УДК 007:681.3.06

ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ ОЦЕНИВАНИЯ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ В УСЛОВИЯХ ОГРАНИЧЕННОЙ ЭКСПЕРИМЕНТАЛЬНОЙ ИНФОРМАЦИИ

Е.В. РЕДЬКО, Т.В. ПОДЛАДЧИКОВА, В.Н. ПОДЛАДЧИКОВ

Для исследования статистических свойств оценок в условиях ограниченной экспериментальной информации в работе предлагается совмещать обработку экспериментальных данных с процедурой имитационного моделирования, имитирующей продолжение эксперимента в тех же условиях. Предложенная методика используется для повышения эффективности процедуры оценивания показателя степени закона распределения энергии солнечных вспышек.

ВВЕДЕНИЕ

При обработке экспериментальных данных для оценки неизвестных параметров широко используется метод наименьших квадратов (МНК). Применение этого метода в условиях недостаточного объема экспериментальных данных и нарушений основных предположений МНК может привести к смещению оценок и не обеспечивает надежность выводов, непосредственно основанных на полученных оценках.

Смещение оценки неизвестного параметра может привести к искажению физического представления об исследуемом процессе. Так, величина показателя степени закона распределения энергии солнечных вспышек, который принято считать степенным, является принципиально важным [1]. Если показатель степени больше двух, то среднее количество энергии, вносимое в солнечную корону вспышками, определяется верхней границей диапазона энергии солнечных вспышек, если меньше двух, то нижней границей.

Как альтернативный способ оценивания рассмотрен бутстреп-метод. Идеей метода является размножение имеющейся выборки, что представляется эффективным способом оценивать параметры распределения в условиях ограниченной экспериментальной информации [3]. Бутстреп-метод позволяет получить асимптотически несмещенные и эффективные оценки [4]. В качестве оценки результата выступает усредненная по N выборкам оценка. При сравнении результатов оценивания с помощью фильтра Калмана и бустреп-метода показано, что последний является более точным [5].

Поэтому для определения смещения оценок в условиях ограниченной экспериментальной информации, предлагается использовать статистическую модель–прототип, имитирующую продолжение эксперимента в тех же условиях.

На основе модели-прототипа проводится обработка данных измерения энергии солнечных вспышек спутником Trace (Transition region and coronal explorer — исследователь короны и переходной области Солнца) и выполняется коррекция смещения оценок МНК показателя степени, изменяющего представление о влиянии энергии солнечных вспышек.

АНАЛИЗ СМЕЩЕНИЯ ОЦЕНКИ ПОКАЗАТЕЛЯ

Плотность $f(x_i^*)$ степенного закона распределения $f(x) = Ax^{-\alpha}$ $(\alpha > 0)$ по экспериментальной выборке аппроксимируется величиной ω_i / Δ_i, где ω_i — относительная частота попадания в i-й интервал группирования; Δ_i — ширина i-го интервала $(x_i, x_i + \Delta_i)$; x_i^* — точка, принадлежащая i-му интервалу, $(i = 1, \ldots, N)$; N — число интервалов группирования.

Неизвестные параметры A и α оцениваются на основе построения модели линейной регрессии вида

$$\ln \tilde{f}(x_i^*) = \ln A - \alpha \ln x_i^* + \varepsilon_i, \qquad (1)$$

где ε_i — ошибка, флуктуационная составляющая модели.

Покажем, что для модели (1) математическое ожидание $E(\varepsilon_i) \neq 0$. Относительная частота ω_i является несмещенной состоятельной оценкой вероятности P_i попадания случайной величины в i-й интервал группирования данных, т.е. $\omega_i = P_i + \varsigma_i$, где ς_i — случайная величина, распределенная по биномиальному закону с параметрами $E(\varsigma_i) = 0$, $\mathrm{var}(\varsigma_i) = E(\varsigma_i^2) = \dfrac{P_i(1-P_i)}{n}$, где n — длина выборки.

