L-momenty s rušivou regresí

Jan Picek, Martin Schindler

e-mail: jan.picek@tul.cz

TECHNICKÁ UNIVERZITA V LIBERCI

ROBUST 2016

Motivation 1

Development of extreme value models with time-dependent parameters in order to estimate (time-dependent) high quantiles of maximum daily air temperatures over Europe in climate change simulations (1961-2100).

Kyselý, Picek and Beranová (2010): Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold, Global and Planetary Change, 72, 55-68

A significant trend is present in climate change simulations (1961-2100) for different scenarios future climate.

Motivation 2 - L-moments

- L-moments are linear combinations of order statistics.
- The concept of L-moments originates from various disconnected results on linear combinations of order statistics, e.g. (Sillitto, 1969, Chernoff et al., 1967, Greenwood et al., 1979)
- J.R.M Hosking (1990) unified the theory of L-moments and provided guidelines for the practical use.
- Since that many applications in hydrology, climatology, quality control (parameter estimation – method L-moments).

Motivation 2 - L-moments

Advantages of method L-moments:

- With small and moderate samples the method L-moments is often more efficient than maximum likelihood *simulation study (Hosking, Wallis, Wood): for all values k of GEV in the range* $-0.5 < k < 0.5$ *and for all sample sizes up to 100, estimates (L-mom) have lower root-mean/square error than the maximum likelihood estimates*
- Usually computationally more tractable than method of maximum likelihood
- Compared to the conventional moments, L-moments have lower sample variances and are more robust against outliers
- The cases in which some of the higher moments fails to exist. *f.e. GEV for* $k < -1/3$ *the third and fourth moments*

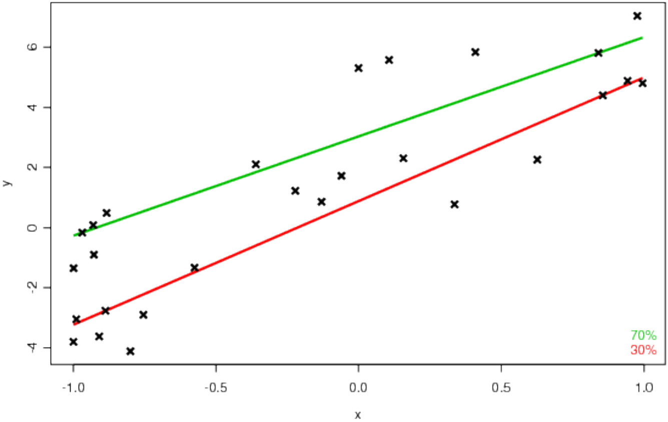
Motivation 1+2 - regression quantiles

Motivation 1 - a significant trend is present in climate change simulations
$\Longrightarrow$ a linear regression model

Motivation 2 - L-moments - linear combinations of order statistics
$\Longrightarrow$ quantiles

linear regression model+ quantiles $\Longrightarrow$ regression quantiles

Regression quantiles



The advantage of this approach is that many aspects of usual quantiles and order statistics are generalized naturally to the linear model.

L-moments

Let $X_1, X_2, \ldots X_n$ are independent, identically distributed random variables with a cumulative distribution function $F(x)$ and a quantile function $Q(u)$.
Let $X_{1:n} \leq X_{2:n} \leq X_{n:n}$ are the order statistics.
Definition:

$$\lambda_r = \frac{1}{r} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} EX_{r-k:r}, \quad r = 1, 2, \ldots$$

$$EX_{j:r} = \frac{r!}{(j-1)!(r-j)!} \int x \left(F(x)\right)^{j-1} \left(1 - F(x)\right)^{r-j} \, \mathrm{d}F(x)$$

