COMPARISON OF STATISTICAL MODELS FOR CPUE STANDARDIZATION BY INFORMATION CRITERIA POISSON MODEL VS. LOG-NORMAL MODEL

Hiroshi SHONO

National Research Institute of Far Seas Fisheries

7-1, 5-chome, Orido, Shimizu-shi, 424-8633, Japan

ABSTRACT

The analysis by generalized linear model (GLM) has been used for the standardization of CPUE, recently. Such calculation has usually been performed through GLM and/or GENMOD procedure of SAS/STAT package assuming that CPUE model with log-normal distribution and/or catch one with Poisson distribution. In order to check about which model is better, log-normal or Poisson, we suggest how to compare the two models (log-normal and Poisson) statistically by information criteria like AIC in this paper. To use the sane response variables enabled statistical comparison between two models, CPUE-lognormal and catch-Poisson. This method is applicable in almost all of GLM models dealing with CPUE standardization.

1. INTRODUCTION

The analysis by generalized linear model (GLM) has been used for the standardization of CPUE, recently. Such calculation has usually been performed through GLM and/or GENMOD procedure of SAS/STAT package assuming that CPUE model with log-normal distribution and/or catch one with Poisson distribution. In order to check about which model is better, log-normal or Poisson, the analysis of CPUE trend and standard residual have often been carried out. However, the comparison of those two models by statistical method seemed not to be performed so far. Therefore, we suggest how to compare the two models (log-normal and Poisson) statistically by information criteria like AIC in this paper. This method is applicable in almost all of GLM models dealing with CPUE standardization.

2. PROCEDURE OF MODEL COMPARISON

A. Calculation in Poisson model

Step A-1. Model selection in Poisson model

Catch model with Poisson distribution is usually defined as follows:

E[Catch] = Effort*exp{(Intercept)+(Year)+ … +(Interactions)},

Catch~ Poisson(?) (2.1)

where E(Catch) is the expectation of catch. We need to set the log(Effort) as the offset. (log shows national logarithm.)

If we assume more than two cases about explanatory variables and/or interactions in above Poisson model, then we can select the best model by stepwise chi-square test based on the value of deviance (Dobson, 1990) or a kind of information criteria. In the case of using stepwise test, we need to assume one full model at first. We suggest using these information criteria if there are not a lot of candidate models, otherwise doing stepwise test, because of practical reason.

In this calculation, over-dispersion parameter f should not be estimated but fixed to 1.0. Because, if estimating the parameter f , then the definition of likelihood and information criteria become very difficulty and complexity.

Remark) If random variable X follows the Poisson distribution with parameter ? , the mean and variance of X show ? and f ? by the over-dispersion parameter, respectively (r.v. X ~ Po(?)? E[X]=? , Var[X]= f ?).

Step A-2. Calculation of Maximum Log-Likelihood in Poisson model

We calculate the MLL (maximum log-likelihood) in a finally selected model (or all candidate models) in step A-1. MLL of Catch-Poisson model with parameter ? is described as follow:

$$\log L(\hat{l} \mid Catch) = \sum_{i=1}^{n} (Catch_i \log \hat{l}_i - Catch_i) - \sum_{i=1}^{n} \log(Catch_i!) \qquad (2.2)$$

$$= (\text{SAS/OUTPUT of GENMOD with Poisson}) - \sum_{i=1}^{n} \log(\Gamma(Catch_i + 1))$$

where

n is the number of observations

G is the gamma function,

$\hat{l}$ is MLE (maximum likelihood estimator) of ? . ($l := (l_1, \cdots, l_n)$)

In an output of GENMOD procedure with Poisson distribution of SAS/STAT package, the second term of right side in formula (2.2) is omitted. Therefore, we must add this term to MLL of SAS/OUTPUT.

B. Calculation in log-normal model

Step B-1. Model selection in log-normal model

Common LN (log-normal) model is defined as follows:

E[log(CPUE+constant)]=(Intercept)+(Year)+ … +(Interactions),

log(CPUE+constant)~ N(m, s^2)) (2.3)

where E[CPUE+constant] is the expectation of CPUE plus constant term and log is the national logarithm.

If there is no zero-catch data, then the constant term is not necessary. In order to compare with Poisson model by information criteria, we must use the same response variables as Poisson case. Therefore, we assume the following model in this paper, although the assumption of model is a little different from the formula (2.3).

E[log(Catch+constant)]=log(Effort)+(Intercept)+(Year)+ … +(Interactions),

$$\log(\text{Catch}+\text{constant}) \sim N(\mu, \sigma^2) \qquad (2.4)$$

where E[ ] shows the expectation. It is necessary to set log(Effort) as the offset.

