ISSN 2045-3345

MODELING UNEMPLOYMENT DURATION IN THE CZECH REPUBLIC FROM LFS

Adam Čabla[I]
University of Economics, Prague, Czech Republic

ABSTRACT

The article deals with the problem of modeling individual unemployment duration in the Czech Republic using data from the CZSO´s Labour Force Surveys (LFS). Standard approach to the interval censored data is used and the hypotheses about different survival distributions for the different groups are tested, thus identifying the more and less endangered strata of the society. Moreover AFT model is used to evaluate the difference between strata. The main aim is to use standard techniques and available software for interval censored data that is outcome from several LFS datasets.

Results indicates, that there may be a connection between unemployment duration and sex, marital status, age and handicap status, but the data were not adjusted to avoid problems with relationship between explanatory variables. On the other hand, the data does not prove connection between unemployment duration and region, education and size of municipality. The first two were sooner found as significant covariates and the difference may be due to the more complicated data structure used.

JEL CLASSIFICATION & KEYWORDS

■ J64 ■ C14 ■ C24 ■ UNEMPLOYMENT DURATION ■ INTERVAL CENSORING ■ NONPARAMETRIC ESTIMATE ■ LABOR FORCE SURVEY

INTRODUCTION

Unemployment is one of the leading problems in economy and thus the object of interest to many people. Standard statistics about unemployment covers mainly unemployment rate and as for the duration of the unemployment the rate of long-term unemployment (over a year) is recorded. Those statistics are recorded by national statistics offices. This is of course reasonable way of thinking about unemployment but the other important matter in the topic is the duration of unemployment.

Some statistics about average duration of unemployment are provided for the US states as well as for the OECD countries (OECD, 2012). Nevertheless those later are somewhat troubling as the development at least for the Czech Republic is unrealistic. The only deeper look at the unemployment duration in the Czech Republic is provided by Jarosova et al (2004) and Jarosova (2006). This article is similar to those two in some respects as the data comes from the same source but are newer.

The main problem in modeling the unemployment duration is the fact that only data available are interval and right censored, so the researcher must deal with it and use the proper techniques of the survival analysis. The nonparametric estimates comes from Turnbull (1976), the tests comparing two or more distribution functions are described by Fay and Shaw (2010), parametric models are dealt with e.g. in Fay (2010). In the text there are several references to the previous research that was made for conferences but not published yet.

Methodology

Data are called censored when the exact value is not known, but this value falls somewhere in an interval (Li, Ri], i = 1, …, n. When Li is minus infinity, than the value is called left censored and when Ri is infinity, than the value is called right censored. When both bounds are known, the value is interval censored. This happens to be a case in used data.

Survival function is the probability of a randomly chosen value to fall beyond the x, so

$$S(x) = P(X>x) = 1 - F(x)$$

The survival function is often of the main interest. In the problem of unemployment duration represents survival function probability, that randomly chosen person´s duration of unemployment is longer than the asked value.

Nonparametric estimate of the survival function is done by the iterative procedure first suggested by Turnbull (1976). These estimates are maximum likelihood estimates, usually used abbreviation is NPMLE. The estimate is done as follows from Klein, Moeschberger (1997):

Let $0 = \tau_0 < \tau_1 < \ldots < t_m$ be a grid of time points which includes all the points L_i, R_i for the points i = 1,…,n. For the *i*th observation define a weight α_{ij} to be 1 if the interval $(\tau_{j-1}, \tau_j]$ is contained in the interval $(L_i, R_i]$ and 0, otherwise. Note that α_{ij} indicates whether event which occurs in the interval $(L_i, R_i]$ could have occurred at τ_j. An initial guess at $S(\tau)$ is made. The algorithm is as follows:

1. Compute the probability of an event occurring at time τ_j, $p_j = S(\tau_{j-1}) - S(\tau_j)$, j = 1,…,m
2. Estimate the number of events which occurred at τ_j by $d_j = \sum_{i=1}^{n} \frac{a_{ij} p_j}{\sum_k a_{ik} p_k}$.
3. Compute the estimated number at risk at time τ_j by $Y_j = \sum_{k=j}^{m} d_k$
4. Compute the updated Product-Limit estimator using the pseudo data found in Steps 2 and 3. If the updated estimate of S is close to the old version of S for all τ_j´s, stop the iterative process, otherwise repeat Steps 1 – 3, using the updated estimate of S.

