Far East Journal of Theoretical Statistics
Volume 40, Number 2, 2012, Pages 107-125
Published Online: October 2012
Available online at http://pphmj.com/journals/fjts.htm
Published by Pushpa Publishing House, Allahabad, INDIA

MODELLING CREDIT RISK FOR PERSONAL LOANS: COX PROPORTIONAL HAZARDS MODEL APPROACH

Okumu Argan Wekesa, Mwalili Samuel and Mwita Peter

Department of Actuarial Science and Statistics
Jomo Kenyatta University of Agriculture and Technology
P. O. Box 62000-00200, Nairobi, Kenya

Abstract

A proportional hazards model approach was adopted to estimate risk of default for loan applicants. A sample of 500 applicants was observed for 36 months. The life of the account is measured from the month, it was opened until the account becomes 'bad' or it is closed or until the end of observation. The account is considered bad if payment is not made for two consecutive months in line with the industry practice. If the account does not miss two payments and is closed or survives beyond the observation period, then it is considered to be censored. The results showed that gender, employment sector and level of education are not significant in credit risk modelling. However, marital status, age, home ownership and duration at residence were found to be significant.

1. Introduction and Literature Review

Background to the study

Traditional credit risk models aim at determining a customer's

2010 Mathematics Subject Classification: 62P05.

Keywords and phrases: survival analysis, Cox proportional hazards model, default risk.

Received September 23, 2011

probability of defaulting on loan repayment. This study used survival analysis method which draws its origin from the study of time to death or occurrence of any other event on life data. The technique has also been applied in engineering to model failure time of components and parts. In credit risk modelling, the event of interest is default, thus modelling time to default on loan obligations.

Modelling of credit risk using survival analysis was first introduced by Narain [32]. Thomas et al. [34] further developed the model. Narain [32] applied the survival model on 24 months of loan data. The result showed that the survival analysis approach provides more detailed and relevant information for credit management than the conventional approaches. Thomas et al. [34] applied the technique by using the accelerated life exponential model to 24 months loan data. The results also showed the approach to be superior to conventional credit scoring methods in that a better credit-granting decision could be made if the score was supported by the estimated survival times. The research by Thomas et al. [34] also did a comparison of exponential, Weibull and Cox non-parametric models with logistic regression and they concluded that survival method was a better modelling tool.

In the literature, a number of techniques have been applied to model credit risk. Orgler [33] applied regression analysis in a model for commercial loans. Wiginton [35] was one of the first to publish credit scoring results using logistic regression. It was compared with discriminant analysis. Leonard [26] also applied logistic regression in evaluating commercial loans. Durand [14] pioneered the use of discriminant analysis for credit scoring. Decision tree and rule was adopted by Makowski [27] and Mehta [30] for credit scoring.

Other techniques include K-Nearest Neighbour Classifiers which were used by Chatterjee and Barcun [8] and Henley and Hand [21]. Baesens et al. [4] studied the use of Bayesian network classifiers to rate borrowers. Linear programming was applied by Hardy and Adrian [19] and compared it with other statistical approaches.

Statement of the problem

During the period between year 2003 and 2010, the Kenyan financial market experienced growing liquidity, which caused banks to rigorously market various loan products. This gave rise to the need to review the banks' credit-granting criteria to reflect the growing volume of loan portfolio and to respond to the year 2008 global credit crunch. However, research on credit risk has surprisingly received insignificant attention from both practitioners and scholars in Kenya. Over the years, banks perpetually used traditional credit scoring techniques to rate loan applicants.

