18th AMSE Applications of Mathematics and Statistics in Economics
International Scientific Conference | Jindřichův Hradec, 2-6 September 2015

MODELLING OF UNEMPLOYMENT DURATION IN THE CZECH REPUBLIC WITH THE USE OF L-MOMENTS

IVANA MALÁ
University of Economics in Prague, Faculty of Informatics and Statistics, Department of Probability and Statistics,
W. Churchill sq. 4, Prague, Czech Republic
email: malai@vse.cz

Abstract
The use of L-moments instead of classical moments is recommended as very effective approach to the quantification of the level, variability, skewness and kurtosis of the distributions as well to the estimation of parameters in the modelling of distributions. In the text the use of sample L-moments as descriptive statistics for censored data are shown. Data dealing with unemployment duration in the Czech Republic are analysed. Lognormal distribution is fitted into subsets of data given by gender and education and these estimates are mixed into an overall distribution. Unknown parameters are estimated with the moment method using L-moments instead of classical moments. All computations are made in R program.

Key words: *moment method of estimation, L-moments, censored data, unemployment duration.*

1. Introduction

Moments are usually used to describe distribution of random variable and quantify level, variability, skewness or kurtosis of the distributions. Moreover large spectrum of estimation methods of distribution parameters based on moments is frequently used due to the easy evaluation (searching for the solutions of equations). Unfortunately these methods give sometimes unsatisfactory results. Sample moments are sensitive to outliers and there exist commonly used distributions without finite moments. L-moments (and other robust analogies of classical moments) provide the characteristics of distribution more robust. The application of L-moments is wider because of weaker assumptions on probability distribution (only finite expected value is required for the existence of all L-moments (Hosking, 1990)). Parameters of distributions are then estimated with the use of moment methods where classical moments are replaced by L-moments (Hosking, 1990, Bílková, 2013). Based on the Hosking´s article large theoretical research was done and a lot of applications were performed (modelling of flood frequencies, precipitation, estimating of extreme quantiles). Extensive use for the modelling of incomes and wages in the Czech Republic was done (Bílková, 2013). In this text this approach is used for time-to-event random variable where there exist incomplete (censored) observations in data. In such a case the usual definition of sample L-moments must be modified, the approach defined in (Wang et al., 2010; Wang, Hutson, 2013) and implemented in R package lmomco (Asquith, 2015) is used in this text.

There is a lot of statistical or econometric models proposed in the huge literature for the modelling of duration of unemployment. In this contribution positively skewed lognormal distribution is fitted into unemployment durations in the Czech Republic with the use of L-moments. In order to obtain more homogenous subgroups and improve the fit, lognormal distribution may be fitted into subgroups given by gender and education and the estimated

distributions mixed into estimates for the unemployment duration in the Czech Republic. The same models were analyzed with the use of maximum likelihood estimates in (Malá, 2013ab) and nonparametric approach in (Čabla, 2012). Properties of maximum likelihood estimates for finite mixtures estimated from censored data are given in (Miyata, 2011).

2. Methodology

L-moments (from linear) were proposed by Hosking in 1990 (Hosking, 1990) and huge research (both theoretical and practical applications) has been done in this field. This rapid development was made possible also thanks to the availability of powerful computer technology.

Suppose T to be a positive value random variable with continuous distribution. Classical central moments are defined as

$$\mu_k = E(T - E(T))^k = \sum_{j=0}^{k} (-1)^j \binom{k}{j} E(T^j)(E(T))^{k-j}.$$

For the existence of finite value of this type of moment the finite value of $E|T|^k$ is necessary. This assumption might be too restrictive in some cases (Cauchy distribution, Pareto distribution, …). *L*-moments provide a possibility how to overcome this problem and construct more robust moment characteristics (Hosking, 1990, Bílková, 2013). For *L*-moments of a probability distribution to be meaningful, we require only that the distribution have finite mean; no higher-order moments need be finite. For standard errors of *L*-moments to be finite, we require finite variance; no higher-order moments need be finite (Hosking, 1990).

