УДК 639.2.053.7: 519. 226.2+234

Робастные функции правдоподобия для оценки параметров моделей «запас — промысел»

Д.А. Васильев, В.К. Бабаян

Всероссийский научно-исследовательский институт рыбного хозяйства и океанографии (ВНИРО, г. Москва)
e-mail: dvasilyev@vniro.ru

Рассмотрены подходы к повышению робастности функций правдоподобия при их использовании для оценки параметров моделей «запас — промысел».

Ключевые слова: модели «запас — промысел», робастные функции правдоподобия.

Введение

В настоящее время для оценки параметров интегральных рыбохозяйственных моделей, включающих в себя разнообразную и разнородную информацию, часто используются приёмы, основанные на максимизации функции правдоподобия. Связано это с тем, что функции правдоподобия позволяют статистически корректно объединить информацию об интересующем нас параметре, заключённую в различных видах используемых данных. Кроме того, функции правдоподобия позволяют применять байесовский анализ и привлекать к анализу иные виды информации вплоть до экспертных оценок. С другой стороны, функции правдоподобия, как классические, так и более сложные, к сожалению, исключительно не робастны.

Результаты и обсуждение

Робастность является ключевым понятием в статистике. Имеются различные определения данного понятия, однако чаще всего под робастностью понимается нечувствительность к малым отклонениям от допущений, лежащих в основе оценки [Huber, 1981; Hampel et al., 1986]. В свою очередь, понятие малого отклонения понимается с двух точек зрения, обе из которых важны: (1) небольшие отклонения для всех точек или (2) значительные отклонения для небольшого числа точек. Второе толкование понятия малых отклонений связано с понятием резко выделяющихся значений (так называемых аутлаеров).

Классические критерии оценки, например, основанные на методе максимального правдоподобия, оптимальны только в том случае, если модель сформулирована корректно, но они могут быть очень чувствительны к аутлаерам. В статистике известны критерии (статистические функции), имеющие более высокую робастность. Среди них: медиана, считающаяся наиболее робастной мерой положения; абсолютное медианное отклонение, часто рассматриваемое в качестве наиболее робастной меры масштаба разброса; критерии, основанные на мерах плотности (например, минимальное расстояние Хелинджера); некоторые непараметрические статистики. Они не являются идеальными, поскольку, например, медиана является робастной статистикой, но имеет более низкую эффективность (если не стремится к строгости формилирок, под эффективностью критерия можно понимать его способность найти наилучшее решение среди группы близких решений). Иногда предлагается комбинировать робастные свойства одних статистических критериев с высокой эффективностью других с помощью так называемых QIF-функций. Эти функции позволяют скомбинировать свойства различных критериев. Например, можно сформулировать адаптивный критерий, по своим свойствам находящийся между средним и медианой, который бы имел эффективность, приближённую к эффективности среднего, но в то же время был достаточно робастным [Park, Lindsay, 1999]. Пусть $X = (X_1, \ldots, X_n)$ является случайной выборкой, а $F_n(\cdot)$ является эмпирической функцией распределения, заданной как

$$Fn(X) = {}^{1}\!/_{n} \sum_{i=1}^{n} I(x_i \leq X),$$

где $I(\cdot)$ — индикаторная функция. QIF-критерий, основанный на расширенной результативной функции $g(\cdot)$, можно определить как:

$$\hat{\theta}_n = \arg\inf_{\theta} Q(\theta; g, F_n) = \arg\inf_{\theta} Q(\theta; g, X).$$

Парк и Линдсей [Park, Lindsay, 1999] сравнили QIF-критерий с другими известными критериями оценки параметра положения, например, с оценкой максимального правдоподобия, средним, медианой, оценкой Ходжеса-Лемана, оптимальной R-оценкой, M-оценкой Хьюбера и оценкой Чан-Хе. Перечисленные выше оценки могут быть представлены следующим образом.

