Нейко М. Нейков

0.1 Обобщено разпределение на Парето

В тази секция ще изложим най-често използвания подход за моделиране на екстремуми, а именно метода на праговете (peaks over threshold).

Нека $X_1, \ldots, X_n$ е редица от случайни величини, разпределени като случайната величина X, с функция на разпределение F. Естествено е да разглеждаме като екстремални стойности тези X_i, който превишават дадено достатъчно високо ниво (праг) u. Да означаим произволен елемент на редицата X_i с X. Тогава поведение на екстремалните стойности на X се характеризира от условната вероятност за всяко $y > 0$:

$$\begin{aligned} P(X > u + y | X > u) &= \frac{P(\{X > u + y\} \cap \{X > u\})}{P(X > u)} \\ &= \frac{P(X > u + y)}{P(X > u)} = \frac{1 - F(u + y)}{1 - F(u)}, \end{aligned} \tag{1}$$

понеже $\{X > u + y\} \cap \{X > u\} = \{X > u + y\}$.

Ако разпределението F беше известно, то от горното уравнение лесно бихме могли да намерим разпределението на наблюденията, превишаващи прага u. Тъй като на практика F е неизвестна, се търсят приближения на (1) за достатъчно големи прагове. Следната теорема дава едно такова приближение.

Теорема 2: Нека $X_1, \ldots, X_n$ е редица от случайни величини, които имат обша функция на разпределение F и $Z_n = \max\{X_1, \ldots, X_n\}$. Да означим с X произволен елемент от редицата X_i, да предположим, че съществуват редици от константи $\{a_n > 0\}$ и $\{b_n\}$, такива че

$$P\left(\frac{Z_n - b_n}{a_n} \leq z\right) \to G(z) \quad \text{когато} \quad n \to \infty,$$

$$G(z) = \exp\left\{-\left[1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, \tag{2}$$

където $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, $-\infty < \mu < \infty$,$\sigma > 0$ и $-\infty < \xi < \infty$. Тогава, за големи u, условната функцията на разпределение на $(X - u)|X > u$, е приблизително равна на

$$H(y) = 1 - \left(1 + \frac{\xi y}{\widetilde{\sigma}}\right)^{-1/\xi} \tag{3}$$

дефинирана за $y : y > 0$ и $(1 + \xi y/\widetilde{\sigma}) > 0$, където

$$\widetilde{\sigma} = \sigma + \xi(u - \mu). \tag{4}$$

Д-во: За големи n имаме

$$\begin{aligned} F_{Z_n}(z) &= P(Z_n \le z) = P(\max(X_1, X_2, \ldots, X_n) \le z) \\ &= P(\{X_1 \le z\} \cap \ldots \cap \{X_n \le z\}) \\ &= P(\{X_1 \le z\}) \ldots P(\{X_n \le z\}) \\ &= F(z)F(z) \ldots F(z) \\ &= F^n(z) \approx \exp\left\{ -\Big[1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}} \right\} \end{aligned}$$

Следователно,

$$n \log F(z) \approx -\Big[1 + \xi\Big(\frac{z-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}.$$

От развитието в ред на Тейлор, при голями стойности на z получаваме

$$\log F(z) \approx -\{1 - F(z)\}.$$

Замествайки този израз в предишния израз получаваме

$$1 - F(u) \approx \frac{1}{n}\Big[1 + \xi\Big(\frac{u-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}$$

за голями u. Аналогично, за $y > 0$ получаваме

$$1 - F(u+y) \approx \frac{1}{n}\Big[1 + \xi\Big(\frac{u+y-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}.$$