For testing hypothesis about equality of survival function are used asymptotic logrank two-sample and k-sample tests with Sun´s scores and exact Wilcoxon two-sample and k-sample tests based on 999 Monte Carlo simulations.

The estimates of means for different groups could not be obtained as there are too many right censored observations, which give the large part of the NPMLE to fall in the last interval ending at plus infinity. The arbitrary setting of the right end is not viable as the data or theory gives no clue

[I] adam.cabla@vse.cz

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

and the estimated mean depends heavily on it and no suitable parametric estimate of this survival function was found yet.

Finally the accelerated failure time (AFT) models for each single covariate were estimated thus evaluating the differences between the unemployment duration for different strata of society. AFT is family of parametric regression models, where the effect of covariate is to change the time scale.

$$log\ (X_i) = \alpha + z_i\beta + \sigma\varepsilon$$

Where α and σ are location and scale parameters, ε is the error with known distribution, z_i is vector of covariates and β is vector of parameters. In using AFT models first important part is to find best fitting distribution of ε (information criteria can be used) and then estimate parameters and test the model using χ^2 test for change of logarithm of likelihood. Exp (b_i) gives the fold of change in time compared to the selected basis.

Data

Data were obtained from the Czech Statistical Office and come from the LFS. This survey is done once per quarter and every person is asked in five following surveys. In one survey approximately 50 – 60 thousand persons are questioned. Five following datasets in the quarters Q1/2010 – Q1/2011 were used.

One of the questions asks for the duration of seeking a job and another one asks for the duration of the current work. Due to the fact that person is questioned during year and a quarter, one can find not only persons, that are looking for a job, but also those, who obtained job in this period and compute the duration of the search retroactively. As the answers to the stated questions are interval, so is the consequent duration. Finally 673 interval censored and 1857 right censored values were obtained. This dataset contains those who found a job and those who keep searching. This may be discussable as the true time to obtaining a job may be lower than the estimate, but the point values are not of the main interest for now. Table 1 shows rounding of some values to the days.

Table 1: Rounding	
Value	**Rounded to days**
1 month	30
1 year	365
18 months	545
Source: own calculations	

Explaining variables were: NUTS region, sex, marital status, age group, handicap status, education according to ISCED and size of municipality. Table 2 contains age groups, Table 3 educational groups according to ISCED and Table 4 arbitrary chosen sizes of municipalities.

Table 2: Age groups		
Value	**Age**	**n_i**
3	15 - 24	214
4	25 – 29	273
5	30 – 34	293
6	35 - 39	304
7	40 - 44	287
8	45 - 49	285
9	50 - 54	329
10+	> 55	473
Source: CZSO, own calculations		

The lowest and the highest age groups were combined to provide sufficient number of cases.

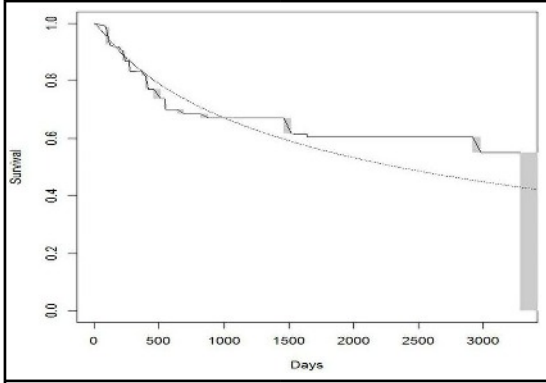
Table 3: ISCED		
Value	**Education**	**n_i**
2	Basic	485
3	High school without leaving exam	1244
4	High school with leaving exam	626
5	University	174
Source: CZSO, own calculations		

Table 4: Size of municipality		
Value	**Dwellers**	**n_i**
Small	< 3 000	918
Medium	3 000 – 30 000	842
Large	> 30 000	769
Source: CZSO, own calculations		

The classes of size of municipalities were determined arbitrary to ensure similar number of cases in each class.

Figure 1 represents the NPMLE of the survival function (solid) for whole dataset and lognormal estimate used later in regressions (dotted). Grey areas represent intervals without known distribution, the line shows the linear interpolation from the two endpoints. Lognormal distribution was found to be best fitting with parameters $\alpha = 7.76$ and $\sigma = 1.92$.

Figure 1: Nonparametric and lognormal estimate of unemployment duration for whole dataset



Source: CZSO, own calculations

NUTS regions

As for the stratification according to the NUTS regions the NMPLE shows some differences, but the both k-sample tests did not rejected null hypothesis at the alpha = 0.05. Wilcoxon test is shown in Table 5, initial interval estimate of the p-value had lower bond under 0.05, so the more accurate one with 99 999 replications had to be done to remove all doubts.

