

MSc Economics

A Tourism Sector Credit Default Model

A Master's Thesis submitted for the degree of
“Master of Science”

supervised by
Alex Stomper

Richard Franz
0452154

Vienna, 14 June 2010



INSTITUT FÜR HÖHERE STUDIEN
INSTITUTE FOR ADVANCED STUDIES
Vienna



CONTINUING
EDUCATION
CENTER

MSc Economics

Affidavit

I, Richard Franz


hereby declare

that I am the sole author of the present Master's Thesis,

A Tourism Sector Credit Default Model

34 pages, bound, and that I have not used any source or tool other than those referenced or any other illicit aid or tool, and that I have not prior to this date submitted this Master's Thesis as an examination paper in any form in Austria or abroad.

Vienna, 14 June 2010


Signature

Contents

1	Introduction	5
2	Econometric Method	5
2.1	Introduction and Definitions	5
2.2	Semiparametric Survival Time Analysis	7
2.3	Testing the Proportionality Assumption	9
3	Data	10
4	Modeling and Estimation	13
4.1	Descriptive Analysis	13
4.2	Modeling	14
4.3	Results	17
5	Conclusion	26
A	Data Tables	30

List of Tables

1	Restructuring Firms vs. normal Firms	11
2	Final Models	18
3	Estimated Models	21
4	Economic Size Variables, Summary	23
5	Raw Variables	30
6	Descriptive Statistic for relevant Variables, by group	31
7	Variables	32
8	Selected Variables as described in Modeling, 4.2	33
9	Economic Size Variables	34

List of Figures

1	Kaplan Meier Survival Function by group	13
2	Nelson Aalen Cumulative Hazard Function by group	14
3	Smoothed Hazard Estimate	14
4	Estimated Hazard	25
5	Estimated Cumulative Hazard	25

Abstract

For bank's share- and stakeholders a borrower's default can lead to significant costs. An appropriate credit risk model is thus needed to ensure that timely enough steps can be taken to avoid failure of firms. A prominent way to model credit risk is survival time analysis which accounts for censored data. This study considers the Austrian hotel sector taking weather, market and macro data into account. As there is a significant difference between firms entering the sample as restructuring cases and firms which are healthy at that time two models are estimated. The final models include macro and weather data and feature an inverted-U shaped estimated hazard rate. This is in-line with the literature and intuition.

1 Introduction

The default of a borrower can be costly for share- and stakeholders. An appropriate credit risk model is therefore needed in order to take appropriate steps against the possibility of significant losses. The literature on credit risk modeling is expanding with many researchers putting emphasis on this topic. A prominent way to model credit risk comes from the area of Biology: survival time analysis. One of the first to introduce this method in economics was Kiefer (1988). This method explicitly accounts for time, exploits time-varying covariates and utilizes more data as pointed out by Shumway (2001). Shumway (2001) also showed that the estimator of a simple static model (e.g. a logit model) is generally inconsistent compared with a hazard model estimator. A higher prediction accuracy of duration models compared to single-period logistic models was also pointed out by Daniele De Leonardis and Roberto Rocci (2008). Duffie et al. (2009) pointed out the importance of considering unexplained variance in a frailty part of the estimation. They also extended the ways of modeling default correlation. A similar approach considering a frailty term was conducted by Daniele De Leonardis and Roberto Rocci (2008).

This paper applies these methods to data about the Austrian hotel sector. The explaining variables are macroeconomic, market and weather variables. The firm specific part is modeled as a frailty term. The dependent variable is the duration of the business relation between the bank and the borrower and is specified in more detail below.

The main source of the data is the Oesterreichische Hotel- und Tourismusbank (OeHT), the Austrian Hotel and Tourism bank. The macroeconomic data was provided by the Oesterreichische Nationalbank, the Austrian national bank, and the weather data by ZAMG.

The outline of this paper is the following: chapter 2 explains the econometric methods used, in chapter 3 I will provide a description of the data, in section 4 the model will be explained and the results will be discussed.

2 Econometric Method

2.1 Introduction and Definitions

As the name survival time analysis suggests, the dependent variable is the time when a subject is at risk to fail at a given point in its life. The method is mostly used in biology where for example medical treatments are tested. The failure of the subject could be death or more positive cure. In this context a subject

is considered to be at risk as soon as it is likely to fail. Surely, if someone has already died he is not at risk anymore. The same applies if the subject has not yet been born. In most cases it is impossible to observe the entire period of the subject when it is at risk. However, this problem can be addressed in survival time analysis. The two concepts are:

- Truncation: Period over which the subject was not observed but is, a posteriori, known to not have failed.
- Censoring: Failure event occurs when subject is not under observation.

The definitions are due to Cleves et al. (2008) and are both relevant for this study as will be described below. The most common type of data incompleteness is the occurrence of right censoring. This is the case when the study ends but the subject has not yet failed but is still at risk. Then the exact point in time will not be known when the subject will eventually fail. Other types as left and interval censoring and truncation are discussed in detail in Hosmer et al. (2008) and Cleves et al. (2008).

The variable to be modeled in this study is the duration of the business relation between the bank and the borrower. The motivation behind this definition of time duration is that the bank is not primarily interested in when the firm defaults but when the credit relation is distressed. This definition relying on business relation is from a bank perspective better to interpret and moreover data is available on these events. A failure is defined as the event when a borrower asks for deferment of an installment for the first time or when a firm actually files for bankruptcy. A failure is thus the end of the business relationship between the bank and a borrower. Hayden (2003) investigated the predictive power of credit risk models based on different default definitions (bankruptcy, restructuring and delay-in-payment) and found that it makes little difference which definition is used. It is assumed that a firm does not withdraw from the market voluntarily so firms want to stay in business forever. This assumption is necessary as beginning from the day the first credit is granted to the borrower the business relation is under risk of failing until a failure occurs. The definition does not depend on the scheduled end of a credit. A definition relying on credit lengths would need to deal with non random failure events. This is the case when the credit is paid back within the scheduled time. There are some important facts for the interpretation of the duration of the business relation to be considered.

- If one credit or many credits are granted and they span the whole observed period it will be always known whether a subject has failed.

- Interval truncation: in the case of a gap in the credit history the subject cannot fail through deferment of the repayment of the credit but through filing for bankruptcy. The subject is still at risk, the business relation can still be considered to be ongoing if there was no bankruptcy. As soon as a new credit is granted it is known that the subject has not failed between. And if the firm had filed for bankruptcy this would be known too.
- The former case can be easily extended to right censoring. As not all future granted credits are known there is at one point in time one date where the most ongoing credit is scheduled to be paid back. Of course, another credit could be granted at a later point in time or the borrower could file for bankruptcy.

2.2 Semiparametric Survival Time Analysis

This section presents the methodology of the analysis below. The most relevant references are Leonardis and Rocci (2008), Duffie et al. (2009) and Shumway (2001). As main econometric references serve Hosmer et al. (2008), Duchateau and Janssen (2008), Kiefer (1988), Gutierrez (2002) and Survival Analysis and Epidemiological Tables in Stata (2002). The following general discussion will be based on those references. A starting point for survival time analysis is the probability distribution

$$F(t) = P(T < t)$$

which defines the probability of a subject living up to a certain time t or less. This probability distribution is associated with a density function

$$f(t) = dF(t)/dt$$

and the survival function $S(t)$ which can be interpreted as the probability of surviving up to a time point t or longer

$$S(t) = 1 - F(t) = P(T \geq t)$$

A useful expression is the hazard function which gives the rate (not the probability) of surviving up to t but failing then. In the discrete case this is given by

$$h(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t \cdot S(t)}$$

where t specifies the starting point of a particular time interval and $t + \Delta t$ specifies the end of this interval. As $\Delta t \rightarrow 0$ the hazard function converges to

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

which describes the instantaneous failure rate if living up to t . This hazard function is usually modeled such that the covariates have a multiplicative effect on the hazard function

$$h(t, \mathbf{x}, \boldsymbol{\beta}) = h_0(t)r(\mathbf{x}, \boldsymbol{\beta})$$

where \mathbf{x} is the vector of covariates, $\boldsymbol{\beta}$ is a vector of coefficients of the covariates, $h_0(t)$ is a nonnegative function depending on time t and $r(\mathbf{x}, \boldsymbol{\beta})$ is a nonnegative function of the regressors. $h_0(t)$ can either take no specified form or can be defined as a parametric function. Popular distributions are the Exponential, Weibull, Gompertz, Lognormal, Log-logistic or Generalized Gamma distribution. $h_0(t)$ is also called the baseline hazard. It is the hazard of the firm when all covariates are zero.