От това следва, че

$$\begin{aligned} P(X > u+y | X > u) &= \frac{1 - F(u+y)}{1 - F(u)} \\ &\approx \frac{n^{-1}\Big[1 + \xi(u+y-\mu)/\sigma\Big]^{-1/\xi}}{n^{-1}\Big[1 + \xi(u-\mu)/\sigma\Big]^{-1/\xi}} \\ &= \Big[1 + \frac{\xi(u+y-\mu)/\sigma}{\xi(u-\mu)/\sigma}\Big]^{-1/\xi} \\ &= \Big[1 + \frac{\xi y}{\tilde{\sigma}}\Big]^{-1/\xi} \end{aligned}$$

Следователно,

$$P(X \le u+y | X > u) = 1 - P(X > u+y | X > u) \approx \begin{cases} 1 - \Big[1 + \frac{\xi y}{\tilde{\sigma}}\Big]^{-1/\xi} & \text{ако} \quad \xi \ne 0 \\ 1 - \exp\big(-\frac{y}{\tilde{\sigma}}\big) & \text{ако} \quad \xi = 0 \end{cases}$$

при условие, че $y > 0$ и $1 + \frac{\xi y}{\tilde{\sigma}} > 0$.

Фамилията от разпределения, дефинирани от (3) се нарича обобщено разпределение на Парето (Generazlized Pareto Distribution, GPD). Теорема 2 гарантира, че ако блок-максимумите имат приблизително разпределение $G(z)$, то превишаванията на прага имат разпределение от GPD. Също така, параметрите на обобщеното разпределение на Парето се определят еднозначно от съответните параметри на GEV разпределението на блок-максимумите. В частност, параметъра ξ в уравнение (3) е равен на съответния му в GEV разпределението. Интересно е да отбележим, че избора на различен по големина размер на блоковете n ще промени параметрите на GEV разпределението, но и тези на съответстващото му обобщено разпределение на Парето. Параметърът за форма ξ има най-голямо влияние за общото поведение на GPD разпределението както е при обобщеното разпределение на екстремалните стойности.

Плътност на обобщеното разпределение на Парето (GPD) с праг u:

$$g(x) = \frac{1}{\sigma}\left[1 + \frac{\xi(x-u)}{\sigma}\right]_+^{-\frac{1}{\xi}-1}$$

където $x > u$, $\sigma > 0$ и ξ са параметрите на мащаба и формата, $[A]_+ = \max(A, 0)$. Параметъра ξ характеризира типа на GP разпределението, както следва: с тежка опашка ако $\xi > 0$, с крайна опашка ако $\xi < 0$ и експоненциален тип ако $\xi = 0$.

Проверката на хипотезата за типа на опашката на разпределението. Проверява се хипотезата за експоненциален тип $\xi = 0$ срещу алтернативната $\xi > 0$ с тежка опашка по данните, след което се формира отношението на правдоподобие. За целта се оценяват параметрите на експоненциалното и GPD разпределениятя по извадката от данните, превишаващи предварително зададен праг, емпиричен квантил на данните, след което се формира теста, основан на отношението на правдоподобие. Да означим тази извадка с up.treshold.dat. Формиранете на теста, основан на отношението на правдоподобие в средата на VGAM библиотеката се постига чрез следните

```
> fit.exp <- vglm(intensity~1, exponential(exp=FALSE,
   location=threshold),data=up.treshold.dat,trace=F,residuals = TRUE)
> fit.gpd <- vglm(intensity~1,gpd(threshold=threshold,lshape="elogit",
   zero=2),data=up.treshold.dat,residuals=TRUE)
> LRT=-2 * (logLik(fit.exp)-logLik(fit.gpd)))
> df=length(coef(fit.gpd))-length(coef(fit.exp))
> chisq(LRT, df = df, lower.tail = FALSE)
```

0.1.1 Оценка на параметрите на обобщеното разпределение на Парето

Нека $y_1, \ldots, y_k$ са стойностите, които превишават прага u. Тогава, за логаритъма на функцията на правдоподобие на GPD за $\xi \neq 0$

$$l(\sigma,\xi) = -k\log\sigma - (1+1/\xi)\sum_{i=1}^{k}\log(1+\xi y_i/\sigma), \tag{5}$$

при ограничението $(1+\xi y_i/\sigma) > 0$ за $i = 1,\ldots,k$; в противен случай $l(\mu,\xi) = -\infty$.