Оценка Ходжеса-Лемана:

$$\underset{i \leq j}{median}\left(\frac{X_i + X_j}{2}\right)$$

рассматривается как компромисс между средним и медианой. Эта оценка является асимптотически эффективной для логистического распределения и имеет асимптотическую относительную эффективность (ARE) относительно среднего не менее 0,86. Эта оценка является также весьма робастной, но может быть неэффективна для некоторых асимметричных распределений.

M-оценка Хьюбера может быть получена решением относительно θ следующего уравнения:

$$\sum_{i=1}^{n} \psi_k\left(\frac{X_i - \theta}{S_n}\right) = 0,$$

где $\psi\kappa(t) = \min\{\kappa, \max(t, -\kappa)\}$, а S_n является робастной оценкой масштаба разброса. В качестве робастного эстиматора (характеристики) масштаба разброса можно использовать абсолютное медианное отклонение AMD:

$$S_n = AMD_n \approx$$
$$\approx 1{,}4826\, \underset{i}{median}\{|\, X_i - \underset{j}{median}(X_j)\,|\}.$$

Критерий Чан-Хе основан на минимизации асимптотической дисперсии на классе всех линейных комбинаций среднего и медианы.

Численные эксперименты показали, что для нормального распределения QIF-критерии могут терять эффективность в случае малых выборок, но почти столь же эффективны, как наименьшая средняя ошибка, для больших выборок. Для загрязненных данных (данных, в некоторые значения которых внесена ошибка) этот критерий оказался наиболее эффективным среди всех перечисленных выше критериев. Результаты показали, что оценки Ходжеса-Лемана и Чан-Хе могут быть неэффективны для асимметричных ошибок.

В целом робастные статистики и процедуры разрабатываются с целью дать возможность избавиться от влияния аутлаеров на результаты оценки. Однако как показывает наш анализ, само понятие аутлаеров не столь однозначно. P. Hampel [2002] рекомендует различать аутлаеры и большие (gross) ошибки. Аутлаеры (или в более общем виде — резко выделяющиеся подструктуры) являются наблюдениями, которые сильно отличаются от большей части наблюдений или, точнее, которые не соответствуют структуре, определяемой большинством наблюдений. Указывается, что понятие аутлаеров является несколько искусственным понятием с размытыми границами и достаточно большой переходной зоной. Однако это понятие является полезным, если не забывать об этом. Так называемые «большие» ошибки возникают в случаях, когда в эксперименте что-то явно было не так, например, в результате поломки оборудования, неправильного выбора категорий и т.д. Естественно, доля «больших» ошибок зависит от качества всех данных, однако даже для данных, считающихся «хорошими» (т.е. данных, полученных с соблюдением всех статистических предосторожностей), как правило, имеется 1—10% «больших» ошибок. Аутлаеры могут являться «большими» ошибками, однако они могут также генерироваться статистическими распределениями с более широкими хвостами. Если известно, что некоторые точки являются точками с «большими» ошибками, то таким точкам в анализе естественно присвоить нулевой статистический вес, однако если это аутлаеры (т.е. результаты, полученные в результате правильно проведённых наблюдений), то это вовсе не значит, что им следует давать полный статистический вес в случае, если в анализе принимается модель нормального распределения. Тут следует рассмотреть возможность использования статистических моделей с более широкими хвостами.

Однако возникает вопрос, какую из моделей с более широкими хвостами использовать? В действительности рассматриваемая выборка может содержать аутлаеры, но может не содержать достаточной информации для однозначного выбора модели распределения. Эта проблема особенно важна, если оценка параметров модели проводится не путём минимизации некоторой функции потерь, которую можно сделать в известной степени «свободной» от выбора функции распределения, а путём максимизации функции правдоподобия, что является необходимым элементом байесовского анализа. Формулировка компонент общей функции потерь может сильно зависеть от свойств конкретных данных. Это может приводить к некоторым теоретическим проблемам в определении статистических весов для компонент функции потерь, например, если возникает необходимость объединения в ней суммы квадратов остатков для одного вида данных с абсолютным медианным отклонением для другого вида данных и т.д. Именно по этой причине часто отдается предпочтение работе с функциями правдоподобия, поскольку в идеале именно таким образом можно избежать проблемы взвешивания различных источников информации. Однако это преимущество часто является иллюзорным, если данные содержат аутлаеры и при этом применяются процедуры снижения их влияния.