We choose the best model by stepwise chi-square (or F) test or a kind of information criteria from among candidate models expressed with the form of the formula (2.4) as well as Step A-1.

Step B-2. Calculation of Maximum Log-Likelihood in log-normal model

We calculate the MLL (maximum log-likelihood) in a finally selected model (or all candidate models) in step B-1. MLL of Catch-LN (log-normal) model with parameter μ and σ^2 in formula (2.4) is described as follow:

$$\log L(\hat{\mu}, \hat{\sigma}^2 \mid Catch + \text{constant}) = -\frac{n}{2}\log(2p\hat{\sigma}^2) \qquad (2.5)$$

$$-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}[\log(Catch_i + \text{constant}) - \hat{\mu}_i]^2 - \sum_{i=1}^{n}\log(Catch_i + \text{constant})$$

$$= (\text{SAS/OUTPUT of GENMOD with normal}) - \sum_{i=1}^{n}\log(Catch_i + \text{constant})$$

where

n is the number of observations,

$\hat{\mu}$ is MLE (maximum likelihood estimator) of μ ($\mu := (\mu_1, \cdots, \mu_n)$),

$\hat{\sigma}^2$ is MLE (maximum likelihood estimator) of σ^2.

Formula (2.5) shows MLL about response variable, Catch plus constant term. However, MLL about Catch is also described as function (2.5), because the probability density function of Catch plus constant term is equal to that of Catch. (We can check this easily by the change of variable from Catch plus constant term to Catch.) Therefore, we can calculation the MLL about Catch variable by formula (2.5) whether the constant term is equal to zero or not, and regardless of magnitude of constant term. In addition, we can decide the value of constant term among the candidate models using various information criteria based on the MLL defined by formula (2.5). For instance, which is better as a constant term, 0.01 or 0.1?

C. Comparison of two models by information criteria.

We choose the best one model with comparing the value of information criteria in these two models that are selected in Poisson and log-normal (or all candidate models in both distributions). Following three criteria AIC (Akaike, 1973), BIC (Schwarz, 1978) and c-AIC (Sugiura, 1978) are well-known, and these are defined as follows:

AIC = - 2*(MLL) + 2*p

BIC = - 2*(MLL) + p*log(n) (2.6)

c-AIC = - 2*(MLL) + 2*n*p / (n- p- 1)

where

p is the number of unknown parameter.

AIC is very popular and widely used. However, c-AIC is more efficient than AIC , especially in small samples (Shono, 2000).

3. AN EXAMPLE OF CPUE S TANDARDIZATION

We compared the two models (Poisson and log-normal) using virtual data in the manual of SAS/STAT package (SAS, 1996). The data used was shown in Table 1. Although the data show the reliability of equipment, we regarded this as CPUE data by longline fisheries.

We compared three models with having the following explanatory variables in each error structure (Poisson and log-normal.)

1. $E[\log(Catch_{ij}+\text{constant})] = \log(\text{Effort}) + (\text{Intercept}) + (\text{Year}_i) + (\text{Season}_j) + (\text{Year*Season})_{ij}$
2. $E[\log(Catch_{ij}+\text{constant})] = \log(\text{Effort}) + (\text{Intercept}) + (\text{Year})_i + (\text{Season})_j$
3. $E[\log(Catch_{ij}+\text{constant})] = \log(\text{Effort}) + (\text{Intercept}) + (\text{Year})_i$

(3.1)

where

Effort: Hooks per 1000,

i(Year): 1983-1990,

j(Season): 1-4 (1: Jan.-Mar., 2: Apr.-Jun., 3: Jul.-Sep., 4: Oct.-Dec.).

log(Effort): offset term.

In addition, we assumed the two constant terms (0.5 and 1.0) in log-normal model. We calculated the value of three information criteria (AIC, BIC and c-AIC) in all models (i.e. nine models in total). The results of model selection and SAS program of GLM calculation are shown in Table 2 and in Appendix, respectively. It is concluded that LN.2-3 on Table 2, in case of log-normal model with having only effect of year (model-3 in formula (3.1)) and 1.0 as a constant term, is the best model by any information criteria in this example.

ACKNOWLEDGEMENT

We acknowledge Dr. K. Hiramatsu, National Research Institute of Far Seas Fisheries, for his useful technical advices.

REFERENCES

AKAIKE, H. (1973): Information theory and an extension of the m,aximum likelihood principle. *2nd International Symposium on Information Theory*. Petrow, B. N. and Csaki, F. (eds.), Akadimiai Kiado, Budapest, p.267-281

DOBSON, A. J. (1990): *An introduction to generalized linear models*. Chapman and Hall. 174pp.

