Survival Analysis Methods for Personal Loan Data

Lyn Thomas

Operations Research

SURVIVAL ANALYSIS METHODS FOR PERSONAL LOAN DATA

MARIA STEPANOVA and LYN THOMAS

Department of Management, University of Southampton, Southampton, United Kingdom, S017 1BJ
M.Stepanova@soton.ac.uk • l.thomas@soton.ac.uk

(Received November 1999; revision received August 2000; accepted October 2000)

Credit scoring is one of the most successful applications of quantitative analysis in business. This paper shows how using survival-analysis tools from reliability and maintenance modeling allows one to build credit-scoring models that assess aspects of profit as well as default. This survival-analysis approach is also finding favor in credit-risk modeling of bond prices. The paper looks at three extensions of Cox's proportional hazards model applied to personal loan data. A new way of coarse-classifying of characteristics using survival-analysis methods is proposed. Also, a number of diagnostic methods to check adequacy of the model fit are tested for suitability with loan data. Finally, including time-by-characteristic interactions is proposed as a way of possible improvement of the model's predictive power.

1. INTRODUCTION

Credit-scoring systems aid the decision of whether to grant credit to an applicant or not. Traditionally, this is done by estimating the probability that an applicant will default. This aim has been changing in recent years towards choosing the customers of highest profit. That change means it now becomes important not only if but when a customer will default (Thomas et al. 1999). It is possible that if the time to default is long, the acquired interest will compensate or even exceed losses resulting from default. Another factor that affects profitability is the cases in which customers close their account early, pay off the loan early by switching to another lender, or for other reasons. Depending on when the actual repayment occurred, the lender will lose a proportion of the interest on the loan.

It has been shown previously by Thomas et al. (1999) and Narain (1992) that survival analysis can be applied to estimate the time to default or to early repayment. Survival analysis is the area of statistics that deals with analysis of lifetime data. Examples of lifetime data can be found in medical or reliability studies, for example, when a deteriorating system is monitored and the time until event of interest is recorded.

The major strength of survival analysis is that it allows censored data to be incorporated into the model. This translates in the consumer credit context as a customer who never defaults, or never pays off early, so an event of interest is not observed. Clearly there is a great amount of such data because, luckily, most of the customers are "good."

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). In his Ph.D. thesis, Lando (1994) introduced a proportional hazards survival-analysis model to estimate the time until a bond defaults, the aim being to use economic variables as covariates.

In credit scoring we look for differences in application characteristics for customers with different survival times. Also, it is possible that there are two or more types of failure outcome. In consumer credit we are interested, in several possible outcomes when concerned with profitability: early repayment, default, closure, etc.

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 the accelerated life exponential model to 24 months of 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 a dimension to the standard approach. The author noted that these methods can be applied to any area of credit operations in which there are predictor variables and the time to some event is of interest.

Thomas et al. (1999) compared the performance of exponential with Weibull's, and Cox's nonparametric 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 payoff.

It was noted by Thomas et al. (1999) that there are several possible ways of improving the performance of the simplest survival-analysis models, such as Weibull's, exponential, or Cox's proportional hazards models.

In this paper we explore three extensions of Cox's proportional hazards model.

Subject classifications: Risk: estimating credit risk for personal loans. Failure models: Survival analysis applied to credit scoring models.
Area of review: FINANCIAL SERVICES.

Section 2 outlines the theory of methods used in the analysis. Section 3 looks at development of the techniques by applying them to personal loan data. The first improvement suggested is to coarse-classify the characteristic variables using survival-analysis techniques rather than the traditional approach. This not only keeps the whole approach consistent, but it means that at no point is it necessary to make arbitrary judgments about what time horizon is critical. In the traditional approach, failure before this time is considered "bad"; failure after it is considered "good." Section 3.1 looks at this new method of coarse-classifying, while §§3.2 and 3.3 apply it to predicting early repayment and default, respectively.

The second improvement is to use diagnostics to test the adequacy of the credit risk, and these are applied to the data in §4. The final improvement is to allow the decrease or increase in the effect of a covariate on the predicted time-to-failure as the loan evolves. Section 5 looks at this improvement, which overcomes the restriction (implicit in the proportional hazards assumption) that the same type of customer is most at risk at all times during the loan duration. Concluding remarks are found in §6.

2. SOME THEORY OF ANALYSIS OF LIFETIME DATA

Let T be the random variable representing time until repayment of a loan ceases,—i.e., time until default or early payoff. Then one way to describe the distribution of T is the hazard function, which is defined as follows:

$$h(t) = \lim_{\delta t \to 0} \left\{ \frac{P(t \leqslant T < t+\delta t | T \geqslant t)}{\delta t} \right\}. \tag{1}$$

This is the probability that an individual defaults, or pays off early, at time t, conditional on his or her having stayed on the books up to that time. Suppose now that one or more further measurements are available for each individual, so that we have a vector of covariates $\mathbf{x}$, e.g., application characteristics. We want to assess the relationship between the distribution of failure time and these covariates. Cox (1972) proposed the following model:

$$h(t; \mathbf{x}) = e^{(\mathbf{x}\beta)} h_0(t), \tag{2}$$

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

It is called the proportional hazards (PH) model because the assumption is that the hazard of the individual with application characteristics $\mathbf{x}$ is proportional to some unknown baseline hazard. Cox (1972) showed that one can estimate β without any knowledge of $h_0(t)$, just by using rank of failure and censored times. Therefore, if $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ are k ordered failure times and $R(t_{(i)})$ is the set of individuals at risk at $t_{(i)}$, then the likelihood function of observed data is:

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(x_{(i)}\beta)}{\sum_{l \in R_{(t_{(i)})}} \exp(x_l \beta)}. \tag{3}$$

Maximum-likelihood estimates of β are then found by maximizing the logarithm of Equation (3) using numerical methods.

Proportional hazards models assume that the hazard functions are continuous. However, credit performance data are normally recorded only monthly so that 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 $t_1, t_2, t_3, \ldots$. The exact likelihood function has to include all possible orderings of tied failures (Kalbfleisch and Prentice 1980), and hence is very difficult computationally.

A number of approximations has been developed. One of these is achieved by replacing Equation (2) by a discrete logistic model (Cox 1972):

$$\frac{h(t; \mathbf{x})}{1 - h(t; \mathbf{x})} = e^{(\mathbf{x}\beta)} \frac{h_0(t)}{1 - h_0(t)}, \tag{4}$$

where

$$h(t, \mathbf{x}) = P(t \leqslant T < t+1 | T \geqslant t). \tag{5}$$

To show that Equation (4) reduces to Equation (2) when time is continuous, note that the general form of discrete hazard would be

$$h(t, \mathbf{x})\delta t = P(t \leqslant T < t+\delta t | T \geqslant t). \tag{6}$$

Then Equation (4) becomes

$$\frac{h(t; \mathbf{x})\delta t}{1 - h(t; \mathbf{x})\delta t} = e^{(\mathbf{x}\beta)} \frac{h_0(t)\delta t}{1 - h_0(t)\delta t}, \tag{7}$$

and taking limit as time interval δt tends to zero gives Equation (2).

