Steps in index standardization using GLMs

(1) Identify response variables. If data were collected using a standardized effort unit (e.g., electrofishing catch/15min sampling event or catch/tow trawl surveys), model numbers caught (not CPUE). If concerned about changes in effort in the dataset, model catch as a function of effort (as an offset) and other covariates. If testing multiple models, make sure the response variables are the same.

(2) If constructing a young of the year (YOY) index, please censor data records that have captured fish larger than those classified as YOY using appropriate months of the year and length cutoffs by region specified in Table 5.3.8 in the assessment report. This will provide a recruitment index, which only captures YOY individuals. Alternatively, if providing an index for sizes classes larger than YOY, please censor records that captured YOY individuals or apportion individuals captured between YOY and non-YOY and censor the YOY fish. Change survey year to model year as necessary; model year begins March 1st.

(3) Identify explanatory variables and associated data type (e.g., categorical, continuous):

- Year will always be included as a categorical explanatory variable in all models.
- Include a small subset of other appropriate variables using the literature and expert judgment, if necessary. Do not include all potential variables - only ones that might be affecting **catchability (not abundance)** or you may standardize away the factors that actually affect trends in abundance.
- Scatterplot each potential covariate…
    - If obvious breaks or groupings appear, (e.g., seasons, depth/habitat categories, etc.) make that a categorical variable. Otherwise, make it a continuous variable. For all categorical variables, check to make sure you have an adequate number of samples in each category or your model will blow up. Lump categories if necessary or meaningful. If not, categories with no samples should be eliminated (data points removed from data set) because the model cannot provide estimates for that factor if there are no observations. If there are only a few observations in that category, try to run the model (if it blows up, you'll have to go back and remove it).
    - If two or more variables are highly (>0.9) or logically correlated, pick the one that makes the most sense biologically; for example, don't include both temperature and dissolved oxygen, or latitude and river system. If desperate, include interaction terms (with anything but year) as an initial test if you're not sure how things will pan out, but don't include interaction terms in the final model because it is nearly impossible to interpret and calculate the final year effect for an index.
    - Check if any factor is orders of magnitude different from others and adjust accordingly (turn 1,000,000 into 1 "million" to be on scale with other measurements in model).

(4) Censor data records from the completed data set, if less than 5% of the catch or less than 5% of the proportion of positive trips occurs for that factor. For example, if less than 5% of the total catch occurs at a station in the data set, censor the records of that station from the data set. This is meant to exclude factor levels in which you may be extremely unlikely to catch your species of interest and may therefore over inflate the number of zero catches in your data set.

(5) Plot histogram of number of animals caught.

- Look at the largest catch sizes. Do those largest catch sizes produce extremely large residuals? Do those largest catches likely reflect population dynamics?
    - Retain large catches if fitting is not a problem and if those observations do not produce large residuals that would be considered outliers.
    - Censor large catches if they produce large residuals considered to be outliers.
- Determine if there is a large gap between # of zeros and next highest bar (e.g., determine if you tend to either catch no animals or a lot of animals).
    - If so, use the delta approach, which models presence-absence with a binomial model and positive tows with a different distribution (usually lognormal).
    - Otherwise, proceed to other generalized/general linear models in next step.

(6) If delta methods are not appropriate, identify what distributional assumptions might be. Plot catch rate vs. variance in catch rate aggregated by each categorical factor and compare pattern with figures at the end of this document. A linear relationship supports an overdispersed Poisson error model, and variance in catch rate proportional to the square of the average catch rate suggests the log-normal and gamma error models. The negative binomial error model implies that the variance in catch rate is a function of both the average catch rate and the square of the average catch rate. Choose from below depending on outcome of mean-variance inspection. Avoid transformations of your response variable or covariates.

- If lognormal or gamma error models are implied, perform the gamma. [If you must use the lognormal, model catch as Gaussian with log link to avoid transforming catch. If you must for some reason model CPUE, use ln(CPUE+min(value/2))~.]
- If Poisson error model is implied, run the basic Poisson model (implying data are probably not overdispersed) and compare with the zero-inflated Poisson using the Vuong test. (Note: you will not be able to compare zero-inflated models with other submodels in step 6).
- If the negative binomial error models are implied, run the basic negative binomial model and compare with the zero-inflated negative binomial using

the Vuong test. (Note: you will not be able to compare zero-inflated models with other submodels in step 6).

(7) Select the appropriate canonical link function (relates mean of response variable to explanatory variables) for the model you've selected. Gamma – inverse. Poisson and negative binomial – log.

(8) If all factors in the final model are not significant, run all submodels and select best model as one with lowest AIC. If too many covariates are included for this to be practical, use stepwise selection of covariates (or better yet, reconsider what covariates you are including).

(9) Evaluate goodness-of-fit.

- Check for overdispersion; if phi is > 2 suggests overdispersion. NA for Poisson model.
- Plot standardized residuals against fitted values for the global model and by factor; presence of pattern may suggest overdispersion, miss-specification of link function, missing covariate, outliers

(10) If desired, perform back-transformation bias correction. Pull out year effects and SEs.

(11) To provide a measure of uncertainty for the standardized index, fit the model using bootstrapped data to provide 95% confidence intervals.

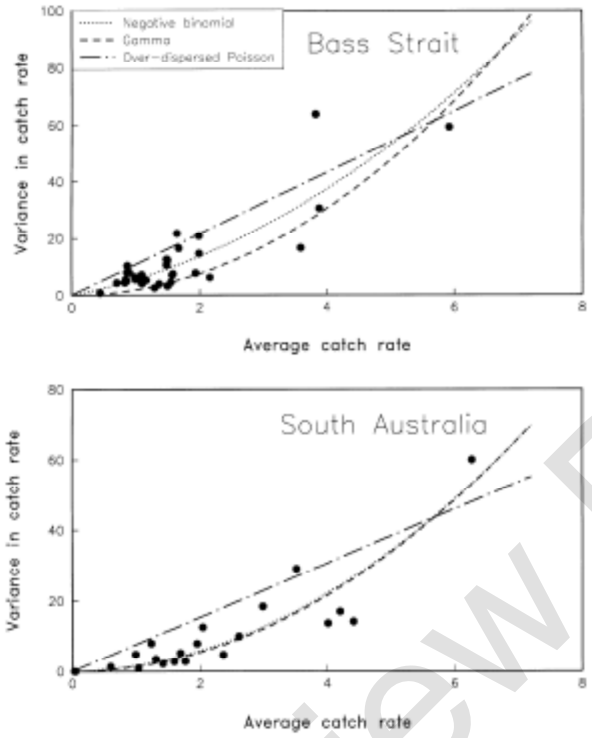

Fig. 4. Plots of the variance in catch rate against the average catch rate. Results are shown separately for the Bass Strait and South Australian zones. The dash–dotted line corresponds to the (over-dispersed) Poisson distribution, the dotted line to the negative binomial distribution, and the dashed line to the log-gamma distribution.

From Dong and Restrepo 1995 ICCAT report

TABLE 1. Variance components of four error distributions.

Distribution	Variance Function	Dispersion
Normal	$V_i(u_i) \equiv 1$	s^2
Poisson	$V_i(u_i) \equiv u_i$	1
Gamma	$V_i(u_i) \equiv u_i^2$	$1/\nu$
Neg. bin.	$V_i(u_i) \equiv u_i + ku_i^2$;	1