При $\xi = 0$ логаритъма на функцията на правдоподобие се редуцира до логаритъма на функцията на правдоподобие на експоненциалното разпределение

$$l(\sigma) = -k\log\sigma - \frac{1}{\sigma}\sum_{i=1}^{k} y_i. \tag{6}$$

Нека $\hat{\sigma}$ и $\hat{\xi}$ са МПО на незивестните параметри σ и ξ.

Квантили на GPD разпределението

Нека параметъра на формата $\xi \neq 0$ на GPD разпределението с параметри (σ,ξ). Да предположим, че GPD е подходямо за анализ на екстремалните стойности на сл. вел. X превишаващи прага u, т.е., $x > u$. От свойствата на условната вероятност за събитието $\{X > x\} \cap \{X > u\} = \{X > x\}$ получаваме

$$P(X > x | X > u) = \frac{P(\{X > x\} \cap \{X > u\})}{P(X > u)} = \frac{P(X > x)}{P(X > u)}.$$

От предположението за GPD на данните превишаващи прага u имаме

$$P(X > x | X > u) \approx \frac{1}{n}\left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}. \tag{7}$$

Следователно

$$P\{X > x\} = P(X > u)P(X > x | X > u) \approx \zeta_u \frac{1}{n}\left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi},$$

Следователно, квантилът x_m, превишаващ всеки m наблюдения е решение на уравнението

$$\zeta_u \frac{1}{n}\left[1 + \xi\left(\frac{x_m - u}{\sigma}\right)\right]^{-1/\xi} = \frac{1}{m}.$$

След преобразуване получаваме

$$x_m = u + \frac{\sigma}{\xi}\left[(m\zeta_u)^{\xi} - 1\right] = u + \frac{\sigma}{\xi}\left[(mp)^{\xi} - 1\right], \tag{8}$$

при условие, че m е достатъчно голямо, за да осигури условието $x_m > u$, $\zeta_u = P(X > u) = p$ е вероятността за превишаване на прага u, x_m представлява максималната стойност на X, която очакваме да наблюдаваме в m измервания.

По аналогия, ако $\xi = 0$, получаваме

$$x_m = u + \sigma \log(m\zeta_u), \tag{9}$$

отново при условие, че m е достатъчно голямо.

По дефиниция, x_m е ниво на обезпеченост (стойност на риска, VaR - value at risk) на m наблюдения. От уравненият за квантилите (8) и (9) следва, че графичното представяне на x_m срещу m в логаритмична скала е подобно на QQ плота за GEV разпределението.

За нуждите на практиката е удачно нивата на обезпеченост да бъдат дадени в годишна скала. В този случай нивото на обезпеченост от N години е нивото, което очакваме да бъде превишено веднъж на всеки N години. Ако имаме n_y наблюдения на година, то това съответства на ниво на обезпеченост на $m = N \times n_y$ наблюдения. Следователно за период на обезпеченост от N години нивото на обезпеченост се дефинира

$$x_N = u + \frac{\sigma}{\xi}\left[(Nn_y\zeta_u)^{\xi} - 1\right] \tag{10}$$

за $\xi \neq 0$ или

$$x_N = u + \sigma \log(Nn_y\zeta_u), \tag{11}$$

за $\xi = 0$.

Замествайки параметрите с техните оценки получаваме емпиричните квантили на разпределение на Парето. За оценка $\hat{\zeta}_u$ на ζ_u се използва отношението k/n, където k е броят на наблюденията, превишаващи прага u, а n е броя на наблюденията. Тази оценка е максимално правдоподобна оценка, понеже понеже броят на превишаванията k от n наблюдения е биномно разпределена случайна величина $\mathrm{Bin}(n, \zeta_u)$. Следователно, $V(\hat{\zeta}_u) \approx \hat{\zeta}_u(1 - \hat{\zeta}_u)/n$.

