DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

EXAMINATION OF THE PROPORTIONAL HAZARD ASSUMPTION IN COX REGRESSION MODEL

by
Ayhan YAĞCI

April, 2018
İZMİR

EXAMINATION OF THE PROPORTIONAL HAZARD ASSUMPTION IN COX REGRESSION MODEL

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science in Statistics. Statistics Program

by
Ayhan YAĞCI

April, 2018
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"EXAMINATION OF THE PROPORTIONAL HAZARD ASSUMPTION IN COX REGRESSION MODEL"** completed by **AYHAN YAĞCI** under supervision of **Asst. Prof. Dr. ÖZGÜL VUPA ÇİLENGİROĞLU** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of the Master of Science.

.....

Asst. Prof. Dr. Özgül VUPA ÇİLENGİROĞLU

Supervisor

.....

Assoc. Prof. Özlem GÜRÜNLÜ ALMA

(Jury Member)

.....

Prof. Dr. Selma GÜRLER

(Jury Member)

Prof. Dr. Kadriye ERTEKİN

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

First of all, I would especially like to thank Asst. Prof. Dr. Özgül VUPA ÇİLENGİROĞLU for her guidance and encouragement throughout the preparation of this thesis. I greatly appreciate the time and effort that she helped me complete my Master degree, as well as her support and advice over the past four years. Certainly, without her supervision, the thesis can not be accomplished.

I would like to thank Dr. Ümit KUVVETLİ who helped me to find a data set and who provided me a lot of suggestions and advice.

I wish to express my deepest gratitude to Behiye YAĞCI, my mother, Süleyman YAĞCI, my father for their encouragement and support during my studies.

I also would like to thank to Tuğçe ERDOĞAN who is as good friends, since started master. Finally, I would like thank to everybody who support and encourage me during my thesis study.

Ayhan YAĞCI

EXAMINATION OF THE PROPORTIONAL HAZARD ASSUMPTION IN COX REGRESSION MODEL

ABSTRACT

The term survival analysis stands for the analysis of data where we have recorded the time period from a defined time origin up to a certain event for a number of individuals. The event is often called a “failure” and the length of time is called the “failure time”. The difference between individuals’ survival periods can be analysed using the nonparametric Kaplan-Meier method, semi parametric life table method and parametric methods. In addition, the most common approach to model covariate effects on survival is the Cox regression (proportional hazards) model where it is necessary that the proportional hazards assumption holds Cox regression model.

The methods used in survival analysis are now applied in many fields, such as biology, engineering, medicine, quality control and credit risk modeling in finance. However, the application of survival analysis in the field of transport in Turkey has not been studied much. We studied a real data taken from İzmir ESHOT (Electricity, Water, Water Gas, Bus and Trolleybus) headquarters consisting of drivers accidents that happened between 2012-2014 years. The outcome variable is the survival time that is the time passed until the first accident happens. The explanatory variables are the district, the traffic jam, the bus capacity, the passenger number, the station number, the bus type, the driver age and the driver experience. We examined the proportional hazards assumption with “log(-log) plots”, “Schoenfeld residual plots”, “expected vs observed plots”, “Arjas plots” and “using time dependent explanatory variables methods”.

Keywords: Survival analysis, Cox regression model, log(-log) plot, Arjas plot, Schoenfeld residual plot

COX REGRESYON MODELİNDE ORANSAL HAZARD VARSAYIMININ İNCELENMESİ

ÖZ

Sağ kalım analizi terimi, belirli bir zaman dilimde belirli bir olaya kadar geçen süreyi kaydettiğimiz verilerin analizini ifade eder. Olay genellikle "başarısızlık" olarak adlandırılır ve zamanın uzunluğu "başarısızlık zamanı" olarak adlandırılır. Bireylerin sağ kalım süreleri arasındaki fark, parametrik olmayan Kaplan-Meier yöntemi, yarı parametrik yaşam tablosu yöntemine ve parametrik yöntemlere göre analiz edilebilir. Ek olarak, sağ kalım üzerindeki değişkenleri modellemek için en yaygın yaklaşım Cox regresyon (orantısal tehlikeler) modelidir. Ancak Cox regresyon modelini kullanmak için orantılı tehlike varsayımının sağlanması gerekmektedir.

Sağ kalım analizinde kullanılan yöntemler, günümüzde biyoloji, mühendislik, tıp, kalite kontrol ve finans alanında kredi riski modellemesi gibi birçok yerde uygulanmaktadır. Türkiye'de ulaşım alanında sağ kalım analizinin uygulanması çok yaygın değildir. Bu çalışmada. 2012-2014 yılları arasında İzmir ESHOT (Elektrik, Su, Havagazı, Otobüs ve Troleybüs) genel müdürlüğünden şoförlerin yaptığı kazalardan alınan gerçek bir veri seti kullanılmıştır. Sonuç değişkeni, ilk kaza gerçekleşene kadar geçen süre olan sağ kalım süresidir. Açıklayıcı değişkenler, kazanın yapıldığı bölge, trafik sıkışıklığı, otobüsün kapasitesi, yolcu sayısı, durak sayısı, otobüs türü, şoförün yaşı ve şoförün deneyimidir. Cox regresyon modelinde oransal hazard varsayımı "log(-log) grafiği", "Schoenfeld artık grafikleri", "beklenen gözlenen grafikleri", "Arjas grafikleri" ve "zamana bağlı açıklayıcı değişken method" yöntemleri kullanılarak incelenmiştir.

Anahtar Kelimeler: Sağkalım analizi, Cox regresyon modeli, log(-log) grafiği, Arjas grafiği, Schoenfeld artık grafiği

CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER ONE
INTRODUCTION

Survival analysis is used by researchers to analyze the elapsed time from the starting point until the occurrence of a defined event. Time can be measured in terms of days, weeks, months or years from the beginning of follow-up of an individual until an event occurs. Event is usually referred to as "failure" because the event of interest is usually death (undesired condition), disease, or another negative individual experience. In some cases, the event may also be positive (e.g. leave from hospital).

Survival analysis utilizes to statistical methods for analyzing survival data. In the past decades, applications of the statistical methods for survival data analysis have been extended beyond biomedical and reliability research to other fields. For example, lifetime of electronic devices, components or systems (reliability engineering), duration of first marriage (sociology), criminology, marketing, health insurance practice, business and economics. One of the reasons for not using classical statistical methods in survival analysis is censoring and the other is time-dependent covariates. There are two important advantages of survival analysis. These are: the need for information on the distribution of survival time and the use of information about censored data. Especially in a long period of follow-up, many subjects drop out. The only information we have about them is the time until the last follow-up. Subject does not experience the event of interest and we don't know the survival time, and this problem is called "censoring". A plus sign as an affix to the censored observation in the literature is used for notating a censoring. In survival analysis, the survival time is denoted by T and it is a nonnegative random variable (T=survival time ($t \geq 0$) and t=specific value for T). The distribution of survival times is usually described by three functions (the survival function, the probability density function and the hazard function). If the distribution of survival time is known to be normal or other known distributions (exponential, lognormal, weibull and gamma etc.), then the parametric statistical methods for survival data are used. If the distribution of survival times is unknown, then the nonparametric statistical methods are used.

Nonparametric or distribution-free methods are quite easy to understand and apply. The Product-Limit (PL or Kaplan Meier) method of estimating the survival function is developed by Kaplan and Meier (1958). The estimator is named after Edward L. Kaplan and Paul Meier. If the data have already been grouped into intervals or the sample size is very large, say in the thousands, or the interest is in a large population, we can use this method.

The most common approach to model covariate effects on survival is the Cox proportional hazards model by Cox (1972). This model takes into account the effect of censored observations. It requires that the data satisfies the proportional hazards assumption for using Cox regression model.

There are several studies based on Kaplan Meier Method and Cox regression proportional hazards assumption in literature. For instance, see Terzi et al. (2005) for an example Cox regression model where was used to determine the prognostic factors that affect survival time in patients with lung cancer. Their results suggested that the model holds proportional hazards assumption with Schoenfeld residuals. Ata et al. (2006) used Cox regression proportional hazards assumption with graphical methods and time dependent covariates. Their data set gathered from a group of patients with lung cancer and they found that the variable does not satisfy the assumption of proportionality. In Yay et al. (2007), the data are taken from 100 women which have just gave a birth in Istanbul University Cerrahpaşa Medical Faculty. These women have been observed throughout the four months after their delivery. The aim of this process is to ascertain the factors that are effective on lactation time with Cox regression method. According to the Cox regression analysis results, the baby's birth weight, first milking time, and baby food feeding variables were found to be significant in the study. Many other researchers use Cox regression and check the proportional hazards assumption. Some of them are: Yılmaz et al. (2013), Karanlık et al. (2006), Vehid et al. (2003), Sertkaya et al. (2003).

The purpose of this thesis is to consider survival analysis based on the proportional hazards (PH) model. In addition, this thesis examines survival methods

(survival analysis, Cox regression model) in the real data set taken from İzmir ESHOT (Electricity, Water, Water Gas, Bus and Trolleybus) headquarter, containing drivers accidents between 2012-2014 years. The event of interest is that the drivers experience the first accident and the survival time is the time passed until the first accident. The explanatory variables are the district, the traffic jam, the bus capacity, the passenger number, the station number, the bus type, the driver age and the driver experience. In this thesis, we examined the proportional hazards assumption with "log(-log) plots", "Schoenfeld residual plots", "expected vs observed plots", "Arjas plots" and "using time dependent explanatory variables methods". The analysis in this thesis is conducted in SPSS 16.0 statistical package.

CHAPTER TWO
SURVIVAL ANALYSIS

2.1 Introduction

Survival analysis involves the consideration of the time between a fixed starting point and a terminating event. Several statistical methods have been proposed for modeling survival analysis data. If the survival time is known then the parametric methods are used. On the other hand, the survival time is sometimes not known and these studies will have censored data, so it is necessary to use Kaplan-Meier method which is a nonparametric method.

> Nonparametric or distribution-free methods are quite easy to understand and apply. These methods are less efficient than parametric methods when survival times follow a theoretical distribution and more efficient when no suitable theoretical distributions are known. The nonparametric methods to analyze survival data are used before attempting to fit a theoretical distribution. If the main objective is to find a model for the data, estimates obtained by nonparametric methods and graphs can be helpful in choosing a distribution. The product-limit (PL) method of estimating the survivorship function developed by Kaplan and Meier (1958). If the data have already been grouped into intervals or the sample size is very large or the interest is in a large population, it may be more convenient to perform a life-table analysis. The PL estimates and life-table estimates of the survivorship function are essentially the same. Many authors use the term life table estimates for the PL estimates. The only difference is that the PL estimate is based on individual survival times, whereas in the life-table method. survival times are grouped into intervals. The PL estimate can be considered as a special case of the life-table estimate where each interval contains only one observation (Lee & Wang. p. 64).

Nonparametric methods for estimating survival functions are shown in Figure 2.1.

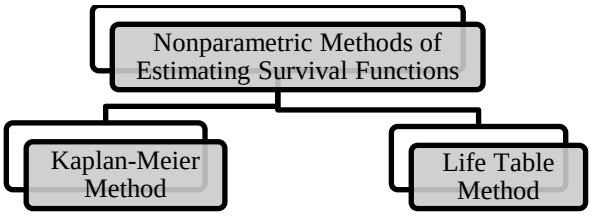

Figure 2.1 Nonparametric methods of estimating survival functions

2.2 Terminology and Notation

In survival analysis, the survival time is denoted by T and it is a nonnegative random variable denoting the time to event of interest (survival time/event time/failure time). T can either be discrete or continuous.

$T = \text{survival time } (T \geq 0) \text{ and } t = \text{specific value for T}$

The random variable for censored data are shown below:

$$\text{Random Variable} = \begin{cases} 1 & \text{if failure} \\ 0 & \text{if censored} \end{cases} \tag{2.1}$$

The functions of survival times are described by three functions. These are;

- $S(t)$: Survival Function (Survivorship Function)
- $f(t)$: Probability Density Function (Density Function)
- $h(t)$: Hazard Function

2.3 Kaplan-Meier Method

The analysis of survival data requires special methods because some observations are censored as the event of interest has not occurred for all subjects. One of these

methods is the Kaplan-Meier survival analysis which is a nonparametric method. Kaplan-Meier survival analysis is a method that involves generating tables and plots of the survival or the hazard function for the event data. Kaplan-Meier method is based on the ideas of conditional probability. Kaplan-Meier survival analysis does not determine the effect of the covariates on either function.

Kaplan-Meier method is also called one-sample nonparametric method. Its survival function can be estimated with Kaplan-Meier plot, life-table or cumulative hazard estimator. It is a very popular method for its advantages, such as, it does not require any mathematical assumption for its hazard function or proportional hazard. It is widely applied for studying survival function of population. It is very commonly used to compare the survival functions among groups. If the sample size is large enough, its estimated survival function represents that of population. It is very popular, because it treats censored data well, particularly right-censoring data (Hu. 2013).

The Kaplan-Meier method is based on the basic idea that the probability of surviving k or more periods from entering the study is a product of the k observed survival rates for each period given by the following:

$$\hat{S}(k) = p_1 \times p_2 \times ... \times p_k \tag{2.2}$$

Where is p_1 the proportion surviving the first period.

- p_1 denotes the proportion of patients surviving at least one year.
- p_2 denotes the proportion of patients surviving the second year after they have survived one year.
- p_3 denotes the proportion of patients surviving the third year after they have survived two years.
- p_k denotes the proportion of patients surviving the k^{th} year after they have survived $k-1$ years.

Therefore, The PL estimate of the probability of surviving any particular number of years from the beginning of study is the product of the same estimate up to the preceding year and the observed survival rate for the particular year, that is,

$$\hat{S}(t) = \hat{S}(t-1)p_t \tag{2.3}$$

In practice, the PL estimates can be calculated by constructing a table like Table 2.1 with five columns following the outline below.

Table 2.1 Summarize table for product limit estimates

(1)	(2)	(3)	(4)	(5)
Survival time.	Rank of observations	Uncensored observations	(n-r)/(n-r+1)	$\hat{S}(t)$

1. Column 1 contains all the survival times, both censored and uncensored, in order from smallest to largest. Affix a plus sign to the censored observation. If a censored observation has the same value as an uncensored observation, the letter should appear first.
2. The second column, labeled $i.$,consists of the corresponding rank of each observation in column 1.
3. The third column, labeled r, pertains to uncensored observations only. Let $r = i$
4. Compute $(n-r)/(n-r+1)$ or p_i, for every uncensored observation $t_{(i)}$ in column 4 to give the proportion of patients surviving up to and then through $t_{(i)}$.
5. In column 5, $\hat{S}(t)$ is the product of all values of $(n-r)/(n-r+1)$ up to and including t. If some uncensored observations are tied, the smallest $\hat{S}(t)$ should be used.

To summarize this procedure, let n be the total number of individuals whose survival times, censored or not, are available. Relabel the n survival times in order

of increasing magnitude such that $t_{(1)} \le t_{(2)} \le ... \le t_{(n)}$. Then let is $\hat{S}(t) = \prod_{t_{(r)} \le t} \frac{n-r}{n-r+1}$, where r runs through those positive integers for which $t_{(r)} \le t$ and $t_{(r)}$ is uncensored. The values of r are consecutive integers $1, 2, ..., n$ if there are no censored observations. If there are censored observations, they are not.

Since every person is alive at the first of the event and no one survives longer than $t_{(n)}$, so $\hat{S}\left(t_{(0)}\right) = 1$ and $\hat{S}\left(t_{(n)}\right) = 0$.

The estimated median survival time is the 50^{th} percentile, which is the value of 50^{th} at $\hat{S}(t) = 0.50$. $\hat{S}(t)$ is a step function that starts at 1 and decreases in steps of $1/n$ (if there aren't ties) to zero. When $\hat{S}(t)$ is plotted versus t, the different percentiles of survival time can be seen from the graph or calculated from $\hat{S}(t)$.

2.4 Censoring

Observations are called censored when the information about their survival time is incomplete. In survival analysis, a data set can be exact or censored. Affix a plus sign to the censored observation in the literature. The endpoint doesn't have to be "death"; it can be any well-defined event time until the event of interest that can be menopause, pregnancy, tumor recurrence, cardiovascular death after some treatment intervention, AIDS for HIV patients, a machine part fails, etc.

Types of censoring are shown in Figure 2.2.

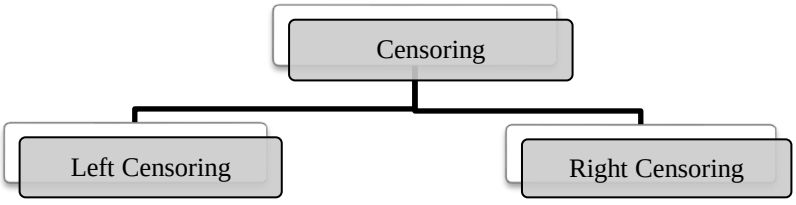

Figure 2.2 Types of censoring

2.4.1 Left Censoring

The event of interest has already occurred before the observation date and it is not known exactly when.

2.4.2 Right Censoring

Right censoring occurs when the study ends while the subject survives but it is not known for exactly how long more.

> Simple approaches researchers might choose to deal with censored data are to set the censored observations to missing or replace the unobserved value of the variable by zero, minimum, maximum, mean value, or a randomly assigned value from the range of possible values. When the censoring is minimal, using one of these approaches can be reasonable. When it is not, these simple solutions can be, however, cause serious bias in estimates and standard errors obtained in subsequent statistical analysis, can discard potentially important information, and can create a sample that is not representative of the population studied (Vermeylen. 2005).

2.5 Functions of Survival Time

First let T denote the survival time. The distribution of survival time can be characterized by three equivalent functions which are "Survival Function", "Probability Density Function" and "Hazard Function". If one of them is known, the others can be obtained.

2.5.1 The Survival Function

Survival function $S(t)$, is defined as the probability that an individual survives longer than time t:

$$S(t) = P(\text{an individual survives longer than } t) = P(T > t) \tag{2.4}$$

From the definition of the cumulative distribution function $F(t)$ of T.

$$S(t) = 1 - P(\text{an individual fails before } t) = 1 - F(t) \tag{2.5}$$

$S(t)$ is a nonincreasing function of time t with the properties: $S(t) = 1$ for $t = 0$ and $S(t) = 0$ for $t = \infty$.

Berkson (1942) recommended a graphic representation of $S(t)$. The graph of $S(t)$ is named the survival curve. The survival function or the survival curve is used to find the 50^{th} percentile (median value) of survival time and to compare survival distributions of two or more groups. The mean is usually used to describe the central tendency of a distribution. However in survival distributions the median is preferable because of a small number of personal with highly long or short lifetimes which reason the mean survival time to be disproportionately large or small (Lee & Wang, p. 9).

2.5.2 The Probability Density Function (Density Function)

The survival time T has a probability density function which is defined as the limit of the probability that an individual fails in the short interval t to $t + \Delta t$ per unit width Δt, or simply the probability of failure in a small interval per unit time. It can be defined as

$$f(t) = \frac{\lim_{\Delta t \to 0} P[\text{an individual dying in the interval } (t, t + \Delta t)]}{\Delta t} \tag{2.6}$$

The graph of $f(t)$ is called the "density curve". The density function has the following two properties:

1. $f(t)$ is a nonnegative function:

$$f(t) \geq 0 \quad \text{for all } t \in (-\infty, \infty) \tag{2.7}$$

2. The area between the density curve and t axis is equal to 1.

$$\int_{-\infty}^{\infty} f(t)\,dt = 1 \tag{2.8}$$

2.5.3 The Hazard Function

The hazard function $h(t)$ of survival time T gives the conditional failure rate. This is defined as the probability of failure during a very small time interval, assuming that the individual has survived to the beginning of the short interval, or as the limiting of the probability that an individual fails in a very short interval, $t + \Delta t$, given that the individual has survived to time t:

$$h(t) = \frac{\lim_{\Delta t \to 0} P\left[\begin{array}{l}\text{an individual fails in the time interval } (t, t+\Delta t) \\ \text{given the individual has survived to t}\end{array}\right]}{\Delta t} \tag{2.9}$$

The hazard function can also be defined in terms of the cumulative distribution function $F(t)$ and the probability density function $f(t)$.

