UNEMPLOYMENT DURATION IN THE CZECH REPUBLIC

Adam Čabla

Abstract

The article deals with the problem of modelling 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. 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 region, number of persons in household, age, sex, education, marital status 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 cooperation with employment bureau, previous job experience, status at last job (employee or employer) or activity before seeking a job.

Key words: unemployment duration, interval censoring, nonparametric estimate, labour force survey

JEL Code: J64, C14, C24

Introduction

Unemployment is one of the leading problems in economy and thus the object of interest of 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 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 censored (and 17 left censored), so the researcher must deal with the fact 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).

1 Brief summary of terms

This chapter introduces the basics of the survival analysis and sources for the practical use of the R packages, dealing with the problem of interval censored data and are used for the computations.

Data are called censored when the exact value is not known, but this value falls somewhere in an interval $(L_i, R_i]$, i = 1, …, n. When L_i is minus infinity, than the value is called left censored and when R_i 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.

The other functions that are often of interest are hazard function and cumulative hazard function. The first one is $h(x) = f(x)/S(x)$ and can be viewed as the "approximate" probability of the gaining a job in the next instant. The second one is

$$H(x) = \int_0^x h(u)du.$$

Nevertheless those two functions are not part of the work.

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.

For testing hypothesis about equality of survival function is used two-sample weighted log-rank test and K-sample weighted log-rank test. The package "interval" in R provides five different score tests, Sun´s scores and permutation form of test were used. These tests provide

estimates of the effect of covariates. If the covariate is a factor, the effect of each value is provided. If the covariate is numeric, the algorithm computes log-rank trend test and again effect of the covariate.

Where it was possible given the enough observations to find mean for different subsamples, it was estimated by nonparametric estimates using Turnbull estimator. The only problem arises with the observations, for which the data were left not interval censored. Theirs right end was set arbitrary to be 3 650 days (10 years), nevertheless this value does not influence the nonparametric mean estimates.

2   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 persons, that obtained job in this period and compute the duration of the search. As the answers to the stated questions are interval, so is the consequent duration. Finally 806 persons for which these intervals could be counted were found. Table 1 contains recounting of some values to the days.

Tab. 1: Rounding

Value	Rounded to days
1 month	30
3 months	90
6 months	180
12 months	365
18 months	545
2 years	730
4 years	1460

Source: CZSO LFS, own calculations

Explaining variables were: NUTS-2 region, number of persons in the household, sex, marital status, age group, handicap, education according to ISCED, employment bureau, job experience, status at last job (employee or employer) and activity before seeking a job. Table 2 contains age groups, Table 3 educational groups according to ISCED, Table 4 type of time is soughed job and Table 5 activity before seeking a job.

Tab. 2: Age groups

Group no.	Age
2	15 - 19
3	20 - 24
4	25 - 29
5	30 - 34
6	35 - 39
7	40 - 44
8	45 - 49
9	50 - 54
10+	55 +

Source: CZSO LFS

Tab. 3: Educational groups according to ISCED

ISCED	Description
1	No education
2	Basic education
3	High-school no leaving exam
4	High-school with leaving exam
5	University

Source: CZSO LFS

Tab. 4: Type of time in soughed job

Code	Time in new job
1	Full-time
2	Part-time or Full-time
3	Part-time
4	Cannot specify

Source: CZSO LFS

Tab. 5: Activity before seeking

Code	Activity
1	Job
2	Study
3	Other (home, retired, maternity leave)

Source: CZSO LFS

The estimated mean unemployment duration in the whole dataset is 399.8 days. The survival function is pictured in the Figure 1. It is quite easily to find nonparametric estimate of median unemployment duration 455 days.