Доверителни интервал на квантилите на GPD разпределението

Стандартните грешки и доверителни интервали за параметрите и x_m се получат по делта метода и максимално правдоподобната оценка на ковариационната матрица на вектора $(\hat{\zeta}_u, \hat{\sigma}, \hat{\xi})$

$$V = \begin{pmatrix} \hat{\zeta}_u(1 - \hat{\zeta}_u)/n & 0 & 0 \\ 0 & v_{1,1} & v_{1,2} \\ 0 & v_{2,1} & v_{2,2} \end{pmatrix}$$

където $v_{i,j}$ за $i,j = 1,2$ са оценките на ковариационната матрица на (σ, ξ). по метода на максималното правдоподобие.

Нека $\hat{\theta} = (\hat{\sigma}, \hat{\xi})$ и $V(\hat{\theta})$ са МПО на неизвестния параметър θ и съответната му ковариационнаа матрица. Тогава МПО на ковариационнта оценка на квантила $\hat{x}_m$ може да бъде получена по делта метода

$$\begin{aligned} Var(x_m) &\simeq \nabla x_m^T V(\theta) \nabla x_m \qquad (12) \\ \nabla x_m^T &= \left(\frac{\partial x_m^T}{\partial \zeta_u}, \frac{\partial x_m^T}{\partial \sigma}, \frac{\partial x_m^T}{\partial \xi} \right) \\ &= \left[\sigma m^{\xi} \zeta_u^{\xi-1}, \xi^{-1}\{(m\zeta_u)^{\xi} - 1\}, -\sigma\xi^{-2}\{(m\zeta_u)^{\xi} - 1\} + \sigma\xi^{-1} (m\zeta_u)^{\xi} \log((m\zeta_u) \right], \end{aligned}$$

оценени в $\hat{\theta} = (\hat{\sigma}, \hat{\xi})$ и $\hat{\zeta}_u$.

По-добри доверителни интервали за квантилите на на GPD разпределението се получават чрез профила на функцията на правдоподобие. От изразите за квантилите (10) и (11) за мащабния параметър σ получаваме представяне

$$\sigma = \begin{cases} \frac{(x_m - u)\xi}{(m\zeta_u)^{\xi} - 1} & \text{ако} \quad \xi \neq 0 \\ \\ \frac{(x_m - u)}{\log(m\zeta_u)} & \text{ако} \quad \xi = 0 \end{cases}$$

Профилни оценки на ξ, респективно на x_m се получават след заместването на този израз във функцията на правдоподобие (5) и определянето на нейния максимум като финкция на ξ при фиксирана стойност на x_m, респективно на x_m при фиксирана стойност на ξ. За начална оценка на x_m може да се използва оценката, получени по делта метода.

Проверка за адекватност на GPD модела

Проверката за адекватност на GPD модела се основава на емпиричните и моделни вероятности плотове, на съответният им квантилен плот и хистограмата на плътността на разпределението. Да разгледаме вариационният ред на наблюденията, надхвърлящи прага u, т.е., $y_{(1)} \leq \ldots \leq y_{(k)}$ и означим с $\widehat{H}$ оценената функция на разпределение. Вероятностният PP плот се строи по точките

$$\{(i/(k+1), \widehat{H}(y_{(i)}), i = 1, \ldots, k\},$$

докато квантилния QQ плот се състои по точките

$$\{(\widehat{H}^{-1}(i/(k+1), y_{(i)}), i = 1, \ldots, k\},$$

където

$$\widehat{H}(y) = 1 - \left(1 + \frac{\hat{\xi} y}{\widehat{\sigma}} \right)^{-1/\hat{\xi}}. \qquad (13)$$

Считаме, че данните следват GPD разпределението, ако точките лежат на диагонала на тези плотове, в противен случай избираме нов праг u и повтаряме процедурата.