A number of studies have been carried out on the issue of credit risk modelling using different approaches. A limited number of studies worldwide have applied survival analysis techniques but none has been applied on the Kenyan or even Eastern African market to analyse credit risk. To this end, the research is intended to model probability of loan default and hazard rates for various risk groups using Cox Proportional Hazards Model. Furthermore, existing credit scoring models classify borrowers into different risk categories but cannot provide any information on when the borrower is likely to default. It is more informative for the lender not only to know the probability of defaulting, but also when the default is likely to happen. This helps to fairly price risks and improves the focus on ultimate profitability. For instance, if the lender knows that a group of loan applicants are bad type, instead of rejecting their applications, then it may grant loans to them at higher interest rate, as long as the term of the loan is shorter than the likely time to default. Thus some "bad" applicants can also be viewed as profitable propositions.

General objective

The broad objective of this research is to use Cox Proportional Hazards Model to generate default probabilities at various points in time. The study also intended to perform significance tests for the covariates captured by the model.

Specific objectives

The specific objectives included:

1. To generate the regression coefficient for each covariate.
2. To generate a correlation matrix of regression coefficients.
3. To test the statistical significance of the each covariate in the resulting Cox Model.
4. To provide a statistical summary for each covariate.

Descriptive methods of time-to-event

Survival analysis is a statistical method for modelling the time to some events for a population of individuals. For example, events may refer to death in medical application, or recidivism of released prisoners in criminology application, or first bought of a new product by customer in marketing studies. The time to the occurrence is termed as survival time or lifetime. In application to credit risk modelling, the events refer to default of a loan and therefore its lifetime refers to time-to-default T.

Default times are subjected to random variation and are thus random variables. To describe their randomness, there are five standard ways:

Density function (PDF), $f_{20}(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T \le t + \Delta t)}{\Delta t}$,

Distribution function (CDF), $F_{20}(t) = \int_0^t f_{20}(u)du = \Pr(T \le t)$,

Survivor function, $S_{20}(t) = \int_t^\infty f_{20}(u)du = 1 - F_{20}(t) = \Pr(T \ge t)$,

Hazard function

$$h_{20}(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \le T \le t + \Delta t \backslash T \ge t)}{\Delta t} = \frac{f_{20}(t)}{S_{20}(t)} = \frac{-d \ln S_{20}(t)}{dt},$$

Cumulative hazard function $H_{20}(t) = \int_0^\infty h_{20}(t)dt = -\ln S_{20}(t)$.

These five formulations are mathematically equivalent but they highlight different aspects of the default time. The distribution function tells us the

probability that default occurs at or before time *t*. Conversely, survivor function is the probability that default does not occur at or before time *t*; in other words, the loan survives (non-default), at least, to time *t*. The interpretation of hazard function is slightly tricky. It is the "rate" that borrower defaults at time *t*, conditional on his staying on the books up to that time. Note that hazard is not a probability and thus can be greater than one.

In survival analysis, we must consider a key analytical problem called *censoring*. In essence, censoring occurs when we have some information about an individual's survival time, but do not know the exact survival time. There are a number of types of censoring, such as random, interval, left and right censoring. In credit scoring application, most of the cases are right censoring.

For example, suppose we follow a group of borrowers for 3 years. If we observe borrower *A* fails to repay at 15th month, then he is certainly classified as a default case and his default time is 15. On the other hand, consider borrower *B*, who repays on time during the whole observed period. We do not know his exact default time but are sure that it must be greater than 36. For such case, borrower *B* is known as a right censored observation. Another example of right censoring could be when borrower *C* repays on time from the 1st month to the 12th month. At the 12th month, we do not have future repayment pattern of borrower *C*. As borrower *B*, we do not know the exact default time of borrower *C*, we only know that it must be greater than 12. This is also a right censoring example.

The Cox proportional hazards model

Introduction

This technique is adopted to assess effect of multiple covariates on survival. It is the most commonly used multivariate survival method.

Model description and derivation

Consider the experimental distribution with parameter λ such that

$$f(t) = \lambda e^{-\lambda t}, \quad t > 0.$$

It can be shown that the hazard function

$$h(t) = \lambda,$$

λ can be estimated using the maximum likelihood method.

Alternatively, we can consider λ to be dependent on a variable X, say. We call X a *covariate* of λ.

