

# USEFUL METHODS FOR DEALING WITH OUTLIERS IN STOCK ASSESSMENTS

SCRS/1994/069

Col.Vol.Sci.Pap. ICCAT, 44 (1) : 151-154 (1995)

Restrepo\*, V.R., J.E. Powers\*\*

\* *University of Miami, Rosenstiel School of Marine and Atmospheric Science, MBF/CUFER, 4600 Rickenbacker Causeway, Miami, Florida 33149, USA*

\*\* *National Marine Fisheries Service, Southeast Fisheries Center, 75 Virginia Beach Drive, Miami, Florida 33149, USA*

## SUMMARY

It is often the case that certain observations appear to be inconsistent with the remainder of the set of data used in stock assessment models. The identification and treatment of these so-called "outliers" during stock assessment group meetings does not usually follow a set of criteria established *a priori*. It is not possible to justify the deletion of apparent outliers from analyses under such subjective circumstances. In this paper, we discuss statistical methods that can be used to detect outliers and methods that may be robust to the presence of outliers.

## RESUME

Il arrive souvent que certaines observations semblent ne pas concorder avec le reste du jeu de données utilisé dans les modèles d'évaluation de stock. L'identification et le traitement de ces points isolés dénommés "outliers" lors des réunions de groupes d'évaluation de stock ne suit généralement pas un ensemble de critères établis *a priori*. Il n'est pas possible de justifier la suppression de points isolés apparents des analyses dans des circonstances aussi subjectives. Le présent document traite de méthodes statistiques qui peuvent servir à détecter les points isolés, et de méthodes qui pourraient s'avérer robustes devant la présence de ces points isolés.

## RESUMEN

Es un caso frecuente que ciertas observaciones aparenten ser incoherentes con el residuo del conjunto de datos utilizado en modelos de evaluación de stock. La identificación y tratamiento de los denominados "outliers" (puntos situados fuera de la tendencia) durante las reuniones del grupo de evaluación de stock no siguen generalmente un conjunto de criterios establecidos *a priori*. No es posible justificar la supresión de aparentes "outliers" de los análisis bajo tales circunstancias subjetivas. En este documento se discuten los métodos estadísticos que pueden utilizarse para detectar "outliers" y métodos que puedan ser sólidos frente a la presencia de "outliers".

## Introduction

Stock assessment working groups at ICCAT and other organizations usually face the difficult problem of making management recommendations based on multiple data sets of different types, sources and quality. Typically, group participants present the various data set (with a discussion of how the data were obtained) and then the group collectively decides on sets of data to be used with a particular assessment model. This initial assessment may become what is known as the "base case", the assessment for which there is group consensus as being the most objective one, given the available scientific information. Alternatively, the initial assessment may reveal nonsensical results or something unexpected about the data and/or the model, in which case another set of data may be used for what becomes a new base case.

The process described above is undesirable from a statistical point of view because the scientists' (sometimes subconscious) perception of what the results ought to be will influence their decisions about what data to include or omit after the initial run. That is, the data are often manipulated *a posteriori* and this cannot be accomplished objectively. On the other hand, the initial assessment may contain observations or fitted values that appear to be inconsistent with the remainder of the set of data or fitted values. These so-called "outliers" may indeed contaminate the rest of the data and lead to an assessment which is not the best one possible.

It was faced with this dilemma that ICCAT's SCRS recommended in its 1992 swordfish report that research be carried out on the "Development of *a priori*, objective criteria for outlier detection and objective criteria for rejection of data on biological or fisheries grounds for the various time series data sets" (ICCAT 1993). This call for methods was again made in last year's swordfish and bluefin tuna reports (ICCAT 1994) and we consider it to be a recommendation that applies to all species reports. The aim of this paper is to more thoroughly discuss the problem of outliers in stock assessment and to present options for dealing with them. This is a necessary first step in responding to SCRS' request.

## The Nature of Outliers

Barnett and Lewis (1994) define an outlier as "*an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data*".

The phrase "appears to be inconsistent" is not meant to be the subjective impression that an individual will form about a particular observation. Instead, the main message in this definition is that a particular observation, *given a set of assumptions*, stands out as coming from a different statistical population. Consider a small sample of observations where all but one are clustered close to each other. The one observation that stands out as being different may well be an "outlier" if one assumes the data come from a normal distribution. On the other hand, it may seem perfectly reasonable if one assumes the data were drawn from a gamma or a lognormal distribution.

