UNIVERSITY OF SOUTHAMPTON

Using survival analysis methods to build credit scoring models

Maria Stepanova

Submitted for the degree of Doctor of Philosophy

Department of Management
Faculty of Social Sciences

January 2001

To my parents.

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL SCIENCE

MANAGEMENT

Doctor of Philosophy

USING SURVIVAL ANALYSIS METHODS TO BUILD CREDIT SCORING MODELS

by Maria Stepanova

Credit scoring systems were originally built to allow organisations to measure how likely an applicant for credit is to default by a certain time in the future. In recent years the objectives of credit scoring models have shifted from choosing the customers presenting the lowest risk, towards choosing the customers offering the highest profitability. This thesis shows how using survival analysis tools from reliability and maintenance modelling allows one to build credit scoring models that assess aspects of profit as well as default.

In particular, this thesis looks at a number of extensions of Cox's proportional hazards model applied to personal loan data that make this technique a consistent and complete method for building a credit scorecard. Firstly, a new way of coarse-classing of characteristics using survival analysis methods is proposed. Secondly, a number of diagnostic methods to check the adequacy of the model fit are tested for suitability for use on loan data.

The inclusion of time-by-characteristic interactions is also proposed in order to account for non-proportional hazards and hence, extend the applicability of Cox's model.

Additionally, behavioural scoring models based on the proportional hazards approach are also developed. In conclusion, this thesis demonstrates how both behavioural and application survival analysis based models can be used to estimate the expected profit from personal loans, and can therefore be used to help lenders to move from default scoring to profit scoring.

Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my most sincere gratitude to my supervisor Professor Lyn C. Thomas for his help, guidance and encouragement in the preparation of this thesis.

I am very grateful to The University of Edinburgh, The University of Southampton and the Committee of Vice-Chancellors and Principals of the Universities of the UK for funding this research.

Many thanks go to my friends and colleagues, especially to Dr. Ann MacNeill, Dr. Guido Giese, Alison Hunter and Dr. Fabian Zuleeg, for their much needed encouragement and for making my years at University a very enjoyable and memorable experience. My special thanks go to Dr. R. Gayler who put me in contact with Professor L.C. Thomas and hence, facilitated the start of this research.

I am indebted to my parents for inspiring me to pursue a doctoral degree by setting an exceptional example.

Chapter 1

Introduction

Credit scoring is one of the most widely used applications of quantitative analysis in business. It has an interesting and dynamic history which we briefly discuss followed by the current practice in credit scoring in the opening section of this chapter.

We then talk about the recent change in the perception of the main aim of credit scoring. That is the growing awareness among lenders about the advantages of scoring for profit rather than risk. A number of authors pointed out the importance of estimating a 'lifetime' of a credit to be able to estimate its profitability and hence, proposed survival analysis as the new modelling approach for credit scoring.

The existing literature about suitability of survival analysis for credit scoring is not extensive. The final section gives a brief overview of the two key papers (Thomas et al. [1999], Narain [1992]). Both authors emphasize the evident need for more research into the 'marriage' of survival analysis and credit scoring, which was the motivation for this thesis. The overview of the upcoming chapters concludes the introduction.

1.1 Current practice in credit scoring

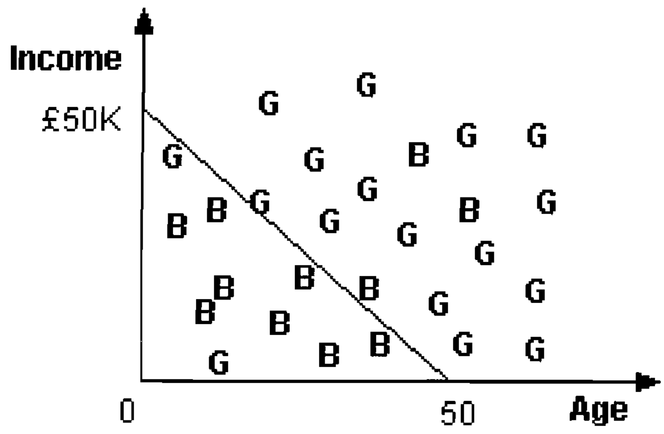
Consumer credit is now a fast-growing multi-billion dollar industry. However, until the second half of this century the decisions about granting credit were made by bankers or shop keepers based on the personal knowledge of their customer, his or her character, capacity and collateral – the so called Three C's of Credit. Nowadays few bankers would know the character of their customers, the capacity to repay can be estimated from their income and expenditure, however there is no guarantee that it is true, and personal loans are normally not secured by any collateral. Thus, the system of Three C's stops working if the number of applicants is too large for the decision maker to know them all personally.

After the second World War the demand for credit increased rapidly, so the need to automate the decision making process became evident. The finance and mail order companies were the first to implement credit scoring systems. They were followed by retailers and organizations issuing travel and entertainment cards and finally, by credit card companies in the 1970s. Lewis [1992] discusses the history and the background of credit scoring in more detail.

As the growth of the industry continues, these automated decision systems need to be constantly improved and made more sophisticated.

Credit scoring is a term used in the consumer credit industry for a collection of statistical or mathematical modelling techniques that aid the decision about granting credit to a new applicant. A credit scorecard is an automated decision system. It returns either a positive or a negative decision on credit applications based on the applicant's characteristics.

The standard approach to building a credit scorecard is to take a sample of past customers – their application forms and the subsequent credit history. Then one needs to define which history is 'good' and which is 'bad'. The most common definition of 'bad' is accounts which default. Default is normally defined as three consecutive payments missed. However, the definition of 'bad' can be related to other events such as bankruptcy or early repayment of a loan depending on the aims of the scorecard. One then builds a model to identify which characteristics, taken from the application form, best separate the 'goods' from the 'bads'. A simplistic example in Figure 1.1 shows that if only two characteristics Age and Income are considered, then a line $Age + Income = 50$ separates the two groups of 'bad' and 'good'.



B - ‘bad’ customer, G - ‘good’ customer

Figure 1.1: *Separation between two groups.*

Historically linear discriminant analysis was used to solve the problem of separating a population into two groups. Hence, it was one of the first methods applied to building credit scoring models. Eisenbeis [1977] discusses statistical difficulties in the theoretical justification and implementation of discriminant analysis in business, finance and economics. One of the major concerns was that an important assumption of discriminant analysis is that the variables used to characterize members of the groups are multivariate normally distributed. However, this is rarely true in credit data, especially since many binary variables are used for the analysis. Reichert et al. [1983] have shown that if one attempts to normalize distributions of variables with some transformation, it has almost no impact on the predictive power of the discriminant model. Hence, the lack of multivariate normality is not considered to be a critical pitfall and the technique is a useful and a successful tool, (Hand and Henley [1997],Thomas [1998]).

Linear regression (1.1) has also been used in credit scoring. It is essentially the same as linear discriminant analysis. The probability p_i of an applicant i being ‘good’ is expressed as a linear combination of the application characteristics $(X_1, X_2, \ldots, X_n)$:

$$p_i = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{1.1}$$

The current industry standard is logistic regression (1.2). This seems more appropriate than linear regression since the left hand side of the logistic regression equation (1.2) can take any value between minus and plus infinity as opposed to left hand side of the linear regression equation(1.1) which has to be between 0 and 1,(Thomas [1998], Hand and Henley [1997]).

$$\log(\frac{p_i}{1-p_i}) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \tag{1.2}$$

In short, all the above techniques try to find the best linear combination of the characteristics which explain the probability of being 'good'. Then the resulting scorecard is used on the new applicant to predict the applicant's probability of being 'good'. If this predicted probability of being 'good' is high, the application is approved and the credit is granted.

Credit scoring is a typical example of a classification problem (Hand [1981, 1997]), thus many other statistical and operational research based approaches were investigated for the use in credit scoring. They include linear programming, neural networks, recursive partitioning algorithms, expert systems and genetic algorithms. Several authors give a comprehensive overview of these techniques and their comparative advantages and disadvantages in relation to credit scoring (Thomas [1998], Hand and Henley [1997], Boyle et al. [1992], Srinivasan and Kim [1987b], Yobas and Crook [1997] and Desai et al. [1997]). Table 1.1 shows a comparison of the classification accuracy between some of the popular approaches to credit scoring. The entries in the table are the percentages of the correctly classified accounts if the acceptance percentage is the same for all the methods. The percentages should be compared only within rows since different authors used different measures of good, and different populations. Recursive partitioning algorithm is the most accurate in Henley [1985] and Srinivasan and Kim [1987b]. Linear regression is the most accurate according to Yobas and Crook [1997] and Boyle et al. [1992] and logistic regression performs best in the Desai et al. [1997] comparison. However, the results are not statistically significant. This means that the choice of one method over others should be determined by what other features it adds to the scorecard. As Hand [1997] points out, the question one should ask is not "Which is the best type of classification rule?" but "Which is the best type of classification rule for my problem?". The nature of the features required can depend on the definition of 'bad' or in other words, on the purpose for which a scoring systems is constructed . The following sections discuss the idea of profit scoring and survival analysis as the approach that brings valuable additions when a lender is concerned with scoring for profitability.

Authors	Lin.Reg.	Log.Reg.	RPA	LP	Neur.Nets	GA
Henley [1985]	43.4	43.3	43.8	-	-	-
Boyle et al. [1992]	77.5	-	75.0	74.7	-	-
Srinivasan and Kim [1987a]	87.5	89.3	93.2	86.1	-	-
Yobas and Crook [1997]	68.4	-	62.3	-	62.0	64.5
Desai et al. [1997]	66.5	67.3	67.3	-	6.4	-

Table 1.1: *A comparison of classification accuracy (% correctly classified accounts when the same number of applications is accepted) for different scoring approaches.*

1.2 Profit scoring

The purpose of credit scoring systems is to aid the decision of whether to grant credit to an applicant or not. Traditionally this was done by estimating the probability that an applicant will default. These objectives are changing in recent years towards choosing the customers offering a chance of highest profit rather than lowest risk. Oliver and Wells [1999] and Leonard [1997] combine profit and risk measures to achieve a more sensitive scoring system. Leonard [1997] shows on the real life example that the relationship between risk and profit is not one-to-one. The 'best' of the 'good' customers, i.e. those who were classified as the lowest risk, may incur losses for the lender because of the low usage on the account or ability to make large non-revolving payments. Customers classified as 'bad' may however bring substantial profits to the lender. It is possible that if time to default is long, the acquired interest or fees will compensate or even exceed losses due to default.

Hence, it now becomes important not only if but when a customer will default, Thomas et al. [1999]. If one is able to model the 'lifetime' of an account, it is then possible to calculate the likely profit as well as the risk associated with the customer. Another factor which affects profitability is when customers close their account early, or repay a loan early, by switching to another lender or for other reasons. In the case of early repayment the lender loses income from the interest that would have been otherwise earned on the loan. So estimating likely time until early repayment may aid the decision about the most appropriate loan term at the time of application.

1.3 Survival analysis approach

It has been shown by Narain [1992] and Thomas et al. [1999] that survival analysis can be applied to estimate the time to default or to early repayment. Survival analysis is the collection of statistical techniques that deal with the analysis of lifetime data. Examples of lifetime data can be found in medical or reliability studies, when a deteriorating system is monitored and the time until some event of interest is recorded. In credit scoring the lifetime is a 'life' of a loan or of a credit card until an event of interest, which can be some level of delinquency, early repayment, or any other event such as a purchase or a payment. Hence, the analogy between survival of machines or people and 'survival' of credit customers is evident.

This approach to using survival analysis to estimate time to default has also been used to model credit risk in the pricing of bonds and other financial investments.

There has been considerable work recently in developing default models to deal with credit risk, see the reviews by Cooper and Martin [1996], Lando [1997], Jarrow and Turnbull [2000]. Lando [1994] in his PhD thesis introduced a proportional hazards survival analysis model to estimate the time until a bond defaults, the aim being to use economic variables as covariates.

Another relevant application was made by Van den Poel and Leunis [1998]. The authors used proportional hazards model to investigate the effect of direct mail messages on the times between the last and the next purchase of a financial product by a customer.

The major strength of survival analysis is that it allows censored data to be incorporated into the model. Censoring is a feature of any lifetime data, since the event of interest can not always be observed. An example from medical studies is when a hospital loses contact with a terminally ill patient and hence the time of death can not be recorded. In reliability studies a machine may fail due to a cause, which was not a cause of interest, resulting in a censored observation.

Censored data occurs in the consumer credit context when a customer never defaults, or never repays early, so one does not observe an event of interest. However, it is clear that 'lifetimes' of these 'good' customers should be incorporated into the analysis and survival analysis allows that.

Another advantage of using survival analysis is that there is no need to choose a time horizon for a definition of 'bad' like in existing regression methods.

The principal problem of survival analysis is assessing the dependence of failure time on the explanatory variables. Thus, in credit scoring one can look for the dependence of a loan lifetime until default, early repayment, or normal completion on application characteristics.

Recall that in consumer credit we are interested in several possible outcomes when concerned with profitability: early repayment, default, closure, etc. Survival analysis allows modelling of such data with two or more events of interest or types of failure. This is called competing risks problem.

The idea of employing survival analysis for building credit scoring models was first introduced by Narain [1992] and then developed further by Thomas et al. [1999].

Narain [1992] applied accelerated life exponential model to 24 months loan data. The author showed that the proposed model estimated the number of failures at each failure time well. Then a scorecard was built using multiple regression and it was shown that a better credit granting decision could be made if the score was supported by the estimated survival times. Thus it was found that survival analysis adds an additional dimension to the standard approach. The author noted that these methods can be applied to any area of credit operations where there are predictor variables available and the time to some event is of interest.

Thomas et al. [1999] compared the performance of exponential, Weibull and Cox's non-parametric models with logistic regression and found that survival analysis methods are competitive with and sometimes superior to the traditional logistic regression approach. Furthermore, the idea of competing risks was employed when two possible outcomes were considered: default and early pay-off.

It was noted that there are several possible ways of improving the performance of the simplest survival analysis models which we are going to explore in this thesis. Chapter 2 outlines the theory of methods used in the analysis. Chapter 3 looks at the development of the techniques by applying them to the personal loan data. One of the suggested improvements is coarse-classing of the characteristic variables using survival analysis techniques rather than using the traditional approach. This is discussed in Section 3.3. Sections 3.5 and 3.6 apply these methods to prediction of early repayment and default respectively. Diagnostics used to test the model adequacy are compared in Section 3.7. The addition of a time by covariate interaction to the model is an important extension as it allows to decrease or increase the effect of a covariate on the predicted time to failure and this is discussed in Chapter 4. An approach to building behavioural scorecards using survival analysis is presented in Chapter 5. Chapter 6 concludes the thesis by summarizing the techniques researched in the earlier chapters and presenting the main advantage of applying survival analysis to credit scoring over other techniques, which is the ability to estimate the expected profit from an applicant.

Chapter 2

Some theory of analysis of lifetime data

2.1 Introduction

The previous chapter has discussed the motivation for applying survival analysis techniques to the building of practical credit scoring models. This chapter gives an overview of the theory behind the survival analysis and practical aspects relevant to the credit scoring purposes. Firstly, the lifetime data is discussed in more detail, including several ways of describing its distribution and its main feature – censoring. Then we explain Cox's proportional hazards model, concentrating on the situation of tied failure times. Ties are likely to occur in credit scoring data and result in a very complex likelihood function. So several available likelihood approximations are discussed and their practical advantages and disadvantages are compared.

Furthermore, we talk about competing risks, the idea that helps one to deal with the situation of two or more types of failure, for example, default and early repayment.

As for any other model, the fitness of a proportional hazards model has to be examined and many graphical and numerical diagnostics methods have been developed. We discuss several most popular ones, which we will compare later in Chapter 3.

Cox's proportional hazards model is proposed as an alternative to the current industry standard, logistic regression. Hence, its performance needs to be compared to logistic regression with respect to the traditional task of separating the population into two groups. The final section of this chapter discusses different ways of assessing classification performance, such as error rate, ROC curves and Gini coefficient.

2.2 Distribution of failure time data

Lifetime or failure time data has been briefly introduced in the first chapter. Such data may be collected for credit scoring purposes by recording, for example, the time from the start of a loan to the end of a loan and the information of why a loan ended; was it a case of default, early repayment or normal closure? One can then choose the event of interest out of these possible outcomes, and consider the lifetimes of loans resulting in all other outcomes censored.

The basis of the survival analysis techniques needed to analyse such failure time data is as follows.

Let T be the random variable representing time until the repayment of a loan ceases, i.e - time until default or early pay off.

The distribution function is then

$$F(t) = P(T \leq t). \tag{2.1}$$

However, in survival analysis the distribution of T is usually described by one of the three following functions.

The survival function is defined as

$$S(t) = 1 - F(t) = P(T \geq t), \tag{2.2}$$

which is the probability that a loan 'survives' past some specified time t.

Another way to describe the distribution of T is the hazard function, which is defined as follows

$$h(t) = lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \right\}. \quad (2.3)$$

This is the probability that a customer defaults, or repays early, at time t, conditional on him/her having stayed on the books up to that time. It is sometimes referred to as the instantaneous failure rate or the instantaneous default rate.

And finally, the cumulative hazard function, which relates to the hazard function, $h(t)$, by

$$H(t) = \int_0^t h(u)du. \quad (2.4)$$

is widely used, for example, for checking the validity of the assumptions of survival models.

Several useful relationships between hazard and survival functions can be derived from (2.3). The conditional probability in the hazard function definition is equal to $P(t \leq T < t + \delta t)/P(T \geq t) = (F(t + \delta t) - F(t))/S(t)$. Hence,

$$h(t) = lim_{\delta t \to 0} \left\{ \frac{F(t + \delta t) - F(t)}{\delta t} \right\} \frac{1}{S(t)} \quad (2.5)$$

$$= \frac{f(t)}{S(t)}, \quad (2.6)$$

where $f(t)$ is the density function. This is equivalent to

$$h(t) = -\frac{d}{dt}\{\log(S(t))\}. \quad (2.7)$$

By integrating both sides of (2.7) one can obtain the relationship between survival and cumulative hazard functions:

$$H(t) = -\log(S(t)), \quad (2.8)$$

$$S(t) = \exp(-H(t)). \quad (2.9)$$

2.3 Censoring

Censoring is an important feature of failure time data that makes methods other then survival analysis unsuitable. There are several types of censoring. Right censoring is when the actual unobserved failure time is greater than the observed. For example, if default is the event of interest and a loan is repaid normally at time T, default could have occurred at $T + x$. Left censoring is when the unobserved failure time is less than the observed censored time, i.e $(T - x)$. This type is very rare. Interval censoring is when the actual failure time is inside some time interval, $x \in (T_1, T_2)$.

Right censoring can be of Type 1, when a study or data collection is terminated at a particular time, or of Type 2, when data collection is terminated upon observing a pre-specified number of failures. Alternatively, it can be random if loans are completed at any time during the observation period.

For the purposes of building a relatively simple model with the help of some software the only thing one has to be absolutely sure about is that the censoring is non-informative. Informative censoring is when censored time is not independent from actual unobserved failure time, that is a loan is removed from the data set because it was deteriorating in its performance. This normally is not the case in credit scoring data.

It seems that if working with credit data for credit scoring purposes the censoring will usually be non-informative, most likely right and random. However, it depends on the event of interest. If the event of interest is default and the observation period is fixed as in our data set, then early repayment or normal repayment can occur at any time and we have right random non-informative censoring. It is right censoring because the the unobserved time of default is greater that the observed loan's lifetime, i.e the start of a loan is always known however the default date may be censored, either because it occurred outside the data collection interval or a loan was competed by normal or early repayment. It is random censoring because normally data used in credit scoring is some sample of past customers, collected during the last two years, for example. During this period of time loans 'leave' the data and hence, are censored at any time not only at the time data is taken.

2.4 Cox's proportional hazard model

Suppose now that on each loan one or more further measurements are available apart from the failure time, so we have a vector of covariates $\mathbf{x}$, such as application characteristics. We want to assess the relationship between the distribution of failure time and these covariates. Cox [1972] proposed the following model

$$h(t;\mathbf{x}) = e^{(\mathbf{x}'\beta)} h_0(t), \tag{2.10}$$

where β is a vector of unknown parameters and $h_0(t)$ is an unknown function giving the hazard for the standard set of conditions, when $\mathbf{x} = 0$. $h_0(t)$ is called the baseline hazard function.

It is called proportional hazards (PH) model because the assumption is that the hazard of the individual with application characteristics $\mathbf{x}$ can be obtained by multiplying the baseline hazard by a function of these characteristics. In other words, it is proportional to the baseline hazard. The strength of this model is that we do not need to know the parametric form of $h_0(t)$ to estimate β. Cox [1972] showed that one can estimate β by using only rank of failure times. So if $t_{(1)} < t_{(2)} < \ldots < t_{(k)}$ are k ordered failure times of individuals with covariates $\mathbf{x}_{(i)}$ and $R(t_{(i)})$ is the set of individuals at risk, i.e. still on the books, at $t_{(i)}$, then the likelihood function of observed data according to Cox's model is

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{k} \frac{P[\text{loan } i \text{ fails at } t_i]}{P[\text{one loan from } R(t_{(i)}) \text{ fails at } t_{(i)}\ ]} \\ &= \prod_{i=1}^{k} \frac{h(t_{(i)}, \mathbf{x}_i)\delta t}{\sum_{l\in R(t_{(i)})} h(t_{(l)}, \mathbf{x}_l)\delta t} \\ &= \prod_{i=1}^{k} \frac{\exp(\mathbf{x}_i'\beta) h_0(t_{(i)})}{\sum_{l\in R(t_{(i)})} \exp(\mathbf{x}_{(l)}'\beta) h_0(t_{(l)})} \\ &= \prod_{i=1}^{k} \frac{\exp(\mathbf{x}_i'\beta)}{\sum_{l\in R(t_{(i)})} \exp(\mathbf{x}_{(l)}'\beta)}. \end{aligned} \tag{2.11}$$

Note that the baseline hazard $h_0(t)$ has cancelled out from this likelihood, so the probability will be the same whatever the form of $h_0(t)$ is.

Also note that in the case of right censoring the likelihood is the same as 2.11, that is the rank vector of censored and failure times is shortened to include only failure times. This is achieved by assuming that censoring occurs straight after the failure (Leemis [1995]), or by arguing that no information can be contributed about β by the time intervals in which no failures occur because $h_0(t)$ might be identically zero in such intervals since it is left unspecified (Cox [1972]).

Hence, only the observed failures contribute to the numerator of (2.11), censored observations do not. However, all the observations contribute to the denominator, because all are at risk of failure (Collett [1994]). Kalbfleisch and Prentice [1980] and Leemis [1995] give derivations and assumptions for this argument.

Maximum likelihood estimates of β are then found by maximizing the logarithm of (2.11) using numerical methods, such as Newton-Raphson algorithm.

Proportional hazards models assume that hazard functions are continuous. However, credit performance data is normally recorded only monthly, so several failures at one time can be observed.

These are tied failure times and the likelihood function must be modified because it is now unclear which individuals to include in the risk set at each failure time $t_1, t_2, t_3, \ldots$. For example, if we have a record of three individuals (i_1, i_2, i_3) failing at t_k we do not know what the true order of failure was between them, and since there are six possible ways to order them, there are six possible risk sets. Hence, the exact likelihood function has to include all possible ordering of tied failures (Kalbfleisch and Prentice [1980]) and becomes very difficult to compute.

A number of approximations have been developed. One of these is achieved by replacing (2.10) by a discrete logistic model (Cox [1972])

$$\frac{h(t;\mathbf{x})}{1-h(t;\mathbf{x})} = e^{(\mathbf{x}\beta)}\frac{h_0(t)}{1-h_0(t)}, \tag{2.12}$$

where

$$h(t,\mathbf{x}) = P(t \leq T < t+1 | T \geq t)). \tag{2.13}$$

To show that (2.12) reduces to (2.10) when the time is continuous note that the general form of discrete hazard, replacing 1 by δt in (2.13), would be

$$h(t, \mathbf{x})\delta t = P(t \le T < t + \delta t | T \ge t)). \tag{2.14}$$

Then (2.12) becomes

$$\frac{h(t; \mathbf{x})\delta t}{1 - h(t; \mathbf{x})\delta t} = e^{(\mathbf{x}\beta)} \frac{h_0(t)\delta t}{1 - h_0(t)\delta t}, \tag{2.15}$$

and taking limit as time interval δt tends to zero gives (2.10).

Let d_i denote the number of failures at t_i and let $R(t_{(i)}; d_i)$ denote the set of all subsets of d_i individuals taken from the risk set $R(t_{(i)})$. $R \in R(t_{(i)}; d_i)$ is then a set of d_i individuals who might have failed at $t_{(i)}$. Let $s_R = \sum_{l \in R} \mathbf{x}_l$ be the sum of the covariate vectors $\mathbf{x}$ over the individuals in the set R. Let D_i denote the set of d_i individuals failing at t_i and $s_{D_i} = \sum_{l \in D_i} \mathbf{x}_l$ is the sum of the covariate vectors of these individuals.

The likelihood function arising from Cox's model is

$$L_{Cox}(\beta) = \prod_{i=1}^{k} \frac{\exp{(s'_{D_i}\beta)}}{\sum_{R \in R(t_{(i)}; d_i)} \exp{(s'_R\beta)}}. \tag{2.16}$$

The other popular approximations were proposed by Breslow [1974] and Efron [1977].

The Efron likelihood is

$$L_E(\beta) = \prod_{i=1}^{k} \frac{\exp{(s'_{D_i}\beta)}}{\prod_{j=1}^{d_i} \left[\sum_{l \in R(t_{(i)})} \exp{(x'_l\beta)} - \frac{j-1}{d_i} \sum_{l \in D_i} \exp{(x'_l\beta)}\right]}. \tag{2.17}$$

The Breslow likelihood is

$$L_B(\beta) = \prod_{i=1}^{k} \frac{\exp{(s'_{D_i}\beta)}}{\left[\sum_{l \in R(t_{(i)})} \exp{(x'_l\beta)}\right]^{d_i}}. \tag{2.18}$$

To illustrate the nature of each approximation let e_j denote $exp(\mathbf{x}_j\beta)$ and assume that we have three subjects all together in the risk set. So if there are no ties, i.e. only one failure occurs at a particular time, the probability that a subject 1 was the one that failed out of the risk set is

$$P_1 = \frac{e_1}{e_1 + e_2 + e_3}. \tag{2.19}$$

Now consider a tied failure case when two subjects 1 and 2 fail at one particular time and that we had three subjects all together in the risk set before that time. Then the probability of this happening is the the sum of the probability that subject 1 fails first and subject 2 fails second, and the probability that subject 2 fails first and subject 1 fails second. Note that the risk set corresponding to the subject that fails second does not include the subject that has already failed. So the exact probability taking into the account all possible ordering of failure is

$$P_{\text{exact}} = \frac{e_1}{e_1 + e_2 + e_3} \frac{e_2}{(e_1 + e_2 + e_3) - e_1} + \frac{e_2}{e_1 + e_2 + e_3} \frac{e_1}{(e_1 + e_2 + e_3) - e_2}. \quad (2.20)$$

For the case of two tied failures there are two terms in this probability corresponding to the two possible orderings of failures. For n tied failures there would be $n!$ terms since there are $n!$ possible ways to order n failures. Now let us look at the three approximations above to this probability for the case of two tied failures.

The simplest approximation is Breslow's,(2.21-2.22), which does not consider different orderings of failure when constructing risk sets and assumes that the two subjects failed at the recorded time simultaneously out of the full risk set. So the denominator is the same for both failures and is the sum over the full risk set. Hence, the two different terms in the exact probability are now replaced by twice the same term.

$$P_{\text{B}} = \frac{e_1 e_2}{(e_1 + e_2 + e_3)^2} + \frac{e_1 e_2}{(e_1 + e_2 + e_3)^2} \quad (2.21)$$

$$= 2 \frac{e_1}{e_1 + e_2 + e_3} \frac{e_2}{e_1 + e_2 + e_3} \quad (2.22)$$

This constant 2, or $n!$ for n ties, can be dropped when estimating parameters of a Cox's model since the estimation process involves only maximising the likelihood not the exact value.

Efron's approximation (2.23-2.24) also does not consider different orderings. It assumes that the two subjects failed sequentially, the first out of the full risk set and the second out of the risk set without the average of the subjects 1 and 2. So the denominator for the second failure is adjusted by taking the average of e_1 and e_2 out of the sum over the risk set.

$$
\begin{aligned}
P_{\mathrm{E}} &= 2\frac{e_1 e_2}{(e_1+e_2+e_3)^2 - \frac{1}{2}(e_1+e_2)(e_1+e_2+e_3)} && (2.23)\\
&= 2\frac{e_1}{e_1+e_2+e_3}\frac{e_2}{(e_1+e_2+e_3)-\frac{1}{2}(e_1+e_2)}. && (2.24)
\end{aligned}
$$

This approximation is closer to the real probability than Breslow's.

And finally, Cox's probability arising from the discrete time model is

$$
\begin{aligned}
P_{\mathrm{Cox}} &= 2\frac{e_1 e_2}{e_1 e_2 + e_2 e_3 + e_3 e_1} && (2.25)\\
&= 2\frac{e_1 e_2}{(e_1+e_2+e_3)^2 - e_1^2 - e_2^2 - e_3^2}. && (2.26)
\end{aligned}
$$

where the denominator in (2.25) is the sum of all possible sets of failures that might have occurred from the risk set. Hence, in the case of two tied failures and three subjects in the risk set the denominator is the sum of the three possible failure pairs. An alternative way to write the denominator, as in (2.26), allows an easier comparison of this approximation with the exact probability. So if the denominator for one of the terms in P_{exact} is $(e_1+e_2+e_3)((e_1+e_2+e_3)-e_1) = ((e_1+e_2+e_3)^2 - e_1^2 - e_1e_2 - e_1e_3)$ the P_{Cox} replaces this with $(e_1+e_2+e_3)^2 - e_1^2 - e_2^2 - e_3^2$.

In summary, if there are n tied failures, all three approximations replace $n!$ different terms of the exact probability with $n!$ identical terms. The difference between them is the form of that one term.

When there are no ties on the event time, i.e. $d_i = 1$, all the approximations reduce to the same expression (2.19).

2.5 Competing risks

So far we defined T as the time until failure. When one is concerned with the profitability of a credit account, several possible outcomes can be of interest. For example, both default on a loan and early repayment of a loan affect profitability.

Hence, it is possible that there are two or more possible risks that 'compete' to be the cause of the failure of a loan. The idea in survival analysis that allows analysis of data with two or more failure types is called 'competing risks', whereby a pair (T,J), where T is a lifetime of a loan and J is the cause of failure, is associated with each loan.

One can work with net or crude lifetimes, (Leemis [1995]). Working with the crude lifetimes means considering lifetimes conditioned on risk j in the presence of all other risks. Here we were only concerned with net lifetimes, i.e. the causes are viewed individually, each risk is considered as if it is the only risk present.

The type-specific hazard function in the context of net lifetimes is then defined as

$$h_j(t) = lim_{\delta t \to 0} \left\{ \frac{P(t \leq T < t + \delta t, J = j | T \geq t)}{\delta t} \right\}. \tag{2.27}$$

for $j = 1, ...m$ types of failure. It is interpreted as the instantaneous failure rate of type j at time t.

If net lifetimes are independent the overall hazard function is

$$h(t) = \sum_{1}^{m} h_j(t), \tag{2.28}$$

since $S(t) = P[T \geq t] = P[T_1 \geq t] \cdot \ldots \cdot P[T_1 \geq t] = \prod_{j=1}^{k} S_{T_j}(t)$.

The assumption of independent lifetimes is common for the reasons of simplicity and because it is impossible to test this assumption. A number of authors have noted and researched this issue (Cox [1959], David and Moeschberger [1978], Cox and Oakes [1984]). It has been shown that different bivariate models can result in the same hazard functions $h_1(t)$ and $h_2(t)$. Specifically, for any model where T_1 and T_2 are dependent there is a model with the same $h_j(t)$s for which T_1 and T_2 are independent (Lawless [1982]). So we feel that assuming time to early repayment and time to default are independent is a satisfactory practical approach to the problem.

Furthermore, Kalbfleisch and Prentice [1980] show that the competing risks data has a likelihood function that can be rearranged into separate components for each type of failure. The component for a failure type j is then exactly the same as if all other types are regarded as censored.

Hence, all the estimation methods for one-failure data can be used for a multiple-failure scenario. This applies to Cox's proportional hazard model as well.

One can estimate time until default T_1 assuming all other observed lifetimes to be censored and separately estimate time until early repayment T_2 assuming all other observed lifetimes to be censored. Hence, survival analysis can be performed separately on T_1 and T_2 assuming they are independent. Then the predicted lifetime of a loan is estimated as $T = \min\{T_1, T_2, \text{term of the loan}\}$, (Thomas et al. [1999]).

2.6 Model diagnostics

Any modelling procedures, such as logistic or linear regression models, for example, are normally followed by the examination of the model's goodness of fit to the data. Proportional hazards model is not an exception.

Residuals are the most popular diagnostic for all modelling approaches including survival analysis. They are usually some form of measure of discrepancy between fitted and predicted values. So if the model is adequate the plot of residuals should not show any unexpected patterns. The simplest residuals are found in linear regression and are calculated as a difference between predicted and actual values. Their plot is expected to be a random scatter about zero, such as shown in Figure 2.2. The examination for randomness is made visually. Several approaches to calculate residuals for survival analysis models have been developed. These residuals are more complicated than those used in linear regression because they have to cope with censoring.

The issues one wants to address in the diagnostics of proportional hazards credit risk models are:

- does the proportional hazards assumption hold?
- do any covariates have to be transformed?

- are there any outliers – individuals with repayment lifetimes greater than expected which might have an unwanted impact on parameter estimates?

- is the effect of any of the covariates time-dependent?

Cox-Snell residual,(Cox and Snell [1968]), is defined as follows

$$r_{C_i} = \exp(\hat{\beta}x_i)\hat{H}_0(t_i) = \hat{H}_i(t_i) = -\log \hat{S}_i(t_i), \tag{2.29}$$

where $\hat{H}_0(t_i)$ is the estimated cumulative baseline hazard, $\hat{H}_i(t_i)$ is the estimated cumulative hazard for the i^{th} individual at time t_i and $\hat{S}_i(t_i)$ is the estimated survivor function of the i^{th} individual at time t_i.

It is proven later that $-logS(t)$ has an exponential distribution with unit mean, no matter what the form of $S(t)$ is. If the model fitted is adequate then the estimated survival function $\hat{S}_i(t_i)$ will be close and will have similar properties to $S(t_i)$. Hence, $-log\hat{S}(t_i) = r_{C_i}$ will be a set of observations from an exponential distribution with unit mean.

If T is the random variable representing the survival time of a subject and S(t) is its survival function , then the random variable $Y = -logS(T) \sim Exp(1)$

Proof (Collett [1994]): Using the formula for the probability density function (p.d.f.) of a random variable which is a function of another random variable ($Y = g(X)$)

$$f_Y(y) = f_X\{g^{-1}(y)\}/|\frac{dy}{dx}|$$

we obtain the formula for the p.d.f of Y in terms of the p.d.f of T, $f_T(t)$:

$$f_Y(y) = f_T\{S^{-1}(e^{-y})\}/|\frac{dy}{dt}|. \tag{2.30}$$

Substituting the expression for Y into the dy/dt

$$\frac{dy}{dt} = \frac{d\{-\log S(t)\}}{dt} = h_T(t) = \frac{f_T(t)}{S(t)}$$

we can express dy/dt in terms of y:

$$\frac{dy}{dt} = \frac{f_T\{S^{-1}(e^{-y})\}}{S\{S^{-1}(e^{-y})\}} = \frac{f_T\{S^{-1}(e^{-y})\}}{e^{-y}}$$

By substituting the derivative in the equation (2.30) we obtain p.d.f of Y:

$$f_Y(y) = e^{(-y)}$$

which is the p.d.f of an exponentially distributed random variable with unit mean.

To test that the residuals have unit exponential distribution, the product-limit estimate (see later, in Section 3.4) of these values is computed. So $log(-log\hat{S}(r_{C_i}))$ is plotted against $log(r_{C_i})$. A straight line with unit slope and zero intercept (Figure 2.1) indicates that the fitted model is correct, i.e the estimated survival function is close to the true survival function of the data. If the plot shows systematic departures from the straight line, or the line does not have an approximately unit slope and zero intercept, the model has to be modified. For example, one may consider including additional covariates.

Hence, the log plot of the Cox-Snell residuals is used for accessing general adequacy of the model.

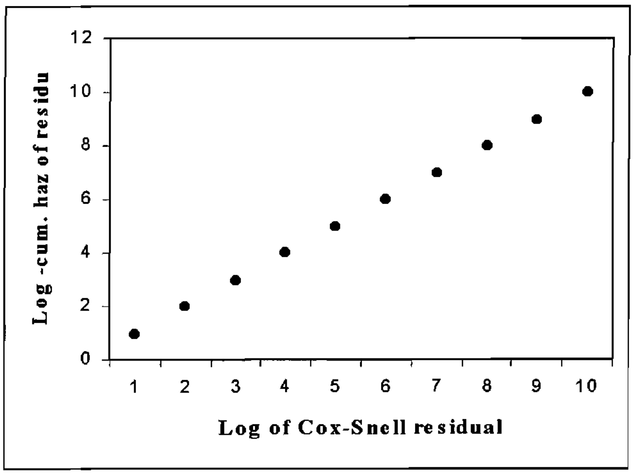


Figure 2.1: *An example of the log plot of the Cox-Snell residuals that indicates an adequate model.*

Martingale residual,(Therneau et al. [1990]), is a transformation of the Cox-Snell residual:

$$r_{M_i} = \delta_i - r_{C_i}, \tag{2.31}$$

where δ_i is a censoring indicator for the i_{th} individual: $\delta = 0$ for censored observations and $\delta = 1$ for non-censored.

