THREE ESSAYS ON THE CROSS-SECTIONAL DISTRIBUTION OF COMPLETED LIFETIMES

Maoshan Tian

Economic Section

Cardiff Business School

Cardiff University

A Thesis Submitted in Fulfilment of the Requirements for the Degree of

Doctor of Philosophy of Cardiff University

August 2019

I would like to dedicate this thesis to my loving parents.

Acknowledgements

I would like to express my sincere gratitude to my supervisor Professor Huw Dixon. He gave me so much help and encouragement to finish this work. I learned a lot under his guidance. I would also like to thank my supervisor Professor Garry Phillips. He taught me a lot about econometric methods. I also would like to thank Professor Patrick Minford's support. My thanks also go to Dr. Guangjie Li who spent long hours in discussion with me. Chapter 1 and chapter 2 are the joint works with my supervisor Professor Huw Dixon.

Abstract

The Distribution of completed lifetimes (*DCL*) is a new estimator defined and derived by Dixon (2012) in the context of the general Taylor price model (GTE). If we have panel data, the *DCL* is an estimator of the cross-sectional distribution of completed lifetimes. It is a new statistic to describe the data alongside the more familiar survival function and hazard function. Chapter 1 focuses on the cross-sectional distribution in relation to survival analysis. The delta method is applied to derive the variance of the of three distribution functions: the distribution of the duration, the cross-sectional distribution of age and the distribution of completed lifetimes. The Monte Carlo method is applied to evaluate the performances of those formulas. The simulation results show that the asymptotic variance formula of the *DCL*, age distribution and distribution of durations perform well when the sample size is above 50. With larger sample sizes, the bias of the variance is reduced.

In chapter 2, the pairs bootstrap method is applied to calculate the variance of the age distribution and the distribution of completed lifetimes (*DCL*). These results are compared with the asymptotic expansion of the variance. The Monte Carlo simulation is applied to investigate and evaluate the properties. In addition, the traditional Fieller's method and the delta method are applied to construct the confidence intervals for the *DCL*. The pairs bootstrap is applied to calculate the bootstrapped version of the variance of *DCL* and the construct the percentile confidence interval. In addition, the bootstrapped Fieller's method and delta method are also shown in this chapter. The numerical methods show that all methods provide the accurate confidence intervals for the *DCL* when the sample size above

200. But Fieller's method and the delta method are superior to the remaining methods when the sample size is below 50.

In chapter 3, we look at the CPI micro data. The CPI micro data are dealt by with different methods. We calculate the frequency and size of the price change. In addition, both the parametric and non-parametric methods are applied. We focus on comparing the survival function and the hazard function to determine whether they follow the same distribution. The null hypothesis is that both the survival function and the hazard function follow the same distribution between different groups. We show that there exist significant differences between the hazard functions and the survival functions before and after the financial crisis. The Cox model and the accelerated failure time model show that the frequency of price change after financial crisis is higher than the frequency of price change before financial crisis.

Table of contents

List of tables

General Introduction

Recently, there has been a growing interest in CPI micro data to investigate the existence of price stickiness. There exist two well-known models in monetary policy: the Taylor contract model and the Calvo contract model. Dixon (2012) derived a general Taylor contract model based on the Taylor contract model. The general Taylor contract model allows for the existence of many sectors, each with a traditional Taylor contract of a particular length. Each sector has a weight reflecting its size, with the sum of weights across all sectors summing to one. The weights represent the cross-sectional distribution of completed lifetimes (*DCL*). These coefficients can be estimated from the survival function and hazard function. It is useful to analyze the properties of the *DCL* estimates such as their variance and the corresponding confidence intervals. An individual price-spell is a sequence of periods during which the price set by a particular seller for a particular good (service) remains the same. Survival analysis can be applied to the duration of price-spells in the dataset. In terms of this, the Kaplan-Meier (KM) estimator can be applied to calculate the survival ratios of the price spell, giving the probability it will survive more than i periods. In addition, the marginal hazard function can be used to calculate the probability of price change in the i-th period conditional on lasting at least i periods. The DCL itself can be estimated using the KM estimator and the marginal hazard function.

Greenwood (1926) derived the variance of the KM estimator by applying the delta method. This method can also be applied to derive the variance of the *DCL* estimates. A potential problem is that the variance may not exist if the *DCL* follows a chi-square distribution.

However, there still exist some value of the mean and the variance which can be named as "pseudo" value. Even the mean and the variance do not exist, the "pseudo" value can be assumed to be the estimator of the mean and the variance since the mean and variance is converge to an area rather than a point.

After obtaining the variance formula of the *DCL* in chapter 1, chapter 2 use the delta method and Fieller's theorem to derive the confidence interval. Fieller (1940, 1954) introduced a transformation method to calculate the confidence interval for the ratio value. Since *DCL* is a ratio estimator, Fieller's theorem can be applied to derive the confidence interval for the *DCL*. In addition, the delta method is also applied to derive the confidence for the *DCL*. In chapter 2, the variance of the *DCL* and some inference obtained from chapter 1 are applied. The bootstrap method is also introduced to chapter 2. Bootstrap is a powerful method when there is no information for the data. The bootstrap method is introduced by Efron (1979). Efron (1981) also introduced the pairs bootstrap method to the case of censored data. By combining Fieller's theorem and Efron's bootstrap method, we showed the bootstrapped Fieller's method which can be derived the confidence interval for *DCL*. The bootstrapped Fieller's method is quite similar to the bootstrapped t statistics. Generally, the original critical value of the t table may not accurate, Therefore, we bootstrapped the test statistics of Fieller's method and create a new 5% and 95% critical value. The new critical values can be applied to construct the confidence interval. The bootstrap method is also powerful at calculating the variance. The bootstrapped variance of the *DCL* is also given and compared with the asymptotic formula derived in chapter 1.

In chapter 3, the CPI micro data is introduced. The CPI micro data is collected from the Office for National Statistics (ONS) website. The CPI micro data of UK consists of millions of monthly price-quote observations over the period 1999 to 2016. In the first part of chapter 3, all the data from 1999 to 2016 are included. All the properties including the frequency, size, and duration of the price change are calculated. The Cox model and the accelerated

failure time model are applied to investigate the heterogeneity effect and the seasonality effect of the price change. In the second part of chapter 3, the data is divided into two groups: the group before the financial crisis and the group after the financial crisis. The log-rank family tests are applied to evaluate the hazard function between the two groups. We find that there exist the significant differences of the hazard function and the survival function between the periods before the financial crisis and after the financial crisis.

Chapter 1

The Cross-sectional Distribution of Completed Lifetimes: Some New Inferences on the Survival Analysis

1.1 Introduction

Survival analysis has a wide range of applications across several disciplines including engineering, medicine and economics. In this chapter, we focus on three related distributions arising from survival analysis: the distribution of durations, the cross-sectional distributions of ages and completed lifetimes (*DCL*). Dixon (2012) introduces a unified framework for modelling the distributions, each of the three distributions can be written in terms of the survival function and the hazard function. In this chapter, we are going to derive the variance of estimates of the three distributions using the Delta method of Greenwood (1926).

Suppose we divide time into discrete periods: days, weeks, months and so on. In economic applications, this will often be driven by the data we have. The survival function gives the probability that an event will last for more than i periods, S_i. Clearly, $S_i \in [0,1]$ and $S_i \geq S_{i+m}$ for $m > 0$. The corresponding hazard function h_i gives the conditional probability

that having survived i periods, the event ends (death or failure). There are two classic methods of estimating this process. The Kaplan-Meier estimator (KM) estimates the survival function, whilst the Nelson-Aalen estimator (NA) estimates the cumulative hazard function (Kaplan and Meier (1958), Nelson (1972) and Aalen (1978)). The properties of both of these estimators have been well studied and in particular their asymptotic variances (see Breslow and Crowley (1974)). Whilst both KM and NA are general non-parametric estimators, they can also be estimated in parametric forms, such as the Cox proportional hazard model.

Starting from the KM and NA estimators, we are able to construct estimators of the three distributions (duration, ages and completed lifetimes) and to derive their asymptotic variances using both the Taylor expansion and delta methods. Theorem 1 derives the asymptotic variance of the distribution of durations. Theorem 2 and its corollary derive the asymptotic variances for the two cross-sectional distributions. In the simulation part, the Monte Carlo method was applied to explore the performance of the estimators and their sensitivity to the right-censored observations. We find that whilst there can be small bias for samples as small as 25, for samples 50 or over there is almost no bias and the results are not sensitive to the right-censored observations for sample sizes of 50 or over.

The three distributions we estimate we believe may have many useful applications. In economics, the estimated cross-sectional distribution of completed durations can be used to calibrate the Generalized Taylor Model of heterogeneous price and/or wage setting in a macroeconomic setting(See Coenen et al. (2008); Dixon and Bihan (2012a); Taylor (1980)). In this setting, the cross-sectional distribution gives the proportion of price or wage setters in the economy who set prices or wages for a particular period of time. However, in demographics, it also gives the distribution of completed lifetimes for those living at a point in time. Dixon and Siciliani (2009) estimate the distribution to obtain the completed waiting times of people on hospital waiting lists. Whilst it is common to calculate the life-expectancy from life tables, our estimates enable the complete distribution of lifetimes to be estimated.

Also, if we are looking at the stock of something at a point in time (unemployed workers, people living in an area, or machines), we can generate the distribution of completed durations (when the unemployed find a job, when people move from an area, when machines will fail). The estimation method can be non-parametric or parametric. The distribution of durations is useful when we want to look at the population over an extended period of time: the distribution of spells of unemployment, the distribution of price spells, the distribution of periods before machines have their first fault and so on.

In section 2 we briefly review the literature on the Kaplan-Meier estimator and Nelson-Aalen estimators of the survival and hazard functions. In section 3 we show how the estimators of the survival and hazard functions are related to estimators of the distribution of duration, the age distribution and the *DCL*. In section 4, the variance of the three distributions are derived in Theorems 1 and 2. In section 5, we use the Monte Carlo method to investigate the accuracy of the analytic formulas of the variance in finite sample sizes. Section 6 concludes.

1.2 Literature Review

The best known non-parametric estimator of the survival function was derived by Kaplan and Meier (1958). As Gillesple and Fisher (1979) explained, the Kaplan-Meier estimator is the limitation of the life table method to calculate the survival function with the increasing of the time intervals thereby tending to be zero. Kaplan and Meier (1958) assumed that lifetime time (the existence of the events or the age of death) was independent with the failure time (hazard rate) and derived the product limit estimator. They derived the variance of the survival function from an alternative way and obtained the same result as Greenwood (1926)'s formula. They also showed the variance formula of the survival function in the large sample size. In addition, the product limit estimators were the consistent estimators. On the other hand, Nelson (1972) applied a graphical method to investigate the failure ratio (hazard

rate) rather than the survival ratio. This graphical method was named as "hazard plotting". It plotted the hazard rate and the cumulative hazard rate depending on the distribution of the hazard function. After that, Aalen (1978) investigated the hazard function and the cumulative hazard function from the theoretical part. The counting process theory was applied to derive the cumulative hazard function. Since both Nelson and Aalen derived the cumulative hazard function, there existed the new estimator named as the Nelson-Aalen estimator. The Nelson-Aalen estimator was the cumulative hazard function. For the asymptotic properties of KM and NA estimators, see Andersen et al. (1993), Fleming and Harrington (1991), Kalbfleisch and Prentice (2002), Fleming and Harrington (1991), Bohoris (1994) and Colosimo et al. (2002). With respect to the parametric method for estimating the survival function and the hazard function, see Cox (1972). In terms of comparing the KM estimators in different groups, see Mantel (1966).

Breslow and Crowley (1974) investigated the life table and the Greenwood formula under the large sample conditions. They also derived the asymptotic normality for the standard life estimators. The estimators were assigned into the vector form which converged to the multivariate normal distribution. The covariance formula for the survival function and cumulative hazard function were derived under large sample conditions. The confidence interval of the KM estimator was introduced by Gillesple and Fisher (1979), Nair (1981), Nair (1984) and Kalbfleisch and Prentice (2002). Kalbfleisch and Prentice (2002) provided a method called the log-log transformation method to guarantee the positive lower bound of the confidence interval of the KM estimator.

Both parametric and non-parametric methods of estimation have been well studied and we know the variances of estimators and related statistical properties. In this chapter, we want to derive the variances of the three related distributions, two of which are cross-sectional. The age distribution, which gives the proportion of observations at a point in time which have particular ages; the cross-sectional distribution of completed lifetimes which gives

the proportion of observations at a point in time which will have a completed lifetime of a particular duration; the distribution of duration, the proportion of observations over the whole period which has completed lifetimes of a particular duration. Our analysis is applicable to a panel of observations, where we observe many agents (people, households, firms, machines) repeatedly over time and also to situations where we observe just a few or even one agent over time. Our framework is one of discrete time, although the analysis easily carries over to continuous time representations.

1.3 Survival Function and Hazard Function

Kaplan and Meier (1958) provided an estimator for the survival function, the Kaplan-Meier estimators $S_i \in [0,1], i = 0,1,2,.....,F$, where F is the maximum duration (this can be arbitrarily large, or may have an obvious empirical value such as the length of the dataset). We can imagine that there is a panel of agents. A spell of time is a period when the agent remains in the same state (remains alive, remains ill, sets the same price). *Failure* occurs when that state changes (death, recovery from illness, price changes). When the state changes, this can either be seen as the same agent continuing in the different states (the firm continues but sets the different price) or a new agent replaces the old (the machine fails and is replaced with a new machine).

If we look across the entire data set, we can count the number of spells that last at least k periods as N_k, and the number of failures in the $k-th$ period is D_k. N_0=N which is the total number of price spells in the sample. The Kaplan-Meier estimator $\hat{S}_i$ of the survival function S_i can be written as:

$$\hat{S}_i = \prod_{k=1}^{i} \frac{N_k - D_k}{N_k} \tag{1.1}$$

$\hat{S}_i$ can be defined as the proportion of spells remaining at $i-th$ period. This formula is also known as the product limited estimator for the survival function. The *cumulative hazard*

function can be named as the Nelson-Aalen estimator. The formula is:

$$\hat{H}_i = \sum_{k=1}^{i} \frac{D_k}{N_k} \tag{1.2}$$

The cumulative hazard function H_i is the summation of the hazard rate in each period until i period. The (marginal) hazard function can be defined as the proportion of failures amongst spells that have lasted i periods :

$$\hat{h}_i = \frac{D_i}{N_i} \tag{1.3}$$

By convention, we set $\hat{S}_0 = 1$ and $D_0 = \hat{h}_0 = 0$ are equal to zero since all spells last at least 0 periods. Since F is the longest spell, $\hat{h}_F = 1$ and $\hat{S}_F = 0$ under the uncensored case. The hazard function can be transformed to the survival function:

$$\hat{S}_i = \prod_{k=1}^{i} (1 - \hat{h}_k) \tag{1.4}$$

Likewise, the survival function can be transformed into the hazard function:

$$\hat{h}_i = \frac{\hat{S}_{i-1} - \hat{S}_i}{\hat{S}_{i-1}}$$

Before deriving our three distributions, we need to define an additional variable $\bar{h}$:

$$\bar{h} = \frac{1}{\sum_{k=0}^{F} \hat{S}_k} = \frac{1}{\hat{S}} \tag{1.5}$$

We define $\hat{S} = \sum_{k=0}^{F} \hat{S}_k$. Note that $\bar{h}$ is the reciprocal of the sum of survival probabilities. Intuitively, in a balanced panel, $\bar{h}$ is the proportion of agents that fail every period. To see this, consider some simple examples. First, failures occur in the first period for all spells. In this case, $\hat{S}_0 = 1$ and $\hat{S}_i = 0$[1] for all $i > 0$. All spells last for one period, and

[1] It should be mention that $\hat{S}_F$ equal to zero in the uncensored case; but $\hat{S}_F$ may not equal to zero when the right-censored problem is considered

$\bar{h} = 1$. Second the example where all spells last for two periods and then fail. In this case we have $\hat{S}_0 = \hat{S}_1 = 1, \hat{S}_i = 0$ for $i \geq 2$. In this case $\bar{h} = 1/2 : 50\%$ of spells fail per period. Hence, with a balanced panel, we can think of $\bar{h}$ as being the proportion of failures each period.

However, if we just have one cohort, we can think of $\bar{h}$ as being the weighted average hazard over $i = 0, 1, ..., F$, where the weights are the proportions surviving to period i divided by the sum of survival probabilities (to ensure the weights add up to 1).

Proposition 1. $\bar{h} = \frac{\sum_{k=0}^{F-1} \hat{S}_k \hat{h}_{k+1}}{\sum_{k=0}^{F} \hat{S}_k} = \frac{1}{\sum_{k=0}^{F} \hat{S}_k} = \frac{1}{\hat{S}}$

All proofs are in the appendix.

1.3.1 The Distribution of Durations

The distribution of durations gives the proportion of spells that survive at least $i-1$ periods and change (or "die") in the $i-th$ period. This is sometimes called the unconditional hazard function. The proportion of spells lasting exactly i periods can be defined as:

$$\hat{a}_i^d = \hat{S}_{i-1} \hat{h}_i \tag{1.6}$$

Clearly, $\hat{a}_i^d > 0$ and for $F > 1$, $1 > \hat{a}_i^d$. Note also that

$$\sum_{i=1}^{F} \hat{a}_i^d = 1$$

Since:

$$\sum_{i=1}^{F} \hat{S}_{i-1} \hat{h}_i = \sum_{i=1}^{F} \left(\hat{S}_{i-1} - \hat{S}_i \right) = \hat{S}_0 = 1$$

The distribution of durations can be treated as applying a particular cohort starting within a specific time frame (as with life tables), or as the distribution of all spells over a possibly long period (as in a balanced panel).

1.3.2 The Age Distribution

The age distribution can be explained as the ratio between the survival function of the price at the time i as:

$$\hat{a}_i^A = \frac{\hat{S}_{i-1}}{\sum_{k=0}^{F} \hat{S}_k} = \frac{\hat{S}_{i-1}}{\hat{S}} = \hat{S}_{i-1}\bar{h} \tag{1.7}$$

Since the survival function is non-increasing, the age distribution is also non-increasing: $\hat{a}_i^A \geq \hat{a}_{i+1}^A$. In addition, it is clear that the summation of the age distribution is equal to 1.

In the case of a balanced panel, we can think of the age distribution as being the cross-sectional distribution of ages across agents at a point in time. However, for a particular cohort, we can also think of it as the proportions of spells from that cohort lasting at least a particular length. This differs from the survival function because the proportions add up to unity (being the survival function pre-multiplied by $\bar{h}$). The survival function does not add up to unity because the events captured are not mutually exclusive. The sum of survival probabilities will exceed one unless all spells last just one period.

The estimates of the cross-sectional age distribution and the distribution of durations are related by the simple equality:

$$\hat{a}_i^A = \hat{a}_i^d \frac{\bar{h}}{h_i}$$

In the case of a constant hazard rate with $F = \infty$, $h_i = \bar{h}$ for all $i = 1, ..., \infty$, the two estimated distributions are identical $\hat{a}_i^A = \hat{a}_i^d$.[2]

1.3.3 The Distribution of Completed Lifetimes

Next, we introduce the less familiar cross-sectional distribution of the complete lifetimes (*DCL*) across agents. The new distribution is derived by Dixon (2012). The *DCL* can be

[2]The cross-section is length biased, so that the probability of observing a spell is proportional to length. The age distribution has an interruption bias, since the spells are incomplete. With a constant hazard, the two biases exactly cancel out.

written as:

$$\hat{a}_i = i\bar{h}\hat{S}_{i-1}\hat{h}_i \tag{1.8}$$

Alternatively, the *DCL* can be written as:

$$\hat{a}_i = i\frac{\hat{S}_{i-1}\hat{h}_i}{\left(\sum_{k=0}^{F}\hat{S}_k\right)} = i\frac{\hat{S}_{i-1}\hat{h}_i}{\hat{S}} \tag{1.9}$$

Clearly, for $i = 1..F,\ 1 > \hat{a}_i > 0$. For $F = 1$, $\hat{a}_1 = 1$. Furthermore, the sum of these estimators is always unity:

Proposition 2. $\sum_{i=1}^{F}\hat{a}_i = 1$. [3]

If we have a balanced panel, we can think of this as the cross-sectional distribution of completed lifetimes. In the case of a single cohort, we can think of *DCL* as being the distribution of completed lifetimes where we take an observation over each of the F periods. In the first period, we have all of the spells. In the second period, the one-period spells drop out and we have the spells with a duration of 2 and above and so on. Hence the i period contracts will be counted i times. Thus the *DCL* for the cohort is given $a_i = \bar{h}i\hat{a}_i^d$. In effect, we can think of the *DCL* as weighting the spells by their length, which as was suggested by Baharad and Eden (2004).

1.3.4 The Three Distributions

The survival function, hazard function, and the three distributions are different ways of describing the same data. They are all linked by identities: for any and all unique survival functions there exists a unique hazard function and the unique functions of the three distributions. These identities hold for the estimators as well. If we take a particular survival function, then we can express the hazard function and the three distributions in terms of

[3] Proportion 2 requires that all the observations are uncensored

Table 1.1 Relationships Among Different Functions and Distributions

	$\hat{S}_i$	$\hat{h}_i$	$\hat{a}_i^d$	$\hat{a}_i^A$	$\hat{a}_i$
$\hat{S}_i$	I	$\prod_{j=1}^{i}\left(1-\hat{h}_j\right)$ for $i=1,2,...,F.$	$1-\sum_{j=1}^{i}\hat{a}_j^d$	$\frac{\hat{a}_{i+1}^A}{\hat{a}_1^A}$	$1-\frac{1}{\sum_{k=1}^{F}\frac{\hat{a}_k}{k}}\sum_{j=1}^{i}\frac{\hat{a}_j}{j}$
$\hat{h}_i$	$\frac{\hat{S}_{i-1}-\hat{S}_i}{\hat{S}_{i-1}}$	I	$\prod_{j=1}^{i-1}\left(1-\hat{h}_j\right)$	$\frac{\hat{a}_i^A-\hat{a}_{i+1}^A}{\hat{a}_i^A}$	$\frac{\hat{a}_i}{i}\left[\sum_{i=1}^{F}\frac{\hat{a}_i}{i}\right]^{-1}$
$\hat{a}_i^d$	$\hat{S}_{i-1}-\hat{S}_i$	$\hat{h}_i\prod_{j=0}^{i-1}\left(1-\hat{h}_j\right)$	I	$\frac{\hat{a}_i^A-\hat{a}_{i+1}^A}{\hat{a}_1^A}$	$\frac{\hat{a}_i}{i\sum_{j=1}^{F}\frac{\hat{a}_j}{j}}$
$\hat{a}_i^A$	$\left[\sum_{i=0}^{F}\hat{S}_i\right]^{-1}\hat{S}_{i-1}$	$\left[\sum_{i=1}^{F}\prod_{j=0}^{i-1}\left(1-\hat{h}_j\right)\right]^{-1}\prod_{j=0}^{i-1}\left(1-\hat{h}_j\right)$	$\frac{1-\sum_{j=1}^{i-1}\hat{a}_j^d}{\sum_{i=1}^{F}i\hat{a}_i^d}$	I	$\sum_{k=1}^{F}\frac{\hat{a}_k}{k}-\sum_{j=1}^{i}\frac{\hat{a}_j}{j}$
$\hat{a}_i$	$i\left[\sum_{i=0}^{F}S_i\right]^{-1}\left(\hat{S}_{i-1}-\hat{S}_i\right)$	$i\prod_{j=1}^{i-1}\left(1-\hat{h}_j\right)\hat{h}_i\left[\sum_{i=1}^{F}\prod_{j=0}^{i-1}\left(1-\hat{h}_j\right)\right]^{-1}$	$\frac{\hat{a}_i^d}{i.\sum_{j=1}^{F}j\hat{a}_j^d}$	$i.\left(\hat{a}_i^A-\hat{a}_{i+1}^A\right)$	I

the survival function. Likewise, if we pick a particular hazard function, we can express the survival function and all three distributions in terms of the particular hazard function. We can also use one of the three distributions to describe the others and the survival function.

The full set of relationships is given in table(1.1). Each column represents the basic function or distribution: $\left\{\hat{S}_i,\hat{h}_i,\hat{a}_i^d,\hat{a}_i^A,\hat{a}_i\right\}$; each row shows how the element can be written in terms of the element of that column. Thus the first row has the different ways of writing the survival probability $\hat{S}_i$ in terms of itself (the indicator I), the hazard function $\hat{h}_i$, and then the three distributions $\hat{a}_i^d$, $\hat{a}_i^A$ and $\hat{a}_i$. The second row shows how we can write the survival function, the hazard (by itself) and the three distributions in terms of hazard function $\hat{h}_i$. The last row expresses the key statistic $\bar{h}$ in terms of all the elements. Note that these identities apply to any and all possible functions. The identities also apply to the estimators if they are unbiased: the estimated values must belong to the set of possible values.

1.4 Asymptotic Variances of the New Statistics

There are two equivalent ways to derive the variance formulas for the three distributions. One is the (multivariate) delta method. Another one is the counting process theory. We will use the delta method since it fits more easily with the discrete time framework. This method is also introduced by Greenwood (1926) to derive the variance of the survival function. Also,

the delta method gives more information with the higher order terms and converges to the true value more quickly than the counting process.

Assume the survival function $\hat{S}_i$ converges to the mean value S_i. It can be expressed as:

$$\sqrt{N_i}[\hat{S}_i - S_i] \overset{a.s.}{\sim} N(0, Var(\hat{S}_i))$$

By Taylor expansion we have:

$$g(\hat{S}_i) = g(S_i) + g'(S_i)(\hat{S}_i - S_i) + O_p((\hat{S}_i - S_i)^2)$$

where $O_p(.)$ is the *asymptotic* or *Bachmann-Landau notation*. From Slutsky's theorem[4], there exist the relationship:

$$\sqrt{N_i}[g(\hat{S}_i) - g(S_i)] \overset{a.s.}{\sim} N(0, [g'(S_i)]^2 Var(\hat{S}_i))$$

1.4.1 Asymptotic Variance for Durations.

We begin with the distribution function of durations. We will then extend this result to cover the distributions of age and *DCL*. Recall the distribution of durations which can be written as:

$$\hat{a}_i^d = \hat{S}_{i-1}\hat{h}_i$$

with variance:

$$Var(\hat{a}_i^d) = Var(\hat{S}_{i-1}\hat{h}_i)$$

[4]Slutsky's theorem states that if there exist two random variables or vectors X_i and Y_i, and those variables or vectors satisfy $X_i \xrightarrow{d.} X$ and $Y_i \xrightarrow{p.} c$, then there exist the relationship:

$$f(X_i, Y_i) \xrightarrow{d.} f(X, c)$$

Where $X_i \xrightarrow{d.} X$ means that X_i converges to the fixed value X in distribution; $Y_i \xrightarrow{p.} c$ means that Y_i converges to the constant point c in probability.

