PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1162 — 2007
Applications of Mathematics and Statistics in Economics

Eva Jarošová

University of Economics, Prague, Czech Republic

ESTIMATING THE UNEMPLOYMENT TIME MEDIAN IN THE CZECH REPUBLIC

1. Introduction

The level of unemployment belongs to the features of state's economy that are given the greatest attention. The number of unemployed and the unemployment rate in the Czech Republic are published quarterly based on the results of the Labour Force Sample Survey (LFSS). The survey is organized by the Czech Statistical Office (CZSO) and is performed at more than 25 thousand households randomly selected from all the Czech Republic. A part of the survey's output can be found on the web site of the CZSO, where the numbers of the unemployed are given not only for the Czech Republic as a whole, but they are also classified by sex, age groups, education and some other characteristics. Classification by job seeking duration is particularly of interest. Unequal intervals represent approximately the time the unemployed had spent seeking a job before they were surveyed. Although these data provide some information on the dynamic feature of unemployment, they do not show directly the time between the beginning and the end of a period of unemployment (TBBE for short).

The aim of our study* was to investigate the distribution of TBBE and to relate this to the various demographic and socio-economic factors. We used the "survival data" approach because no matter what scheme of observing unemployed individuals is chosen, the study is always terminated before all individuals have found their job and so the corresponding observations are censored. Aside from comparing different groups of the unemployed we were interested in whether there were any

* The paper was written with the financial support of Internal Grant Agency of the University of Economics, Prague, grant No IG 410044.

noticeable changes in the distribution over successive years. Some papers dealing with both parametric and semi-parametric models applied to data on TBBE were published abroad. Unfortunately, the sample of unemployed and the observation scheme were not fully described.

The absence of longitudinal data for the unemployed in our country was what made the investigation difficult. The question was whether the LFSS data could be used for that purpose anyway. These data are basically cross-sectional and do not enable observing unemployed individuals for a specific time period. On the top of it, data on TBBE are heavily censored and the type of censoring is rather complex. These features create an interesting problem for analysis. Solution of the problem required primarily the suitable way of selection of the unemployed from the LFSS database so that the absence of longitudinal data might be compensated. Considering the aim of the study and the character of the LFSS data a parametric model, namely the Accelerated Failure Time model was used. The median as the characteristic of TBBE distribution in different groups of the unemployed was computed by means of the model and its dependence on time throughout the years 2000-2004 was examined. S-Plus was used to estimate the parameters of the model.

2. Data set and variables

The data come from the LFSS that takes place quarterly. Households from the whole Czech Republic are randomly selected and form a rotating panel, i.e. one fifth of the households is substituted by new ones before the next survey. The number of visits of the household is limited up to five. Demographic and social data and additional answers concerning the economical activity for each member of the household over fifteen years of age are recorded during the interview. The unemployed members state the time elapsed from the beginning of seeking a job by the time of the visit. This time is recorded only in a categorized form. Since 2002 the last day of employment for persons with work experience is recorded yielding a little bit better specification of TBBE.

The unemployed having been found out during the first visit of the household, whereas surveys in the period from 2000 to 2004 were considered, formed the original sample. The results of their second interview were added from the LFSS database for the labour status of the individuals after a lapse of further 3 months to be determined. It follows that at the first visit all data on TBBE for unemployed are right-unbounded whereas after the second visit some of them are known to lie in a bounded interval. It was expected that the absence of longitudinal data might be settled in this way. Long-term unemployed who had been seeking job for more than 2 years were dropped from the data set, so were those aged more than 65 years, leaving 6141 individuals in the sample, 2790 males and 3351 females.

TBBE was the dependent variable. Lower and upper limits of intervals 0-1, 1-3, 3-6, 6-12, 12-18, 18-24, possibly modified using information about the last day of employment, entered the model. Their values were taken either from the second interview, if given, or derived from the first interview, with respect to the 3-month difference.

The following explanatory variables were used during our investigation: SEX, EDU (highest educational attainment, (4 levels), AGE (age in 5-year brackets, 10 levels), AREA (statistical territorial units, 8 levels), INV (reduced capacity to work, 3 levels), MAR (marital status, 4 levels), CHLD (children, 2 levels), HIS (activity status before seeking job, 6 levels), Q (index of survey, 1 to 19), and SES (seasonal component, 1 to 4).