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{2.10}$$

The definition of the hazard function in survival analysis is as shown below.

$$h(t)=\lim_{\Delta t\to 0}\frac{P(t<T\le t+\Delta t\,|,T>t)}{\Delta t}=\lim_{\triangle t\to 0}\frac{P(t<T\le t+\Delta t)}{\Delta t P(T>t)} \tag{2.11}$$

$$=\lim_{\Delta t\to 0}\frac{P(t<T\le t+\Delta t)}{\triangle t S(t)}=\lim_{\Delta t\to 0}\frac{F(t+\Delta t)-F(t)}{\Delta t S(t)}=\frac{F'(t)}{S(t)}=\frac{f(t)}{S(t)} \tag{2.12}$$

The cumulative hazard function is defined as below:

$$H(t)=\int_0^t h(x)dx \tag{2.13}$$

$$H(t)=-\log S(t) \tag{2.14}$$

Thus, at $t=0$, $S(t)=1$, $H(t)=0$ and $t=\infty$, $S(t)=0$, $H(t)=\infty$. The cumulative hazard function can take any value between zero and infinity.

2.6 The Relationship Among Survival Functions $f(t)$, $S(t)$ and $h(t)$

2.6.1 The Relationship Between $f(t)$ and $S(t)$

The relationship between $f(t)$ and $S(t)$ is as shown as:

$$f(t)=\frac{d}{dt}F(t)=\frac{d}{dt}(1-S(t))=-\frac{d}{dt}S(t) \tag{2.15}$$

$$S(t)=1-F(t)=1-\int_{-\infty}^{t} f(s)ds \tag{2.16}$$

$$S(t)=\int_{-\infty}^{\infty} f(s)ds-\int_{-\infty}^{t} f(s)ds=\int_{t}^{\infty} f(s)ds \tag{2.17}$$

2.6.2 The Relationship Between $S(t)$ and $h(t)$

The relationship between $S(t)$ and $h(t)$ is as shown as:

$$h(t) = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\ln(S(t)) \tag{2.18}$$

$$H(t) = -\ln(S(t)) \tag{2.19}$$

$$\ln(S(t)) = -\int_{-\infty}^{t} h(s)ds \tag{2.20}$$

$$S(t) = \exp\left[-\int_{-\infty}^{t} h(s)ds\right] \tag{2.21}$$

2.6.3 The Relationship Between $h(t)$ and $f(t)$

The relationship between $h(t)$ and $f(t)$ is as shown as:

$$f(t) = h(t)S(t) \tag{2.22}$$

$$h(t) = \frac{f(t)}{\int_{t}^{\infty} f(s)ds} \tag{2.23}$$

$$f(t) = h(t)\exp-\left[\int_{-\infty}^{t} h(s)ds\right] = h(t)\exp\left[-H(t)\right] \tag{2.24}$$

2.7 Nonparametric Methods for Comparing Survival Distributions

In survival analysis, estimating the survival probability of a population is important. In addition, investigators want to compare the survival experiences of

different groups. In such cases, the differences between groups can be illustrated by drawing survival curves obtained from the Kaplan-Meier method, but this will only give a rough comparison and does not reveal whether the differences are statistically significant or not. For this reason nonparametric methods for comparing survival distributions are used.

2.8 The Comparison of Two Survival Distributions

Lee & Wang (2003) describe the comparison of two survival distributions. According to this description, suppose that there are n_1 and n_2 patients who receive treatments I (1) and II (2), respectively. Let $x_1,...,x_{r_1}$ be the r_1 failure observations and $x^+_{r_1+1},...,x^+_{n_1}$ the $n_1 - r_1$ censored observations in group I. In group II, let $y_1,...,y_{r_2}$ be the r_2 failure observations $y^+_{r_2+1},...,y^+_{n_2}$ the $n_2 - r_2$ censored observations. That is, at the end of the study $n_1 - r_1$ patients who received treatment I and $n_2 - r_2$ patients who received treatment II are still alive. Suppose that the observations in group I are samples from a distribution with survival function $S_1(t)$ and the observations in group II are samples from a distribution with survival function $S_2(t)$. Then the null and alternative hypothesis are:

$H_0 : S_1(t) = S_2(t)$ (treatment I and II are equally effective)

$H_1 : S_1(t) > S_2(t)$ (treatment I is more effective than II)

or

$H_1 : S_1(t) < S_2(t)$ (treatment II is more effective than I)

or

$H_1 : S_1(t) \neq S_2(t)$ (treatment I and II are not equally effective)

Comparison of two survival curves can be done using one of the statistical hypothesis tests called the Log Rank test, Gehan's Generalized Wilcoxon Test and Mantel Haenszel Test.

When there are no censored observations, nonparametric tests can be used to compare two survival distributions. For example, the Wilcoxon (1954) test or the Mann-Whitney (1947) U-test test.

2.8.1 Log-Rank Test

Log rank test is used to test the null hypothesis that there is no difference between the population survival curves (i.e. the probability of an event occurring at any time point is the same for each population). An approximately chi-square test is used to test the significance of a mathematical expression involving the observed and expected number of the event of interest.

The test statistic is calculated as follows:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \tag{2.25}$$

where "O_1" and "O_2" are the total number of observed events in groups I and II, respectively, "E_1" and "E_2" are the total number of expected events. The statistic χ^2 follows a chi-square 1 degree of freedom.

E_2 is calculated as follows:

$$E_2 = \sum_{i=1}^{k} \frac{d_i}{r_i} r_{2i} \tag{2.26}$$

Where r_{2i} is the number alive from group II at the time of event i, d_i is the number of failures and E_1 can be calculated as $n - E_2$, where n is the total number of events.

2.8.2 Gehan's Generalized Wilcoxon Test

In Gehan's generalized Wilcoxon test every observation x_i or x_i^+ in group I is compared with every observation y_j or y_j^+ in group II and score U_{ij} is given to the result of every comparison.

Define U_{ij};

$$U_{ij} = \begin{cases} +1 & \text{if} \quad x_i > y_j \quad \text{or} \quad x_i^+ \geq y_j \\ 0 & \text{if} \quad x_i = y_j \text{ or } x_i^+ < y_j \quad \text{or} \quad y_j^+ < x_i \quad \text{or} \quad \left(x_i^+, y_j^+\right) \\ -1 & \text{if} \quad x_i < y_j \quad \text{or} \quad x_i \leq y_j \end{cases} \tag{2.27}$$

and calculate the test statistic W, where

$$W = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} U_{ij} \tag{2.28}$$

The test statistic W under H_0 can be considered approximately normally distributed with mean 0 and variance V(W).

Since W has an asymptotically normal distribution with mean zero and variance in $V(W) = \dfrac{n_1 n_2 \sum_{i=1}^{n_1+n_2} U_i^2}{(n_1+n_2)(n_1+n_2-1)}$, $Z = W / \sqrt{V(W)}$ has standard normal distribution.

2.8.3 Mantel Haenszel Test

The Mantel-Haenszel (1959) test is generally used for comparing survival times between two groups when adjustments for other prognostic factors are necessary. The test has been used in many clinical and epidemiological studies as a method of

controlling the effects of confounding variables. To use the Mantel-Haenszel test, the data are stratified by the confounding variable and cast into a sequence of 2×2 tables, one for each stratum.

Let s be number of strata, n_{ji} be the number of individuals in group j, $j=1,2$ and stratum i, $i=1,...,s$ and d_{ji} be the number of deaths (or failures) in group j and stratum i. For each of the s strata, the data can be represented by a 2×2 contingency table.

Table 2.2 Contingency table for two groups

Group	Number of Deaths	Number of Survivors	Total
1	d_{1i}	$n_{1i}-d_{1i}$	n_{1i}
2	d_{2i}	$n_{2i}-d_{2i}$	n_{2i}
Total	D_i	S_i	T_i

In Table 2.2, the number of deaths in two groups are shown by d_{ji}, ($j=1,2$) total number of deaths are shown by D_i, total number of survivors are shown by S_i, total number of deaths and number of survivors are shown by T_i.

The chi-square test statistic is evaluated as:

$$X^2=\frac{\left[\sum_{i=1}^{s}d_{1i}-\sum_{i=1}^{s}E\left(d_{1i}\right)\right]^2}{\sum_{i=1}^{s}Var\left(d_{1i}\right)} \tag{2.29}$$

where $E\left(d_{1i}\right)=\frac{n_{1i}D_i}{T_i}$ and $Var\left(d_{1i}\right)=\frac{n_{1i}n_{2i}D_iS_i}{T_i^2\left(T_i-1\right)}$.

This statistic follows the chi-square distribution with 1 degree of freedom. So decision rule, a computed chi-square value larger than the chi-square table value for the significance level chosen indicates a significant difference in survival between two groups.

2.9 An Example for Kaplan-Meier Method

Suppose that the survival times (in months since transplant) for fourteen patients who received bone marrow transplants are shown in following Table 2.3. We can apply the Kaplan-Meier method for this data set.

Table 2.3 Summary data for marrow transplants patients

Survival Time (months) (+ : censored observation)	Gender (1:women, 0:Men)
3	1
4	0
+4	1
6	1
8	0
10	1
+10	0
11	0
12	0
+14	1
16	1
+17	0
18	1
20	0

The following table shows the steps applied in the Table 2.1.

Table 2.4 Calculations for the Kaplan-Meier estimate of the survival function for marrow transplants patients

Survival Time	Rank of observations	Uncensored observations	$(n-r)/(n-r+1)$	$\hat{S}(t)$
3	1	1	13/14=0.93	0.93
4	2	2	12/13=0.92	0.93x0.92=0.86
+4	3			
6	4	4	10/11=0.91	0.86x0.91=0.78
8	5	5	9/10=0.90	0.78x0.90=0.70
10	6	6	8/9=0.89	0.70x0.89=0.62
+10	7			
11	8	8	6/7=0.86	0.62x0.86=0.53
12	9	9	5/6=0.83	0.53x0.83=0.45
+14	10			
16	11	11	3/4=0.75	0.45x0.75=0.33
+17	12			
18	13	13	1/2=0.50	0.33x0.50=0.17
20	14	14	0.00	0.17x0.00=0.00

- First column contains all the survival times, both censored and uncensored, in order from smallest to largest. Affix a plus sign to the censored observation.
- The second column contains the corresponding ranks of each observation in first column.
- The third column contains uncensored observations.
- The fourth column contains $(n-r)/(n-r+1)$ for every uncensored observation.
- The last column contains the product of all values of $(n-r)/(n-r+1)$.

According to Table 2.4 median survival time is 12 months because $\hat{S}(t)=0.45$ is the value nearly closest to 0.5. As a result, survival times (in months since transplant) for the fourteen patients who received bone marrow transplants have 12 months median survival time. We can also see this result in the following figure.

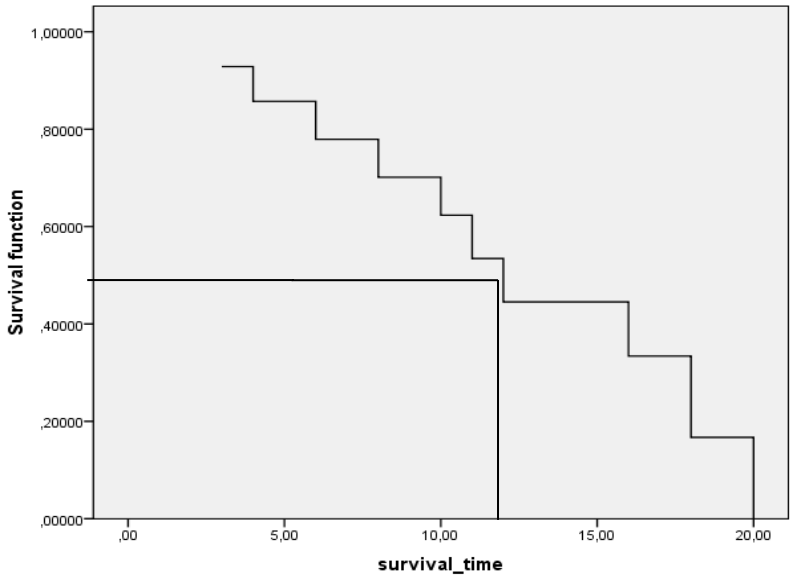

Figure 2.3 Survival time versus survival function values

2.10 An Example for the Log-Rank Test

A hypothetical data set is given by Bewick (2004), which is given in Table 2.5. For this data set the event is the death of the patient, and so the censored data are those where the patient survived or the outcome is unknown. The null hypothesis is that the survival distributions are the same in the two treatment groups and the alternative hypothesis is that the survival distributions are different in the two groups.

Table 2.5 Survival time, age and outcome for a group of patients diagnosed with a disease and receiving one of two treatments

Patient Number	Survival Time (days)	Outcome	Treatment	Age (years)
1	1	died	2	75
2	1	died	2	79
3	4	died	2	85
4	5	died	2	76
5	6	unknown	2	66
6	8	died	1	75
7	9	survived	2	72
8	9	died	2	70
9	12	died	1	71
10	15	unknown	1	73
11	22	died	2	66
12	25	survived	1	73
13	37	died	1	68
14	55	died	1	59
15	72	survived	1	61

Table 2.6 shows the calculation of the expected number of deaths for treatment group 2.

Table 2.6 Calculations for the log-rank test to compare treatments for the data in Table 2.5

Survival Time (days)	Treatment group	Number known to be alive (r_i)	Deaths (d_i)	Risk for death (d_i / r_i)	Number known to be alive from treatment group 2 (r_{2i})	Expected number of events in treatment group 2 (E_2)
0						
1	2	15	2	2/15=0.133	8	8x0.133=1.07
4	2	13	1	1/13=0.177	6	6x0.177=0.46
5	2	12	1	1/12=0.083	5	5x0.083=0.42
6+	2	11	0	0/11=0	4	4x0=0
8	2	10	1	1/10=0.1	3	3x0.1=0.3
9	1	9	1	1/9=0.111	3	3x0.111=0.333
9+	2	8	0	0/8=0	2	2x0=0
12	2	7	1	1/7=0.143	1	1x0.143=0.143
15+	1	6	0	0/6=0	1	1x0=0
22	1	5	1	1/5=0.2	1	1x0.2=0.2
25+	2	4	0	0/4=0	0	0x0=0
37	1	3	1	1/3=0.133	0	0x0.133=0
55	1	2	1	½=0.5	0	0x0.5=0
72+	1					
						Total E_2 =2.92

For the data in Table 2.6, the total number of expected deaths for treatment group 2 is calculated as 2.92 and the total number of observed deaths is 10, giving a total number of expected deaths for treatment group 1 of 10 – 2.92 = 7.08. The value of the test statistic is therefore calculated as follows:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(4 - 7.08)^2}{7.08} + \frac{(6 - 2.92)^2}{2.92} = 4.59$$

χ^2 table value 3.84 is and the log rank chi-square value is 4.59 so we reject the null hypothesis. It can be said that there is a difference between the treatments 1 and 2 with respect to for survival times at the level of α=0.05.

CHAPTER THREE
COX REGRESSION

3.1 Cox Regression Model

The most common approach to model covariate effects on survival is the Cox (proportional hazards) model by Cox (1972). The Cox proportional hazards model is a specialized regression method. The main aim of a regression model is to analyze relationships among variables. The analysis is carried out through the estimation and predictions of a relationship. The Cox proportional hazards model takes into account the effect of censored observations. Therefore, the purpose of the Cox proportional hazards method is to investigate the effects of several independent variables on the probability of survival.

In a Cox proportional hazards regression model, the measure of effect is the hazard rate which is the risk of failure, given that the participant has survived up to a specific time. In other words, the hazard rate is the measure of the risk of occurrence of an event at a point in time. A probability must lie in the range 0 to 1. However, the hazard represents the expected number of events per one unit of time. It means that the hazard in a group can exceed 1.

In a Cox proportional hazards regression model, we are interested in comparing groups with respect to their hazards. Hazard ratio is used and this is analogous to an odds ratio in the setting of logistic regression analysis. The hazard ratio is the ratio of the total number of observed to expected events in two independent comparison groups. It is given as:

$$\mathrm{HR} = \frac{\sum \mathrm{O}_{\mathrm{treatment},t} / \sum \mathrm{E}_{\mathrm{treatment},t}}{\sum \mathrm{O}_{\mathrm{control},t} / \sum \mathrm{E}_{\mathrm{control},t}} \tag{3.1}$$

The interpretation of the hazards ratio is then the risk of event in the group in the numerator (treated) as compared to the risk of event in the group in the denominator (control, placebo, the other treatment group).

Finally a Cox proportional hazard regression model compares the hazards of the two treatment groups (treated and control or treatment and control groups) and allows several variables to be taken into account. In addition, there are several important assumptions for appropriate use of the Cox proportional hazards regression model including.

1. Independence of survival times between district individuals in the sample.
2. A multiplicative relationship between the predictors and hazard.
3. A constant hazard ratio over time.

In literature, there are many definitions about Cox regression model.

> The Cox model is based on a modelling approach to the analysis of survival data. The purpose of the model is to simultaneously explore the effects of several variables on survival. The Cox model is a well-recognised statistical technique for analysing survival data. When it is used to analyse the survival of patients in a clinical trial, the model allows us to isolate the effects of treatment from the effects of other variables. The model can also be used. a priori, if it is known that there are other variables besides treatment that influence patient survival and these variables cannot be easily controlled in a clinical trial. Using the model may improve the estimate of treatment effect by narrowing the confidence interval. Survival times now often refer to the development of a particular symptom or to relapse after remission of a disease, as well as to the time to death (Walters. 2009).

> The log rank test is used to test whether there is a difference between the survival times of different groups but it does not allow other explanatory variables to be taken into account. Cox's proportional hazards model is analogous to a multiple regression model and enables the difference between survival times of particular groups of patients to be tested while allowing for other factors. In this model, the response (dependent) variable is the “hazard”. The hazard is the probability of dying (or experiencing the event in question)

given that patients have survived up to a given point in time or the risk for death at that moment (Bewick, Cheek & Ball, 2005).

The cox proportional hazard regression model is also called a semi-parametric model because there are no assumptions about the shape of the baseline hazard function.

3.1.1 The Formula for the Cox Proportional Hazards Model

The Cox proportional hazard regression model is written as follows:

$$h(t,X) = h_0(t)\exp\left(\sum_{i=1}^{p}\beta_i X_i\right) \tag{3.2}$$

Where $h(t,X)$ is the expected hazard at time t and where $X = (X_1, X_2, ..., X_p)$ are the explanatory/predictor/risk factor variables. The X's are time-independent for example smoking status, gender, weight.

Notice that the predicted hazard is the product of the baseline hazard $h_0(t)$ and the exponential function of the linear combination of the independent variables. Thus, the independent variables have a multiplicative effect on the predicted hazard.

The model can also be shown as:

$$\frac{h(t,X)}{h_0(t)} = \exp\left(\sum_{i=1}^{p}\beta_i X_i\right) \tag{3.3}$$

We can take the natural logarithm (ln) of each side of the Cox proportional hazards regression model, to produce the following which relates the log of the relative hazard to a linear function of the predictors.

$$\ln\left\{\frac{h(t,X)}{h_0(t)}\right\} = \left(\sum_{i=1}^{p}\beta_i X_i\right) = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \quad (3.4)$$

The estimated coefficients in the Cox proportional hazards regression model. β_1, for example, represent the change in the expected log of the hazard ratio relative to a one unit change in X_1, holding all other predictors constant.

The antilog of an estimated regression coefficient $\exp(\beta_i)$, produces a hazard ratio if the hazard ratio for a predictor is close to 1 then that predictor does not affect survival. If the hazard ratio is less than 1, then the predictor is protective (i.e. associated with improved survival) and if the hazard ratio is greater than 1, then the predictor is associated with increased risk (or decreased survival).

3.1.2 Components of model

The components of the Cox proportional hazards model are the product of two quantities.