A problem we face in stock assessments is that the set of assumptions, implicit and explicit, is very complex. We are not just dealing with the choice between two types of univariate distributions. We are dealing with highly structured situations and often

overparameterized models. With tuned virtual population analyses (VPA) such as the ADAPT methodology used in many ICCAT species groups (Gavaris 1988, Powers and Restrepo 1992) we have catch data compiled, substituted and raised from many sources, and then assigned to ages by some model; we have relative abundance data also from many sources and which have already undergone sophisticated analyses (usually generalized linear models - GLMs); and, we also have a large number of assumptions such as the catch equation being deterministic and assumed selectivity values for various age groups in the last year of data. What appears to be an outlier with data-model set would not necessarily be an outlier with different assumptions in the model.

The statistical theory developed by others so far for detecting outliers covers much ground in terms of univariate analyses and simple structured situations such as linear regression (see Barnett and Lewis 1994, Weisberg 1985) and less ground for more structured situations with non-linear optimization. However, much research has been devoted in structured situations to develop *robust* methods for analyses, rather than procedures to detect outliers with the objective of deleting them from subsequent analyses. A robust method is one that is reasonably insensitive to (small) departures from the idealized assumptions being made. As we will show later, ICCAT species groups already incorporate a number of features in the analyses that may make the assessments more robust than they would be otherwise.

### Outlier Detection

Before getting into statistical methods for outlier detection, it is important to note that assessment working groups already do much in the way of protecting the analyses against potential outliers. Take for example the case of indices of relative abundance obtained from GLMs. It sometimes happens that new management measures are introduced in a given year such that the CPUE observations collected after that year do not possess the same characteristics as earlier ones. Perhaps the gear restrictions changed or some areas were closed for a few months, or fish under a given size were not fully reported as caught and were discarded, instead. In such cases, most scientists would opt for separating the relative abundance data into two distinct series or excluding the latter part from analysis. If the entire data set were used as a single time series of relative abundance, it would not be surprising to obtain poor results. In our opinion, there is no statistical procedure that is superior to this common-sense practice: If there is reasonable cause based on external fishery or biological evidence to dismiss some observations or to treat them differently during analyses, then this should be agreed to prior to the initial assessment. This may sound obvious but we feel it must be emphasized here. The nature of stock assessment meetings is such that a great number of things must be accomplished during a brief period of time. In doing so, we tend to forget asking ourselves the question: Are the data consistent with the assumptions we will make during analysis? For instance, scientists may have computed a very precise time series of relative abundance based on 100,000 observations. But, what if the fish or the fleet behavior is such that a systematic change in catchability over time is suspected? Then the data should generally not be included in a base case analysis that assumes equal catchability, regardless of the precision.

Statistical methods to detect outliers in structured models typically involve some analysis of the residuals. The most common procedure is to compute "studentized residuals", i.e., residuals that are scaled such that they all have more or less constant variance. In a simple linear regression, a studentized residual ( $r_i$ ) is computed as (Barnett and Lewis 1994, Weisberg 1985):

$$r_i = \frac{e_i}{\sqrt{\text{Var}(e_i)}} = \frac{e_i}{\sigma\sqrt{1-h_{ii}}}$$

where  $e_i$  is the residual from the fit,  $\sigma$  is the square root of the mean squared error, and

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}$$

The test used to detect an outlier is usually based on the statistic for the maximum absolute studentized residual:

$$t = \max |r_i|$$

When  $t$  is significantly large, the observation leading to it can be considered as an outlier. Note that because one is testing whether the largest  $r_i$  is an outlier without knowing *a priori* which observation will result in the largest  $r_i$ , one is really making  $n$  tests: one for each data point. This has an impact on the significance level that should be used because the probability that a  $t$  statistic with  $n-2$  degrees of freedom exceeds a given critical value is much smaller than the probability that the largest of  $n$  independent  $t$  tests exceeds the same value. Statistical texts recommend choosing critical levels based on the "Bonferroni inequality" which conservatively reduces the probability of incorrectly labeling at least one residual as an outlier. For the simple linear regression case, the critical value should be the  $[(\alpha/n) \times 100\%]$  point of  $t$ . So, for a case involving 100 residuals, a critical value corresponding to  $(0.05/100 = 0.0005)$  for each test would give an overall probability of incorrectly labeling an observation as an outlier of no more than 5%.

Tests similar to that using studentized residuals above have been developed for general linear models (e.g. Ellenberg 1976). The test for the presence of an outlier can be conducted by comparing the value of the maximum absolute studentized residual against

$$h_{\alpha} = \sqrt{\frac{(n-p)F}{n-p-1+F}}$$

where  $p$  is the number of parameters estimated and  $F$  is the upper  $\alpha/n$  point of the  $F$ -distribution with 1 and  $(n-p-1)$  degrees of freedom. Note that the use of the  $\alpha/n$  point also

invokes the Bonferroni inequality. Additional tests, based on maximum likelihood theory, have been developed to detect the presence of an outlier in general linear models and time series analyses (Barnett and Lewis 1994, Pankratz 1991). Many of these involve carrying out the regression  $n$  times and finding the maximum of the ratio between the residual sum of squares with all of the data and the sum of squares with one observation excluded from analysis (see Barnett and Lewis 1994). Such methods require much more computation time and usually yield the same results as those based on studentized residuals.