Fig. 1: Nonparametric estimate of unemployment duration
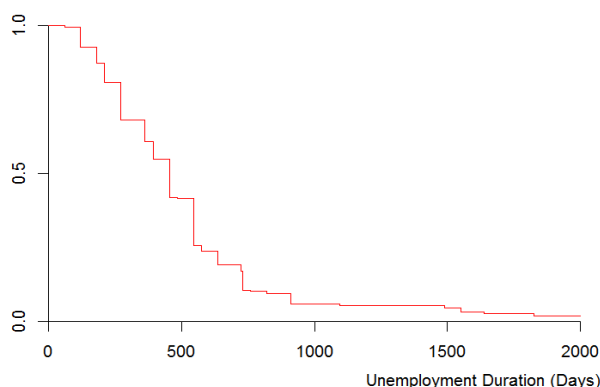
Source: CZSO LFS, own calculations

3   Results

In this chapter results from the tests and mean and median estimates are provided.

3.1   NUTS2 regions

As for the region, the k-sample log-rank test´s p-value is 0.00001. The null hypotheses of the equality of survival distributions is rejected at $\alpha = 0.05$. The results are shown in the Figure 1 suggesting, that shorter duration of unemployment are e.g. in Prague, Central Bohemia and Southeast regions and longer durations are in Northwest, Moravian-Silesian, or Southwest region.

Fig. 2: K-sample log-rank test for regions

```
            Asymptotic Logrank k-sample test (permutation form), Sun's scores

data:  my.surv by NUTS2
Chi Square = 37.2071, p-value = 4.285e-06
alternative hypothesis: survival distributions not equal

                 n Score Statistic*
NUTS2=Prague    37        11.597693
NUTS2=CentBoh   65        12.827553
NUTS2=Southw   129        -6.419514
NUTS2=NorthW    86       -34.646865
NUTS2=NorthE   128         1.628777
NUTS2=SouthE   143        13.443898
NUTS2=CenM     102         9.895483
NUTS2=MorSil   115        -8.327025
* like Obs-Exp, positive implies earlier failures than expected
```

Source: CZSO LFS, own calculations

The mean and median estimates are provided in the Table 6 except for Central Moravia, for which nonparametric estimate could not be obtained due to data insufficiency. The results are the same.

261

Tab. 6: Mean and median estimates of unemployment duration for regions in days

Region	Prague	Central Bohemia	Southwest	Northwest	Northeast	Southeast	Central Moravia	Moravian-Silesian
Mean e.	288	329	427	550	384	366	NA	438
Median e.	210	360	360	365	360	270	NA	360

Source: CZSO LFS, own calculations

3.2 Number of persons

Figure 3 and Table 7 shows that higher the number of persons in household the shorter the unemployment duration. That holds up to five persons in household. For six and more persons the unemployment duration is on the similar level as for the three to four persons in household. The regression models are not significant at $\alpha = 0.05$.

Fig. 3: K-sample log-rank test for number of persons in household

```
        Asymptotic Logrank k-sample test (permutation form), Sun's scores

data:  Surv(left, right, type = "interval2") by PocOs2
Chi Square = 14.2578, p-value = 0.01405
alternative hypothesis: survival distributions not equal

               n Score Statistic*
PocOs2=four  211     19.9687173
PocOs2=one    65    -19.9067136
PocOs2=two   196     -8.1409752
PocOs2=three 236      0.7212141
PocOs2=five   79      6.7647060
PocOs2=6plus  18      0.5930515
* like Obs-Exp, positive implies earlier failures than expected
```

Source: CZSO LFS, own calculations

Tab. 7: Mean and median estimates of unemployment duration for number of persons

Number of persons	One	Two	Three	Four	Five	Six or more
Mean e.	523	427	398	382	370	389
Median e.	365	360	360	360	360	270

Source: CZSO LFS, own calculations

3.3 Sex

For the covariate sex the test rejected null hypothesis (survival functions for men and woman are equal) at $\alpha = 0.05$ as it is shown in Figure 4. Men generally find new job sooner. The Figure 5 shows the difference – as far as for approximately 250 days unemployed, men´s and women´s survival function are very close, but between those looking for a job longer, the man are more successful in getting it sooner.

3.4 Marital status

The singles find jobs slightly sooner than married and both of these groups find job sooner than widowed or divorced people. The Figure 6 shows the difference between survival functions that is confirmed by p-value of the k-sample log-rank test (0.0046), and Table 9 between the mean and median estimates.

