Paper SD-004

Developing Business Failure Prediction Models Using SAS® Software

Oki Kim, Statistical Analytics

ABSTRACT

The credit crisis of 2008 has changed the climate in the investment and finance industry. The importance of developing accurate business failure prediction models has become more evident now than at any other time in recent history. Using various statistical techniques, business failure prediction models attempt to estimate the bankruptcy probability of a firm using a set of covariates. One powerful method is survival analysis. This paper applies two survival analysis techniques in developing business failure prediction models using SAS® system's PROC LIFETEST and PROC PHREG. Results are presented using data from publicly traded firms covering the first five quarters of the current recession. Through the work presented in this paper, the reader will learn how to estimate the survivor and hazard functions, to fit the Cox proportional hazard model, and to draw meaningful statistical inferences.

INTRODUCTION

Business failure prediction (BFP) has been a salient topic in finance for both researchers and practitioners for decades. Using various statistical techniques, BFP models attempt to estimate the bankruptcy probability of a firm using a set of covariates such as financial ratios, market-related variables, or the type of industry. Given the credit crisis of 2008, accurate BFP models are needed now more than ever. Recently, the Associated Press reported that bankruptcy filings among publicly traded companies surged 74 percent in 2008 ("Bankruptcy filings").

Gepp and Kumar illustrate the benefits of accurate BFP models for various entities and individuals (Gepp and Kumar 2008). Possible uses include: helping to avoid lending to (or investing in) businesses likely to fail, early identification of failing businesses by regulatory bodies, and more accurate scoring models for rating agencies.

Anyone interested in business failure prediction should ask the following questions:

- What is the probability that a business will survive beyond a particular time?
- Given that a business has survived up to time t, what is the firm's potential to fail?
- What covariates influence bankruptcy probability and the duration of time before bankruptcy?
- How do specific covariate values affect survival probability?

BANKRUPTCY DATA

An important consideration in developing a failure model is choosing an origin point, and length, of the study. For a failure prediction study on marriage, one might use the date of marriage as a natural starting point. In a medical study, the date of surgery or perhaps diagnosis may be appropriate. In business failure prediction (BFP) models, choosing a starting point is rather arbitrary. Economic downturns expose businesses to serious stress and can lead to failure for some firms. A US recession began in December 2007 according to the Business Cycle Dating Committee of the National Bureau of Economic Research (NBER): http://www.nber.gov/. The period of time used in this analysis starts from the beginning of the first quarter of 2008 and runs through the end of the first quarter of 2009 (five quarters of daily bankruptcy data). Public companies that filed Chapter 11 bankruptcy during the study period, and their main surviving competitors, are included. Additionally, the firm specific covariates used in this study come from publically available data just prior to the start of the study period (i.e. the last quarter of 2007).

Every domestic and foreign public company is required to file periodic reports, such as annual and 10-K reports, and other information with the U.S. Securities and Exchange Commission (SEC): http://www.sec.gov/. This information is available to the public via EDGAR database - the Electronic Data Gathering, Analysis, and Retrieval system. You can find a complete list of filings by company name, ticker symbol, Standard Industrial Classification Codes (SIC), or Central Index Key (CIK). In this study, once a bankrupt firm was identified during the study period, the firm's SIC code was used to select other non-bankrupt firms in the same industry code. Only publicly traded U.S. companies on the New York, American, and NASDAQ stock exchanges are considered. Companies that either merged or did not fail are considered censored.

Additional data sources used for this study include: Standard & Poor's Compustat file, LexisNexis database, BankruptcyData.com, Chapter11Library.com, American Bankruptcy Institute, and the Federal Deposit Insurance Corporation (FDIC http://www.fdic.gov/). In order to compute financial and market ratios, quarterly financial

statements were downloaded from Hoover's (a Dun & Bradstreet company). Historical prices were obtained from Standard & Poor's web site, Yahoo! Finance, or MSN Money. Note that the current government's Financial Regulatory Reform plan might bring slight changes to the roles of the U.S. Securities and Exchange Commission, the Federal Deposit Insurance Corporation, or other agencies. Thus, it is possible that such changes might affect how we collect the bankruptcy information in the future.

