Survival Analysis of Corporate Defaults in the Czech Construction Sector

Martina Novotná
VŠB - Technical University of Ostrava
Faculty of Economics, Department of Finance
Sokolská třída 33, Ostrava
Czech Republic
e-mail: martina.novotna@vsb.cz

Abstract
Survival analysis is a statistical technique which can be used for the analysis of time to event data. The purpose of this study is to provide a specific application of survival analysis in the area of credit risk. The aim of this paper is to analyse the time to default and explore the effect of selected variables on time to corporate failures. Survival analysis in this study is based on modelling the time interval between foundation of the company and its bankruptcy. In this paper, two survival models are estimated by the Cox proportional hazard model, including a model with quadratic terms. The overall results of the analysis suggest there are certain financial variables with a significant effect on survivorship of Czech construction firms. The empirical results provide evidence that selected indicators of return, coverage, turnover and liquidity can be considered as key variables in the hazard of corporate bankruptcy. Thus, the main contribution of this paper is in examining the survivor data of the Czech construction companies, and in identifying variables with a significant effect on time to their corporate failures.

Keywords: Cox proportional model, duration, hazard rate, failure, survival analysis
JEL codes: G30, G32, G33

1. Introduction

The paper aims at the analysis of survival data of Czech construction companies, where the time between the foundation of the company and its failure is modelled by the means of survival analysis. The main objective of this study is to use the Cox proportional hazards models and to estimate the survival and hazard functions. The failure of companies in this study is determined by the occurrence of a bankruptcy during the observed time span. The bankruptcy of companies is usually the basis of credit score models, which are statistically derived models of the prediction of credit risk. Among all the studies on scoring models, we can mention the study by Altman (1968) and the model known as the Altman´s model or Z- score model. The approach of survival analysis can be seen as an alternative way to examine the survivor data. For example Kelly et al. (2015) focus on corporate liquidations in Ireland, Lonzada et al. (2014) model time to default on a personal loan portfolio. As they state in their article, due to the continuous monitoring of risk over time, survival models are being proposed in financial risk management as alternative tools. Their empirical study is illustrated on a credit data from a Brazilian commercial bank and their results show that the attention should be paid to continuous checking of the validity of requirements for use of the available models. Among other studies, Agarwal and Audretsch (2001) focus on the effect of the size of a time on its survival. In their study, they find that smaller companies face a lower likelihood of survival when compared to larger companies. However, they suggest that general pronouncements are hazardous, because the role of the size changes over the industry cycle and with the technological demands of that industry.

In this paper, the empirical analysis of survival time on corporate data is provided. For the purposes of the analysis of time to event, it is suggested to use the regression models that are appropriate for survivor data (Hosmer et al., 2008). As Hosmer et al. (2008, p. 3) state, the most important differences between the outcome variables modelled via linear and logistic regression analyses and the time variable is the fact that we may only observe the survival time partially. If the time until the occurrence of the event is not important, the event can be analysed as a binary outcome using the logistic regression model (Harrell, 2010, p. 389). As Harrell (2010) points out, survival analysis is used to analyse the data in which the time until event is of interest. The input variable is the time until the event, or duration time. The survival analysis allows the response to be incompletely determined for some subjects, perhaps we are not able to follow all observations in the dataset. For example, some companies are still alive after the observation time, or they might be lost to follow-up. As we face the problem of incomplete information, we need to analyse the data using the specialised survival techniques. The analysis involves censoring mechanism, when we define the censored and uncensored observations. For example, Hosmer et al. (2008, p. 18) define a censored observation as one whose value is incomplete due to random factors for each subject. If no responses are censored, standard regression models for continuous responses could be used to analyse the failure times (Harrell, 2010). Based on the assumptions about the distribution of failure times, we can use parametric, semiparametric and nonparametric modelling. In this paper, the focus is paid to the application of semiparametric methods such as Cox proportional hazards model. The main principles of this approach and used methodology are described in the chapter two of this paper. The empirical analysis and the examination of the construction sector are provided in the chapter three, where the multivariable survivor model is estimated by the means of Cox model. The attention is paid to the interpretation of the hazard ratios and practical implications of the model. Finally, overall results and recommendations are summarized in the conclusion of this article.