Theorem 1: Assume that we have the estimates of the hazard function $\hat{\mathbf{h}} \in [0,1]^F$ and the survival function $\hat{\mathbf{S}} \in [0,1]^F$ where the notation $\hat{\mathbf{h}} = (h_0, h_1, h_2, ...h_F)$ and $\hat{\mathbf{S}} = (\hat{S}_0, \hat{S}_1, ...\hat{S}_{F-1}, \hat{S}_F)$. The variance of the estimators $\hat{a}_i^d$ of the distribution of durations are given by:

$$\widehat{Var(\hat{a}_i^d)} = (\hat{S}_{i-1}\hat{h}_i)^2[\sum_{k=1}^{i-1} \frac{D_k}{N_k(N_k - D_k)} + \frac{N_i - D_i}{N_i D_i}] \tag{1.10}$$

To derive the variance of *DCL*, some additional formulae need to be derived. In equation (1.9), it can be seen that it is the product of the constant value i, and three random variables $\hat{S}_i$, $\hat{h}_i$ and $\bar{h}$. As Breslow and Crowley (1974) showed that the survival function and the hazard function followed the normal distribution asymptotically. The diagonal variance-covariance matrix of the hazard function is the variance formula for the hazard function while the off-diagonal terms are all equal to zero. It means that the hazard function is independent in different periods. On the other hand, they show the covariance for the survival function did not equal to zero.

We adopt a different method to derive the covariance of $\hat{S}_i$ and $\hat{S}_j$ for $i < j$. Recall the Taylor expansion for $\hat{S}_i$ and $\hat{S}_j$:

$$exp(ln\hat{S}_i) = exp(lnS_i) + (ln\hat{S}_i - lnS_i)exp(lnS_i) + O_p((ln\hat{S}_i - lnS_i)^2) \tag{1.11}$$

$$exp(ln\hat{S}_j) = exp(lnS_j) + (ln\hat{S}_j - lnS_j)exp(lnS_j) + O_p((ln\hat{S}_j - lnS_j)^2) \tag{1.12}$$

Rearranging equation (1.11) and (1.12) two equations:

$$\hat{S}_i - S_i = S_i(ln\hat{S}_i - lnS_i) + O_p((ln\hat{S}_i - lnS_i)^2) \tag{1.13}$$

$$\hat{S}_j - S_j = S_j(ln\hat{S}_j - lnS_j) + O_p((ln\hat{S}_j - lnS_j)^2) \tag{1.14}$$

If we multiply equation (1.13) with (1.14) and take the expectation:

$$\begin{aligned} Cov(\hat{S}_i,\hat{S}_j) &= E[\hat{S}_i - S_i)(\hat{S}_j - S_j)] \\ &\approx S_i S_j E[(ln\hat{S}_i - lnS_j)(ln\hat{S}_j - lnS_j)] \\ &= S_i S_j Cov(ln\hat{S}_i, ln\hat{S}_j) \\ &= S_i S_j Cov(\sum_{k=1}^{i} ln(1-\hat{h}_k), \sum_{k=1}^{j} ln(1-\hat{h}_j)) \\ &= S_i S_j Var(\sum_{k=1}^{i} ln(1-\hat{h}_k) \end{aligned} \tag{1.15}$$

The delta method is applied to derive the covariance of the KM estimators in equation (1.15). Since the hazard function $\hat{h}_k$ is assumed to follow the binomial distribution and it is independent in each period, the $Cov(\hat{h}_m,\hat{h}_n) = 0$ for $m \neq n$. Therefore, $Cov(\sum_{k=1}^{i} ln(1-\hat{h}_k), \sum_{k=1}^{j} ln(1-\hat{h}_j)) = Var(\sum_{k=1}^{i} ln(1-\hat{h}_k))$ when $i < j$. We have $Var[\sum_{k=1}^{i} ln(1-\hat{h}_k)] = \sum_{k=1}^{i} \frac{D_k}{N_k(N_k-D_k)}$ which is shown in the proof of theorem 1. Applying the large sample properties of the maximum likelihood estimator, the estimator of the covariance between $\hat{S}_i$ and $\hat{S}_j$ can be written as:

$$\widehat{Cov}(\hat{S}_i,\hat{S}_j) = \hat{S}_i\hat{S}_j[\sum_{k=1}^{i} \frac{D_k}{N_k(N_k-D_k)}] \; for \; i < j \tag{1.16}$$

After deriving the covariance, we can use the delta method of ratio variable to derive the formula of the *DCL*. At this point, the delta method has been applied twice. The first, the delta method is applied to derive the variance of the distribution of durations. In the second step, we treat the a_i as the ratio distribution $\hat{x}_i/\hat{y}$ with $\hat{x}_i = i\hat{S}_{i-1}\hat{h}_i$ and $\hat{y} = \sum_{k=0}^{F} \hat{S}_k$. Followed by this, we can apply the delta method for the ratio estimator $\hat{x}_i/\hat{y}$ to approximate $\hat{x}_i$ and $\hat{y}$ at the mean value x_i and y:

$$\frac{\hat{x}_i}{\hat{y}} \approx \frac{x_i}{y} + \frac{\hat{x}_i - x_i}{y} - \frac{x_i}{y^2}(\hat{y} - y)$$

Take the expectation on both sides, it can be seen that:

$$E[\frac{\hat{x}_i}{\hat{y}}] \approx \frac{x_i}{y} \tag{1.17}$$

Therefore, the variance of the ratio estimator $\hat{a}_i = \hat{x}_i/\hat{y}$ is:

$$Var(\frac{\hat{x}_i}{\hat{y}}) \approx \frac{Var(\hat{x}_i)}{y^2} + \frac{x_i^2}{y^4}Var(\hat{y}) - 2\frac{x_i}{y^3}Cov(\hat{x}_i, \hat{y}) \tag{1.18}$$

Apply the large sample properties of the maximum likelihood estimator, replace x_i by $\hat{x}_i$ and y by $\hat{y}$ where $\hat{x}_i = i\hat{S}_{i-1}\hat{h}_i$ and $\hat{y} = \hat{S}$.[5] First, note that the variance of $\hat{S}$ is:

$$Var(\hat{S}) = Var(\sum_{i=0}^{F}\hat{S}_i) = \sum_{i=0}^{F}Var(\hat{S}_i) + 2\sum_{i\neq j}Cov(\hat{S}_i, \hat{S}_j) \tag{1.19}$$

In addition, the covariance of iS_ih_i and S_j can be derived as:

$$Cov(i\hat{S}_{i-1}\hat{h}_i, \hat{S}_j) = iCov(\hat{S}_{i-1} - \hat{S}_i, \hat{S}_j) = i[Cov(\hat{S}_{i-1}, \hat{S}_j) - Cov(\hat{S}_i, \hat{S}_j)]$$

In other words, the $Cov(i\hat{S}_{i-1}\hat{h}_i, \hat{S}_j)$ can be transformed to the covariance of $Cov(i\hat{S}_{i-1}\hat{h}_i, \hat{S})$:

$$\begin{aligned} Cov(i\hat{S}_{i-1}\hat{h}_i, \sum_{k=1}^{F}\hat{S}_k) &= i[Cov(\hat{S}_{i-1}, \sum_{k=1}^{F}\hat{S}_k) - Cov(\hat{S}_i, \sum_{k=1}^{F}\hat{S}_k)] \\ &= i[\sum_{k=1}^{F}Cov(\hat{S}_{i-1}, \hat{S}_k) - \sum_{k=1}^{F}Cov(\hat{S}_i, \hat{S}_k)] \end{aligned} \tag{1.20}$$

Substituting the equation (1.10), (1.15), (1.19) and (1.20) into equation (1.18), we have:

Theorem 2 The variance of the *DCL* can be defined as:

$$\widehat{Var(\hat{a}_i)} = i^2\frac{\widehat{Var(\hat{S}_{i-1}\hat{h}_i)}}{\hat{S}^2} + i^2\frac{\hat{S}_{i-1}^2\hat{h}_i^2\widehat{Var(\hat{S})}}{\hat{S}^4} - 2i^2\frac{\hat{S}_{i-1}\hat{h}_i\widehat{Cov}(\hat{S}_{i-1}\hat{h}_i, \hat{S})}{\hat{S}^3} \tag{1.21}$$

[5]The maximum likelihood estimator $\hat{S}_i$ is close to the mean value of S_i in large sample size. the S_i can be replaced by $\hat{S}_i$ in Greenwood formula. At this point, we replace x_i by $\hat{x}_i$ and y by $\hat{y}$

For $i = 1, 2, ...F$.

Since the variance of the *DCL* has been derived, the variance formula of the age distribution can be given.

Corollary 1 The variance of the age distribution is given by:

$$\widehat{Var(\hat{a}_i^A)} = \frac{\widehat{Var(\hat{S}_{i-1})}}{\hat{S}^2} + \frac{\hat{S}_{i-1}^2 \widehat{Var(\hat{S})}}{\hat{S}^4} - 2\frac{\hat{S}_{i-1}\widehat{Cov}(\hat{S}_{i-1}, \hat{S})}{\hat{S}^3} \tag{1.22}$$

For $i = 1, 2, ..., F$.

1.4.2 Right-Censored Problem and the Non-parametric Maximum Likelihood Estimation

The KM estimators can be derived from the non-parametric maximum likelihood estimators (NPMLE). The NPMLE gives the same results as the product limited estimator (PL). NPMLE is the numerical method to solve out the hazard function. The maximum likelihood function for the survival function can be written as:

$$L = \prod_{i=1}^{F} [\hat{S}_{i-1} - \hat{S}_i]^{D_i} \hat{S}_i^{N_i - D_i} \tag{1.23}$$

Since $S_0 = 1$, this maximum likelihood formula can be applied to estimate the survival function for F periods. This function can be modified if we replace the survival function by the hazard function:

$$\hat{S}_i = \prod_{k=0}^{i} (1 - \hat{h}_k)$$

Define the value N_i to be the total number of the observations at the risk in the period i. Then the NPMLE function can be rewritten as:

$$L = \prod_{i=1}^{F} (1-\hat{h}_i)^{N_i - D_i} \hat{h}_i^{(D_i)} \tag{1.24}$$

Take the first order derivatives with respect to h_i:

$$\frac{\partial lnL}{\partial \hat{h}_i} = -\frac{N_i - D_i}{1-\hat{h}_i} + \frac{D_i}{\hat{h}_i} = 0$$

$$\hat{h}_i = \frac{D_i}{N_i}$$

Since the NPMLE provides the estimator of the hazard function, the survival function can be calculated as $\hat{S}_i = \sum_{k=1}^{i}(1-\hat{h}_k)$. The age distribution, distribution of duration and DCL can be calculated and expressed as functions of $\hat{S}_i$ and $\hat{h}_i$.

Now we introduce the concept of the censored spells. The left-censored spells occur when the starting point cannot be observed, being outside the period of observation, also known as the sample period. However, the endpoint is included in the sample period. The right-censored spells are where the endpoint cannot be observed but the start can be. The KM estimator is usually only applied after excluding left-censored spells, so we just consider the implications of right-censored data. The maximum length of a spell is F periods. Define the T_j as the true lifetime of a spell $j \in (1,2,.....N)$. N is the total number of the observations in the sample size. The observed lifetime t_i can be defined as:

$$t_j = min(T_j, C_j) \quad and \quad \omega_j = I(T_j \leq C_j) \quad j = 1,2,.....N_j.$$

Where the C_j means the censored time of the observation for the j-th observation; T_j is the survival time of the j-th observation; The observed lifetime t_j is the minimum of C_j and T_j.

ω_j is the uncensored dummy coefficient. If the observation t_j is right-censored, ω_j is equal to 0. If he observation t_j is uncensored, ω_j is equal to 1. If the censored time is less than the survival time, the observation is right-censored:

$$C_j < T_j,\ t_j = C_j\ (right\ \ censored) \quad and \quad \omega_j = 0$$

Otherwise, if the observation is uncensored:

$$T_j \leq C_j,\ t_j = T_j\ (uncensored) \quad and \quad \omega_j = 1$$

1.5 Monte Carlo Simulation

The variance formulas of the *DCL*, age distribution and distribution of duration have been derived by the delta method. In this section, we are going to investigate the properties of those formulas. Depending on the simulation, the bias of the analytic variances can be evaluated. The data is generated from the continuous time exponential distribution function. The sample sizes are $N = 25$, $N = 50$, $N = 100$ and $N = 200$. Both the cases of uncensored and right-censored observations are considered in the simulations. We assume that the data is collected in discrete time. We collect the raw continuous time data and transfer them into intervals defined as $(0, r_1], (r_1, r_2], ...(r_{k-1}, r_k], ...(r_{F-1}, r_F]$.[6] For $t_j \in (0, r_1]$ we set duration $t_j = r_1$. For $t_j \in (r_{k-1}, r_k]$ we set $t_j = r_k$ and so on. After that, we can count the number of the observations locating in each interval. The number of the observations locating in k-th interval can be defined as D_k when all the observations are uncensored. Therefore, the KM estimators and the hazard functions can be calculated for each interval. The simulation process is:

[6]Now the interval $(0, r_1]$ can be defined as the "first" period. Following by this rule, $(r_{k-1}, r_k]$ means the $k-th$ period.

Step 1: assume the observed duration is $t_j = min(T_j, C_j)$. The sample sizes are $N = 25$, $N = 50$, $N = 100$ and $N = 200$. The results are reported separately. Both the lifetime time T_j and the censored time C_j follow the exponential distribution. The censored time and the lifetime have the probability density functions (PDF):

$$p(C_j) = 0.5exp(-0.5C_j) \quad p(T_j) = 2exp(-2T_j) \tag{1.25}$$

Therefore, they have the survival function for each r_k-th period:

$$p(C_j > r_k) = exp(-0.5r_k) \quad p(T_j > r_k) = exp(-2r_k) \tag{1.26}$$

j is the $j-th$ observation where $j \in (1, 2, ..., N)$. N is the total sample size. There exists the $\frac{2}{2+0.5}$ [7] uncensored proportion of the total observations depending on the PDF. For the uncensored problems, we can just generate the survival time $t_j = T_j$ and assume they are all uncensored with the right-censored coefficient $\omega_j = 1$ for all the j. In other words, the observations can be written as $(T_j, 1)$ for all the j. For the right-censored problem, we also need to generate the censored time C_j. If $T_j < C_j$, the $j-th$ observation is uncensored and we assign a parameter $\omega_j = 1$ to the $j-th$ observation and define it as (T_j, ω_j). If $C_j < T_j$, it means the observation is right-censored. Therefore, we written it as (C_j, ω_j) with $\omega_j = 0$. After that, the survival data are allocated into F categories. In other words, different survival periods are transformed into some fixed period groups. In the first case, F is chose to be 5. We divide the observations into five regions: $(0, 0.1]$, $(0.1, 0.2]$, $(0.2, 0.3]$, $(0.3, 0.5]$, and $(0.5, \infty)$. This can be known as case 1. In case 2, another five regions are generated: $(0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$, and $(0.8, \infty)$.

[7] Since the parameter of the exponential distribution of Censored time and observed time are 0.5 and 2, separately. The right-censored proportion of the total sample can be known as $\frac{0.5}{2+0.5}$. The algebra is shown by Efron (1981).

Step 2: The formula (2.3), (1.22) and (2.14) are applied to calculate the variance of the *DCL*, age distribution and duration of distribution for each period.

Step 3: Repeat step 1 and step 2 by M times. M is chosen to be 10000. One important thing is that there may exist zero observations in one of the 5 intervals in the simulated samples. If there exists such a zero, this sample is eliminated and another sample is re-simulated again until we have 10000 samples in which all 5 intervals are non-empty. Following Kiviet and Phillips (2014)'s method, The real value of the variance of *DCL* can be calculated by:

$$Var(\hat{a}_i)_{true} = \sum_{m=1}^{M} (\hat{a}_{i,m} - \frac{\sum_{m=1}^{M} \hat{a}_{i,m}}{M})^2/(M-1) \tag{1.27}$$

The true value of the variance of *age* distribution[8] can be calculated by:

$$Var(\hat{a}_i^A)_{true} = \sum_{m=1}^{M} (\hat{a}_{i,m}^A - \frac{\sum_{m=1}^{M} \hat{a}_i^A}{M})^2/(M-1) \tag{1.28}$$

The true value of the variance of *duration* distribution can be calculated by:

$$Var(\hat{a}_i^d)_{true} = \sum_{m=1}^{M} (\hat{a}_{i,m}^d - \frac{\sum_{m=1}^{M} \hat{a}_{i,m}^d}{M})^2/(M-1) \tag{1.29}$$

Where the subscript m in $a_{i,m}$, $a_{i,m}^d$ and $a_{i,m}^A$ means the m-th Monte Carlo simulation. In other words, we collect M estimators of $a_{i,m}$,$a_{i,m}^A$ and $a_{i,m}^d$ and calculate the variance of them[9]. Equation (1.27), (1.28) and (1.29) can be known as the real variance of the three distributions depending on the properties of the Monte Carlo simulations. The benchmark real variances are applied to compare with the analytic variance derived by delta method whether the approximation results are close to the real value. Since the formula $\hat{S}_{i-1}\hat{h}_i = \hat{S}_{i-1} - \hat{S}_i$ is

[8]The age distribution is the special case of the *DCL*, so we only provide the empirical results of *DCL*.

[9]In the simulation result, the coefficient i of equation (1.27) is ignored in the simulation process. The reason is that i is a constant parameter for each a_i

replaced in the variance formula to calculate the true value of the variance of the distribution of duration, so the first period of the variance of *DCL* can be known as the variance of the age distribution. In addition, the final period variance of *DCL* is also a special case of the variance of the age distribution when all the observations are uncensored.

Table (1.2) reports the simulation results for the uncensored data for case 1. Those data are divided into $(0,0.1]$, $(0.1,0.2]$, $(0.2,0.3]$, $(0.3,0.5]$, and $(0.5,\infty)$. In table (1.2), all the right-censored parameters ω_j are all equal to 1, which means that the observations are all uncensored. As we can see from table (1.2), when the sample size is equal to 25, there exists a slight bias for the variance. When the sample size is increased to 50, the approximation formula of the variance performs very well for all the regions. When the sample size is increased to either N=100 or N=200, the gap between the benchmark value and the analytic value of the variance are reduced. With the increase of the sample size, the approximation value is closer to the true value when the observations are uncensored.

Table 1.2 The Variance of the Duration of Case 1 when All the Observations Are Uncensored. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a^d_{0.1})$	$Var(a^d_{0.2})$	$Var(a^d_{0.3})$	$Var(a^d_{0.5})$	$Var(a^d_{\infty})$
25	5.6497	4.6783	3.7561	5.6901	9.1525
50	2.9913	2.5801	2.1244	2.9637	4.6877
100	1.4844	1.2554	1.0757	1.4705	2.2919
200	0.7451	0.6293	0.5404	0.7406	1.1594
Approximation Value					
N	$E[\widehat{Var(a^d_{0.1})}]$	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.3})}]$	$E[\widehat{Var(a^d_{0.5})}]$	$E[\widehat{Var(a^d_{\infty})}]$
25	5.6880	4.9125	4.2650	5.6862	8.8820
50	2.8987	2.4790	2.0904	2.9041	4.5610
100	1.4677	1.2533	1.0613	1.4640	2.3022
200	0.7367	0.6293	0.5318	0.7355	1.1578

Note: $Var(a^d_i)$ is the benchmark value calculated from formula (1.28); $E[\widehat{Var(a^d_i)}]$ is the variance calculated from formula (2.14).

In table (1.3), the data are simulated by the same process. There still exists a slight bias for the variance of the *DCL* when the sample size is $N = 25$. When the sample size is

Table 1.3 The Variance of the *DCL* of Case 1 when All the Observations Are Uncensored. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a_{0.1})$	$Var(a_{0.2})$	$Var(a_{0.3})$	$Var(a_{0.5})$	$Var(a_{\infty})$
25	0.7378	0.51417	0.3606	0.4858	0.46588
50	0.36525	0.27085	0.19608	0.24835	0.2327
100	0.1777	0.1299	0.0986	0.1228	0.1137
200	0.0879	0.0646	0.0494	0.0616	0.0575
Approximation Value					
N	$E[\widehat{Var(a^d_{0.1})}]$	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.3})}]$	$E[\widehat{Var(a^d_{0.5})}]$	$E[\widehat{Var(a^d_{\infty})}]$
25	0.7719	0.5478	0.4084	0.4752	0.4460
50	0.3644	0.2638	0.1951	0.2418	0.2256
100	0.1782	0.1305	0.0982	0.1221	0.1140
200	0.0877	0.0647	0.0489	0.0613	0.0572

Note: $Var(a_i)$ is the benchmark value calculated from formula (1.27); $E[\widehat{Var(a_i)}]$ is the variance calculated from formula (2.3).

increased to 50, all the approximated results are improved and they are all close to the true value. With respect to N=100 and N=200, the approximation of the variance tends to be closer to the real variance. However, it can be found that the approximation of the variances do not always overestimate the true value. Sometimes it underestimates the true variance of the *DCL*. In conclusion, the analytic variance formula of *DCL* is reduced with the increase of the sample size.

Next, the right-censored observations are considered. Table (1.4) shows the the variance of the distribution of duration. Compared with the benchmark value, there exists a slight bias in the variance calculated from the analytic formula when the sample size N=25. When sample size is increased to 50, the analytic variance performs well. When the sample size tends to be a larger (N=100 and N=200), the empirical results show that the approximations of the variances are nearly the same as the true values. In conclusion, the analytic formula of the variance can capture the true value even when the sample size is small(N=25). The asymptotic variance may overestimate or underestimate the true value.

Table 1.4 The Variance of the Duration of Case 1 when the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a^d_{0.1})$	$Var(a^d_{0.2})$	$Var(a^d_{0.3})$	$Var(a^d_{0.5})$	$Var(a^d_{\infty})$
25	5.4770	4.7128	4.0016	6.1382	9.5321
50	2.8680	2.5773	2.2555	3.4038	5.0703
100	1.4566	1.3187	1.1746	1.7109	2.5091
200	0.7497	0.6525	0.5807	0.8130	1.2539
Approximation Value					
N	$E[\widehat{Var(a^d_{0.1})}]$	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.3})}]$	$E[\widehat{Var(a^d_{0.5})}]$	$E[\widehat{Var(a^d_{\infty})}]$
25	5.6045	5.0796	4.7601	6.4884	9.4990
50	2.8519	2.5617	2.2987	3.3076	4.9242
100	1.4442	1.3024	1.1618	1.6589	2.4841
200	0.7248	0.6537	0.5837	0.8349	1.2496

Note: $Var(a^d_i)$ is the benchmark value calculated from formula (1.28); $E[\widehat{Var(a^d_i)}]$ is the variance calculated from formula (2.14).

Table 1.5 The Variance of the *DCL* of Case 1 when the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a_{0.1})$	$Var(a_{0.2})$	$Var(a_{0.3})$	$Var(a_{0.5})$	$Var(a_{\infty})$
25	0.6767	0.4910	0.3696	0.5094	0.6005
50	0.3311	0.2565	0.2006	0.2770	0.3102
100	0.1645	0.1290	0.1032	0.1381	0.1545
200	0.0832	0.0631	0.0506	0.0656	0.0772
Approximation Value					
N	$E[\widehat{Var(a^d_{0.1})}]$	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.3})}]$	$E[\widehat{Var(a^d_{0.5})}]$	$E[\widehat{Var(a^d_{\infty})}]$
25	0.6798	0.5162	0.4174	0.5094	0.5824
50	0.3345	0.2557	0.2032	0.2660	0.2992235
100	0.1648	0.1285	0.1025	0.1345	0.1513
200	0.0811	0.0636	0.0511	0.0676	0.0763

Note: $Var(a_i)$ is the benchmark value calculated from formula (1.27); $E[\widehat{Var(a_i)}]$ is the variance calculated from formula (2.3).

Table (1.5) shows the simulation results of the *DCL* variance. When the sample size is extremely small (N=25), the approximations of the variances are still quite accurate. When the sample size is increased to 50, all the analytic variances are improved. They are all close to the true value. When the sample size is large (N=100 and N=200), the analytic variances

are very close to the true value. Therefore, the analytic formula of the variance of the *DCL* works well when the right-censored observations exist.

Table (1.6) to (1.9) present the variance of the distribution of durations and *DCL* under the alternative assumption case 2 where all the data are assigned into the five regions: $(0,0.2]$, $(0.2,0.4]$, $(0.4,0.6]$, $(0.6,0.8]$, and $(0.8,\infty)$. In terms of the uncensored case, all the right-censored coefficients $\omega_j = 1$. When the right-censored observations are considered, the $p(C_j) = 0.5exp(-0.5C_j)$ and $p(T_j) = 2exp(-2T_j)$ are generated. This process is the same as the case 1. In table (1.6) and (1.7), the analytic formula of the variance can give an accurate approximation for the true value even in the extremely small sample size (*N*=25). Both the variance of the *DCL* and duration are either overestimated or underestimated without a unique conclusion. When the sample size tends to be a large number, they are nearly unbiased from the true value.

When there exist the right-censored observations in the samples, the same results can be concluded in table (1.8) and table (1.9).