Let d_i denote the number of failures at t_i. 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 $s_R = \sum_{l \in R} x_l$ be the sum of the covariate vectors $\mathbf{x}$ over the individuals in 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 covariate vectors of these individuals.

The likelihood function arising from the Cox's model is

$$L_{\text{Cox}}(\beta) = \prod_{i=1}^{k} \frac{\exp(s'_{D_i}\beta)}{\sum_{R \in R(t_{(i)}; d_i)} \exp(s'_R \beta)}. \tag{8}$$

The summation in the denominator makes this difficult to calculate, so easier approximations were proposed by Breslow (1974) and Efron (1977).

The Efron likelihood is

$$L_E(\beta) = \prod_{i=1}^{k} \frac{\exp(s'_{D_i}\beta)}{\prod_{j=1}^{d_i} \left[\sum_{l \in R(t_{(i)})} \exp(x'_l \beta) - \frac{j-1}{d_i} \sum_{l \in D_i} \exp(x'_l \beta) \right]}. \tag{9}$$

The Breslow likelihood is

$$L_B(\beta) = \prod_{i=1}^{k} \frac{\exp(s'_{D_i}\beta)}{\left[\sum_{t \in R(t_{(i)})} \exp(x'_l \beta) \right]^{d_i}}. \tag{10}$$

2.1. Competing Risks

So far we have defined T as time until failure. It is possible that there are two or more possible risks that "compete" to be the cause of failure, e.g., default and early repayment. One can estimate time until default T_1, assuming all other observed lifetimes to be censored; and, separately, time until early repayment T_2, assuming all other observed lifetimes to be censored. Therefore, survival analysis can be performed separately on T_1 and T_2. Then predicted lifetime of the loan is estimated as $T = \min\{T_1, T_2,$ term of the loan$\}$ (Thomas et al. 1999).

2.2. Model Diagnostics

As in other standard procedures, when the proportional hazards model is fitted, one should examine model fit. Residuals are the most popular diagnostic, and they are usually some form of measure of discrepancy between fitted and predicted values. Therefore, if the model is adequate, the plot of residuals should not show any unexpected patterns. The simplest residuals are from linear regression and are calculated as a difference between predicted and actual values. Their plot is expected to be a random scatter about zero. 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 credit-risk models are: Does the proportional hazards assumption hold, do any covariates have to be transformed, and are there any outliers—individuals with repayment lifetimes greater than expected—that might have an unwanted impact on parameter estimates?

Cox-Snell residuals (Cox and Snell 1968) are defined as follows:

$$r_{C_i} = \exp(\hat{\beta} x_i)\widehat{H}_0(t_i) = \widehat{H}_i(t) = -\log \widehat{S}_i(t_i), \tag{11}$$

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

It can be shown (Collett 1994) that $-\log S(t)$ 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 $\widehat{S}_i(t_i)$ will be close and will have similar properties to $S(t_i)$. Hence $-\log \widehat{S}(t_i) = r_{C_i}$ will be a set of observations from an exponential distribution with unit mean. To test that the residuals have unit exponential distribution, the Kaplan-Meier estimate of these values is computed. Therefore, $\log(-\log \widehat{S}(r_{C_i}))$ is plotted against $\log(r_{C_i})$. A straight line with unit slope and zero intercept indicates that the fitted model is correct.

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

$$r_{M_i} = \delta_i - r_{C_i}. \tag{12}$$

It 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. r_{M_i} is plotted against the rank order of time. Ideally, it should not exhibit any pattern if the model is adequate.

Deviance residual, proposed by Therneau et al. (1990), is defined as:

$$r_{D_i} = sgn(r_{M_i})[-2\{r_{M_i} + \delta_i \log(\delta_i - r_{M_i})\}]. \tag{13}$$

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

If $\mathbf{x_i} = (x_{i1}, \ldots, x_{ip})'$ is a vector of covariates for the ith individual, R_i is the risk set, i.e., a set of indices of individuals 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

$$r_{ik} = x_{ik} - E(x_{ik}|R_i). \tag{14}$$

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_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, a value for every covariate. Schoenfeld residuals are plotted against rank order of time for their corresponding covariate. These plots are then used in investigating whether the particular covariate needs a transformation or whether there is an indication of time dependency.

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

3. SURVIVAL ANALYSIS TECHNIQUES APPLIED TO LOAN DATA

The survival-analysis techniques were applied to personal loan data from a major U.K. financial institution. The data set consisted of the application information of 50,000 personal loans, together with the repayment status for each month of the observation period of 36 months. Application characteristics available in the data set are found in Table 1. The status variable observed whether they had defaulted, paid off to term, paid off early, or the loan was still open. The borrowers were all U.K. 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 loans were needed were summarized in Table 2.

Table 1. Application characteristics used in the analysis.

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

3.1. Coarse-Classifying Using the 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. Then, each "band" or "group" is replaced with a binary dummy variable.

Traditional approaches of finding the suitable splits involve looking at the good-bad ratio or related measures for different values of the characteristic, and then grouping values with similar good-bad ratios. 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 survival-analysis modeling 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 to use the survival-analysis approach to coarse-classify the variables.

The following method was used for 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 hazard model to these binary variables.
4. Chart parameter estimates for all bands.
5. Choose the splits based on similarity of parameter estimates.

For discrete characteristics such as purpose of the loan, a binary variable is created for each attribute of the characteristic and then the method is the same as for a continuous characteristic.

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

3.1.1. Example. There are 27 different purposes of a loan in the data. Some purposes are very rare and hence have very few observations. These were combined with other more frequent purposes, so that the proportional hazards model was fitted to 22 binary variables indicating 22 purpose groups. Figure 1 shows a chart of 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 parameters, i.e., purposes with highest risk of early repayment, the second one has purposes with the middle values of parameters, and the third one has purposes with the lowest parameters.

Table 2. Purposes of personal loan.

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

Figure 1. PH parameter estimates for purpose of loan for early repayment.