This may be surprising finding for the sooner research showed significant differences between regions, but it may be due to the high ratio of right censored observations that makes more than a half of the estimate of survival function fall in the last interval ending at plus infinity.

Sex

Difference between sexes is well known and was confirmed by the presented tests in Table 6. Both tests suggest that women´s unemployment duration is generally longer than men´s. This fact is confirmed by AFT model in Table 7. The parameter estimate for women is 0.2906, which transforms

Table 5: Wilcoxon k-sample test for the NUTS regions		
p - value =	**0.079**	
alternative hypothesis: survival distributions not equal		
NUTS	**n**	**Score Statistics ***
Southwest	293	13.88
Central Moravia	343	9.39
Moravian Silesian	383	3.79
Southeast	456	-0.30
Central Bohemia	237	-0.39
Northwest	375	-4.94
Prague	89	-9.09
Northeast	353	-12.94
* like Obs - Exp, positive implies earlier failure than expected		
p - value estimated from 99 999 Monte Carlo replications		
99 percent confidence interval on p-value:		
0.077	0.082	
Source: CZSO, own calculations		

in 1.34 fold longer time of unemployment compared to men´s unemployment duration.

Table 6: Asymptotic logrank and Wilcoxon 2-sample tests for the Sex		
Asymptotic logrank 2-sample test		
p - value =	**0.007**	
alternative hypothesis: survival distributions not equal		
Sex	**n**	**Score Statistics ***
Man	1204	34.04
Woman	1325	-34.04
* like Obs - Exp, positive implies earlier failure than expected		
Exact Wilcoxon 2-sample test (permutation form)		
p - value =	**0.006**	
alternative hypothesis: survival distributions not equal		
Sex	**n**	**Score Statistics ***
Man	1204	-30.67
Woman	1325	30.67
* like Obs - Exp, positive implies earlier failure than expected		
p - value estimated from 999 Monte Carlo replications		
99 percent confidence interval on p-value:		
< 0.001	0.018	
Source: CZSO, own calculations		

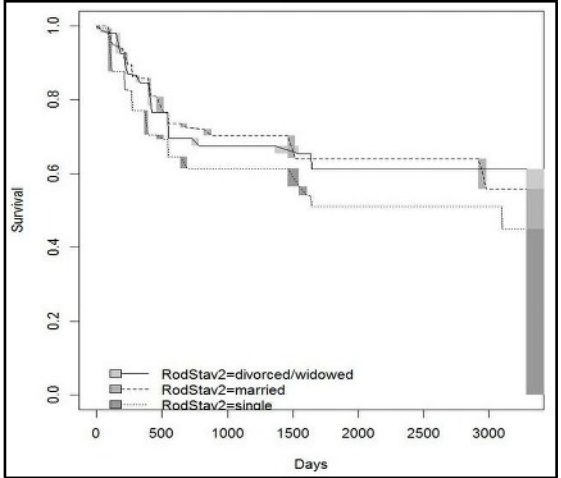
Table 7: AFT model for sex	
Coefficients:	
Intercept =	7.29
Scale =	1.55
Woman =	0.2906
Loglik (model) =	-2250.9
Loglik (intercept only) =	-2255.1
Chisq =	**8.46**
on 1 degrees of freedom	
p =	**0.0036**
Source: CZSO, own calculations	

Marital status

Marital status was confirmed by both tests to influence the unemployment time. Single finds new job generally sooner than divorce or widowed and those than married. These findings were confirmed by AFT model, both outputs are in Table 8. Time of unemployment is 1.61 fold longer for divorced or widowed compared to single´s time of unemployment and 1.81 fold longer for married. This seems to be contrary to the previous findings claiming that unemployment time is longer for divorced/widowed than for married. Figure 2 shows NPMLE of survival functions that suggests, that time of unemployment is very similar for these two categories and only significantly different survival function is the one for single.