В качестве оценки плотности распределения $\tilde{f}(x_i^*)$ используется отношение $\dfrac{\omega_i}{\Delta_i}$, т.е. $\tilde{f}(x_i^*) = \dfrac{\omega_i}{\Delta_i} = \dfrac{P_i + \varsigma_i}{\Delta_i}$. Или

$$\tilde{f}(x_i^*) = f(x_i^*) + \eta_i. \qquad (2)$$

Здесь $\eta_i = \xi_i + \dfrac{\varsigma_i}{\Delta_i}$, где $\xi_i = \dfrac{P_i}{\Delta_i} - f(x_i^*)$ — составляющая ошибки, обусловленная заменой предела $\lim \dfrac{P_i}{\Delta_i} = f(x_i^*)$ при $\Delta_i \to 0$ отношением $\dfrac{P_i}{\Delta_i}$. Математическое ожидание $E(\eta_i)$ ошибки модели (2) определяется ошибкой выбора точки x_i^* на интервале группирования и равно

$$E(\eta_i) = \frac{P_i}{\Delta_i} - f(x_i^*).$$

Рассмотрим смещение ошибки ε_i модели (1) при больших n.

Логарифм правой части выражения (2) имеет вид

$$\ln\left(f(x_i^*)+\eta_i\right)=\ln f(x_i^*)+\ln\left(1+\frac{\eta_i}{f(x_i^*)}\right).$$

Используя разложение в ряд Тейлора логарифмической функции и ограничиваясь двумя членами разложения, получим для больших n асимптотическое представление

$$\ln\left(1+\frac{\eta_i}{f(x_i^*)}\right)\sim\frac{\eta_i}{f(x_i^*)}-\frac{{\eta_i}^2}{2f^2(x_i^*)}.$$

Следовательно, флюктуационная составляющая модели (1) имеет вид

$$\varepsilon_i\sim\frac{\eta_i}{f(x_i^*)}-\frac{{\eta_i}^2}{2f^2(x_i^*)}=\frac{\xi_i}{f(x_i^*)}+\frac{\varsigma_i}{\Delta_i f(x_i^*)}-\frac{\left(\xi_i+\frac{\varsigma_i}{\Delta_i}\right)^2}{2f^2(x_i^*)}.$$

Математическое ожидание ошибки ε_i равно

$$E(\varepsilon_i)\sim E\left(\frac{\xi_i}{f(x_i^*)}\right)-E\left(\frac{{\xi_i}^2}{2f^2(x_i^*)}\right)-E\left(\frac{{\varsigma_i}^2}{2\Delta^2 f^2(x_i^*)}\right)=$$

$$=\left(\frac{P_i}{\Delta_i f(x_i^*)}-1\right)-\frac{1}{2}\left(\frac{P_i}{\Delta_i f(x_i^*)}-1\right)^2-\frac{P_i(1-P_i)}{2n{\Delta^2}_i f^2(x_i^*)}\neq 0.$$

Таким образом, флюктуационная составляющая модели (1) смещена, что обусловливает смещение оценки показателя степени.

ОЦЕНКА ХАРАКТЕРИСТИК РАСПРЕДЕЛЕНИЯ НА ОСНОВЕ ИМИТАЦИОННОЙ МОДЕЛИ

Описание модели

В соответствии с моделью (1) определим оценку показателя α по данным измерения энергии солнечных вспышек спутником Trace.

Длина экспериментальной выборки составляет $n=1215$; минимальное и максимальное выборочные значения равны $x_{\min}=14{,}8523$, $x_{\max}=2515{,}57$.

Весь интервал наблюдения $(x_{\min},x_{\max})$ первоначально разбивается на 49 равных интервалов группирования. После объединения тех интервалов, в которых оказывается менее пяти выборочных значений, число интервалов группирования составляет $N=9$.

На основе МНК была получена оценка показателя степени $\hat{\alpha}=1{,}828$. Непосредственные выводы о действительных значениях показателя степени на основе этой оценки в условиях смещения оценок МНК, обусловленного смещением ошибок ε_i модели (1) могут оказаться ненадежными.