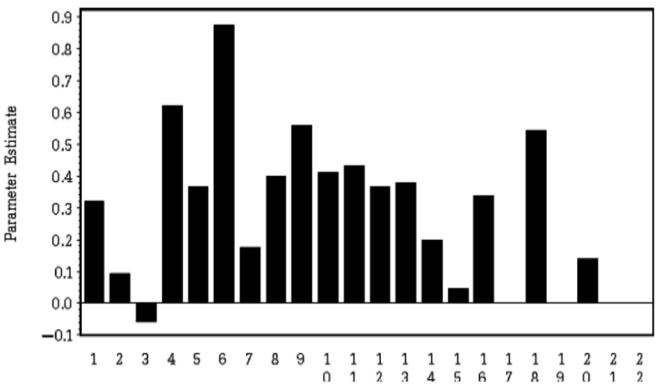


Figure 2. PH parameter estimates for purpose of the loan for default.

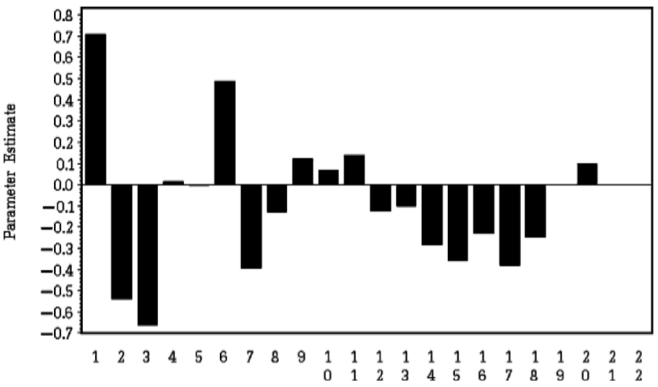


Figure 2 shows a chart of parameter estimates from PH regression predicting default for 22 purpose groups. Notice that the most risky purposes for early repayment are not the most risky purposes for default. This illustrates the importance of doing variable splits independently for each failure type.

The same method was used to achieve the best grouping on terms of the loan. It will be shown later that this variable is best used to segment the population, so that a scorecard is built on each segment. The best segmentation for predicting early repayment is to take 6-month term by itself, 12- and 18-month terms together, 24- and 30-month terms together, and more than 30 months together. It was found that for models predicting default, it is best not to group any of the terms of the loan together, so one builds a separate scorecard for each loan term.

3.2. Predicting Early Repayment

In line with the competing risk approach discussed in §2.1, the loans that are paid off early are considered "failures" while all others are considered censored. The results are presented using a comparison of Cox's proportional hazards model (PH) with a logistic regression approach (LR) under two criteria:

1. Estimating which loans will be paid off early within the first 12 months (Table 3, 1st year).

2. Estimating which loans, which are still repaying after 12 months, will pay off early within the next 12 months (Table 3, 2nd year).

ROC curves were also produced to compare performance of the models under the above criteria (Figures 4–5).

Two separate LR models were built on the training sample for each of these definitions. One PH model was fitted to the times until early payoff, considering all other outcomes to be censored.

To compare LR and PH models, the latter were measured under two criteria whose two definitions are as follows:

1. PH model gives the ordering of relative likelihood to pay off early, i.e., for each customer there is a "score" that 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 number of predicted "bads" equals actual number of "bads" in some holdout sample, i.e., 2,928 customers paid off early in the first year, so the 2,928 with the highest probability of failure in the first year under PH are predicted to be "bad."

3. The numbers of "bads" and "goods" correctly classified by the models in the holdout sample are compared (see Table 3).

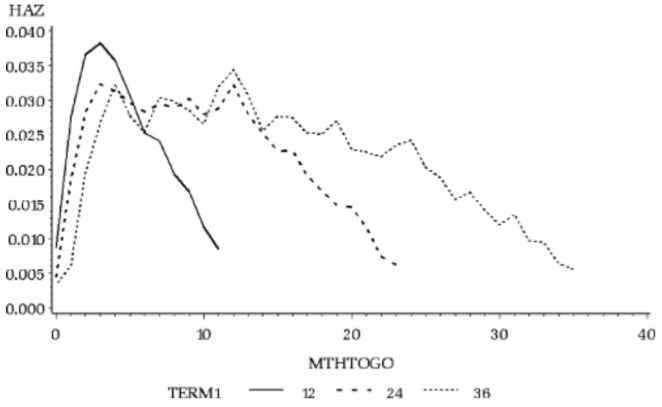
In the case of nonsegmented models, LR slightly outperforms PH. This is not surprising because only one PH model was used to classify according to both definitions, as opposed to two LR models fitted specifically to each definition. Notice that the PH model loses a lot of power in the second year. On investigating this further, it was realized that the term of the loan was a very significant characteristic that correlated strongly with the other variables, including time to early repayment. When the models were segmented by term (see Table 3) and their performance compared in the same ways as for nonsegmented scorecards, the results for the PH model were much better. In the segmented case, PH performs better than LR in the second year. This suggests that the time to early repayment is strongly affected by the term of the loan. This is supported by Figure 3, which suggests 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

Table 3. Comparison of the number of correctly classified accounts for the PH and LR models predicting early repayment.

		1st Year				2nd Year			
		Act. G and Pred. G*	Act. G but Pred. B	Act. B but Pred. G	Act. B and Pred. B	Act. G and Pred. G	Act. G but Pred. B	Act. B but Pred. G	Act. B and Pred. B
	Actual Nos.	11,964	0	0	2,928	6,274	0	0	1,825
PH	Nonsegmented	9,802	2,162	2,162	**766**	4,843	1,431	1,431	**394**
	Segmented by Term	9,765	2,199	2,199	**729**	4,981	1,293	1,293	**532**
LR	Nonsegmented	9,820	2,144	2,144	**784**	4,984	1,289	1,289	**536**
	Segmented by Term	9,768	2,196	2,196	**732**	5,000	1,273	1,273	**552**

*act. G and pred. G = actual Good and predicted Good by the model.

Figure 3. Hazard rate for early repayment for different terms of the loan, plotted versus number of months remaining to final repayment.



using the age of the loan, one can only translate this into time-until-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 are produced. Figure 4 shows the ROC curves for PH and LR models without segmentation, and confirms the poor performance of the PH in the second year. Figure 5 shows the ROC curve results when the data are segmented by term. Figure 5a corresponds to loans of 12 and 18 months. Only early repayment in the first 12 months can be observed for this segment because there are too few loans with the 18-month term. Figures 5b and 5d are the results for loans with terms of 24 and 30 months, 5b being the ROC curve for early repayment in the first year of the loan and 5d early repayment in the second year of the loan. Figures 5c and 5e are the similar ROC curves for loans that were to be repaid in three or more years. The PH and LR ROC curves are very similar to one another.

