SCRS/2015/029

SOME CONSIDERATIONS FOR CPUE STANDARDIZATION; VARIANCE ESTIMATION AND DISTRIBUTIONAL CONSIDERATIONS

Matthew V. Lauretta, John F. Walter[1], and Mary C. Christman

SUMMARY

Two stage statistical models, such as the delta lognormal, that explicitly model the distribution of the proportion positive and the non-zero observations are widely used for CPUE standardization. Estimation of the variance of the index is obtained as the variance of the product of two random variables. Many current treatments assume or explicitly test for the independence of the two components and use a covariance term to estimate the index variance. Subsequent work indicates that this is incorrect, that much of the code used to estimate the covariance should be replaced and the two components are, under most situations, independent such that the covariance is zero. This allows for an exact variance estimate based on Goodman (1960). Existing code should be revised to reflect this development. Most CPUE treatments also assume a delta-lognormal model when other distributions may be more appropriate and may obviate the need for a two-stage model. We present alternatives to two-stage models that implicitly assume a lognormal distribution for the positive observations, for cases when other distributions may be more appropriate. We also present a set of decision rules for selecting the appropriate discrete distribution with examples of simulated data that demonstrate the various distributional forms, along with statistical codes, with the goal of improving CPUE modeling.

KEY WORDS

CPUE standardization, delta lognormal model, variance

[1] U.S. Department of Commerce National Marine Fisheries Service, Southeast Fisheries Science Center Sustainable Fisheries Division 75 Virginia Beach Drive. Miami, Florida 33149 USA Contribution SFD-2009/013
Email: John.f.walter@noaa.gov

1. Introduction

Many catch per unit effort standardizations employ a two-stage modeling approach where the proportion positive ($\hat{P}_i$) by year is modeled separately from the catch when positive ($\hat{C}_i$). Then the index for year i ($\hat{I}_i = \hat{P}_i\hat{C}_i$) is obtained as the product of the two components under the assumption that the dependent variable is a mixture distribution of a binomial and another non-zero distribution (lognormal, truncated poisson, etc). These types of models are variously called two-stage or hurdle models, of which the delta-lognormal model commonly used in ICCAT is a specific example. For convenience we will use the term two-stage in this paper.

To obtain an estimator of the variance of the two stage model index several estimators have been proposed. An early development of the delta-lognormal estimator by Lo et al ( ) employed a variance estimator, that uses the covariance between the two components, under the assumption that the two components were functionally related:

$$\hat{V}(\hat{I}) \approx \hat{P}^2\hat{V}(\hat{C}) + \hat{C}^2\hat{V}(\hat{P}) + 2\hat{C}\hat{P}Cov(\hat{C},\hat{P}) \qquad (1)$$

Note that this variance estimator is an approximation for the full variance estimator where the last four terms have been removed.

Further consideration of this variance estimate by led to a proposal (Walter and Ortiz, 2011) to test the significance of the correlation between ($\hat{C}_i$). and ($\hat{P}_i$) and, if significant, use the variance estimator proposed by Lo et al 1992 (1) and, if nonsignificant, use Goodman's exact estimator for the variance of the product of two random variables:

$$\hat{V}(\hat{I}) = \hat{P}^2\hat{V}(\hat{C}) + \hat{C}^2\hat{V}(\hat{P}) - \hat{V}(\hat{P})\hat{V}(\hat{C}) \qquad (2)$$

Subsequently Christman, *pers comm*, determined through simulation that the estimator of Lo et al 2002 and the proposed two-step method of Walter and Ortiz (2012) are unnecessary as, under most situations, the two parts ($\hat{C}_i$) and ($\hat{P}_i$) are independent which would make the Goodman (1960) proposal the most appropriate variance estimator.

