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 255, 2011

Beata Jackowska[*], *Ewa Wycinka*[**]

ANALYSIS OF THE LAST EMPLOYMENT PERIOD OF THE UNEMPLOYED: THE APPLICATION OF THE COX MODEL

Abstract. This paper presents the implementation of survival analysis (event history analysis) methods in the examination of the employment period at the last place of work of people who have been unemployed for a long time. At first, the Kaplan-Meier method is used for the identification and categorization of the determinants of the last job period. The next step is constructing the Cox proportional hazards model. The model fit is made with the use of the likelihood ratio test and information criteria. We pay special attention to the properties of the Cox model in order to improve its application in socio-economic surveys.

Key words: survival analysis, event history analysis, survival time, time of duration, risk, Kaplan-Meier estimator, Cox model, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)

I. INTRODUCTION

Survival analysis involves modelling a process of developing a certain unit between two specified time points: the beginning and the end of the process. This time is called the survival time or the time of duration. One of the most important groups of this statistical method is regression analysis used in the examination of the direction and power of the influence of independent variables on the time of duration. The Cox proportional hazards model is commonly used in the survival analysis regression model (Cox 1972). This model is widely employed in demography, social sciences and medicine (see e.g. Collett 1994, Klein and Moeschberger 1997).

In this paper we present an attempt of the implementation of the Cox model in the examination of the employment period at the last place of work of people who have been unemployed for a long time. Special attention is paid to the problem of variable selection because the lack of rules in this process could lead to random results. Variable selection depends not only on variable implementation and the removal algorithm but also on the valuation of the property criterion.

[*] Ph. D., Chair of Statistics, University of Gdańsk.
[**] Ph. D., Chair of Statistics, University of Gdańsk.

As regards duration, unemployment is usually classified as follows: short-term unemployment (up to 3 months), temporary unemployment (from 3 to 6 months), long-term unemployment (from 6 to 12 months) and serious long-term unemployment (exceeding 12 months) (GUS 2004). Serious long-term unemployment (exceeding 12 months) is an grave socio-economic problem as it tends to reduce the opportunity to find a new job. Therefore, it seems to be a challenge to conduct such a policy which would increase the number of long-term unemployed people who find new jobs. Among the long-term unemployed, there are people with long job seniority: their experience and qualifications are wasted economic potential.

The constructed model of employment at the last place of work of people who are long-term unemployed could be applied in the examination of layoff, the stability (fluency) of unemployment and the study of the outflow of long-term unemployed workers from the job market. It might also help to take preventive action at an adequate moment, aimed at appropriate target groups.

II. THE DESCRIPTION OF THE STUDY AND DATA CHARACTERISTICS

To carry out the research, we used some results of the sampling survey concerning unemployed people registered in District Labour Office in Gdańsk in December 2003 (Banaszkiewicz 2004).

At first, the Kaplan-Meier method was used for the evaluation of the survivor function of the length of the last employment period. The value of the estimator assesses the probability that at a particular moment of time an employee does not lose his/her job. Subsequently, we compared the survivor curves for groups identified on the basis of different values of the variables. On the basis of the Gehan test for the equity of survivor curves, the levels of significant variables determining the total length of the employment time and the length of the last time of employment were obtained. For the length of the last time of employment, the significant subsequent variables included gender, age, education, the number of previous registrations at the Labour Office, the kind of job and the reason for losing the job.

It follows from the Kaplan-Meier analysis that people who work longer are women, people who had not been unemployed before and workers who lost their job because of reduction. Part-time workers and support workers have the shortest periods of employment. Employees with tertiary education have the shortest time of employment in one company, while employees with secondary comprehensive education – the longest one. For men, the risk of dismissal is the

highest during the first months of employment; later it constantly decreases. The risk for women does not change during one employment period.

Table 1. The structure of the population of the long-term unemployed registered in the District Labour Office, Gdańsk: Socio-demographic characteristics (sampling survey 01.12.2003)

Age	Women	Men
	%	
18 – 24	15	17
25 – 34	24	21
35 – 44	26	16
45 – 54	29	34
55 – 59	6	11
60 and more	0	1
Education	%	
Elementary	28	35
Basic vocational	31	44
General secondary	16	5
Technical secondary and post-secondary	17	10
Tertiary	9	7
Sequence of registration	%	
1	48	44
2	32	28
3 or more	19	28

	Women	Men
Direct employment before registration at District Labour Office	%	
yes	64	63
no	36	37
Job offers	%	
1	62	31
2	16	22
3 and more	22	48
Type of last job	%	
Regular job at public company	38	34
Regular job at private company	51	51
Own business activity	5	5
Part-time job	6	8
Support work	0	2
Reason for dismissal	%	
Company liquidation	23	26
Reduction	33	29
Quitting	18	18
Other	26	27

Source: Own elaboration.

