Minimal Adequate Model of Unemployment Duration in the Post-Crisis Czech Republic

Adam Čabla[1] | *University of Economics in Prague, Czech Republic*

Abstract

Unemployment is one of the leading economic problems in a developed world. The aim of this paper is to identify the differences in unemployment duration in different strata in the post-crisis Czech Republic via building a minimal adequate model, and to quantify the differences.

Data from Labour Force Surveys are used and since they are interval censored in nature, proper methodology must be used. The minimal adequate model is built through the accelerated failure time modelling, maximum likelihood estimates and likelihood ratio tests.

Variables at the beginning are sex, marital status, age, education, municipality size and number of persons in a household, containing altogether 29 model parameters. The minimal adequate model contains 5 parameters and differences are found between men and women, the youngest category and the rest and the university educated and the rest. The estimated expected values, variances, medians, modes and 90th percentiles are provided for all subgroups.

Keywords

Unemployment duration, Labour Force Survey, minimal adequate model, interval censoring, AFT model

JEL code

J64, C24

INTRODUCTION

Unemployment is one of the leading problems in economy in a developed world and thus rightly the object of interest to many people. As usual, the problem of unemployment is statistically described by an unemployment rate and, regarding the duration of unemployment, a rate of long-term unemployment, i.e. the proportion of those who are unemployed longer than one year to all unemployed, is used (Eurostat, 2015b; CZSO, 2015).

Statistics about average unemployment duration for selected countries are provided by OECD (OECD. Stat, 2015). A deeper look at the unemployment duration in the Czech Republic was provided e.g. in Jarošová et al. (2004), Jarošová (2006) and more recently in Čabla (2014, 2015) and Malá (2013, 2014).

Main findings in previously cited papers are the changing role of possible explanatory variables. During the crisis the unemployment duration was influenced by sex, marital status, number of persons in household and education. After the crisis the unemployment duration was influenced only by sex and education, so we can see diminishing importance of marital status and number of persons in household.

[1] Department of Statistics and Probability, Nám. W. Churchilla 4, 130 67 Prague 3, Czech Republic. E-mail: adam.cabla@vse.cz.

The current paper newly provides multivariable model how to deconstruct possible dependence between these variables and to confirm or reject previous findings via more sound methodology.

Data are obtained from the Labour Force Survey (LFS), which is a large household sample survey providing quarterly results on labour participation of people aged 15 and over as well as on persons outside the labour force (Eurostat, 2015a).

The main problem in modeling the unemployment duration and obtaining its characteristics lies in the fact, that the data from LFS are censored. A researcher must consider it and use proper methodology based on a survival analysis. Some deeper methodological sources are provided in Čabla (2012).

The current paper offers an evaluation of the post-crisis data, specifically the year 2014. The main aim is to provide the minimal adequate model of the unemployment duration.

1 DATA

Data come from the LFS from quarters Q4/2013–Q1/2014. The LFS is conducted quarterly and 20% of the participants are changed every quarter. In other words, each participant takes part in five consecutive surveys. One survey includes approximately 50–60 thousand of participants.

One of the questions refers to the duration of a job search and another one the duration of current job. As a person is questioned over a year and a quarter, one can find those, who obtained a job in this survey period and compute the search duration. All participants were checked on their entry to the LFS and in the end of their participation. As the answers to the stated questions are interval censored, so is the consequent duration. Finally, 673 of participants who found a job were found.

It is important to keep in mind that the paper deals only with the unemployment duration of the participants, who were unemployed to begin with (unlike being economically inactive) and then found a job.

Possible explanatory variables for a model building, their shortcuts and base values are in the Table 1. Codes and numbers of observations for each category of the explanatory variables are in the annex in the Tables A1–A6. The Unemployment duration is given in months.