CHOOSING VARIABLES

Numerous researchers have applied sets of covariates to determine their influence on bankruptcy probability and duration of time before bankruptcy (Altman 1968; Gepp and Kumar 2008; LeClere 2005; Shumway 2001; Zmijewski 1984). Covariates used include both financial ratios and market-related variables. A set of financial variables used in one study included the following ratios: working capital to total assets, retained earnings to total assets, earnings before interest and taxes to total assets, market equity to total liabilities, and sales to total assets (Altman 1968). Another study included the ratio of net income to total assets, the ratio of total liabilities to total assets, and the ratio of current assets to current liabilities (Zmijewski 1984). Additionally, Shumway used both market-driven and accounting variables in developing hazard models. The market variables included market size, past stock returns, and the standard deviation of stock returns (Shumway 2001).

In addition to financial and market-related variables considered above, you can also examine whether the type of industry is a significant factor in the failure of businesses. In order to group the data by industry sector, the Global Industry Classification Standard (GICS®) methodology was utilized. GICS® assigns the company's industry, industry group, and sector and it is jointly developed and maintained by Standard & Poor's and MSCI Barra. The GICS classification system currently consists of 10 sectors: Consumer Discretionary, Consumer Staples, Energy, Financials, Health Care, Industrials, Information Technology, Materials, Telecommunication Services, and Utilities ("Global Industry Classification Standard").

A set of covariates used in the current study includes a combination of financial and market variables as well as the industry type. Covariates include: industry sector (SECT), working capital to total assets ratio (WCTA), earnings before Interest and taxes to total assets ratio (EBITTA), total equity to total liabilities ratio (TETL), sales to total assets ratio (STA), net income to total assets ratio (NITA), total liabilities to total assets ratio (TLTA), current assets to current liabilities ratio (CACL), price to earnings ratio (PE), price to book value ratio (PBV), and the common logarithm of market capitalization (LMC).

NON-PARAMETRIC SURVIVAL ANALYSIS

In this section, I describe how to estimate survival probabilities and graph survival curves using the Kaplan-Meier, or product limit, method. Note that Kaplan-Meier curves do not adjust for covariates. In addition, the product limit estimate of the survivor function is computed for each of the two industry types. Then, the statistical test of the equality of two survival curves is described. Finally, the estimation of hazard function is presented. Keep in mind that the survival and hazard descriptor functions provide a better understanding of what actually happened rather than making predictions.

In the following statements, PROC LIFETEST is invoked to compute the product limit estimate of the survivor function for each type of two industry sectors. METHOD=KM specifies that Kaplan-Meier estimates are computed. You can also specify as METHOD=PL for product limit, which is the default setting. PLOTS=(S) requests a plot of the estimated survivor function versus time (S for survival). In the TIME statement, which is required, the failure time variable (SURVT) is followed by the censoring variable (STATUS, with the value 1 indicating a business failure time and the value 0 indicating a censored time). The variable SECT (type of industry) is specified in the STRATA statement.

```
%LET LIB=CLIENT;
%LET DSN=BFM;

PROC LIFETEST DATA=&LIB..&DSN METHOD=KM PLOTS=(S);
TIME SURVT*STATUS(0);
STRATA SECT;
RUN;
```

OVERALL COMPARISON VS. COMPARISON OVER TIME

The descriptive measures provide overall comparison of the two sectors. The output from PROC LIFETEST contains the summary statistics of the survival times and the distribution of event and censored observations between the two sectors. In addition, the average survival time and the average hazard rate, for each stratum, can be obtained by running the SQL procedure below. The macro is called for each sector. Note that the average hazard rate is computed by dividing the total number of failures by the sum of the observed survival times.

```
%MACRO SIMPLE(SEC);
 PROC SQL;
   RESET NOPRINT;
   SELECT AVG(SURVT) INTO:MEAN FROM &LIB..&DSN WHERE SECT=&SEC;
   SELECT SUM(SURVT) INTO:SUM_TIME FROM &LIB..&DSN WHERE SECT=&SEC;
   SELECT SUM(STATUS) INTO:N_FAIL FROM &LIB..&DSN WHERE SECT=&SEC;
 QUIT;
 DATA DESCSTAT;
   SECTOR=&SEC;
   MEAN_TIME=&MEAN;
   H_RATE= &N_FAIL/&SUM_TIME;
 RUN;
 PROC SQL;
   SELECT SECTOR, MEAN_TIME AS AVG_SURV_TIME, H_RATE AS AVG_HAZARD_RATE
   FROM DESCSTAT;
 QUIT;
%MEND SIMPLE;

%SIMPLE(1)
%SIMPLE(2)
```

Below is an example of the output of descriptive measures. Average survival times are 337 days and 359 days for sector one and sector two, respectively.

