8-2014

Analyses of 2002-2013 China's Stock Market Using the Shared Frailty Model

Chao Tang
East Tennessee State University

Analyses of 2002-2013 China's Stock Market Using the Shared Frailty Model

---

A thesis

presented to

the faculty of the Department of Mathematics and Statistics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Chao Tang

August 2014

Yali Liu, Ph.D., Chair

Robert Price Jr, Ph.D.

Edith Seier, Ph.D.

Keywords: correlated recurrent event, shared frailty model, stock market

ABSTRACT

Analyses of 2002-2013 China's Stock Market Using the Shared Frailty Model

by

Chao Tang

This thesis adopts a survival model to analyze China's stock market. The data used are the capitalization-weighted stock market index (CSI 300) and the 300 stocks for creating the index. We define the recurrent events using the daily return of the selected stocks and the index. A shared frailty model which incorporates the random effects is then used for analyses since the survival times of individual stocks are correlated. Maximization of penalized likelihood is presented to estimate the parameters in the model. The covariates are selected using the Akaike information criterion (AIC) and the variance inflation factor (VIF) to avoid multicollinearity. The result of analyses show that the general capital, total amount of a stock traded in a day, turnover rate and price book ratio are significant in the shared frailty model for daily stock data.

TABLE OF CONTENTS

ABSTRACT 2

LIST OF TABLES 5

LIST OF FIGURES 6

1 INTRODUCTION 7

2 UNDERGRADUATE RESEARCH 10

3 DATA DESCRIPTION 11

4 DEFINITION OF SURVIVAL TIME 13

4.1 Survival Time in the Undergraduate Research 13

4.2 Recurrent Event 13

5 SURVIVAL MODELS 16

5.1 Survival and Hazard Functions 16

5.2 Kaplan-Meier Model 17

5.3 Cox Proportional Hazard Model 17

5.4 Shared Frailty Model and Estimation 19

5.4.1 The Shared Frailty Model 19

5.4.2 Maximization of the Penalized Likelihood 20

5.4.3 Spline-Based Approximation 21

5.4.4 Smoothing Parameter 22

6 DATA ANALYSIS 23

7 CONCLUSIONS 34

8 FUTURE WORK 35

BIBLIOGRAPHY 36

APPENDICES 38
1 DATA PROCESS IN SAS 38
2 DATA ANALYSIS IN R 41
VITA 48

LIST OF TABLES

1 The industry categories and the number of the 219 stocks in each industry 11

2 Table of Recurrent Event of Aviation Stock 600115.SH 15

3 Table of Correlation of Covariates 25

4 Analyses Results for Possible Models 26

5 Significant Covariates in Possible Models 27

6 Covariates and its VIF (Excluded stock CC) 28

7 Estimation of Shared Frailty Model 29

8 Illustration Example of Stock 600151.SH and Stock 600660.SH 31

9 Estimated Probability of the Event of Stock 600151.SH and Stock 600660.SH 32

LIST OF FIGURES

1 Plot of Aviation Stock 600115.SH vs. CSI 300 14

2 Survival Function of Kaplan-Meier Model 23

3 Scatter Plot of Covariates Matrix 24

4 Estimated Baseline Hazard Function of the Shared Frailty Model . . 30

5 Plot of the Estimated Probability of the Event for Stock 600151.SH . 32

1 INTRODUCTION

Survival analysis is widely used in many different fields such as medical studies, biological studies, clinical studies, etc. It deals with the survival data which have one or more events. In recent years, economists have started applying this statistic method to the stock market, currency market, and macro and micro economics phenomena. Using a survival model, economists are able to study the behaviors of the derivate and market and finally provides investment advice. With the growth of interest in the stock market, a right investment decision is the priority for the investor. This thesis analyzes the stock market using a survival model and provides the investment suggestions.

In the stock market, the composite index of stock market is calculated based on the prices of the selected stocks. The stock market index provides an overall view of the market. Investors usually use the index as a tool to observe the stock market. Most of the time, investors not only observe the composite index, but also focus on the stocks with good performance. There are a variety of measurements to assess the stock's performance. The most popular one is the stock price. However, the price of stock is a measurement for an individual stock. The rising and falling of stock prices could not represent the relationships between the behaviors of the individual stocks and the stock market. Indeed, the behaviors of an individual stock are affected by the operating of a company and the stock market. Instead of using the stock prices as a measurement, comparing the daily return of an individual stock and the composite index interprets the performance of the stock well. Another aspect affecting the performance of a stock is the company's operation. Some covariates, such as the

price earning ratio, free cash flow and asset-liability ratio, are important covariates in an analysis. It is necessary to obtain the effects of the influence covariates for evaluating a company's operation. Within a proper time period, the behaviors of the stocks are therefore predictable using survival models and investment advice could be concluded.

Some remarkable works have been done. In 2006, Yuanlin used a survival model to analyze the trend of the Shenzhen stock market composite index under different transaction systems [9]. The different transaction systems have distinct restrictions on the behaviors of the composite index. The study of the composite index provides an overall view of the stock market for investors. To investigate the good performing stock in the market, investors use different measures in the analyses. Another application of survival analysis evaluated the commercial bank stocks using the ability of loan repaying as a measurement [8]. Using an appropriate measure, the behaviors of the stocks can be presented well. The covariates of the stocks are important in analysis, and they reveal the operation of company. The certain covariates may lead the company to going bankrupt, being acquired or going private [3].

The data used in this thesis were collected from Shanghai and Shenzhen Stock Exchanges. The CSI 300 (SHA:000300) and a total of 219 stocks from different industries are used as the stock data. The information of each selected stock was collected from the day it was listed. This thesis presents a survival analysis of the stock data.

The thesis is organized as follows. In section 2, a summary of undergraduate research related to the analysis of the stock market is presented. Section 3 provides

the details of the data. Section 4 defines the recurrent event and correlated survival times using the daily return of the selected stock and their market index, CSI 300. The shared frailty model and the estimation of parameters based on the penalized log-likelihood are described in section 5. The results of data analyses are presented in section 6. Finally, a discussion of conclusions and future work end the thesis.

2 UNDERGRADUATE RESEARCH

We have presented a preliminary survival analysis of China's stock market in 2011 as an undergraduate research project [2]. Using the data of 298 stocks which are traded in Shanghai and Shenzhen Stock Exchanges of China, the Kaplan-Meier (KM) and the Cox's proportional hazards models were adopted for analyses. The survival time we defined was the number of days between the date that the stock price reached its peak and the date that the stock has a loss of specific percentages from its peak in 2011. In the analysis, each stock experienced only one survival event, so the survival times are assumed to be independent in the Cox proportional hazards model. As a result, the profit, asset liability ratio, market capitalization, and industry category were significant covariates when the loss is 15% or 20% in the Cox proportional hazards model. In the current research, we compare the daily return of the individual stocks and CSI 300 to define the survival times. The Kaplan-Meier and the shared frailty model are used in the data analyses.