3. Model

The unemployed enter the retrospective study at various times and are under observation for different lengths of time. Data are collected up to the date of the second visit. Type of censoring that arises due to the way of recording is rather complex. The individuals enter the study during the chosen time period (24 months), the timing of the beginning of unemployment is interval-censored and we observe only that it is in some of the intervals mentioned above. It means that even if the exact time for the end of unemployment were observed, the exact duration of unemployment would be known only to lie in an interval. Although the theoretical solution of this double censoring exists and consists in the proper formulation of the likelihood function, available statistical software does not handle corresponding procedures. To make use of existing programs, the data are considered as interval--censored.

The Accelerated Failure Time model takes the distribution of $Y = \ln T$, where T is time of unemployment, given a vector of covariates $\mathbf{x}$, to be of the form

$$S(y \mid \mathbf{x}) = S_0\left(\frac{y - \boldsymbol{\beta}^T \mathbf{x}}{b}\right), \quad -\infty < y < \infty,$$

where $\boldsymbol{\beta}^T \mathbf{x}$ is a location parameter, $b > 0$ is a scale parameter, $S_0()$ is a fully specified survival function defined on $(-\infty;\infty)$ and independent of $\mathbf{x}$.

Another way to express the model is

$$Y = \boldsymbol{\beta}^T \mathbf{x} + bZ,$$

where Z is a random variable with survival function $S_0(z)$.

The survival function for T given $\mathbf{x}$ is correspondingly of the form

$$S(t \mid \mathbf{x}) = S_0^*\left[\left(\frac{t}{\exp(\boldsymbol{\beta}^T\mathbf{x})}\right)^{1/b}\right],\ t>0,\ S_0^*(t) = S_0(\ln t).$$

The covariates effectively alter the time scale which is the reason for the name of the model. If $\exp(\boldsymbol{\beta}^T\mathbf{x}) > 1$, the effect of the covariate vector is to decelerate time, if $\exp(\boldsymbol{\beta}^T\mathbf{x}) < 1$, the effect is to accelerate it.

Logistic distribution for Z (i.e. log-logistic distribution for T) with the survival function $S_0(z) = \dfrac{1}{1+\exp(z)}$ was chosen based on our previous studies, see [1; 2]. The median of T equals $\exp(\boldsymbol{\beta}^T\mathbf{x})$.

The vector of covariates $\mathbf{x}$ included variables mentioned above. Variable Q denoting the order of the survey together with factor SES was included so that the effect of time throughout the period 2000-2004 could be examined. Since S-Plus supports smoothing splines as a part of survival data model fitting, the spline term for variable Q was included. After the ML estimates of median over the period 2000-2004 based on our model were computed, seasonal components derived from model coefficients were subtracted from the linear predictor $\boldsymbol{\beta}^T\mathbf{x}$ dependent on a given quarter.

4. Results

First the AFT model for males and females together was fitted. All categorical factors except for CHLD and interaction AGE*HIST were clearly significant in the model. The presence of SEX*CHLD, MAR*CHLD, Q*SEX interactions was also checked but proved insignificant. Continuous covariate Q indexing the survey does not affect TBBE significantly but highly significant seasonal effects exist. It is apparent in Fig. 1. The median drops in every 2nd and 4th quarter. This behaviour coincides with a periodic character of the number of the unemployed who seek work for 0-3 months, no matter if they have found their job or not (Fig. 2, males and females together). The larger number of these short-term unemployed causes the decrease of the median. The increase of short-term unemployed entering the study is higher in the 1st and 3rd quarter, affecting results belonging to the following quarters.

Further two separate models for males and females were considered to reveal some differences in the shape of time dependence between the two categories. Smooth curves in Fig. 1 result from fitting a smoothing spline for variable Q and removing seasonal components from predicted values. Dependence of median over time period 2000-2004 does not show any evident trend, yet there are slightly opposite trends for both sex categories at the end of 2004.

