Institute of Actuaries of Australia

Optimal Groups using the Akaike Information Criterion

Prepared by George Pierides and Chul Park

Presented to the Institute of Actuaries of Australia
XVth General Insurance Seminar 16-19 October 2005

The Institute of Actuaries of Australia
Level 7 Challis House 4 Martin Place
Sydney NSW Australia 2000
Telephone: +61 2 9233 3466 Facsimile: +61 2 9233 3446
Email: actuaries@actuaries.asn.au Website: www.actuaries.asn.au

Optimal Groups using the Akaike Information Criterion

1. Abstract

In any initial data analysis, it is necessary to gain an understanding of the data at a basic level. What is there, what isn't, what idiosyncrasies arise from the insurer's processing of the data, is the data in an appropriate form for modeling?

A good starting point is to construct histograms of the data. The histograms are constructed from raw data. The most common question asked when constructing histograms is "How many bins should I use". Through the use of the Akaike Information Criterion we will show how an optimal number of bins is calculated and in turn use this approach to estimate an optimal number of homogenous groupings in attempts to classify rating variable categories in a pricing exercise. The basic idea for the method is obtained from Sakamoto(1985) & later by Taylor (1987).

Keywords: bins, optimal groupings, histogram, pricing, Akaike Information Criterion

2. Introduction

Following the work of Sakamoto(1985), the Akaike Information Criterion(AIC) is a basis of comparison and selection among several models. The AIC was introduced by Akaike (1973). In his paper Akaike showed the importance of the Kullback-Leibler (1951) information quantity in statistics and derived AIC as its estimator.

The AIC is a basis of comparison and selection among several statistical models. As we all know the goodness of fit of parameters of a model can be calculated by the expected log likelihood, namely, the larger the expected log likelihood the better the explanation. The log likelihood can be regarded as an estimator of the expected log likelihood.

The mean expected log likelihood is the quantity defined as the mean , with respect to the data x, of the expected log likelihood of the maximum likelihood model. That is, the larger the mean expected log likelihood the better the fit to the model. At first sight, it would seem that the mean expected log likelihood can be estimated by the maximum log likelihood. The maximum log likelihood, however, is shown to be a biased estimator of the mean expected log likelihood. The maximum log likelihood has a general tendency to over estimate the true value of the mean expected log likelihood. This tendency is more prominent for models with larger number of free parameters. This means that if we choose the model with the largest maximum log likelihood, a model with an unnecessarily large number of free parameters is likely to be chosen.

In looking at the relationship between the bias and the number of free parameters of a model, it is found that

(maximum log likelihood of a model) – (number of free parameters of the model)

is an asymptotically unbiased estimator of the mean expected log likelihood. As defined by Akaike (1987) his AIC estimator of Kullback –Leibler information is

AIC = -2 x (maximum log likelihood of the model)+ 2 x(number of free parameters of the model)

Or denoted as $AIC(k) = -2\ \lambda(\theta_k) + 2\ K$

is Akaike's proposed criterion for model selection. A model which minimises the AIC (denoted by MAICE) is the most significant model considered appropriate.

When there are several models whose values of the maximum likelihood are about the same level, we would choose the model with the smallest number of free parameters. In this sense AIC realises, the principle of parsimony.

3. Constructing an optimal histogram?

As Sakamoto(1985) defined, let $x_{1}, \ldots, x_n$ be a sample of n measurements from a certain population and let $x_{(1)}$ and $x_{(n)}$ be the smallest and largest values in the sample. An interval $[x_{(1)}-0.5p, x_{(n)}+0.5p]$ is divided into g classes, where p is the precision of each score.

One way to determine the category size g is to take $[2n^{1/2}-1]$, where the operator [] represents the integer part as this is used as the first set of bins.

Using the initial set of bins a histogram is constructed. The AIC is calculated to check the goodness of fit of various models to the initial frequency table. Using this approach will allow the derivation of an optimal frequency histogram.

So consider where c_1 and c_2 are the number of bins to be grouped in the two ends of the histogram. Also let r be the number of bins that are grouped in the middle section.

Sakamoto(1985) defined a model as being Model(c_1, r, c_2).