3 DATA DESCRIPTION

The CSI 300 (SHA:000300) data from January 4th 2002 to March 5th 2013 are used as the market index in this thesis. The index is complied by the China Securities Index Company, which reflects the price performance of China A share market [1]. The index includes 300 stocks traded in Shanghai and Shenzhen Stock Exchanges. In our data, there are a total of 219 stocks with 27 different industries. Table 1 presents the number of stocks in each industry. The metal mining industry has 42 stocks, the most among all industries.

Table 1: The industry categories and the number of the 219 stocks in each industry

Industry	Number of Stocks	Industry	Number of Stocks
Aviation	4	Auto Spare Part	4
Building Material	4	Capital Market	9
Chemistry	12	Commercial Bank	14
Comprehensive Finance	3	Computer	4
Construction	8	Drink	9
Electric	4	Family	6
Foodstuff	8	Highway and Railway	2
Insurance	3	Machinery	12
Media	6	Metal Mining	42
Multiple Retail	4	National Defense	4
Ocean Carriage	4	Oil Gas Coal	20
Pharmacy	9	Specialty Retail	5
Spinning Clothing	3	Traffic	14
Water Affairs	2		

A total of 10 covariates are considered in the stock data: turnover rate (TR), price-earning ratio (PER), price-book ratio (PBR), price-to-sales ratio (PSR), price-cash-flow ratio (PCFR), the total amount of a stock traded per day (AMO), A share or B share circulation market value (Stock CMV), aggregate market value (AMV), A share or B share circulation of capital (Stock CC), and general capital (GenC). The turnover rate (TR) represents the ratio of the number of times stock is sold associated with the stock held by company, as known inventory turnover ratio. The price-earning ratio (PER) is a valuation ratio corresponding to share price of a company and its earnings on each share. The price-book ratio (PBR) is a ratio of closing price of a stock and its book value from last quarter. The price-to-sales ratio (PSR) is calculated by dividing the market capitalization of a company by its overall sales in annual. The price-cash-flow ratio (PCFR) represents the ratio of a company's market capitalization and its annual cash flow. The stock circulation market (Stock CMV) value is the amount of a stock value in the market, and the aggregate market value (AMV) is the total value of the stock of a company. The stock circulation of capital (Stock CC) is the amount of capitalization in the market, and the general capital (GenC) is the total capitalization of a company.

Some of the covariates have very large values, i.e., general capital. To avoid different scales, we have standardized each covariate, so the coefficients of covariates will not have an extremely small value in analysis.

4 DEFINITION OF SURVIVAL TIME

4.1 Survival Time in the Undergraduate Research

In previous work [2], the stock data were collected from January 4th, 2011 to October 28th, 2011. The accumulative return was the measure to define survival time. The event occurs when a stock plunged certain percentages (-15%, -20% and -31.9% to be specific). Therefore, the survival time is the number of trading days from the date of its highest price of a stock to the date with a certain loss. Right censoring occurs when the stock price never dropped by the specified percentage. The event for individual stock can only occur once within the entire study period. Meanwhile, under the independence assumption, the cox proportional hazards model is an appropriate approach for the data.

4.2 Recurrent Event

In this thesis, the data started at the day when each stock was listed. In contrast to prior work [2], the daily return, which is calculated by dividing the current stock price by its price of the previous trading day, is the measure used to define the event. Considering CSI 300 data, the study period is from January 4th, 2002 to March 5th, 2013. During a time period, the survival time begins with the date of an individual stock with a daily return higher than CSI 300, and the end of this survival period is the date of its daily return lower than the index. The duration time of the individual stock with higher daily return than the index is the survival time for a single time period. In the entire study period, the event occurs more than once for an individual

stock and the survival times are correlated for each stock. This leads to correlated recurrent survival times. For the last recurrent event of an individual stock, the censoring occurs when the daily return of an individual stock is still higher than CSI 300 at the last day of the study period (March 5th, 2013).

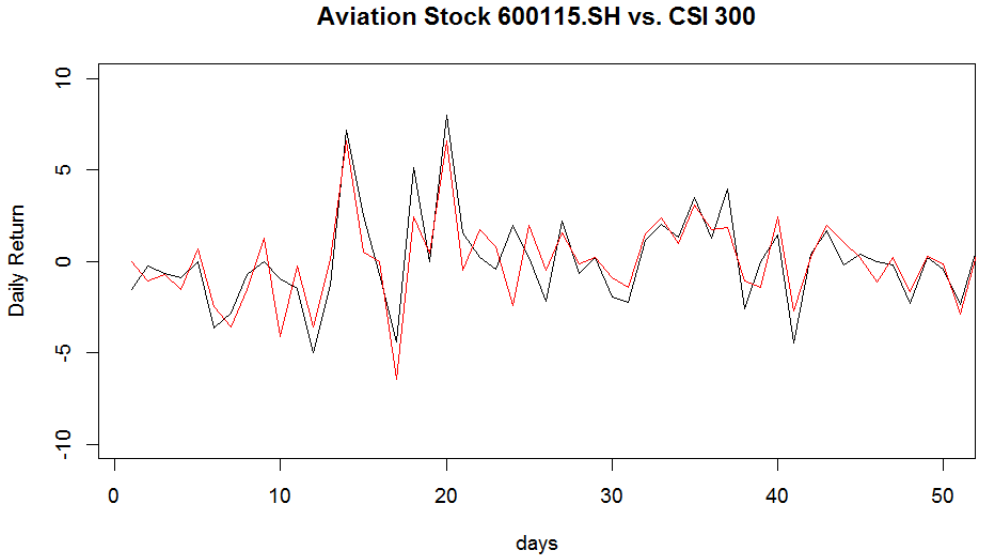

Figure 1: Plot of Aviation Stock 600115.SH vs. CSI 300

To illustrate the construction of the recurrent survival data, Figure 1 is produced using the performance of a stock (600115.SH) in aviation industry and CSI 300 during a short time period. The black line represents the daily return of stock (600115.SH) and red line represents the daily return of the index. When the black line is above the red line, the stock's daily return is higher than the index. The number of days under this condition is defined as one of the survival times. Corresponding to Figure 1, Table 2 provides a brief example of the list of recurrent event with survival times.

Table 2: Table of Recurrent Event of Aviation Stock 600115.SH

Date begin	Time begin	Time end	Survival time (days)
07JAN2002	2	5	3
14JAN2002	7	9	2
17JAN2002	10	11	1
23JAN2002	14	16	2
28JAN2002	17	19	2

The aviation stock (600115.SH) has 5 survival time periods in the first 19 days of the study. The survival times are 3, 2, 1, 2 and 2 days respectively. The first survival period started at January 7th, 2002, which is the second day in the entire study period, and it lasted for three trading days. Therefore, the survival time of this period is three. Similar interpretation is suitable for the other four survival periods listed in Table 2.

A total of 219 stocks in 27 industries are involved in this study. Because of the occurrence of censoring and correlated recurrent survival times of an individual stock, the shared frailty model is appropriate and will be adopted for data analysis. In section 5, we discuss the survival models and the estimation methods used in this thesis.

5 SURVIVAL MODELS

The survival distribution can be presented through the survival and hazard functions. In this section, we first review the basic survival concepts, the Kaplan-Meier (KM) model [7] and the Cox proportional survival model [7].