For example, we can express λ as follows:

1. $\lambda = \beta_0 + \beta_1 X, \ \lambda > 0$

2. $\frac{1}{\lambda} = \beta_0 + \beta_1 X, \ \lambda > 0.$

In these two models, we can have situations, where $\lambda < 0$ depending on the values of X. And yet we have been given that $\lambda > 0$.

A better model for which $\lambda > 0$ is

$$\log_e \lambda = \beta_0 + \beta_1 X \Rightarrow \lambda = e^{\beta_0 + \beta_1 X} > 0, \quad -\infty < X < \infty.$$

Instead of considering only one covariate, we can take $X_1, X_2, X_1, ..., X_P$ such that

$$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + B_P X_P}.$$

In the matrix form,

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_P X_P = (\beta_0 \quad \beta_1 \quad \beta_2 \quad \beta_3 \cdots \beta_P) \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_P \end{pmatrix}.$$

We can also express

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_P X_P$$

as

$$\beta_0 + (\beta_1 \quad \beta_2 \quad \beta_3 \cdots \beta_P) \begin{pmatrix} 1 \\ X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_P \end{pmatrix} = \beta_0 + \underline{\beta}' \underline{X},$$

where

$$\underline{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_P \end{pmatrix}$$

and

$$\underline{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_P \end{pmatrix},$$

$\underline{\beta}$ and $\underline{X}$ and are column vectors,

$\beta' =$ Transpose of $\underline{\beta}$ which is $(\beta_1 \quad \beta_2 \quad \beta_3 \cdots \beta_P)$.

Therefore,

$$\lambda = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_P X_P}$$

$$= e^{\beta_0 + \underline{\beta}' \underline{X}}$$

$$= e^{\beta_0} e^{\underline{\beta}' \underline{X}}.$$

Also, we can write

$$\lambda = \lambda_0 e^{\underline{\beta}' \underline{X}}.$$

Since λ is the hazard function for the exponential distribution, we can generalize the notion as

$$h(t) = e^{\beta_0} e^{\underline{\beta}' \underline{X}}.$$

Based on the above notion, we have Cox's Proportional Hazards Model given as

$$h(t) = h_0(t) e^{\underline{\beta}' \underline{X}},$$

$$h_0(t) = e^{\beta_0},$$

where

$H_i(t) =$ the hazard function at time t,

$h_0(t) =$ the baseline hazard function at time t,

$\underline{\beta} =$ a vector of parameters,

$\underline{X} =$ a vector of covariates.

Remarks

1. The Cox model assumes a common shape of the hazard for all individuals.
2. $h(t)$ is proportional to $h_0(t)$, i.e., $h(t) = Kh_0(t)$, where K is the constant of proportionality. In this case, the constant of proportionality is $e^{\beta' X}$.
3. Cox's contribution was to figure out how to estimate the β's without specifying $h_0(t)$.
4. Cox proportional hazard model is semi-parametric because $h_0(t)$ is arbitrary or not specific. Hence $h_0(t)$ is non-parametric. The exponential component $e^{\beta' X}$ is parametric.
5. In general, a proportional hazard model can be expressed in the form

$h(t) = h_0(t)\psi(X_1 + X_2 + \cdots + X_p)$, where the multiplier $\psi(\underline{X})$ does not depend on time and it is positive.

6. $S(t) = e^{-\int_0^t h(u)\,du}$

But

$$h(t) = h_0(t)e^{\beta' X} \Rightarrow h(u) = h_0(u)e^{\beta' X},$$

$$S(t) = e^{-\int_0^t [h_0(u)e^{\underline{\beta}'\underline{X}}]du}$$

$$= e^{-\int_0^t [h_0(u)\,du]e^{\underline{\beta}'\underline{X}}}$$

$$= e^{-\int_0^t h_0(u)\,du}\, e^{\underline{\beta}'\underline{X}}$$

$$= S_0(t)e^{\underline{\beta}'\underline{X}}$$

$$= e^{\int_0^t [h_0(u)\,du]e^{\beta' X}}$$

$$S(t) = [S_0(t)]e^{\beta' X}.$$

Construction of partial likelihood

Suppose a sample of size 4 has the following individuals named *A*, *B*, *C* and *D*. Let their failure times be t_A, t_B, t_C and t_D, respectively.