A second consideration for much of the modeling is the appropriate distribution for the dependent variable. Many papers assume a lognormal distribution for the positive catch, often even when the catch consists of a discrete number of fish, which would more appropriately be modeled with a discrete distribution, of which there are three logical choices, the binomial (also called the logistic model), the poisson, or the negative binomial distribution. In this paper we present the case that many CPUE treatments would be more appropriately modeled with a discrete distribution assumption where effort is treated as a linear model intercept offset, and in many cases, these distributional forms might obviate the need for a two stage model (as they can allow zeros). We present a simple decisional rule for selecting the appropriate discrete distribution and examples of simulated data that demonstrate the various distributional forms. Furthermore, we outline cases where, depending on the magnitude of zero inflation and data overdispersion, it may be appropriate to use a zero-inflated model; and when a two-stage model would be most appropriate in cases when the sampling occurs in two phases where the first phase of sampling is used to identify the presence of the fish and second phase of sampling represents the capture of individuals (e.g., spotter plane searches followed by purse seining of schools).

In this paper we develop the two concepts, above, and propose updated R and SAS code to obtain the variance for two-stage and discrete CPUE standardization models.

2. Materials and methods

2.1. The covariance estimator proposed by Lo et al (2002) does not actually estimate the correct covariance

Christman (*pers comm*) has pointed out that the estimate of the covariance proposed by Lo et al (2012) and used by Walter and Ortiz (2012) does not estimate the covariance of ($\hat{C}_i$) and ($\hat{P}_i$) *within* a year, but rather the covariance

between annually varying estimates of ($\hat{C}_i$) and ($\hat{P}_i$). The correct covariance in eq(1) would be the covariance between estimates of ($\hat{C}_i$) and ($\hat{P}_i$) within a year, given below:

$$C\hat{o}v(C,P) = \sum_{i=1}^{n}(C_i - u_c)(P_i - u_p)/n-1 \qquad (4)$$

Where u_c and u_p are the respective means of C and P within a year and *n* is the sample size.

Lo et al (2002) proposed an estimator that used the Pearson correlation between the annual estimates of C and P calculated across years, which does not estimate (4):

$$Cov(\hat{C},\hat{P}) \approx \hat{\rho}_{\hat{C},\hat{P}}[\sigma(\hat{C})\sigma(\hat{P})]$$

where σ standard error of the year effect predictions of C and P. The use of the correlation across years is incorrect and Christman (*pers comm)* argues that any perceived covariance is a simply a relationship between annual estimates of ($\hat{C}_i$) and ($\hat{P}_i$).

2.2. Under the assumption of simple random sampling, the two components ($\hat{C}_i$) and ($\hat{P}_i$) and independent, obviating the need for the covariance term.

The covariance term in (2) assumes that ($\hat{C}_i$) and ($\hat{P}_i$) are parameters from a common joint, mixture distribution. If this distribution is randomly sampled, then the covariance between the two components is zero. This can be shown by a simple example borrowed from Christman (*pers comm*) where we create a mixture distribution that is the product of a lognormal random variable $C = LogN(u * a, \sigma^2)$ and a Bernoulli random variable $P = \text{binomial}(n, a)$ where *n*=1. The product of the two, $I = P * C$ is then the distributional assumption of a delta-lognormal models. Further we impose a functional relationship such that the mean of the lognormal distribution is a function of the binomial probability (*a*) under a situation where increasing abundance means that we encounter a positive observation more often and, when we do, the positive catch rate is higher. If we repeatedly sample from each distribution and calculate the covariance and the correlation, we see that both have expected values of zero (Figure 1). Hence the covariance term in eq (1) is zero. Christman (*pers comm*) takes the example further to evaluate the situation where both components are functionally dependent upon some covariate, say an environmental factor and shows that the covariance is also zero.