5.1 Survival and Hazard Functions

This thesis only considers right censored data. Right censoring occurs when a subject drops off before an event occurs, or the study ends before the event has occurred. Let T_i denote a positive random variable which represents the real survival time with probability density function (*p.d.f*), $f(t)$, and cumulative distribution function (*c.d.f*), $F(t) = P(T \leq t)$. Then, the survival function at time t is

$$S(t) = P(T > t) = \int_t^\infty f(u)\,\mathrm{d}u = 1 - F(t) \tag{1}$$

which gives the probability that the event of interest has not occurred by time t. The hazard function of T is

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t < T \leq t + dt | T > t)}{dt}. \tag{2}$$

It represents the instantaneous failure rate at time t. The relationship among the hazard function, density function, and survival function is

$$\lambda(t) = \frac{1}{S(t)} \lim_{dt \to 0} \frac{1}{dt} \int_t^{t+dt} f(u)\,\mathrm{d}u = \frac{f(t)}{S(t)}. \tag{3}$$

It follows that

$$S(t) = \exp\left\{ - \int_0^t \lambda(u)\,\mathrm{d}u \right\} = \exp(-\Lambda(t)) \tag{4}$$

and

$$f(t) = \lambda(t) \exp(-\Lambda(t)) \tag{5}$$

where $\Lambda(t) = \int_0^t \lambda(u)\,\mathrm{d}u$ is the cumulative hazard function. The results show that the density, survival, and hazard function provide the distribution of T alternatively and equivalently.

5.2 Kaplan-Meier Model

The Kaplan-Meier model [7] is a non-parametric method and it provides an overall view of the data. Assume the ordered survival times are $t_1 < t_2 < ... < t_i < ... < t_r$ where r represent the number of the distinct failure times. Let n_j be the number of individuals at risk and d_j be the number of death at time t_j. If the censoring observations occur at the same time t_j as the failures occur, then n_j includes the censored observations. For $t_i \leq t < t_{i+1}$, the Kaplan-Meier estimator is

$$\hat{S}(t) = \prod_{j=1}^{i} \frac{n_j - d_j}{n_j} \tag{6}$$

The Kaplan-Meier model estimates an overall survival function $S(t)$ with product of conditional probabilities. It incorporates both censored and uncensored individuals, however, it does not include any covariates in the model.

5.3 Cox Proportional Hazard Model

The Cox proportional hazard model [7] is a common used model for right censored data with covariates in survival analysis. The model specifies the hazard function for the i^{th} individual as

$$\lambda_i(t) = \lambda_0(t) \exp\{\beta' X_i\} \tag{7}$$

where X_i is the covariates vector, and β is a vector of regression parameters associated with the covariates. The baseline hazard function $\lambda_0(t)$ represents the risk when all covariates are zero. The Cox model is a semi-parametric method because there is no assumption for the distribution of baseline survival time, i.e., it allows that the baseline hazard function $\lambda_0(t)$ takes any positive form. The hazards ratio for the i^{th} and the j^{th} individuals is

$$HR_{i,j} = \frac{\lambda_i(t)}{\lambda_j(t)} = \exp\{\beta'(X_i - X_i)\} \tag{8}$$

The likelihood function of the Cox proportional hazard function is used to estimate the parameters, β, in the model. Let C_i be the right-censoring times. Define $Y_i = \min(T_i, C_i)$ which is the observed survival data. The censoring indicator is $\delta_i = I_{(T_i \leq C_i)}$. The likelihood function is

$$L = \prod_{i=1}^{n} \left\{ \lambda_0(t_i) e^{\beta' X_i} \right\}^{\delta_i} \exp\left\{ - \int_0^{t_i} \lambda_0(u) e^{\beta' X_i} \, \mathrm{d}u \right\}. \tag{9}$$

However, the baseline hazard function $\lambda_o(t)$ is hard to be estimated. The partial likelihood function is therefore proposed for estimation. The partial likelihood function is

$$PL(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\{\beta' X_i\}}{\sum_{l \in R(t_i)} \exp\{\beta' X_l\}} \right]^{\delta_i}, \tag{10}$$

where $R(t_i) = \{j : t_j \geq t_i\}$ is the risk set at t_i, i.e., the set of subjects who have not had an event by time t_i. In the Cox proportion hazards model, the parameter, β, is estimated using the Newton-Raphson procedure [6]. The iteration of Newton-Raphson step continues until convergence is obtained. In software R, the Breslow method is used to estimate the baseline hazard function, $\lambda_0(t)$. Between the distinct

failure times, the baseline hazard function is assumed to be the same. The estimator is

$$\tilde{\lambda}_0(t) = \frac{d_i}{(t_{i+1} - t_i) \sum_{l \in R(t_i)} \exp\left\{\hat{\beta}' X_l\right\}}, \tag{11}$$

where $t_i < t < t_{i+1}$.

5.4 Shared Frailty Model and Estimation

In resent years, the shared frailty model is commonly used to deal with recurrent event times [6]. The model considers an unobserved random effect within the subject and right-censored data in the recurrent event. In this thesis, we will apply this model to the stock data.

5.4.1 The Shared Frailty Model

For the j^{th}(j = 1, ..., n_i) individual of the i^{th} group, define T_{ij} as the recurrent event times and C_{ij} as the right-censoring times in the study. The observed survival time, Y_{ij}, is $\min(T_{ij}, C_{ij})$ and the censoring indicators are $\delta_{ij} = I_{(Y_{ij}=T_{ij})}$. The hazard function for a shared frailty model can be written as

$$\lambda_{ij}(t) = v_i \lambda_0(t) \exp(\beta X_{ij}) \tag{12}$$

where $\lambda_0(t)$ is the baseline hazard function, β is the vector of regression parameters, X_{ij} denotes the covariate vector, and v_i is the random effect associated with the i^{th} group. In addition, the $v_i's$ have an independent and identical distribution. A common distribution for v_i is a gamma distribution with $\mathbf{E}(v_i) = 1$ and $\mathbf{Var}(v_i) = \theta$,

i.e., $v_i \sim \Gamma(\frac{1}{\theta}, \frac{1}{\theta})$ [7]. The marginal log-likelihood function is

$$\begin{aligned} l(\Phi) &= \sum_{i=1}^{G} \left\{ \left[\sum_{j=1}^{n_i} \delta_{ij} \ln \lambda_{ij}(Y_{ij}) \right] - (\frac{1}{\theta} + m_i) \ln \left[1 + \theta \sum_{j=1}^{n_i} \Lambda_{ij}(Y_{ij}) \right] \right. \\ &\quad \left. + I_{\{m_i \neq 0\}} \sum_{k=1}^{m_i} \ln(1 + \theta(m_i - k)) \right\}, \end{aligned} \tag{13}$$

where $m_i = \sum_{j=1}^{n_i} I_{(\delta_{ij}=1)}$ which denotes the number of recurrent events, Λ_{ij} is the cumulative hazards function for the j^{th} individual in the i^{th} group, and $\Phi = (\lambda_0(\cdot), \beta, \theta)^T$.

5.4.2 Maximization of the Penalized Likelihood

We now introduce the steps of estimation methods of the shared frailty model. First, the maximization of penalized likelihood function estimates the parameters in the model. Then, the spline-based method approximates the baseline hazard function $\lambda_0(t)$. Last, the smoothing parameter used in the model is estimated.

