Application of the Cox regression model with time dependent parameters to unemployment data

Petr Volf[1]

Abstract. The contribution deals with the application of statistical analysis of the process of events, with the intensity described by a generalized version of Cox regression model. Namely, we study a case where the impact of covariates changes in time. Therefore the model with time dependent parameters should be applied. A method of model components non-parametric estimation is recalled, the flexibility of result is assessed with a goodness-of-fit test based on martingale residuals. The application concerns to the real data representing the job opportunities development and reduction, during a given period. They record the period of changes characterized by the increased employees fluctuation and staff reduction. Hence, the risk of leaving the company is changing. In particular, the risk of older persons increases, while the fluctuation concerns more the people with shorter time spent with the company. Both these covariates are considered and their impact to the risk analyzed.

Keywords: statistical survival analysis, unemployment study, Cox regression model, time-dependent parameters, goodness-of-fit.

JEL classification: C41, J64
AMS classification: 62N02, 62P25

1 Introduction

The paper studies the problem of flexible modelling of process of events in the framework of statistical survival analysis. As a rule, the model for events occurrence is based on the notion of counting process (registering, i.e. counting observed events) and its intensity given by corresponding model of the hazard rate. The running time could be the calendar time as well as the individual time of the object after its entrance to the study. The hazard rate can, moreover, depend on a set of covariates influencing the risk of events. Even these covariates can change their values during the time, the basic assumption states that at time t the values of covariates are known (are observable) up to t. Thus, they can also depend on the history of running counting process of events (by so called "endogenous" covariates). The most popular model of hazard rate in survival analysis is the Cox regression model based on the assumption of multiplicative influence of covariates to the hazard rate. Naturally, the model has a number of variants and generalizations. One direction leads to a more flexible regression part, generalizing its linear form to a nonlinear or to fully non-parametric. Another generalization, which is used here, corresponds to possible change of covariates impact in time and utilizes time-varying regression parameters (c.f. Murphy and Sen, 1991).

Naturally, a set of models for hazard rate is much richer, from parametric ones, for instance based on the Weibull distribution with its scale parameter being a function of covariates, up to fully nonparametric models (cf. the idea of "doubly cumulative hazard rate" in McKeague and Utikal, 1991). Another often used option is the additive Aalen model consisting of a sum of components, each representing the influence of one covariate. The "accelerated failure time" (AFT) model is popular as well, based on the assumption that the influence of covariate can accelerate (or slow down, on the contrary) the flow of individual object's time, i.e. it speeds up or slows down its growing, ageing, degeneration etc. Approaches and methods of event-data modeling are described in a number of papers and monographs, e.g. in Andersen et al (1993), Kalbfleisch and Prentice (2002).

The outline of the paper is the following: The next section introduces the Cox model variant with time-dependent parameters and the method of its analysis. Namely, a method of its non-parametric estimation is presented. Then, in part 3, the data are presented and their analysis provided. The application analyzes the process of job opportunities development and reduction, during a given period, in a company. The example is taken from Kadane and Woodworth (2004, see also "Data case K"), as their data are interesting by recording, after an initial interval of stable growth, the period of changes characterized by the increased employees fluctuation and staff reduction. Two covariates are available, namely the age of employees and the time spent with the company. Both are developing in time, too. It is due the non-stable conditions in the company that a standard Cox model does not describe the process sufficiently. The impact of covariates to the risk of leaving the company is changing during the period of observation. That is why the model with time-dependent regression parameters has to be utilized. In

[1]Department of Stochastic Informatics, ÚTIA AV ČR, Pod vodárenskou věží 4, 182 08 Praha 8, Czech Republic, volf@utia.cas.cz

particular, the risk of older persons increases (which could be taken as a feature of discrimination), while the fluctuation concerns more the people with shorter time with company. Finally, in part 4, a method of a goodness-of-fit test based on martingale residuals is described and the flexibility of model is assessed with its aid.