L-moments

The first four L-moments are

$$\lambda_1 = EX = \int_0^1 Q(u)du$$

$$\lambda_2 = \frac{1}{2}E(X_{2:2} - X_{1:2}) = \int_0^1 Q(u)(2u-1)du$$

$$\lambda_3 = \frac{1}{3}E(X_{3:3} - 2X_{2:3} + X_{1:3}) = \int_0^1 Q(u)(6u^2 - 6u + 1)du$$

$$\lambda_4 = \frac{1}{4}E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}) = \int_0^1 Q(u)(20u^3 - 30u^2 + 12u - 1)du$$

L-moments

EXAMPLE: L-moments of some distribution:

Uniform (a, b)	$\lambda_1 = \frac{1}{2}(a+b), \lambda_2 = \frac{1}{6}(b-a), \tau_3 = 0, \tau_4 = 0$
Normal $\mathcal{N}(\mu, \sigma^2)$	$\lambda_1 = \mu, \lambda_2 = \frac{\sigma}{\pi}, \tau_3 = 0, \tau_4 = 0.1226$
Gumbel	$F(x) = \exp[-\exp(-(x-\xi)/\alpha)]$
	$\lambda_1 = \xi + \alpha\gamma, \lambda_2 = \alpha \log 2, \tau_3 = 0.1699,$ $\tau_4 = 0.1504, \gamma = 0.5772...$ const.
Generalized extreme value (GEV)	$F(x) = \exp[-(1 - k(x-\xi)/\alpha)^{\frac{1}{k}}]$ $\lambda_1 = \xi + \alpha(1 - \Gamma(1+k))/k,$ $\lambda_2 = \alpha(1 - 2^{-k})\Gamma(1+k)/k,$ $\tau_3 = 2(1-3^{-k})/(1-2^{-k}) - 3, \tau_4 = \ldots$ $k > -1,\ \Gamma(.)$ denotes gamma function

L-moments

Estimations of L-moments – Sample $L-$moment:

$$l_r = \binom{n}{r}^{-1} \sum\sum_{1\le i_1<i_2<\ldots<i_r\le n}\ldots\sum r^{-1}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}X_{i_{r-k}:n},$$

$r = 1, 2, \ldots, n.$
in particular:

$$l_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad l_2 = \frac{1}{2}\binom{n}{2}^{-1} \quad \sum\sum_{i>j}(X_{i:n} - X_{j:n})$$

$$l_3 = \frac{1}{3}\binom{n}{3}^{-1}\sum\sum_{i>j>k}\sum(X_{i:n} - 2X_{j:n} + X_{k:n})$$

$$l_4 = \frac{1}{4}\binom{n}{4}^{-1}\sum\sum_{i>j>k>l}\sum\sum(X_{i:n} - 3X_{j:n} + 3X_{k:n} - X_{l:n})$$

L-moments

Parameter estimation – method L-moments

Uniform (a, b)	$\hat{a} = l_1 - 3l_2, \hat{b} = l_1 + 3l_2$
Normal $\mathcal{N}(\mu, \sigma^2)$	$\hat{\mu} = l_1, \hat{\sigma} = \pi^{1/2} l_2$
Gumbel	$F(x) = \exp[-\exp(-(x-\xi)/\alpha)]$ $\hat{\xi} = l_1 - \hat{\alpha}\gamma, \hat{\alpha} = l_2/\log 2$ $\gamma = 0.5772...$ const.
Generalized extreme value (GEV)	$F(x) = \exp[-(1 - k(x-\xi)/\alpha)^{\frac{1}{k}}]$ $z = 2/(3+t_3) - \log 2/\log 3,$ $\hat{k} = 7.8590z + 2.9554z^2,$ $\hat{\alpha} = l_2\hat{k}/[(1-2^{-\hat{k}})\Gamma(1+\hat{k})],$ $\hat{\xi} = l_1 + \hat{\alpha}[\Gamma(1+\hat{k}) - 1]/\hat{k}$

Regression quantiles

Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}, \qquad (1)$$

where $\mathbf{Y}$ is an $(n \times 1)$ vector of observations, $\mathbf{X}$ is an $(n \times (p+1))$ matrix, $\boldsymbol{\beta}$ is the $((p+1) \times 1)$ unknown parameter $(p \geq 1)$and $\mathbf{E}$ is an $(n \times 1)$ vector of i. i. d. errors with a cumulative distribution function F.