Table 1.6 The Variance of the Duration of Case 2 when All the Observations Are Uncensored. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a^d_{0.2})$	$Var(a^d_{0.4})$	$Var(a^d_{0.6})$	$Var(a^d_{0.8})$	$Var(a^d_{\infty})$
25	8.511	6.4731	4.6370	2.9440	6.0651
50	4.4362	3.4353	2.5537	1.7365	3.2366
100	2.1816	1.6982	1.2514	0.8812	1.5987
200	1.1384	0.8610	0.6392	0.4491	0.8074
Approximation Value					
N	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.4})}]$	$E[\widehat{Var(a^d_{0.6})}]$	$E[\widehat{Var(a^d_{0.8})}]$	$E[\widehat{Var(a^d_{\infty})}]$
25	8.4295	6.5750	4.9049	3.6899	6.1695
50	4.3413	3.3608	2.4783	1.7592	3.1459
100	2.1879	1.7030	1.2538	0.8834	1.5953
200	1.0990	0.8550	0.6298	0.4459	0.8012

Note: $Var(a^d_i)$ is the benchmark value calculated from formula (1.28); $E[\widehat{Var(a^d_i)}]$ is the variance calculated from formula (2.14).

Table 1.7 The Variance of the DCL of Case 2 when All the Observations Are Uncensored. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a_{0.2})$	$Var(a_{0.4})$	$Var(a_{0.6})$	$Var(a_{0.8})$	$Var(a_{\infty})$
25	2.3456	1.1895	0.6760	0.3829	0.5396
50	1.2091	0.6276	0.3794	0.2329	0.2956
100	0.5764	0.3054	0.1860	0.1176	0.1462
200	0.2987	0.1535	0.0938	0.0602	0.0731
Approximation Value					
N	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.4})}]$	$E[\widehat{Var(a^d_{0.6})}]$	$E[\widehat{Var(a^d_{0.8})}]$	$E[\widehat{Var(a^d_{\infty})}]$
25	2.4358	1.2334	0.7190	0.4755	0.5430
50	1.2085	0.6222	0.3667	0.2325	0.2828
100	0.5841	0.3095	0.1845	0.1174	0.1440
200	0.2884	0.1540	0.0925	0.0594	0.0727

Note: $Var(a_i)$ is the benchmark value calculated from formula (1.27); $E[\widehat{Var(a_i)}]$ is the variance calculated from formula (2.3).

Table 1.8 The Variance of the Duration of Case 2 when the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a^d_{0.2})$	$Var(a^d_{0.4})$	$Var(a^d_{0.6})$	$Var(a^d_{0.8})$	$Var(a^d_{\infty})$
25	7.9283	6.9220	5.1855	3.6686	6.3243
50	4.2300	3.7601	3.0929	2.3271	3.9246
100	2.1853	1.8781	1.5767	1.2560	1.9910
200	1.0669	0.9379	0.7799	0.6350	0.9996
Approximation Value					
N	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.4})}]$	$E[\widehat{Var(a^d_{0.6})}]$	$E[\widehat{Var(a^d_{0.8})}]$	$E[\widehat{Var(a^d_{\infty})}]$
25	8.1582	7.0511	6.0630	5.5234	7.5641
50	4.2223	3.6962	3.0363	2.4997	3.8673
100	2.1348	1.8690	1.5411	1.2321	1.9587
200	1.0734	0.9371	0.7749	0.6247	0.9865

Note: $Var(a^d_i)$ is the benchmark value calculated from formula (1.28); $E[\widehat{Var(a^d_i)}]$ is the variance calculated from formula (2.14).

Table 1.9 The Variance of the DCL of Case 2 when the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3

True Value					
N	$Var(a_{0.2})$	$Var(a_{0.4})$	$Var(a_{0.6})$	$Var(a_{0.8})$	$Var(a_{\infty})$
25	1.9352	1.1768	0.7175	0.4503	0.5984
50	1.0521	0.6580	0.4406	0.2989	0.3947401
100	0.5269	0.3224	0.2233	0.1618	0.2025
200	0.2539	0.1603	0.1108	0.0813	0.1017
Approximation Value					
N	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.4})}]$	$E[\widehat{Var(a^d_{0.6})}]$	$E[\widehat{Var(a^d_{0.8})}]$	$E[\widehat{Var(a^d_{\infty})}]$
25	1.8620	1.1101	0.7612	0.5964	0.6855
50	1.0220	0.6266	0.4140	0.3027	0.3806
100	0.5159	0.3197	0.2164	0.1557	0.1970
200	0.2561	0.1594	0.1091	0.0798	0.1000

Note: $Var(a_i)$ is the benchmark value calculated from formula (1.27); $E[\widehat{Var(a_i)}]$ is the variance calculated from formula (2.3).

1.6 Conclusion

In this chapter, we use the delta method to derive the variance of the distributions of durations, ages and completed lifetimes (*DCL*). The *DCL* and age distributions are cross-sectional, although they can be applied to the case of a single cohort of data rather than a panel. Depending on the asymptotic approximation of the variance, we provide the analytic formula to calculate the variance of the three distributions. The asymptotic variance derived from delta method is straightforward since it is the same derivation as the Greenwood formula. In addition, the covariance between different survival lengths is derived in a clearer way compared with Breslow and Crowley (1974).

The data is simulated to investigate the accuracy of the asymptotic variances of the three distributions. There are two cases of simulation considered in this chapter. The observations are assumed to follow the exponential distribution. Depending on the Monte Carlo results, the analytic formulas of the variance of the *DCL*, age distribution and the distribution of the

durations give more accurate results as the sample size increases. In other words, the bias between the approximations and the true values are reduced as increase of the sample size.

For the further study, it will be useful to see if the bootstrap corrected variance can provide a better result compared with the asymptotic formula in the small sample size. Another extension is to derive the confidence interval for the *DCL*.

1.7 Appendix.

1.7.1 Proof of Proposition 1

To see why, note that

$$\begin{aligned}
\bar{h} &= \frac{1}{\sum_{i=0}^{F}\hat{S}_i}\sum_{i=0}^{F-1}\hat{S}_i\hat{h}_{i+1} \\
&= \frac{1}{\sum_{i=0}^{F}\hat{S}_i}\left(\frac{\hat{S}_0-\hat{S}_1}{\hat{S}_0}+\hat{S}_1\left(\frac{\hat{S}_1-\hat{S}_2}{\hat{S}_1}\right)+...\hat{S}_{F-1}\right) \\
&= \frac{1}{\sum_{i=0}^{F}\hat{S}_i}\left(1-\hat{S}_1+\left(\hat{S}_1-\hat{S}_2\right)+\left(\hat{S}_2-\hat{S}_3\right)+..(\hat{S}_{F-2}-\hat{S}_{F-1})+\hat{S}_{F-1}\right) \\
&= \frac{1}{\sum_{i=0}^{F}\hat{S}_i}
\end{aligned}$$

Proof of Proposition 2

In proportion 2, it requires that all the observations are uncensored. To see why Proposition 2 holds, note that

$$\begin{aligned}
\sum_{i=1}^{F}\hat{a}_i &= \bar{h}\sum_{i=1}^{F} i\hat{S}_{i-1}\hat{h}_i \\
&= \bar{h}\sum_{i=1}^{F} i\left(\hat{S}_{i-1}-\hat{S}_i\right) \\
&= \bar{h}\left[\sum_{i=1}^{F}\left(\hat{S}_{i-1}-\hat{S}_i\right)+\sum_{i=2}^{F}\left(\hat{S}_{i-1}-\hat{S}_i\right)+..\sum_{i=j}^{F}\left(\hat{S}_{i-1}-\hat{S}_i\right)+\hat{S}_{F-1}\right] \\
&= \bar{h}\left[\hat{S}_0+\hat{S}_1+\hat{S}_2..+\hat{S}_F\right] \\
&= 1
\end{aligned}$$

1.7.2 Proof of Theorem 1

The first-order Taylor expansion can be applied to derive the variance of the distribution of duration. Recall the distribution of duration formula $\hat{a}_i^d = \hat{S}_{i-1}\hat{h}_i$, the Taylor series for $\hat{a}_i^d$ at $\hat{S}_{i-1}$=S_{i-1} and $\hat{h}_i$=h_i is[10]:

$$\hat{S}_{i-1}\hat{h}_i \approx S_{i-1}h_i + h_i(\hat{S}_{i-1} - S_{i-1}) + S_{i-1}(\hat{h}_i - h_i)$$

Recall the equation 1.11 and 1.12, we can use the same method get the equations as below:

$$\hat{S}_{i-1} - S_{i-1} = S_i(ln\hat{S}_{i-1} - lnS_{i-1}) + O_p((ln\hat{S}_{i-1} - lnS_{i-1})^2) \tag{1.30}$$

$$\hat{h}_i - h_i = h_i(ln\hat{h}_i - lnh_i) + O_p((ln\hat{h}_i - lnh_i)^2) \tag{1.31}$$

Depending on the equation (1.30) and (1.31), equation (1.30) can be expressed as:

$$\hat{S}_{i-1}\hat{h}_i - S_{i-1}h_i \approx S_{i-1}h_i[(ln\hat{S}_{i-1} - lnS_{i-1}) + (ln\hat{h}_i - lnh_i)]$$

Hence the variance $Var(a_i^d)$ can be rewritten as:

$$Var(\hat{a}_i^d) = Var(\hat{S}_{i-1}\hat{h}_i) \cong (S_{i-1}h_i)^2[Var(\ln\hat{S}_{i-1}) + Var(\ln\hat{h}_i)]$$

Depending on the large sample property of the maximum likelihood estimator, the KM estimator $\hat{S}_{i-1}$ converges to the true value S_{i-1} and the marginal hazard function $\hat{h}_i$ converges to the true value h_i. Therefore, $Var(\hat{a}_i^d)$ can be written as:

$$Var(\hat{a}_i^d) \cong (\hat{S}_{i-1}\hat{h}_i)^2[Var(ln\hat{S}_{i-1}) + Var(ln\hat{h}_i)]$$

[10]The hazard function can be estimated by the maximum likelihood method, and the KM estimator consists of the hazard function. Depending on the properties of the maximum likelihood estimator, it can be known that the KM estimator $\hat{S}_{i-1}$ converges to the true value S_{i-1} and the marginal hazard function $\hat{h}_i$ converges to the true value h_i. At this point, we show that those result can be derived from the delta method

This approximation assumes that $\hat{S}_{i-1}$ is independent with $\hat{h}_i$, the covariance between the $\ln\hat{S}_{i-1}$ and $\ln\hat{h}_i$ is zero.

The logarithm version of the survival function can be written as:

$$\ln\hat{S}_{i-1} = \sum_{k=1}^{i-1} \ln(1-\hat{h}_k)$$

Assume the D_i follows the binomial distribution with parameters N_i and $\hat{h}_i$. Therefore, $Var(D_i) = N_i\hat{h}_i(1-\hat{h}_i)$. It can be shown that $Var(\hat{h}_i) = Var(\frac{D_i}{N_i}) = \hat{h}_i(1-\hat{h}_i)/N_i$. By applying the first-order Taylor expansion:

$$ln(\hat{h}_i) = lnh_i + (\hat{h}_i - h_i)\frac{1}{h_i} + O_p((\hat{h}_i - h_i)^2)$$

$$ln(1-\hat{h}_i) = ln(1-h_i) + (\hat{h}_i - h_i)\frac{1}{1-h_i} + O_p((\hat{h}_i - h_i)^2)$$

To rearrange the formula:

$$ln(\hat{h}_i) - lnh_i \cong (\hat{h}_i - h_i)\frac{1}{h_i}$$

$$ln(1-\hat{h}_i) - ln(1-h_i) \cong (\hat{h}_i - h_i)\frac{1}{1-h_i}$$

It can be assumed that the observations are independent Bernoulli distribution and they are independent with each other. Then the variance of $\hat{h}_i$ can be written as:

$$Var(\hat{h}_i) = Var(1-\hat{h}_i) = \frac{\hat{h}_i(1-\hat{h}_i)}{N_i}$$

Apply the large sample properties of the maximum likelihood estimator:

$$\begin{aligned} Var(ln(\hat{h}_i)) &\cong \frac{1}{\hat{h}_i^2}\frac{\hat{h}_i(1-\hat{h}_i)}{N_i} \\ &\cong \frac{N_i - D_i}{N_i D_i} \end{aligned}$$

$$\begin{aligned} Var(ln(1-\hat{h}_i)) &\cong \frac{1}{(1-\hat{h}_i)^2}\frac{\hat{h}_i(1-\hat{h}_i)}{N_i} \\ &\cong \frac{D_i}{N_i(N_i - D_i)} \end{aligned}$$

We have a formula for the exponential function:

$$Var(\hat{S}_{i-1}\hat{h}_i) = (\hat{S}_{i-1}\hat{h}_i)^2[Var(ln\hat{S}_{i-1}) + (ln\hat{h}_i)]$$

Therefore:

$$\widehat{Var(\hat{S}_{i-1}}\hat{h}_i) = (\hat{S}_{i-1}\hat{h}_i)^2[\sum_{k=1}^{i-1}\frac{D_k}{N_k(N_k - D_k)} + \frac{N_i - D_i}{N_i D_i}] \tag{1.32}$$

Chapter 2

The Confidence Interval of Cross-Sectional Distribution of Completed Lifetimes and the Pairs Bootstrap

2.1 Introduction and Literature Review

2.1.1 Introduction

Dixon (2012) and Dixon and Bihan (2012b) introduce the general Taylor model into the monetary economics. In the general Taylor model, there exists a cross-sectional distribution named as the distribution of the completed lifetimes (*DCL*). Since the *DCL* is a new distribution, there is no method to investigate the confidence interval for it. The confidence interval of *DCL* is very important since the confidence interval can be applied to check whether the estimator of the *DCL* is rejected or not. In this chapter, we introduce the numerical method-Fieller's method to construct the confidence interval of the *DCL*. We also introduce

the delta method to derive the confidence interval of the *DCL*. In addition, the bootstrap Fieller's method and bootstrap delta method are also introduced and applied to construct the confidence interval of the *DCL*. Monte Carlo simulation is applied to compare the accuracy of those methods. Another contribution of this chapter is to investigate whether those confidence intervals of the *DCL* obtained from different methods are valid or not.

2.1.2 Literature Review

The confidence interval of the *DCL* and age distribution can be calculated by Fieller (1954)'s method and delta method. In chapter 1, the variance of the distribution of duration, age distribution and the *DCL* have been derived. In chapter 2, the variance formulas are applied in constructing the confidence interval of *DCL*. In Fieller's method, the ratio variable is transformed to a linear function. The confidence interval of the ratio variable can be obtained by solving out the linear function. Fieller (1932) investigated and derived the general cumulative distribution formula for the ratio distribution $w = \frac{x}{y}$. Both x and y followed the normal distribution and correlated with each other. The skulls of those ratio distributions were plotted. Fieller (1954) focused on the distribution of the ratio $w = \frac{x}{y}$ where x and y were independent with each other. Since *DCL* is a ratio distribution, it is worth to introduce the Fieller's method to construct the confidence interval of it. With respect to the delta method, it is a robust method even the distribution of the ratio variable is unknown.

Alternatively, the probably density function of the *DCL* can be derived directly. Marsaglia (1965) studied the density function of the ratio variable. The ratio variable can be written as $m = \frac{c+x}{d+y}$. Both x and y were random variables following the normal distribution; c and d were constant variables. Cedilnik et al. (2004) derived the probability density of the ratio variable $w = x/y$ when x and y followed the bivariate normal distributions. They focused on the shape of the density of the distribution w. They defined the shape estimator and investigated it under three conditions: (a) the sign of shape estimator was positive; (b) the

sign of shape estimator was negative; (c) shape estimator was equal to zero. They gave the density function when x and y were perfectly correlated with each other (the correlation coefficient is equal to 1 or -1). In terms of the existence of the moment of the ratio Variable, see Cedilnik et al. (2006). However, it is hard to derive the density function of the *DCL* since the expression of the *DCL* is complicated. Therefore, we focus on the confidence interval rather than the density function of *DCL*.

Comparing with the traditional numerical method, the bootstrap method introduced by Efron (1979) is used widely in the statistical inference. This method was applied by Efron (1981) in the survival analysis when right-censored observations could exist. The bootstrap sample was applied to calculate the variance of Kaplan-Meier (KM) estimator. Those results were compared with the Greenwood formula which was the analytic variance formula of the KM estimator. Efron (1981) showed that the bootstrap variance of the KM estimator calculated from the bootstrap sample was close to the value of the Greenwood formula. Therefore, we also introduce bootstrap method to investigate the confidence interval of *DCL*.

Hwang (1995) introduced the bootstrap method into the ratio variable. The bootstrap method could be applied to derive the confidence interval of the ratio variable. The numerator and denominator could follow any distributions. Both the parametric and non-parametric bootstrap method were applied to construct the confidence interval. In addition, the empirical sizes were evaluated between the Fieller's method and the bootstrap method. Both methods provided the accurate empirical sizes. When the variables in the ratio formula do not follow the normal distributions, the bootstrap method can be applied to improve it.

There are some studies which focus on the comparison of the Fieller's method with any other numerical methods. As Polsky et al. (1997) and Briggs et al. (1999) showed, the Fieller's method and the parametric bootstrap method were suitable for constructing the confidence interval of the ratio variable. Fan and Zhou (2007) suggested that the Fieller's method, the standard bootstrap method and the bootstrap percentile method provided quite

accurate confidence intervals of the ratio variables when the numerator and denominator followed the different distributions. Depending on the simulation work of Wang and Zhao (2008), they suggested that the bootstrap Fieller's method provided more accurate confidence interval of the ratio variable. In Bebu et al. (2016)'s simulation studies, the Fieller's method provided a more accurate confidence interval of the ratio variable than the remaining methods. Also see Cox (1990) and Gardiner et al. (2001). In this chapter, those methods are applied to investigate the confidence interval of the *DCL*.

2.2 The Properties of Age Distribution and *DCL*

Breslow and Crowley (1974) showed that the survival function $\hat{S}_i$ with $i = 0,1,2,.....,F$ followed the normal distribution, the vector $V = (\hat{S}_0, \hat{S}_1, ..., \hat{S}_F)$ follows the asymptotic multivariate normal distribution: $V \overset{a.s.}{\sim} MN(E(V), \Sigma)$. Where Σ is the variance-covariance matrix for the vector V, and the vector $E(V)$ is the mean value of each element in vector V. With respect to the age distribution, it can be written as $\hat{a}_i^A = \frac{\hat{S}_i}{\sum_{k=0}^{F} \hat{S}_k}$. We define a new variable $\hat{S} = \sum_{k=0}^{F} \hat{S}_k$. $\hat{S}$ follows the asymptotic normal distribution $\hat{S} \overset{a.s.}{\sim} N(\mu_{\hat{S}}, \sigma_{\hat{S}}^2)$. Now $\mu_{\hat{S}}$ is $E(\hat{S})$ and $\sigma_{\hat{S}}^2$ is the variance of $\hat{S}$. The age distribution is a ratio distribution $\hat{a}_i^A = \frac{\hat{S}_i}{\hat{S}}$. There exists the relationship that $\hat{S}_i$ and $\hat{S}$ follow the asymptotic multivariate normal distribution:

$$\begin{bmatrix} \hat{S}_i \\ \hat{S} \end{bmatrix} \overset{a.s.}{\sim} MN\left(\begin{bmatrix} \mu_{\hat{S}_i} \\ \mu_{\hat{S}} \end{bmatrix}, \begin{bmatrix} \sigma_{\hat{S}_i}^2 & \sigma_{\hat{S}_i,\hat{S}} \\ \sigma_{\hat{S}_i,\hat{S}} & \sigma_{\hat{S}_i}^2 \end{bmatrix} \right)$$

Where $\mu_{\hat{S}_i}$ is the mean value of the $\hat{S}_i$; $\mu_{\hat{S}} = \mu_{\hat{S}_1} + \mu_{\hat{S}_2} + ... + \mu_{\hat{S}_F}$; the variance of $Var(\hat{S}_i) = \sigma_{\hat{S}_i}^2$; the variance of $\hat{S}$ is $Var(\hat{S}) = \sigma_{\hat{S}}^2$; $Var(\hat{S}) = \sigma^2 = \sum_{k=1}^{F} \sigma_{\hat{S}_k}^2 + 2\sum_{k=1}^{F} \sum_{j=1}^{F} \sigma_{\hat{S}_k,\hat{S}_j}^2$ and $k \neq j$; $\sigma_{\hat{S}_i,\hat{S}}$ is the covariance of $\hat{S}_i$ and $\hat{S}$; $Cov(\hat{S}_i, \hat{S}) = \sigma_{\hat{S}_i,\hat{S}} = \sum_{k=1}^{F} Cov(\hat{S}_i, \hat{S}_k)$. The variance and the covariance formula of the survival function are found from equation (2.1) and (2.2).

The survival function can be known as the Kaplan-Meier estimator (KM estimator).The variance of the KM estimator was derived by Greenwood (1926). The Greenwood formula of the KM estimator can be written as:

$$\widehat{Var(\hat{S}_i)} = \hat{S}_i^2[\sum_{k=1}^{i} \frac{D_k}{N_k(N_k - D_k)}] \tag{2.1}$$

It is crucial to investigate the covariance among the survival functions. It is clear that the survival function $\hat{S}_i$ is correlated with $\hat{S}_j$. Breslow and Crowley (1974) investigated the large sample properties of the hazard function and the survival function. They found that the off-diagonal variance-covariance matrix of the hazard functions are all equal to zero. If the hazard function $\hat{h}_i$ with $i = 1, 2, ..., F$ are collected by the vector $\hat{h}$, the joint distribution of the vector h follows the Gaussian distribution asymptotically. They also show that the vector of survival function $V = (\hat{S}_0, ...\hat{S}_F)$ converge weakly to the Gaussian process. They derived the asymptotic covariance of the survival functions between different periods. Tsai et al. (1987) wrote a literature review to discuss the covariance properties of KM estimators. In chapter 1, the Taylor expansion method have been applied to derive the covariance of the KM estimators in another way:

$$\widehat{Cov}(\hat{S}_i, \hat{S}_j) = \hat{S}_i\hat{S}_j[\sum_{k=1}^{i} \frac{D_k}{N_k(N_k - D_k)}] \; for \; i < j \tag{2.2}$$

Define the estimator of the *DCL* as $\hat{a}_i = \frac{i\hat{S}_{i-1}\hat{h}_i}{\sum_{k=0}^{F}\hat{S}_k}$. The distribution of duration can be defined as $\hat{a}_i^d = \hat{S}_{i-1}\hat{h}_i$. Since $\hat{S}_{i-1}\hat{h}_i = \hat{S}_{i-1} - \hat{S}_i$, both $\hat{S}_i$ and the summation of the survival function $\hat{S} = \sum_{k=0}^{F}\hat{S}_k$ follow the asymptotic normal distribution.In addition, $\hat{S}_{i-1} - \hat{S}_i$ and $\hat{S}$ follow the multivariate normal distribution asymptotically:

$$\begin{bmatrix} \hat{S}_{i-1}\hat{h}_i \\ \hat{S} \end{bmatrix} \overset{a.s.}{\sim} MN\left(\begin{bmatrix} \mu_{\hat{S}_{i-1}\hat{h}_i} \\ \mu_{\hat{S}} \end{bmatrix}, \begin{bmatrix} \sigma^2_{\hat{S}_{i-1}\hat{h}_i} & \sigma_{\hat{S}_{i-1}h_i,\hat{S}} \\ \sigma_{\hat{S}_{i-1}\hat{h}_i,\hat{S}} & \sigma^2_{\hat{S}} \end{bmatrix} \right)$$

Where $\sigma^2_{\hat{S}_{i-1}\hat{h}_i} = Var(\hat{S}_{i-1}\hat{h}_i)$; $\sigma_{\hat{S}_{i-1}\hat{h}_i,\hat{S}}$ is the covariance between $\hat{S}_{i-1}\hat{h}_i$ and $\hat{S}$; $Cov(i\hat{S}_{i-1}\hat{h}_i,\hat{S}) = i\sigma_{\hat{S}_{i-1}\hat{h}_i,\hat{S}} = i[\sum_{k=1}^{F} Cov(\hat{S}_{i-1},\hat{S}_k) - \sum_{k=1}^{F} Cov(\hat{S}_i,\hat{S}_k)]$.

The covariance between $i\hat{S}_{i-1}\hat{h}_i$ and $\hat{S}_j$ can be derived as:

$$Cov(i\hat{S}_{i-1}\hat{h}_i,\hat{S}_j) = iCov(\hat{S}_{i-1}-\hat{S}_i,\hat{S}_j) = i[Cov(\hat{S}_{i-1},\hat{S}_j) - Cov(\hat{S}_i,\hat{S}_j)]$$

In other words, the $Cov(i\hat{S}_{i-1}\hat{h}_i,\hat{S}_j)$ can be transformed to the covariance of $Cov(i\hat{S}_{i-1}\hat{h}_i,\hat{S})$:

$$\begin{aligned} Cov(i\hat{S}_{i-1}\hat{h}_i,\sum_{k=1}^{F}\hat{S}_k) &= i[Cov(\hat{S}_{i-1},\sum_{k=1}^{F}\hat{S}_k) - Cov(\hat{S}_i,\sum_{k=1}^{F}\hat{S}_k)] \\ &= i[\sum_{k=1}^{F} Cov(\hat{S}_{i-1},\hat{S}_k) - \sum_{k=1}^{F} Cov(\hat{S}_i,\hat{S}_k)] \end{aligned}$$

Note that $Cov(\hat{S}_i,\hat{S}) = \sigma_{\hat{S}_i,\hat{S}} = \sum_{k=1}^{F} Cov(\hat{S}_i,\hat{S}_k)$ where $\hat{S} = (S_0,S_1,.....S_F)$. Using equation (2.2), we can calculate $Cov(\hat{S}_i,\hat{S})$.

The variance of *DCL* is derived in chapter 1 which can be written as:

$$\widehat{Var(\hat{a}_i)} = i^2\frac{\widehat{Var(\hat{S}_{i-1}\hat{h}_i)}}{\hat{S}^2} + i^2\frac{\hat{S}^2_{i-1}\hat{h}^2_i\widehat{Var(\hat{S})}}{\hat{S}^4} - 2i^2\frac{\hat{S}_{i-1}\hat{h}_i\widehat{Cov}(\hat{S}_{i-1}\hat{h}_i,\hat{S})}{\hat{S}^3} \tag{2.3}$$

After holding those properties, Fieller's method can be applied to construct the confidence interval.