2 Cox model with time-dependent parameters

Suppose that $i = 1, 2, ..., n$ objects are examined, each during time period $[0, t_i]$, where the final moment t_i is either the time of observed event (count) or the time of censoring. Random censoring from the right side is considered, as a rule. Let $N_i(t)$ denote the counting process of $i-th$ object, i.e. then $N_i(0) = 0$ and jumps to 1 at t_i, provided the event is observed. This is denoted by the indicator $\delta_i = 1$, while in the case of censoring by $\delta_i = 0$. Further, let $Y_i(t)$ on $t \geq 0$ denote the process indicating whether the object i is at t observed (then $Y_i(t) = 1$) or not ($Y_i(t) = 0$). Standardly $Y_i(t) = 1$ on $[0, t_i]$. Finally, denote by $N(t) = \sum_{i=1}^{n} N_i(t)$ the process counting events observed on all subjects. One of basic assumption is that the censoring mechanism does not depend on times of events and does not contain any information on the count times distribution (the censoring is independent and uninformative).

The behavior of $N_i(t)$ is governed by its intensity process, $\lambda_i(t) = h_i(t) \cdot Y_i(t)$, where $h_i(t)$ is the hazard rate of the event occurrence for subject i at time t. The Cox model specifies its form as

$$h_i(t) = h_0(t) \cdot \exp(\boldsymbol{\beta}' \boldsymbol{x}_i(t)), \tag{1}$$

where $\boldsymbol{x}_i(t) = (x_{i,1}(t), ..., x_{i,K}(t)'$, are values of K covariates (they may depend on time), $\boldsymbol{\beta}$ corresponding K-variate vector of regression parameters, and $h_0(t)$ is a common baseline hazard rate.

Standardly, the estimation of regression parameters is performed in the generalized maximal likelihood framework and is also described elsewhere (c.f. again Andersen et all, 1993, Kalbfleisch and Prentice, 2002). The full likelihood function can be written as

$$L = \prod_{i=1}^{n} h_i(t_i)^{\delta_i} \cdot \exp\{-\int_0^{t_i} h_i(t) dt\},$$

where $h_i(t)$ from (1) includes both unknown parameters as well as unspecified baseline hazard function. Estimates of regression parameters are obtained by the maximization of corresponding partial likelihood

$$L_p = \prod_{i=1}^{n} \Big(\frac{\exp(\boldsymbol{\beta}' \boldsymbol{x}_i(t_i))}{\sum_{k=1}^{n} \exp(\boldsymbol{\beta}' \boldsymbol{x}_k(t_i)) Y_k(t_i)} \Big)^{\delta_i}.$$

Then, the Breslow-Crowley estimator of increments of cumulative baseline hazard function $H_0(t) = \int_0^t h_0(s)\, ds$ is

$$\Delta \hat{H}_0(t) = \frac{dN(t)}{\sum_{k=1}^{n} \exp(\boldsymbol{\beta}' \boldsymbol{x}_k(t)) Y_k(t)},$$

which is nonzero just at times of observed events, i.e. at $t = t_i$ with $\delta_i = 1$.

However, let us assume that the data indicate that the impact of covariates is changing during observation period and therefore the time-dependent parameters, i.e. functions $\boldsymbol{\beta} = \boldsymbol{\beta}(t)$, should be considered. It opens a question of their flexible estimation. The problem is solved quite similarly as in the standard regression model case: Either the functions are approximated by certain functional types (polynomial, combination of basic functions, regression splines) or constructed by a smoothing method, similar to moving window or kernel regression approach. The method adopted in Murphy and Sen (1991) is of such a type. Another method is based on the Bayes approach and treats coefficients $\boldsymbol{\beta}(t)$ as a random dynamic sequence with Gauss prior model of its development.

3 Real data example

The data are taken from the Statlib database: http://lib.stat.cmu.edu/datasets/caseK.txt, the "Case K" data. The data contain the records of all persons employed by a firm during the period of observation, from 1.1.1900 to 31.1.1995, namely their dates of birth, dates when persons were hired by the company and when they have left it, either voluntarily or were forced to leave (dismissed). There were together 412 people, from them 96 were fired, 108 left voluntarily, the rest, 208 employees, were still with the company at the end of data collection period. The time considered is the calendar time, in days, from 1 to 1857, the end of study is also the fixed time of censoring, namely $C = 1857$ is the upper bound for each personal record (it is so called type I censoring by fixed value). The number of employees at the study beginning was 163, it means that 249 people have joined the company during the followed period, i.e. their history in company started at certain $t_{0i} > 0$, thus changing the risk set of the study. Hence corresponding processes $Y_i(t) = 0$ on $(0, t_{0i})$.