Table 1 Explanatory variables, their coding and base value

Explanatory variable	Shortcut	Base value
Number of persons in the household	PocOD	2
Sex	Pohl	1
Marital status	RodStav	1
Age group	VekSk	2
Education according to the ISCED scale	ISCED	3
Municipality size	MuniSize	2

Source: CZSO, own construction

2 METHODOLOGY

The main feature of the dataset is that the variable of interest – the unemployment duration, has only interval or right censored values. Standard methodology for dealing with censored variables is survival analysis.

2.1 Probability distribution in survival analysis

The main form of a description of a probability distribution in survival analysis, is a survival function. The survival function gives the probability that random variable T exceeds the specified time t.

$$S(t) = P(T > t) = 1 - F(t). \tag{1}$$

The second description of a probability distribution that is often used in survival analysis, is a hazard function. The hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t. In any analysis survival function can be transformed to hazard function or vice versa.

$$h(t) = l\ \underset{\Delta t \to 0}{im} \frac{P(t \le T < t + \Delta t \big| T \ge t}{\Delta t} = -\frac{dS(t)/dt}{S(t)} = \frac{f(t)}{S(t)}. \tag{2}$$

Being more specific here, the random variable T is the time of looking for a job of an unemployed person. The survival function $S(t)$ is the probability, that an unemployed person has not found a job at time t and finally the hazard function $h(t)$ is the instantaneous potential that an unemployed person will search for a job given that he has not found it up to time t (Kleinbaum and Klein, 2012).

2.2 Interval censoring

Data are called censored when exact value is not known, but they are known to fall within some interval (*Li*, *Ri*]. If only *Li* is known, than it is the case of right censoring. If only *Ri* is known, than it is the case of left censoring. If both *Li* and *Ri* are known, than it is interval censored variable.

Survival analysis is most detailed for the cases of right censoring, which usually occurs because the experiment ends before the specific event occurs. Here it means that when a person drops out from the LFS before finding a job, he can be assumed right censored as we can know only that the duration of his unemployment is longer than some specific time period *Li* (Kleinbaum and Klein, 2012).

2.3 Accelerated Failure Time model

Accelerated Failure Time (AFT) model is parametric survival model in which survival time is assumed to follow a known distribution. The underlying assumption of the AFT models is that the effect of covariates is multiplicative with respect to survival time.

The regression model is considered in the form:

$$\log(T) = \mu + \mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon, \tag{3}$$

where μ is an intercept, σ is a scale parameter, $\boldsymbol{\beta}$ is a vector of regression parameters, $\boldsymbol{x}$ is a vector of explanatory variables and ε is an error term with a known distribution. This can be used in

$$S(t) = P\left[\left(\mu + x'\beta' + \sigma\varepsilon\right) \ge \log(t)\right] = P\left(\varepsilon \ge \frac{\log(t) - \mu - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right), \tag{4}$$

$$S(t\big|\mathbf{x}) = \mathrm{S}_0\left(\frac{\log(t) - \mu - \mathbf{x}'\boldsymbol{\beta}}{\sigma}\right), \tag{5}$$

where S_0 is an independent survival function of the distribution of ε and $\boldsymbol{x}'\boldsymbol{\beta}$ defines the location of T, referred to as an accelerated factor. It can be formulated with respect to the random variable T instead of $\log(T)$:

$$T = \exp\left(\mu + \mathbf{x}'\boldsymbol{\beta}\right)\exp(\sigma\varepsilon), \tag{6}$$

$$S(t) = P(T \ge t) = P\left[\exp\left(\mu + \mathbf{x}'\boldsymbol{\beta} + \sigma\varepsilon \ \ge t\right)\right], \tag{7}$$

$$S(t|\mathrm{x}) = S_0\left[t\exp\left(-\mathbf{x}'\boldsymbol{\beta}\right)\right]. \tag{8}$$

In the AFT models, the effect of explanatory variables is such that if exp($\boldsymbol{x}'\boldsymbol{\beta}$) > 1, the effect of vector $\boldsymbol{x}$ is to decelerate the survival process and if exp($\boldsymbol{x}'\boldsymbol{\beta}$) < 1, the effect of vector $\boldsymbol{x}$ is to accelerate the survival process. The individual terms exp(b_m) indicates the multiplicative effect of a 1-unit change of the explanatory variable x_m on the time scale.