- $h_0(t)$ is called the baseline hazard. Involves t but not X. The baseline hazard function describes the risk for individuals with independent variables $X = (X_1, X_2, ..., X_p)$ equal to zero but otherwise it is unspecified.
- Exponential of the sum of β_i and X_i. Notice that this term does not involve t.

The fact that the baseline hazard is an unspecified function and semi-parametric nature of the model are reasons for Cox model's popularity.

3.2 Proportional Hazards

Hazard functions should be strictly parallel. This ratio is constant over time.

$$\text{Hazard Ratio}_{i,j} = \frac{h_i(t)}{h_j(t)} = \frac{\cancel{\lambda_0(t)}e^{\beta_1 x_{i1}+\ldots+\beta_k x_{ik}}}{\cancel{\lambda_0(t)}e^{\beta_1 x_{j1}+\ldots+\beta_k x_{jk}}} = e^{\beta_1(x_{i1}-x_{j1})+\ldots+\beta_1(x_{ik}-x_{jk})} \quad (3.5)$$

$h_i(t)$ Hazard for person i (eg a smoker)

$h_j(t)$ Hazard for person j (eg a non-smoker)

According to the formula, hazard ratio for smoker vs. non-smoker is constant over time.

3.3 Evaluating the Proportional Hazards Assumption

The Cox regression model should be checked to make sure that it satisfies a proportional hazard assumption so we can examine the proportional hazard assumption with

- Log(-log) plots,
- Schoenfeld residual plots,
- Expected vs Observed plots,
- Arjas plots and
- Using time dependent explanotary variables methods.

3.3.1 Log-(Log) Plot

This plot displays the log-minus-log of the survival function against the survival time. The log(-log) plot is one way to assess graphically whether the assumption of proportional hazards is acceptable. For the assumption to hold then the log-log plot should show the separate lines as approximately parallel to each other and the distance should be constant over time.

3.3.2 Schoenfeld Residuals Plot

Schoenfeld residuals are named after David Schoenfeld who wrote about them in 1982. These are also known as partial residuals which should be independent of time. Schoenfeld residuals are calculated from the difference between observed and expected covariates. Schoenfeld residuals against the time are plotted and this plot helps check the proportional hazards assumption. If proportional hazard assumption holds, Schoenfeld residuals against time plot should be randomly scattered around zero. If Schoenfeld residuals against time show a non-random pattern, this is evidence of violation of the proportional hazards assumption.

Suppose n individuals are indexed by $i = 1, 2, ..., n$ and that each has a p-dimensional vector of covariates $X_i = \left(X_{i1}, ..., X_{ip}\right)'$. The proportional hazards regression model specifies that the hazard function of the i^{th} individual is $h_i(t) = h_0(t)\exp\left(\beta' X_i\right)$ where β is a vector of p parameters and $h_0(t)$ is an arbitrary function.

Let D be the indices of the individuals who failed and let R_i be the indices of those under observation when the i^{th} individual fails. Using partial likelihood arguments one can estimate β by maximizing the likelihood function of the following model (Cox, 1975): for $i \in$ D. an index $m \in R_i$ is selected with probability

$$\exp\left(\beta' X_m\right) \Big/ \sum_{k \in R_i} \exp\left(\beta' X_k\right) \tag{3.6}$$

In this model X_i is a random variable with

$$E\left(X_{ij} \middle| R_i\right) = \sum_{k \in R_i} X_{kj} \exp\left(\beta' X_k\right) \Big/ \sum_{k \in R_i} \exp\left(\beta' X_k\right) \tag{3.7}$$

Difference between observed and expected covariates is what defines the partial residual.

$$\hat{r}_{ik} = X_{ik} - E\left(X_{ij} \middle| R_i\right) \tag{3.8}$$

3.3.3 Expected versus Observed Plots

The expected and the observed plots are obtained from different calculations. Kaplan-Meier curves are used to obtain the "observed" plots and Cox proportional hazards model is used to obtain the "expected" plots. After obtaining these plots, the sets of plots are compared. If they are close, then the proportional hazards assumption holds; if not, the assumption is violated.

3.3.4 Arjas Plots

The estimated cumulative hazard against the number of failures are plotted and this plot is called Arjas plot and helps check the proportional hazards assumption checking. If the Cox regression model is correct, it is expected that the slope of the graph is close to one and the plot is approximately linear.

3.3.5 Using Time Dependent Variable

After the final model of significant explanatory variables is created, it is needed to examine the proportional hazards assumption. A variable*time interaction term is added to the model and is tested for significance. If the interaction term is found to be insignificant, then the proportional hazards assumption is acceptable. The extended cox model is defined as follows:

$$h\left(t, x\right) = h_0\left(t\right)\exp\left[\beta x + \delta\left(x \times g\left(t\right)\right)\right] \tag{3.9}$$

There are different options for g (t) function which are $g\left(t\right)$, $g\left(t\right) = \log t$, etc

Null hypothesis is $H_0 : \delta = 0$

Alternative hypothesis is $H_1 : \delta \neq 0$

If the null hypothesis is not rejected, the model is reduced to a proportional hazard model with a single variable and the interaction term can be dropped from the model. So the proportional hazards assumption holds.

CHAPTER FOUR
APPLICATION

4.1 Introduction

The purpose of this thesis is to consider survival analysis based on the proportional hazards (PH) model. In addition, this thesis examines survival methods (survival analysis, Cox regression model) in the real data set taken from İzmir ESHOT (Electricity, Water, Water Gas, Bus and Trolleybus) headquarter, containing drivers accidents between 2012-2014 years. The aim of the ESHOT headquarters is to provide passenger satisfaction and economic, safe, comfortable, continuous and reliable public transport service.

The event of interest is the first accident that the drivers experience and the survival time is the time passed until the first accident. The outcome variable is the survival time.

The explanatory (independent) variables are district where the accident happens, the traffic jam, the capacity of bus, the number of passengers, the number of stations, the type of bus, the age of the driver and the experience of driver.

Drivers who had no accidents during the specified date were identified as censored and coded as zero (0) in SPSS 16.0 version. The number of total drivers is 1636. The number of censored data, i.e. the number of drivers who don't have any accidents, is 298 (18.2%). Since the survival times of censored variables are unknown, numbers are generated for this data in the range of known survival times from the uniform distribution and these survival times are used in this study. As well as the last route information used by drivers in other variables for censored data see, Vermeylen (2005).

The quantitative and the qualitative explanatory variables are investigated by descriptive statistics and frequency tables. Descriptive statistics for survival time (in days) are shown in Table 4.1.

Table 4.1 Descriptive statistics for survival time recorded by day

	Minimum	Maximum	Median	Mean	Standard Deviation
Survival Time	2	729	268	295.71	214.42

The minimum and maximum values of the outcome variable are 2 and 729, respectively. The median and mean values of the outcome variable are 268 and 295.71, respectively. The standard deviation value of the outcome variable is 214.42. The histogram of survival time is shown in Figure 4.1

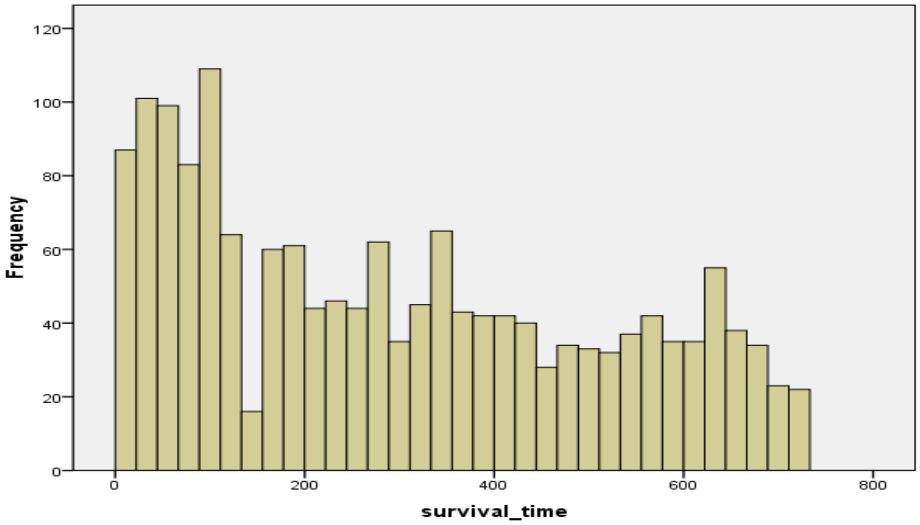

Figure 4.1 Survival time histogram

When the Figure 4.1 is examined, it is seen that survival time does not proper the specific known distributions as (normal. uniform and exponential).

The Kolmogorov-Smirnov test is used to decide if the sample comes from a population with a specific known distribution (normal, uniform or exponential). The test results for survival time are shown in Table 4.2.

Table 4.2 Kolmogorov-Smirnov test results for survival time

	Kolmogorov-Smirnov
	p-value
normal distribution test for survival time	0.000
uniform distribution test for survival time	0.000
exponential distribution test for survival time	0.000

The hypotheses for the Kolmogorov-Smirnov test are as follows;

H_0 : The data follow a specified known distribution (normal, uniform or exponential)

H_1: The data don't follow a specified known distribution (normal, uniform or exponential)

According to the Table 4.2, all the null hypotheses are rejected ($p < 0.05$). In this situation, the survival time distribution does not follow a specified known distribution (normal, uniform or exponential).

After the distribution of the survival time is examined, the explanatory variables are investigated by frequency tables and graphs.

4.1.1 District

The variable "district" is categorized by ESHOT headquarters as "1", "2" and "3". We have not been given more information about this variable. The following table and figure show the frequencies of the data according to the "district" levels.

Table 4.3 Frequencies for all district variables

Category	Frequency	Percent %
1	348	21.3
2	657	40.2
3	631	38.6
Total	1636	100.0

According to the Table 4.3, 348 of 1636 data is in the "district 1". 647 of 1636 data is in the "district 2" and 631 of 1636 data is in the "district 3". The bar chart of the variable "district" is shown as:

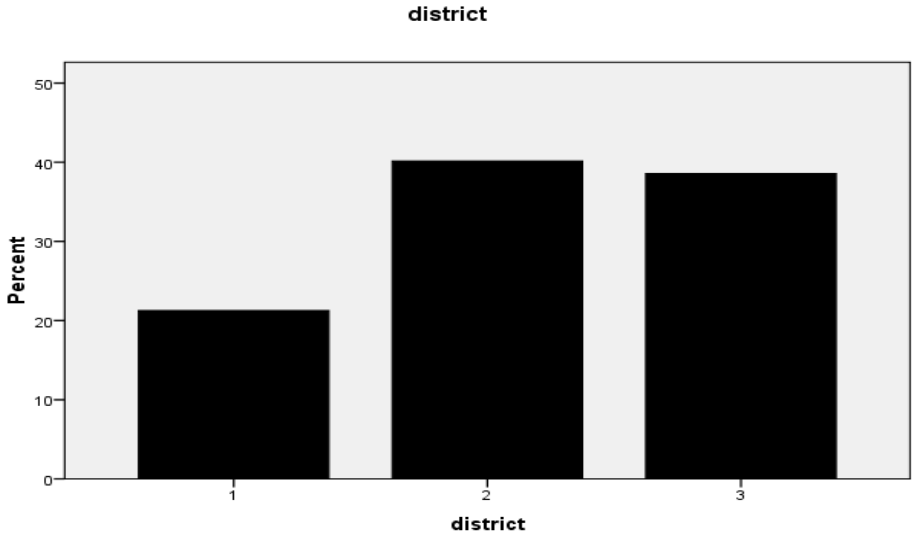

Figure 4.2 Bar charts with percentages of district variable

According to the Table 4.3 and the Figure 4.2, the lowest frequency in the period until the accident happen is in the first district with 21.3 percent and until the time that accident happens, the highest frequency value is found for “district 2” with 40.2 percent.

4.1.2 Traffic Jam

The variable “traffic jam” is categorized by ESHOT headquarters into “light”, “moderate” and “heavy”. We have not been given more information about this variable. The following table and figure show the frequencies of the data according to the “traffic jam”.

Table 4.4 Frequencies for all traffic jam variables

Category	Frequency	Percent %
Light	376	23.0
Moderate	847	51.8
Heavy	413	25.2
Total	1636	100.0

According to the Table 4.4, 376 of 1636 data is in the light traffic jam, 847 of 1636 data is in the moderate traffic jam and 413 of 1636 data is in the heavy traffic jam. The bar chart of the variable “traffic jam” is shown as:

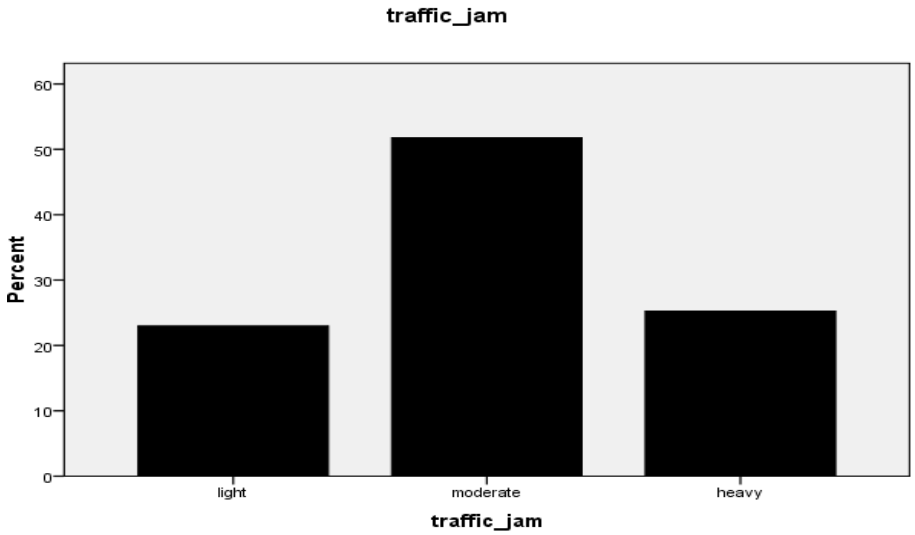

Figure 4.3 Bar charts with percentages of traffic jam variable

According to the Table 4.4 and the Figure 4.3, the lowest frequency in the period until the accident happen is in the “light traffic” jam with 23.0 percent and until the time that accident happens, the highest frequency value is found for “moderate traffic jam” with 51.8 percent.

4.1.3 Bus Capacity

The variable “bus capacity” is categorized by ESHOT headquarters as “empty”, ”normal” and “full”. We have not been given more information about this variable. The following table and figure show the frequencies of the data according to the “bus capacity”.

Table 4.5 Frequencies for all bus capacity variables

Category	Frequency	Percent %
Empty	632	38.6
Normal	737	45.0
Full	267	16.3
Total	1636	100.0

According to the Table 4.5, 632 of 1636 data is in the empty bus capacity, 737 of 1636 data is in the normal bus capacity and 267 of 1636 data is in the full bus capacity. The bar chart of the variable “bus capacity” is shown as:

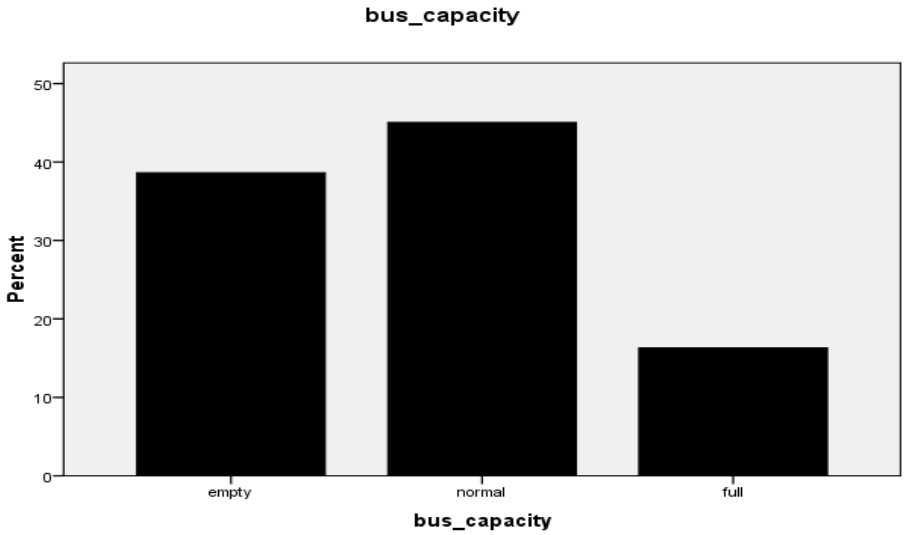

Figure 4.4 Bar charts with percentages of bus capacity variable

According to the Table 4.5 and the Figure 4.4, the lowest frequency in the period until the accident happen is in the "full bus capacity" with 16.3 percent and until the time that accident happens, the highest frequency value is found for "normal bus capacity" with 45.0 percent.

4.1.4 Station Number

This variable was continuous values and gave information on the station number used in the study. In order to be used in the study, this variable was divided into two parts. The first part was named "57 and less than 57 station number" and second part was named "greater than 57 station number". The following table and figure show the frequencies of the data according to the "station number".

Table 4.6 Frequencies for all station number variables

Category	Frequency	Percent %
57 and less than 57	838	51.2
Greater than 57	798	48.8
Total	1636	100.0

According to the Table 4.6, 838 of 1636 data is in the "57 and less than 57 station number". 798 of 1636 data is in "greater than 57 station number". The bar chart of the variable "station number" is shown as:

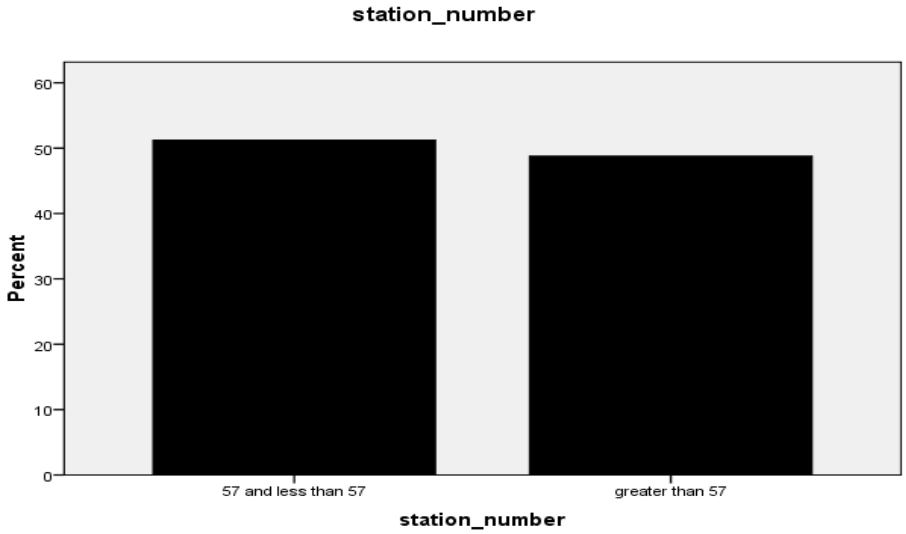

Figure 4.5 Bar charts with percentages of station number variable

According to the Table 4.6 and the Figure 4.5, the lowest frequency in the period until the accident happen is in “greater than 57 station number” with 48.8 percent and until the time that accident happens the highest frequency value is found for “57 and less than 57 station number” with 51.8 percent.

4.1.5 Bus Type

The variable “bus type” is categorized by ESHOT headquarters as “big”, and “others”. We have not been given more information about this variable. The following table and figure show the frequencies of the data according to the bus type.