SAS statistical software was used to fit both the PH and the LR models with procedures PHREG and LOGISTIC, respectively. There are three options of treatment of ties available in the PHREG procedure: "Breslow," "Efron," and "discrete," which correspond to three different approximations of exact likelihood as discussed in §2. The SAS statistical package recommends "discrete" for the data that contain large number of ties. Table 4 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 the Breslow approximation. The smaller value of the log-likelihood statistic indicates better fit to the data. Therefore, these log-likelihood values suggest that the discrete approximation has given a much better fit in all four cases, but there was almost no difference in parameter estimates and no difference in the number of correctly classified accounts between the two methods. This suggests that the Breslow approach is a good approximation, and because it is by far the fastest method of the three, it was used for the majority of calculations.

3.3. Predicting Default

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

The results are presented using the comparison of Cox's proportional hazards model with a logistic regression approach, again under the two criteria:

1. Estimating which loans will default within the first 12 months (Table 5, 1st year).
2. Estimating which loans, which are still repaying after 12 months, will default within the next 12 months (Table 5, 2nd year).

Table 5 shows the results on a holdout sample using a cutoff where the number of "bads" predicted agrees with the number of "bads" in the sample. The results suggest there is little difference between LR and PH in either the first or the second year, and that segmentation has a less dramatic improvement on PH results under the default criterion that it did under the early repayment criterion.

These results are confirmed by the ROC curves of Figures 6–7. Figure 6 shows that without segmenting on

Figure 4. ROC curves for PH and LR predicting early repayment.

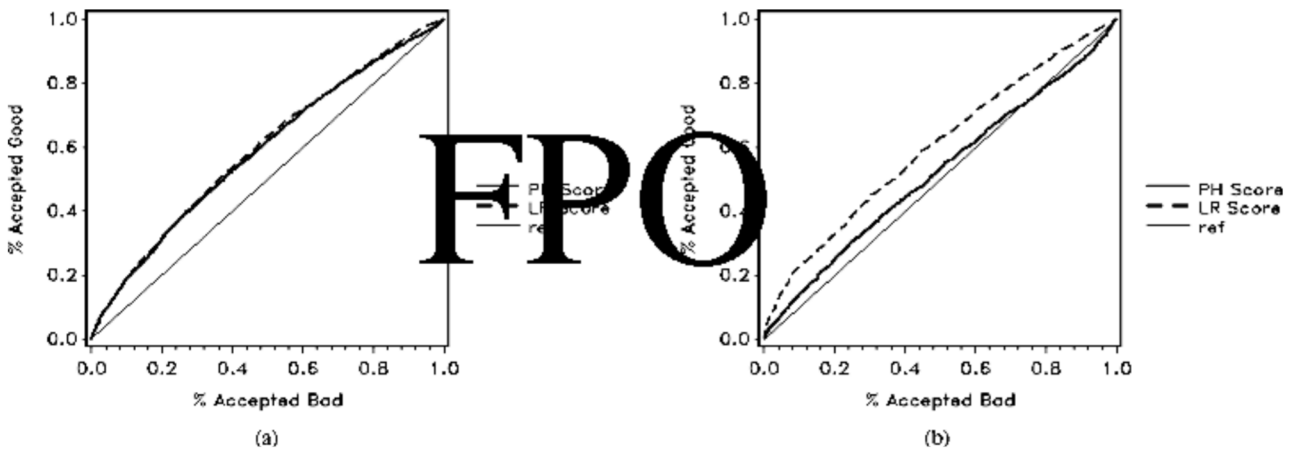
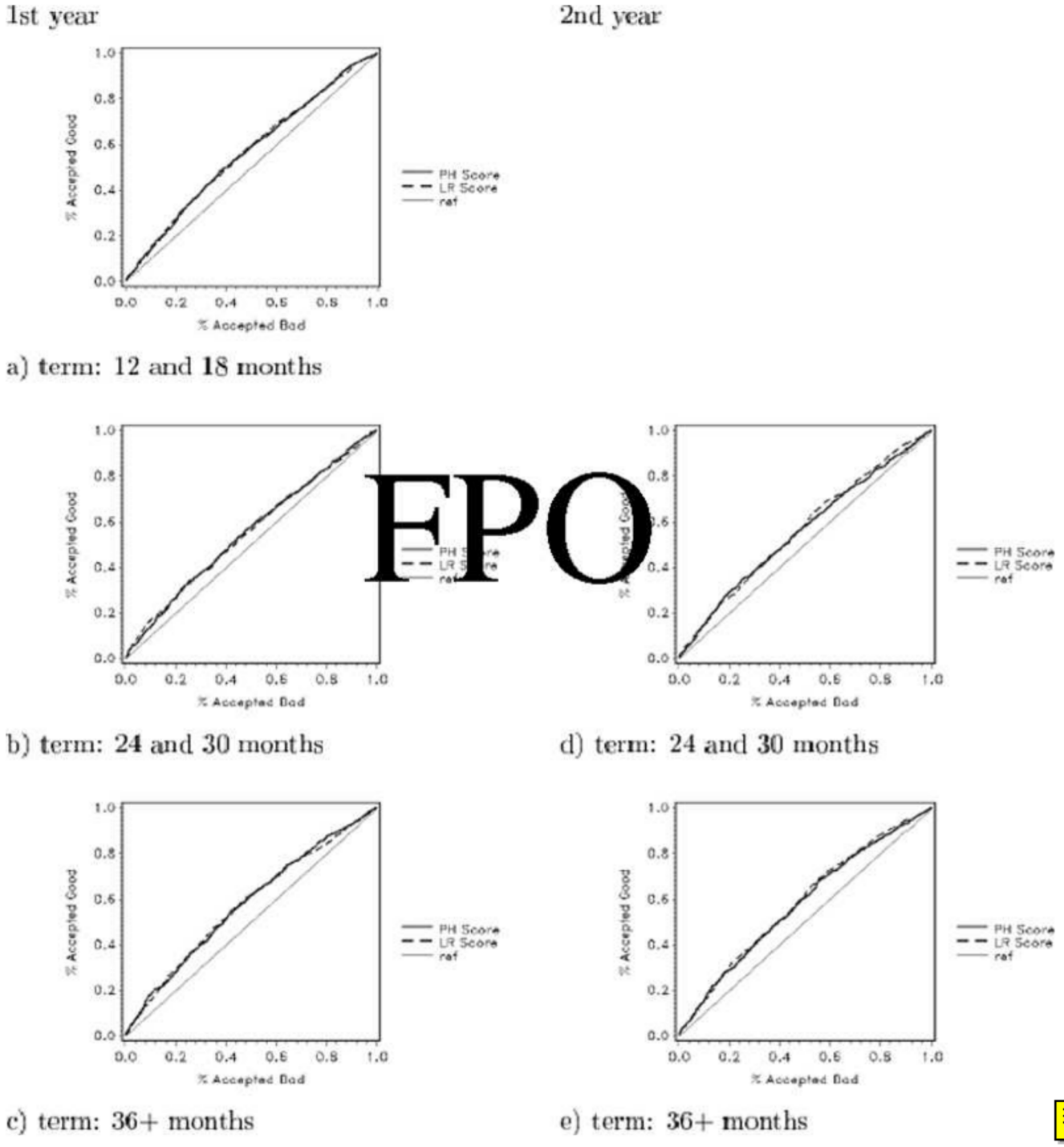


Figure 5. ROC curves for PH and LR predicting early repayment for different loan terms.



the term of the loan, LR and PH give very similar results in both the first (a) and the second (b) years.

