Application of Cox Proportional Hazard Model to the Stock Exchange Market

Jiayi Ni

Jiayi Ni is a graduate student at Ball State University majoring in Statistics and will graduate in May 2009. This article was completed under the supervision of Dr. Munni Begum as part of a course in Research Methods in Mathematics and Statistics offered in the Fall 2008.

Abstract

Survival analysis is widely used in mechanical research, engineering and many other fields. This paper introduces the properties and modeling methods for survival data, then fits a Cox Proportional Hazards Model for stock data in the Shanghai Security Market.

Introduction

By September 17th, the Shanghai stock exchange's benchmark index had plunged 64% since the start of the year, reaching a 52-week low and crashing past the 2000 points barrier to close at 1986.64. The market is thus worth less than a third of its value at its peak in mid-October 2007, when it reached 6395.76. This marks the most rapid decline of any major market, even in such an internationally gloomy year.

Though it is an overall crash of a stock market, differences still exist among individual stocks. Some experienced wild ups and downs in price, while others rapidly fell down, almost straight down to half of their highest prices. This paper aims to find out what the main factors are that influence price performances of quoted companies, and what kind of companies are more likely to survive this meltdown. Dismissing the macro factors, such as a change of the stamp tax on stock trading and macro economy regulation and control, my study focuses on the financial data of each individual stock.

The data come from the SSE 50 index of the Shanghai Security Market. The basics of survival analysis and the definition of survival time of stocks is are introduced first. This is followed by an introduction of theories of modeling survival data and the necessary foundations for Cox Proportional Hazards

Model. The estimation parameters using a Cox Proportional Hazards Model for the stock data in Shanghai Security Market are given later. The paper ends with a summary section.

Introduction of Survival Data and Definition of Stock Survival Time

The object of survival analysis is data in the form of times from a well-defined time origin to an end point. The end point could be the occurrence of some particular events or a particular time point. In medical research the time origin will often correspond to the recruitment of an individual in an experimental study, such as a clinical trial to compare two or more treatments. The outcome of interest is the duration until an event occurs; that is, an analysis of the time until an event occurs. Such events include the time to respond to treatment, relapse-free survival time, time to death, time to device failure, and time to regain mobility. More generally, survival times can also be observed in other application areas, such as the time taken by an individual to complete a task in a psychological experiment, the storage times of seeds held in a seed bank, or the lifetimes of industrial or electronic components.

One reason that survival data are not amenable to standard statistical procedures used in data analysis is that it generally is not symmetrically distributed. Typically, a histogram constructed from the survival times of a group of similar individuals will have a longer "tail" to the right of the interval that contains the largest number of observations. Also, survival time is positive while a normal distribution is defined on the entire real line. Thus, the normal distribution assumption is not valid. This difficulty could be resolved by first transforming the data to give a positive and more symmetric distribution, however, a more satisfactory approach is to adopt an alternative distributional model for the original data.

The main feature of survival data that renders standard methods inappropriate is that survival times are frequently censored. An individual survival time is said to be censored when the end-point of interest has not been observed. Censoring is classified into two types. In Type I censoring, the number of uncensored observations is a random variable, while in Type II censoring, the number of uncensored observations is fixed in advance.

In this study, I am interested in the survival time of stocks. The time origin is defined as the date when a stock's price reached its highest point in this year. The end-point is the date its price dropped to below 40% of that price for the first time. The number of days between these two dates is then the survival time of a stock. As an example, the study times of eight stocks are shown in Figure 1. The length of my study is 8 months, from Jan 1 to Aug 31 in 2008. The time of entry to study is represented by a dot. Stocks 1, 4, 5 and 8 died (D) during the study period, because their prices had dropped below 40% of their highest prices prior to the end of study. Stocks 3 and 6 were alive (A), which means their prices were still above 40% of their highest when the study period ended. Stocks 2 and 7 had stopped trading for some reason at the time points labeled "L". They were lost-to-follow-up. The survival times

