**Nour Mohamed Eletreby**

**Task 2 Completed**

**A research about Survival Analysis**

**First: Introduction**

Survival analysis is known as time-to-event data accounts for the dynamic aspect of time by considering the probability of an event occurring over a specific time interval. Here are various examples to illustrate the concepts:

-Medical Research: In a clinical trial, researchers might be interested in studying the time it takes for patients to recover from a particular disease, experience a relapse, or develop a side effect. This time duration is crucial for assessing the efficacy of a treatment.

-Engineering: Engineers may want to analyze the time it takes for a piece of machinery or equipment to fail. This information helps in preventive maintenance and understanding the reliability of the equipment.

-Finance: In finance, time-to-event data can be used to model the time it takes for a bond to default or for a stock to reach a certain price level. This is essential for risk assessment and investment strategies.

**Second: Definition**

Survival analysis is a statistical method used to analyze and interpret time-to-event data. In various fields such as medical research, epidemiology, finance, and engineering, it plays a crucial role in understanding the time until a specific event of interest occurs. This event could be anything from a patient's recovery or relapse, a product's failure, or the occurrence of an economic event.

Time-to-event data are unique because they involve two key components:

Time: This represents the duration or time elapsed until the event occurs or until the observation is censored

Event: This is the specific outcome or event of interest. It could be a positive event (e.g., recovery, success, death) or usually a negative event (e.g., failure, relapse, default).

There are three primary goals of survival analysis, to estimate and interpret survival functions from the survival data; to compare survival, and to assess the relationship of explanatory variables to survival time.

**Third : Characteristics**

-Time-to-Event Data: Survival analysis focuses on time-to-event or time-to-failure data, where the primary interest is in understanding when an event of interest occurs. This event can be anything from a patient's recovery or death in medical research to machinery failure in engineering.

-Censored Data: One of the defining features of survival analysis is the presence of censored data. Censoring occurs when some observations do not experience the event of interest during the study period or are lost to follow-up before the event occurs. Survival analysis accounts for these censored observations in its calculations.

-Survival Function: The survival function (often denoted as S(t)) is a fundamental concept in survival analysis. It represents the probability that an individual or item will survive beyond a specific time point, t. The survival function provides insights into the survival or event occurrence probabilities over time.

-Hazard Function: The hazard function (often denoted as λ(t) or h(t)) describes the instantaneous rate at which events occur at a given time, t, for individuals who have survived up to that point. It measures the risk of experiencing the event at any given moment.

-Survival Curves: Survival analysis produces survival curves (Kaplan-Meier curves) that visually depict the probability of survival over time. These curves illustrate how the probability of an event changes as time progresses. Survival curves can be used to compare different groups or conditions.

-Covariates and Risk Factors: Survival analysis allows for the inclusion of covariates or independent variables to assess their impact on the time to the event. This is often done through techniques like the Cox proportional hazards model, which estimates the hazard ratio for each covariate.

-Right-Censoring: In many real-world applications, data are often right-censored, meaning that events of interest have not occurred by the end of the study period. Survival analysis accommodates this type of censoring and provides methods to estimate survival probabilities in the presence of censored data.

Examples: Clinical Trials: In a clinical trial testing a new cancer treatment, patients are followed for a certain period to observe their survival times. However, at the end of the study, some patients may still be alive without experiencing the event (death). Their data are right-censored because the event (death) did not occur during the study period.

-Time-Varying Covariates: Survival analysis can incorporate covariates that change over time, reflecting the dynamic nature of risk factors. This is important when studying diseases or conditions where the impact of covariates may evolve.

-Non-Parametric and Parametric Models: Survival analysis offers both non-parametric methods (e.g., Kaplan-Meier estimator) and parametric models (e.g., exponential, Weibull) to describe and predict survival data. The choice of model depends on the data distribution and assumptions.

-Cohort Studies and Clinical Trials: Survival analysis is commonly used in cohort studies and clinical trials to assess the effectiveness of treatments, interventions, or exposures in terms of their impact on the time to an event.