2.3. Modeling catch data using discrete distributions

In many cases, fisheries catch data represent the count of individuals captured in each sample, and therefore discrete models are appropriate for obtaining mean, variance, and confidence interval estimates. Effort is appropriately modeled as an intercept offset in the generalized linear models, assuming the mean catch shifts according to the amount of fishing effort. This treatment of the data differs from modeling the catch rate as a continuous variable, which then uses a log-transformation to scale the data to meet normality assumptions, and requires the use of the delta-method to account for zeros. The discrete model approach treats zeros and positive observations as one distribution of counts, which the reduces two-stage model to a single regression, and more accurately reflects the discrete nature of the data. When taking this approach, the critical first step is to graph the frequency of counts in a histogram to gain a sense of the range, mode, and form of the distribution. Here we present examples of simulated discrete distributions which can be used as a guide for choosing the appropriate model (**Figure 2**). There is also a set of general rules of thumb that can assist in the selection of discrete models based on the observed annual mean and variance of catches (Young and Young 1998):

Binomial distribution: mean > variance
Poisson distribution: mean = variance
Negative binomial distribution: mean < variance

The binomial distribution (**Figures 2a to 2x**) is appropriate when modeling two potential outcomes, e.g., the presence or absence of a species in the catch. Often this situation arises when a species is rare or when one is only

concerned with the probability of encountering a species. Sometimes, the catch of more than one individual is rare enough that those observations can be treated as presence data along with the single observations, and the logistic regression (binomial model) adequately models means and variance trends over time (**Figure 2x**).

The Poisson distribution assumption should be used when the mean and variance are approximately equal (**Figures 2x to 2x**), although it has been our experience that this situation is rarely the case when modeling fisheries catch data. Rather, we prefer the negative binomial model, which is more flexible in the distributional form (**Figures 2x to 2x**), but can also take the shape of the Poisson distribution (**Figure 2x**), albeit, at the cost of an additional parameter, the variance scaling parameter or overdispersion parameter. In general, the negative binomial is an useful for modeling catch data, which often demonstrate considerable overdispersion, or large variance that is not accurately modeled under alternative distribution assumptions.

There are cases when the negative binomial model fails to account for both the large amount of zero catches, and a large range (and/or variance) of positive observations (e.g., **Figure 2x**). For some of these cases, a zero-inflated model is appropriate (**Figure 2x**), while for others a two-stage model may be more appropriate (**Figure 2x**). A zero-inflated model can be used to allocate the expected proportion of zeros based on the discrete distribution while allowing for a proportion of zero counts assumed to be structural and separate from the sample distribution by adding an additional parameter, the zero-inflation parameter or probability of additional zero counts. A two-stage model is appropriate when the positive catch observations are expected to occur from a separate process than the zero observations, i.e. some "hurdle" exists that must be overcome before a positive catch is observed. One classic example is catch data of schooling fish sampled by first identifying the presence of a school (e.g., spotter plane surveys), and then obtaining a catch of individuals in the school (e.g., purse seine hauls of identified schools). Therefore, there is information on the relative abundance of a species in both the proportion of samples that positively identified the presence of the species (e.g., number of schools), and the number of fish captured in the positive samples (e.g., size of the schools). The two-stage model is not, however, exclusive to this situation. In some cases, the range of positive observations is so large that a log-transformation is most appropriate to accurately model the mean and variance (**Figure 2x**). It is in these situations, that the delta-lognormal model or other delta-transformation model is more appropriate (e.g., Lo et al. 1992).

A comparison of the goodness-of-fit to the observed data provides a good method for model selection (Young and Young 1998). This test can be conducted on the data, aggregated across years and, ideally by year as each year is likely to have different mean. This should be done prior to the index standardization to avoid duplication of effort. We provide R code for data exploration, frequency histogram creation, and discrete distribution goodness-of-fit tests in **Appendix 3**. We provide SAS code for the generalized linear model standardizations using the various discrete distributions in **Appendix 4**, and note the simplification of estimating yearly least-square means, standard deviations, and confidence intervals using these models compared to the delta-method.

3. Results and discussion

Based upon the derivations in 1.1 and 1.2 we recommend that the variance for two stage estimators be calculated according to Goodman (1960) exact estimator, eq (2). We have proposed SAS and R code to do so in Appendix 1. This assumes that the two components are independent as shown briefly in 1.2 and more formally in Christman (*in prep*). This finding obviates the cumbersome testing of the significance of the correlation proposed by Walter and Ortiz 2012 and eliminates the need for the covariance estimator proposed by Lo et al (2002).