2.3 Confidence Interval for the *DCL*

2.3.1 The Fieller's Method

In this section, the confidence interval (CI) of *DCL* is given. The *DCL* can be written as $a_i = \frac{iS_{i-1}h_i}{\sum_{k=0}^{F} S_k}$ which is known as a ratio distribution. Define the numerator $iS_{i-1}h_i = x_i$ and the denominator $\sum_{k=0}^{F} S_k = y$, the *DCL* can be written as $a_i = x_i/y$ at for the i-th period.

In the *DCL* formula, the denominator is always above zero. Normally, the mean value of the summation of survival function should be a positive value. If the denominator is close to zero, the CI for x_i/y is not accurate.

Fieller (1940, 1954) gave a method to derive the confidence interval of the ratio of two random variables. In Fieller's method, it requires that the numerator and the denominator follow the bivariate normal distribution asymptotically. Franz (2007) gave a guidance to use the Fieller's method. In the *DCL* formula, there exist the relationships between $\hat{x}_i$ and $\hat{y}$:

$$\begin{bmatrix} \hat{x}_i \\ \hat{y} \end{bmatrix} \overset{a.s.}{\sim} MN\left(\begin{bmatrix} x_i \\ y \end{bmatrix}, \begin{bmatrix} \sigma^2_{\hat{x}_i} & \sigma_{\hat{x}_i\hat{y}} \\ \sigma_{\hat{x}_i\hat{y}} & \sigma^2_{\hat{y}} \end{bmatrix} \right)$$

Therefore, the $a_i = x_i/y$ satisified the requirement of Fieller's theorem. The *DCL* $a_i = x_i/y$ can be modified as:

$$x_i = a_i y$$

As Fieller had explained, the bivariate distribution function $v = f(x_i, y)$ with a given value a_i existed in the linear function $x_i = a_i y$. It can be rewritten as:

$$x_i - a_i y = 0$$

Replace the x_i and y by the estimator $\hat{x}_i$ and $\hat{y}$ [1],separately. Now there exists the new relationship:

$$\hat{x}_i - a_i\hat{y} \overset{a.s.}{\sim} N(0, Var(\hat{x}_i - a_i\hat{y}))$$

Thus a new statistic can be written as:

$$H(a_i) = \frac{\hat{x}_i - a_i\hat{y}}{(a_i^2\sigma^2_{\hat{y}} - 2a_i\sigma_{\hat{x}_i\hat{y}} + \sigma^2_{\hat{x}_i})^{\frac{1}{2}}} \tag{2.4}$$

[1]The estimator $\hat{x}_i = i\hat{S}_{i-1}\hat{h}_i$, and the estimator $\hat{y} = \sum_{k=0}^{F}\hat{S}_k$

Where $H(a_i)$ follows the student-t distribution with the degree of freedom d. If both $\hat{x}_i$ and $\hat{y}$ are approximated as the normal distribution asymptotically, the degree of freedom d can be assumed as infinite. As Buonaccorsi (2005) showed, there exists a relationship:

$$P[H(a_i) < t_{1-\alpha/2}(d)] = 1 - \alpha/2$$

Where α is the type I error; $t_{1-\alpha/2}(d)$ is the critical value in the student-t table; d is the degree of freedom. There exists the relationship:

$$P[L(a_i) < 0] = 1 - \alpha/2$$

The function $L(a_i)$ can be defined as:

$$L(a_i) = (\hat{x}_i - a_i\hat{y})^2 - t_{1-\alpha/2}(d)^2(a_i^2\sigma_{\hat{y}}^2 + 2a_i\sigma_{\hat{x}_i\hat{y}} + \sigma_{\hat{x}_i}^2) \tag{2.5}$$

Open the bracket, equation(2.5) can be written as:

$$L(a_i) = F_1 - 2F_2a_i + F_3a_i^2 \tag{2.6}$$

As can be seen from the formula (2.6), the function $L(a_i)$ is a quadratic form of a_i. Where $F_1 = \hat{x}_i^2 - t_{1-\alpha/2}(d)^2\sigma_{\hat{x}_i}^2$; $F_2 = \hat{x}_i\hat{y} - t_{1-\alpha/2}(d)^2\sigma_{\hat{x}_i\hat{y}}$; $F_3 = \hat{y}^2 - t_{1-\alpha/2}(d)^2\sigma_{\hat{y}}^2$. To find out the confidence interval of a_i, some notations need to be defined: $B = F_2^2 - F_1F_3$; $B_1 = (F_2 - B^{1/2})/F_3$; $B_2 = (F_2 + B^{1/2})/F_3$. As Buonaccorsi (2005) showed, the confidence interval of a_i can be written as $[B_1, B_2]$ when both F_2 and B are not less than zero. If B is not less than zero while $F_2 > 0$, the confidence interval is $(-\infty, B_2]$ and $[B_1, \infty)$. If both B and F_2 are less than zero, the confidence interval of a_i is $(-\infty, \infty)$. The confidence interval of a_i can be written as:

$$B_{1,2} = (F_2 \pm (F_2^2 - F_1F_3)^{1/2})/F_3 \tag{2.7}$$

By replacing $\hat{x}_i = i\hat{S}_{i-1}\hat{h}_i$ and $\hat{y} = \sum_{k=0}^{F}\hat{S}_k$ into F_1, F_2 and F_3, we can get:

$$F_1 = (i\hat{S}_{i-1}\hat{h}_i)^2 - t_{1-\alpha/2}(d)^2 i^2 Var(\hat{S}_{i-1}\hat{h}_i)$$

$$F_2 = i\hat{S}_{i-1}\hat{h}_i(\sum_{k=0}^{F}\hat{S}_k) - t_{1-\alpha/2}(d)^2 iCov(\hat{S}_{i-1}\hat{h}_i, \sum_{k=0}^{F}\hat{S}_k)$$

$$F_3 = (\sum_{k=0}^{F}\hat{S}_k)^2 - t_{1-\alpha/2}(d)^2 Var(\sum_{k=0}^{F}\hat{S}_k)$$

In the statistical area, researchers prefer the bounded confidence interval to the unbounded confidence interval. See Guiard (1989) for the unbounded confidence interval of the ratio variable.

2.3.2 Delta Method

Delta method can be also known as the Taylor expansion method. The KM estimator $\hat{S}_i$ follows the asymptotic normal distribution:

$$\sqrt{N_i}[\hat{S}_i - S_i] \overset{a.s.}{\sim} N(0, \sigma^2_{\hat{S}_i})$$

Where N_i is the sample size at i-th period; $\sigma^2_{\hat{S}_i}$ is the Greenwood formula of the KM estimator. Depending on the Taylor expansion:

$$\sqrt{N_i}[g(\hat{S}_i) - g(S_i)] \overset{a.s.}{\sim} N(0, [g'(S_i)]^2\sigma^2_{\hat{S}_i})$$

$g'(S_i)$ can be replaced by $g'(\hat{S}_i)$ since $[g'(S_i)]^2\sigma^2_{\hat{S}_i}$ is an estimator of $Var[g(\hat{S}_i)]$. At this point, the KM estimator is defined as $\hat{S}_i = exp(-\sum_{k=1}^{i}\hat{h}_i)$. The Taylor series of $g(\hat{S}_i)$ can be written

as:

$$g(\hat{S}_i) = g(S_i) + g'(S_i)(\hat{S}_i - S_i) + O_p((\hat{S}_i - S_i)^2)$$

Delta method is also available for the ratio distribution $\hat{a}_i = \hat{x}_i/\hat{y}$. The first-order Taylor expansion of $\hat{a}_i = \hat{x}_i/\hat{y}$ approximated at the $\hat{x}_i = x_i$ and $\hat{y} = y$ is:

$$\frac{\hat{x}_i}{\hat{y}} \approx \frac{x_i}{y} + \frac{\hat{x}_i - x_i}{y} - \frac{x_i}{y^2}(\hat{y} - y)$$

In terms of the variance:

$$Var(\frac{\hat{x}_i}{\hat{y}}) \approx \frac{1}{y^2}\sigma^2_{\hat{x}_i} + \frac{{x_i}^2}{y^4}\sigma_{\hat{y}^2} - 2\frac{x_i}{y^3}\sigma_{\hat{x}_i\hat{y}}$$

This variance formula can be rewritten as:

$$Var(\frac{\hat{x}_i}{\hat{y}}) \approx (\frac{x_i}{y})^2(\frac{\sigma^2_{\hat{x}_i}}{x^2} + \frac{\sigma^2_{\hat{y}}}{y^2} - 2\frac{\sigma_{\hat{x}_i\hat{y}}}{x_i y})$$

The 1-$\alpha/2$-th point-wise CI of a_i obtained from delta method can be defined as:

$$B_{1,2} = \hat{a}_i \pm t_{1-\alpha/2}(d)\sqrt{Var(\frac{\hat{x}_i}{\hat{y}})} \tag{2.8}$$

Where $Var(\frac{\hat{x}_i}{\hat{y}}) = Var(\hat{a}_i)$ can be calculated from equation (2.3). Even when the $\hat{x}_i/\hat{y}$ may not follow the normal distribution, it is worth to investigate the CI derived by delta method since delta method is robust. Depending on the estimator $\hat{a}_i$, the σ_{x_i} and σ_y are replaced by $\sigma_{\hat{x}_i}$ and $\sigma_{\hat{y}}$ in equation (2.8). In addition, x_i and y are replaced by $\hat{x}_i$ and $\hat{y}$, respectively. If the denominator is close to zero, the delta method is not available.

2.4 Bootstrap

The bootstrap method is a resampling method given by Efron (1979). The pairs bootstrap for the survival analysis is introduced by Efron (1981).

2.4.1 Pairs Bootstrap and Right-Censored Data

According to Efron (1981)'s paper, there exist the variable X_n obtained from the unknown population distribution **D** with n=$1,2,.....N$. The sample $X=(X_1,X_2,.....,X_N)$ follow the sample distribution $\hat{\mathbf{D}}$. The parameter $\theta(\mathbf{D})$ can be estimated by $\hat{\theta}(\hat{\mathbf{D}})$. In other words, $\hat{\theta}(\hat{\mathbf{D}})$ can be any estimators such as OLS, maximum likelihood, KM estimator or *DCL*. The bootstrap standard deviation of $\hat{\theta}$ can be defined as $\hat{\sigma}^*$. The detail of the non-parametric bootstrap process is: Step 1. Assign the equivalent probability $1/N$ to each X_n. Then the bootstrap sample $X^*=(X_1^*,X_2^*,.....,X_N^*)$ can be drawn from the sample $X=(X_1,X_2,.....,X_N)$ independently. More specially, the bootstrap sample $X^*=(X_1^*,X_2^*,.....,X_N^*)$ is resampled from the original sample $X=(X_1,X_2,.....,X_N)$ with $X_n^*=X_j$; $j=1,2,.....N$.. After that, the bootstrap empirical distribution $\hat{\mathbf{D}}^*$ can be generated. Step 2. Depending on the bootstrap sample X^* and distribution $\hat{\mathbf{D}}^*$, we can get the bootstrap estimator $\hat{\theta}^*$. To repeat the bootstrap process of the pairs observations for B times, there exist the number of B estimators $\hat{\theta}_b^*$ for $b=1,2,.....,B$. The bootstrap standard deviation can be calculated as below:

$$\sigma_{bootstrap}=\sqrt{\sum_{b=1}^{B}\frac{(\hat{\theta}_b^*-\sum_{b=1}^{B}\hat{\theta}_b^*/B)^2}{B-1}} \tag{2.9}$$

The non-parametric bootstrap method can be available for the pairs observations. Assume there exist the observation sample of pairs $(X,Y)=((X_1,Y_1),(X_2,Y_2),.....,(X_N,Y_N))$. In other words, a unique X_n corresponds to a unique Y_n. To bootstrap those pairs observations, the bootstrap sample (X^*,Y^*) can be directly resampled from the sample (X,Y). More specially, the observation of the pair (X_n,Y_n) in the sample (X,Y) is assigned by the equivalent

probability 1/N and resampled to obtain the bootstrap sample (X^*, Y^*). The bootstrap sample can be written as $(X^*, Y^*) = ((X_1^*, Y_1^*), (X_2^*, Y_2^*), ..., (X_N^*, Y_N^*))$ with (X_n^*, Y_n^*)=(X_j, Y_j); $j = 1, 2, ...N$. This method can be known as the pairs bootstrap method. The bootstrap empirical distribution $\hat{\mathbf{D}}^*$ can be generated by the bootstrap pairs sample (X^*, Y^*). The KM estimators and the *DCL* can be calculated from the bootstrapped sample. The formulas are given in the next section.

With respect to the survival observations, the maximum length of the events is assumed to last F periods. N_k means the total observations survived at the $k-th$ period with $k = 0, 1, ..., F$.[2] Define the T_n as the lifetime of the n-th event while $n = 1, 2, ...N$. The observed lifetime can be defined as:

$$t_n = min(T_n, C_n) \quad and \quad \omega_n = I(T_n \leq C_n) \quad n = 1, 2, ...N.$$

Where the C_n is the censored time; ω_n is the right-censored coefficient. If the n-th observation is uncensored, ω_n is equal to 1. Otherwise, it is equal to zero.

$$T_n \leq C_n,\ t_n = T_n\ \ (uncensored)\ \ \omega_n = 1$$

Otherwise,

$$C_n \leq T_n,\ t_n = C_n\ \ (right\ \ censored)\ \ \omega_n = 0$$

Now the bootstrap method can be linked to the survival analysis. We assume that the survival time follow the rule $t_1 < t_2 < ..... < t_N$. Therefore, the survival data can be rewritten as $((t_1, \omega_1), (t_2, \omega_2), ..., (t_N, \omega_N))$. If the right-censored observations are not considered, there exist the relationships $\omega_n = 1$ and the $t_n = T_n$ with $n = 1, 2, ...N$. The sample $((T_1, 1), (T_2, 1), ..., (T_N, 1))$ can be written as $(T_1, T_2, ..., T_N)$. The bootstrap sample can

[2]Note that $N_0 = N$ which is the total number of observations in the sample.

be written as $T_1^*, T_2^*, ..., T_N^*$. After that, repeat this process for B times, the bootstrap sample can be applied to calculate the KM estimator.

When the right-censored observations exist, some ω_n may not equal to 1. To bootstrap (t_n, ω_n): Step 1. The bootstrap sample can be written as $(T_1^*, \omega_1^*), (T_2^*, \omega_2^*), ..., (T_N^*, \omega_N^*)$. It means that the bootstrap sample is drawn independently from the original sample. It assigns the $1/N$ probability for each pair (t_n, ω_n); Step 2. The bootstrap sample is applied to calculate KM and *DCL* estimators. More details can be found in section 2.4.2; Step 3. Repeat the step 1 and 2 for B times to get the bootstrap KM and *DCL* estimators; Step 4. The bootstrap KM and *DCL* estimators are applied to calculate the bootstrap standard deviations and variances.

2.4.2 Bootstrap the Variance of the *DCL*

In this section, the bootstrap method is applied to derive the variance of *DCL* and compared with the benchmark value and the analytic results derived in chapter 1. Since the survival time is not exactly the same for each Monte Carlo simulation, the survival data can be rearranged and divided into special F-categories. In other words, the data can be assigned into some intervals such as $(0, r_1], (r_1, r_2], ...(r_{k-1}, r_k], ...(r_{F-1}, r_F]$. The value of r_k is depending on the value of the observation. After this, a new sample can be generated. Define the D_k to be the number of the observations in the k-th interval with $k = 1, 2, ..., F$. More specially, we can count how many observations are located in $(r_{k-1}, r_k]$ and define the number of them as D_k. N_k is the total number of the observations from interval $(r_{k-1}, r_k]$ to interval $(r_{F-1}, r_F]$. If all the observations are uncensored, we also have the relationships $N_i = \sum_{k=i}^{F} D_k$. Depending on this rule, we can bootstrap the sample and get the bootstrap value N_k^* and D_k^*. The bootstrap value N_k^* and D_k^* can be applied to calculate the bootstrap KM estimator and the Nelson-Aalen (NA) estimator. The bootstrap KM estimator can be written as:

$$\hat{S}_i^* = \prod_{k=1}^{i} \frac{N_k^* - D_k^*}{N_k^*} \tag{2.10}$$

There exist the relationships $D_k^* = N_k^* - N_{k+1}^*$ under the uncensored situation. The bootstrap NA estimator can be written as:

$$\hat{H}_i^* = \sum_{k=1}^{i} \frac{D_k^*}{N_k^*} \tag{2.11}$$

The $\hat{H}_i^*$ is the bootstrap NA estimator which is also known as the cumulative hazard function. With respect to the bootstrap version of the marginal hazard rate $\hat{h}_i^*$, it can be written as:

$$\hat{h}_i^* = \frac{D_i^*}{N_i^*} \tag{2.12}$$

Since we have both the bootstrap KM and NA estimators, they can be applied to calculate the distribution of duration, the age distributions and the *DCL*. The distribution of duration can be written as:

$$\hat{a}_i^d = \hat{S}_{i-1}\hat{h}_i$$

It can be explained as the proportion of observations survive for $(i-1)$ periods and die at the i-th period. Therefore, the bootstrap version of the distribution of the duration can be written as:

$$\hat{a}_i^{d*} = \hat{S}_{i-1}^*\hat{h}_i^* \tag{2.13}$$

The variance of the $\hat{a}_i^d$ is derived in chapter 1, it can be written as:

$$\widehat{Var(\hat{a}_i^d)} = (\hat{S}_{i-1}\hat{h}_i)^2[\sum_{k=1}^{i-1} \frac{D_k}{N_k(N_k - D_k)} + \frac{N_i - D_i}{N_i D_i}] \tag{2.14}$$

While for the bootstrap variance of the a_i^{d*}, it can be written as:

$$Var(\hat{a}_i^{d*}) = \sum_{b=1}^{B} [\hat{a}_{i,b}^{d*} - \frac{\sum_{b=1}^{B} \hat{a}_{i,b}^{d*}}{B}]^2/(B-1) \tag{2.15}$$

Where $\hat{a}^d_{i,b}$ is the distribution of duration in the b-th bootstrap re-sample process. B is the total number of the bootstrap resample process. The age distribution can be written as:

$$\hat{a}^A_i = \frac{\hat{S}_i}{\sum_{k=0}^{F} \hat{S}_k} \tag{2.16}$$

The bootstrap age distribution can be written as:

$$\hat{a}^{A*}_i = \frac{\hat{S}^*_i}{\sum_{k=0}^{F} \hat{S}^*_k} \tag{2.17}$$

With respect to the *DCL*:

$$\hat{a}_i = \frac{i\hat{S}_{i-1}\hat{h}_i}{\sum_{k=0}^{F} \hat{S}_k} \tag{2.18}$$

The bootstrap *DCL* can be written as:

$$\hat{a}^*_i = \frac{i\hat{S}^*_{i-1}\hat{h}^*_i}{\sum_{k=0}^{F} \hat{S}^*_k} \tag{2.19}$$

To calculate the bootstrap variance of the *DCL*, the formula is:

$$Var(\hat{a}^*_i) = \sum_{b=1}^{B} [\hat{a}^*_{i,b} - \frac{\sum_{b=1}^{B} \hat{a}^*_{i,b}}{B}]^2/(B-1) \tag{2.20}$$

It means that the bootstrap variance of the *DCL* a^*_i can be calculated from each bootstrap sample for $b = 1, 2, ..., B$. Define $\hat{S}^*_b = S^*_{0,b} + S^*_{1,b}, ... + S^*_{F,b}$, the variance formula of of the KM estimator in bootstrapped version can be written as:

$$Var(S^*_i) = \sum_{b=1}^{B} [\hat{S}^*_{i,b} - \frac{\sum_{b=1}^{B} \hat{S}^*_{i,b}}{B}]^2/(B-1) \tag{2.21}$$

Therefore:

$$Var(\hat{S}^*) = \sum_{b=1}^{B} [\hat{S}^*_b - \frac{\sum_{b=1}^{B} \hat{S}^*_b}{B}]^2/(B-1) \tag{2.22}$$

The covariance formula of the bootstrap version can be written as:

$$\begin{aligned} Cov(\hat{S}^*_{i-1}\hat{h}^*_i, \sum_{k=1}^{F}\hat{S}^*_k) &= Cov(\hat{S}^*_{i-1}, \sum_{k=1}^{F}\hat{S}^*_k) - Cov(\hat{S}^*_i, \sum_{k=1}^{F}\hat{S}^*_k) \\ &= \sum_{k=1}^{F} Cov(\hat{S}^*_{i-1}, \hat{S}^*_k) - \sum_{k=1}^{F} Cov(\hat{S}^*_i, \hat{S}^*_k) \end{aligned}$$

We have the bootstrap covariance formula:

$$Cov(\hat{S}^*_{i-1}, \hat{S}^*_k) = \sum_{b=1}^{B}[\hat{S}^*_{i-1,b} - \frac{\sum_{b=1}^{B}\hat{S}^*_{i-1,b}}{B}][\hat{S}^*_{k,b} - \frac{\sum_{b=1}^{B}\hat{S}^*_{k,b}}{B}]/(B-1) \tag{2.23}$$

2.4.3 The Bootstrap CI of the Ratio Variables

The bootstrap Fieller's method means that the bootstrap sample can be applied to construct the confidence interval of Fieller's method. The bootstrap Fieller's method was applied by Hwang (1995) and Wang and Zhao (2008). The bootstrap estimators of the $\hat{S}^*_{i-1}$ and $\hat{h}^*_i$ are applied. Define the $\hat{x}^*_i = i\hat{S}^*_{i-1}\hat{h}^*_i$ and $\hat{y}^* = \hat{S}^* = \sum_{k=0}^{F}\hat{S}^*_k$. Following Fieller's theorem, the bootstrap formula of the $H(a_i)^*$ can be written as:

$$H(a_i)^* = \frac{\hat{x}^*_i - \hat{a}_i\hat{y}^*}{(\hat{a}^2_i\sigma^2_{\hat{y}^*} + 2\hat{a}_i\sigma_{\hat{x}^*_i\hat{y}^*} + \sigma^2_{\hat{x}^*_i})^{\frac{1}{2}}} \tag{2.24}$$

$H(a_i)^*$ can be known as the bootstrap version of the $H(a_i)$. The bootstrap Fieller's method is applied to find out the $\alpha/2$-th and (1-α/2)-th smallest value of $H(a_i)^*$ in the bootstrap sample. Assume the original sample has been bootstrapped B times. The $H(a_i)^{\alpha/2*}$ and $H(a_i)^{1-\alpha/2*}$ are the $\alpha/2$-th and (1-α/2)-th smallest value of the bootstrap statistic $H(a_i)^*$, respectively. "$^{\alpha/2*}$" means the $\alpha/2$'s smallest value in the bootstrap sample. It should be mention that there exists the relationship $t_{1-\alpha/2}(d) = -t_{\alpha/2}(d)$. In other words, the ratio statistic satisfied $H(a_i)^{1-\alpha/2} = -H(a_i)^{\alpha/2}$. However, it may not exactly follow the symmetric rule when the sample size is not big enough. At this point, $H(a_i)^{1-\alpha/2*}$ is replaced by$-H(a_i)^{\alpha/2*}$ when

the upper bound of the confidence interval of a_i is calculated. There exists the relationship:

$$P[-H(a_i) < H(a_i)^{\alpha/2*}] = \alpha/2 \tag{2.25}$$

Alternatively, it can be expressed as:

$$P[H(a_i) > H(a_i)^{1-\alpha/2*}] = \alpha/2 \tag{2.26}$$

Solving out the quadratic formula by applying the same method in section 2.3.1, the bootstrap confidence interval of $\hat{a}_i$ can be constructed. The two sided bootstrapped confidence intervals from Fieller's method can be written as:

$$(B_1^*, B_2^*)$$

The lower bound B_1^* of the confidence interval $\hat{a}_i$ is:

$$B_1^* = (M_2 - (M_2^2 - M_1 M_3)^{1/2})/M_3 \tag{2.27}$$

$$M_1 = (i\hat{S}_{i-1}\hat{h}_i)^2 - (H(a_i)^{1-\alpha/2*})i^2 Var(\hat{S}_{i-1}\hat{h}_i)$$

$$M_2 = i\hat{S}_{i-1}\hat{h}_i(\sum_{k=0}^{F}\hat{S}_k) - (H(a_i)^{1-\alpha/2*})^2 iCov(\hat{S}_{i-1}\hat{h}_i, \sum_{k=0}^{F}\hat{S}_k)$$

$$M_3 = (\sum_{k=0}^{F}\hat{S}_k)^2 - (H(a_i)^{1-\alpha/2*})^2 Var(\sum_{k=0}^{F}\hat{S}_k)$$

The upper bound B_2^* of the confidence interval $\hat{a}_i$ is:

$$B_2^* = (M_5 + (M_5^2 - M_4 M_6)^{1/2})/M_5 \tag{2.28}$$

$$M_4 = (i\hat{S}_{i-1}\hat{h}_i)^2 - (H(a_i)^{\alpha/2*})i^2 Var(\hat{S}_{i-1}\hat{h}_i)$$

$$M_5 = i\hat{S}_{i-1}\hat{h}_i(\sum_{k=0}^{F}\hat{S}_k) - (H(a_i)^{\alpha/2*})^2 iCov(\hat{S}_{i-1}\hat{h}_i, \sum_{k=0}^{F}\hat{S}_k)$$

$$M_6 = (\sum_{k=0}^{F}\hat{S}_k)^2 - (H(a_i)^{\alpha/2*})^2 Var(\sum_{k=0}^{F}\hat{S}_k)$$

In the equation (2.27) ,the critical value of the t statistic $t_{1-\alpha/2}(d)$ is replaced by the bootstrap value $H(a_i)^{1-\alpha/2*}$. In terms of the equation (2.28), the $t_{\alpha/2}(d)$ is replaced by the bootstrap value $H(a_i)^{\alpha/2*}$. It should be mention that $H(a_i)^{\alpha/2*}$ applied to calculate the upper bound B_2^* due to the bootstrap properties. The bootstrapped-t value may not be the symmetric case. In other words, it is exactly different from the t distribution when the sample size is small. Therefore, both the $H(a_i)^{\alpha/2*}$ and $H(a_i)^{1-\alpha/2*}$ must be applied to construct the confidence intervals for the ratio a_i.