-Time-Dependent Effects: Survival analysis allows for the assessment of how the impact of covariates on survival changes over time. This is particularly relevant in cases where the hazard rate varies with time.

-Cure Models: In some situations, not all individuals will experience the event of interest (e.g., cancer remission). Survival analysis can accommodate "cure models" to estimate the proportion of individuals who will never experience the event.

**Fourth: How to model survival data and the analysis of survival data**

-Data Collection and Preparation:

Gather relevant data, including the time-to-event (survival time), event status (e.g., event occurred or censored), and any covariates (independent variables).

Clean and preprocess the data, handling missing values and outliers as necessary.

To illustrate: In survival analysis, when the data take values of 0 or 1, they typically represent the status of the event of interest. This binary coding is commonly used to indicate whether an event has occurred or not for each observation within a given time frame. In survival analysis, a status of 1 signifies that the event has occurred, while a status of 0 indicates that the event has not occurred within the specified observation period, and the data are right-censored for that observation.

-Exploratory Data Analysis (EDA):

Conduct descriptive analysis to understand the distribution of survival times, visualize survival curves, and assess the impact of covariates.

Use techniques such as Kaplan-Meier survival curves and log-rank tests to compare survival distributions among different groups.

-Select a Survival Model:

Choose an appropriate survival model based on the characteristics of your data. Common models include:

Kaplan-Meier Estimator: Non-parametric method for estimating survival curves.

Cox Proportional Hazards Model: A widely used semi-parametric model that assesses the effects of covariates on the hazard rate.

-Model fitting

-Asses model fit

-Predication and Inference

-Interpret and report.

**Fifth: Kaplan-Meier Estimator and Cox Proportional Hazards Model**

The Kaplan-Meier estimator, often simply referred to as the Kaplan-Meier curve, is a non-parametric statistical method used in survival analysis to estimate the probability of an event occurring over time. It is a fundamental tool for analyzing time-to-event data, where the event of interest could be anything from patient survival to equipment failure.

The Kaplan-Meier estimator is used to calculate the survival function, which provides estimates of the probability that an individual or item will survive beyond a specific time point.

Key characteristics of the Kaplan-Meier estimator and its curve:

The estimator accommodates right-censored data, which is common in survival analysis when not all events have occurred by the end of the study.

It provides survival probabilities at specific time points, making it easy to estimate the probability of survival at any desired time.

The Kaplan-Meier curve can be used to compare survival distributions among different groups or strata by plotting multiple curves on the same graph.

The estimator does not make assumptions about the underlying probability distribution of survival times, making it a non-parametric method.

It is widely used in clinical trials, epidemiology, engineering, and other fields to visualize and analyze time-to-event data.

How it works?

-Data Preparation: You start with a dataset containing information about survival times (time-to-event data) and the status of each observation (whether the event occurred or not, usually coded as 1 for event occurred and 0 for right-censored). This data can be used to construct a survival curve.

-Estimation: The Kaplan-Meier estimator calculates the survival probabilities at specific time intervals. It does this by taking into account the observed survival times and the number of individuals or items still at risk (i.e., have not experienced the event) at each time point. The estimator is a product-limit formula that iteratively updates the survival probabilities based on the observed data.

-Construction of Survival Curve: The survival probabilities obtained from the Kaplan-Meier estimator are used to construct a stepwise survival curve. This curve represents how the probability of survival changes over time. It typically starts at 1 (indicating 100% survival at time zero) and gradually declines as time progresses or events occur.

-Visualization: The Kaplan-Meier survival curve is often visualized graphically. The x-axis represents time, and the y-axis represents the estimated survival probability. The curve may have steps or stair-like segments, which indicate when events occurred. It can also include confidence intervals to express the uncertainty in the estimates.

The log-rank test is a statistical test commonly used alongside Kaplan-Meier survival analysis to compare the survival curves of different groups or categories. It's used to determine if there are statistically significant differences in survival times among these groups.

Data Preparation: You start with survival data that includes information about survival times and the status of each observation (event occurred or right-censored). You also have one or more groups or categories (e.g., treatment groups, patient subgroups) that you want to compare in terms of survival.