2. Methodology description

The primary objective of this paper is to use survival analysis on the corporate data to estimate the survival and hazard functions. Survival analysis is an approach that allows working with censored data and modelling the time to an event, such as a corporate failure. To model the time to event, two time points must be clearly defined, the beginning point and an endpoint when the event of interest occurs. Then, the survival time is the distance on the time scale between these two points (Hosmer et al., 2008). When applying the survival analysis, we deal with the process of censoring the data. It comes from the fact that we can face the problem of incomplete observation of time. It usually occurs when the observation begins at the defined time and terminates before the outcome of interest is observed. The most common type of censoring is right censoring, because the incomplete observations occur in the right tail of the time axis. The estimated survival function incorporates all the information available, both uncensored (event times) and censored observations. In this chapter, the elementary terminology and relations of survival analysis are described. Firstly, the attention will be paid to survival and hazard functions, and then the Cox proportional model will be presented.

2.1 Survival and Hazard Functions

The survival function evaluated at time t can be considered as the probability that a subject will live for at least time *t* (Gourieroux and Jasiak, 2007). It takes values between 0 and 1 and is decreasing in *t*. At t = 0 the survival function is equal to 1 and decreases toward zero as *t* goes to infinity (Cleves et al., 2010).

The term survival function, *S*, is given by

$$S(t) = 1 - F(t) = \Pr(T > t)\ , \tag{1}$$

where *T* is a nonnegative random variable denoting the time to a failure event. As Cleves et al. (2010, p. 7) show, the survivor function is the reverse cumulative distribution of *T*:

$$F(t) = \Pr(T \leq t). \tag{2}$$

Using the survival function, we can estimate the probability of surviving beyond time *t*. In other words, we can estimate the probability that there is no failure event prior to *t*.

The density function *f(t)* can be obtained both from *S(t)* or *F(t)*:

$$f(t) = \frac{dF(t)}{dt} = \frac{d}{dt}\{1 - S(t)\} = -S'(t)\ . \tag{3}$$

The hazard function or rate *h(t)* at time *t* can be explained as the probability that the company will default very shortly after reaching time *t*, provided that it reaches time *t* (Gourieroux and Jasiak, 2007). Cleves et al. (2010) explain the hazard rate as the conditional failure rate or the intensity function. As they emphasize, the hazard rate represents the instantaneous rate of failure with 1/t units. Said differently, it is the probability that the failure event occurs in a given interval, conditional upon the subject having survived to the beginning of that interval, divided by the width of the interval (Cleves et al., 2010):

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t + \Delta t > T > t | T > t}{\Delta t} = \frac{f(t)}{S(t)}. \tag{4}$$

The hazard function can range from zero (no risk) to infinity (the certainty of failure at that instant) and can be decreasing, increasing, or constant, or it can even take on other different shapes.

The relationship between the hazard and the survival function can be described as

$$h(t) = \frac{f(t)}{S(t)}. \tag{5}$$

Gourieroux and Jasiak (2007) use the duration dependence to describe the relationship between the exit rate and the time spent in a given state by a subject. It is determined by the form of the hazard function. For example, the positive duration dependence in a sequence of failure events occurring randomly in time means that the more time elapsed since the last failure event, the greater the probability of an instantaneous occurrence of another failure. There are three types of duration dependence: (i) negative, associated with decreasing hazard functions, (ii) positive, associated with increasing hazard functions, and (iii) there can be absence of duration dependence, when there is no relationship between the exit rate and the duration.

2.2 Cox Proportional Hazards Model

Analysis of survival data can be based on parametric, semiparametric and nonparametric modelling. While parametric models require assumptions about the distribution of failure times, semiparametric models are parametric in the sense that the effect of the covariates is assumed to take a certain form (Cleves et al., 2010). In other words, they are semiparametric models in terms that no parametric form of the survival function is specified, yet the effects of covariates are parametrized to modify the baseline survivor function. In general the baseline survival function is the function for which all covariates are equal to zero in a certain way. In the Cox model specifically, we assume that the covariates multiplicatively shift the baseline hazard function (Cleves et al., 2010). The form of the Cox model can be formulated as

