lambda}_{LSE} = 2.06 \times 10^{-3}$	0.035	70.482
$\hat{\lambda}_{MLE} = 2.39 \times 10^{-3}$	0.047	70.379

The results we have here are quite obvious: $\hat{\lambda}_{LSE}$ is the value for which *RSS* is minimal, so any other parameter value yields greater *RSS*.

Likewise, $\hat{\lambda}_{MLE}$ is the value that maximizes $\Lambda(\widehat{\boldsymbol{\Theta}})$ (and minimizes $-2\Lambda(\widehat{\boldsymbol{\Theta}})$).

PSEUDO-RANDOM NUMBER SAMPLING

Pseudo-Random Number Sampling

Quite often in statistics and simulation we need to obtain samples of random numbers distributed according to a given probability distribution.

Modern mathematical software is equipped with *pseudo-random number generator* producing *uniformally* distributed samples.

To generate samples drawn from other distributions we need to resort to *pseudo-random number sampling* techniques.

Pseudo-Random Number Sampling

The most common technique is *inverse transform sampling* (*Smirnov transform*, *inverse transformation method*).

Inverse transformation sampling takes uniform samples of a number u between 0 and 1, *interpreted as a probability*, and then returns the largest number *x* from the domain of the distribution $P(X)$ such that

$$P(-\infty < X < x) \leq u.$$

Pseudo-Random Number Sampling

Computationally, this method involves computing the quantile function of the distribution — in other words, computing the cumulative distribution function (*cdf*) of the distribution and then inverting that function.

For a continuous distribution we need to integrate the probability density function (*pdf*) of the distribution or to obtain quantile function in an explicit form, which is impossible to do analytically for most distributions (including the normal distribution).

Pseudo-Random Number Sampling

Let X be a random variable whose distribution can be described by the *cdf* F_X. We want to generate values of X which are distributed according to this distribution.

The inverse transform sampling method works as follows:

- Generate a random number u from the standard uniform distribution in the interval $[0,1]$.
- Find the inverse of the desired *cdf*, $F_X^{-1}(x)$.
- Compute $X = F_X^{-1}(u)$. The computed random variable X has distribution $F_X(x)$.

Ex.: Suppose we have a random variable $U \sim Unif(0,1)$ and a *cdf* of Weibull distribution

$$F(x) = 1 - e^{-\left(\frac{x}{\eta}\right)^{\beta}}.$$

In order to perform an inversion we need to express x in terms of $U = F(x)$.

Pseudo-Random Number Sampling

$$F(x) = 1 - e^{-\left(\frac{x}{\eta}\right)^{\beta}} \qquad F(x) = U$$

$$1 - e^{-\left(\frac{x}{\eta}\right)^{\beta}} = U$$

$$e^{-\left(\frac{x}{\eta}\right)^{\beta}} = 1 - U$$

$$\left(\frac{x}{\eta}\right)^{\beta} = -\ln(1 - U)$$

$$x = \eta \cdot (-\ln(1 - U))^{\frac{1}{\beta}}$$

Pseudo-Random Number Sampling

Once the samples of components' failure times are generated, we can obtain the sample of the system failures, providing that the system configuration is known.

For example, if the *series system* consists of m components and for each of them failure time samples $X^{\langle i \rangle}$ $(i = 1..m)$ of size n were generated, then we can get the sample Y of system's failure times as follows:

$$Y_j = \min_{i \in (1,m)} X_j^{\langle i \rangle}, j = 1..n$$

Pseudo-Random Number Sampling

For the *parallel hot redundancy system* of m components we get:

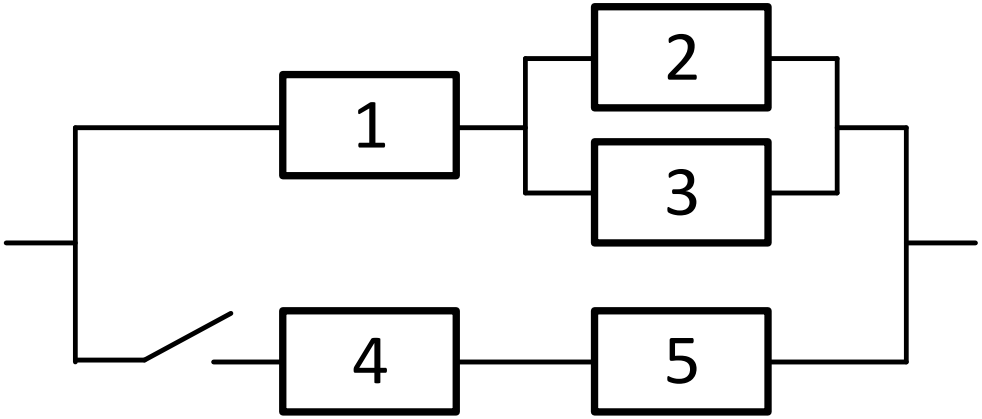
$$Y_j = \max_{i \in (1,m)} X_j^{\langle i \rangle}$$

For the *cold standby redundancy system* of m components we get:

$$Y_j = \sum_{i=1}^{m} X_j^{\langle i \rangle}$$

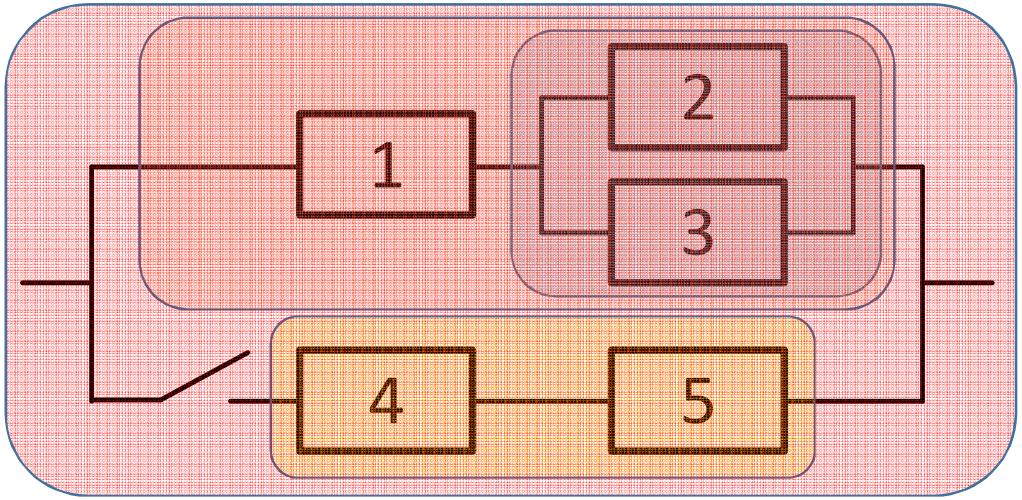
Pseudo-Random Number Sampling

Ex.: For the system with RBD shown below provide an algorithm of generating the sample of n system failures.



We assume the failure times of components 1 – 5 are collected into samples X1 – X5, respectively.

Pseudo-Random Number Sampling



$$\max(X2_j, X3_j)$$

$$\min\left[X1_j, \max(X2_j, X3_j)\right]$$

$$\min(X4_j, X5_j)$$

$$Y_j = \min\left[X1_j, \max(X2_j, X3_j)\right] + \min(X4_j, X5_j)$$

MODEL SELECTION

Model Selection

Model selection is the task of selecting a statistical model from a set of candidate models, given data.

Given candidate models of similar predictive or explanatory power, the simplest model is most likely to be the best choice (*Occam's razor*).

In its most basic forms, model selection is one of the fundamental tasks of scientific inquiry. Determining the principle that explains a series of observations is often linked directly to a mathematical model predicting those observations.