Whether a fully parametric or semiparametric model is chosen depends on the interpretation of the dependent variable. Fully parametric models try to achieve two goals at the same time: (1) describe the basic underlying distribution of survival time (error component) and (2) characterize how the distribution changes as a function of covariates (systematic component).

If a prediction of life-length is required the full parametric specification is the appropriate way to model the problem. However, if it is enough to state whether some factors reduce or increase the risk of failure a semiparametric approach suffices. It is not necessary to define $h_0(t)$. Instead, the analysis is based on the hazard ratio

$$HR(t, \mathbf{x}_i, \mathbf{x}_j) = \frac{h(t, \mathbf{x}_i, \boldsymbol{\beta})}{h(t, \mathbf{x}_j, \boldsymbol{\beta})} = \frac{h_0(t)r(\mathbf{x}_i, \boldsymbol{\beta})}{h_0(t)r(\mathbf{x}_j, \boldsymbol{\beta})} = \frac{r(\mathbf{x}_i, \boldsymbol{\beta})}{r(\mathbf{x}_j, \boldsymbol{\beta})}$$

where the subscripts refer to individuals i and j . Thus the hazard ratio only depends on the function $r(\mathbf{x}, \boldsymbol{\beta})$ as $h_0(t)$ cancels. The hazard functions are assumed to be proportional, i.e. their ratio is constant over survival time. This concept was introduced by Cox (1972). Cox also suggested to use an exponential parametrization of the covariate part $r(\mathbf{x}, \boldsymbol{\beta})$ of the hazard function

$$h(t, \mathbf{x}, \boldsymbol{\beta}) = h_0(t)e^{\mathbf{x}'\boldsymbol{\beta}}$$

This specification ensures that the parameters of interest $\boldsymbol{\beta}$ can take values in an infinite parameter space and that no constraints need to be imposed during

the maximum likelihood estimation process.

If the proportionality assumption of the hazard fails one possible transformation of the non-proportional variable is (Survival Analysis and Epidemiological Tables in Stata (2002))

$$z_i(t) = z_i g(t)$$

where $g(t)$ can take any form, usually $g(t) = t$ or $g(t) = \ln(t)$. $g(t)$ can vary continuously over time. Therefore variables with this imposed structure are called continuous time varying covariates.

The hazard rate can be written as

$$h(t) = h_0(t) \exp \{ \beta_1 x_1 + \dots + \beta_k x_k + g(t)(\gamma_1 z_1 + \dots + \gamma_m z_m) \}$$

with z_l being the l 'th continuously time varying variable with coefficient γ_l .

A frailty model includes an additional unobservable covariate z_i in the hazard function

$$h_f(t, \mathbf{x}_i, \boldsymbol{\beta}) = z_i h(t, \mathbf{x}_i, \boldsymbol{\beta})$$

where subscript f denotes the hazard function with a frailty term z_i .

The frailty distribution must be chosen such that the hazard function h_f is positive. Usually a Gamma distribution is chosen with mean 1 and variance θ which needs to be estimated. If now $z_i > 1$ the individual has a larger than average hazard. It is said to be more "frail". If however $z_i < 1$ the subject is less frail.

2.3 Testing the Proportionality Assumption

The main assumption of the proportional hazard model is proportionality of the hazard. Given the logged hazard function,

$$\ln [h(t, x, \beta)] = \ln [h_0(t)] + x\beta$$

with only one dichotomous covariate the log hazard would take for $x = 0$: $\ln [h_0(t)]$ and for $x = 1$: $\ln [h_0(t)] + \beta$. The parameter β represents a proportional shift in the hazard. For a non-dichotomous covariate, the log hazard difference between two individuals with values of $x + c$ and x would be $c\beta$ at any point in time. If the proportional hazard assumption fails different modifications of the hazard function can be taken into account such as continuously varying covariates. Alternatively the model could be separately estimated for different subgroups.

Different tests have been suggested to verify if the proportionality assumption holds or not. In this paper the test is based on the assumption that $\beta_j(t) = \beta$ for all t . It has been shown that $\mathbb{E}(s_{j*}) + \hat{\beta} \approx \beta(t_j)$ where s_{j*} denotes the scaled Schoenfeld residuals and $\hat{\beta}$ is the estimated coefficient from the cox model. Therefore a plot or derived test of $s_{j*} + \beta$ versus some function of time provides an assessment of the proportionality assumption. In this study the function $g(t) = t$ and $g(t) = \ln(t)$ is chosen. See Hosmer et al. (2008) and Survival Analysis and Epidemiological Tables in Stata (2002) for further details.

3 Data

The explanatory variables used for estimation are based on yearly Austrian macroeconomic, market and weather data between 1977 and 2008. The data has been obtained from the Oesterreichische Hotel und Tourismusbank, the Oesterreichische Nationalbank and ZAMG. Macro data was used for example in the studies of Shumway (2001), Carling et al. (2004), Hazak and Maennasoo (2007) and Castro (2008). The motivation for the inclusion of macro variables is straightforward: default risks should be driven down by a stronger economy. Kaniowski et al. (2008), who also looked at the Austrian hotel sector, also referred to market data. The weather variables are added to the model as it can be expected that those significantly influence the success of this industry. Although many papers point out the importance of accounting data this approach is not taken in this study as the available balance sheet data is very sparse.

Overall there are 2180 firms in the sample available for the analysis where 352 (16.1%) fail. A failure occurs when a borrower asks for deferment of an installment for the first time or when the firm files for bankruptcy. The average yearly failure rate is about 1.4 % which seems reasonable. In total there are 26665 time points available for estimation with the longest spell lasting 27 years, the average spell 12.2 years and a minimum spell of 1 year.

There is however a potential difference in the failure rate between two groups of firms as some firms were already in a restructuring process when they appear in the sample. A priori one would assume a higher failure rate for the latter firms. Table 1 summarizes the difference between the two groups. This difference between the restructuring cases and the other firms will be measured based on an indicator variable 'group'. The variable 'group' takes the value 1 if the company was healthy at the moment when it was taken into the sample and 0 if it was already a restructuring case.

Apart from relevance a key factor for including a variable in the model is the

	Normal	Restructuring	All
Number of firms	2038	142	2180
Number of failures	304	48	352
% failures	14.92 %	33.80 %	16.14 %

Table 1: Restructuring Firms vs. normal Firms

availability of data. For the macro data there is no missing data problem. In total out of the 12 available macro variables 6 were chosen, 2 out of the 35 market and 3 out of the 31 weather variables available for each month.

For some variables annual data is available for two “seasons”, summer and winter. The summer season lasts from May to October and the winter season from November to April. The summer values of these variables are indicated by the suffix ‘_sf’ and the winter variables by the suffix ‘_wf’. The market variables were already provided classified whereas the monthly weather data was summarized in two seasonal variables, i.e. taking the averages over the respective months. There are three types of firms: firms only operating during winter, firms only opening in summer and firms which are open in both seasons. Firms classified to be only open in summer do not include any winter variables, i.e. the winter variable has a value of 0. Analogously winter firms do not include any summer variables.

Table 5 in appendix A lists the preselected raw variables used for model building, a short description and an indication whether there are years without data for each subject.

The real long (ltireal) and short term interest rates (stireal) are used to construct a spread between the two. This approach was used for example by Carling et al. (2004). A positively sloping yield curve would be an indication for a strong future economy, a negative for an economic downturn. As Carling et al. (2004) state banks will have strong incentives to renegotiate loan terms or to refrain from calling loans of firms at risk when better economic conditions are expected. Firms might also be more committed to avoid failing if the future outlook is better.

Real private consumption (pcr) and disposable income (pyr) is used to construct a ratio between these two variables. This indicates how much of the disposable income is spent on consumption. Holidays is a part of consumption and a good part of the Austrian hotel sector is relying on domestic demand. Thus this ratio could give an indication of how demand and the failure rate changes when the share of disposable income used for consumption changes.