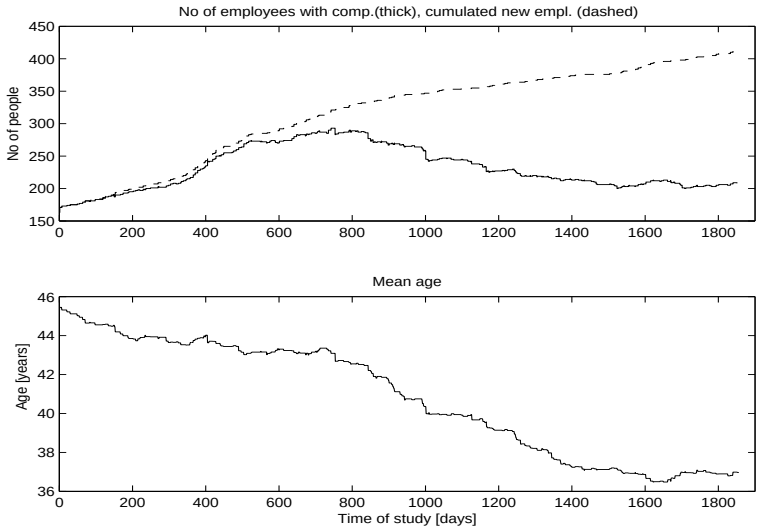

Figure 1 Development of number of employees during the study period (above). Thick curve shows actual number of employees, dashed displays the increase of newly hired people, i.e. the difference means the number of people who already left. The lower plot shows the development of average age of employees.

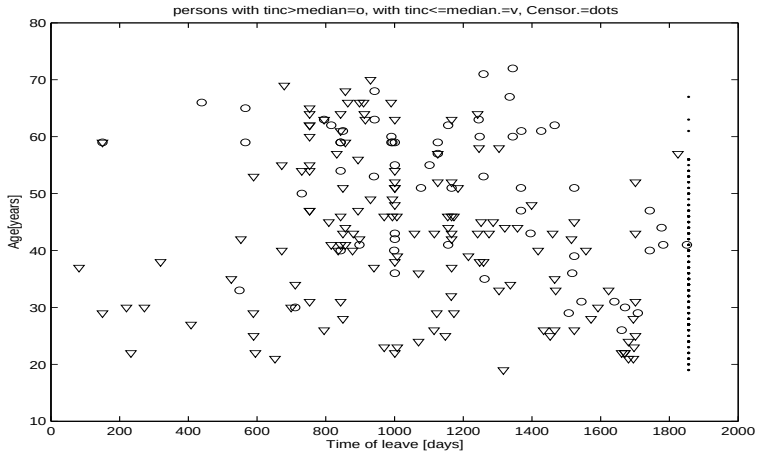

Figure 2 Times when employees have left the company vers. their age; circles for $tinc >$ its median, triangles for $tinc \leq$ median, dots denote censored records for people staying with company at the study termination.

The age of employees and the time spent with company (named here *"tinc"*) were taken as covariates, both were time-dependent. It was expected that both can influence the risk of leaving the company. The age varied from 20 to 70 years, its median was 39, while the time spent with the company varied from only 11 days to more than 41 years (15086 days), with median 1322 days. The changes in the company can be traced already from Figure 1, upper plot. The period of higher intensity of employees fluctuation (which could be taken as an indicator of certain non-stable conditions in the firm) started at about day 800 and lasted almost another 800 days. Figure 2 then shows graphically times of leaves, their dependence on age and also on the time spent with company $(tinc)$.

The analysis of these data appeared also in Kadane and Woodworth (2004). They used the Cox model and Bayes method mentioned above to analyze time changes of the impact of age to employees forced dismissals, the aim of their study was to explore whether older employees were discriminated having higher rate of dismissal. To do it, they explored the probability of regression parameter (of hazard of dismissals on age) being significantly positive, in Bayes sense. However, they did not take into account possible inter-dependence between the risk of forced and voluntary terminations. Such a dependence was studied in Volf (2018) in the framework of competing risks model, the mutual positive dependence was proved, simultaneously decreasing with the age. The interpretation was that the risk of being fired might lead the employees to decision to leave the company voluntarily, in time, and that such a preference concerned more the younger people, thus decreasing the occurrence of forced dismissals for them. However, it was due computational complexity that in Volf (2018) just constant Cox parameters and fully parametric model were considered. The present study does not distinguish between the ways of employment