Suppose $t_C < t_D < t_A < t_B = t_1 < t_2 < t_3 < t_4$.

Just before the occurrence of the first failure, all four individuals were at risk. In other words, the risk set at t_1 was:

$$R(t_1) = \{A, B, C, D\}.$$

Similarly, the risk set at t_2, t_3 and t_t was:

$$R(t_2) = \{A, B, D\},$$

$$R(t_3) = \{A,\ B\},$$

$$R(t_4) = \{B\}.$$

The probability of C failing at time t_1, given the risk set $R(t_1)$ can be looked at in terms of the hazard $h_c(t)$ with respect to all of them.

$$\text{Thus P(C failing)} = \frac{h_c(t_1)}{h_A(t_1) + h_b(t_1) + h_c(t_1) + h_p(t_1)}$$

If we incorporate the General Proportional Hazards model; i.e., $h(t) = h_0(t)\psi(x)$, then the formula

$$= \frac{h_0(t)\psi_c(t_1)}{h_0(t)\psi_A(t_1) + h_0(t)\psi_0(t_1) + h_0(t)\psi_0(t_1) + \cdots}$$

$$= \frac{\psi_0(t_1)h_0(t)\psi_0(t_1)}{\psi_A(t_1) + \psi_B(t_1) + \psi_c(t_1) + \psi_D(t_i)}$$

$$= \frac{\psi_1}{\psi_1 + \psi_2 + \psi_2 + \psi_4}$$

Probability of individual D defaulting at time t_2 given $R(t_2)$

$$= \frac{\psi_2}{\psi_2 + \psi_3 + \psi_4}$$

Probability of A defaulting at time t_3 given $R(t_3) = \frac{\psi_3}{\psi_3 + \psi_4}$

Probability of B defaulting at time t_t given $R(t_4) = \frac{\psi_4}{\psi_4} = 1$

Partial likelihood is the product

$$= \frac{\psi_1}{\psi_1 + \psi_2 + \psi_3 + \psi_4} \cdot \frac{\psi_2}{\psi_2 + \psi_3 + \psi_4} \cdot \frac{\psi_3}{\psi_3 + \psi_4} \cdot \frac{\psi_4}{\psi_4}$$

$$= \prod_j \frac{\psi_j}{\sum_{j \varepsilon R(t_j)} \psi_j}.$$

In the Cox Proportional Hazards Model:

$$\psi_i = e^{\beta' X_j}.$$

$$\text{Thus partial likelihood} = \prod_j \left[\frac{e^{\beta' X_j}}{\sum_{j \varepsilon R(t_j)} e^{\beta' X_j}} \right].$$

β can be estimated using the maximum likelihood estimation method. If the partial likelihood is denoted by L, then

$$\frac{d \log L}{d\beta} = 0.$$

2. Methodology

Research design and data

This study was carried on data provided by a leading commercial bank in Kenya. The loan borrowers included in the study were randomly picked from the bank's databank such we had a sample of 500 borrowers. The sample was drawn from a portfolio of personal loans whose maturity was 36 months or more and whose commencement date was January 2005. The subjects were observed through December 2007.

The event of interest is loan default denoted by 1. Censored state is denoted by 0. The covariates under analysis included:

(1) Customer's gender (gen) defined as $\begin{Bmatrix} 0 = \text{female} \\ 1 = \text{male} \end{Bmatrix}$.

(2) Customer's marital status (mar) defined as $\begin{Bmatrix} 0 = \text{single} \\ 1 = \text{married} \end{Bmatrix}$.