4. References

Goodman, L. A.1960. On the exact variance of products. Journal of the American Statistical Association. 55(292): 708- 713.

Lo, N. C. H., L. D. Jacobson, and J. L. Squire. 1992. Indices of relative abundance from fish spotter data based on delta-lognormal models. Can. J. Fish. Aquat. Sci. 49: 251 5-2526.

Walter, J. & Ortiz, M. 2012. Derivation of the delta-lognormal variance estimator and recommendation for approximating variances for two-stage cpue standardization models. Collect. Vol. Sci. Pap. ICCAT, 68, 365-369.

Young, L.J. and J.H. Young. 1998. Statistical Ecology. Springer. Pgs. 1-74.

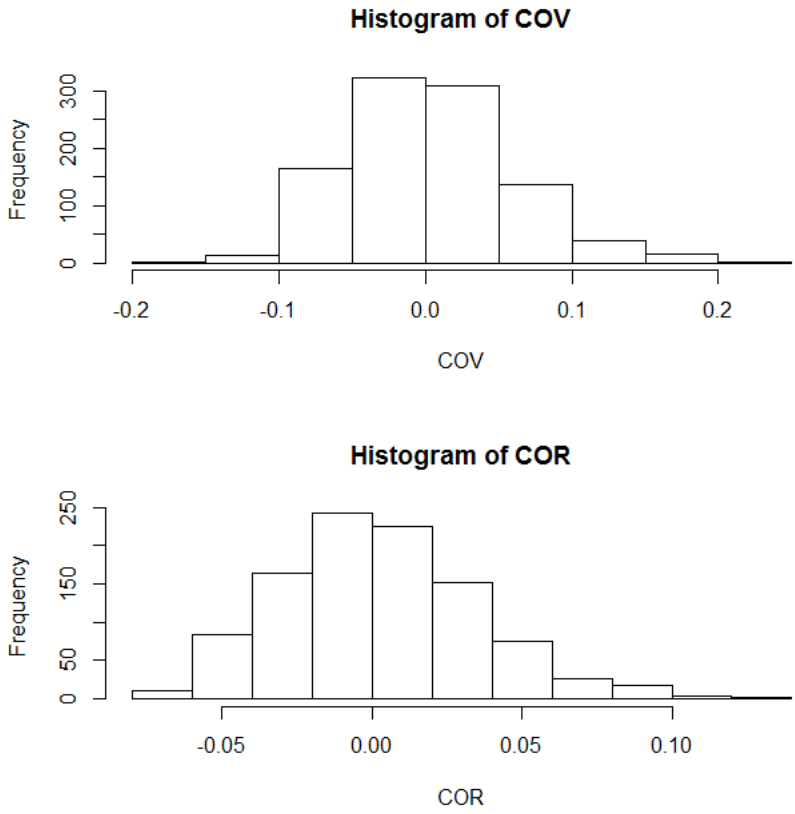

Figure 1. Histogram of 1000 estimates of the covariance and the correlation between the proportion positive and the lognormal mean for the a mixture distribution of a lognormal random variable $C = LogN(u * a, \sigma^2)$ and a Bernoulli random variable $P = \text{binomial}(n, a)$ where n=1. The product of the two, $I = P * C$ is then the distributional assumption of a delta-lognormal models.