It follows from the previous that the AFT models can be used to model dependence of random variable T on a vector of explanatory variables with clear and simple description of this dependence, which is very convenient for the use in survival analysis.

Model is estimated via maximizing likelihood function. In the case of censoring the additional assumption is that censored times are independent of each other and of actual survival times, which should be fulfilled in the dataset of unemployed. For interval censoring the likelihood function is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n}\left[F(u_i) - F(v_i)\right] = \prod_{i=1}^{n}\left[S(v_\mathrm{i}) - S(u_i)\right], \tag{9}$$

where $S(t;\boldsymbol{\theta})$ is the parametric survival function and u_i and v_i are defined by

$$u_i = \frac{\log(R_i) - \mu - \mathbf{x}_\mathrm{i}'\boldsymbol{\beta}}{\sigma}, \tag{10}$$

$$v_i = \frac{\log(L_i) - \mu - \mathbf{x}_\mathrm{i}'\boldsymbol{\beta}}{\sigma}. \tag{11}$$

Equation 9 shows that under the interval censoring each observation contributes two pieces of information to the likelihood, $S(L_i, \boldsymbol{\theta}_i)$ and $S(R_i, \boldsymbol{\theta}_i)$, which follows the same distributional function (Liu, 2012).

2.3.1 Log-logistic distribution

The results (see below) indicates, that the unemployment duration follows log-logistic distribution. AFT model for this distribution has the survival function:

$$S(t;\mathbf{x},\boldsymbol{\beta},\sigma) = \left[1 + \exp\left(\frac{\log(t) - \mathbf{x}'\boldsymbol{\beta} - \mu}{\sigma}\right)\right]^{-1}, \tag{12}$$

and characteristics of log-logistic distribution are:

$$E(X) = \frac{\alpha b}{\sin(b)}, \tag{13}$$

$$Var(X) = \alpha^2\left(2b/\sin(2b) - b^2/\sin^2(2b)\right), \tag{14}$$

$$Mode(X) = \alpha\left(\frac{\beta - 1}{\beta + 1}\right)^{1/\beta}, \tag{15}$$

$$F^{-1} = \alpha\left(\frac{p}{1-p}\right)^{1/\beta}, \tag{16}$$

where

$$\alpha = \exp(\mu)\exp(\mathbf{x}'\boldsymbol{\beta}), \tag{17}$$

$$\beta = \frac{1}{\sigma}, \tag{18}$$

$$b = \frac{\pi}{\beta} \tag{19}$$

(Liu, 2012).

2.4 Minimal adequate model

The idea of a minimal adequate model is based on the principle of parsimony called sometimes Occam´s razor. In regard to statistical modeling it states (among other things), that models should have as few parameters as possible.

The process starts with a maximal model, i.e. model containing all the possible explanatory variables with all the possible values of interest. Than, it is simplified step by step, first removing the explanatory variables one by one and then merging the values within the remaining variables.

Results in this paper were obtained by removing explanatory variables based on the p-values of log-likelihood ratio tests comparing the maximal model and the model without the variable of interest – the variable with the largest p-value was removed, if the p-value was greater than 0.05, otherwise it would be considered to significantly reduce the likelihood.

When there were no variables, which would meet the criteria, left, the categories of the remaining variables were merged in a similar way. As they were ordinal categories, the two values next to each others were always merged (Crawley, 2013).

2.5 Likelihood ratio test

In the process of building minimal adequate model from the maximal model it is important to make comparisons of models and choosing which explanatory variables should be dropped. In the current paper the process is based on the likelihood ratio test and is similar to the backward selection method known from regression.