Figure 7 shows the ROC curve results when the data are segmented by term, for the three most common terms of the loan. Figures 7a and 7d correspond to loans of 24 months, with 7a being the ROC curve for default in the first year of the loan, and Figure 7d default in the second year of the loan. Figures 7b and 7e are the similar ROC curves for loans with the term of 36 months, and Figures 7c and 7f are the ROC curves for loans that were to be repaid in 48 months. It should be noted that segmentation by term of the loan has less effect in predicting default than early repayment because default rate is independent of term of the loan and early repayment is not.

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

Table 4. Comparison of the models fit for two different methods for the treatment of ties—discrete and Breslow.

	−2Log(Likelihood)	
Term Months	Discrete	Breslow
12–18	28,956.193	57,041.291
24–30	38,374.029	72,658.827
36–48–60	39,189.753	70,345.125
6	1,111.873	1,775.261

explains why segmentation works so much better for early repayment than for default.

The hazard functions plot in Figure 3 suggests that early repayment is influenced most by time left to the maturity of the loan. However, logistic regression and proportional hazards are using time from the start of the loan. If one segments by term, then the time from the start is a linear transformation of time to maturity.

4. COMPARISON OF MODEL DIAGNOSTIC METHODS

Cox-Snell residuals were calculated from Martingale residuals, as discussed in §2.2. To examine whether Cox-Snell residuals have unit exponential distribution, the Kaplan-Meier estimate of the survivor function is obtained and log-log transformation of these values is plotted against log of the corresponding residual (Figure 8).

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 observations should be considered at all because repayment after one month is not necessarily a typical or normal feature of a personal loan portfolio. If we ignore these observations, we can conclude that the model fits data well.

Martingale residuals were plotted against rank order of time (Figure 9). The values appear in two bands, one representing uncensored observations and another representing censored ones. This is because Martingale residuals are always negative for the censored observations. The scatter of the points within a band increases with rank order of time. It is expected because calculation of r_{M_i} involves an estimated survivor function that close to one for earlier times, hence its logarithm is close to zero and r_{M_i} will be clustered around one for uncensored and around zero for censored observations.

Deviance residuals are very similar in appearance to Martingale residuals.

There are no clear outliers. Because 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 the number of observations is fewer. Because of the large number of observations, the explainable patterns are clearly visible and overshadow any other systematic features or outliers.

Schoenfeld residuals (sample plot is in Figure 10) were plotted to investigate the following:

1. whether any covariates need to be transformed;
2. whether the effect of a covariate on the survival time changes over time.

This diagnostic is very laborious when the number of covariates is as large as in our data. The plots do not show any signs of time dependency or transformation.

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

5. TIME-DEPENDENT EFFECTS OF COVARIATES

We return to the model formulation to show one of the extensions of the model and see how it works on the example. Suppose we have just one covariate:

$x_1 = 1$ if purpose of the loan is refinance,

$x_1 = 0$ otherwise.

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

$$h(t; x_1) = e^{(\beta_1 x_1)} h_0(t). \tag{15}$$

If it is not, refinance $(x_1 = 0)$

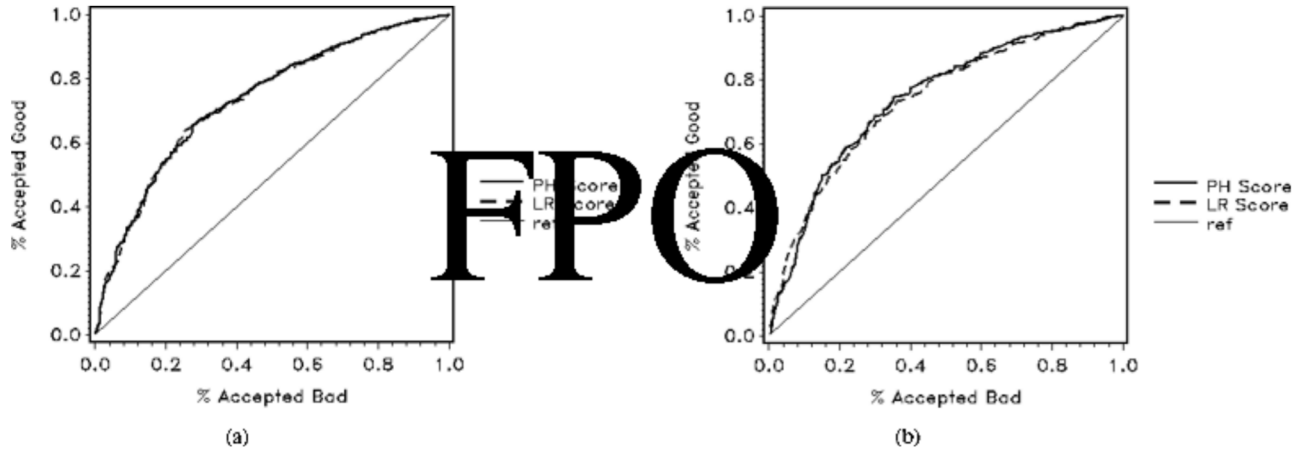
$$h(t; x_1 = 0) = h_0(t). \tag{16}$$

Table 5. Comparison of the number of correctly classified accounts for the PH and LR models predicting default.

		1st Year				2nd Year			
		Act. G and Pred. G*	Act. G but Pred. B	Act. B but Pred. G	Act. B and Pred. B	Act. G and Pred. G	Act. G but Pred. B	Act. B but Pred. G	Act. B and Pred. B
	Actual Nos.	14,495	0	0	397	7,915	0	0	184
PH	Nonsegmented	14,145	351	351	**46**	7,747	168	168	**16**
	Segmented by Term	14,149	346	346	**51**	7,752	163	163	**21**
LR	Nonsegmented	14,145	350	350	**47**	7,748	166	166	**18**
	Segmented by Term	14,145	351	351	**46**	7,752	162	162	**22**

*act. G and pred. G = actual Good and predicted Good by the model.

Figure 6. ROC curves for PH and LR predicting default.