(3) Customer's age at application (age) date given as

$$\begin{Bmatrix} 0 = \text{below 35 years} \\ 1 = 35 \text{ years and above} \end{Bmatrix}.$$

(4) Customer's employment sector (emp) such that $\begin{Bmatrix} 0 = \text{private sector} \\ 1 = \text{public sector} \end{Bmatrix}$.

(5) Home ownership (home) denoted by $\begin{Bmatrix} 0 = \text{tenant} \\ 1 = \text{owner occupier} \end{Bmatrix}$.

(6) Highest level of education (edu) defined as $\begin{Bmatrix} 0 = \text{non-graduate} \\ 1 = \text{university graduate} \end{Bmatrix}$.

(7) Duration at current residence (res) given as $\begin{Bmatrix} 0 = \text{less than 5 years} \\ 1 = 5 \text{ years and above} \end{Bmatrix}$.

The model structure

Hazard rate at time *i* is given by

$$h_i(t) = h_0(t)e^{(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K)},$$

$\beta's =$ relative risks or hazard ratios (HR).

For this research,

$$h_i(t) = h_0(t)e^{(\beta_{gen} X_i + \beta_{mar} X_i + \beta_{age} X_i + \beta_{emp} X_i + \beta_{home} X_i + \beta_{edu} X_i + \beta_{resi} X_i)}.$$

3. Results, Discussions, Conclusions and Recommendations

Cox regression

Table 3.1

Case Processing Summary

		N	Percent
Cases available in analysis	Event[a]	104	20.8%
	Censored	394	78.8%
	Total	498	99.6%
Cases dropped	Cases with missing values	0	.0%
	Cases with negative time	0	.0%
	Censored cases before the earliest event in a stratum	2	.4%
	Total	2	.4%
Total		500	100.0%

a. Dependent Variable: survival time in months

Table 3.2

Categorical Variable Codings[c,d,e,f,g,h,i]

		Frequency	(1)[b]
gen[a]	0=female	198	1
	1=male	302	0
mar[a]	0=single	200	1
	1=married	300	0
age[a]	0=below 35 years	358	1
	1=35 years and above	142	0
emp[a]	0=private sector	299	1
	1=public sector	201	0
home[a]	0=tenant	408	1
	1=owner occupied	92	0
edu[a]	0=non-graduate	352	1
	1=university graduate	148	0
res[a]	0=less than 5 years	186	1
	1=5 years and above	314	0

a. Indicator Parameter Coding

b. The $(0, 1)$ variable has been recoded, so its coefficients will not be the same as for indicator $(0, 1)$ coding

c. Category variable: gen (gender)

d. Category variable: mar (marital status)

e. Category variable: age (age at application date)

f. Category variable: emp (employment sector)

g. Category variable: home (home ownership)

h. Category variable: edu (highest level of education)

i. Category variable: res (duration at current residence).

Table 3.3

Omnibus Tests of Model Coefficients

2 Log Likelihood
1242.293

Omnibus Tests of Model Coefficients[a,b]

–2 Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	Df	Sig.	Chi-square	Df	Sig.	Chi-square	Df	Sig.
1203.882	37.345	7	.000	38.412	7	.000	38.412	7	.000

a. Beginning Block Number 0, initial Log Likelihood function: –2 Log likelihood: 1242.293

b. Beginning Block Number 1. Method = Enter

Table 3.4

Variables in the Equation

	B	SE	Wald	Df	Sig.	Exp(B)	95.0% CI for Exp(B)	
							Lower	Upper
Gen	.271	.199	1.864	1	.172	1.312	.889	1.937
Mar	.587	.198	8.819	1	.003	1.799	1.221	2.652
Age	.562	.246	5.236	1	.022	1.755	1.084	2.840
Emp	.391	.212	3.384	1	.066	1.478	.975	2.242
Home	.819	.322	6.467	1	.011	2.268	1.207	4.264
Edu	.129	.225	.327	1	.567	1.137	.732	1.767
Res	.520	.197	6.936	1	.008	1.682	1.142	2.477

