A C T A U N I V E R S I T A T I S L O D Z I E N S I S
FOLIA OECONOMICA 269, 2012

Wioletta Grzenda[*]

BAYESIAN EXPONENTIAL SURVIVAL MODEL IN THE ANALYSIS OF UNEMPLOYMENT DURATION DETERMINANTS

Abstract. The primary objective of the work is to identify demographic and socio-economic factors influencing the unemployment duration in the recent period in Poland. Different approaches to the problem have been applied. In this paper we have used a survival parametric model in Bayesian approach. The following determinants have been concerned in the model: sex, marital status, education level, information about continuing an education, region of Poland, and age of respondent. The empirical analysis is based on "Household budgets in 2008" survey of Central Statistical Office and indicates the main factors influencing unemployment duration.

Key words: unemployment; survival exponential model; Bayesian inference; MCMC method.

I. INTRODUCTION

The significance of unemployment results from its economic, social and political aspects. To investigate the unemployment determinants, event history models (Drobnič and Frątczak, 2001) and logit models (Daras and Jerzak, 2005; Collier, 2003) are usually applied. Another approach, based on standardised unemployment rates can be found in (Socha and Sztanderska, 2000). The most frequently reported factors related to unemployment are: sex, age, education status and place of living. Besides these, many other determinants are considered such as: the techniques of searching for a job, the number of received job offers, the period of unemployment benefit, minimal salary rates, etc.

The primary objective of this work is to identify the demographic and socio-economic features, which influence the unemployment. A Bayesian exponential survival model was applied to analyse the determinants affecting the duration of unemployment period.

[*] Ph.D., Institute of Statistics and Demography, Warsaw School of Economics.

II. THE SCOPE OF RESEARCH

The empirical analysis is based on "Household budgets in 2008" survey of Central Statistical Office. According to the aim of this research, we take into consideration unemployed persons, who were looking for a job and were ready to take a job (Eurostat). As different factors can influence unemployment depending on its duration, a decision has been made to consider only persons who were unemployed maximally for 24 months. In this way we chose 2512 individuals. 214 of them already found a job and waited for starting work – for these persons an event holds, while the others are censored.

In a model, a dependent variable is time defined as the number of months of unemployment. The characteristics of independent variables that potentially may have an impact on the unemployment duration has been discussed below.

The first potential determinant is *sex*: 1 – man (49.56%), 2 – woman (50.44%). One can expect that higher chances for finding a job have men, then women, who more time devote to their families.

Marital status is one of the factors considered by employers when hiring new employees. Hence, it is important to examine if unmarried people have more chance of finding a job. *Marital status* was encoded as follows: 1 – unmarried (49.32%), 2 – married (42.40%), 3 – separated (1.04%), 4 – a widower, a widow (2.07%), 5 – divorced (5.18%).

We can suppose that education status is one of the most important determinants influencing the chance of finding a job. *Education level* was encoded as follows: 1 – higher (10.19%), 2 – post-secondary (2.95%), 3 – secondary professional (21.38%), 4 – secondary general (12.18%), 5 – basic vocational (34.63%), 6 – primary school (18.67%). A related factor potentially influencing unemployment periods is whether a respondent continues education. The latter variable takes two values: 1 – yes (8.16%), 2 – no (91.84%).

The regions of Poland differ in the economic and technological development, hence we can suppose that the residents of the west and central region of Poland have more chance of finding a job, then the residents of the remaining regions. *Region of Poland* was defined as follows: 1 – central (province: łódzkie, mazowieckie) (18.95%), 2 – southwest (province: dolnośląskie, opolskie) (10.91%), 3 – south (province: małopolskie, śląskie) (14.81%), 4 – northwest (province: wielkopolskie, zachodniopomorskie, lubuskie) (17.12%), 5 – north (province: kujawsko-pomorskie, warmińsko-mazurskie, pomorskie) (17%), 6 – east (province: lubelskie, podkarpackie, świętokrzyskie, podlaskie) (21.22%).

Next determinant which has been taken into consideration is *age* (min=17, max=66). It is important to examine if young persons have more chance of finding a job.

III. RESEARCH METHOD

In this paper we have used a Bayesian survival exponential model. The Bayesian methods combine subjective prior knowledge with the information acquired from the data by using Bayes' theorem (Bolstad, 2007; Bernardo and Smith, 2004; Gelman et al., 2000).

The proposed exponential model is one of the most important models in survival analysis (Blossfeld et al., 1989; Blossfeld and Rohwer, 1995). This survival parametric model will be presented in Bayesian approach. The Bayesian analysis of survival parametric models has been discussed in many works (Ibrahim et al., 2001).