For the construction of L-moments, random sample **T** is ordered. Denote (for any given k) ordered sample of the size k $(T_{1:k}, T_{2:k}, ..., T_{k:k})$, where $T_{1:k} \le T_{2:k} \le ... \le T_{k:k}$. The first index refers to rank and the second to the sample size. For the sample of the size of n we use ordered random vector $(T_{1:n}, T_{2:n}, ..., T_{n:n})$ with values $(t_{1:n}, t_{2:n}, ..., t_{n:n})$. Define for $k = 1, 2, ...$ the k-th *L*-moment λ_k as ((Hosking, 1990) or (Bílková, 2013)) as

$$\lambda_k = \frac{1}{k} \sum_{j=0}^{k-1} (-1)^j \binom{k-1}{j} E\left(T_{k-j:k}\right), \tag{1}$$

and sample *L*-moment l_k by the sample version of (1)

$$l_k = \binom{n}{k}^{-1} \sum_{1 \le i_1 < i_2 < ... < i_k \le n} \frac{1}{k} \sum_{j=0}^{k-1} (-1)^j \binom{k-1}{j} t_{i_{k-j}:n}. \tag{2}$$

From (1) it follows that L-moments are really linear functions of the moments of ordered sample (and from (2) sample L-moments are linear functions of values in ordered sample).

We will use two-parametric distribution $LN(\mu, \sigma^2)$, this distribution was used in (Malá, 2013a,b). In (Hosking, 1990) the moment method is used to estimate parameters of the

distribution. In our problem two equation are to be sold with respect to two unknown parameters μ and σ^2. Using (Hosking, 1990) or (Bílková, 2013) we obtain equation

$$\begin{aligned} l_1 &= \exp(\mu + \sigma^2) \\ l_2 &= \left(2\Phi\left(\frac{\sigma}{\sqrt{2}}\right) - 1\right)\exp(\mu + \sigma^2/2), \end{aligned} \tag{3}$$

where Φ is the cumulative probability function of standard normal distribution. These equations are to be solved by numeric methods or by polynomial approximation (Bílková, 2013).

Let T is a continuous nonnegative time-to-event random variable. For the evaluation of *l*-moments in case of the presence of censored data the formula (2) must be changed. Let us suppose only right censored data, where we have a sample of *n* pairs of $(T_i, \delta_i), i = 1, 2, ..., n,$ with $\delta_i = 1$ for noncensored value (exact value) and $\delta_i = 0$ for censored datum (if $T > T_i$). For right censored data Kaplan-Meier estimator (called also product limit estimator) of survival function is defined as $S(t) = P(T > t) = 1 - F(t),$ where *F* is cumulative distribution function of *T*, and this function is used instead of distribution function. It follows directly from properties of *F*, that survival function is continuous, non-increasing function with $S(0) = 1$ and $\lim_{t\to\infty} S(t) = 0.$ For ordered (distinct) values $t_1 \le t_2 \le t_3 \le \cdots \le t_n$ and $1 \le i \le n$ define by n_i number of observations "at risk" just prior to time t_i and d_i number of events at time t_i. The Kaplan-Meier estimator of survival function has a close form (Klein, Moeschberger, 1998)

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}, t > 0. \tag{4}$$

Variance of this estimate can be estimated by

$$est\left[D(\hat{S}(t))\right] = \hat{S}(t)^2 \sum_{t_i \le t} \frac{d_i}{n_i(n_i - d_i)}. \tag{5}$$

In (Wang et al., 2010) the method of evaluating sample *L*-moments is given as (instead of formula (2))

$$l_k = \sum_{j=1}^{n} T_{(j:n)} u_{j(k)}, \tag{6}$$

where $T_{(0)} = 0,\ p = k - r,\ q = r + 1,\ B_{p,q}(.)$ is incomplete beta function and

$$u_{j(k)} = \sum_{r=1}^{k} \frac{1}{k}(-1)^r \binom{k-1}{r} \left[B_{p,q}\left(1 - \hat{S}(t_{j:n})\right) - B_{p,q}\left(1 - \hat{S}(t_{j-1:n})\right)\right]. \tag{7}$$

The estimator is implemented in the R package lmomco (Asquit, 2015). In (Wang et al., 2010) are given regularity conditions for consistency of estimates and for asymptotic normality. In the text bootstrap is used in order to estimate standard deviations of parameters, in this text we use Monte Carlo estimation described below.