The parameters, β, are estimated using the maximization of the penalized log-likelihood. The penalized log-likelihood function for the shared frailty model is

$$pl(\Phi) = l(\Phi) - \kappa \int_0^\infty \lambda_0''(t)^2 \, \mathrm{d}t \tag{14}$$

where $l(\Phi)$ and $\lambda_0(t)$ were defined in section 5.4.1 and κ is a positive smoothing parameter associated with the smoothness of the functions. The term, $\lambda_0''(t)$, is the second derivate of the baseline hazard function and it is approximated by a sum of polynomial functions of first order. A flexible shape of the hazard function is allowed in this approximation.

The maximization of the penalized likelihood estimates the parameter using the robust Marquardt algorithm. This algorithm combine the Newton-Raphson algorithm

and the steepest descent algorithm [4]. By using a squared transformation, the variance of frailties and the spline coefficients are restricted to be positive. This assures that the hazard functions are positive at all stage of the algorithm. The iteration is

$$\Phi^{(r+1)} = \Phi^{(r)} - \zeta(\widetilde{H}^{(r)})^{-1}\Delta(pl(\Phi^{(r)})) \tag{15}$$

The ζ equals to 1 by default, but it can be modified for improving the likelihood at each iteration. The $\widetilde{H}$ denotes a diagonal-inflated Hessian matrix(a square matrix of second-order partial derivatives of a function). The penalized log-likelihood gradient at the r^{th} iteration is explained by the term $\Delta(pl(\Phi^{(r)}))$. When the difference between two consecutive log-likelihoods is smaller than 10^{-4}, the iteration will stop. In this case, the coefficients are stable and the gradient is smaller than 10^{-6}. The inverse of the final Hessian matrix provides the standard errors of the estimates.

5.4.3 Spline-Based Approximation

After estimating the covariate coefficients in the shared frailty model, we approximate the baseline hazard function $\lambda_0(t)$. Notes it may not have an analytical solution. Using the spline method, the baseline hazard function can be approximated, i.e.,

$$\widetilde{\lambda}_0 = \sum_{i=1}^{m} \eta_i M_i \tag{16}$$

where $m = Q + 2$, Q is the number of knots, η_i's are the control points and M_i represents cubic M-splines [5]. The spline-based approximation allows arbitrary hazard function. The number of knots is suggested to be between 4 and 20. In the approximation, the more knots are used, the closer to the true hazard function.

5.4.4 Smoothing Parameter

The maximization of a likelihood cross-validation criterion is used for estimating the smoothing parameter, κ [4]. The parameter is obtained by minimizing the function

$$\bar{V}(\kappa) = \frac{1}{Q}[tr(\hat{H}_{pl}^{-1}\hat{H}_l - pl(\hat{\Phi}_\kappa))]$$

where Q is the number of knots, $\hat{H}_{pl}$ is the converged Hessian matrix of the penalized log-likelihood, $\hat{H}_l$ is the converged Hessian matrix of the log-likelihood, and $pl(\hat{\Phi}_\kappa)$ represents the maximum penalized log-likelihood at the final.

6 DATA ANALYSIS

Statistical package software, R, is used to perform the analyses. Considering an overall survival function (without any covariate involved), the Kaplan-Meier model displays an overall view of the stock data.

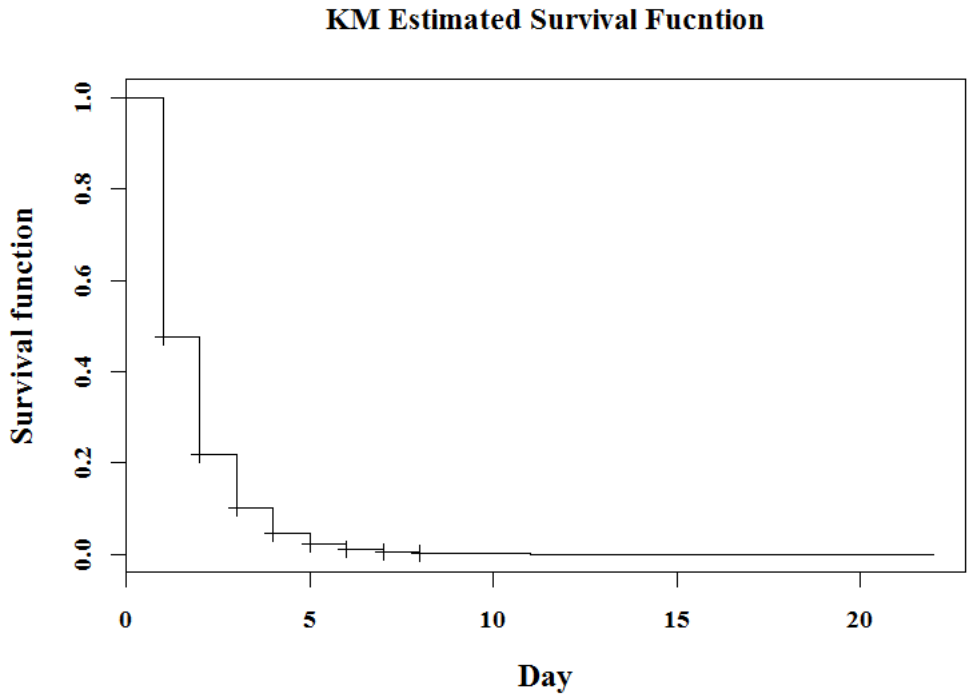

Figure 2: Survival Function of Kaplan-Meier Model

Figure 2 presents the results from the Kaplan-Meier estimation. Regardless of correlation among survival times of each stock, the survival rate drops rapidly in 5 days, which means the hazard rates increase fast when the stock has survived a week or longer. There are only few stocks whose survival times sustain longer than 10 days. The Kaplan-Meier model only provides an overall view of the data, however, the ef-

fects of covariates provides reliable information for a company's operation. Therefore, it is necessary to analyze the data using a model which incorporates the covariates.

Before analyzing the stock data with the covariates, the multicollinearity of these covariates is necessarily to be checked. Figure 3 is the scatter plot of the covariates.