Авторы [Chen, Fournier, 1999] исследовали «смешанное» распределение как способ получения робастного распределения, которое было бы менее чувствительно к аутлаерам и одновременно хорошо работало в отсутствие аутлаеров. «Смешанное» распределение может увеличивать величину своих хвостов в соответствии с наличием аутлаеров в данных. Они использовали распределение, в котором ошибки подчиняются либо закону распределения $N(0, \sigma^2)$ с вероятностью $1 - \rho$, либо распределению с более толстыми хвостами с вероятностью ρ. Авторы использовали функцию плотности вероятности вида:

$$\rho(x) = \frac{2}{\sqrt{\pi g \sigma}}[1 + \frac{x^4}{(g\sigma)^4}]^{-1},$$

где параметр g сможет использоваться для настройки ширины распределения с толстыми хвостами. Увеличение значения g утолщает хвосты, увеличивая тем самым вероятность присутствия экстремальных значений, далеких от среднего. Для этого двухкомпонентного «смешанного» распределения нормальное распределение, имеющее большую априорную вероятность $(1 - \rho)$, описывает «правильные» наблюдения, а распределение с жирными хвостами, имеющее малую априорную вероятность (ρ), описывает «проблемные» наблюдения (аутлаеры). Значение ρ было принято равным 0,05. Соответствующая функция правдоподобия для наблюдений может быть представлена в виде:

$$L(\beta, \sigma \mid X, Y) =$$
$$= \prod_{i=1}^{N} \{\frac{1-\rho}{\sqrt{2\pi\sigma}} \exp[-\frac{(Y_i - f_i(X_1, \ldots, X_k; \beta))^2}{2\sigma^2}] +$$
$$+ \frac{2\rho}{\sqrt{\pi g \sigma}}[1 + \frac{(Y_i - f_i(X_1, \ldots, X_k; \beta))^4}{(g\sigma)^4}]^{-1}.$$

Проблемой в данном подходе является выбор ρ, поскольку при различных его значения можно получить существенно разные результаты, а информации для его объективного выбора в общем случае нет.

В целом большинство подходов к робастизации функций правдоподобия нацелено на улучшение описания аутлаеров выбранной статистической моделью. Но остается вопрос: помогает ли это реально снизить влияние аутлаеров на результат анализа? Ф. Хэмпел [Hampel, 2002] подчеркивает, что наиболее часто применяемым способом является замена исходной параметрической модели, (например, модели нормального распределения) другими специальными и более сложными моделями. Строго говоря, эти модели столь же нереалистичны, что и исходная модель. Правда, если эти модели выбраны разумно, то они могут работать, но судить об этом можно только на основании теории робастности.

Менее противоречивым способом является выделение аутлаеров и присвоение им пониженных статистических весов в рамках анализа с помощью какого-либо простого закона распределения. Критерием для нахождения аутлаеров в многомерном случае может служить расстояние Махаронобиса:

$$(y-\hat{\mu})^{\mathrm{T}}\hat{\Sigma}^{-1}(y-\hat{\mu})$$

для вектора наблюдений y, где $\hat{\mu}$ и $\hat{\Sigma}$ — оценки вектора положения и ковариационной матрицы.

Предложен также метод робастизации оценок максимального правдоподобия с использованием так называемых Ψ-функций правдоподобия.

Метод выделения и придания понижающих статистических весов аутлаерам на основе Ψ-функций правдоподобия состоит в следующем. Пусть $f\ (y,\ \theta)$ является статистической моделью для имеющихся наблюдений $\{y_i$: $i = 1, \ldots, n\}$, где $f\ (y,\ \theta)$ является функцией плотности относительно некоторой несущей меры $\nu(dy)$, а Θ является пространством для параметров θ. Логарифмическая функция правдоподобия определяется как сумма логарифмов функций плотности:

$$L(\theta)=\frac{1}{n}\sum_{i=1}^{n} l(y_i,\theta),$$

где $l(y,\ \theta) = \log f(y,\ \theta)$. Без потери общности можно усовершенствовать метод максимального правдоподобия для учёта ситуаций, когда данные могут включать в себя аутлаеры.

