Chapter 7
Probabilities of Default: Continuous-Time Hazard Models

INTRODUCTION

Continuous-time hazard models in a credit risk context describe the survival time of a borrower or loan T_i ($T_i \geq 0$), which is known as time to default, as a random variable in continuous form. The survival time is generally measured from loan origination. Alternatively, you may measure survival time from the date of incorporation of a firm, if this information is available, or the time of the first observation. In our empirical analyses, we use the latter as the mortgage data set observes the loans after origination and to ensure that the PD models are calibrated with regard to the in-sample default rate.

CENSORING

Censoring is an important characteristic of survival analysis data. There are three types of censoring: right-censoring, left-censoring, and interval-censoring.

An observation on a variable T is right-censored if all that you know about T is that it is greater than some value. For example, suppose T is a firm's age at default, and you only know that the firm survived up to the age of 50 but do not know when it will default, as the firm is no longer observed. This situation is called right-censored at age 50.

An observation on a variable T is left-censored if all that you know about T is that it is smaller than some value. To illustrate this, suppose you are doing a study to analyze when people started smoking. Some smokers might not know the exact start time but can give an upper bound. For example, someone mentions that he started smoking when he was younger than 16 years. This is an example of left-censoring.

An observation on a variable T is interval-censored if all that you know about T is that it is bigger than some value and smaller than some value. In other words, the time of the event is situated in a continuous interval. Let's reconsider the smoking example. For example, someone mentions that he started smoking when he was younger than 16 years, but older than 12 years. This is an example of interval censoring.

Basically, the problem of censoring is that data is missing. In other words, we don't know the precise value of our target variable, which is the timing of the event, but we can specify a lower bound, upper bound, or both. Classical regression approaches always assume a precise value for the target. Hence, new techniques may be used to deal with censored data.

Right-censoring is a key concern for credit instruments that have a limited lifetime. Banks stop the collection of loan or borrower performance data at loan maturity or the early prepayment of a loan if borrowers no longer require the credit or refinance with a different lender. As a result, the observed outcome for a given time interval is:

$$\delta_{it} = 1_{\{T_i \le t\}} = \begin{cases} 1, & \text{if } T_i \le t, \text{ i.e., default} \\ 0, & \text{otherwise (i.e., censoring)} \end{cases}$$

Exhibit 7.1 shows the various outcomes.

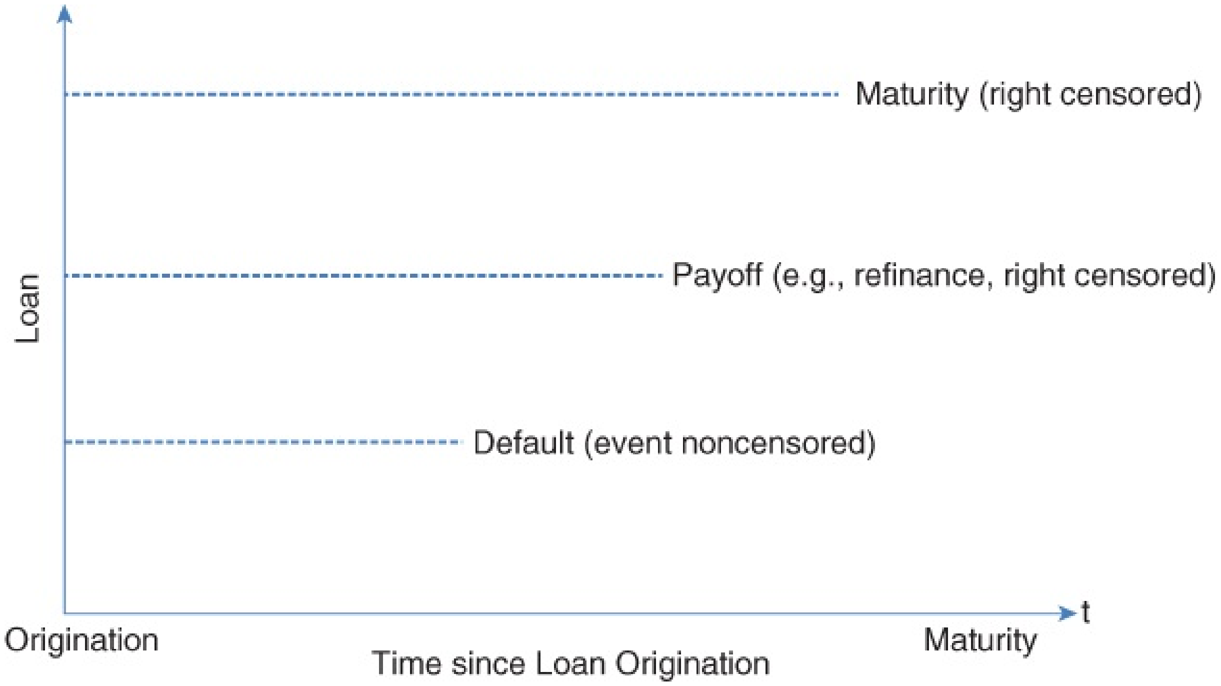

Maturity (right censored)
Payoff (e.g., refinance, right censored)
Default (event noncensored)
Loan
Origination
Time since Loan Origination
Maturity
t

Exhibit 7.1 Observation Credit Outcomes: Default or Censoring

Examples in the literature that have applied survival models include Quigley and Van Order (1991), Belotti and Crook (2009), Malik and Thomas (2010), Tong , Mues, and Thomas (2012), and Dirick, Bellotti, Claeskens, and Baesens (2015).

Generally speaking, following functions may represent the random variable T:

- The probability density function (PDF) $f_i(t)$, with cumulative density function $F_i(t) = \int_{-\infty}^{t} f_i(u)du$
- The survival function $S_i(t) = 1 - F_i(t)$
- The hazard rate $\lambda_i(t) = \lim_{\Delta t \to 0, \Delta t > 0} \quad \frac{1}{\Delta t} P(t \le T_i < t + \Delta t | T_i \ge t) = \frac{f_i(t)}{S_i(t)}$
- The cumulative hazard rate $\int_0^t \lambda_i(u)du = \int_0^t \frac{f_i(u)}{1-F_i(u)} du = -\log\ (1 - F_i(t)) = -\log\ \ S_i(t)$

with borrower $i = 1 \ldots, I$. The functions may be transformed into each other, and SAS offers a number of approaches to estimate these functions, which are discussed latter. We consider three approaches: nonparametric methods (section on life tables), semiparametric methods (section on Cox proportional hazard model), and parametric methods (section on accelerated failure time model).

Survival models control for censoring. We focus on right-censoring, which means that borrowers or loans are no longer observed for the reasons given.

LIFE TABLES

Life table models estimate a survival function, which measures the probability of survival based on the consideration of past default times (i.e., default has occurred within the observation window) and censoring times (i.e., default has not occurred within the observation window). The probability of default follows as one minus the probability of survival. Two approaches are common: Kaplan-Meier analysis (the product limit method) and the actuarial method.

Kaplan-Meier Analysis

Kaplan-Meier (KM) analysis is a first method for survival analysis. The KM estimator is also known as the product limit estimator. It is basically a nonparametric maximum likelihood estimator for the survival probability $S(t)$ as follows:

$$\hat{S}(t) = \hat{S}(t-1)(1 - \frac{d_t}{n_t}) = \hat{S}(t-1)(1 - \lambda(t)) = \prod_{j:t_j \leq t}(1 - \frac{d_j}{n_j})$$

If there is no censoring, then the KM estimator for time t, $\hat{S}(t)$, is simply the proportion of observations in the sample with event times greater than t. If there is censoring, we start by ordering the event times in ascending order $t_1 \leq t_2 \leq ... \leq t_T$. At each time t_j, there are n_j individuals who are at risk of the event. ”At risk” means that they have not undergone the event, nor have they been censored prior to t_j. In other words, they will either undergo the event or become censored after t_j. Assume now that d_j represents the number of individuals who will default at t_j.

The KM estimator is then calculated as follows: $\hat{S}(t)$ equals $\hat{S}(t-1)$ times one minus d_t divided by n_t. This is very intuitive because it basically says that in order to survive time t, you must survive $t-1$ and cannot die during time t. d_t divided by n_t represents the hazard for time t. The expression can now be worked out recursively. This will bring us to the last term in the expression. That is, $\hat{S}(t)$ equals the product across all j whereby t_j is less than or equal to t, of one minus d_j divided by n_j. The latter term is the conditional probability of surviving to time $t_j + 1$, given that the subject has survived to time t_j. For further details, we refer to the SAS documentation for PROC LIFETEST (Details/Computational Formulas).

Let's illustrate the Kaplan-Meier estimate with an example. In Exhibit 7.2 we have a data set of 10 customers: C1 through C10. The second column denotes the time of default or censoring. The third column indicates whether the customer is a defaulter or a censored observation.

Exhibit 7.2 Example for Kaplan-Meier Analysis

ID	Time of Default or Censoring	Default or Censored
C1	6	Default
C2	3	Censored
C3	12	Default
C4	15	Censored
C5	18	Censored
C6	12	Default
C7	3	Default
C8	12	Default
C9	9	Censored
C10	15	Default

We can now complete the table shown in Exhibit 7.3.

Exhibit 7.3 Example for Kaplan-Meier Analysis (cont.)

Time	At Risk at t n_t	Defaulted at t d_t	Censored at t	$S(t)$
0	10	0	0	1
3	10	1	1	0.9
6	8	1	0	0.9 * 7/8 = 0.79
9	7	0	1	0.79 * 7/7 = 0.79
12	6	3	0	0.79 * 3/6 = 0.39
15	3	1	1	0.39 * 2/3 = 0.26
18	1	0	1	0.26 * 1/1 = 0.26

At time 0, all customers are still alive. Hence, the number of customers at risk, n_0, equals 10. The number of customers that defaulted, d_0, equals zero. The number of customers that were censored also equals zero. This results in a survival probability of one. The first default occurs at time 3. We started with 10 customers, so n_3 equals 10. Customer C7 defaulted so d_3 equals one. Customer C2 was censored. This results in a survival probability of 0.9. The next default happens at time 6. We still have eight customers, so n_6 equals eight. Customer C1 defaulted. No one was censored. In order to survive time 6, a customer must first survive time 3 and cannot default during time 6. Hence, this results in a survival probability of $0.9 * 7/8$ or 0.79. At time 9, we still have seven customers at risk, so n_9 equals seven. No one defaults, but customer C9 was censored. Hence, the survival probability becomes $0.79 * 7/7$, or thus 0.79. In other words, because censoring occurred only during time 9, the survival probability remains unaffected. At time 12, we still have six customers at risk, three of which will default. The survival probability thus becomes $0.79 * 3/6$ or 0.39. At time 15, we start with three customers. Customer C10 defaulted and customer C4 was censored. Hence, the survival probability becomes $0.39 * 2/3$ or 0.26. Finally, at time 18, no one defaults, so that the survival probability remains 0.26.

Actuarial Method

If there are many unique event times, we recommend using a life table or actuarial method to group the event times into intervals. The survival probability, $\hat{S}(t)$, then equals:

$$\hat{S}(t) = \prod_{j:t_j \leq t} (1 - \frac{d_j}{n_j - c_j/2})$$

Basically, this formula assumes that censoring occurs uniformly across a time interval. Because we started with n_j at the beginning of the time interval and ended with n_j minus c_j, the average number at risk equals $(n_j + (n_j - c_j))/2$ or $n_j - c_j/2$, which corresponds to the denominator in the expression. For further details, we refer to the SAS documentation for PROC LIFETEST (Details/Computational Formulas).

Reshaping the Data

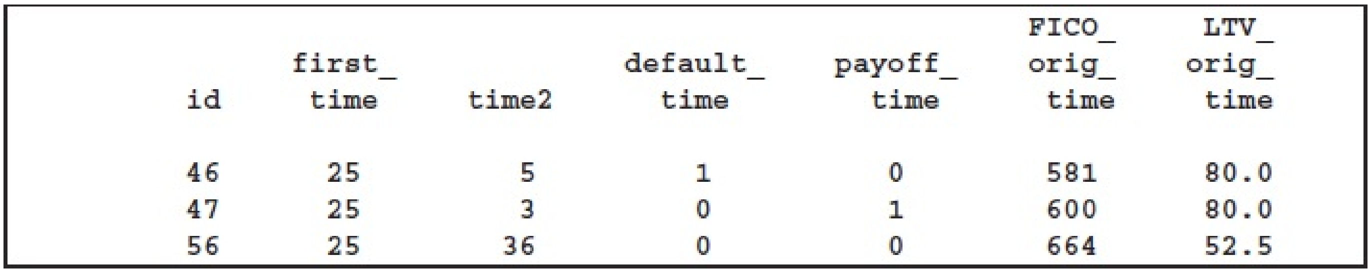
Life tables do not generally condition on observable information (i.e., are nonparametric) and require a cross-sectional form (i.e., one observation per loan). Our mortgage data have to be reshaped by keeping the last observation and computing the time in months since the first observation time. Hence, the default indicator default-time is zero if an observation is censored and one if a default occurs.