There are right censored and interval censored data in the analysed dataset of unemployment durations. Interval censoring means that only an interval $(L,R\rangle$ is known and information whether the event occurred in this interval or didn´t occur up to the end of this interval. The method proposed above in this contribution is basically intended for right censored data and for this reason the straightforward operation was done to obtain only complete and right censored data. For interval censored observations (those, who found a job in $(L,R\rangle$) an exact "pseudo" complete value was generated from uniform distribution on the interval (L,R). The process was done repeatedly in order to obtain estimate of the distribution of sample *l*-moment (and estimate of standard deviation). Due to extensive time of computations only 200 replications were performed (for the purpose of the illustration).

For all computations R program was used (RCORE TEAM, 2013). Estimation of *l*-moments was performed with the package lmomco (Asquith, 2015). Working with sample *l*-moments when the method is used repeatedly for samples of thousands of observations (as in this analysis) is time demanding.

3. Data and Results

Data dealing with unemployment duration in the Czech Republic in 2008 and 2010 (combined samples in order to obtain large datasets and due to nonsignificant differences between years (not shown in this text)) were used. Data were analyzed in (Malá, 2013ab) or (Čabla, 2012). Data from the Labour Force Sample Survey (LFSS), that is performed quarterly by the Czech Statistical Office (CZSO, 2015), cover four consecutive quarters in 2008 and 2010. The overall set of unemployed people (with unemployment duration up to two years) was divided into subgroups defined according to sex (two levels) and education (four levels: basic education (ISCED97 = 1, 2), lower secondary education (without baccalaureate, ISCED97 = 2), upper secondary education (ISCED97 = 3, 4), tertiary education (ISCED97 = 5, 6)). Sample sizes and number of censored and noncensored observations are given in the Table 1, all samples are large with exception of smaller subgroups of tertiary educated unemployed individuals. Data are heavy (interval) censored with only small percentage of reemployments (numbers after the slash in the Table 1).

Table 1. Sample sizes and counts of still unemployed/new jobs.

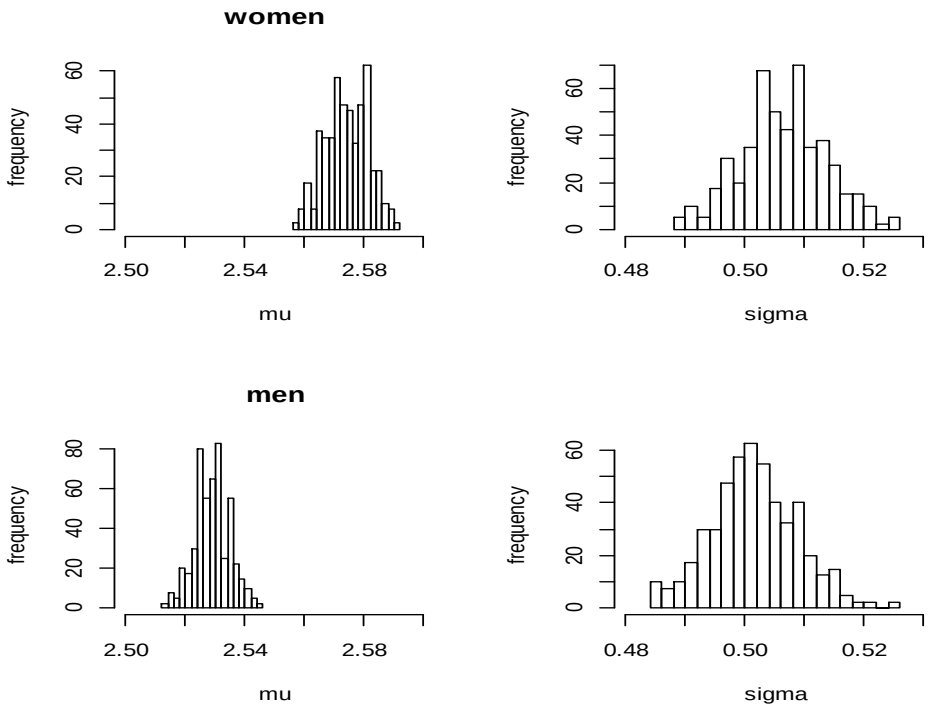
subgroup	sample size	censored/ uncensored	subgroup	sample size	censored/ uncensored
men	3 207	2 433/774	women	3 513	2 706/807
basic	1 138	1 065/ 73	secondary	3 071	2 340/731
upper secondary	2 018	1 510/508	tertiary	493	224/269