Next, the statistically significant variables were used to construct the Cox proportional hazards model.

III. THE COX PROPORTIONAL HAZARDS MODEL

The distribution of the survival time T can be identified by the hazard function (rate) which is also called force of mortality and is given as follows:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}. \tag{1}$$

The value of this function gives instantaneous potential for an event to occur at a particular moment. In the event history analysis, the influence of particular factors is frequently evaluated by the semiparametric proportional hazards model (PH). The Cox PH model is usually given in terms of the hazards model formula:

$$\lambda\left(t \mid x_1, x_2, \ldots, x_k\right) = \lambda_0(t) \cdot \exp\left(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k\right). \quad (2)$$

The explanatory variables $x_1, x_2, \ldots, x_k$ may be quantitative or categorized. The function $\lambda_0(t)$ is called the baseline hazard. If all the X's are equal to zero, the Cox model reduces to the baseline hazard function.[1] Another important property of the Cox model is that the baseline hazard could be unspecified. It is a property that makes the Cox model a semiparametric model. In such a case, we are not able to estimate the hazard for particular individuals, but we can calculate the hazard ratio, that is the hazard for one individual compared with the hazard for another individual. The estimators $\hat{\beta}_i$ of the parameters of the Cox model are obtained by the partial maximum likelihood method (see e.g. Cox and Oakes 1984 or Lawless 2003).

The hazard ratio for individuals *i* and *j* is calculated as follows:

$$\frac{\lambda\left(t \mid x_{1i}, x_{2i}, \ldots, x_{ki}\right)}{\lambda\left(t \mid x_{1j}, x_{2j}, \ldots, x_{kj}\right)} = \exp\left(\beta_1(x_{1i} - x_{1j}) + \beta_2(x_{2i} - x_{2j}) + \ldots + \beta_k(x_{ki} - x_{kj})\right). \quad (3)$$

Particularly, when the value of the explanatory variable x_i increases by a unit (with the fixed values of the other variables), the hazard ratio (and risk) changes $\exp(\beta_i)$-times: if $\exp(\beta_i) > 1$, the risk increases; if $\exp(\beta_i) < 1$, the risk decreases.

The Cox proportional hazards model requires the acceptance of two assumptions (Blossfeld, Golsch and Rohwer 2007):

1) At a particular moment of time, the hazard for one individual is proportional to the hazard of any other individual with the same set of explanatory variables;

2) The hazard ratio for any two individuals is independent of time (proportional hazard assumption).

There are three general approaches to assessing the proportional hazard assumption:

- Graphical techniques (comparing two groups $-\ln\left(-\ln\left(\hat{S}(t)\right)\right)$, where $\hat{S}(t)$ is the estimated survivor curve),

[1] For the implementation of the Cox model, qualitative explanatory variables can be changed into a set of indicator variables. One attribute of each variable is omitted. This attribute is called the reference group.

- Goodness-of-fit tests (Schoenfeld residuals tests),
- Assessing the proportional hazard assumption with the use of time-dependent covariates.

The assessment of the model, that is its parameters in a global sense, is based on the likelihood ratio test (which is a global test; see Klein and Moeschberger 1997). The alternative hypothesis assumes that at least one explanatory variable is significant. Assessing a single variable is based on the Wald test (which is a local test; see Klein and Moeschberger 1997).[2] These techniques are used to build the most appropriate model of survival.

