IJST EXPLORE International Journal of Science and Technology
CENTRE OF PROFESSIONAL RESEARCH PUBLICATIONS
ISSN 2049-7318

International Journal of Science and Technology Volume 3 No. 12, December, 2014

On Model Selection Criterion in Capture –Recapture Experiments with Sparse Data

Danjuma Jibasen* Ezra Gayawan**
* Department of Statistics and Operations Research, Modibbo Adama University of Technology, Yola, Nigeria
** Department of Mathematical Sciences, Redeemer's University, Redemption city, Nigeria

ABSTRACT

Model selection involving sparse data is always difficult, this is because with sparse data, quite different models can appear to fit adequately with highly diverse point, it is also almost impossible to test the underlying assumptions and select the "best" model". Some ecological as well as epidemiological experiments result in sparse data. In this paper we propose a modified Akaike information criterion (call it AIC$_J$) for selecting models in capture-recapture experiments resulting in sparse data. The proposed criterion was compared with the Akaike Information Criterion (AIC) and likelihood ratio test G^2. It was found that AIC$_J$ performed well in comparison to AIC and G^2. We therefore recommend that instead of concluding that model selection criterion may give misleading results for sparse data, or even that selection criterion does not exist, AIC$_J$ can be used. The proposed criterion is based on the link between interval estimation, hypothesis testing and model selection

Keywords: *Akaike Information Criterion, Likelihood ratio test, hypothesis testing, confidence interval estimation, sparse data*

1. INTRODUCTION

Model selection is an important issue in almost any practical data analysis. Here, we are faced with the question of what model should be used to best approximate reality, given the data at hand. The basic assumption is that good data, relevant to the issue are available and these have been collected in an appropriate manner. A common problem is variable selection in regression, given a large group of covariates including some higher order terms, one needs to select a subset to be included in the regression model, (Pan, 2001). Model selection is needed in capture-recapture (C-R) experiments, where a large number of models are usually identified. Such a choice of models may be necessary to handle the dependence between capture occasions and the extreme heterogeneity that is often observed among individuals. In C-R experiments, several model selection methods have been used. These among others includes; discriminant analysis (Otis *et al.*,1978), Likelihood ratio tests (LRTs) where available Akaike Information Criterion (AIC) (Burnham, Anderson and White, 1994) Bayes' information criterion (BIC) IWGDMF (1995) and bootstrap – based LRTs (Norris and Pollock, 1996). These selection strategies are useful, but the experimenters' knowledge of the biological situation plays an important role in the choice of the "best model".

Model selection is a bias verses variance trade off and this is the principle of parsimony; models with too few parameters have high variance, whereas models with many parameters may seem to fit the data (Forster, 2000). In the same token models with few parameters may have wider confidence interval length (CIL) compare to a model with many parameters. Model with narrower CIL are usually preferred (Mood, *et al*., 19774; Otis *et al*., 1978). This is the connection between interval estimation, hypothesis testing and model selection.

In earlier work in C-R experiments, a natural approach to interval estimation was to assume that $\widehat{N}$ had a large sample (asymptotic) normal distribution and to use $\widehat{N} + 1.96(S.E)$ as a 95% confidence interval. This approach, according to (IWGDMF, 1995) gives misleading results, because the actual distribution of $\widehat{N}$ in practice is skewed, for virtually all models. Hence IWGDMF (1995) suggested that $\widehat{N}$ be suitably transformed to a random variable. Chao (1989) proposed using log$_e$ $(\widehat{N} - r)$, where r is the number of different animals caught, (that is, the number of different people on all lists, in epidemiologic studies). Chao used the transform $\widehat{N}$ to construct 95% confidence interval for sparse data. In her words, "with sparse data, it is impossible to test the underlying assumptions and select the "best" model" Chao (1989). Hence, in her analysis of the turtle mud data, the proposed model was selected based on the simulation results not on any model selection criterion. She added that it is impossible to apply the model selection procedures given in Otis *et al*., (1978) to the turtle mud data. Agresti (1994) also, assert that with sparse data quite different models can appear to fit adequately with highly diverse point and interval estimates of N, hence, he, added that knowledge of the biological context may provide some guidance in choosing a model and estimate of N. These and other views form the bases for this work. We therefore proposed a model selection criterion based on the confidence interval for the transformed $\widehat{N}$, given in Chao (1989).