Source: own computations.

From the Table 1 we obtain estimates of mixing probabilities in the mixture (relative frequencies of subgroups in the sample) as 0.477 (men) and 0.523 (women) and 0.169 (basic education), 0.457 (lower secondary education), 0.300 (upper secondary education) and 0.073 (for tertiary education). Estimates of parameters from the Table 2 define estimated distributions for components. These component distributions are mixed into one mixture with estimated desity function $\hat{f}(t)$ ($K=2$ for the model with subgroups given by gender, $K=4$ for education)

$$\hat{f}(t)=\sum_{j=1}^{K}\hat{\pi}_j\,\hat{f}_{LN,j}\left(t;\hat{\mu},\hat{\sigma}^2\right),$$

where $\hat{\pi}_j$ are estimated mixing proportions (weights of components) given above and $\hat{f}_{LN,j}\left(t;\hat{\mu},\sigma^2\right)$, $j=1,2$ ($j=1,\ldots,4$) are densities of estimated lognormal distributions (estimated parameters are given in the Table 2).



Source: own computations.

Figure 1 Histogram of estimated parameters for man and women

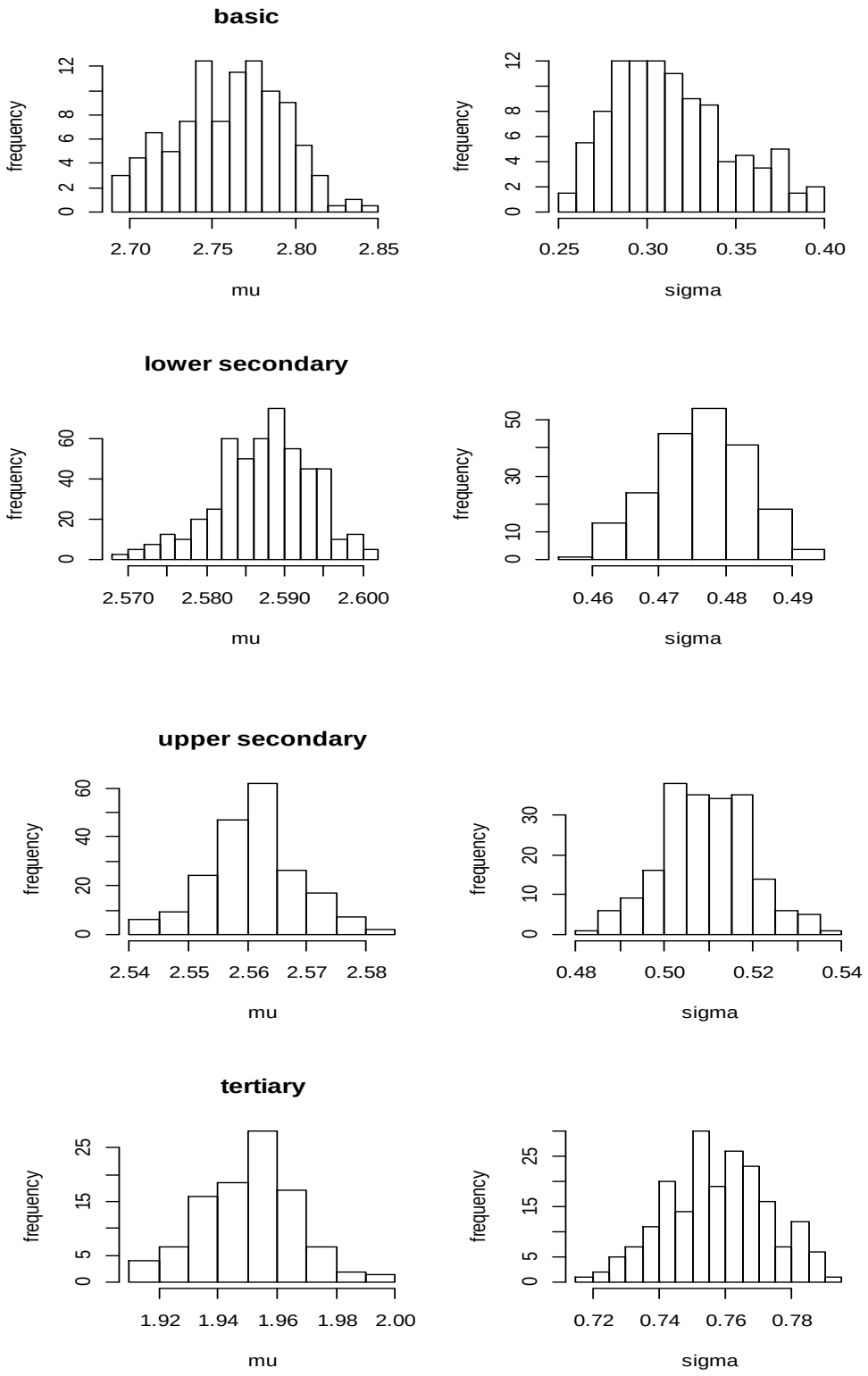
In the Table 2 estimates of parameters are shown for all subgroups. Parameters μ are means of logarithms of unemployment spell. Comparable values for men and women and decreasing values with education are well visible. Values of standard deviations are comparable with two exceptions, lower for basic education and higher for tertiary education. This fact reflects wider spectrum of possible solutions for tertiary educated people and probably also smaller sample size in this subgroup. Estimated medians quantify gender gap