In the likelihood ratio test the null hypothesis is:

$$\mathrm{H}\ : \boldsymbol{\theta} = \hat{\boldsymbol{\theta}}, \tag{20}$$

For all parameters in $\hat{\boldsymbol{\theta}}$, or:

$$\mathrm{H}\ : \theta = \hat{\theta}_m, \tag{21}$$

for a single component in θ, and the statistic:

$$\Lambda = 2\log\left(L(\hat{\boldsymbol{\theta}})\right) - 2\log\left(L(\boldsymbol{\theta})\right), \tag{22}$$

has asymptotically $\chi^2(m)$ distribution. $L(\boldsymbol{\theta})$ is the likelihood function for the model without one or more parameters and $L(\hat{\boldsymbol{\theta}})$ is the likelihood function for the model containing all parameters (Liu, 2012).

3 RESULTS

All calculations were done in MS Excel and R software, specifically package interval (Therneau, 2013), (Fay, 2013).

The first important thing to do is to identify plausible distribution. The nine distributions for the whole dataset (i.e. without explanatory variables) were estimated and the one with maximum likelihood is used further. There were two distributions with similar likelihoods – log-normal and log-likelihood, the later with a little greater likelihood. The same distribution was used by Jarošová et al. (2004). Distributions and corresponding log-likelihoods are in the Table A7 in the Annex.

Since all explanatory variables are ordinal or nominal variables, the second thing to do is to consider which value will create baseline distribution – $\mathbf{x} = \mathbf{0}$. The selected values are presented in the Table 1 as base values, it means that baseline distribution is for a single man in the age of 21–25 years with secondary education without graduation and living in a household of 2 persons in a municipality with size 1 000–9 999 inhabitants.

The maximal model is presented in the Table A8 in the Annex, with three variables without significant values – number of persons in household, marital status and municipality size. They were omitted from the model one by one as described in chapter 2.5. The order of drop-outs, log-likelihoods and p-values are in the Table A9 in the Annex. Model without insignificant variables is in the Table A10 in the Annex.

The third step here is to remove insignificant values by merging them. The process is described in the Table A11 in the Annex. Note that the final model has log-likelihood – 976.3 with 5 parameters, whereas the maximum model has log-likelihood – 961.9 but with 29 parameters, and the model without explanatory variables has log-likelihood – 990.8.

Minimal adequate model is in the Table 2. Other thing being equal unemployment duration for women is longer than for men by 17.1%, for the youngest category (16–20 years) is shorter than for the rest of unemployed by 28.5% and is shorter by 32.8% for unemployed with university education in comparison to the rest.

Table 2 Minimal adequate model

Variable	Code	Parameter estimate bi	S.E.	p-value	exp(bi)
Intercept		2.220	0.0506	0.000	–
Pohl	2	0.158	0.0645	0.014	1.171
VekSk	1	–0.336	0.1087	0.002	0.715
ISCED	5	–0.398	0.0974	0.000	0.672
Log(scale)		–0.799	0.0378	0.000	–

Source: CZSO, own construction

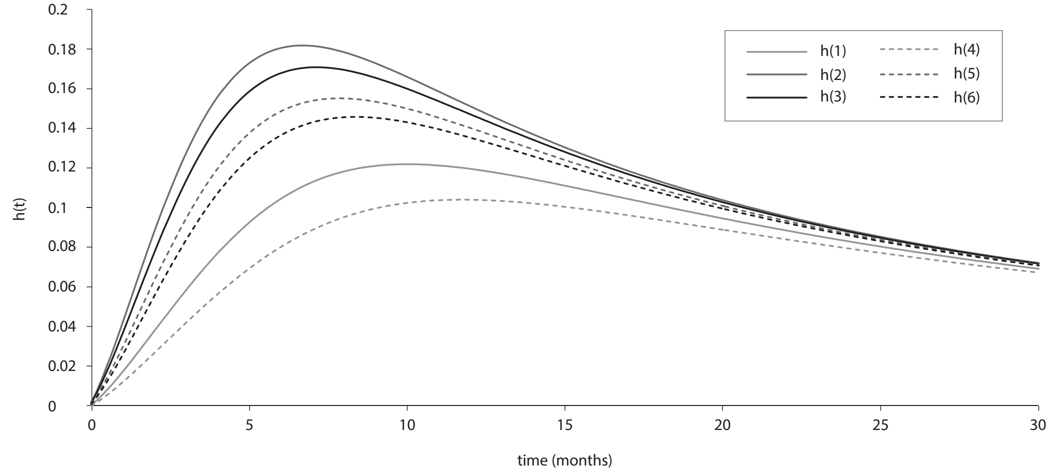
Table 3 Effects of combinations on time scale