Table 3.5

Correlation Matrix of Regression Coefficients

	Gen	Mar	Age	Emp	Home	Edu
Mar	–.075					
Age	–.039	.018				
Emp	–.066	.026	–.078			
Home	–.100	.050	.041	.028		
Edu	–.019	–.001	–.021	–.044	–.065	
Res	–.038	–.078	–.006	–.045	.025	.052

Table 3.6

Survival Table

Time	Baseline Cum Hazard	At mean of covariates		
		Survival	SE	Cum Hazard
14	.000	.998	.002	.002
15	.001	.992	.004	.008
16	.002	.987	.005	.014
17	.003	.980	.006	.021
18	.004	.973	.007	.028
19	.006	.960	.008	.041
20	.008	.945	.010	.057
21	.010	.932	.011	.071
22	.012	.918	.012	.086
23	.014	.906	.013	.099
24	.016	.895	.014	.111
25	.019	.878	.016	.130
26	.022	.860	.017	.151
27	.026	.838	.018	.177
28	.028	.822	.019	.196
29	.030	.815	.019	.205
30	.032	.803	.020	.219
31	.033	.794	.021	.231
32	.034	.791	.021	.234
36	.035	.787	.021	.240

Table 3.7

Covariate Means

	Mean
Gen	.396
Mar	.398
Age	.715
Emp	.598
Home	.817
Edu	.705
Res	.373

Discussions, conclusions and recommendations

Table 3.1 gives a summary of case processing showing that 20.8% of the sampled loan applicants defaulted at some point in time over the observation period of 36 months. 78.8% survived there payment schedule and 0.4% settled their accounts before the earliest default was recorded. Table 3.2 provides summary statistics for each covariate in terms of sample size or frequency. Each covariate is well defined and recoded and uniqueness. Table 3.3 gives tests of the model coefficients. Overall, the model is significant since chi-square of 37.345 has significance of 0.000 at 5% level.

From Table 3.4, we obtain the β for each covariate, thus Cox regression model can be inferred. Based on the coefficients, the model is represented by

$$h_i(t) = h_0(t)e^{(0.271gen_i+0.581mar_i+0.562age_i+0.391emp_i+0.819home_i+0.129edu_i+0.520res_i)}.$$

Based on Wald test statistics, four covariates (mar, age, home and res) were significant at 95% confidence interval. The remaining three covariates (gen, emp and edu) proved insignificant. This implies the insignificant covariates should be dropped from loan application of the check list.

The correlation matrix of regression coefficients (Table 3.5) gives an indicator of the best combination of covariates that would minimise credit risk. The combination is given by covariates with negative correlation since such would serve as risk diversifiers.

The survival table (Table 3.6) provides baseline cumulative hazard at each point of incidence and corresponding survival rate. Survival rate declines with time while cumulative hazard increases. The covariate means (Table 3.7) indicate the average value for the hazard associated with each covariate. For instance, gen having the value of 0.396 implies that the hazard ratio for female to male is 1:0.4. Thus female borrowers were approximately two times more risky than male borrowers. The visual impressions for survival and hazard rates are shown by Figures 3.1 and 3.2, respectively.

Further research may be done on the same data set using other techniques like logistic regression and multivariate Kaplan-Meier estimator. It would interest to compare the results.