SECTOR	AVG_SURV_ TIME	AVG_HAZARD_ RATE
1	336.895	.001562256
2	358.922	.001529637

The comparison over time can be described by the survival curves. The Kaplan-Meier method is used to estimate the survival curves. Figure 1 shows the estimated survivor functions for sectors one and two. The survival rates of sector one decrease rapidly in roughly 250 days. Shapes of the curves of two sectors are quite different. The sector two curve decreases more rapidly initially than the sector one curve, but the role is reversed in the later period. Additionally, the median survival times are 336 days and 414 days for the sector one and sector two, respectively. The median survival time is the time at which the survival probability is 0.5 for each group. This statistic is provided in a table of quartile estimates of summary statistics from the output of PROC LIFETEST.

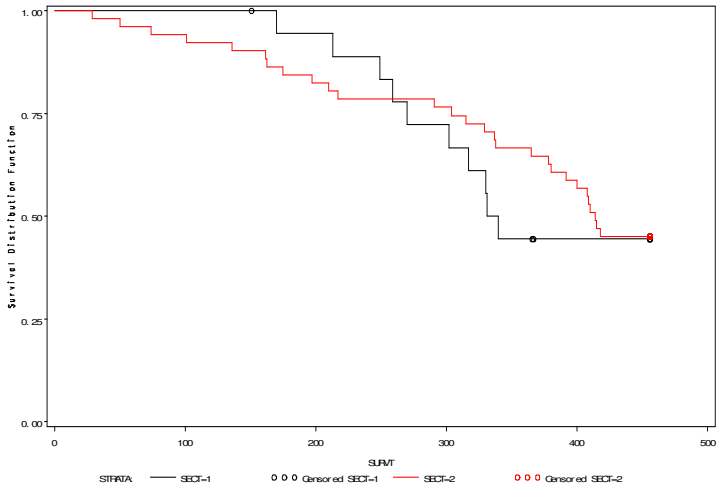

Figure 1. Estimated survivor curves for the two sectors

HOMOGENEITY TESTS ACROSS STRATA

After the multiple survival curves are estimated, the next step is to conduct homogeneity test across strata. The null hypothesis is that there is no difference among the survival curves. The statistical tests often employed are the log-rank test and the Wilcoxon test. The log-rank test statistic is approximately chi-square distributed in large samples with K-1 degrees of freedom, where K denotes the number of groups being compared. The Wilcoxon test allows early failures to receive more weights than later failures whereas the log-rank test gives equal weight to failures at all

failure time. For more detailed discussion on the statistical tests of multiple survival curves, refer to the references (Hosmer and Lemeshow 1999; Kleinbaum and Klein 2005).

Results of the homogeneity tests across two industry types are given in the output below. The null hypothesis is that the two survival curves are the same. All three tests and the corresponding p-values indicate no evidence of a significant difference among the survival curves for the two types of industry.

The LIFETEST Procedure
Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.1635	1	0.6859
Wilcoxon	0.1599	1	0.6892
-2Log(LR)	0.0011	1	0.9736

HAZARD FUNCTION

In the statements below, PROC LIFETEST is invoked to compute the life table estimates by specifying METHOD=LT or LIFE. The hazard function represents the instantaneous rate of failure. Given that an individual survived up to time t, the hazard rate indicates the potential to fail. You might be familiar with the default rate, terminology used in credit risk modeling. Note that the hazard function is not a probability whereas the survival function is. In the INTERVALS=option, you can specify the time intervals used in the model. PLOTS=(H) requests a plot of the estimated hazard function over time.

```
PROC LIFETEST DATA=&LIB..&DSN METHOD=LT INTERVALS=(0 TO 500 BY 90) PLOTS=(H);
TIME SURVT*STATUS(0);
RUN;
```

Figure 2 shows the plot of the estimated hazard function in my example. It indicates that the failure rate increases roughly linearly over the five quarter period.