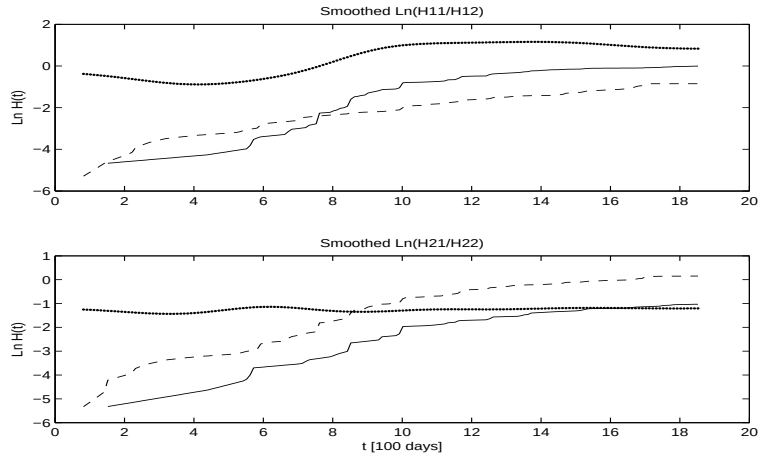

Figure 3 Above: Logarithms of cumulated hazard rates for persons with age$>$ 39 years (full curve) and age$\leq$ 39 (dashed), and their smoothed difference (thick). Below, similarly for $tinc > 1322$ days (full), for $tinc \leq 1322$ (dashed), and their smoothed difference (thick).

termination, on the other hand, the regression parameters are allowed to vary in time, thus making the model rather general.

3.1 Results

Figure 3 presents preliminary graphical test of whether the impact of covariates to the risk of leave the company can be modelled by the proportional hazard model. Namely, we observe difference of logarithms of plain cumulated hazard rates estimated from data with age grater or smaller than 39 years (upper graph) and from data with time with company greater or smaller than 1322 days (both being the medians of covariates). While in the lower plot the curves are approximately parallel, the upper plot shows intersecting curves, thus indicating that the impact of age was changing during the observation period. Hence, we decided to use the Cox model variant with time-dependent "parameters" $\beta_k(t)$, $k = 1, 2$, for both covariates, though Figure 3 does not indicate its necessity in the case of dependence on the time with company. Initial rough estimates of $\beta_k(t)$ were obtained via the moving window method (a variant of the method of sieves described e.g. in Murphy and Sen, 1991). Then, this rough estimates were smoothed with the aid of a kernel smoothing. Both estimates are displayed in Figure 4. It is seen that also $\beta_2(t)$ shows interesting variability and decrease in the second part of observation period. Prevailingly positive β_1 means that the risk of leaving the company increased with larger age, simultaneously $\beta_2 < 0$ indicates higher fluctuation among people with shorter time spent in the company.

Finally, Figure 5 shows the estimate of corresponding cumulated baseline hazard rate. It is seen that at the right end the baseline CHR increases sharply, which is the consequence of the fact that the impact of age decreased in the end part of observation period. It is seen both from Figure 2 and from the shape of estimated $\beta_1(t)$.

4 Goodness-of-fit tests

The selection and evaluation of a reasonable model is just one step of statistical analysis. The goodness-of-fit tests should follow. In the framework of intensity models for lifetime data, the goodness-of-fit tests are often based on the analysis of residual process (martingale residuals). The residual process is defined as the difference between estimated cumulative intensity and observed counting process of failures (see for instance Andersen et al, 1993), formally

$$R(t) = \hat{A}(t) - N(t) = \hat{A}(t) - A(t) - M(t),$$

where $M(t)$ is a martingale, $A(t) = \int_0^t \sum_{i=1}^n h_i(s)Y_i(s)ds$ is the cumulative intensity process (simultaneously it is the variance process of $M(t)$) and $\hat{A}(t) = \int_0^t \sum_{i=1}^n \hat{h_i}(s)Y_i(s)ds$ is its estimate. In the framework of Cox model version (1), $\hat{h_i}(s)ds \sim \Delta\hat{H}_0(s) \cdot \exp(\hat{\boldsymbol{\beta}}'(s)\,\boldsymbol{x}_i(s))$ at counts points $s = t_i$.