With respect to the delta method, the bootstrap-t confidence interval is also available. The ratio statistic derived from the delta method can be treated as:

$$H(a_i)^*_{delta} = \frac{\hat{a}_i^* - \hat{a}_i}{\sqrt{Var(\hat{a}_i^*)}} \tag{2.29}$$

There exists the relationship:

$$P[H(a_i)_{delta} < t_{1-\alpha/2}(d)] = 1 - \alpha/2$$

By applying the bootstrap-t percentile method, it can be rewritten as:

$$P[-H(a_i)_{delta} < H(a_i)_d^{\alpha/2*}] = \alpha/2$$

Alternatively, it can be expressed as:

$$P[H(a_i)_{delta} > H(a_i)_{delta}^{1-\alpha/2*}] = \alpha/2$$

Where $H(a_i)_{delta}^{\alpha/2*}$ and $H(a_i)_{delta}^{1-\alpha/2*}$ are the $\alpha/2$-th and (1-$\alpha/2$)-th smallest value in the bootstrap sample. The two sided confidence intervals are:

$$(B_{1,delta}^*, B_{2,delta}^*)$$

Where the formula for the $B_{1,delta}^*$ is:

$$B_{1,delta}^* = \hat{a}_i - H(a_i)_{delta}^{1-\alpha/2*} \sqrt{Var(\frac{\hat{x}_i}{\hat{y}})} \tag{2.30}$$

Where the formula for the $B_{2,delta}^*$ is:

$$B_{2,delta}^* = \hat{a}_i - H(a_i)_{delta}^{\alpha/2*} \sqrt{Var(\frac{\hat{x}_i}{\hat{y}})} \tag{2.31}$$

$Var(\frac{\hat{x}_i}{\hat{y}})$ is the variance of DCL which can be calculated from equation (2.3). $B_{1,delta}^*$ is the lower bound; $B_{2,delta}^*$ is the upper bound; Since the bootstrap-t critical value may not be the exactly symmetric relationship, the bootstrap-t critical value $H(a_i)_{delta}^{1-\alpha/2*}$ and $H(a_i)_{delta}^{\alpha/2*}$ need to be aoolied in the equation (2.30) and (2.31), respectively.

2.5 Monte Carlo and Pairs Bootstrap Simulation

2.5.1 The Pairs Bootstrap of the Variance

In this part, the Monte Carlo simulation and pairs bootstrap are applied generate and resample the sample. Those samples are applied to evaluate the analytic results of the variance, the bootstrapped variance and the benchmark value. The benchmark value of variance means the variance is calculated from the Monte Carlo simulation directly. In other words, the benchmark value of the variance of *DCL* can be treated as the true value of the variance. Following Kiviet and Phillips (2014)'s method, the true value of the variances for *DCL* can be defined as:

$$Var(\hat{a}_i)_{true} = \sum_{m=1}^{M} (\hat{a}_{i,m} - \frac{\sum_{m=1}^{M} \hat{a}_{i,m}}{M})^2/(M-1) \tag{2.32}$$

Where the $\hat{a}_{i,m}$ means the estimator of the *DCL* from the m-th simulation sample with $m = 1, 2, ..., M$. The true value of the variance of age distribution can be calculated by:

$$Var(\hat{a}_i^A)_{true} = \sum_{m=1}^{M} (\hat{a}_{i,m}^A - \frac{\sum_{m=1}^{M} \hat{a}_i^A}{M})^2/(M-1) \tag{2.33}$$

Where the $\hat{a}_{i,m}^A$ means the estimator of the age distribution from the m-th simulation sample with $m = 1, 2, ..., M$. The true value of the variance of the distribution of duration can be calculated by:

$$Var(\hat{a}_i^d)_{true} = \sum_{m=1}^{M} (\hat{a}_{i,m}^d - \frac{\sum_{m=1}^{M} \hat{a}_{i,m}^d}{M})^2/(M-1) \tag{2.34}$$

Where the $\hat{a}_{i,m}^d$ means the estimator of the distribution of duration from the m-th simulation sample with $m = 1, 2, ..., M$. To generate the survival data, we follow the steps as below: Step 1: The observations are assumed to follow the exponential distribution functions. The probability density function of lifetime of the n-th observation can be written as:

$$p(T_n) = 2exp(-2T_n) \tag{2.35}$$

The probability density function of the n-th censoring time can be written as:

$$p(C_n) = 0.5exp(-0.5C_n) \tag{2.36}$$

The survival function for each r_k-th period can be defined as:

$$p(C_n > r_k) = exp(-0.5r_k) \quad p(T_n > r_k) = exp(-2r_k) \tag{2.37}$$

Where $n = 1,2,.....,N$ and $k = 1,2,.....,F$. N is the total number of the observations in the sample and F is the maximum length of the period. If the uncensored observations exist, the survival time of observations T_k is applied to draw the sample directly. Otherwise, the T_k is compared with the C_k when the right-censored observations exist. In addition, those observations are assigned into five categories: $(0,0.1]$, $(0.1,0.2]$, $(0.2,0.3]$, $(0.3,0.5]$ and $(0.5,\infty)$. This can be known as case 1. In case 2, the observations are allocated in another five categories: $(0,0.2]$, $(0.2,0.4]$, $(0.4,0.6]$, $(0.6,0.8]$ and $(0.8,\infty)$. Therefore, the variance of the duration and the *DCL* can be calculated by the asymptotic formulas.

Step 2: The pairs bootstrap is applied to resample the original sample.

Step 3: Repeat the step 2 by $B = 1000$ times. After bootstrap $B = 1000$ times, the bootstrapped variance can be calculated directly from the bootstrap sample by applying the equation (2.15) and (2.20).

Step 4: Repeat step1, step 2 and step 3 for $M = 5000$ times. The true value of the variance can be calculated directly by by applying the equation (2.32) and (2.34).

It means that there are 5000 original samples. Each sample is bootstrapped by 1000 times. Therefore, there are 5000 bootstrap variances of the duration and the *DCL*.[3] The mean value of the bootstrap variance is calculated and compared with the true value and the analytic formula.

[3]The parameter i is ignored when the *DCL* is calculated since i is a constant.

Table 2.1 The Variance of the Distribution of Duration for Case 1 When the Uncensored Observations Exist in the Samples. All the Results Are Multiplied by 10^3 and the Sample Size N=50

True Value					
N	$Var(a^d_{0.1})$	$Var(a^d_{0.2})$	$Var(a^d_{0.3})$	$Var(a^d_{0.5})$	$Var(a^d_{\infty})$
50	3.0323	2.5584	2.1298	2.9756	4.6111
Asymptotic Approximation Value					
N	$E[\widehat{Var(a^d_{0.1})}]$	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.3})}]$	$E[\widehat{Var(a^d_{0.5})}]$	$E[\widehat{Var(a^d_{\infty})}]$
50	2.9014	2.4764	2.0986	2.9100	4.5568
Pairs Bootstrap Value					
N	$E[\widehat{Var(a^{d*}_{0.1})}]$	$E[\widehat{Var(a^{d*}_{0.2})}]$	$E[\widehat{Var(a^{d*}_{0.3})}]$	$E[\widehat{Var(a^{d*}_{0.5})}]$	$E[\widehat{Var(a^{d*}_{\infty})}]$
50	2.8839	2.4491	2.0544	2.8913	4.5362

Note: $Var(a^d_i)$ is the benchmark value calculated from formula (2.34); $E[\widehat{Var(a^d_i)}]$ is the variance of the distribution of duration calculated from formula (2.14); $E[\widehat{Var(a^{d*}_i)}]$ is the bootstrap variance of the distribution of duration calculated from formula (2.15).

Table 2.1 is the result for the variance of the duration. The sample size is chosen to be $N = 50$. The variances calculated from the analytic formula and the bootstrap variance of the distribution of duration is close to the benchmark true value.

Table 2.2 The Variance of the *DCL* for Case 1 When the Uncensored Observations Exist in the Samples. All the Results Are Multiplied by 10^3 and the Sample Size N=50

True Value					
N	$Var(a_{0.1})$	$Var(a_{0.2})$	$Var(a_{0.3})$	$Var(a_{0.5})$	$Var(a_{\infty})$
50	0.3703	0.2660	0.1970	0.2488	0.2308
Asymptotic Approximation Value					
N	$E[\widehat{Var(a_{0.1})}]$	$E[\widehat{Var(a_{0.2})}]$	$E[\widehat{Var(a_{0.3})}]$	$E[\widehat{Var(a_{0.5})}]$	$E[\widehat{Var(a^d_{\infty})}]$
50	0.3654	0.2634	0.1961	0.2424	0.2260
Pairs Bootstrap Value					
N	$E[\widehat{Var(a^*_{0.1})}]$	$E[\widehat{Var(a^*_{0.2})}]$	$E[\widehat{Var(a^*_{0.3})}]$	$E[\widehat{Var(a^*_{0.5})}]$	$E[\widehat{Var(a^{d*}_{\infty})}]$
50	0.3819	0.2688	0.1954	0.2419	0.2270

Note: $Var(a_i)$ is the benchmark value calculated from formula (2.32); $E[\widehat{Var(a_i}]$ is the variance of the *DCL* calculated from formula (2.3); $E[\widehat{Var(a^*_i)}]$ is the bootstrap variance of *DCL* calculated from formula (2.20).

With respect to the table 2.2, it is the empirical results of variance of the DCL. The asymptotic approximation variance provides an accurate result which is close to the benchmark

variance. The bootstrap variance still shows a good performance and provided an accurate result.

Table 2.3 The Variance of the Distribution of Duration for Case 2 When the Uncensored Observations Exist in the Samples. All the Results Are Multiplied by 10^3 and the Sample Size N=50

True Value					
N	$Var(a^d_{0.2})$	$Var(a^d_{0.4})$	$Var(a^d_{0.6})$	$Var(a^d_{0.8})$	$Var(a^d_{\infty})$
50	4.4982	3.4382	2.4551	1.7393	3.2032
Asymptotic Approximation Value					
N	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.4})}]$	$E[\widehat{Var(a^d_{0.6})}]$	$E[\widehat{Var(a^d_{0.8})}]$	$E[\widehat{Var(a^d_{\infty})}]$
50	4.3283	3.3684	2.4803	1.7644	3.1552
Pairs Bootstrap Value					
N	$E[\widehat{Var(a^{d*}_{0.2})}]$	$E[\widehat{Var(a^{d*}_{0.4})}]$	$E[\widehat{Var(a^{d*}_{0.6})}]$	$E[\widehat{Var(a^{d*}_{0.8})}]$	$E[\widehat{Var(a^{d*}_{\infty})}]$
50	4.3015	3.3551	2.4518	1.6998	3.1360

Note: $Var(a^d_i)$ is the benchmark value calculated from formula (2.34); $E[\widehat{Var(a^d_i)}]$ is the variance of the distribution of duration calculated from formula (2.14); $E[\widehat{Var(a^{d*}_i)}]$ is the bootstrap variance of the distribution of duration calculated from formula (2.15).

Table 2.4 The Variance of the *DCL* for Case 2 When the Uncensored Observations Exist in the Samples. All the Results Are Multiplied by 10^3 and the Sample Size N=50

True Value					
N	$Var(a_{0.2})$	$Var(a_{0.4})$	$Var(a_{0.6})$	$Var(a_{0.8})$	$Var(a_{\infty})$
50	1.2222	0.6283	0.3640	0.2299	0.2903
Asymptotic Approximation Value					
N	$E[\widehat{Var(a_{0.2})}]$	$E[\widehat{Var(a_{0.4})}]$	$E[\widehat{Var(a_{0.6})}]$	$E[\widehat{Var(a_{0.8})}]$	$E[\widehat{Var(a_{\infty})}]$
50	1.1981	0.6220	0.3659	0.2323	0.2827
Pairs Bootstrap Value					
N	$E[\widehat{Var(a^*_{0.2})}]$	$E[\widehat{Var(a^*_{0.4})}]$	$E[\widehat{Var(a^*_{0.6})}]$	$E[\widehat{Var(a^*_{0.8})}]$	$E[\widehat{Var(a_{\infty})}]$
50	1.2458	0.6313	0.3635	0.2220	0.2804

Note: $Var(a_i)$ is the benchmark value calculated from formula (2.32); $E[\widehat{Var(a_i}]$ is the variance of the *DCL* calculated from formula (2.3); $E[\widehat{Var(a^*_i)}]$ is the bootstrap variance of *DCL* calculated from formula (2.20).

Tale (2.3) and (2.4) are the empirical results for case 2. It can be seen that the approximation variance calculated from the analytic formula and the bootstrap method are nearly the same as the true variance. In addition, both the bootstrap method and the asymptotic

expansion method provide the accurate results of the variance. Table (2.5), (2.6), (2.7) and (2.8) show the empirical results of the variance when the right-censored observations exist. Depending on the empirical results, both the bootstrap variance and the variance calculated from the analytic formula are close to the benchmark variance. Therefore, both the bootstrap method and the delta method can be applied to calculate the variance of the distribution of duration and *DCL*.

Table 2.5 The Variance of the Distribution of Duration for Case 1 When the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3 and the Sample Size N=50

True Value					
N	$Var(a^d_{0.1})$	$Var(a^d_{0.2})$	$Var(a^d_{0.3})$	$Var(a^d_{0.5})$	$Var(a^d_{\infty})$
50	2.9092	2.6885	2.2432	3.3141	4.9893
Asymptotic Approximation Value					
N	$E[\widehat{Var(a^d_{0.1})}]$	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.3})}]$	$E[\widehat{Var(a^d_{0.5})}]$	$E[\widehat{Var(a^d_{\infty})}]$
50	2.8537	2.5708	2.2852	3.2736	4.9094
Pairs Bootstrap Value					
N	$E[\widehat{Var(a^{d*}_{0.1})}]$	$E[\widehat{Var(a^{d*}_{0.2})}]$	$E[\widehat{Var(a^{d*}_{0.3})}]$	$E[\widehat{Var(a^{d*}_{0.5})}]$	$E[\widehat{Var(a^{d*}_{\infty})}]$
50	2.8130	2.5166	2.2020	3.2107	4.8270

Note: $Var(a^d_i)$ is the benchmark value calculated from formula (2.34); $E[\widehat{Var(a^d_i)}]$ is the variance of the distribution of duration calculated from formula (2.14); $E[\widehat{Var(a^{d*}_i)}]$ is the bootstrap variance of the distribution of duration calculated from formula (2.15).

2.5.2 The Confidence Interval of *DCL*

In this section, the CI of the Fieller's method, Delta method, bootstrapped Fieller's method, bootstrapped delta method and percentile bootstrap CI have been investigated. The data generating process is the same as that of section 2.5.1:

Step 1: In this section, we focus on the CI rather than the variance. The data generating method of the section 2.5.1 is applied to generate the data. The sample sizes are $N = 25$, $N = 50$, $N = 200$ and $N = 400$. The Fieller's statistic $H(a_i)$ and the delta statistic $H(a_i)_{delta}$

Table 2.6 The Variance of the *DCL* for Case 1 When the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3 and the Sample Size N=50

True Value					
N	$Var(a_{0.1})$	$Var(a_{0.2})$	$Var(a_{0.3})$	$Var(a_{0.5})$	$Var(a_{\infty})$
50	0.3333	0.2651	0.1957	0.2686	0.2978
Asymptotic Approximation Value					
N	$E[\widehat{Var(a_{0.1})}]$	$E[\widehat{Var(a_{0.2})}]$	$E[\widehat{Var(a_{0.3})}]$	$E[\widehat{Var(a_{0.5})}]$	$E[\widehat{Var(a^d_{\infty})}]$
50	0.3311	0.2590	0.2030	0.2649	0.2991
Pairs Bootstrap Value					
N	$E[\widehat{Var(a^*_{0.1})}]$	$E[\widehat{Var(a^*_{0.2})}]$	$E[\widehat{Var(a^*_{0.3})}]$	$E[\widehat{Var(a^*_{0.5})}]$	$E[\widehat{Var(a^{d*}_{\infty})}]$
50	0.3438	0.2592	0.1976	0.2583	0.2892

Note: $Var(a_i)$ is the benchmark value calculated from formula (2.32); $E[\widehat{Var(a_i)}]$ is the variance of the *DCL* calculated from formula (2.3); $E[\widehat{Var(a^*_i)}]$ is the bootstrapped variance of the *DCL* calculated from formula (2.20).

Table 2.7 The Variance of the Distribution of Duration for Case 2 When the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3 and the Sample Size N=50

True Value					
N	$Var(a^d_{0.2})$	$Var(a^d_{0.4})$	$Var(a^d_{0.6})$	$Var(a^d_{0.8})$	$Var(a^d_{\infty})$
50	4.2584	3.6246	2.9629	2.2189	3.9568
Asymptotic Approximation Value					
N	$E[\widehat{Var(a^d_{0.2})}]$	$E[\widehat{Var(a^d_{0.4})}]$	$E[\widehat{Var(a^d_{0.6})}]$	$E[\widehat{Var(a^d_{0.8})}]$	$E[\widehat{Var(a^d_{\infty})}]$
50	4.2027	3.6509	2.9927	2.4578	3.8309
Pairs Bootstrap Value					
N	$E[\widehat{Var(a^{d*}_{0.2})}]$	$E[\widehat{Var(a^{d*}_{0.4})}]$	$E[\widehat{Var(a^{d*}_{0.6})}]$	$E[\widehat{Var(a^{d*}_{0.8})}]$	$E[\widehat{Var(a^{d*}_{\infty})}]$
50	4.1015	3.5506	2.8382	2.1796	3.5195

Note: $Var(a^d_i)$ is the benchmark value calculated from formula (2.34); $E[\widehat{Var(a^d_i)}]$ is the variance of the distribution of duration calculated from formula (2.14); $E[\widehat{Var(a^{d*}_i)}]$ is the bootstrap variance of the distribution of duration calculated from formula (2.15).

Table 2.8 The Variance of the *DCL* for Case 2 When the Right-Censored Observations Exist in the Samples. All the Results Are Multiplied by 10^3 and the Sample Size N=50

True Value					
N	$Var(a_{0.2})$	$Var(a_{0.4})$	$Var(a_{0.6})$	$Var(a_{0.8})$	$Var(a_{\infty})$
50	0.9954	0.5977	0.4060	0.2753	0.3829
Asymptotic Approximation Value					
N	$E[\widehat{Var(a_{0.2})}]$	$E[\widehat{Var(a_{0.4})}]$	$E[\widehat{Var(a_{0.6})}]$	$E[\widehat{Var(a_{0.8})}]$	$E[\widehat{Var(a_{\infty})}]$
50	1.0010	0.6215	0.4148	0.3040	0.3801
Pairs Bootstrap Value					
N	$E[\widehat{Var(a_{0.2})}]$	$E[\widehat{Var(a_{0.4})}]$	$E[\widehat{Var(a_{0.6})}]$	$E[\widehat{Var(a_{0.8})}]$	$E[\widehat{Var(a_{\infty})}]$
50	0.9851	0.5876	0.3793	0.2563	0.3255

Note: $Var(a_i)$ is the benchmark value calculated from formula (2.32); $E[\widehat{Var(a_i)}]$ is the variance of the *DCL* calculated from formula (2.3); $E[\widehat{Var(a_i^*)}]$ is the bootstrap variance of the *DCL* calculated from formula (2.20).

are calculated for each Monte Carlo simulation. The significant level α is chosen to be 10 percent.

Step 2: The pairs bootstrap is applied to get the bootstrap value from the Fieller's formula $H(a_i)^*$ and delta method $H(a_i)^*_{delta}$. The confidence interval of the bootstrap Fieller's method and delta method can be calculated by the bootstrap sample.

Step 3: Repeat the step 2 by $B = 1000$ times. Therefore, there exist 1000 $H(a_i)^*$. Rearrange the $H(a_i)^*$ from the smallest value to the largest value, and find out the $(\frac{\alpha}{2}B+1)$-th and $((1-\frac{\alpha}{2})B+1)$-th smallest value of $H(a_i)^*$. Since the significant level α is 10 percent, $(\frac{\alpha}{2}B+1)$-th smallest value of $H(a_i)^*$ is applied to be the lower bound and the $((1-\frac{\alpha}{2})B+1)$-th smallest value of $H(a_i)^*$ is applied to the upper bound of the confidence interval of *DCL*. The $((1-\frac{\alpha}{2})B+1)$-th smallest value of $H(a_i)^*$ can be known as $H(a_i)^{1-\alpha/2*}$. The $(\frac{\alpha}{2}B+1)$-th smallest value of $H(a_i)^*$ can be known as $H(a_i)^{\alpha/2*}$. In terms of the delta method, the bootstrapped version of the critical value $H(a_i)^{1-\alpha/2*}_{delta}$ and $H(a_i)^{\alpha/2*}_{delta}$ are calculated.

Step 4: Repeat step 1-3 by $M = 5000$ times when sample size $N = 25$ and 50. Repeat step 1-3 by $M = 1000$ when the sample size $N = 200$ and 400. Each Monte Carlo simulation is followed by the 1000 bootstrap simulations. After that, the simulation result can be applied

to calculate the average value of the lower bound and upper bound. The benchmark value of the CI can be calculated depending on the Monte Carlo method. The Monte Carlo results of a_i can be arranged from the smallest value to the largest value. After that, the $\alpha/2$-th and $1-\alpha/2$-th smallest value are chosen to be the benchmark value of lower bound and upper bound of the CI, separately.

Table (2.9), (2.10), (2.11) and (2.12) list the different confidence intervals of the *DCL* in different sample sizes. Table (2.9) shows the CI of the *DCL* when $N = 25$. The first column is the benchmark CI which is simulated from the Monte Carlo method. There are five methods to calculate the confidence interval of the *DCL*. Those confidence intervals are compared with the benchmark CI. Fieller's method always provides the accurate results for the upper bound of the confidence interval while there exist the bias in the lower bound. In the first three periods, Fieller's method is superior to the delta method while both methods perform well in the last two periods. In terms of the bootstrap method, they give the significant bias for the confidence interval of the *DCL*. It should be mention that there exists the negative value of the lower bound in case 2 when the right-censored observations exist in the samples. The negative value is due to the extremely small sample size in that category. Kalbfleisch and Prentice (2002) provided a log-log transformation method to derive the confidence interval of the KM estimator. Since the *DCL* is between 0 and 1, the log-log transferred confidence interval provided by Kalbfleisch and Prentice (2002) can be considered for *DCL* when the negative value of the confidence interval exists.

In table (2.10), the Fieller's method provides a quite accurate confidence interval over the five categories when the sample size is 50. The delta method does not work very well compared with the Fieller's method in the first three periods while both of them give the quite accurate results in the last two periods. In terms of the bootstrap method, the confidence intervals provided by bootstrap Fieller's and delta method are not accurate when the sample

Table 2.9 The 90% Confidence Interval of *DCL* with Sample Size N=25

Case 1 with Sample Size N=25. All the Observations Are Uncensored.

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.1}$	[0.02062, 0.10667]	[0.01646, 0.10584]	[0.01169, 0.09924]	[0.01804, 0.10485]	[0.01980, 0.10125]	[0.02402, 0.10724]
$a_{0.2}$	[0.01149, 0.08642]	[0.01047, 0.08541]	[0.00801, 0.08185]	[0.01521, 0.09703]	[0.01568, 0.09454]	[0.01829, 0.08632]
$a_{0.3}$	[0.01124, 0.07317]	[0.00637, 0.07166]	[0.00551, 0.07001]	[0.01111, 0.07711]	[0.01123, 0.07629]	[0.01778, 0.07293]
$a_{0.5}$	[0.02105, 0.09091]	[0.01722, 0.08849]	[0.01757, 0.08804]	[0.02040, 0.10063]	[0.01913, 0.09911]	[0.02388, 0.08955]
a_{∞}	[0.06849, 0.13953]	[0.06773, 0.13807]	[0.07140, 0.14067]	[0.07362, 0.14051]	[0.07430, 0.14208]	[0.06716, 0.13465]

Case 1 with Sample Size N=25. The Right-Censored Observations Exist in the Samples.

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile bootstrap*
$a_{0.1}$	[0.01107, 0.09598]	[0.01381, 0.09569]	[0.00965, 0.08980]	[0.01253, 0.10774]	[0.01529, 0.10336]	[0.01746, 0.09254]
$a_{0.2}$	[0.01122, 0.08136]	[0.00865, 0.08201]	[0.00604, 0.07816]	[0.01502, 0.09456]	[0.01554, 0.09174]	[0.01761, 0.07792]
$a_{0.3}$	[0.01157, 0.06855]	[0.00386, 0.06949]	[0.00272, 0.06745]	[0.01013, 0.07323]	[0.01039, 0.07230]	[0.01808, 0.06546]
$a_{0.5}$	[0.01414, 0.08784]	[0.01234, 0.08584]	[0.01221, 0.08479]	[0.01954, 0.09457]	[0.01953, 0.09448]	[0.02143, 0.08235]
a_{∞}	[0.04076, 0.11813]	[0.03875, 0.11882]	[0.04102, 0.12001]	[0.04947, 0.12869]	[0.04938, 0.12980]	[0.04111, 0.10716]

Case 2 with Sample Size N=25. All the Observations Are Uncensored.

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.2}$	[0.05714, 0.21429]	[0.05971, 0.22160]	[0.04882, 0.20544]	[0.06357, 0.23158]	[0.06549, 0.22440]	[0.05835, 0.21162]
$a_{0.4}$	[0.03125, 0.14754]	[0.03192, 0.14697]	[0.02808, 0.14064]	[0.03889, 0.16783]	[0.03921, 0.16468]	[0.03547, 0.14265]
$a_{0.6}$	[0.01613, 0.10448]	[0.01413, 0.10203]	[0.01392, 0.10025]	[0.02122, 0.11206]	[0.02113, 0.11164]	[0.02495, 0.10081]
$a_{0.8}$	[0.01429, 0.07692]	[0.00339, 0.07491]	[0.00499, 0.07522]	[0.00949, 0.07555]	[0.00831, 0.07500]	[0.02223, 0.07458]
a_{∞}	[0.03448, 0.11268]	[0.03078, 0.10907]	[0.03632, 0.11267]	[0.04023, 0.11550]	[0.04077, 0.11843]	[0.03742, 0.10768]

Case 2 with Sample Size N=25. The Right-Censored Observations Exist in the Samples

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.2}$	[0.04773, 0.17778]	[0.04679, 0.18681]	[0.03760, 0.17264]	[0.05635, 0.21014]	[0.05810, 0.20098]	[0.04691, 0.15944]
$a_{0.4}$	[0.02620, 0.12713]	[0.02213, 0.13103]	[0.01772, 0.12372]	[0.03480, 0.15870]	[0.03531, 0.15360]	[0.02915, 0.11220]
$a_{0.6}$	[0.01593, 0.09521]	[0.00719, 0.09678]	[0.00593, 0.09357]	[0.01884, 0.10451]	[0.01908, 0.10344]	[0.02379, 0.08378]
$a_{0.8}$	[0.01623, 0.07427]	[−0.00153, 0.07848]	[−0.00081, 0.07755]	[0.00938, 0.07230]	[0.00939, 0.07250]	[0.02411, 0.06578]
a_{∞}	[0.02154, 0.09413]	[0.00964, 0.09728]	[0.01315, 0.09883]	[0.02512, 0.09997]	[0.02505, 0.10143]	[0.02940, 0.08132]

Table 2.10 The 90% Confidence Interval of *DCL* with Sample Size N=50

Case 1 with Sample Size N=50. All the Observations Are Uncensored.