Also related to the consumption-income ratio is the unemployment rate (urx). Higher unemployment would lead to less disposable income and it seems reasonable to assume that people will first cut unnecessary expenditures such as

holidays.

Including the real GDP expenditure (yer) as growth rate should capture the overall constitution of the economy.

Days with nice weather (Wschoen_sf'ms, Wschoen_wf'mw) would influence the failure rate via higher or lower demand for accommodation. The same applies for rainy days in the summer (Wn1_sf'ms). In the winter (Wn1_wf'mw) rain could also mean snow what would directly influence the snow covering (Wschdeck_wf'mw) which would again increase demand. Here also the difference between stock and flow variables can be pointed out. Snow would be a stock variable usually only melting starting from spring and sun as well as rain representing flow variables.

The only indicator for competition is the number of firms in a region (Mfhs_sf, Mfhw_wf). The effects could be two sided. More firms could indicate a more attractive region, but, clearly higher competition reduces the residual demand of any individual firm.

Simply because of the definition of the variables they are exogenous. However, in the default literature endogeneity is not the biggest concern. Most publications as for example Hazak and Maennasoo (2007), Shumway (2001), Maennasoo (2007), Carling et al. (2004), Castro (2008) and Leonardis and Rocci (2006) use also accounting data which is subject to endogeneity.

Out of the pool of raw variables some sensible variables were derived: ratios, percentage changes, two- and three-year moving averages from the deviation of the longterm median and mean. Ratios and percentage changes are expressed in per cent. The moving average deviation from the median or mean can be interpreted as an indicator of whether the previous period before a default occurred was exceptionally good or bad. Moreover only lagged data was considered since it is important that the output of the analysis can be used to predict defaults. Descriptive statistics of the derived variables considered in the model can be found in table 6 in appendix A. These variables are a subsample of the variables stated in table 7 in appendix A. The variable selection is described in detail in section 4.2 just below.

Additionally the average altitude of the postal code region is known where the firm is located. This information is used to construct a dummy variable 'plzhighlow' which takes the value of 1 if the altitude is more than 1000 meters and 0 else.

The variables were checked for outliers but no outliers were found.

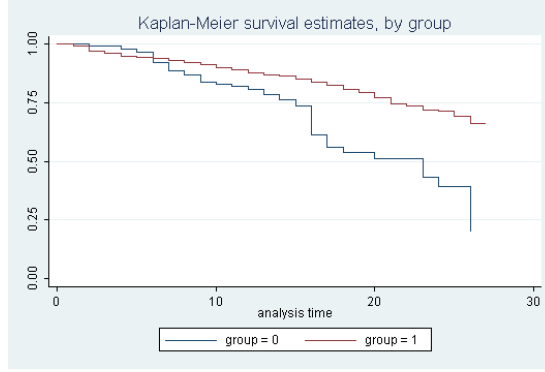


Figure 1: Kaplan Meier Survival Function by group

4 Modeling and Estimation

4.1 Descriptive Analysis

A first step in the analysis is to estimate the survivor function. The Kaplan-Meier estimator for survival at time t_j is (Kiefer (1988) and Hosmer et al. (2008))

$$\hat{S}(t_j) = \prod_{i=1}^j \frac{n_i - h_i}{n_i}$$

where h_j is the number of completed spells of duration t_j , for $j = 1, \dots, K$, K is the number of distinct completed durations and n_j is the number of spells which are not completed or censored before time t_j . Or in other words: the number of survivors less the ones lost due to censoring at time t_j . Figure 1 shows the estimates for the two groups of firms where 0 refers to the restructuring firms and 1 to the normal firms. It is visible from the plot that there is a difference between the groups: the restructuring cases are more at risk.

Another way to look at the estimates is based on $H(t)$, the cumulative hazard function. Following Hosmer et al. (2008) the Nelson-Aalen estimator is given by

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

where d_i is the number of events up to point i and n_i the number of individuals at risk at t_i . Figure 2 plots the group specific estimated cumulative hazard function, again suggesting differences between the two.

Figure 3 plots the estimated hazard function for the two groups of financially distressed and normal firms. The plots are based on the estimation method of Klein and Moeschberger (1997). They show a skewed inverted U-shape of the hazard function. This shape is typical for hazard functions of credit risk data,

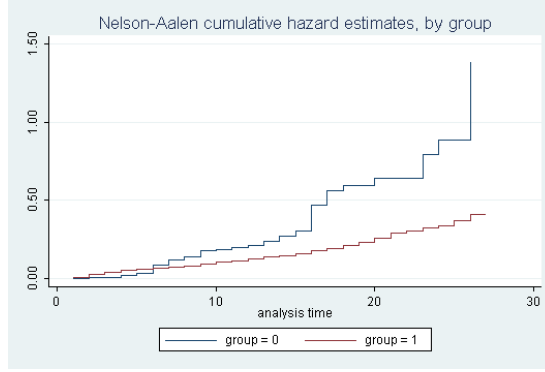


Figure 2: Nelson Aalen Cumulative Hazard Function by group

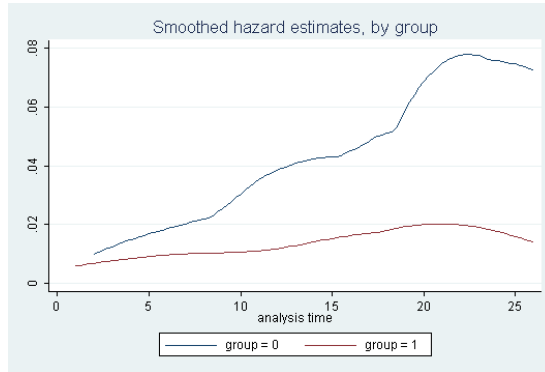


Figure 3: Smoothed Hazard Estimate

consistent with the idea that negative shocks accumulate over time. After some years firms drop then out of the market. The maximum hazard occurs quite late though, between 20 and 25 years. Successful firms surviving this period observe a decreasing risk of getting financially distressed. Kaniovski et al. (2008) and Hazak and Maennasoo (2007) refer to the same findings although usually the maximum is earlier in time.

The Wilcoxon (Breslow-Gehan) test can be used to verify whether the survivor functions of the two groups are the same or not. This test is appropriate when hazard functions vary disproportionally but censoring patterns are similar among groups. The null hypothesis of equal survival functions is clearly rejected with a p-value of 0 (test statistic: $\chi^2(1) = 27.91$).

4.2 Modeling

The model selection is based on a procedure suggested by Hosmer et al. (2008) and is adopted to consider time varying covariates. The steps are the following:

1. For all variables separately as described in table 7 in Appendix A a univariate semiparametric survival analysis is performed. The significance of the

parameters is obtained and the proportionality assumption is tested.

2. From step 1 those variables are selected that are statistically significant explanatory variables with a p-value of 0.2 or less. Within the explanatory variable groups the variable with the highest significance level is selected. A variable group is for example the interest_spread with its variations like lagged or moving average deviation from the mean. To keep the model easy to interpret those variables are preferred which are proportional. If a simpler variant of the variable is significant and proportional this one is chosen. Table 8 lists the relevant variables selected by this procedure omitting the others for the sake of brevity.
3. A large model is then estimated, using all preselected variables from step 2. Those variables which are not proportional by themselves are directly considered as continuously time varying covariates. Starting from this model variables are excluded step by step. The exclusion criterion is a significance level of 0.05% in general and 0.10% for the models considering only the restructuring firms. The more general significance level is used for the financially distressed firms as only few individuals and observed failures are available for estimation. To ensure that no important confounders are dropped from the model it is required that no coefficient changes by a too large degree. For this a value of 20 % is employed following Hosmer et al. (2008).
4. Variables which turn out to be not proportional any more in the multivariate model are transformed by multiplying by the term $g(t)$. Thus they are considered as continuously time varying covariates in order to fulfill the proportionality assumption.
5. If still one variable does not fulfill the proportionality assumption by itself it is dropped from the model.
6. The next step is to consider possible multiplicative effects as there could be interdependencies between the explanatory variables. This refers to a possible effect between the altitude of the firm and the snow covering during the winter season. A priori a multiplicative effect seems sensible. It can be assumed that firms higher in altitude are more prone to the effects of snow fall than hotels in lower regions. Here altitude could be considered as an indicator for winter tourism, so skiing and the like which requires snow.
7. Thereafter all previously dropped variables are added back into the model

to verify if they are still insignificant or not. If they turn out to be significant at this step the procedure continues from step 4 until no further variables can be added to the model.

