DE GRUYTER OPEN

Folia Oeconomica Stetinensia
DOI: 10.1515/foli-2015-0027

WYDZIAŁ NAUK EKONOMICZNYCH I ZARZĄDZANIA
UNIWERSYTET SZCZECIŃSKI

DURATION MODELS IN LOAN MANAGEMENT

Julian A. Vasilev, Ph.D. Associate Professor

Varna University of Economics
Department of Informatics
Knyaz Boris I, 77, Varna, Bulgaria
e-mail: vasilev@ue-varna.bg

Received 5 January 2015, Accepted 27 April 2015

Abstract

The purpose of this study is to estimate the future duration of a loan contract on the basis of several factors. The main methodology consists of a brief explanation of a survival analysis and a thorough application of a survival analysis in loan management. A real dataset from a credit institution (situated in Varna) is used. All contracts were signed for 30 days but some contracts were ended earlier, others – later. The main research question concerns the following statement. We may try to predict future loan duration by making an econometric model describing the dependency between the loan duration (as a dependent variable) and several independent variables. The dataset is analysed by calculating life tables, applying the Kaplan-Maier method and using Cox regression within SPSS. It is has been proved that the main covariates affecting loan duration are the variables: born in the region, month of birth and age. The formulated conclusions are valid for the analysed credit institution. This work provides a methodology for adapting duration models in credit institutions. The presented methodology (in this paper) may be applied over the dataset of other credit institutions (including banks) for loan duration prediction.

Keywords: survival analysis, SPSS, the Wilcoxon test, the Kaplan-Maier method, Cox regression.

JEL classification: C12, C52, C63, C87.

Introduction

Duration models are a branch of statistics where special kinds of events are studied. An event happens. The expected time to happen in an event is the duration. Life expectancy at birth is studied by duration models. Popular terms are: duration models, duration modelling, event history analysis. After several times of signing contracts, some of them "survive", others are finished. The "lifetime" of a contract is an interesting study. Another branch of statistics studies are probability models. The probability of an event to happen is studied. Duration models try to predict the duration (the life) of a contract (from the date of signing until the date of finishing, if a contract is finished).

A survival function (in medicine) or reliability function (in mechanical engineering) are studied. In statistics the hazard function and cumulative hazard functions are described. Future lifetime and expected future lifetime are calculated. Left and right censoring are used to solve problems in survival analysis. The Kaplan-Maier estimator is one of the estimators for estimating the survival function. In this study this method is used to calculate the duration of a loan.

Loan management concerns different issues. Credit institutions (including banks) have datasets. These datasets contain data about loans. In most cases credit institutions just register transactions concerning given loans and returned loans. A further analysis may be made in the context of creating customer profiles (Vasilev, 2014a). In this article it has been proved that the possibility of returning a loan depends on the sum of the contract, gender, age, year and month of signing the contract. A time series analysis is another technique that may be applied over a dataset from a credit institution. By using a time series analysis it can be stated (Vasilev, 2014b) that incoming money flows in credit institutions do not depend on the amount of given loans. The probability of returning a loan is studied by binary probability models (Vasilev, 2015). Calculating the probability of returning a loan is a difficult task. The author (Vasilev, 2015) assumes that specific data fields concerning the contract (month of signing, year of signing, given sum) and data fields concerning the borrower of the loan (month of birth, year of birth (age), gender, region, where he/she lives) may be independent variables in a binary logistics model with a dependent variable "the probability of returning a loan". It has been proved that the month of signing a contract, the year of signing a contract, the gender and the age of the loan owner do not affect the probability of returning a loan. It has been proved (Vasilev, 2015) that the probability of returning a loan depends on the sum of the contract, the remoteness of the loan owner and the month of birth. The probability of returning a loan increases with the increase

of the given sum, decreases with the proximity of the customer, increases for people born in the beginning of the year and decreases for people born at the end of the year.

All previous studies concerning analysing the dataset of credit institutions focus on a time series analysis, binary probability models, one-factor analysis, and a correlation analysis. We assume that a survival analysis may be used for the prediction of future loan durations.