Kaplan-Meier Curves: First, you use the Kaplan-Meier estimator to calculate the survival curves for each group. This involves estimating the survival probabilities over time for each group.

Log-Rank Test: The log-rank test is a hypothesis test that assesses whether there is a statistically significant difference between the survival curves of the groups. It does this by comparing the observed number of events and censored observations in each group with the expected numbers if there were no differences between the groups. The test calculates a test statistic, often denoted as "χ²" (chi-squared), which follows a chi-squared distribution.

Hypothesis Testing: The null hypothesis (H0) of the log-rank test is that there are no differences in survival between the groups, i.e., the survival curves are the same. The alternative hypothesis (H1) is that there are differences in survival between the groups.

Statistical Significance: If the calculated test statistic is large enough (i.e., exceeds a critical value determined by the chi-squared distribution), it suggests that the survival curves are significantly different between at least one pair of groups. In such cases, you reject the null hypothesis and conclude that there are significant differences in survival

**Now I will talk about cox model**

This model is used to investigate the relationship between one or more covariates (independent variables) and the hazard of an event occurring over time. In essence, it allows you to examine how these covariates influence the risk or hazard of an event.

Here are the key components and features of the Cox proportional hazards model:

-Hazard Function: The Cox model is a semi-parametric model that describes the hazard function. The hazard function, denoted as λ(t) or h(t), represents the instantaneous rate at which events occur at a given time t for individuals who have survived up to that time. It characterizes the risk of the event happening at any moment in time.

-Proportional Hazards Assumption: The central assumption of the Cox model is that the hazard ratios for the covariates are constant over time. In other words, the effect of covariates on the hazard is proportional and does not change with time. This assumption simplifies the modeling process.

-Covariates: The Cox model allows you to include one or more covariates that you believe may influence the hazard of the event. These covariates can be continuous or categorical variables.

-Partial Likelihood Estimation: The Cox model estimates the regression coefficients for the covariates using a partial likelihood approach. This means that it maximizes the likelihood of observing events for individuals who experienced the event, adjusting for the individuals who are still at risk (have not experienced the event) at each time point.

-Interpretation: The estimated coefficients in the Cox model represent the log hazard ratios. A hazard ratio greater than 1 indicates an increased risk (positive effect) associated with a covariate, while a hazard ratio less than 1 indicates a decreased risk (negative effect).

-Non-Parametric Baseline Hazard: The Cox model does not assume a specific distribution for the baseline hazard function. Instead, it estimates the hazard ratios while leaving the baseline hazard unspecified. This makes it flexible and applicable to a wide range of data.

-Censored Data: Like other survival analysis methods, the Cox model can handle censored data, where not all events have occurred by the end of the study.

-Model Validation: It's important to assess the proportional hazards assumption and the overall fit of the model, typically through graphical methods and statistical tests

**What are the assumptions of this model?**

-Proportional Hazards Assumption: This is the central assumption of the Cox model. It states that the hazard ratios for the covariates are constant over time. In other words, the effect of each covariate on the hazard of an event is the same at any point in time. Violation of this assumption can lead to biased parameter estimates and incorrect conclusions.

-Independence of Censoring: The model assumes that censoring is independent of the event occurrence and covariates. In other words, the reason for censoring should not be related to the survival time or the covariates. For example, if individuals with more severe disease are more likely to drop out of a study, this could violate the assumption.

-Linearity in Log-Hazard: The Cox model assumes a linear relationship between the logarithm of the hazard and the covariates. This implies that the log hazard increases or decreases linearly with changes in the covariate values. If this assumption is violated, the model may not adequately capture the true relationship.

-No Measurement Error: The model assumes that there is no measurement error in the covariates. In reality, measurement errors can occur and can lead to biased parameter estimates if not properly addressed.

-Non-Interaction: The model assumes that there are no interactions between covariates in their effects on the hazard. In other words, the impact of each covariate is assumed to be independent of the other covariates in the model. Including interaction terms in the model can relax this assumption when necessary.