We assume that the first column of $\mathbf{X}$ is $\mathbf{1}_n$, i.e. the first component of $\boldsymbol{\beta}$ is an intercept.

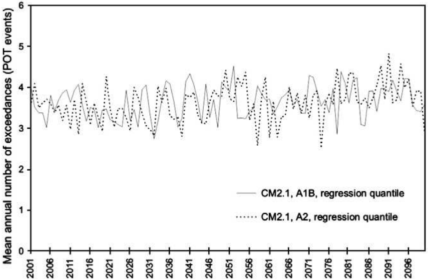
R. Koenker and G. Basset (1978) defined the α-regression quantile $\widehat{\boldsymbol{\beta}}(\alpha)$ $(0 < \alpha < 1)$ for the model (1) as any solution of the minimization

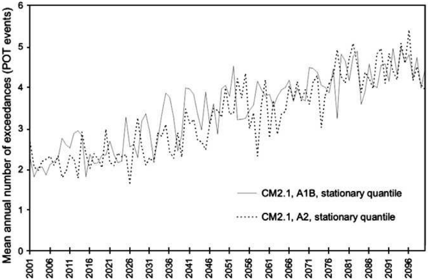
$$\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i'\mathbf{t}) := \min, \quad \mathbf{t} \in \mathbb{R}^{p+1}, \qquad (2)$$

where

$$\rho_\alpha(x) = x\psi_\alpha(x),\ x \in \mathbb{R}^1 \text{ and } \psi_\alpha(x) = \alpha - I_{[x<0]},\ x \in \mathbb{R}^1. \qquad (3)$$

Regression quantiles





Mean annual number of exceedances above the threshold (averaged over gridpoints) for the 95 regression quantile and the 95% quantile.

Regression quantiles

Computation of $\widehat{\boldsymbol{\beta}}$ can be expressed as a parametric linear programming problem with m_n distinct solutions as α goes from zero to one. That is, there will be m_n breakpoints $\{\tau_i\}$, for $i = 1, \cdots, m_n$. Each $\widehat{\boldsymbol{\beta}}_n(\tau_i)$ is characterized by a specific subset of $p+1$ observations.

Portnoy (1991) – $m_n = \mathbf{O}(n \log n)$ in probability.

R – package *quantreg*

Regression quantiles

The dual linear program to the RQ-problem:

$$\begin{aligned} \mathbf{Y}'\hat{\mathbf{a}}_n(\alpha) &:= \max \\ \mathbf{X}'\hat{\mathbf{a}}_n(\alpha) &= (1-\alpha)\mathbf{X}'\mathbf{1}_n \\ \hat{\mathbf{a}}_n(\alpha) &\in [0,1]^n, \quad 0 < \alpha < 1. \end{aligned} \tag{4}$$

It defines the vector of regression rank scores
$\hat{\mathbf{a}}_n(\alpha, \mathbf{Y}) = \hat{\mathbf{a}}_n(\alpha) = (\hat{a}_{n1}(\alpha), \ldots, \hat{a}_{nn}(\alpha))'$ in the linear model.
The regression rank scores are

- continuous, piecewise linear in α, invariant with respect to the shift in location and scale and also regression invariant, i.e.,

$$\hat{\mathbf{a}}_n(\alpha, \mathbf{Y} + \mathbf{X}\mathbf{b}) = \hat{\mathbf{a}}_n(\alpha, \mathbf{Y}) \qquad \forall \mathbf{b} \in \mathbf{R}^p$$

- From the duality between $\hat{\boldsymbol{\beta}}(\alpha)$ and $\hat{\mathbf{a}}_n(\alpha)$: $\forall \alpha \in (0,1)$ for $i = 1, \ldots, n$

$$\hat{a}_{ni}(\alpha) = \begin{cases} 1 & \text{if } Y_i > \sum_{j=0}^{p} x_{ij}\hat{\beta}_j(\alpha), \\ \\ 0 & \text{if } Y_i < \sum_{j=0}^{p} x_{ij}\hat{\beta}_j(\alpha) \end{cases}$$