1. The essence of survival analysis

Survival analysis is applied when the time-to-event is interesting for a researcher. Some popular examples are life expectancy at birth and marriage duration. Usually different types of factors influence the monitored event. Survival analysis in most cases focuses on a hazard function. An event is expected to happen. In loan management the event is finishing a contract. Although all contracts in the studied credit institution are signed for 30 days, some of them finish on time, others – earlier, others – later. Some of the contracts do not finish at all. The hazard rate (Bian, 2015) is the instantaneous probability of a given event occurring at any point in time. It can be plotted against time on the X axis, forming a graph of the hazard rate over time. The hazard function describes this plotted line. The hazard ratio is called "relative risk" in SPSS.

Using empirical data we may make a graph of the hazard function. Depending on the shape of the line, the analysis may be nonparametric, semi-parametric or non-parametric. The distribution of the time to event variables is described by life tables. Life tables divide the observed period into small intervals of time. The probability for each "small" interval is calculated. Three types of variables are used: (1) a duration variable, (2) a status variable (usually binary) and (3) a categorical variable. The main assumption of the survival analysis is that the probability of an event (the probability of finishing a loan contract in time) depends only on time. We expect the same behaviour of cases (signed contracts) in different "small" periods of time. This is a basic assumption that has to be proved or rejected.

2. Analysing the dataset in SPSS

The dataset contains 726 records from a credit institution, situated in the Varna region. We have data for the whole population of the studied credit institution. The fields of the dataset may be grouped into the following two groups: (1) information about the contract: month of signing, sum of contract, finished (a Boolean variable), duration (in days) and (2) information about the borrower: born in the region (a Boolean variable), month of birth, gender and age.

The main research question concerns the following statement. We may try to predict future loan duration by making an econometric model (a duration) describing the dependency between the loan duration (as a dependent variable) and several independent variables. The dataset is analysed by calculating life tables, applying the Kaplan-Maier method and using Cox regression within SPSS.

2.1. Calculating life tables

These fields are enough to make the survival analysis (Figure 1) in SPSS (Analyze/ Survival/Life Tables).

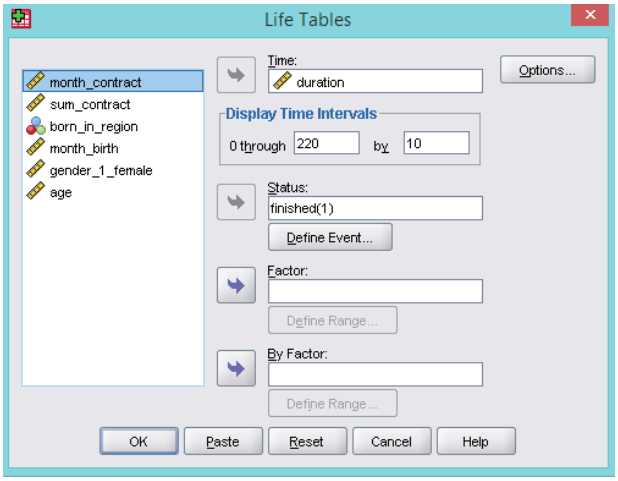

Fig. 1. Generating Life Tables in SPSS

Source: screenshot in SPSS.

The "duration" variable is put into the time field. Time intervals are from 0 to 220 with a step of 10. The status variable is "finished".

Interval start time (Table 1) is the beginning of each interval. The "number of terminal events" shows the number of contracts that are finished. When we look at the survival table in the SPSS output we may conclude that with the increase of the duration of a contract decreases the number of terminal events. "Proportion terminating" is the ratio of terminal events to the number exposed to risk. "Proportion surviving" is one minus "proportion terminating". "Probability density" is the probability of meeting the terminal event (finishing the contract) during the interval. The probability of finishing a contract within 50 days is 53% per cent.

The concentration of terminal events is in the first intervals of time. It is obvious that there is a negative correlation between the interval start time and the number of terminal events (Figure 2).