Hence, the residual process is constructed from observed data, its properties depend on properties of estimators. In the case without regression, as well as in the Aalen additive regression model, residual processes are the martingales, too (Volf, 1996). In some other cases, as are the Cox model or the accelerated failure time model, the behavior of estimates, and therefore of residuals, is more complicated. Notice also that in the Cox model framework the sum over all subjects yields $\hat{A}(t) = N(t)$ directly, that is why the residuals are as a rule formulated more

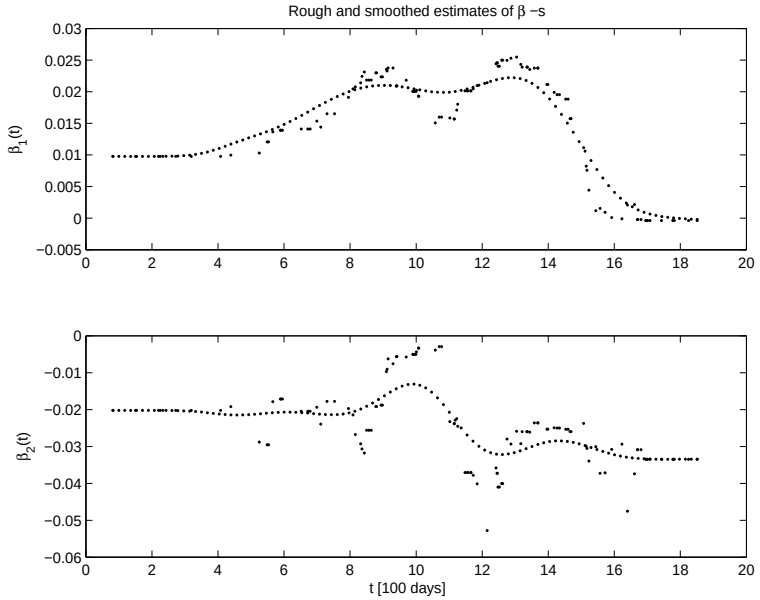

Figure 4 $\beta_{1,2}(t)$ estimated by a mowing window method, at data time points (scattered dots), then their kernel-smoothed versions (smooth curves).

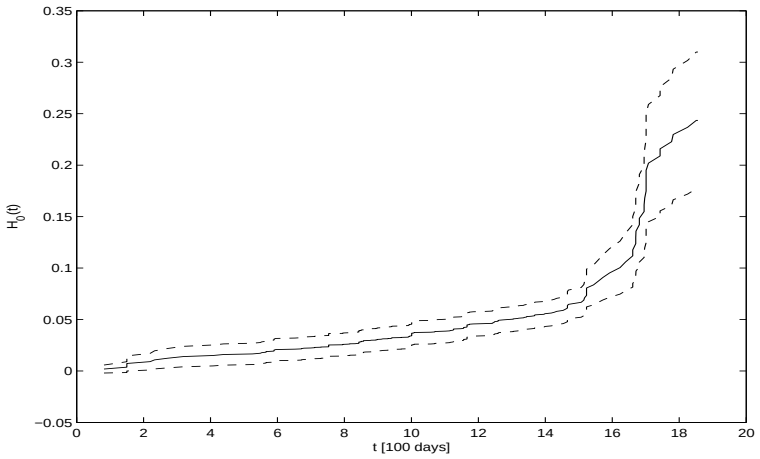

Figure 5 Estimated cumulated baseline hazard rate, with point-wise 95% confidence intervals

generally, namely

$$dR(t) = \sum_{i=1}^{n} W_i(t) \cdot (dN_i(t) - d\hat{A}_i(t)),$$

with some convenient weight processes $W_i(t)$. The simplest choice takes $W_i(t) = 1[i \in S]$, i.e. the indicator of a set of indices (a stratum) $S \subset \{1, 2, ..., n\}$. Then stratified residuals are obtained.

As it has been already said, in the Cox model case the residual processes are not the martingales, though, after a proper standardization, they asymptotically tend to Gauss processes. Therefore, a random generation of would-be residual processes under the hypothesis of model fit is possible. By such a generation we obtain a sample of 'ideal' residual processes. Then, certain characteristics of generated residuals are compared with the same characteristics obtained from the data. Anyway, practical tests of Cox model fit are often performed just graphically, comparing visually how far are residuals in group S from zero line, or, equivalently, $\hat{A}_S(t) = \sum_{i \in S} \hat{A}_i(t)$ from $N_S(t) \sum_{i \in S} N_i(t)$, as in Arjas (1988). We shall use the same method in the present case.