If $x_1 = 1$,

$$h(t; x_1 = 1) = e^{\beta_1} h_0(t). \tag{17}$$

e^{β_1} is called relative hazard. Notice that it is independent of time. Therefore, no matter how long a loan stays on the books, if the purpose is refinance, it will always be considered to be more likely to pay off early when compared to other loans. This is questionable. To check this, define a variable $x_2 = x_1 t$, which represents an interaction of refinance indicator with time. Then add this variable to the model:

$$h(t; \mathbf{x}) = e^{(\beta_1 x_1 + \beta_1 x_2)} h_0(t). \tag{18}$$

Notice that now the relative hazard for loans on refinance to others is $e^{(\beta_1 + \beta_2 t)}$, which depends on 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 because the ratio of the hazards is not constant. Stablein et al. (1981) actually fitted the time-by-covariate interaction in the model to account for nonproportional hazard functions.

5.1. Test for Time-Dependency

In credit data the number of predictor variables, or characteristics, is usually large. As a result, before including a time-by-characteristic interaction in the model, it's desirable to screen all the characteristics for possibility of the time-dependent effect. A large number of graphical and numerical tests were developed for testing for time dependency. Ng'andu (1997) compares the performance of the five most popular numerical tests:

1. time-dependent covariate method (described in §5);
2. Harrel's linear correlation test;
3. weighted residuals score test;
4. score process; and
5. omnibus test;

for different scenarios of nonconstant hazard ratio.

He finds that 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 the time-dependent covariate because it is close to the time-dependent covariate test in power, and is also computationally simple. It is based on Fisher's z-transform of the Pearson correlation between Schoenfeld residuals of the model and rank order of time. Schoenfeld residuals can be requested as an additional output from PHREG in SAS. The test statistics for testing the hypothesis of $\rho = 0$ is

$$Z = \rho \sqrt{(n_u - 2)/(1 - \rho^2)}, \tag{19}$$

where ρ is the correlation between Schoenfeld's residuals (Schoenfeld 1982) and failure time order and n_u is the total number of uncensored observations.

Z is a normal deviate and tends to be positive if the hazard ratio for high values of the covariate increases over time, and it tends to be negative if this hazard ratio decreases over time.

5.2. A Model with Time-by-Characteristic Interactions

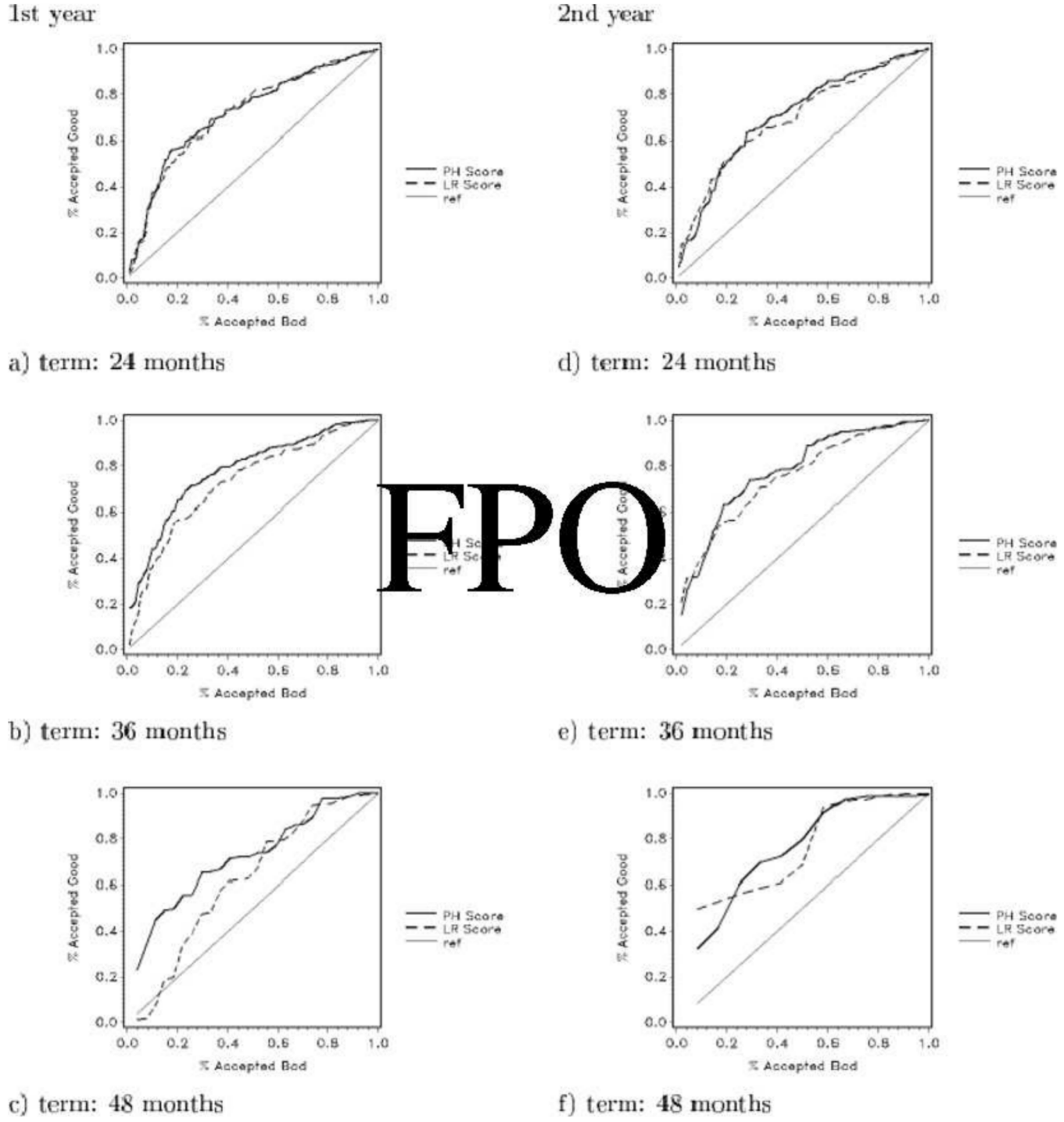
Schoenfeld residuals were calculated when fitting the proportional hazards model to predict early repayment. Harrel's Z-test was then performed for all the covariates. The statistic is a normal deviate, so its value should be compared with normal distribution tables to test for significance.

Harrel's Z-test suggested several possible time-by-covariate interactions. We chose only a few of them to illustrate the idea of time-by-covariate interaction.

Parameter estimates from proportional hazards regression predicting early repayment when no time-by-covariate interactions are compared with parameter estimates when time-by-covariate interactions are included. Consider, for example, PURPE02, which is an indicator for a medium-risk group of loan purposes.