Table 4.7 Frequencies for all bus type variables

Category	Frequency	Percent %
Big	1216	74.3
Others	420	25.7
Total	1636	100.0

According to the Table 4.7, 1216 of 1636 data is in the “big bus type” and 420 of 1636 data is in the “other bus type”. The bar chart of the variable “bus type” is shown as:

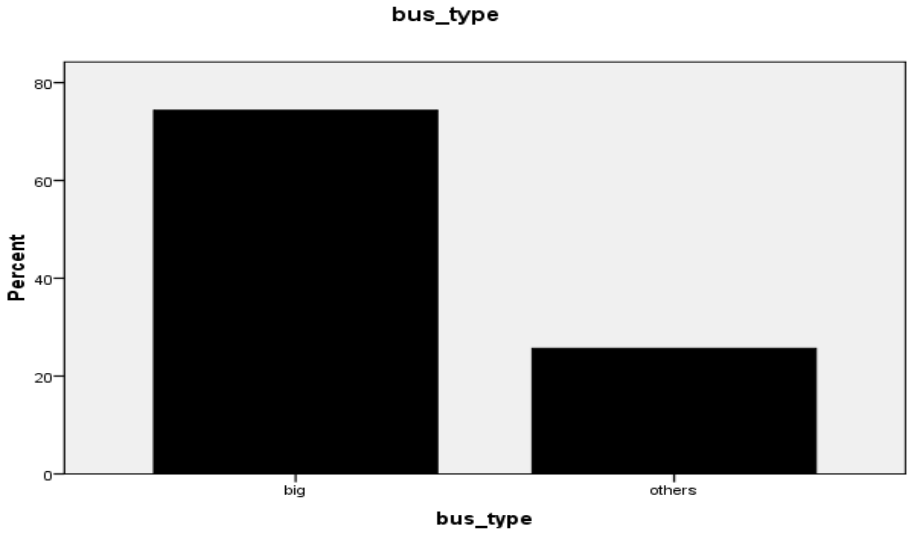

Figure 4.6 Bar charts with percentages of bus type variable

According to the Table 4.7 and the Figure 4.6, the lowest frequency in the period until the accident happen is “other bus type” with 25.7 percent and until the time that accident happens, the highest frequency found for “big bus type” with 74.3 percent.

4.1.6 Passenger Number

The variable “passenger number” is categorized by ESHOT headquarters as “low”, and “high”. We have not been given more information about this variable. The following table and figure show the frequencies of the data according to the “passenger number”.

Table 4.8 Frequencies for all passenger number variables

Category	Frequency	Percent %
Low	570	34.8
High	1066	65.2
Total	1636	100.0

According to the Table 4.8, 570 of 1636 data is in the “low passenger number” and 1066 of 1636 data is in the “high passenger number”. The bar chart of the variable “passenger number” is shown as:

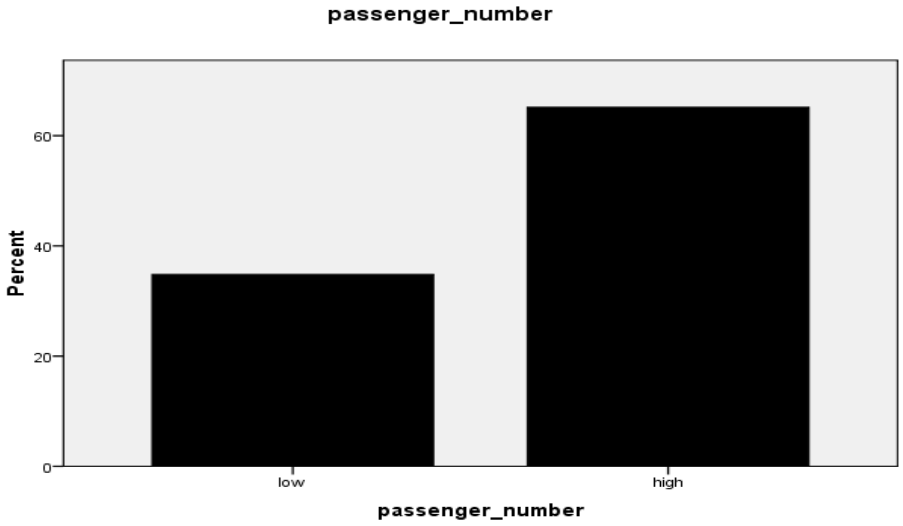

Figure 4.7 Bar charts with percentages of passenger number variable

According to the Table 4.8 and the Figure 4.7, the lowest frequency in the period until the accident happen is “low passenger number” with 34.8 percent and until the time that the accident happens, the highest frequency is found for “high passenger number” with 65.2 percent.

4.1.7 Driver Experience

The driver experience variable was continuous and this variable provides information on the driver experience used in the study but in order to be used in the study, this variable was divided into two parts with 35 months median value. First part was named “35 and less than 35 months driver experience” and second part was named “greater than 35 months driver experience”. The following table and figure show the frequencies of the data according to the “driver experience”.

Table 4.9 Frequencies for all driver experience variables

Category	Frequency	Percent %
35 and less than 35 months	821	50.2
Greater than 35 months	815	49.8
Total	1636	100.0

According to Table 4.9. 821 of 1636 data is in “35 and less than 35 months driver experience” and 815 of 1636 data is in the “greater than 35 months driver experience”. The bar chart of the variable “driver experience” is shown as:

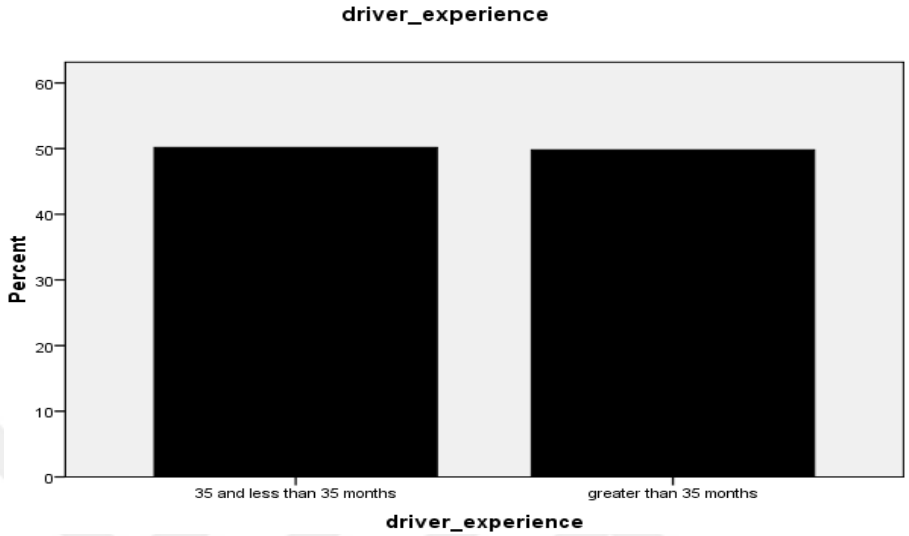

Figure 4.8 Bar charts with percentages of driver experience variable

According to the Table 4.9 and the Figure 4.8, until the accident happen driver experience variable have very close values but “35 and less than 35 months driver experiences” are higher frequencies than “greater than 35 months driver experience”.

4.1.8 Driver Age

The driver age variable was continuous and this variable provides information on the driver age used in the study but in order to be used in the study, this variable was divided into two parts with 33 median values. First part named “33 and less than years old driver age” and second part named “greater than 33 years old driver age”. The following table and figure show the numbers of the data according to the “driver age”.

Table 4.10 Frequencies for all driver age variables

Category	Frequency	Percent %
33 and less than 33 years old	699	42.7
Greater than 33 years old	937	57.3
Total	1636	100.0

According to the Table 4.10, 699 of 1636 data is in “33 and less than 33 years old driver age” and 937 of 1636 data is in the “greater than 33 years old driver age”. The bar chart of the variable “driver age” is shown as:

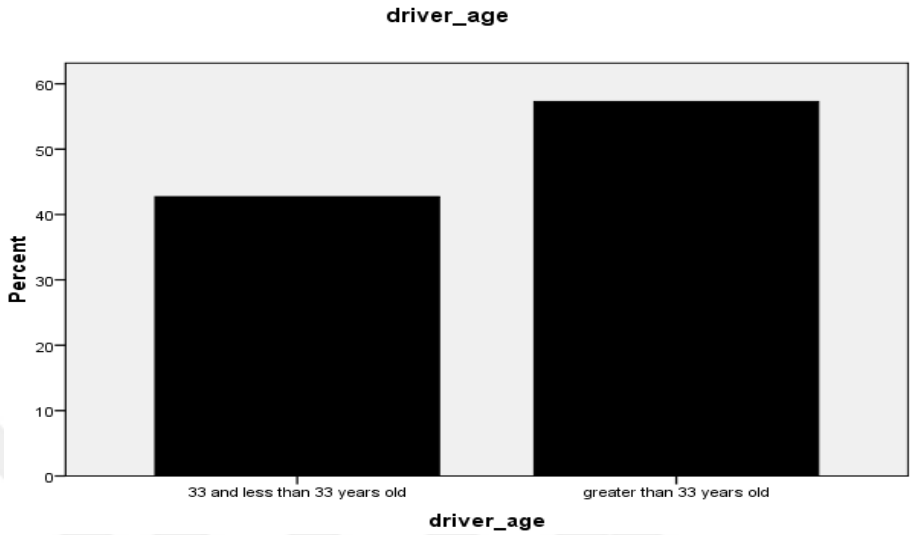

Figure 4.9 Bar charts with percentages of driver age variable

According to the Table 4.10 and the Figure 4.9, the lowest frequency in the period until the accident happen is “33 and less than 33 years old driver age” with 42.7 percent and until the time that accident happens, the highest frequency is found for “greater than 33 years old driver age” with 57.3 percent.

4.2 Kaplan-Meier Results

Median survival is a frequently method used reported in all survival studies. The median survival is calculated as the smallest survival time for which the survival function is less than or equal to 0.5.

In this thesis the Kaplan Meier analysis is used to estimate the survival curves for independent variables. These independent variables are “district”, “traffic jam”, “bus capacity”, “station number”, “bus type”, “passenger number”, “driver experience” and “driver age”.

4.2.1 District

The variable “district” is categorized as “1”, “2” and “3”. The following Table 4.11 shows the frequencies of the uncensored and censored data according to the “district” variable.

Table 4.11 Summary cases for all district variables

District	Total n	Uncensored		Censored	
		n	Percent %	n	Percent %
1	348	303	87.1	45	12.9
2	657	525	79.9	132	20.1
3	631	510	80.8	121	19.2
Overall	1636	1338	81.8	298	18.2

The rate of uncensored data in the first district is 87.1 percent, the rate of uncensored data in the second district is 79.9 percent, the rate of uncensored data in the last district is 80.8 percent and finally the rate of uncensored data in the overall district is 81.8 percent. Additionally, the rate of censored data in the first district is 12.9 percent, the rate of censored data in the second district is 20.1 percent, the rate of censored data in the last district is 19.2 percent and the rate of censored data in the overall district is 18.2 percent. Median survival times for all district levels are shown in Table 4.12.

Table 4.12 Medians for survival time for all district variables

District	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	286	22.29	242.29	329.70
2	339	19.06	301.63	376.36
3	282	15.44	251.72	312.27
Overall	312	10.20	292.00	331.99

The estimated median survival time is 286 days for category “district 1” and 339 days for category “district 2” and 282 days for category “district 3”. In addition, the best median survival time for “district” variable is in the category “district 2”. The survival curve is shown in Figure 4.10 for “district” variable.

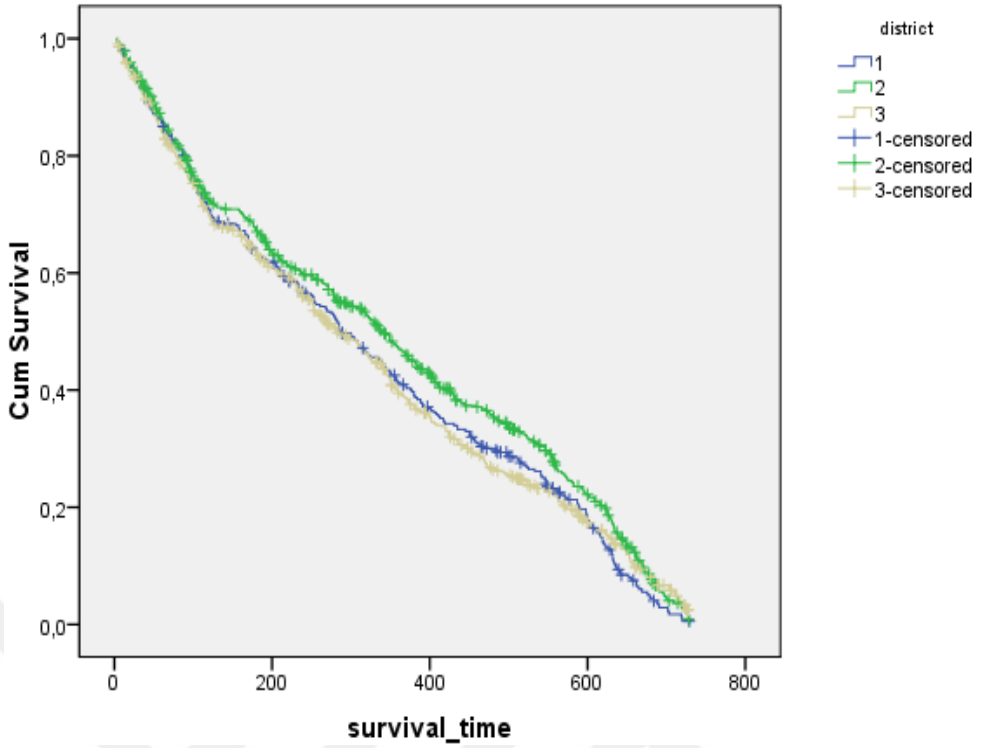

Figure 4.10 Survival plot for district variable

The little pluses show when someone was censored in the plot. As the figure shows the survival curves don't differ, but this is insufficient to conclude that the survival times are the same in the three groups of district levels. For this reason the log-rank test is used to test the null hypothesis that there is no difference between the populations in the probability of an event at any time point. The analysis is based on the times of event. Table 4.13 is shown log-rank results for "district" variable.

Table 4.13 Test of equality of survival distributions for the different levels of district variable

	Chi-Square test value	p-value
Log Rank (Mantel-Cox)	5.504	0.064

H_0 : The survival distributions are the same in the three groups of district levels

H_1: The survival distributions are different in the three groups of district levels

The test statistic for the log-rank test is 5.504. The test statistic follows a chi-square distribution and so we find the critical value in the table of critical values for the χ^2 distribution for $df = k-1 = 3-1 = 2$ and $\alpha = 0.05$ (k is the number of district

levels). The critical value is 5.991 and H_0 hypothesis is not rejected ($\chi^2 < 5.991$). We don't have statistically significant evidence (p=0.064>0.05) at the level of $\alpha = 0.05$ to show that the survival distributions (times) are different in the three groups of district levels. In another words, there is no difference among groups of district levels for survival time.

4.2.2 Traffic Jam

The variable "traffic jam" is categorized as "light", "moderate" and "heavy". The following Table 4.14 shows the frequencies of the uncensored and censored data according to the "traffic jam" variable.

Table 4.14 Summary cases for all traffic jam variables

		Uncensored		Censored	
Traffic jam	Total n	n	Percent %	n	Percent %
Light	376	284	75.5	92	24.5
Moderate	847	708	83.6	139	16.4
Heavy	413	346	83.8	67	16.2
Overall	1636	1338	81.8	298	18.2

The rate of uncensored data in the "light traffic jam" is 75.5 percent, the rate of uncensored data in the "moderate traffic jam" is 83.6 percent. the rate of uncensored data in the "heavy traffic jam" is 83.8 percent and finally the rate of uncensored data in the "overall traffic jam" is 81.8 percent. Additionally the rate of censored data in the "light traffic jam" is 24.5 percent, the rate of censored data in the "moderate traffic jam" is 16.4 percent, the rate of censored data in the "heavy traffic jam" is 16.2 percent and the rate of censored data in the "overall traffic jam is 18.2 percent. Median survival times for all traffic jam levels are shown in Table 4.15.

Table 4.15 Medians for survival time for all traffic jam variable

			95% Confidence Interval	
Traffic jam	Estimate	Std. Error	Lower Bound	Upper Bound
Light	326	19.38	288.01	363.98
Moderate	291	15.23	261.13	320.86
Heavy	324	22.18	280.52	367.47
Overall	312	10.20	292.00	331.99

The estimated median survival time is 326 days for category “light traffic jam”, 291 days for category “moderate traffic jam” and 324 days for category “heavy traffic jam”. In addition, the best median survival time for “traffic jam” variable is in the category “light traffic jam”. The survival curve is shown in Figure 4.11 for “traffic jam” variable.

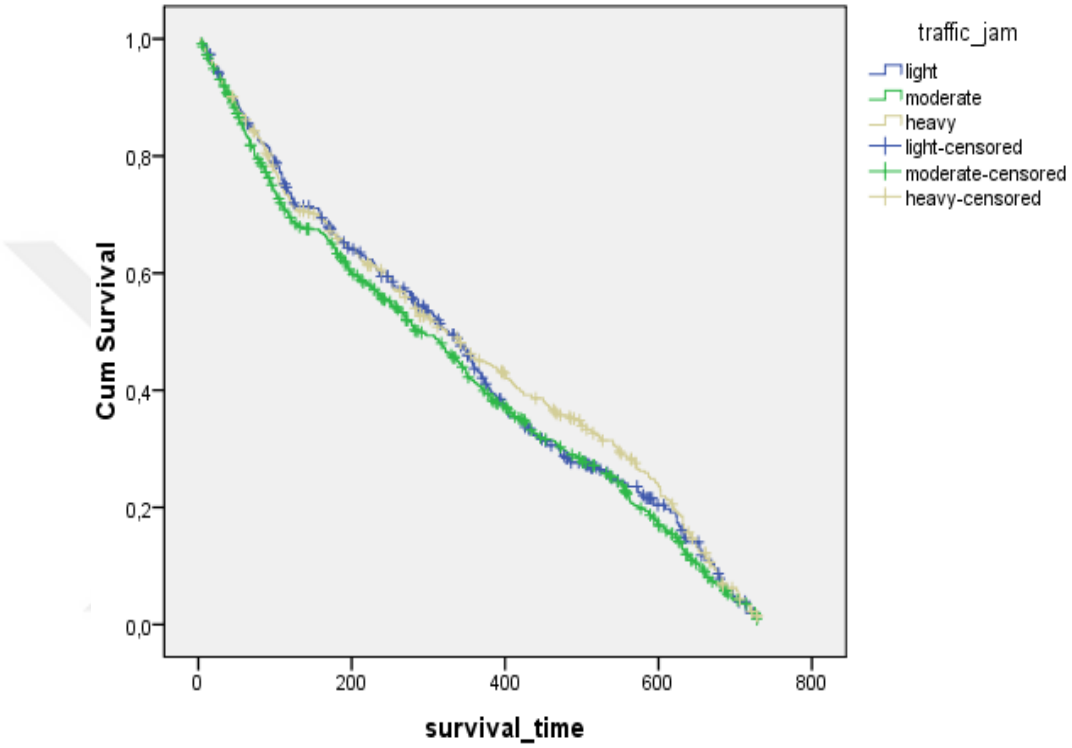

Figure 4.11 Survival plot for traffic jam variable

The little pluses show when someone was censored in the plot. As the figure shows the survival curves don’t differ. However conceptually, we must show the compare the log rank test results for the different levels of “traffic jam” variable. Table 4.16 is shown log-rank results for “traffic jam” variable.

Table 4.16 Test of equality of survival distributions for the different levels of traffic jam variable

	Chi-Square test value	p-value
Log Rank (Mantel-Cox)	3.82	0.148

H_0 : The survival distributions are the same in the three groups of traffic jam levels

H_1: The survival distributions are different in the three groups of traffic jam levels

The test statistic for the log-rank test is 3.82. The test statistic follows a chi-square distribution and so we find the critical value in the table of critical values for the χ^2 distribution for $df = k-1 = 3-1 = 2$ and $\alpha = 0.05$ (k is the number of traffic jam levels). The critical value is 5.991 and H_0 hypothesis is not rejected ($\chi^2 < 5.991$). We don't have statistically significant evidence (p=0.148>0.05) at the level of $\alpha = 0.05$ to show that the survival distributions (times) are different in the three groups of traffic jam levels. In another words, there is no difference among groups of traffic jam levels for survival time. We can also see this outcome from the survival plot in Figure 4.11.