The respective model is defined as

$$p(1) = p(2) = p(3) = \ldots = p(c_1) = \vartheta(1)$$
$$p(c_1 + (j-2)r + 1) = \ldots = p(c_1 + (j-1)r) = \vartheta(j)$$
$$p(c - c_2 + 1) = \ldots = p(c) = \vartheta(\frac{c - c_1 - c_2}{r} + 2)$$

using the fact that $\sum_{i=1}^{c} p(i) = 1$

$$c_1\vartheta(1) + r\sum_j \vartheta(j) + c_2\vartheta(\frac{c - c_1 - c_2}{r} + 2) = 1$$

calculating the maximum log likelihood will give

$$AIC(c_1, r, c_2)$$
$$= (-2)\left[\left(\sum_{i=1}^{c_1} n(i)\right)\log\left(\sum_{i=1}^{c_1}\frac{n(i)}{c_1 n}\right) + \sum_{j=2}^{(c-c_1-c_2/r)+1}\left\{\left(\sum_{i=c_1+(j-2)r+1}^{c_1+(j-1)} n(i)\right)\log\left(\sum_{i=c_1+(j-2)r+1}^{c_1+(j-1)r}\frac{n(i)}{rn}\right)\right\}\right.$$
$$\left. + \left(\sum_{i=c-c_2+1}^{c} n(i)\right)\log\left(\sum_{i=c-c_2+1}^{c}\frac{n(i)}{c_2 n}\right)\right] + 2\left(\frac{c-c_1-c_2}{r}+1\right)$$

Let $n'(j)$ be the frequencies in the jth interval and c' be the number of bins in the histogram as a result of grouping initial classes resulting in

$$AIC(c_1, r, c_2) = (-2)\left\{n'(1)\log\frac{n'(1)}{c_1 n} + \sum_{j=2}^{c'} n'(j)\log\frac{n'(j)}{rn} + n'(c')\log\frac{n'(c')}{c_2 n}\right\} + 2(c'-1)$$

We will use the chi-squared statistic to test the application of the AIC in selecting the optimal number of bins. Given this we will need to define the chi-squared statistic, as it is the best known goodness of fit statistic. It can be used for both continuous and discrete sample data. To calculate the chi-squared statistic, you must first break up the x-axis domain into several bins.

The chi-squared statistic is then defined as:

$$\chi^2 = \sum_{i=1}^{n}\frac{\left(N_i - E\ )^2\right)}{E_i}$$

Where

N_i = the observed number of sample in the ith bin.
E_I = the expected number of samples in the ith bin.
N = The number of bins

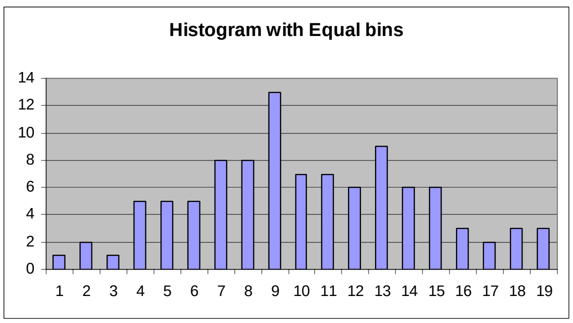
We know that a weakness of the chi-squared statistic is that there are no clear guidelines for the selecting the number and location of the bins. In some situations, you can reach different conclusions from the same data depending on how you specified the bins. We recognise the existence of other non-parametric statistics that could be used but for the purposes of this paper we will only use chi-squared.

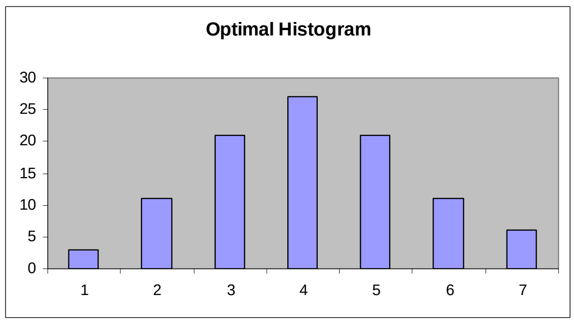
The following example illustrates when the optimal number of bins are calculated by using the AIC the most appropriate distribution is selected. In the example below a sample of 100 observations have been simulated from the normal distribution.

Below is the summary statistics of the sample of 100 observations.

Descriptive Statistics of sample	
Mean	29.79658
Standard Error	0.542975
Median	29.5755
Standard Deviation	5.429754
Sample Variance	29.48222
Kurtosis	-0.4755
Skewness	0.08972
Range	24.76618
Minimum	17.1121
Maximum	41.87827
Sum	2979.658
Count	100

The graph on the left is the histogram based on equal bin lengths, the graph on the right is based on the AIC optimal number of bins.