Martingale residual can be interpreted as the difference between the observed number of failures for an individual in the interval $(0, t_i)$ and the expected number of failures according to the model, or as excess failures. These residuals have some properties similar to linear models, such as they sum to zero for any t: $\sum r_{M_i} = 0$ and their expected value is zero: $E(r_{M_i}) = 0$ asymptotically, (Therneau et al. [1990]).

To check for the departures from proportional hazards r_{M_i} is plotted against the rank order of time. Ideally it should not exhibit any pattern if the model is adequate (Figure 2.2).

A plot of martingale residuals against values of a particular covariate can be used to decide whether the variable needs to be transformed, so if the plot is non-linear one may consider replacing the original covariate x by $log(x)$, x^2 etc. However, if most of the covariates are binary, which is common in credit scoring, these plots are not useful.

Index plots can also be created by plotting the Martingale residuals against the observation number. A relatively large value of the residual indicates that the model does not fit well to the corresponding observation.

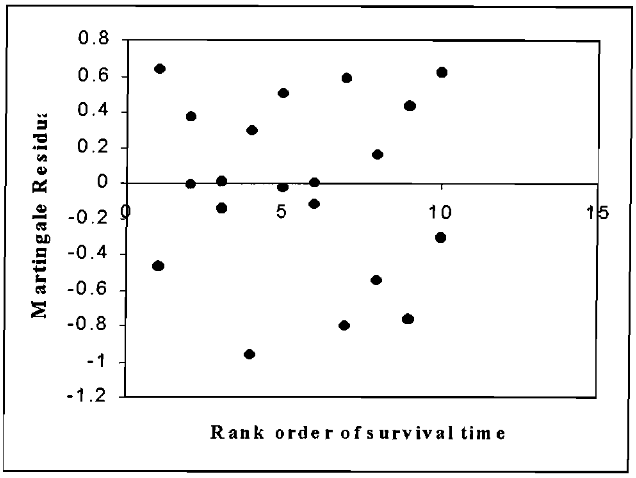


Figure 2.2: *An example of the Martingale residual plot that indicates an adequate model.*

One drawback in the use of martingale residuals is that they have a skewed distribution. For Cox's proportional hazard model it has a maximum of 1 and a minimum of $-\infty$. To check the accuracy of the prediction for the individual subjects, i.e. to detect outliers, the transformation of the residual to achieve a more normal shaped distribution is desirable.

The deviance residual is one such transformation of the martingale residual, which makes it more symmetrically distributed about zero and hence, their plots are sometimes easier to interpret.

Deviance residual proposed by Therneau et al. [1990] is defined as

$$r_{D_i} = sgn(r_{M_i})[-2\{r_{M_i} + \delta_i log(\delta_i - r_{M_i})\}]^{1/2}. \quad (2.32)$$

The log function increases martingale residuals close to one, while the square root contracts large negative values, hence normalising the shape of their distribution. The *sgn* function ensures the deviance residual has the same sign as the Martingale residual.

If $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})'$ is a vector of covariates for the i^{th} individual, $R(t_i)$ is the risk set, i.e. a set of indices of the individual who are still repaying at time t_i, then the **Schoenfeld residuals**, Schoenfeld [1982], at t_i are defined as the vector $r_i = (r_{i1}, \ldots, r_{ip})$, where

$$r_{ik} = x_{ik} - E(x_{ik}|R(t_i)). \tag{2.33}$$

So the Schoenfeld residual is the difference between the observed value of the covariate $\mathbf{x_i}$ and its expected value conditional on the risk set $R(t_i)$. That means that individuals who are unlikely to fail relative to the risk set, i.e. their covariate value is similar to those in the risk set, will have a small absolute value of the residual. Individuals who are likely to fail relative to those at risk will have a large absolute value of the residual.

The main difference of this residual from the others is that it has a vector of values for each individual, with a value for every characteristic or covariate of this individual. These residuals are uncorrelated with each other and $E(r_i) \simeq 0$ if the fitted model is adequate. So if the proportional hazards assumption holds the plot of the Schoenfeld residuals against the rank order of time for their corresponding covariate should be centred on zero.

Suppose a covariate effect has a time trend $g(t_i)$. Then it was shown by Schoenfeld [1982] that the sign of $E(r_{ik})$ depends on the the sign of $g(t_i)$. Hence, the plots of r_{ik} against the rank order of time are used in investigating whether there is an indication of time dependency, $g(t_i)$, for a particular covariate.

It is important to note that none of the above diagnostics make any assumptions about the distribution of loan lifetimes.

We further discuss applications of the above residuals when building a proportional hazards model on a data set in Chapter 3.

2.7 Comparing survival model with logistic regression

Assessment of the scorecard performance is an important part of the scorecard building process. One has to choose the performance measure depending on the objectives of the scoring system.

We will need to compare systems built using survival analysis both between themselves, and with the industry standard logistic regression. Hence, we will be comparing the systems in relation to the standard objective of classifying applicants into two groups, 'good' and 'bad'.

There are a number of ways to measure the classifying performance of a model. They differ with respect to the sample on which the misclassification rate is calculated. For example, if one checks the misclassification on the training sample used for building the model the result will be better than if the misclassification rate is calculated on the sample that was put aside for testing before the model building. This is because the model has taken into the account particular features of the sample it was built on. The testing sample is usually referred to as a hold-out sample.

This method is appropriate when the amount of data is large, so the quality of the model is not reduced by splitting the data into the training and the hold-out sample. There are also methods which can cope with a limited amount of data, such as cross-validation and bootstrapping.

The cross-validation method involves building a scorecard on all data but one observation, then testing the scorecard on this observation and repeating this process of leave one out for all of the data in turn. This method is often referred to as the leave-one-out method.

2.7.1 Error rate

Let us assume that the application characteristics $\mathbf{x} = (x_1, \ldots, x_2)$ are continuous, and let $f(x)$ be the distribution of the application characteristics. Denote the probability of being 'good' with the application characteristics $\mathbf{x}$ as $p(G|x)$ and the probability of being 'bad' with $\mathbf{x}$ as $p(B|x)$.

Then the Bayes rate is the optimal error rate given one has a complete knowledge of the above distributions:

$$e(Bayes) = \int \min\{p(B|x), p(G|x)\} f(x) dx. \tag{2.34}$$

When a credit scoring system is built, say on a sample Y of n customers, the probabilities $p(B|x)$ and $p(G|x)$ are estimated and the two regions of the values of the application characteristics are defined A_G and A_B. If the applicants characteristics fall into A_G the applicant is classified as 'good', and if into A_B - as 'bad'.

The true error rate for this system is then

$$e_Y(true) = \int_{A_G} f(x)p(B|x)dx + \int_{A_B} f(x)p(G|x)dx. \quad (2.35)$$

$e_Y(true)$ is the rate of misclassifying the new applicant if the system was applied to an infinite test set. However, since we only ever have finite samples we can only obtain an estimator of $e_Y(true)$ using some sample Y^*:

$$e_Y(Y^*) = \int_{A_G \cap Y^*} f(x)p(B|x)dx + \int_{A_B \cap Y^*} f(x)p(G|x)dx. \quad (2.36)$$

As was mentioned earlier, if one uses the training sample Y to calculate $e_Y(Y)$ the resulting estimate will be optimistically biased because the system has taken into the account all the features of the training sample which are not necessarily typical of the population. Hence, it is desirable to use some hold-out sample Y^* which is independent of the Y to estimate the actual error rate. The expected value of $e_Y(Y^*)$ over the data which excludes Y is equal to the actual error rate $e_Y(true)$.

To calculate the $e_Y(Y^*)$ for the credit scoring system that was built on the training sample Y we compare the the actual classes of 'good' and 'bad' with the predicted classes on the hold-out sample Y^*. Then the simplest way to summarise the results is to produce a confusion matrix shown in Table 2.1.

		Actual Class	
		Good	Bad
Predicted	Good	a	b
Class	Bad	c	d

Table 2.1: *A Confusion matrix for two groups: Good/Bad.*

The error rate can be estimated from this table, produced using a hold-out sample, by

$$\frac{b+c}{a+b+c+d}. \tag{2.37}$$

Different scorecards can be compared by comparing the number of correctly classified accounts, i.e. a or d. Also, misclassified 'bad' and 'good' rates are given by $\frac{b}{b+d}$ and $\frac{c}{a+c}$. Their complements are called sensitivity $\frac{a}{a+c}$ and specificity $\frac{d}{b+d}$.

Note that the confusion matrix method requires specification of a cut-off to separate the sample into the two classes of 'good' and 'bad'. This is not always desirable. For example, one may want to look at the range of cutoffs to choose the appropriate one taking the misclassification rate into the account.

2.7.2 Separation measures, ROC curves and Gini coefficient

There are several methods in statistics which can measure how different are the characteristics of two populations. These are called separation measures, and in the case of credit scoring they can help to get an idea of how good the scorecard is at separating populations of 'bad' and 'good'. These measures do not require a choice of a specific cut-off. So they are diagnostics concerned with a scoring system in general. One such measure is the Kolmogorov-Smirnov statistic given by

$$KS = max_i(P_g(i) - P_b(i)), \tag{2.38}$$

where $P_g(i)$ and $P_b(i)$ are the cumulative proportions of 'good' and 'bad' with a score i respectively. The KS statistic gives the maximum distance between the cumulative proportions of 'good' and 'bad' (see Figure 2.3).

Receiver Operating Characteristics (ROC) curve is another method that does not require a choice of a cut-off. A common ROC curve used in the credit scorecard construction is a plot of the percentage of 'good' accepted against the percentage of 'bad' accepted, i.e. correctly classified 'goods' against incorrectly classified 'bads', or sensitivity against one minus specificity, for all possible values of a score (Hand [1997] and Wilkie [1992]).

Figure (2.4) shows an example of ROC curves where the higher, dotted, curve indicates the better scorecard.

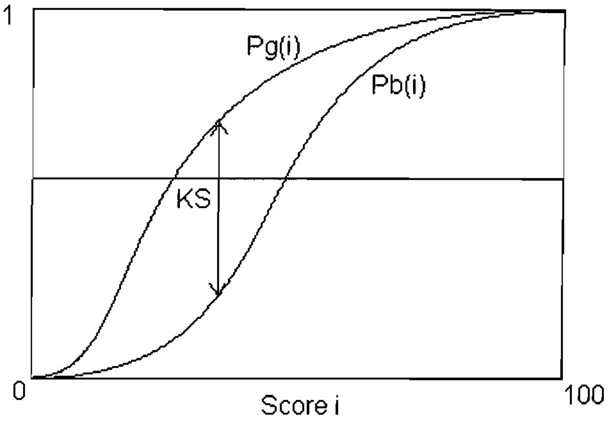


Figure 2.3: *The cumulative proportions of 'bad' and 'good' and the KS measure.*

The best possible ROC curve would go up the vertical axes all the way and then go parallel to the horizontal axes, so that all the accepted customers were correctly classified as good. The ROC curve along the diagonal would correspond to equal numbers of correctly and incorrectly classified accounts among the accepts for each score, which indicates that a scorecard is not discriminating between 'good' and 'bad' customers at all.

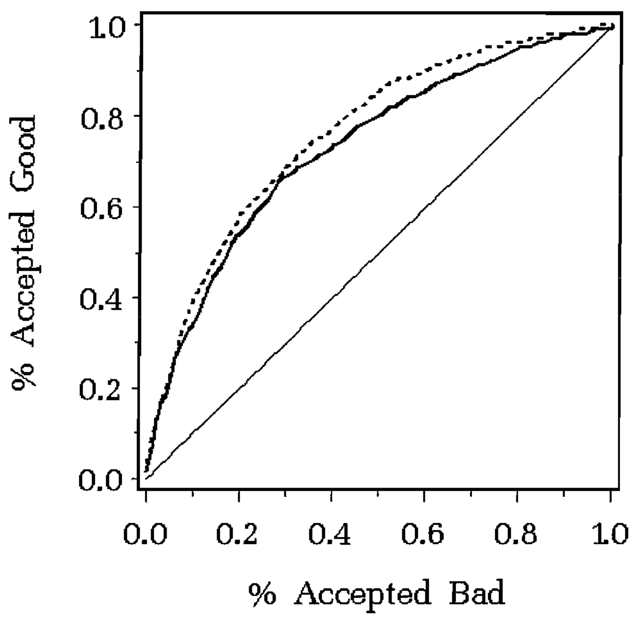


Figure 2.4: *An example of ROC curves.*

One can also summarise the information given by a ROC curve about the performance of a scorecard over all the cut-offs in one number – Gini coefficient. It is calculated as twice the area between the ROC curve and the diagonal, so the perfect scorecard has a Gini of 1 and the worst scorecard a Gini of 0.

To obtain Gini coefficient from the quantities used in building a ROC curve, first the area under the curve can be calculated as

$$A = \sum_i (P_g(i) + P_g(i-1)))/2 * (P_b(i) - P_b(i-1))). \qquad (2.39)$$

Then the Gini coefficient is $G = (A - 1/2) * 2$.

Cox's proportional hazards model gives a relative order of failure, i.e. of becoming 'bad', for a set of applicants. To compare the performance of a PH model with that of a logistic regression model we need to transform the estimated ordering into the two predicted classes. If the actual number of 'bads' in the sample is B, this can be achieved by considering B applicants who are most likely to fail according to the estimated survival function as predicted 'bad'.

Then the confusion matrices and the ROC curves can be constructed as for logistic regression.

Chapter 3

Survival analysis techniques applied to loan data

3.1 Introduction

In the earlier chapters we talked about motivation for using survival analysis as well as the theory and formulae needed to understand how to build a Cox's Proportional hazards model for the credit scoring purpose, how to examine its fitness and compare its performance with a logistic regression model. We now apply these techniques to personal loan data from a major UK financial institution. Several improvements are suggested over the current application of survival analysis in credit scoring. We suggest a method of coarse-classing the characteristics using Cox's proportional hazards model. Its advantage over the traditional methods is that it uses time to an event rather than a definition of 'bad' with an arbitrary time-horizon. We then illustrate the effect of the application characteristics on time to an event by looking at the hazard functions for early repayment and default. Two proportional hazards models, one predicting time to early repayment, another – time to default, are built and compared with analogous logistic regressions. The comparison shows that the survival analysis model measuring default risk is as good as the existing methods. In Chapter 6 we will also show how a PH model can be used to calculate the expected profit from a loan which existing default scorecards would not do.

Finally, a variety of graphical diagnostics, described in the previous chapter, are compared to each other by examining the fitness of the early repayment model.

3.2 Personal loan data

The data set was supplied by an anonymous UK financial institution and consisted of the application information on 50,000 personal loans together with the repayment status for each month of the observation period of 36 months. The application characteristics available in the data set are found in the Table 3.1. The status variable observed whether they had defaulted, paid off to term, paid off early, or was a loan still open. The definition of default used is three or more months delinquent. The definition of early repayment used is closure of a loan before its agreed term, where the indicator for closure was supplied with the data. The borrowers were all UK consumers, who had applied to the bank for a loan. Their repayment terms varied from 6 to 60 months and the various purposes for which the loan was needed are summarised in Table 3.7.

No	Characteristic	No	Characteristic
1	Customer Age	9	Home Phone No Given
2	Amount of Loan	10	Insurance Premium
3	Account Closing Date	11	Loan Type (single or joint)
4	Years at Current Address	12	Marital Status
5	Years with Current Employer	13	Account Opening Date
6	Customer Gender	14	Term of Loan
7	Number of Dep. Children	15	Home Ownership
8	Frequency Paid	16	Purpose of Loan

Table 3.1: *Application characteristics used in the analysis.*

The 'survival' time of a loan was calculated using the opening and the closing dates of the loan available in the data set. Two censoring indicators were created for each of the two types of failure under the consideration, default and early repayment, using the status of a loan at the end of the observation period:

1. CENSOR - the censoring indicator for default, i.e. 3 or more payments missed;
2. CENSORE - the censoring indicator for early repayment, i.e. the loan is repaid at least 1 months before the term.

The censoring indicator is usually a binary variable that takes the value of 0 if the observation is censored and the value of 1 otherwise. So that when estimating the risk of default, the survival times of the loans whose status was not default were considered censored. When estimating risk of early repayment, the survival times of the loans whose status was not early repayment were considered censored.

Entries with zero survival time were excluded from the analysis as well as entries with improbable values of other characteristic.

For example, the Age of a customer was limited to be between 18 and 85 because to get credit you have to be 18 and it is very unusual for people over 85 to apply for loans. So the observations with unlikely combinations of the Purpose of a Loan and the Age of an Applicant, such as 112 years old and a loan for a motorcycle, were deleted from the analysis assuming that the data were entered wrongly.

3.3 Coarse classing using survival analysis approach.

To ensure that credit scoring systems are robust, i.e. predictive rather than descriptive of data, continuous characteristics such as Age are usually split into 'bands' and the values of discrete characteristics with many values are grouped. This procedure is called coarse classing. Then each 'band' or 'group' is replaced with a binary dummy variable to be used as covariates in a model.

The traditional approaches of finding the suitable splits involve looking at the odds to be good, log(odds), or related measures for the different values of a characteristic and then grouping the values with similar odds (Lewis [1992]). Inherent in these approaches is the choice of a time horizon, so that defaults before that time horizon are 'bad', while ones that default after it or do not default at all are 'good'.

When using the survival analysis modelling techniques it seems more appropriate to use an approach that avoids the need to use such an arbitrary time horizon. It is also the case that if survival analysis models are being built to estimate both default and early repayment risk, then one will want to band the variables differently for the different risks. For these reasons it seems more appropriate to try and use the survival analysis approach in order to coarse classify the variables.

The following method was used for the continuous characteristics:

1. Split the characteristic into 15 to 20 equal bands.

2. Create a binary variable for each band.

3. Fit Cox's proportional hazards model to these binary variables.

4. Chart the parameter estimates obtained in step 3 for all bands.

5. Choose the splits based on similarity of the parameter estimates.

Example 3.3.1 looks at both the standard and the survival analysis approaches by coarse-classing one of the continuous variables.

For discrete characteristics, such as the Purpose of a Loan, a binary variable is created for each attribute of the characteristic and then the method is as for a continuous characteristic (Example 3.3.2).

Note that it is important to do separate splits for every type of failure considered. Effect of the characteristics on the failure time differs substantially for early repayment and default, for example.

There are a number of statistics that are used to compare coarse groupings created using different classing methods. We will use these statistics to compare the survival analysis classing with the log-odds one in the example below.

Suppose that a continuous characteristic was coarse classed into a bands with g_i 'good' and b_i 'bad' customers in the band $i = 1 \dots a$. Also let g and b be the total numbers of 'good' and of 'bad' in the sample. Then if one assumes that there is no difference in good/bad ratio between a class and the whole population, i.e. the classing does not reflect the effect of the characteristic on the good/bad ratio, the expected number of 'good' customers in the class is $\hat{g}_i = (g_i + b_i)g/(g + b)$ and of 'bad' – $\hat{b}_i = (g_i + b_i)b/(g + b)$. Good classing would separate the values into classes which differ the most from one another in the good/bad rate. The χ^2 statistic is designed to measure how different are these good/bad ratios in the different classes:

$$\chi^2 = \sum_{i=1}^{a} \frac{(g_i - \hat{g}_i)^2}{\hat{g}_i} + \frac{(b_i - \hat{b}_i)^2}{\hat{b}_i}. \tag{3.1}$$

The larger value of the statistic indicates a better split. For a formal significance test for a difference between good/bad ratios in the different classes one would use the fact that the above χ^2 statistic has a χ^2 distribution with $a - 1$ degrees of freedom.

F-statistic is defined as follows:

$$F = \sum_{i=1}^{a} (\frac{g_i}{g} - \frac{b_i}{b}) \log(\frac{g_i b}{b_i g}). \tag{3.2}$$

It measures how different are the $p(x|G)$ and $p(x|B)$, density functions that a good/bad has application data $\mathbf{x}$. The larger values of F indicate the larger difference and hence, a better split.

To calculate Somer's D Concordance statistic the classes need to be ordered with respect to the good/bad ratio, so that the first has the lowest 'good' rate and the last – the highest. Then the D-statistic measures how likely is that when you pick a 'good' customer from 'goods' and a 'bad' from 'bads' at random, the bad's attribute of the characteristic will fall into the lower class of that characteristic than the good's attribute:

$$D = \sum_{i=1}^{a} \frac{(\sum_{j<i} b_j) g_i - (\sum_{j<i} g_j) b_i)}{gb}. \tag{3.3}$$

The higher values of the statistics indicate a better split.

3.3.1 Example of coarse-classing a continuous characteristic

To illustrate coarse-classing of a continuous characteristic consider the characteristic the Amount of a Loan. It was split into 10 bands with approximately equal number of observations in each band. Then a proportional hazards model was fitted to the time to early repayment with covariates being binary variables corresponding to each band. The histogram of the parameter estimates is shown in Figure 3.1.

Looking at the histogram, it seems reasonable to group the first two bands, then the next three, the next two, and the last two together. We know the minimum and the maximum amounts in each band, so the limits of a class combining several bands will be constructed by taking the minimum of the first band and the maximum of the last one in the class. Notice that it is desirable to group approximately the same number of bands together, so that the resulting classes are of comparable sizes. The histogram shows only nine bars because the first split, the fine-classing, is exhaustive and hence, the model estimates $n - 1$ parameters.

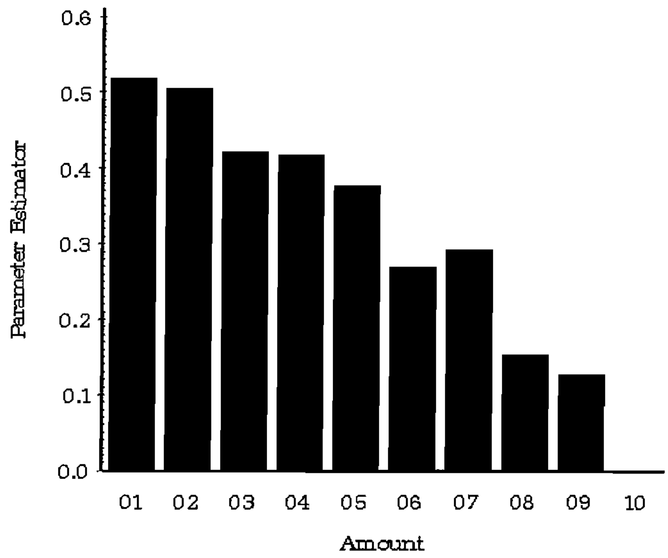


Figure 3.1: *Parameter estimates for the Amount of a Loan from the early repayment model.*

class	N obs	Min(Amount)	Max(Amount)
1	7132	400	800
2	10350	850	1850
3	6684	1900	2950
4	9652	3000	7900

Table 3.2: *Coarse classing of the Amount of a Loan using survival analysis.*

The final split of the Amount (see Table 3.2) is not exhaustive since we left out the last decile, which is equivalent to giving this band a parameter of zero in the model.

To see how the survival coarse-classing approach compares to the standard one, the Amount of a Loan was coarse-classed using both methods. The standard method involved fine-classing by splitting the values of the Amount of a Loan into 10 intervals of similar size. Then the log-odds were calculated for each interval: $log(g_i b/b_i g)$ where g_i and b_i are numbers of 'good' and 'bad' cases in the interval i and g and b are the total numbers of 'good' and 'bad' respectively. The definition of 'bad' was early repayment before the end of a loan. Finally, the intervals with similar log-odds values were grouped. The fine-classing histogram of log-odds is shown in Figure 3.2 and the resulting coarse-classing is in Table 3.3.

class	N obs	Min(Amount)	Max(Amount)
1	4241	400	500
2	8344	550	1000
3	8889	1050	2000
4	12345	2050	7900

Table 3.3: *Coarse-classing of the Amount of a Loan using log-odds method.*

It can be seen that the histograms of the PH parameter estimates and the log-odds are slightly different and may suggest different groupings of the fine-classing intervals.

We can now use the statistics discussed earlier to compare the two alternative coarse-classings of the Amount. Table 3.4 gives values of χ^2, F and D statistics for the two classing options.

All three statistics have similar values for both groupings, however log-odds method values are slightly higher and hence, indicate a better classing. This is expected because the log-odds method uses the same good/bad definition as the statistics and the survival analysis method does not use a good/bad definition at all. So one should not be discouraged by the difference in the values but encouraged by how close the survival analysis is to the traditional method which is designed to give the best split.

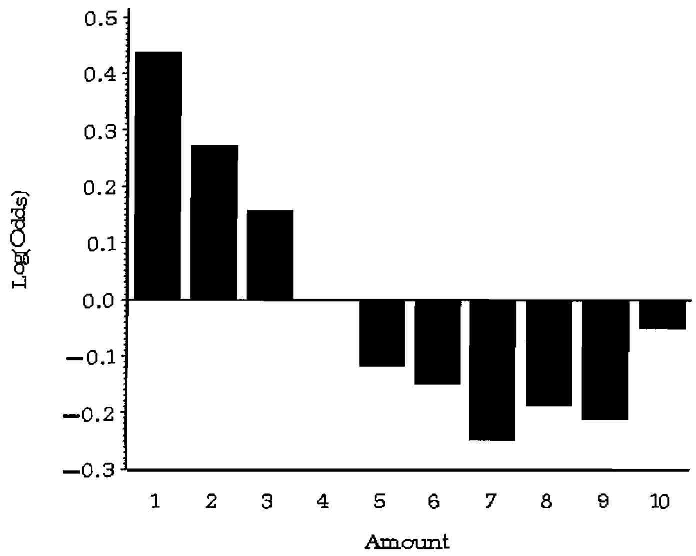


Figure 3.2: *Log-odds for the Amount of a Loan for early repayment before the end of a loan.*

Method used for classing	Statistic χ^2	F	D
Survival Analysis	0.107	0.042	0.107
Log-odds	0.140	0.046	0.112

Table 3.4: *Statistics comparing the two classing methods for the definition of 'bad': early repayment before the end of a loan.*

Method used for classing	Statistic χ^2	F	D
Survival Analysis	0.326	0.056	0.123
Log-odds	0.316	0.049	0.114

Table 3.5: *Statistics comparing two classing methods for the definition of 'bad': early repayment in the first 12 months.*

However, if one was to use a different definition of good/bad to that used in the log-odds classing, the survival analysis method is expected to perform better while the log-odds approach would lose some of its power. In other words, the survival analysis method's performance is not connected to one particular good/bad definition and this is one of its advantages over the log-odds method. This is demonstrated in the Table 3.5, where χ^2, F and D statistics were calculated using the following definition of bad: 'bad' are those who repaid early in the first 12 months of the life of a loan.

The values of the statistics are higher for the survival analysis based classing then for the log-odds classing. As expected, the survival analysis method for coarse-classing outperforms the log-odds method when the good/bad definition is different from the one used in the log-odds classing.

We have also tested alternative groupings, one based on the survival analysis approach and another on the log-odds approach, by fitting two proportional hazards model to the time to early repayment and using two alternative sets of binary variables, representing the Amount of a Loan, as the covariates. The models were of the form:

$$h(\text{Open}) = exp(\beta_1 \text{Amount}_1 + \beta_2 \text{Amount}_2 + \beta_3 \text{Amount}_3 + \beta_4 \text{Amount}_4) h_0 \quad (3.4)$$

where Open is the time to early repayment and $\text{Amount}_1 \ldots \text{Amount}_4$ are the indicator variables for the different classes of the Amount of a Loan constructed by either the survival analysis based coarse-classing or the standard coarse-classing approach.

The two models were compared using the corresponding values of the -2 Log Likelihood statistic, Table 3.6. The values of the statistic are similar, with the survival analysis classed model slightly better than the standard one, but not significantly.

Model	-2 Log Lik
using survival analysis classing	224086.040
using log-odds classing	224131.596

Table 3.6: *The -2 Log Likelihood statistic for the models using different classing methods.*

To summarise, the survival analysis based approach to coarse classing performed as well as the standard one based on the log-odds. However, one may prefer using the survival analysis based classing when developing a scorecard using survival analysis for the reasons of not having to make the assumption about a time horizon for the good/bad definition, and because of the consistency and the convenience of using the same software as for the main model building.

3.3.2 Example of coarse-classing a discrete characteristic

There are 27 different purposes of loans in the data. Some purposes are very rare and hence, have a very few observations in the data set. These were combined with the other more frequently occurring purposes, so that the proportional hazards model was fitted to 22 binary variables indicating 22 purpose groups (Step 3 of the coarse-classing procedure above). Figure 3.3 shows a chart of the parameter estimates from the proportional hazards model predicting early repayment for each group. Then three binary indicator variables are created, so that one has purposes with the highest parameter estimates, i.e. purposes with the highest risk of early repayment, second one - purposes with the middle values of parameter estimates, and third one - with the lowest values.

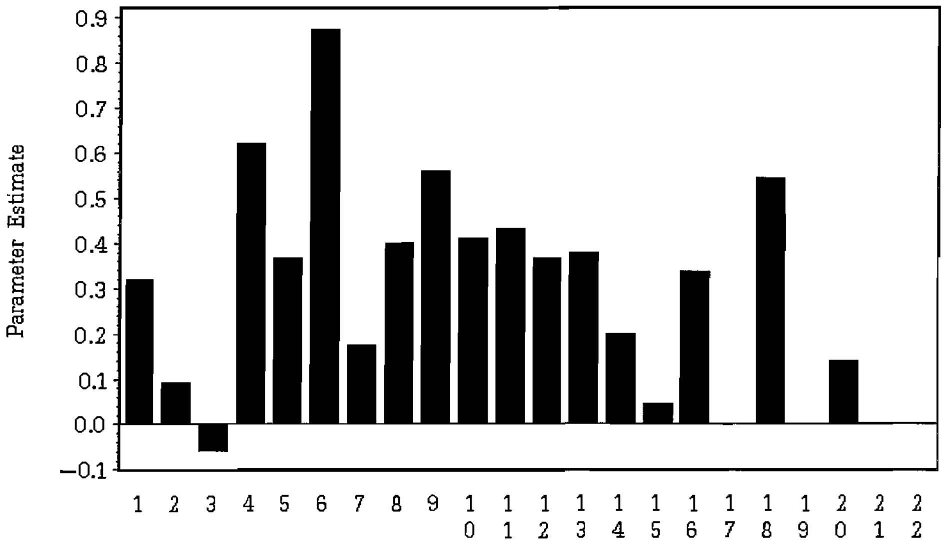


Figure 3.3: *Parameter estimates for the Purpose of a Loan from the early repayment PH model.*

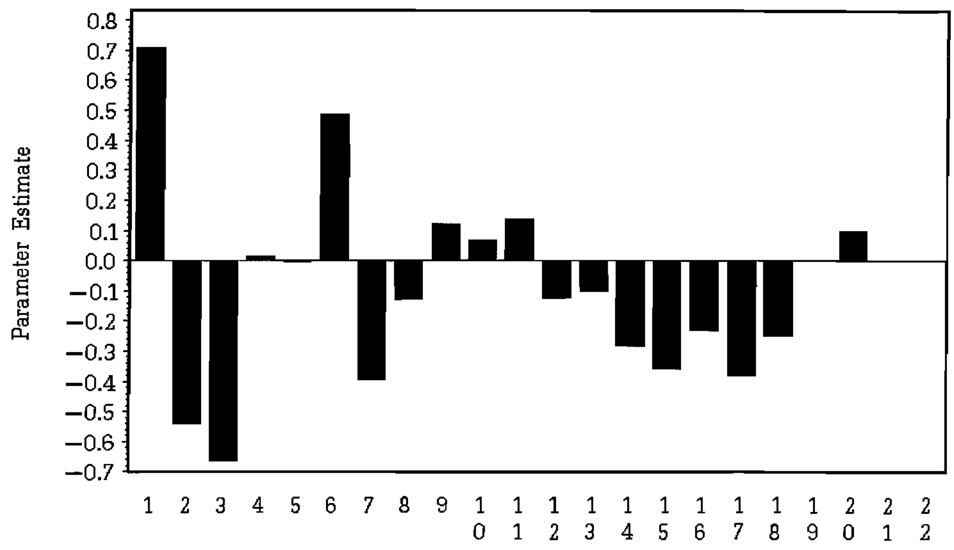


Figure 3.4: *Parameter estimates for the Purpose of a Loan from the default PH model*

No	Loan Purpose	No	Loan Purpose
1	Account Standard, Re-mortgages, Graduate Loan, Refinance	12	Weddings
2	Caravan, Motor Caravan	13	Boat, Motor Cycle
3	New Car	14	Car Over 3 Yr Old
4	Car Repair	15	Car Under 3 Yr Old
5	Electrical	16	Furniture
6	General Living	17	Kitchen Units
7	Home Improvement	18	Musical Instrument
8	Honeymoon, Holiday	19	Other Specific
9	Mixed Purchases	20	Other Vehicles
10	Others	21	Van
11	Redecoration	22	not specified

Table 3.7: *Purposes of a personal loan.*

Figure 3.4 shows a chart of the parameter estimates from the proportional hazards regression predicting default for 21 purpose groups. Notice that the most risky purposes for early repayment are not the most risky purposes for default. This illustrates the importance of doing coarse classing independently for each failure type.

Coarse-classing of the Term of a Loan

The same method was used to achieve the best grouping of the values of the Term of a Loan. This characteristic was considered for the segmentation of the scorecard because it was the most significant one in the preliminary PH regression. Further analysis discussed in Section 3.4.1 confirmed that the Term of a loan is best used to segment the population.

The best segmentation for predicting early repayment is to take 6 months term by itself, 12 and 18 months terms together, 24 and 30 months terms together and more than 30 months terms together. It was found that for the model predicting default it is best not to group values of the Term of a Loan and so, the segmenting variable has eight values corresponding to eight different terms.

3.4 Hazard function and how it is affected by covariates

The hazard function which was defined in Chapter 2 is a useful tool for data exploration. Plotting the hazard function shows how the instantaneous default or early repayment rate changes over time. Hence, it can be used for getting a general idea about default or early repayment patterns in a portfolio as well as looking for the effects of marketing campaigns or policy changes on customer attrition.

Figure 3.5 shows a plot of a default hazard rate against the time on the books. It was calculated on the personal loan data described in Section 3.2 using the product-limit method: if $t_1 < \ldots < t_k$ are k event times, n_j is the number of customers in the risk set just prior to t_j when d_j customers fail, the product limit estimate of the hazard function at time t_j is

$$\hat{h}(t_j) = \frac{d_j}{n_j} \tag{3.5}$$

However, most statistical packages including SAS only calculate the product-limit estimate of the survival function:

$$\hat{S}(t_j) = \prod_{j=1}^{i}(1 - \frac{d_j}{n_j}). \tag{3.6}$$

To get the hazard function estimate from (3.6) one simply needs to divide (3.6) by $S(t_{j-1})$ and solve for d_j/n_j :

$$\hat{h}(t_j) = \frac{d_j}{n_j} = 1 - \frac{\hat{S}(t_j)}{\hat{S}(t_{j-1})}; \tag{3.7}$$

When examining the default hazard rate plot (Figure 3.5), the first three months have to be ignored because the definition of default is more than three months delinquent.

The highest default rate is at the beginning of the loans and it then decreases with time. This may support the well-known concept that "if they go 'bad' they go 'bad' early". For this particular data set 'early' means from 18 months to 2 years as it can be seen from the plot in Figure 3.5.

Early repayment hazard, calculated analogously to the default hazard and on the same data, is shown on Figure 3.6. It can be seen that the hazard rate of early repayment is low at the very beginning, which is sensible because not many people repay a loan straight away, then it gradually increases and then flattens out.

Note that both hazard plots are calculated using the whole sample which was not segmented by Term or any other characteristic, so these hazards are averaged between the loans of different terms. Hazard plots for the data segmented by the Term of a Loan provide additional information about hazard rates for loans of the same term and are discussed in the following section.

Figure 3.5: *Default hazard rate.*

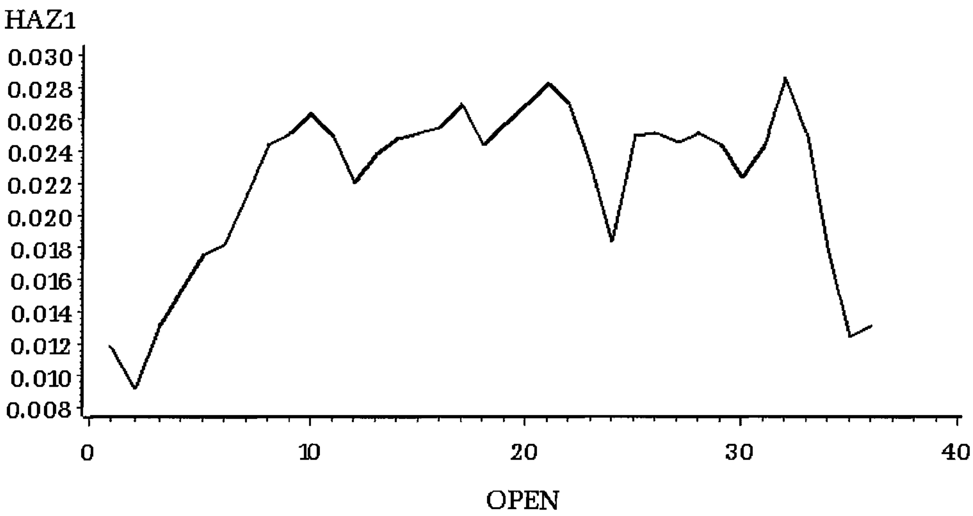


Figure 3.6: *Early repayment hazard rate.*

3.4.1 Segmentation of the hazard rate and the 'reversed' time scale

When data, as in our case, contains loans with a range of terms, from 6 months to 72 months, it is advisable to look at the hazard rates for each term separately. This highlights the features of default or early repayment rates in relation to the start and the end of a loan, rather than to the time on the books in general, which were otherwise hidden by averaging over all terms.