$$h(t|\mathbf{x}) = h_0(t)\exp(\mathbf{x}\boldsymbol{\beta}_x), \tag{6}$$

where $\boldsymbol{\beta}_x$ are the regression coefficients and $h_0(t)$ is the baseline function. In this model, we do not make any assumptions about $h_0(t)$, however at a cost of a loss in efficiency. As Hosmer et al. (2008) point out, the baseline hazard function can be seen as a generalization of the intercept or constant term found in parametric regression models. The Cox model (6) is the most used form of the hazard function which was first proposed by Cox in 1972. The term proportional hazards (PH) refers to the fact that the hazard functions are multiplicatively related (Hosmer et al., 2008, p. 70). The regression coefficients can be estimated by the partial maximum likelihood method, which is described for example by Gourieroux and Jasiak (2007, p. 99). Cleves et al. (2010) use the term relative hazard for $\exp(\mathbf{x}\boldsymbol{\beta}_x)$, and the log relative hazard, or risk score, for $\mathbf{x}\boldsymbol{\beta}_x$.

To verify the specification of $\mathbf{x}\boldsymbol{\beta}_x$ and an adequate parametrization of the model, we can use tests called tests of the proportional-hazard assumptions (P-H assumptions). In this study, the tests are based on the analysis of residuals. As to the fact that the proportional hazards model to censored survival data is fit using the partial likelihood, the calculation of residuals differs from the usual regression models. For this reason, various approaches have been developed for the purposes of Cox proportional model. The residuals used in this study are based Schoenfeld residuals, for more details see for example Hosmer et al. (2008), Cleves et al. (2010), Harrel (2010). Since the survival models estimate the time to event, the explained variation should be assessed after the development of the model. The measures of explained variation for use with censored survival data differ from the traditional concept of variation using the index of determination. Roysten (2006) proposed a measure with the character of explained variation in proportional hazards models which can be used as an adjusted index of determination in PH models.

3. Empirical Study and Model Estimation

The survival analysis in this paper is used on the data of selected Czech companies from the construction sector. For the purposes of the analysis, the data about the companies were extracted from the Bisnode Magnusweb database[1] and from the government portal Justice.cz[2]. The sample comprises data of 4546 companies, including 665 failrures. For the purposes of the analysis, the dates of two types of events are essential: the date of company foundation (t=0) and the date of the bankruptcy (t =1). The companies are observed during the period 1988 – 2015 and they were founded during the period 1988 – 2005. The end of the study is March 15, 2015. If the company did not bankrupt until this date, or if the company was not registered in the database any more, it is assumed to be a censored observation. Otherwise, the observation is uncensored. Each record documents the time span of a particular company and 24 quantitative variables (financial analysis ratios of activity, profitability, liquidity and solvency observed at the end of the particular years).

3.1 Cox Proportional Hazards Model Estimation

The survival analysis in this paper is based on the Cox proportional hazards model and we analyze the impact of the selected variables on time to corporate failure. In the first step, the individual coefficients are estimated to determine variables with a significant impact on the hazard rate. There are various methods for the model development and the selection of influential variables. For example, Hosmer (2008) suggests purposeful or stepwise selection of covariates. Using the univariable analysis in this study, we can determine significant variables at the 20 percent level. The statistical significance is based on the Wald test of the null hypothesis, $H_0 : \beta_x = 0$ versus $H_1 : \beta_x \neq 0$. Results show that there are seven significantly important covariates on time to failure (Table 1).

Table 1: Univariable Survival Analysis

Financial ratio	Variable	Coef.	Std. error	z	P>|z|	95% confidence interval	
Logarithm of total assets	*lnta*	0.25918	0.03572	7.26	0.000	0.18918	0.32918
Return on assets	*roa*	-0.00126	0.00085	-1.48	0.140	-0.00292	0.00041
Coverage of long-term assets	*cla*	-0.012871	0.00108	-11.87	0.000	-0.01500	-0.01075
Interest coverage	*ic*	-0.00005	0.00003	-1.85	0.064	-0.00011	$3.13.10^{-6}$
Total assets turnover	*ta_turn*	-0.17911	0.048851	-3.67	0.000	-0.27485	-0.08336
Current ratio	*cr*	-0.00313	0.00128	-2.46	0.014	-0.00564	-0.00063
Cash ratio	*cash*	-0.00341	0.00157	-2.17	0.030	-0.00650	-0.00033

Source: author’s calculations

[1] Bisnode Magnusweb [online database]. Available from: http://www.bisnode.cz/ [cit. 2015-03-15].