Пусть Ψ является растущей дифференцируемой функцией. Ψ-функция правдоподобия определяется как:

$$\bar{L}_{\Psi}(\theta)=\frac{1}{n}\sum_{i=1}^{n}\Psi(l(y_i,\theta))-b_{\Psi}(\theta)$$

для наблюдений $\{y_i\}$, где:

$$b_{\Psi}(\theta)=\int \Psi^{*}(l(y,\theta)\nu(dy)$$

и

$$\Psi^{*}(z)=\int_{0}^{z}\exp(s)\frac{\partial\Psi}{\partial s}(s)ds.$$

Эстиматор (критерий оценивания) на основе максимума Ψ-функции правдоподобия, называемый Ψ-эстиматором, определяется как:

$$\hat{\theta}_{\Psi}=\arg\max_{\theta\in\Theta}\{\bar{L}_{\Psi}(\theta)\}.$$

Если $\Psi(z) = z$, то метод Ψ-функций правдоподобия сводится к обычному методу максимального правдоподобия, поскольку $b\Psi(\theta) = 1$. Известно, что метод максимального правдоподобия является выборочным аналогом метода минимизации отклонения Кулбака-Лейблера от действительного распределения в статистической модели. Отклонение Кулбака-Лейблера определяется по частоте отклонений $f(y)$ от $g(y)$ (плотностей относительно $\nu(dy)$) как:

$$KL(f,g)=\int_{y} g(y)\log\frac{g(y)}{f(y)}\nu(dy)$$

и, следовательно:

$$KL(f(\bullet,\theta),g)=-E_g\{l(y,\theta)\}$$

с учётом того, что постоянные множители в θ опускаются. Поэтому логарифмическая функция правдоподобия $L(\theta)$ эквивалентна $E_g\{l(y,\ \theta)\}$, если математическое ожидание E_g заменить на эмпирическую функцию наблюдений таким образом, что максимум функции правдоподобия будет эквивалентен минимуму отклонений Кулбака-Лейблера.

Другим подходом является использование лог-логистической функции в качестве генерирующей функции:

$$\Psi_{\eta}(z) = \log \frac{1+\exp(z+\eta)}{1+\exp(\eta)},$$

$$\Psi^*_{\eta}(z) = \exp(z) + \exp(\eta)\log \frac{1}{\exp(z)+\exp(\eta)}.$$

Тогда функция весов будет иметь вид:

$$\psi_{\eta}(z) = \frac{\partial}{\partial z}\Psi(z) = \frac{1}{1+\exp(-z-\eta)}.$$

Параметр η называется параметром насыщения.

Другим вариантом выбора генерирующей функции является:

$$\Psi_{\beta}(z) = \frac{\exp(\beta z)-1}{\beta};$$

$$\Psi^*_{\beta}(z) = \frac{\exp\{(1+\beta)z}{1+\beta},$$

где $\beta > 0$. Эстиматор, основанный на $\Psi\beta(z)$ и называемый β-эстиматором, получается за счёт максимизации целевой функции вида:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{f(y_i,\theta)^{\beta}}{\beta} - \int \frac{f(y,\theta)^{1+\beta}}{1+\beta}\,d\nu(y).$$

Этот подход фактически совпадает с М-эстиматором [Huber, 1981].

Если записать:

$$\rho(y,\theta) = \Psi(l(y,\theta)) - b_{\Psi}(\theta),$$

то Ψ-эстиматор эквивалентен М-эстиматору для ρ.