Plots of the default hazard rates for 12, 24 and 36 months loans against time are shown in Figure 3.7. All three lines seem to be very similar and show no new features to the ones observed on the averaged default hazard rate plot. The 12 months hazard is very similar to the 24 months hazard 'cut' at 12 months, and the 36 months hazard also just adds another 12 months to the familiar picture of the 24 months hazard. However, the 12 months hazard is slightly higher than the 24 months one because shorter loans would usually be associated with lower amounts and hence, would be available for more 'risky' applicants. These plots tell us that default hazard rate does not depend on time to the start or to the end of the loan, but on time in general. Also, one can see that there is almost no decline in the hazard rate for the 12 months loans and a very slight decline for the 24 months loans, which gives a more clear idea of what 'early' means for this data from the phrase "if they go 'bad' they go 'bad' early".

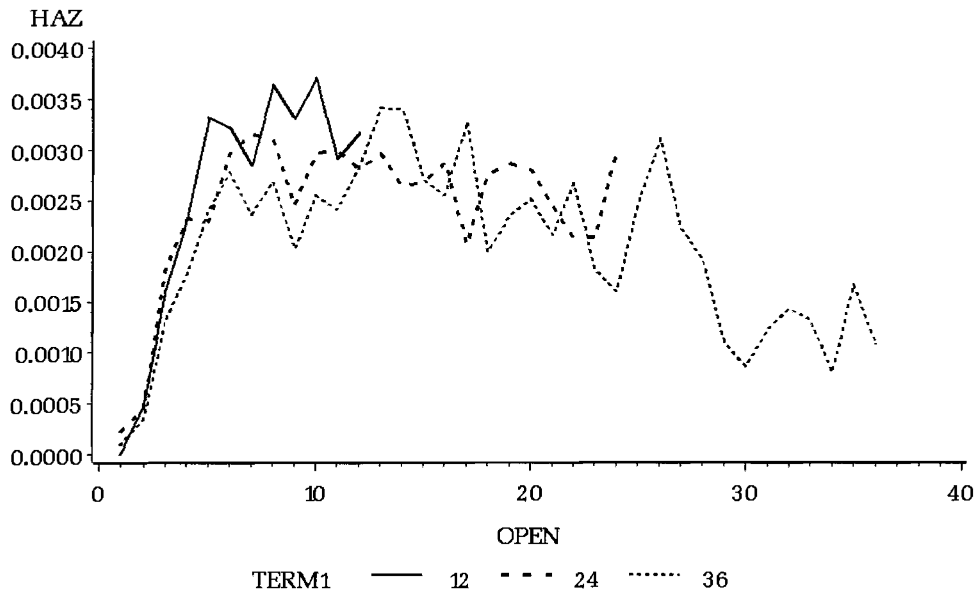


Figure 3.7: *Default hazard rate for three different terms of loans .*

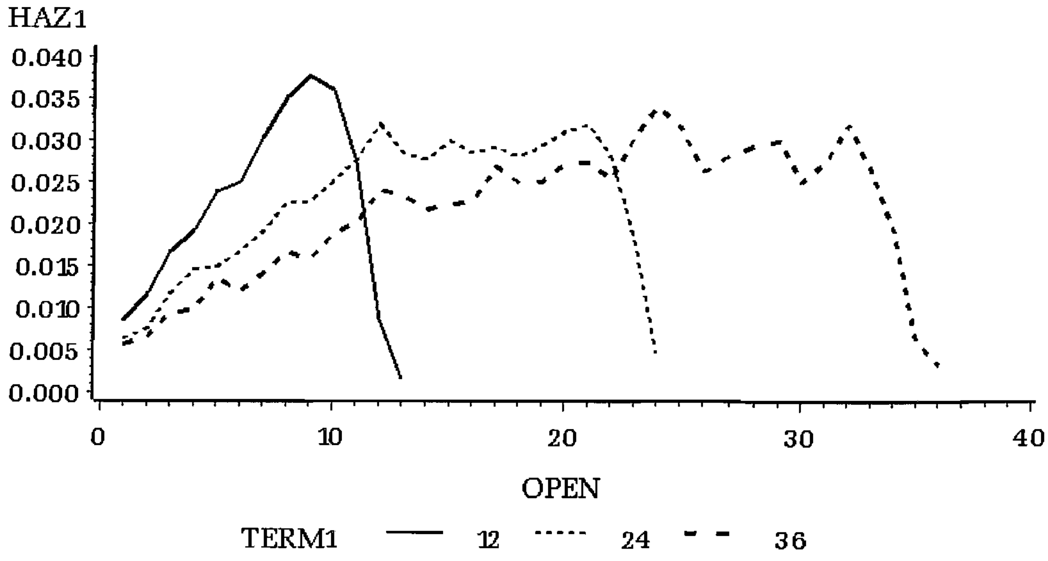


Figure 3.8: *Early repayment hazard for three different terms of loan.*

Early hazard rate plots for 12, 24 and 36 months are shown in Figure 3.8. The hazard rates for the three different loan terms do not seem to be part of one general pattern as for default. Early repayment hazard rate rises through the course of a loan, reaching its maximum three to four months before the end of a loan and then falls rapidly for all three terms shown. These feature may be seen more easily if the time scale is 'reversed', i.e. transformed from *time from the start* of a loan into *time left to maturity* of a loan.

Plotting the hazard rates against *time left to maturity* of a loan (Figure 3.9) made the interpretation easier. It can be seen that all the hazards have a peak about three to four months to a loan's maturity at the same rate. This high repayment rate is believed to be caused by the lenders themselves when they offer a re-financing option to their customers at about 3 months to completion. Other peaks at 12, 24 and 36 months are probably due to the large number of customers repaying at the 1, 2 or 3 year mark of the duration of a loan. The overall picture is such that the hazard rates for 24 and 36 months loans, for example, climb up to some level at about 12 months to a loan's end and then follow the same pattern till the end.

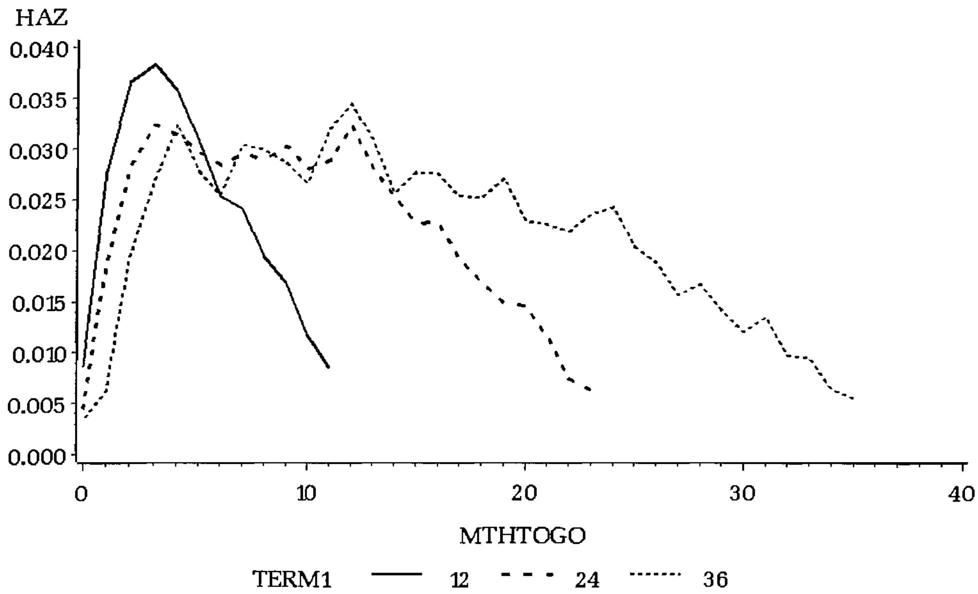


Figure 3.9: *Early repayment Hazard rate for different terms of loans, plotted against number of months remaining to final repayment*

This idea of the transformed time scale can be used in the graphical exploration of data as well as in fitting a survival analysis model.

Hazard plots against time to maturity in Figure 3.9 suggest that early repayment is influenced most by the time left to the maturity of a loan. However, logistic regression and proportional hazards are using the time from the start of a loan. If one segments the model by term, the *time from the start* is a linear transformation of the *time to maturity*, hence it is equivalent to fitting the model to the *time to maturity*. Indeed, the *time to maturity* is the difference between the *time from the start of a loan* and the term of the loan. So if one segments the models by the Term of a Loan, the *time to maturity* is the *time from the start* minus a constant (Term). This is why segmenting by the Term of a Loan is so useful for early repayment in this data set.

3.4.2 Effect of application characteristics on the hazard rate

We have plotted hazard rates for different terms of loans. In the same fashion one can look at the hazard rates for different values of any other characteristic, such as the Purpose of a Loan, the Amount, or the Age of an Applicant, etc. This can be a part of pre-modelling data exploration to help in understanding of which characteristics are likely to be significant and whether they have a positive or a negative effect on the survival time, i.e whether they increase or decrease the hazard rate.

Figure 3.10 shows that default hazard rate for the high risk purpose group containing refinance is much higher then for all other purposes. Hence, we should expect a positive and a significant value of β for the corresponding indicator variable in the proportional hazards model.

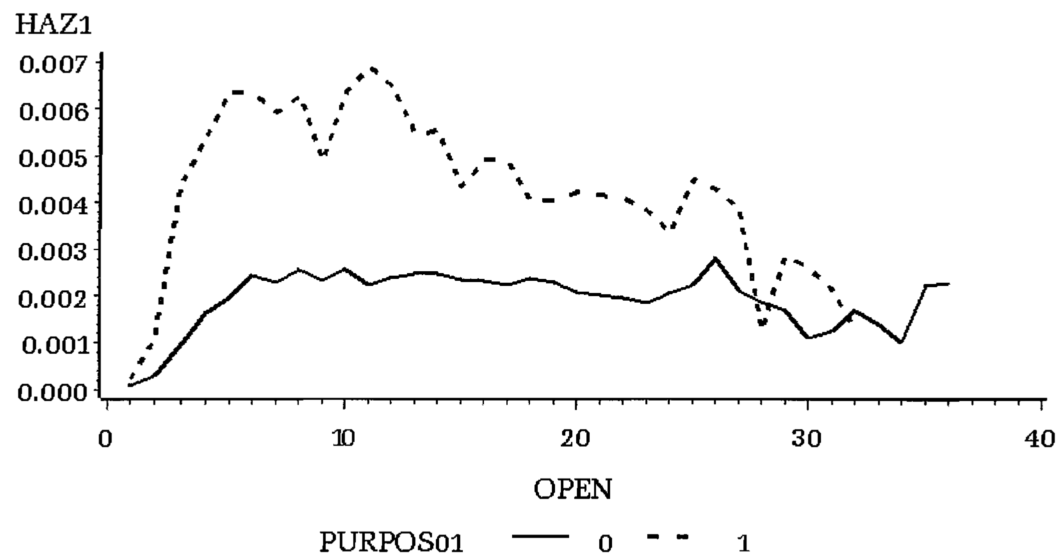


Figure 3.10: *Default hazard rate for two groups of loan purposes*

Another use of such plots is assessing whether the proportionality assumption – the assumption needed for Cox's model to be valid, Section 2.4 – holds. So one has to check that the hazards do not cross, i.e. proportional to each other. The hazards on Figure 3.10 can be assumed to satisfy this assumption. Although they cross and then return to the original position at one point at the end of the observation time period, the amount of data there is small, so this can be regarded as a random fluctuation.

Other methods of checking the proportionality assumption include various residuals which are discussed in detail in Section 2.6 and time dependent covariates in Chapter 4.

3.5 Predicting early repayment

3.5.1 Model construction

Cox's proportional hazard model was fitted to loans' survival time until early pay-off. In line with the competing risk approach discussed in Section 2.5, the lifetimes of the loans that are paid off early are considered 'failures' while all others are considered censored.

Models were built on a training sample and tested on a holdout. Tables 3.8, 3.9, 3.10 show the indicator variables that were included into the early repayment model as covariates (or explanatory variables). These indicator variables were created from the continuous and the discrete characteristics using the survival analysis based coarse classing method from Section 3.3.

Indicator var	Applicants age between
age01=1	18 and 21
age02=1	21 and 25
age03=1	25 and 33
age04=1	33 and 39
age05=1	39 and 57
agem=1	missing
	Amount of a loan between
amount01=1	500 and 800
amount02=1	800 and 1850
amount03=1	1850 and 2950
amount04=1	2950 and 8000

Table 3.8: *Indicator variables constructed from continuous characteristics, part I.*

Indicator var	Years at current address between
currad01=1	0 and 1.4
currad02=1	1.4 and 3.8
currad03=1	3.8 and 7.4
currad04=1	7.4 and 11.7
currad05=1	11.7 and 24.2
curradm=1	missing
	Years with current employer between
currem01=1	0.4 and 2.1
currem02=1	2.1 and 2.9
currem03=1	2.9 and 6.6
currem04=1	6.6 and 12.4
currem05=1	12.4 and 99
curremm=1	missing
	Insurance premium between
inprem01=1	27 and 45
inprem02=1	46 and 105
inprem03=1	105 and 270
inprem04=1	270 and 458
inprem05=1	more that 458

Table 3.9: *Indicator variables constructed from continuous characteristics, part II.*

Indicator var	Purposes of loans
purpe01=1	6,10,23,4
purpe02=1	16,8,12,13,11,24,19,26
purpe03=1	1,27,3,22,20
	Applicant's Gender
gender01=1	female
	Frequency of applicant's salary payment
freqpa01=1	weekly
freqpa02=1	missing
	Number of dependent children
depkid01=1	1 to 3
depkid01=1	4 or more
	Home owenership
homown01=1	own
	Home telephone
hometel01=1	yes
	Loan type
joints01=1	single
	Applicant's marital status
wedded01=1	single or divorced
wedded02=1	married or widowed

Table 3.10: *Indicator variables constructed from discrete characteristics.*

The general form of the proportional hazards model fitted was:

$$h(Open) = exp(\beta_1 * Var_1 + \ldots + \beta_k * Var_k) * h_0, \tag{3.8}$$

where $Var_1, \ldots, Var_2$ are the indicator variables constructed from application characteristics and loan specifications and $\beta_1 \ldots \beta_k$ are the parameters to be estimated.

All the variables were included in the model to allow for a consistent comparison of the PH models between themselves and with a logistic regression model. However, in practice, one may choose to use a selection mechanism, such as provided by stepwise, backward or forward options in SAS PHREG procedure, to reduce the number of covariates. In the forward selection method the procedure starts with finding the most significant covariate and adding it to the model. The process is repeated until there are no covariates left which are significant according to the specified entry significance level. In the backward selection method the procedure starts with all the covariates in the model, then finds the least significant one among them and removes it from the model. The process is repeated until there are no covariates left in the model with the significance level less than the specified staying level. In the forward and the backward selections once the variable is added or removed from the model it stays in or out respectively.

The stepwise selection is similar to the forward selection except that the covariates that have been added to the model can be removed if their significance level drops below the staying level.

SAS PHREG procedure output is shown in Appendix A. It can be seen that the Term of a Loan is the most significant of the characteristics.

Recall that in Section 3.4.1 we have shown that the relationship between early repayment and time left to maturity of a loan is much stronger than that between early repayment and time from the start of a loan. The *time to maturity* is the difference between the *time from the start* of a loan and the term of the loan. So if one segments the models by the Term of a Loan, the *time to maturity* is the *time from the start* minus the Term of a Loan (constant). Hence, fitting the segmented by term model to the *time from the start of a loan to its early repayment* is equivalent to fitting the model to *time from the early repayment of a loan to its maturity* .

The PHREG output showing the parameter estimates for the segmented models is in Appendix A.

The next section compares both the segmented and the non-segmented proportional hazards models to the segmented and the non-segmented logistic regression models respectively.

3.5.2 Comparison of Proportional Hazards models with Logistic Regression models

The results presented compare Cox's proportional hazards model (PH) with the logistic regression approach (LR) under the two criteria:

1. Estimating which loans will be paid off early within the first 12 months (Table 3.11, 1st year).
2. Estimating which loans which are still repaying after the first 12 months will pay off early within the next 12 months (Table 3.11, 2nd year).

		1st year				2nd year			
		G-pG	G-pB	B-pG	B-pB	G-pG	G-pB	B-pG	B-pB
	Actual Nos	11964	0	0	2928	6274	0	0	1825
PH	Non-segment	9802	2162	2162	**766**	4843	1431	1431	**394**
	Segm by Term	9765	2199	2199	**729**	4981	1293	1293	**532**
LR	Non-segment	9820	2144	2144	**784**	4984	1289	1289	**536**
	Segm by Term	9768	2196	2196	**732**	5000	1273	1273	**552**

G-pG - 'good' predicted as 'good'; G-pB - 'good' predicted as 'bad';
B-pG - 'bad' predicted as good; B-pB - 'bad' predicted as 'bad'.

Table 3.11: *Predicting early repayment (Personal loan data).*

Two separate LR models were built for each of these definitions. One PH model was fitted to the times until early pay off, considering all other outcomes as censored.

To compare the LR and the PH models, the latter was measured under the two criteria above in the following way:

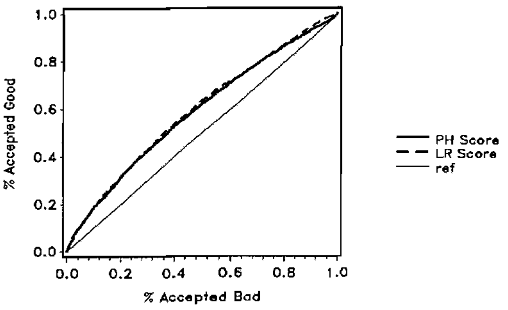
1. PH model gives an ordering of the relative likelihood to pay off early, i.e. for each customer there is a 'score' which reflects the estimated likelihood to pay off early relative to others.

2. The cutoff is then chosen in both the PH and the LR models, so that the number of predicted 'bads' equals the actual number of 'bads' in some hold-out sample, i.e. 2928 customers that are most likely to pay off early according to the PH model are considered 'bad'.

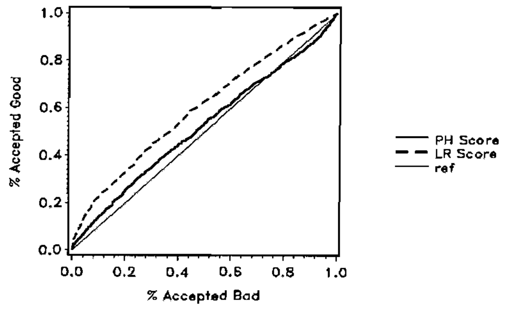
3. The numbers of 'bads' and 'goods' correctly classified by the models in the hold-out sample are then compared between the models (see Table 3.11).

In the case of non-segmented models, LR slightly outperforms PH. This is not surprising since only one PH model was used to classify loans according to both definitions as opposed to two LR models fitted specifically to each definition. Notice that the PH model loses a lot of power in the second year. On investigating this further, it was realised that the Term of a Loan was a very significant characteristic that correlated strongly with the other variables including the time to early repayment. When the models were segmented by Term, see Table 3.11 and their performance compared in the same way as for the non-segmented ones, the results for PH model were much better. In the second year in the segmented case the PH model performs better than the LR model. This suggests that the time to early repayment is strongly affected by the Term of a Loan. This is supported by the discussion of Section 3.4.1 and Figure 3.9 which suggest that the critical measure for early repayment of a loan is how much longer until maturity, rather than its current duration. Thus if one is using the duration of a loan in a PH model, one can only translate this into the time until the maturity for a set of loans if they have the same maturity.

To show that these results are not artifacts of the cut-off chosen, ROC curves for each scorecard were produced. Figure 3.11 shows the ROC curves for the PH and the LR models without segmentation and confirms the poor performance of the PH in the second year. Figure 3.12 shows the ROC curve results when the data is segmented by term. a) corresponds to loans of 12 and 18 months. Only early repayment in the first twelve months can be observed for this segment, since there are too few loans with the 18 months term; b) and d) are the results for loans with terms of 24 and 30 months, b) being the ROC curve for early repayment in the first year of a loan and d)early repayment in the second year of a loan; c) and e) are the similar ROC curves for loans which were to be repaid in 3 or more years. The PH and the LR ROC curves are very similar to one another.



a)



b)

Figure 3.11: *ROC curves for PH and LR predicting early repayment: a) 1st year, b) 2nd year*

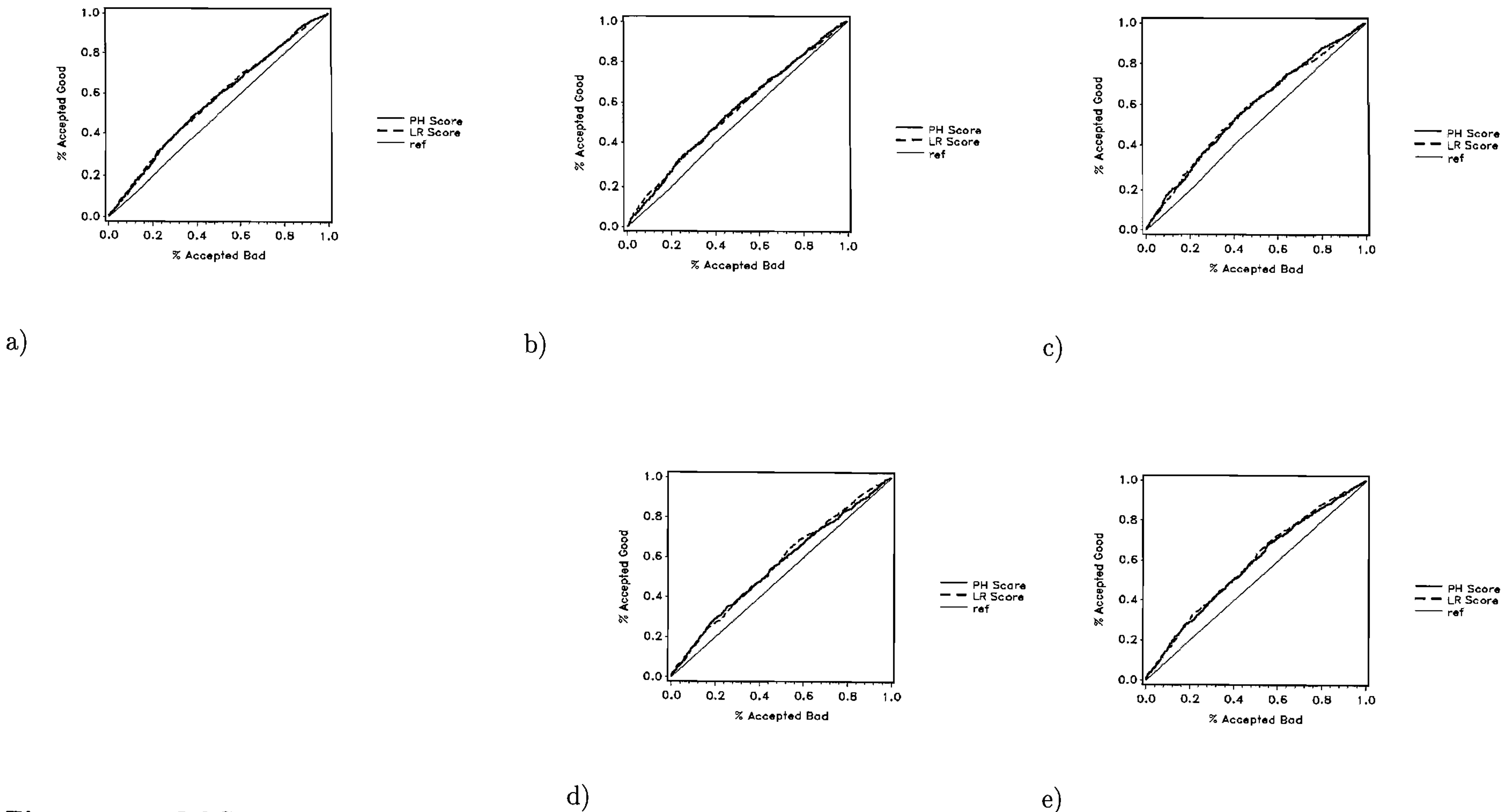


Figure 3.12: *ROC curves for PH and LR predicting early repayment for different loan terms: a)-c) 1st year, d)-e) 2nd year*

Note that the LR segmented model is worse than the non-segmented one. Sometimes this happens because although segmentation should result in a better model, since it builds a different model for each segment, it also has much less data for each of these models, so the performance may be affected, (Banasik et al. [1996]). However, in our case the difference is large enough to deserve an investigation into its causes. Note that we will only be concerned with the logistic regression models for the 1st year in this investigation.

The choice of a cutoff based on matching the number of predicted 'bad' to the number of real 'bad' for each the term sub-groups can be a cause of the problem if the real number of 'bad' in one of the sub-groups in the holdout sample does not reflect the expected 'bad' rate for this sub-group. In other words, it seems more or less risky than expected.

Hence, three different methods of choosing a cut-off were tried out:

1. by matching marginal 'bad' rates between the term sub-groups ;

2. by matching 1-specificity (horizontal axes of a ROC curve) ;

3. by matching sensitivity (vertical axes of a ROC curve).

However, none of the above cut-offs improved the match rate.

It was noticed that if one treats the scores from the four segmented models as scores on the same scale and chooses the cut-off as in the non-segmented case by matching the total number of predicted and actual 'bad' in the whole hold-out sample, then the segmented model performs better than the non-segmented. The number of correctly classified 'bad' is 792, not 732 as in Table 3.11. If one now looks where this cut-off has 'cut' the term sub-groups (Table 3.12), it is clear that the 12 and 18 months group is very different from the rest. It is much more risky then the others, since only about 50 percent of the actual 'good' and 'bad' were accepted as opposed to 70 to 98 percent in the other sub-groups, but the actual number of 'bad' in the hold-out sample does not reflect this. That is why choosing the cut-off based on the actual numbers of 'bad' in the sub-groups resulted in a poor prediction.

Term Sub-group	1-Specificity	Sensitivity
6 mths	0.732	0.818
12 and 18 mths	0.501	0.596
24 and 30 mths	0.875	0.892
36,48 and 60 mths	0.980	0.983

Table 3.12: *Percentages of accepted 'bad' (1-specificity) and accepted 'good' (sensitivity) in the term sub-groups corresponding to the cut-off chosen by matching the bad-rate in the whole hold-out sample.*

Furthermore, if one compares predictive power of the non-segmented model on the term sub-groups rather than on the whole sample with the segmented model it can be seen they are comparable, see Gini coefficients in Table 3.13. In fact, the segmented model wins in three out of the four term sub-groups.

	Gini coefficient	
Term Sub-group	Non-segmented	Segmented
6 mths	0.0500	0.1629
12 and 18 mths	0.1091	0.1272
24 and 30 mths	0.1042	0.1059
36,48 and 60 mths	0.1546	0.1443

Table 3.13: *Gini coefficients for the non-segmented and the segmented models applied to the term sub-groups.*

Thus the predictive power of both models is worse if they are applied to each of the term sub-groups separately rather than the whole sample. This can be explained by the fact that a large part of the discrimination is based on discriminating between the terms which is only possible if the score is applied to the whole sample. In other words, the term sub-groups are very different from one another in terms of risk, which can be illustrated by comparing their scores (Table 3.14).

	Statistic	
Term Sub-group	Mean	Std. Dev
6 mths	-1.705	0.246
12 and 18 mths	-1.175	0.227
24 and 30 mths	-1.46	0.247
36,48 and 60 mths	-1.846	0.291

Table 3.14: *Moments of the segmented score for different term sub-groups.*

The hold-out method is based on the assumption that there is enough data to choose a hold-out sample which would represent the whole population well. When a holdout sample is segmented into n sub-samples, then one is hoping to have enough data for n holdout samples that are similar to the respective subgroups in the population. This is clearly more difficult to achieve than in the case of one hold-out sample. It was seen in the case of our data, where the sub-holdout of 12 and 18 months loans was not representative of the 12 and 18 months loan population. This sub-group is riskier than others which is indicated by the lower mean score in Table 3.14, but the number of 'bad' in the hold-out is not enough to reflect this.

In summary, it has been demonstrated that the segmented LR model for early repayment in the 1st year performs worse than the non-segmented because the hold-out sample is not representative of the population for one of the segments. If the cut-off is chosen by matching the number of 'bad' on the whole hold-out rather than each segment, the segmented model performs slightly better than the non-segmented. So if the segmented model is to be used in the 1st year, care should be taken while choosing the cut-off, alternatively the non-segmented PH and LR scorecards can be used for predicting early repayment in the 1st year. However, in the 2nd year we would definitely use the segmented PH scorecard since it is much better than the non-segmented one.

SAS statistical software was used to fit both PH and LR models, using the procedures PHREG and LOGISTIC respectively. There are three ways of the treatment of ties available in PHREG procedure: 'Breslow', 'Efron' and 'Discrete', which correspond to the three different approximations of the exact likelihood as discussed in Chapter 2. The SAS statistical package recommends 'Discrete' for the data that contains a large number of ties.

	-2Log(Likelihood)	
Term Sub-group	Discrete	Breslow
6 mths	1111.873	1775.261
12 and18 mths	28956.193	57041.291
24 and 30 mths	38374.029	72658.827
36, 48 and 60 mths	39189.753	70345.125

Table 3.15: *-2Log(Likelihood) statistics from the early repayment models.*

Table 3.15 calculates the log-likelihood values obtained by fitting the proportional hazards model to the data (segmented into four groups, terms of 6, 12 and 18 months, 24 and 30 months and 3 or more years) using the 'Discrete' method and Breslow's approximation. The smaller value of the log-likelihood statistic indicates a better fit to the data. So these log-likelihood values suggest that the discrete approximation has given a much better fit in all four cases, however there was almost no difference in the parameter estimates and no difference in the number of correctly classified accounts between the two methods.

R.Peto, who proposed an approximation differing from Breslow's only by a constant, (discussion of Cox [1972]), notes that the differences between his, Cox's and the exact probability "is two orders of magnitude less than the random variation which is being analysed". In addition, Breslow's approximation is the fastest method out of the three and hence, it was used for the majority of the calculations.

3.6 Predicting default

Methods analogous to those used to predict early repayment were also used to predict default.

The results presented compare Cox's Proportional Hazards model with the logistic regression approach again under the two criteria:

1. Estimating which loans will default within the first 12 months (Table 3.16, 1st year).

2. Estimating which loans which are still repaying after the first 12 months will default within the next 12 months (Table 3.16, 2nd year).

		1st year				2nd year			
		G-pG	G-pB	B-pG	B-pB	G-pG	G-pB	B-pG	B-pB
	Actual Nos	14495	0	0	397	7915	0	0	184
PH	Non-segment	14145	351	351	**46**	7747	168	168	**16**
	Segm by Term	14149	346	346	**51**	7752	163	163	**21**
LR	Non-segment	14145	350	350	**47**	7748	166	166	**18**
	Segm by Term	14145	351	351	**46**	7752	162	162	**22**

G-pG - 'good' predicted as 'good'; G-pB - 'good' predicted as 'bad';
B-pG - 'bad' predicted as 'good'; B-pB - 'bad' predicted as 'bad'.

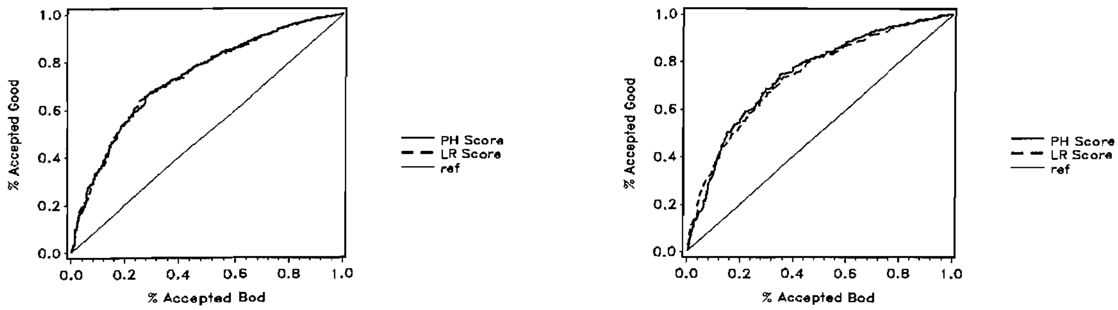
Table 3.16: *Predicting default (Personal loan data).*

Table 3.16 shows the results on a holdout sample using a cut-off where the number of 'bads' predicted agrees with the number of 'bads' in the sample. The results suggest there is a little difference between LR and PH in either the first or the second year and that the segmentation has a less dramatic improvement on PH results under the default criterion that it did under the early repayment criterion.

The ROC curves in Figures 3.13-3.14 agree with Table 3.16. Figures 3.13 shows that without segmenting by the Term of a Loan LR and PH give very similar results in both the first, a), and the second, b), years.

Figure 3.14 shows the ROC curve results when the data is segmented by term, for the three most common terms of loans. a) and d) correspond to loans of 24 months a) being the ROC curve for default in the first year of a loan and d) default in the second year of a loan; b) and e) are the similar ROC curves for loans with the term of 36 months and c) and f) are the ROC curves for loans which were to be repaid in 48 months. It should be noted that segmentation by the Term of a

Loan has less effect in predicting default than early repayment because default rate is independent of the Term of a Loan and early repayment is not.



a) b)

Figure 3.13: *ROC curves for PH and LR predicting default: a) 1st year b) 2nd year*

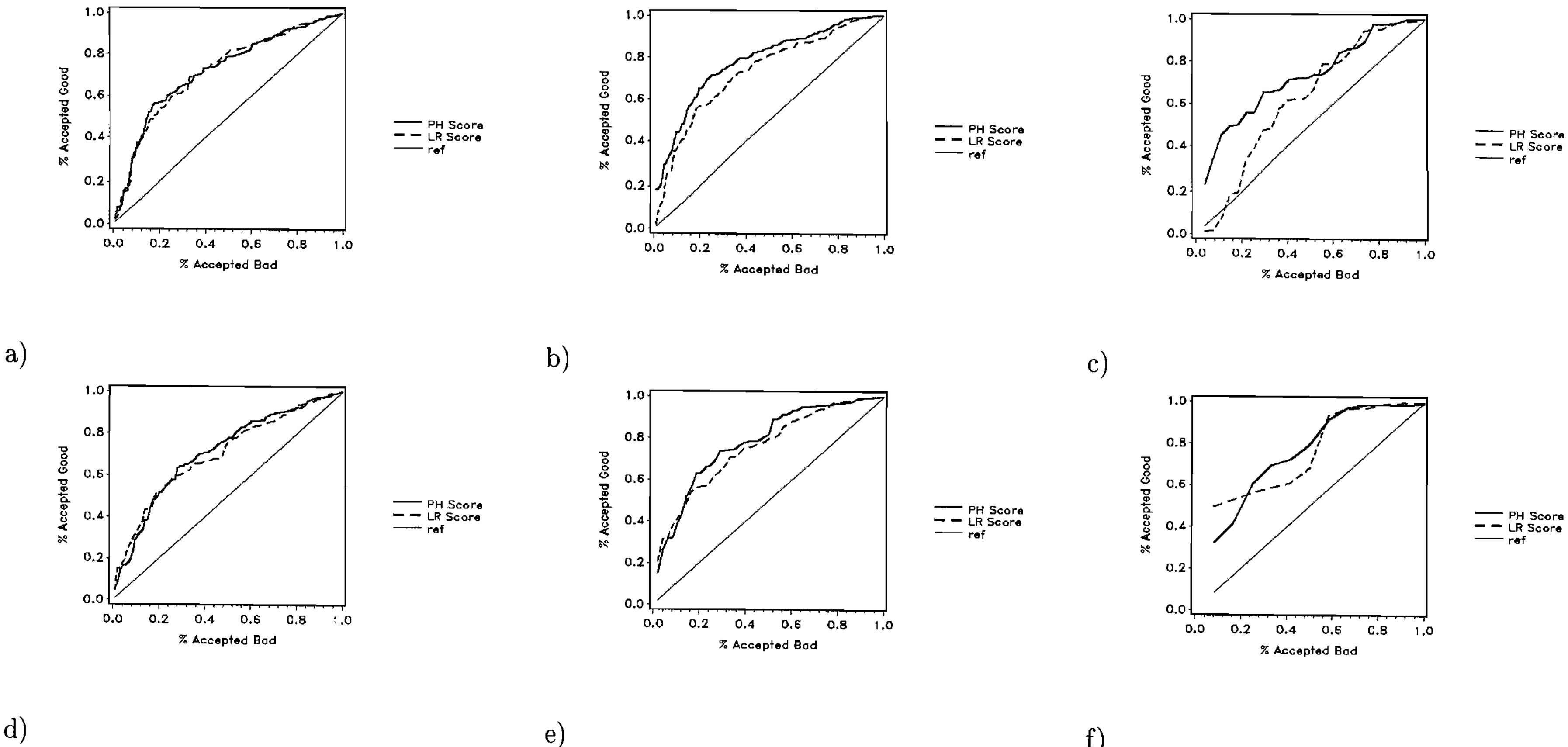


Figure 3.14: *ROC curves for PH and LR predicting default for different loan terms: a)-c) 1st year, d)-f) 2nd year*

It seems reasonable that default is a function of present and past conditions, but that early repayment also takes into the account how much longer a loan is to exist and how much more would be needed to pay it off now. Hence, early repayment has a strong relationship with the remaining time to maturity of a loan and thus, with the actual term of a loan. This explains why segmentation works so much better for early repayment than for default.

3.7 Comparison of model diagnostic methods

Martingale and Schoenfeld residuals can be requested as part of the output from the PHREG procedure in SAS.

Cox-Snell residuals were calculated from martingale residuals as discussed in Section 2.6. To examine whether Cox-Snell residuals have unit exponential distribution the product-limit estimate of the survivor function $\hat{S}(r_{C_i})$ was obtained and log-log transformation of these values was plotted against log of the corresponding residual. The $-\log S(r_{C_i})$ is exponentially distributed with unit mean irrespective of the form of $S(r_{C_i})$. So we check that r_{C_i} also has an exponential distribution by plotting $\log(r_{C_i})$ against $\log(-\log \hat{S}(r_{C_i}))$ – the log of the quantity known to be exponentially distributed with unit mean.