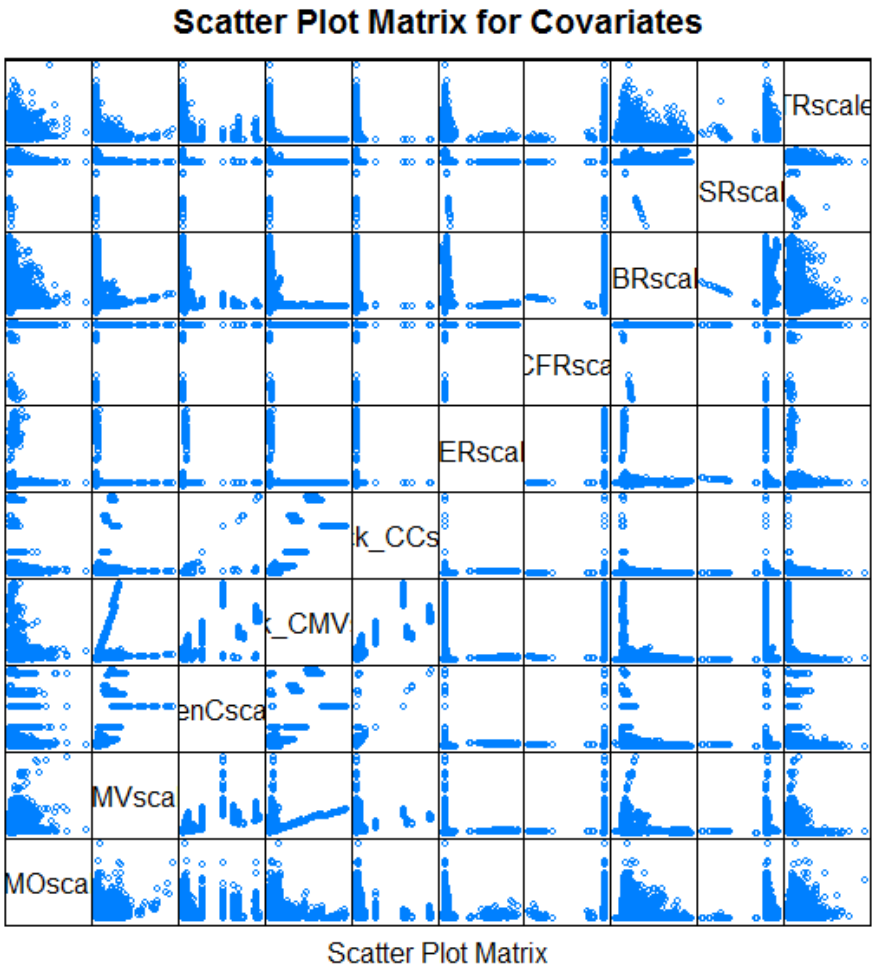

Figure 3: Scatter Plot of Covariates Matrix

Table 3: Table of Correlation of Covariates

Variable	AMO	AMV	GenC	StockCMV	StockCC	PER	PCFR	PBR	PSR	TR
AMO	1.000									
AMV	0.329	1.000								
GenC	0.177	0.815	1.000							
StockCMV	0.206	0.569	0.581	1.000						
StockCC	0.072	0.496	0.695	0.862	1.000					
PER	0.033	-0.001	-0.005	-0.005	-0.005	1.000				
PCFR	-0.018	0.0002	0.003	-0.007	0.002	0.002	1.000			
PBR	0.210	0.0194	-0.059	-0.013	-0.065	0.029	-0.022	1.000		
PSR	0.091	0.0235	-0.001	0.0003	-0.017	-0.023	0.005	0.231	1.000	
TR	0.377	-0.030	-0.046	-0.096	-0.084	0.015	0.007	0.150	0.033	1.000

Table 3 provides the correlations between each pair of the covariates. According to Table 3, there are three pairs of highly correlated covariates, which are the stock circulation market value (Stock CMV) and share circulation of capital (Stock CC), the aggregate market value (AMV) and general capital (GenC), and the stock circulation of capital (Stock CC) and general capital (GenC).

We assume there is an unobserved random effect within each stock. The shared frailty model is an appropriate approach to estimate the random effect and its variance. Using the package, *frailtypack* and *usdm*, in R, several analyses are performed. We start to identify the multicollinearity by using the Akaike information criterion (AIC) and the variance inflation factor (VIF) of several possible models. Table 4

Table 4: Analyses Results for Possible Models

Covariates Included	log-likelihood	AIC Value	R^2	VIF value
GenC, AMV and StockCMV	-74199.2	148416.4 (*)	0.4094	1.2014 (*)
GenC and StockCMV	-74200.38	148416.76	0.5883	1.5293
GenC and StockCC	-74200.26	148416.52	0.5758	1.4959
StockCMV	-74204.93	148423.86	0.9012	5.3218

presents the log-likelihoods, the AIC and the VIF values when different covariates are included in the models. Other covariates included in the model are the turnover rate (TR), price-earning ratio (PER), price-book ratio (PBR), price-to-sales ratio (PSR), price-cash-flow ratio (PCFR), and the total amount of a stock traded per day(AMO). The VIF measures the increase of the variance of an estimated model (coefficient) due to the collinearity. As a common rule, the survival model with the VIF value higher than 5 has the multicollinearity. The minimum of AIC and the

smallest VIF indicate the same model, which has the general capital (GenC), aggregate market value (AMV) and stock circulation market value (Stock CMV) and all others as the covariates. Meanwhile, all possible models identify the same significant covariates.

Table 5: Significant Covariates in Possible Models

Models Include	GenC, AMV and StockCMV (*)	GenC and StockCMV	GenC and StockCC	StockCMV
GenC	-0.0258	-0.0167	-0.0190	NA
AMO	0.0480	0.0498	0.0499	0.0489
TR	0.0308	0.0302	0.0303	0.0305
PBR	-0.0134	-0.0130	-0.0130	-0.0124
AMV	0.0116	NA	NA	NA
StockCMV	-0.0008	-0.0011	NA	-0.0057
PER	-0.0001	-0.0001	-0.0002	-0.0002
PSR	0.0052	0.0053	0.0053	0.0052
PCFR	-0.0025	-0.0025	-0.0026	
StockCC	NA	NA	0.0024	-0.0025

Table 5 gives a summary of the coefficients of covariates in all possible models. All covariate coefficients remain the same sign and their values are similar in the different models. In addition, the VIF values presented in Table 6 are smaller than 5 for the covariates in the optimal model after excluding the stock circulation capital (stock CC).

The AIC and VIF values designate the same covariate, stock circulation capital (stock CC), which should be excluded. Therefore, the shared frailty model (*) in

Table 6: Covariates and its VIF (Excluded stock CC)

Covariates	VIF
AMO	1.470366
AMV	4.000649
GenC	3.645724
StockCMV	1.627245
PER	1.005562
PCFR	1.001270
PBR	1.105222
PSR	1.044550
TR	1.234149

Table 4, including 9 covariates, the turnover rate (TR), price-earning ratio (PER), price-book ratio (PBR), price-to-sales ratio (PSR), price-cash-flow ratio (PCFR), the total amount of a stock traded per day (AMO), general capital (GenC), aggregate market value (AMV) and stock circulation market value (Stock CMV), is adopted for further analysis.

The stock data are clustered by individual stock. The shared frailty model with the random effect corresponding to each stock involves all covariates except the ones with multicollinearity.