The whole model selection process is based only on those observations without missing values of any variables.

The group variable is not proportional by itself: a p-value of 0.01 rejects the proportionality assumption, as can be seen from table 8 in appendix A. It is therefore not reasonable to estimate an overall model containing all firms. Thus another way of achieving proportionality is chosen: estimating two models, one for the restructuring firms only (group 0) and one for the healthy firms only (group 1).

For group 0 the basic model takes the following form of the hazard rate h_{g0} in covariate form

$$h_{g0}(t, \mathbf{x}, \boldsymbol{\beta}) = h_{g0,0}(t)(\mathbf{x}_{g0,d}\boldsymbol{\beta}_{g0,d} + g(t)(\mathbf{x}_{g0,c}\boldsymbol{\beta}_{g0,c})) \quad (1)$$

$\mathbf{x}_{g0,d}$ includes : interest_spread_l
 rat_pcr_pyr_lmed2
 yer_plmea2
 Wschoen_w_lwf
 Wschoen_s_lmed3_sf
 Wschdeck_w_lmed2_wf
 Wn1_w_lmed3_wf
 logMfhw_wf_l
 plzhighlow

$\mathbf{x}_{g0,c}$ includes : urx_l
 Wn1_s_lmed3_sf
 logMfhs_sf_l

where $g(t)$ denotes the continuous time varying covariate effect (either $g(t) = t$ or $g(t) = \ln(t)$), $\mathbf{x}_{g0,d}$ are the discrete time varying covariates of group 0 with related $\boldsymbol{\beta}_{g0,d}$, $\mathbf{x}_{g0,c}$ represents the continuous time varying covariates with related $\boldsymbol{\beta}_{g0,c}$ and $h_{g0,0}$ stands for the baseline hazard function of group 0.

Although the altitude (plzhighlow) is not significant for group 0 according to table 8 in appendix A it is included into the model in order to allow for possible interdependencies among the variables as discussed above. Moreover as the logged and lagged number of summer firms (logMfhs_sf_l) is included in the basis model, its insignificant winter pendant (logMfhw_wf_l) is also added back into the model.

For group 1 ($g1$) the basic model takes the form

$$h_{g1}(t, \mathbf{x}, \boldsymbol{\beta}) = h_{g1,0}(t)(\mathbf{x}_{g1,d}\boldsymbol{\beta}_{g1,d} + g(t)(\mathbf{x}_{g1,c}\boldsymbol{\beta}_{g1,c})) \quad (2)$$

$\mathbf{x}_{g1,d}$ includes : interest_spread_lmed3
 Wschoen_w_l_wf
 Wschoen_s_lmed3_sf
 Wschdeck_w_lmed2_wf
 Wn1_w_lmed2_wf
 logMfhw_wf_l
 Wn1_s_lmed3_sf
 plzhighlow

$\mathbf{x}_{g1,c}$ includes : urx_l
 rat_pcr_pyr_l
 yer_pl
 logMfhs_sf_l

4.3 Results

Based on the variable selection process described in section 4.2 the two models (equations 1 and 2) can be simplified. The final models are depicted in table 2 for two ways of modeling the continuous time varying covariates: $g(t) = t$ or $g(t) = \ln(t)$. The coefficients of the models are listed in table 3. On the basis of the Akaike information criterion the logarithmic modeling approach of the continuously varying covariates is preferred over the formulation $g(t) = t$. It should be pointed out that in the models of group 1, $g1_t$ and $g1_{\ln(t)}$, the variable Wn1_s_lmed3_sf turned out to be non proportional when included in the whole model. Therefore it was added to the continuous time varying covariate part. Moreover no market data variable turned out to be significant. The number of variables in the model seems reasonable in the light of a suggestion of Hosmer et al. (2008) who proposed as a rough guide to not include more than one covariate for each 10th failure.

The pseudo R^2 for the estimated models are low. For group 0 the pseudo R^2 has a value of 0.066 for model $g0_t$ and 0.086 for model $g0_{\ln(t)}$. The models of group 1 observe even lower pseudo R^2 with 0.026 for model $g1_t$ and 0.028 for model $g1_{\ln(t)}$. A possible explanation for these low pseudo R^2 are potentially missing firm specific covariates. It is reasonable to think that substantial reasons for a firm's default can be found within the firm.

As discussed in the section Econometric Method (2) a way to consider unex-

Model g0_t: group 0, $g(t) = t$

$$h_{g0,t}(t, \mathbf{x}, \boldsymbol{\beta}) = h_{g0,t,0}(t)(\mathbf{x}_{g0,t,d}\boldsymbol{\beta}_{g0,t,d})$$

$\mathbf{x}_{g0,(t),d}$: rat_pcr_pyr_lmed2
 yer_plmea2
 Wschoen_w_l_wf
 plzhighlow

Model g0_{ln(t)}: group 0, $g(t) = \ln(t)$

$$h_{g0,\ln(t)}(t, \mathbf{x}, \boldsymbol{\beta}) = h_{g0,\ln(t),0}(t)(\mathbf{x}_{g0,\ln(t),d}\boldsymbol{\beta}_{g0,\ln(t),d} + g(t)(\mathbf{x}_{g0,\ln,c}\boldsymbol{\beta}_{g0,\ln,c}))$$

$\mathbf{x}_{g0,\ln(t),d}$: interest_spread_l $\mathbf{x}_{g0,\ln(t),c}$: urx_l
 rat_pcr_pyr_lmed2
 yer_plmea2

Model g1_t: for group 1, $g(t) = t$

$$h_{g1,t}(t, \mathbf{x}, \boldsymbol{\beta}) = h_{g1,t,0}(t)(\mathbf{x}_{g1,t,d}\boldsymbol{\beta}_{g1,t,d} + g(t)(\mathbf{x}_{g1,t,c}\boldsymbol{\beta}_{g1,t,c}))$$

$\mathbf{x}_{g1,t,d}$: interest_spread_lmed3 $\mathbf{x}_{g1,t,c}$: Wn1_s_lmed3_sf
 Wschoen_w_l_wf yer_pl
 Wschdeck_w_lmed2_wf
 plzhighlow

Model g1_{ln(t)}: group 1, $g(t) = \ln(t)$

$$h_{g1,\ln(t)}(t, \mathbf{x}, \boldsymbol{\beta}) = h_{g1,\ln(t),0}(t)(\mathbf{x}_{g1,\ln(t),d}\boldsymbol{\beta}_{g1,\ln(t),d} + g(t)(\mathbf{x}_{g1,\ln,c}\boldsymbol{\beta}_{g1,\ln,c}))$$

$\mathbf{x}_{g1,\ln(t),d}$: interest_spread_lmed3 $\mathbf{x}_{g1,\ln(t),c}$: rat_pcr_pyr_l
 Wschoen_w_l_wf Wn1_s_lmed3_sf
 Wschdeck_w_lmed2_wf
 Wn1_w_lmed2_wf
 plzhighlow

Table 2: Final Models

plained firm specific influences is to take a frailty specification into account. The estimation of firm-level frailty terms however fails because of insufficient data. An estimation of zip-code-level frailty terms was therefore performed, but the frailty terms turned out to be insignificant in all models. Model $g1_t$ features a p-value of the frailty of 0.3892 and the model $g1_{\ln(t)}$ features a p-value of 0.391 of the frailty part. In the case of models $g0_t$ and $g0_{\ln(t)}$ the frailty estimation failed because flat or non-continuous areas were detected in the specification. In a model without covariates a frailty specification does not result in a significant frailty term on the zip code level.

The models were also checked for influential observations but no evidence could be found. For this the score residuals of the variables in the model were plotted against the variables themselves. This procedure follows again Hosmer et al. (2008).

The interpretation of the results can be most directly be seen for dichotomous variables. If a hazard rate of a dichotomous variable was estimated to be 1.6 then a firm with the variable value of 1 has a failure rate 1.6 times that of a firm with a variable value of 0 throughout the study period. Alternatively this could be expressed as a 60% larger failure rate for a firm observing a 1 over the study period.