Figure 3.15 shows Cox-Snell residual plots for the early repayment model segmented by term, hence there are four plots, one for each term subgroup. Each circle or dot represents one observation – a loan lifetime – from the data. This is true for all the residual plots which follow.

The plotted points are close to the straight line with unit slope and zero intercept if the observations with the lowest residuals are ignored. There are only a few of those and they correspond to the loans with the shortest lifetime – one month. It is arguable whether these observation should be considered at all since repayment after one month is not necessarily a typical or normal feature of a personal loan portfolio. These are likely to be 'dirty' data in that they reflect the loans that were never really taken up. If we ignore these observations we can conclude that the model is adequate.

Martingale residuals for the early repayment model segmented by the Term of a Loan were plotted against the rank order of time. The plots corresponding to each of the four term subgroups are shown in the Figure 3.16.

The size of a loan's residual relates to the model accuracy, so that large positive value indicates that the loan failed (repaid early) sooner than expected and large negative value indicates a failure later than expected from the model.

The values appear in two bands, one representing uncensored observations, another - censored ones. This is because Martingale residuals are always negative for the censored observations. The scatter of the points within a band increases with the rank order of time. It is expected that since the calculation of r_{M_i} involves estimating a survivor function which is close to one for early times, and hence, its logarithm is close to zero, then r_{M_i} will be clustered around 1 for uncensored and around 0 for censored observations.

There are no clear outliers. Therneau et al. [1990] note that it is almost impossible to detect outliers of the 'failed sooner than expected' type because there are so many observations clustered around 1. The deviance residual transformation should help to reduce this problem, since it is designed to make the residual distribution more normal (Section 2.6).

Deviance residuals, Figure 3.17, are very similar in appearance to Martingale residuals. Still, there appears to be no outliers.

Since the number of observations is very large, it is doubtful that these plots can be as useful in identifying problems with the model as in medical studies, where a number of observations is fewer. Because of the large number of observations the patterns explained by the nature of the residual are clearly visible and overshadow any other systematic features or outliers.

Schoenfeld residuals were plotted to investigate whether the effect of a covariate on the survival time changes over time, i.e. whether the proportional hazard assumption holds.

Figures 3.18, 3.19 show examples of Schoenfeld residuals plots against the rank order of time for two covariates, FREQPA01 and CURRAD01, for the early repayment model segmented by the Term of a Loan.

Since all the covariates are binary, the residual is either $1 - E(x_i|R(t_i))$ when $x_i = 1$ or $-E(x_i|R(t_i))$ when $x_i = 0$. Hence, there are two lines of points on the plot.

This diagnostic is very laborious when the number of covariates is as large as in our data, because there is a plot corresponding to every covariate for each term subgroup.

There appears to be an approximately equal number of the Schoenfeld residuals at equal distances from 0. The plots do not show any signs of time-dependency or of non-proportionality.

However, it is plausible in credit scoring that some characteristics are more important as predictors of failure at the beginning of a loan and lose their significance later. Therefore more diagnostic methods, graphical and numeric, were employed to investigate a possibility of a time-dependent covariate effect, (Section 4.1).

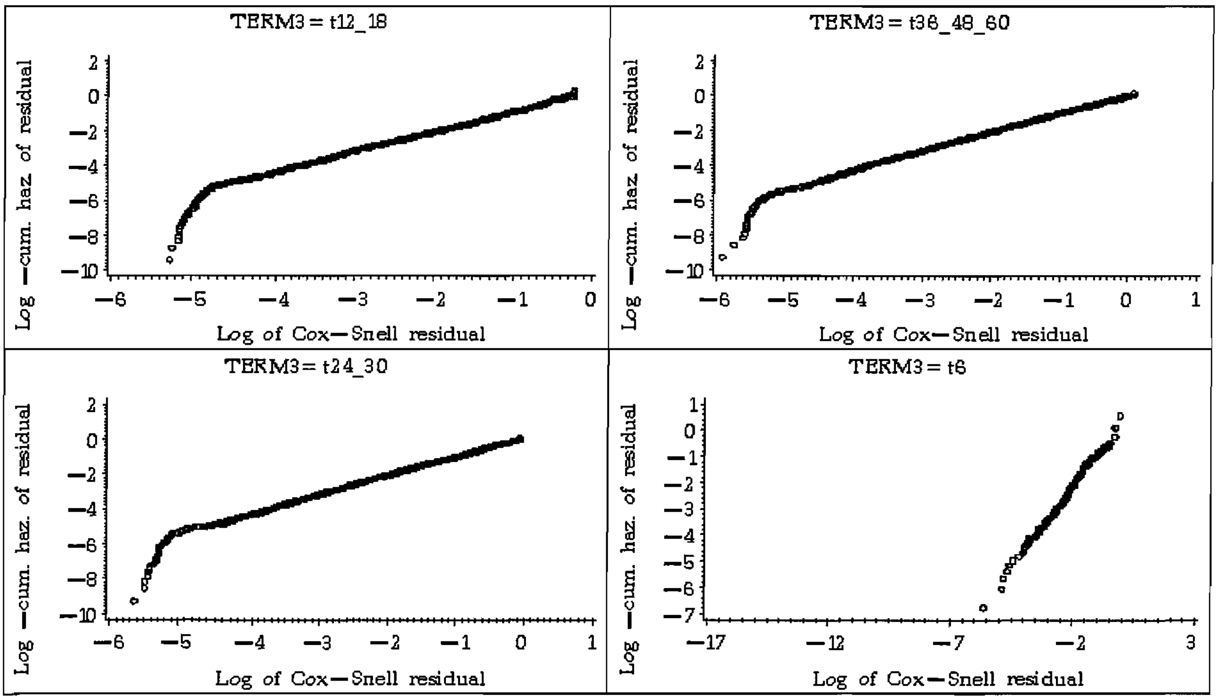


Figure 3.15: *Cox-Snell residuals*

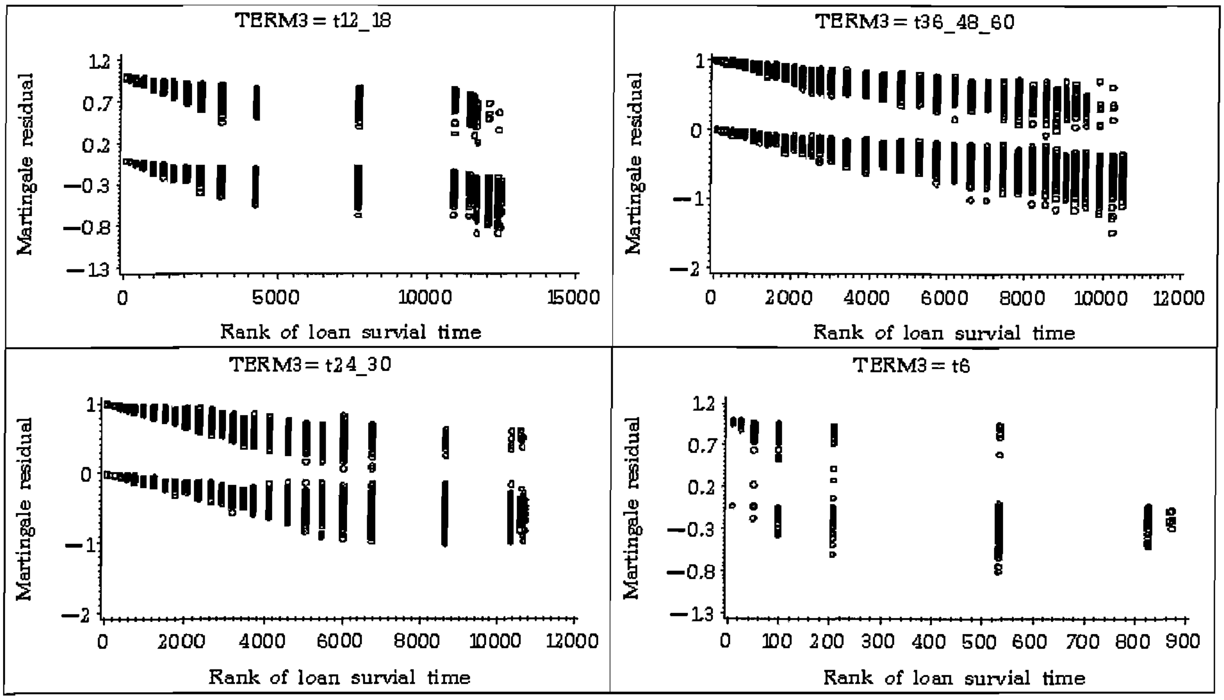


Figure 3.16: *Martingale residuals*

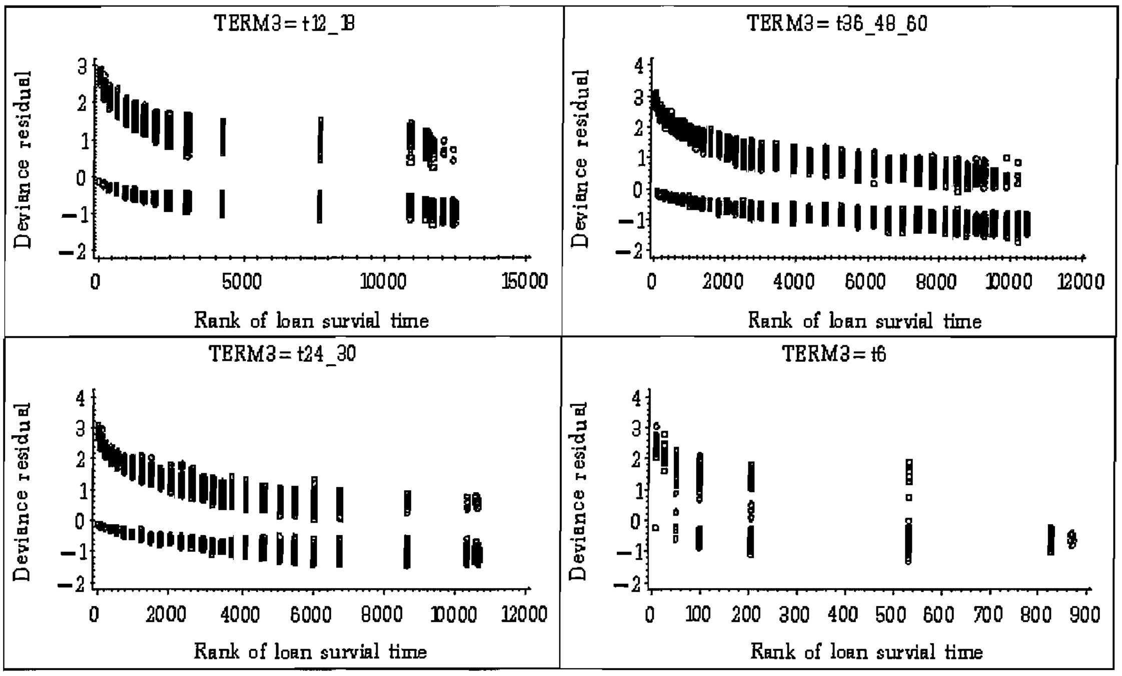


Figure 3.17: *Deviance residuals*

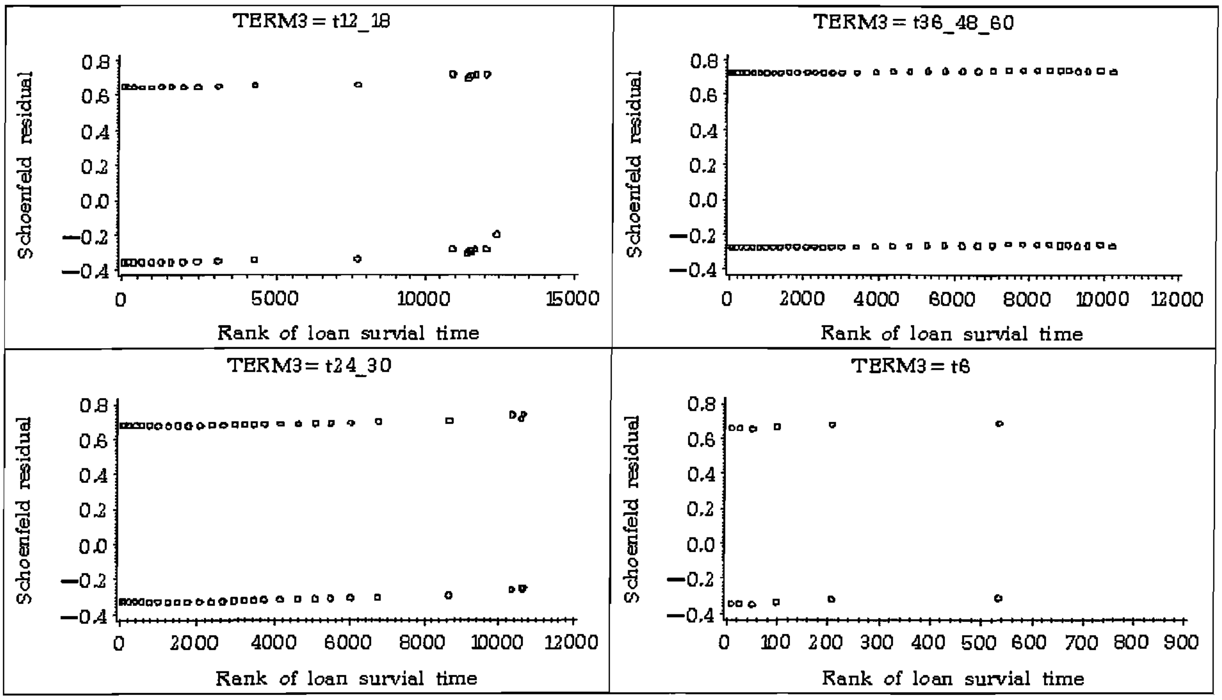


Figure 3.18: *Schoenfeld residuals for FREQPA01*

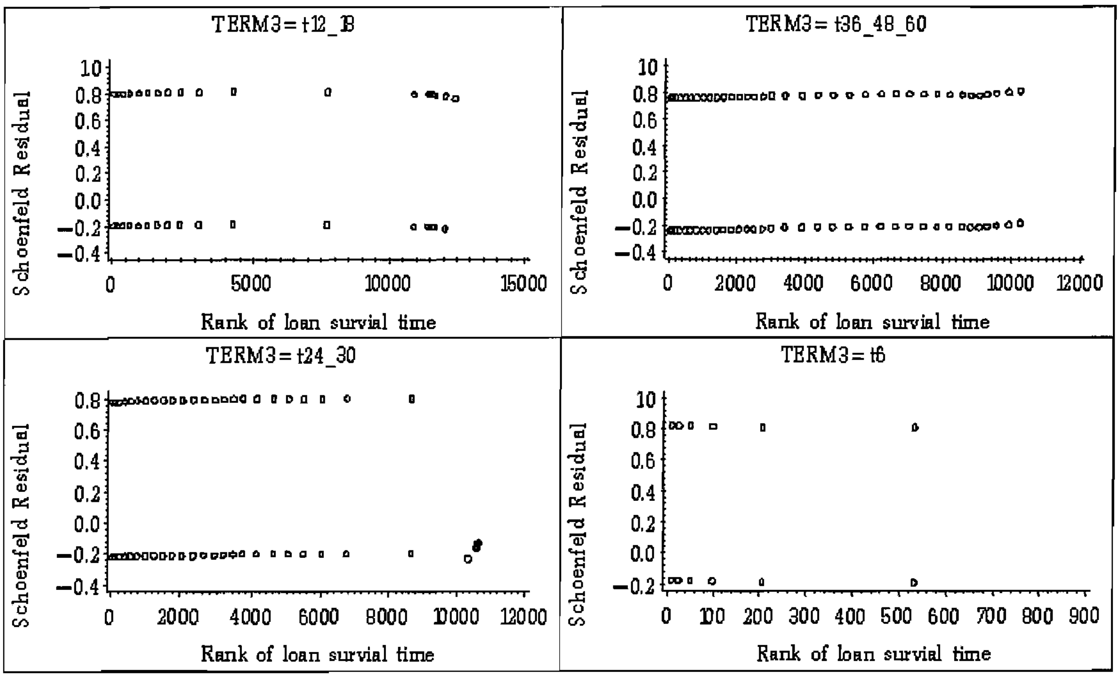


Figure 3.19: *Schoenfeld residuals for CURRAD01 (0-1.5 years at current address)*

3.8 Conclusions

This chapter demonstrated how survival analysis can be used to build credit scoring models. The summary of the results and conclusions is as follows:

First, the general conclusions:

- Proportional hazards coarse-classing of characteristics has a more consistent performance over the different definitions of 'bad' than the traditional log-odds method, because it uses time to an event, rather than a specific 'bad' definition.
- Plots of hazard functions provide a powerful visual tool that allows one to examine the probability of an event, such as default, over time, can suggest a segmenting characteristic, and provide preliminary assessment of whether a proportional hazards model is suitable.
- Proportional hazards models are competitive in their classification performance with logistic regression (see data specific conclusions below).
- Residual plots should be used to examine model fitness, outliers and possible time-dependency. Cox-Snell residual is the most popular in literature and was found to be suitable for credit scoring data. Martingale and Deviance residuals are not very informative in the case of large data sets and credit data sets are normally large enough to cause this problem. Schoenfeld residual is a very laborious diagnostic in the case of many characteristics, which is often the case in credit data.

The conclusions specific to the data set (further research is needed to see if these are general):

- The hazard rate of default is almost constant for about two years and then decreases slowly. It is consistent with the common notion "if they go 'bad' they go 'bad' early" if one believes that two years is 'early'.

- Features of the hazard rate of early repayment were seen more clearly after segmenting the data by the Term of a Loan. It slowly rises reaching its peak at 3 to 4 months to the loan's end and then falls. This suggested that early repayment is influenced most by the time left to maturity of a loan and that models predicting time to early repayment should be segmented by the Term of a Loan.

- ROC curves showed that the early repayment PH model performs very closely to the LR model when predicting early repayment in the first year of a loan. However, the segmented model should be used in the second year.

- In the case of predicting default, the PH and the LR models give very similar results. The term-segmented PH outperforms the segmented LR for 36 months loans.

- Residual plots for the early repayment PH model show that the model fits data well except for loans with very short repayment times, which might be not valid data anyway. There is no indication of either outliers or noticeable time-dependency.

Chapter 4

Time-Dependent Effects of the Covariates

In this chapter we return to the model formulation to show one of the extensions of Cox's model – time-by-covariate interaction – and to illustrate it on the personal loan data set from Chapter 3. Suppose we have just one binary covariate x_1:

$$\begin{aligned} x_1 &= 1 \quad \text{if purpose of a loan is refinance,} \\ x_1 &= 0 \quad \text{otherwise.} \end{aligned}$$

Cox's model gives the hazard of a customer to default at a time t as a baseline hazard multiplied by some function of a covariate value,

$$h(t; x_1) = e^{(\beta_1 x_1)} h_0(t). \tag{4.1}$$

If the purpose of a loan is not refinance, then $x_1 = 0$, and the hazard is equal to the baseline hazard :

$$h(t; x_1 = 0) = h_0(t). \tag{4.2}$$

If however the purpose is refinance, then $x_1 = 1$, and the hazard is e^{β_1} times higher than the baseline:

$$h(t; x_1 = 1) = e^{\beta_1} h_0(t). \tag{4.3}$$

e^{β_1} is called the relative hazard. Notice that it is independent of time under Cox's proportional hazards model. So no matter how long a loan stays on the books, if the purpose is refinance, it will always be considered more likely to default when compared to other loans. This may not really be the case. To check this a variable $x_2 = x_1 t$ is defined, which represents an interaction of the refinance indicator with time. Then this variable is added to the model (4.1):

$$h(t; \mathbf{x}) = e^{(\beta_1 x_1 + \beta_1 x_2)} h_0(t). \tag{4.4}$$

Notice that now the relative hazard for loans on refinance to others is $e^{(\beta_1 + \beta_2 t)}$, which depends on time and hence, allows for a change in the effect of refinance on default over time.

This approach was proposed by Cox [1972] as a test for the assumption of proportionality. If the time-by-covariate interaction is significant the assumption does not hold, since the ratio of the hazards is not constant. Stablein et al. [1981] actually fitted the model including a time-by-covariate interaction to a data set to account for non-proportional hazard functions. The authors showed that the inclusion of a time-by-covariate interaction is not only a means for testing for non-proportional hazards but also a modelling technique for when hazards are indeed non-proportional.

4.1 Tests for time-dependency

In consumer credit data the number of predictor variables, or characteristics, is usually large. So before including a time-by-covariate interaction in the model, it is desirable to screen all the characteristics for a possible time-dependent effect. A large number of graphical and numerical tests were developed for testing for time-dependency and are discussed in the following two sections.

4.1.1 Harrel's test and other numerical tests.

Ng'andu [1997] compares performance of the five most popular numerical tests for different scenarios of non-constant hazard ratio:

1. time-dependent covariate method (described above);

2. Harrel's linear correlation test ;
3. weighted residuals score test;
4. score process;
5. omnibus test;

To explain what each of these tests does, let us write down a general form of a model including time-dependency as:

$$h_1(t) = exp(\beta x_1 + \theta g(t) x_1) h_0(t), \tag{4.5}$$

where $\beta + \theta g(t)$ can be also seen as one time-dependent coefficient $\beta(t)$.

The idea of the omnibus test is to assume that the unknown functional form of time dependency can be represented by a piece-wise constant function (Pettitt and Daud [1990]). So that if one chooses n time intervals, then time dependency in (4.5) is of the form $g(t) = c_k$, $k = 1 \ldots n$ where c_k are unknown constants. Then the hypothesis of all $c_k = 0$ is tested using a specially constructed statistic proposed by Moreau et al. [1985].

The score process $L_{ij}(\beta_j, t)$ for a covariate x_j and subject i used for test 4 is constructed by taking a derivative of the partial likelihood function (2.11) with respect to β_j (jth component of the vector β) (Therneau et al. [1990]). Then Therneau et al. [1990] showed that the assumption of proportionality might be rejected for a particular covariate x_j if the supremum of the sum of the score processes for all the subjects for this covariate – $\sup_t \sum_i L_{ij}(\beta_j, t)$ – has a 'large' value.

Weighted residual score test (Grambsch and Therneau [1994]) is based on the Schoenfeld residuals, see Section 2.6. Grambsch and Therneau [1994] showed that if one wants to test for a time-dependent coefficient $\beta(t)$, in other words non-proportional hazards in Cox's model, the function $\beta(t)$ can be visualised by smoothing a plot of appropriately scaled Schoenfeld residuals.

Harrel's test is based on Fisher's Z-transform of the Pearson correlation between Schoenfeld residuals of a model under consideration and rank order of time. Fisher's Z-transform is a method for normalizing distribution of correlation coefficients. It is used here to construct a normal test statistic from the Schoenfeld residuals, which can be requested as an additional output from SAS procedure PHREG. The test statistics for testing the hypothesis of $\rho = 0$ is

$$Z = \rho\sqrt{(n_u - 2)/(1 - \rho^2)}, \tag{4.6}$$

where ρ is the correlation between Schoenfeld residuals and failure time order and n_u is the total number of uncensored observations.

So after the Fisher's transformation Z is a normal deviate, hence to test the hypothesis one can compare its value with the normal statistic. Recall that the sign of the expected value of the Schoenfeld residual depends on the sign of the time trend $g(t)$ of the covariate under consideration (Schoenfeld [1982]), Section 2.6). Hence, if trend is positive, the hazard ratio increases over time: $h_1(t)/h_0(t) = exp(\beta x_1 + \theta g(t)x_1)$ and also, the correlation of the residual with rank order of time is positive ($\rho > 0$). This means Z tends to be positive if the hazard ratio for the covariate increases over time and it tends to be negative if this hazard ratio decreases over time.

The first test, which is the time-dependent covariate method, was discussed at the start of this chapter.

According to Ng'andu [1997], the three best tests are the time-dependent covariate method, the weighted residuals score test and Harrel's linear correlation test.

Harrel's test was chosen as a screening test for including time-dependent covariates because it is close to the time-dependent covariate test in power, Ng'andu [1997], and also computationally simple compared to other tests.

4.1.2 Graphical tests

There are a number of graphical tests for time-dependency. Hess [1995] has given a comprehensive overview of the most popular methods and their variations for checking the assumption of Cox's model, that the effect of a covariate does not change over time, hence the hazards are proportional.

Rewriting the assumption of the proportionality of the hazard functions in different forms suggests a number of possible plots to test for its validity. Table 4.1 outlines some of the methods reviewed by Hess [1995] together with the relative formulae to explain the expected pattern if the assumption holds.

	Plot	Formulae	Expected Shape
1	$\log \frac{H_1(t)}{H_0(t)}$ vs t	$\log \frac{H_1(t)}{H_0(t)} = \log(exp(\beta)) = \beta$	horizontal line
2	$\log \frac{h_1(t)}{h_0(t)}$ vs t	$\log \frac{h_1(t)}{h_0(t)} = \log(\exp(\beta)) = \beta$	horizontal line
3	$H_1(t) vs H_0(t)$	$H_1(t) = \exp(\beta)H_0(t)$	straight line with slope $\exp(\beta)$
4	$\log H_0(t)$ and $\log H_0(t)$ vs t	$\log H_1(t) = \beta + \log H_0(t)$	two parallel lines

Table 4.1: *Graphical tests for the assumption of proportional hazards.*

All covariates in our data were binary and each was considered one at a time, so Cox's proportional hazard model for a particular covariate can be rewritten as:

$$h_1(t) = \exp(\beta)h_0(t) \tag{4.7}$$

or in terms of the cumulative hazard function as:

$$\int_0^t h_1(t)du = \exp(\beta)\int_0^t h_0(t)du \tag{4.8}$$

$$H_1(t) = exp(\beta)H_0(t). \tag{4.9}$$

The cumulative hazard function is used in graphical tests more extensively than the hazard function because it is a negative logarithm of the survival function, hence is simpler to compute. Note that $\log(H_1(t)/H_0(t)) = log(h_1(t)/h_0(t)) = \beta$ if hazards are proportional. However, if the proportional hazards assumption does not hold, i.e there is a time-dependency, then $log(H_1(t)/H_0(t)) = \log \int_0^t \exp \beta(t) du$ but $\log(h_1(t)/h_0(t)) = \beta(t)$. Thus the difference between plotting the log of the hazard ratio to the log of the cumulative hazard ratio is that the functional form of $\beta(t)$ can be inferred from the former, but not from the latter graph.

These tests will be compared and assessed on the example in Section 4.2.

Apart from the visual examination of the plots one can fit a linear regression to the one-line plots which would add a helpful numerical indication that can assist comparison of the plots when the number of covariates is large.

4.2 A model with time-by-characteristic interactions

A smaller version of the proportional hazards model from Section 3.6 is introduced here. It is fitted to the time to default using only ten most significant characteristics selected using stepwise procedure as covariates. This model is used to illustrate the concept of time-dependency. The parameters from the model together with the corresponding $\mathcal{X}^2$ statistics and significance probabilities are shown in Table 4.2

Covariate	Parameter estimate	Wald $\mathcal{X}^2$	Pr > $\mathcal{X}^2$
PURPOS01	1.095	330.478	0.0001
HOMOWN01	-0.431	53.924	0.0001
FREQPA01	0.671	145.761	0.0001
CURRADD04	-0.448	47.859	0.0001
WEDDED01	0.351	37.562	0.0001
CUREMP04	-0.770	73.637	0.0001
CUREMP05	-1.104	51.721	0.0001
CUREMP03	-0.394	33.044	0.0001
PURPOS02	0.388	22.981	0.0001
CURADD05	-0.883	18.163	0.0001

Table 4.2: *Parameter estimates from the proportional hazards model estimating time to default.*

The graphical tests and Harrel's test described in Section 4.1 were then applied to these characteristics to examine whether their effect on default changes with time.

Plots 1-4 from Table 4.1 for the binary variable PURPOS01, which is an indicator for loans on refinance, are shown in Figure 4.1. It can be seen that plots of $\log(H_1(t)/H_0(t))$ and $\log(h_1(t)/h_0(t))$ vs t are not horizontal lines as expected under the proportional hazards assumption. However, $H_1(t)$ against $H_0(t)$ (H-H plot) and the plot of $\log H_0(t)$ and $\log H_0(t)$ vs t do not suggest any departure from the proportional hazards.

Hess [1995] votes for the plot of $\log(H_1(t)/H_0(t))$ against t because it gives a direct assessment of the proportional hazards assumption. Also it is preferred to the two line plots, such as $\log H_1(t)$ and $\log H_0(t)$ against time, because the visual assessment of the constancy of a single curve is easier than the visual assessment of parallelism between two curves.

Figure 4.2 gives the same four plots for another indicator related to the purpose of a loan, PURPOS02, which divides the loans into vehicle related and others. For this variable all plots but H-H plot exhibit patterns suggesting violation of the proportional hazards assumption: Plots 1 and 2 are clearly not horizontal lines and Plot 4 shows that the hazards cross.

According to the plots we conclude that for both covariates there is an indication of a time-by-covariate interaction. Note that the directions of the time-dependencies of PURPOS01 and PURPOS02 are different. If loans on refinance become less risky as time goes by, since the hazard ratio decreases, vehicle loans become more risky since the hazard ratio rises.

To compare the appearance of the graphs for covariates with and without time dependency consider now covariate CUREMP04 with the graphical diagnostics shown in Figure 4.3. CUREMP04 is an indicator variable for 8.5 to 17.49 years with the current employer. According to other tests discussed later its effect on default is not time-dependent and the graphical tests seem to agree with that. The plot of $\log(H_1(t)/H_0(t))$ does not have a trend like for PURPOS01 and PURPOS02 and is close to a horizontal line. The peak at the beginning should be ignored since the definition of default is 3 or more months delinquent, hence the plots should be considered from 3 months onwards. The same can be said about $\log(h_1(t)/h_0(t))$ plot – it does not exhibit a trend. Plots of logs of cumulative hazards are close to parallel, ignoring the first three months, and H-H plot is close to a straight line.

It can be seen that examining all the possible graphical diagnostics visually takes a long time when the number of covariates is large. The time-saving alternative to the visual examination of the plots is to calculate $\log(H_1(t)/H_0(t))$ for all of the ten covariates and use linear regression (4.10) to examine the presence of a trend.

$$\log \frac{H_1(t)}{H_0(t)} = a + b * t, \quad \text{where} \quad t \geq 4. \tag{4.10}$$

One then needs to examine a table of regression estimates instead of ten plots. However, the results may be deceptive if there are non-typical very large or very small values that will influence the regression fit.

The parameter estimates, a and b, from the linear regressions of the form (4.10) and their significance for our ten covariates are presented in Table 4.3.

Covariate	a	$Pr > T$	b	$Pr > T$
PURPOS01	1.255	0.0001	-0.015	0.0001
HOMOWN01	-0.488	0.0001	-0.006	0.0001
FREQPA01	0.391	0.0001	0.006	0.0001
CURRADD04	-0.680	0.0001	0.005	0.0001
WEDDED01	0.725	0.0001	-0.005	0.0001
CUREMP04	-0.646	0.0001	-0.003	0.0603
CUREMP05	-1.800	0.0001	0.025	0.0001
CUREMP03	-0.732	0.0001	0.029	0.0001
PURPOS02	-0.311	0.0001	0.017	0.0001
CURADD05	-1.134	0.0001	0.006	0.3971

Table 4.3: *Linear regression parameter estimates fitted to the* $log(H_1(t)/H_0(t))$ *for the ten covariates in the model.*

It can be seen that all the covariates but CURADD05 and CUREMP04 have highly significant trends (b coefficients) in the log of the cumulative hazard ratio according to these results. However, we know that, for example, the effects of PURPOS01 and of CUREMP03 might be time-dependent according to other tests (Harrel's test and the time-dependent covariate method which are discussed later in the section) and they have larger parameter estimates for the trend, so one should judge by the magnitude of the parameter estimate rather than by its significance.

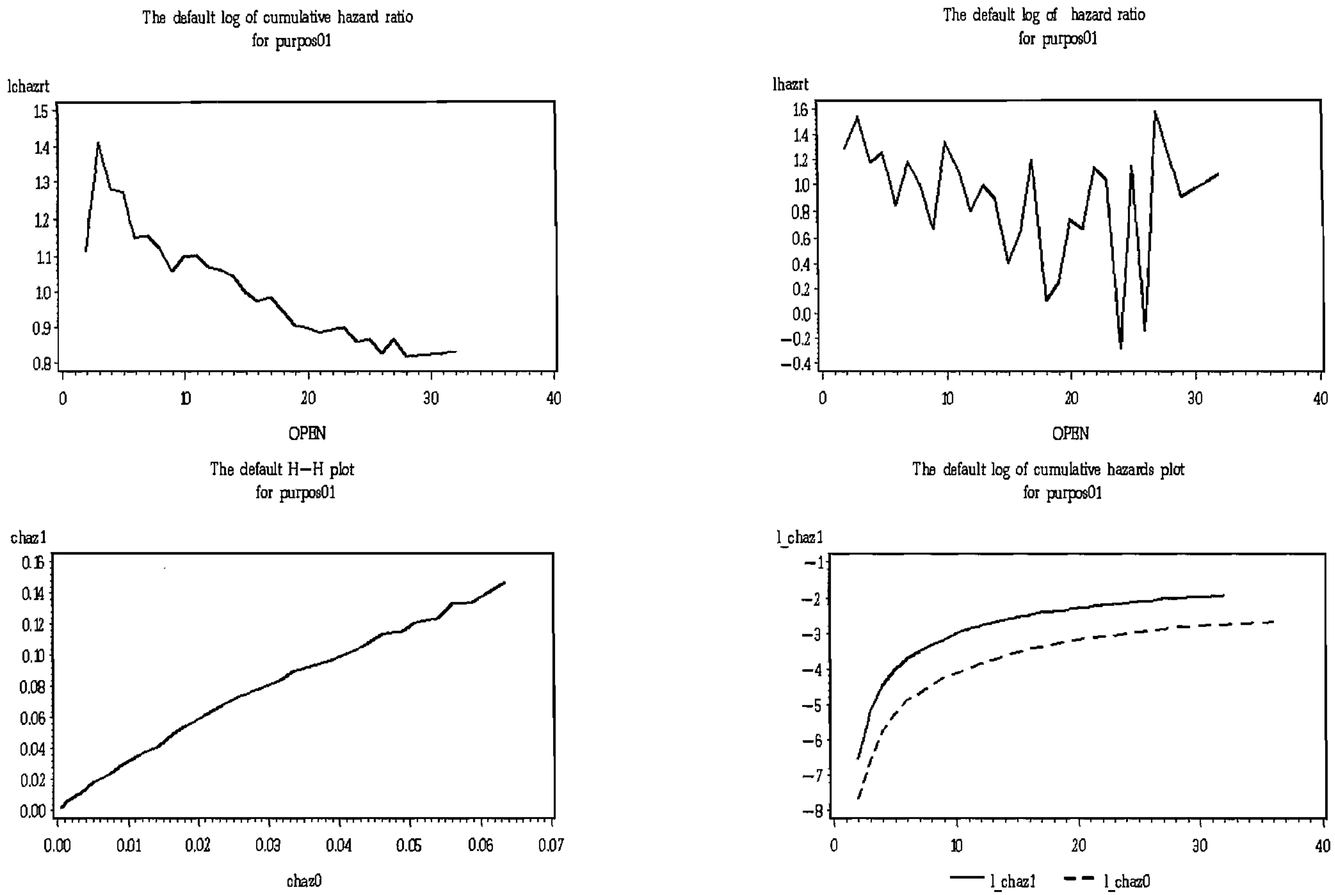


Figure 4.1: *Graphical diagnostics for PURPOS01 : the indicator for refinance.*

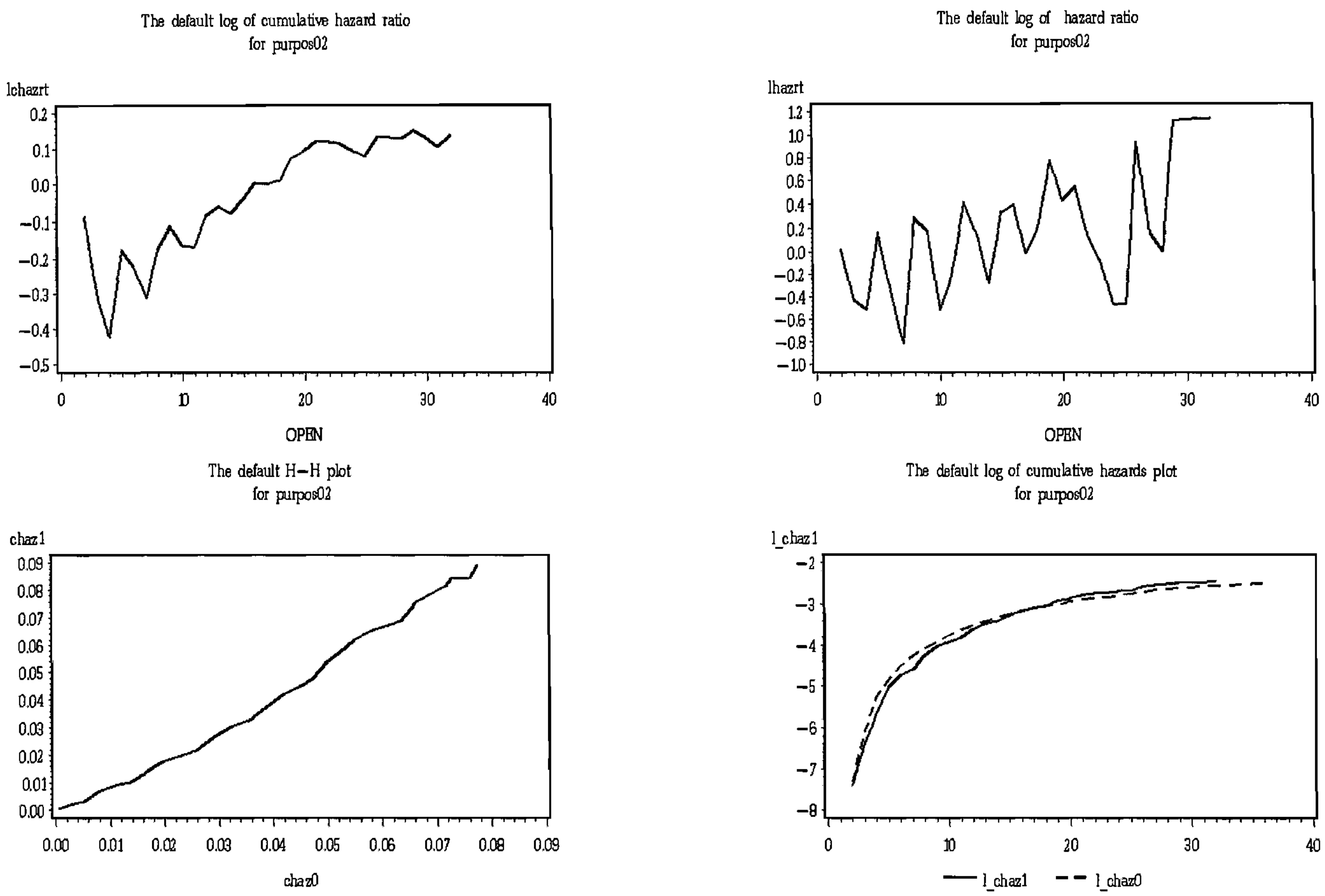


Figure 4.2: *Graphical diagnostics for PURPOS02 : the indicator for vehicle related loans.*