Fig. 4: K-sample log-rank test for sex

```
        Asymptotic Logrank two-sample test (permutation form), Sun's scores

data:  my.surv by Sex2
Z = -2.5172, p-value = 0.01183
alternative hypothesis: survival distributions not equal

            n Score Statistic*
Sex2=woman 365        -28.13091
Sex2=man   440         28.13091
* like Obs-Exp, positive implies earlier failures than expected
```

Source: CZSO LFS, own calculations

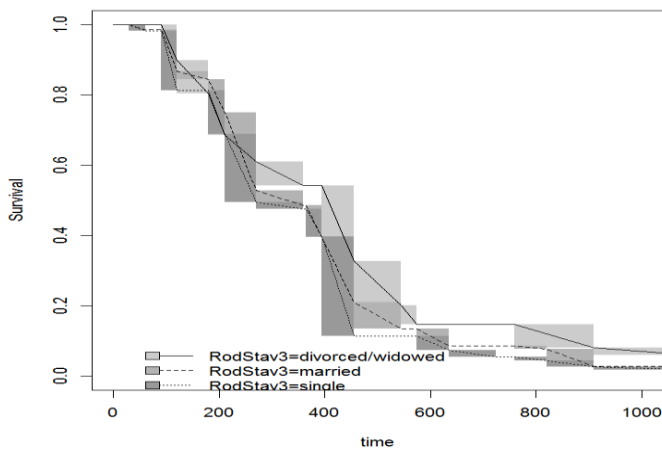**Fig. 5: NPMLE survival functions for men and women**
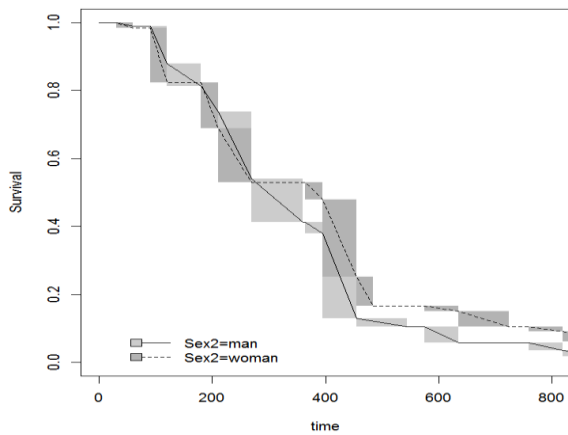
Source: CZSO LFS, own calculations

Tab. 8: Mean and median estimates of unemployment duration for sex

Sex	Man	Woman
Mean e.	377	431
Median e.	360	360

Source: CZSO LFS, own calculations-

Fig. 6: NPMLE survival functions for single, married and divorced/widowed

Source: CZSO LFS, own calculations

Tab. 9: Mean and median estimates of unemployment duration for marital status

Sex	Single	Married	Divorced/Widowed
Mean e.	372	404	598
Median e.	360	360	365

Source: CZSO LFS, own calculations

3.5 Age group

The Figure 7 shows the results of the test, that shows difference between survival functions for different age groups. From these results we can see, that the shortest duration of unemployment applies for people of age 20 – 24, than 25 – 29 followed by group of people with age 30 – 34. On the other side with the longest duration are groups with age 45 – 49, than 40 – 44 and 35 – 39. Surprisingly age group 10+ has quite shorter duration of unemployment. Table 10 shows the mean and median nonparametric estimates, where the mean duration of unemployment for the group with age 35 – 39 is surprisingly high.