Table 7 presents the results of estimation from model (*) in Table 4. The estimation of the variance of the random effect, θ, is 0.00626 with a standard error of 0.0008. In the table, "SE coef (H)" is the standard error estimated by inverting the Hessian matrix, "SE coef (HIH)" is the standard error estimated using the matrix product $H^{-1}IH^{-1}$ where H^{-1} is the inverse of the Hessian matrix and I the Fisher Information matrix, and the Z value refers to the Wald test statistic. The random

Table 7: Estimation of Shared Frailty Model

	coef	exp(coef)	SE coef (H)	SE coef (HIH)	z	p
GenC	-0.0258	0.9744	0.00819	0.0082	-3.1591	0.0016(*)
AMV	0.0116	1.0116	0.00751	0.0075	1.5437	0.1227
StockCMV	-0.0008	0.9991	0.00403	0.0041	-0.2010	0.8407
AMO	0.0481	1.0492	0.00372	0.0037	12.9161	0.0000(*)
TR	0.0308	1.0313	0.00331	0.0033	9.3145	0.0000(*)
PER	-0.0001	0.9999	0.00297	0.0030	-0.0435	0.9653
PBR	-0.0134	0.9867	0.00343	0.0034	-3.9028	0.0001(*)
PSR	0.0052	1.0052	0.00354	0.0035	1.4616	0.1438
PCFR	-0.0025	0.9975	0.00306	0.0030	-0.8174	0.4137

effect, an important parameter, is needed to be verified. If the random effect is not significant, the model will become the Cox proportional hazards model. This can be done by testing the hypothesis that $\theta = 0$ using a modified Wald test for unobserved heterogeneity. The test statistic is $W(\theta) = 0.00626/0.0008 = 7.825 > 1.64$, where 1.64 is the critical value for a normal one-sided test. Therefore, heterogeneity occurs in the data, which indicates that the random effect, v_i, is necessary in the shared frailty model. Results in Table 7 show that there are four significant(at significant level $\alpha = 0.05$) covariates: the general capital (GenC), total amount of a stock traded per day (AMO), turnover rate (TR) and price book ratio (PBR). The general capital and price book ratio have negative effects in the hazard function, whereas the total amount of a stock traded per day or the turnover rate have positive effects. A stock with higher value of the general capital and price book ratio has less hazards in survival period, when other covariates are fixed. And the stock with higher value of the

total amount traded per day or higher turnover rate has more hazards.

Figure 4 displays the baseline hazard function of the shared frailty model when using the 9 covariates in model (*) in Table 4. The two dash lines are the confidence bands of the baseline survival function at significant level, $\alpha = 0.05$.

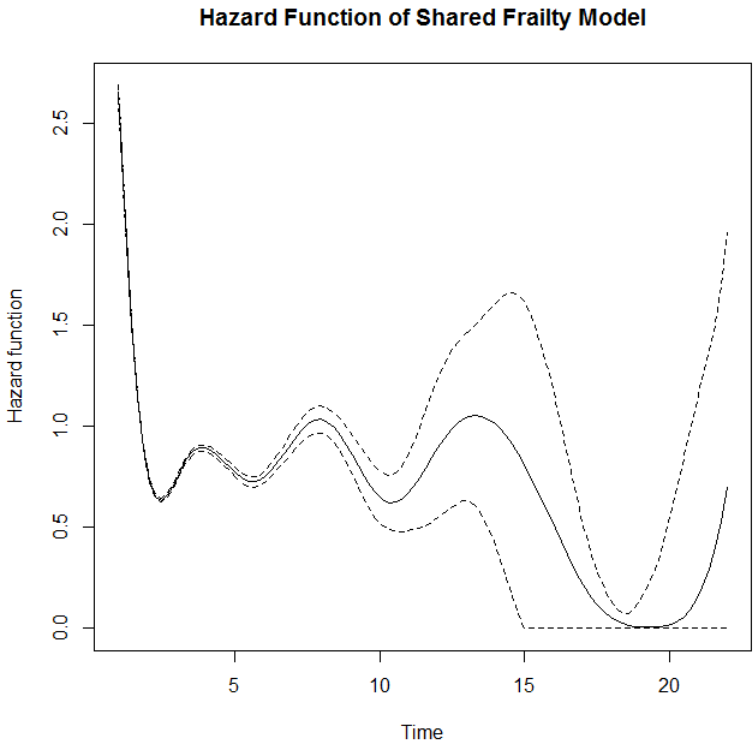

Figure 4: Estimated Baseline Hazard Function of the Shared Frailty Model

To compare the performance of the stocks in the market, we obtain their hazard functions. Note that the hazard function is $\lambda_{ij}(t) = v_i \lambda_0(t) \exp(\beta X_{ij})$. In the hazard function, the variance of the random effect, v_i, is very small (0.00626) and the baseline hazard function, $\lambda_0(t)$, is the same for each stock. The difference in the hazard functions also depends on the difference in the vector of covariates, X_{ij}. For two

stocks, we obtain the multiplications of the covariate values and their coefficients. Note we are using the hazard functions of the stocks. The better performing stock has a smaller value of $\beta^T X_{ij}$ compared to the other.

Table 8: Illustration Example of Stock 600151.SH and Stock 600660.SH

December 31, 2012	Coefficient	Covariate of 600151.SH (Standardized Value)	Covariate of 600660.SH (Standardized Value)
GenC	-0.0258	1,250,179,897 (0.446)	2,002,986,332 (1.773)
AMO	0.0481	32,677,508 (-0.496)	111,857,192 (0.078)
TR	0.0308	0.6423 (-0.489)	0.652 (-0.485)
PBR	-0.0134	2.5371 (-0.980)	2.8189 (-0.867)
AMV	0.0116	6,675,960,650 (-0.483)	17,566,190,132 (0.963)
StockCMV	-0.0008	5,112,915,058 (-0.258)	17,217,227,885 (1.638)
PER	-0.0001	451.3379 (3.597)	11.6132 (-0.623)
PSR	0.0052	2.7966 (-0.549)	1.8129 (-1.045)
PCFR	-0.0025	2356 (-0.063)	4,306 (1.238)
$\beta^T X$		-0.305	-0.067

Table 8 presents the values of the covariates of Stock 600151.SH and Stock 600660.SH at December 31, 2012. Based on the hazard function, the hazard ratio (ratio of two hazards function) of two auto spare part stocks can be obtained for prediction, $\exp\left[(-0.305) - (-0.067)\right] = 0.788$. With the hazard ratio less than 1, the stock 600151.SH is more likely to survival after the next trading day, January 4, 2013.

Last, a prediction based on the model (*) in Table 4 is preferred for completing the analyses. The "Frailtypack" package in R can estimate the probability of the event of a stock at different time point (days). Table 9 and Figure 5 provide an example of the estimated probability of the event of two auto spare part stocks at different days.