Interpreting a continuously scaled covariate is a bit more involved and can be best seen by taking the logged difference between two firms. Assuming for the moment only one covariate the difference between two firms with the variable values of $(x + c)$ and (x) for their respective covariates can be written as (Hosmer et al. (2008))

$$\begin{aligned} [h(t, x + c, \beta) - h(t, x, \beta)] &= \{\ln [h_0(t)] + (x + c)\beta\} - \{\ln [h_0(t)] + (x)\beta\} \\ &= (x + c)\beta - x\beta \\ &= c\beta \end{aligned}$$

Then the difference between the two firms is equal to the change of the variable of interest times the coefficient. This can be expressed in terms of the hazard when taking the exponent

$$\hat{HR}(c) = e^{c\hat{\beta}}$$

This concept easily extends to the multivariate case. In general the link between the estimated coefficient, $\hat{\beta}$, of a variable and the hazard ratio, \hat{HR} , is

$$e^{\hat{\beta}} = \hat{\text{HR}}$$

For continuously time varying covariates the factor $g(t)$ must be considered in the hazard rate. There is in most cases no straightforward interpretation. Assuming for simplicity only one time continuous time varying covariate it follows that

$$\begin{aligned} [h(t, x + c, \beta) - h(t, x, \beta)] &= \{\ln[h_0(t)] + g(t)(x + c)\beta\} - \{\ln[h_0(t)] + g(t)x\beta\} \\ &= g(t)(x + c)\beta - g(t)x\beta \\ &= g(t)c\beta \end{aligned}$$

Thus the difference between the two firms is equal to the change of the variable of interest times the coefficient and the $g(t)$ term which depends on t . Expressing this as the estimated hazard ratio yields

$$\hat{\text{HR}}(c) = e^{g(t)c\hat{\beta}}$$

What follows is a discussion of the variables included in the final models preferred by the Akaike information criterion, $g0_{\ln(t)}$ for group 0 and $g1_{\ln(t)}$ for group 1. The estimation is based on a semiparametric approach. Thus only relative statements between individuals concerning their failure rates can be given. It is not possible to state when a failure might occur.

The model preferred by the Akaike information criterion for group 0, $g0_{\ln(t)}$, includes three discrete time varying covariates and one continuously time varying covariate. The former three variables are the lagged interest spread, `interst_spread_l`, the lagged two period moving average deviation from the median of the personal consumption income ratio, `rat_pcr_pyr_lmed2` and the lagged two period moving average deviation from the mean of GDP growth, `yer_plmea2`. The unemployment rate, `urx_l`, enters as continuously time varying covariate. No weather variables enter this model. Three variables, the consumption income ratio (`rat_pcr_pyr_lmed2`), the unemployment rate (`urx_l`) and GDP growth (`yer_plmea2`), have a hazard ratio larger than one. This means that the firm is more likely to fail with larger variable values. The moving average deviation of the median of the consumption income ratio (`rat_pcr_pyr_lmed2`) as well as the deviation in the GDP growth (`yer_plmea2`) show the opposite signs of what is expected by intuition as discussed in section 3. The sign is related here to the size of the hazard rate. When the hazard rate is larger than one the risk of failing increases relatively, if the hazard rate is smaller than one it decreases relatively.

model	$g0_t$		$g0_{\ln(t)}$		$g1_t$		$g1_{\ln(t)}$	
group	0		0		1		1	
tvc specification, $g(t)$	t		$\ln(t)$		t		$\ln(t)$	
r^2_p	0.066		0.086		0.026		0.028	
AIC	334.936		327.966		3276.464		3272.277	
No of subjects	137		137		1847		1847	
No of failures	43		43		247		247	
No of obs	1614		1614		21463		21463	
	$\hat{H}R$	pvalue	propp	$\hat{H}R$	pvalue	propp	$\hat{H}R$	pvalue
Macro Data								
interest_spread_l				0.66	0.04	0.91		
interest_spread_lmed3							1.62	0.00
rat_pcr_pyr_l							0.85	0.51
rat_pcr_pyr_lmed2	1.50	0.00	0.53	1.41	0.00	0.83		
urx_l				1.66	0.00	0.11		
yer_pl							1.02	0.01
yer_plmea2	1.72	0.03	0.55	1.73	0.04	0.93	0.70	
Weather Data								
Wschoen_w_l_wf	1.14	0.02	0.40				1.08	0.00
Wschdeck_w_lmed2_wf							0.88	0.00
Wn1_w_lmed2_wf							0.88	0.06
Wn1_s_lmed3_sf				1.04	0.00	0.16	1.34	0.00
General								
plzhighlow	0.48	0.05	0.86				0.67	0.01
							0.71	0.04
							0.72	0.62

Table 3: Estimated Models

Abbreviations: $\hat{H}R$ = estimated hazard ratio, pvalue = significance (p) value of $\hat{H}R$ (t-statistic), propp = univariate significance (p) value of the variable to fulfill the proportionality assumption: H_0 : the variable fulfills the proportionality assumption (distribution: $\chi^2(1)$), hazard ratios in bold are multiplied with $g(t)$

The hazard rate of the unemployment rate (`urx_l`) is in general difficult to interpret as it is modeled continuously with $g(t) = \ln(t)$. The interest spread (`interest_spread_l`) shows a hazard ratio smaller than one. Thus a higher interest spread is associated with a smaller risk of firm failures. This is in-line with the intuitive expected effect of a positively sloped yield curve discussed in section 3.

Table 4 shows the main effects of the significant covariates in economic magnitude for the models preferred by the Akaike information criterion, table 9 in appendix A shows further details. For each variable $\bar{\mu}$ is the mean of the firm-specific means, l is the overall minimum, h the overall maximum and $\bar{\sigma}$ the mean of the firm-specific variances. The economic magnitude of the correlation between the variables and the risk of failure are described by:

- $h(\bar{d}l)$: the estimated hazard of an otherwise identical firm but with an variable value of the lowest level of this variable l in comparison with an firm with the mean value $\bar{\mu}$. Omitting the otherwise identical variable values this is in terms of the logged hazard ratio

$$\begin{aligned} [h(t, l, \beta) - h(t, \bar{\mu}, \beta)] &= \{\ln[h_0(t)] + (l)\beta\} - \{\ln[h_0(t)] + (\bar{\mu})\beta\} \\ &= l\beta - \bar{\mu}\beta \\ &= (l - \bar{\mu})\beta \end{aligned}$$

Or expressed in terms of the hazard of an firm with a value of the variable of $\bar{\mu}$ in comparison with a firm with a value of the variable of l

$$\hat{\text{HR}}(l - \bar{\mu}) = e^{(l - \bar{\mu})\hat{\beta}}$$

- $h(\bar{d}h)$: the estimated hazard of an otherwise identical firm but with an variable value of h in comparison with an firm with the value of the mean level $\bar{\mu}$. Again omitting the otherwise identical variable values this is in terms of the logged hazard ratio

$$[h(t, h, \beta) - h(t, \bar{\mu}, \beta)] = (h - \bar{\mu})\beta$$

Expressed in terms of the hazard of an firm with a value of the variable of h in comparison with a firm with a value of the variable of $\bar{\mu}$

$$\hat{\text{HR}}(h - \bar{\mu}) = e^{(h - \bar{\mu})\hat{\beta}}$$

- σ^+ : the estimated hazard of an otherwise identical firm but with an variable value of the mean level plus one standard deviation, $\bar{\mu} + \bar{\sigma}$, in comparison

model	$g^0_{\ln(t)}$		$g^1_{\ln(t)}$	
	σ^-	σ^+	σ^-	σ^+
	Macro Data			
interest_spread_l	59%	37%		
interest_spread_lmed3			40%	65%
rat_pcr_pyr_l			11%	10%
rat_pcr_pyr_lmed2	38%	62%		
urx_l	26%	36%		
yer_pl				
yer_plmea2	38%	62%		
	Weather Data			
Wschoen_w_l_wf			13%	15%
Wschdeck_w_lmed2_wf			19%	16%
Wn1_w_lmed2_wf			14%	12%
Wn1_s_lmed3_sf			16%	19%