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.1}$	[0.02632, 0.08861]	[0.02579, 0.08775]	[0.02338, 0.08474]	[0.02602, 0.08828]	[0.02685, 0.08708]	[0.02691, 0.09022]
$a_{0.2}$	[0.02051, 0.07333]	[0.01925, 0.07190]	[0.01796, 0.07024]	[0.02195, 0.08082]	[0.02217, 0.07994]	[0.02112, 0.07392]
$a_{0.3}$	[0.01563, 0.06061]	[0.01391, 0.05936]	[0.01343, 0.05861]	[0.01674, 0.06731]	[0.01681, 0.06692]	[0.01652, 0.06080]
$a_{0.5}$	[0.02874, 0.08046]	[0.02775, 0.07890]	[0.02784, 0.07870]	[0.02905, 0.08681]	[0.02840, 0.08601]	[0.02873, 0.08056]
a_{∞}	[0.08219, 0.13158]	[0.08090, 0.13065]	[0.08263, 0.13202]	[0.08301, 0.13013]	[0.08344, 0.13073]	[0.08139, 0.13083]

Case 1 with Sample Size N=50. The Right-Censored Observations Exist in the Samples.

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.1}$	[0.02520, 0.08429]	[0.02419, 0.08320]	[0.02195, 0.08035]	[0.02357, 0.08645]	[0.02459, 0.08528]	[0.02566, 0.08500]
$a_{0.2}$	[0.01735, 0.07021]	[0.01745, 0.06957]	[0.01608, 0.06777]	[0.02059, 0.07923]	[0.02084, 0.07825]	[0.01824, 0.07006]
$a_{0.3}$	[0.01244, 0.05913]	[0.01174, 0.05787]	[0.01113, 0.05695]	[0.01501, 0.06631]	[0.01513, 0.06583]	[0.01436, 0.05868]
$a_{0.5}$	[0.02416, 0.07832]	[0.02320, 0.07651]	[0.02308, 0.07606]	[0.02636, 0.08439]	[0.02635, 0.08434]	[0.02460, 0.07713]
a_{∞}	[0.05650, 0.11335]	[0.05595, 0.11315]	[0.05714, 0.11394]	[0.05853, 0.11685]	[0.05841, 0.11715]	[0.05559, 0.11052]

Case 2 with Sample Size N=50. All the Observations Are Uncensored.

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.2}$	[0.07639, 0.19130]	[0.07786, 0.19163]	[0.07207, 0.18399]	[0.07793, 0.19344]	[0.07883, 0.19143]	[0.07757, 0.19380]
$a_{0.4}$	[0.04698, 0.12931]	[0.04660, 0.12854]	[0.04446, 0.12549]	[0.04915, 0.13682]	[0.04930, 0.13599]	[0.04770, 0.12973]
$a_{0.6}$	[0.02778, 0.09023]	[0.02597, 0.08875]	[0.02572, 0.08793]	[0.02938, 0.09790]	[0.02933, 0.09770]	[0.02847, 0.09013]
$a_{0.8}$	[0.01481, 0.06429]	[0.01259, 0.06244]	[0.01324, 0.06264]	[0.01581, 0.07108]	[0.01521, 0.07062]	[0.01700, 0.06409]
a_{∞}	[0.04762, 0.10345]	[0.04561, 0.10158]	[0.04832, 0.10359]	[0.04966, 0.10524]	[0.04992, 0.10634]	[0.04702, 0.10206]

Case 2 with Sample Size N=50. The Right-Censored Observations Exist in the Samples.

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.2}$	[0.06702, 0.17022]	[0.06812, 0.17220]	[0.06283, 0.16506]	[0.06924, 0.18168]	[0.07041, 0.17961]	[0.06678, 0.16610]
$a_{0.4}$	[0.04002, 0.12051]	[0.03935, 0.12119]	[0.03680, 0.11754]	[0.04456, 0.13374]	[0.04484, 0.13248]	[0.03969, 0.11654]
$a_{0.6}$	[0.02069, 0.08728]	[0.02010, 0.08667]	[0.01937, 0.08521]	[0.02564, 0.09881]	[0.02573, 0.09835]	[0.02290, 0.08377]
$a_{0.8}$	[0.01117, 0.06658]	[0.00804, 0.06483]	[0.00839, 0.06458]	[0.01359, 0.07051]	[0.01355, 0.07072]	[0.01690, 0.06352]
a_{∞}	[0.02770, 0.09321]	[0.02670, 0.09127]	[0.02874, 0.09254]	[0.03409, 0.09916]	[0.03398, 0.10017]	[0.02919, 0.08701]

Table 2.11 The 90% Confidence Interval of *DCL* with Sample Size N=200 and 400. All the Observations Are Uncensored.

Case 1 with Sample Size N=200

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.1}$	[0.03927, 0.07176]	[0.03894, 0.06977]	[0.03831, 0.06907]	[0.03949, 0.06906]	[0.03957, 0.06891]	[0.03949, 0.07218]
$a_{0.2}$	[0.03125, 0.05714]	[0.03092, 0.05731]	[0.03058, 0.05692]	[0.03175, 0.05882]	[0.03179, 0.05875]	[0.03154, 0.05739]
$a_{0.3}$	[0.02479, 0.04741]	[0.02437, 0.04728]	[0.02423, 0.04711]	[0.02526, 0.04884]	[0.02527, 0.04881]	[0.02493, 0.04773]
$a_{0.5}$	[0.04067, 0.06676]	[0.04019, 0.06597]	[0.04018, 0.06593]	[0.04027, 0.06776]	[0.04010, 0.06758]	[0.04067, 0.06691]
a_{∞}	[0.09568, 0.12045]	[0.09506, 0.11999]	[0.09547, 0.12036]	[0.09610, 0.11919]	[0.09622, 0.11932]	[0.09559, 0.12032]

Case 1 with Sample Size N=400

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.1}$	[0.04313, 0.06405]	[0.04283, 0.06455]	[0.04251, 0.06421]	[0.04280, 0.06454]	[0.04285, 0.06449]	[0.04326, 0.06432]
$a_{0.2}$	[0.03486, 0.05342]	[0.03447, 0.05311]	[0.03430, 0.05292]	[0.03492, 0.05376]	[0.03493, 0.05374]	[0.03495, 0.05352]
$a_{0.3}$	[0.02802, 0.04367]	[0.02775, 0.04401]	[0.02768, 0.04392]	[0.02822, 0.04470]	[0.02822, 0.04469]	[0.02807, 0.04374]
$a_{0.5}$	[0.04425, 0.06172]	[0.04392, 0.06217]	[0.04391, 0.06215]	[0.04427, 0.06264]	[0.04427, 0.06264]	[0.04436, 0.06174]
a_{∞}	[0.09948, 0.11674]	[0.09897, 0.11660]	[0.09917, 0.11679]	[0.09901, 0.11670]	[0.09901, 0.11671]	[0.09939, 0.11670]

Case 2 with Sample Size N=200

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.2}$	[0.10124, 0.15619]	[0.09999, 0.15586]	[0.09847, 0.15411]	[0.10015, 0.15564]	[0.10028, 0.15542]	[0.10167, 0.15711]
$a_{0.4}$	[0.06471, 0.10669]	[0.06478, 0.10565]	[0.06418, 0.10494]	[0.06550, 0.10696]	[0.06553, 0.10690]	[0.06485, 0.10741]
$a_{0.6}$	[0.04035, 0.07214]	[0.04048, 0.07194]	[0.04038, 0.07176]	[0.04147, 0.07352]	[0.04147, 0.07351]	[0.04070, 0.07232]
$a_{0.8}$	[0.02549, 0.05047]	[0.02501, 0.05030]	[0.02513, 0.05037]	[0.02617, 0.05224]	[0.02616, 0.05226]	[0.02553, 0.05045]
a_{∞}	[0.06301, 0.09108]	[0.06241, 0.09055]	[0.06305, 0.09110]	[0.06312, 0.09148]	[0.06310, 0.09151]	[0.06277, 0.09097]

Case 2 with Sample Size N=400

DCL	*True CI*	*Fieller*	*Delta*	*bootstrapFieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.2}$	[0.10623, 0.14644]	[0.10709, 0.14646]	[0.10631, 0.14561]	[0.10709, 0.14641]	[0.10714, 0.14634]	[0.10644, 0.14701]
$a_{0.4}$	[0.07038, 0.09821]	[0.07045, 0.09936]	[0.07014, 0.09901]	[0.07083, 0.09997]	[0.07085, 0.09994]	[0.07049, 0.09829]
$a_{0.6}$	[0.04550, 0.06753]	[0.04519, 0.06750]	[0.04513, 0.06741]	[0.04571, 0.06817]	[0.04571, 0.06816]	[0.04566, 0.06762]
$a_{0.8}$	[0.02903, 0.04730]	[0.02877, 0.04670]	[0.02883, 0.04759]	[0.02936, 0.04754]	[0.02936, 0.04755]	[0.02899, 0.04731]
a_{∞}	[0.06725, 0.08676]	[0.06674, 0.08664]	[0.06705, 0.08693]	[0.06712, 0.08708]	[0.06711, 0.08709]	[0.06711, 0.08666]

Table 2.12 The 90% Confidence Interval of *DCL* with Sample Size N=200 and 400. There Exist the Right-Censored Observations in the Samples

Case 1 with Sample Size N=200

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.1}$	[0.03572, 0.06771]	[0.03668, 0.06604]	[0.03609, 0.06538]	[0.03750, 0.06502]	[0.03756, 0.06486]	[0.03580, 0.06801]
$a_{0.2}$	[0.02912, 0.05522]	[0.02920, 0.05545]	[0.02883, 0.05503]	[0.03007, 0.05699]	[0.03011, 0.05692]	[0.02939, 0.05544]
$a_{0.3}$	[0.02350, 0.04634]	[0.02265, 0.04610]	[0.02248, 0.04589]	[0.02361, 0.04781]	[0.02362, 0.04777]	[0.02365, 0.04664]
$a_{0.5}$	[0.03589, 0.06461]	[0.03640, 0.06342]	[0.03635, 0.06332]	[0.03720, 0.06474]	[0.03720, 0.06474]	[0.03602, 0.06466]
a_{∞}	[0.07095, 0.10152]	[0.07190, 0.10066]	[0.07218, 0.10089]	[0.07204, 0.10109]	[0.07201, 0.10112]	[0.07104, 0.10143]

Case 1 with Sample Size N=400

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.1}$	[0.04032, 0.06153]	[0.04030, 0.06095]	[0.04000, 0.06063]	[0.04012, 0.06110]	[0.04017, 0.06105]	[0.04053, 0.06181]
$a_{0.2}$	[0.03283, 0.05140]	[0.03269, 0.05121]	[0.03250, 0.05100]	[0.03316, 0.05191]	[0.03317, 0.05189]	[0.03289, 0.05149]
$a_{0.3}$	[0.02648, 0.04223]	[0.02583, 0.04238]	[0.02574, 0.04228]	[0.02634, 0.04314]	[0.02634, 0.04313]	[0.02646, 0.04232]
$a_{0.5}$	[0.03979, 0.05987]	[0.04026, 0.05937]	[0.04023, 0.05932]	[0.04067, 0.05996]	[0.04067, 0.05996]	[0.03983, 0.05995]
a_{∞}	[0.07671, 0.09738]	[0.07673, 0.09709]	[0.07686, 0.09721]	[0.07678, 0.09721]	[0.07677, 0.09722]	[0.07679, 0.09737]

Case 2 with Sample Size N=200

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.2}$	[0.08999, 0.14353]	[0.09109, 0.14336]	[0.08966, 0.14170]	[0.09083, 0.14412]	[0.09100, 0.14392]	[0.08978, 0.14383]
$a_{0.4}$	[0.05901, 0.10155]	[0.05912, 0.10069]	[0.05841, 0.09984]	[0.06002, 0.10232]	[0.06007, 0.10223]	[0.05927, 0.10175]
$a_{0.6}$	[0.03824, 0.07233]	[0.03704, 0.07139]	[0.03681, 0.07107]	[0.03823, 0.07354]	[0.03824, 0.07351]	[0.03828, 0.07180]
$a_{0.8}$	[0.02213, 0.05196]	[0.02208, 0.05133]	[0.02213, 0.05132]	[0.02352, 0.05423]	[0.02351, 0.05425]	[0.02375, 0.05188]
a_{∞}	[0.04750, 0.08045]	[0.04718, 0.08024]	[0.04771, 0.08067]	[0.04810, 0.08171]	[0.04441, 0.04778]	[0.04732, 0.07984]

Case 2 with Sample Size N=400

DCL	*True CI*	*Fieller*	*Delta*	*bootstrap Fieller*	*bootstrap Delta*	*Percentile Bootstrap*
$a_{0.2}$	[0.09631, 0.13555]	[0.09771, 0.13453]	[0.09698, 0.13372]	[0.09758, 0.13470]	[0.09763, 0.13462]	[0.09642, 0.13580]
$a_{0.4}$	[0.06540, 0.09423]	[0.06459, 0.09392]	[0.06423, 0.09351]	[0.06505, 0.09459]	[0.06507, 0.09456]	[0.06545, 0.09410]
$a_{0.6}$	[0.04273, 0.06625]	[0.04184, 0.06608]	[0.04172, 0.06592]	[0.04244, 0.06698]	[0.04245, 0.06697]	[0.04261, 0.06647]
$a_{0.8}$	[0.02732, 0.04760]	[0.02663, 0.04745]	[0.02666, 0.04744]	[0.02735, 0.04861]	[0.02735, 0.04862]	[0.02734, 0.04751]
a_{∞}	[0.05261, 0.07566]	[0.05247, 0.07584]	[0.05273, 0.07606]	[0.05286, 0.07641]	[0.05284, 0.07643]	[0.05246, 0.07557]

size is small. Table (2.11) and (2.12) show the results when the sample size is 200 and 400. In the large sample size, all the five methods provide very similar results.

2.6 Conclusion

In this chapter, the Filler's method, the delta method, bootstrapped Fieller's method, bootstrapped delta method and the bootstrap percentile method are applied to derive the CI of the *DCL*. In addition, the pairs bootstrap method is applied to derive the variances of the distribution of the duration and the *DCL*. Since the *DCL* is a new inference in the survival analysis, the variance derived by the asymptotic expansion in chapter 1 is compared with the bootstrap variance. The simulation studies show that the pairs bootstrap can provide the accurate results for the variances of the distribution of duration and the *DCL*. The empirical results show the both the bootstrap variance and the variance derived in chapter 1 are close to the benchmark value.

In terms of the confidence interval of the *DCL*, the empirical results show that the CI of *DCL* derived by Fieller's method is superior to the remaining methods when the sample size is below 50. The bootstrap Fieller's method, the bootstrap delta method and the bootstrap percentile method do not work well when the sample size is below 50. However, all the five methods provide the accurate confidence intervals of the *DCL* when the sample size is above 200. In addition, the bootstrap variance of duration and *DCL* are close to the benchmark values even when the sample size is 50.

Chapter 3

Price Change before the Financial Crisis and after the Financial Crisis: Some Evidence from the CPI Micro Data in the UK

3.1 Introduction and Literature Review

In the past decade, many economists focused on the price rigidity and stickiness which is an important transmission mechanism in the macroeconomic models. Accordingly, the different types of models have been developed to study the sticky price. There are two types of models named as the Calvo family model and the Taylor family model which can capture the effect of the price stickiness. The proportion of the price change, a parameter in the Calvo family model, is calculated by the Kaplan-Meier estimator (survival function) and the Nelson-Aalen estimator (hazard function). In other words, the survival analysis has been applied in the CPI micro data study. In this chapter, we are going to focus on three topics: firstly, we are going

to describe the UK CPI micro data. All the properties are given including the frequency of the price change, the size of the price change in each division; the data cover the period from 1999 to 2016. Secondly, the non-parametric test is introduced to investigate the two-sample problem. The CPI micro data are divided into two groups: before financial crisis (before 2008, named as the BF group) and after the financial crisis (after 2008, named as the AF group). The log-rank family tests are applied in order to compare the hazard function and survival function between the two groups. The null hypothesis is that the hazard function and the survival function of the two groups come from the same distribution; the last, the Cox model and the accelerated failure time model are applied to evaluate the heterogeneity and seasonality effect of the price changes.

3.1.1 The CPI Micro Data

A far-researching study of CPI micro-data is given by Bils and Klenow (2004). They applied the unpublished data, obtaining from US Bureau of Labor Statistics, to calculate the duration of the price changes from 1995 to 1997. The data included hundreds of the categories of the goods and the services. Depending on this study, the average duration of the price changes was around 3.3 months, and the median duration was 4.3 months including sales. Excluding sales, the median duration was about 5.5 months. After that, Nakamura and Steinsson (2008) used the US Bureau of Labor Statistics CPI micro data to evaluate the properties of the price adjustment. They found that the median frequency of the price adjustment was around 20 percent each month including sales. The median frequency of the price change without sales was only around 12 percent each month. The truncation and censored problems had been considered.

Bunn and Ellis (2009) applied the survival analysis to investigate the duration of the price adjustments in the United Kingdom. These CPI micro data were obtained from the Office for National Statistics (ONS). The data included 11 million price quotes from 1996 to 2006.The

frequency of the price changes was 19 percent each month. In other words, the mean duration of the price adjustments was around 5 months. They also found that the duration of the price adjustments varied over time.

Carroll et al. (2011) applied the instrumental variables and the Kalman filter methods to capture the degree of the price stickiness in the total consumption growth among different countries. They used an auto-regressive moving average (ARMA) model to study the aggregate consumption data. It was a state-dependent model using the quarterly data of the non-durable goods and services. Kumar and Owen (2013) used the same methods to investigate how the financial crisis affected the consumption. However, they got some different results by adding the data including the financial crisis period. Therefore, it is worth to investigate how hazard function changed before 2008 and after 2008.

Cavallo (2016) applied the scraped data to investigate the price adjustments. These scraped data were the daily data collected from the Internet. Compared with previous results, this study showed that there existed the longer duration of the price change in the scraped data. In addition, they found that the time average effect and the imputations of the missing data were the important reasons which caused the bias of the duration and the size of the price adjustments.

However, most of the scholars investigated the CPI micro data before the financial crisis. In our research, the UK CPI micro data cover the period from 1999 to 2016. In this chapter, we are going to evaluate and compare the hazard function between the AF group and the BF group. We want to find out how the difference in the duration of price adjustment between the two groups (periods). In this chapter. the basic statistics are applied to evaluate the frequency, duration and the implied duration of price adjustments. In addition, the Cox model and accelerated failure time model are applied to investigate the effects of heterogeneity and the seasonality in the data. The non-parametric methods were introduced to study the difference in the hazard function between the groups.

3.1.2 Non-Parametric Test for Two-Sample Comparison

There are many non-parametric methods, which can be applied to make the comparison between two samples or groups. Welch (1938) derived a new method and applied it to investigate the properties between two groups. This method was named as the Welch's t test. Especially, the mean value of the two groups can be compared even there existed the different variances and sample sizes. In Welch's t test, the two groups might have the different population mean and sample size. However, the two groups must follow the normal distribution. In terms of the two-sample t test, there existed the same variances in the two groups. The null hypothesis was that whether the mean μ_1 from group 1 was equal to the mean μ_2 from group 2. Since then, many papers had been published to improve those methods.

Satterthwaite (1946) derived a reasonable formula for the chi-square test and gave an approximation for the degree of freedom. This approximate can be directly applied to calculate the degree of freedom in Welch's t test. The requirement of this method was that different groups had the different variances. Otherwise, the degree of freedom should be $n_1 + n_2 - 2$. This approximation method could give a valid degree of freedom for both chi-square and Welch's t test.

Welch (1947) focused on the application of student-t test with unequal sample size and variance. He also gave the series approximation for the inverse degree of freedoms. This approximation of the degree of freedom was the same as Satterthwaite (1946)'s result. The properties of Welch's t test when the sample size was small was investigated by Welch (1951).

Wang (1971) investigated the type I error of the Welch's t test. Depending on the simulation, the empirical size of the test was close to the nominal significance level. Therefore, it can be concluded that student t table can be directly applied in the Welch's t test. To compare the Welch's t test and any other two-sample tests, see Yuen (1974) and Zimmerman and Zumbo (1993). In the Welch's t test, the assumption that the data followed the normal

distributions could be ignored when the sample size is large enough. In the small sample size situation, it may cause the bias problem.

Alekseyenko (2016) developed a multivariate Welch's t test. This multivariate Welch's t test could be applied in the two-sample problem with unbalanced sample size and heteroscedasticity effect. This new Welch's t test was derived from the pairwise distance matrix.

Wilcoxon (1945) derived a Wilcoxon sign-rank test. It was a non-parametric test. It can be applied to investigate whether the pairs were draw from the same population. In the Wilcoxon sign-rank test, the data were not required to follow the normal distribution.

Mann and Whitney (1947) derived the Mann-Whitney U test which was known as the Wilcoxon rank sum test. The Mann-Whitney U test was different from the sign-rank test. The former can be applied in the independent obervations while the later can be applied in the dependent observations. Even the Mann-Whitney U test did not require the normality assumption for the observations, it could be approximated as the asymptotic normal distribution. The Mann-Whitney U test could be applied when the data did not include the censored observations.

However, the point tests are not suitable for survival analysis. The reason is that the point tests cannot capture the effects of right censored observations. To capture the effects of right censored observations, the log-rank test should be considered. The log-rank test was derived by Mantel and Haenszel (1959). The log-rank test can be applied to evaluate whether hazard function and KM estimator followed the same distribution over the whole periods. The null hypothesis of the log-rank test is that the hazard functions are the same between among the groups. Even the log-rank test and its family tests are designed to apply in the censored data, they are still available in the uncensored data. It should be mention that the CPI micro data is interval censored. Since the prices of goods and services are collected once per month, the exact date of price changes cannot be known. For the interval censored data, the log-rank

test may provide a bias result. The weighted log-rank test is also considered since it is more robust than the log-rank test.

Mantel and Haenszel (1959) used the 2×2 contingency table to explain the log-rank test. After that, Peto and Peto (1972) derived several rank tests to study the properties of the different groups. Those tests could be applied when there existed the right censored observations in the data. They found that the log-rank test was suitable for the lifetime table. The local power of the log-rank test was greater than any other rank test when the data included the right censored observations.Peto and Peto (1972) derived a type of the weighted log-rank test named as Peto and Peto's log-rank test. It gave more weight to the early observations. Prentice (1978) extended Peto and Peto's log-rank test by modifying the weighted coefficient. Finkelstein (1986) provided a method to fit the data in the proportional hazard model.

In terms of choosing the variance for the log-rank test, see Morton (1978) and Brown (1984). In terms of the comparison of the different log-rank family tests with equal or unequal sample size, see Latta (1981), Heinze et al. (2003),

Gehan (1965) derived a generalized Wilcoxon test to study the differences between different groups when the right censored observations existed. Breslow (1970) extended Gehan's generalize Wilcoxon test to evaluate the differences in the survival functions. Martinez and Naranjo (2010) compared the survival cures by applying the generalized Wilcoxon test and the log-rank test. They found that the power of those tests were related to the proportional hazard assumptions and the lateness of the separation of the survival curves between the two groups.

Tarone and Ware (1977) indicated that the only difference between the log-rank test and the modified Wilcoxon test was the weighted coefficient. The log-rank test can be thought as a special case of the weighted log-rank test with the weighted coefficient equal to 1. Fleming et al. (1980) derived a modified Kolmogorov-Smirnov (KS) test which can be applied in the

uncensored data. Fleming et al. (1987) derived the supremum version of the log-rank test which was related with the log-rank test and Wilcoxon test.

3.2 Properties of the CPI Micro Data

In this chapter, we use the CPI micro data obtained from the UK Office for National Statistics (ONS) Virtual Micro-data Laboratory (VML). The ONS is the UK government's statistics department. The VML is a facility of the ONS. ONS provides different economic data for the researchers. The data includes over 30 million observations from January 1999 to December 2016. These monthly price quotes are collected "locally" by the ONS officers. The data can be known as the interval censored data. More generally, the price quotes can be changed each day or each week while they are collected once per month by the ONS.

3.2.1 Introduction for the UK CPI Micro Data

In terms of the UK CPI micro data, the observations are divided into 13 sections which are associated with the COICOP4 and the COICOP5 codes. The COICOP4 codes including 5-6 digit numbers are applied to classify the goods and the services. After 2015, ONS also reports the COICOP5 codes which give more details for observations. There exist 13 sections in the UK CPI data corresponding to the COICOP4 code. There exist multi-observations for the same items since the prices are collected from different shops and regions. There exist 12 regions to collect the price including London, Scotland, Wales and Northern Ireland etc. There also exist the shop codes for the different brands of shops such as Tesco and Sainsbury.

There exist the weighted coefficients for the shop, the region, the item and the COICOP4 division (COICOP weight). The weighted coefficients can be applied to calculate the weighted price (aggregate price). Therefore, the different levels of the indices can be

generated. The weighted coefficients of different regions are named as stratum weight in the data.

There exist some indicators for the price quotes. An indicator "S" gives the information that the observation is on "sale". In addition, if the price quote of observation is missing, an indicator factor "M" is assigned to the observation. An indicator "T" is assigned to the observation when the observation is temporary out of the stock.