2. HYPOTHESIS TESTING, INTERVAL ESTIMATION AND AIC

Consider, a typical two-sided hypothesis about the population mean μ, for σ^2 unknown. The hypothesis under consideration is; H$_o$: $\mu = \mu_o$ verses H$_a$: $\mu \neq \mu_o$. To obtain this test as usual, we could use the generalized likelihood ratio principle or find some statistic, say,

$$t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \sim t_{n-1} \qquad (1)$$

where the critical region is given as; reject H$_o$ if $t > k$. where, $k = t_{\alpha/2}{}^{n-1}$. This is equivalent to constructing a (1-α) 100% region;

$$(\bar{x} \pm t_{\frac{\alpha}{2}}\, s/\sqrt{n}) \qquad (2)$$

Thus the null hypothesis is rejected, if the confidence region does not contain μ_o.

It is clear that the interval estimation and hypothesis testing rely heavily on $s/\sqrt{n}$, the standard error.

Model selection can be based on hypothesis testing or through the use of information. Test procedures do not penalize for over parameterization (Sclove, 1978, see Gayawan and Ipinyomi, 2009). Whereas, information criteria are selection criterions which balance model fit and its complexity.

Akaike's procedures are called information-theoretic because they are Kullback-Leibler (K-L) information (Burnham and Anderson, 2004). K-L information is a measure (a 'distance' in an heuristic sense) between conceptual reality, *f*, and approximating model, *g*, and is defined for continuous functions as the integral.

$$l(f,g) = \int f(x) log_e\left(\frac{f(x)}{g(x|\theta)}\right) dx \qquad (3)$$

where *f* and *g* are n-dimensional probability distributions. K-L information denoted by *I(f,g)*, is the 'information' lost when model g is used to approximate reality, *f*. The analyst seeks an approximating model that loses as little information as possible. AIC is based on this principle.

The general AIC is based on the quantity

$$AIC = -2log_e(L) + 2K \qquad (4)$$

where, $log_e(L)$ is the log-likelihood evaluated at the maximum likelihood estimates of the parameters, and k is the number of parameters in the model. The first term is a measure of how well the model fits the data, and the second term is a penalty for the addition of parameters (and hence, model complexity). The model for which AIC is minimum is selected as the best for the empirical data at hand.

In least squares estimation with normally distributed errors for all R models in the set, and apart from an additive constant, AIC can be expressed as

$$AIC = nlog_e(\hat{\sigma}) + 2k \qquad (5)$$

Where,

$$\hat{\sigma}^2 = \frac{\sum \hat{e}_i{}^2}{n}$$

and $\hat{e}_i$ are the estimated residuals from the fitted model. In this case the number of estimable parameters, *k*, is the total number of parameters in the model, including the intercept and σ^2. Thus, AIC is easy to compute from the results of least squares estimation in the case of linear models or from results of a likelihood-based analysis in general, (Burnham and Anderson, 2001).

As powerful as it is, AIC has been criticized because of its inconsistent tendency to over fit models, (Gayawan and Ipinyomi, 2009). Hence, researchers modify the general AIC or compare it with other model selection criteria.

Burnham *et al.*, (1994) and Burnham and Anderson (2004) compare AIC and Bayes' Information Criterion (BIC) for capture-recapture experiments; Gayawan and Ipinyomi (2009) compare the relative performance of AIC, and adjusted R Square in selecting fertility models. Pan (2001) gave an AIC suitable for generalized estimating equations, Hossain (2002) developed a modified version of AIC for statistical model selection where the parameter of interest is restricted to be in the range [a, b]. Sanni and Jolayemi (2009) modified an AIC for contingency table.