Model Selection

The mathematical approach of model selection decides among a set of *candidate models*; this set must be chosen by the researcher.

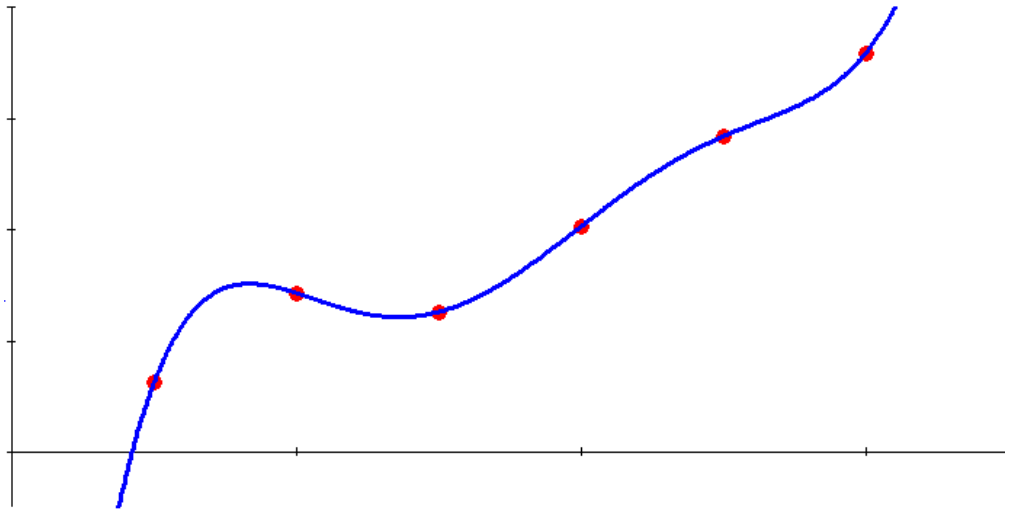
Once the set of candidate models has been chosen, the statistical analysis allows selecting the best of these models.

What is meant by “best” is controversial.

Model Selection

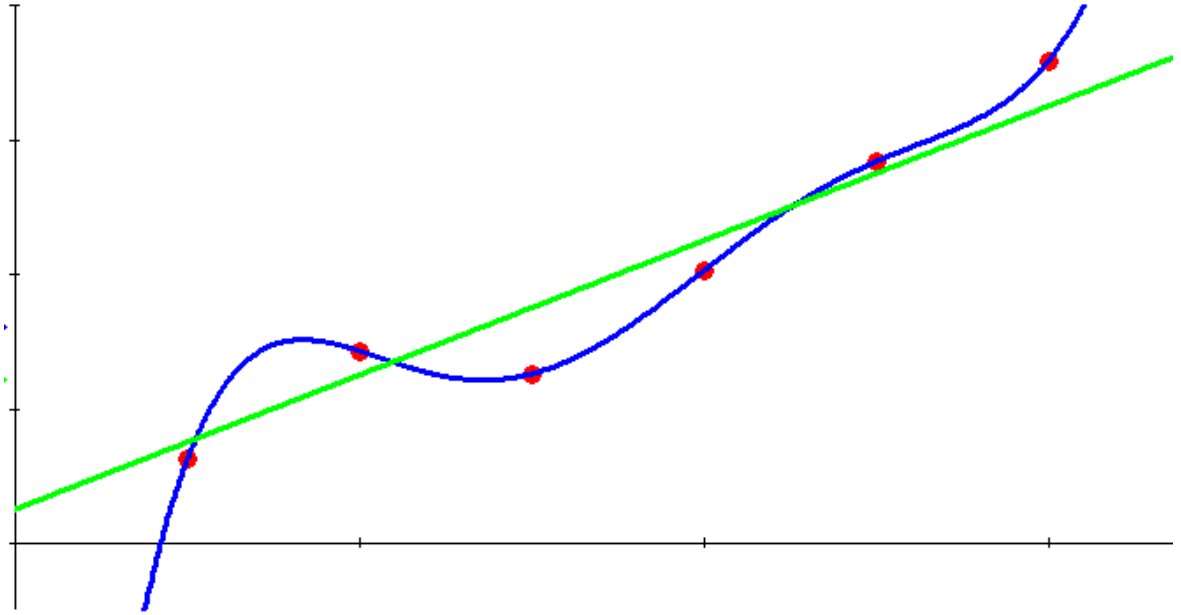
A good model selection technique will *balance goodness of fit with simplicity*.