[2] Justice.cz [online]. Available from: https://or.justice.cz/ias/ui/rejstrik [cit. 2015-03-15].

Using the univariable analysis, we can identify influential variables which can be used to fit a multivariable model in the next step. The final multivariable model contains five variables which are statistically significant at the level of 0.05 (Table 2). The overall significance of the model is tested by the log partial likelihood ratio test, where the value of the test is G = 207.63, and the G statistic follows chi-square distribution with 5 degrees of freedom. Since the p-value for the test is less than 0.000, at least one of the coefficients in the model is significantly associated with survival time. As it is evident from the table (Table 2), all five variables are statistically significant at the level of significance of 0.05, based on the Wald statistic. Using the estimated coefficients, we can identify the relationship between each variable and survival time. It is evident that an increase in the following four variables *roa, cla, ta_turn, cr* decreases the hazard, while the hazard is increased by the increase in *lnta*.

Table 2: Multivariable Survival Model

Financial variable	Variable	Coef.	Std. error	z	P>|z|	95% confidence interval	
Logarithm of total assets	*lnta*	0.35148	0.04001	8.9	0.000	0.273072	0.42990
Return on assets	*roa*	-0.38846	0.05030	-7.72	0.000	-0.48705	-0.28988
Coverage of long-term assets	*cla*	-0.01490	0.00125	-11.90	0.000	-0.01736	-0.01245
Total assets turnover	*ta_turn*	-0.29249	0.06819	-4.30	0.000	-0.42581	-0.15918
Current ratio	*cr*	-0.00356	0.00137	-2.60	0.009	-0.00624	-0.00088

Source: author's calculations

To interpret the results, we can use the exponentiated individual coefficients that represent the ratio of the hazards for a 1-unit change in the corresponding covariate. The hazard ratios are shown in the table below (Table 3).

Table 3: Hazard Ratios

Financial variable	*lnta*	*roa*	*cla*	*ta_turn*	*cr*
Hazard ratio	1.42117	0.67810	0.98521	0.74640	0.99645

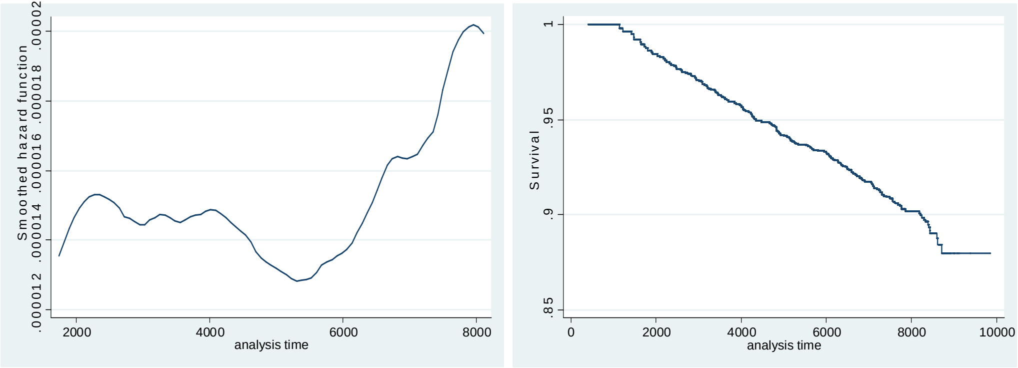
Source: author's calculations

For example, a 1-unit increase in *roa* decreases the hazard by 32.2%. From the economic point of view, the results are consistent with theoretical assumptions. The higher the return on assets, the coverage of long-term assets, the turnover of total assets and the current liquidity ratio, the lower the hazard of bankruptcy. As can be seen, the final estimated model includes the ratios of profitability, activity and liquidity. The size of the company is another significantly important factor in the model, however with an opposite impact on the hazard. When transformed to the logarithm of total assets, the higher the variable, the higher the hazard rate. In conclusion, the model implies that larger companies face a higher probability to corporate failure. It is likely to be a specific attribute of the Czech construction sector and may be explained by the stage of industry life cycle, technological demands or other factors, such as suggested by Agarwal and Aaudretsch (2001).