Разбира се, проверката на модела става чрез стандартните методи-qq-plot, графика на вероятностите и хистограма на плътността на разпределението. QQ-plot-а се прави върху наредената статистика на наблюденията, надхвърлящи прага u: $y_{(1)} \leq \ldots \leq y_{(k)}$ и оцененият ни вече модел $\widehat{H}$-графиката се състои от точките

$$\{(\widehat{H}^{-1}(i/(k+1), y_{(i)}), i = 1, \ldots, k\},$$

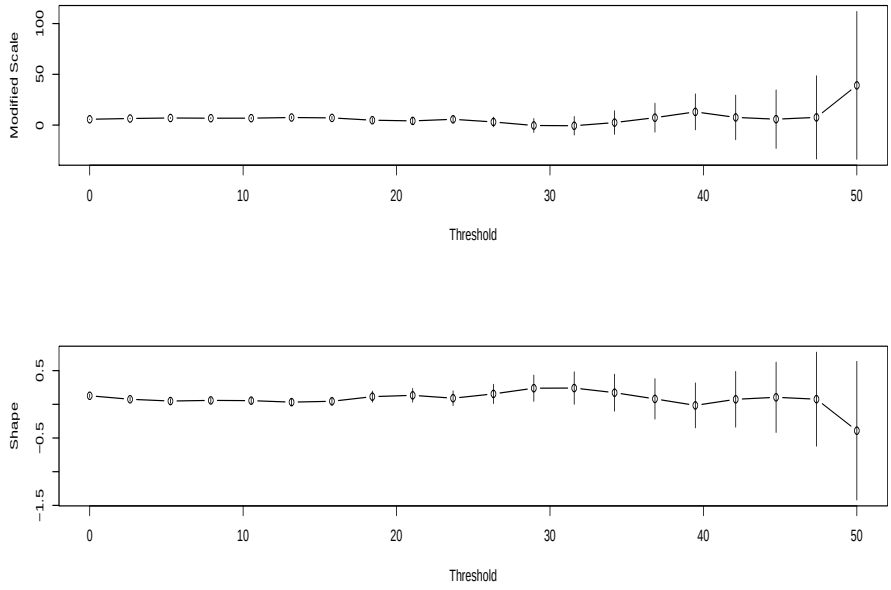
където

$$\widehat{H}^{-1}(y) = u + \frac{\widehat{\sigma}}{\widehat{\xi}}\left[y^{-\widehat{\xi}} - 1\right].$$

Сега, след като разяснихме основните постановки при моделирането с метода на прага за стационарни и независими времеви редове е време да преминем към малко по-реалистичният случай на зависимост на случайните наблюдения, участващи във времевия ред.

Това е съвсем нормално, тъй като на практика почти няма случай, в който да няма някакъв вид зависимост между наблюденията-като прост пример ще спомена валежите-ясно е, че по време на циклон, вероятността за няколко поредни дни с екстремални валежи е много голяма, което от своя страна изобщо не отговаря на досега приетите условия за моделиране на екстремални данни. Точно заради това ни е нужно по някакъв начин да отчетем тази зависимост между наблюденията, за да можем да използваме при статистическото моделиране на данните принципа на максималното правдоподобие като метод за оценяване на стойностите на параметрите на разпределението.