More complex models will be better able to adapt their shape to fit the data (for example, a fifth-order polynomial can exactly fit six points).



Model Selection

However, the additional parameters may not represent anything useful. (Perhaps those six points are really just randomly distributed about a straight line.)



Model Selection

The most straightforward technique of model selection is to prefer the model with maximum likelihood (log-likelihood) score.

If $\Lambda_1(\widehat{\boldsymbol{\Theta}}_1)$ and $\Lambda_2(\widehat{\boldsymbol{\Theta}}_2)$ are the values of log-likelihood functions for two candidate models with estimated parameter vectors $\widehat{\boldsymbol{\Theta}}_1$ and $\widehat{\boldsymbol{\Theta}}_2$, then the model with greater log-likelihood value should be preferred.

For various reasons, in statistics usually estimate the value $-2\Lambda(\widehat{\boldsymbol{\Theta}})$, and select the model with minimal score.

Model Selection

Though this technique accounts for goodness of the fit, it doesn't take into account model complexity.

In order to do that, the *Akaike information criterion (AIC)* was introduced, which is an estimator of the relative quality of statistical models for a given set of data.

Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models.

Model Selection

Suppose that we have a statistical model of some data.

Let k be the number of estimated parameters in the model.

Let $\widehat{\Lambda}$ be the maximum value of the log-likelihood function for the model.

Then the AIC value of the model is the following:

$$AIC = 2k - 2\widehat{\Lambda}$$

Model Selection

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value.

Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a *penalty* that is an increasing function of the number of estimated parameters.

The penalty discourages overfitting, because increasing the number of parameters in the model almost always improves the goodness of the fit.

Model Selection

Note that AIC tells nothing about the absolute quality of a model, only the quality relative to other models.

Thus, if all the candidate models fit poorly, AIC will not give any warning of that.

Hence, after selecting a model via AIC, it is usually good practice to validate the absolute quality of the model.

Various developments of the AIC were introduced later. Based on ideas from statistics and information theory, researchers suggested different penalties for the number of parameters.

Bayesian information criterion (BIC):

$$BIC = k \ln n - 2\widehat{\Lambda}$$

Hannan–Quinn information criterion (HQC):

$$HQC = 2k \ln(\ln n) - 2\widehat{\Lambda}$$

where n is the sample size.

Model Selection

Another approach for model selection is to apply various statistical goodness-of-the-fit tests.

Generally, they are performed in order to validate the model (as a part of hypothesis testing), i.e. to determine whether the model is applicable to the given data.

However, these tests can be applied to assess the relative quality of models.

Cramér–von Mises Test

Let $x_1, x_2, \ldots, x_n$ be the observed values, *in increasing order*, and $F(x)$ is the *cdf* of the model under the test. Then, the statistic is:

$$CM = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F(x_i) \right]^2$$

The smaller the value of CM is, the better the quality of the model.

Anderson-Darling Test

Let $x_1, x_2, \ldots, x_n$ be the observed values, *in increasing order*, and $F(x)$ is the *cdf* of the model under the test. Then, the statistic is:

$$A^2 = -n - S,$$

where

$$S = \sum_{i=1}^{n} \frac{2i-1}{n} \left[\ln\left(F(x_i)\right) + \ln\left(1 - F(x_{n+1-i})\right) \right]$$

Again, smaller value of A^2 indicates better model.