We avoid left-censoring by assuming that all loans start from the first observation period onward and generate the new time stamp time2. Note that banks may also consider a time stamp from loan origination if loans are observed since origination.

```
PROC SORT DATA=data.mortgage;
BY id;
RUN;
DATA lifetest_temp1;
SET data.mortgage;
time2 = time-first_time+1;
BY id;
RETAIN id;
IF LAST.id THEN indicator=1;
RUN;
DATA lifetest_temp2;
SET lifetest_temp1;
IF indicator = 1 OR default_time =1;
RUN;
DATA lifetest;
SET lifetest_temp2;
BY id;
RETAIN id;
IF FIRST.id THEN output;
```

```
RUN;
```

Exhibit 7.4 shows three borrower observations as an example. The time stamp time2 shows the time to default since the first observation period. This is contrary to the panel data set applied for discrete-time hazard models, where the time stamp time indicated the absolute time and loans were generally originated at different times. The first loan has defaulted in the fifth period after first observation, the second loan has been paid off in the third period after first observation, and the third loan is observed until the last observation period. The second and third loans are considered to be right-censored in the analyses shown in Exhibit 7.2.



					FICO_	LTV_
	first_		default_	payoff_	orig_	orig_
id	time	time2	time	time	time	time
46	25	5	1	0	581	80.0
47	25	3	0	1	600	80.0
56	25	36	0	0	664	52.5

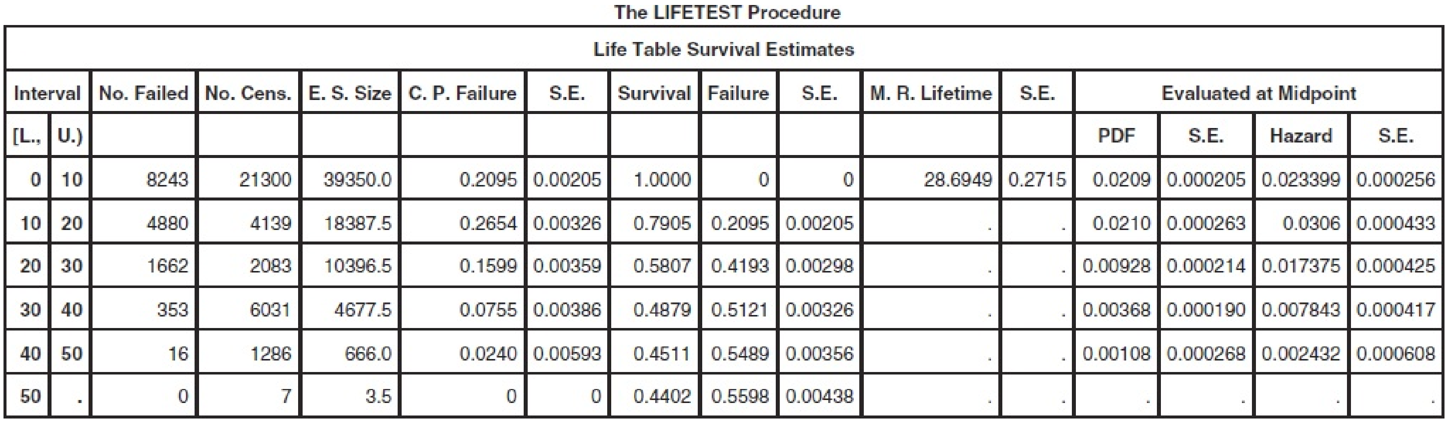
Exhibit 7.4 Cross-Sectional Data

Model Estimation Using PROC LIFETEST

In a first step, the survival function can be estimated with PROC LIFETEST using the product-limit method (also known as the Kaplan-Meier method, option method=PL in the PROC LIFETEST statement) or the life table method (also called the actuarial method, option method=LT in the PROC LIFETEST statement). We apply the life table method in the following:

```
ODS GRAPHICS ON;
PROC LIFETEST DATA=lifetest METHOD=LT INTERVALS=(0 TO 50 BY 10)
PLOTS=(ALL);
TIME time2*default_time(0);
RUN;
ODS GRAPHICS OFF;
```

The output presented in Exhibit 7.5 shows descriptive statistics for discrete time intervals including the numbers for failed observations, censored observations, and effective sample size (i.e., number of observations at the beginning of the time interval less 50 percent of the number of censored observations). We assume that censoring occurs uniformly across the time interval. Furthermore, the values for a number of functions are estimated (for METHOD = LT, detailed mathematical formulas are available in the SAS documentation for PROC LIFETEST):



The LIFETEST Procedure

Life Table Survival Estimates

Interval		No. Failed	No. Cens.	E. S. Size	C. P. Failure	S.E.	Survival	Failure	S.E.	M. R. Lifetime	S.E.	Evaluated at Midpoint			
[L.,	U.)											PDF	S.E.	Hazard	S.E.
0	10	8243	21300	39350.0	0.2095	0.00205	1.0000	0	0	28.6949	0.2715	0.0209	0.000205	0.023399	0.000256
10	20	4880	4139	18387.5	0.2654	0.00326	0.7905	0.2095	0.00205	.	.	0.0210	0.000263	0.0306	0.000433
20	30	1662	2083	10396.5	0.1599	0.00359	0.5807	0.4193	0.00298	.	.	0.00928	0.000214	0.017375	0.000425
30	40	353	6031	4677.5	0.0755	0.00386	0.4879	0.5121	0.00326	.	.	0.00368	0.000190	0.007843	0.000417
40	50	16	1286	666.0	0.0240	0.00593	0.4511	0.5489	0.00356	.	.	0.00108	0.000268	0.002432	0.000608
50	.	0	7	3.5	0	0	0.4402	0.5598	0.00438	.	.	.	.	.	.

Exhibit 7.5 Life Table Model

- The default probability, which is labeled Conditional Probability of Failure. The estimator is the number of default events over the effective sample size for a given time interval. The standard deviation is computed for the default rate assuming a Bernoulli distribution of default events given the estimated default rate.
- The survival rate, which is labeled Survival. The estimator is one for the first time period and one minus the estimated default probability of the interval times the estimated survival rate in the previous time interval. The cumulative failure rate (Failure) is one minus the survival rate.
- The probability density function, which is labeled PDF. The estimator is the estimated survival rate times the estimated default probability relative to the length of the time interval.
- The hazard rate, which is labeled Hazard. The estimator is the estimated default rate over the midpoint of the survival rate relative to the length of the time interval.

The PLOTS command generates a number of plots, including the three estimated functions: PDF, survival function, and hazard rate (see Exhibits 7.6, 7.7, and 7.8).