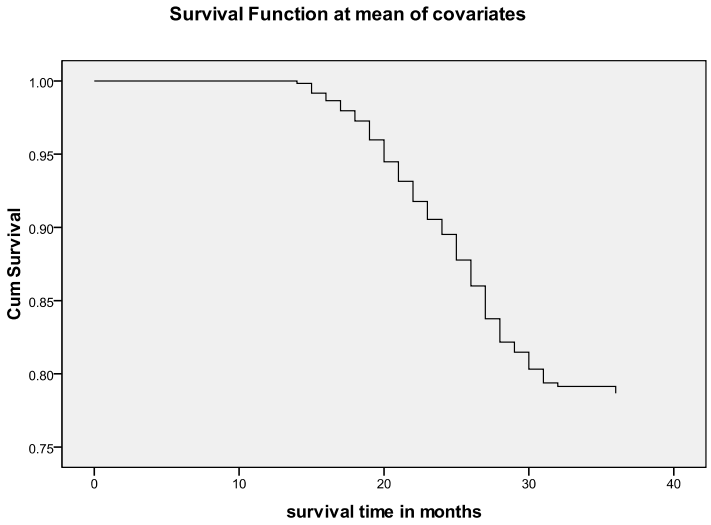

Figure 3.1

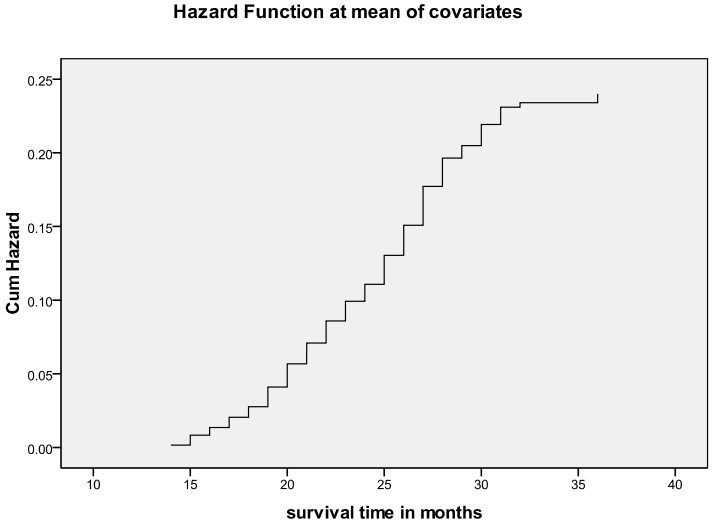

Figure 3.2

References

[1] A. Agresti, Categorical Data Analysis, Wiley, New York, 2002.

[2] P. D. Allison, Logistic Regression Using SAS System: Theory and Application, SAS Institute Inc., Cary, NC, 1999.

[3] B. Baesens, T. Van Gestel, M. Stepanova, D. Van den Poel and J. Vanthienen, Neural network survival analysis for personal loan data, Journal of the Operational Research Society 56 (2005), 1089-1098.

[4] B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens and J. Vanthienen, Benchmarking state-of-the-art classification algorithms for credit scoring, Journal of the Operational Research Society 54 (2003), 627-635.

[5] D. Bamber, The area above the ordinal dominance graph and the area below the receiver operating graph, J. Math. Psych. 12 (1975), 387-415.

[6] J. Banasik, J. N. Crook and L. C. Thomas, Not if but when will borrowers default, Journal of the Operational Research Society 50(12) (1999), 1185-1190.

[7] M. Boyle, J. N. Crook, R. Hamilton and L. C. Thomas, Methods for credit scoring applied to slow payers, L. C. Thomas, J. N. Crook and D. B. Edelman, eds., Credit Scoring and Credit Control, Oxford University Press, Oxford, 1992, pp. 75-90.

[8] S. Chatterjee and S. Barcun, A nonparametric approach to credit screening, J. Amer. Statist. Assoc. 65 (1970), 150-154.

[9] J. B. Copas and F. Heydari, Estimating the risk of reoffending by using exponential mixture models, J. Roy. Statist. Soc. Ser. A 160 (1997), 237-252.

[10] D. R. Cox, Partial likelihood, Biometrika 62(2) (1975), 269-276.

[11] R. B. D'Agostino and B. H. Nam, Evaluation of the performance of survival analysis models: discrimination and calibration measures, N. Balakrishnan and C. R. Rao, eds., Advances in Survival Analysis, Elsevier, London, 2004, pp. 1-25.