The situation with structured nonlinear methods like many we use in stock assessments is much more complex than that of the linear regression case. It is possible to use a similar procedure to detect an outlier, but we are only aware of research on the accuracy of such tests in very simple nonlinear regressions. In addition, we should emphasize that most tests are used to determine the presence of a *single outlier*, e.g. the observation yielding the maximum studentized residual. Tests for two or more outliers are more complicated and their power may be more severely affected by correlation among residuals. In addition, the presence of some outliers may mask the presence of others. Thus, we believe outlier detection tests to have potential but limited usefulness in stock assessment applications.

#### Accommodation Methods

Robust methods are those that are relatively insensitive to departures from the basic assumptions being made. One of the main objectives of robust procedures is for the parameter estimates to retain desirable statistical properties over a range of plausible distribution types. In the case of outliers, methods that allow for valid inferences to be drawn from the data and which are not seriously distorted by outliers are called *accommodation* procedures (see Barnett and Lewis 1994). As such, these methods are said to be robust to the presence of outliers.

There are many examples of procedures and algorithms which can be used to accommodate outliers in structured models such as the ones used in stock assessments. The ones presented in this section do not conform an exhaustive list, but are rather ones which we think can be very useful in working group meetings.

Two of the most robust estimation procedures widely used today are known as least median of squares (LMS) regression (Rousseeuw 1984) and least absolute value (LAV) regression (Bloomfield and Steiger 1983). As opposed to least squares regression where the *sum* of squared residuals is minimized, LMS minimizes the *median* of the squared residuals; LAV minimizes the sum of the absolute residuals. Simulation studies conducted in many linear and nonlinear applications have indicated that these procedures can be quite robust to the presence of outliers. In some simple cases, LMS produces good parameter estimates when even up to 50% of the data are contaminated. In a recent application of a linearized mortality estimation model, Chen and Paloheimo (1994) found that LMS and LAV were superior to least squares when catch or CPUE data had unequal variances or contained outliers. Results of their study cannot be generalized to the more complex case of an age-structured, nonlinear VPA, but robust regression holds such promise that it is worth investigating. LAV regression may also work well with nonequilibrium production models (Prager 1994) and age-structured methods (Kimura and Scott 1993). We recommend that

such possibilities be explored through simulation.

Transformations are typically used to homogenize variances or to linearize relationships. But they can also be used to provide some protection against outliers. For example, a transformation of the independent data can be used in the objective function being minimized in order to downweight extreme values. Logarithmic transformations have the effect of dampening large signals, but other transformations with more extreme downweighting can be found (see Seber and Wild 1989).

In many regression applications, the errors do not have equal variances. Weighted least squares can be used to include *a priori* information about unequal variances in the regression (Weisberg 1985). While most applications of weighted regression are intended to provide "minimum variance estimates" of the parameters or to reduce possible biases, the method can also have the effect of accommodating outliers. For example, examination of index of abundance values over time with their associated confidence limits often reveals what appear to be extremely high or low values associated with large standard errors. If the estimates of standard errors of the indices is used during the assessment, then the values with large errors will be given lower weights and will be less influential on the overall results. The ADAPT software used in ICCAT species groups (STAATS, Powers and Restrepo 1992) offers an option for weighting *individual observations* of relative abundance. Although this option has not been used by the SCRS during recent years, it may be very appropriate when external variances are available. Note that because of the inherent differences in aggregation and scale of the surveys used to collect index data, the external variances cannot be used to weight between indices. An alternative weighting option often used in ADAPT applications is that of iterative reweighting, where *each series* of relative abundance (as opposed to each individual observation) receives an overall weight, depending on the residual error for the series. We are not sure whether iterative reweighting provides a degree of protection against an entire relative abundance time series being an "outlier", but there is a strong possibility in the sense that the series may not conform to the model's assumptions. This is something that can also be explored via simulation.