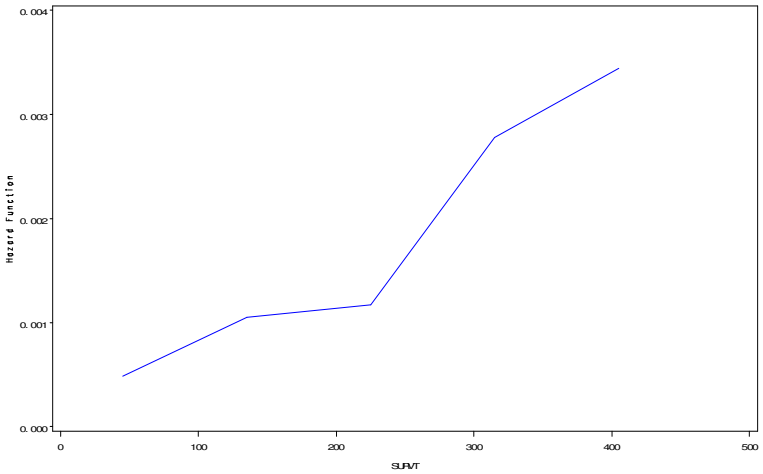

Figure 2. Estimated hazard function over time (in days)

SEMI-PARAMETRIC COX PROPORTIONAL HAZARD MODEL

In this section, I describe how to develop BFP models using a Cox proportional hazards (PH) model. A brief review of Cox PH model is presented, followed by variable selection methods, model adequacy assessment, and interpretation of a fitted model. SAS PROC PHREG fits survival data using a set of covariates. By employing the procedure and regression, the effect of covariates on bankruptcy probability and duration of time before bankruptcy can be evaluated.

REVIEW OF COX PROPORTIONAL HAZARDS (PH) MODEL

The Cox PH model expresses the individual's hazard at time t with a set of covariates as shown in the following formula:

$$H(t) = h_0(t) \times \exp(ß_i X_i + ß_2 X_2 + \ldots + ß_p X_p)$$

One prominent feature of the Cox PH model, compared to other statistical BFP techniques such as discriminant and logistic models, is that it is a semi-parametric technique. No assumptions are made about the baseline hazard

function, $h_0(t)$. Thus, the hazard rate is a function of the covariates, which are time-independent. Another prominent feature of the Cox PH model is that it considers censoring observations. The Cox model accounts for survival times, thus, it uses more information than the logistic model.

When estimating the model parameters, the maximum likelihood (ML) estimates are derived by maximizing a likelihood function. The Cox likelihood is based on the observed order of events rather than the joint distribution of events. Thus, the Cox likelihood is called a "partial" likelihood. In addition, a set of covariates in the Cox PH model can be time-dependent (or time-varying) covariates. The extended Cox model can be employed in such a case. For more detailed discussion on the Cox PH model, refer to the references (Hosmer and Lemeshow 1999; Kleinbaum and Klein 2005).

MODEL DEVELOPMENT

The variable selection methods in a proportional hazard regression are similar to the methods in linear or logistic regression. Below, I make an example of each of the following methods: backward elimination, stepwise selection, and best subset selection.

Backward elimination begins by including all covariates in the model and testing each covariate for statistical significance. Any covariate that is not significant is removed from the model. To execute the backward elimination, you can specify the SELECTION=BACKWARD option in the MODEL statement of PROC PHREG as shown below. The option SLSTAY=0.1 indicates that a level of significance of 10 percent is chosen for retaining covariates in the model. In my BFP application, the backward elimination process identified two covariates, LMC (log of market capitalization) and WCTA (ratio of working capital to total assets), that affect the survivorship of the publicly traded companies during the study period.

```
PROC PHREG DATA=&LIB..&DSN;
MODEL SURVT*STATUS(0)= SECT PE PBV LMC WCTA EBITTA TETL STA NITA CACL/ RL
SELECTION=BACKWARD SLSTAY=0.1;
RUN;
```