Table 8: Asymptotic logrank and Wilcoxon k-sample tests and AFT model for the Marital Status		
Asymptotic logrank k-sample test		
p - value =	**< 0.001**	
Sex	**n**	**Score Statistics ***
Single	785	53.65
Divorced/Widowed	515	-11.03
Married	1158	-42.62
* like Obs - Exp, positive implies earlier failure than expected		
Exact Wilcoxon k-sample test (permutation form)		
p - value =	**0.001**	
NUTS	**n**	**Score Statistics ***
Single	785	49.22
Divorced/Widowed	515	-8.93
Married	1158	-40.29
* like Obs - Exp, positive implies earlier failure than expected		
p - value estimated from 999 Monte Carlo replications		
99 percent confidence interval on p-value:		
< 0.001	0.005	
AFT model		
Coefficients:		
Intercept =	7.31	
Scale =	1.22	
Married =	0.1225	
Single =	-0.4733	
Loglik (model) =		-2195.5
Loglik (intercept only) =		-2209.7
Chisq =	**28.43**	
on 2 degrees of freedom		
p =	**< 0.0001**	
Source: CZSO, own calculations		

Figure 2: NPMLE of the survival functions for Marital Status



Source: CZSO, own calculations

Age group

Unemployment time differs significantly according to age groups. Wilcoxon k-sample test and AFT model both pictures the same order of groups according to unemployment time. The shortest time of unemployment have people with age of 15 – 34 year followed by those with age 50 – 54 and then with age 35 – 49. The longest time of unemployment was found in the group with age 55 and more. This order is different from previous examination especially for the oldest group and may be result of including those who did not find a job during the follow-up period. Findings are summarized in Table 9 and Table 10.

Table 9: Wilcoxon k-sample test and AFT model for Age groups

Exact Wilcoxon k-sample test (permutation form)			AFT model	
p - value =	**0.001**		**Coefficients:**	
alternative hypothesis: survival distributions not equal			**Intercept =**	8.53
Age group	**N**	**Score Statistics ***	**Scale =**	1.99
15 - 24	214	25.11	**15 - 24**	-1.2493
25 - 29	273	34.05	**25 - 29**	-1.2786
30 - 34	293	18.13	**30 - 34**	-0.8784
35 - 39	304	-11.39	**35 - 39**	-0.3206
40 - 44	287	-5.62	**40 - 44**	-0.4694
45 - 49	285	-18.16	**45 - 49**	-0.2281
50 - 55	329	0.69	**50 - 55**	-0.5663
> 55	473	-42.82	Loglik (model) =	-2195.5
* like Obs - Exp, positive implies earlier failure than expected			Loglik (int. only) =	-2209.7
p - value estimated from 999 Monte Carlo replications			Chisq =	**28.43**
99 percent confidence interval on p-value:			on 2 degrees of freedom	
< 0.001	0.005		p =	**< 0.0001**

Source: CZSO, own calculations

Table 10: Folds of the time of unemployment for age groups compared to group 10+

Age	**Fold**
15 - 24	0,287
25 – 29	0,278
30 – 34	0,415
35 - 39	0,726
40 - 44	0,625
45 - 49	0,796
50 - 54	0,568
> 55	1

Source: CZSO, own calculations

Handicap status

It is well known, that handicapped have longer time of unemployment which was confirmed by the tests and the difference was numerated by the AFT model. Both of these are in Table 11. Time of unemployment of people with handicap is 3.38 fold of time of people without handicap.

Education according to ISCED

Even though plot of NPMLE of survival functions for different groups suggests, that those with higher education have shorter unemployment duration, formal tests did not confirm that. Table 12 contains Wilcoxon test and Figure 3 shows the plot. It seems that difference between educational groups diminishes around 1 500 days of unemployment.

Table 11: Asymptotic logrank and Wilcoxon 2-sample test and AFT model for Handicap status

Asymptotic logrank 2-sample test		
p - value =	**< 0.001**	
Handicap	**n**	**Score Statistics ***
No	2234	52.17
Yes	295	-52.17
Exact Wilcoxon 2-sample test (permutation form)		
p - value =	**0.002**	
Handicap	**n**	**Score Statistics ***
No	2234	45.09
Yes	295	-45.09
p - value estimated from 999 Monte Carlo replications		
99 percent confidence interval on p-value:		
< 0.001	0.011	
AFT model		
Coefficients:		
Intercept =	6.69	
Scale =	1.63	
Handicap	1.75	
Loglik (model) =		-2231.2
Loglik (intercept only) =		-2255.1
Chisq =	**47.8**	
p =	**< 0.0001**	