3. METHODOLOGY

The philosophy behind model selection and hypothesis is that large variance yields a bad model; this is same as with confidence interval estimation. That is, the information lost is high. Thus, AIC, hypothesis testing and interval estimation are functions of the standard error or the error sum of squares (SSE) likewise the proposed criterion. The confidence interval in Chao (1989) treats $log_e(\hat{N} - r)$ as approximately normal random variable which gives a 95% confidence interval as,

$$\{r + \frac{(\hat{N}-r)}{c}, r + (\hat{N} - r)c\} \qquad (6)$$

Where,

$$C = exp[1.96\{log_e(1 + \frac{var(\hat{N})}{(\hat{N}- r)^2})\}^{\frac{1}{2}}]$$

r is the number of different individuals captured. This interval bound depends on *c*; if *c* is large the interval will certainly be wide. This is the crux of the proposed criterion. The proposed modified criterion is

$$AIC_J = 2\,log_e(c) + 2k \qquad (7)$$

Where *k* is the number of parameters in the model and *c* is as defined above. This criterion too works as the AIC and BIC, where the smaller the value the better the model. This criterion can be used to assess models with the same number of parameters as well. In the examples below, AIC_J was applied to different models where the numbers of variables are unavailable to us. We hence assigned k = 0, this was compared to the classical AIC.

4. EXAMPLES

4.1 Rabbits Redux 1

Otis, et al (1978), described a survey in which snowshoe hares were repeatedly captured in 6 consecutive trapping days. This has been tabulated and analyzed elsewhere (Cormack (1986); Coull and Agresti (1994) and Dorazio and Royle, 2003). Agresti (1994), fit log-linear and latent class models to the data and chose the “best model” using likelihood ratio statistic G^2. He observed that with small sample sizes, a model selection procedure may suggest a model that is much simpler than one that truly represents reality. The summary of the results of Agresti analysis is represented alongside AIC_J in Table 1.

Table 1: Comparison of the analysis of Results of Agresti (1994) with AIC_J

Model	Estimate of N	G^2	AIC_J
Log-linear			
Mutual independence	75.1	58.3	1.1152
Two-factor quasi-symmetric	90.5	50.7	1.8556
No three-factor interaction	104.5	32.4	2.0475
Latent class			
Quasi-symmetric (L = 2)	77.3	47.7	1.1946
Ordinary (L = 2)	85.2	41.2	1.4772
Ordinary (L = 3)	81.3	33.1	1.9692

The likelihood ratio test (G^2) and AIC_J both favor the mutual independence model among the loglinear models. Similarly, among the Latent classes both selection criterion selects and favors the Quasi-symmetric.

4.2. Rabbits Redux 2

Dorazio and Royle (2003) analyzed the Snowshoe hares data using classes of mixture models, including beta-binomial, logistic-normal, latent-class. Model selection was based on AIC and G^2. Here we present a summary of their analysis, together with AIC_J as shown in Table2.

Table 2: Comparison of the analysis of Results of Dorazio and Royle (2003) with AIC_J

Model	Estimate of N	G^2	AIC	AIC_J
Beta-binomial	90.8	62.0	66.0	3.7863
Logistic-normal M_h	91.7	61.7	65.7	2.4971
Logistic-normal M_{t+h}	91.9	52.8	66.8	2.5375
Latent-class M_h	76.7	58.5	64.5	1.5622
Latent-class M_{t+h}	76.4	49.7	65.7	1.4565
Latent-class M_{txh}	85.2	41.2	67.2	1.6063

It can be observed from Table 2 that within the logistic –normal classes G^2, AIC and AIC_J favors the model M_h. Within latent classes G^2 and AIC selected M_h with a population size estimates of 76.7 whereas AIC_J favors M_{t+h} with population size of 76.4. If population estimation is our aim (population estimation is the goal of most capture-recapture experiments), AIC_J also performs well.

4.3 Mud Turtle Data

Chao (1989) reported the capture-recapture experiment of Illinois mud turtle which was originally conducted by Bickham and Gallaway (1980), with 99 distinct turtles out of 104 captures. Chao developed a model for the data criticizing the use of Darroch maximum likelihood estimator (MLE) because the data is sparse. The summary of the work of Chao is compared with AIC_J in Table 3.