-No Omitted Variables: The Cox model assumes that all relevant covariates are included in the model. Omitting important covariates can lead to omitted variable bias and incorrect conclusions.

-Proper Model Specification: The model assumes that the functional form of the covariate effects is correctly specified. Using the wrong functional form can result in biased parameter estimates.

-Constant Hazard Ratios over Time: The proportional hazards assumption implies that the hazard ratios do not change over time. If the hazard ratios vary over time (non-proportional hazards), alternative modeling techniques may be more appropriate.

**Sixth: The survival analysis and cumulative distribution**

Survival analysis and the cumulative distribution function (CDF) are closely related concepts in statistics, especially when dealing with time-to-event data.

Survival Function (S(t)): In survival analysis, the survival function represents the probability that an individual or item will survive beyond a specific time point, t. It's denoted as S(t). Mathematically, S(t) = P(T > t), where T is the random variable representing the time to an event. In other words, S(t) gives you the probability that an event has not occurred by time t.

Cumulative Distribution Function (CDF) of T (F(t)): The CDF of a random variable T gives you the probability that T is less than or equal to a specific value, t. It's denoted as F(t). Mathematically, F(t) = P(T ≤ t). The CDF provides the cumulative probability distribution of the time-to-event variable.

The survival function and the CDF are complementary to each other.

**Mathematically, there's a direct relationship between the survival function and the CDF: S(t) = 1 - F(t). Which means the sum of survival function and cdf = 1**

**Seventh: The difference between survival analysis and regression analysis**

**Type of Data:**

Survival Analysis: Survival analysis is used for time-to-event data, where the primary interest is in understanding the time it takes for an event of interest to occur. This event could be, for example, patient survival, equipment failure, disease recurrence, or time to a specific outcome.

Regression Analysis: Regression analysis is used to model the relationship between one or more independent variables (predictors or covariates) and a dependent variable (the outcome or response). It is typically applied to continuous or categorical outcomes and does not specifically focus on time-to-event data.

**Outcome Variable:**

Survival Analysis: The primary outcome variable in survival analysis is the survival time or time-to-event, often represented as a random variable T. The goal is to estimate the survival function, hazard function, or other time-dependent measures.

Regression Analysis: In regression analysis, the outcome variable can take various forms, including continuous (e.g., linear regression), binary (e.g., logistic regression), or count (e.g., Poisson regression). The goal is to model and understand the relationship between predictors and the outcome variable.

**Focus on Time:**

Survival Analysis: Survival analysis explicitly accounts for time and is used to estimate quantities such as survival curves, hazard rates, and median survival times. It is particularly suited for analyzing data where the timing of events is critical.

Regression Analysis: While regression models can include time as a predictor, they do not inherently focus on modeling time-dependent outcomes and are more general in their application.

**Finally: Conclusion**

In conclusion, survival analysis is a powerful statistical methodology used to analyze time-to-event data, making it invaluable for understanding and predicting outcomes in various fields, including medicine, engineering, and finance. It accounts for censored data, models time-dependent risks, and provides critical insights into the dynamics of events over time. By estimating survival probabilities, hazard rates, and comparing survival curves, survival analysis equips researchers and practitioners with the tools to make informed decisions and improve outcomes for individuals and systems.

**Survival Analysis on Employee Turnover**

**Introduction**

Employee turnover is the percentage of employees that leave your organization during a given time period. Organizations typically calculate turnover rates annually or quarterly. Employee turnover is a crucial metric for measuring the performance of human resources departments or human resource management apps.

**Why is survival analysis important on employee turnover?**

Survival analysis can be of significant importance in studying employee turnover for several reasons:

-Time-to-Event Analysis: Employee turnover is an event-based phenomenon. Survival analysis allows HR professionals and researchers to analyze the time it takes for employees to leave the organization. This is crucial because it provides insights into the timing and patterns of turnover.

-Identifying Risk Factors: Survival analysis helps identify the factors or variables that contribute to employee turnover. By modeling the survival time as a function of various predictors (e.g., job satisfaction, salary, tenure), organizations can pinpoint which factors have the most significant impact on turnover rates.