The stepwise selection method is a combination of forward selection and backward elimination. The forward selection adds covariates to the model whereas the backward elimination removes covariates from the model. To produce a stepwise regression analysis, you can specify the SELECTION= STEPWISE option in the MODEL statement of PROC PHREG below. The options SLENTRY=0.25 and SLSTAY=0.15 indicate that 25 percent and 15 percent significance levels are chosen for a variable to enter into the model and to stay in the model, respectively. In this example, the stepwise selection process chose LMC (log of market capitalization) and WCTA (ratio of working capital to total assets). Recall that the stepwise selection method has some problems such as collinearity and biased regression coefficients.

```
PROC PHREG DATA=&LIB..&DSN;
MODEL SURVT*STATUS(0)= SECT PE PBV LMC WCTA EBITTA TETL STA NITA CACL/ RL
SELECTION=STEPWISE SLENTRY=0.25 SLSTAY=0.15;
RUN;
```

The best subset selection process in a proportional hazard regression is similar to the method in linear regression. The global score chi-square statistic is used as a test criterion. To carry out the best subset selection, the option SELECTION=SCORE is specified in the MODEL statement of PROC PHREG below. By specifying BEST=3 option, the procedure identifies the three best models with the largest score test value for each number of covariates.

```
PROC PHREG DATA=&LIB..&DSN;
MODEL SURVT*STATUS(0)= SEC PE PBV LMC WCTA EBITTA TETL STA NITA CACL/ RL
SELECTION=SCORE BEST=3;
RUN;
```

Below is an example of how SAS PHREG procedure displays a table of the regression models selected by score criterion. The models are listed in descending order of their score chi-square values within each model size. Among all models containing two covariates (`Number of Variables=2`), the model that contains the variables LMC and WCTA has the largest score value (37.6252), the model that contains the variables SECT and LMC has the second-largest score value (33.6505), and the model that contains the variables LMC and CACL has the third-largest score value (29.2975). Only a partial output is shown below.

```
Regression Models Selected by Score Criterion

 Number of       Score
 Variables  Chi-Square  Variables Included in Model

    1       25.5309        LMC
    1       15.5428        EBITTA
    1       7.1175         WCTA
```

```
2        37.6252        LMC WCTA
2        33.6505        SECT LMC
2        29.2975        LMC CACL

3        39.6176        PBV LMC WCTA
3        38.6942        LMC WCTA EBITTA
3        38.5758        LMC WCTA TETL
```

MODEL ASSESSMENT

Once you have fitted data with a set of covariates using the Cox model, you will want to carry out an assessment of model adequacy using various statistical tests. In this section, I describe the statistical tests on the overall goodness of fit, the assumption of proportional hazard, and the evaluation of residuals.

In the following statements, PROC PHREG is used to fit the Cox proportional hazards model and to analyze the effects of the financial and market variables on the survival of the company. The response variable in a hazard model is the time of non-failure firms and an indicator variable for censorship status (with the value 1 indicating a failure time and the value 0 indicating a censored time). SURVT*STATUS(0) is specified in the MODEL statement. The values of SURVT are considered censored if the value of STATUS is 0; otherwise, they are considered failure times. Two covariates are included in the model as predictors of survival time: LMC (log of market capitalization) and WCTA (ratio of working capital to total assets).

```
PROC PHREG DATA=&LIB..&DSN;
MODEL SURVT*STATUS(0)= LMC WCTA /RL ;
RUN;
```

OVERALL GOODNESS-OF-FIT

One measure of overall goodness-of-fit test is partial likelihood-ratio test. After invoking PROC PHREG, you can obtain the likelihood-ratio chi-square statistic from the Model Fit Statistics table. The output produces the values of -2 log likelihood for fitting a model without covariates and for fitting a model with all covariates (295.210 - 264.436 = 30.774). This statistic is also shown in the table Testing Global Null Hypothesis: BETA=0. The null hypothesis is that all variables included in the model have coefficients of 0. The likelihood-ratio test statistic equals 30.7742 with 2 degrees of freedom. Thus, the null hypothesis is rejected (p<0.0001).