Selecting the most appropriate model may be difficult. Information criteria provide an approach for comparing the fit of the models with different numbers of explanatory variables (Collett 1994). The best model has the lowest value of the information criterion. The most frequent criteria are the following:

- Akaike Information Criterion:

$$AIC = -2\ln \hat{L}_k + 2k\,, \tag{4}$$

- Bayesian Information Criterion (Schwartz Criterion):

$$BIC = -2\ln \hat{L}_k + k\ln(n)\,, \tag{5}$$

where k – number of parameters in model, n – number of observations, $\hat{L}_k$ – likelihood function for model with k-variables

IV. THE RESULTS OF THE ESTIMATION OF THE COX PROPORTIONAL MODEL

To construct the Cox proportional model, we used the determinants of the last job length obtained in the Kaplan-Meier analysis. All the variables employed in this model are dummy variables: all the qualitative variables with more than two categories were changed into a set of dummy (indicator) variables. The number of indicator variables is smaller than the number of categories of the initial variable by 1. Otherwise, the variables would be dependent. The omitted factors create a reference group for which the hazard function is evaluated when all the explanatory variables are equal to zero. As a result, 11 indicator variables were obtained (Table 2).[3]

[2] Testing a set of simple hypotheses is not an equivalent to the pooled hypothesis inference (Lovell bias problem).

[3] The method of coding the explanatory variable and the choice of reference group influence the form of the model. In this study we selected a reference group for which the Kaplan Meier plot was the most distant from any other curve.

Table 2. Results of the estimation of the Cox model for $k = 11$ indicator variables

Variables	Atribute	Parameter Estimates b_i	Standard error	Wald Test	*p-value*	Odds ratio e^b	95% confidence interval for odds ratio
Gender	Reference group: "woman"						
	Man	0.2621	0.1284	4.1670	0.0412	1.2997	1.0105-1.6717
Age	Reference group: „45 and more"						
	18-24	0.7771	0.2259	11.8331	0.0006	2.1751	1.3970-3.3866
	25-34	0.4738	0.1722	7.5745	0.0059	1.6061	1.1461-2.2507
	35-44	0.3007	0.1567	3.6835	0.0550	1.3508	0.9936-1.8363
Education	Reference group: "other than tertiary"						
	Tertiary	0.5286	0.2284	5.3543	0.0207	1.6965	1.0842-2.6546
Sequence of registration	Reference group: "first time"						
	Second	0.4344	0.1490	8.5052	0.0035	1.5441	1.1531-2.0676
	Third or more	0.8995	0.1667	29.1219	0.0000	2.4584	1.7733-3.4084
Type of last work	Reference group: "regular job in public company"						
	Regular work at private company or own business activity	0.8261	0.1513	29.8167	0.0000	2.2843	1.6982-3.0728
	Part-time job and support job	1.9800	0.2896	46.7367	0.0000	7.2429	4.1056-12.7776
Reason for dismissal	Reference group: "quitting" or "other"						
	Company liquidation	-0.2975	0.1511	3.8775	0.0489	0.7427	0.5523-0.9986
	Reduction	-0.4545	0.1444	9.8985	0.0016	0.6348	0.4783-0.8425

Source: Calculated with Statistica.

We implemented the information criteria to find the smallest significant model which would be the best one. The values of *AIC* and *BIC* calculated for the models created with the descending method (from the full model to the basic model) are presented in Figure 1. The *AIC* favours the model with 11 variables (Table 2), while the *BIC* points to the model with 6 variables (Table 3).[4]

[4] The *AIC* criterion tends to indicate models with a larger set of explanatory variables than the *BIC* criterion.

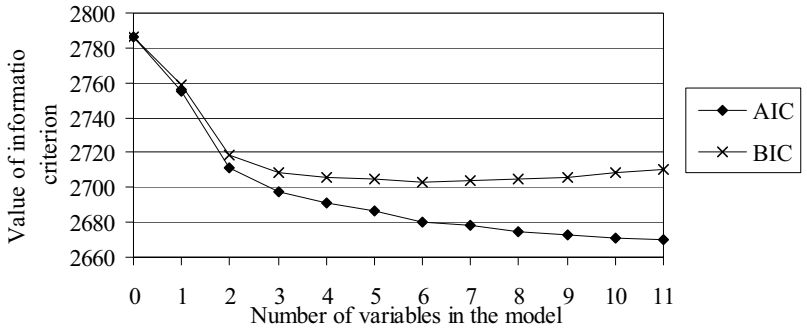

Figure 1. *AIC* and *BIC* information criteria for the best model with *k* variables, calculated for models constructed with the use of the descending method

Source: Own elaboration.

The model with 6 explanatory variables excludes the variable "reason for dismissal". Moreover, the division into age groups "35-44" and "45 and more" was not significant, which changed the reference group for this variable.