Table 9: Estimated Probability of the Event of Stock 600151.SH and Stock 600660.SH

Survival Days	Probability of 600151.SH	Probability of 600660.SH
2	0.7555215	0.7561427
3	0.8708815	0.8713570
4	0.9420166	0.9423127
5	0.9736136	0.9737851
6	0.9867438	0.9868460
7	0.9938531	0.9939087
8	0.9975533	0.9975794
9	0.9989964	0.9990087
10	0.9994957	0.9995024

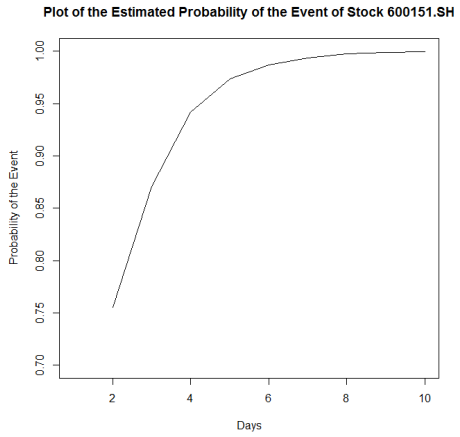

Figure 5: Plot of the Estimated Probability of the Event for Stock 600151.SH

The covariate values of these two stocks (600151.SH and 600660.SH) are taken as they were at the last event day, December 31, 2012 shown in Table 8. The standardized values of these covariates are used in the prediction. The estimated probabilities of the event of the stock 600151.SH are generally smaller than the probabilities of the stock 600660.SH at different survival days. The hazards ratio and the estimated probability both demonstrate that the stock 600151.SH performs better than the stock 600660.SH. Note that the two selected stocks are from the same industry category. This could explain why the estimated probabilities in Table 9 are so close. If we select 2 stocks from different industry categories, the difference might be obvious. Figure 5 illustrates the trend of the estimated probability of the event of the stock 600151.SH. We are interested in the probability of the event after the stock survived 2 days or more. As illustrated, the probabilities are close to 1 after the stock survived 6 days, which means the the stock is not likely to perform better than CSI 300 more than 6 days. For short term investment in stock market, it might be a good time to sell the stock when it has performed better than the index for one week.

7 CONCLUSIONS

This thesis is an application of correlated survival analysis in stock market. The data used are the capitalization-weighted stock market index (CSI 300, stock ticker symbol SHA:000300) and the 300 stocks for creating the index. In order to study the performance of the stocks, the daily return is used as a measure for comparing the selected stocks and CSI 300. The survival times are the duration of time when the daily return of an individual stock is higher than the index. The shared frailty model associated with a gamma distribution of random effect is adopted for analysis. The covariates are selected using the Akaike information criterion (AIC) and variance inflation factor (VIF) to avoid the multicollinearity. The results show that the random effect, general capital (GenC), total amount of a stock traded per day (AMO), turnover rate (TR) and price book ratio (PBR) are significant at level 0.05. The good performing stock has a smaller value of $\beta^T X_{ij}$ compared to others. The probability, that the daily returns of the stock turn out to be lower than the return of the CSI 300, can be predicted and some investment advice can be provided.

8 FUTURE WORK

More work will be done in the future. Based on the package, "frailtypack", the diagnosis of model does not have a residual analysis, but only provids the diagnosis of the random effect in the model. Test of goodness-of-fit is needed to assess the model.

In this thesis, we use daily return in analyses. Weekly data or even monthly data might be used to compare the difference among models and the significance of covariates. Lastly, the categories of stock can be considered as a sub-cluster which is appropriate in the nest frailty model.

BIBLIOGRAPHY

[1] China Securities Index CO., LTD, CSI300 Index Methodology. Accessed Apr 3, 2014. `http://www.csindex.com.cn`.

[2] C. Deng, C. Tang, and Q. Tang. (2011). Survival Analysis of 2011 Chinas Stock Market. Undergraduate Research. East Tennessee State University, Unpublished manuscript.

[3] Q. He, T.T. Chong, L. Li, and J. Zhang. (2010). A Competing Risks Analysis of Corporate Survival. Financial Management, **39**, 1697–1718.

[4] D.W. Marquardt. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. Journal of Society for Industrial and Applied Mathematics, **11**, 431–441.

[5] J.O. Ramsay. (1988). Monotone Regression Splines in Action. Statistical Science, **3**, 425-461.

[6] V. Rondeau, D. Commenges, and P. Joly. (2003). Maximum Penalized Likelihood Estimation in a Gamma-Frailty Model. Lifetime Data Analysis, **9**, 139–153.

[7] V. Rondeau, Y. Mazroui, and J.R. Gonzalez. (2012). Frailtypack: An R Package for the Analysis of Correlated Survival Data with Frailty Models Using Penalized Likelihood Estimation or Parametrical Estimation. Journal of Statistical Software, **47**, 1-28.

[8] X. Wang. (2008). Research on Financial Credit Risk Assessment of Stated-owned Commercial Banks Based on the Cox Model of Survival Analysis. Master's thesis. Southeast University, Nanjing, China.

[9] Y. Wang. (2007). Duration Analysis on Shenzhen Stock Market Component Index Based on Survival Model. Shandong Ecomomy, **132**, 79–81.

APPENDICES

1 DATA PROCESS IN SAS

In data process, the codes are used in SAS version 9.4 (July 10th, 2013). The macro program in the codes splits the individual stocks from the original data and merge with the index data.

```
%macro dataProcess(filename, a0);
proc import out= a1;
            datafile= ‘‘Z:\&filename..xlsx"
            dbms=EXCEL REPLACE;
     range=‘‘ &filename.$";
     gentnames=YES;
     mixed=NO;
     scantext=YES;
     usedate=YES;
     scantime=YES; run;
proc sort data=a1; by date; run;
data a1; set a1;
      rename Turnover = TurnoverStock;run;
data new;
     merge a1 &a0;by date; run;
data new; set new;
```

```
     If date = .  then delete;
 If OP = .  then delete;run;
proc freq data = new;
      tables code  / out= CodeTable norow nocum nopercent; run;
proc sort data=codeTable; by code; run;
proc sql;
    select code from codetable;
    select code into : code1 - :code &sqlobs
    from codetable; quit;
%let N = &sqlobs;
%do i = 1 %to &N;
%put &&code&i; %end;
%do i=1 %to &N;
      data &filename&i; set new;
         where code =‘‘& &code&i";run; %end;
%do i=1 %to &N;
proc export data= &filename&i
            outfile= ‘‘ Z: \&filename&i..csv"
            dbms=CSV REPLACE;
     putnames=YES; run;
%end; quit;
data codetable; set codetable;
stock =‘‘&filename";
```

```
index = _N_;
drop PERCENT; run;
proc append base = StockCode data = codetable force; run;
%mend dataProcess;
data Stockcode; Set Stockcode;
     If Index = .  then delete; run;
/*export the stockcode file*/
proc export data=  Stockcode
            outfile= `` Stockcode.csv"
            dbms=CSV REPLACE;
     putnames=YES; run;
```