3.1.1 Survival and Hazard Functions Estimates

The overall estimated survival function for the data is shown in Figure 1 (a). The estimated hazard function shows the probability that the failure event occurs in a given interval (kernel smoother is applied) and decreases with time. As we can see from the graph, Figure 1 (b), the hazard rates change meaning that the risk of failure is not constant over time.

Figure 1: Survival and Smoothed Hazard Functions Estimates



a) Survival function

(b) Smoothed hazard function

Source: author’s calculations

3.1.2 Model Verification

The assumption of proportional-hazards specification is based on the analysis of residuals. Based on the variable-by-variable tests and the combined test, the overall proportional-hazards (PH) assumption is not violated at the significance level of 0.05 (Table 4).

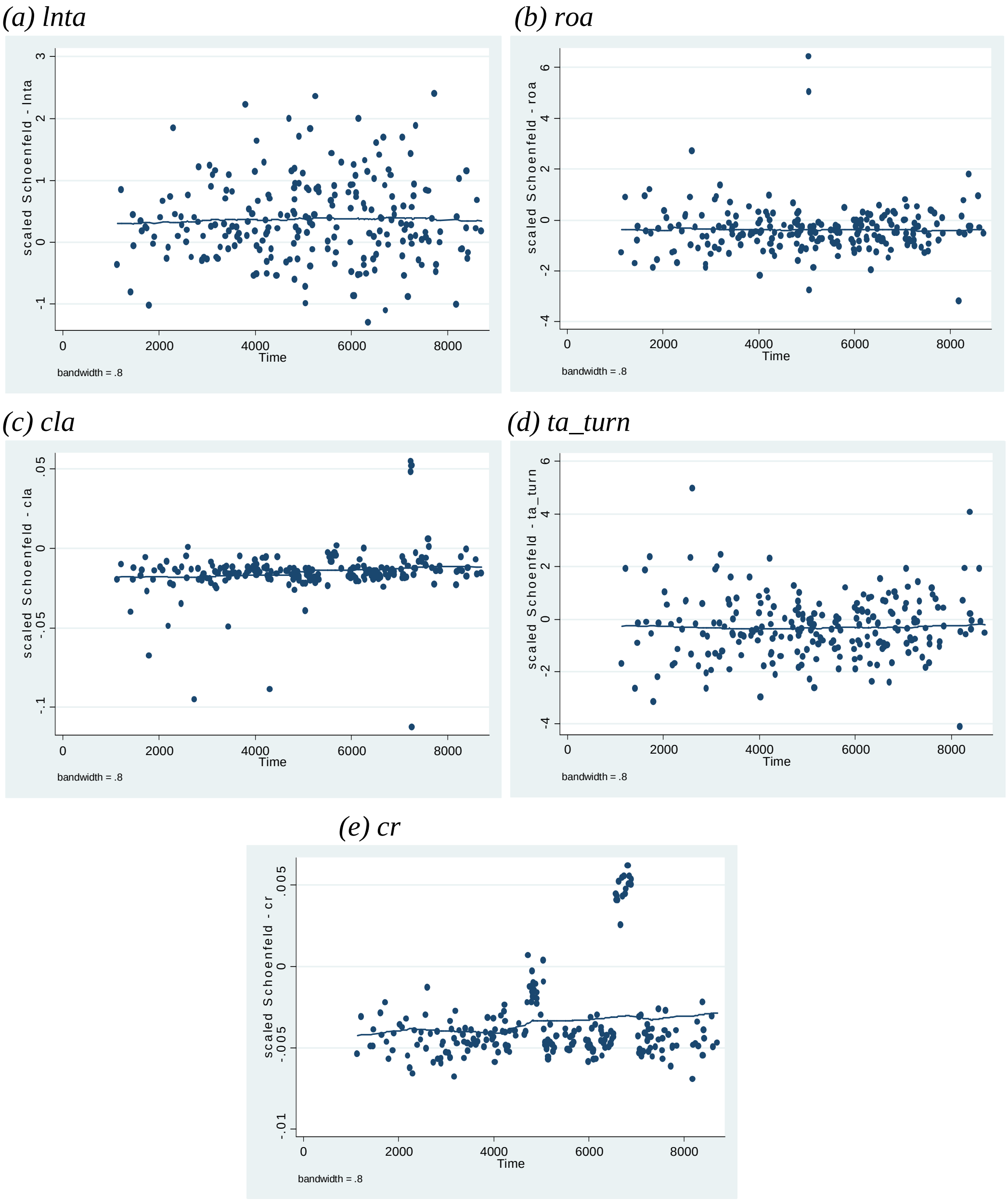
Table 4: The Overall Test of the PH Assumptions