N.	Sex	Age group	ISCED	exp(x´β)	E(X)	Var(X)	Median	x0.9	Mode	Obs.
1	Male	> 20	2–4	1	13.174	598.443	9.207	24.736	5.955	245
2	Male	> 20	5	0.672	8.848	269.975	6.184	16.614	3.996	29
3	Male	16–20	2–4	0.715	9.414	305.617	6.580	17.677	4.256	34
4	Female	> 20	2–4	1.171	15.429	820.842	10.783	28.970	6.974	280
5	Female	> 20	5	0.787	10.363	370.307	7.243	19.458	4.684	54
6	Female	16–20	2–4	0.837	11.026	419.193	7.706	20.703	4.984	31

Source: CZSO, own construction

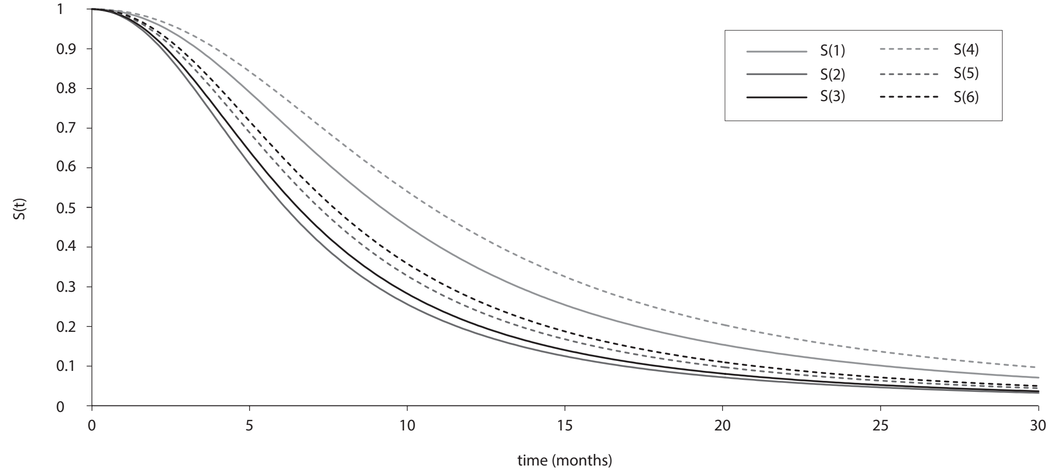
The baseline distribution is now for men older than 20 years without university education. For baseline distribution parameters and characteristics are α = 9.207, β = 2.223, b = 1.413, E(X) = 13.174, Var(X) = 598.443, median = 9.207, $x_{0.9}$ = 24.736 and mode = 5.955. Characteristics for all 6 possible combinations are in the Table 3 – note that 2 combinations (for men and for women) of university education and age group 16 – 20 years are impossible. Figures 1 and 2 contain hazard functions and survival functions of all possible combinations, respectively.

Figure 1 Hazard functions for all six possible combinations



Source: CZSO, own calculations

Figure 2 Survival functions for all six possible combinations



Source: CZSO, own calculations

4 DISCUSSION

4.1 Data limitations

At the present paper the status of economic activity is checked at the entry into the survey and at the exit. It is much easier to find the participants who found a job this way, but it means that there are possible omitted cases – firstly the situation in which a participant finds a job in between and then loses it, secondly the situation in which participants lose jobs and then find and lastly the situation when they find a job, lose it and find it again. These cases are possible but not very likely, so their omission should not change the overall results.