Для определения вероятностных характеристик оценок в условиях ограниченной экспериментальной информации предлагается использовать

статистическую модель-прототип, имитирующую продолжение эксперимента в тех же условиях по ансамблю реализаций. Вычисленные по экспериментальным данным оценки параметров распределения рассматриваются как один из результатов, которые могут быть получены при моделировании.

В предлагаемой модели генерируется выборка случайных величин распределенная по степенному закону, таким образом, что длина выборки, минимальное и максимальное выборочные значения полностью соответствуют экспериментальным данным. Алгоритм обработки предполагает также первоначальное разбиение на 49 интервалов с последующим объединением тех интервалов, в которых оказалось менее пяти выборочных значений.

Выбор допустимого диапазона действительных значений показателя степени

Экспериментальная выборка рассматривается как одна из реализаций имитационной модели. Поэтому обязательным требованием к имитационной модели-прототипу является выбор такого интервала $(\alpha_{\min}, \alpha_{\max})$, чтобы для каждого действительного значения показателя степени $\alpha \in (\alpha_{\min}, \alpha_{\max})$ существовала ненулевая вероятность совпадения результатов моделирования с результатами обработки экспериментальных данных. С возможным совпадением полученных при моделировании и при обработке экспериментальных данных оценки $\hat{\alpha} = 1{,}828$ можно связать лишь вероятность нуль. Поэтому при выборе интервала $(\alpha_{\min}, \alpha_{\max})$ следует исходить из требования существования ненулевой вероятности попадания полученной при моделировании оценки в некоторый интервал $(\hat{\alpha}_1, \hat{\alpha}_2)$, внутри которого находится оценка $\hat{\alpha} = 1{,}828$. Кроме того, диапазон выбираемых действительных значений α ограничивается условием существования ненулевой вероятности сокращения числа интервалов группирования до $N = 9$ после их объединения при первоначальном разбиении на 49 равных интервалов.

Вероятность сокращения числа интервалов до 9 оценивается как относительная частота появления этого события в 1000 реализациях выборки случайных величин, распределенных по степенному закону при различных действительных значениях α, взятых с шагом $0{,}01$.

Как показали результаты моделирования, число интервалов группирования после объединения уменьшается с ростом действительного значения α. Наименьшее значение показателя степени, при котором хотя бы в одной из 1000 реализаций число интервалов после объединения сократилось до 9, равно $\alpha = 1{,}71$. Максимальное значение показателя, при котором еще с ненулевой вероятностью число интервалов после объединения сокращается до 9, равно $\alpha = 2{,}32$.

Относительная частота попадания оценки показателя $\hat{\alpha}$ в интервал $(\hat{\alpha}_1, \hat{\alpha}_2)$, включающий точку $\hat{\alpha} = 1{,}828$, определялся по 1000 реализациям для всех значений $\alpha \in (1{,}71; 2{,}32)$, взятых с шагом $0{,}01$, для двух случаев: $(\hat{\alpha}_1, \hat{\alpha}_2) = (1{,}825; 1{,}835)$; $(\hat{\alpha}_1, \hat{\alpha}_2) = (1{,}8; 1{,}86)$.

Как показали результаты моделирования, минимальное значение α, при котором еще существует ненулевая вероятность попадания оценки

в интервал $(1{,}825;\ 1{,}835)$ равно $\alpha_{\min} = 1{,}85;$ а в интервал $(1{,}8;\ 1{,}86)$ равно $\alpha_{\min} = 1{,}82.$

На рис. 1 приведена зависимость от $\alpha^{(k)}$ значений оценок условной вероятности $\hat{P}_1^{(k)} = [\hat{\alpha} \in (1{,}825;\ 1{,}835) / \alpha = \alpha^{(k)} \in (1{,}825;\ 2{,}32)]$, на рис. 2 приведена зависимость от $\alpha^{(k)}$ значений оценок условной вероятности $\hat{P}_2^{(k)} = [\hat{\alpha} \in (1{,}8;\ 1{,}86) / \alpha = \alpha^{(k)} \in (1{,}825;\ 2{,}32)].$