The PHREG Procedure
Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	295.210	264.436
AIC	295.210	268.436
SBC	295.210	271.711

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	30.7742	2	<.0001
Score	37.6252	2	<.0001
Wald	34.0074	2	<.0001

Another measure of model performance may be "some measure analogous to R^2", as shown in the formula below, where "L_p is the log partial likelihood for the fitted model with p covariates and L_0 is the log partial likelihood for model zero with no covariates" (Hosmer and Lemeshow 1999). N represents the number of observations. Keep in mind that this measure does not explain the proportion of variability of the response variable by the explanatory variables as in the linear regression. However, it provides a pseudo measure of association between the response variable and covariates.

$$R^2 = 1 - \exp((L_0-L_p)*2/n)$$

You can simply compute generalized R^2 using the formula, given L_p, L_0, and n. In my application, this statistic is passed to another routine, so I created the following macro to carry out the calculation. Passing parameter by reference code is not provided here. In this example, the generalized R^2 approximately equals to 0.5849. The model seems to provide a good fit.

```
%MACRO RSQUARE(N);
PROC PHREG DATA=&LIB..&DSN;
 MODEL SURVT*STATUS(0)= LMC WCTA /RL ;
 ODS OUTPUT FITSTATISTICS=STATOUT;
RUN;
```

```
PROC SQL;
 SELECT WITHOUTCOVARIATES INTO: LO FROM STATOUT WHERE CRITERION='-2 LOG L';
 SELECT WITHCOVARIATES INTO: LP FROM STATOUT WHERE CRITERION='-2 LOG L';
QUIT;
DATA GRSQ;
 X=((-&LO)-(-&LP))*2/&N;
 R=1-EXP(X);
RUN;
PROC SQL; SELECT R AS GEN_RSQUARE FROM GRSQ; QUIT;
%MEND RSQUARE;
```

TESTING PH ASSUMPTION

The key assumption in the Cox model is the proportional hazard (PH) assumption. It assumes that the hazard ratio for the two individuals with the different covariate values is independent of time. In order to assess the PH assumption, a statistical test can be carried out by finding the correlation between the Schoenfeld residuals for a given covariate and the ranked individual failure time (Kleinbaum and Klein 2005).

In the PROC PHREG statements below, the Schoenfeld residuals for each covariate are obtained in the output dataset RES. The residuals for LMC and WCTA are named as RLMC and RWCTA, respectively, in the RESSCH= option of the OUTPUT statement. In the following SQL and RANK procedures, the Schoenfeld residuals for a failure event are selected and the ranked survival time variable is given a name T_Rank. Finally, PROC CORR is carried out to obtain the correlations between the Schoenfeld residuals and the ranked failure time (T_Rank) for each covariate used in the model.

```
PROC PHREG DATA=&LIB..&DSN;
MODEL SURVT*STATUS(0)= LMC WCTA /RL ;
OUTPUT OUT=RES RESSCH=RLMC RWCTA;
RUN;
PROC SQL;
CREATE TABLE FAIL AS
SELECT * FROM RES WHERE STATUS=1;
QUIT;
PROC RANK DATA=FAIL OUT=RANKED TIES=MEAN;
VAR SURVT;
RANKS T_RANK;
RUN;
PROC CORR DATA=RANKED NOSIMPLE;
VAR RLMC RWCTA;
WITH T_RANK;
RUN;
```

Below is an output of testing the PH assumption in my example. The p-values for LMC and WCTA are 0.2399 and 0.9459, respectively. Results indicate that there is no relation between the residuals and survival time, thus both covariates meet the PH assumption. Note that you can employ a stratified Cox model or time-dependent variables if the PH assumption is not satisfied (Hosmer and Lemeshow 1999; Kleinbaum and Klein 2005).