Regression quantiles

Jurečková and Picek (2014) introduced the averaged regression quantile

$$\bar{B}_n(\alpha) = \bar{\mathbf{x}}_n^\top \widehat{\boldsymbol{\beta}}_n(\alpha), \qquad \bar{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{ni} \tag{5}$$

and studied its properties and relations to other statistics. Some properties of $\bar{B}_n(\alpha)$ are surprising: $\bar{B}_n(\alpha)$ is asymptotically equivalent to the $[n\alpha]$-quantile of the location model.

$$n^{1/2}\left[\bar{\mathbf{x}}_n^\top(\widehat{\beta}_n(\alpha) - \beta) - E_{[n\alpha]:n}\right] = \mathcal{O}_p(n^{-1/4}) \tag{6}$$

as $n \to \infty$, where $E_{1:n} \leq \ldots \leq E_{n:n}$ are the order statistics corresponding to $E_1, \ldots, E_n$.

So, the averaged regression α-quantile is asymptotically equivalent to the location α-quantile.

$\Longrightarrow$ we can generalize the *method of L-moments* in the regression context.

Generalization of L-moments in linear regression model

Substitution into

$$\lambda_r = \frac{1}{r}\sum_{k=0}^{r-1}(-1)^k\binom{r-1}{k}EX_{r-k:r}, \quad r = 1, 2, \dots$$

of a standard expression for the expected value of an order statistic (e.g. David and Nagaraja, 2003).

$$EX_{j:r} = \frac{r!}{(j-1)!(r-j)!}\int x\,(F(x))^{j-1}\,(1-F(x))^{r-j}\ \mathrm{d}F(x)$$

yields a classical L-functional representation,

$$\lambda_k = \int_0^1 F^{-1}(u)P^*_{k-1}(u)du$$

Generalization of L-moments in linear regression model

$$\lambda_k = \int_0^1 F^{-1}(u) P_{k-1}^*(u) du,$$

where

$$P_k^*(u) = \sum_{j=0}^{k} p_{k,j}^* u^j,$$

with

$$p_{k,j}^* = (-1)^{k-j} \binom{k}{j} \binom{k+j}{j}$$

P_k^* is the so-called shifted Legendre polynomial

L-moments is a special case of L-estimator $\Longrightarrow$ we can define L-moments in linear regression model as special L-estimators.

Generalization of L-moments in linear regression model

Koenker and Portnoy (1987) and Gutenbrunner and Jurečková (1992) generalized L-statistics to the linear model integrating the regression quantile process with respect to a suitable signed measure ν on $(0,1)$:

$$\int_0^1 \widehat{\boldsymbol{\beta}}_n(\alpha) d\nu(\alpha) = \int_0^1 \widehat{\boldsymbol{\beta}}_n(\alpha) J(\alpha)\, d\alpha$$

and showed the L-statistic's asymptotic normality. J is the density of ν.

Taking for J the shifted Legendre polynomial $P^*_{r-1}(u)$ we get:

$$\lambda^R = \int_0^1 \bar{B}_n(\alpha) P^*_{r-1}(u)\, d\alpha = \frac{1}{n} \sum_{i=1}^{n} Y_i \left(-\int_0^1 \hat{a}'_{ni}(\alpha) P^*_{r-1}(u)\, d\alpha \right). \quad (7)$$

$r = 1, 2, \ldots, n.$

Generalization of L-moments in linear regression model

Sample $L-$moments based on averaged regression quantiles:
First we create subsample

$$Z_i^* := \bar{\mathbf{x}}_n^\top \left[\widehat{\beta}(\tau_i)\right], \quad i = 1, \ldots, m_n. \tag{8}$$