Как видно из рис. 1 и 2, в обоих случаях наиболее вероятно попадание в интервал $(\hat{\alpha}_1; \hat{\alpha}_2)$ при $2{,}06 < \alpha < 2{,}14.$

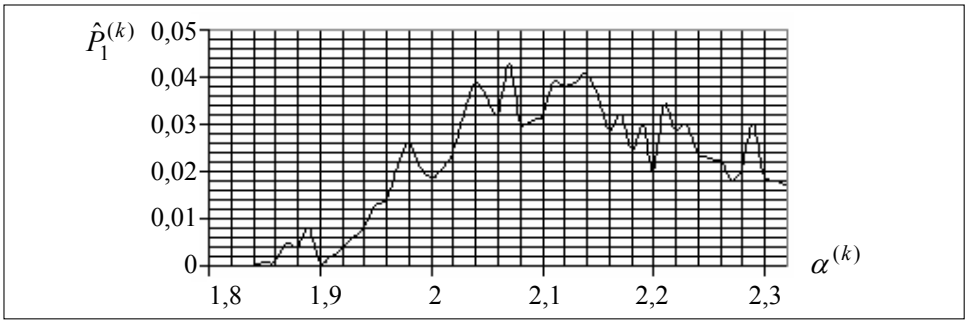

Рис. 1. Оценка условной вероятности при $\hat{\alpha} \in (1{,}825;\ 1{,}835)$

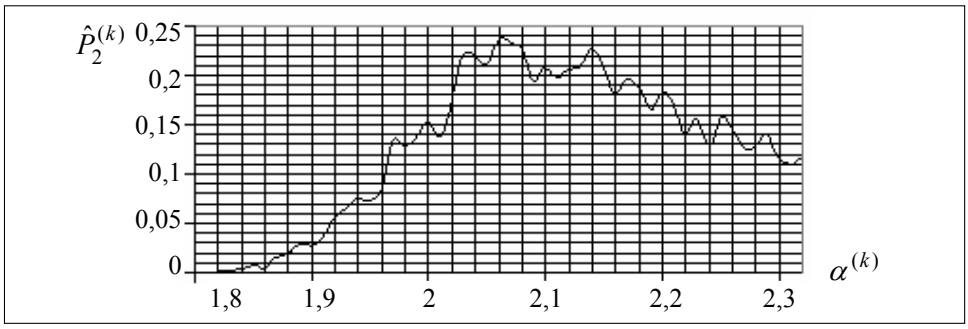

Рис. 2. Оценка условной вероятности при $\hat{\alpha} \in (1{,}8;\ 1{,}86)$

АПОСТЕРИОРНОЕ РАСПРЕДЕЛЕНИЕ ПОКАЗАТЕЛЯ СТЕПЕНИ

Вероятность принадлежности действительного значения α интервалу $(\alpha^{(k)}; \alpha^{(k+1)})$ при условии, что его оценка оказалась в интервале $(\hat{\alpha}_1; \hat{\alpha}_2)$ рассчитывается по формуле вероятностей гипотез Байеса [2]:

$$P[\alpha \in (\alpha^{(k)}; \alpha^{(k+1)}) / \hat{\alpha} \in (\hat{\alpha}_1; \hat{\alpha}_2)] =$$

$$= \frac{P[\alpha \in (\alpha^{(k)}; \alpha^{(k+1)})] P[\hat{\alpha} \in (\hat{\alpha}_1; \hat{\alpha}_2) / \alpha \in (\alpha^{(k)}; \alpha^{(k+1)})}{\sum_{j=0}^{L-1} (P[\alpha \in (\alpha^{(j)}; \alpha^{(j+1)})] P[\hat{\alpha} \in (\hat{\alpha}_1; \hat{\alpha}_2) / \alpha \in (\alpha^{(j)}; \alpha^{(j+1)})]}.$$

Здесь $P[\alpha \in (\alpha^{(k)}, \alpha^{(k+1)})]$ — априорная вероятность принадлежности действительного значения показателя степени α интервалу $(\alpha^{(k)}, \alpha^{(k+1)})$.