4.2.3 Bus Capacity

The variable "bus capacity" is categorized as "empty", "normal" and "full". The following Table 4.17 shows the frequencies of the uncensored and censored data according to the "bus capacity" variable.

Table 4.17 Summary cases for all bus capacity variables

		Uncensored		Censored	
Bu capacity	Total n	n	Percent %	n	Percent %
Empty	632	490	77.5	142	22.5
Normal	737	642	87.1	95	12.9
Full	267	206	77.2	61	22.8
Overall	1636	1338	81.8	298	18.2

The rate of uncensored data in the "empty bus capacity" is 77.5 percent, the rate of uncensored data in the "normal bus capacity" is 87.1 percent, the rate of uncensored data in the "full bus capacity" is 77.2 percent and finally the rate of uncensored data in the "overall bus capacity" is 81.8 percent. Additionally the rate of censored data in the "empty bus capacity" is 22.5 percent, the rate of censored data in the "normal bus capacity" is 12.9 percent, the rate of censored data in the "full bus capacity" 22.8 percent and the rate of censored data in the "overall bus capacity" is

18.2 percent. Median survival times for all bus capacity levels are shown in Table 4.18.

Table 4.18 Medians for survival time for all bus capacity variables

Bus capacity	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Empty	309.000	315.000	16.593	282.478
Normal	315.000	316.000	13.973	288.613
Full	312.000	282.000	25.916	231.204
Overall	312.000	10.202	292.004	331.996

The estimated median survival time is 309 days for category "empty bus capacity", 315 days for category "normal bus capacity" and 312 days for category "full bus capacity". In addition, the best median survival time for "bus capacity" variable is in the category "normal bus capacity". The survival curve is shown in Figure 4.12 for "bus capacity" variable.

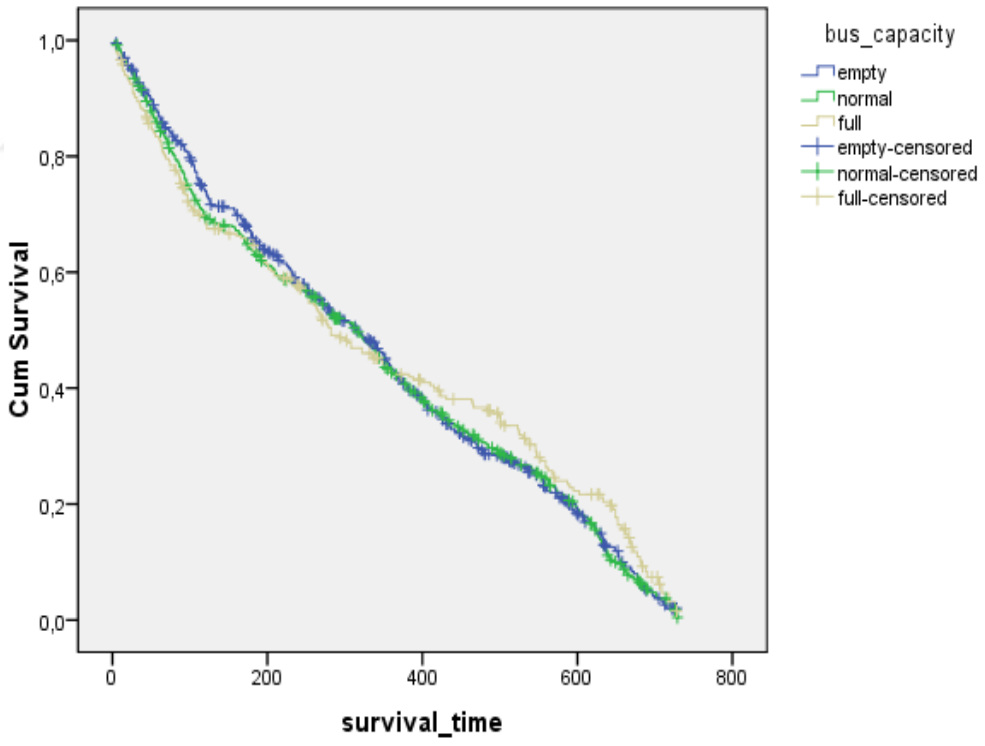

Figure 4.12 Survival plot for bus capacity variable

The little pluses show when someone was censored in the plot. As the figure shows the survival curves don't differ. However conceptually, we must show the

compare the log rank test results for the different levels of "bus capacity" variable. Table 4.19 is shown log-rank results for "bus capacity" variable.

Table 4.19 Test of equality of survival distributions for the different levels of bus capacity variable

	Chi-Square test value	p-value
Log Rank (Mantel-Cox)	1.439	0.487

H_0 : The survival distributions are the same in the three groups of bus capacity levels

H_1: The survival distributions are different in the three groups of bus capacity levels

The test statistic for the log-rank test is 1.439. The test statistic follows a chi-square distribution and so we find the critical value in the table of critical values for the χ^2 distribution for $df = k-1=3-1=2$ and $\alpha = 0.05$ (k is the number of bus capacity levels). The critical value is 5.991 and H_0 hypothesis is not rejected $\left(\chi^2 < 5.991\right)$. We don't have statistically significant evidence $\left(p = 0.487 > 0.05\right)$ at the level of $\alpha = 0.05$ to show that the survival distributions (times) are different in the three groups of bus capacity levels. In another words, there is no difference among groups of bus capacity levels for survival time. This is confirmed by looking at the plot of the survival curves Figure 4.12, which both drop down initially at the same rate.

4.2.4 Station Number

This station number variable is categorized as "57 and less than 57 station number" and "greater than 57 station number". The following Table 4.20 shows the frequencies of the uncensored and censored data according to the "station number" variable.

Table 4.20 Summary cases for all station number variables

Station number	Total n	Uncensored		Censored	
		n	Percent %	n	Percent %
57 and less than 57	838	713	85.1	125	14.9
Greater than 57	798	625	78.3	173	21.7
Overall	1636	1338	81.8	298	18.2

The rate of uncensored data in the "57 and less than 57 station number" is 85.1 percent, the rate of uncensored data in the "greater than 57 station number" is 78.3 percent and finally the rate of uncensored data in the "overall station number" is 81.8 percent. Additionally the rate of censored data in the "57 and less than 57 station number" is 14.9 percent. the rate of censored data in the greater than "57 station number" is 21.7 percent and the rate of censored data in the "overall station number" is 18.2 percent. Median survival times for all station number levels are shown in Table 4.21.

Table 4.21 Medians for survival time for all station number variables

Station number	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
57 and less than 57	309.000	15.100	279.405	338.595
Greater than 57	315.000	14.052	287.457	342.543
Overall	312.000	10.202	292.004	331.996

The estimated median survival time is 309 days for category "57 and less than 57 station number" and 315 days for category "greater than 57 station numbers". In addition, the best median survival time for station number variable is in the category "greater than 57 station number". The survival curve is shown in Figure 4.13 for station number variable.

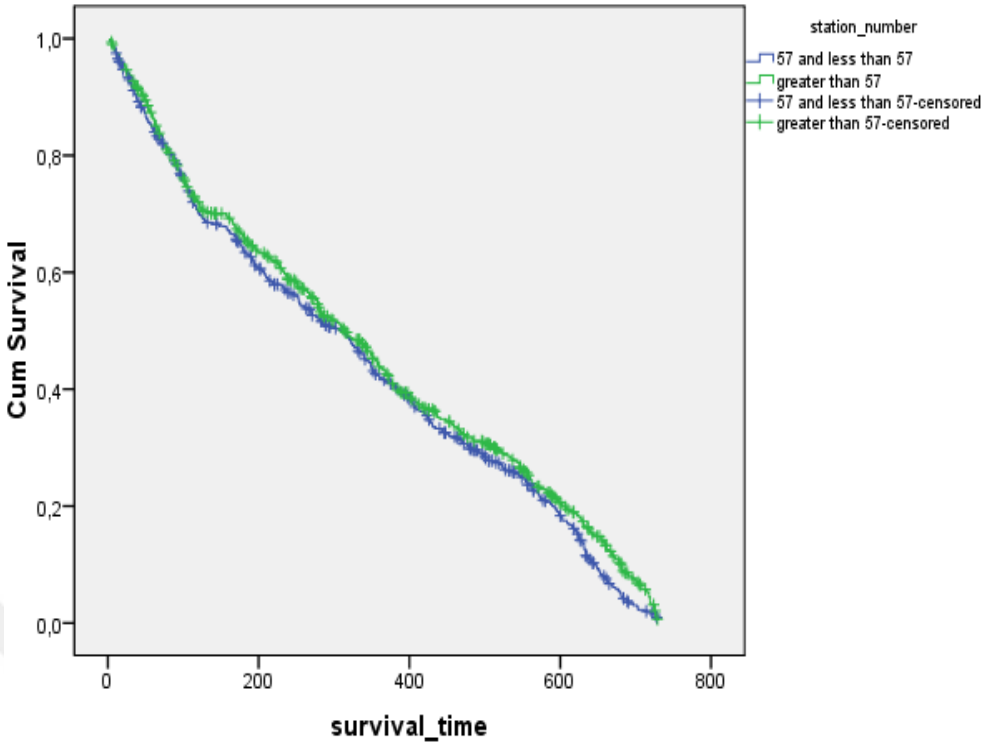

Figure 4.13 Survival plot for station number variable

Even though the Figure 4.13 survival plots for "station numbers" appear to be very close we reject the null hypothesis for the log-rank results. Table 4.22 is shown log-rank results for "station number" variable.

Table 4.22 Test of equality of survival distributions for the different levels of station number variable

	Chi-Square test value	p-value
Log Rank (Mantel-Cox)	3.860	0.049

H_0 : The survival distributions are the same in the three groups of station number levels

H_1: The survival distributions are different in the two groups of station number levels

χ^2 table value is 3.84 is and log- rank chi-square value is 3.86 so we fail to the null hypothesis and the p-value of the log-rank test is 0.049. We reject the null hypothesis and conclude that there is a difference in "station number" levels until the accident happen.

4.2.5 Bus Type

The variable “bus type” is categorized as “big” and “others”. The following Table 4.23 shows the frequencies of the uncensored and censored data according to the “bus type” variable.

Table 4.23 Summary cases for all bus type variables

Bus type	Total n	Uncensored		Censored	
		n	Percent %	n	Percent %
Big	1216	1034	85.0	182	15.0
Others	420	304	72.4	116	27.6
Overall	1636	1338	88.2	298	18.2

The rate of uncensored data in the “big bus type” is 85.0 percent, the rate of uncensored data in the “others bus type” is 72.4 percent and finally the rate of uncensored data in the “overall bus type” is 88.2 percent. Additionally the rate of censored data in the “big bus type” is 15.0 percent, the rate of censored data in the “others bus type” is 27.6 percent and the rate of censored data in the “overall bus type” is 18.2 percent. Median survival times for all bus type levels are shown in Table 4.24.

Table 4.24 Medians for survival time for all bus type variables

Bus type	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Big	195.000	8.669	178.008	211.992
Others	572.000	8.426	555.485	588.515
Overall	312.000	10.202	292.004	331.996

The estimated median survival time is 195 days for category “big bus type” and 572 days for category “other bus type”. In addition, the best median survival time for “bus type” variable is in the category “others bus type”. The survival curve is shown in Figure 4.14 for “bus type” variable.

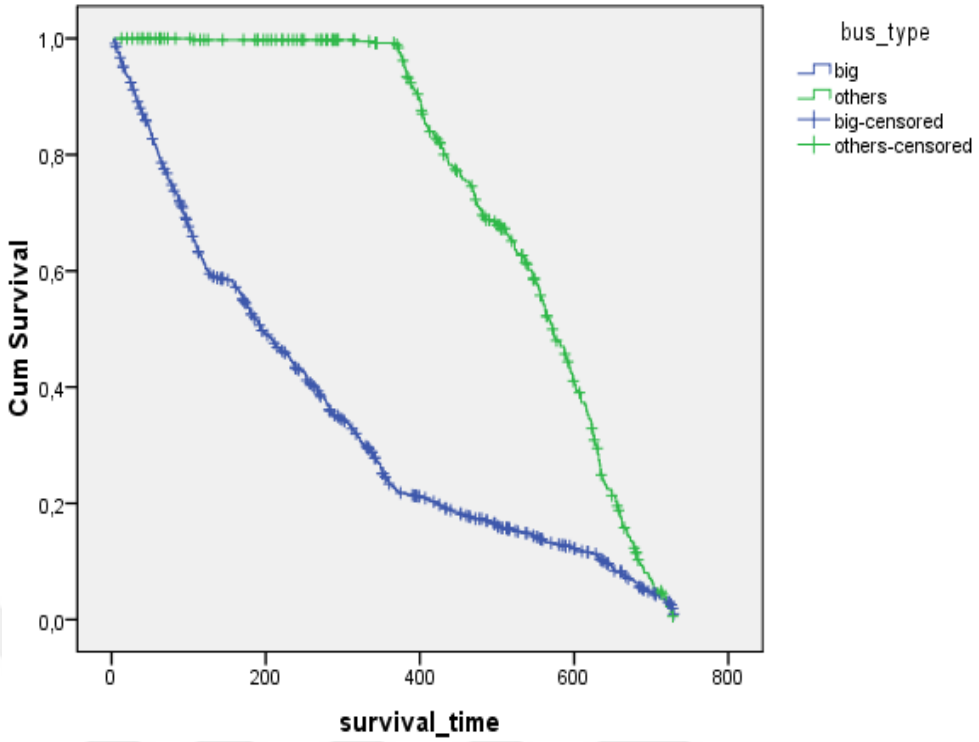

Figure 4.14 Survival plot for bus type variable

From the Figure 4.14, it can be seen that the "big" and "other" type of buses are obviously greatly different survival times. Table 4.25 is shown log-rank results for "bus type" variable.

Table 4.25 Test of equality of survival distributions for the different levels of bus type variable

	Chi-Square test value	p-value
Log Rank (Mantel-Cox)	273.100	0.000

H_0 : The survival distributions are the same in the two groups of bus type levels

H_1: The survival distributions are different in the two groups of bus type levels

χ^2 table value is 3.84 and log-rank chi-square value is 273.1 so we fail to the null hypothesis and the significant value $(p = 0.00 < 0.05)$ so it can be said that there is a difference between the "bus type" variable until the accident happen. We can also see in the survival plot Figure 4.14 that there is significant difference between the bus types levels until the accident happen.

4.2.6 Passenger Number

The variable “passenger number” is categorized as “low” and “high”. The following Table 4.26 shows the frequencies of the uncensored and censored data according to the “passenger number” variable.

Table 4.26 Summary cases for all passenger number variables

Passenger number	Total n	Uncensored		Censored	
		n	Percent %	N	Percent %
Low	570	447	78.4	123	21.6
High	1066	891	83.6	175	16.4
Overall	1636	1338	88.2	298	18.2

The rate of uncensored data in the low passenger number is 78.4 percent, the rate of uncensored data in the “high passenger number” is 83.6 percent and finally the rate of uncensored data in the “overall passenger number” is 88.2 percent, Additionally the rate of censored data in the “low passenger number” is 21.6 percent, the rate of censored data in the “high passenger number” is 16.4 percent and the rate of censored data in the “overall passenger number” is 18.2 percent. Median survival times for all passenger number levels are shown in Table 4.27.

Table 4.27 Medians for survival time passenger number variable

Passenger number	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
Low	319.000	14.717	290.155	347.845
High	305.000	13.337	278.860	331.140
Overall	312.000	10.202	292.004	331.996

The estimated median survival time is 319 days for category “low passenger number” and 305 days for category “high passenger number”. In addition, the best median survival time for “passenger number” variable is in the category “low passenger number”. The survival curve is shown in Figure 4.15 for “passenger number” variable.

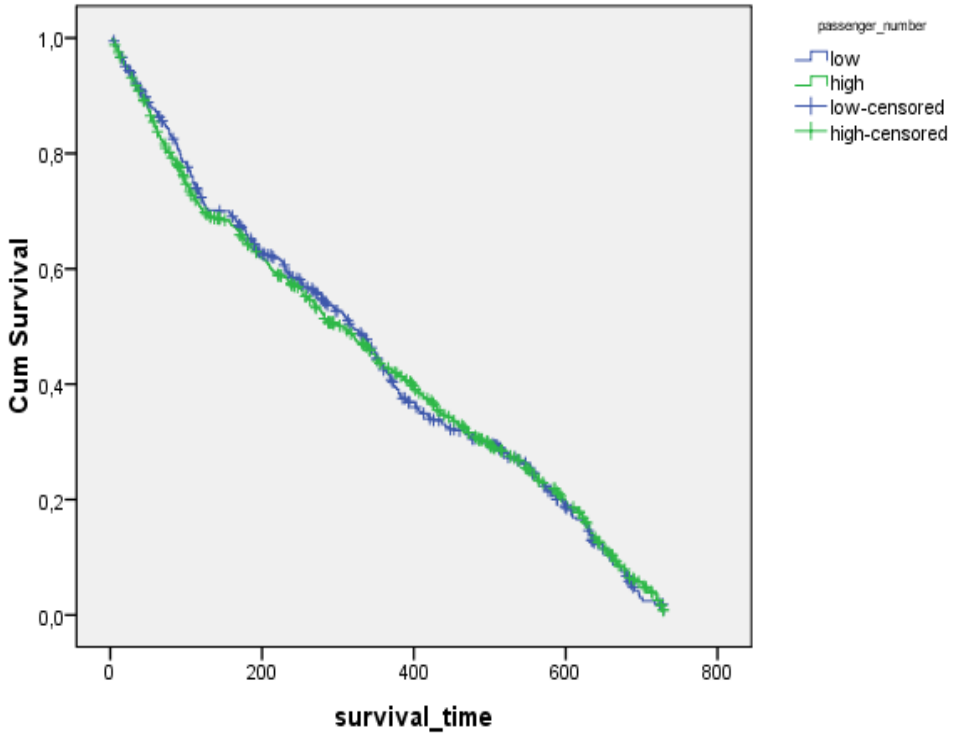

Figure 4.15 Survival plot for passenger number variable

Figure 4.15 presents survival curves were constructed, which did not show statistically significant differences of survival between "low" and "high" passenger numbers. Table 4.28 is shown log-rank results for "passenger number" variable.

Table 4.28 Test of equality of survival distributions for the different levels of passenger number variable

	Chi-Square test value	p-value
Log Rank (Mantel-Cox)	0.008	0.927

H_0: The survival distributions are the same in the two groups of passenger number levels

H_1: The survival distributions are different in the two groups of passenger number levels

The test statistic for the log-rank test is 0.008. The test statistic follows a chi-square distribution and so we find the critical value in the table of critical values for the χ^2 distribution for $df = k - 1 = 2 - 1 = 1$ and $\alpha = 0.05$ (k is the number of

passenger number levels). The critical value is 5.991 and H_0 hypothesis is not rejected $\left(\chi^2 < 5.991\right)$.We don't have statistically significant evidence $\left(p = 0.927 > 0.05\right)$ at the level of $\alpha = 0.05$ to show that the survival distributions (times) are different in the three groups of passenger number levels. In another words, there is no difference among groups of passenger number levels for survival time. This is confirmed by looking at the plot of the survival curves Figure 4.12, which both drop down initially at the same rate.

4.2.7 Driver Experience

The driver experience variable is categorized as "35 and less than 35 months driver experience" and "greater than 35 months driver experience". The following Table 4.29 shows the frequencies of the uncensored and censored data according to the "driver experience" variable.