Fig. 7: K-sample log-rank test for age groups

```
        Asymptotic Logrank k-sample test (permutation form), Sun's scores

data:  my.surv by VekSk2
Chi Square = 23.4261, p-value = 0.002858
alternative hypothesis: survival distributions not equal

             n Score Statistic*
VekSk2=10+  80        2.370218
VekSk2=7    74       -13.465333
VekSk2=6    82       -12.659559
VekSk2=9    94       -11.384217
VekSk2=4   114        13.017424
VekSk2=3   139        22.314050
VekSk2=8    83       -14.003607
VekSk2=5   103         9.775006
VekSk2=2    36         4.036019
* like Obs-Exp, positive implies earlier failures than expected
```

Source: CZSO LFS, own calculations

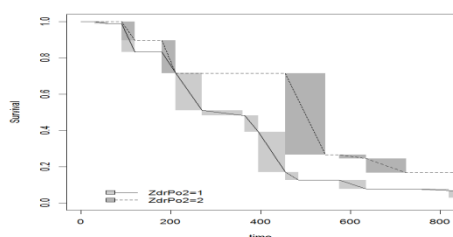**Tab. 10: Mean and median estimates of unemployment duration for age groups**

Age	15 – 19	20 – 24	25 – 29	30 – 34	35 – 39	40 – 44	45 – 49	50 – 54	≥ 55
Mean e.	352	336	351	376	573	472	473	448	388
Median e.	270	270	270	270	365	360	360	360	360

Source: CZSO LFS, own calculations

3.6 Handicap

Duration of unemployment for people with handicap is longer. The survival function is shown in the Figure 8, mean and median estimates in the Table 11. About 20 % of handicapped people look for a job longer than 635 days.

Fig. 8: NPMLE survival functions for handicapped (2) and non-handicapped (1)

Source: CZSO LFS, own calculations

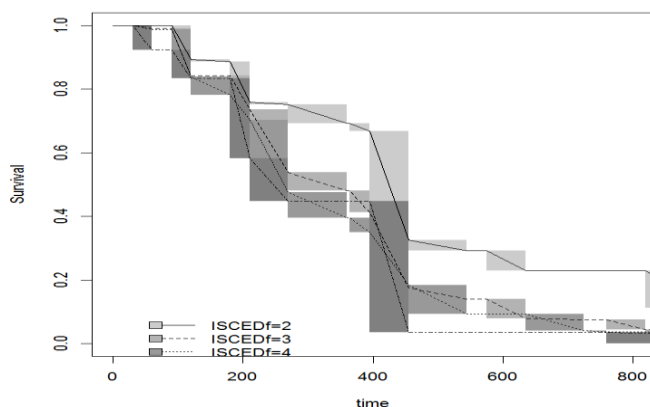**Tab. 11: Mean and median estimates of unemployment duration for handicapped**

Handicap	Yes	No
Mean e.	740	388
Median e.	395	360

Source: CZSO LFS, own calculations

3.7   Education according to ISCED

Results of K-sample log-rank test and plot of survival functions (Figure 9) show, that there is a negative relationship between level of education and unemployment duration ($p = 2.7e^{-07}$). The Table 12 contains mean and median estimates, sadly without estimate for the university education.

Fig. 9: NPMLE survival functions for educational groups

Source: CZSO LFS, own calculations

Tab. 11: Mean and median estimates of unemployment duration for educational groups

Educational grop	2	3	4	5
Mean e.	582	400	368	NA
Median e.	365	360	270	NA

Source: CZSO LFS, own calculations

3.8   Other covariates

Other covariates did not according to K-sample log-rank tests influenced estimated survival function. The only one worthy mentioning is cooperation with unemployment bureau. Those who cooperated with it and were receiving unemployment benefits seemed to find new job sooner than other two groups (cooperation with UB without unemployment benefits and non-cooperation with UB). But this result was influenced by the fact that unemployment benefits are time-limited and those, who did not find a job soon enough, moved to the second group. When those who cooperated with UB were taken as one group, no difference in survival functions was find.

Conclusion

According to findings, the unemployment duration is influenced by the region, number of persons in household, sex, marital status, age group, being handicapped and education. But these results are little tricky as there are relationships between these covariates. For example older people are generally less educated than younger, their marital status is more often to be divorced/widowed and lives in households with no one else or one other person. 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.

Acknowledgment

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 >.

Contact

Adam Čabla

Katedra statistiky a pravděpodobnosti, VŠE v Praze

nám. W. Churchilla 4

130 67 Praha 3

adam.cabla@vse.cz