$P[\hat{\alpha} \in (\hat{\alpha}_1; \hat{\alpha}_2) / \alpha \in (\alpha^{(k)}, \alpha^{(k+1)})$ — вероятность попадания оценки показателя степени в интервал $(\hat{\alpha}_1; \hat{\alpha}_2)$ при условии, что его действительное значение принадлежит интервалу $(\alpha^{(k)}, \alpha^{(k+1)})$, определялось как среднее арифметические рассчитанных при моделировании значений вероятностей попадания оценок $\hat{\alpha}$ в интервал $(1{,}825; 1{,}835)$ при условиях, что его действительные значения принадлежат границам этого интервала, т.е. $\alpha = \alpha^{(k)}$ и $\alpha = \alpha^{(k+1)}$.

На рис. 3, 4 показана оценка апостериорной плотности распределения показателя степени α, если $\hat{\alpha}$ оказалась в интервалах $(1{,}825; 1{,}835)$ и $(1{,}8; 1{,}86)$ соответственно.

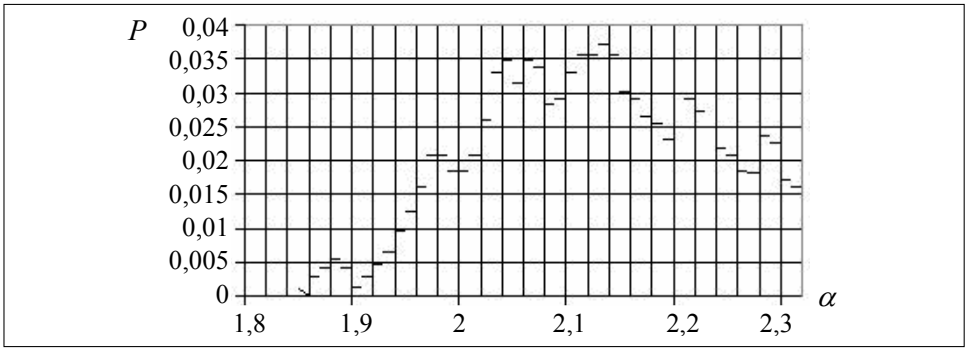

Рис. 3. Оценка апостериорной плотности распределения показателя степени α, если $\hat{\alpha} \in (1{,}825; 1{,}835)$

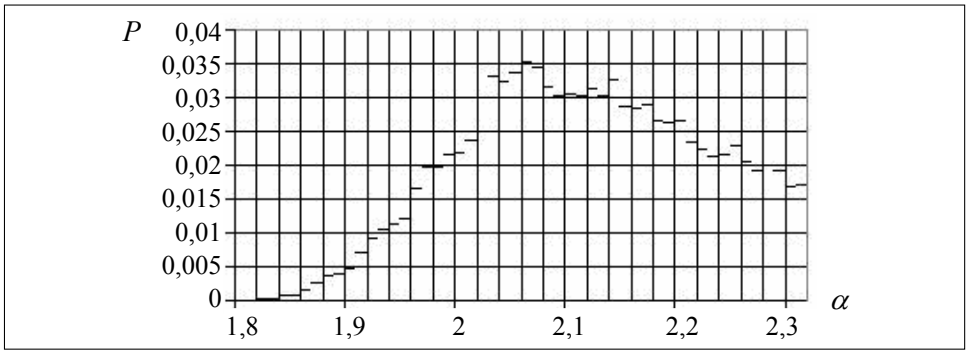

Рис. 4. Оценка апостериорной плотности распределения показателя степени α, если $\hat{\alpha} \in (1{,}8; 1{,}86)$

Из рис. 3 и 4 видно, что в обоих случаях плотность распределения резко возрастает от нуля до значений, превышающих $0{,}03$, при изменении α от $1{,}82$ до $2{,}03$. Максимальных значений плотность распределения достигает на интервале $\alpha \in (2{,}03; 2{,}15)$. При $\alpha > 2{,}15$ плотность распределения снижается и достигает значения $0{,}016 - 0{,}17$ при $\alpha = 2{,}32$.