[12] R. H. Davis, D. B. Edelman and A. J. Gammerman, Machine-learning algorithms for credit-card applications, IMA J. Manag. Math. 4(1) (1992), 43-51.

[13] V. S. Desai, D. G. Conway, J. N. Crook and G. A. Overstreet, Credit-scoring models in the credit-union environment using neural networks and genetic algorithms, IMA J. Manag. Math. 8(4) (1997), 323-346.

[14] D. Durand, Risk Elements in Consumer Installment Financing, National Bureau of Economic Research, New York, 1941.

[15] V. T. Farewell, Mixture models in survival analysis: are they worth the risk? Canad. J. Statist. 14(3) (1986), 257-262.

[16] J. W. Gamel, I. W. McLean and S. H. Rosenberg, Proportion cured and mean log survival time as functions of tumour size, Stat. Med. 9 (1990), 999-1006.

[17] D. J. Hand and W. E. Henley, Statistical classification methods in consumer credit scoring: a review, J. Roy. Statist. Soc. Ser. A 160 (1997), 523-541.

[18] J. A. Hanley and B. J. McNeil, The meaning and use of the area under a receiver operating characteristic (ROC) curve, Radiology 143 (1982), 29-36.

[19] W. E. Hardy, Jr. and J. L. Adrian, A linear programming alternative to discriminant analysis in credit scoring, Agribusiness 1(4) (1985), 285-292.

[20] W. E. Henley, Statistical aspects of credit scoring, Ph.D. thesis, Open University, Milton Keynes, U.K., 1995.

[21] W. E. Henley and D. J. Hand, A *k*-nearest-neighbour classifier for assessing consumer credit risk, Statistician 65 (1996), 77-95.

[22] W. E. Henley and D. J. Hand, Construction of a *k*-nearest-neighbour credit-scoring system, IMA J. Manag. Math. 8 (1997), 305-321.

[23] P. Kolesar and J. L. Showers, A robust credit screening model using categorical data, Management Science 31(2) (1985), 123-133.

[24] A. Y. C. Kuk and C. H. Chen, A mixture model combining logistic regression with proportional hazards regression, Biometrika 79(3) (1992), 531-541.

[25] M. G. Larson and G. E. Dinse, A mixture model for the regression analysis of competing risks data, Applied Statistics 34 (1985), 201-211.

[26] K. J. Leonard, Empirical Bayes analysis of the commercial loan evaluation process, Statist. Probab. Lett. 18(4) (1993), 289-296.

[27] P. Makowski, Credit scoring branches out, Credit World 75 (1985), 30-37.

[28] R. A. Maller and S. Zhou, Testing for sufficient follow-up and outliers in survival data, J. Amer. Statist. Assoc. 89 (1994), 1499-1506.

[29] R. A. Maller and X. Zhou, Survival Analysis with Long-term Survivors, Wiley, New York, 1996.

[30] D. Mehta, The formulation of credit policy models, Management Science 15(2) (1968), 30-50.

[31] J. H. Myers and E. W. Forgy, The development of numerical credit evaluation system, J. Amer. Statist. Assoc. 58 (1963), 799-806.

[32] B. Narain, Survival analysis and the credit granting decision, L. C. Thomas, J. N. Crook and D. B. Edelman, eds., Credit Scoring and Credit Control, Oxford, 1992.

[33] Y. E. Orgler, A credit scoring model for commercial loans, Journal of Money, Credit and Banking 2(4) (1970), 435-445.

[34] L. C. Thomas, J. N. Crook and D. B. Edelman, Credit Scoring and Credit Control, Oxford University Press, Oxford, 1999.

[35] J. C. Wiginton, A note on the comparison of logit and discriminant models of consumer credit behavior, Journal of Financial and Quantitative Analysis 15(3) (1980), 757-770.