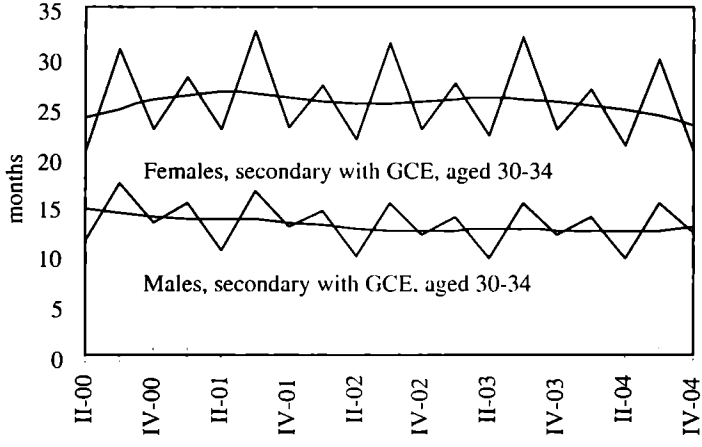

Fig. 1. Changes of median of TBBE over the period 2000-2004

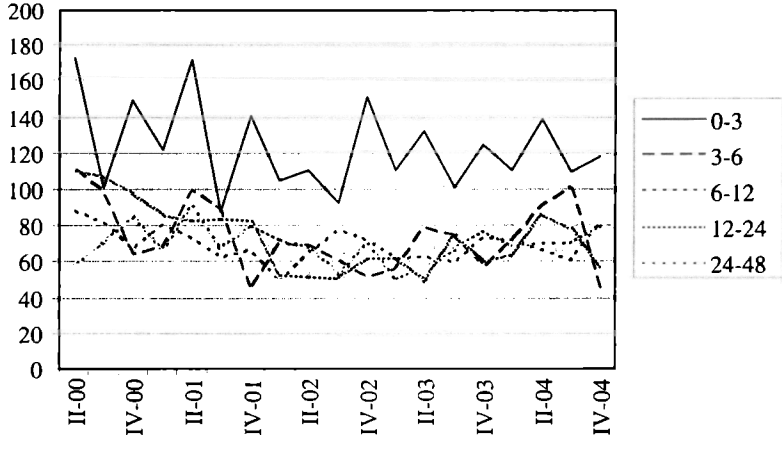

Fig. 2. Changes in number of unemployed by duration of seeking job

Though the curves in Fig. 1 belong to the unemployed with secondary education with GCE, aged 30-34, the trend in median would be the same for other categories according to the model. For example, the median of unemployed with basic education is 2.4 (for males) or 1.6 (for females) times higher every quarter, the median of unemployed with university education is 1.4 (for males) or 1.6 (for females) times lower as follows from the estimated parameters. Similarly the differences between various age groups can be interpreted. Median is lowest for unemployed aged 20-24, 1.01 and 1.7 times the median for 30-34, for males and females respectively. Table with parameter estimates is not shown for shortage of place.

5. Discussion

The main drawback of the given solution is the lack of precision of the data entering the model. Though the item in the LFSS questionnaire relating to the start of seeking a job assumes an answer in the form of the calendar date, only intervals in place of exact data are available in the LFSS database. Aside from less preciseness of model parameter estimates another problem arises, though the consequences may not be substantial when periods up to two years are considered. Namely, variable AGE represents age at the time of the first interview, i.e. after different lengths of seeking period and not the age at the beginning of the follow-up period as it properly should do. The absence of exact data makes deriving of this age impossible. Further on, if the duration of seeking were accurate at least to a quarter of the year, other variable representing the beginning of a period of unemployment could be included in the model instead of variable Q and changes of median over years would be more easily interpretable.

References

[1] Jarosova E., Mala I., Esser M., Popelka J., "Modelling Time of Unemployment via Log-location-scale Model", [in:] Antoch J. (ed.), *COMPSTAT 2004*, [CD-ROM], Physica-Verlag, Heidelberg 2004.

[2] Jarosova E., Mala I., "Modelling Time of Unemployment in the Czech Republic", [in:] M. Kováčová (ed.), *Aplimat 2005. Part II*, Slovak University of Technology, Bratislava 2005.

[3] Lawless J.F., *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, Hoboken, NJ, 2003.

[4] Venables W.N., Ripley B.D., *Modern Applied Statistics with S*, Springer-Verlag, New York 2002.

SZACOWANIE MEDIANY CZASU POZOSTAWANIA BEZ PRACY W REPUBLICE CZESKIEJ

Streszczenie

Parametryczny model AFT (*Accelerated Failure Time*) jest stosowany w celu odniesienia różnych demograficznych i socjoekonomicznych cech bezrobotnych w Republice Czeskiej do czasu między początkiem i końcem okresu pozostawania bez pracy. Zmiany mediany czasu w okresie 2000--2004 są sprawdzane przez sklejane wygładzanie i podejście regresyjne do dekompozycji sezonowej. Badanie jest oparte na danych z Badania Ekonomicznej Aktywności Ludności organizowanego przez Czeski Urząd Statystyczny.

Słowa kluczowe: dane cenzurowane, model przyśpieszonego czasu awarii AFT (*Accelerated Failure Time*), sklejane wygładzanie, podejście regresyjne do sezonowości.