The more likely and thus problematic case is omitting of cases of participants, who were not in the labor force and then found a job. This is necessary because for these cases it is impossible to calculate the duration of their unemployment in the sense used throughout the paper.

Reader should still keep in mind that unemployment duration is calculated only for those, who found a job within a five quarter period. It is not unemployment duration at some specific time point and not an unemployment duration of those who did not find a job, neither.

4.2 Results in the context of previous research

The results presented here are in line with the results presented in Čabla (2015), where unemployment duration was different for variables sex and ISCED, but not for age groups without the youngest group 16–20 years. These results were obtained by several models with only one explanatory variable in each, whereas here I confirm those results using multivariable model.

The age group 16–20 years is in different situation than the rest of unemployed – they have usually secondary education with or without graduation (ISCED = 3 or 4) and in my hypothesis are looking for a job soon after the end of their education in the age of. The end of tertiary education is usually more dispersed and people leaving universities do not form a specific age group.

It presents a shift from the crisis situation (year 2010), in which marital status and age played significant role (Čabla, 2012). But note that the crisis results were obtained from the data set containing those who did not found a job either, so the results are not directly comparable.

The use of log-logistic distribution for modeling unemployment duration is not usual in last years but the distribution was used in the further past by Jarošová et al. (2004). The second distribution which fits the data is log-normal.

4.3 Further research

The presented results are just a part of a research of unemployment duration. They describe the post-crisis situation on the limited dataset of those who found a job. The focus should now move on the direct comparison of pre-crisis, crisis and post-crisis situation. It also would be useful to make similar research on the dataset containing all the unemployed at the beginning and research at specific time points to compare the results from these.

CONCLUSION

An unemployment is one of the leading economic problems and the paper contributes to our understanding of the problem. The paper identifies and quantifies the differences in the unemployment duration in different strata in the post-crisis Czech Republic via building a minimal adequate model.

The unemployment duration is described as a survival function $S(t)$ and hazard function $h(t)$ of log-logistic distribution as a part of accelerated failure time model with explaining variables. The estimated expected values, variances, medians, modes and 90th percentiles are provided for all subgroups.

The variables in the maximal model are sex, marital status, age, education municipality size and number of persons in a household and the model contains 29 parameters. The model is reduced in a backward selection manner with the use of likelihood ratio test – first the explanatory variables are reduced and then the categories of remaining variables are merged. The minimal adequate model contains five parameters – two of the log-logistic distribution and three describing the differences between men and women, the youngest and the rest and those with university education and the less educated.

The duration of unemployment is longer for women by 17.1%, shorter for the youngest category by 28.5% and for people with university education by 32.8%. The findings are limited to the group of those who found a job during the selected period, i.e. in the last quarter of year 2013 and year 2014. This is in line with previous findings but there is a need to make more direct comparison to the findings from previous time periods.

ACKNOWLEDGMENT

The paper was processed with contribution of long term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague. The support of the grant scheme IG410045 by the University of Economics IGA is gladly acknowledged.

References

CRAWLEY, M. J. *The R book*. 2nd Ed. Chichester: Wiley, 2013c, xxiv, 1051 p. ISBN 9780470973929.

CZSO. *Statistics VDB* [online]. 2015. [cit. 7.11.2015]. <https://vdb.czso.cz/vdbvo2/faces/en/index.jsf?page=statistiky#katalog=30853>.

ČABLA, A. *Unemployment duration in the Czech Republic.* In: The 6th International Days of Statistics and Economics, Conference Proceedings, Prague, 13.9.2012–15.9.2012, pp. 257–267. ISBN 978-80-86175-86-7.

ČABLA, A. Unemployment Duration before and during The Economic Crisis in the Czech Republic. *Acta Aerarii Publici*, 2014, Vol. 11, pp. 19–26. ISSN 1336-8818.

ČABLA, A. *Unemployment Duration in the Czech Republic After the Economic Crisis.* In: Applications of Mathematics and Statistics in Economics – AMSE, Jindřichův Hradec, 2.9.2015–6.9.2015 [CD ROM], Prague: University of Economics, Oeconomica Publishing House, 2015, 11 p. ISBN 978-80-245-2099-5.