Table 3: Comparison of the analysis of Results of Chao (1989) with AIC$_J$

Model	Method	Estimate of N	95% conf. Interval	AIC$_J$
M_t	Darroch's MLE	1009	473, 2313	1.7782
M_t	Chao's (f_1=94, f_2 =5) bias corrected	805	399, 1762	1.7135
M_t	Chao's (f_1=94, f_2=5) non-bias corrected	946	438, 2215	1.8312
M_h	Chao's (f_1=94, f_2=5)	983	453, 2306	1.8299

Chao concludes that it is impossible to apply the model selection procedures given in Otis *et al*. to decide on a more appropriate model, hence, selected model M_h with population estimate of 983 based on the simulation results. Though she said, “if equal catchability is a reasonable assumption, there are about 800 turtle in the habitat”. The proposed AIC$_J$ selects Chao’s M_t (f_1=94, f_2=5) bias corrected, estimating a population size of 805, this model incidentally has the narrowest confidence interval length compare to the others. With all fairness AIC$_J$ perform well here.

5. CONCLUSION

Population size estimation is the aim of most C-R experiments, this study has therefore shown that AIC$_J$ perform credibly well in comparison to AIC and G^2, and even more advantageous because AIC$_J$ can be computed with the aid of a scientific calculator not necessarily with a computer program. It simplicity is appealing.

REFERENCES

Agresti, A. (1994). Simple capture-recapture models permitting unequal catchability and variable sampling effort. *Biometrics.* **50,**494-500.

Bickham, J. W. and Gallaway, B.J. (1980). A status report on the studies of the taxonomy of the Illinois mud turtle (Kinosternon flavescens spooneri) with supplementary notes on its distribution and ecology. Byran, Texas. LGL Ecological Research Associates, Inc

Burnham, K. P., Anderson, D. R., and White, G. C. (1994). Evaluation of the Kullback-Leibler discrepancy for model selection in open population capture-recapture models. *Biomet J* **36**, 299-315.

Burnham, K.P. and Anderson, D.R. (2001).Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, **28**. 111-119

Burnham, K.P. and Anderson, D.R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research* **33** 261-304

Chao, A.(1989). Estimating size of population size for sparse data in capture-recapture experiments. *Biometrics* **45**,427-438.

Cormack, R.M. (1985). Examples of the use of GLIM to analyze capture-recapture data. In Statistics in Ornithology, B.J.T. Morgan and P.M. North (eds), 243-273 New York. Springer-Verlag.

Coull, B.A. and Agresti, A. (1999). The use of mixed logit models to reflect heterogeneity in capture-recapture studies. *Biometrics* **55**, 294-301.

Dorazio, R.M. and Royle, J. A. (2003). Mixture models for estimating the size of a closed population when capture rates among individuals. *Biometrics* **59**,351-364.

Forster, M.R. (2000). Key concepts in model selection: Performance and generalizability *Journal of Mathematical Psychology* **44**, 205-231

Gayawan, E and Ipinyomi, R.A. (2009). A Comparison of Akaike, Schwarz and R Square Criteria for Model Selection Using Some Fertility Models. *Australian Journal of Basic and Applied Sciences* **3(4)** 3524-3530

Hossain, M.Z. (2002). Modified Akaike Information Criterion (MAIC) for Statistical Model Selection. *Pak. J. Statist.* **18(3)** 383-393.

International Working Group for Disease Monitoring and Forecasting (IWGDMF) (1995a). Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development. *Am. J. Epidemiol*. **142**,1047-58.

Mood, A. M., Graybill, F. A. and Boes, D. C. (1974). *Introduction to the Theory of Statistics*. New York: McGraw-hill

Norris, J. L. and Pollock, K.H. (1996a). Nonparametric MLE under two capture-recapture models with heterogeneity. *Biometrics* 52, 639-649.

Otis, D. L., Burnham, K. P., White, G. C., and Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monograph* **62**, 1-135

Pan, W. (2001) Akaike’s Information Criterion in Generalized Estimating Equations. Biometrics, **57** 120-125.

Sanni, O. O. M., and Jolayemi, E. T. (2009). On Selection of an Optimal Model using AIC Approach. *Journal of Mathematical Sciences* **20**, 161-170

Sclove, S. T. (1987). Application of model selection criteria to some problems in multivariate analysis Psychometrika, 52(3): 333-343