Suppose we have independent identically distributed survival times $\mathbf{y} = (y_1,\ldots,y_n)'$; with each y_i, $i = 1,...,n$ having an identical exponential distribution with parameter λ. The censoring indicators we denote by $\mathbf{v} = (v_1,\ldots,v_n)'$, where $v_i = 0$ if y_i is right censoring and $v_i = 1$ if y_i is failure time, $i = 1,...,n$. The density function for y_i is $f(y_i \mid \lambda) = \lambda \exp(-\lambda y_i)$, the survival function $S(y_i \mid \lambda) = \exp(-\lambda y_i)$. In regression models we have one more additional element – a matrix of independent variables $\mathbf{X}$ $(n \times p)$. Let $\mathbf{x}_i'$ denote *i*th row of the matrix, then $D = (n, \mathbf{y}, \mathbf{X}, \mathbf{v})$ is the observed data.

Let $\lambda_i = \varphi(\mathbf{x}_i'\boldsymbol{\beta})$, where $\mathbf{x}_i$, $(p \times 1)$ is a vector of covariates, $\boldsymbol{\beta}$, $(p \times 1)$ is a vector of regression coefficients and φ is a known function. For $\varphi(\mathbf{x}_i'\boldsymbol{\beta}) = \exp(\mathbf{x}_i'\boldsymbol{\beta})$, we have the following likelihood function:

$$
\begin{aligned}
L(\boldsymbol{\beta} \mid D) &= \prod_{i=1}^{n} f(y_i \mid \lambda_i)^{v_i} S(y_i \mid \lambda_i)^{(1-v_i)} = \\
&= \prod_{i=1}^{n} \left[\exp(\mathbf{x}_i'\boldsymbol{\beta}) \exp(-y_i \exp(\mathbf{x}_i'\boldsymbol{\beta}))\right]^{v_i} \left[\exp(-y_i \exp(\mathbf{x}_i'\boldsymbol{\beta}))\right]^{(1-v_i)} = \\
&= \exp\left\{\sum_{i=1}^{n} v_i \mathbf{x}_i'\boldsymbol{\beta}\right\} \exp\left\{-\sum_{i=1}^{n} y_i \exp(\mathbf{x}_i'\boldsymbol{\beta})\right\}.
\end{aligned} \tag{1}
$$

Often for regression coefficients $\boldsymbol{\beta}$ we choose uniform improper prior or a normal prior. In our model we take a *p*–dimensional normal prior $N_p(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ for $\boldsymbol{\beta}$, where $\boldsymbol{\mu}_0$ denotes the prior mean vector, and $\boldsymbol{\Sigma}_0$ denotes the prior covariance matrix. Then the posteriori distribution for $\boldsymbol{\beta}$ is given by

$$
p(\boldsymbol{\beta} \mid D) \propto L(\boldsymbol{\beta} \mid D) p(\boldsymbol{\beta} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \tag{2}
$$

$p(\boldsymbol{\beta} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ denotes multivariate normal density with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$.

IV. MODEL ESTIMATION

Estimation and verification of all the models has been performed using SAS system. In order to obtain objectively correct results, we have used a priori distributions that have a minimal impact on a posteriori distribution. Moreover, we have neither results of statistical modeling for the investigated time period nor the data for the entire country. Still, only credible information may improve the quality of model estimation. Therefore, non-informative independent normal prior distributions have been used for all regression parameters to estimate all the models: $p(\boldsymbol{\beta}) \sim N(\mathbf{0}, 10^6 \mathbf{I})$.

The estimated models have been evaluated to assess the convergence of generated Markov chains. Inference in Bayesian analysis under unchecked convergence for some model parameters may result in wrong conclusions. Using Geweke's test (Geweke, 1992) we have found that there is no indication that the Markov chain has not converged for all the parameters of investigated models, at any significance level.

The same result has been obtained for Heidelberger-Welch test (Heidelberger and Welch, 1983), which consists of two parts i.e. a stationarity test and a halfwidth test. The halfwidth test additionally reports whether the sample size is adequate to meet the required accuracy for the mean estimate.

Thus, it can be assumed that the obtained posterior samples are appropriate for statistical inference. The results of model estimation have been summarized in table 1.

Table 1. Posterior sample mean and interval statistics

Parameter	Mean	Highest Probability Density Interwal ($\alpha = 0.05$)		Exp(Mean)	Exp(-Mean)
Intercept	4.8315	4.1880	5.4644	125.399	0.008
Sex 1	–0.3547	–0.6374	–0.0850	0.701	1.426
Education 1	–0.9193	–1.4357	–0.3769	0.399	2.508
Education 2	–0.2906	–1.2897	0.6852	0.748	1.337
Education 3	–0.6864	–1.1603	–0.2497	0.503	1.987
Education 4	–0.5770	–1.1029	–0.0260	0.562	1.781
Education 5	–0.3488	–0.8063	0.0671	0.706	1.417
Region 1	–0.4681	–0.8869	–0.0566	0.626	1.597
Region 2	0.1815	–0.4396	0.7919	1.199	0.834
Region 3	–0.3878	–0.8294	0.0572	0.679	1.474
Region 4	–0.3229	–0.7676	0.1152	0.724	1.381
Region 5	–0.2975	–0.7364	0.1300	0.743	1.346
Age	0.0150	0.00262	0.0263	1.015	0.985

Basing on the highest probability density interval (Bolstad, 2007), statistically significant variables are sex, age and at least one level of other variables.