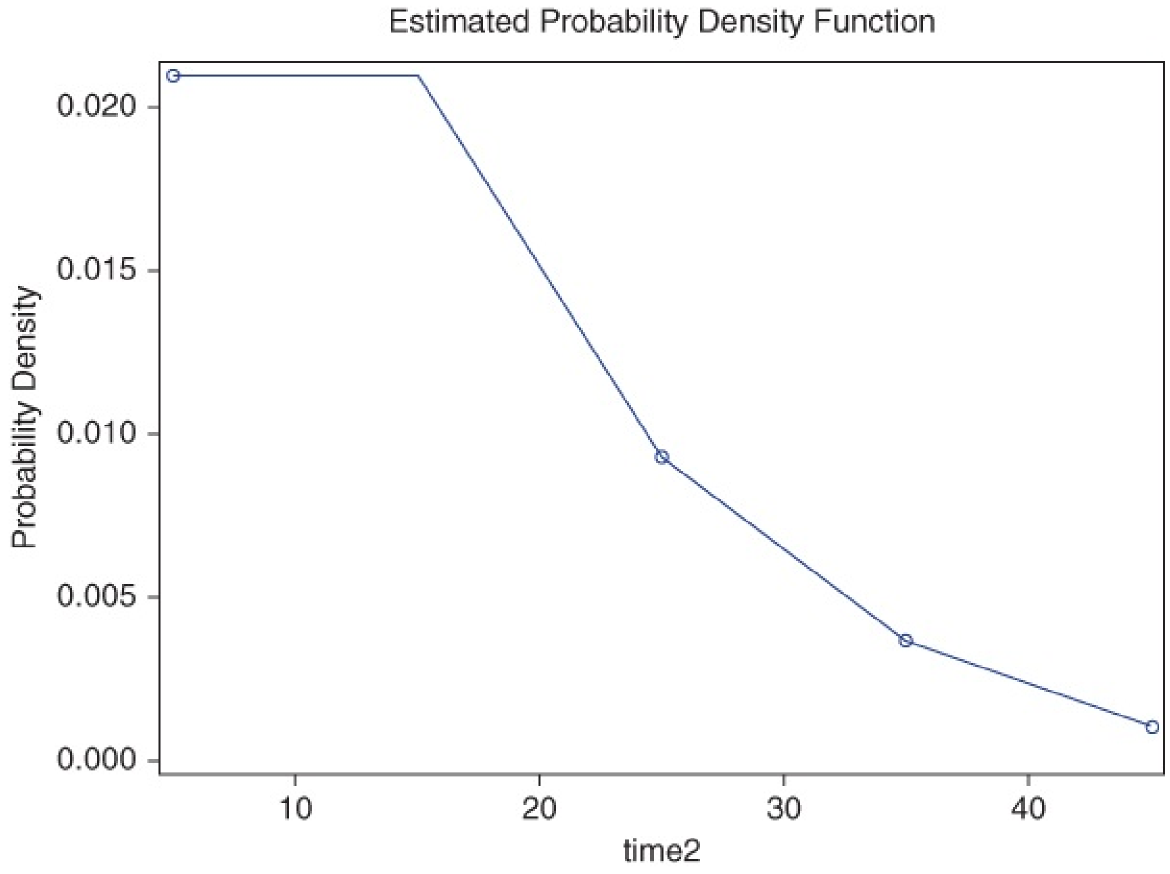


Exhibit 7.6 PDF plot

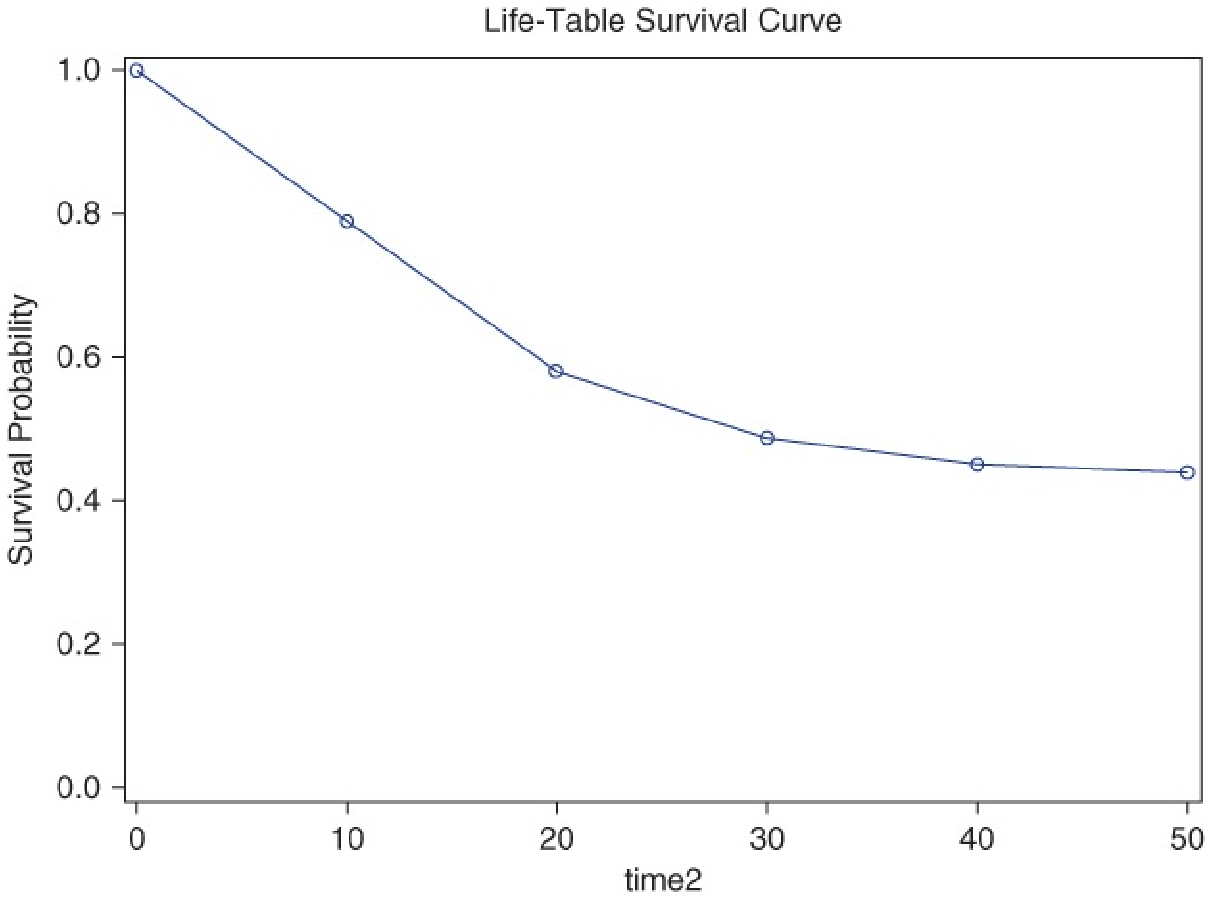


Exhibit 7.7 Survival Function Plot

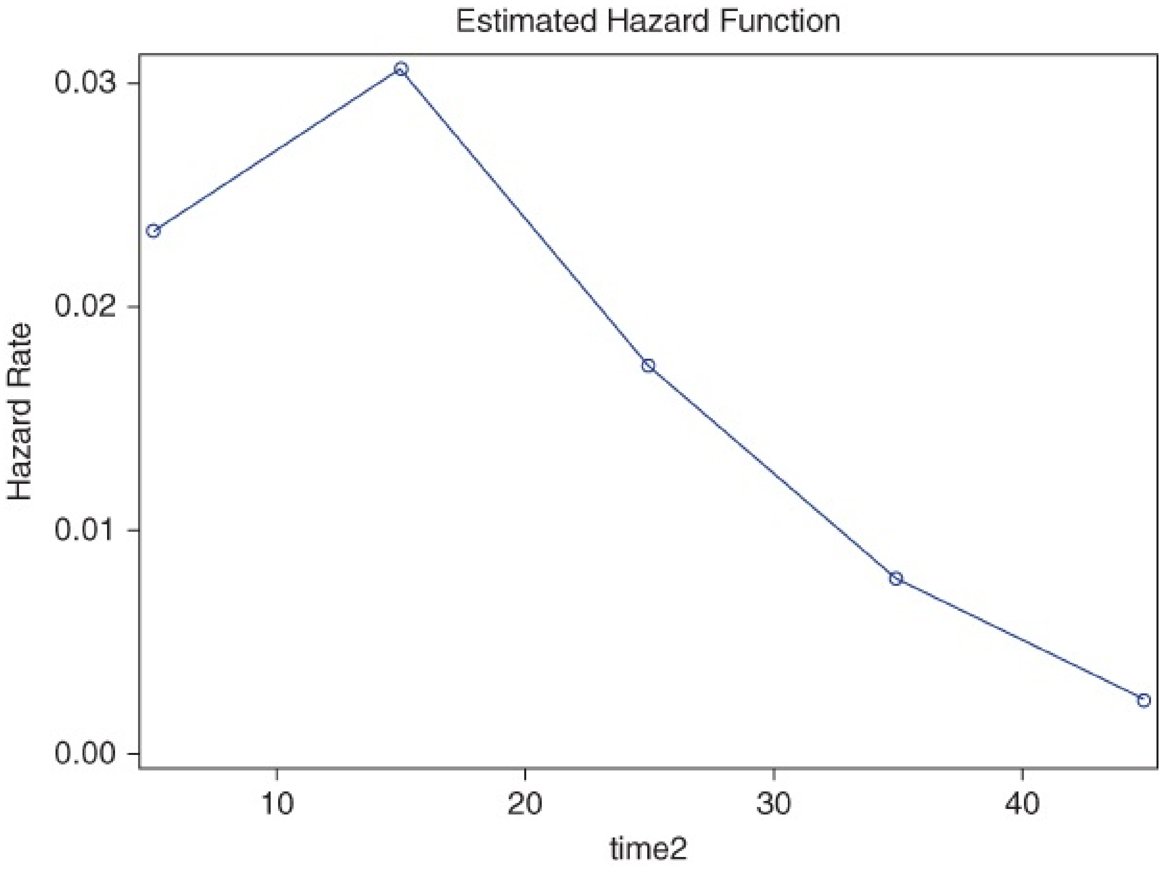


Exhibit 7.8 Hazard Rate Plot

Controlling for Information in Nonparametric Models

Life tables are nonparametric and as such do not condition on control information such as borrower, collateral, loan, or economic variables. Despite these limitations, there is a simple way to include (even time-varying) observable information by means of stratification. To demonstrate this point, we now stratify the results into five groups of equal size (i.e., 20 percent of all observations) using PROC RANK and the FICO score as discriminatory variable.

```
PROC RANK DATA=lifetest OUT=lifetest2 GROUPS=5;
VAR FICO_orig_time;
RANKS FICO_orig_time_rank;
RUN;
ODS GRAPHICS ON;
PROC LIFETEST DATA=lifetest2 METHOD=LT INTERVALS=(0 TO 50 BY 10)
PLOTS=(SURVIVAL);
TIME time2*default_time(0);
STRATA FICO_orig_time_rank;
RUN;
ODS GRAPHICS OFF;
```

We do not show the output for this model but present the survival plot, which is produced with the ODS GRAPHICS ON/ODS GRAPHICS OFF statements before and after the actual PROC LIFETEST code (see Exhibits 7.9).

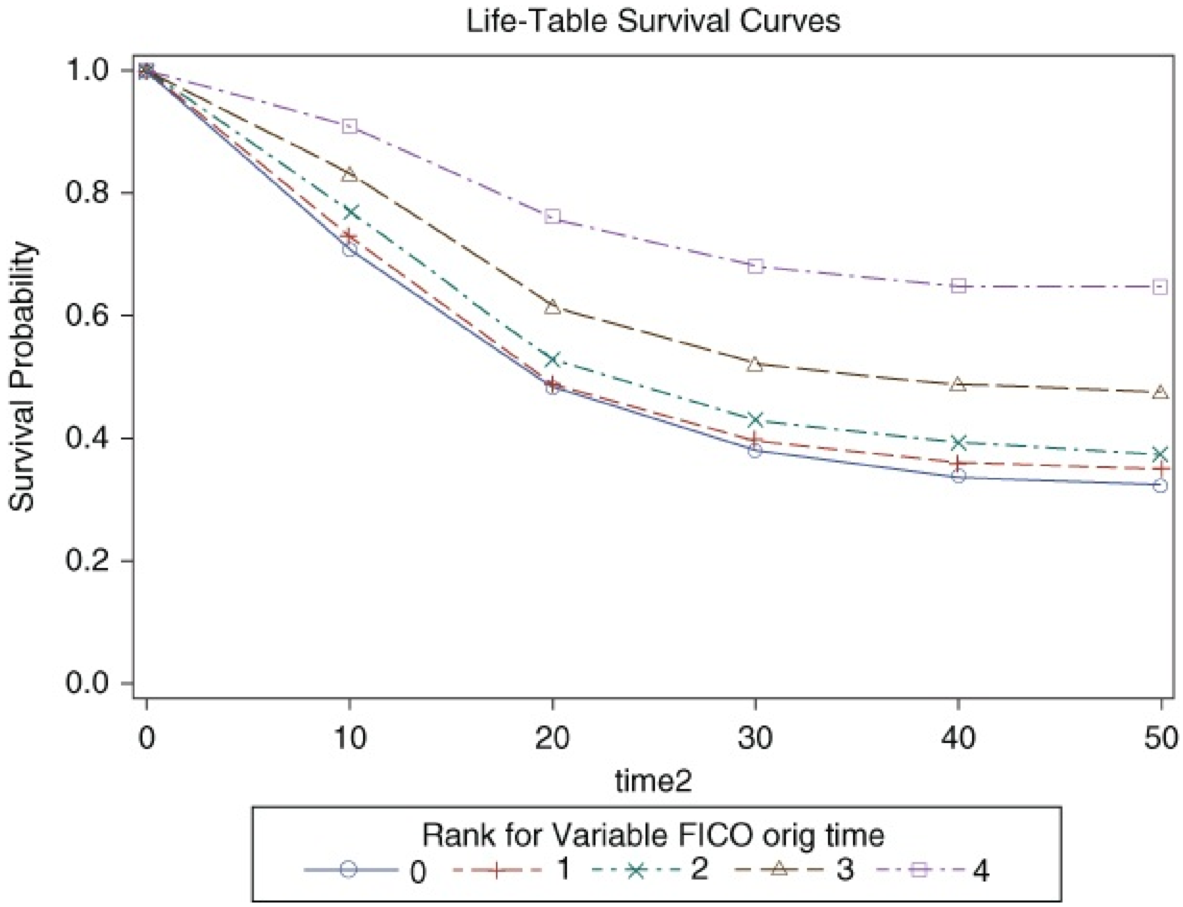


Exhibit 7.9 Survival Plot

It is clear that the FICO score and survival probability are positively correlated. A low FICO score has a greater default risk, which results in a low survival probability; a high score has less risk, and high survival probability.

Test of Equality over Groups

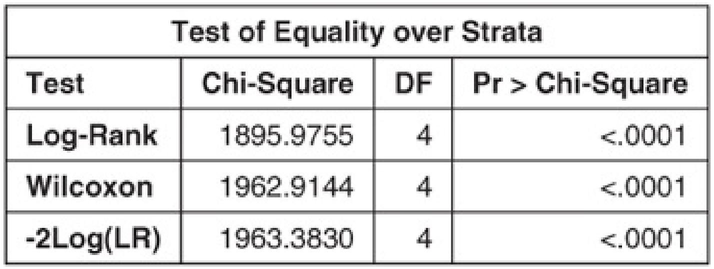
As mentioned, the Kaplan-Meier estimator does not account for the presence of covariates. It is a very useful tool for exploring and describing your survival data. However, other techniques are needed for building predictive survival analysis models, which is discussed in what follows.

A first extension of Kaplan-Meier analysis is to statistically test the equivalence of survival curves of different samples.

- H_0: The survival curves are statistically the same.
- H_1: The survival curves are statistically different.

As an example, suppose you want to test whether the survival curves for males and females are the same in a default setting. The null hypothesis then reads as follows: The survival curves for males and females are statistically the same. The alternative hypothesis is the following: the survival curves for males and females are statistically different. Various test statistics can be used to evaluate this, such as the log-rank test (sometimes also referred to as the Mantel-

Haenszel test), the Wilcoxon test, and the likelihood ratio statistic. These three tests are closely related and usually give the same results. (See Exhibit 7.10.) They enable you to verify some basic insights about the survival data, which might further be elaborated on in the subsequent analysis. The STRATA statement reports these tests. This is useful in exploring data to see whether survival probabilities are significantly different across different segments.



Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	1895.9755	4	<.0001
Wilcoxon	1962.9144	4	<.0001
-2Log(LR)	1963.3830	4	<.0001

Exhibit 7.10 PROC LIFETEST: Test of Equality over Groups

Estimation of Survival Probabilities

With PROC LIFETEST we can export the survival probabilities using the OUTSURV command.

```
PROC LIFETEST DATA=lifetest METHOD=LT INTERVALS=(1 TO 102 BY 1)
   OUTSURV=SURVIVAL;
TIME time2*default_time(0);
RUN;
```

We convert the time stamp so that the measurement time is at the end of the period as OUTSURV computes the survival likelihood at the beginning of the period:

```
DATA survival;
SET survival;
time2=time2-1;
RUN;
```

Estimation of Default Probabilities

We have shown earlier that the conditional default probabilities between time t_1 and t_2 (PD_{t_1,t_2}), which are the quantities of interest for capital regulation, may be computed from the survival functions as the difference $S(t_1) - S(t_2)$ over $S(t_1)$:

$$PD_{t_1,t_2} = \frac{S(t_1) - S(t_2)}{S(t_1)}$$

We will apply variations of this formula later to derive discrete time default probabilities from the survival probability functions of continuous-time hazard models.

We can now compute default probabilities as follows: ($S(t1) - S(t2)/S(t1)$):

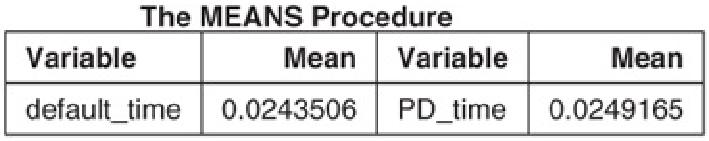
```
DATA survival2(WHERE=(PD_time NE .));
SET survival;
IF time2>=1 THEN PD_time=(lag(survival)-survival)/lag(survival);
IF time2 =1 THEN PD_time=1-survival;
KEEP time2 PD_time;
RUN;
```

Finally, we add the default probabilities to the panel data set via match merging:

```
PROC SORT DATA=lifetest_temp1;
BY time2;
RUN;
PROC SORT DATA=survival2 NODUPKEY;
BY time2;
RUN;
DATA probabilities;
MERGE lifetest_temp1(IN=a) survival2;
BY time2;
IF a;
RUN;
```

Calibration of Life Tables

A PROC MEANS for the default indicator and the default probabilities shows that the mean of these in-sample PD estimates approximately matches the default rate. (see Exhibit 7.11).



The MEANS Procedure

Variable	Mean	Variable	Mean
default_time	0.0243506	PD_time	0.0249165

Exhibit 7.11 Calibration of Life Tables: Comparison of Default Indicators and Estimated Default Probabilities

```
PROC MEANS DATA=probabilities MEAN NOLABELS;
VAR default_time PD_time;
RUN;
```

COX PROPORTIONAL HAZARDS MODELS

Cox proportional hazards (CPH) models following Cox (1972) are regression models that link the survival time under consideration of censoring with predictive covariates. Hence, they are more flexible than the life table approach and stratification technique in controlling for observable information.

CPH models link the hazard rate with a baseline hazard rate and a transformation of the linear predictor (i.e., a linear combination of parameters with explanatory variables that exclude an intercept) as follows:

$$\lambda(t|\boldsymbol{x}_i) = \lambda_0(t)\exp\ (\boldsymbol{\beta}'\boldsymbol{x}_i)$$

where the baseline hazard rate $\lambda_0(t)$ is not specified ($\lambda_0(t) \geq 0$). CPH models are hence called semiparametric. We restrict the covariates to idiosyncratic information for now. From this equation, it becomes obvious that the covariates in a CPH model have a proportional impact on the hazard rate regardless of the base level of the hazard rate. In other words, if we take the ratio of the hazards of two individuals (i and j), we can see that the baseline hazard divides away, creating an expression that is independent of time as you can see in Exhibit 7.12.