Table 4: Economic Size Variables, Summary

Abbreviations: $\bar{\mu}$ = mean of mean of firms, $\bar{\sigma}$ = mean of standard dev. of firms
 σ^- = hazard of $\bar{\mu}$ firm if one $\bar{\sigma}$ subtracted, σ^+ = analogous to σ^- if $\bar{\sigma}$ is added,
 .% = . % smaller hazard, .% = . % larger hazard

with an firm with the value of the mean level $\bar{\mu}$. In terms of the logged hazard ratio

$$\begin{aligned} [h(t, \bar{\mu} + \bar{\sigma}, \beta) - h(t, \bar{\mu}, \beta)] &= (\bar{\mu} + \bar{\sigma} - \bar{\mu})\beta \\ &= \bar{\sigma}\beta \end{aligned}$$

In terms of the hazard

$$\hat{\text{HR}}(\bar{\sigma}) = e^{\bar{\sigma}\hat{\beta}}$$

- σ^- : the estimated hazard of an otherwise identical firm but with an variable value of the mean level minus one standard deviation, $\bar{\mu} - \bar{\sigma}$, in comparison with an firm with the mean level of this variable $\bar{\mu}$. In terms of the logged hazard ratio

$$\begin{aligned} [h(t, \bar{\mu} - \bar{\sigma}, \beta) - h(t, \bar{\mu}, \beta)] &= (\bar{\mu} - (\bar{\sigma} - \bar{\mu}))\beta \\ &= -\bar{\sigma}\beta \end{aligned}$$

Or in terms of the hazard

$$\hat{\text{HR}}(-\bar{\sigma}) = e^{-\bar{\sigma}\hat{\beta}}$$

In the case of the interest spread (`interest_spread_l`) the hazard, which has a hazard ratio of less than one, the firm with a variable value of $\bar{\mu}$ has a 237% higher failure rate in comparison with a firm with a variable value of l . The failure rate is 51% smaller in comparison with a firm with a variable value of h . The effect of a positive standard deviation change is a reduction in the failure rate by 37%. A decrease by one standard deviation is associated with an increase of the failure rate by 59%. The other variables can be interpreted analogously.

In terms of the standard deviation the consumption income ratio (`rat_pcr_pyr_lmed2`) and GDP (`yer_plmea2`) has roughly the opposite economic size compared with the interest spread (`interest_spread_l`). The size of the consumption income ratio and GDP is moreover almost the same. The effect of one standard deviation change of the unemployment rate (`urx_l`) is less pronounced.

Model $g1_{\ln(t)}$ contains five non-continuous covariates: the lagged three period moving average deviation from the median of the interest spread (`interest_spread_lmed3`), the altitude (`plzhighlow`), the lagged two period moving average deviation from the median of precipitation in winter (`Wn1_w_lmed2_wf`), the lagged two period moving average deviation from the median of snow covering (`Wschdeck_w_lmed2_wf`) and the lagged nice weather days during winter (`Wschoen_w_lwf`). Two variables enter into the continuous part: the lagged consumption income ratio (`rat_pcr_pyr_l`) and the lagged three period moving average deviation from the median of precipitation during summer (`Wn1_s_lmed3_sf`).

The interest spread (`interest_spread_lmed3`), precipitation during summer (`Wn1_s_lmed3_sf`) and nice weather days during winter (`Wschoen_w_lwf`) feature a hazard rate larger than one, thus increasing the rate of failure with larger values. Four weather variables enter the model. More sunny days during the winter season increase the risk of failure (`Wschoen_w_lwf`), more snow reduces the risk (`Wschdeck_w_lmed2_wf`), as does more precipitation which translates into snow fall in the winter (`Wn1_w_lmed2_wf`). In the summer less rain is preferred (`Wn1_s_lmed3_sf`). Thus the estimated hazard ratios go well along with the intuition given in section 3. Firms in high altitude regions feature less failure. The hazard of the consumption income ratio (`rat_pcr_pyr_l`) corresponds to intuition that increasing the consumption rate and thus possibly demand for vacations reduces the risk of failure. However, the results for the interest spread (`interest_spread_lmed3`) are counter-intuitive.

Table 4 shows the main effects of the significant covariates (for more details see 9 in appendix A). The variable with the most substantial effect when a change in the value of one standard deviation occurs on the risk of a firm default is the interest spread (`interest_spread_lmed3`). It is also of the opposite sign compared

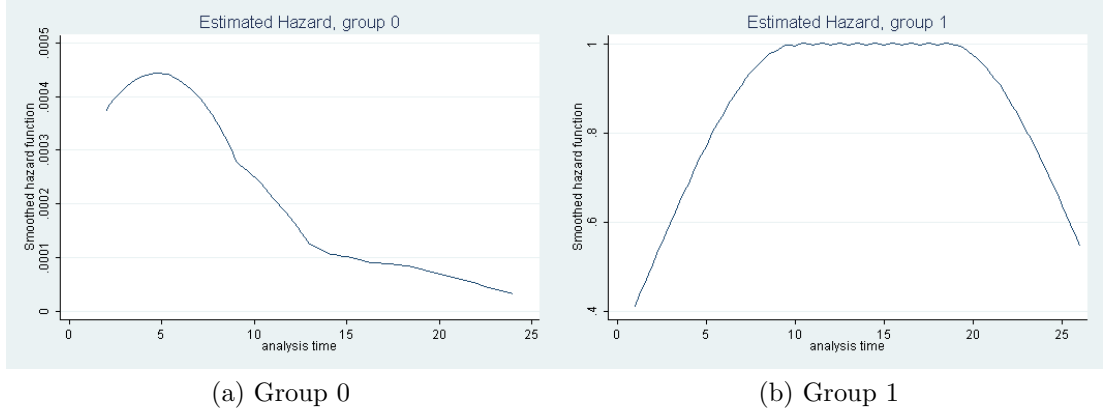


Figure 4: Estimated Hazard

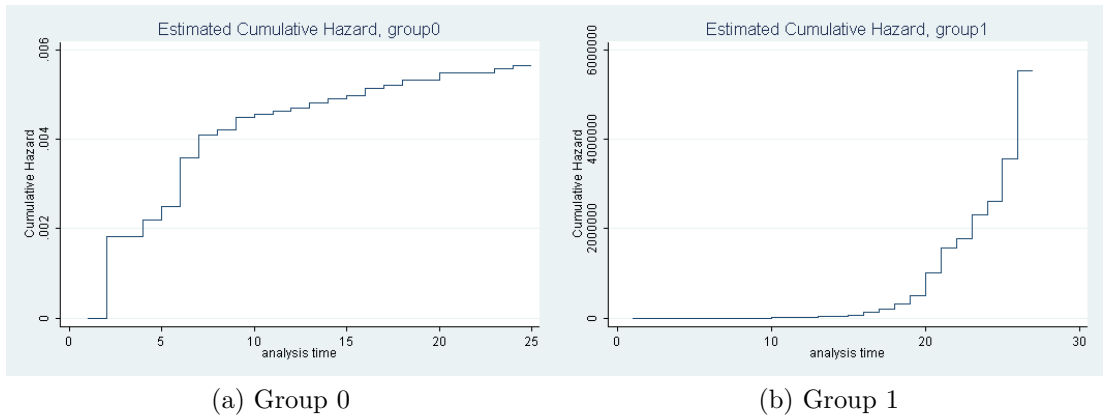


Figure 5: Estimated Cumulative Hazard

to the interest spread (`interest_spread.l`) of the model $g0_{\ln(t)}$. The consumption income ratio (`rat_pcr_pyr.l`) is of much smaller size and again of the opposite sign compared to the model of group 0. For the weather variables the economic size of a one deviation change is roughly around 15% in both directions. Table 9 in appendix A lists all economic effects of this model.

It is interesting to notice that within the groups the coefficients of the non continuously time varying covariates differ just marginally across the continuous time varying specifications, $g(t)$. This indicates some robustness choosing $g(t)$. Only for the variable `yer_plmea2` the coefficient varies more substantially between the models $g0_t$ and $g0_{\ln(t)}$, but the variation of roughly 10% is still reasonably moderate.