There exist the records of the start date and end date for each observation. This is a very useful information when the censored and truncation problems are considered. Currently, some observations are still tracked and included in the CPI framework so that the end date is 999999. Since ONS may change the CPI framework each year, some goods and services may not be included in the CPI framework. In such cases, the old observations are replaced by the new observations.

For each observation, both the price and log price are given in the data. Those values can be applied to calculate the frequency and the size of the price changes.

3.2.2 Censored and Truncated Data

In this section, we use the whole period of the data as an example to explain the censored and truncated problems. In terms of the CPI micro data, the price quotes start from January 1999 and end in December 2016. This period can be defined as the observed window. In terms of each observation, it has a quote date which is the recorded date of the price quote. However, each observation also has the "start date" and the "end date". The "start date" of the price quotes means the earliest recorded date of the price quotes. The "start date" can be earlier than the observed window. With respect to the "end date" of each price quote, it can be excluded or included in the observed window. If the "end date" exceeds the December 2016, the final state whether the price adjustments happen cannot be observed. In the survival analysis, there exist the truncation and censored observations.

The left truncation means that the start date of the observation lies within the observed window. However, its state is not recorded since it does not satisfy the requirements. In the data, the price quote must satisfy three requirements then it can be defined as the left truncation. The first, the beginning date of the last state of the price quote is recorded in the observed window; the second, the last record of the price quote is earlier than its end date and included in the observed window; finally, this last record of the price quote is earlier than its end date. Under these conditions, it can be defined as the left truncation price spell.

The right truncation means that the observation must be observed in the observed window. If the end date exceed the observed window, those observations are excluded. More specially, the last record of the price quote is the same as its end date. According to the CPI micro data, if all the price changes are assumed to be observed in the observed window, it can be known as the right-truncation data.

The left-censored observation means that the end date of the observations is included in the observed window, while the start date is outside the observed window. In other words, the start date is unknown while the failure date is known. In the CPI micro data, some start dates of the price quotes are earlier than the January 1999 while the end dates are included in the observed window. Under these conditions, those price quotes can be defined as the left-censored observations since the exact length of the price spell is unknown.

The right-censored observation means that the start date of the price quote can be observed in the observed window while the end date exceeds the observed window. In other words, the price adjustment does not happen in the observed window, and it may happen outside the observed window. This can be known as the right-censored observation.

3.2.3 Missing Data and the Imputation of the Substitution Effect

In terms of the missing data, it means that the price quotes of the observation are missing in some period and exist later. The missing indicator is "M" in the CPI micro data. Some

goods are temporal missing in the CPI micro data which are assigned by the indicator "T". For instance, the cheery are out of the market in the winter. In addition, the color of goods may be changed such as the clothes. Depending on this, the price may be missing for several months. There are many methods to deal with this problem. The missing price can be solved by the imputed price. In the US, the of Labor Statistics (2015) used the mean value of the price changes obtained from the related categories as the missing value of the price quote. This can be known as the cell-related imputation method. There exists another method to deal with the missing price. If the price quote is missing in a special period, the price can be assumed as the same as the its last record until the price quote exists again. As Cavallo (2016) showed that this method replacing its missing price by its last price gives a quite similar result as Nakamura and Steinsson (2008)'s method. Therefore, we directly use its last record to be the missing price quote.

In addition, there also exists the indicator "S" in the data. The indicator "S" means the observations are on sales. Many works show that the data including sales may affect the accuracy of the duration and frequency, so we replace the sale price by its previous record (the same method to deal with missing price). Another point is that some price quotes may be increased by 200 percent or 300 percent in one month. In reality, it is unreasonable to see a huge increase in the price of goods and services. If the prices are increased more than 130 percent, They are deleted from the data. Furthermore, If the prices are decreased more than 75 percent, they are also deleted from the data.

3.2.4 The Duration, the Size and the Frequency of the Price Change

The duration of the price spell means that how long the price spell lasts. More specially, it means how long the price does not change. There exist the 13 divisions in UK CPI framework. In addition, the CPI micro data include the financial crisis period. Therefore, the duration for each division before the financial crisis (before January 2008) and after the financial crisis

(after January 2008) are reported, separately. In other words, the original data is divided into AF group (from January 2008 to December 2016) and BF group (from January 1999 to December 2007).

To calculate the frequency of the price changes, we need to define some variables as Veronese et al. (2005) did. The frequency for item j in division m at time t can be defined as $f_{m,j,t}$ where $m = 1,2,.....,M$, $j = 1,2,.....,J_m$ and $t = 1,2,.....,T$; J_m is the number of items in m-th division. Define the price for item j in division m at time t as $p_{m,j,t}$. We also define some indicators as below:

$$K_{m,j,t}= \begin{cases} 1 & If\ both\ price_{m,j,t}\ and\ price_{m,j,t-1}\ exist \\ 0 & If\ price_{m,j,t}\ exists\ but\ price_{m,j,t-1}\ does\ not\ exist \end{cases}$$

$$I_{m,j,t}= \begin{cases} 1 & p_{m,j,t} \neq p_{m,j,t-1} \\ 0 & p_{m,j,t} = p_{m,j,t-1} \end{cases}$$

Therefore, the frequency for item j at division m can be written as:

$$f_{m,j} = \frac{\sum_{t=2}^{T} I_{m,j,t}}{\sum_{t=2}^{T} K_{m,j,t}} 1 \tag{3.1}$$

The frequency for division m can be written as:

$$f_m = \frac{\sum_{j=1}^{J_m}\sum_{t=2}^{T} I_{m,j,t}}{\sum_{j=1}^{J_m}\sum_{t=2}^{T} K_{m,j,t}} \tag{3.2}$$

[1] It should be noted that the first record of each product is ignored since it is just an indicator to calculate the duration. Therefore, t starts from 2 rather than 1

the overall frequency can be written as:

$$f = \frac{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} I_{m,j,t}}{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} K_{m,j,t}} \tag{3.3}$$

The frequency of the price increases can be calculated as:

$$f^{increase} = \frac{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} I^{ins}_{m,j,t}}{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} K_{m,j,t}} \tag{3.4}$$

where:

$$I^{ins}_{m,j,t} = \begin{cases} 1 & price_{m,j,t} > price_{m,j,t-1} \\ 0 & price_{m,j,t} \leq price_{m,j,t-1} \end{cases}$$

The frequency of the price decreases can be calculated as:

$$f^{decrease} = \frac{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} I^{des}_{m,j,t}}{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} K_{m,j,t}} \tag{3.5}$$

where:

$$I^{des}_{m,j,t} = \begin{cases} 1 & price_{m,j,t} < price_{m,j,t-1} \\ 0 & price_{m,j,t} \geq price_{m,j,t-1} \end{cases}$$

The size of price changes is another important theorem. The size of price increases can be known as the average price increases. The size of price increases can be calculated as:

$$size^{increase} = \frac{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} I^{ins}_{m,j,t}(lnp_{m,j,t} - lnp_{m,j,t-1})}{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} K_{m,j,t}} \tag{3.6}$$

The size of price decreases can be calculated as:

$$size^{decrease} = \frac{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} I^{des}_{m,j,t}(lnp_{m,j,t} - lnp_{m,j,t-1})}{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} K_{m,j,t}} \tag{3.7}$$

We also calculate the absolute size of price changes:

$$size^{abs.} = \frac{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} I_{m,j,t}(|lnp_{m,j,t} - lnp_{m,j,t-1}|)}{\sum_{m=1}^{M}\sum_{j=1}^{J_m}\sum_{t=2}^{T} K_{m,j,t}} \tag{3.8}$$

The duration of item j belonging to the division m can be defined as:

$$d_{m,j} = \frac{1}{f_{m,j}} \tag{3.9}$$

Since the scraped data has been widely applied in CPI micro data study, the duration can be calculated in the continuous time formula. The relationship between the duration and the frequency of price changes can be defined as:

$$d'_{m,j} = -\frac{1}{ln(1-f_{m,j})} \tag{3.10}$$

Equation (3.10) can be known as the implied duration. The median implied duration can be calculated as:

$$d'^{median}_{m,j} = \frac{ln(0.5)}{ln(1-f_{m,j})} \tag{3.11}$$

In table (3.1)[2], the durations of price changes are reported. Before 2008, the overall duration is 5.75 month. The shortest duration is 4.54 in division 1 (food and soft drink). The longest duration is 9.05 in division 6 (health). After 2008, the overall duration is 5.47. Compared with the BF group (before 2008), the durations of divisions 1, 7 , 8 and 99 are increased while the remaining durations of the divisions are decreased in AF group (after 2008).

In table (3.2), the mean and the median of the implied duration are reported. The mean of implied duration are lower than the duration in discrete time. The overall median of the

[2]In this research, we do not include division 10 since we only have the records of division 10 before 2008.

Table 3.1 The Mean Duration for each Division in Discrete Time Version

Before the Financial Crisis(2008)		
Division	*mean duration*	*COICOP weight*
01 *Food and soft drink*	4.54	0.103
02 *Alcohol and tobacco*	4.62	0.042
03 *Clothing and footwear*	6.09	0.071
04 *Energy*	7.37	0.120
05 *Householdequipment*	6.78	0.09
06 *Health*	9.05	0.028
07 *Transport*	7.1	0.153
08 *Communication*	5.05	0.032
09 *Recreation and culture*	5.4	0.148
11 *Restaurants and hotels*	7.07	0.123
12 *Miscellaneous goods*	7.26	0.096
99 *Unallocated goods*	7.51	0.05
Overall	5.75	0.975
After the Financial Crisis(2008)		
Division	*mean duration*	*COICOP weight*
01 *Food and soft drink*	4.8	0.103
02 *Alcohol and tobacco*	4.19	0.042
03 *Clothing and footwear*	5.41	0.071
04 *Energy*	6.71	0.120
05 *Household equipment*	6.04	0.09
06 *Health*	8.2	0.028
07 *Transport*	7.38	0.153
08 *Communication*	5.26	0.032
09 *Recreation and culture*	4.82	0.148
11 *Restaurants and hotels*	6.49	0.123
12 *Miscellaneous goods*	6.83	0.096
99 *Unallocated goods*	8.76	0.05
Overall	5.47	0.975

Note: There exist 13 divisions in the UK CPI micro data. The results of division 10 are excluded since there are no records for division 10 after 2008. The first column shows the divisions. The second column reports the mean duration for each division. The third column reports the COICOP weight for each division.

implied duration is 5.23 before 2008 while the overall median duration decreases to 4.95 after 2008.

In table (3.3), the size of the price changes is reported. The 4-th column reports the absolute size of price changes. Before 2008, the overall price increase is 1.87%. It can be

Table 3.2 Implied Duration: The Mean and Median Duration for each Division in Continuous Time Version.

Before the Financial Crisis(2008)		
Division	*mean duration*(*continuous time*)	*median duration*(*continuous time*)
01 *Food and soft drink*	4.01	2.79
02 *Alcohol and tobacco*	4.1	2.84
03 *Clothing and footwear*	5.58	3.86
04 *Energy*	6.86	4.75
05 *Household equipment*	6.27	4.34
06 *Health*	8.54	5.92
07 *Transport*	6.59	4.57
08 *Communication*	4.53	3.14
09 *Recreation and culture*	4.88	3.38
11 *Restaurants and hotels*	6.56	4.55
12 *Miscellaneous goods*	6.75	4.68
99 *Unallocated goods*	7	4.85
Overall	5.23	3.63
After the Financial Crisis(2008)		
Division	*mean duration*	*median duration*
01 *Food and soft drink*	4.28	2.97
02 *Alcohol and tobacco*	3.67	2.54
03 *Clothing and footwear*	4.89	3.39
04 *Energy*	6.2	4.3
05 *Household equipment*	5.52	3.82
06 *Health*	7.69	5.33
07 *Transport*	6.87	4.76
08 *Communication*	4.74	3.29
09 *Recreation and culture*	4.3	2.98
11 *Restaurants and hotels*	5.98	4.14
12 *Miscellaneous goods*	6.32	4.38
99 *Unallocated goods*	8.25	5.72
Overall	4.95	3.43

Note: The first column shows the divisions. The second column reports the mean duration (continuous time) for each division. The mean duration (continuous time) can be known as the implied duration. The third column reports the median duration (continuous time) for each division.

seen that the highest price increase is 2.29% in division 1 (food and soft drink) and 3(clothing and footwear). The lowest price increase is 0.86% in division 2 (alcohol and tobacco). The overall price decrease is 1.35%. In terms of the empirical results, The overall size of the price increases is higher than the overall size of the price decreases in the UK CPI data before 2008.

After 2008, the overall size of price increases is 2.03% while the overall size of price decreases is 1.42%. The highest price increase is 2.82% in division 3 (clothing and footwear). The overall size of the price increases is higher than the overall size of the price decreases after 2008. There exist the highest absolute size of price changes in division 3 in both AF and BF groups.

In table (3.4), the frequency of the price changes, the frequency of price increases and the frequency of price decreases are reported. Before 2008, the overall frequency of price changes is 17.39%. The overall frequency of price increases is 8.68% while the overall frequency of price decreases is 7.16%. The highest frequency of price changes is 22.03% in division 1 while the lowest frequency of price changes is 11.05% in division 6. In addition, the frequency of price increases is higher than the price decreases in most divisions except the division 3, 8 and 9. Depending on the result, the frequency of price changes is dominated by the price increases.

After 2008, the overall frequency of the price change is 18.28%. Compared with the BF group, both the frequencies of the price increases and decreases are increased. The frequency of price changes is still dominated by the price increases. Another point is that the price changed more frequently after 2008.

Table 3.3 The Size of the Price Change

Before the Financial Crisis(2008)			
Division	*size of increase*	*size of decrease*	*size of abs. change*
01 *Food and soft drink*	0.0229	0.0155	0.0384
02 *Alcohol and tobacco*	0.0086	0.0053	0.0139
03 *Clothing and footwear*	0.0229	0.0189	0.0418
04 *Energy*	0.017	0.0109	0.0279
05 *Household equipment*	0.0168	0.0127	0.0295
06 *Health*	0.0106	0.0072	0.0178
07 *Transport*	0.0138	0.008	0.0218
08 *Communication*	0.0173	0.0197	0.037
09 *Recreation and culture*	0.0204	0.0172	0.0376
11 *Restaurants and hotels*	0.0115	0.0065	0.018
12 *Miscellaneous goods*	0.015	0.0099	0.0249
99 *Unallocated goods*	0.0124	0.0051	0.0175
Overall	0.0187	0.0135	0.0322
After the Financial Crisis(2008)			
Division	*size(increase)*	*size(decrease)*	*size of abs. change*
01 *Food and soft drink*	0.0212	0.0147	0.0359
02 *Alcohol and tobacco*	0.0141	0.0084	0.0225
03 *Clothing and footwear*	0.0282	0.0213	0.0495
04 *Energy*	0.0156	0.0102	0.0258
05 *Household equipment*	0.0191	0.013	0.0321
06 *Health*	0.0119	0.0075	0.0194
07 *Transport*	0.0125	0.0081	0.0206
08 *Communication*	0.0264	0.0185	0.0449
09 *Recreation and culture*	0.0236	0.018	0.0416
11 *Restaurants and hotels*	0.0122	0.0068	0.019
12 *Miscellaneous goods*	0.018	0.0116	0.0296
99 *Unallocated goods*	0.008	0.0048	0.0128
Overall	0.0203	0.0142	0.0345

Note: This table reports the size of the price adjustments in different divisions. The first column shows the divisions. The second column reports the size of price increases. The third column reports the size of price decreases. The last column reports the absolute size of price changes.

Table 3.4 The Frequency of the Price Change

Before the Financial Crisis(2008)			
Division	$frequency$	$Frequency(increase)$	$Frequency(decrease)$
01 *Food and soft drink*	0.2203	0.1121	0.0953
02 *Alcohol and tobacco*	0.2165	0.1397	0.06
03 *Clothing and footwear*	0.1642	0.0728	0.0757
04 *Energy*	0.1357	0.0712	0.0493
05 *household equipment*	0.1475	0.0675	0.0647
06 *Health*	0.1105	0.0576	0.0373
07 *Transport*	0.1408	0.0758	0.0464
08 *Communication*	0.1980	0.0675	0.1084
09 *Recreation and culture*	0.1851	0.0802	0.0872
11 *Restaurants and hotels*	0.1414	0.0827	0.0408
12 *Miscellaneous goods*	0.1377	0.0687	0.0519
99 *Unallocated goods*	0.1332	0.0803	0.0364
Overall	0.1739	0.0868	0.0716
After the Financial Crisis(2008)			
Division	$frequency$	$Frequency(increase)$	$Frequency(decrease)$
01 *Food and soft drink*	0.2083	0.1089	0.0921
02 *Alcohol and tobacco*	0.2387	0.154	0.0791
03 *Clothing and footwear*	0.1848	0.0919	0.0862
04 *Energy*	0.1490	0.0872	0.0558
05 *Household equipment*	0.1656	0.0.0899	0.0696
06 *Health*	0.122	0.0641	0.0484
07 *Transport*	0.1355	0.0766	0.0507
08 *Communication*	0.1901	0.0879	0.0886
09 *Recreation and culture*	0.2075	0.1021	0.0987
11 *Restaurants and hotels*	0.1541	0.1027	0.0444
12 *Miscellaneous goods*	0.1464	0.0801	0.0585
99 *Unallocated goods*	0.1142	0.0643	0.0414
Overall	0.1828	0.099	0.077

Note: This table reports the frequency of price changes in different divisions between the BF and the AF groups. The first column shows the divisions. The second column reports the frequency of the price changes. The third column reports the frequency of the price increases. The last column reports the frequency of the price decreases.

3.2.5 Heterogeneity and Seasonality effect for the CPI Micro Data

In this part, the Cox model is applied to regress the CPI micro data in the UK. The data has been divided into two groups: before 2008 (1999-2007, BF group) and after 2008 (2008-2016, AF group). We use the dummy variable dmf_n to distinguish the two groups: $dmf_n = 0$ for BF group and $dmf_n = 1$ for AF group; Where the subscript n is the n-th observations with $n = 1, 2, ..., N$; N is the sample size. The Cox model can be defined as:

$$h(t) = h_0(t)exp(\sum_{i=1}^{12} \alpha_i dms_{i,n} + \sum_{j=1}^{11} \beta_j dmm_{j,n} + \delta dmf_n) \tag{3.12}$$

Where $h_0(t)$ is the baseline hazard function; the $dms_{i,n}$ is the division dummy with $i = 1, 2, ...13$ and $n = 1, 2, ..., N$. There exist 12 divisions[3] in the UK CPI framework. Define $dms_{1,n}$ is division 1 (food and non-alcoholic beverages); $dms_{2,n}$ is division 2 (alcohol and tobacco); $dms_{3,n}$ is division 3 (clothing and footwear); $dms_{4,n}$ is division 4 (energy); $dms_{5,n}$ is division 5 (household equipment); $dms_{6,n}$ is division 6 (health); $dms_{7,n}$ is division 7 (transport); $dms_{8,n}$ is division 8 (communication); $dms_{9,n}$ is division 9 (recreation and culture); $dms_{11,n}$ is division 11 (restaurants and hotels); $dms_{12,n}$ is division 12 (miscellaneous goods). $dms_{13,n}$ is division 99 (Unallocated goods). In UK CPI micro data, the division 10 is excluded since there are no records of division 10 in BF group. In addition, the division 99 $dms_{13,n}$ is chosen as the reference group.

The $dmm_{j,n}$ is the monthly dummy variable of price changes with $j = 1, 2, ...12$ and $n = 1, 2, ..., N$. In particular, the monthly dummy variables can be defined as:$dmm_{1,n}$ for January; $dmm_{2,n}$ for February;$dmm_{3,n}$ for March; $dmm_{4,n}$ for April; $dmm_{5,n}$ for May; $dmm_{6,n}$ for June; $dmm_{7,n}$ for July; $dmm_{8,n}$ for August; $dmm_{9,n}$ for September; $dmm_{10,n}$ for October; $dmm_{11,n}$ for November; $dmm_{12,n}$ for December. December $dmm_{12,n}$ is chosen as the reference group.

[3] Actually, there exist 13 divisions including the division 99. But we do not use the division 10 since there are no record for division 10 after 2008

The results of Cox model are shown in table (3.5). The significant coefficient of the dmf_n means that there exist the different hazard ratios between the BF and AF group. The coefficient of dmf_n=0.0195 means that the hazard ratio of AF group is higher than the BF group. In other words, the duration of price spells is shorter and the frequency of price changes is higher in AF group. The coefficients of divisions are significant expect the coefficient of division 4 $dms_{4,n}$. Therefore, there exist the heterogeneity effects of price adjustments across different divisions. In terms of the seasonal dummy variables, the coefficient of $dmm_{2,n}$ is insignificant while the remaining seasonal dummy variables are all significant. Therefore, there also exist the seasonality effects of price changes across different months.

However, the proportional hazard test shows that the proportional hazard assumption is violated by the Cox regression model. The proportional hazard test is introduced by Grambsch and Therneau (1994). The null hypothesis is that the proportional hazard assumption holds. In table (3.6), the results of proportional hazard tests are reported. It can be seen that the null hypothesis is rejected depending on the p-value. In other words, the proportional hazard assumption is violated. Therefore, we also apply the accelerated failure time model (AFT model) to regress the data. The AFT model does not require the proportional hazard assumption. The AFT model can be written as:

$$S(t) = S_0(exp(\sum_{i=1}^{12} \theta_i dms_{i,n} + \sum_{j=1}^{11} \eta_j dmm_{j,n} + \gamma dmf_n)t) \tag{3.13}$$

Where the $dms_{i,n}$ is the division dummy variable with $i = 1,2,.....,12$; $dmm_{j,n}$ is the seasonal dummy variable with $j = 1,2,.....,11$; the subscript n means the n-th observations with $n = 1,2,.....,N$. $S_0(.)$ can be defined as the baseline survival function. Alternatively, log-linear AFT model can be expressed as:

$$log(T_n) = \omega + \sum_{i=1}^{12} \theta_i dms_{i,n} + \sum_{j=1}^{11} \eta_j dmm_{j,n} + \gamma dmf_n + \sigma e_n \tag{3.14}$$

Where $log(T_n)$ is the log transformed lifetime; ω is the intercept; σ is the scale parameter; e_n is the random error term. In the AFT model, the T_n must be assumed to follow some distribution such as log-logistic, Weibull and log normal distribution. To selecting the most suitable AFT model, the log likelihood value can be applied to evaluate the performances of different distributions. At this point, T_n is assumed to follow either the log-logistic or Weibull distributions.

The empirical result of the AFT model with log-logistic distribution is shown in table (3.7). The empirical result of the AFT model with Weibull distribution is shown in table (3.8). It can be seen that all the coefficients of the divisions and the monthly dummy variables are significant. It means that there exist the cross-sectional effects across the different divisions and the months. In AFT model, the negative coefficient of dmf_n indicates that the frequency of the price adjustments in AF group is higher than the price adjustments in BF group. The log likelihood value of the log-logistic distribution is higher than the log likelihood value of the Weibull distribution. Therefore, the AFT model with log-logistic distribution is more suitable for the data.

Table 3.5 Cox Regression for the CPI Micro Data in the UK

$dummies$	$coefficients$	$exp(coefficients)$	$std.$	$p-value$
dmf_n	0.0195	1.0197	0.00115	0
$dms_{1,n}$	0.4970	1.6437	0.0105	0
$dms_{2,n}$	0.4927	1.6368	0.0107	0
$dms_{3,n}$	0.2899	1.3363	0.0105	0
$dms_{4,n}$	0.1786	1.1956	0.0111	0
$dms_{5,n}$	0.2249	1.2522	0.0106	0
$dms_{6,n}$	−0.0002	0.9998	0.0118	0.986
$dms_{7,n}$	0.1157	1.1227	0.0111	0
$dms_{8,n}$	0.3620	1.4362	0.0148	0
$dms_{9,n}$	0.4188	1.5202	0.0106	0
$dms_{11,n}$	0.2031	1.2251	0.0106	0
$dms_{12,n}$	0.1480	1.1595	0.0107	0
$dmm_{1,n}$	0.2002	1.2216	0.0024	0
$dmm_{2,n}$	−0.0019	0.9981	0.0024	0.436
$dmm_{3,n}$	0.3576	1.4299	0.0026	0
$dmm_{4,n}$	0.3097	1.3631	0.0025	0
$dmm_{5,n}$	0.4045	1.4986	0.0026	0
$dmm_{6,n}$	0.4068	1.5020	0.0026	0
$dmm_{7,n}$	0.3570	1.4291	0.0026	0
$dmm_{8,n}$	0.2994	1.3490	0.0026	0
$dmm_{9,n}$	0.2560	1.2917	0.0025	0
$dmm_{10,n}$	0.2669	1.3059	0.0026	0
$dmm_{11,n}$	0.2550	1.2905	0.0026	0

Note: This is the empirical result from the Cox model. dmf_n is the dummy variable which is applied to distinguish the BF and the AF groups. We define $dmf_n = 1$ for the AF group and $dmf_n = 0$ for the BF group. $dms_{i,n}$ is the division dummy variable for division i; $dmm_{j,n}$ is the monthly dummy for j-th month. We choose division 99 as reference division dummy and December as reference monthly dummy. dms_{10} is not included since there are no records of division 10 in the CPI micro data after 2008.

Table 3.6 Proportional Hazard Assumption

dummies	*correlation coefficient*	$p-value$
dmf_n	0.0193	0
$dms_{1,n}$	−0.0195	0
$dms_{2,n}$	−0.0074	0
$dms_{3,n}$	−0.0054	0
$dms_{4,n}$	−0.0111	0
$dms_{5,n}$	−0.0099	0
$dms_{6,n}$	−0.0082	0
$dms_{7,n}$	−0.0039	0
$dms_{8,n}$	−0.0041	0
$dms_{9,n}$	−0.0150	0
$dms_{11,n}$	−0.0139	0
$dms_{12,n}$	−0.0087	0
$dmm_{1,n}$	0.0263	0
$dmm_{2,n}$	−0.0282	0
$dmm_{3,n}$	−0.0113	0
$dmm_{4,n}$	0.0118	0
$dmm_{5,n}$	0.0012	0
$dmm_{6,n}$	−0.0075	0
$dmm_{7,n}$	0.0096	0
$dmm_{8,n}$	0.0209	0
$dmm_{9,n}$	0.0281	0
$dmm_{10,n}$	0.0209	0
$dmm_{11,n}$	0.0192	0
Global	*NA*	0

Note: This is the proportional hazard test introduced by Grambsch and Therneau (1994). The last row "Global" means whether there exists the proportional hazard assumption for the whole model including all the dummy variables. The last column reports the p value of the Chi-square test.