In order to check the fit of proposed model, we stratified the data randomly, many times, selecting a subgroup, strata S, and displaying $R_S(t) = N_S(t) - \hat{A}_S(t)$. It is seen that the residual processes oscillate around zero level, thus assessing a good fit of the model. As the residual processes should, approximately, represent a Gauss processes with zero means and growing variances given by cumulated intensities $A_S(t)$, that is why also the variability of

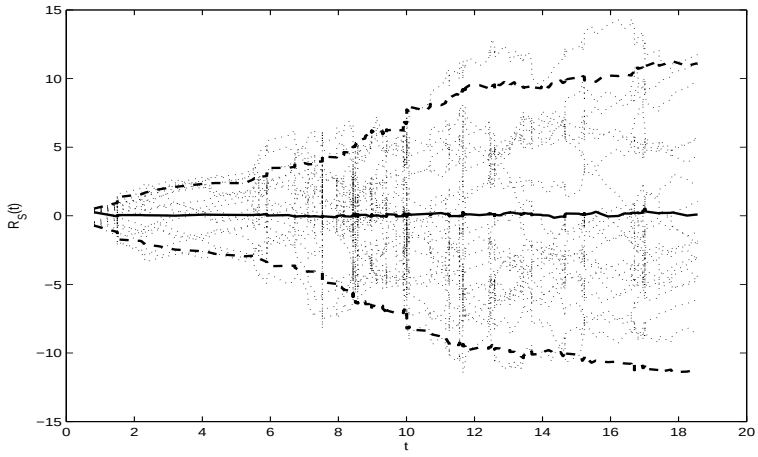

Figure 6 Graphical goodness of fit test: Empirical point-wise median and 5%, 95% quantiles from residual processes computed in 1000 randomly selected data sub-samples. Some residual processes are visualized (dotted).

processes in Figure 6 increases with growing time.

5 Concluding remarks

The contribution had in fact two goals. The first purpose was to present a real case of count data with regression such that the impact of regressors to the risk of count is changing in time. Then also the model had to reflect this phenomenon. Hence, the selection of model and presentation of methodology of its evaluation was the second purpose of the paper. From a set of convenient models the Cox model with time-dependent parameters was chosen, as the Cox model ranks among the most popular ones in the field of statistical survival analysis, also due its reasonable interpretation. The model suitability was confirmed by an appropriate goodness-of-fit test.

Acknowledgement: The research was supported by the grant No. 18-02739S of the Grant Agency of the Czech Republic.

References

[1] Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazard model. *J. Amer. Statist. Assoc.,* 83, 204–212.

[2] Andersen, P.K., Borgan, O., Gill, R.D., Keiding, N. (1993). *Statistical Models Based on Counting Processes.* New York: Springer.

[3] Kadane, J.B., Woodworth, G.G. (2004). Hierarchical Models for Employment Decisions. *Journal of Business and Economic Statistics*, 22, 182–193.

[4] Kalbfleisch, J.D., Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

[5] McKeague, I. W., Utikal, K. J. (1991). Goodness-of-fit tests for additive hazards and proportional hazards models. *Scand.J. Statist.*, 18, 177-195.

[6] Murphy, S.A., Sen, P.K. (1991). Time-dependent coeffisients in a Cox-type regression model. *Stoch. proc. and Applications,* 39, 153–180.

[7] Volf, P. (1996). Analysis of generalized residuals in hazard regression models. *Kybernetika,* 32, 501–510.

[8] Volf, P. (2018). Problem of competing risks with covariates: Application to an unemployment study. In *Proceedings of the 36th International Conference Mathematical Methods in Economics,* FM VŠE Jindřichův Hradec, 624–629.

[9] Data case K: Downloaded 04.03.2019 from *http://lib.stat.cmu.edu/datasets/caseK.txt*

Ekonomická fakulta
Faculty of Economics

Jihočeská univerzita v Českých Budějovicích
University of South Bohemia in České Budějovice

Conference Proceedings

37th International Conference on Mathematical Methods in Economics 2019

České Budějovice | September 11–13, 2019

Published by:
University of South Bohemia in České Budějovice, Faculty of Economics
Studentská 13, 370 05 České Budějovice, Czech Republic

Editors: Michal Houda, Radim Remeš

The publication has not passed the language correction.

ISBN 978-80-7394-760-6