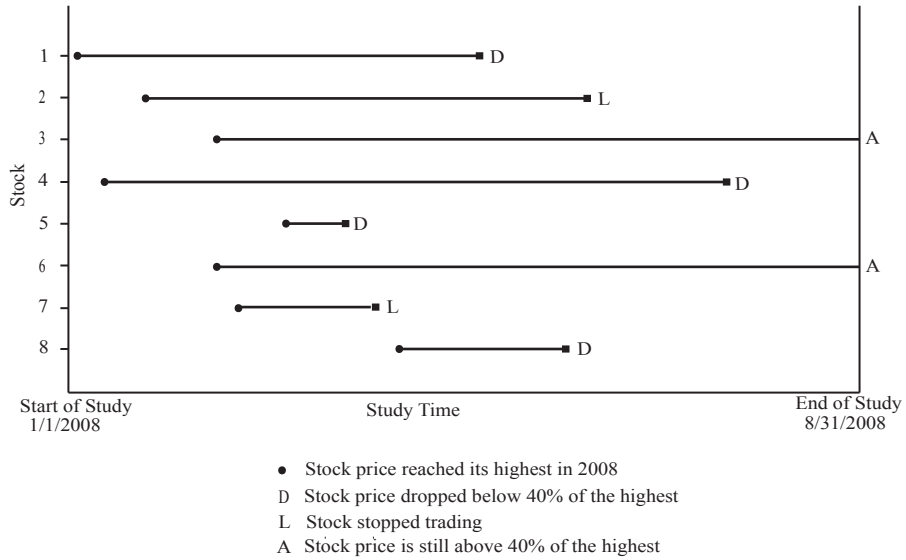

Figure 1: Study time for eight stocks

are recorded for stocks 1, 4, 5 and 8, while the survival times of the remaining stocks are censored (Type I).

Some Theory of Modeling Survival Data

Let T be the random variable representing the survival time of stocks. Then one way to describe the distribution of T is the hazard function, which is defined as follows:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta T | T > t)}{\Delta t}$$

This is the probability that the closing price of a stock drops below 40% of its highest price, at time t, conditioning on its having stayed above 40% of the starting price up to that time. In other words, it is the instantaneous rate of death of a stock at time t, given that it survives up to time t. Most of the time, the distribution of data is unknown. We may expect survival times to depend on the outcome of several explanatory variables; these may be collected together in vector form, X. Cox (1972) proposed the following model:

$$h_X(t) = h_0(t)e^{\beta^t X},$$

where X is the explanatory variables, β is the vector of unknown regression coefficients and $h_0(t)$ is the baseline hazard, which is an unknown function giving the hazard for the standard set of conditions, when the values of all the explanatory variables equal 0.

It is assumed that there are no ties in the n observed lifetimes at first. Suppose that k death times are uncensored and the remaining $n - k$ are censored. The Cox Proportional Hazards Method allows estimating β without any

knowledge of the baseline function, just by using rank of death and uncensored times. Therefore, if $t_{(1)}, t_{(2)}, \ldots, t_{(k)}$ are k ordered death times and R_j, the risk set, is the set of subjects which have survived until $t_{(j)-}$, immediately prior to the jth survival time, then the Cox conditional likelihood function of observed data is:

$$\begin{aligned} L_C(\beta) &= \prod_{j=1}^{k} \frac{P(i \text{ dies at time } t_{(j)} | i \text{ survives until } t_{(j)-})}{P(\text{a death in } R_j \text{ at time } t_{(j)})} \\ &= \prod_{j=1}^{k} \frac{e^{\beta^T X_{(}j)}}{\sum_{l \in R_j} e^{\beta^T X_l}} \end{aligned}$$

If there are d_j repetitions of an observed death time $t_{(j)}$, the ties will record possibly different covariate values, denoted as $X_{(j)1}, X_{(j)2}, \ldots X_{(j)d_j}$. Thus, the likelihood function simplifies to:

$$L_C(\beta) = \prod_{j=1}^{k} \frac{e^{\beta^T S_j}}{(\sum_{l \in R_j} e^{\beta^T X_l})^{d_j}},$$

where

$$S_j = \sum_{l=1}^{d_j} X_{(j)l}$$

Maximum likelihood estimates of β are then found by maximizing the logarithm of this equation using numerical methods.

Fit Cox Proportional Hazard Model for Stock Data

Data were collected from stocks in the SSE 50 Index. The index selects the 50 largest stocks of good liquidity and representativeness in the Shanghai security market by scientific and objective method. Referring to the semi-annual report of each company, 6 factors were considered at first, including earning per share (EPS), net asset per share (NAPS), cash flow per share (CFPS), return on equity (ROE), growth rate of operating profit (GROP), and the percentage of released non- floating shares (RNF) by the end of study. Also stocks are divided into 14 sectors by industry (Table 1). The Communication Device sector is selected as the reference for the sectors and one dummy variable is created for each of the other 13 sectors.

No.	Industry	No.	Industry
1	Finance	8	Automobile manufacturing
2	Metal machining	9	Machinery material
3	Electricity and heating power	10	Consumer products
4	Port traffic	11	Commerce
5	Petroleum and chemical industry	12	Coal industry
6	Mass media, internet and hardware	13	Medication
7	Real estate	14	Communication device

Table 1 *Industry Sectors of Stocks*

	EPS	NAPS	CFPS	ROE	GROP	RNF
EPS	1					
NAPS	0.61993	1				
	<0.0001					
CFPS	0.26422	0.30271	1			
	0.0666	0.0345				
ROE	0.76797	0.18287	0.04381	1		
	<0.001	0.2085	0.765			
GROP	0.40809	0.08774	-0.05762	0.50313	1	
	0.0036	0.5488	0.6941	0.0002		
RNF	0.01724	0.05994	-0.05831	0.02346	-0.0755	1
	0.9064	0.6825	0. 6907	0.8729	0.6061	

Table 2 *Pearson Correlation Coefficients and Prob* $> |r|$ *under* $H_0 : Rho = 0$

If two factors are highly correlated, then only one of them will be included in the model to reduce the multicollinearity among the factors. Table 2 shows the Pearson Correlation Coefficients matrix. EPS and NAPS, EPS and ROE are highly correlated, with P-values both less than 0.0001, but NTAPS and ROE are not. The Cox Proportional Hazards Model fit is improved after dropping EPS.