V. SUMMARY AND CONCLUSIONS

We obtained that among variables chosen to model: sex, marital status, education level, information about continuing an education, region of Poland and the age at the moment of research, only two variables have been determined to be statistically insignificant: marital status and information about continuing an education.

In the case of first determinant, previous assumptions that unmarried people have more chance of finding a job, were not confirmed. Information about continuing an education has turned out to be statistically insignificant, but we can state that by improving the education status, one can increase the chances for finding a job in the future. We obtained that the individuals, who had education higher than primary, have more chance of finding a job. The persons having a secondary professional education have 98.7% more chance of finding a job comparing to those who have attended primary schools only, the persons having a higher education have this chance more than twice as high as the members of the former group. Our research confirms the previous assumption and is consistent with the results of other studies (Daras and Jerzak, 2005).

The results for sex variable also confirm our previous speculations, we obtained that men have 42.6% more chance of finding jobs than women. The results of other research (Daras and Jerzak 2005; Socha and Sztanderska, 2000) indicate worse situation of women in the labor market, even if women are better educated and are more actively searching for a job. Moreover employers are more likely to hire men, than women, due to the role women play in their families i.e. they more frequently take care of children.

One of important factors influencing the unemployment duration is age; we obtained that the chances of finding a job decrease by about 1.5% as the age of a respondent increases by one year. According to other researchers (Daras and Jerzak, 2005) the unemployment rate is also dependant of the age, the lowest chances for finding a job have persons aged over 44.

The results for region of Poland variable are that only one level of this variable is statistically significant being central region i.e. provinces łódzkie i mazowieckie. We obtained that the mean unemployment duration for the residents of this region is shorter by 37.4% then for the residents of the east region.

The model applied in this article enables the identification of demographic and socio-economic factors influencing the unemployment duration. The advantage of survival models is the fact they include all the history of an individual. But these models demand data, which are frequently not provided by commonly made surveys.

REFERENCES

Bernardo J.M., Smith A.F.M. (2004), *Bayesian Theory*, John Wiley & Sons, New York.
Blossfeld H.P., Hamerle A., Mayer K. (1989), *Event history analysis, Statistical theory and application in the social sciences*, Hillsdale, NJ: L. Erlbaum.
Blossfeld H.P., Rohwer G. (1995), *Techniques of event history modeling, New approaches to causal analysis*, Hillsdale, NJ: L. Erlbaum.
Bolstad W.M. (2007), *Introduction to Bayesian statistics*, John Wiley & Sons, New York.
Collier W. (2003), The impact of demographic and individual heterogeneity on unemployment Duration: A regional study, *Studies in Economics*, 0302.
Daras T., Jerzak M. (2005), Wpływ cech społeczno-demograficznych osób bezrobotnych na możliwość znalezienia pracy, *Materiały i Studia*, Zeszyt nr 189.
Drobnič S., E. Frątczak (2001), Employment patterns of married women in Poland, *Careers of couples in contemporary society*, New York.
Gelman A., Carlin J.B., Stern H.S., Rubin D.B. (2000), *Bayesian data analysis*, Chapman & Hall/CRC, London.
Geweke J. (1992), Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In Bernardo J., Berger J., Dawiv A., Smith A., *Bayesian Statistics*, 4, 169-193.
Heidelberger P., Welch, P. (1983), Simulation run length control in the presence of an initial transient, *Operation Research*, 31, 1109 –1144.
Ibrahim J.G., Chen M-H, Sinha D. (2001), *Bayesian survival analysis*, Springer-Verlag, New York.
Socha M., Sztanderska U. (2000), *Strukturalne podstawy bezrobocia w Polsce*, PWN, Warszawa.

Wioletta Grzenda

BAYESOWSKI WYKŁADNICZY MODEL PRZEŻYCIA W ANALIZIE DETERMINANT DŁUGOŚCI CZASU POZOSTAWANIA BEZ PRRACY

Celem niniejszego opracowania jest identyfikacja czynników demograficznych oraz społeczno-ekonomicznych wpływających na długość czasu pozostawania bez pracy. Zbiór danych wykorzystany w badaniu pochodzi z badań Głównego Urzędu Statystycznego „Budżety Gospodarstw Domowych 2008".

Do analizy determinant długości czasu pozostawania bezrobotnym wykorzystano bayesowski wykładniczy model przeżycia. W estymacji modelu wykorzystano metody Monte Carlo oparte na łańcuchach Markowa, a w szczególności próbnik Gibbsa.

W wyniku przeprowadzonej analizy otrzymano, że wśród wybranych do modelowania zmiennych objaśniających: płeć, stan cywilny, poziom wykształcenia, informacja o tym, czy respondent nadal się dokształca, region Polski, który zamieszkuje respondent oraz wiek w momencie badania, tylko dwie okazały się statystycznie nieistotne: stan cywilny oraz informacja o tym, czy respondent nadal się dokształca.