Table 3. Results of the estimation of the Cox model for $k = 6$ indicator variables

Variables	Attribute	Parameter estimates b_i	Standard error	Wald Test	*p-value*	Odds ratio e^b	95% confidence interval for odds ratio
Age	Reference group: "35 and more"						
	18-24	0.6620	0.2167	9.3278	0.0023	1.9386	1.2676-2.9647
	25-34	0.4902	0.1532	10.2403	0.0014	1.6326	1.2092-2.2043
Sequence of registration	Reference group: „first"						
	Second	0.4328	0.1428	9.1779	0.0025	1.5415	1.1651-2.0395
	Third and more	0.8757	0.1610	29.5820	0.0000	2.4006	1.7509-3.2913
Type of last job	Reference group: "Regular job at public company"						
	Regular job at private company or own business activity	0.8406	0.1498	31.4966	0.0000	2.3179	1.7282-3.1088
	Part time job or support work	1.9730	0.2860	47.5964	0.0000	7.1919	4.1060-12.5971

Source: Calculated with Statistica.

V. CONCLUSIONS

The constructed proportional hazards model allows us to estimate the distribution of the last employment time which depends on the socio-demographic variables of an employee. The significant variables include age, previous unemployment and the kind of job. Moreover, the expanded model takes into consideration gender and the reason for dismissal. The use of the Cox model enables us to identify not only the groups of high risk of being dismissed, but also the time points when the risk of layoff is the highest. The applied model may be used for identifying proper target groups which should benefit from a variety of programs designed to combat unemployment. Such programs should be aimed not only at the unemployed, but also at those who face the high risk of layoff.

The construction of the Cox model requires special attention to be paid to its adequacy in a particular situation, which has been stressed in this paper. Emphasis has been put on the consequences of explanatory variable coding and related different estimates of their influence on dependent variables. The most effective coding method is changing the qualitative variables into a set of dummy variables with the reference group selected on the basis of the observation of the Kaplan-Meier plot. The attribute of the variable for which the survivor curve is at the largest distance to any other should be treated as a reference group. If the survivor curve for the reference group is sufficiently distant from the other attributes, the Cox model parameters are significant. Constructing the Cox model is based on the compromise between the wish to have the best adjusted model and the wish to have the simplest model. This is the reason why the information criteria should be employed to select the final set of explanatory variables.

REFERENCES

Banaszkiewicz D. (2004), *Statystyczna analiza sytuacji społeczno-ekonomicznej osób długotrwale bezrobotnych w powiecie gdańskim w roku 2003*, rozprawa doktorska, Sopot.

Blossfeld H. P., Golsch K., Rohwer G. (2007), *Event History Analysis with Stata*, Lawrence Erlbaum Associates, New Jersey.

Collett D. (1994), *Modelling Survival Data in Medical Research*, Chapman & Hall, London.

Cox D. R. (1972), Regression Models and Life Tables, *Journal of the Royal Statistical Society*, B. 34, p. 187220.

Cox D., Oakes D. (1984), *Analysis of Survival Data*, Chapman & Hall, London.

GUS (2004), *Bezrobocie rejestrowane I-IV kwartał 2003 r*., Główny Urząd Statystyczny, Warszawa.

Klein J., Moeschberger M. (1997), *Survival Analysis: Techniques for Censored and Truncated Data*, Springer Science + Business Media, New York.

Lawless J. (2003), *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New Jersey.

Beata Jackowska, *Ewa Wycinka*

WYKORZYSTANIE MODELU COXA DO BADANIA CZASU PRACY U OSTATNIEGO PRACODAWCY OSÓB BEZROBOTNYCH

W artykule wykorzystano metody analizy przeżycia (analizy historii zdarzeń) do badania czasu pracy u ostatniego pracodawcy osób długotrwale bezrobotnych. W pierwszym kroku posłużono się analizą Kaplana-Meiera w celu identyfikacji oraz efektywnej kategoryzacji zmiennych determinujących długość okresu ostatniej pracy. W drugim kroku skonstruowano model proporcjonalnego hazardu Coxa. Przy konstrukcji modelu wykorzystano następujące kryteria: test istotności modelu oparty na ilorazie wiarygodności oraz kryteria informacyjne. Szczególną uwagę poświęcono własnościom modelu Coxa w celu zapewnienia poprawności jego stosowania w zagadnieniach społeczno-ekonomicznych.