Pearson Correlation Coefficients
Prob > |r| under H0: Rho=0

	RLMC	RWCTA
T_RANK	0.19534	0.01139
Rank for Variable SURVT	0.2399	0.9459

ANALYSIS OF RESIDUALS

Martingale and deviance residuals can be obtained by specifying RESMART and RESDEV in an OUTPUT statement in the following PROC PHREG statements. The RESOUT data set contains linear predictors, Martingale residuals, and deviance residuals for the Cox proportional hazards regression analysis. By invoking PROC GPLOT, the plots of Martingale residuals versus linear predictors and deviance residuals versus linear predictors can be generated. The resulting plots are not shown here. In my example, an outlier is detected in the Martingale residual plot. However, the deviance residual plot does not seem to reveal an apparent outlier.

```
PROC PHREG DATA=&LIB..&DSN NOPRINT;
MODEL SURVT*STATUS(0)= LMC WCTA /RL ;
OUTPUT OUT=RESOUT XBETA=XB RESMART=MART RESDEV=DEV;
RUN;
```

```
SYMBOL COLOR=BLUE VALUE=DOT HEIGHT=1.0;
PROC GPLOT DATA=RESOUT;
PLOT MART*XB;
RUN;
PROC GPLOT DATA=RESOUT;
PLOT DEV*XB;
RUN; QUIT;
```

INTERPRETATION OF A FITTED MODEL

After you have settled on assessing the adequacy of the model that seems a good-fit, you can carry out statistical inferences of a fitted model. The output below is produced by running PROC PHREG with two covariates, LMC (log of market capitalization) and WCTA (ratio of working capital to total assets). The RL (RISKLIMITS) option in the MODEL statement provides 95% confidence intervals for the hazard ratio estimates.

```
PROC PHREG DATA=&LIB..&DSN;
MODEL SURVT*STATUS(0)= LMC WCTA /RL ;
OUTPUT OUT=S_OUT SURVIVAL=S;
RUN;
```

The PHREG Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
LMC	1	-0.57207	0.11612	24.2713	<.0001	0.564	0.449	0.709
WCTA	1	-0.97880	0.38719	6.3906	0.0115	0.376	0.176	0.803

Results of the analysis indicate that both covariates appear to add significantly to the model. The p-values of the parameter estimates for the regression coefficients are highly significant ($p<0.0001$ and $p=0.0115$ for LMC and WCTA, respectively). In addition, both covariates appear to be associated with a decrease in the risk to fail. The hazard ratio for the effect of the covariate is obtained by exponentiating the estimated regression coefficient of that covariate. In my example, the estimated hazard ratio for LMC covariate is $e^{-0.57207} = 0.564$. The estimated hazard ratio for WCTA covariate is $e^{-0.97880} = 0.376$.

When interpreting a fitted proportional hazards model, examining the scale of covariates is recommended: continuous or categorical. In my example, both covariates are continuous. Thus, the hazard ratio is the ratio of hazard rates for an increase of one unit of the variable. For example, the hazard ratio estimate for LMC is 0.564, meaning that an increase of one unit in the log of market capitalization will shrink the hazard rate by 43.6%. That is, the company that is 10 times bigger in terms of market capitalization will reduce the risk of failing by 43.6%. The hazard ratio estimate for WCTA is 0.376, meaning that an increase of one unit in the ratio of working capital to total assets will shrink the hazard rate by 62.4%. Further analysis shows that a decrease in the risk of failing by 9.30% when a firm increases its liquidity ratio by 10%.

Along with a point estimate, a 95% confidence interval for the hazard ratio is also given. A 95% confidence interval for LMC is given by the range of values 0.449-0.709. This interval includes the point estimate of 0.564 and does not contain the null value of 1. The interval width equals 0.260 (0.709-0.449). Also, a 95% confidence interval for WCTA is given by the range of values 0.176-0.803. This interval also includes the point estimate of 0.376 and does not contain the null value of 1. The interval width equals 0.627 (0.176-0.803).

PREDICTED SURVIVAL CURVES FOR SPECIFIC COVARIATE VALUES

Once the Cox proportional hazards regression analysis results are obtained, the predicted survival curves for specific covariate values can be produced. In the following PROC SQL statements, a new SAS data set CIN is created containing a set of covariate values. In my example, the specific covariate values of a firm, a financial services company that was reported to go bankrupt in 2009, are 2.2861 and -0.0329 for LMC and WCTA, respectively.