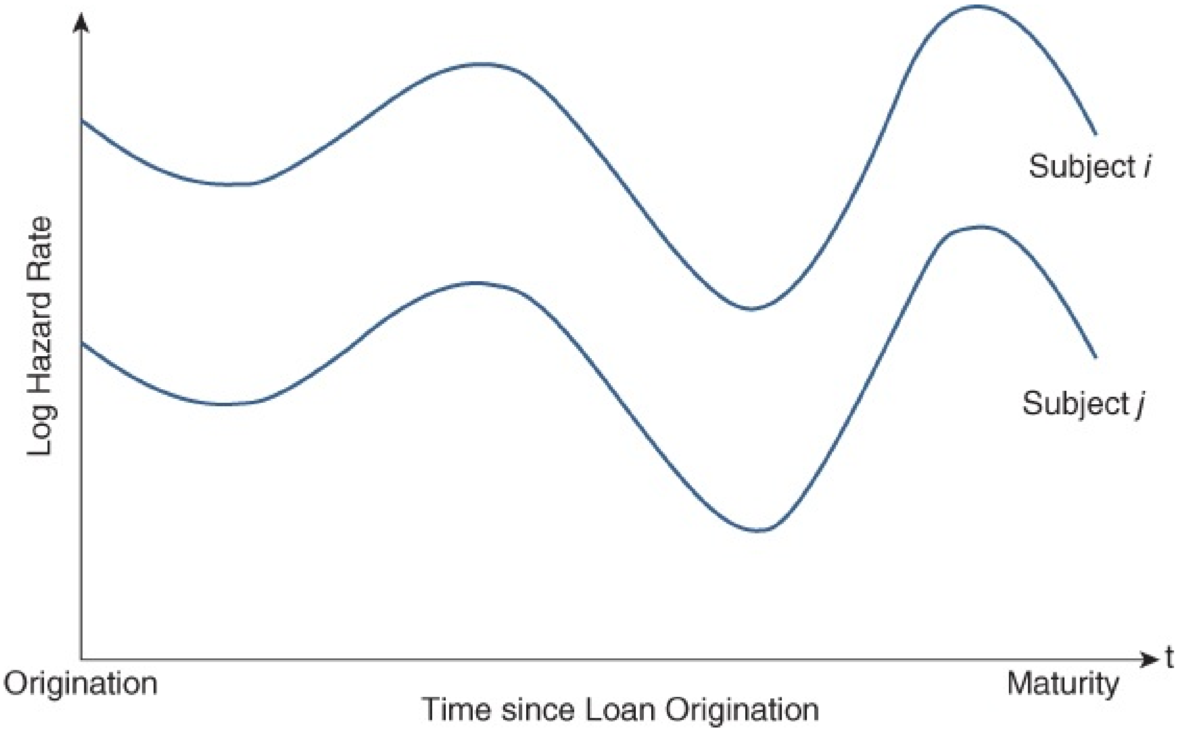


Exhibit 7.12 Proportional Hazards

This implies that the hazard of any individual is a fixed proportion of the hazard of any other individual; hence, the name proportional hazards. Put differently, the subjects most at risk at any one time remain the subjects most at risk at any other time, as you can see depicted in Exhibit 7.12.

The proportional hazards assumption can be tested by inspecting the survival functions for subgroups from the life table analysis. A good indication that the assumption holds is verifying that the survival functions do not cross.

The survival function of a CPH model is:

$$\begin{aligned} S(t|\boldsymbol{x}_i) &= \exp\left(-\int_0^t \lambda(u|\boldsymbol{x}_i)du\right) \\ &= \exp\left(-\int_0^t \lambda_0(u)\exp\,(\boldsymbol{\beta}'\boldsymbol{x}_i)du\right) \\ &= S_0(t)^{\exp\,(\boldsymbol{\beta}'\boldsymbol{x}_i)} \end{aligned}$$

with $S_0(t) = \exp\left(-\int_0^t \lambda_0(u)du\right)$

The probability density function of the survival time T is given by:

$$f(t|\boldsymbol{x}_i) = \lambda(t|\boldsymbol{x}_i)\ S(t|\boldsymbol{x}) = \lambda_0(t)\exp\,(\boldsymbol{\beta}'\boldsymbol{x}_i)\ S_0(t)^{\exp\,(\boldsymbol{\beta}'\mathbf{x}_1)}$$

If we now introduce an indicator variable for borrower i, δ_i, whereby $\delta_i = 0$ if censoring occurs and $\delta_n = 1$ if default occurs, then the likelihood for all observations becomes:

$$\begin{aligned} L(\lambda_0(t), \boldsymbol{\beta}, \boldsymbol{x_i}) &= \prod_{i=1}^{N} f_i(t_i)^{\delta_i} \; S_i(t_i)^{1-\delta_i} \\ &= \prod_{=1}^{N} [\lambda_0(t_i) \exp\ (\boldsymbol{\beta}' \boldsymbol{x}_{ii})]^{\delta_i} \exp\ \left[- \int_0^{t_i} \lambda_0(u) \exp\ (\boldsymbol{\beta}' \boldsymbol{x}_{ii}) du \right] \end{aligned}$$

Partial Likelihood

Let's now discuss how the β parameters of a proportional hazards regression model can be estimated using the idea of partial likelihood (Cox, 1972, 1975). Suppose we have I individuals with i ranging from one to I. Each individual has three characteristics: x_i is the vector of covariates, t_i is the time of the event or censoring, and δ_i is one if the individual is uncensored and zero if the individual is censored. We start by ranking all the events of the noncensored subjects (t_1 up to t_T). Given the fact that one subject has event time t_i, the probability that this subject has inputs x_j is then given by:

$$\frac{\lambda(t_i \boldsymbol{x}_i)\Delta t}{\sum_{l \in R(t_i)} h(t_i, \boldsymbol{x}_l)\Delta t}$$

where $R(t_i)$ represents the subjects that are at risk at time t_i. Since the baseline hazard $\lambda_0(t_i)$ occurs in both the numerator and the denominator, it will cancel out. Hence, this gives us the following expression:

$$\frac{\exp\ (\boldsymbol{\beta}' \boldsymbol{x}_i)}{\sum_{l \in R(t_i)} \exp\ (\boldsymbol{\beta}' \boldsymbol{x}_l)}$$

which is independent of the baseline hazard.

The partial likelihood function then becomes:

$$\prod_{j=1}^{I} \frac{\exp\ (\boldsymbol{\beta}' \boldsymbol{x}_i)}{\sum_{l \in R(t_i)} \exp\ (\boldsymbol{\beta}' \boldsymbol{x}_l)}$$

Note that, for ease of notation, we assumed that individual j with covariates x_j has event time t_j. The β parameters can be optimized using the Newton-Raphson algorithm. It is important to observe how the censored observations enter the partial likelihood function. They will be included in the risk sets $R(t_j)$ until their censoring time.

Also, it is important to note that the β parameters can be estimated without having to specify the baseline hazard $\lambda_0(t)$. Furthermore, it can be shown that the partial likelihood estimates are consistent, which means that they converge to the true values as the sample increases, and are asymptotically normal. Moreover, the partial likelihood estimates depend only on the ranks of the event times and not on the numerical values. An important assumption made in deriving the partial likelihood function is that there are no tied event times. However, in many real-life settings, time is measured in a discrete way so that ties are likely to occur.

There are four common ways to deal with tied event times: the exact method, two approximations, and the discrete method. In case there are no ties, all four methods give the same estimates. The exact method assumes that ties occur because of imprecise time measurements and treats time as continuous. Hence, it considers all possible orderings of the event times and constructs a likelihood term for each ordering. When there are three ties, six orderings are possible, and thus six terms are added to the partial likelihood function. Obviously, this procedure is very time consuming for heavily tied data. It is recommended only when few ties occur. Two popular approximations are the Breslow and Efron likelihood methods. The Breslow likelihood method works well if ties occur rarely. Empirical evidence has shown that the Efron approximation is often superior to the Breslow approximation. A fourth option is the discrete method whereby time is treated in a discrete way so that multiple events can occur at the same time. Although this can also be addressed using the partial likelihood approach, it will become very time consuming for large, heavily tied survival datasets.

Cox Proportional Hazards Model in SAS Using PROC PHREG

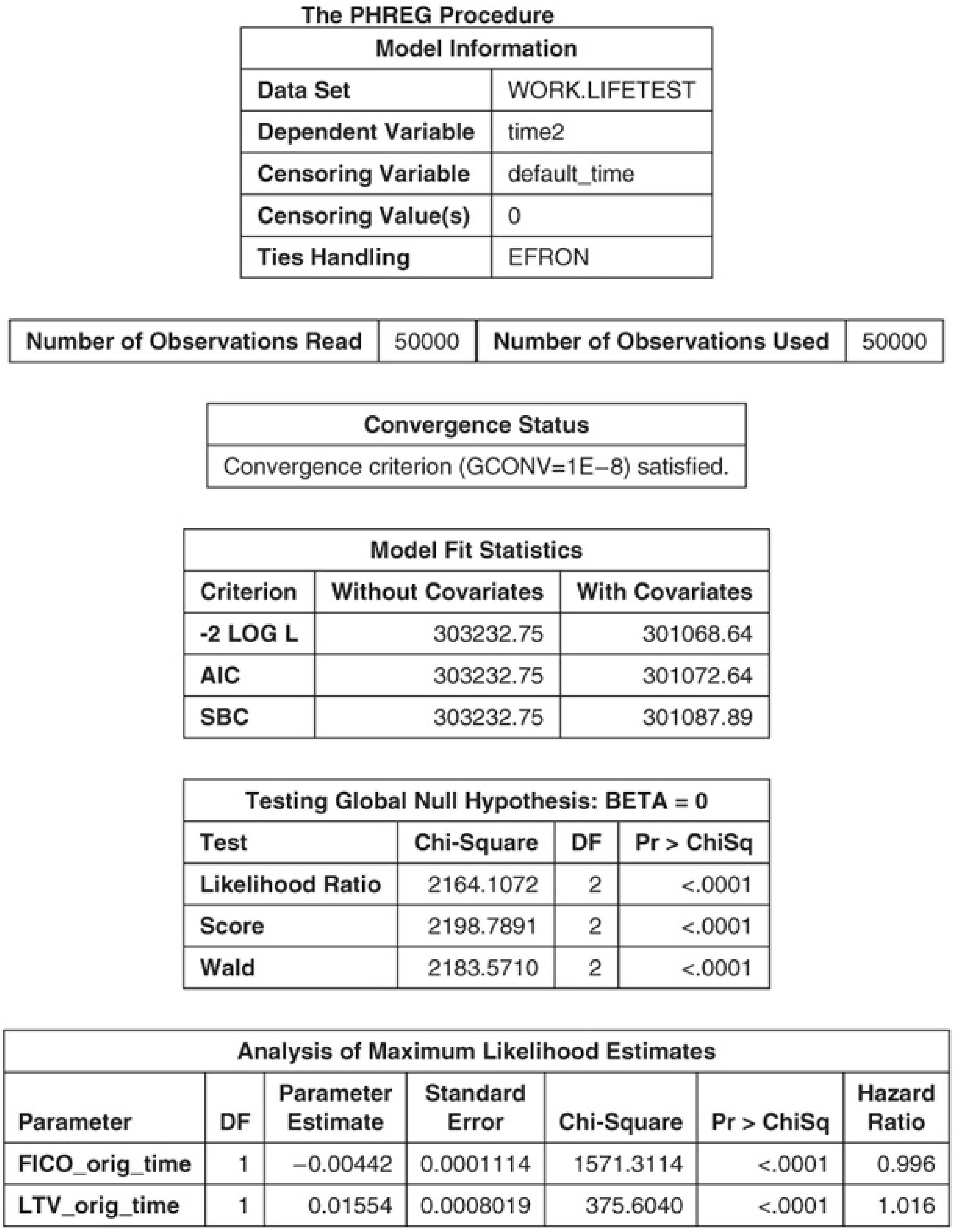
CPH models can be estimated in SAS using life table data (see previous discussion) if estimated for information that is not time-varying (e.g., credit information at origination).

We first create a data set called covariates with the FICO score and the LTV ratio to plot the survival function for each. We then run PROC PHREG, including a command to generate the survival functions:

```
ODS GRAPHICS ON;
PROC PHREG data=lifetest PLOTS(OVERLAY)=SURVIVAL;
MODEL time2*default_time(0)=FICO_orig_time LTV_orig_time / TIES=EFRON;
BASELINE COVARIATES=covariates_orig_time / ROWID=set;
RUN;
ODS GRAPHICS OFF;
```

The structure of PROC PHREG is very similar to PROC LOGISTIC. The dependent variable in the model is the (survival) observation time and an indicator variable, which indicates default (coded by one), or whether the borrower is no longer observed (i.e., censored, coded by zero). The censoring state is specified in brackets and the two variables connected by "*".

The model output is presented in Exhibit 7.13.