And more.

**More Information about the dataset I chose**

This Employee Turnover dataset is a real dataset shared from Edward Babushkin's blog used to predict an Employee's risk of quitting (with a Survival Analysis Model). Edward Babushkin is an HR Analyst from Moscow, Russia. This data is for a specific company but its name is not mentioned.

Now I will list the name of the variables with their explanations:

1-stag - Experience (time)

2-event - Employee turnover : A binary variable (0 and 1) , a status of 1 signifies that the event has occurred, while a status of 0 indicates that the event has not occurred within the specified observation period

3-gender - Employee's gender, female(f), or male(m)

4-age - Employee's age (year)

5-industry - Employee's Industry

6-profession - Employee's profession

7-traffic - From what pipelene employee came to the company. You contacted the company directly (after learning from advertising, knowing the company's brand, etc.) - advert You contacted the company directly on the recommendation of your friend - NOT an employee of this company-recNErab You contacted the company directly on the recommendation of your friend - an employee of this company - referal You have applied for a vacancy on the job site - youjs The recruiting agency brought you to the employer - KA Invited by the Employer, we knew him before the employment - friends The employer contacted you on the recommendation of a person who knows you - rabrecNErab The employer reached you through your resume on the job site - empjs

8-coach - Presence of a coach (training) on probation

9-head\_gender - head (supervisor) gender

10-greywage - The salary does not seem to the tax authorities. Greywage in Russia or Ukraine means that the employer (company) pay

11-way - Employee's way of transportation

12-extraversion - Extraversion score

13-independ - Independend score

14-selfcontrol - Selfcontrol score

15-anxiety - Anxiety score

16-novator - Novator score

**The sample size is 1129 (we have 1129 rows)**

***Descriptive Statistics***

**First: The categorical variables**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **event** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 0 | 558 | 49.4 | 49.4 | 49.4 |
| 1 | 571 | 50.6 | 50.6 | 100.0 |
| Total | 1129 | 100.0 | 100.0 |  |

Figure 1: frequency table for the event

Comment: There are 558 employees who did not left the company (no turnover) with percentage 49.4 percent while 571 employees left the company (turnover) with percentage 50.6 percent

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **gender** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | f | 853 | 75.6 | 75.6 | 75.6 |
| m | 276 | 24.4 | 24.4 | 100.0 |
| Total | 1129 | 100.0 | 100.0 |  |

Figure 2: frequency table for the gender

Comment: There are 853 females employees in the company with 75.6 percent while 276 males in the company with 276 males employees in the company with 24.4 “percentage of females is much bigger”

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **industry** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Agriculture | 15 | 1.3 | 1.3 | 1.3 |
| Banks | 114 | 10.1 | 10.1 | 11.4 |
| Building | 41 | 3.6 | 3.6 | 15.1 |
| Consult | 74 | 6.6 | 6.6 | 21.6 |
| etc | 94 | 8.3 | 8.3 | 29.9 |
| HoReCa | 11 | 1.0 | 1.0 | 30.9 |
| IT | 122 | 10.8 | 10.8 | 41.7 |
| manufacture | 145 | 12.8 | 12.8 | 54.6 |
| Mining | 24 | 2.1 | 2.1 | 56.7 |
| Pharma | 20 | 1.8 | 1.8 | 58.5 |
| PowerGeneration | 38 | 3.4 | 3.4 | 61.8 |
| RealEstate | 13 | 1.2 | 1.2 | 63.0 |
| Retail | 289 | 25.6 | 25.6 | 88.6 |
| State | 55 | 4.9 | 4.9 | 93.4 |
| Telecom | 36 | 3.2 | 3.2 | 96.6 |
| transport | 38 | 3.4 | 3.4 | 100.0 |
| Total | 1129 | 100.0 | 100.0 |  |