Figure 5 shows the estimated cumulative hazard for the two groups. Figure 4 the estimated hazard which also has a skewed inverted U-shape as the descriptive hazard estimate (figure 3) in case of the financially distressed group. However, comparing the descriptive hazard (figure 3) and the estimated hazard

(figure 4) some differences are immanent. Firstly the estimated hazard for group 0 (figure 4a) takes very high values early-on which indicates that failures tend to occur within the first years after the bank has provided a credit. Therefore the restructuring firms should be monitored heavily especially within the first years after paying out the credit. The estimated hazard for group 1 (figure 4b) shows an almost symmetric inverted U with a very long plateau. This leads to the conclusion that the risk of a group 1 firm failing is highest between year 8 and 20 but remains steady in this time slot. Within this group, the estimated cumulative hazard increases more steeply during the start of the observation period (figure 5a) than that of group 1 (figure 5b). Basically the estimated hazard rate of the model is closer to what is described by Kaniovski et al. (2008) and Hazak and Maennasoo (2007) than what can be seen from the descriptive estimates (figure 3).

5 Conclusion

For a bank's share- and stakeholders a borrower's default can be very costly. An appropriate credit risk model is thus needed to ensure that timely steps can be taken to avoid failure of firms. A prominent way to model credit risk is survival time analysis. This study considers the Austrian hotel sector taking weather, market and macro data into account. As there is a significant difference between firms which entered the sample in a healthy state and firms which were financially distressed two models are estimated using a variable selection procedure based on suggestions of Hosmer et al. (2008).

It turns out that for the restructuring firms only macro data enter the model. For the - at the start of the study - healthy group macro and weather data have significant explanatory power. The estimated hazard rates are inverted U-shaped. However there are significant differences between the two groups: the hazard of the restructuring firms is especially high at the beginning, but within the healthy firms instead low at the beginning. For the healthy firms the hazard remains almost constant for an extended period of time before it is low again on the long term end. This result is intuitive and in-line with the literature.

Concerning the fit of the model a pseudo R^2 was calculated which is between 0.03 und 0.09 and thus not particularly high. This low value could be explained as a consequence of not including any firm specific variables in the models. An attempt to address this issue through frailty estimates was not successful because the estimation failed computationally. This result is a natural starting point for an extension of this study, including firm specific covariates and filling missing

data gaps where possible. It seems also interesting to investigate the mechanisms behind the macroeconomic hazard ratios which are in some cases counter-intuitive at first sight.

References

- [1] Kenneth Carling, Tor Jacobson, Jesper Linde, and Kasper Roszbach. Corporate credit risk modelling and the macroeconomy. *Journal of Banking and Finance*, 31(3):845–868, 2004.
- [2] Christian E. Castro. Estimating a financial distress rating system for spanish firms with a simple hazard model. *Working Paper*, 2008.
- [3] Mario Alberto Cleves, William Gould, and Roberto Gutierrez. *An introduction to survival analysis using Stata*. Stata Press, College Station, Texas, second edition, 2008.
- [4] David Roxbee Cox. Regression models and life tables. *Journal of Royal Statistical Society, Series B*(34):187–220, 1972.
- [5] Luc Duchateau and Paul Janssen. *The Frailty Model*. Springer, New York, 2008.
- [6] Darrell Duffie, Andreas Eckner, Guillaume Horel, and Leandro Saita. Frailty correlated default. *Journal of Finance*, LXIV(5):2089–2123, 2009.
- [7] Roberto G. Gutierrez. On frailty models in stata, <http://fmwww.bc.edu/repec/usug2001/uk7.pdf>, accessed on 14/03/2010.
- [8] Evelyn Hayden. Are credit scoring models sensitive with respect to default definitions? evidence from the austrian market. *SSRN Working Paper*, 2003.
- [9] Aaro Hazak and Kadri Maennasoo. Indicators of corporate default. an eu based empirical study. *Eesti Pank, Bank of Estonia, Working Paper Series*, 10, 2007.
- [10] David W. Hosmer, Stanley Lemeshow, and Susanne May. *Applied Survival Analysis*. Wiley-Interscience, New Jersey, second edition, 2008.
- [11] Nicholas M. Kiefer. Economic duration data and hazard functions. *Journal of Economic Literature*, XXVI:646–679, 1988.
- [12] John P. Klein and Melvin L. Moeschberger. *Survival Analysis*. Springer, New York, second edition, 2005.
- [13] Daniele De Leonardis and Roberto Rocci. Assessing the default risk by means of a discrete-time survival analysis approach. *Applied Stochastic Models in Business and Industry*, 24:291–306, 2008.

- [14] Kadri Maennasoo. Determinants of firm sustainability in estonia. *Eesti Pank, Bank of Estonia, Working Paper Series*, 4, 2007.
- [15] Egon Smeral Serguei Kaniovski, Michael Peneder. Determinants of firm survival in the austrian accommodation sector. *Tourism Economics*, 14(3):527–543, 2008.
- [16] Tyler Shumway. Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business*, 74(1):101–124, 2001.
- [17] StataCorp. *Stata Statistical Software: Release 8.0*. Stata Press, College Station, Texas, 2003.

A Data Tables

Variable	Description	nm	am	sm
<i>Macro Data</i>				
ltireal	Long term interest rates, real	2180	0	0
stireal	Short term interest rates, real	2180	0	0
pcr	Private consumption, real	2180	0	0
pyr	Private disposable income, real	2180	0	0
urx	Unemployment rate	2180	0	0
yer	GDP expenditure, real	2180	0	0
<i>Weather Data</i>				
Wschoen_sf'ms	Days, nice weather (cc<50% per day), ms	829	228	1123
Wschoen_wf'mw	Days, nice weather (cc<50% per day), mw	822	228	1130
Wschdeck_wf'mw	Day, snow, mw	866	219	1095
Wn1_sf'ms	Days, rain, ms	889	205	1087
Wn1_wf'mw	Days, rain, mw	892	204	1084
<i>Macro Data</i>				
Mfhs_sf	Number of firms, all categories, summer	294	283	1603
Mfhw_wf	Number of firms, all categories, winter	292	284	1604

Table 5: Raw Variables

Abbreviations: nm = never missing, am = always missing, sm = sometimes missing, *_s* = summer season (May - October), *_w* = winter season (November - April), 'ms = mean on summer season months, 'mw = mean on winter season months, cc = cloud cover of sky

Variable	Unit	μ	σ	min	1%	50%	99%	max
<i>group = 0: 137 subjects, 43 failures, 1611 observations</i>								
interest_spread_lmed3	moving average absolute deviation	-0.02	1.00	-2.46	-2.45	0.13	1.53	1.69
interest_spread_l	absolute	1.37	1.19	-1.55	-1.29	1.66	3.05	3.05
plzhighlow	binary	0.74	0.44	0.00	0.00	1.00	1.00	1.00
rat_pcr_pyr_lmed2	moving average ratio deviation	0.69	1.37	-2.84	-2.67	0.66	3.07	3.50
rat_pcr_pyr_l	ratio	92.24	1.66	88.43	88.65	91.92	94.85	94.85
urx_l	percentage	3.87	0.65	1.50	2.61	3.94	5.17	5.17
Wn1_s_lmed3_sf	moving average absolute deviation	-0.04	0.70	-2.56	-1.89	0.00	1.53	2.06
Wn1_w_lmed2_wf	moving average absolute deviation	0.07	1.04	-2.67	-2.00	0.00	2.67	3.83
Wschdeck_w_lmed2_wf	moving average absolute deviation	-0.09	2.67	-12.00	-5.92	-0.08	7.33	10.08
Wschoen_w_l_wf	moving average absolute deviation	10.08	3.04	0.00	0.00	10.00	16.67	18.50
yer_pl	percentage	2.49	1.05	0.09	0.19	2.37	4.06	4.06
yer_plmea2	moving average percentage deviation	-0.01	0.85	-1.54	-1.38	-0.12	1.45	1.53
<i>group = 1: 1844 subjects, 247 failures, 21402 observations</i>								
interest_spread_lmed3	moving average absolute deviation	-0.03	0.96	-2.46	-2.45	0.11	1.50	1.69
interest_spread_l	absolute	1.35	1.16	-1.55	-1.29	1.65	3.05	3.05
plzhighlow	binary	0.71	0.45	0.00	0.00	1.00	1.00	1.00
rat_pcr_pyr_lmed2	moving average ratio deviation	0.55	1.36	-2.84	-2.67	0.47	3.07	3.50
rat_pcr_pyr_l	ratio	92.06	1.67	88.43	88.65	91.88	94.85	94.85
urx_l	percentage	3.93	0.66	1.07	2.61	3.94	5.17	5.17
Wn1_s_lmed3_sf	moving average absolute deviation	-0.01	0.64	-2.56	-1.61	0.00	1.44	2.22
Wn1_w_lmed2_wf	moving average absolute deviation	0.12	1.05	-3.42	-2.00	0.00	2.67	3.83
Wschdeck_w_lmed2_wf	moving average absolute deviation	-0.14	2.55	-17.17	-6.00	0.00	6.33	11.08
Wschoen_w_l_wf	moving average absolute deviation	9.96	3.21	0.00	0.00	10.17	17.00	20.33
yer_pl	percentage	2.54	1.02	0.09	0.19	2.57	4.06	5.13
yer_plmea2	moving average percentage deviation	0.02	0.85	-1.62	-1.37	-0.06	1.44	1.53