Some assessment methods, both age structured and unstructured, yield results characterized by a "retrospective" pattern. The pattern is one of apparent over- or under-estimation of stock sizes or fishing mortality rates and may be due to multiple data-specific causes (Sinclair et al. 1990). Outliers, e.g. from misreported catches, can be one of the causes for a retrospective patterns. It is possible that LMS regression can alleviate retrospective symptoms, although, to our knowledge, this possibility has not been explored. Another option is the use of "shrinking" (ICES 1991) by which a penalty is incorporated in the objective function if the estimates of, say  $F_t$ , in the last year of data differ by a large factor from the mean estimate of previous  $F_s$ . Again, the usefulness of shrinking in terms of accommodating outliers could be explored through simulation.

#### Discussion

The practice of excluding observations subjectively and *a posteriori* from stock assessments without objective criteria is very undesirable. This was recognized by the SCRS early in the 1980's when age-structured models started to be tuned with multiple relative

abundance indices. At the time, a rule was suggested whereby an index was excluded from further analyses when the correlation between the fishing mortality predicted by the VPA and the observed index effort fell below a pre-specified value with some probability. However, it became difficult to impose such criteria in practice and, therefore, almost all indices available at the group meetings ended up being used and given equal weights. Hence, conflicting trends were rationalized by the preponderance of evidence.

As we have discussed above, there are statistical methods which can be used to detect individual outliers and, if necessary, these should be used. However, most of these methods have been developed for simple models such as simple linear regression and for the detection of a single outlier. We deal with much more complex nonlinear models in stock assessments and more than one outlier may be present. Our recommendation to ICCAT species working groups is not to use these methods blindly in all situations. Instead, we recommend that prior to the initial assessment, the scientists involved carefully scrutinize the data available and their conformance to the model with which they will be analyzed. If there are data (time series, individual observations) that are known on biological or fishery grounds to be highly suspect, then they should not be included in the initial analysis, if possible. If two indices for the same age groups are indicative of conflicting trends, it would be best to decide which one not to use based on biological or fishery grounds. But we realize that making such choices will not always be possible. Therefore, we suggest that external variances might be used to weight individual observations and iterative reweighting be used to weight between indices. We also recommend that assessment methodologies in use or being introduced be examined for their robustness to the presence of outliers. Robust minimization procedures such as least median of squared and least absolute value regression should be explored, as they appear to hold good promise in accommodating outliers in stock assessment applications.

#### Acknowledgments

Support for this work was provided through the Cooperative Unit for Fisheries Education and Research (CUFER) by National Oceanic and Atmospheric Administration (NOAA) Cooperative Agreement NA90-RAH-0075.

#### Literature Cited

- Barnett, V., and T. Lewis. 1994. *Outliers in Statistical Data*. Third edition. J. Wiley, New York.
- Bloomfield, P., and W.L. Steiger. 1983. *Least Absolute Deviations: Theory, Applications and Algorithms*. Birkhauser Verlag, Boston.
- Chen, Y., and J.E. Paloheimo. 1994. Robust regression approach to estimating fish mortality rates with a cohort-based model. *Trans. Amer. Fish. Soc.* 123: 508-518.
- Ellenberg, J.H. 1976. Testing for a single outlier from a general linear regression. *Biometrics* 32: 637-645.
- Gavaris, S. 1988. An adaptive framework for the estimation of population size. CAFSAC (Can. Atl. Fish. Sci. Adv. Comm.) Res. Doc. 88/29.
- ICCAT 1993. International Commission for the Conservation of Atlantic Tunas Report for the biennial period 1992-1993, Part I.
- ICCAT 1994. International Commission for the Conservation of Atlantic Tunas Report for the biennial period 1992-1993, Part II.
- ICES 1991. The Working Group on Methods of Fish Stock Assessment. ICES CM1991 / Assess:25.
- Kimura, D.K., and G.P. Scott. 1993. Length-based separable sequential population analysis as applied to swordfish, *Xiphias gladius*. ICCAT SCRS 93/51.
- Pankratz, A. 1991. *Forecasting with Dynamic Regression Models*. J. Wiley, New York.
- Powers, J.E., and V.R. Restrepo. 1992. Additional options for age-sequenced analysis. ICCAT Coll. Vol. Sci. Pap. 39: 540-553.
- Prager, M.H. 1994. A suite of extensions to a nonequilibrium surplus-production model. *Fish. Bull.* 92:374-389.
- Rousseeuw, P.J. 1984. Least median of squares regression. *J. Amer. Stat. Assoc.* 79: 871-880.
- Seber, G.A.F., and C.J. Wild. 1989. *Nonlinear Regression*. J. Wiley, New York.
- Sinclair, A., D. Gascon, R. O'Boyle, D. Rivard, and S. Gavaris. 1990. Consistency of some Northwest Atlantic groundfish stock assessments. NAFO SCR, Doc. 90/96.
- Wiseberg, S. 1985. *Applied Linear Regression*. Second edition. J. Wiley, New York.