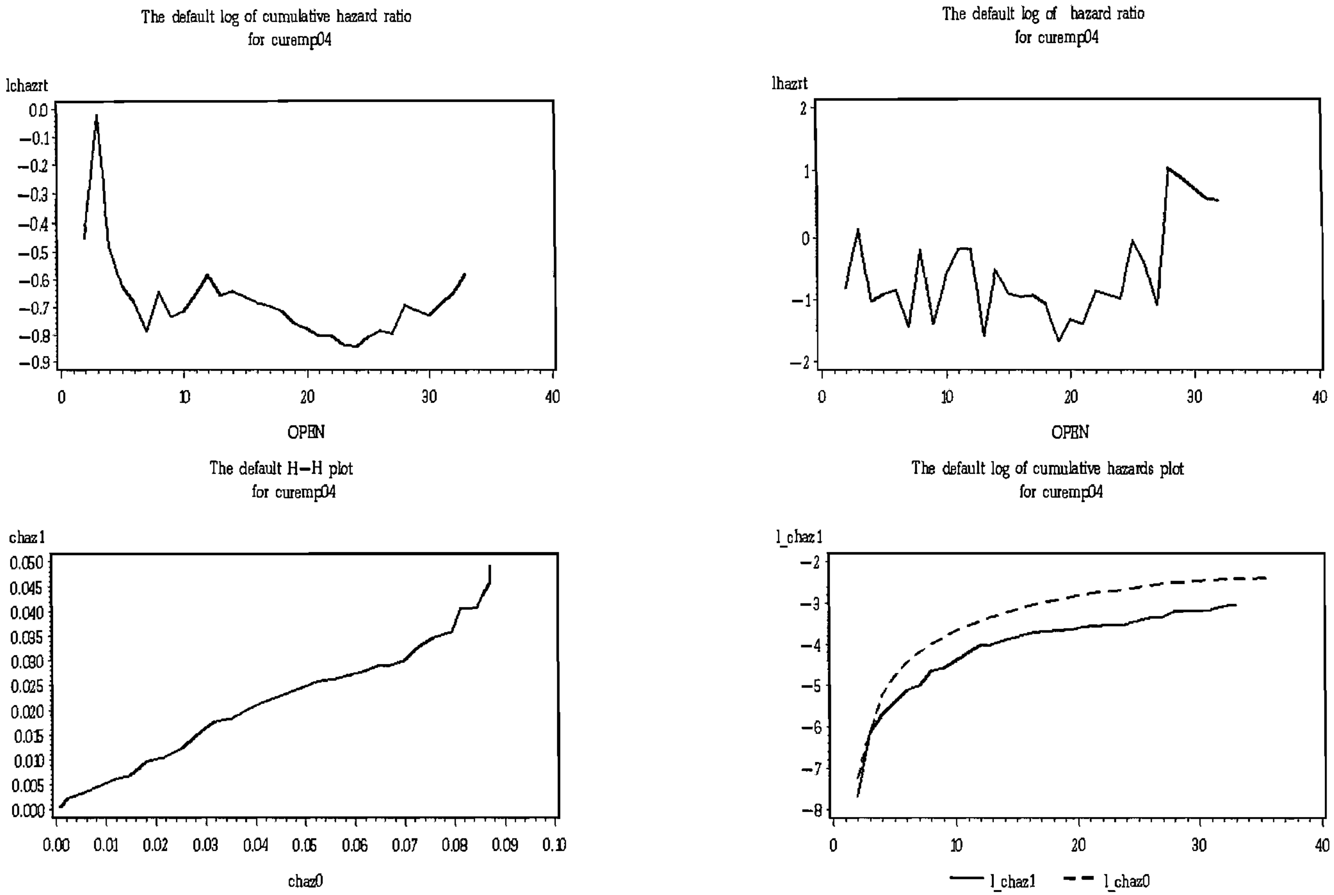


Figure 4.3: *Graphical diagnostics for CUREMP04 : the indicator for 8.5 to 17.49 years with the current employer.*

Following the comparison of the graphical tests we have performed Harrel's Z-test for all the covariates. First, Schoenfeld residuals were obtained when the proportional hazards model (Table 4.2) was fitted. Then the Z-statistic was calculated according to (4.6).

Variable	Z-statistic
PURPOS01	-19.4182
HOMOWN01	-7.8415
FREQPA01	7.6008
CURADD04	6.0507
WEDDED01	-8.1209
CUREMP04	-4.7463
CUREMP05	4.0371
CUREMP03	17.2405
PURPOS02	11.9752
CURADD05	-0.3187

Table 4.4: *Harrel's Z-statistic values for the ten covariates.*

Harrel's Z-test suggested that the hazard ratios for PURPOS01, HOMOWN1, WEDDED01 and CUREMP03 decrease over time and the hazard ratios for FREQPA01, CURRADD04, CUREMP05, CUREMP03 and PURPOS02 increase over time. The larger absolute values indicate a stronger time trend. Examples of a negative and a positive trend were also illustrated with the graphical tests. It can now be seen that graphical tests supply information about the direction and the shape of the trend. However, they do not give a clear indication of the trend magnitude which can be obtained by comparing values of the Z-statistics from Harrel's test.

For the sake of demonstrating the connection between the test results and the model with time-dependent covariates we have included a time-dependent covariate corresponding to each out of the ten covariates in the original model, so that the model has now 20 covariates in total.

The parameter estimates are shown in Tables 4.5-4.6

Covariate	Parameter estimate	Wald $\mathcal{X}^2$	Pr > $\mathcal{X}^2$
PURPOS01	1.409875	126.33331	0.0001
PRP01*T	-0.028081	8.30476	**0.0040**
HOMOWN01	-0.240712	3.93890	0.0472
HMO01*T	-0.016224	3.19398	**0.0739**
FREQPA01	0.613699	28.56601	0.0001
FEQ01*T	0.004762	0.30298	0.5820
CURADD04	-0.504642	14.07506	0.0002
CRA04*T	0.004795	0.23074	0.6310
WEDDED01	0.598879	25.06824	0.0001
WDD01*T	-0.021113	5.53145	**0.0187**

Table 4.5: *Parameter estimates from the PH model with ten covariates and ten time-dependent covariates representing interactions of a each covariate with time, Part I.*

Covariate	Parameter estimate	Wald $\mathcal{X}^2$	Pr > $\mathcal{X}^2$
CUREMP04	-0.732291	15.59530	0.0001
CRE04*T	-0.002838	0.04036	0.8408
CUREMP05	-1.390540	18.19428	0.0001
CRE05*T	0.023167	1.07541	0.2997
CUREMP03	-0.836101	32.73235	0.0001
CRE03*T	0.036467	12.41908	**0.0004**
PURPOS02	0.225265	1.75240	0.1856
PRP02*T	0.013097	1.18913	0.2755
CURADD05	-0.835429	3.80506	0.0511
CRA05*T	-0.003590	0.01292	0.9095

Table 4.6: *Parameter estimates from the PH model with ten covariates and ten time-dependent covariates representing interactions of a each covariate with time, Part II.*

Looking at the parameter estimates of the time-dependent covariates and their significance one can see that they agree with the trend directions suggested by Harrel's test but the covariate with the most significant time trend is CUREMP03, which had the second largest value of the Z-statistic. This may be because of the correlation between covariates and between time-components influencing the significance level when all are included in the model. So if one performs Harrel's test one should select a group of covariates of a reasonable size with the largest values of the Z-statistic keeping in mind that the order in terms of the time-trend significance may not be the same as the order according to the Z-statistic.

To illustrate the role of the time-by-covariate interaction, we consider one of the covariates, the indicator for the loans with the purpose of refinance, PURPOS01.

The downward trend in the $\log(H_1(t)/H_0(t))$ plot, Figure 4.1, is quite clear, which means the refinance group is a high risk group in terms of default compared to the others at the start of a loan, but becomes less so as time goes on.

It is interesting to compare parameter estimates from the proportional hazards regression predicting default when there are no time-by-covariate interactions, with parameter estimates when time-by-covariate interactions are included.

Consider, for example, PURPOS01, the indicator for the loans with the purpose of refinance. Proportional hazards regression gave estimate of $\beta_1 = 1.095$, i.e. the hazard to default for a customer on refinance is $e^{1.095} = 2.989$ times higher than for others. When the time-by-characteristic interaction was added and the model refitted, the estimates were $\beta_1 = 1.410$ and $\beta_2 = -0.02$. Hence, at $t = 1$ month into a loan the hazard for a customer on refinance is $e^{1.410-0.02} = 4.015$ times higher than for others. After 18 months, at $t = 18$, the hazard to default is $e^{(1.410-0.02*18)} = 2.858$ times higher for customers on refinance than for others. Hence, it can be seen that the parameter estimate from the model without time-dependency represents an average indication of the effect of the covariate while the addition of the time-by-covariate interaction allows for the changing relative hazard and reflects that change in the model.

Consider now CUREMP03, which is an indicator for 4.50 to 8.49 years with a current employer, it is one of the five indicators representing the continuous characteristic. It has parameter estimates $\beta_1 = -0.836$ and $\beta_2 = 0.036$ in the model which includes time-by-characteristic interactions. Note that a 'neighbouring' indicator is CUREMP04 with $\beta_1 = -0.732$ and $\beta_2 = -0.003$. So after 3 months, for example, the hazard to default for those 4.50 to 8.49 years with the current employer is $e^{(-0.836+0.036*3)} = 0.483$ times the hazards for others and the hazard to default for those 8.50 to 17.49 years with the current employer is $e^{(-0.732-0.003*3)} = 0.477$ times others. So at 3 months into a loan the groups of people 4.5 to 8.49 years and 8.5 to 17.49 years with the current employer are similar in terms of risk of early repayment. 12 months into a loan CUREMP03 is notably more risky than CUREMP04 (see graph of relative hazards in Figure 4.4). In fact, after about a year the time-dependent relative hazard is equal to that from the non-time-dependent model (shown as dotted and solid lines for CUREMP03 and CUREMP04 respectively in Figure 4.4).

Thus the inclusion of the time-dependent variables in this case shows that the difference in effects of different characteristics increases with time, after a year the time-dependent coefficient matches the non-time-dependent one and continues to increase or decrease.

This result relates to the later one, in Chapter 5, where it is shown that an application score seems to be important later on in the course of a loan, while behavioural information is predictive at the start.

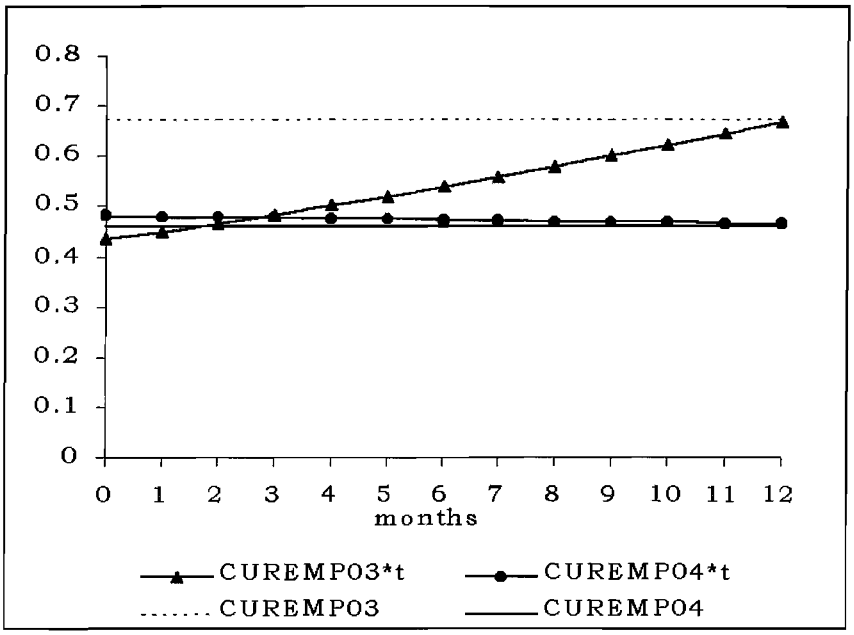


Figure 4.4: *Time-dependent and non-time dependent parameter estimates for CUREMP03 and CUREMP04.*

To compare the predictive power of the models with and without time-dependent covariates the ROC curves were plotted for the two definitions, (Figure 4.5) :

- 'bads' are those who defaulted in the first 4 months;

- 'bads' are those who survived the first 4 months and defaulted in the next 4 months.

However, for this particular data set the time trends are not strong enough to make a big difference for the predictive power of the model.

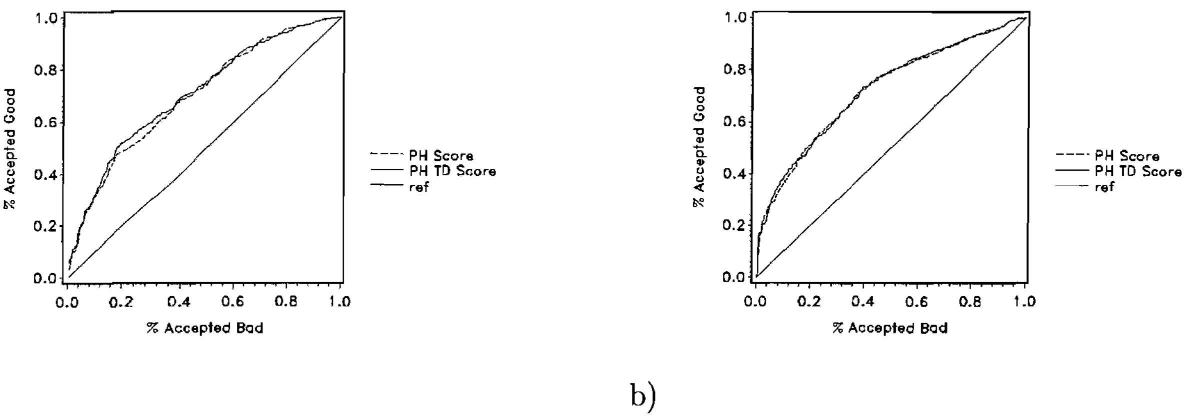


Figure 4.5: *ROC curves comparing models with and without time-dependent covariates in predicting default: a) defaut in the first 4 months; b) survived the first 4 months, default in the next 4 months.*

To investigate how the performance of the models with and without time-dependent covariates differs when a time-by-covariate interaction is larger than in our data a data set was simulated. Among two covariates in this data set, one was made to have a time-dependent effect on the survival time. The baseline hazard was chosen to be constant at 0.2, for simplicity. Then the following two models were fitted to the training portion of the data set, one without and one with the time-by-covariate interaction:

$$h(t; x, y) = exp(ay + bx)h_0. \tag{4.11}$$

$$h(t; x, y) = exp(a_1 y + (b_1 + c_1 t)x)h_0. \tag{4.12}$$

The estimated coefficients are shown in Table 4.7 for model 4.11 and in Table 4.8 for model 4.12.

Covariate	Parameter estimate	Wald $\mathcal{X}^2$	Pr > $\mathcal{X}^2$
y	0.819	179.722	< 0.0001
x	3.397	17097.991	< 0.0001

Table 4.7: *Parameter estimates from Model 4.11.*

Covariate	Parameter estimate	Wald $\mathcal{X}^2$	Pr > $\mathcal{X}^2$
y	0.954	2340.862	< 0.0001
x	1.574	1552.211	< 0.0001
x*t	0.446	3158.638	< 0.0001

Table 4.8: *Parameter estimates from Model 4.12.*

The predictive power of both models was then investigated on the holdout sample. A number of good/bad definitions were constructed to cover a period of 24 months:

1. 'bads' are those who defaulted in the first 4 months;
2. 'bads' are those who survived the first 4 months and defaulted in the next 2 months.
3. 'bads' are those who survived the first 6 months and defaulted in the next 2 months.

   ...

11. 'bads' are those who survived the first 22 months and defaulted in the next 2 months.

The Gini coefficient plot in Figure 4.6 shows that for the first 10 months the models perform very similar to each other, but after that the model allowing for the time-by-covariate interaction is predicting better than the ordinary model. The difference in performance increases with time.

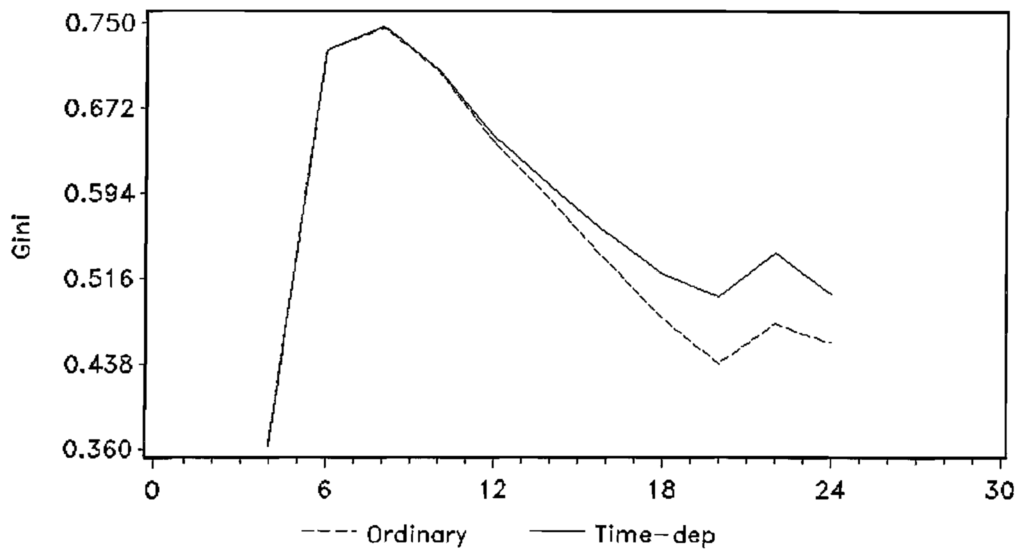


Figure 4.6: *Gini coefficients for models 4.11 and 4.12 for the definitions (1 -11).*

The time at which the difference starts to show will depend on the magnitude of the time-dependency of the effect of a covariate.

4.3 Conclusion

This chapter examined several graphical and numerical tests for time-dependency of the effect of a covariate. Harrel's Z-test was found to be the most appropriate. Hence, the most significant time-by-characteristic interactions suggested by this test were included in the model. It was noted that if one performs the Harrel's test one should select a reasonable size group of covariates with the largest values of the Z-statistic keeping in mind that the order in terms of the significant time-trend may not be the same as the order according to the Z-statistic.

The model built in this chapter has shown that including time-by-characteristic interaction in credit scoring models adds another dimension – flexibility to reflect an increase or a decrease of the effect of a characteristic over the duration of a loan. The simulation has shown that the model with such interactions predicts better than the one without.

Chapter 5

PHAB Scores: Proportional Hazards Analysis Behavioural Scores

5.1 Introduction

Once an application has been approved the lender is interested in monitoring the performance and the behaviour of a customer. A behavioural score is constructed using the information about the customer's activity, payments, purchases etc, which are normally recorded monthly.

The lender may wish to predict future payment amounts or purchase frequency based on the most recent behavioural score. Account management strategies, such as marketing campaigns, increase or decrease of credit limit are often driven by the behavioural score. At present techniques similar to those applied for the application scoring are used for constructing the behavioural scoring systems. Logistic regression is the most common technique, and as well as application characteristics used in application scoring, performance variables are also used to construct the score. We have seen in the earlier chapters that survival analysis can be successfully applied to estimate the time to default or to early repayment at the time of application. In this chapter we develop techniques based on Cox's proportional hazards model incorporating behavioural data, such as information about monthly balance and delinquency, in order to develop survival analysis approaches to behavioural scoring. Firstly, we outline the proposed technique for building behavioural scoring models in terms of proportional hazards formulae in Section 5.2. The specifics of the actual model building process for a proportional hazards analysis behavioural score (PHAB score) are explained in Section 5.3. Sections 5.4 to 5.6 present an example of building such models on a behavioural data set from a major UK financial institution. We discuss three PHAB models which use different behavioural variables and the PH application score as the covariates to predict default, (Section 5.5) and one model which uses the latest available behavioural variables and the previous behavioural score as the covariates, hence accumulating the behavioural information over time (Section 5.6). We compare the performance of these models with that of the application score and between each other. Then two PHAB models, the simplest and the best, are compared with the behavioural scores built using logistic regression (Section 5.7). Finally, we successfully attempt to simplify the PHAB model by smoothing the parameter estimates over time without loosing too much predictive power in Section 5.8.

5.2 Proportional hazards model for behavioural scoring

In consumer lending a lender has a set of customers on the books with their application scores. As information on the performance of the customer becomes available, it is then desirable to include this information in the model to improve predictions of time to default. Existing behavioural scoring systems based on logistic regression do not consider how long a loan has been running. The assumption is that once the initial new applicant phase is passed all the information that affects whether a loan will default in the next 6 months or 12 months is in the current or recent behaviour of a loan. This indicative behaviour is assumed to be independent of how long a loan has been running. Thus loans of all ages are lumped together. In the survival analysis approach one is trying to estimate when a loan will default, or rather how much longer it will survive. A loan that fails after 9 months has a further lifetime of 6 months when considered at 3 months into a loan but only 5 months at 4 months into a loan.

Thus survival analysis builds a different behavioural model for each age of a loan, using the customer's behaviour up to that month to predict their remaining time to default.

Let $h(t)$ be a hazard rate of default of a loan; $\mathbf{x}$ – application data, then the proportional hazards model, (Cox [1972]), for the application score is

$$h^0(t) = exp(\mathbf{x}'b(0))h_0^0(t) \tag{5.1}$$

where $h_0^0(t)$ is a baseline hazard for the population of applicants; $b(0)$ is a vector of unknown parameters.

s months after the start of a loan a lender has behavioural data $\mathbf{y(s)}$ on the consumer and the PHABS model estimates the remaining time to failure from s by the hazard function

$$h^s(t) = exp(\mathbf{x}'b(s) + \mathbf{y}'(s)c(s))h_0^s(t) \tag{5.2}$$

where $h_0^s(t)$ is a baseline hazard for the population of customers who are still on the books at the period s, i.e. those who defaulted or paid off early or closed normally before s are excluded from the population; $b(s)$ and $c(s)$ are vectors of unknown parameters.

Figure 5.1 depicts the case, namely that the behavioural score is built to predict the remaining time to default at some point in the course of a loan and incorporates application information as well as performance information such as monthly balance and any partial delinquency.

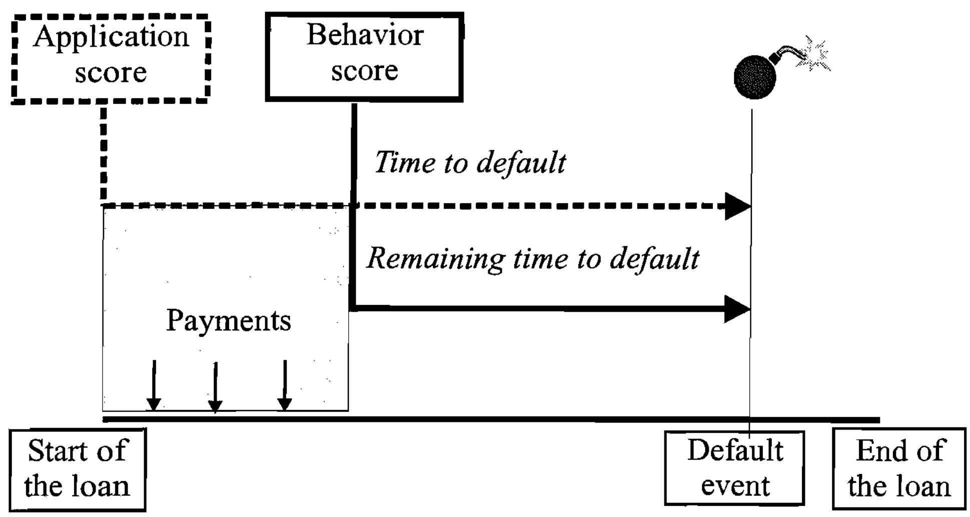


Figure 5.1: *Application and behavioural scores.*

5.3 Modelling approach

The modelling process had two model building stages and one diagnostic stage, so the data was split into three parts: two training and one holdout samples.

First, the application score was built using only application characteristics as covariates in Cox's proportional hazards model on a first training sample. Model (5.1) was fitted to time to default and lifetimes of the customers who did not default were considered censored using the approach suggested in Thomas et al. [1999], Thomas and Stepanova [1999] and developed further in Chapters 2 and 3.

The second step was to build a PHAB score for each month of the duration of a loan. Now the application information was included in the model in the form of the application score from the first step and behavioural information in the form of available behavioural data. The dependent variable was remaining time to default at the month of observation (Table 5.1).

The holdout sample was then scored with the application score and all the PHAB scores so that their performance may be evaluated and compared.

Mth	Data Set	Response	Covariates for Models I	Resulting Score
	training	T	application variables	Ascore
4	holdout-D_3	T - 4	Ascore + $BVAR_4$	PHABS_4
5	holdout-D_4	T - 5	Ascore + $BVAR_5$	PHABS_5
6	holdout-D_5	T - 6	Ascore + $BVAR_6$	PHABS_6
			...	
36	holdout-D_{35}	T - 36	Ascore + $BVAR_{36}$	PHABS_{36}

D_i - a set of customers whose loan has ceased (either by default, early or normal repayment) by month i; T - time to default; Ascore - application score; PHABS - Proportional hazards analysis behavioural score; $BVAR_i$ - behaviour variables for month i.

Table 5.1: *Building PHAB score models.*

5.4 Behavioural Data Set

We have used data provided by a UK financial institution. It contained 11 500 customers with their application characteristics and subsequent performance variables for 36 months. Table 5.2 describes the application data that was available and Table 5.3 describes the performance data.

Expected Month End Balance for month I was calculated by multiplying the Loan Instalment by a number of months left to the end of a loan:

$$\text{Expected Balance}_I = ((\text{Term} - I) * \text{Loan Instalment}).$$

1	Account Opening Date	9	Time with Current Employer
2	Account Closing Date	10	Residential Status
3	Amount of Loan	11	Marital Status
4	Term of Loan	12	Frequency Paid
5	Purpose of Loan	13	Number of Dep. Children
6	Loan instalment	14	Age of customer
7	Total to be repaid	15	Net Income
8	Time at Current Address	16	Occupation Code

Table 5.2: *Application characteristics in the behavioural data set.*

1	Delinquency balance
2	Delinquency status
3	Current Month end Balance
4	Worst status

Table 5.3: *Performance variables in the behavioural data set.*

Then two variables were constructed for use as predictors of customer behaviour, Balance Difference:

$$\text{Balance Difference}_I(BD_I) = \text{Balance}_I - \text{Expected Balance}_I,$$

which indicates how far the current balance is from the expected balance. If the difference is positive, the customer is behind the schedule in repaying a loan, hence may be more likely to default than the one with zero or negative difference. Negative differences may occur if the payments are ahead of schedule.

The second variable constructed was Balance Difference Difference:

$$\text{Balance Difference Difference}_I(BDD_I) = BD_I - BD_{I-1},$$

which indicates if there is any short term trend in Balance Differences. If the value is positive, for example, the Balance Difference has increased from one month to another, this means that the customer has fallen further behind the schedule in the last month.

We have also used the variable Delinquency Status, DL_I, as is.

5.5 Three behavioural models

The data was split into three samples (two training and one holdout) of approximately equal size.

Stepwise proportional hazards model, phr [1998], using Breslow likelihood approximation was used to build an application score on the first sample, which is summarised in the Table 5.4.

Since this was an exploratory analysis to see if PHABS made sense, we did not apply any transformations to the variables nor did we coarse-classify the first five variables which were continuous. The remaining five variables are binary and are attributes that were found to be of importance by the stepwise procedure.

PHAB scores were then built on the second training sample for each month of the duration of a loan, from 4 to 36. The model for the i^{th} month is fitted to the remaining time to default (Time to Default$-i$), e.g if a customer defaulted in month 12 and the model is for month 4, remaining time to default is 8 months. Predictor variables were the application score and the performance variables for the i^{th} month.

Variable Name	Parameter Estimate	Significance
Amount of Loan	$9.98 * 10^{-5}$	0.0414
Term of Loan	-0.0278	0.0017
Months with current Employer	-0.0059	0.0005
Months at Current Address	-0.0032	0.0017
Net Income	-0.0004	0.0540
Living w/Parents or Renting	0.4389	0.0188
0-1 dependent children	-0.3158	0.1204
Married	-0.5479	0.0048
High Risk Purposes		
(such as Refinance)	0.9014	0.0001
Low Risk Occupation Code	-0.4414	0.0151

Table 5.4: *Application characteristics and their coefficients in the application scoring proportional hazards model.*

Several performance variables and their combinations were tried out as predictors in the model to see which ones result in a model with a better fit to the data. The following three models will be discussed in this chapter:

	Dependent Variable	Predictors
Model 1	Remaining Time to Default	Balance Difference (BD_I), App. Score
Model 2	–"–	Balance Difference and Balance Difference Difference (BD_I, BDD_I), App. Score
Model 3	–"–	Last Month Delinquency status ($DL_(I-1)$), App. Score

Table 5.5: *Three behavioural models.*

The Log-likelihood statistic, which indicates how well a model fits the data, was plotted over time for each of the models. Figure 5.2 shows that Model 3 was the best, followed by models 2 and 1, which were very close but 2 was slightly better.

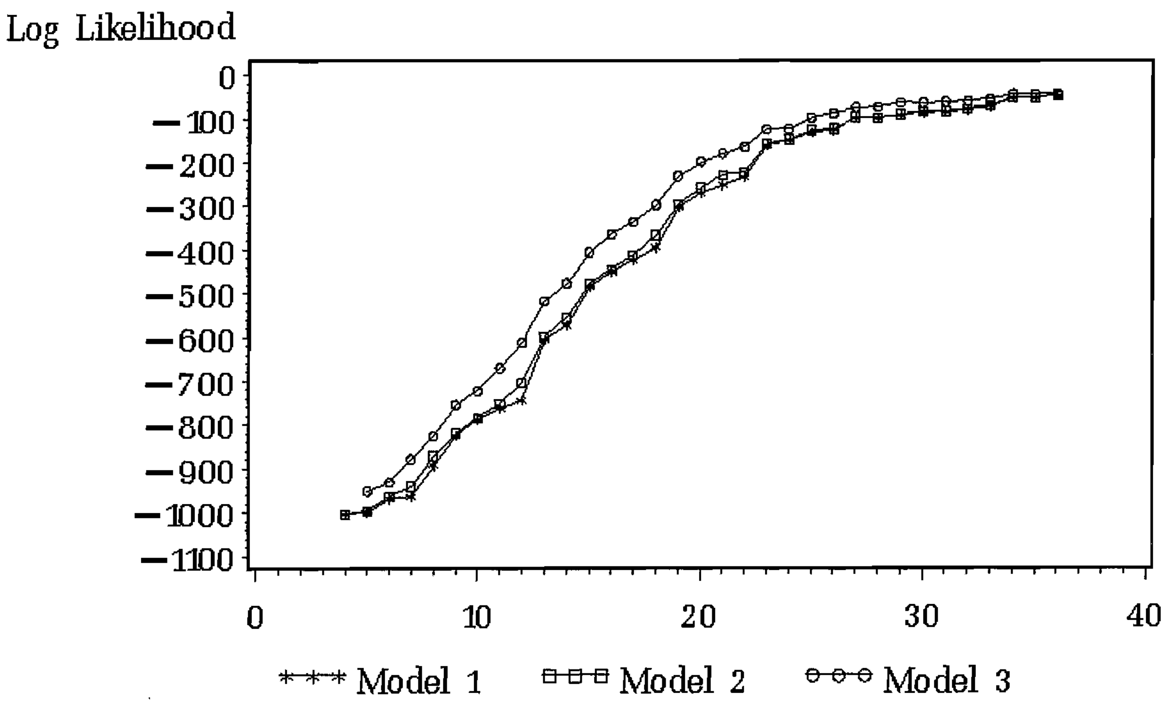


Figure 5.2: *-2 Log-Likelihood statistic for the three models.*

The third model was expected to be the best because the behavioural covariate it is using, Last Month Delinquency Status, is closely related to default which is the outcome we are trying to predict (see the default definition in Section 3.2).

These three models are described in detail in the following sections. We examine parameter estimates over time and compare the predictive power of each of the three PHAB scores with the proportional hazards based application score over time using ROC curves for two different definitions of 'good' and 'bad'. The definitions are given in Table 5.6.

Definition 1	'bad'	default in the next 12 months
	'good'	other
Definition 2	'bad'	default before the end of a loan
	'good'	other

Table 5.6: *Good-bad definitions.*

The score that was used in building the ROC curves can be calculated according to the proportional hazards model:

$$H^s(t) = exp(\mathbf{x}b(s) + \mathbf{y}(s)c(s))H_0^s(t) \tag{5.3}$$

where $H_0^s(t)$ is nonparametric baseline cumulative hazard obtained together with $b(s)$ and $c(s)$ parameter estimates from fitting the model on the training sample. This model is equivalent to 2.10 as was shown in Section 4.1.2.

For Definition 2, $H(\text{term})$ is used as a score, where term is a period of time remaining until a loan ends. For Definition 1, in which the forecasting time horizon is fixed at 12 months, we are interested in $H(12)$ but as this is monotone in $exp(\mathbf{x}b(s) + \mathbf{y}(s)c(s))$, it is sufficient to use $exp(\mathbf{x}b(s) + \mathbf{y}(s)c(s))$ as the score.

Only the ROC curves for Definition 2 are shown since the curves for Definitions 1 and 2 were very similar.

UNIVERSITY OF SOUTHAMPTON LIBRARY

5.5.1 Model 1: Application Score and Balance Difference

Plots of the parameter estimates for Model 1 are shown in Figures 5.3-5.4. It can be seen that the parameter estimate for the balance difference is higher at the beginning and then decreases with time after about 22 months while the parameter estimate for the application score increases with time. This result appears to show how the importance of application and behavioural scores changes over time.

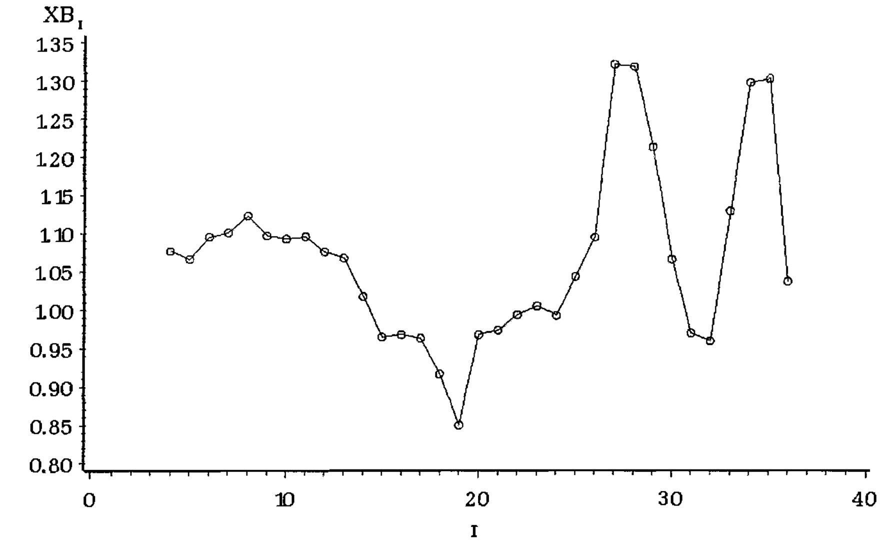


Figure 5.3: *Parameter estimates for the application score from Model 1.*

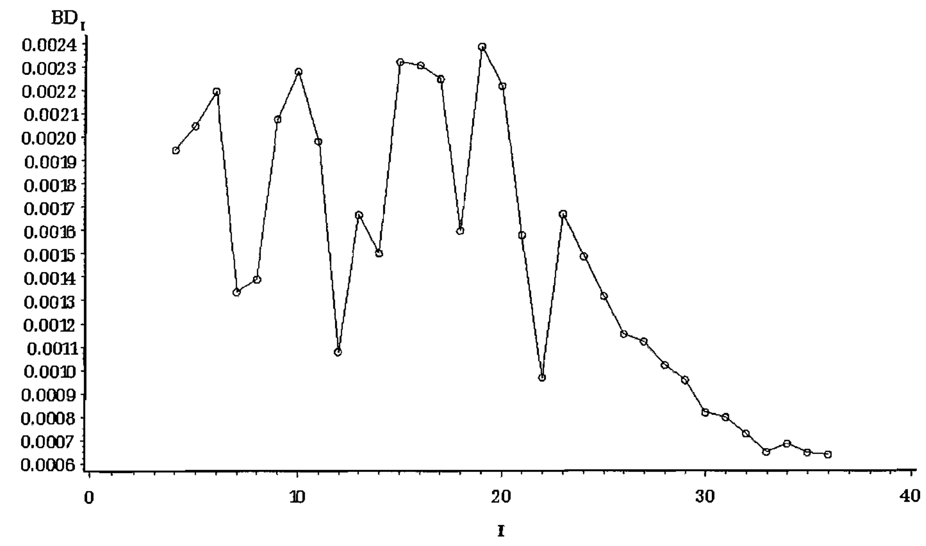


Figure 5.4: *Parameter estimates for BD_i from Model 1.*

As opposed to what would have been expected intuitively, early default could not be predicted well by the application score since if it could, these applications would be rejected. However, these early defaulters should be identified quite easily using behavioural variables as customers with large balance difference, i.e. customers who are falling behind at the early stages.