Table 6: Descriptive Statistic for relevant Variables, by group

Variable	Description	Variants
<i>Macro Data</i>		
interest_spread	Spread between long term and short term interest rates, real	_l, _lmed2, _lmed3, _lmea2, _lmea3
rat_pcr_pyr	Ratio private consumption and disposable income, real	_l, _lmed2, _lmed3, _lmea2, _lmea3
urx	unemployment rate	_l, _lmed2, _lmed3, _lmea2, _lmea3
yer	GDP expenditure, real	_pl, _plmed2, _plmed2
<i>Weather Data</i>		
Wschoen_w	Days with nice weather (cc<50% per day), average over months	_l_wf, _lmed2_wf, _lmed3_wf, _lmea2_wf, _lmea3_wf
Wschoen_s	Days with nice weather (cc<50% per day), average over months	_l_sf, _lmed2_sf, _lmed3_sf, _lmea2_sf, _lmea3_sf
Wschdeck_w	Days with snow, average over months	_l_wf, _lmed2_wf, _lmed3_wf, _lmea2_wf, _lmea3_wf
Wn1_s	Days with rain, average over months	_l_sf, _lmed2_sf, _lmed3_sf, _lmea2_sf, _lmea3_sf
Wn1_w	Days with rain, average over months	_l_wf, _lmed2_wf, _lmed3_wf, _lmea2_wf, _lmea3_wf
<i>Macro Data</i>		
Mfhs_sf	Number of firms, all categories, summer	log and lagged
Mfhw_wf	Number of firms, all categories, winter	log and lagged

Table 7: Variables

Abbreviations: _l = lagged one period, _lmed2 = lagged two year moving average of deviation from the median, _lmed3 = as _lmed2 but three years, _lmea2 = as _lmed2 but with mean, _lmea3 = as _lmea2 but three years, _pl = percentage change, lagged, _plmed2 = as _lmed2 but percentage change, _plmea2 = as _plmed2 but with mean, _wf = winter firm: season November to April, _sf = summer firm: season May to October

Variable	overall					group = 0					group = 1				
	pvalue	NoS	NoO	NoF	propp	pvalue	NoS	NoO	NoF	propp	pvalue	NoS	NoO	NoF	propp
	Macro Data														
interest_spread_l	0.00	2125	26563	344	0.05	0.02	142	1826	48	0.43	0.00	1983	24737	296	0.09
interest_spread_lmed3	0.00	2092	26484	343	0.76	0.24	142	1826	48	0.95	0.00	1950	24658	295	0.77
rat_pcr_pyr_l	0.01	2125	26563	344	0.00	0.06	142	1826	48	0.87	0.00	1983	24737	296	0.00
rat_pcr_pyr_lmed2	0.12	2101	26512	343	0.67	0.00	142	1826	48	0.89	0.80	1959	24686	295	0.97
urx_l	0.00	2125	26563	344	0.00	0.00	142	1826	48	0.33	0.00	1983	24737	296	0.00
yer_pl	0.00	2101	26512	343	0.00	0.00	142	1826	48	0.01	0.02	1959	24686	295	0.01
yer_plmea2	0.05	2092	26484	343	0.02	0.10	142	1826	48	0.44	0.17	1950	24658	295	0.03
	Weather Data														
Wschoen_w_l_wf	0.00	1982	23930	298	0.61	0.04	137	1611	43	0.20	0.00	1845	22319	255	0.39
Wschdeck_w_lmed2_wf	0.00	1987	23903	303	0.16	0.09	136	1616	41	0.59	0.00	1851	22287	262	0.17
Wn1_w_lmed2_wf	0.05	2010	24404	310	0.64	0.34	138	1655	43	0.71	0.12	1872	22749	267	0.63
Wn1_s_lmed3_sf	0.00	2011	24222	306	0.52	0.00	139	1606	43	0.07	0.00	1872	22616	263	0.33
	General Data														
plzhighlow	0.38	2073	26429	339	0.72	0.22	142	1826	48	0.56	0.54	1931	24603	291	0.59
group	0.00	2180	26665	352	0.01	-	-	-	-	-	-	-	-	-	-

Table 8: Selected Variables as described in Modeling, 4.2

NoS = number of subjects, NoF = number of firms, pvalue = significance (p) value of $\hat{\text{HR}}$ (t-statistic), propp = univariate significance (p) value of the variable to fulfill the proportionality assumption: H_0 : the variable fulfills the proportionality assumption (distribution: $\chi^2(1)$)

model	$g0_t$			$g0_{\ln(t)}$			$1g1_t$			$1g1_{\ln(t)}$		
	$\bar{\mu}$	l	h	$\bar{\sigma}$	$\bar{\mu}$	l	h	$\bar{\sigma}$	$\bar{\mu}$	l	h	$\bar{\sigma}$
	Macro Data											
interest_spread_l					1.35	-1.55	3.05	1.11				
interest_spread_lmed3									-0.03	-2.46	1.69	0.89
rat_pcr_pyr_l												
rat_pcr_pyr_lmed2	0.48	-2.84	3.50	1.40	0.48	-2.84	3.50	1.40				
urx_l					3.98	3.14	4.79	0.60				
yer_pl									2.54	0.09	5.13	1.03
yer_plmea2	-0.01	-1.54	1.53	0.88	-0.01	-1.54	1.53	0.88				
	Weather Data											
Wschoen_w_l_wf	10.10	0.00	18.50	1.81					9.99	0.00	20.33	1.79
Wschdeck_w_lmed2_wf									-0.05	-17.17	11.08	1.34
Wn1_w_lmed2_wf												
Wn1_s_lmed3_sf									0.01	-2.56	2.22	0.58
	$h(\bar{dl})$	$h(\bar{dh})$	$\bar{\sigma}-$	$\bar{\sigma}+$	$h(\bar{dl})$	$h(\bar{dh})$	$\bar{\sigma}-$	$\bar{\sigma}+$	$h(\bar{dl})$	$h(\bar{dh})$	$\bar{\sigma}-$	$\bar{\sigma}+$
	Macro Data											
interest_spread_l					237%	51%	59%	37%				
interest_spread_lmed3									69%	129%	35%	54%
rat_pcr_pyr_l												
rat_pcr_pyr_lmed2	74%	240%	43%	76%	68%	182%	38%	62%				
urx_l					71%	83%	26%	36%				
yer_pl									5%	5%	2%	2%
yer_plmea2	56%	131%	38%	61%	57%	133%	38%	62%				
	Weather Data											
Wschoen_w_l_wf	73%	201%	21%	27%					54%	122%	13%	15%
Wschdeck_w_lmed2_wf					792%	76%	19%	16%				
Wn1_w_lmed2_wf									792%	76%	19%	16%
Wn1_s_lmed3_sf									57%	38%	14%	12%
					6%	9%	2%	2%	38%	91%	16%	19%

Table 9: Economic Size Variables

Abbreviations: $\bar{\mu}$ = mean of firms, l = overall min of firms, h = overall max of firms, $\bar{\sigma}$ = mean of standard dev. of firms
 $h(\bar{dl})$ = hazard of \bar{l} firm compared with $\bar{\mu}$ firm, $h(\bar{dh})$ = hazard of \bar{h} firm compared with $\bar{\mu}$ firm, $\bar{\sigma}-$ = hazard of $\bar{\mu}$ firm if one $\bar{\sigma}$ subtracted, $\bar{\sigma}+$ = analogous to $\bar{\sigma}-$ if $\bar{\sigma}$ is added, $\%$ = . % smaller hazard, $\%$ = . % larger hazard