2 DATA ANALYSIS IN R

In data analysis, the codes are used in R version 3.0.3. The "frailtypack" is used for modeling the data. The required packages are "reshape2", "splines", "survival", "survC1", "MASS", "boot", "frailtypack" and "lattice". All given comments are preceded by the # symbol.

```
#Read stockcode and combine stock‘s names
   file0 = "Z:\ Stockcode.csv"
   stockcode = read.csv(file0,head=TRUE,sep=‘’,")
   stockcat=stockcode[,1]
   index=stockcode[,3]
   stockname = paste(stockcat, index, sep=‘’ ")
#Read stock files
   filename = paste("Z:\ ‘’, stockname, ".csv", sep =" ")
   filenameN = length(filename)
#Loop for analysis
   #f <- function(i){
   stockdata <- data.frame()
   for (i in 1: filenameN){
   #for (i in 4: 5){
   #i = 4
      stock <- read.csv(filename[i], header=T, sep=",")
```

```
#define index closing price:
     #stock = stock[stock$Weekday == "F", ]
     C<- 0
     stock = subset(stock, stock$Closing_Price >C)
     CPindex = stock$Closing_Price
#calc IndexPCR
     n = length(CPindex)
     IndexPCR = (CPindex[-1]/CPindex[-n]-1)*100
     stock$IndexPCR = c(0,IndexPCR)
#Create Count of Date
     stock$nDate = (1:(length(stock$Date)))
#Record beat or not
     stock$Beat = as.numeric(stock$PCR > stock$IndexPCR)
     stock$Beat1 = c(NA, stock$Beat[-length(stock$Beat)])
     stock$Bindex = as.numeric(stock$Beat==stock$Beat1)
#Record Survival Time
     temp<-subset(stock, stock$Bindex==0)
     temp$nDate1 = c(temp$nDate[-1],NA)
     temp$survtime=c(temp$nDate[-1]-temp$nDate[-length(temp$nDate)],NA)
#Subset  Survival Time Greater Than 1
     #C <- 1
     #SurvTable <- subset(temp, temp[,40] > C)
     #SurvTable[1:10,]
```

```
 #Seperate Beat
      C<-1
      Surv0<- subset(temp, temp[,35] == C)
      Surv0[1:5,]
 # A stock or B stock
      Stock_CMV=is.numeric(Surv0$A_Stock_CMV)*as.numeric
(Surv0$A_Stock_CMV)
      Stock_CMV = Stock_CMV +is.numeric(Surv0$ B_Stock_CMV)
*as.numeric(Surv0$ B_Stock_CMV)
      Stock_CC = is.numeric(Surv0$A_Stock_CC) *as.numeric
(Surv0$A_Stock_CC)
      Stock_CC = Stock_CC +is.numeric(Surv0$ B_Stock_CC)
*as.numeric(Surv0$ B_Stock_CC)
      tmp = cbind(rep("B", length(Stock_CMV)), rep("A",
length(Stock_CMV)))
      Stock_CMV_Type = tmp[,max(is.numeric(Surv0$A_Stock_CMV)+1)]
      Surv0$Stock_CMV = Stock_CMV
      Surv0$Stock_CC = Stock_CC
      Surv0$Stock_Type = Stock_CMV_Type
      Surv0$Categ = stockcat[i]
      Surv0=subset(Surv0, Surv0$TurnoverStock > 0)
      CeIndicator = abs(as.numeric(stock$nDate[nrow(stock)] ==
Surv0$nDate[nrow(Surv0)])-1)
```

```
      Surv0$Censor = c(rep("1",(nrow(Surv0)-1)),CeIndicator)
      Surv0$ID = rep(i, nrow(Surv0))
      Surv0=Surv0[(!(is.na(Surv0$PBR))),]
      Surv0=Surv0[(!(is.na(Surv0$PER))),]
      Surv0=Surv0[(!(is.na(Surv0$PCFR))),]
      Surv0=Surv0[(!(is.na(Surv0$PSR))),]
      Surv0=Surv0[(!(is.na(Surv0$AMV))),]
      Surv0=Surv0[(!(is.na(Surv0$AMO))),]
      Surv0=Surv0[(!(is.na(Surv0$TR))),]
      Surv0=Surv0[(!(is.na(Surv0$GenC))),]
      Surv0=Surv0[(!(is.na(Surv0$Stock_CC))),]
      Surv0=Surv0[(!(is.na(Surv0$Stock_CMV))),]
      Surv0$Censor[nrow(Surv0)] = 0
      stockdata = rbind(stockdata, Surv0)
 }
      stockdata$Censor = as.numeric(stockdata$Censor)
      stockdata$PER = as.numeric(stockdata$PER)
      stockdata$PBR = as.numeric(stockdata$PBR)
      stockdata$PSR = as.numeric(stockdata$PSR)
      stockdata$PCFR = as.numeric(stockdata$PCFR)
      stockdata$AMOscale = scale(stockdata$AMO, center = TRUE,
scale = TRUE)
      stockdata$AMVscale = scale(stockdata$AMV, center = TRUE,
```

```
scale = TRUE)
      stockdata$GenCscale = scale(stockdata$GenC, center = TRUE,
scale = TRUE)
      stockdata$Stock_CMVscale = scale(stockdata$Stock_CMV,
center = TRUE, scale = TRUE)
      stockdata$Stock_CCscale = scale(stockdata$Stock_CC,
center = TRUE, scale = TRUE)
      stockdata$PERscale = scale(stockdata$PER, center = TRUE,
scale = TRUE)
      stockdata$PCFRscale = scale(stockdata$PCFR, center = TRUE,
scale = TRUE)
      stockdata$PBRscale = scale(stockdata$PBR, center = TRUE,
scale = TRUE)
      stockdata$PSRscale = scale(stockdata$PSR, center = TRUE,
scale = TRUE)
      stockdata$TRscale = scale(stockdata$TR, center = TRUE,
scale = TRUE)
#Kaplan-Meier model
      fitKM=survfit(Surv(survtime, Censor) ~ 1 , stockdata,
se.fit=F)
      summary(fitKM)
      win.graph(width=8, height=6, pointsize=14)
      title= paste("KM Estimated Survival Fucntion")
```

```
      plot(fitKM,  xlab = "Day", ylab = "Survival function",
cex.lab=1.3, font.lab = 7, font.axis = 7, font.main = 7, main=title)
#Covariates Scatter Plot & Matrix
   covariate.mat = as.matrix(stockdata[, c(46:55)])
   cor(covariate.mat, use="complete.obs")
   splom(~ covariate.mat,cex=.5, pscales = 0,  main = "Scatter
Plot Matrix for Covariates")
   cor(covariate.mat, use="complete.obs")
#Shared Frailty Model
mod.shared<-frailtyPenal(formula = Surv(survtime, Censor) ~
cluster(ID) + GenCscale + AMVscale+ StockCMVscale  + AMOscale
+ TRscale + PERscale + PBRscale + PSRscale  + PCFRscale,
Frailty = TRUE, data = stockdata, n.knot = 10, kappa1 = 1,
cross.validation = TRUE)
#Summary and Plot of Shared Frailty Model
      summary(mod.shared, level = 0.95)
      plot(mod.shared, type.plot = "hazard", main = "Hazard Function
of Shared Frailty Model", level = 0.95, conf.bands = TRUE)
#Prediction
      datapred = data.frame( AMOscale =0, AMVscale=0, GenCscale=0,
Stock_CMVscale=0, PERscale=0, PCFRscale=0, PBRscale=0, PSRscale=0,
TRscale=0)
      datapred[1,] = c(-0.1837737, -0.1649021, -0.09315761,
```

```
-0.05133359, -0.09979914, 0.02425171, -0.6805865, -0.08478979,
 -0.472343)
      predm = prediction(mod.shared, datapred, 1, 1, 20 )
      plot(predm)
```

VITA

CHAO TANG

Education:	B.S. Statistics, North China University of Technology, Beijing, China 2012 M.S. Mathematical Sciences, East Tennessee State University, Johnson City, Tennessee 2014
Professional Experience:	Graduate Assistant, East Tennessee State University, Johnson City, Tennessee, 2012-2014