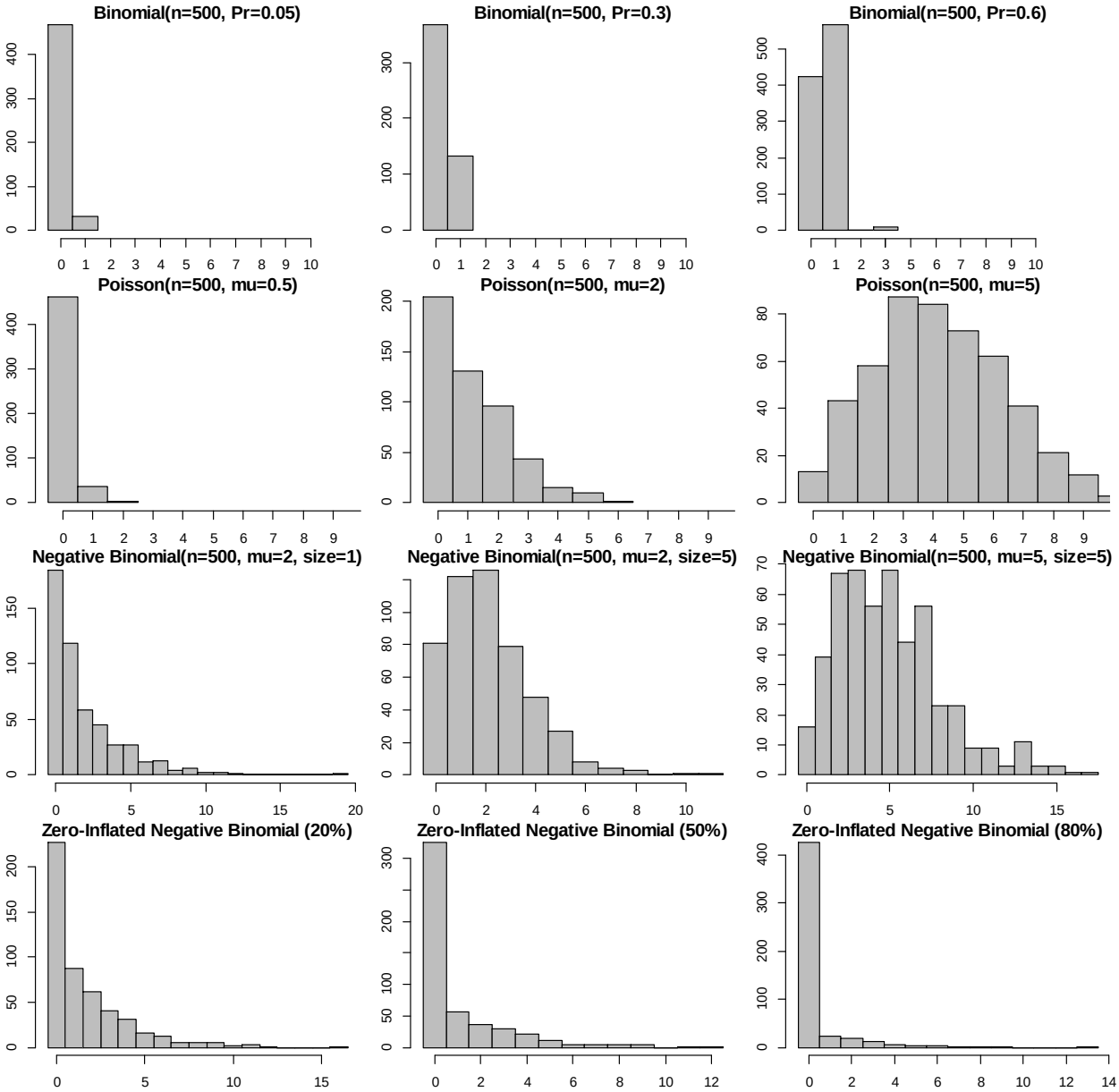

Figure 2. Examples of discrete distributions from simulated data.

Appendix 1. R code

```
gooodman.se <- function(p,var_p, c,var_c ){(var_p*c^2+var_c*p^2-var_p*var_c)^.5}

var_i= var_p* cpue ^2+var_c* ppos ^2-var_p*var_c)
```

Appendix 2. SAS code for Two-stage model variance estimation

```
var_i=var_p*cpue**2 +        var_c*ppos**2 - var_p* var_c;      *Goodman exact estimator
```

where var_i is the annual variance of the index, ppos is the annual estimate of the proportion positive (lsmeans in SAS) for the proportion positive, var_p is the variance of the of the proportion positive and cpue is the back-calculated (lsmeans in SAS) lognormal component and var_c is the variance for the lognormal component.