Therefore application and behaviour information complement each other, but their importance changes over time. At the early stages behaviour variables are more important and at the later stages application information becomes more predictive.

Figure 5.5 compares Model 1 PHAB score with the proportional hazards based application score. In this and all other ROC curves figures in this chapter a) corresponds to the results 6 months into a loan, b) to 10 months, c) to 14 months, d) to 18 months, e) to 22 months and f) to 26 months.

These ROC curves show that the addition of just one behavioural variable results in a definite improvement in the performance over the application score.

Note that after 10 months or so the PHAB scores are less predictive than the application score for the least risky customers (i.e left end of the ROC curve). This is probably because behavioural variables tend to be predictive only when there is some minimal level of delinquency present. The reason the ROC curve for the application score changes is because the population is getting smaller as people 'complete' a loan one way or another. This is demonstrated in column "Data set" of the Table 5.1, which shows that by each month i some number of loans D_i has ceased because of default, repayment or for other reasons. The same application score is used every month but as the holdout sample changes and gets smaller, the appearance of the ROC curve also changes.

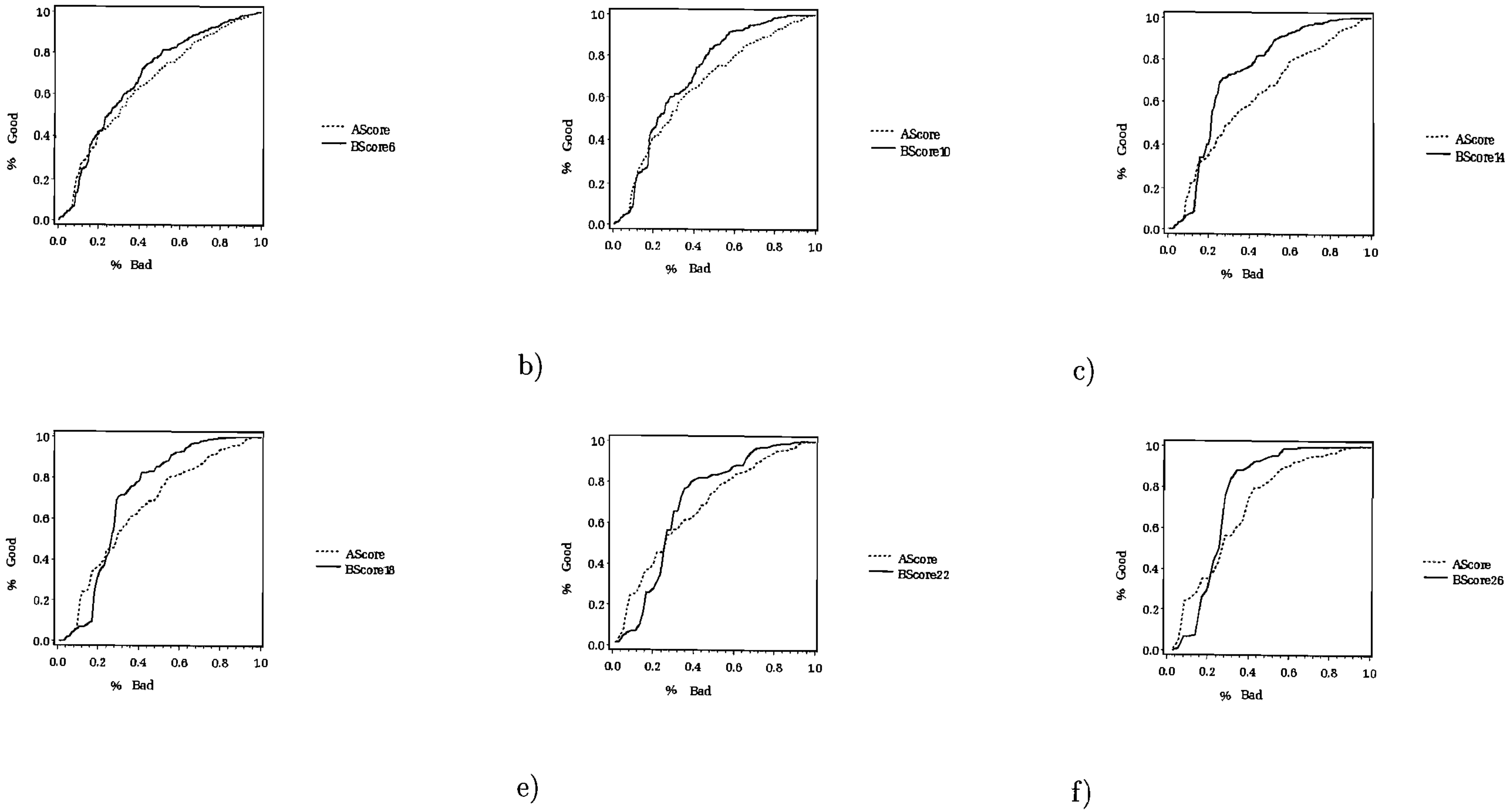


Figure 5.5: *ROC curves for Model 1.*

5.5.2 Model 2: Application Score, Balance Difference and balance Difference Difference

Plots of the parameter estimates for Model 2 are shown in Figures 5.6-5.8. It can be seen that as in Model 1 the parameter estimate for the balance difference is higher at the beginning and then decreases with time while the parameter estimate for application score increases with time. This illustrates again the fact that the behavioural information is most important earlier on and the application information later on.

However, the argument is slightly more subtle than in the case of Model 1 because of the interaction between BD and BDD. In particular the average coefficient of the latter at -0.003 is larger in absolute value than the typical 0.002 coefficient of the former. One could rewrite $0.002BD_t - 0.003(BD_t - BD_{t-1})$ as $0.003BD_{t-1} - 0.001BD_t$. Since there are strong correlations between BD_t and BD_{t-1} this expression is approximately the $0.002BD_t$, which is similar to the result in Figure 5.4 of Model 1. The fact that the actual weighting on BD_t in this expression is negative is difficult to explain. The peaks in the coefficient of BDD_t at 6,11 and 24 months are an artifact of the large number of loans completing at these times.

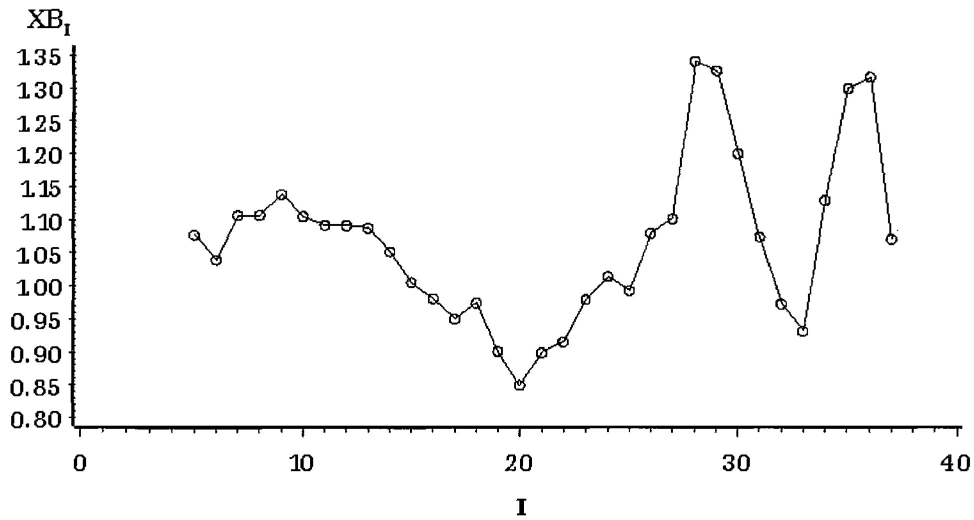


Figure 5.6: *Parameter estimates for the application score from Model 2.*

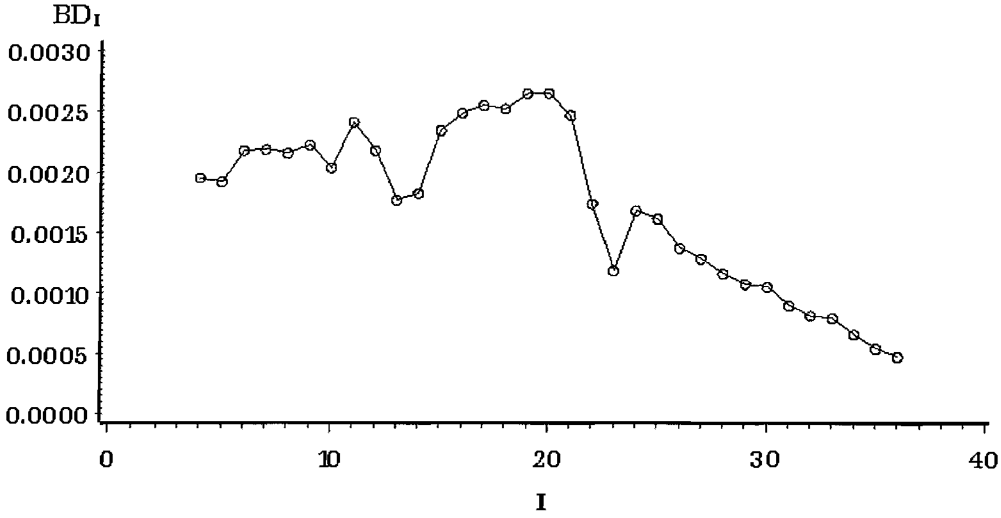


Figure 5.7: *Parameter estimates for BD_i from Model 2.*

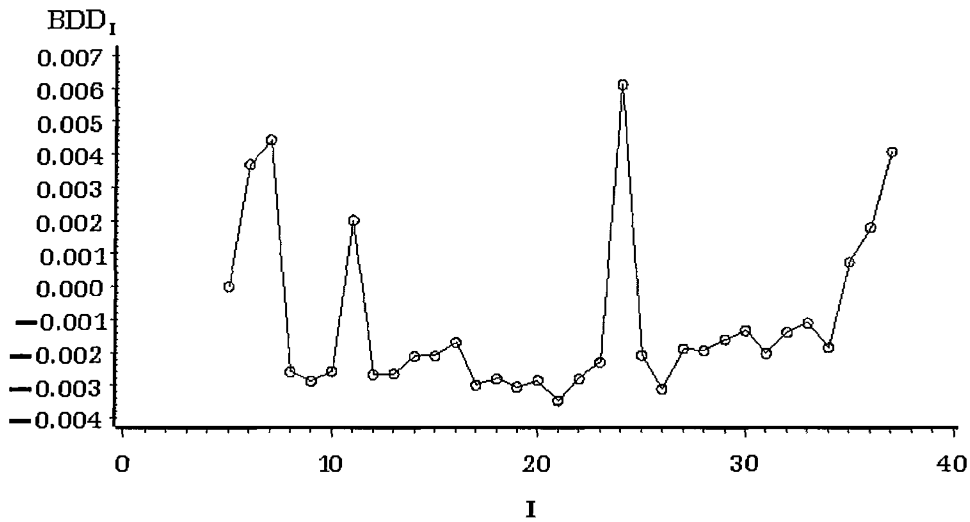


Figure 5.8: *Parameter estimates for BDD_i from Model 2.*

Figure 5.9 shows the ROC curves comparing Model 2 PHAB score with the proportional hazards based application score. The curves are similar to those from Model 1, so they support the finding that the PHABS perform better than the application score.

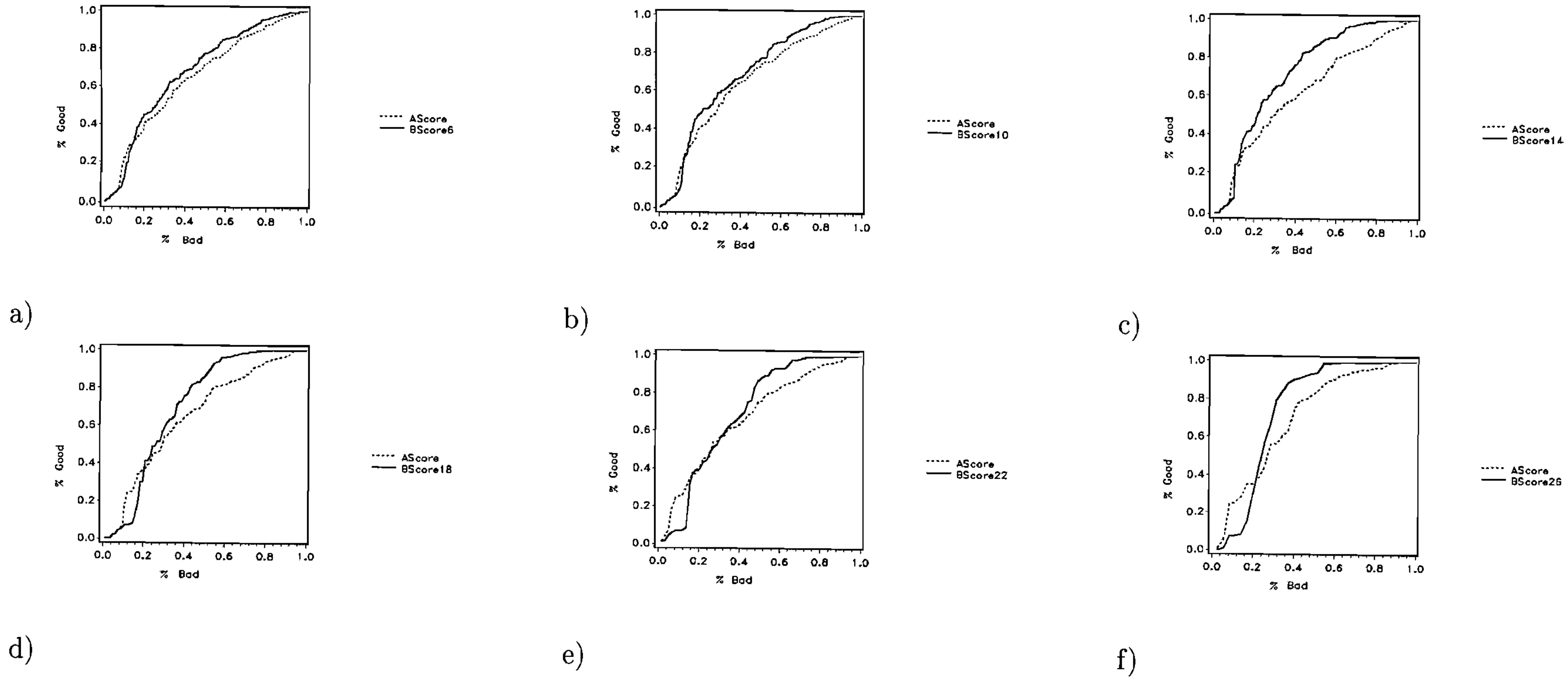


Figure 5.9: *ROC curves for monthly behavioural PH models compared with the application PH model from Model 2.*

5.5.3 Model 3: Application Score and Last Month Delinquency Status

The plots of the parameter estimates for the last of the three models (Figure 5.10 and 5.11) show strong interdependence between the application score and the last month delinquency status. The value of the parameter estimate for DL_i is much larger than those of the behavioural variables from Models 1 and 2. This suggests that the DL_i is a strong predictor for default. It 'competes' with the application score in its importance throughout the duration of a loan since the two plots are almost mirror images of each other. Obviously, early on the DL_i estimate is low as there are few cases with 'bad' status. DL_i becomes very important in the second year of a loan (months 12 to 24) but as in the first two models the application score effect increases over time.

Figure 5.12 compares Model 3 PHAB score with the application score results. The performance of this model is much better than the application score and than the other two PHAB scores (Models 1 and 2). However, it is logical to assume that most delinquencies, especially the serious ones, do not recover and hence, prediction of default using last month delinquency indicator may be pointless as it comes too late.

The detailed comparison of the three models can now be summarised. The best model is Model 3, but its usefulness is arguable because the prediction may come too late. The next best is Model 2 followed very closely by Model 1. Hence, one may prefer Model 1 to Model 2 for its relative simplicity.

In the sections that follow we use the simplest PHAB score – Model 1 in the comparison with an accumulated model and when smoothing the parameter estimates, but both Model 1 and Model 2 are compared with an alternative logistic regressions model.

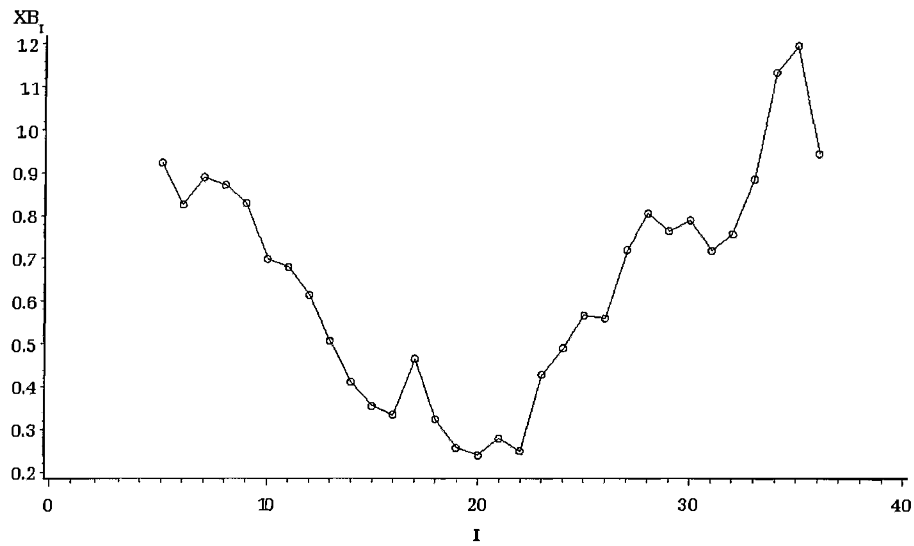


Figure 5.10: *Parameter estimates for the application score from Model 3.*

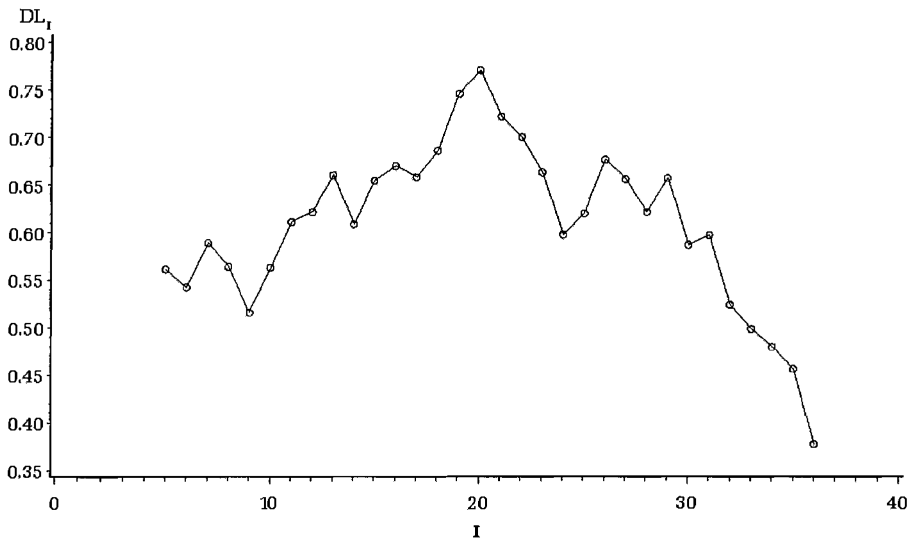


Figure 5.11: *Parameter estimates for DL_i from Model 3.*

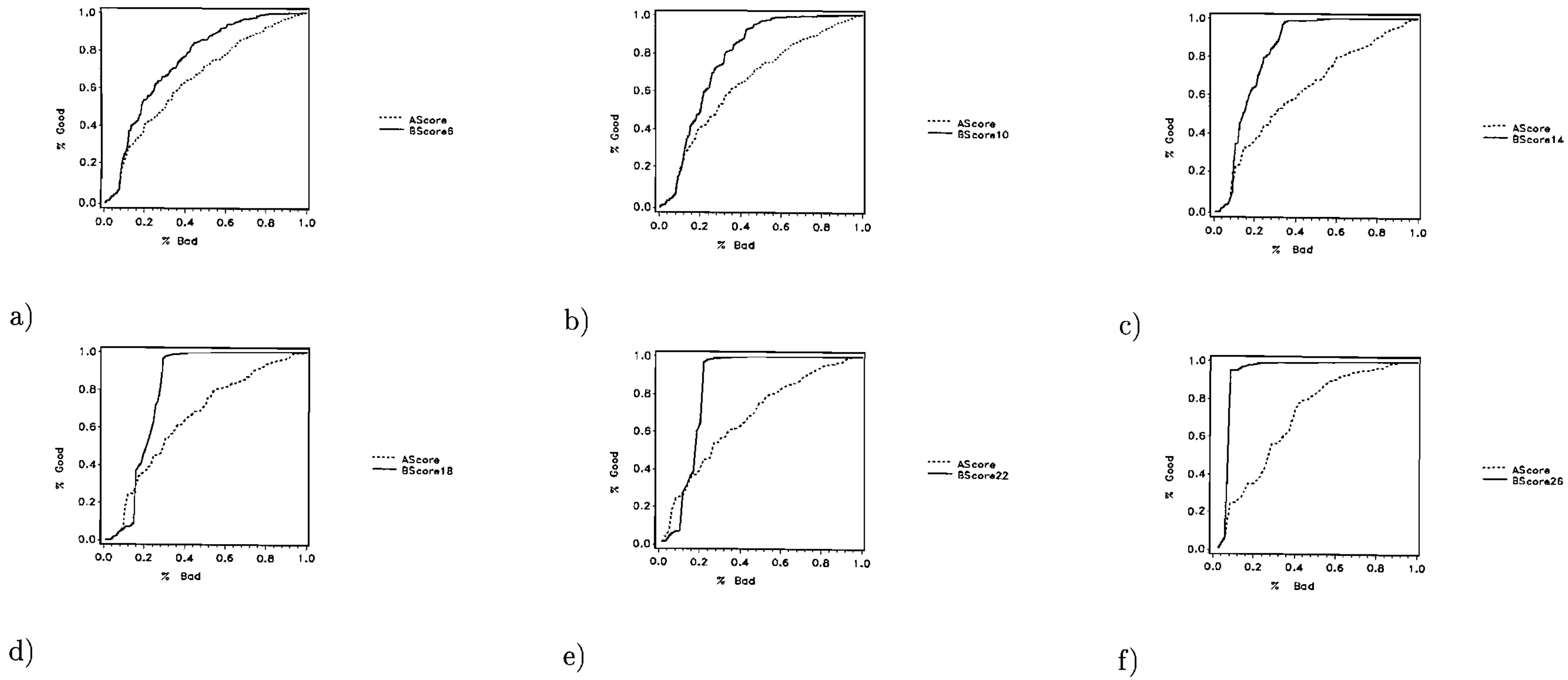


Figure 5.12: *ROC curves for the monthly behavioural PH compared with the application PH score from Model 3.*

5.6 Accumulated Model

The first extension of the PHAB scores described above is the accumulated model. The score for the i^{th} month is now built by taking a behavioural score from the $(i-1)^{st}$ month instead of the application score and a performance variable for the i^{th} month as the covariates for the model. This way the score accumulates the information from all the previous months and the performance variable adds the latest data. This model building approach is summarised in Table 5.8.

To examine this approach we have built a PHAB score using one behavioural variable Balance Difference and the accumulated behavioural score as the covariates. It will be referred to as the accumulated model or Model 4:

	Dependent Variable	Predictors
Model 4	Remaining Time to Default	Balance Difference (BD_I), PHAB score(I-1)

Table 5.7: *The accumulated behavioural model.*

Figure 5.13 compares log-likelihood statistics plots for the accumulated and the non-accumulated score models (Models 1 and 4).

It was found that accumulating behavioural information into one of the covariates results in high correlation between the covariates in the model, see the cross-plot of the parameter estimates in Figure 5.14, which in turn results in the unstable parameter estimates (Figures 5.15 and 5.16) and slightly lower log-likelihood statistics.

Further comparisons of the accumulated with the non-accumulated model were performed using ROC curves, which are shown in Figure 5.18. The accumulated model performs better than the non-accumulated in predicting default before the end of a loan. However, the plots of Gini coefficients against time for both models (Figure 5.17) show that the performance of the accumulated model is less uniform during the course of a loan than that of the non-accumulated model. Therefore the non-accumulated approach remains the preferred one.

Mth	Data Set	Response	Covariates for Models II	Resulting Score
	training	T	application variables	Ascore
4	holdout-D_3	T - 4	Ascore + $BVAR_4$	Bscore_4
5	holdout-D_4	T - 5	$\text{Bscore}_4 + BVAR_5$	Bscore_5
6	holdout-D_5	T - 6	$\text{Bscore}_5 + BVAR_6$	Bscore_6
			...	
36	holdout-D_{35}	T - 36	$\text{Bscore}_{35} + BVAR_{36}$	Bscore_{36}

D_i - a set of customers who defaulted in month i; T - time to default; Ascore - application score; Bscore - behavioural score; $BVAR_i$ - behaviour variables for month i.

Table 5.8: *Building an accumulated behavioural score.*

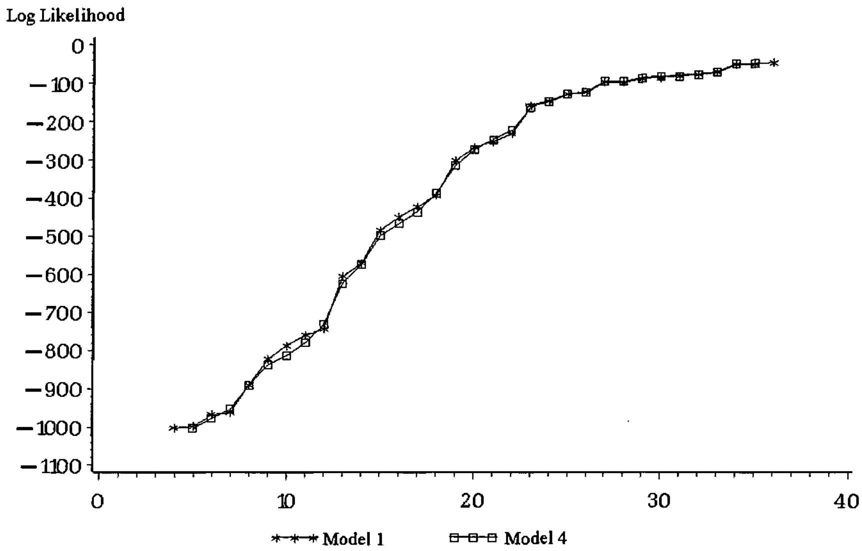


Figure 5.13: *The Log-likelihood statistic for Models 1 and 4.*

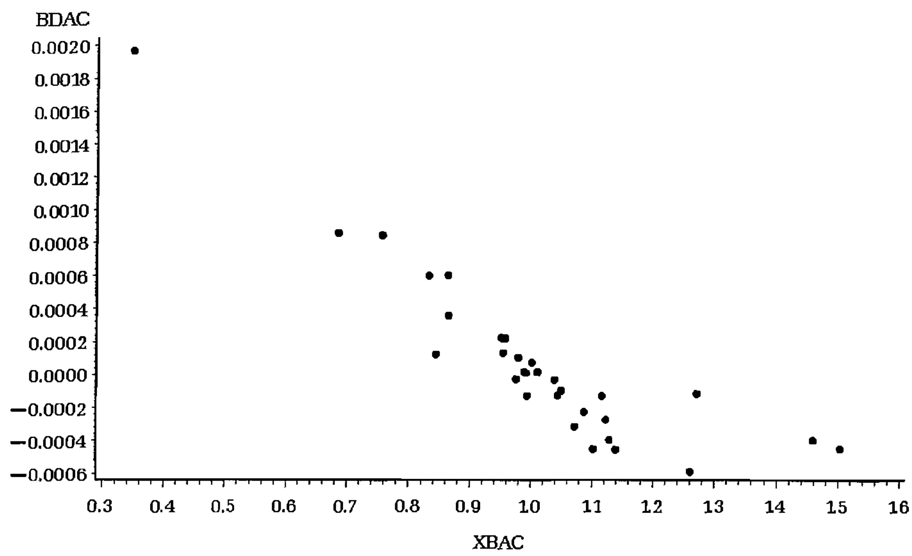


Figure 5.14: *Cross-plot between the the two covariates from Model 4: Accumulated Behavioural Score and* BD_i.

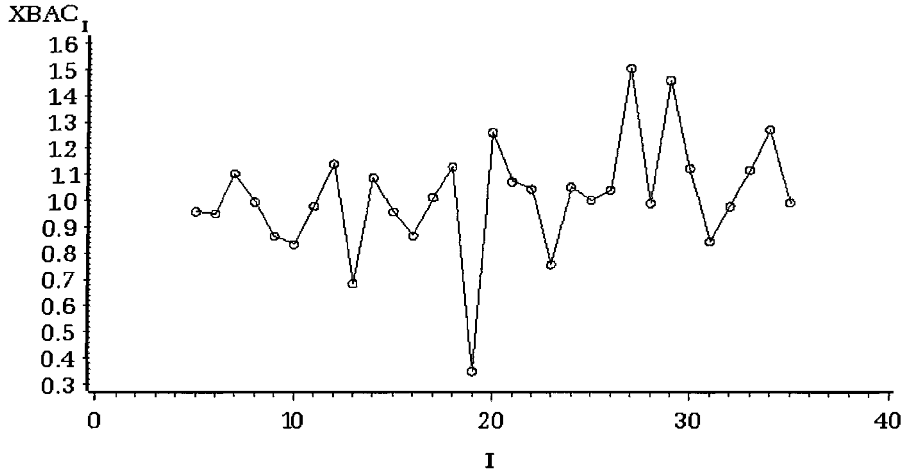


Figure 5.15: *Parameter estimate for the Accumulated Behaviour Score from Model 4.*

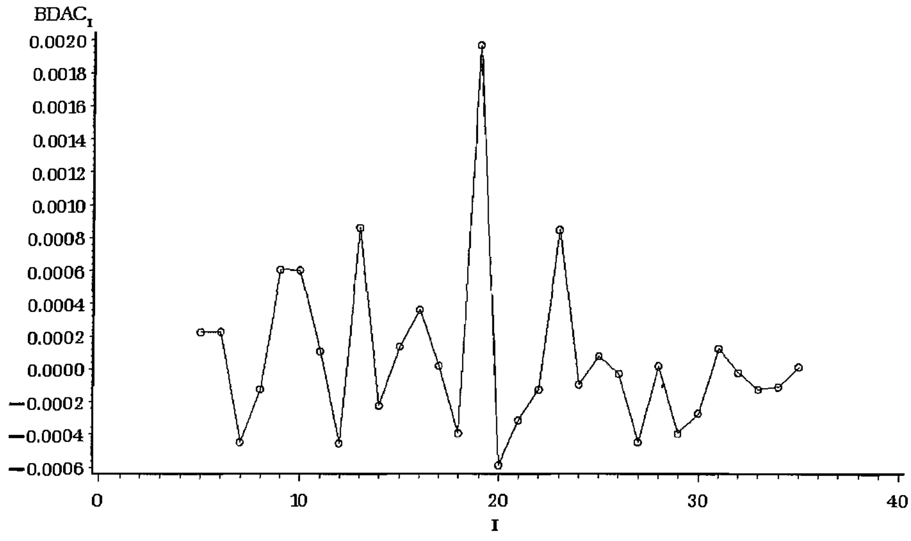


Figure 5.16: *Parameter estimate for the BD_i from Model 4.*

Note that, while the ROC curves suggest that Model 4 is better than Model 1, the log-likelihood statistic was higher for Model 1 than for Model 4. This is because this statistic was calculated on the training sample , not on the holdout as the ROC curves. So the higher value of the log-likelihood statistic simply indicates that Model 1 was a slightly better fit to the training sample than Model 4.

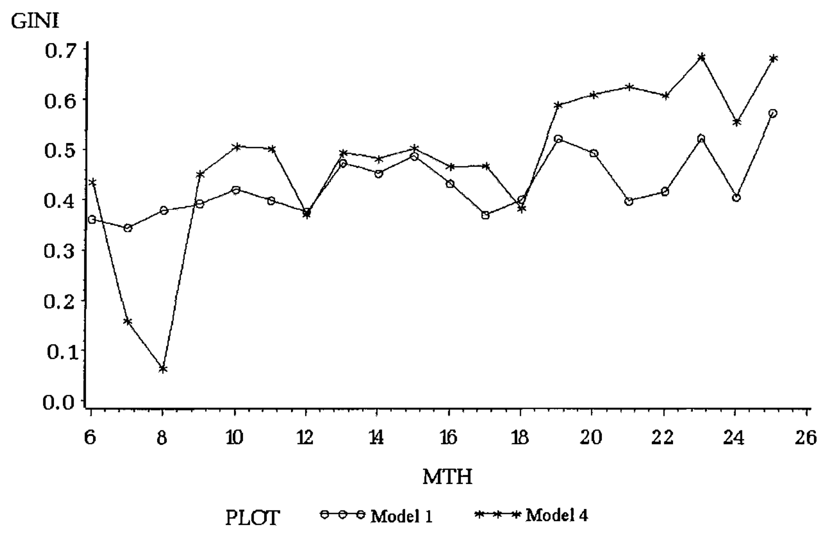


Figure 5.17: *Gini coefficients for Model 1 and Model 4.*

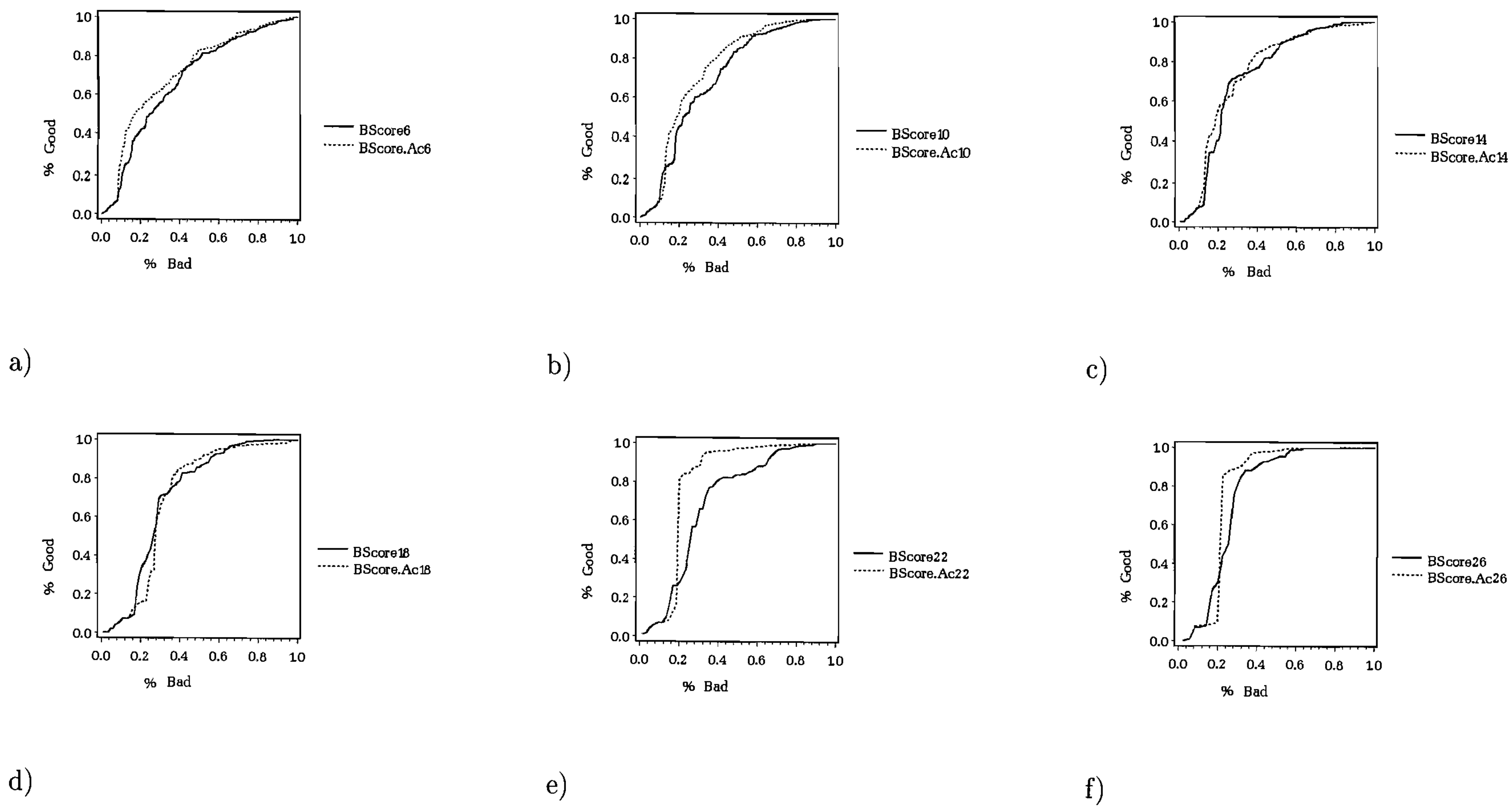


Figure 5.18: *ROC curves comparing the accumulated (Model 4) and the non-accumulated (Model 1) models.*

5.7 Comparison of proportional hazards behavioural models to logistic regression models

An alternative application score and a set of behavioural scores for each month of the lifetime of a loan were built using the logistic regression. Two different binary responses were used corresponding to the good-bad definitions used earlier for the ROC curves for the proportional hazards models. The resulting sets of the logistic regression models were then compared with the proportional hazards models using ROC curves.