(0.7 months) and positive impact of education (difference of more than 8 months between basic and tertiary education). All quartile deviations are comparable with values from 3 to 4 months.

Table 2. Estimated parameters of lognormal distribution and estimates of median unemployment spell and quartile deviation of this duration for all subgroups; mean (standard deviation).

subgroup	**men**	**women**	**basic**	**lower secondary**	**upper secondary**	**tertiary**
$\hat{\sigma}$	0.5016	0.5064	0.3141	0.4763	0.509	0.757
	(0.007)	(0.007)	(0.033)	(0.007)	(0.010)	(0.015)
$\hat{\mu}$	2.5290	2.574	2.760	2.587	2.561	1.950
	(0.005)	(0.007)	(0.032)	(0.006)	(0.007)	(0.016)
median	12.54	13.12	15.81	13.30	12.95	7.03
(months)	(0.07)	(0.09)	(0.51)	(0.08)	(0.10)	(0.12)
q	4.32	4.57	3.38	4.35	4.54	3.75
(months)	(0.08)	(0.09)	(0.43)	(0.08)	(0.11)	(0.07)

In the Figure 1 the distributions of μ and σ are shown for men and women and in the Figure 2 for four levels of education (based on only 200 replications of procedure due to the extensive computational time). The range of estimates is large from (by my opinion) at least three reasons: heavy censoring (from simulations follows that in case in heavy censoring the quality of estimates is poor), long intervals (L,R) making construction of complete observation more variable and variability of sample L-moments.



Source: own computations.