Figure 3 : frequency table for industry

Comment: I can see the highest percentage of employees in the retail industry with total 289 employees with 25.6 percent while the least percentage of employees in the HoReCa industry- HoReCa is a short for hotel / restaurant / catering and encompasses the whole food service industry- with total 11 employees with 1 percent (all of this within the company)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **profession** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | Accounting | 10 | .9 | .9 | .9 |
| BusinessDevelopment | 27 | 2.4 | 2.4 | 3.3 |
| Commercial | 23 | 2.0 | 2.0 | 5.3 |
| Consult | 25 | 2.2 | 2.2 | 7.5 |
| Engineer | 15 | 1.3 | 1.3 | 8.9 |
| etc | 37 | 3.3 | 3.3 | 12.1 |
| Finanñe | 17 | 1.5 | 1.5 | 13.6 |
| HR | 757 | 67.1 | 67.1 | 80.7 |
| IT | 74 | 6.6 | 6.6 | 87.2 |
| Law | 7 | .6 | .6 | 87.9 |
| manage | 22 | 1.9 | 1.9 | 89.8 |
| Marketing | 31 | 2.7 | 2.7 | 92.6 |
| PR | 6 | .5 | .5 | 93.1 |
| Sales | 66 | 5.8 | 5.8 | 98.9 |
| Teaching | 12 | 1.1 | 1.1 | 100.0 |
| Total | 1129 | 100.0 | 100.0 |  |

Figure 4 : frequency table for profession

Comment: I can see the highest percentage of profession in this country is HR , 757 employees with 67.1 percent while the least percentage of profession is PR , 6 employees with 0.5 percent

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **traffic** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | advert | 33 | 2.9 | 2.9 | 2.9 |
| empjs | 248 | 22.0 | 22.0 | 24.9 |
| friends | 118 | 10.5 | 10.5 | 35.3 |
| KA | 67 | 5.9 | 5.9 | 41.3 |
| rabrecNErab | 211 | 18.7 | 18.7 | 60.0 |
| recNErab | 39 | 3.5 | 3.5 | 63.4 |
| referal | 95 | 8.4 | 8.4 | 71.8 |
| youjs | 318 | 28.2 | 28.2 | 100.0 |
| Total | 1129 | 100.0 | 100.0 |  |

Figure 5: frequency table for traffic

Comment: I can see the highest percentage for traffic 318 persons with 28.2 percent is youjs - youjs The recruiting agency brought you to the employer- while the least percentage 33 persons with 2.9 percent is advert - advert You contacted the company directly on the recommendation of your friend-

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **coach** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | my head | 314 | 27.8 | 27.8 | 27.8 |
| no | 683 | 60.5 | 60.5 | 88.3 |
| yes | 132 | 11.7 | 11.7 | 100.0 |
| Total | 1129 | 100.0 | 100.0 |  |

Figure 6: frequency table for coach

Comment: the highest percentage of employees have -no present for coach- 683 employees with 60.5 percent while the least percentage 132 employees have -yes present for coach- with 11.7 percent

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **head\_gender** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | f | 545 | 48.3 | 48.3 | 48.3 |
| m | 584 | 51.7 | 51.7 | 100.0 |
| Total | 1129 | 100.0 | 100.0 |  |

Figure 7: frequency table for head\_gender

Comment: The lowest percentage of head gender for each employee in the company is females 545 employees with 48.3 percent while the highest percentage of head gender for each employee is males employees 584 with 51.7 percent - head\_gender :head (supervisor) gender-

Figure 8: frequency table for greywage

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **greywage** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | grey | 127 | 11.2 | 11.2 | 11.2 |
| white | 1002 | 88.8 | 88.8 | 100.0 |
| Total | 1129 | 100.0 | 100.0 |  |

Comment: greywage - The salary does not seem to the tax authorities. Greywage in Russia or Ukraine means that the employer (company) pay , I can see the grey- The salary does not seem to the tax authorities - with least percentage 11.2 percent (127 employees) while white with highest percentage 88.8 percent (1002 employees)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **way** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | bus | 681 | 60.3 | 60.3 | 60.3 |
| car | 331 | 29.3 | 29.3 | 89.6 |
| foot | 117 | 10.4 | 10.4 | 100.0 |
| Total | 1129 | 100.0 | 100.0 |  |