The PHREG Procedure

Model Information	
Data Set	WORK.LIFETEST
Dependent Variable	time2
Censoring Variable	default_time
Censoring Value(s)	0
Ties Handling	EFRON

Number of Observations Read	50000	Number of Observations Used	50000

Convergence Status
Convergence criterion (GCONV=1E−8) satisfied.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	303232.75	301068.64
AIC	303232.75	301072.64
SBC	303232.75	301087.89

Testing Global Null Hypothesis: BETA = 0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	2164.1072	2	<.0001
Score	2198.7891	2	<.0001
Wald	2183.5710	2	<.0001

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
FICO_orig_time	1	−0.00442	0.0001114	1571.3114	<.0001	0.996
LTV_orig_time	1	0.01554	0.0008019	375.6040	<.0001	1.016

Exhibit 7.13 CPH Model

The output is similar to PROC LOGISTIC with the distinction that rank-correlation measures between default probability and default events are no longer provided and a hazard ratio is included. The hazard ratio minus 100 percent shows the increase in default risk for a unit change in the explanatory variable.

Remember, we included only information that is, not time-varying, that is the FICO score and LTV ratio at origination. As in the PROC LOGISTIC results, FICO has a negative impact on the hazard rate and LTV a positive impact.

Similar to the PROC LIFETEST statement, various plots (here the survival function) can be generated with the (1) ODS GRAPHICS ON/ODS GRAPHICS OFF statements before and after the actual PROC PHREG code, (2) the PLOTS (OVERLAY) command whereby (OVERLAY) indicates that all curves are displayed in a single chart, and (3) the BASELINE command. The COVARIATES statement indicates reference values for the plot of the survival function given. The ROWID command indicates the names of the various reference value sets. We have created the following data set in SAS:

```
DATA covariates_orig_time;
INPUT set $ FICO_orig_time LTV_orig_time;
DATALINES;
high 600 90
low 800 60
;
```

If a data set is not explicitly specified, then a single survival function is plotted based on the average default values for the metric variables and the reference categories for the categorical variables as specified by the class statement. The survival functions for the two observations included in the data set covariates are depicted in Exhibit 7.14.

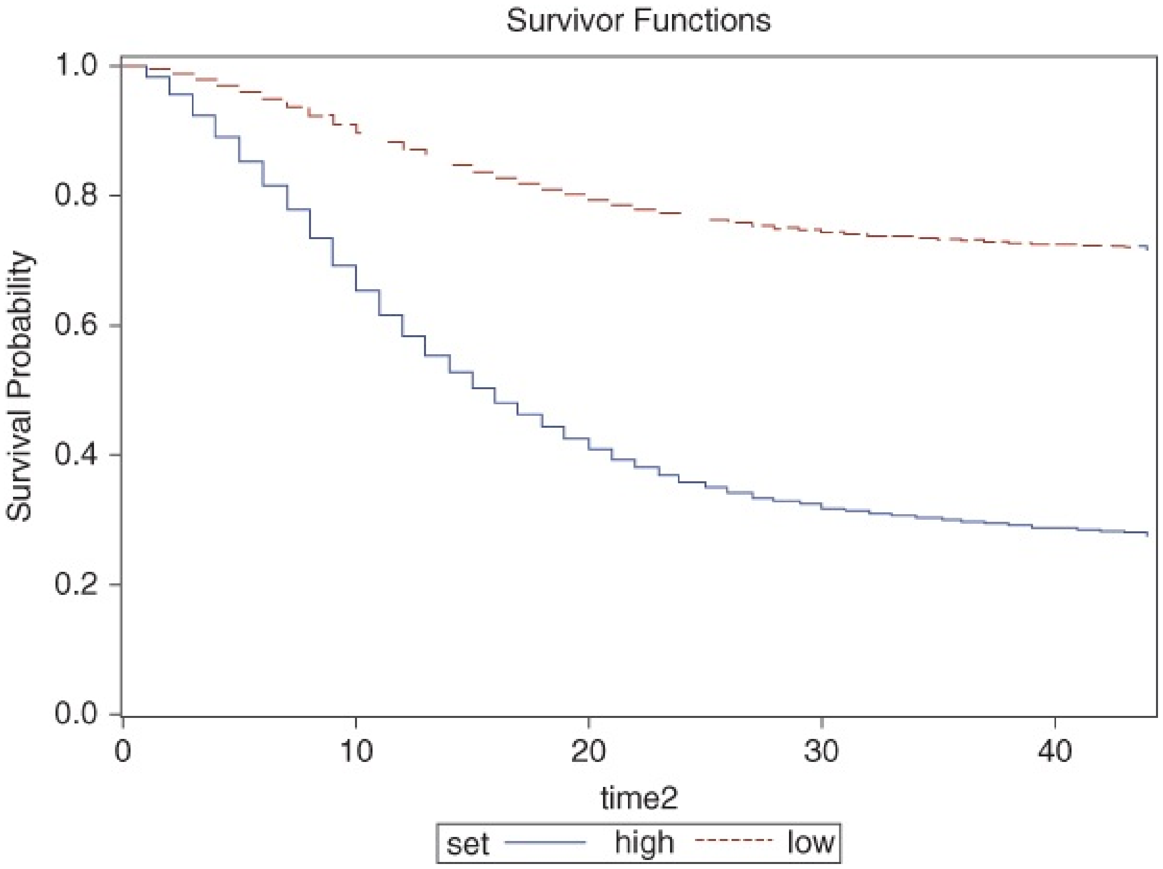


Exhibit 7.14 Survival Plot

As an extension, the baseline hazard function can be parameterized and the parameters estimated. Popular examples are the exponential, Weibull, and Gompertz distributions.

Furthermore, PROC SURVEYPHREG estimates clustered standard errors by assuming dependence between observations of the same cluster unit. Clustering variables can be defined by the CLUSTER command.

Time-Varying Information

Time-varying information is a key concern in credit risk modeling. CPH models are able to accommodate time-varying explanatory variables as discrete-time hazard models:

$$\lambda(t|x_{it-1}) = \lambda_0(t)\exp\,(\boldsymbol{\beta}'\boldsymbol{x}_{it-1})$$

There are two main ways to include time-varying covariates: first, the aggregation of time-varying information and second, counting process data. We discuss both in what follows. SAS also offers an option for inclusion of time-varying covariates by programming statements. However, this assumes that the time variation occurs along the life cycle of a loan (time stamp time2) and not along the line of the economy (time stamp time). As a result, we do not discuss this technique in more detail.

Time-Varying Covariates: Aggregation of Time-Varying Information

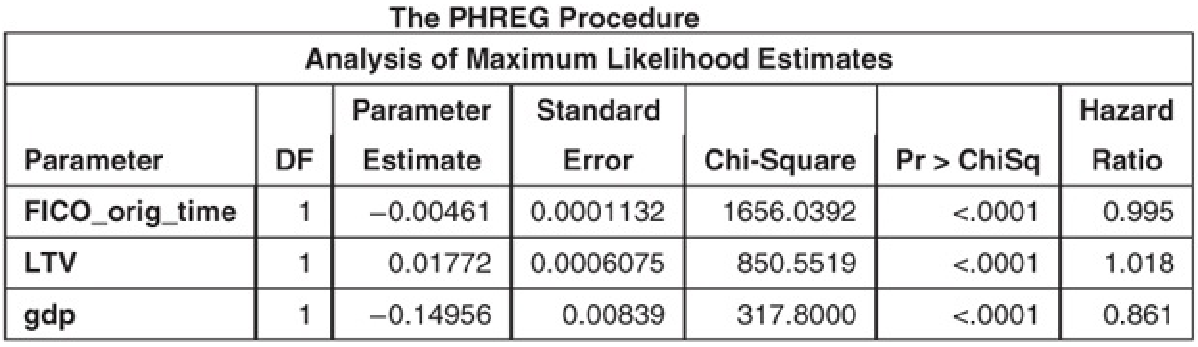
In a first setting one can aggregate time-varying information to compute a moment of the distribution of the covariates per borrower for the period during which a subject is at risk using a PROC MEANS statement or, alternatively, using the last value that is observed (we use this as an example, as the PROC MEANS by borrower can be time consuming):

```
PROC SORT DATA=data.mortgage OUT=mortgage;
BY DESCENDING id;
RUN;
PROC SORT DATA=mortgage OUT=moment NODUPKEY;
BY id;
RUN;
DATA moment(KEEP=id LTV gdp);
SET moment;
RENAME LTV_time=LTV;
RENAME gdp_time=gdp;
RUN;
PROC SORT DATA=lifetest;
BY id;
RUN;
DATA lifetest2;
MERGE lifetest(IN=a) moment;
BY id;
RUN;
```

We then run the PROC PHREG for the lifetime data set

```
PROC PHREG data=lifetest2;
MODEL time2*default_time(0)=FICO_orig_time LTV gdp/ TIES=EFRON;
RUN;
```

which gives the parameter estimates presented in Exhibit 7.15.



The PHREG Procedure

Analysis of Maximum Likelihood Estimates						
Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
FICO_orig_time	1	−0.00461	0.0001132	1656.0392	<.0001	0.995
LTV	1	0.01772	0.0006075	850.5519	<.0001	1.018
gdp	1	−0.14956	0.00839	317.8000	<.0001	0.861

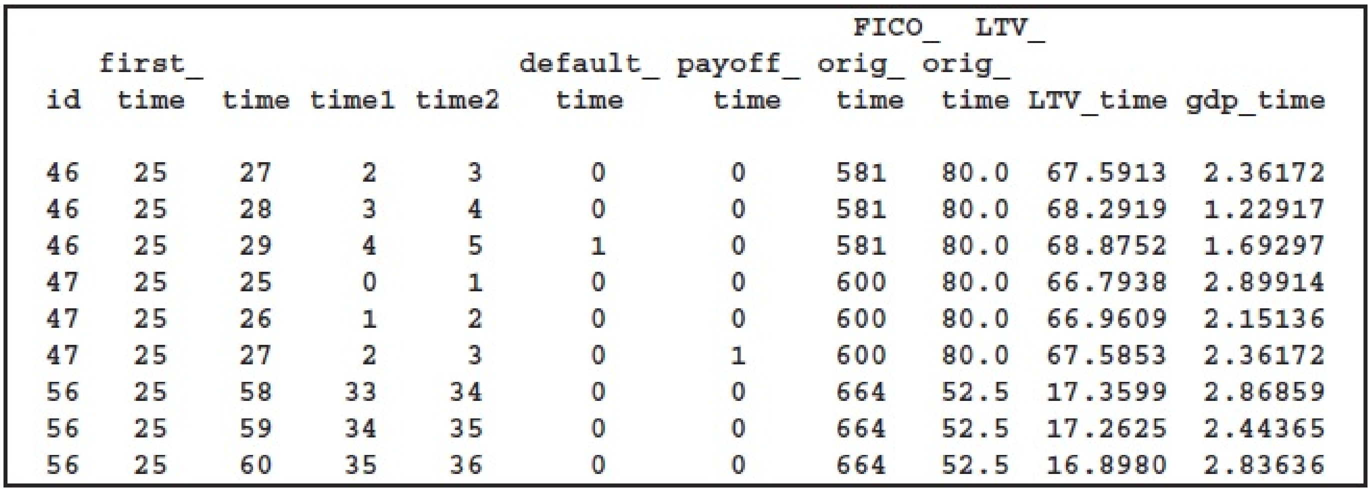
Exhibit 7.15 CPH Model

Time-Varying Covariates: Counting Process Data

The counting process data style of input requires the data to be in panel form (as for discrete-time hazard models) and two additional time stamps relative to the first observation time. These time stamps are (1) the time from first loan observation (alternatively, loan origination) to the beginning of an observation period, and (2) the time from the first loan observation to the end of an observation period.

```
DATA phreg;
SET data.mortgage;
time1 = time-first_time;
time2 = time-first_time+1;
RUN;
```

The data set looks as shown in Exhibit 7.16.



```
                                                       FICO_   LTV_
     first_                         default_  payoff_  orig_  orig_
id    time   time  time1  time2      time      time    time   time  LTV_time  gdp_time

46     25     27      2      3         0         0      581   80.0   67.5913   2.36172
46     25     28      3      4         0         0      581   80.0   68.2919   1.22917
46     25     29      4      5         1         0      581   80.0   68.8752   1.69297
47     25     25      0      1         0         0      600   80.0   66.7938   2.89914
47     25     26      1      2         0         0      600   80.0   66.9609   2.15136
47     25     27      2      3         0         1      600   80.0   67.5853   2.36172
56     25     58     33     34         0         0      664   52.5   17.3599   2.86859
56     25     59     34     35         0         0      664   52.5   17.2625   2.44365
56     25     60     35     36         0         0      664   52.5   16.8980   2.83636
```

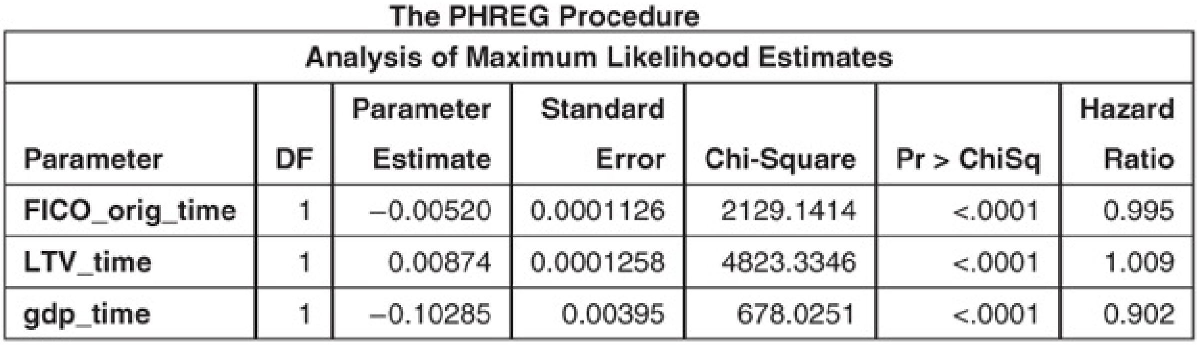
Exhibit 7.16 Counting Process Data

We now include one additional variable into our model, the GDP growth rate, and change the MODEL statement in PROC PHREG to accommodate the panel structure of the data with the two time stamps. The dependent variable in the MODEL statement is the two time stamps separated by a comma in brackets and an indicator variable, which indicates default (coded by one) or whether the borrower is no longer observed (i.e., censored, coded by zero). The censoring state is specified in brackets. The time stamps and censoring indicator are connected by “*”.