Total	Event	Censored	Percent Censored
50	36	14	28

Table 3 *Summary of the Number of Event and Censored Values*

Test	Chi-Square	DF	Pr > Chi-Sq
Likelihood Ratio	32.3178	18	0.0202
Score	36.6816	18	0.0058
Wald	27.3037	18	0.0735

Table 4 *Testing Global Null Hypothesis:* $BETA = 0$

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > Chi-Sq	Hazard Ratio
NAPS	1	0.27293	0.13080	4.3541	0.0369	1.314
CFPS	1	0.27323	0.11764	5.3948	0.0202	1.314
ROE	1	-0.01281	0.04108	0.0973	0.7551	0.987
GROP	1	0.00720	0.00387	3.4684	0.0626	1.007
RNF	1	-0.00862	0.01422	0.3678	0.5442	0.991
X1	1	-1.36957	1.29089	1.1256	0.2887	0.254
X2	1	-0.31155	1.13154	0.0758	0.7831	0.732
X3	1	0.55925	1.25178	0.1996	0.6550	1.749
X4	1	-1.71112	1.24645	1.8846	0.1698	0.181
X5	1	0.74952	1.29799	0.3334	0.5636	2.116
X6	1	0.86980	1.17420	0.5487	0.4588	2.386
X7	1	2.19749	1.34612	2.6649	0.1026	9.002
X8	1	1.30418	1.31235	0.9876	0.3203	3.685
X9	1	1.23638	1.19320	1.0737	0.3001	3.443
X10	1	-3.67439	1.71484	4.5912	0.0321	0.025
X11	1	-1.56261	1.45913	1.1469	0.2842	0.210
X12	1	-13.31468	1197	0.0001	0.9911	0.000
X13	1	-0.81313	1.49864	0.2944	0.5874	0.443

Table 5 *Analysis of Maximum Likelihood Estimates*

Among 50 stocks, 36 died and 14 were censored. The percent censored is 28%. Table 4 shows the Global Null Hypothesis: all β equal to 0. The P-value of the Likelihood Ratio test is less than 0.05, so the null hypothesis is rejected at the 5% level of significance. SAS phreg procedure gives out the Maximum Likelihood Estimates for β and tests the significance for each individual variable (Table 5). P-values of NAPS, CFPS and GROP are less than 0.1, so at a significance level of 0.1, the effects of these three factors are significant, while that of ROE and RNF are not. The dummy variables are included to exempt industry effects. So the Cox Model is fitted as shown below:

$$\begin{aligned}\text{Survival Time} = {} & 0.27293\text{NAPS} + 0.27323\text{CFPS} + 0.00720\text{GROP} - 1.36957X_1 \\ & - 0.31155X_2 + 0.55925X_3 - 1.71112X_4 + 0.74952X_5 \\ & + 0.86980X_6 + 2.19749X_7 + 1.30418X_8 + 1.23638X_9 \\ & - 3.67439X_{10} - 1.56261X_{11} - 13.31468X_{12} - 0.81313X_{13}\end{aligned}$$

SAS also calculates the hazard ratio of each parameter, which is the natural base e to the power of the parameter estimate, and is the relative hazard (the risk of death increased) corresponding to a unit change in the corresponding variable. For example, the hazard ratio of CFPS is $e^{0.27323} = 1.314$.

Notice that the standard error for X_{12} is very large, which means that this variable is not reliable and there is correlation between Sector 12 and other

sectors. One way to solve this problem is do reclassification. Combine this sector with another sector of similar industry characteristics, and then fit the model again. The modification will not be detailed in this paper.

Conclusion

This paper introduces some theories and modeling methods in survival analysis and applies the Cox Proportional Hazards Model to analyze stock survival times. The regression model tells us that release of more and more non-floating shares is not a main cause of the nose-diving in stock market, and at this time return on equity also does not affect the stock price a lot, while NAPS, CFPS and GROP are positively related to stock survival times. It is quoted companies' good financial condition, high liquidity and growth rate of earning capacity that make their stocks survive longer in the market.

References

[1] D. Collett, Modeling Survival in Medical Research, 2nd ed. CRC Press (2003).

[2] P. J. Smith, Analysis of Failure and Survival Data, Chapman & Hall/CRC (2002).

[3] M. Stepanova and L. Thomas. 2002. Survival Analysis Methods for Personal Loan Data, *Operations Research* **50** (2002) 277-289.

[4] Semiannual report of quoted companies, ⟨`http://www.zhicheng.com/dxf/200806.html`⟩.