Then we plug averaged regression quantiles into the usual estimators of L-moments statistics, i.e.,

$$l_r^{AR} = \binom{m_n}{r}^{-1} \sum_{1 \le i_1 < i_2 < \ldots < i_r \le m_n} \sum \ldots \sum r^{-1} \sum_{k=0}^{r-1} (-1)^k \binom{r-1}{k} Z^*_{i_{r-k}:m_n},$$

$r = 1, 2, \ldots.$

Generalization of L-moments in linear regression model

First sample averaged regression L-moments may be written as

$$l_1^{AR} = \frac{1}{m_n}\sum_{i=1}^{m_n} Z_i^*, \quad l_2^{AR} = \frac{1}{2}\binom{m_n}{2}^{-1} \sum_{i>j}\sum (Z_{i:m_n}^* - Z_{j:m_n}^*).$$

$$l_3^{AR} = \frac{1}{3}\binom{m_n}{3}^{-1} \sum\sum_{i>j>k}\sum (Z_{i:m_n}^* - 2Z_{j:m_n}^* + Z_{k:m_n}^*)$$

$$l_4^{AR} = \frac{1}{4}\binom{m_n}{4}^{-1} \sum\sum_{i>j>k>l}\sum\sum (Z_{i:m_n}^* - 3Z_{j:m_n}^* + 3Z_{k:m_n}^* - Z_{l:m_n}^*)$$

Numerical illustration

The performance of the proposed sample averaged regression L-moments in the regression model

$$Y_i = \beta_0 + \beta_1 x_i + E_i, \quad i = 1, \ldots, n, \tag{9}$$

is studied on the simulated values.

- The chosen values of the parameter β are $\beta_0 = -1$, $\beta_1 = -2.$, the errors were generated from the the normal and GEV distributions.
- vector $x_1, \ldots, x_n$ was generated from the uniform distribution on the interval (-5,30) and was fixed for all simulations.
- 10 000 replications of the linear regression model were simulated for each case, and the sample averaged regression L-moments $l_r^{AR},\ r = 1, 2, 3$ were computed and used to estimate the parameters.
- location parameter of errors E_i were assumed to be known and equal to zero.

Numerical illustration

Case	MSE	mean	median
		σ^2	
ARQ,$n = 100$	0.3312	3.859	3.818
errors, $n = 100$	0.3320	4.019	3.970

Table: Mean, median and MSE of 10 000 estimated parameters based on the sample L-moments for model (9) with errors simulated from the Normal distribution $N(\mu = 0, \sigma^2 = 2)$

Case	MSE	mean	MSE	mean
	α		k	
ARQ, $n = 100$	0.0234	1.0520	0.0176	-0.4417
errors, $n = 100$	0.0205	1.0160	0.0157	-0.4660

Table: Mean and MSE of 10 000 estimated parameters based on the sample L-moments for model (9) with errors simulated from GEV distribution $GEV(\xi = 0, \alpha = 1, k = -0.5)$

Conclusions

- We can use the proposed L-moments as tool for parametr estimation if a significant trend is present in our dataset.
- With small and moderate samples this method might be more efficient than maximum likelihood method.
- We can use the known results for L-estimators for statistical inference (eg. construction confidence intervals).
- L-estimators (moments) based on regression rank scores (dual solution of parametric linear programming problem - regression quantiles) could be used for testing purposes - future work.

Thank you for your attention!

Bibliography

Hosking, J.R.M. (1990),
L-moments: Analysis and Estimation of Distribution Using Linear Combinations of Order Statistics.
J. Roy. Statist. Soc. Ser. B, **52**, 105–124.

Jurečková, J., Picek, J. (2014),
Averaged regression quantiles.
Springer Proceedings in Mathematics & Statistics, 2014, vol. 68, Chap. 12, pp. 203-216

Koenker R. and Bassett G. (1978).
Regression quantiles.
Econometrica, **46**, 33-50.

Kyselý J., Picek J., Beranová R. (2010).
Estimating extremes in climate change simulations using the peaks-over-threshold method with a non-stationary threshold.
Global and Planetary Change, **72**, 55-68.