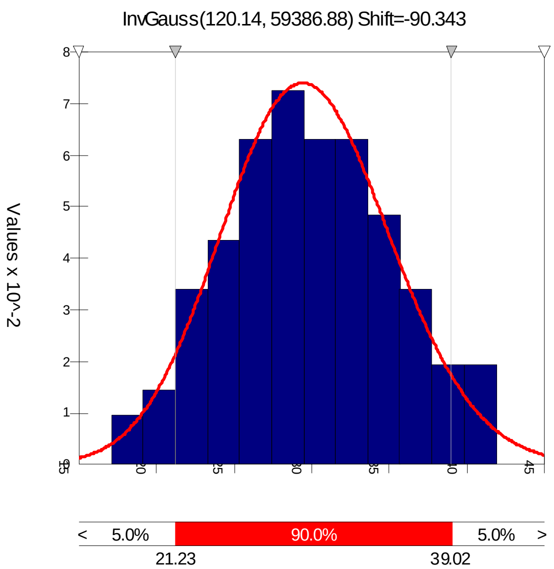




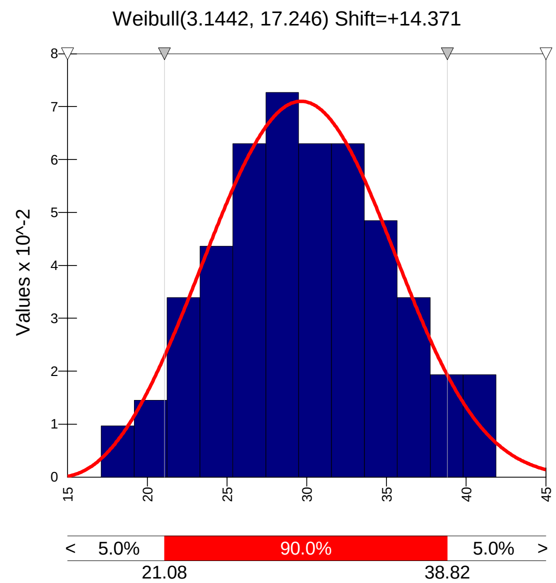
We used two other methods in calculating the bins when attempting to fit the distribution of the simulated set of values. The methods used were the equal probabilities & equal intervals.

In all, three methods were used, including the AIC option to define an optimal number bins. The graphs below show how the fitted distribution altered based on the approach selected and that the AIC bins provided the best fit to the sampled data.

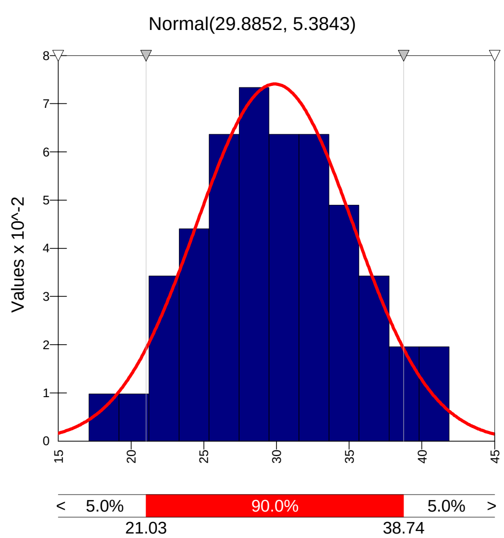
Graph 1: The InvGauss distribution was selected when attempting to fit the simulated normal distributed data. The equal probability approach was used to select the number of bins.



Graph 2: The Weibull distribution was selected when attempting to fit the simulated normal distributed data. The equal bin approach was used.



Graph 3: Based on using the AIC optimal bins. The normal distribution was selected when attempting to fit the simulated normal distributed data.



4. Using the AIC to obtain an optimal number of homogenous groupings.

Similar to the application of the AIC for selection of optimal number of bins, we have further looked at how we could use such an approach to identify homogenous groups. This, for example, is of use when trying to group postcodes or selecting sum insured bands in a pricing review. Often you would use statistical techniques such as cluster analysis which requires specification of the number of clusters required. It is often difficult deciding how many groupings one should use. An optimal number of clusters/bands or groupings can be established by applying a similar approach as to finding the optimal number of bins using the AIC.

Dayton (1998) proposed the use of AIC to identify optimal subsets of means or proportions based on independent groups. This is similar to conducting an analysis of variance and testing the hypothesis of equality of the means. When testing a large number of groups and for non-parametric alternatives it would involve testing a large number of pairs.

An example of another practical use of AIC approach is to group individual policyholder age into homogeneous age bands in the context of private motor pricing. The recommended approach to produce relativity for policyholder age is fitting a continuous polynomial for the most of ages, except where one should allow for age specific characteristics.

The most common approach in modelling risk premiums is to fit individual frequency and size models for each claim type then combine to form a risk premium model. If the individual frequency and size model adopted a polynomial function for age relativity, then the final risk premium relativities for age is also smooth without any further fitting process. This final smoothed relativity for age is ideal for rating purposes.