Figure 2 Histogram of estimated parameters for subgroups defined by education

4. Conclusion

In the contribution the *L*-moments are used for the estimation of parameters of lognormal distributions in case of heavy interval censored data. These moments are considered to be superior to classical moments (Hosking, 1990, Bílková, 2014) especially in case of the modelling of skewed data with heavy tails. In the text a modification of sample L-moments according to (Wang et al., 2010) is treated if right censored data are included in analyzed data and moment method is used to estimate parameters of two-parametric lognormal distribution from unemployment duration. Modelling of distribution of unemployment duration seems to be very suitable application for L-moment method as it is highly positively skewed heavy tailed random variable. The three parametric version of lognormal distribution was fitted as well, but all shift parameters were estimated by negative values. In the problem of unemployment the attention is frequently paid to very short unemployment durations (as well as to long-term unemployment) and negative values of shift parameter might misrepresent results for example for low quantiles (in comparison with incomes by (Bílková, 2014). The mixture of distributions (and more homogenous subgroup than it is expected in the whole population) could compensate for the omission of shift parameter.

As stated above, unemployment duration and the modelling of its distribution seems to be very suitable problem for the use of L-moments. Data are heavy censored, skewed and with heavy tails even if only unemployment durations shorter than two years are taken into analysis. Interval censored data were transformed to right censored with the use of random times with uniform distribution on censoring interval. Comparing with results based on maximum likelihood estimation in (Malá, 2013) estimated median durations by *l*-moments are shorter than MLE estimates of parameters. This interesting gap will be the subject of further research.

Regardless of estimated medians (or other characteristics) the known relationship between gender or education and position of analyzed groups of the unemployed on the labor market was shown (but quantified in different figures). The positive impact of education is well visible using median duration. The situation for men is better than for women, but this gender gap is not as large as it could be expected (result in accordance with (Malá, 2013)).

Acknowledgements

The support of the project “Doba nezaměstnanosti po krizi” number 87/2015 from the Faculty of Informatics and Statistics of the University of Economics, Prague is gladly acknowledged.

References

1. ASQUITH, W. H. 2015. lmomco: L-moments, censored L-moments, trimmed L-moments, L-comoments, and many distributions. R package version 2.1.3. Retrieved from: http://www.cran.r-project.org/package=lmomco
2. BÍLKOVÁ, D. 2013. Modeling of Wage Distribution in Recent Years in the Czech Republic Using L-moments and the Prediction of Wage Distribution by Indistry. E & M Ekonomie a Management, vol. 16, iss. 4, pp. 42-54.
3. CZSO. 2015. Czech Statistical Office. Retrieved from: http://www.czso.cz
4. ČABLA, A. 2012. Modeling Unemployment Duration in the Czech Republic from LFS. Research Journal of Economics, Business and ICT [online], vol. 7, pp. 1-5. ISSN 2045-3345.

5. HOSKING, J. R. M. 1990. L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. Journal of the Royal Statistical Society. Series B (Methodological), vol. 52, pp. 105-124.
6. KLEIN J.P., MOESCHBERGER, M.L. 1998. Survival Analysis, Techniques for Censored and Truncated Data. Springer. ISBN 038795399X.
7. MALÁ, I. 2013a. Použití konečných směsí pravděpodobnostních rozdělení pro modelování rozdělení doby nezaměstnanosti v České republice. Acta Oeconomica Pragensia, vol. 21, iss. 5, pp. 47-63.
8. MALÁ, I. 2013b. Finite Mixtures of Lognormal and Gamma Distributions. In: *International Days of Statistics and Economics.* [online] Prague, 19.09.2013 – 21.09.2013. Slaný: Melandrium, 2013, pp. 924–936. ISBN 978-80-86175-87-4.
9. MIYATA, Y. 2011. Maximum likelihood estimators in finite mixture models with censored data. Journal of Statistical Planning and Inference, vol. 141, iss. 1, pp. 56-64.
10. R CORE TEAM 2013: R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2012. Retrieved from: http://www.r-project.org/.
11. WANG, D., HUTSON, A. D., MIECZNIKOWSKI, J. C. 2010. L-moment estimation for parametric survival models given censored data. Statistical Methodology, vol. 7, iss. 6, pp. 655-667.
12. WANG, D., HUTSON, A. D. 2013. Joint confidence region estimation of L-moment ratios with an extension to right censored data. Journal of Applied Statistics, vol. 40, iss. 2, pp. 368-379.