Table 1. A part of the survival tables in SPSS output

Interval start time	Number exposed to risk	Number of terminal events	Proportion terminating	Proportion surviving	Probability density
0	726	115	0.16	0.84	0.16
10	611	77	0.13	0.87	0.11
20	534	67	0.13	0.87	0.09
30	467	23	0.05	0.95	0.03
40	444	18	0.06	0.94	0.04
50	416	16	0.04	0.96	0.02
60	400	12	0.03	0.97	0.02
70	388	5	0.01	0.99	0.01
80	383	8	0.02	0.98	0.01
90	375	0	0.00	1.00	0.00
100	375	0	0.00	1.00	0.00
110	374	1	0.00	1.00	0.00
120	365	1	0.00	1.00	0.00

Source: own calculation in SPSS.

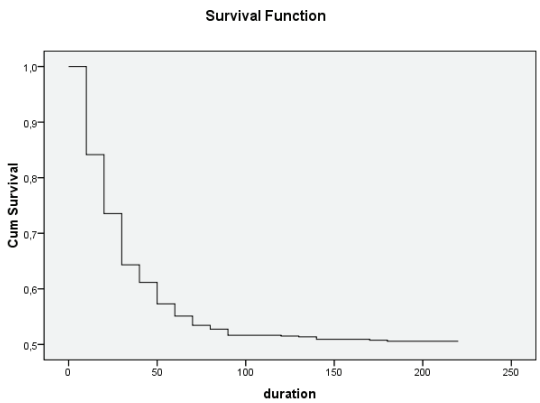

Fig. 2. Survival function

Source: own calculation in SPSS.

Life tables may be extended by adding a factor which may influence the loan duration. We may check if there are significant differences in the survival function between male and female.

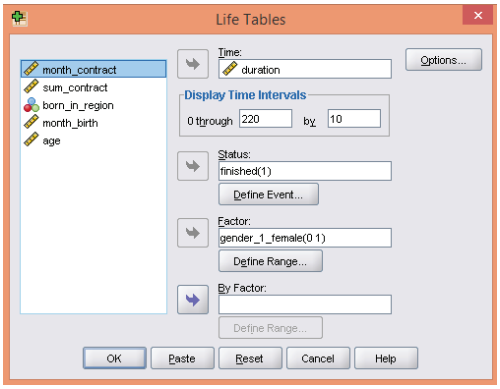

Fig. 3. Generating Life Tables in SPSS with factor "gender"

Source: screenshot in SPSS.

The Wilcoxon test is used to compare distributions of the survival function among males and females. The significance value of the test (0.124) is greater than 0.05. It means that the survival curves are not significantly different among males and females (Figure 4).

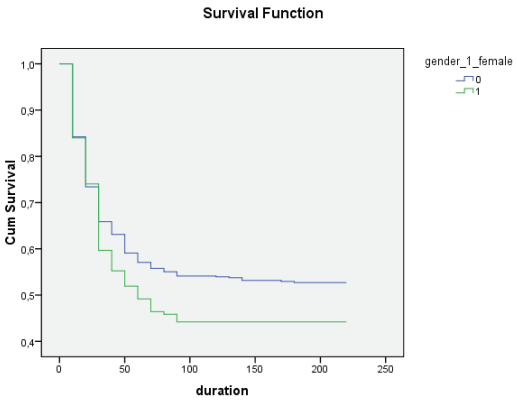

Fig. 4. Survival functions for male and female

Source: own calculation in SPSS.

Even though there are not significant differences among males and females (in terms of loan duration), it is obvious from the graph that if a loan is expired, the probability to be returned is higher for women than for men. The lower graph (Figure 4) is for female borrowers. For expired contracts with a passed duration of more than 50 days, more contracts with a male borrower "survive" – they are not finished. The same conclusion may be made by looking at

the second survival table, provided in the SPSS output. Women are more willing to return their overdue loans than men.

A similar check may be made for the factor "born in the region". The Wilcoxon test (we have checked "Pairwise" in the "Compare levels of first factor" panel in the options menu) is used to compare distributions of the survival function among people born in the region and people not born in the region. The significance value of the test (0.000) is less than 0.05. It means that the survival curves are significantly different among people born in the region and people not born in the region.