SAS(1996): SAS/STAT Changes and Enhancements through Release 6.11.

SCHWARZ, G. (1978): Estimating the dimension of a model. *Ann. Statist.*, **6**:

p.461-464

SHONO, H. (2000): Efficiency of finite correction of AIC. *Fisheries Science*, **66**: p.608-610

SUGIURA, N. (1978): Further analysis of the data by Akaike's information criterion and the finite corrections. *Common. Statist. –Theor. Meth.*, **7**: p.13-26

Table 1. *Virtual data for CPUE standardization by longline fisheries (loosely based on the data, SAS, 1996, p290)*

No. of data	Year	Month	Hooks	Catch	No. of data	Year	Month	Hooks	Catch
1	1983	1	1000	1	49	1987	1	5749	2
2	1983	2	1010	0	50	1987	2	5682	4
3	1983	3	1026	0	51	1987	3	6460	3
4	1983	4	1192	0	52	1987	4	6681	3
5	1983	5	1238	0	53	1987	5	7215	3
6	1983	6	1374	1	54	1987	6	6650	8
7	1983	7	1356	0	55	1987	7	7094	2
8	1983	8	1358	4	56	1987	8	6600	6
9	1983	9	1594	0	57	1987	9	7649	3
10	1983	10	1786	1	58	1987	10	7615	9
11	1983	11	1885	2	59	1987	11	6733	4
12	1983	12	1981	2	60	1987	12	6540	10
13	1984	1	1044	0	61	1988	1	6680	4
14	1984	2	1266	1	62	1988	2	7646	6
15	1984	3	1533	1	63	1988	3	8139	4
16	1984	4	1510	2	64	1988	4	7829	4
17	1984	5	1818	2	65	1988	5	8220	3
18	1984	6	2210	7	66	1988	6	7671	5
19	1984	7	2003	2	67	1988	7	7120	3
20	1984	8	2489	4	68	1988	8	7293	4
21	1984	9	2841	2	69	1988	9	8045	5
22	1984	10	3236	3	70	1988	10	8567	3
23	1984	11	3104	5	71	1988	11	7682	6
24	1984	12	2919	3	72	1988	12	7048	3
25	1985	1	2849	3	73	1989	1	7369	2
26	1985	2	3119	2	74	1989	2	7270	6
27	1985	3	3596	7	75	1989	3	8124	1
28	1985	4	4003	5	76	1989	4	7636	5
29	1985	5	4067	5	77	1989	5	7512	5
30	1985	6	3690	2	78	1989	6	7049	4
31	1985	7	3509	2	79	1989	7	7286	2
32	1985	8	3653	9	80	1989	8	7624	2
33	1985	9	4186	3	81	1989	9	7623	2
34	1985	10	4562	2	82	1989	10	7970	5
35	1985	11	4277	2	83	1989	11	7569	1
36	1985	12	3838	2	84	1989	12	7156	10
37	1986	1	4253	3	85	1990	1	7404	3
38	1986	2	4242	5	86	1990	2	7447	8
39	1986	3	5119	5	87	1990	3	7951	12
40	1986	4	5281	7	88	1990	4	8065	7
41	1986	5	5163	4	89	1990	5	7742	3
42	1986	6	4977	2	90	1990	6	7109	2
43	1986	7	4663	4	91	1990	7	7229	4
44	1986	8	5465	4	92	1990	8	7279	3
45	1986	9	5314	5	93	1990	9	7366	0
46	1986	10	5485	4	94	1990	10	7955	6
47	1986	11	5688	6	95	1990	11	7044	6
48	1986	12	5403	0	96	1990	12	3929	3

Table 2. *Results of model selection by three information criteria (AIC, BIC and c-AIC) using the virtual data on Table 1.*

Model	Explanatory variables	n	p	LL(SAS)	constant	MLL	AIC	BIC	c-AIC
Po.-1	Year,Season,Year*Season	96	32	142.3134	318.6423	-176.329	416.6578	498.717	450.1817
Po.-2	Year,Season	96	11	125.8096	318.6423	-192.833	407.6654	435.8733	410.8083
Po.-3	Year	96	8	124.3941	318.6423	-194.248	404.4964	425.0112	406.1516
LN.1-1	Year,Season,Year*Season	96	32	-66.4273	113.2055	-179.633	423.2655	505.3246	456.7893
LN.1-2	Year,Season	96	11	-82.2179	113.2055	-195.423	412.8467	441.0545	415.9896
LN.1-3	Year	96	8	-83.1693	113.2055	-196.375	408.7495	429.2643	410.4047
LN.2-1	Year,Season,Year*Season	96	32	-45.2627	130.9323	-176.195	416.3899	498.4491	449.9137
LN.2-2	Year,Season	96	11	-61.0725	130.9323	-192.005	406.0095	434.2174	409.1524
LN.2-3	Year	96	8	-62.2655	130.9323	-193.198	402.3955	422.9103	404.0507

Remark) Constant term is set to 0.5 in the models for LN.1-1~ 1-3, set to 1.0 in the model for LN.2-1~ 2-3.