The remaining code is identical to the life table data:

```
ODS GRAPHICS ON;
PROC PHREG DATA=phreg PLOTS(OVERLAY)=SURVIVAL;
MODEL (time1,time2)*default_time(0)=FICO_orig_time LTV_time gdp_time
  TIES=EFRON;
BASELINE COVARIATES=covariates_time/ ROWID=set;
RUN;
ODS GRAPHICS OFF;
```

The counting process style of input invoked by the MODEL statement assumes that every line in the panel data is a stand-alone line. This corresponds to the assumption that every observation and subject is observed for one period in which default or nondefault occurs. This situation is comparable with discrete-time hazard models. Exhibit 7.17 shows the parameter estimates of the model and Exhibit 7.18 the survival plots for the upturn and downturn sets of covariates.



The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
FICO_orig_time	1	−0.00520	0.0001126	2129.1414	<.0001	0.995
LTV_time	1	0.00874	0.0001258	4823.3346	<.0001	1.009
gdp_time	1	−0.10285	0.00395	678.0251	<.0001	0.902

Exhibit 7.17 CPH Model

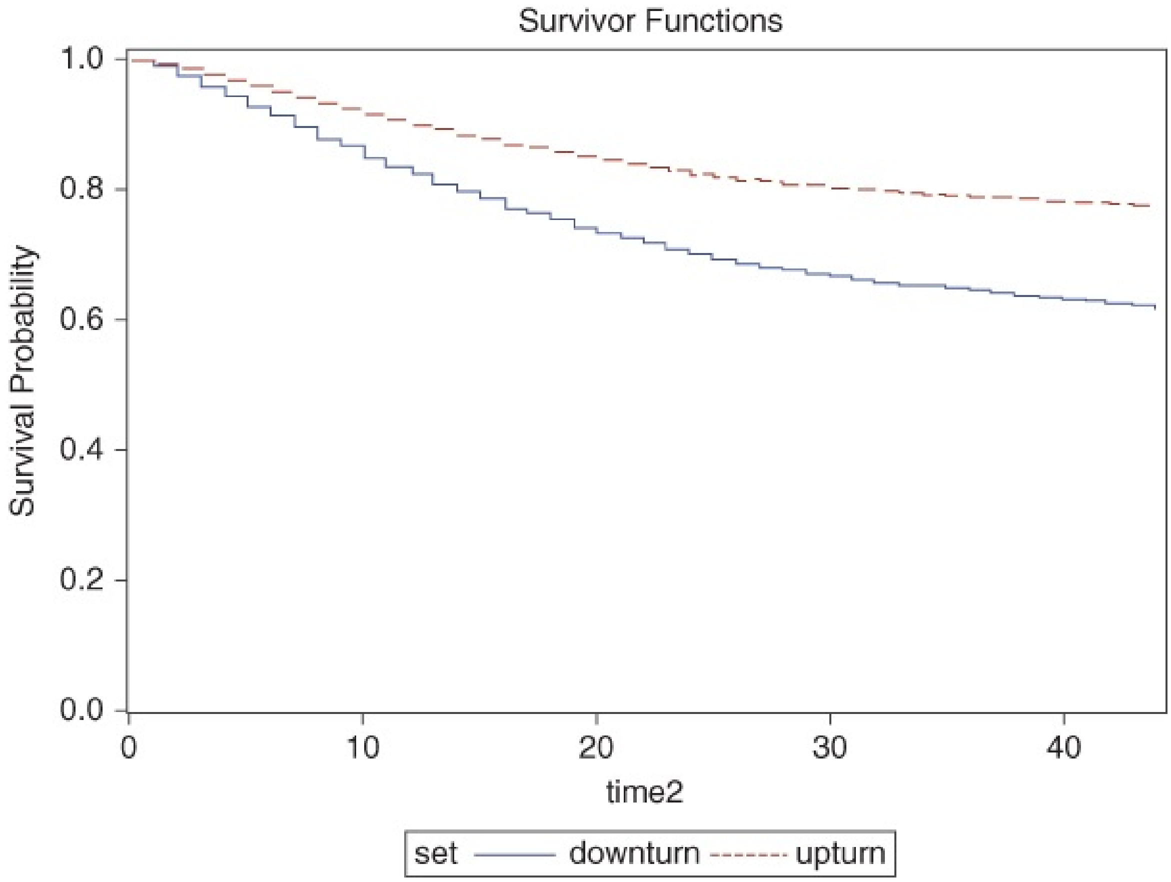


Exhibit 7.18 Survival Plot

The data set covariates are the basis of an analysis for two borrowers corresponding to an economic downturn (GDP growth of -3 percent) and an economic upturn state (GDP growth of 3 percent):

```
DATA covariates_time;
INPUT set $ FICO_orig_time LTV_orig_time gdp_time;
DATALINES;
downturn 800 60 -3
upturn 800 60 3
;
```

Estimation of Survival Probabilities

Computing default probabilities is not accommodated by design in CPH models. Remember, to simplify the parameter estimation, CPH models are based on the maximization of the partial likelihood that relates to the parameters of the explanatory variables and not to the baseline hazard function. To enable the computation of the probability density function, survival function, and hazard rate, SAS has added the BASELINE command to PROC PHREG, where the baseline hazard rate is estimated in a second stage using the approximate likelihood provided in Breslow (1974) as a default technique or the optional Kaplan and Meier (1958) technique. The survival probabilities are then estimated as follows:

$$\hat{S}(t|\boldsymbol{x}_\mathbf{i}) = \exp\ \left(-\int_0^t \hat{\lambda}_0(u)\exp\ (\hat{\boldsymbol{\beta}}'\boldsymbol{x}_\mathbf{i})du\right)$$

The BASELINE command is computationally expensive as it computes the probability of default for all interactions of the explanatory variables and time2. Different ways to reduce the complexity exist. For example, you may:

- Generate mutually exclusive subsets of data using PROC SURVEYSELECT.
- Estimate the full model using all data multiple times but with different subsets for the COVARIATES and OUT commands in the BASELINE specification.
- Append the various subsets from the OUT commands in the BASELINE specification to obtain a complete collection of default probabilities.
- Select the observations that reflect explanatory variables and time stamps (time2) of the original data set.

Alternatively, you may restrict the data set to observations with a certain time stamp (here time2) and/or draw a random sample using PROC SURVEYSELECT:

```
DATA phreg2;
SET phreg;
WHERE time2<=10;
KEEP time1 time2
default_time FICO_orig_time LTV_time gdp_time;
RUN;
PROC SURVEYSELECT DATA=phreg2 SAMPRATE=0.05 OUT=PHREG3 SEED=12345;
RUN;
```

PROC SURVEYSELECT randomly draws 5 percent of all remaining observations. The command SEED = 12345 fixes the random experiment so that the random draw results in the same outcome if executed another time. We now create a new data set called “survival”, which stores the survival probabilities for the first three periods (note that this step may take a few minutes despite the small sample size):

```
PROC PHREG data=phreg3;
MODEL (time1,time2)*default_time(0)=FICO_orig_time LTV_time gdp_time /
  TIES=EFRON;
BASELINE COVARIATES=phreg3 OUT=survival
SURVIVAL=SURVIVAL;
RUN;
```

Estimation of Default Probabilities

We then compute the PD as the first-order difference of the cumulative hazard rate of the current observation and the observation in the previous period ($S(t1) - S(t2)/S(t1)$):

```
DATA survival2(WHERE=(PD_time NE .));
SET survival;
IF time2>=1 THEN PD_time=(lag(survival)-survival)/lag(survival);
IF time2 =1 THEN PD_time=1-survival;
```

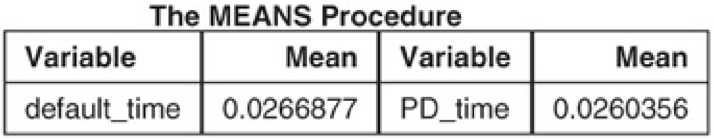
```
KEEP FICO_orig_time LTV_time gdp_time time2 PD_time;
RUN;
```

Finally, we merge the default probabilities of the initial estimation data set for all observations:

```
PROC SORT DATA=phreg3;
BY time2 FICO_orig_time LTV_time gdp_time;
RUN;
PROC SORT DATA=survival2 nodupkey;
BY time2 FICO_orig_time LTV_time gdp_time;
RUN;
DATA probabilities;
MERGE phreg3(IN=a) survival2;
BY time2 FICO_orig_time LTV_time gdp_time;
IF a;
RUN;
```

Calibration of CPH Models

We test the calibration of the CPH model with a comparison of the mean estimated default probability by the CPH model and the default rate in the sample. PROC MEANS for the default indicator and default probabilities shows that the mean of the in-sample PD estimates approximately matches the default rate (See Exhibit 7.19).



The MEANS Procedure

Variable	Mean	Variable	Mean
default_time	0.0266877	PD_time	0.0260356

Exhibit 7.19 Calibration of CPH Models: Comparison of Default Indicators and Estimated Default Probabilities

```
PROC MEANS DATA=probabilities MEAN NOLABELS;
VAR default_time PD_time;
RUN;
```

Note that the default rate for the data subset (based on the restriction $time2 \leq 10$ and the random sampling using PROC SURVEYSELECT) is a little higher than for the complete data set.

ACCELERATED FAILURE TIME MODELS

Accelerated failure time (AFT) models are parametric survival models that link the (log) transformed survival time to a linear predictor of the sum of parameter-weighted covariates:

$$\log\ (T_i) = \boldsymbol{\beta}'\boldsymbol{x}_i + \sigma\epsilon$$

with T_i the time to failure. We restrict the covariates to idiosyncratic information for now. The models are called AFT models as they assume that the parameters $\beta_1, \cdots, \beta_p$ and hence the impact of the independent variables are multiplicative on the event time. The parameters indicate how strongly the survival time accelerates or decelerates when a covariate changes by one unit. Krüger et al. (2015) apply AFT models.

The survival time is generally positive, and the transformation of the log-link function is defined from minus infinity to infinity, which matches the range of the model-implied estimated dependent variable that results from the weighted sum of the predictors and the volatility-weighted residual.

PROC LIFEREG fits these models for failure time data that can be uncensored, right-censored, left-censored, or interval-censored. The models for the dependent variable consist of a linear effect composed of the covariates and a random disturbance term. Common distributions for the residual are the exponential, Weibull (two parameters), lognormal, log-logistic, and gamma (three parameters) distributions. There are both graphical and likelihood procedures to choose the appropriate distribution.

Graphical Procedures

Let's reconsider some of the relationships that were introduced earlier. Remember that we have $\lambda(t)$ equals minus the derivative of $log S(t)$ to t, or stated differently, minus $log S(t)$ is equal to the integral from 0 to t of $\lambda(u)du$. Because of this relationship, the log survivor function is commonly referred to as the cumulative hazard function, which can be interpreted as the sum of the risks that are faced when going from time 0 to time t. If the survival times are exponentially distributed, then the hazard is constant and the cumulative hazard rate is equal to $\lambda * t$. Hence a plot of $-log\ (S(t))$ versus t should yield a straight line through the origin at 0. Similarly, it can be shown that if the survival times are Weibull distributed, then a plot of $log\ (-log\ (S(t)))$ versus $log\ (t)$ yields a straight line, not through the origin. In the case of a lognormal distribution, a plot of $N^{-1}(1 - S(t))$ versus $log\ (t)$ should yield a straight line, whereby N^{-1} represents the inverse, cumulative, standard normal distribution. Finally, in the case of a log-logistic distribution, a plot of $log\ ((1 - S(t))/S(t))$ versus $log\ (t)$ should be a straight line. Notice that in all these cases, the survival probabilities $S(t)$ to be plotted can be obtained from a Kaplan-Meier analysis. The plots of $-log\ (S(t))$ and $log\ (-log\ (S(t)))$ can be easily asked for in SAS using PROC LIFETEST with the PLOTS(LS,LLS) option. (See Exhibits 7.20 and 7.21.)