Для получения Ψ-эстиматора $\theta\Psi$ предложен следующий алгоритм. Оценка $\theta^{(k)}$ на шаге k преобразуется в оценку $\theta^{(k+1)}$ на следующем шаге по формуле:

$$\hat{\theta}^{(k+1)} = \arg\max_{\theta}\{\frac{1}{n}\sum_{i=1}^{n}\psi(l(y_i,\theta^{(k)}))l(y_i,\theta) - -b_{\Psi}(\theta)\},$$

где $k = 0,1,\ldots$, начиная с некоторого разумно выбранного начального приближения θ_0. Тогда выражение для оценки $\theta^{(k+1)}$ можно записать как:

$$\frac{1}{n}\sum_{i=1}^{n}\psi(l(y_i,\theta^{(k)}))S(y_i,\theta) = = E_{\theta}\{\psi(l(y,\theta))S(y,\theta)\},$$

где $\psi(z) = (\partial/\partial z)\Psi(z)$ и $S(y, \theta) = (\partial/\partial\theta)\,\ell(y, \theta)$.

Выводы

Использование функций правдоподобия для оценки параметров рыбохозяйственных моделей является ключевым элементом в процедуре оценки запасов, однако их практическое применение должно основываться на использовании достаточно робастной модели для выявления наблюдений, содержащих большие ошибки, и их коррекции.

Литература

*Cadigan N.G., Myers R.*A. 2001. A Comparison of Gamma and Lognormal Maximum Likelihood Estimators in a Sequential Population Analysis // Canadian Journal of Fisheries and Aquatic Sciences. № 58. P. 560–567.

Chen Y., Fournier D. 1999. Impacts of Atypical Data on Bayesian Inference and Robust Bayesian Approach in Fisheries // Canadian Journal of Fisheries and Aquatic Sciences. № 56. P. 525–1533.

Deriso R.B., Quinn T.J. II, Neal P.R. 1985. Catch-Age Analysis with Auxiliary Information // Canadian Journal of Fisheries and Aquatic Sciences. № 42, N 4. P. 815–824.

Doubleday W.G. 1976. A Least Squares Approach to Analysing Catch at Age Data // ICNAF Res. Bull. 12. P. 69–81.

Fournier D.A., Hampton J., Sibert J.R. 1998. MULTI FAN-CL: a Length-Based, Age-Structured Model for Fisheries Stock Assessment, with Applications to South Pacific Albacore // Canadian Journal of Fisheries and Aquatic Sciences. № 55. P. 1–12.

Gudmundsoon G. 1986. Statistical Considerations in the Analysis of Catch-at-Age Observations // J. Cons. Int. Explor. Mer. № 43. P. 83–90.

Hampel F. 2002. Some Thoughts about Classification // Research Report No. 102. January 2002. Seminar fur Statistik Eidgenossische Technische Hochschule (ETH) CH-8092, Zurich, Switzerland. 19 p.

Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. 1986. Robust Statistics: The Approach Based on Influence Functions. New York: Wiley.

Huber P. J. 1981. Robust Statistics. NY: John Wiley & Sons.
Park Ch., *Lindsay B. G.* 1999. Robust Estimation and Tests based on Quadratic Inference Function. Technical Report #99–02. May, 1999. Center for Likelihood Studies, Department of Statistics, The Pennsylvania State University. PA 16802. 42 p.
Passarin K. 2004. Robust Bayesian estimation. Università Dell'insubria, Facoltà di Economia, 11/2004.
Shertzer K.W, *Prager M.* H. 2002. Least Median of Squares: A Suitable Objective Function for Stock Assessment Models? // Canadian Journal of Fisheries and Aquatic Sciences. № 59. P. 1474–1481.
Tsou T.— Sh., *Richard R. M.* 1995. Robust Likelihoods // Journal of American Statistical Association. March 1995. Vol. 90. No. 429. P. 316–320.
Vasilyev D. 2004. Winsorization: Does It Help in Cohort Models? // ICES CM2004/K:45.
Vasilyev D. 2005. Key Aspects of Robust Fish Stock Assessment. M.: VNIRO Publishing. 105 p.

Robust Likelihood Functions for Estimation of Parameters of Stock-Fishery Models

D. Vasilyev, V. Babayan

Federal Research Institute of Fisheries and Oceanography (VNIRO, Moscow)
e-mail: dvasilyev@vniro.ru

Some approaches to increase the robustness of likelihood functions in their application for estimation of parameters of stock assessment models are considered.

Key words: stock assessment models, robust likelihood functions.