Figure 9: frequency table for the way employees come to work

Comment: I can see that the highest percentage of employees use bus to work (681 employees) with 60.3 percent while the least percentage of employees come to work on foot (117 employees) with 10.4 percent

**Second : the quantitative variables**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | |
|  | N | Minimum | Maximum | Mean | Std. Deviation |
| stag | 1129 | .39 | 179.45 | 36.6275 | 34.09660 |
| age | 1129 | 18.00 | 58.00 | 31.0670 | 6.99615 |
| extraversion | 1129 | 1.00 | 10.00 | 5.5924 | 1.85164 |
| independ | 1129 | 1.00 | 10.00 | 5.4780 | 1.70331 |
| selfcontrol | 1129 | 1.00 | 10.00 | 5.5973 | 1.98010 |
| anxiety | 1129 | 1.70 | 10.00 | 5.6656 | 1.70918 |
| novator | 1129 | 1.00 | 10.00 | 5.8796 | 1.90402 |
| Valid N (listwise) | 1129 |  |  |  |  |

Figure 10: descriptives for quantitative variables

Comment: the stag “time variable” has a minimum of 0.39 and maximum 179.45 with average 36.63 units of time and standard deviation 34.1 , the age variable has a minimum of 18 ad maximum of 58 with average 36.6 units and standard deviation 7 , the extraversion variable has minimum of 1 and maximum 10 with average 5.6 and standard deviation 1.9 , the independ variable has minimum 1 an d maximum 10 with average 5.5 and standard deviation 1.7 the selfcontrol variable has a minimum of 1 and maximum of 10 with average 5.6 and standard deviation 1.99 , the anxiety variable has a minimum of 1.7 and maximum 10 with average 5.7 and standard deviation 1.7 , the novator variable has a minimum of 1 and maximum 10 with average 5.9 and standard deviation 1.9

***Survival Analysis***

First: Kaplan-Meier

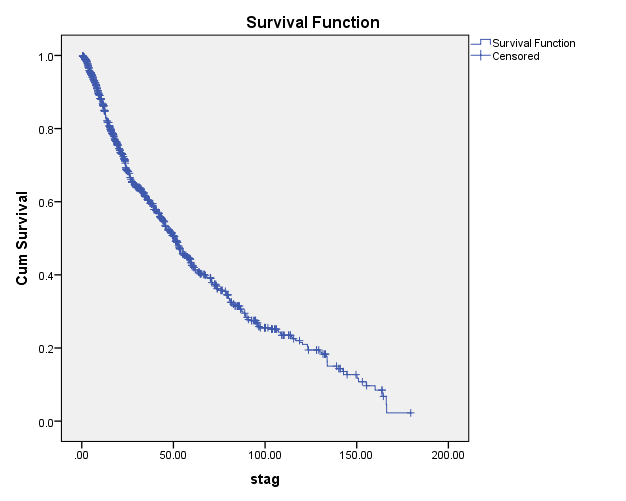


Figure 11: Kaplan Meier Curve for the data

Comment: This curve shows the cumulative survival probabilities; we can see that the curve is steeper which indicates a very high rate of employees’ turnover within the company as seen it falls from 1 fastly till it reaches zero we can see here in the curve that 50 percent (median more than 0.5) last is likely that the employees will last in their jobs 60 units of time

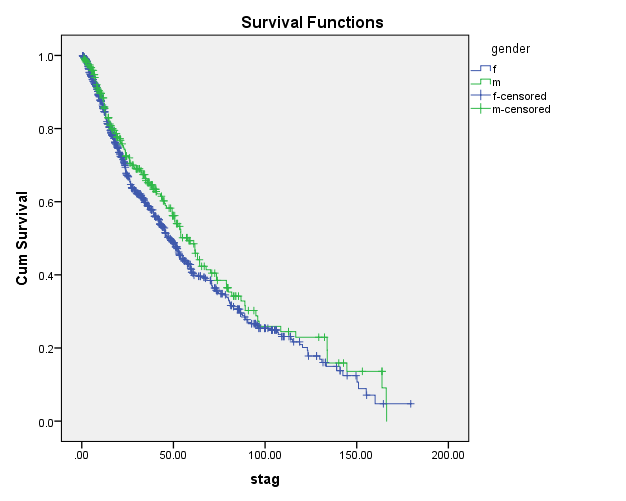


Figure 12: log rank curve between males and females

Comment: Here the survival probabilities for males and the survival probabilities for females are crossed at most times except for a small period the survival probability for males was higher than the survival probabilities for females; therefore we can assume no significant difference.