Figure 5.19 and Figure 5.20 show the ROC curves comparing Models 1 and 2 respectively with the corresponding LR models using Definition 2. As in Section 5.5 results for Definitions 1 and 2 were very similar, so only Definition 2 results are presented.

It is expected that the LR scores should perform better than the PH scores because the LR models were fitted specifically to the definition of 'good' and 'bad' while the PH model is fitted to the time to default.

Surprisingly it can be seen (Figure 5.19 and 5.20) that the PHAB scores' performance improves over time. After about two years the PHAB scores clearly outperform the LR scores.

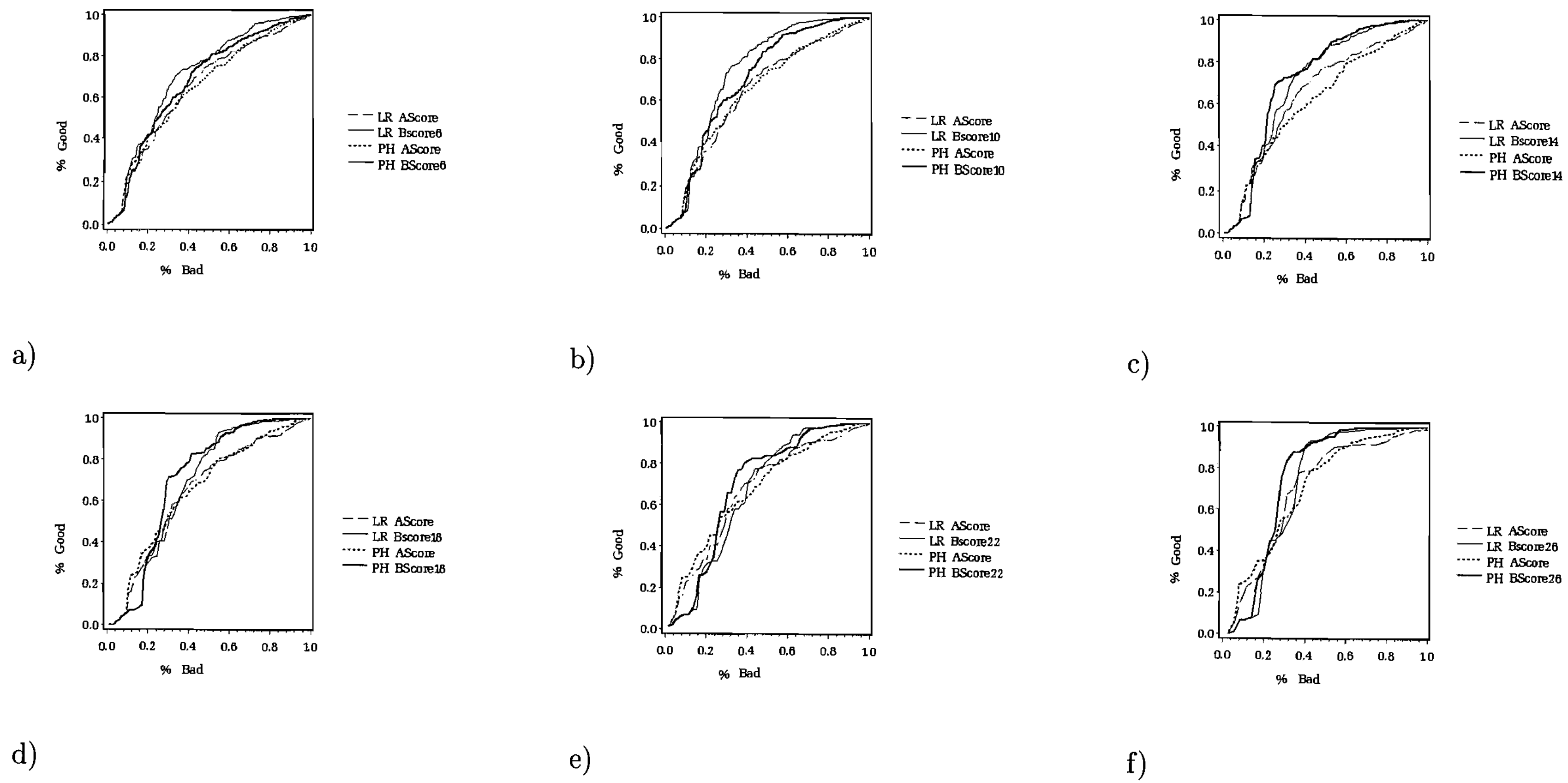


Figure 5.19: *ROC curves comparing Model 1 with Logistic regression.*

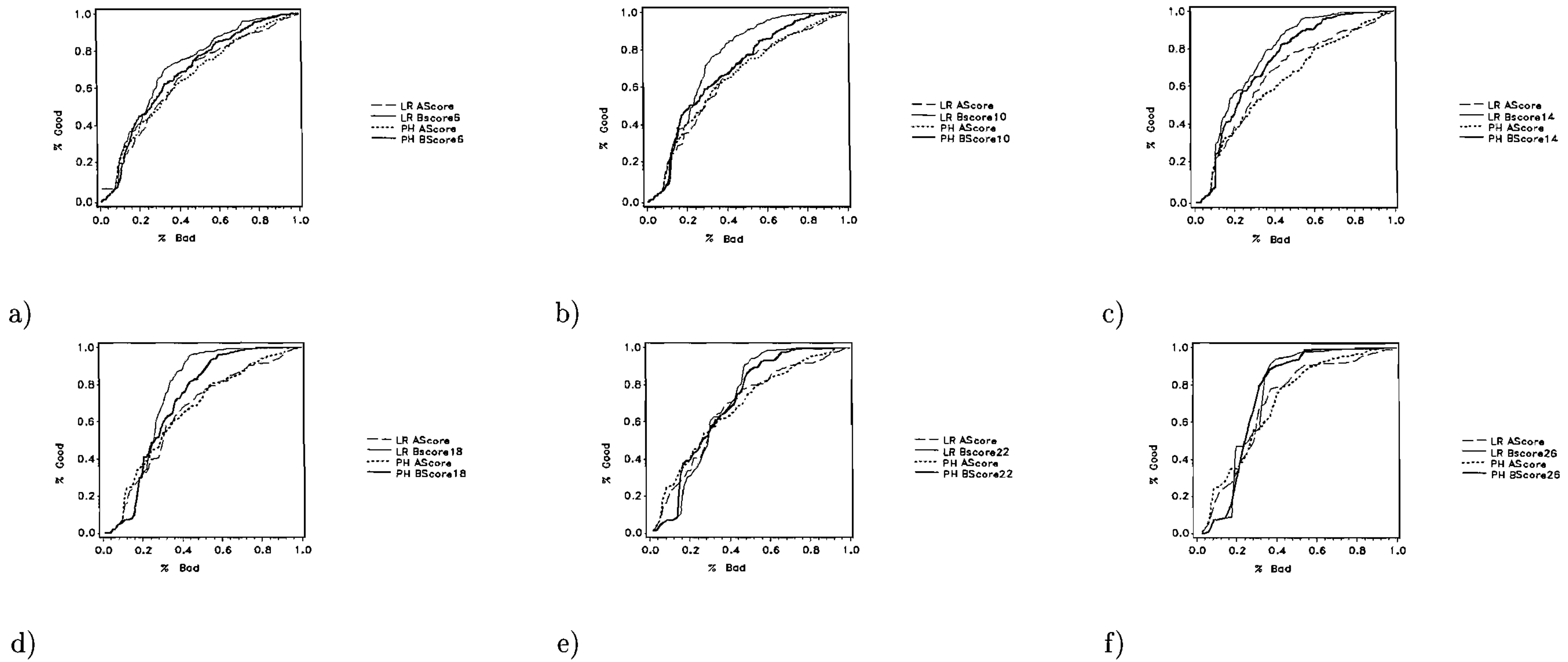


Figure 5.20: *ROC curves for the monthly behavioural (Model 2) and the application PH scores compared with the logistic regression scores.*

5.8 In search of a more robust model

In the previous sections we have explored and compared a number of models all of which require recalculating PH parameter estimates every month. This may be considered as not very practical or not 'user-friendly'. There also may be a danger of overfitting the data and hence, building a descriptive rather than a predictive model.

Two approaches of smoothing the parameter estimates, one by taking the average of the estimates over the model building time interval (32 months in our case) and another by fitting a linear trend to them, are described below.

5.8.1 Averaged Model

The first and the crudest simplification of the PHAB score was achieved by taking the averages of the parameter estimates from Model 1 for both the application score and the behavioural variable over 32 months, so that we have only one model for all the months. The average values were 1.0682 and 0.0015 for XB and BD respectively (Figures 5.21 and 5.22). This one model was then used to score the holdout sample, so that the predictive power of the averaged PHAB score and the original Model 1 PHAB score can be compared. From the ROC curves (Figure 5.23) we conclude that the averaged model is slightly worse than the original PHAB.

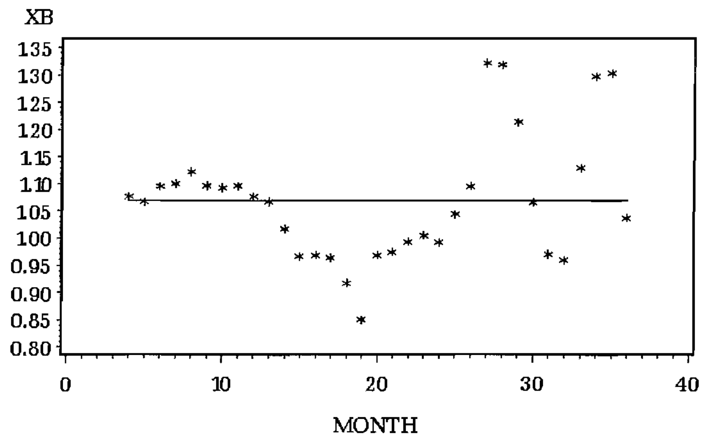


Figure 5.21: *The average value for the XB parameter estimate from Model 1.*

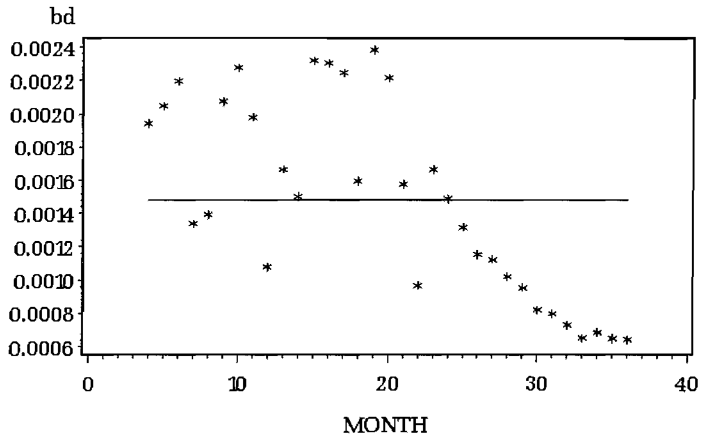


Figure 5.22: *The average for the BD parameter estimate from Model 1.*

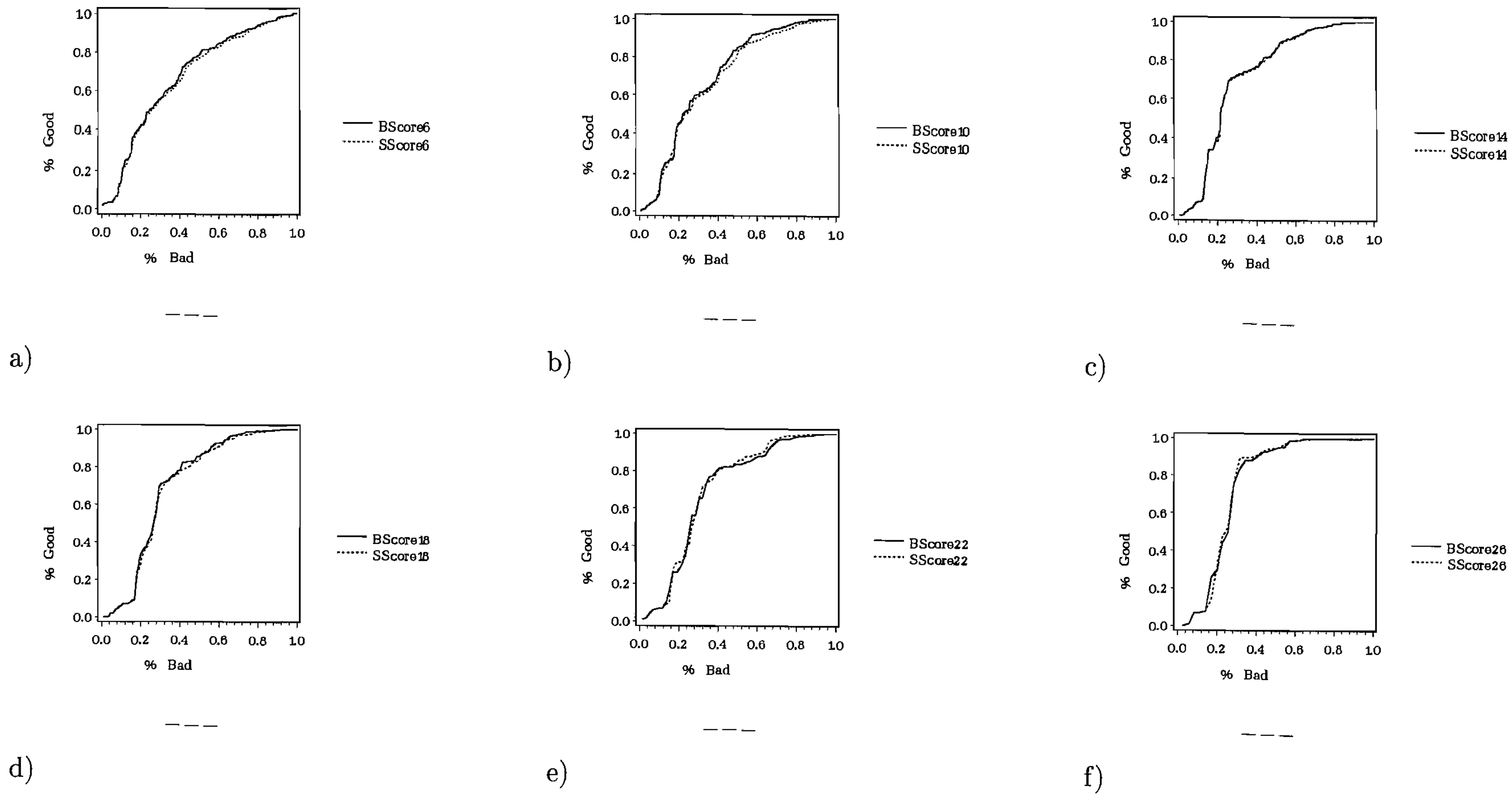


Figure 5.23: *ROC curves comparing Model 1 with the Averaged model.*

5.8.2 Linear Smoothed Model

The comparison of the averaged and the original models above is encouraging in the sense that the difference between their performance is not too big. Hence, if we improve our smoothing slightly the difference may disappear.

Linear approximation was chosen as the next simplest smoothing. The trends shown in Figures 5.24 and 5.25 were fitted using linear regression on the values of the parameter estimates against time. The equations of the trends are:

$$\begin{aligned} \text{XB}_{trend} &= 1.0105 + 0.0029 * \text{Month}; \\ \text{BD}_{trend} &= 0.0024 + 0.000045 * \text{Month} \end{aligned}$$

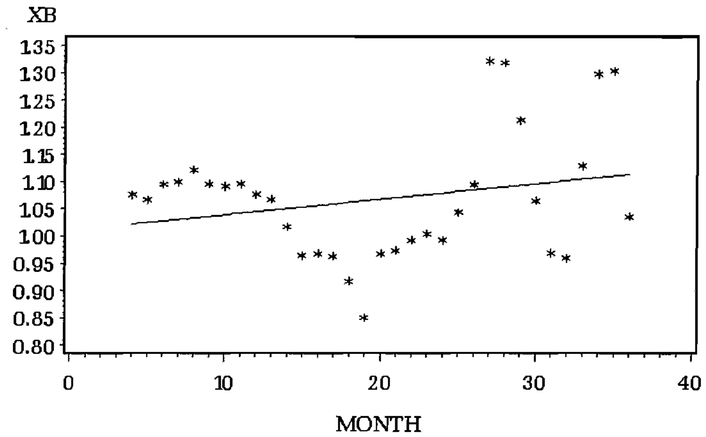


Figure 5.24: *Linear trend for the XB parameter estimate from Model 1.*

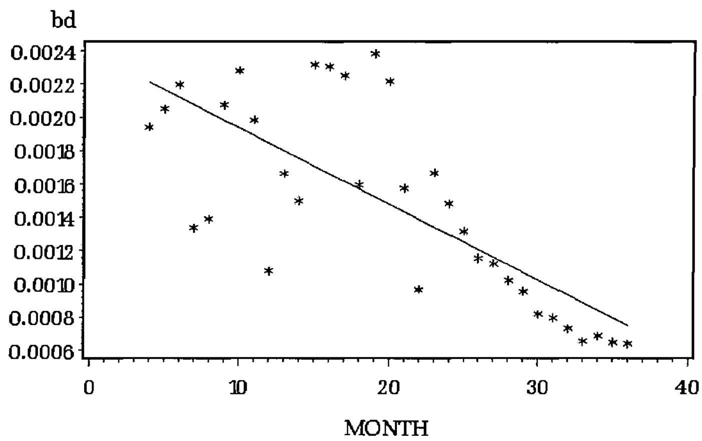


Figure 5.25: *Linear trend for the BD parameter estimate from Model 1.*

The holdout sample was then scored using the values of these trends for each month to compare the performance of the smoothed model and the original Model 1. The ROC curves (Figure 5.26) show almost no difference at all between the models. One can not say which is better since no one model has the ROC curve consistently above the other.

In summary, the smoothing experiment showed that it is possible to simplify the PHAB score to various degrees depending on the aims and the resources of a lender. One can trade off the slight loss of the predictive power for the extreme simplicity, or choose a bit more laborious linear smoothing to match the performance of the original PHAB score.

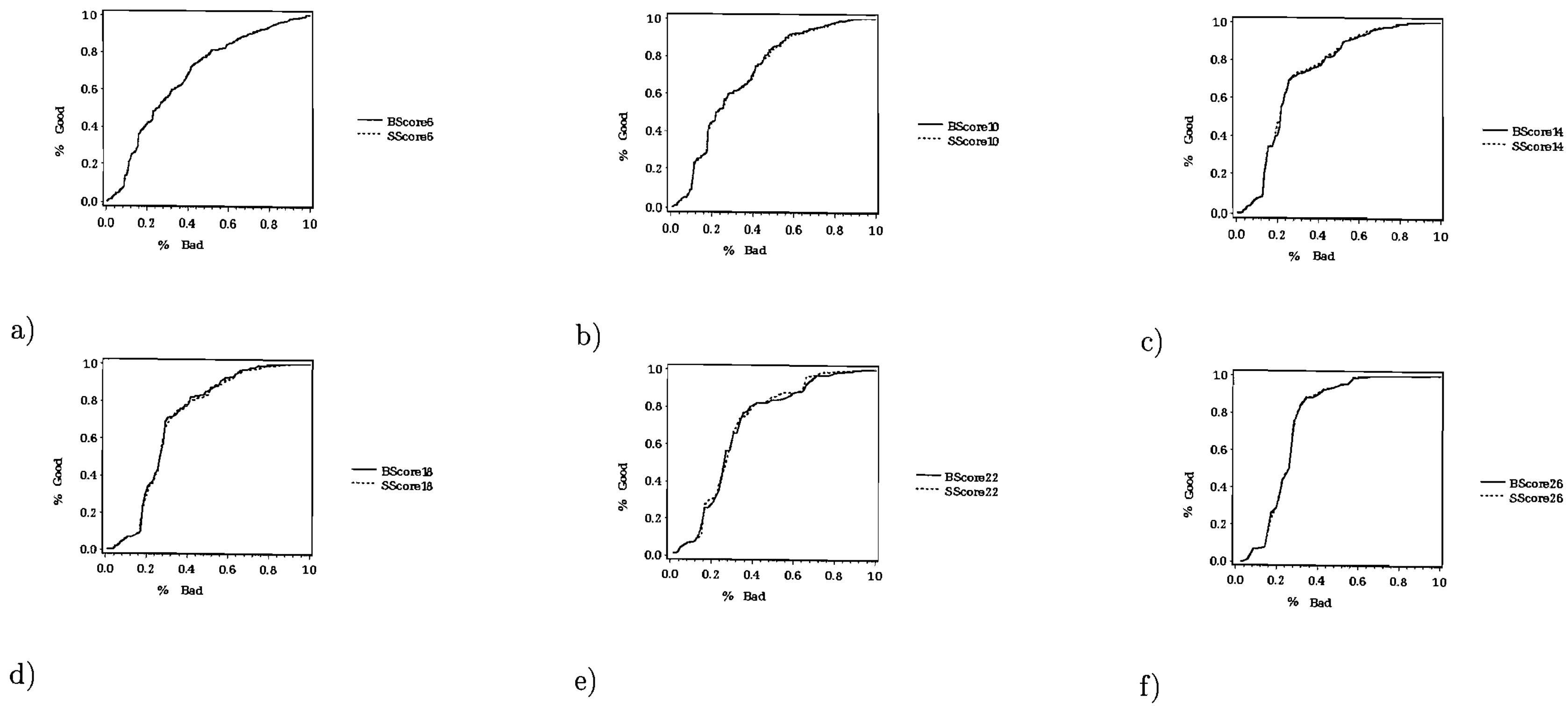


Figure 5.26: *ROC curves comparing Model 1 with the Linearly Smoothed model.*

5.9 Conclusion

This chapter developed techniques for building behavioural scoring systems using proportional hazard regression.

A number of different models were discussed and compared with each other and with the application score. It was found that adding only one behavioural variable, Balance Difference specifically, results in a definite performance improvement over the application score. Also the plots of the parameter estimates for the application and the behavioural variables highlighted an interesting relationship between these variables, namely that the behavioural information is predictive at the early stages and the application information – at the later stages of the course of a loan.

In addition, a model which accumulated the behavioural information rather than using the latest only was discussed and examined. It was found to be slightly better but less uniform in its performance than the non-accumulated PHABS.

The ROC curve analysis has shown that the PHAB scores are competitive with the traditional logistic regression scores, especially after about 2 years into the course of a loan.

It was demonstrated on the example of the simplest PHAB model that one does not have to use the full set of the parameter estimates for every month of a loan's duration. The linear approximation of these parameter estimates is believed to be a more robust and simple to use model which matches the original PHAB in its performance.

Chapter 6

Conclusion

The aim of this thesis was to develop and research further the application of survival analysis to credit scoring and to demonstrate the advantages of using these new techniques and the trade-offs that have to be made by analysing real-life data. This concluding chapter summarises the main results of the research and brings us to the main advantage of using survival analysis for credit scoring – the ability to estimate customer's profit profile.

6.1 How to build a Proportional Hazards scorecard?

This thesis identified two developments that improve the present application of Cox's proportional hazards model to building of credit scoring models and that make it a competitive alternative to logistic regression.

Firstly, a new coarse-classing approach for characteristics in credit scoring data was developed. It uses survival analysis and has an advantage over the traditional log-odds method in that it does not require an arbitrary time horizon for the definition of 'bad'.

Secondly, we have demonstrated how a number of residual tools can be used to examine the fitness of proportional hazards models and we have discussed the advantages and disadvantages of each of these tools.

6.2 Is Proportional Hazards model as good as Logistic Regression in risk assessment?

The ROC curves (Figures 3.11-3.14) showed that the survival analysis model's performance is very close to that of the current industry standard approach, i.e. logistic regression, when used for the traditional purpose of classifying applicants into two groups. Thus, a lender who will adopt proportional hazards techniques for profit scoring will not loose any substantial predictive performance in scoring for risk.

6.3 How to overcome the restriction of the proportional hazards?

We have showed that proportional hazards model is a good alternative to logistic regression, however its main assumption is that hazards are proportional. It is not always the case, since very risky applicants become much less risky if they stay on the books for a long enough time. This is normally referred to as an interaction of a characteristic with time.

Several tests designed to detect such interactions in the data set were considered and Harrel's Z-test was found to be the most appropriate. We have illustrated the concept by building a model which included time-by-covariate interactions. This extension made the model much more flexible since it allowed the effect of a covariate on the predicted time to failure to increase or decrease as a loan evolves.

However, the interactions in the data were not large enough to result in a visible improvement in the performance or the time-dependent model over the non-time-dependent one. The analysis of a simulated data set helped to understand this issue and showed that the time dependent model performs better than the non-time-dependent one and that the improvement increases with time.

6.4 Are PHAB scores as good as they sound?

Behavioural scoring is as important as application scoring, especially when lenders want to focus on the profitability of the existing customers.

Proportional hazards behavioural (PHAB) scores were constructed, as an alternative to the logistic regression behavioural scores, using the survival analysis techniques applied for application scoring in the first half of the thesis. PHAB scores were expected not only to be competitive but to have a significant advantage over the traditional behavioural scores.

The reason for that is that the existing behavioural scoring systems based on logistic regression do not consider how long the loan has been running. Thus loans of all ages are lumped together. In the survival analysis approach one is trying to estimate when a loan will default, or rather how much longer will it survive. Hence, survival analysis techniques allow one to build a different behavioural model for each age of a loan, using customers' behavioural data up to that month to predict their remaining time to default.

The ROC curve analysis has shown that the PHAB scores are competitive with the traditional logistic regression scores, especially after about 2 years into a loan.

In addition, plots of the PHAB's parameter estimates suggested that the importance of application and behavioural scores changes over time – behavioural data is more important at the beginning and application data is more important later in the course of a loan (Figures 5.3-5.4).

6.5 Profit Calculation using PHAB scores

The main advantage of using survival analysis in credit scoring is that it enables one to estimate the 'survival' probability of a loan over time, i.e. the probability of receiving each of the monthly repayments. This allows one to estimate the expected profit from a loan, which is an important addition to scoring techniques since lenders are now moving from scoring only for risk to scoring for profitability, (Hopper and Lewis [1992], Thomas [1992], Leonard [1997]).

Proportional hazards models estimate survival probability for every customer. This was used in Chapter 3 to perform the traditional task of dividing customers in two groups of 'goods' and 'bads'. However, there are more uses of the information provided by PHAB scores. For example, one may use the survival probability profile they give to calculate the expected profit from a loan. At the time of application the profit would be:

$$\text{Profit(Application Time)} = \sum_{i=3}^{T+2} S_i \frac{a}{(1+r)^{i-2}} - L, \tag{6.1}$$

where

S_i is a survival probability to month i, i.e. the probability that a customer is still repaying a loan and has not defaulted at month i (estimated using (5.1));

a is the monthly repayment amount (instalment);

L is the amount of a loan;

T is the term of a loan and r is the monthly interbank lending rate.

In words, the expected profit from a loan is the sum of the present values of the instalments each multiplied by the probability of receiving it (the loan's survival probability), less the loan's amount. The summation starts from month 3 because the definition of default is 3 or more months delinquent, which is used in the survival function estimation. It stops at $Term+2$ to allow for delinquency starting in period $T-1$.

The expected profit at month K can be similarly calculated, namely

$$\text{Profit(Month K)*} = \sum_{i=1}^{T+2-K} S_i \frac{a}{(1+r)^i} - (1+r)^K \sum_{j=K}^{T} \frac{a}{(1+r')^j} - (B_K - (T-K)a), \tag{6.2}$$

where S_i is survival probability estimated using (5.2), r' is the monthly interest that the customer is paying to the lender and B_K is the actual balance at month K.

The first term is the expected present value of the repayments, where S_i is the survival probability that a customer will still be repaying a loan at month $K+i$. The second term is the present value of the capital still outstanding. $\sum_{j=K}^{T} \frac{a}{(1+r')^j}$

is the part of the initial loan amount that is yet to be repaid, priced at the start of a loan. The $(1 + r)^K$ transforms this into the value K months into the loan. This is only correct if repayments are up to date at month K. The third term allows for the situation when this is not true. The difference is between the actual balance and the expected balance if the repayments were on schedule. So if the repayments are behind schedule, more is owed than is expected, and this difference, expressed in terms of its value at K, is subtracted from the expected profit.

The expected profit was plotted against the application score, i.e. $\mathbf{x}'\beta$ from (5.1) in Figure 6.1. We assumed for simplicity that all loans are up to date and hence, ignored the last term in (6.2). Profit increases as $\mathbf{x}'\beta$ decreases, i.e. as risk decreases which is the expected effect risk would have on the return.

Figures 6.2-6.4 show similar plots but for time periods 6, 10 and 14 months into a loan using (6.2). Hence, the PHAB scores are useful as indicators of both risk and profit.

It can be seen that the profit curves for different terms of the loan cross. This shows that shorter duration loans are more profitable than longer duration loans when given to low scorers and shorter loans are more profitable for high scorers.

This means that one has to look at both the term of a loan and the score when ranking loans of similar amount.

The formula used to calculate profit can be altered to include a time-dependent interest rate and hence, to incorporate economic conditions into the model. Alternatively, the interest rate can be included as a covariate when estimating the survival function. This is similar to the approach of Duffie and Singgleton [1999], where the interest rate process in pricing of bonds is replaced by a default adjusted process incorporating the interest rate and the default hazard function.

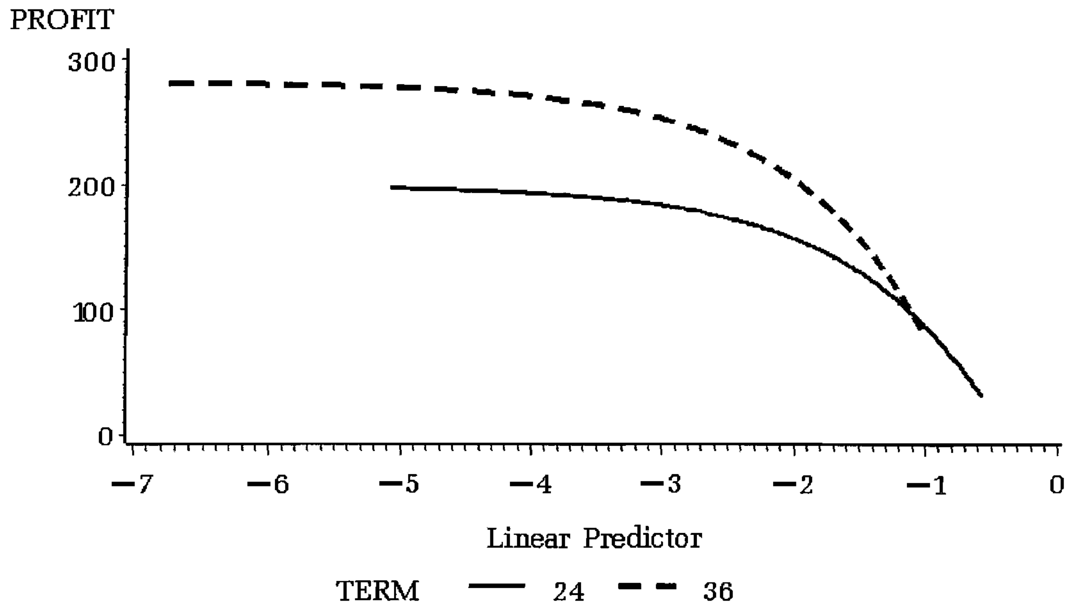


Figure 6.1: *The expected profit from personal loans of £2000 against the application score.*

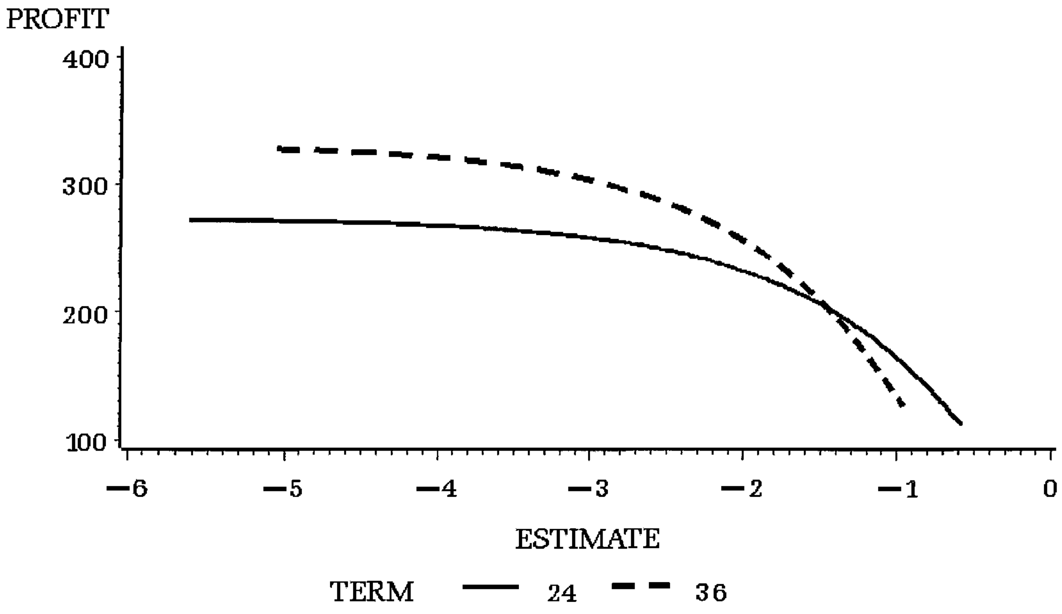


Figure 6.2: *The expected profit from personal loans of £2000 against the PHAB score for month 6.*

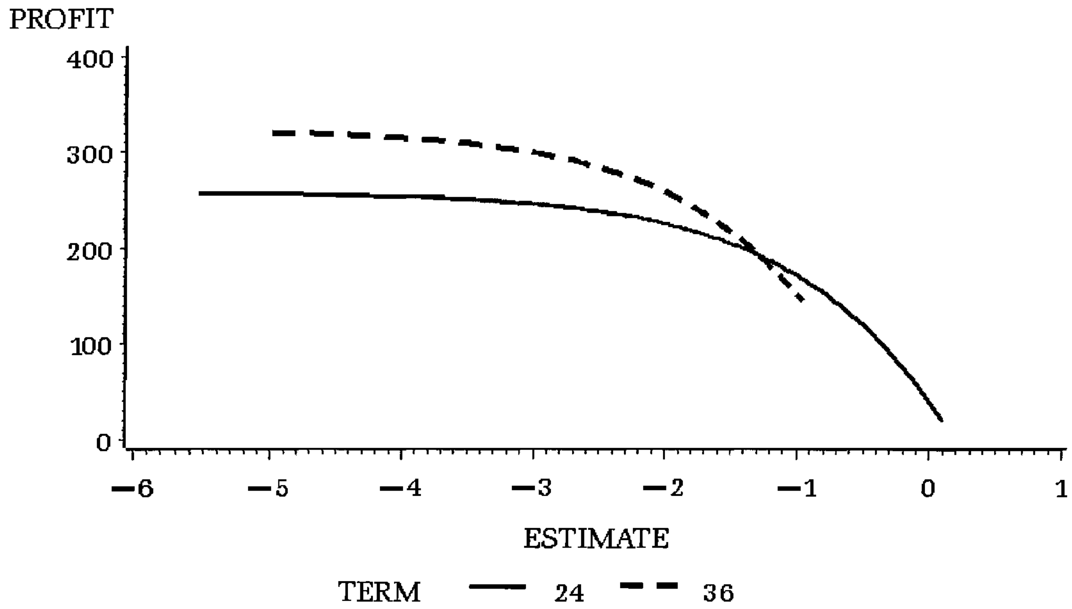


Figure 6.3: *The expected profit from personal loans of £2000 against the PHAB score for month 10.*

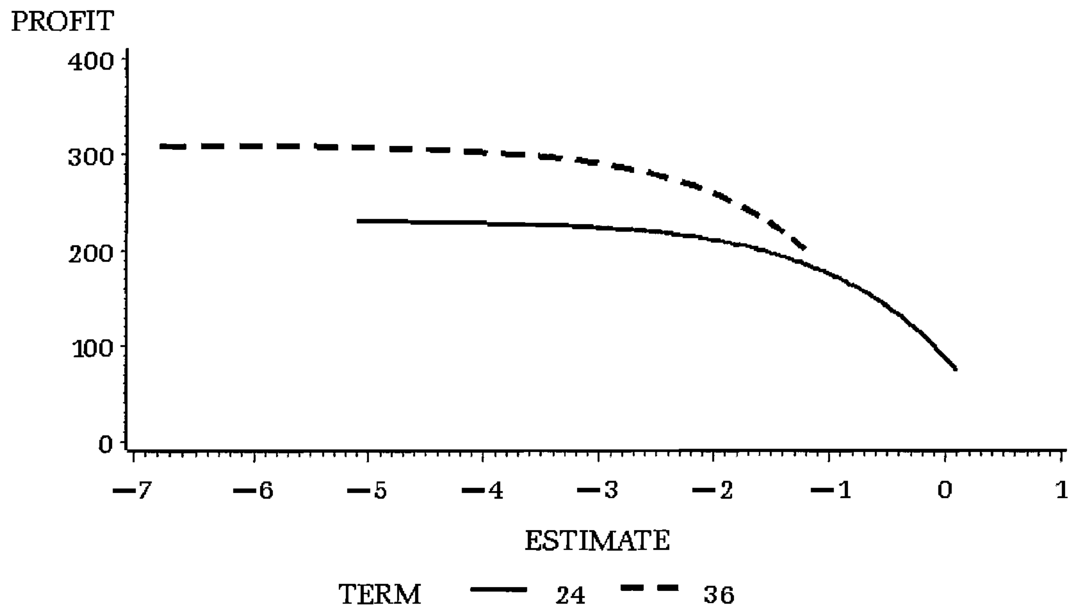


Figure 6.4: *The expected profit from personal loans of £2000 against the PHAB score for month 14.*

6.6 Contribution to knowledge

This thesis has developed a consistent strategy for building application and behavioural scoring models based on survival analysis tools, specifically – proportional hazards model. All the stages of the model building process, from coarse-classing of characteristics through to model diagnostics, can now be done using survival analysis techniques without borrowing any tools from the traditional approaches, such as logistic regression, and hence there is no need for specifying an arbitrary time horizon for the outcome. These survival analysis based application and behavioural scoring models were shown to be as good as the industry standard logistic regression when scoring for risk. However, this is using only a part of the information survival model provides. It was also shown that lenders can use the estimated time to an event, such as default, to estimate expected profit of a customer and hence to score for profit.