Var	rho	chi2	df	Prob>chi2
lnta	0.03734	0.39	1	0.5331
roa	0.01390	0.07	1	0.7980
cla	0.22469	7.51	1	0.0062
ta_turn	0.04069	0.51	1	0.4773
cr	0.18835	0.13	1	0.7191
Global		8.44	5	0.1338

Source: author’s calculations

The proportional-hazards assumption of individual covariates can be assessed by the use of graphs. The graphs of all covariates included in the model are shown in the following figure (Figure 2). The curves are roughly linear with a nonzero slope for all covariates which means there is no need of covariates transformation.

Figure 2: Tests of PH Assumptions



(a) lnta *(b) roa* *(c) cla* *(d) ta_turn* *(e) cr*

Source: author's calculations

The explained variation of the model measured by the adjusted index of determination R^2 equals 0.368519 (SE = 0.032674). The greatest contribution to the explained variation is carried by covariates *cla* and *lnta*, followed by *ta_turn*, *roa* and *cr*.

3.2 Cox Model Modification

As to the previous results, the PH model contains five continuous variables *cla, lnta, ta_turn, roa* and *cr*. In the next step, we fit the model considering the quadratic effects of covariates. The final model contains two quadratic forms, *qcla* and *qcr*, in addition to the previous model in Chapter 3.1. We can see results in the following table (Table 5).

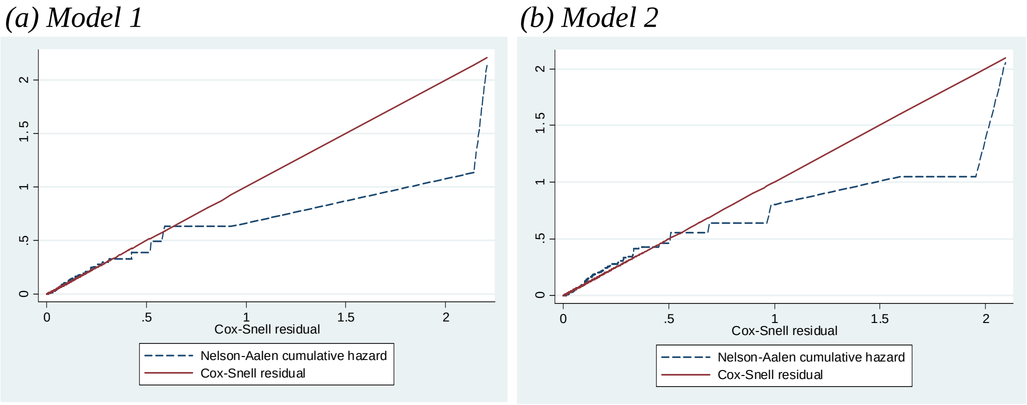
Table 5: Multivariable Survival Model with Quadratic Terms

Financial variable	Variable	Coef.	Std. error	z	P>|z|	95% confidence interval	
Logarithm of total assets	*lnta*	0.382032	0.04007	9.53	0.000	0.30349	0.46057
Return on assets	*roa*	-0.318292	0.048928	-6.51	0.000	-0.41419	-0.22240
Coverage of long-term assets	*cla*	-0.029276	0.004738	-6.18	0.000	-0.03856	-0.01999
Q. Coverage of long-term assets	*qcla*	-0.00004	0.000014	-3.21	0.001	-0.00007	-0.00002
Total assets turnover	*ta_turn*	-0.264507	0.06440	-4.11	0.000	-0.39073	-0.13829
Current ratio	*cr*	-0.632771	0.103885	-6.09	0.000	-0.83638	-0.42916
Q.Current ratio	*qcr*	-0.086747	0.029916	-2.90	0.004	-0.14538	-0.028113