Table 4.29 Summary cases for driver experience variables

Driver experience	Total n	Uncensored		Censored	
		n	Percent %	n	Percent %
35 and less than 35 months	821	725	88.3	96	11.7
Greater than 35 months	815	613	75.2	202	24.8
Overall	1636	1338	88.2	298	18.2

The rate of uncensored data in the "35 and less than 35 months driver experience" is 88.3 percent, the rate of uncensored data in the "greater than 35 months driver experience" is 75.2 percent and finally the rate of uncensored data in the "overall driver experience" is 88.2 percent. Additionally the rate of censored data in the "35 and less than 35 months driver experience" is 11.7 percent, the rate of censored data in the "greater than 35 months driver experience" is 24.8 percent and the rate of censored data in the "overall driver experience" is 18.2 percent. Median survival times for all driver experience levels are shown in Table 4.30.

Table 4.30 Medians for survival time driver experience variable

Driver experience	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
35 and less than 35 months	309.000	16.458	276.743	341.257
Greater than 35 months	314.000	12.985	288.550	339.450
Overall	312.000	10.202	292.004	331.996

The estimated median survival time is 309 days for category "35 and less than 35 months driver experience" and 314 days for category "greater than 35 months driver experience". In addition, the best median survival time for "driver experience variable" is in the category "greater than driver experience". The survival curve is shown in Figure 4.16 for "driver experience" variable.

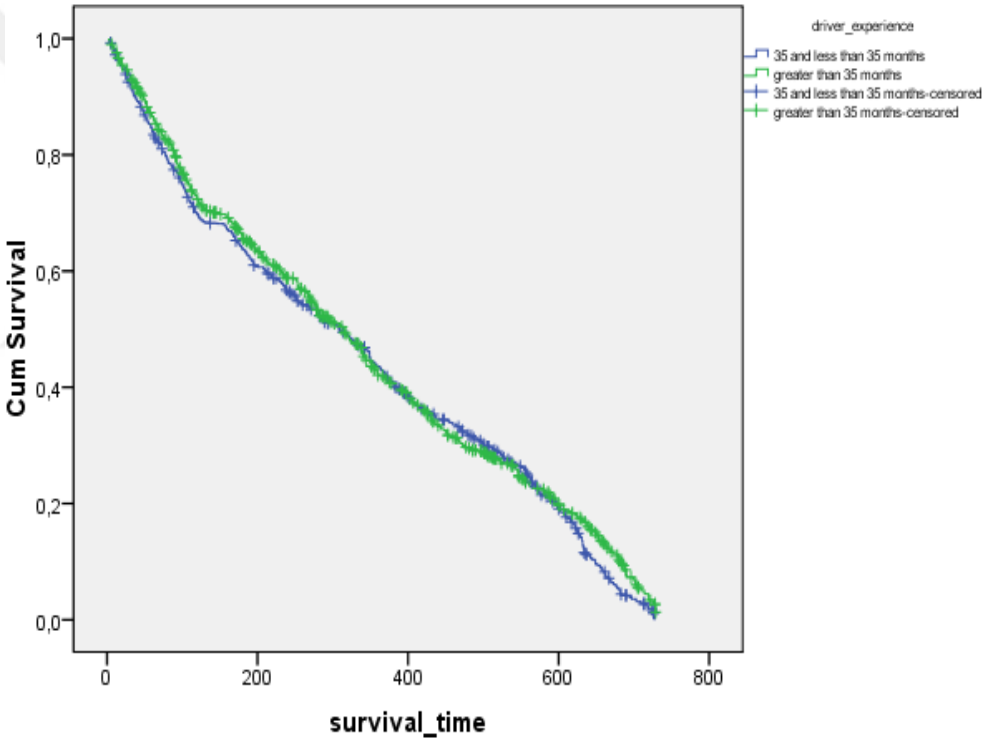

Figure 4.16 Survival plot for driver experience variable

The little pluses show when someone was censored in the plot. As the figure shows the survival curves don't differ. However conceptually, we must show the compare the log rank test results for the different levels of "driver experience" variable. Table 4.31 is shown log-rank results for "driver experience" variable.

Table 4.31 Test of equality of survival distributions for the different levels of driver experience variable

	Chi-Square test value	p-value
Log Rank (Mantel-Cox)	2.048	0.152

H_0: The survival distributions are the same in the two groups of driver experience levels

H_1: The survival distributions are different in the two groups of driver experience levels

The test statistic for the log-rank test is 2.048. The test statistic follows a chi-square distribution and so we find the critical value in the table of critical values for the χ^2 distribution for $df = k-1 = 2-1 = 1$ and $\alpha = 0.05$ (k is the number of driver experience levels). The critical value is 3.84 and H_0 hypothesis is not rejected $\left(\chi^2 < 3.84\right)$. We don't have statistically significant evidence $\left(p = 0.152 > 0.05\right)$ at the level of $\alpha = 0.05$ to show that the survival distributions (times) are different in the three groups of driver experience levels. In another words, there is no difference among groups of driver experience levels for survival time. We can also see this result in the survival plot Figure 4.16.

4.2.8 Driver Age

The driver age variable is categorized as "33 and less than years old driver age" and "greater than 33 years old driver age". The following Table 4.32 shows the frequencies of the uncensored and censored data according to the "driver age" variable.

Table 4.32 Summary cases for driver age variables

		Uncensored		Censored	
Driver age	Total n	n	Percent %	n	Percent %
33 and less than 33 years old	699	589	84.3	110	15.7
Greater than 33 years old	937	749	79.9	188	20.1
Overall	1636	1338	88.2	298	18.2

The rate of uncensored data in the “33 and less than 33 years old driver age” is 84.3 percent, the rate of uncensored data in the “greater than 33 years old driver age” is 79.9 percent and finally the rate of uncensored data in the “overall driver age” is 88.2 percent. Additionally the rate of censored data in the “33 and less than 33 years old driver age” is 15.7 percent, the rate of censored data in the “greater than 33 years old driver age” is 20.1 percent and the rate of censored data in the “overall driver age” is 18.2 percent. Median survival times for all driver age levels are shown in Table 4.33.

Table 4.33 Medians for survival time driver age variable

		Std.	95% Confidence Interval	
Driver age	Estimate	Error	Lower Bound	Upper Bound
33 and less than 33 years old	317.000	15.285	287.041	346.959
Greater than 33 years old	301.000	13.453	274.632	327.368
Overall	312.000	10.202	292.004	331.996

The estimated median survival is 317 days for category “33 and less than 33 years old driver age” and 301 days for category “greater than 33 years old driver age”. In addition, the best median survival time for “driver age” variable is in the category “33 and less than 33 years old driver age”. The survival curve is shown in Figure 4.17 for “driver age” variable.

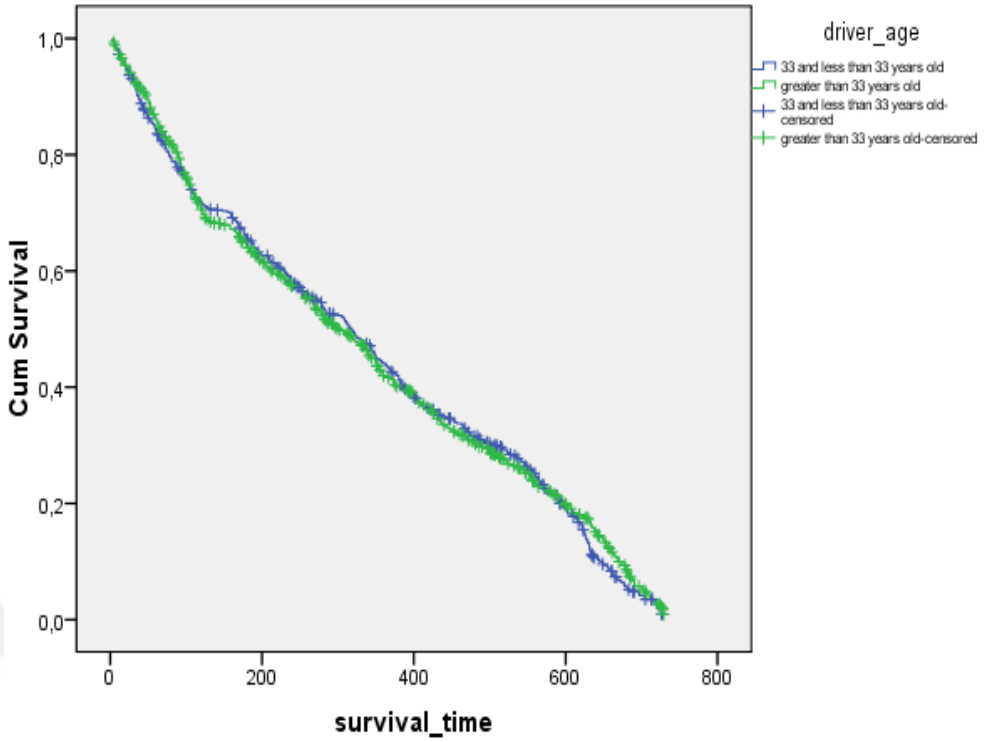

Figure 4.17 Survival plot for driver age variable

The little pluses show when someone was censored in the plot. As the figure shows the survival curves don't differ. However conceptually, we must show the compare the log rank test results for the different levels of "driver age" variable. Table 4.34 is shown log-rank results for "driver age" variable.

Table 4.34 Test of equality of survival distributions for the different levels of driver age variable

	Chi-Square test value	p-value
Log Rank (Mantel-Cox)	0.388	0.533

H_0 : The survival distributions are the same in the two groups of driver age levels

H_1: The survival distributions are different in the two groups of driver age levels

The test statistic for the log-rank test is 0.388.The test statistic follows a chi-square distribution and so we find the critical value in the table of critical values for the χ^2 distribution for $df = k-1 = 2-1 = 1$ and $\alpha = 0.05$ (k is the number of driver age levels). The critical value is 3.84 and H_0 hypothesis is not rejected $\left(\chi^2 < 3.84\right)$.

We don't have statistically significant evidence $(p = 0.388 > 0.05)$ at the level of $\alpha = 0.05$ to show that the survival distributions (times) are different in the three groups of driver age levels. We can also see in the survival plot Figure 4.17 that there is no difference between the "driver age" until the accident happens.

As a result, it cannot be said that there is a difference between the levels of "district", "traffic jam", "bus capacity", "passenger number", "driver experience", "driver age" variables until the time of the first accident. It can also be said that there is a difference between the levels of "station number" and "bus type" variables until the time of the first accident.

4.3 Cox Regression Model Application

The purpose of this thesis is to consider survival analysis based on the proportional hazards (PH) model (Cox 1972) in the real data set taken from İzmir ESHOT headquarters drivers accidents between 2012-2014 years.

The first key assumption in the Cox model is that of proportional hazards must have hazard functions that are proportional over time (i.e. constant relative hazard or the hazard ratio between two groups remains constant over time). We examinated the Cox regression proportional hazard assumptions with "log(-log) plots", "Schoenfeld residual plots", "expected vs observed plots", "Arjas plots", and "using time dependent explanatory variables methods". If the Cox regression proportional hazards assumption is violated, the simple Cox model is invalid and more sophisticated analyses are required or Kaplan-Meier analysis should be used.

Table 4.35 A section from the data set

identification_no	censor	district	bus_capacity	traffic_jam	station_number	bus_type	passenger_number	driver_experience	driver_age	survival_time
1	1	1	1	1	1	1	2	2	2	98
2	0	1	1	2	1	1	1	2	2	202
3	0	2	3	2	2	2	1	2	2	47
4	1	2	2	3	2	2	1	2	2	673
5	0	2	1	1	1	2	2	2	2	294
6	0	3	1	1	1	2	1	2	2	190
7	0	3	1	1	2	1	1	2	2	332
8	1	3	1	2	2	1	1	2	2	109
9	0	2	1	2	2	1	2	2	2	258
10	1	3	3	3	1	1	2	2	2	253
11	1	3	1	1	2	1	1	2	2	16
12	1	3	1	1	1	1	1	2	2	105
13	1	3	3	3	1	1	2	2	2	56
14	1	3	1	2	2	1	1	2	2	75
15	1	2	2	2	2	1	1	2	2	223
16	1	3	1	1	2	2	2	2	2	623
17	1	2	2	2	1	1	2	2	2	158

Summary table for categorical variables codings and frequencies is shown in Table 4.36.

Table 4.36 Summary table for categorical variables codings and frequencies

Variable levels and codes (R:Reference)	Frequencies(n)
District	
1=district 1	348
2=district 2	657
3=district 3 (R)	631
Bus capacity	
1=empty	632
2=normal	737
3=full (R)	267
Traffic jam	
1=light	376
2=moderate	847
3=heavy (R)	413
Station number	
1=57 and less than 57	838
2=greater than 57 (R)	798
Bus type	
1=big	1216
2=others (R)	420
Passenger number	
1=low	570
2=high (R)	1066
Driver experience	
1=35 and less than 35 months	821
2=greater than 35 months (R)	815
Driver age	
1=33 and less than 33 years old (R)	699
2=greater than 33 years old	937

According to Table 4.36, the district levels are coded with “district 1” as “1” and frequency of district 1 is 348, “district 2” as “2” and frequency of district 2 is 657. “district 3” as “3” and frequency of district 3 is 631. The bus capacity levels are coded with “empty” as “1” and frequency of empty is 632, “normal” as “2” and frequency of normal is 737, “full” as “3” and frequency of full is 267. The traffic jam levels are coded with “light” as “1” and frequency of light is 376 “moderate” as “2” and frequency of moderate is 847 “heavy” as “3” and frequency of heavy is 413. The station number levels are coded with “57 and less than 57” as “1” and frequency of 57 and less than 57 is 838 “greater than 57” as “2” and frequency of greater than is 57 is 798. The bus type levels are coded with “big” as “1” and frequency of big is 1216, “others” as “2” and frequency of others is 420. The passenger numbers levels are coded with “low” as “1” and frequency of low is 570, “high” as ”2” and frequency of high is 1066. The driver experience levels are coded with “35 and less than 35 months” as “1” and frequency of 35 and less than 35 months is 821. “greater than 35 months” as”2” and frequency of greater than 35 months is 815. The driver age levels are coded with “33 and less than 33 years old” as “1” and frequency of 33 and less than 33 years old is 699. “greater than 33 years old” as”2” and frequency of greater than 33 years old is 937.

SPSS knows that for example district variable “1” represents “district 1”, and so on…

There are a lot of outputs from SPSS but the following table probably contains all calculations. Cox regression model output is shown in Table 4.37.

Table 4.37 Variables in the equation for Cox regression model

	β	SE	Wald	df	p-value	Exp(β)	95% CI for Exp(β)	
							Lower	Upper
district			23.178	2	0.000			
district(1)	0.286	0.085	11.445	1	0.001	1.331	1.128	1.571
district(2)	-0.135	0.073	3.414	1	0.065	0.873	0.756	1.008
bus_capacity			16.054	2	0.000			
bus_capacity(1)	0.121	0.107	1.288	1	0.256	1.129	0.916	1.392
bus_capacity(2)	0.310	0.084	13.508	1	0.000	1.363	1.156	1.608
traffic_jam			23.048	2	0.000			
traffic_jam(1)	0.416	0.115	13.148	1	0.000	1.516	1.211	1.898
traffic_jam(2)	0.399	0.084	22.483	1	0.000	1.491	1.264	1.758
station_number	0.208	0.057	13.348	1	0.000	1.231	1.101	1.377
bus_type	1.251	0.071	314.797	1	0.000	3.494	3.043	4.011
passenger_number	0.236	0.074	10.295	1	0.001	1.266	1.096	1.463
driver_experience	0.239	0.064	14.156	1	0.000	1.270	1.121	1.439
driver_age	0.011	0.064	0.033	1	0.857	1.012	0.893	1.146

According to Table 4.37, "β" column shows the value of parameter estimations "SE" and "df" columns show their standard errors and degree of freedoms, respectively. The "Wald" column shows the output of the Wald test. The Wald test is a way to find out if explanatory variables in a model are statistical significant. "Significant" means that they add something to the model; variables that add nothing can be deleted without affecting the model in any meaningful way. The null hypothesis for the test is "some parameter"="some value". The value could be zero. If the null hypothesis is rejected, it suggests that the variables in question can be removed without much harm to the model fit. If the Wald test shows that the parameters for certain explanatory variables are zero, remove the variables from the model**.** If the test shows the parameters are not zero, include the variables in the model. The Wald test is usually talked about in terms of chi-squared, because the sampling distribution is usually known. For this reason, you can take into; to the p-value of this test. Compare the p-values to the standard significance level of "Exp(β)" column gives the exponential value of the beta coefficient, which variables with positive coefficients are associated with increased hazard and decreased survival times, i.e. as the predictor increases the hazard of the event increases and the

predicted survival duration decreases. Negative coefficients indicate decreased hazard and increased survival times. Exp(β) is the ratio of hazard rates that are one unit apart on the predictor. The last column is the confidence interval for Exp(β).

For “district” variable, the reference category is “district 3” which means that all the other categories are compared to the levels containing “district 3” category value. The p-value of “district 1” category is 0.001 so there is an evidence of a risk the time until the accident happen in either district levels. The “district 2” category has 87% of the risk of failure compared with the “district 3” category. i.e., risk is reduced by 13% but it is not significant because of p-value 0.65 is greater than 0.05 and CI for Exp(β) includes 1 (null value). Thus, time until the accident happen is 1.331 times higher in the “district 1” level as compared to the “district 3” level.

For “bus capacity” variable, the reference category is “full” which means that all the other categories are compared to the levels containing “full” category value. In the unadjusted model, there is a decrease risk of time until the accident happen compared to “empty bus capacity” and in “normal bus capacity” as compared to “full bus capacity” (hazard ratios of 1.129 and 1.363, respectively).

For “traffic jam” variable, the reference category is “heavy” which means that all the other categories are compared to the levels containing “heavy” category value. In the unadjusted model, time until the accident happen is 1.516 times higher in the “light traffic jam” level as compared to the “heavy traffic jam”. In addition, time until the accident happen is 1.491 times higher in the “moderate traffic jam” level as compared to the “heavy traffic jam” category.

For “station number” variable, the reference category is “greater than 57” which means that compared to the levels containing “greater than 57” category value. In the unadjusted model, time until the accident happen is 1.231 times higher in the “57 and less than 57 station number” as compared to the “greater than 57 station number”.

For “bus type” variable the reference category is “others” which means that compared to the levels containing “others” category value. Time until the accident happen hazard is 3.494 times higher in a “big type” of bus than “others type” of buses. In addition CI for Exp(β) is proper and this result is interpretable.

For “passenger number” variable the reference category is “high” which means that compared to the levels containing “high” category value. Time until the accident happen hazard is 1.266 times higher in a “low passenger number” than “high passenger number”.

For “driver experience” variable the reference category is “greater than 35 months” which means that compared to the levels containing “greater than 35 months” category value. Time until the accident happen is 1.27 times higher in a “35 and less than 35 months” than “greater than 35 months”.

For “driver age” variable the reference category is “33 and less than 33 years old” which means that compared to the levels containing “33 and less than 33 years old” value. “33 and less than 33 years old” drivers are 1.012 more likely to time until the accident happen at any specific time than a “greater than 33 years old” drivers. This difference is not statistically significant.

This Cox regression model with all variables is statistically significant with p-value is 0.000. We can also see this result the following table.

Table 4.38 Cox regression model with all variables results

	Chi-square	df	p-value
Cox regression model with all variables	374.956	11	0.000

Before the Cox regression model building, it is necessary to check whether the variables are independent of each other. The Spearman correlation coefficient and chi-square independence test can be used for this check. The Spearman’s rho correlation coefficients and the uncorrelated variables which are emphasized are shown in Table 4.39.