Апостериорные вероятности неравенства $\alpha > 2$, для обоих рассматриваемых случаев также отличаются незначительно:

$$\hat{P} = [\alpha > 2 / \hat{\alpha} \in (1{,}825; 1{,}835)] = 0{,}868,$$

$$P = [\alpha > 2 / \hat{\alpha} \in (1{,}8; 1{,}86)] = 0{,}853.$$

Полученные оценки апостериорного распределения показателя степени незначительно отличаются при изменении длины интервала, внутри которого находится оцененное по экспериментальным данным значение α и позволяют делать обоснованные выводы о диапазоне вероятных значений показателя степени по полученной по экспериментальным данным оценке.

Таким образом, с высокой вероятностью можно утверждать, что показатель степени больше 2 и среднее количество энергии, вносимое в солнечную корону вспышками, определяется верхней границей энергии солнечных вспышек.

ВЫВОДЫ

В данной работе были рассмотрены статистические свойства оценок параметров в условиях ограниченной выборки и доказана смещенность оценок, которые можно получить используя МНК. Доказано, что вклад в смещение оценок вносит линеаризация исходной модели и ограниченное количество экспериментальной информации. В работе использован и описан метод построения имитационной модели прототипа, позволяющий выполнить коррекцию смещения полученных оценок.

В предлагаемой модели генерируется выборка случайных величин распределенная по степенному закону, таким образом, что длина выборки, минимальное и максимальное выборочные значения полностью соответствуют экспериментальным данным. Алгоритм обработки предполагает также первоначальное разбиение на 49 интервалов с последующим объединением тех интервалов, в которых оказалось менее пяти выборочных значений.

Данная модель использована для определения действительного значения показателя степени экспоненциального закона распределения энергии солнечных вспышек. Для этого было построено апостериорное распределение вероятностей значения показателя степени α, если $\hat{\alpha}$ оказалась в интервалах (1,825; 1,835) и (1,8; 1,86) соответственно.

Результаты показали, что в обоих случаях плотность распределения резко возрастает от нуля до значений, превышающих $0{,}03$, при изменении α от $1{,}82$ до $2{,}03$. Максимальных значений плотность распределения достигает на интервале $\alpha \in (2{,}03; 2{,}15)$. При $\alpha > 2{,}15$ плотность распределения снижается и достигает значения $0{,}016 - 0{,}017$ при $\alpha = 2{,}32$.

На основе графиков плотности вероятностей сделан окончательный вывод о том, что показатель степени имеет действительное значение больше 2 и среднее количество энергии, вносимое в солнечную корону вспышками, определяется верхней границей энергии солнечных вспышек.

ЛИТЕРАТУРА

1. *Crosby N.B., Aschwanden M.J., Dennis B.R.* Frequency Distribution and correlation of solar X-Ray flare parameters // Solar Physics. — 1993. — № 143. — P. 275–299.
2. *Смирнов Н.В., Дунин-Барковский И.В.* Курс теории вероятностей и математической статистики. — М.: Наука, 1969. — 512 с.
3. *Орлов А.И.* Эконометрика. — М.: Экзамен, 2002. — 576 с.
4. *Гаев Л.В.* Рандомизированная обработка результатов имитационных экспериментов. — http://www.gpss.ru/immod'03/017.html.
5. *Prasad R.M., Sinha A.K.* Two-stage bootstrap algorithms for parameter estimation // International Journal of Systems Science. — 1977. — **8**, № 12. — P. 1365–1374.

Поступила 15.05.2009