In the BASELINE statement, an input dataset CIN is specified in the COVARIATES option and an output data set COUT is specified in the OUT option, which contains the adjusted survival estimates (here I name it S in the SURVIVAL option). In addition, the lower and upper 95% confidence limits for these estimates are requested in the LOWER and UPPER options. By default, the output data set also contains the survival estimates using the mean values of various sets of covariates. You can stop this by specifying the NOMEAN option. By Running the PROC GPLOT statements, a plot of the adjusted survival estimates versus survival time is generated. The predicted survival function for this company is presented in Figure 3.

```
PROC SQL;
CREATE TABLE CIN(LMC NUM, WCTA NUM);
INSERT INTO CIN (LMC, WCTA)
VALUES (2.2861,-0.0329);
QUIT;
PROC PHREG DATA=&LIB..&DSN;
MODEL SURVT*STATUS(0)= LMC WCTA /RL ;
BASELINE COVARIATES=CIN OUT=COUT SURVIVAL=S LOWER=SL UPPER=SU /NOMEAN;
RUN;
PROC GPLOT DATA=COUT;
PLOT S*SURVT;
RUN; QUIT;
```

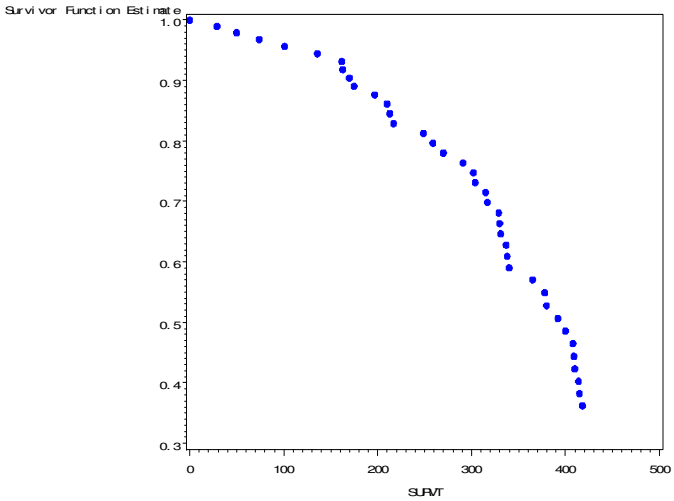

Figure 3. Predicted survival function for specific covariate values

CONCLUSION

With the recent financial crisis, developing accurate business failure prediction models has become more important. The statistical methods and procedures explained in this paper include survival analysis techniques. The Cox PH model is powerful and popular due to its semi-parametric technique and the use of censoring observations. You can also employ other statistical techniques and procedures such as logistic analysis and SAS® system's PROC LOGISTIC or discriminant analysis and SAS® system's PROC CANDISC/PROC DISCRIM. One recommended way to evaluate the prediction accuracy is to compare the classification results of each statistical method. This further analysis will be presented in a future paper.

REFERENCES

Altman, E. I. 1968. Financial Ratios, Discriminant Analysis, and the Prediction of Corporate Bankruptcy. *Journal of Finance* 23: 589-609.

"Bankruptcy filings surged in 2008." Associated Press. Yahoo! Finance.
13 Jan. 2009 <http://finance.yahoo.com/news/Bankruptcy-filings-surged-in-apf-14047705.html>.

Gepp, A. and K. Kumar. 2008. The Role of Survival Analysis in Financial Distress Prediction. *International Research Journal of Finance and Economics* 16 (2008)

"Global Industry Classification Standard (GICS®)." Standard & Poor's. August 2006

Hosmer JR., D. W. and S. Lemeshow. 1999. *Applied Survival Analysis: Regression Modeling of Time to Event Data*. New York: John Wiley & Sons.

Kleinbaum, D. G. and M. Klein. 2005. *Survival Analysis: A Self-Learning Text*. 2nd ed. New York: Springer.

LeClere, M. J. 2005. Time-Dependent and Time-Invariant Covariates within a Proportional Hazards Model: A Financial Distress Application. *Review of Accounting & Finance* 4 (4): 91.

Shumway, T. 2001. Forecasting Bankruptcy More Accurately: A Simple Hazard Model. *The Journal of Business* 74 (1): 101-124.

SAS Institute Inc., 2004. SAS/STAT ® 9.1 User's Guide, Volumes 1-7, Cary, NC: SAS Institute Inc., 2004.

Zmijewski, M.E. 1984. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. *Journal of Accounting Research* 22: 59-82.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Oki Kim
Statistical Analytics, LLC
E-mail: oki-kim@statistical-analytics.com
Web: http://www.statistical-analytics.com/