EUROSTAT. *European Union Labour Force Survey* [online]. 2015. [cit. 7.11.2015]. <http://ec.europa.eu/eurostat/web/microdata/european-union-labour-force-survey>.

EUROSTAT. *LFS Database – Eurostat* [online]. 2015. [cit. 7.11.2015]. <http://ec.europa.eu/eurostat/web/lfs/data/database>.

FAY, M. P. *Package "interval". R Project* [online]. 2013. [cit. 17.11.2013]. <http://cran.r-project.org/web/packages/interval/interval.pdf>.

JAROŠOVÁ, E. Modelovani delky trvani nezamestnanosti. *Statistika*, 3/2006, pp. 240–251.

JAROŠOVÁ, E., MALÁ, I., ESSER, M., POPELKA, J. Modelling time of unemployment via log-location-scale model. In: ANTOCH, J. eds. *COMPSTAT 2004* [CD-ROM], Prague, 23.8.2004–27.8.2004. Heidelberg: Physica-Verlag, 2004, pp. 1255–1262. ISBN 3-7908-1554-3.

KLEINBAUM, D. G., KLEIN, M. *Survival Analysis: A Self-Learning Text.* 3rd Ed. New York: Springer-Verlag, 2012. ISBN 978-1-4419-6645-2.

KLEIN, J. P., MOESCHBERGER, M. L. *Survival Analysis: Techniques for Censored and Truncated Data.* New York: Springer-Verlag, 1997.

LIU, X. *Survival analysis: models and applications.* Peking: Higher Education Press, 2012, xii, 446 p. ISBN 0470977159.

MALÁ, I. Použití konečných směsí pravděpodobnostních rozdělení pro modelování rozdělení doby nezaměstnanosti v České republice. *Acta Oeconomica Pragensia*, 2013, Vol. 21, No. 5, pp. 47–63. ISSN 0572-3043. eISSN 1804-2112.

MALÁ, I. *The Use of Finite Mixture Model for Describing Differences in Unemployment Duration.* In: AMSE [CD ROM]. Jerzmanovice, 27.8.2014–31.8.2014. Wroclaw: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, 2014, pp. 164–172. ISBN 978-83-7695-421-9.

OECD. Stat. *Average duration of unemployment* [online]. 2015. [cit. 7.11.2015]. <https://stats.oecd.org/Index.aspx?DataSetCode=AVD_DUR>.

THERNEAU, T. *Package "survival". R Project* [online]. 2013. [cit. 17.11.2013]. <http://cran.r-project.org/web/packages/survival/survival.pdf>.

TURNBULL, B. W. The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data. *Journal of the Royal Statistical Society,* B38, 1976, pp. 290–295.

ANNEX

Table A1 Coding and observations of Number of persons in the households

Code	Observations
1	63
2	181
3	179
4	183
5	49
6	13
7	4
8	1

Source: CZSO, own construction

Table A2 Coding and observations of Sex

Code	Meaning	Observations
1	Male	308
2	Female	365

Source: CZSO, own construction

Table A3 Coding and observations of Marital Status

Code	Meaning	Observations
1	Single	319
2	Married	252
3	Widowed	13
4	Divorced	89

Source: CZSO, own construction

Table A4 Coding observations observations of Age group

Code	Meaning	Observations
1	Age 16–20	65
2	21–25	116
3	26–30	81
4	31–35	65
5	36–40	109
6	41–45	61
7	46–50	72
8	51–55	58
9	56–60	39
10	Age > 60	7

Source: CZSO, own construction

Table A5 Coding and observations of ISCED

Code	Meaning	Oservations
2	Primary education	54
3	Secondary without graduation	302
4	Secondary with graduation	234
5	Terciary	83

Source: CZSO, own construction

Table A6 Coding and observations of Municipality Size

Code	Meaning	Observations
1	Population < 1 000	133
2	1 000–9 999	219
3	10 000–49 999	170
4	50 000–99 999	71
5	Population > 100 000	80