А както знаем, основна предпоставка за това е наблюденията ни да са независими, което ни позволява да запишем коректно функцията на правдоподобие на отделните наблюдения. Без да навлизаме в подробности, относно теоритичните работи върху поведението на екстремуми на зависими редици от случайни величини, ще формулираме чисто практичен начин за "отстраняване"на зависимостта при моделирането по метода чрез използването на праг. Да допуснем, че сме избрали подходящ праг, данните над който бихме искали да моделираме с обощеното разпределение на Парето. Често срещана ситуация в практиката е няколко съседни елемента на редицата от данни да надвишават зададения праг, т.е., екстремалните стойноти да формират клъстери. За да редуцираме очевидната зависимост в данните се използва максималната стойност на данните във всеки от клъстерите. Ясно е, че клъстерите трябва да са отделени във времето. Така формираната редица от екстремални стойности можем да смятаме като редица от слабо зависими стойности на случайни в88



Фигура 1: Примерна графика за установяване на прага u (данни за валежи)-вижда се, че за случая подходяща стойност за прага е 30, тъй като до тази стойност графиката е сравнително константна

0.2 Заключение

В първата част са описани някои от най-често използваните подходите за статистическо моделиране на максимумите на редици от случайни величини: блок-максимум метода, основан на обобщеното разпределение на екстремалните стойности GEV; метода на праговете на екстремалност, чрез обобщеното разпределение на Парето.

Във втората част на работа е направен анализ на реални данни вълнението в Черно море по моделни данни от SWAN модела, а също така за температури и валежи за станция Кнежа. По този начин са илюстрирани възможностите на програмнта система. Първо, за максимални температури, е използван модел на екстремумите като времеви ред, като времето е единствения предиктор, от който могат да зависят параметрите на разпределението на екстремалните стойности. За намирането на подходящ, променящ се с времето праг, който да описва трендовете през различните сезони, се използва квантилна регресия, която впоследствие използваме при GPD-модела на температурите. Понеже данните формират нестационарен времеви ред са анализирани чрез Поасонов точков процес , методика на моделирането с GPD разпределението.

Библиография

[1] de Melo Mendes B., V. and Pericchi, L. P. (2009) Assessing Conditional Extremal Risk of Flooding in Puerto Rico, Stochastic Environmental Research and Risk Assessment, vol. 23,

[2] Coles, S. (2001). An Introduction to Statistical Modelling of Extreme Values, Springer

[3] De Haan L., Ferreira A., Extreme value theory-an introduction, Springer,2006

[4] Embrechts P.,Klüppelberg C., Mikosch T., Modeling extremal events for Insurance and Finance, Springer,1997

[5] Eastoe E. and Tawn J. (2009).Modelling non-stationary extremes with applications to surface level ozone,J.R.Statistical Soc. C,58,Part 1,pp. 25-45

[6] Heffernan J. and Tawn J. (2004)A conditional approach for multivariate extreme values,J.R.Statistical Soc. B.,66,Part3,pp 497-546

[7] Jenkinson, A. F. (1955). The frequency distribution of the annual maximum (or minimum) values of meteorological elements

[8] Koenker R.,Quantile regression,Cambridge University Press,2005

[9] oenker R.,Quantile regression in R:a vignette,documentation to the *quantreg* package for R.

[10] otz S. and Nadarajah, S. (2000). *Extreme value distributions-theory and applications.* Imperial College Press.

[11] eadbetter M., Lindgren G. and Rootzen, H. (1986). Extremes and related properties of random sequences and processes. Springer.

[12] cNeil, A., Frey, R. and Embrechts, P. (2005). Quantitative risk mangement, Springer,Princeton University Press.

[13] eiss R. and Thomas D. (, (2007). Statistical analysis of extreme values-with applications to Insurance, Finance, Hydrology and other fields, Birkhäuser.

[14] mith R.,Statistics of extremes,with applications in environment,insurance and finance,webnotes, www.stat.unc.edu/postscript/rs/semstatrls.pdf

[15] tephenson, A. and Gilleland, E. (2006). Software for the analysis of extreme events:The current state and future directions, Extremes, vol. 8, pp. 7–109, DOI 10.1007/s10687-006-7962-0

[16] ee, T. and Stephenson, A. (2007). Vector generalized linear and additive extreme value models, Extremes, vol. 10, pp. 1–19, DOI 10.1007/s10687-007-0032-4