Table 2. Survival table with grouping variable "born in the region"

First order controls	Interval start time	Number exposed to risk	Number of terminal events	Proportion terminating	Probability density
Born in region "0" (no)	0	199	17	0.09	0.009
	10	182	16	0.09	0.008
	20	166	16	0.10	0.008
	30	150	5	0.03	0.003
	40	145	11	0.08	0.006
	50	134	5	0.04	0.003
	60	129	2	0.02	0.001
	70	127	2	0.02	0.001
	80	125	3	0.02	0.002
Born in region "1" (yes)	10	527	98	0.19	0.019
	20	429	61	0.14	0.012
	30	368	51	0.14	0.010
	40	317	18	0.06	0.003
	50	299	17	0.06	0.003
	60	282	11	0.04	0.002
	70	271	10	0.04	0.002
	80	261	3	0.01	0.001

Source: own calculation in SPSS.

It is obvious (Table 2) that the number of terminal events decreases with the increase of the loan duration. The "proportion terminating" ratio differs significantly for people born in the region and people not born in the region. People born in the region have higher values of "proportion terminating" of loans compared in pairwise intervals of duration. We may conclude that people born in the region of the credit institution are more willing to return their loans than people born in remote places.

2.2. The Kaplan-Maier method

The Kaplan-Maier method is a method to compare the distribution of time-to-event variables by a grouping factor. The probability of finishing a contract should depend only on the time after signing a contract without any covariates effects. Contracts signed in different months of the year should "behave" similarly. As a time variable we use the "duration" variable. As a status variable we use the "finished" variable. For stratification we use a categorical variable – the "gender" variable. The null hypothesis states that the survival function is the same between male and female borrowers. The analysis in made in SPSS (Analyze/Survival/Kaplan-Maier).

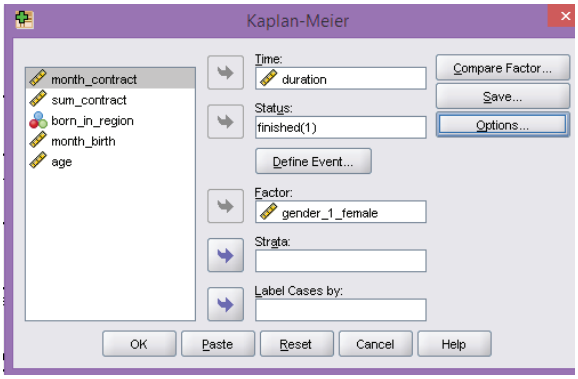

Fig. 5. Kaplan-Maier parameters – step one

Source: screenshot in SPSS.

From the "Compare Factor" button we chose: Log rank, Breslow and Tarone-Ware statistics. From the "Options" button we chose: Survival tables, Mean median survival, survival plots and hazard plots. SPSS gives an output with several tables. Some of them are very large. We summarized the results. The total number of contracts is 726 (181 for female and 545 for male borrowers). 101 female borrowers finished their contracts (55.8% of women finished their contracts). 257 male borrowers finished their contracts (47.2% of men finished their contracts). A fact we proved before. The overall comparison table provides tests of the equality (or inequality) of survival times (duration of contracts) among groups (males and females). The Log Rank (Mantel-Cox) test, the Breslow (Generalized Wilcoxon) test and the Tarone-Ware test have significance over 0.05. It means that there is no statistically significant difference in survival time (the duration of loans) between males and females.

Another check is made. Now the factor is the variable "born in the region" (Figure 6).

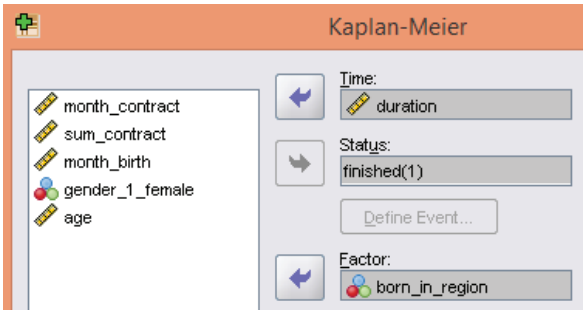

Fig. 6. Kaplan-Maier parameters – step two

Source: screenshot in SPSS.