Source: author's calculations

As can be seen in the table (Table 5), the general interpretation of the effect of covariates did not change when compared to the previous model (Table 2). The value of log partial likelihood ratio test is *G* = 407.64, and the *G* statistic follows chi-square distribution with 7 degrees of freedom. Since the *p*-value for the test is less than 0.000, at least one of the coefficients in the model is significantly associated with survival time. All the estimated coefficients are significant at 0.05 level. The explained variation of the model measured R^2 equals 0.541556 (SE = 0.024892). The greatest contribution to the explained variation is carried by covariates *cla* and *qcr,* followed by *lnta, qcla, ta_turn, roa, cr.*

Figure 3: Goodness of Fit


(a) Model 1

(b) Model 2

Source: author's calculations

In the figure above (Figure 3), we plot the Nelson-Aalen cumulative hazard estimator for Cox-Snell residuals. We can see some variability about the 45°, particularly in the right-hand tail. This is the reason of the reduced effective sample caused by prior failures and censoring. It is evident that the second model with quadratic forms fits better when compared to the first model.

4. Conclusion

The paper was devoted to the analysis of corporate failures using the survival analysis. In this study, the survival analysis was carried out to estimate the survival and hazard functions of the Czech construction sector. The first chapter provided some introduction about the use of survival analysis in corporate failure prediction and explained the use of censored and uncensored data. In the next chapter, the attention was paid to a brief description of methodology. For the reason that some methods are very specific, you can find some relevant literature for more details and derivations. Finally, the application on a data sample of the Czech companies was carried out using Cox proportional hazards model.

Two models were estimated and the outputs are summarized in tables (Table 2, Table 5). Both models contain five covariates; and the second model includes quadratic terms of *cla* and *cr* in addition. The models suggest that the higher the return on assets, the coverage of long-term assets, the turnover of total assets and the current liquidity ratio, the lower the hazard of bankruptcy. The size of the company is another significantly important factor in the model, however with an opposite impact on the hazard. In conclusion, the model implies that larger companies face a higher probability to corporate failure. It is likely to be a specific aspect of the Czech construction sector and may be explained by the stage of industry life cycle, technological demands or other factors, such as suggested by Agarwal and Audretsch (2001). The possible explanation of the unusual result may be associated with a decrease in public investments and a decline in housing construction in the Czech Republic during the observed period, which is crucial primarily for large construction companies.

Both models were verified to access the fit of the model. The results show that the consideration of quadratic forms increased the fit of the model. In conclusion, the survival analysis is a useful method for the analysis of censored and uncensored data. Suggestions for further research include the use of parametric models, which are more flexible and they can overcome the problems of relatively poor fit of the Cox model.

Acknowledgement

This paper was supported within Operational Programme Education for Competitiveness – Project No. CZ.1.07/2.3.00/20.0296 and SGS Project VSB – TU Ostrava 2015/75 “Application of generalized linear models in insurance and finance.”

References

ALTMAN, E.I. (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, vol. 23, no. 4, pp. 189–209.

AGARWAL, R., AUDRETSCH, D.B. (2001). Does Entry Size Matter? The Impact of the Life Cycle and Technology on Firm Survival. *The Journal of Industrial Economics*, vol. 49, no. 1, pp. 21–43.

CLEVES, M., GOULD, W., GUTIERREZ, R.G., MARCHENKO, Y.V. (2010). *An Introduction to Survival Analysis Using Stata*. 3rd ed. College Station, Tex: Stata Press.

GOURIEROUX, CH., JASIAK, J. (2007). *The Econometric of Individual Risk*. New Jersey: Princeton.

HARRELL, F.E. (2010). *Regression Modeling Strategies. With Applications to Linear Models, Logistic Regression, and Survival analysis*. New York: Springer-Verlag.

HOSMER, D.W., LEMESHOW, S., MAY, S. (2008). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: Wiley.

KELLY, R., BRIEN, E. O., STUART, R. (2015). A long-run survival analysis of corporate liquidations in Ireland. *Small Business Econ*, vol. 44, pp. 671–683.

LOUZADA, F., CANCHO, V. D., OLIVEIRA, M. R., BAO, Y. (2014). Modeling Time to Default on a Personal Loan Portfolio in Presence of Disproportionate Hazard Rates. *Journal of Statistics Applications and Probability*, vol. 3, no. 3, pp. 1–11.

ROYSTON, P. (2006). Explained variation for survival models. *The Stata Journal*, vol. 6, no. 1, pp. 83–96.