|  |  |  |  |
| --- | --- | --- | --- |
| **Overall Comparisons** | | | |
|  | Chi-Square | df | Sig. |
| Log Rank (Mantel-Cox) | 2.348 | 1 | .125 |
| Figure 13: log rank test between males and females | | | |

Comment: There is no significant difference between the 2 groups males and females (the highlighted number); we can see the test agrees with the curve

Second: Cox model

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | | |
|  | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
| Lower | Upper |
| gender | .126 | .106 | 1.425 | 1 | .233 | 1.135 | .922 | 1.397 |
| age | .022 | .006 | 12.034 | 1 | .001 | 1.022 | 1.010 | 1.035 |
| head\_gender | -.055 | .088 | .391 | 1 | .532 | .946 | .796 | 1.125 |
| anxiety | -.041 | .026 | 2.525 | 1 | .112 | .960 | .912 | 1.010 |

Figure 14: Cox model for 4 variables in interest

Comment: I care in my analysis about these 2 categorical variables gender and head gender and 2 quantitative variables age and anxiety; from this model we can see that only 1 variable that is significant (the highlighted number) which is the age only while the other 3 variables are insignificant so I will remove them and re form the model.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | | |
|  | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
| Lower | Upper |
| age | .022 | .006 | 13.036 | 1 | .000 | 1.022 | 1.010 | 1.035 |

Figure 15: Cox model (final model)

Comment: So it is clear now that the age variable is significant and hence this will be our model

I can see the hazard ratio 1.022 is greater than 1 which means age is associated with positive slopes ; for every increase in 1 unit in age the hazard of turnover increase by 22% as seen B=0.22

Now I will make proportional hazard rate (check the assumptions)

Proportional hazards assumptions are a fundamental aspect of survival analysis, particularly in Cox proportional hazards regression. These assumptions are essential for the validity and reliability of the results obtained from this statistical method. The Cox proportional hazards model assumes that the hazard rates of different groups being compared are proportional over time.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | |
|  | B | SE | Wald | df | Sig. | Exp(B) |
| age | .026 | .009 | 8.589 | 1 | .003 | 1.026 |
| T\_COV\_\*age | .000 | .000 | .348 | 1 | .555 | 1.000 |

|  |  |
| --- | --- |
| **Covariate Means** | |
|  | Mean |
| T\_COV\_ | 19.453 |
| age | 30.334 |
| T\_COV\_\*age | 577.842 |

Figure 16&17: Proportional hazard rates

Comment: Null Hypothesis (H0): The hazard ratios for the covariates are constant over time. In other words, there is no violation of the proportional hazards assumption. (assumptions validated)

Alternative Hypothesis (H1): The hazard ratios for the covariates are not constant over time. This implies that the proportional hazards assumption is violated. (assumptions not validated)

It is clear here the p value for T COV is 0.348 which is more than 0.05 , therefore do not reject. Hence, we can assume that the assumptions are validated.

**Finally: Conclusion**

In summary, this survival analysis conducted in SPSS provided valuable insights into the topic of employee turnover. My results suggest that the age variable is a major contributor here in the model. These findings contribute to our understanding of the analysis of this model.

This analysis serves as a foundation for further research and underscores the importance of employee turnover analysis. We hope that this study will inform decision-making.

Finally, survival analysis is a powerful tool for studying employee turnover patterns, identifying risk factors, and developing effective retention strategies. By applying this methodology, organizations can make data-driven decisions to reduce turnover, improve employee satisfaction, and enhance overall workforce stability and productivity.