The Log Rank (Mantel-Cox) test, the Breslow (Generalized Wilcoxon) test and the Tarone-Ware test have significance below 0.05. It means that there is a significant difference in survival time (the duration of loans) between people born in the region and people not born in the region. The overall mean is 210 days with a standard error of 6.831. The mean for people born in the region is 195 days with a standard error of 8.033. The mean for people not born in the region is 249 days with a standard error of 12.575. We may conclude that people born in the region return their loans in a shorter time, than people born in remote places.

2.3. Cox regression

Cox regression is used for predicting an event on the basis of the values of several factors (covariates). Cox regression tries to find a linear relationship between survival times and factors (Analyze/Survival/Cox Regression). In the time filed we have put "duration", in the status field "finished", in the covariates list: month of contract, sum of contract, born in the region, month of birth, gender and age. We have put "gender" and "born in the region" as categorical

Table 3. SPSS output of Cox regression – version 1

Variables in the Equation								
	B	SE	wald	Df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							lower	upper
month_contract	–0.030	0.019	2.426	1	0.119	0.971	0.935	1.008
sum_contract	0.000	0.000	2.979	1	0.084	1.000	1.000	1.000
born_in_region	–0.425	0.130	10.665	1	0.001	0.654	0.507	0.844
month_birth	–0.043	0.016	7.533	1	0.006	0.958	0.928	0.988
gender_1_female	–0.231	0.118	3.821	1	0.051	0.794	0.629	1.001
age	–0.016	0.006	7.793	1	0.005	0.984	0.973	0.995

Source: own calculation in SPSS.

variables. We have set the field "CI for exp(B)" to 95% in the "Model statistics panel" of the Cox regression options.

It is clear that the sum of the contract does not affect the duration (the B value is zero). The month of contract, the sum of contract and gender have a p-value greater than 0.05. It means that these variables should not be included in a Cox regression model. Statistically significant covariates ($p < 0.05$) are: born in the region, month of birth and age. We used the Cox regression again including as covariates only those covariates with $p < 0.05$ (Table 4).

Table 4. SPSS output of Cox regression – version 2

Variables in the Equation								
	B	SE	wald	df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							lower	upper
born_in_region	–0.396	0.129	9.382	1	0.002	0.673	0.522	0.867
month_birth	–0.044	0.016	7.703	1	0.006	0.957	0.928	0.987
age	–0.014	0.006	5.937	1	0.015	0.986	0.976	0.997

Source: own calculation in SPSS.

Exp(B) is interpreted as the predicted change in the hazard for a unit increase in the predictor. Born in the region is a binary covariate. The value of Exp (B) for "born in the region" means that the probability of finishing a loan for a person not born in the region is 0.673 times less than that of a person born in the region. The Exp(B) value of "month of birth" is 0.957. It means that the probability of finishing a loan decreases with 100% – (100% × 0.957) = 4.3% for each next month of the year a borrower is born. The Exp(B) value of "age" is 0.986. It means that the probability of finishing a loan decreases with 100% – (100% × 0.986) = 1.4% for each year a borrower has lived.

Numeric examples were created in MS Excel to demonstrate the results from the research (Figure 7). The employee in the credit institution enters the variables: born in the region, month of birth and age. As a result he/she has information about the probability of finishing the loan.

Comparison 1 shows that people born in the region have a higher probability of returning loans than people not born in the region. For instance a borrower (not born in the region, born in January, 20 years old) has a 58% probability of returning a loan. For instance a borrower (born in the region, born in January, 20 years old) has a 67% probability of returning his/her loan.

Comparison 2 is for two borrowers born in the region at the age of 20. The first one is born in February. His probability of returning a loan is 68%. The second one (born in March) has a probability of returning a loan of 69%. People born at the end of the year are more willing to return their loans than people born in the beginning of the year.