Appendix 3. Distribution Fitting and Goodness-of-fit Tests for Discrete Catch Models in R

```
count=rnbinom(1000,mu=2,size=1)
n=length(count)
hist=hist(count,breaks=seq(-0.5,max(count)+0.5,1),col=8)

# POISSON AND NEGATIVE BINOMIAL MODEL FITTING
theta=c(mu=1)
pois_negLL=function(theta)
                {
                -sum(dpois(count,theta[1],log=T))
                }
fit=optimize(pois_negLL,lower=0,upper=max(count))
fit
lines(seq(0,100),n*dpois(seq(0,100),fit$minimum),lwd=2,type='b',pch=18,col=3)

theta2=c(mu=1,k=1)
NB_negLL=function(theta2)
                {
                -sum(dnbinom(count,mu=theta2[1],size=theta2[2],log=T))
                }
fit2=optim(theta2,NB_negLL)
fit2
lines(seq(0,100),n*dnbinom(seq(0,100),mu=fit2$par[1],size=fit2$par[2]),lwd=2,type='b',pch=18,col=4)

# MODEL SELECTION
AIC=rbind(2+2*fit$objective,2*2+2*fit2$value)
dAIC=AIC-min(AIC)
colnames(dAIC)=c('dAIC')
rownames(dAIC)=c('Poisson','Neg Binom')
dAIC

# GOODNESS-OF-FIT TESTS
pois_chisq=chisq.test(hist$counts,p=dpois(seq(0,max(hist$mids)),fit$minimum),rescale.p=T)
pois_chisq
NB_chisq=chisq.test(hist$counts,p=dnbinom(seq(0,max(hist$mids)),mu=fit2$par[1],size=fit2$par[2]),rescale.p=T)
NB_chisq
```

Appendix 4. SAS code for Discrete Catch Regression Models

Logistic Regression Model in SAS for Discrete Presence/Absence Data

```
proc glimmix data=analysis;
nloptions maxiter=500;
class year month area;
ln_effort=log(effort)
cpue=catch/effort
model catch = year season area target / dist=binomial link=logit ddfm=kr s;
random year*month;
lsmeans year / ilink cl;
output out=GLMM_out pred(ilink) pearson;
id year month area target effort ln_effort catch cpue;
run;
```

Poisson Regression Model in SAS for Discrete Count Data

```
proc glimmix data=analysis;
nloptions maxiter=500;
class year month area;
ln_effort=log(effort)
cpue=catch/effort
model catch = year season area target / dist=poisson link=log offset=ln_effort ddfm=kr s;
random year*month;
lsmeans year / ilink cl;
output out=GLMM_out pred(ilink) pearson;
id year month area target effort ln_effort catch cpue;
run;
```

Negative Binomial Regression Model in SAS for Discrete Count Data

```
proc glimmix data=analysis;
nloptions maxiter=500;
class year month area;
ln_effort=log(effort)
cpue=catch/effort
model catch = year season area target / dist=negbin link=log offset=ln_effort ddfm=kr s;
random year*month;
lsmeans year / ilink cl;
output out=GLMM_out pred(ilink) pearson;
id year month area target effort ln_effort catch cpue;
run;
```

Testing the Zero-Inflated Negative Binomial Model for Discrete Count Data in R using glmmADMB package

Webpages: http://glmmadmb.r-forge.r-project.org/, http://glmmadmb.r-forge.r-project.org/glmmADMB.html

```
#requires R package "glmmADMB"

library(glmmADMB)
ln_effort=log(effort)
nbinom=glmmadmb(catch~year+season+area+target+offset(ln_effort),family='nbinom',link='log')
summary(nbinom)
ZI_nbinom=glmmadmb(catch~year+season+area+target+offset(ln_effort),family='nbinom',link='log',zeroInflation=T
RUE)
summary(ZI_nbinom)
AICtab(nbinom,ZI_nbinom)
```