Source: CZSO, own construction

Table A7 Log-likelihoods of selected distributions for whole population

Distribution	Log-likelihood
Loglogistic	– 990.8
lognormal	– 991.2
Weibull	– 1 040.8
exponential	– 1 055
t	– 1 154.9
rayleigh	– 1 197.7
logistic	– 1 197.9
gaussian	– 1 262.7

Source: CZSO, own construction

Table A8 Maximal model (statistically significant parameters are bold)

Variable	Code	Parameter estimate	S.E.	p-value
Intercept		**2.46876**	**0.1098**	**0.000**
PocOD	1	–0.12615	0.1296	0.330
	3	0.12841	0.0893	0.150
	4	0.15176	0.0952	0.111
	5	0.16093	0.1417	0.256
	6	0.29967	0.2510	0.233
	7	–0.13028	0.4008	0.745
	8	0.58231	0.6471	0.368
Pohl	**2**	**0.18639**	**0.0679**	**0.006**
RodStav	2	–0.18822	0.1109	0.090
	3	–0.43506	0.2599	0.094
	4	–0.10593	0.1303	0.416
VekSk	**1**	**–0.41861**	**0.1310**	**0.001**
	3	–0.08182	0.1206	0.497
	4	0.00721	0.1364	0.958
	5	0.09595	0.1345	0.476
	6	–0.05861	0.1623	0.718
	7	0.15552	0.1599	0.332
	8	**0.36252**	**0.1697**	**0.033**
	9	0.26462	0.1951	0.170
	10	–0.23308	0.3661	0.524
ISCED	2	0.12389	0.1227	0.313
	4	–0.07346	0.0752	0.328
	5	**–0.41784**	**0.1090**	**0.000**
MuniSize	1	0.03356	0.0923	0.716
	3	–0.00589	0.0866	0.946
	4	–0.02842	0.1130	0.801
	5	–0.17781	0.1113	0.110
Log(scale)		**–0.82385**	**0.0379**	**0.000**

Source: CZSO, own construction

Table A9 Dropouts and log-likelihood

Dropped variable	Log-likelihood	p-value
Maximal model	–961.9	NA
Municipality Size	–963.7	0.482
Number of Persons in a Household	–968.2	0.318
Marital Status	–969.5	0.363

Source: CZSO, own construction

Table A10 Model without insignificant variables (statistically significant parameters are bold)

Variable	Code	Parameter estimate	S.E.	p-value
Intercept		**2.27551**	**0.0939**	**0.000**
Pohl	**2**	**0.17609**	**0.0658**	**0.007**
VekSk	**1**	**–0.33995**	**0.1292**	**0.009**
	3	–0.09822	0.1181	0.406
	4	–0.03978	0.1262	0.753
	5	–0.01608	0.1131	0.887
	6	–0.17116	0.1360	0.208
	7	–0.00102	0.1234	0.993
	8	**0.17244**	**0.1338**	**0.198**
	9	0.03999	0.1595	0.802
	10	–0.60855	0.3373	0.071
ISCED	2	0.08661	0.1224	0.479
	4	–0.11623	0.0741	0.117
	5	**–0.44643**	**0.1076**	**0.000**
Log(scale)		**–0.81103**	**0.0379**	**0.000**

Source: CZSO, own construction

Table A11 Merging variables

Step	Variable	Values merged	Log-likelihood	p-value
Start	NA	NA	–969.5	NA
1	ISCED	2+3	–969.8	0.401
2	Age Group	4+5	–969.8	0.471
3		3+4+5	–970.0	0.513
4		2+3+4+5	–970.2	0.558
5		8+9	–970.4	0.589
6		7+8+9	–970.9	0.593
7		2+3+4+5+6	–971.6	0.559
8		2–9	–972.9	0.459
9		2–10	–974.6	0.332
10	ISCED	2–4	–976.3	0.228

Source: CZSO, own construction