Table 4.39 Spearman correlation results for ordinal variables

			bus_capacity	traffic_jam	station_number	passenger_number	driver_experience	driver_age
Spearman's rho	bus_capacity	Correlation Coefficient	1.000	0.578**	0.001	0.480**	-0.056*	-0.044
		Sig. (2-tailed)	.	0.000	0.957	0.000	0.023	0.077
		N	1636	1636	1636	1636	1636	1636
	traffic_jam	Correlation Coefficient	0.578**	1.000	-0.165**	0.467**	-0.055*	-0.064**
		Sig. (2-tailed)	0.000	.	0.000	0.000	0.027	0.010
		N	1636	1636	1636	1636	1636	1636
	station_number	Correlation Coefficient	0.001	-0.165**	1.000	-0.139**	0.079**	0.067**
		Sig. (2-tailed)	0.957	0.000	.	0.000	0.001	0.007
		N	1636	1636	1636	1636	1636	1636
	passenger_number	Correlation Coefficient	0.480**	0.467**	-0.139**	1.000	-0.041	-0.020
		Sig. (2-tailed)	0.000	0.000	0.000	.	0.096	0.429
		N	1636	1636	1636	1636	1636	1636
	driver_experience	Correlation Coefficient	-0.056*	-0.055*	0.079**	-0.041	1.000	0.440**
		Sig. (2-tailed)	0.023	0.027	0.001	0.096	.	0.000
		N	1636	1636	1636	1636	1636	1636
	driver_age	Correlation Coefficient	-0.044	-0.064**	0.067**	-0.020	0.440**	1.000
		Sig. (2-tailed)	0.077	0.010	0.007	0.429	0.000	.
		N	1636	1636	1636	1636	1636	1636

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

Spearman correlation is a non-parametric test and nonparametric version of the Pearson correlation coefficient that is used to measure the degree of association between two variables. It was developed by Spearman, thus it is called the Spearman rank correlation. Spearman rank correlation test does not assume any assumptions about the distribution of the data and is the appropriate correlation analysis when the variables are measured on a scale that is ordinal, interval or ratio. Then null and alternative hypothesis are

$H_0 : \rho = 0$ (there is no relationship between the variables)

$H_1 : \rho \neq 0$ (there is a relationship between the variables)

According to decision rule, when

The p-value is greater than $\alpha = 0.05$, then the null hypothesis is not rejected.

The p-value is less than. $\alpha = 0.05$ we can reject the null hypothesis.

According to Table 4.39, uncorrelated variables which groups of variables suitable for the cox model are “bus capacity” - “station number”, “bus capacity” - “driver age” “passenger number” - “driver experience”, and “passenger number” - “driver age”.

Models for these variables which are uncorrelated should be determined in order and the assumption of the cox regression proportional hazards assumption for the model which is significant should be done.

The independence examination of the nominal variables can be done by the chi-square test.

Table 4.40 District vs all variables chi-square independence results (p-values)

	bus capacity	traffic jam	station number	bus type	passenger number	driver experience	driver age
district	0.000	0.000	0.045	0.003	0.000	0.001	0.001

Table 4.41 Bus type vs all variables chi-square independence results (p-values)

	district	bus capacity	traffic jam	station number	passenger number	driver experience	driver age
bus type	0.003	0.000	0.002	0.662*	0.000	0.001	0.010
*: nonsignificant at the level of α=0.05							

The following table contains the results of Cox regression model with uncorrelated variables. This table is the main output for this procedure.

Table 4.42 Uncorrelated variables cox regression significance results

Cox Model	p-value
Bus capacity and station number	0.154
Bus capacity and drive age	0.647
Passenger number and driver experience	0.356
Passenger number and driver age	0.820
Bus type and station number	0.000*
*<0.05	

"Bus capacity-station number", "bus capacity-driver age"," passenger number-driver experience" and "passenger number-driver age" Cox regression models are not statistically significant according to Table 4.42. On the other hand, "bus type-station number" regression model is significant at the level of α=0.05. So that the assumption of a Cox regression proportional hazards for "bus type-station number" cox model should be checked. Graphical methods and time dependent explanatory variable methods can be used to examine the cox regression proportional hazards assumption. These methods are shown in Figure 4.18.

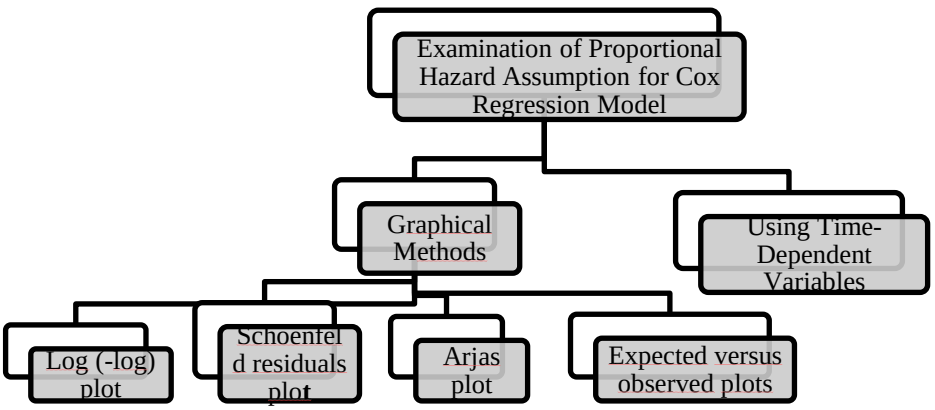

Figure 4.18 Examination of proportional hazards assumption for cox regression model

4.4 Examination of Cox Regression Proportional Hazards Assumption for Station Number and Bus Type Variables

4.4.1 Station Number Variable

4.4.1.1 Log(-log) Plot for Station Number Variable

Log(-log) plot for station number variable is shown in Figure 4.19.

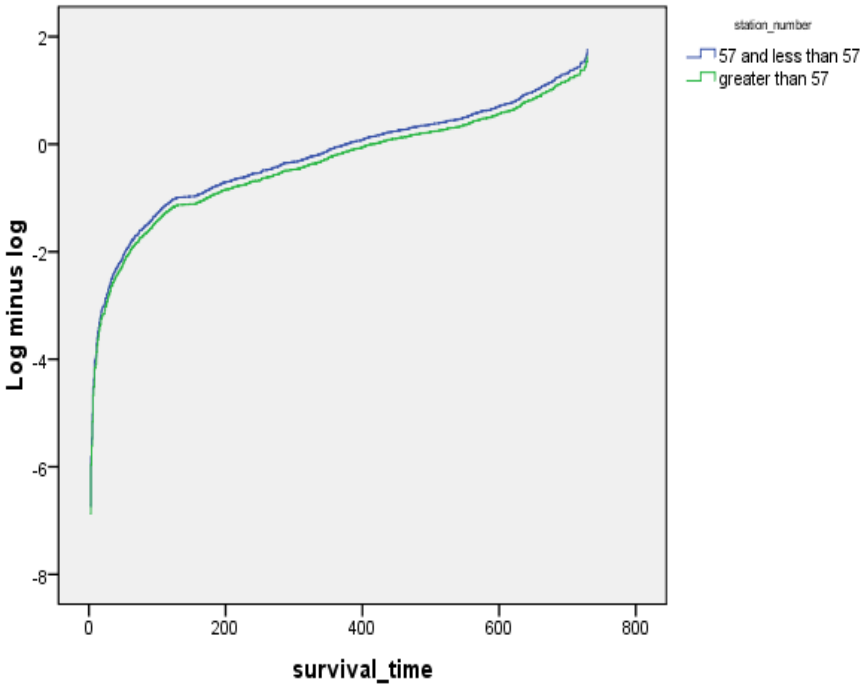

Figure 4.19 Log(-log) plot for station number variable

Log(-log) plot of "station number" variable for the Cox regression proportional hazards assumption show the separate lines as approximately parallel to each other and lines' distance is constant over time. According to Figure 4.19, log(-log) plot for "station number" variable satisfies the Cox regression proportional hazards assumption.

4.4.1.2 Expected Versus Observed Plots for Station Number Variable

Expected survival functions estimated from cox regression vs survival time for "station number" variable is shown in Figure 4.20.

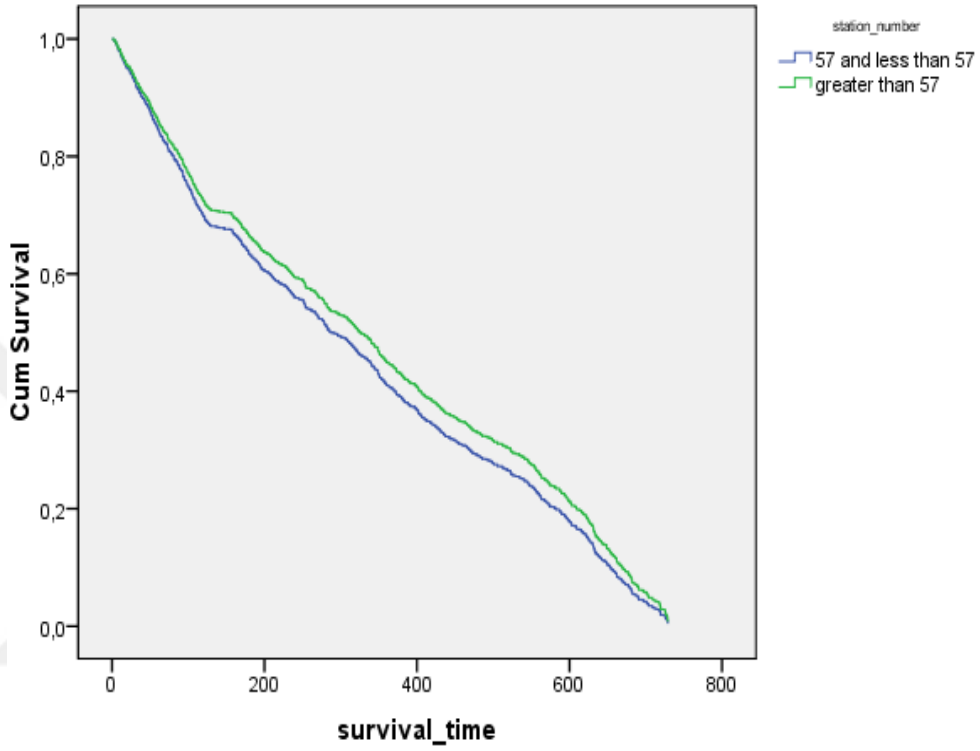

Figure 4.20 Expected survival functions estimated from cox regression vs survival time for station number variable plot

Observed survival functions estimated from Kaplan-Meier vs survival time for "station number" variable is shown in Figure 4.21.

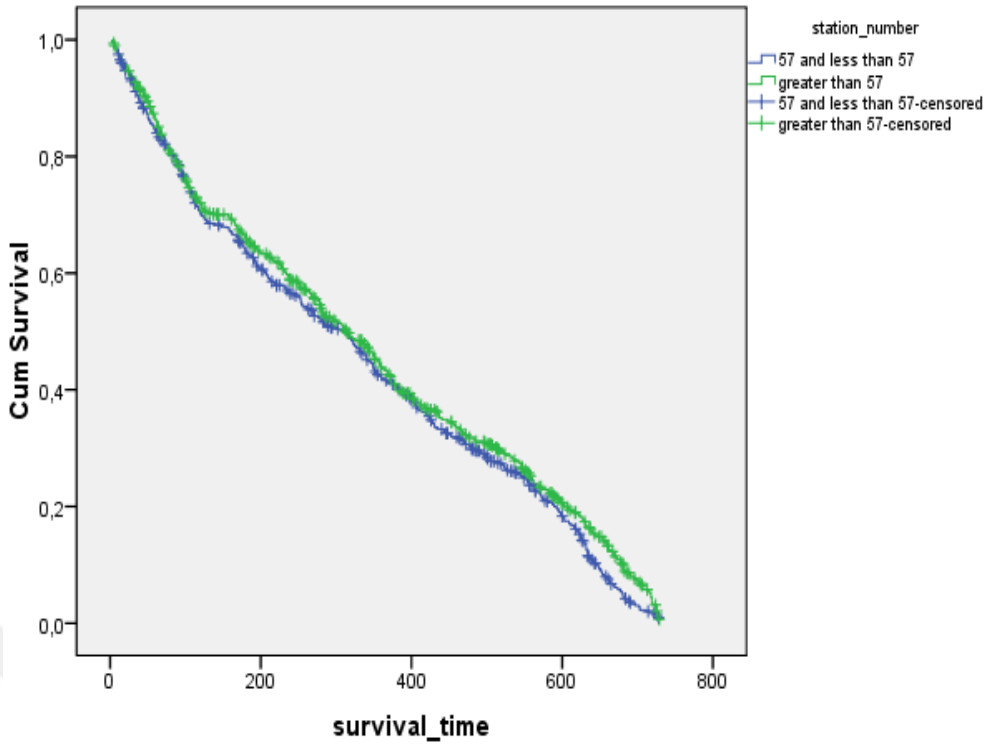

Figure 4.21Observed survival functions estimated from Kaplan-Meier vs survival time for station number variable plot

When the expected survival curve in Figure 4.20 and observed survival curve in Figure 4.21 are close together, the Cox regression proportional hazards assumption has not been violated. According to these graphs, "station number" variable satisfies of the Cox regression proportional hazards assumption.

4.4.1.3 Arjas Plot for Station Number Variable

Arjas plot for "station number" variable is shown in Figure 4.22.

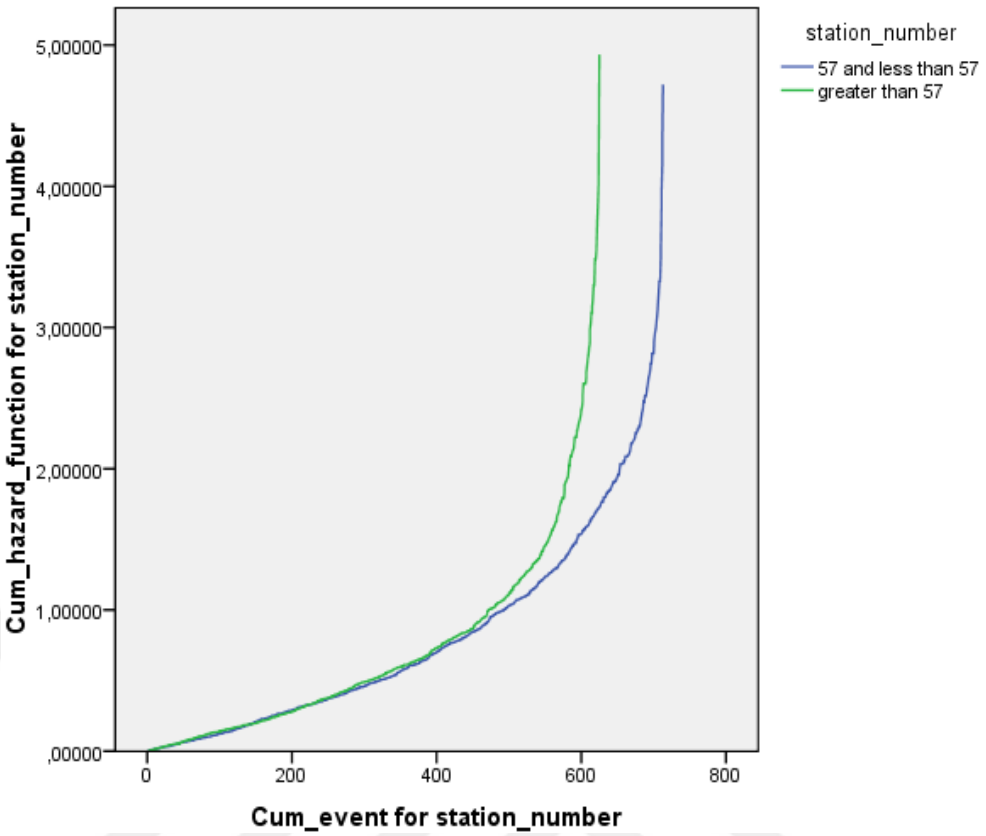

Figure 4.22 Arjas plot for station number variable

If the the cox regression proportional hazards assumption satisfies, it is expected that the slope of the Arjas graph is close to one and approximately linear. According to Figure 4.22, station number variable violates the proportional hazards assumption.

4.4.1.4 Adding Time Dependent Variable for Station Number Variable

The product of variable T_ and variable station number is shown with T_COV_ in SPSS. Then we build a cox regression model with T_COV_ and "station number" variable, Table 4.43 is shown this models results.

Table 4.43 Variables in the equation for time dependent station number variable

	p-value
T_COV_	0.146
Station number	0.974

T_COV was removed from the model so that the model is not time dependent. The assumption of proportional hazards satisfied according to the time-dependent explanatory variable addition method.

Results obtained from checking the Cox regression proportional hazards assumption using various methods can be summarized as in the following Table 4.44 for station number variable.

Table 4.44 Results obtained from checking the cox regression proportional hazards assumption using various methods

	Log(-log) plot	Schoenfeld residuals plot	Arjas plot	Expected versus observed plots	Using time dependent variables
Station number	✓	Don't use	X	✓	✓
✓	Cox regression proportional hazards assumption satisfied				
X	Cox regression proportional hazards assumption violated				

According to these all results, "station number" variable violates the proportional hazards assumption only according to the Arjas plot. On the other hand, the other methods conclude that the proportional hazards assumption holds. However according to Table 4.45 does not give significant results, it would be more accurate to interpret the results of Kaplan-Meier.

Table 4.45 Variables in the equation for station number variable

	β	SE	Wald	df	Sig.	Exp(β)	95.0% CI for Exp(β)	
							Lower	Upper
station_number	0.108	0.055	3.838	1	0.05	1.114	1.000	1.240

4.4.2 Bus Type Variable

4.4.2.1 Log(-log) Plot for Bus Type Variable

Log(-log) plot for "bus type" variable is shown in Figure 4.23.

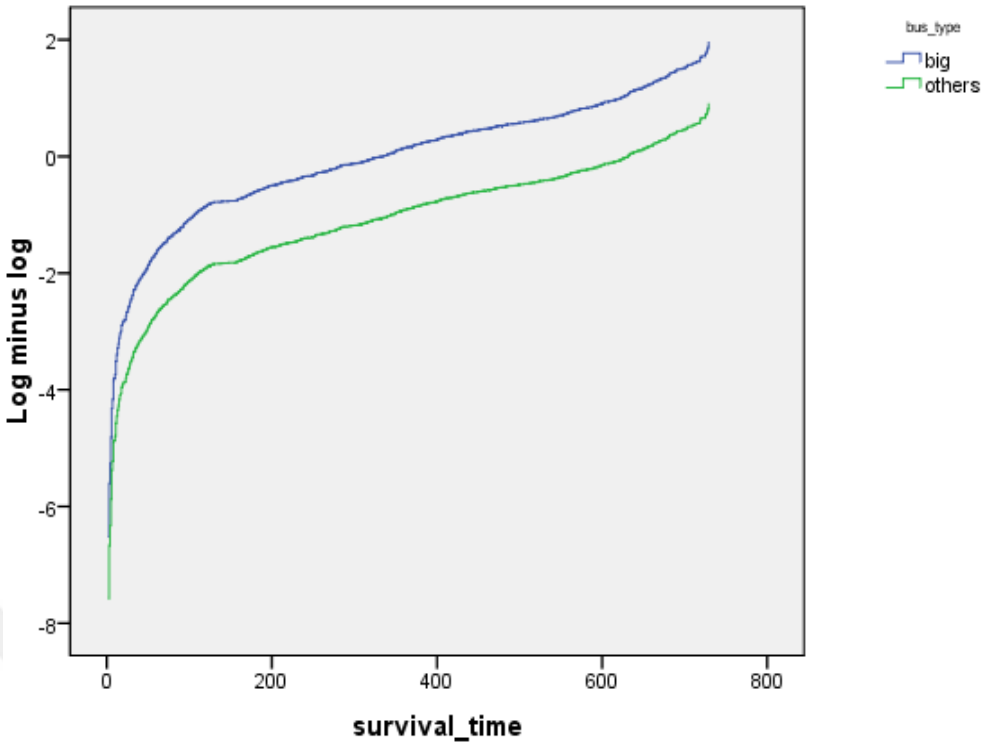

Figure 4.23 Log(-log) plot for bus type variable

Log(-log) plot of “bus type” variable for the Cox regression proportional hazards assumption show the separate lines are approximately parallel to each other and lines distance is constant over time. According to Figure 4.23, log(-log) plot for “bus type” variable satisfies the cox regression proportional hazards assumption.

4.4.2.2 Expected versus Observed Plots for Bus Type Variable

Expected survival functions estimated from cox regression vs survival time for “bus type” variable is shown in Figure 4.24.