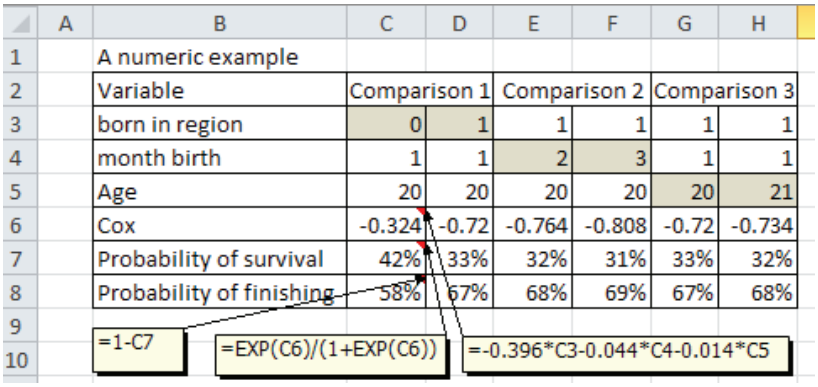

Fig. 7. Numeric examples in MS Excel

Source: own calculation in MS Excel.

Comparison 3 is for two borrowers born in the region, born in January, the first one is 20 years old, the second one – 21 years old. The probability of finishing the loan for the first one is 67%, for the second one – 68%. Older people are more willing to return their loans than younger ones.

Conclusions

Duration models have wide applications. This study focuses on the application of duration models in loan management. The duration models in this study are used for the prediction of loan duration. The Kaplan-Maier method was used to estimate the survival function. Life tables are used. The main assumption of the survival analysis is that the probability of an event (the probability of finishing a loan contract in time) depends only on time.

The initial research questions concerns the idea that the loan duration depends on several covariates. The main assumption is the following. Duration models may describe formally the duration of a loan (the "life" of a contract) with several factors. Using data from a real dataset and statistical methods the initial research assumption is proved.

Several hypotheses (by using life tables within survival analysis) are proved. With the increase of the duration of a contract decreases the number of terminal events. The survival curves are not significantly different among male and female borrowers. If a loan is expired, the probability to be returned is higher for women than for men. Women are more willing to return

their overdue loans than men. The survival curves are significantly different among people born in the region and people not born in the region.

Applying the Kaplan-Maier method we have proven that the duration of loans does not differ significantly between male and female borrowers. There is a significant difference in survival time (the duration of loans) between people born in the region and people not born in the region. People born in the region return their loans in a shorter time, than people born in remote places.

Using the Cox regression it has been proved that the month of contract, the sum of contract and gender are not factors in a Cox regression model. Statistically significant covariates are: born in the region, month of birth and age. A numeric example in MS Excel is given to illustrate the Cox regression and the relative importance of each covariate. It has been proven that:

(1) People born at the end of the year are more willing to return their loans than people born at the beginning of the year;
(2) People born in the region return their loans in a shorter time, than people born in remote places;
(3) People born in the region of the credit institution are more willing to return their loans than people born in remote places;
(4) Older people are more willing to return their loans than younger ones.

Future research may extend the current paper. This paper gives an example for life tables with a grouping variable "gender". Future research may focus on calculating life tables with other grouping variables – for instance: "sum of contract" or "age group".

References

Bian, H. (2015). *Survival analysis using SPSS*, http://core.ecu.edu/ofe/StatisticsResearch/Survival%20Analysis%20Using%20SPSS.pdf(19.05.2015).

Garth A. (2008). *Analysing data using SPSS*. Sheffield Hallam University, https://students.shu.ac.uk/lits/it/documents/pdf/analysing_data_using_spss.pdf (19.05.2015).

Gujarati D. (2004). *Basic econometrics. 4th edition*. The McGraw-Hill Companies, http://egei.vse.cz/english/wp-content/uploads/2012/08/Basic-Econometrics.pdf (19.05.2015).

Vasilev, J. (2014). Creating a customer profile in a credit institution. *International Journal of Advanced Research in Computer Science and Software Engineering*, 4 (1): 1108–1112, www.ijarcsse.com/docs/papers/Volume_4/1_January2014/V4I1-0564.pdf (19.05.2015).

Vasilev, J. (2014). Time series analysis in loan management systems. *Theoretical and Applied economics*, 21 (3): 57–66, http://store.ectap.ro/articole/962.pdf (19.05.2015).

Vasilev, J. (2015). Calculating the probability of returning a loan with binary probability models. *Romanian Statistical Review*, 4: 55–71, www.revistadestatistica.ro/wp-content/uploads/2015/01/RRS_04_2014_A5.pdf (19.05.2015).