Proportional hazards regression gave an estimate of $\beta_1 = 0.157$, i.e., the hazard to pay off early for a customer from

Figure 7. ROC curves for PH and LR predicting default for different loan terms.



the medium-risk purpose group is $e^{0.157} = 1.17$ times higher than for others. When time-by-characteristic was added and the model refitted, the estimates were $\beta_1 = 0.32$ and $\beta_2 = -0.01$. Hence at $t = 1$ month into the loan, the hazard for a customer from the medium-risk purpose group is $e^{0.31} = 1.36$ times higher than for others. After 18 months, at $t = 18$, the hazard to pay off early is $e^{(0.32-0.01*18)} = e^{0.14} = 1.15$ times higher for medium-risk purpose customers than for others.

Therefore, including time-by-characteristic interaction in the credit scoring for loan data adds another dimension—flexibility to reflect an increase or decrease of the effect of a characteristic with the age of the loan.

6. CONCLUSION

Credit scoring is one of the most successful applications of quantitative analysis in business.

Figure 8. Cox-Snell residuals.

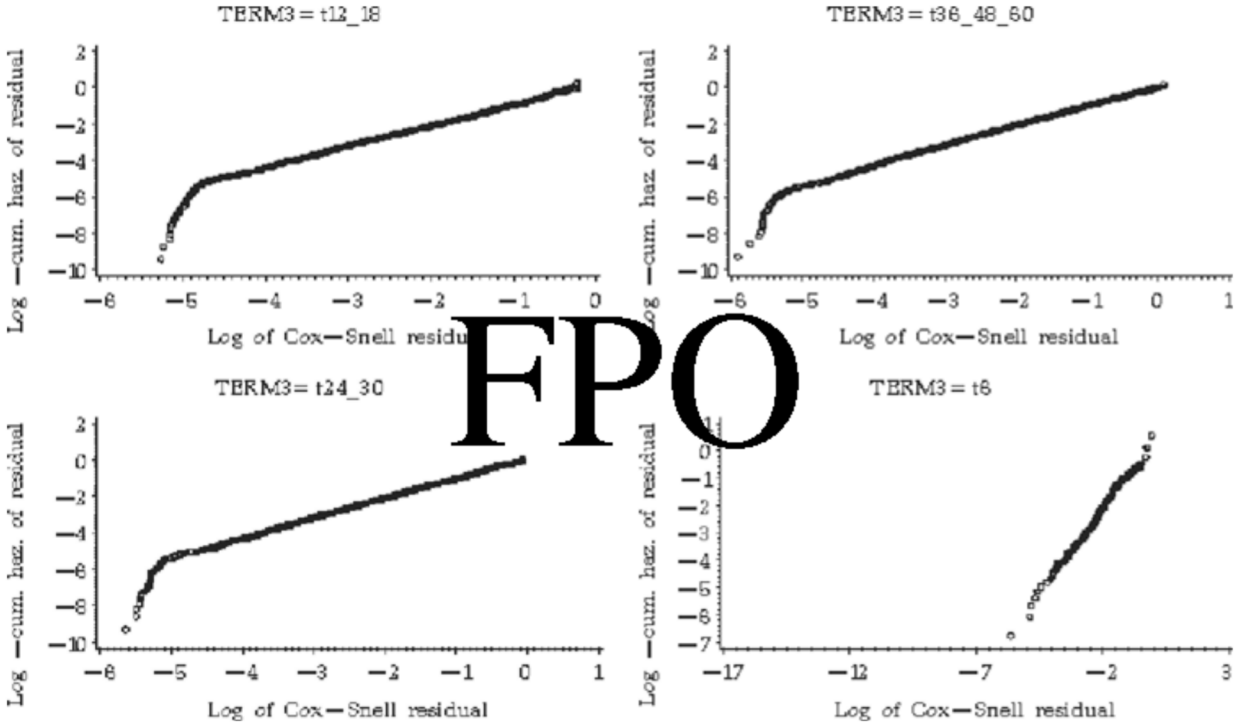


Figure 9. Martingale residuals.