If the current rating system that an insurer has can allow relativities for individual age, then the final smoothed relativity without any grouping should be used. However, it is often required in practice that the continuous relativities to be grouped into age bands mostly due to the inflexibility of rating systems. We have used AIC approach to group continuous relativities into optimal number of groups.

The following summarises the steps we have taken and presents our findings.

- From ages 16 to 100, a smoothed final risk premium relativity for age is simulated and sorted in an order.
- Basic parameters for AIC method are calculated. (eg. Interval width, Initial number of intervals, minimum/maximum points and initial histogram)
- Then using an iterative process on different combination of groups, the optimal combination of groups with the lowest AIC was found.
- Compare the relativities suggested by AIC groupings to that of smoothed relativity.

AIC values for some models of the Policyholder Age example.

Rank	Model	AIC	AIC-minAIC
1	(4,2,8)	191.1974	0
2	(4,1,8)	191.6099	0.4125
3	(4,2,6)	191.7835	0.5861
4	(11,5,2)	191.8737	0.6763
5	(4,1,7)	191.8796	0.6822
6	(10,1,7)	191.9160	0.7186
7	(4,6,8)	192.0251	0.8277
8	(5,6,7)	192.0368	0.8394
9	(10,2,6)	192.2324	1.0349
10	(11,6,1)	192.2366	1.0392
11	(4,7,7)	192.4745	1.2771
12	(2,2,8)	192.7216	1.5242
13	(4,2,4)	192.7269	1.5295
14	(6,5,7)	192.7354	1.5380
15	(11,4,3)	192.7364	1.5390
16	(10,3,2)	192.7784	1.5810
17	(5,6,1)	192.7969	1.5995
18	(5,3,7)	192.8144	1.6170
19	(5,5,8)	192.8414	1.6440

The next graph compares two relativities;

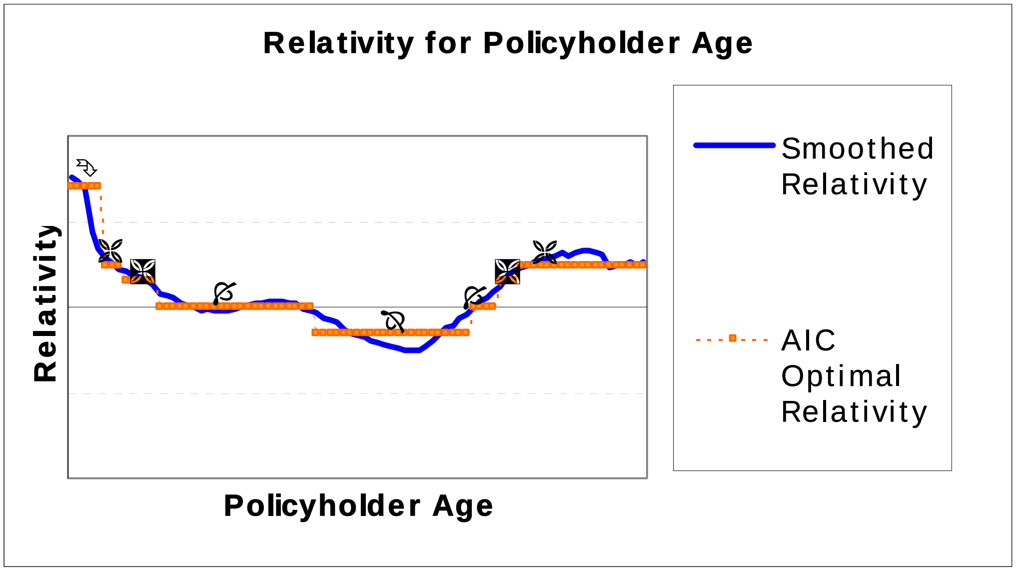


Figure 1 Shows the optimal number of groupings identified by the AIC.

In the exercise, the optimal AIC was found to have five distictive groups.

5. Conclusion

In this paper, we have described the use of the Akaike Information Criterion to select the optimal number of bins when constructing a histogram. Through simulation of data it was also shown that by using the bins from the AIC approach improved the accuracy of estimating the appropriate distribution for a particular data set. An example was also provided as to how the AIC could be used to provide optimal groupings for particular rating variables in a pricing analysis.

References:

1. Dayton, C (1998) Best Subsets using the Information criterion, University of Maryland.
2. Kullback, S & Leibler, R.A. (1951). On information and sufficiency. Annuals of Mathematical Statistics.
3. Sakamoto, Y. (1985). Categorical Data Analysis by AIC. Kluwer Academic Publishers, Dordrecht, Holland, London, Boston, 1985),
4. Taylor, C.C., (1987). Akaike's Information Criterion and the Histogram. Biometrika, 74, 636-639,