Table 3.7 AFT Regression for the CPI Micro Data in the UK with the Log-logistic Distribution

$dummies$	$coefficients$	$std.$	$p-value$
$intercept$	1.8252	0.0107	0
dmf_n	−0.0096	0.0011	0
$dms_{1,n}$	−0.7122	0.0106	0
$dms_{2,n}$	−0.5778	0.0108	0
$dms_{3,n}$	−0.3768	0.0107	0
$dms_{4,n}$	−0.3189	0.0112	0
$dms_{5,n}$	−0.3523	0.0107	0
$dms_{6,n}$	−0.0895	0.0121	0
$dms_{7,n}$	−0.1743	0.0112	0
$dms_{8,n}$	−0.4431	0.0148	0
$dms_{9,n}$	−0.5939	0.0107	0
$dms_{11,n}$	−0.3640	0.0107	0
$dms_{12,n}$	−0.2558	0.0108	0
$dmm_{1,n}$	−0.1246	0.0024	0
$dmm_{2,n}$	−0.0773	0.0026	0
$dmm_{3,n}$	−0.3721	0.0027	0
$dmm_{4,n}$	−0.2752	0.0026	0
$dmm_{5,n}$	−0.3844	0.0026	0
$dmm_{6,n}$	−0.4057	0.0026	0
$dmm_{7,n}$	−0.3155	0.0026	0
$dmm_{8,n}$	−0.2309	0.0026	0
$dmm_{9,n}$	−0.1678	0.0026	0
$dmm_{10,n}$	−0.1983	0.0026	0
$dmm_{11,n}$	−0.1980	0.0026	0
$log(scale)$	−0.5541	0.0004	0
log-likelihood=-8945459			

Note: This is regression result of the AFT model with log-logistic distribution. ω is the intercept in the linear AFT model.

Table 3.8 AFT Regression for the CPI Micro Data in the UK with the Weibull Distribution

$dummies$	$coefficients$	$std.$	$p-value$
$intercept$	2.4157	0.0109	0
dmf_n	−0.0320	0.0011	0
$dms_{1,n}$	−0.5366	0.0108	0
$dms_{2,n}$	−0.5845	0.0111	0
$dms_{3,n}$	−0.3390	0.0108	0
$dms_{4,n}$	−0.1701	0.0114	0
$dms_{5,n}$	−0.2431	0.0109	0
$dms_{6,n}$	−0.0417	0.0122	0
$dms_{7,n}$	−0.1244	0.0114	0
$dms_{8,n}$	−0.4412	0.0152	0
$dms_{9,n}$	−0.4619	0.0109	0
$dms_{11,n}$	−0.2024	0.0109	0
$dms_{12,n}$	−0.1505	0.0110	0
$dmm_{1,n}$	−0.3242	0.0025	0
$dmm_{2,n}$	−0.0130	0.0025	0
$dmm_{3,n}$	−0.4787	0.0027	0
$dmm_{4,n}$	−0.4423	0.0026	0
$dmm_{5,n}$	−0.5418	0.0026	0
$dmm_{6,n}$	−0.5361	0.0026	0
$dmm_{7,n}$	−0.4929	0.0027	0
$dmm_{8,n}$	−0.4347	0.0026	0
$dmm_{9,n}$	−0.3897	0.0026	0
$dmm_{10,n}$	−0.3979	0.0027	0
$dmm_{11,n}$	−0.3812	0.0027	0
$log(scale)$	0.0328	0.0004	0
log-likelihood=-9304401			

Note: This is regression result of the AFT model with Weibull distribution. The intercept is ω in the linear AFT model. The log(scale) value is the log value of the scale variable σ.

3.3 Survival Analysis

In this section, the statistical techniques of survival analysis are reviewed. In addition, the log-rank family tests are introduced. The estimator of the survival function can be known as the Kaplan-Meier (KM) estimator. It can be assumed that there exist N_i price spells at the i-th period with $i = 1, 2, ...F$, and the maximum length of the price spell lasts F period. The number of price spells disappearing at the i-th period can be defined as D_i. Therefore, the KM estimator for i-th period can be defined as:

$$\hat{S}_i = \prod_{k=1}^{i} \frac{N_k - D_k}{N_k} \tag{3.15}$$

The hazard function can be estimated by the Nelson-Aalen (NA) estimator. It can be defined as:

$$\hat{H}_i = \sum_{k=1}^{i} \frac{D_k}{N_k} \tag{3.16}$$

For the marginal hazard function, it can be written as:

$$\hat{h}_i = \frac{D_i}{N_i} \tag{3.17}$$

The KM estimator and hazard function are the coefficients of sectors existing in the Calvo model.

3.3.1 Log-Rank Test

The log-rank family test can be applied to investigate whether the marginal hazard functions follow the same distribution between different groups. The log-rank test is the joint test which is designed for the survival analysis. The log-rank test can be derived from the 2×2 contingency table. However, the log-rank test can be derived related to Fieller's method. In this section, the Fieller's method is applied to derive the log-rank test. In addition, the

weighted log-rank tests are introduced for the two-sample comparison. One advantage of the log-rank family test is that they do not require the proportional hazard assumption. Another advantage of the log-rank test is that it is quite robust. Moreover, the weighted log-rank test is more robust than the log-rank test.

Assume that there exist two groups: group 1 and 2. Group 1 includes the price spells before the financial crisis. Group 2 includes the price spells after the financial crisis. The maximum period is F months in each group. In the i-th month, the number of price spells can be defined as N_i, g with $g = 1,2$ and $i = 1,2,.....,F$. In other words, N_i is the number of the price spells surviving at i-th month. The $D_{i,g}$ is the number of the price spells disappear in i-th month for group g. The hazard function of group 1 and 2 can be defined as $\hat{h}_{i,1} = D_{i,1}/N_{i,1}$ and $\hat{h}_{i,2} = D_{i,2}/N_{i,2}$, respectively. The group 1 and group 2 can be combined and create an overall sample. In this overall sample, the price spells survive at i-th period is $N_i = N_{i,1} + N_{i,2}$ and the number of the price spells disappearing at the i-th period is $D_i = D_{i,1} + D_{i,2}$ with $i = 1,2,.....,F$. The hazard function of the overall sample can be defined as $\hat{h}_i = D_i/N_i$. If the observations of the two groups are drawn from the same distribution, they should have the same distribution of the observations in the overall sample. To compare $\hat{h}_{i,1}$ with $\hat{h}_{i,2}$, it is equivalent to test whether the $\hat{h}_{i,g}$ and $\hat{h}_i$ follow the same distribution.

The null hypothesis is $H_0 : \hat{h}_{i,1} = \hat{h}_{i,2}$. However, the null hypothesis can be also written as $H_0 : \hat{h}_{i,1} = \hat{h}_i$ or $H_0 : \hat{h}_{i,2} = \hat{h}_i$ with $i = 1,2,.....,F$. These null hypotheses are the same. To see this, the null hypothesis $H_0 : \hat{h}_{i,1} = \hat{h}_i$ can be used as an example. It can be rewritten as:

$$D_{i,1}/N_{i,1} = D_i/N_i \tag{3.18}$$

Substituting $D_i = D_{i,1} + D_{i,2}$ and $N_i = N_{i,1} + N_{i,2}$ into the equation (3.18):

$$\frac{D_{i,1}}{N_{i,1}} = \frac{D_{i,1} + D_{i,2}}{N_{i,1} + N_{i,2}} \tag{3.19}$$

To simplify this function by deleting the denominator:

$$D_{i,1}(N_{i,1}+N_{i,2}) = N_{i,1}(D_{i,1}+D_{i,2}) \tag{3.20}$$

Delete the same variable:

$$D_{i,1}N_{i,2} = D_{i,2}N_{i,1} \tag{3.21}$$

When $N_{i,1}$ and $N_{i,2}$ are not equal to zero:

$$D_{i,1}/N_{i,1} = D_{i,2}/N_{i,2} \tag{3.22}$$

This is the null hypothesis of the log-rank test. Therefore, the modified null hypothesis is the same as the original null hypothesis. The equation (3.18) can be rewritten as $H_0 : D_{i,1} = \frac{D_i}{N_i}N_{i,1}$. On the right-hand side, $\frac{D_i}{N_i}N_{i,1}$ can be known as the mean value of the hypergeometric distribution. Therefore, the $D_{i,1}$ can be assumed to follow the hypergeometric distribution with mean $E_{i,1} = \frac{D_i}{N_i}N_{i,1}$ and variance $Var(D_{i,1}) = \frac{D_i(N_{i,1}/N_i)(1-N_{i,1}/N_i)(N_i-D_i)}{N_i-1}$ with $i = 1,2,3,...F$. To simplify the formula, the variance $V_{i,1} = Var(D_{i,1})$ can be defined with $i = 1,2,.....,F$. The log-rank test for the *i*-th period can be written as:

$$log-rank\ statistic = \frac{(D_{i,1} - \frac{D_i}{N_i}N_{i,1})^2}{Var(D_{i,1})} \tag{3.23}$$

Mantel and Haenszel (1959) showed that the log-rank test can be compared with the chi-square critical value with 1 degree of freedom[4]. In addition, the log-rank test can be written as:

$$log-rank\ statistic = \frac{(D_{i,1} - \frac{D_i}{N_i}N_{i,1})}{\sqrt{Var(D_{i,1})}} \tag{3.24}$$

[4]The degree of freedom is the number of groups minus one in log-rank test

Now the log-rank statistic follows the standard normal distribution asymptotically. As Mantel and Haenszel (1959) argued that the log-rank test should include all the periods rather than the single period since the some period may be rejected. The overall test is to test whether $\sum_{i=1}^{F}(D_{i,1} - \frac{D_i}{N_i}N_{i,1})$ is less than the critical value or not. The overall test is stronger than the single period test. Depending on this, the log-rank test can be rewritten as:

$$log - rank\ statistic = \frac{[\sum_{i=1}^{F}(D_{i,1} - \frac{D_i}{N_i}N_{i,1})]^2}{\sum_{i=1}^{F} Var(D_{i,1})} \tag{3.25}$$

Equation(3.25) can be compared with chi-square critical value with $(F-1)$ degrees of freedom. It can be rewritten as:

$$log - rank\ statistic = \frac{\sum_{i=1}^{F}(D_{i,1} - \frac{D_i}{N_i}N_{i,1})}{\sum_{i=1}^{F}\sqrt{Var(D_{i,1})}} \tag{3.26}$$

Equation(3.26) follows the standard normal distribution asymptotically. The null hypothesis is that the $D_{i,1}/N_{i,1} = D_{i,2}/N_{i,2}$ with $i = 1,2,.....,F$. The null hypothesis can be also rewritten as $\hat{S}_{i,1} = \hat{S}_{i,2}$ with $i = 1,2,.....,F$. The log-rank test can be simplified as:

$$log - rank\ statistic = \frac{\sum_{i=1}^{F}(D_{i,1} - E_{i,1})}{\sum_{i=1}^{F}\sqrt{V_{i,1}}} \tag{3.27}$$

One important thing is that the log-rank test can be viewed as the score test came from the partial likelihood function of the Cox model. If the sample includes right-truncated observations, the Cox model can be applied to regress it and provides the score test result.

3.3.2 Weighted Log-Rank Test

The weighted log-rank family tests including the generalized Wilcoxon test are more robust than the log-rank test. Generally, the log-rank test can be thought as the special case of the weighted log-rank test when the weighted coefficient is equal to 1 for each period. The

general form of the weighted log-rank test can be written as:

$$weighed\ log-rank\ statistic = \frac{\sum_{i=1}^{F} W_i(D_{i,1} - E_{i,1})}{\sum_{i=1}^{F} \sqrt{Var(W_i(D_{i,1} - E_{i,1}))}} \tag{3.28}$$

Where W_i is the weighted coefficient. When the weighted coefficient $W_i = 1$, the weighted log-rank test is the same as the log-rank test. The advantage of the weighted log-rank test is that it can assign more weight to the important period (or group) depending on the different data. Marsaglia (1965) and Breslow (1970) derived the generalized Wilcoxon test belonging to the weighted log-rank test. They assumed that there existed the weighted coefficient $W_i = N_i$. In other words, Gehan and Breslow's weighted coefficient pays more attention to the early period since the number of observations N_i go down with the increasing of i. Tarone and Ware (1977) also provided a weighted coefficient $W_i = \sqrt{N_i}$. Tarone and Ware's weighted coefficient also pay more attention to the early period while this effect is weaker than Gehan and Breslow's method. Peto and Peto (1972) also provided a weighted coefficient for the log-rank test. They applied the KM estimator to be the weighted coefficient and defined it as $W_i = \prod_{k=1}^{i}(1 - \frac{D_i}{N_i+1})$. The Peto-Peto's weighted coefficient also pays more attention to the early periods. Peto-Peto's weighted log-rank test is more robust than any other weighted log-rank test when there exist many censored observations. Fleming et al. (1987) provided a weighted coefficient $W_i = S_{i-1}^{r}(1 - S_{i-1})^{q}$. The r and q can be any values. If $r = 1$ and $q = 0$, it is the Peto-Peto weighted coefficient; if $r = 0$ and $q = 0$, it is the regular log-rank test; When $r > q$, the test pays more attention to the early difference. Since the micro data has few price spells in the later period, so it should pay more attention to the early difference between the AF and BF groups. The Peto-Peto and Fleming-Harrington's weighed coefficient are applied in the CPI micro data.

In addition, the Mann-Whitney U test is also known as the Wilcoxon rank sum test. If all the observations are uncensored, the Mann-Whitney U test can be applied to evaluate

the difference between the two groups. The details of the U test can be found at Mann and Whitney (1947).

3.3.3 Empirical Results

There are different ways to deal with the CPI micro data compared with section (3.2). In this section, all the price spells are included even the price increased more than 130% or decreased more than 75%. But all the sales are still excluded.

In case 1, the price quotes of the CPI micro data are applied. This data is divided into two groups: the BF group includes the uncensored price spells starting from February 2003 to December 2016. This group can be assumed as the treatment group. It means that both the quote date and the start date of the price spells are February 2003. The second group includes the uncensored price spells starting from February 2008 to December 2016. The AF group can be known as the control group including the financial crisis period (AF group). In the AF group, both the quote date and the start date of the price spells are February 2008. All the price spells start from February so that they have the same seasonality effect. In other words, the seasonality effect does not affect the results. The motivation is that most of the price spells start from February each year in the UK CPI micro data.

In the BF control group, there exist 18327 price spells. The mean duration of the price spells is around 6.5 months. In the AF control group, the total number of the price spells is 23295. The mean duration of the price adjustment is 5 months in the AF group. [5]

Table (3.9) show the different two-sample tests for the two groups. The p value of U test which is known as the Wilcoxon rank sum test is close to 0. Therefore, the AF and the BF groups do not have the same median. Depending on the regular log-rank test, the chi-square value is equal to 473 and p value is close to zero. With respect to the Peto-Peto log-rank test, the test statistic is 240 and p value is close to zero. Modified the Fleming's weighed

[5]The results may be different from section(3.2) since we do not delete the unreasonable price increases and decreases.

Table 3.9 Non-Parametric Tests for the Hazard Function in Case 1

	$Test-value$	$P-value$
$Wilcoxon\ Rank\ Test$	232004912	0
$Log-rank\ Test$	473	0
$Peto-Peto\ log-rank\ test(r=1)$	240	0
$Weighted\ Log-rank\ test(r=0.5)$	390	0

coefficient with $r = 0.5$, the null hypothesis is still rejected. Therefore, the null hypothesis of the log-rank family tests is rejected. In other words, the hazard functions of AF and BF groups follow the different distributions.

In case 2, the 16 years UK CPI micro data is applied. The data are still divided into two groups: the first group includes the observations from January 1999 to December 2007. The first group is named BF group including 1697620 price spells. The Second group includes the observations from January 2008 to December 2016. The second group is named as AF group which including 2335756 price spells. The right-censored data are considered in case 2. However, as Dixon (2012) showed, if the right-censored data is dealt with the traditional way, the KM estimator does not fit the real situation of the UK's economy. Therefore, the right-censored data are treated as the uncensored data. In table (3.10), the p value of U

Table 3.10 Non-Parametric Tests for the Hazard Function in Case 2

	$Test-value$	$P-value$
$Wilcoxon\ Rank\ Test$	$2.035349*10^{12}$	0
$Log-rank\ Test$	1256	0
$Peto-Peto\ log-rank\ test(r=1)$	2252	0
$Weighted\ Log-rank\ test(r=0.5)$	2320	0

statistic is 0. The regular log-rank test statistic is equal to 1256 and the p value is equal to zero. The p value of the Peto-Peto weighted statistic is still close to 0. Therefore, there exist the different hazard functions between the BF and AF groups.

3.4 Conclusion

In this chapter, we calculate the frequency and the size of price changes by using the UK CPI micro data. The log-rank family tests are introduced. Depending on the empirical results, the AF and the BF groups have the different hazard functions and survival functions when the seasonality effect is deleted or included. In addition, the Cox model and AFT model applied to regress the data from January 1999 to December 2016 in section (3.2). The empirical results show that the price changed more frequently after the financial crisis. In section (3.3), the CPI micro data is evaluated by the non-parametric method. The CPI micro data is dealt with two ways: one is the cohort study; the other is the data without deleting any price spells. The log-rank family tests show that the hazard functions and survival functions follow different distributions between the AF and the BF groups.

References

Aalen, O. O. (1978). Non parametric inference for a family of counting processes. *Annals of Statistics*, 6:701–726.

Alekseyenko, A. V. (2016). Multivariate welch t-test on distances. *Bioinformatics*, 32(23):3552–3558.

Andersen, P. K., Ørnulf Borgan, Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer.

Baharad, E. and Eden, B. (2004). Price rigidity and price dispersion: Evidence from micro data. *Economic Modelling*, 7:613–641.

Bebu, I., Luta, G., Mathew, T., Kennedy, P. A., and Agan, B. K. (2016). Parametric cost-effectiveness inference with skewed data. *Computational Statistics and Data Analysis*, 94:210–220.

Bils, M. and Klenow, P. J. (2004). Some evidence on the importance of sticky prices. *Journal of Political Economy*, 33:1–26.

Bohoris, G. A. (1994). Comparison of the cumulative-hazard and Kaplan–Meier estimators of the survivor function. *IEEE Transactions on Reliability*, 43(2):230–232.

Breslow, N. (1970). A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika*, 57(3):579–594.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics*, 2:437–453.

Briggs, A. H., Mooney, C. Z., and Wonderling, D. E. (1999). Constructing confidence intervals for cost-effectiveness ratios: an evaluation of parametric and non-parametric techniques using Monte Carlo simulation. *Statistics in Medicine*, 18:3245–3262.

Brown, M. (1984). On the choice of variance for the log rank test. *Biometrika*, 71(1):66–74.

Bunn, P. and Ellis, C. (2009). Price-setting behaviour in the United Kingdom: A microdata approach. *Bank of England Quarterly Bulletin*.

Buonaccorsi, J. P. (2005). Fieller's theorem. *Encyclopedia of Environmetrics*, pages 773–775.

Carroll, C. D., Slacalek, J., and Sommer, M. (2011). International evidence on sticky consumption growth. *The Review of Economics and Statistics*, 93(4):1135–1145.

Cavallo, A. (2016). Scraped data and sticky prices. *Working Paper*.

Cedilnik, A., Košmelj, K., and Blejec, A. (2004). The distribution of the ratio of jointly normal variables. *Metodološki zvezki*, 1(1):99–108.

Cedilnik, A., Košmelj, K., and Blejec, A. (2006). Ratio of two random variables: A note on the existence of its moments. *Metodološki zvezki*, 3(1):1–7.

Coenen, G., Mohr, M., and Straub, R. (2008). Fiscal consolidation in the euro area: Long-run benefits and short-run costs. *Economic Modelling*, 25:912–932.

Colosimo, E., Ferreira, F., Oliveira, M., and Sousa, C. (2002). Empirical comparisons between Kaplan-Meier and Nelson-Aalen survival function estimators. *Journal of Statistical Computation and Simulation*, 72(4):299–308.

Cox, C. (1990). Fieller theorem, the likelihood and the delta method. *Biometrics*, 46(3):709–718.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Dixon, H. (2012). A unified framework for using micro-data to compare dynamic time-dependent price-setting models. *BE Journal of Macroeconomics (Contributions)*, 12:1–43.

Dixon, H. and Bihan, H. L. (2012a). Generalised taylor and generalised calvo price and wage setting: Micro-evidence with macro implications. *Economic Journal*, 122(560):532–554.

Dixon, H. and Bihan, H. L. (2012b). Generalised taylor and generalised calvo price and wage setting: Micro-evidence with macro implications. *Economic Journal*, 122:532–554.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association*, 76(374):312–319.

Fan, M. and Zhou, X. (2007). A simulation study to compare methods for constructing confidence intervals for the incremental cost-effectiveness ratio. *Statistics in Medicine*, 7:57–77.

Fieller, E. C. (1932). The distribution of the index in a normal bivariate population. *Biometrika*, 24:175–185.

Fieller, E. C. (1940). The biological standardization of insulin. *Supplement to the Journal of the Royal Statistical Society*, 7(1):1–64.

Fieller, E. C. (1954). Some problems in interval estimation. *Journal of Royal Statistical Society. Series B (Methodological)*, 16(2):175–185.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics*, 42(4):845–854.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Process and Survival Analysis*. John Wiley & Sons, Inc.

Fleming, T. R., Harrington, D. R., and O'Sullivan, M. (1987). Supremum versions of the log-rank and generalized wilcoxon statistics. *Journal of the American Statistical Association*, 82(397):312–320.

Fleming, T. R., O'Fallon, J. R., O'Brien, P. C., and Harrington, D. P. (1980). Modified kolmogorov-smirnov test procedures with application to arbitrarily right-censored data. *Biometrics*, 36(4):607–625.

Franz, V. H. (2007). Ratios: A short guide to confidence limits and proper use. *Working Paper*.

Gardiner, J. C., Huebner, M., Jetton, J., and Bradley, C. J. (2001). On parametric confidence intervals for the cost-effectiveness ratio. *Biometrical Journal*, 43(3):283–296.

Gehan, A. (1965). Generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*, 52:203–223.

Gillesple, M. J. and Fisher, L. (1979). Confidence bands for the Kaplan-Meier survival curve estimate. *Annals of Statistics*, 7:920–924.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81:515–526.

Greenwood, M. (1926). The natural duration of cancer. *Reports on Public Health and Medical Subjects*, 33:1–26.

Guiard, V. (1989). Some remarks on the estimation of the ratio of the expectation values of a two-dimensional normal random variable (correction of the theorem of milliken). *Biometika*, 31(6):681–697.

Heinze, G., Gnant, M., and Schemper, M. (2003). Exact log-rank tests for unequal follow-up. *Biometrics*, 59(4):1151–1157.

Hwang, J. T. G. (1995). Fieller's problems and resampling techniques. *Statistica Sinica*, 5:161–171.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data, 2nd Edition*. John Wiley & Sons, Inc.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

Kiviet, J. F. and Phillips, G. D. A. (2014). Improved variance estimation of maximum likelihood estimators in stable first-order dynamic regression models. *Computational Statistics and Data Analysis*, 76:424–448.

Kumar, S. and Owen, B. (2013). International evidence on sticky consumption growth. *Working Paper*.

Latta, R. B. (1981). A Monte Carlo study of some two-sample rank tests with censored data. *Journal of the American Statistical Association*, 76(375):713–719.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18(1):50–60.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, 50(3):163–170.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719–748.

Marsaglia, G. (1965). Ratios of normal variables and ratios of sums of uniform variables. *Journal of the American Statistical Association*, 60(309):193–204.

Martinez, R. L. M. C. and Naranjo, J. D. (2010). A pretest for choosing between log rank and wilcoxon tests in the two-sample problem. *International Journal of Statistics*, 68(2):111–125.

Morton, R. (1978). Regression analysis of life tables and related nonparametric tests. *Biomttrika*, 65(2):329–333.

Nair, V. N. (1981). Plots and tests for goodness of fit with randomly censored data. *Biometrika*, 68:99–103.

Nair, V. N. (1984). Confidence bands for survival functions with censored data: A comparative study. *Technometrics*, 26:265–275.

Nakamura, E. and Steinsson, J. (2008). Five facts about prices: A re-evaluation of menu cost models. *Quarterly Journal of Economics*, 123(4):1415–1464.

Nelson, W. (1972). Theory and application of hazard plotting for censored failure data. *Technometrics*, 14:945–965.

of Labor Statistics, B. (2015). The consumer price index. *Bureau of Labor Statistics*.

Peto, R. and Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society*, 135(2):186–207.

Polsky, D., Glick, H. A., Willke, R., and Schulman, K. (1997). Confidence intervals for cost-effectiveness ratios a comparison of four methods. *Health Economics*, 6(3):243–252.

Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, 68(1):167–179.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, 2(6):110–114.

Tarone, R. E. and Ware, J. (1977). On distribution-free tests for equality of survival distributions. *Biometrika*, 64(1):156–160.

Taylor, J. B. (1980). Aggregate dynamics and staggered contracts. *Journal of Political Economy*, 88(1):1–23.

Tsai, W. Y., Jewell, N. P., and Wang, M. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74(4):883–886.

Veronese, G., Fabiani, S., Gattulli, A., and Sabbatini, R. (2005). Consumer price behaviour in Italy: Evidence from micro CPI data. *Working Paper*.

Wang, H. and Zhao, H. (2008). A study on confidence intervals for incremental cost-effectiveness ratios. *Biometrical Journal*, 50:505–514.

Wang, Y. Y. (1971). Probabilities of the type I errors of the welch tests for the Behrens-Fisher problem. *Journal of the American Statistical Association*, 66(335):605–608.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *J*, 29:350–362.

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika*, 24:28–35.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38:330–336.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, 61(1):165–170.

Zimmerman, D. W. and Zumbo, B. D. (1993). Rank transformations and the power of the student t test and welch's t test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology*, 47(3):523–539.