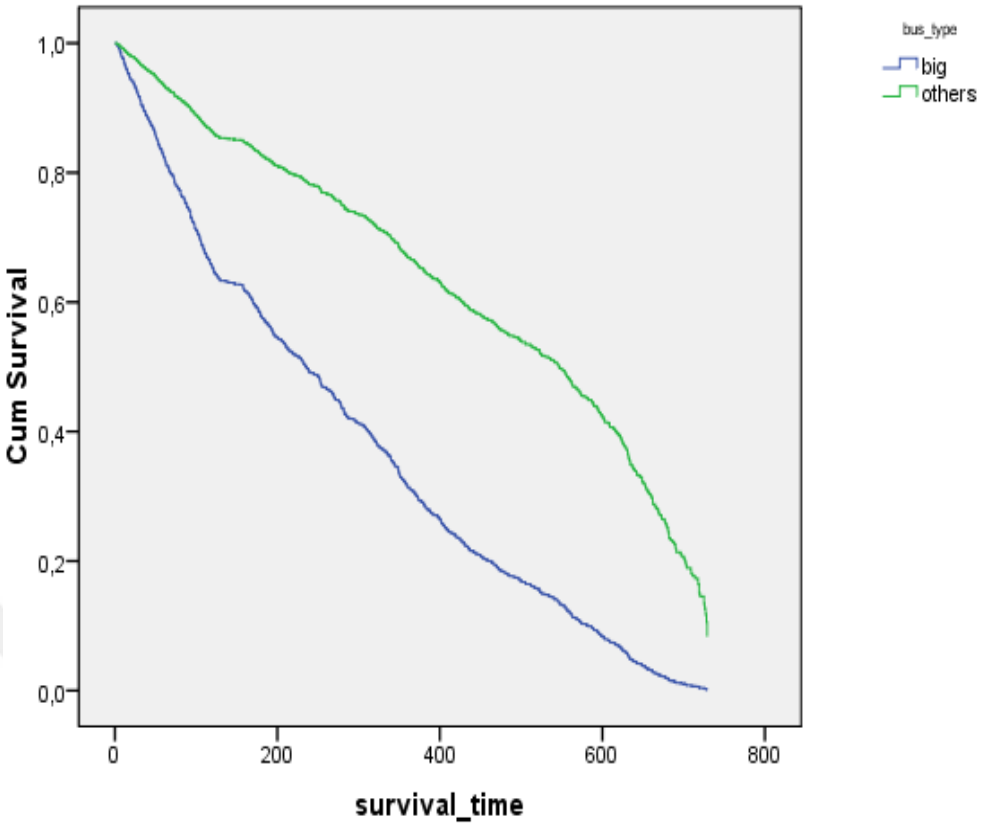

Figure 4.24 Expected survival functions estimated from cox regression vs survival time for bus type variable plot

Observed survival functions estimated from Kaplan-Meier vs survival time for "bus type" variable is shown in Figure 4.25.

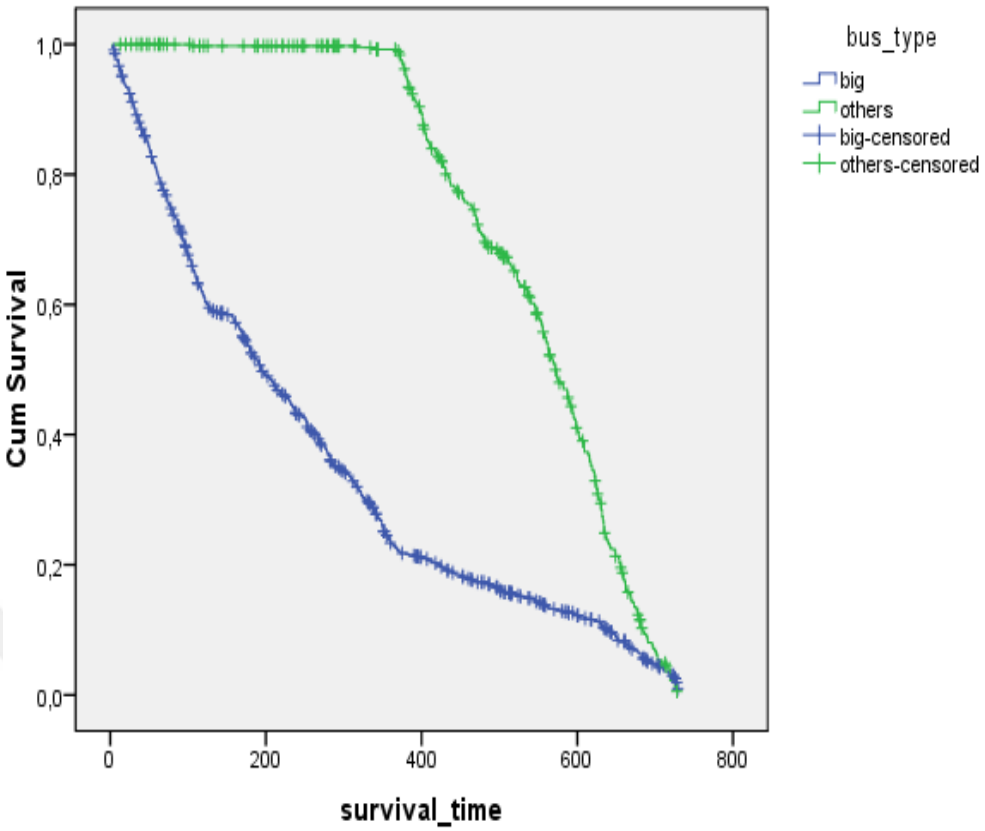

Figure 4.25 Observed survival functions estimated from Kaplan-Meier vs survival time for bus type variable plot

When the expected survival curve in Figure 4.24 and observed survival curve in Figure 4.25 are close together, the cox regression proportional hazards assumption has not been violated. According to these graphs, “bus type” variable is violated of the cox regression proportional hazards assumption.

4.4.2.3 Arjas Plot for Bus Type Variable

Arjas plot for “bus type” variable is shown in Figure 4.26.

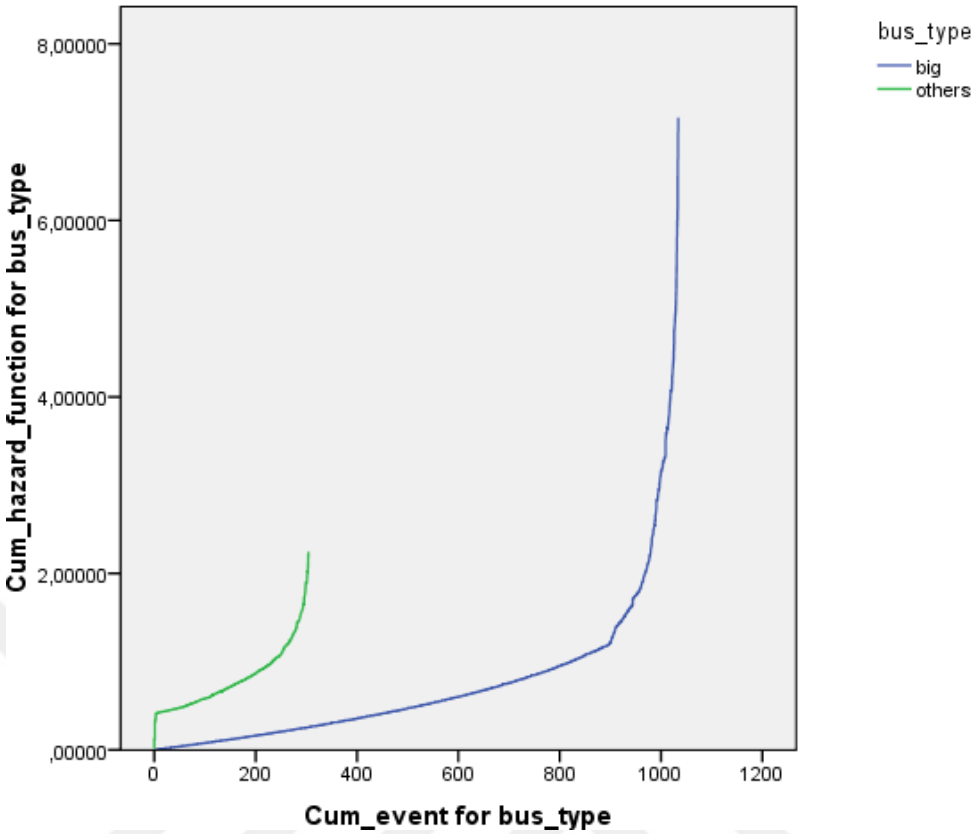

Figure 4.26Arjas plot for bus type variable

If the the cox regression proportional hazards assumption satisfies it is expected that the slope of the Arjas graph is close to one and approximately linear. According to Figure 4.26 “bus type” variable is violated the proportional hazards assumption.

4.4.2.4 Adding Time Dependent Variable for Bus Type Variable

The product of variable T_ and variable “bus type” is shown with T_COV_ in SPSS. Then we build a cox regression model with T_COV_ and “bus type” variable. Table 4.46 is shown this models results.

Table 4.46 Variables in the equation for time dependent bus type variable

	p-value
T_COV_	0.000
Bus type	0.000

T_COV was not removed from the model so that the model became time dependent. The assumption of proportional hazards is violated according to the time-

dependent explanatory variable addition method. According to these results, "bus type" variable violated the cox regression proportional hazards assumption.

Results obtained from checking the Cox regression proportional hazards assumption using various methods can be summarized as in the following Table 4.47 for bus type variable.

Table 4.47 Results obtained from checking the cox regression proportional hazards assumption using various methods

	Log(-log) plot	Schoenfeld residuals plot	Arjas plot	Expected versus observed plots	Using time-dependent variables
Bus type	✓	Don't use	X	X	X
✓	Cox regression proportional hazards assumption satisfied				
X	Cox regression proportional hazards assumption violated				

According to these all results, "bus type" variable is violates the proportional hazards assumption for the Arjas plot, expected versus observed plots and using time dependent variable method.

4.5 Discussion and Conclusion

The event of interest was that the drivers experience the first accident and the survival time was the time passed until the first accident. The explanatory variables (the district, the traffic jam, the capacity of bus, the number of passengers, the number of stations, the type of bus, the age of the driver and the experience of driver) used for the event of interest were first investigated by descriptive statistics and frequency tables. Specified distributions (normal, uniform or exponential) were checked for the event of interest with the Kolmogorov-Smirnov test. Data includes censored survival times in this thesis, so the Kaplan Meier method was first applied and the relationships among the variables were examined by the "log-rank" test and median survival times were interpreted.

Before the cox regression model building, it is necessary to check whether the variables are independent of each other. The Spearman correlation coefficient values and chi-square test can be used for this check and the uncorrelated variables which

are emphasized. The cox results of the models built according to the correlation results were examined. The cox regression proportional hazards assumption was checked for the bus type and station number variables. The cox regression model with the station number was found significant. As a result, it would be more accurate to use Kaplan-Meier's results for this real data set for showing the other variables.

CHAPTER FIVE
CONCLUSION

Survival analysis is generally defined as a set of methods for analyzing data where the outcome variable is the time until the occurrence of an event of interest. The event can be death, occurrence of a disease, accident, marriage, divorce, breakdown of a product etc. The time to event or survival time can be measured in days, weeks, years, etc. It depends on the event. In survival analysis, observations are called censored when the information about their survival time is incomplete. Censoring is an important issue in survival analysis; it represents a particular type of missing data.

Unlike ordinary regression models, survival methods correctly incorporate information from both censored and uncensored observations in estimating important model parameters. The outcome variable in survival analysis is composed of two parts: one is the time to event and the other is the event status, which records if the event of interest occurred or not. The probability density, survival and hazard functions are key concepts in survival analysis for describing the distribution of event times. The survival function gives, for every time, the probability of surviving (or not experiencing the event) up to that time. The hazard function gives the potential that the event will occur, per time unit, given that an individual has survived up to the specified time. In survival analysis, the median is generally a better measure of central location than the mean because of the distribution of the survival time.

It is generally of interest in survival studies to describe the relationship of a factor of interest to the time to event, in the presence of several covariates, such as age, gender, race, etc. A number of models are available to analyze the relationship between covariates and outcome variable with the survival time. Methods include parametric, nonparametric and semiparametric approaches. Parametric methods assume that the underlying distribution of the survival times follows certain known probability distributions (normal, exponential, weibull, gamma and lognormal). Model parameters are usually estimated using an appropriate modification of

maximum likelihood. A nonparametric estimator of the survival function, the Kaplan Meier method is widely used to estimate and graph survival probabilities as a function of time. It can be used to obtain univariate descriptive statistics for survival data, including the median survival time, and compare the survival experience for two or more groups of subjects. The log-rank test is available to test for overall differences between estimated survival curves of two or more groups of subjects. This test is a type of chi-square test, a widely used test in practice, and in reality it is a method for comparing the Kaplan-Meier curves estimated for each group of subjects.

A popular regression model for the analysis of survival data is the Cox proportional hazards regression model. It allows testing for differences in survival times of two or more groups of interest, while allowing to adjust for covariates of interest. The Cox regression model is a semiparametric model, making fewer assumptions than typical parametric methods but more assumptions than nonparametric methods. In particular, and in contrast with parametric models, it makes no assumptions about the shape of the so-called baseline hazard function.

The Cox regression model provides useful and easy to interpret information regarding the relationship of the hazard function to independent variables. While a nonlinear relationship between the hazard function and the independent variables is assumed, the hazard ratio comparing any two observations is in fact constant over time in the setting where the independent variables do not vary over time. This assumption is called the proportional hazards assumption and checking if this assumption is met is an important part of a Cox regression analysis. There are many methods to check this assumption.

In this thesis, we examined the proportional hazards assumption with

- Log(-log) plots,
- Schoenfeld residual plots,
- Expected vs Observed plots,

- Arjas plots and
- Using time dependent explanatory variables methods.

In addition, we applied survival methods (survival analysis, Cox regression model) in the real data set taken from İzmir ESHOT headquarter, containing drivers accidents between 2012-2014 years. The event of interest was that the drivers experience the first accident and the survival time was the time passed until the first accident. The independent (explanatory) variables were the district, the traffic jam, the bus capacity, the passenger number, the station number, the bus type, the driver age and the driver experience. These independent variables were investigated by frequency tables and graphs. The Kolmogorov-Smirnov test was used to decide it sample comes from a population with a specific distribution (normal, uniform or exponential) and survival time distribution does not follow a specified distribution (normal. uniform or exponential) for this data set with p-values were 0.000 for all results in Chapter 4 Table 4.2. After the Kaplan-Meier analysis was used to estimate the survival curves for independent variables. We had also identified the variable category that has the best median of survival time.

Before fitting a Cox regression model, it was necessary to check whether the variables were independent of each other. The Spearman correlation coefficient and chi-square independence test can be used for this check. Models for these variables which were uncorrelated determined in order and the assumption of the cox regression proportional hazards assumption for the model which was significant was made in Chapter 4 Table 4.42. The cox regression proportional hazards assumption was checked for the “bus type” and “station number” variables.

Table 5.1 Summary table for station number and bus type variables cox regression proportional hazards assumption

	Log (-log) plot	Schoenfeld residuals plot	Arjas plot	Expected versus observed plots	Using Time-Dependent Variables
Station number	✓	Don’t use	X	✓	✓
Bus Type	X	Don’t use	X	X	X
✓	Cox regression proportional hazards assumption satisfied				
X	Cox regression proportional hazards assumption violated				

The cox regression proportional hazards assumption was checked for the “bus type” and “station number” variables. The cox regression model with the station number was found significant for holding the proportional hazards assumption. But it didn’t hold for “bus type” variable. As a result, the cox regression model with a single variable (station number) was built. In addition, the results of Kaplan-Meir were interpreted for the other variables.

The log-rank test was used to compare the Kaplan-Meier curves estimated for each group of subjects. According to these results, the time among groups until the accident happens was different for “station number” and “bus type” variables. The estimated median survival time was 309 days for category “57 and less than 57 station number” and 315 days for category “greater than 57 station numbers”. In addition, the best median survival time for “station number” variable was in the category “greater than 57 station number”. The estimated median survival time was 195 days for category “big bus type” and 532 days for category “other bus type”. In addition, the best median survival time for “bus type” variable was in the category “other bus type”.

Finally, the survival time passed until the first accident was found to depend on “station number” variable with regard to Cox proportional hazards assumption. In addition, the survival analysis and examination Cox regression proportional hazards assumptions in the field of transportation for Turkey is an original contribution to literature.

REFERENCES

Arjas, E. (1988). A graphical method for assessing goodness of fit in Cox's proportional hazards model. *Journal of the American Statistical Association*, *83*(401), 204-212.

Ata, N., Karasoy, D., ve Sözer, M. T. (2008). Orantısız hazardlar için parametrik ve yarı parametrik yaşam modelleri. *İstatistikçiler Dergisi*, *1*(3), 125-134.

Ata, N., Karasoy, D., ve Sözer, M. T. (2007). Orantılı tehlike varsayımının incelenmesinde kullanılan yöntemler ve bir uygulama. *Eskişehir Osmangazi Üniversitesi Müh. Mim. Fak. Dergisi*, *20*(1), 57-80.

Bulut, V. (2011). *Türkiye'de işsizlik süresini etkileyen faktörlerin yaşam çözümlemesi ile incelenmesi.* Hacettepe Üniversitesi, Yüksek Lisans Tezi, Ankara.

Harris, S. (2009). *Additional regression techniques*. Retrieved June 20, 2017, from www.edshare.soton.ac.uk/id/document/9437

Kardiyen, F., & Kaygisiz, G. (2011). Kırmızı ışık kural ihlali nedeni ile meydana gelen trafik kazalarının değerlendirilmesi. *Ankara Trafik Vakfı, 42,* 1-13.

Kay, R. (1977). Proportional hazard regression models and the analysis of censored survival data. *Applied Statistics*, 227-237.

Klein, J. P., & Moeschberger, M. L. (2005). *Survival analysis techniques for censored and truncated data*. (2nd ed.). Springer Science & Business Media.

Le, C. T. (1997). *Applied survival analysis*. Wiley.

Lee, E. T., & Wang, J. (2003). *Statistical methods for survival data analysis* (3th ed.). John Wiley & Sons.

Roy, T. (2016). *Extended cox model (time dependent covariates) and its application.* Retrieved June 20, 2017, from http://www.phusewiki.org/docs/2016_Trivandrum_SDE/5_Tapas%20Roy.pdf

Sainani, K. (n.d.). *Introduction to cox regression*. Retrieved June 20, 2017, from www.pitt.edu/~super4/33011-34001/33091-33101.ppt

Sertkaya, D., Ata, N., ve Sözer, M. T. (2005). Yaşam çözümlemesinde zamana bağlı açıklayıcı değişkenli cox regresyon modeli. *Ankara Üniversitesi Tıp Fakültesi Mecmuası*, *58*(1), 153-158.

Smith, T., & Smith, B. (2001). Survival analysis and the application of cox's proportional hazards modeling using SAS. In *Proceedings of the twenty-sixth annual SAS user's group international conference* (244-246). Cary (NC): SAS Institute Inc.

Sullivan L. (n.d.). *Survival analysis.* Retrieved December 1, 2017, from http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Survival/BS704_Survival_print.html

Terzi, Y., Cengiz, M. A., ve Bek, Y. (2005). Cox regresyon modelinde oransal hazard varsayımının artıklarla incelenmesi ve akciğer kanseri hastaları üzerinde uygulanması. *Türkiye Klinikleri Journal of Medical Sciences*, *25*(6), 770-775.

Vermeylen F. (2005). *Censored data*. Retrieved June 20, 2017, from https://www.cscu.cornell.edu/news/statnews/stnews67.pdf

Yay, M., Çoker, E., ve Uysal, Ö. (2007). Yaşam analizinde cox regresyon modeli ve artıkların incelenmesi. *Cerrahpaşa Tıp Dergisi*, *38*(4), 139-145.