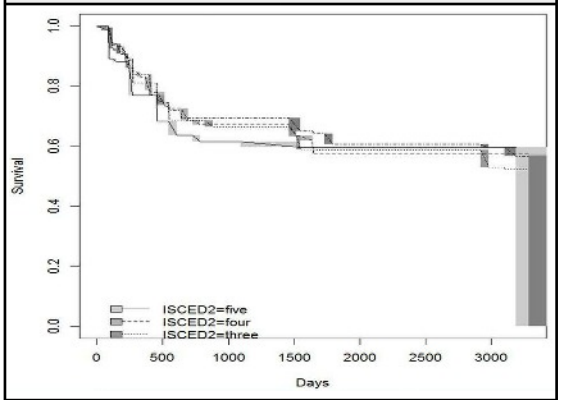
Source: CZSO, own calculations

Table 12: Exact Wilcoxon k-sample test for ISCED groups

Exact Wilcoxon k-sample test (permutation form)		
p - value =	**0.493**	
alternative hypothesis: survival distributions not equal		
ISCED	**n**	**Score Statistics ***
2	485	-5.96
3	1244	4.94
4	626	-4.79
5	103	5.81
* like Obs - Exp, positive implies earlier failure than expected		
p - value estimated from 999 Monte Carlo replications		
99 percent confidence interval on p-value:		
0.451	0.534	

Source: CZSO, own calculations

Figure 3: NPMLE of survival functions for ISCED educational classes



Source: CZSO, own calculations

Municipality size

Estimated time of unemployment does not differ by municipality sizes. Logrank and Wilcoxon tests are in Table 13.

Table 13: Asymptotic logrank and exactz Wilcoxon k-sample tests for the Municipality Size		
Asymptotic logrank 2-sample test		
p - value =	**0.128**	
alternative hypothesis: survival distributions not equal		
Municipality size	**n**	**Score Statistics ***
Small	918	-2.72
Medium	842	22.56
Large	769	-19.84
* like Obs - Exp, positive implies earlier failure than expected		
Exact Wilcoxon k-sample test (permutation form)		
p - value =	**0.124**	
alternative hypothesis: survival distributions not equal		
Municipality size	**n**	**Score Statistics ***
Small	918	-1.89
Medium	842	20.10
Large	769	-19.84
* like Obs - Exp, positive implies earlier failure than expected		
p - value estimated from 999 Monte Carlo replications		
99 percent confidence interval on p-value:		
0.098	0.152	
Source: CZSO, own calculations		

Conclusion

According to the findings, the unemployment duration is influenced by the sex, marital status, age group and handicap status. But these results are little tricky as there are relationships between these covariates. For example older people´s marital status is more often to be divorced/widowed. There is still a need for deeper look at the relationship between these covariates and next step in the analysis shall be removal of the influence of these relationships. The second problem arises from the available data, more than a half of them is right censored while remaining part is interval censored, which makes NPMLE. Note that the dependent variable is time of unemployment before finding new one or just the time of unemployment mixed together.

References

Belyaev, Yuri and Bengt Kristrom. 2010. Approach to Analysis of Self-Selected Interval

Data. Technical report SLU, Department of Forest Economics 90183 Umea, Sweden.

Fay, Michael P., Shaw Pamela A., “Exact and Asymptotic Weighted Logrank Tests for Interval Censored Data: The interval R Package.” Journal of Statistical Software Vol. 36, Issue 2, Aug 2010

Fay, Michael. “Interval Censored Data Analysis.” July 2010. < http://user2010.org/tutorials/Fay_1.pdf>

Giolo, Suely Ruiz. “Turnbull´s Nonparametric Estimator for Interval-Censored Data”. Technical Report. Aug. 2004. < http://www.ms.uky.edu/~mai/splus/IcensEM.pdf>

Jarosova, Eva. “Modelovani delky trvani nezamestnanosti.” Statistika 3/2006: 240 – 251.

Jarosova E., Mala I., Esser M., Popelka J. “Modelling time of Unemployment via Log-location-scale Model.” COMPSTAT 2004 Symposium: 1 – 8.

Klein, John P., Moeschberger, Melvin L. “Survival Analysis: Techniques for Censored and Truncated Data. New Yourk: Springer-Verlag New Your, Inc., 1997.

OECD. “Average duration of unemployment.” OECD.StatExtracts. 31 July 2012. <http://stats.oecd.org/Index.aspx?QueryId=25006>

Turnbull, B. W. “The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data.” Journal of the Royal Statistical Society B38 1976: 290 – 295.

Zhou Wenchao, Belyaev Yuri, Kristrom Bengt. “iwtp: Software for Analysis of Self-Selected Interval Data”. 29 Feb. 2012. < http://finzi.psych.upenn.edu/R/library/iwtp/doc/manual.pdf >