6.7 Suggestions for further research

Survival analysis is an extensive collection of techniques for analysing lifetime data. However, all the techniques were developed for either medical or reliability data. Credit scoring data has features that make some techniques non-applicable or not particularly useful. For example, a large number of observations makes some residual diagnostics non-informative. Hence, further research could focus on exploring usefulness of other survival models and techniques when applied to credit scoring data. Accelerated life models (see, for example, Collett [1994]), assuming some parametric form of the baseline hazard other than Weibull and exponential, can be used if data suggests a suitable distribution. Such models would be applicable if the proportional hazards assumption does not hold.

Also, accelerated life models seem to be a suitable framework for incorporating economic conditions into the scorecard. They differ from proportional hazards models in that the effect of the covariates is multiplicative on time rather than on the hazard function, so a covariate specifying some economic perturbation can accelerate (or decelerate) the time to default.

As we mentioned in the conclusion to Chapter 3, early repayment seems to depend more on the time left to maturity of a loan rather than the time from the start of a loan. Further research is needed to investigate if this is true for other data sets and hence, if early repayment models should be fitted to the time to maturity.

The profit model (6.2), proposed earlier in this chapter, can be made more sophisticated by replacing the constant interest rate r by a time series $r(t)$ to reflect economic conditions. Another way to include a time-dependent interest rate into the model is to use it as a covariate when estimating the survival probability of a credit.

Bibliography

The phreg procedure. In *SAS/STAT Software: Changes and Enhancements through Release 6.12*, pages 807–884. Cary, NC: SAS Institute Inc, 1998.

J.L. Banasik, J.N. Crook, and L.C. Thomas. Does scoring a subpopulation make a difference? *The international review of retail, distribution and consumer research*, 6(2):180–195, 1996.

M. Boyle, J.N. Crook, R. Hamilton, and L.C. Thomas. Methods for credit scoring applied to slow payers. In L.C. Thomas, J.N. Crook, and D.B. Edelman, editors, *Credit Scoring and Credit Control*, pages 75–90. Oxford: Oxford Univeristy Press, 1992.

N.E Breslow. Covariance analysis of censored survival data. *Biometrics*, 30:89–99, 1974.

D. Collett. *Modelling Survival Data in Medical Research.* Chapman & Hall, 1994.

I. Cooper and M. Martin. Default risk and derivative products. *Applied Mathematical Finance*, 3:53–74, 1996.

D.R. Cox. The analysis of exponentially distributed life-times with two types of failure. *Journal of the Royal Statistical Society, Series B*, 21:411–421, 1959.

D.R. Cox. Regression models and life-tables (with discussion). *Journal of Royal Statistical Society,Series B*, 74:187–220, 1972.

D.R. Cox and D Oakes. *Analysis of survival data.* Chapman and Hall, London, 1984.

D.R. Cox and E.J. Snell. A general definition of residuals (with discussion). *Journal of Royal Statistical Society, Series B*, 30:248–275, 1968.

H.A. David and M.L Moeschberger. *The theory of competing risks.* Griffin, London, 1978.

V.C. Desai, D.G. Convay, J.N. Crook, and G.A. Overstreet. Credit scoring models in the credit union enviroment using neural networks and genetic algorithms. *IMA Journal of Mathematics applied in Business and Industry*, 8:323–346, 1997.

D. Duffie and K.J Singgleton. Modeling term structures of defaultable bonds. *Review of financial studies*, 12(4):687–720, 1999.

B. Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association*, 72:557–565, 1977.

R.A. Eisenbeis. Pitfalls in the application of discriminant analysis in business, finance, and economics. *The Journal of Finance*, 32:875–900, 1977.

P.M. Grambsch and T.M. Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515–526, 1994.

D.J. Hand. *Discrimination and classification.* Wiley, Chichester, 1981.

D.J Hand. *Construction and assessment of classification rules.* Wiley, Chichester, 1997.

D.J Hand and W.E. Henley. Statistical classification methods in consumer credit scoring: a review. *Journal of Royal Statistical Society A*, 160:523–541, 1997.

W.E Henley. *Statistical aspects of credit scoring.* PhD thesis, Open University, 1985.

K.R. Hess. Graphical methods for assessing violations of the proportional hazards assumption in cox regression. *Statistics in Medicine*, 14:1706–1723, 1995.

M.A. Hopper and E.M. Lewis. Development and use of credit profit measures for account management. *I M A Journal of Mathematics Applied in Business and Industry*, 4:3–17, 1992.

R.A. Jarrow and S.M Turnbull. The intersection of market and credit risk. *Journal of Banking and Finance*, 24:271–299, 2000.

J.D. Kalbfleisch and R.L. Prentice. *The Statistical Analysis of Failure Time Data.* Wiley, New York, 1980.

D. Lando. *Three essays on contingent claims pricing.* PhD thesis, Cornell University, 1994.

D. Lando. Modelling bonds and derivatives with credit risk. In M. Dempster and S. Pliska, editors, *Mathematics of Derivative Securities*, pages 369–393. Cambridge Univeristy Press, 1997.

J.F Lawless. *Statistical models and methods for lifetime data.* Wiley, Nee York, 1982.

M.L. Leemis. *Reliability: probabilistic models and statistical methods.* Prentice-Hall, New Jersey, 1995.

K.J. Leonard. Behavior scores to predict profitability. In *Credit Scoring and Credit Control V*, 1997.

E.M. Lewis. *An Introduction to credit scoring.* Athena Press, San Rafael, California, 1992.

T. Moreau, J. O'Quigley, and M. Mesbach. A global goodness-of-fit statistic for the proportional hazards model. *Applied Statistics*, 34:212–218, 1985.

B. Narain. Survival analysis and the credit granting decision. In L.C. THOMAS, J.N. CROOK, and EDELMAN D.B., editors, *Credit Scoring and Credit Control*, pages 109–121. OUP, 1992.

N.H. Ng'andu. An empirical comparison of statistical tests for assessing the proportional hazards assumption of cox's model. *Statistics in Medicine*, 16:611–626, 1997.

R.M. Oliver and E. Wells. Portfolio cutoff strategies with profit, risk and volume constraints. In *Credit Scoring and Credit Control VI*, 1999.

A.N. Pettitt and I.B. Daud. Investigating time dependence in cox's proportional hazards model. *Applied Statistics*, 39:313–329, 1990.

A.K. Reichert, C. Cho, and G.M. Wagner. An examination of the conceptual issues involved in developing credit-scoring models. *Journal of Business and Economic Statistics*, 1(2):101–114, 1983.

D. Schoenfeld. Partial residuals for the proportional hazards regression model. *Biometrika*, 69:239–241, 1982.

V. Srinivasan and Y.H. Kim. The biermann-hausman credit granting model: a note. *Management Science*, 33:1361–1362, 1987a.

V. Srinivasan and Y.H. Kim. Credit granting: a comparative analysis of classification procedures. *Journal of Finance*, 42:665–683, 1987b.

D.M. Stablein, W.H. Carter Jr., and J.W. Novak. Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials*, 2:149–159, 1981.

T.M. Therneau, P.M. Grambsch, and T.R. Fleming. Martingale-based residuals for survival models. *Biometrika*, 77:147–160, 1990.

L.C. Thomas. Dividing credit card costs fairly. *I M A Journal of Mathematics Applied in Business and Industry*, 4:19–33, 1992.

L.C. Thomas. Methodologies for classing applicants for credit. In D.J. Hand, editor, *Statistics in Finance*, pages 83–100. London:Arnold, 1998.

L.C. Thomas, J. Banasik, and J.N. Crook. Not if but when will borrowers default. *Journal of Operational Research Society*, 50:1185–1190, 1999.

L.C. Thomas and M. Stepanova. Survival analysis methods for personal loan data. In *Credit Scoring and Credit Control VI*, 1999.

D. Van den Poel and J. Leunis. Database marketing modelling for financial services using hazard rate models. *The international review of retail, distribution and consumer research*, 8(2):243–257, April 1998.

A.D. Wilkie. Measures for comparing scoring systems. In L.C. Thomas, J.N. Crook, and D.B. Edelman, editors, *Statistics in Finance*, pages 123–138. Oxford: Clarendon Press, 1992.

M.B. Yobas and J.N. and Crook. Credit scoring using neural and evolitionary techniques. Technical report, Credit Research Centre, University of Edinburgh, 1997.

Appendix A

Proportional Hazards Regression output

A.1 Time to Early Repayment: Segmented by TERM

```
---------------------------- TERM3=t12_18 ----------------------------

                            The PHREG Procedure

Data Set: CREDITMY.CR97B Dependent Variable: OPEN Censoring
Variable: CENSORE Censoring Value(s): 0 Ties Handling: BRESLOW


                        Summary of the Number of
                        Event and Censored Values

                                                        Percent
                Total         Event      Censored       Censored

                12485          3121          9364          75.00


                Testing Global Null Hypothesis: BETA=0

                Without        With
Criterion     Covariates    Covariates    Model Chi-Square

-2 LOG L      57218.398     57041.291     177.108 with 41 DF
(p=0.0001) Score                .             .         181.330 with 41
DF (p=0.0001) Wald                 .            .          179.543 with
41 DF (p=0.0001)


                Analysis of Maximum Likelihood Estimates

                  Parameter    Standard      Wald          Pr >         Risk
  Variable DF      Estimate       Error   Chi-Square   Chi-Square      Ratio

  AGE01     1      0.039054     0.10925     0.12779       0.7207      1.040
  AGE02     1     -0.036766     0.09655     0.14501       0.7034      0.964
  AGE03     1     -0.182478     0.08843     4.25853       0.0391      0.833
  AGE04     1     -0.127609     0.09251     1.90270       0.1678      0.880
  AGE05     1     -0.290536     0.08278    12.31806       0.0004      0.748
```

AGEM	1	-0.278780	0.40159	0.48190	0.4876	0.757
AMOUNT01	1	0.227839	0.56847	0.16064	0.6886	1.256
AMOUNT02	1	0.215478	0.56822	0.14381	0.7045	1.240
AMOUNT03	1	0.156442	0.57102	0.07506	0.7841	1.169
AMOUNT04	1	0.056072	0.57080	0.00965	0.9217	1.058
CURRAD01	1	0.120558	0.07988	2.27773	0.1312	1.128
CURRAD02	1	0.119336	0.07987	2.23220	0.1352	1.127
CURRAD03	1	-0.025337	0.07969	0.10110	0.7505	0.975
CURRAD04	1	-0.015183	0.08415	0.03255	0.8568	0.985
CURRAD05	1	-0.015418	0.07769	0.03939	0.8427	0.985
CURRADM	1	-0.327385	0.19562	2.80076	0.0942	0.721
CURREM01	1	-0.023022	0.07787	0.08740	0.7675	0.977
CURREM02	1	-0.100932	0.09950	1.02902	0.3104	0.904
CURREM03	1	-0.156289	0.07608	4.22050	0.0399	0.855
CURREM04	1	-0.053890	0.08003	0.45346	0.5007	0.948
CURREM05	1	-0.115073	0.08457	1.85164	0.1736	0.891
CURREMM	1	0.046570	0.13133	0.12575	0.7229	1.048
GENDER01	1	0.040408	0.03815	1.12161	0.2896	1.041
FREQPA01	1	0.050358	0.04010	1.57709	0.2092	1.052
FREQPA02	1	-0.030717	0.10739	0.08182	0.7748	0.970
DEPKID01	1	0.104999	0.04577	5.26286	0.0218	1.111
DEPKID02	1	0.176585	0.18393	0.92178	0.3370	1.193
HOMOWN01	1	-0.071666	0.04294	2.78605	0.0951	0.931
HOMOWN02	1	0.077827	0.07830	0.98805	0.3202	1.081
HOMTEL01	1	-0.023082	0.07478	0.09528	0.7576	0.977
INPREM01	1	-0.370264	0.11580	10.22396	0.0014	0.691
INPREM02	1	0.059008	0.04533	1.69480	0.1930	1.061
INPREM03	1	0.166285	0.06282	7.00568	0.0081	1.181
INPREM04	1	0.266927	0.26275	1.03208	0.3097	1.306
INPREM05	1	0.879638	0.56862	2.39311	0.1219	2.410
JOINTS01	1	0.060568	0.05032	1.44855	0.2288	1.062
WEDDED01	1	0.071533	0.13452	0.28279	0.5949	1.074
WEDDED02	1	0.141289	0.13514	1.09301	0.2958	1.152
PURPE01	1	0.444302	0.06104	52.98828	0.0001	1.559
PURPE02	1	0.093074	0.04719	3.88937	0.0486	1.098
PURPE03	1	-0.043397	0.10985	0.15608	0.6928	0.958

---------------------------- TERM3=t24_30 ----------------------------

The PHREG Procedure

Data Set: CREDITMY.CR97B Dependent Variable: OPEN Censoring Variable: CENSORE Censoring Value(s): 0 Ties Handling: BRESLOW

Summary of the Number of
Event and Censored Values

Total	Event	Censored	Percent Censored
10703	4060	6643	62.07

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	72871.747	72658.827	212.920 with 40 DF (p=0.0001)
Score	.	.	214.724 with 40 DF (p=0.0001)
Wald	.	.	212.297 with 40 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Risk Ratio
AGE01	1	0.285902	0.10777	7.03754	0.0080	1.331
AGE02	1	0.233941	0.09597	5.94220	0.0148	1.264
AGE03	1	0.149551	0.08893	2.82786	0.0926	1.161
AGE04	1	0.110749	0.09201	1.44867	0.2287	1.117
AGE05	1	0.013119	0.08310	0.02492	0.8746	1.013
AGEM	1	-0.661627	0.36791	3.23401	0.0721	0.516
AMOUNT01	1	0.133386	0.27382	0.23729	0.6262	1.143
AMOUNT02	1	0.113633	0.26194	0.18819	0.6644	1.120
AMOUNT03	1	0.130416	0.26137	0.24897	0.6178	1.139
AMOUNT04	1	0.058583	0.26231	0.04988	0.8233	1.060
CURRAD01	1	0.092990	0.07649	1.47797	0.2241	1.097
CURRAD02	1	0.036063	0.07644	0.22256	0.6371	1.037
CURRAD03	1	0.025034	0.07614	0.10809	0.7423	1.025
CURRAD04	1	-0.052179	0.07994	0.42602	0.5139	0.949
CURRAD05	1	-0.053620	0.07490	0.51254	0.4740	0.948
CURRADM	1	-0.095882	0.15971	0.36043	0.5483	0.909
CURREM01	1	0.115243	0.07174	2.58034	0.1082	1.122
CURREM02	1	0.082607	0.08780	0.88515	0.3468	1.086
CURREM03	1	0.054533	0.06993	0.60818	0.4355	1.056
CURREM04	1	-0.010951	0.07379	0.02202	0.8820	0.989
CURREM05	1	0.157692	0.07582	4.32585	0.0375	1.171
CURREMM	1	0.352808	0.12251	8.29372	0.0040	1.423
GENDER01	1	-0.028997	0.03486	0.69199	0.4055	0.971
FREQPA01	1	0.038246	0.03627	1.11219	0.2916	1.039
FREQPA02	1	-0.238398	0.09489	6.31184	0.0120	0.788
DEPKID01	1	0.026417	0.03887	0.46197	0.4967	1.027
DEPKID02	1	0.241710	0.16337	2.18905	0.1390	1.273
HOMOWN01	1	-0.053614	0.03803	1.98719	0.1586	0.948
HOMOWN02	1	-0.028322	0.07176	0.15578	0.6931	0.972
HOMTEL01	1	-0.027243	0.06847	0.15832	0.6907	0.973
INPREM01	0	0	.	.	.	.
INPREM02	1	-0.077323	0.11333	0.46548	0.4951	0.926
INPREM03	1	0.015439	0.03883	0.15811	0.6909	1.016
INPREM04	1	0.015512	0.06347	0.05973	0.8069	1.016
INPREM05	1	0.048182	0.13375	0.12978	0.7187	1.049
JOINTS01	1	0.075711	0.04056	3.48426	0.0620	1.079
WEDDED01	1	0.129753	0.11282	1.32273	0.2501	1.139
WEDDED02	1	0.091297	0.11303	0.65241	0.4193	1.096
PURPE01	1	0.440902	0.04865	82.13348	0.0001	1.554
PURPE02	1	0.188839	0.03752	25.33605	0.0001	1.208
PURPE03	1	-0.062126	0.06049	1.05486	0.3044	0.940

-------------------------- TERM3=t36_48_60 --------------------------

The PHREG Procedure

Data Set: CREDITMY.CR97B Dependent Variable: OPEN Censoring
Variable: CENSORE Censoring Value(s): 0 Ties Handling: BRESLOW

Summary of the Number of
Event and Censored Values

Total	Event	Censored	Percent Censored
10588	3998	6590	62.24

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square

-2 LOG L   70550.242   70345.125   205.116 with 40 DF
(p=0.0001) Score   .   .   206.633 with 40
DF (p=0.0001) Wald   .   .   204.980 with
40 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Risk Ratio
AGE01	1	0.459781	0.12173	14.26727	0.0002	1.584
AGE02	1	0.246952	0.10368	5.67303	0.0172	1.280
AGE03	1	0.075780	0.09572	0.62672	0.4286	1.079
AGE04	1	0.052433	0.09797	0.28642	0.5925	1.054
AGE05	1	-0.062555	0.09129	0.46958	0.4932	0.939
AGEM	1	-0.031690	0.27676	0.01311	0.9088	0.969
AMOUNT01	1	0.322280	0.24103	1.78786	0.1812	1.380
AMOUNT02	1	0.267682	0.09679	7.64837	0.0057	1.307
AMOUNT03	1	0.258824	0.07859	10.84535	0.0010	1.295
AMOUNT04	1	0.161222	0.06754	5.69885	0.0170	1.175
CURRAD01	1	0.165079	0.08472	3.79652	0.0514	1.179
CURRAD02	1	0.130781	0.08467	2.38581	0.1224	1.140
CURRAD03	1	0.009591	0.08520	0.01267	0.9104	1.010
CURRAD04	1	0.031313	0.08816	0.12616	0.7224	1.032
CURRAD05	1	0.035758	0.08453	0.17894	0.6723	1.036
CURRADM	1	0.081136	0.15990	0.25746	0.6119	1.085
CURREM01	1	0.094103	0.07288	1.66703	0.1967	1.099
CURREM02	1	-0.016489	0.09159	0.03241	0.8571	0.984
CURREM03	1	-0.025606	0.07013	0.13332	0.7150	0.975
CURREM04	1	-0.016178	0.07302	0.04909	0.8247	0.984
CURREM05	1	0.020650	0.07486	0.07610	0.7827	1.021
CURREMM	1	0.107407	0.13131	0.66903	0.4134	1.113
GENDER01	1	0.033838	0.03689	0.84162	0.3589	1.034
FREQPA01	1	-0.031725	0.03817	0.69077	0.4059	0.969
FREQPA02	1	-0.238427	0.09854	5.85441	0.0155	0.788
DEPKID01	1	0.080638	0.03784	4.54112	0.0331	1.084
DEPKID02	1	0.104927	0.14537	0.52102	0.4704	1.111
HOMOWN01	1	0.056836	0.04114	1.90839	0.1671	1.058
HOMOWN02	1	0.023936	0.07581	0.09968	0.7522	1.024
HOMTEL01	1	0.065078	0.07966	0.66737	0.4140	1.067
INPREM01	0	0	.	.	.	.
INPREM02	1	0.511307	0.50499	1.02517	0.3113	1.667
INPREM03	1	0.017560	0.09098	0.03725	0.8469	1.018
INPREM04	1	0.046206	0.04707	0.96357	0.3263	1.047
INPREM05	1	0.039157	0.04191	0.87275	0.3502	1.040
JOINTS01	1	0.020458	0.03793	0.29093	0.5896	1.021
WEDDED01	1	0.205619	0.12332	2.78017	0.0954	1.228
WEDDED02	1	0.222456	0.12252	3.29667	0.0694	1.249
PURPE01	1	0.261777	0.04813	29.57610	0.0001	1.299
PURPE02	1	0.163734	0.03977	16.95195	0.0001	1.178
PURPE03	1	-0.101299	0.04776	4.49815	0.0339	0.904

------------------------------ TERM3=t6 ------------------------------

The PHREG Procedure

Data Set: CREDITMY.CR97B Dependent Variable: OPEN Censoring Variable: CENSORE Censoring Value(s): 0 Ties Handling: BRESLOW

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored

872	137	735	84.29

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	1832.668	1775.261	57.407 with 39 DF (p=0.0289)
Score	.	.	68.980 with 39 DF (p=0.0022)
Wald	.	.	63.646 with 39 DF (p=0.0076)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Risk Ratio
AGE01	1	-0.526370	0.53421	0.97087	0.3245	0.591
AGE02	1	-0.712137	0.50559	1.98393	0.1590	0.491
AGE03	1	-0.338535	0.44890	0.56873	0.4508	0.713
AGE04	1	-0.263725	0.47906	0.30306	0.5820	0.768
AGE05	1	-0.150551	0.42813	0.12366	0.7251	0.860
AGEM	1	-0.285607	1.46207	0.03816	0.8451	0.752
AMOUNT01	1	15.386638	2050	0.0000563	0.9940	4812068
AMOUNT02	1	15.300354	2050	0.0000557	0.9940	4414275
AMOUNT03	1	14.625705	2050	0.0000509	0.9943	2248345
AMOUNT04	1	14.768773	2050	0.0000519	0.9943	2594160
CURRAD01	1	0.208929	0.41066	0.25883	0.6109	1.232
CURRAD02	1	0.374555	0.40592	0.85143	0.3561	1.454
CURRAD03	1	-0.063028	0.43357	0.02113	0.8844	0.939
CURRAD04	1	0.767883	0.39931	3.69794	0.0545	2.155
CURRAD05	1	0.032393	0.42242	0.00588	0.9389	1.033
CURRADM	1	1.546056	0.95207	2.63701	0.1044	4.693
CURREM01	1	-0.344305	0.29490	1.36315	0.2430	0.709
CURREM02	1	-0.265023	0.38643	0.47036	0.4928	0.767
CURREM03	1	-0.475545	0.29174	2.65693	0.1031	0.622
CURREM04	1	-0.677051	0.34607	3.82757	0.0504	0.508
CURREM05	1	-0.804698	0.36725	4.80123	0.0284	0.447
CURREMM	1	-0.767862	0.64335	1.42455	0.2327	0.464
GENDER01	1	-0.181897	0.19395	0.87961	0.3483	0.834
FREQPA01	1	0.131296	0.20484	0.41084	0.5215	1.140
FREQPA02	1	0.150790	0.52848	0.08141	0.7754	1.163
DEPKID01	1	0.235765	0.24066	0.95978	0.3272	1.266
DEPKID02	1	1.510282	0.57526	6.89259	0.0087	4.528
HOMOWN01	1	-0.304488	0.21800	1.95079	0.1625	0.738
HOMOWN02	1	-0.076861	0.39993	0.03694	0.8476	0.926
HOMTEL01	1	0.383125	0.36646	1.09304	0.2958	1.467
INPREM01	1	0.120926	0.21670	0.31139	0.5768	1.129
INPREM02	1	0.012892	0.32633	0.00156	0.9685	1.013
INPREM03	1	1.466409	0.48364	9.19334	0.0024	4.334
INPREM04	0	0	.	.	.	.
INPREM05	0	0	.	.	.	.
JOINTS01	1	0.229898	0.29773	0.59623	0.4400	1.258
WEDDED01	1	-0.188637	0.63797	0.08743	0.7675	0.828
WEDDED02	1	-0.286707	0.63611	0.20315	0.6522	0.751
PURPE01	1	0.679643	0.31083	4.78096	0.0288	1.973
PURPE02	1	0.267419	0.27232	0.96434	0.3261	1.307
PURPE03	1	1.051138	0.47856	4.82441	0.0281	2.861

A.2 Time to Early Repayment: Non-segmented

The PHREG Procedure

Data Set: CREDITMY.CR97B Dependent Variable: OPEN Censoring

Variable: CENSORE Censoring Value(s): 0 Ties Handling: BRESLOW

Summary of the Number of
Event and Censored Values

Total	Event	Censored	Percent Censored
34648	11316	23332	67.34

Testing Global Null Hypothesis: BETA=0

Criterion Without Covariates With Covariates Model Chi-Square

-2 LOG L 224396.124 223478.131 917.993 with 44 DF
(p=0.0001) Score . . 941.088 with 44
DF (p=0.0001) Wald . . 924.574 with
44 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Risk Ratio
AGE01	1	0.218889	0.06387	11.74447	0.0006	1.245
AGE02	1	0.113977	0.05597	4.14728	0.0417	1.121
AGE03	1	-0.004090	0.05150	0.00631	0.9367	0.996
AGE04	1	-0.007699	0.05328	0.02088	0.8851	0.992
AGE05	1	-0.124447	0.04851	6.58173	0.0103	0.883
AGEM	1	-0.335910	0.18890	3.16227	0.0754	0.715
AMOUNT01	1	0.235964	0.07476	9.96331	0.0016	1.266
AMOUNT02	1	0.227027	0.07017	10.46692	0.0012	1.255
AMOUNT03	1	0.229695	0.06865	11.19378	0.0008	1.258
AMOUNT04	1	0.145619	0.06443	5.10755	0.0238	1.157
CURRAD01	1	0.130582	0.04575	8.14844	0.0043	1.139
CURRAD02	1	0.098864	0.04572	4.67580	0.0306	1.104
CURRAD03	1	0.008050	0.04579	0.03091	0.8604	1.008
CURRAD04	1	-0.002141	0.04782	0.00200	0.9643	0.998
CURRAD05	1	-0.010046	0.04510	0.04962	0.8237	0.990
CURRADM	1	-0.095104	0.09419	1.01952	0.3126	0.909
CURREM01	1	0.056517	0.04218	1.79508	0.1803	1.058
CURREM02	1	-0.008190	0.05276	0.02410	0.8766	0.992
CURREM03	1	-0.045744	0.04094	1.24866	0.2638	0.955
CURREM04	1	-0.038921	0.04300	0.81924	0.3654	0.962
CURREM05	1	0.017105	0.04445	0.14806	0.7004	1.017
CURREMM	1	0.186393	0.07249	6.61223	0.0101	1.205
GENDER01	1	0.008321	0.02093	0.15802	0.6910	1.008
FREQPA01	1	0.021916	0.02177	1.01366	0.3140	1.022
FREQPA02	1	-0.179293	0.05686	9.94335	0.0016	0.836
DEPKID01	1	0.067611	0.02316	8.51969	0.0035	1.070
DEPKID02	1	0.190717	0.09194	4.30273	0.0381	1.210
HOMOWN01	1	-0.030493	0.02307	1.74673	0.1863	0.970
HOMOWN02	1	0.014071	0.04300	0.10706	0.7435	1.014
HOMTEL01	1	0.003352	0.04222	0.00630	0.9367	1.003
INPREM01	1	-0.242513	0.09691	6.26236	0.0123	0.785
INPREM02	1	0.032475	0.03877	0.70178	0.4022	1.033
INPREM03	1	0.051642	0.02992	2.97965	0.0843	1.053
INPREM04	1	0.048298	0.03555	1.84533	0.1743	1.049
INPREM05	1	0.035777	0.03880	0.85015	0.3565	1.036
JOINTS01	1	0.058038	0.02406	5.81825	0.0159	1.060
WEDDED01	1	0.133026	0.07020	3.59051	0.0581	1.142
WEDDED02	1	0.146577	0.07019	4.36124	0.0368	1.158
PURPE01	1	0.380527	0.02949	166.53640	0.0001	1.463
PURPE02	1	0.147643	0.02338	39.86898	0.0001	1.159

PURPE03	1	-0.074027	0.03501	4.47143	0.0345	0.929
TERM01	1	-0.623756	0.09054	47.45946	0.0001	0.536
TERM02	1	-0.863262	0.09342	85.38486	0.0001	0.422
TERM03	1	-1.005892	0.09583	110.16920	0.0001	0.366

A.3 Time to Default: non-segmented

The PHREG Procedure

Data Set: CREDITMY.WK1 \\ Dependent Variable: OPEN \\ Censoring Variable: CENSOR \\ Censoring Value(s): 0 \\ Ties Handling: BRESLOW\\

Summary of the Number of
Event and Censored Values

Total	Event	Censored	Percent Censored
34648	1394	33254	95.98

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	27804.726	26691.192	1113.533 with 41 DF (p=0.0001)
Score	.	.	1152.725 with 41 DF (p=0.0001)
Wald	.	.	1066.757 with 41 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Risk Ratio
AGE01	1	-0.269382	0.11871	5.14969	0.0233	0.764
AGE02	1	-0.343628	0.10818	10.08963	0.0015	0.709
AGE03	1	-0.364341	0.11795	9.54149	0.0020	0.695
AGE04	1	-0.451764	0.13060	11.96610	0.0005	0.637
AGE05	1	-0.608729	0.15481	15.46132	0.0001	0.544
AGE06	1	-0.618968	0.40281	2.36117	0.1244	0.538
AMOUNT01	1	0.178986	0.10015	3.19415	0.0739	1.196
AMOUNT02	1	0.640693	0.17609	13.23875	0.0003	1.898
CURADD01	1	-0.114471	0.09696	1.39389	0.2377	0.892
CURADD02	1	-0.321211	0.09192	12.21174	0.0005	0.725
CURADD03	1	-0.425863	0.10586	16.18338	0.0001	0.653
CURADD04	1	-0.709043	0.09608	54.46244	0.0001	0.492
CURADD05	1	-1.027027	0.22311	21.19010	0.0001	0.358
CURADD06	1	-0.707675	0.23102	9.38353	0.0022	0.493
CUREMP01	1	-0.182657	0.09483	3.71044	0.0541	0.833
CUREMP02	1	-0.392496	0.08769	20.03552	0.0001	0.675
CUREMP03	1	-0.586363	0.09304	39.71983	0.0001	0.556
CUREMP04	1	-0.907697	0.11104	66.81988	0.0001	0.403
CUREMP05	1	-1.163392	0.17000	46.83564	0.0001	0.312
CUREMP06	1	-0.346176	0.20940	2.73298	0.0983	0.707
GENDER01	1	-0.100541	0.06095	2.72102	0.0990	0.904
FREQPA01	1	0.647060	0.06022	115.45990	0.0001	1.910
FREQPA02	1	0.886438	0.15183	34.08609	0.0001	2.426
DEPKID01	1	0.219537	0.06803	10.41441	0.0013	1.246
DEPKID02	1	0.346733	0.25797	1.80656	0.1789	1.414

HOMOWN01	1	-0.329396	0.06405	26.44543	0.0001	0.719
HOMOWN02	1	-0.220099	0.12245	3.23109	0.0723	0.802
HOMTEL01	1	-0.263810	0.10336	6.51389	0.0107	0.768
INPREM01	1	0.637909	0.10319	38.21795	0.0001	1.893
INPREM02	1	0.511085	0.07263	49.51720	0.0001	1.667
INPREM03	1	0.739772	0.11738	39.71703	0.0001	2.095
INPREM04	1	0.431300	0.16212	7.07802	0.0078	1.539
INPREM05	1	0.439056	0.20159	4.74355	0.0294	1.551
JOINTS01	1	0.371621	0.07519	24.42757	0.0001	1.450
WEDDED01	1	0.239690	0.07324	10.70923	0.0011	1.271
WEDDED02	1	0.083785	0.19396	0.18660	0.6658	1.087
PURPOS01	1	1.055140	0.06277	282.56579	0.0001	2.872
PURPOS02	1	0.351996	0.08239	18.25177	0.0001	1.422
PURPOS03	1	-0.035385	0.13821	0.06555	0.7979	0.965
TERM01	1	-0.036799	0.09568	0.14791	0.7005	0.964
TERM02	1	0.187143	0.14958	1.56523	0.2109	1.206
TERM03	0	0	.	.	.	.

A.4 Results of the test for time-by-characteristic interaction

OBS	_NAME_	_LABEL_	T12_18	T24_30	T36_48_6	T6
42	Z1	Z-statistic for AGE01	0.96407	3.11723	2.92220	0.4691
43	Z2	Z-statistic for AGE02	-2.47498	0.75595	-1.23693	-0.0038
44	Z3	Z-statistic for AGE03	-4.54529	-3.67380	0.15197	-0.9298
45	Z4	Z-statistic for AGE04	-0.03661	-1.80679	-1.26378	-0.0667
46	Z5	Z-statistic for AGE05	4.02323	1.96061	0.99392	-0.0952
47	Z6	Z-statistic for AGEM	-0.27564	-1.36400	-1.06867	-0.2587
48	Z7	Z-statistic for AMOUNT01	-1.71950	-0.48622	-0.28243	1.8250
49	Z8	Z-statistic for AMOUNT02	0.39952	-1.96234	-4.02650	-1.5428
50	Z9	Z-statistic for AMOUNT03	2.39164	1.35059	-1.64540	-0.4348
51	Z10	Z-statistic for AMOUNT04	-0.09864	0.92095	3.53058	-0.7520
52	Z11	Z-statistic for CURRAD01	-3.50615	-5.61950	-2.55947	0.3019
53	Z12	Z-statistic for CURRAD02	-1.20786	-0.59476	0.18336	-0.0065
54	Z13	Z-statistic for CURRAD03	-1.20147	0.41260	1.34558	0.5172
55	Z14	Z-statistic for CURRAD04	1.73816	1.97402	-0.77766	-2.0524
56	Z15	Z-statistic for CURRAD05	2.96093	3.93291	3.24661	0.7030
57	Z16	Z-statistic for CURRADM	-0.57851	-3.78587	-3.18673	0.6902
58	Z17	Z-statistic for CURREM01	-2.08013	1.05081	0.90232	-1.0010
59	Z18	Z-statistic for CURREM02	0.56383	-0.66019	0.72723	-0.4235
60	Z19	Z-statistic for CURREM03	-1.85181	0.06004	0.92253	0.2113
61	Z20	Z-statistic for CURREM04	0.86218	-0.17344	0.14129	0.4570
62	Z21	Z-statistic for CURREM05	2.69784	-0.29515	-0.62701	0.3786
63	Z22	Z-statistic for CURREMM	2.99812	0.24717	-2.85846	1.9192
64	Z23	Z-statistic for GENDER01	1.45055	0.20825	-0.50683	-0.7350
65	Z24	Z-statistic for FREQPA01	3.97853	9.00855	6.52022	0.1660
66	Z25	Z-statistic for FREQPA02	0.63266	-0.60663	-2.17336	0.8432
67	Z26	Z-statistic for DEPKID01	-0.60106	-1.62230	0.23928	-1.1719
68	Z27	Z-statistic for DEPKID02	0.13516	0.87384	-0.97782	1.4186
69	Z28	Z-statistic for HOMOWN01	-3.32696	-7.49313	-3.82417	-0.8616
70	Z29	Z-statistic for HOMOWN02	-0.25164	-1.56416	-2.41389	1.6493
71	Z30	Z-statistic for HOMTEL01	1.67806	-0.65622	0.71020	-3.2474
72	Z31	Z-statistic for INPREM01	1.06259	.	.	-0.4495
73	Z32	Z-statistic for INPREM02	2.48734	-1.20845	0.01151	-0.1482
74	Z33	Z-statistic for INPREM03	2.38192	4.99509	-2.02767	-0.1562
75	Z34	Z-statistic for INPREM04	1.08687	0.84549	1.99370	-14.4456
76	Z35	Z-statistic for INPREM05	1.78992	1.39365	1.16953	-5.7545
77	Z36	Z-statistic for JOINTS01	0.95138	1.62449	-2.35678	-1.1688
78	Z37	Z-statistic for WEDDED01	-2.12868	-0.59347	-1.16223	-1.4504
79	Z38	Z-statistic for WEDDED02	-0.43350	-0.73637	-1.65420	0.8701
80	Z39	Z-statistic for PURPE01	-0.01285	0.42922	0.69987	0.2439
81	Z40	Z-statistic for PURPE02	-1.35235	-5.55275	-6.18552	-0.6567
82	Z41	Z-statistic for PURPE03	1.66952	1.12390	5.86987	-1.7228