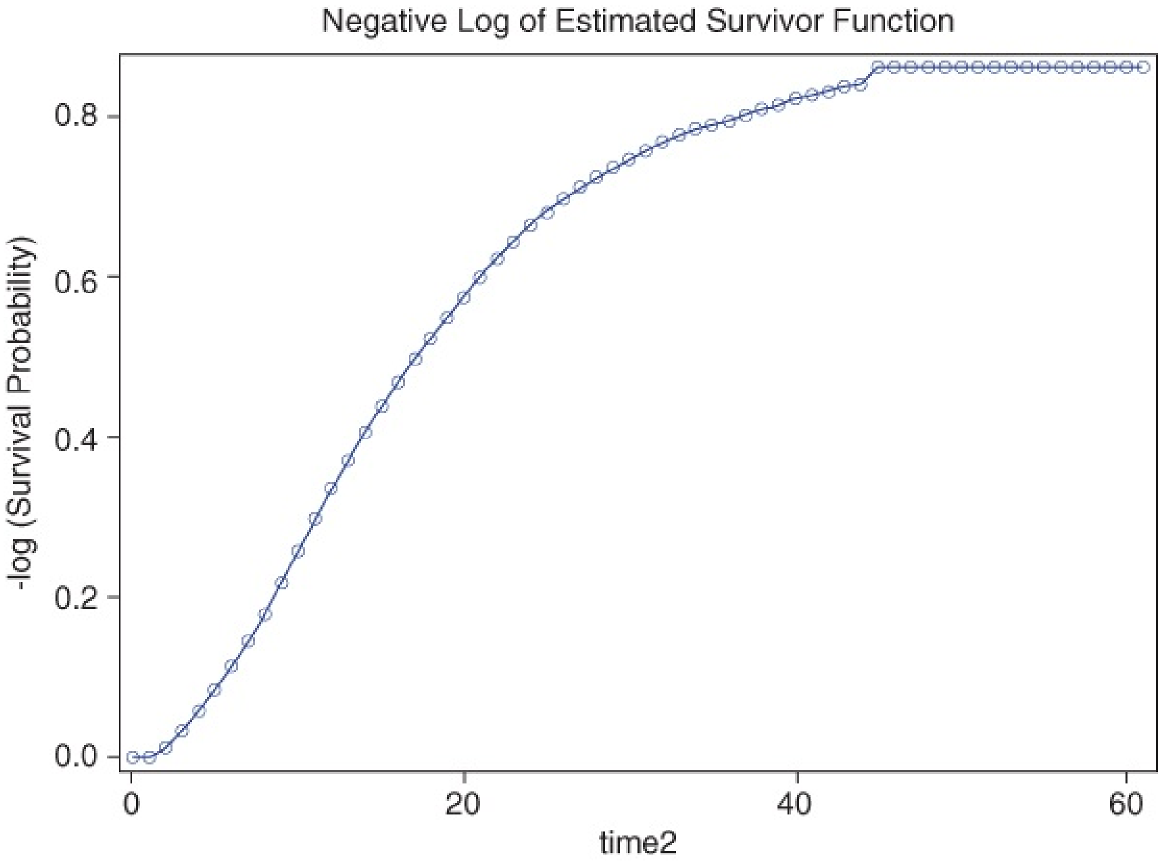


Exhibit 7.20 Graphical Procedures: Negative Log of Estimated Survivor Functions versus Time

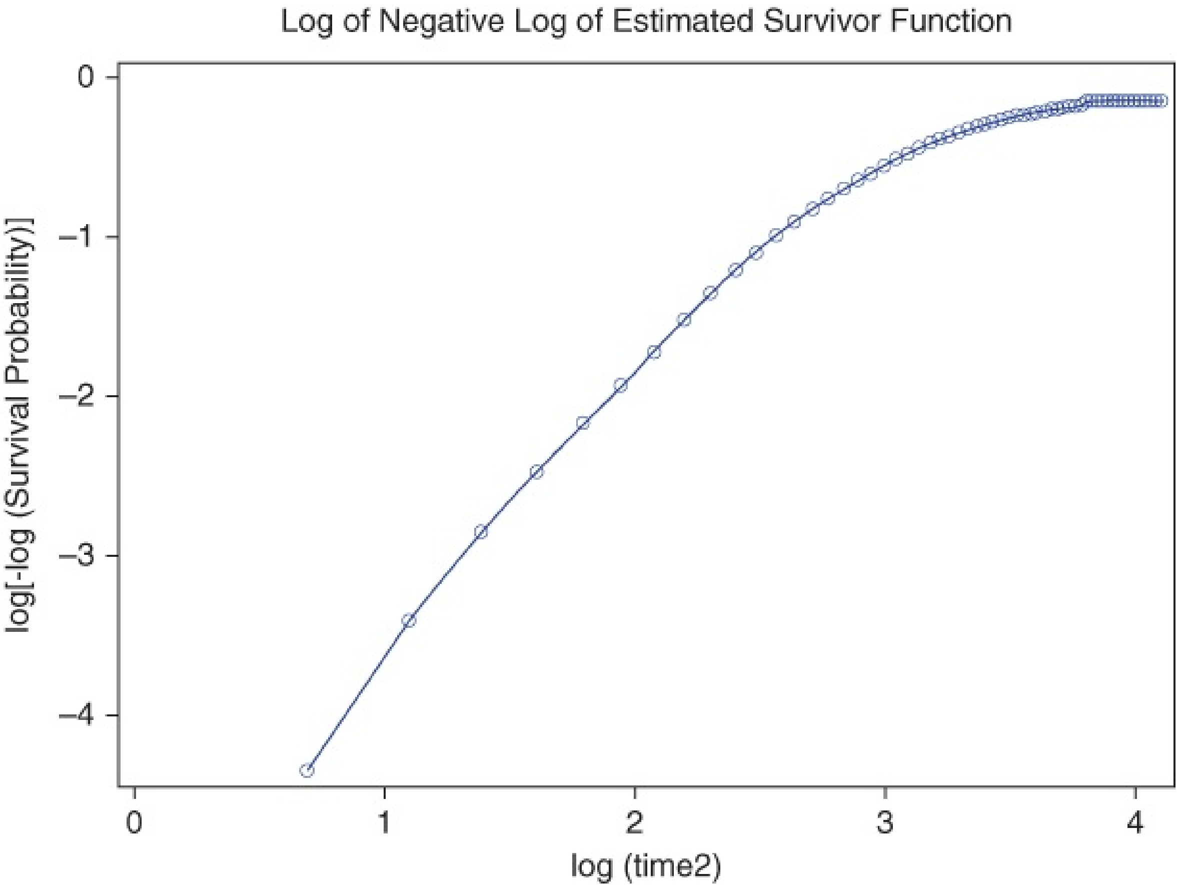


Exhibit 7.21 Graphical Procedures: Log of Negative Log of Estimated Survivor Functions versus the Log of Time

```
ODS GRAPHICS ON;
PROC LIFETEST DATA=lifetest METHOD=LT INTERVALS=(1 to 102 BY 1)
  PLOTS=(LS LLS);
TIME time2*default_time(0);
RUN;
ODS GRAPHICS OFF;
```

Likelihood Procedure

A more formal approach to evaluate model fit is based on a likelihood procedure. More specifically, the likelihood ratio test statistic can be used to compare models and test if one model is a special case of another. Let's start from the generalized gamma distribution, which you can see defined right here.

$$f(t) = \frac{\beta}{\Gamma(\kappa)\theta}\left(\frac{t}{\theta}\right)^{\kappa\beta-1} \exp\left(-\left(\frac{t}{\theta}\right)^{\beta}\right)$$

Note that it has three parameters: β, θ, and κ. If we now define $\sigma = 1/(\beta\sqrt{\kappa})$ and $\delta = 1/\sqrt{\kappa}$, then the Weibull, exponential, standard gamma, and lognormal distributions are all special versions of the generalized gamma distribution as follows: If σ equals δ, we have a standard gamma distribution; if δ equals one, we have a Weibull distribution; if both σ and δ equal one, we have an exponential distribution; and if δ equals zero, we have a lognormal distribution.

We can now use this to perform a likelihood ratio test. Let L_{full} be the likelihood of the full model (such as a generalized gamma distribution) and L_{red} be the likelihood of the reduced or specialized model (such as an exponential distribution). If both are very similar, then, of course, the reduced or specialized model will be selected. More formally, a chi-square statistic can be computed as $-2log\ (L_{\text{red}}/L_{\text{full}})$. The degrees of freedom correspond to the number of reduced parameters. You can see the various options listed in Exhibit 7.22:

Parameters	**Degrees of Freedom**
Freedom	
Exponential versus Weibull	1
Exponential versus standard gamma	1
Exponential versus generalized gamma	2
Weibull versus generalized gamma	1
Lognormal versus generalized gamma	1
Standard gamma versus generalized gamma	1

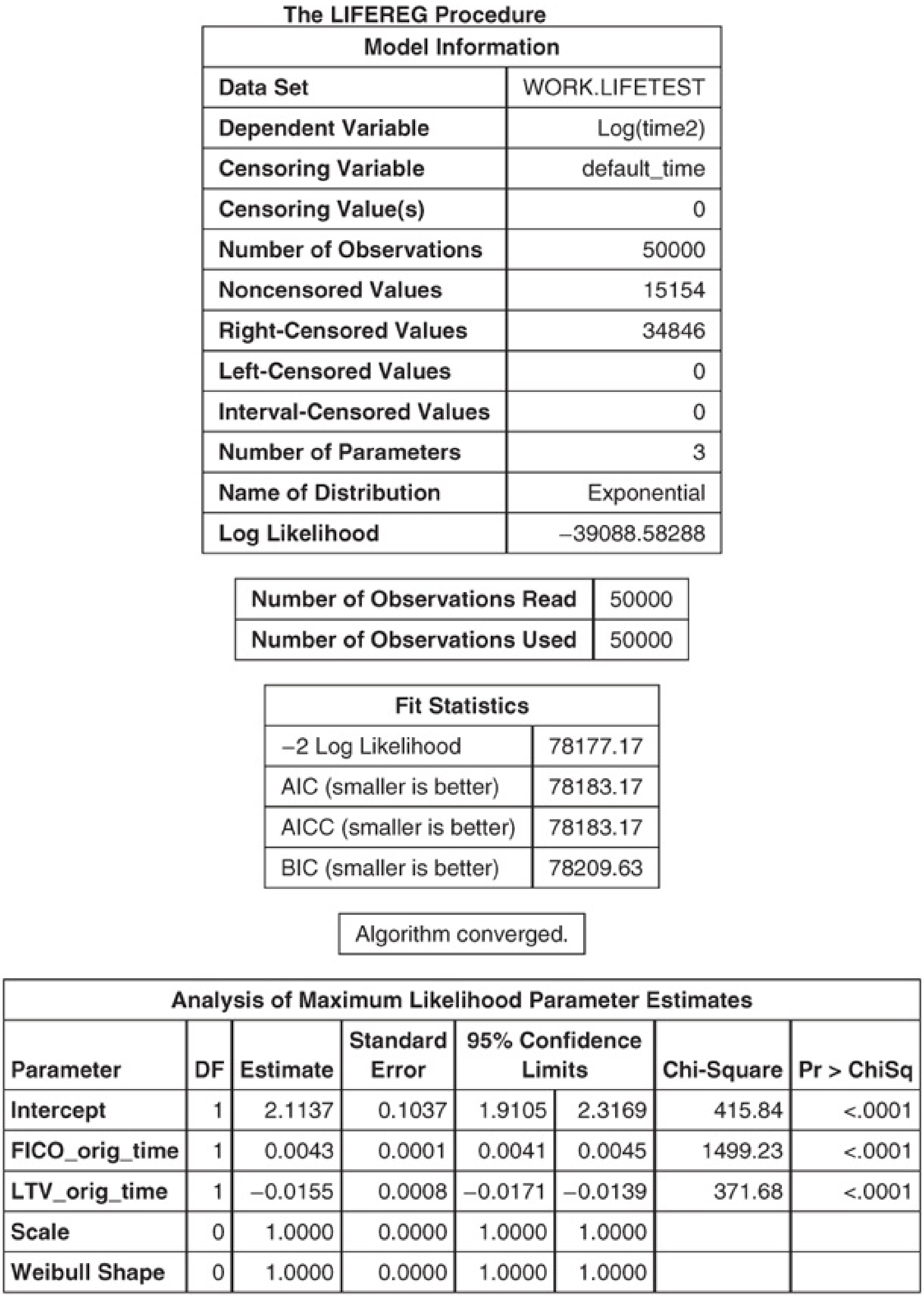
Exhibit 7.22 Degrees of Freedom for Likelihood Ratio Test

Accelerated Failure Time Models with PROC LIFEREG

As a simple first example, we choose an AFT model with an exponential distribution function for the residual (i.e., command / D = EXPONENTIAL in the MODEL statement). The exponential model has a constant hazard rate and the SAS code is:

```
    PROC LIFEREG DATA = lifetest;
    MODEL time2*default_time(0)
      = FICO_orig_time LTV_orig_time / D = EXPONENTIAL;
RUN;
```

We obtain the parameter estimates shown in Exhibit 7.23, which refer to the dependent variable of the survival time that is transformed by the natural logarithm. Contrary to PROC LOGISTIC (with the dependent variable default_time descending) and PROC PHREG, the parameter estimates have to be interpreted in the opposite way: An increase in the FICO score implies a higher survival time and hence lower probability of default. A higher LTV ratio at origination implies a lower survival time and hence higher probability of default. A higher GDP growth rate implies a higher survival time and hence lower probability of default.



The LIFEREG Procedure

Model Information	
Data Set	WORK.LIFETEST
Dependent Variable	Log(time2)
Censoring Variable	default_time
Censoring Value(s)	0
Number of Observations	50000
Noncensored Values	15154
Right-Censored Values	34846
Left-Censored Values	0
Interval-Censored Values	0
Number of Parameters	3
Name of Distribution	Exponential
Log Likelihood	−39088.58288

Number of Observations Read	50000
Number of Observations Used	50000

Fit Statistics	
−2 Log Likelihood	78177.17
AIC (smaller is better)	78183.17
AICC (smaller is better)	78183.17
BIC (smaller is better)	78209.63

Algorithm converged.