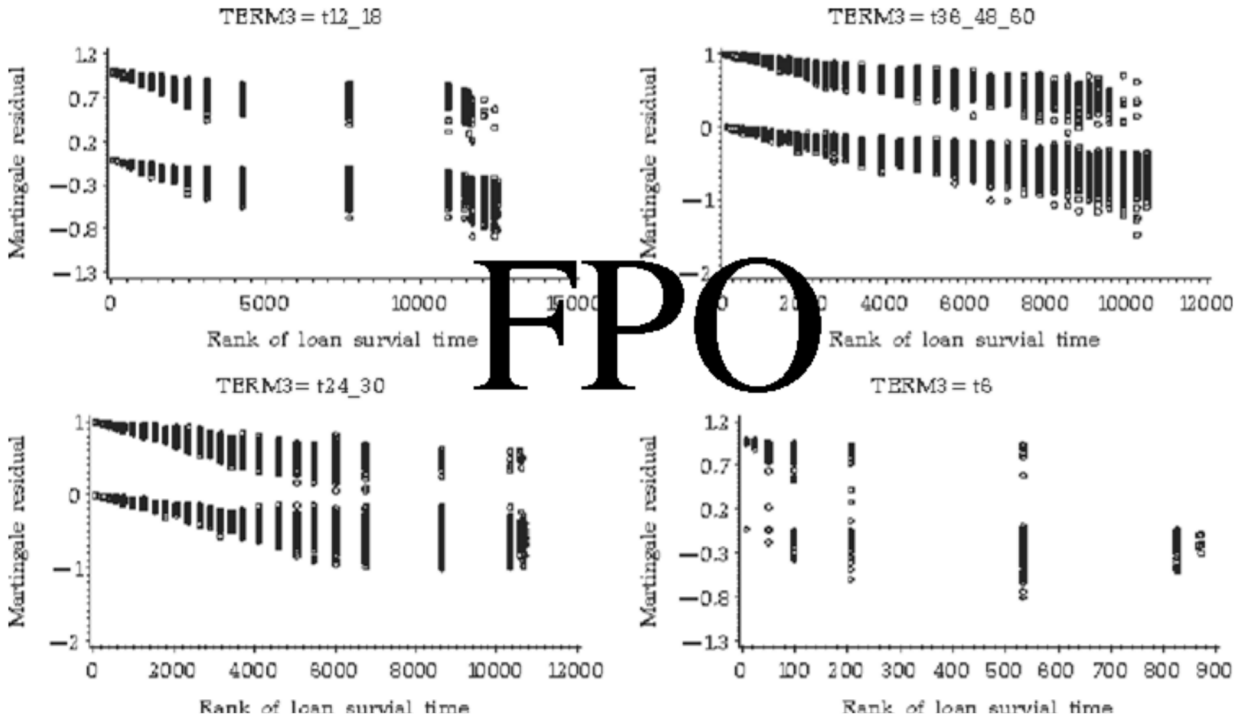
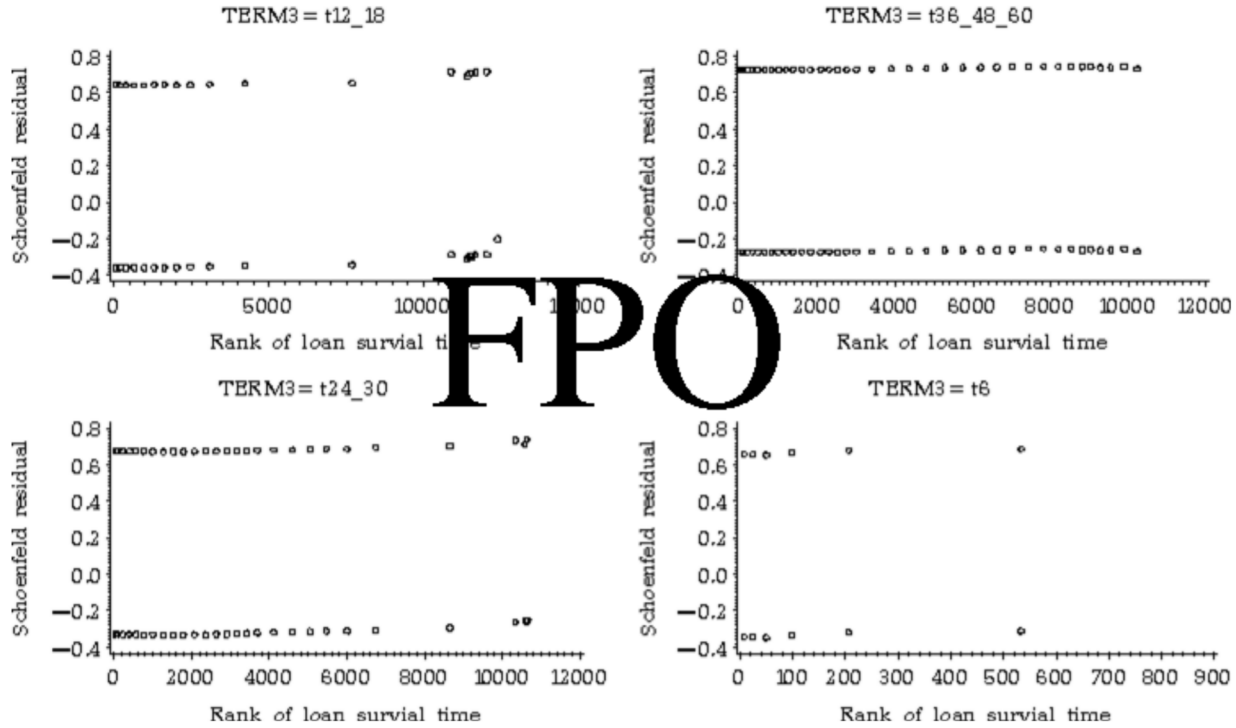


Figure 10. Schoenfeld residuals for FREQPA01.



This paper identifies three developments that improve the present application of Cox's proportional hazards model to building credit-scoring models. Firstly, it develops a new coarse-classifying approach for the characteristics in credit scoring. Secondly, it explains how the residual tools can be used for examining fitness of the model, and discusses pluses and minuses of each of these tools. Finally, the paper expands the use to the time-dependent models to overcome the restriction of the proportional hazards.

Data analysis, specifically match-rate tables and ROC curves, supports the idea that survival-analysis models are competitive with the industry standard logistic regression approach when used for the traditional purpose of classifying applicants into two groups.

Segmenting continuous-characteristic variables and regrouping discrete ones using survival-analysis techniques are more appropriate than the traditional method of using good-bad ratio, if one wishes to avoid choosing an arbitrary time horizon. It is important to do such segmentation separately for all types of failure under consideration because the attributes of the most risky individuals depend on the type of failure.

Diagnostics methods to test the model adequacy were compared. Plots of Cox-Snell, Martingale, deviance, and Schoenfeld residuals were examined for the data under consideration and all suggested that the model fits well. Cox-Snell residuals are the easiest to interpret when analyzing loan data.

Several tests for time-dependency of the effect of a covariate were considered, and Harrel's Z-test was found to be the most appropriate. Time-by-characteristic interactions suggested by Harrel's test were included in the model. This extension allows the effect of a covariate on the predicted time-to-failure to increase or decrease as the loan evolves.

ACKNOWLEDGMENTS

The first author is grateful to the Universities of Edinburgh and Southampton for their financial support during this work. The authors are grateful to the referees for their useful comments on the paper.

REFERENCES

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

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

Cooper, I., M. Martin. 1996. Default risk and derivative products. *Appl. Math. Finance* **3** 53–74.

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

——, E. J. Snell. 1968. A general definition of residuals (with discussion). *J. Royal Statist. Society Series B* **30** 248–275.

Efron, B. 1977. The efficiency of cox's likelihood function for censored data. *J. Amer. Statist. Assoc.* **72** 557–565.

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

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

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

——. 1997. Modelling bonds and derivatives with credit risk. In *Mathematics of Financial Derivatives*. M. Dempster, S. Pliska, (eds.), Cambridge University Press, 369–393.

Narain, B. 1992. Survival analysis and the credit granting decision. In *Credit Scoring and Credit Control*. L. C. Thomas, J. N. Crook, D. B. Edelman (eds.), OUP, 109–121.

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

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

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

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

Thomas, L. C., J. Banasik, J. N. Crook. 1999. Not if but when loans default. *J. Oper. Res. Soc.* **50**.

Annotations from opre119.pdf

Page 1

Annotation 1; Date: 2/27/2002 2:05:40 PM
au: OK as recast?

Page 2

Annotation 1; Date: 2/27/2002 2:06:10 PM
au: OK?

Page 4

Annotation 1; Date: 2/27/2002 2:06:29 PM
au: correct?

Page 7

Annotation 1; Date: 2/27/2002 2:08:26 PM
au: shouldn't it be 7a and 7f for 48 mos?

Page 9

Annotation 1; Date: 2/27/2002 2:09:11 PM
au: please clarify subject and verb here

Page 12

Annotation 1; Date: 2/27/2002 2:09:41 PM
au: pub. location?

Annotation 2; Date: 2/27/2002 2:09:56 PM
au: pub. location?

Page 13

Annotation 1; Date: 2/27/2002 2:10:13 PM
au: pub. location?

Annotation 2; Date: 2/27/2002 2:10:30 PM
au: page nos.