APPENDIX. SOURCE CODE OF SAS PACKAGE FOR CPUE STANDARDIZATION BY LONGLINE FISHERIES USING VIRTUAL DATA O N TABLE 1.

```
** Model comparison (Poisson vs. Log-Normal) ** ;
** by Hiroshi SHONO (NRIFSF,Japan) 2000/4/16 ** ;
**  1. Data step ** ;
option linesize=120 pagesize=200 ;
data example ;
infile 'c:\NewWork\IOTC2001\example2.prn' ;
input number year month hooks catch ;
*** Remark) the process of reading the virtual data on Table 1.
effort = hooks / 1000 ;
cpue = catch / effort ;
catch1 = catch + 0.5 ;
catch2 = catch + 1 ;
Leffort = log(effort) ;
Lcatch1 = log(catch1) ;
Lcatch2 = log(catch2) ;
LLcatch = log(gamma(catch+1)) ;
season = 0 ;
if (month= 1 or month= 2 or month= 3) then season=1 ;
if (month= 4 or month= 5 or month= 6) then season=2 ;
if (month= 7 or month= 8 or month= 9) then season=3 ;
if (month=10 or month=11 or month=12) then season=4 ;
proc means data = example sum ; var LLcatch Lcatch1 Lcatch2 ;
*** Remark) calculation of additional term of log-likelihood ;
run ; quit ;
** 2. Proc step ** ;
** A. Poisson model ** ;
** calculation Po.-1 ** ;
proc genmod data = example ;
*   make 'obstats' out=glmoutP1 ; class year season ; model catch = year season year*season
/ dist = poisson
link = log
offset = Leffort
noscale
type1
type3 ;
* obstats ;
run ;
quit ;
** calculation Po.-2 ** ;
proc genmod data = example ;
* make 'obstats' out=glmoutP2 ;class year season ;
model catch = year season
/ dist = poisson
link = log
offset = Leffort
noscale
type1
type3 ;
* obstats ; run ; quit ;
** calculation Po.-3 ** ;
proc genmod data = example ;
*   make 'obstats' out=glmoutP3 ;
class year season ;
model catch = year
/ dist = poisson
link = log
offset = Leffort
noscale
type1
type3 ;
* obstats ; run ; quit ;
** B. Log-Normal model ** ;
** B-1. Constant term is set to 0.5 ** ;
** calculation LN.1-1 ** ;
proc genmod data = example ;
*   make 'obstats' out=glmout11 ;
class year season ;
model Lcatch1 = year season year*season
/ dist = normal
link = identity
offset = Leffort
type1
type3 ;
* obstats ;
run ; quit ;
** calculation LN.1-2 ** ;
proc genmod data = example ;
*  make 'obstats' out=glmout12 ;
class year season ;
model Lcatch1 = year season
/ dist = normal
link = identity
offset = Leffort
type1
type3 ;
*   obstats ;
run ;
quit ;
** calculation LN.1-3 ** ;
proc genmod data = example ;
*   make 'obstats' out=glmout13 ;
class year season ;
model Lcatch1 = year
/ dist = normal
link = identity
offset = Leffort
type1
type3 ;
*   obstats ;
run ; quit ;
** B-2. Constant term is set to 1. ** ;
** calculation LN.2-1 ** ;
proc genmod data = example ;
*   make 'obstats' out=glmout21 ;
class year season ;
model Lcatch2 = year season year*season
/ dist = normal
link = identity
offset = Leffort
type1
type3 ;
*   obstats ;
run ; quit ;
** calculation LN.2-2 ** ;
proc genmod data = example ;
* make 'obstats' out=glmout22 ;
class year season ;
model Lcatch2 = year season
/ dist = normal
link = identity
offset = Leffort
type1
type3 ;
* obstats ;
run ; quit ;
** calculation LN.2-3 ** ;
proc genmod data = example ;
*   make 'obstats' out=glmout23 ;
class year season ;
model Lcatch2 = year
/ dist = normal
link = identity
offset = Leffort
type1
type3 ;
* obstats ;
run ; quit ;
```