Analysis of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	2.1137	0.1037	1.9105	2.3169	415.84	<.0001
FICO_orig_time	1	0.0043	0.0001	0.0041	0.0045	1499.23	<.0001
LTV_orig_time	1	−0.0155	0.0008	−0.0171	−0.0139	371.68	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		
Weibull Shape	0	1.0000	0.0000	1.0000	1.0000		

Exhibit 7.23 LIFEREG Model

We show in Exhibit 7.23 that the exponential distribution is a special case of the popular Weibull distribution with the two parameters Scale and Weibull shape being equal to one. Note that CPH models are identical to AFT models if the baseline hazard rate is Weibull distributed in a full likelihood estimation. The parameters can be estimated by maximum likelihood. The constituents of the likelihood are the hazard rate, the survival function, and the probability density function.

These functions are for the exponential distribution:

- Hazard rate: $\lambda_i = \exp\,(-\boldsymbol{\beta}'\boldsymbol{x_i})$
- Survival function: $S_i(t_i) = \exp\,(-\exp\,(-\boldsymbol{\beta}'\boldsymbol{x_i} * t_i))$
- Probability density function: $f_i(t_i) = \exp\,(-\boldsymbol{\beta}'\boldsymbol{x_i})\exp\,(-\exp\,(\boldsymbol{\beta}'\boldsymbol{x_i}t_i))$

The hazard rate, survival function, and probability density function are derived in a similar way for the other distributions. The likelihood becomes:

$$\begin{aligned} L(\boldsymbol{\beta}, \boldsymbol{x}) &= \prod_{i=1}^{I} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i} \\ &= \prod_{i=1}^{I} \lambda_i^{\delta_i} S(t_i) \\ &= \prod_{i=1}^{I} (\exp\,(-\boldsymbol{\beta}'\boldsymbol{x_i}))^{\delta_i} \exp\,(-\exp\,(-\boldsymbol{\beta}'\boldsymbol{x_i} * t_i)) \end{aligned}$$

SAS then maximizes the logarithm of this likelihood.

Estimation of Default Probabilities

The estimation of default probabilities is not straightforward for AFT models in PROC LIFEREG. This is similar to the SAS implementation of the Cox proportional hazard model in PROC PHREG. We follow the methodology discussed previously and estimate the default probabilities from the survival probabilities.

First, we compute the hazard rate (λ_i). Remember that the hazard rate is time-invariant with regard to the life cycle (i.e., the baseline hazard rate) in the exponential model. This is also referred to as the memoryless property of the exponential distribution. Second, we compute default probabilities as the first-order difference of the cumulative hazard rate at the end of the current observation period and at the beginning of the current observation period. This is the same as the difference between the survival function at the beginning of the current period and the survival function at the end of the current period relative to the survival function at the beginning of the current period ($S_i(t1) - S_i(t2)/S_i(t1)$). Note that it is important that $S_i(t1)$ and $S_i(t2)$ are based on the same realizations of the time-varying covariates, as otherwise negative default probabilities are likely to result if the covariates indicate a lower risk at the end of a period than at the beginning of a period. We now compute the survival probabilities, which is somewhat different from the Cox proportional hazard model where we estimated default probabilities based on the PROC PHREG-generated survival probabilities:

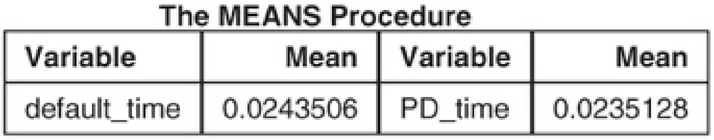
$$\hat{S}(t_i) = \exp\,(-\exp\,(-(\hat{\beta}_0 + \hat{\beta}_1 * \text{FICO\_orig\_time} + \hat{\beta}_2 * \text{LTV\_orig\_time}) * t_i))$$

This is implemented in the following data step:

```
DATA probabilities;
SET phreg;
xbeta=2.0998+0.0044*FICO_orig_time-0.0159*LTV_orig_time;
lambda = EXP(-xbeta);
S1 = 1-CDF('EXPONENTIAL', time1 ,1/lambda);
S2 = 1-CDF('EXPONENTIAL', time2 ,1/lambda);
PD_time = (S1-S2)/(S1);
RUN;
```

Calibration of AFT Models: Comparison of Default Indicators and Estimated Default Probabilities

A PROC MEANS for the default indicator and default probabilities shows that the mean of the in-sample PD estimates approximately matches the default rate. (See Exhibit 7.24.)



The MEANS Procedure

Variable	Mean	Variable	Mean
default_time	0.0243506	PD_time	0.0235128

Exhibit 7.24 Calibration of AFT Models: Comparison of Default Indicators and Estimated Default Probabilities

```
PROC MEANS DATA=probabilities MEAN NOLABELS;
VAR default_time
PD_time;
RUN;
```

Time-Varying Covariates

The estimation of time-varying covariates for AFT models is as challenging as it is for Cox proportional hazard models, and you may choose to refer to some of the solutions presented in the previous section on the CPH models. When estimating default probabilities based on time-varying information, care must be taken in the presence of time-varying covariates, in the sense that the survival probabilities are based on the same covariates, and no lag function is applied to derive the survival probability at the beginning/end of the prior period.

EXTENSION: MIXTURE CURE MODELING

A key assumption in survival analysis is that, in the long run, everyone will experience the event. In a medical setting, everybody will die at some point and the survival probability becomes 0. This is not true in a credit-risk modeling setting because a large part of the population never defaults. Hence, the survival probability of the population will not be zero, but levels out at some value. You can see this visualized in the figure shown in Exhibit 7.25.

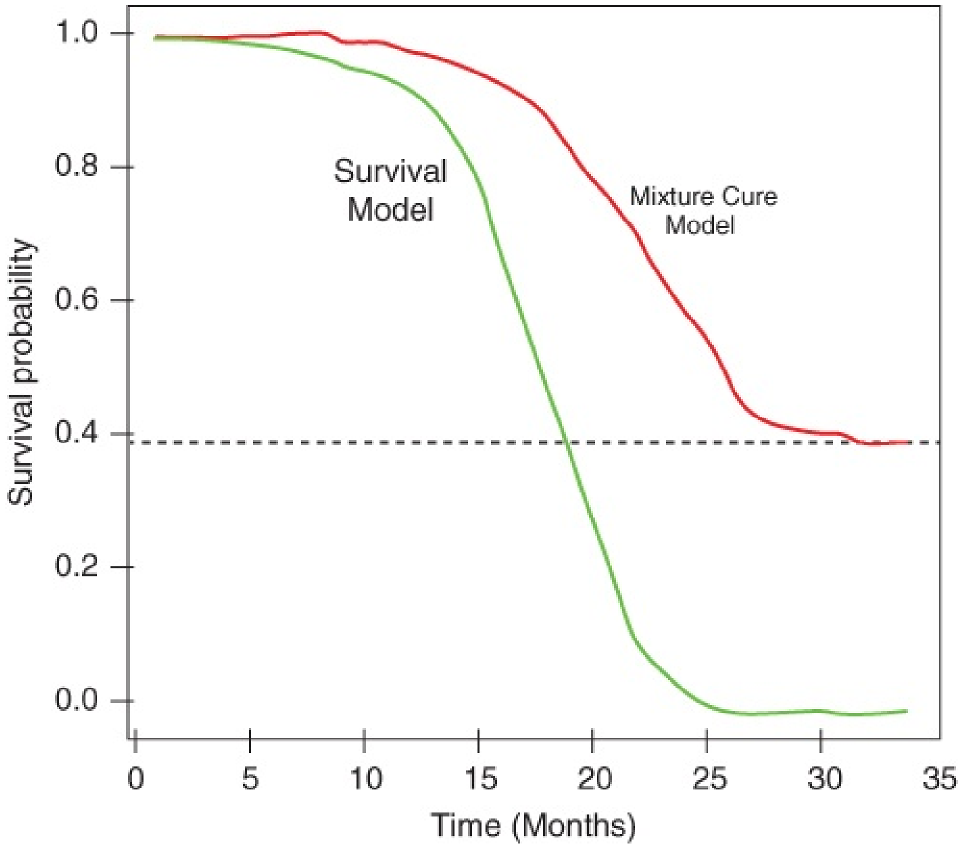


Exhibit 7.25 Mixture Cure Modeling

The survival model function goes to zero and the mixture cure model function goes to 0.4. Mixture cure modeling offers a solution to this problem by modeling distinct subpopulations.

Let us continue the example of a time-to-default prediction to illustrate the idea of mixture cure modeling. Let $Y = 1$ when an account is susceptible to default and $Y = 0$ otherwise. Let x and z be customer characteristics. You can think of characteristics such as age, income, marital status, and so on. The mixture cure model then becomes $S(t|x, z) = \pi(z)S(t|Y = 1, x) + 1 - \pi(z)$. As you can see, this model has two components. $\pi(z)$ is the probability of being susceptible to default, given characteristics z. It represents the incidence component and can be modeled using a binary logistic regression model, or any other classification technique. $S(t|Y = 1, x)$ is the latency model component and can be modeled using any survival analysis technique such as proportional hazards regression.

The parameters of the mixture cure model can then be estimated by formulating a combined likelihood function, which can be optimized by using the expectation maximization (EM) algorithm. This makes the mixture cure model computationally more intensive than the other survival models previously discussed. For more information, refer to Tong et al. (2012) and Dirick, Claeskens, and Baesens (2015).

DISCRETE-TIME HAZARD VERSUS CONTINUOUS-TIME HAZARD MODELS

In practice, discrete-time hazard models (in particular logit and probit regression models) are far more popular than continuous-time models. These models directly estimate the probability of default for the given data periodicity. Recent advances in regulation, in particular loan loss provisions under the International Financial Reporting Standards (IFRS 9), require the estimation of lifetime and life cycle effects (i.e., the baseline hazard rates). These effects can be included through controls that indicate the life cycle stage. Examples are lifetime dummies, time since origination, or time to maturity. Furthermore, interactions with other information may be added to the models. We generally recommend using these models as a first step.

Continuous-time hazard models may provide for more efficient and accurate default risk estimation, but come at the cost of complexity and loss of transparency. The analysis of these trade-off effects is an important consideration for the credit analyst. We recommend using these models in advanced applications.

PRACTICE QUESTIONS

1. Form five categories for the current LTV ratio. Estimate a life table model that is stratified by these five current LTV categories. Perform a statistical test for the equivalence of survival curves in the five samples.
2. Estimate a CPH model for PDs based on the FICO score and the LTV ratio at origination. Compute the survival function for a loan with a FICO score of 670 and an LTV ratio at origination of 90 percent. What is the PD for a loan five years from origination? Use data set mortgage.
3. Draw a random sample of 5,000 loans. Estimate two CPH models: one based on the FICO score and the LTV ratio at origination and another based on FICO score, the LTV ratio at origination, and the macro variable GDP growth. Estimate the probabilities of default, plot the default rate, and calculate the mean of the estimated default probabilities for both models by time. Use data set mortgage.
4. Estimate the probabilities of default for an AFT model based on the interest rate at origination. Use data set mortgage.

REFERENCES

Bellotti, T., and J. Crook. 2009. “Credit Scoring with Macroeconomic Variables Using Survival Analysis.” *Journal of the Operational Research Society* 60 (12): 1699–1707.

Breslow, N. 1974. “Covariance Analysis of Censored Survival Data.” *Biometrics* 30: 89–99.

Cox, D. R. 1972. “Regression Models and Life Tables (with Discussion).” *Journal of the Royal Statistical Society* 34: 187–220.

Cox, D. R. 1975. “Partial Likelihood.” *Biometrika* 62 (2): 269–276.

Dirick, L., T. Bellotti, G. Claeskens, and B. Baesens. 2015. “The Prediction of Time to Default for Personal Loans Using Mixture Cure Models: Including Macro-economic Factors.” *Proceedings of the Credit Scoring and Credit Control XIII Conference*.

Dirick, L., G. Claeskens, and B. Baesens. 2015. “An Akaike Information Criterion for Multiple Event Mixture Cure Models.” *European Journal of Operational Research* 241 (2): 449–457.

Kaplan, E. L., and P. Meier. 1958. “Nonparametric Estimation from Incomplete Observations.” *Journal of the American Statistical Association* 53 (282): 457–481.

Krüger, S., T. Oehme, D. Rösch, and H. Scheule. 2015. “Expected Loss Over Lifetime.” *Working Paper, University of Regensburg and University of Technology Sydney*.

Malik, M., and L. C. Thomas. 2010. “Modelling Credit Risk of Portfolio of Consumer Loans.” *Journal of the Operational Research Society* 61: 411–420.

Quigley, J. M., and R. Van Order. 1991. “Defaults on Mortgage Obligations and Capital Requirements for US Savings Institutions: A Policy Perspective.” *Journal of Public Economics* 44 (3): 353–369.

Tong, E. N., C. Mues, and L. C. Thomas. 2012. “Mixture Cure Models in Credit Scoring: If and When Borrowers Default.” *European Journal of Operational Research* 218 (1): 132–139.