Title: Survival **Survival analysis: A Comprehensive Overview and Applications**

**Introduction**Survival analysis is a statistical method used to analyze time-to-event data, where the event of interest could be death, failure, recovery, or any other occurrence. It is widely employed in various fields such as medicine, finance, engineering, and social sciences. This report provides a comprehensive overview of survival analysis, its key concepts, and its applications in different domains.

**Key Concepts in Survival Analysis**- Survival Time and Survival Function  
Survival time refers to the time it takes for an event to occur. The survival function, also known as the survival probability, estimates the probability that an individual or a group survives beyond a certain time point.

- Hazard Function  
The hazard function, also called the instantaneous failure rate, characterizes the probability of experiencing the event of interest at a given time, given that the individual or group has survived up to that time. It is a fundamental concept in survival analysis and provides insights into the rate of event occurrence.

- Censoring  
Censoring occurs when the event of interest is not observed for some individuals within the study period. It can be either right-censoring, left-censoring, or interval-censoring. Various statistical methods are available to handle censored data appropriately.

- Survival Distributions and Curves  
Survival distributions, such as the Kaplan-Meier estimator and parametric models like Weibull, exponential, and log-normal, describe the probability distribution of survival times. Survival curves visually represent the estimated survival probabilities over time based on the available data.

**Types of Survival Analysis**  
- Nonparametric Survival Analysis  
Nonparametric methods, such as the Kaplan-Meier estimator and the log-rank test, make minimal assumptions about the underlying distribution and provide valuable insights into survival probabilities between different groups or treatments.

- Cox Proportional Hazards Model  
The Cox proportional hazards model is a widely used semi-parametric model that accounts for multiple covariates affecting survival. It estimates hazard ratios to quantify the impact of predictors on survival while assuming a proportional hazards assumption.

**Comparison between cox proportional hazards, logistic regression, and simple linear regression:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Cox regression** | **Logistic regression** | **Simple linear regression** |
| **Outcome** | T=Time to event, continuous, positive | Y=indicator of event, binary (0,1): yes, no | Continuous variable |
| **What we model** | Hazard ratio | Odds ratio | Slope and intercept |
| **units** | time | unitless | Depends on continuous variable |
| **Interpretation in terms of** | Hazard ratios (e β), between two groups (after controlling for other covariates) | Odds ratios (e β), between two groups (after controlling for other covariates) | The mean of a continuous variable. |
| **Type of model** | Semiparametric Fully parametric of model Form of baseline hazard Form of (log) odds (ho(t)) not specified fully specified through β’s Estimated only hazard ratios between reference and other groups. | Fully parametric of model Form of baseline hazard Form of (log) odds (ho(t)) not specified fully specified through β’s | Parametric model that assumes linear relationship between predictors and the outcome. |
| **Assumptions** | -Independent observations -Independent observations Censoring independent of time to event  -Proportional hazard (rates) = hazard ratio between two groups constant over time | -Independent observations. | -normality  -homoscedasticity  -linearity  -autocorrelation |

**Applications of Survival Analysis**  
- Medical Research  
Survival analysis plays a crucial role in clinical trials, studying disease progression, and evaluating treatment effectiveness. It helps estimate survival rates, identify risk factors, and determine optimal treatment strategies.

- Financial Risk Assessment  
Survival analysis is valuable in assessing credit risk, predicting loan default, and estimating the lifetime of financial products. It aids in modeling event probabilities and determining appropriate risk mitigation strategies.

- Engineering and Reliability Analysis  
Survival analysis is used in engineering to evaluate the reliability and lifetime of mechanical systems, analyze failure modes, and optimize maintenance strategies. It enables proactive decision-making to prevent system failure.

- Social Sciences and Event History Analysis  
Survival analysis is applied in social sciences to study various events such as marriage, divorce, unemployment, and migration. It helps identify factors influencing event occurrence and estimate the duration until event realization.

**Conclusion**  
Survival analysis provides a powerful framework for analyzing time-to-event data, enabling researchers to understand and predict event occurrences. Its applications span across diverse fields, contributing to improved decision-making processes, risk assessment, and understanding of event dynamics. As more sophisticated techniques and data become available, survival analysis continues to evolve and find new applications in an increasingly complex world.

**Survival analysis on breast cancer**

**Introduction:**

This analysis will be used to do a study on breast cancer.

**Data**:272 breast cancer patients (as rows), 1570 columns. Network built using only gene expression. Meta data includes patient info, treatment, and survival.

Each node is a group of patients like each other. Flares (left) represent sub-populations that are distinct from the larger population. (One differentiating factor between the two flares is estrogen expression (low = top flare, high = bottom flare)). Bottom flare is a group of patients with 100% survival. Top flare shows a range of survival – very poor towards the tip (red), and very good near the base (circled).

The circled group of good survivors have genetic indicators of poor survivors (i.e., low ESR1 levels, which is typically the prognostic indicator of poor outcomes in breast cancer) – understanding this group could be critical for helping improve mortality rates for this disease. Why this group survived was quickly analyzed by using the Outcome Column (here Event Death - which is binary - 0,1) as a Data Lens (which we term Supervised vs Unsupervised analyses).

***1)Descriptive statistics***

First for the categorical variables :

Table 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **eventdeath** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 0 | 195 | 71.7 | 71.7 | 71.7 |
| 1 | 77 | 28.3 | 28.3 | 100.0 |
| Total | 272 | 100.0 | 100.0 |  |

Comment:

* There are 195 persons didn’t died or censored with a percentage 71.7%, and there are 77 people who died with probability 28.3% .

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Chemo table 2** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 0 | 165 | 60.7 | 60.7 | 60.7 |
| 1 | 107 | 39.3 | 39.3 | 100.0 |
| Total | 272 | 100.0 | 100.0 |  |

Comment:

* There are 165 persons who didn’t take the chemotherapy with percentage 60.7% and there are 107 people didn’t take chemotherapy with percentage 39.3%.

Second for continuous and discrete variables :

Table 3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Descriptive Statistics** | | | | | | |
|  | N | Minimum | Maximum | Mean | Std. Deviation | Variance |
| age | 272 | 26 | 53 | 44.05 | 5.465 | 29.861 |
| timerecurrence | 272 | .271047 | 18.340862 | 7.25043254 | 4.177461512 | 17.451 |
| esr1 | 272 | -1.511655 | .596177 | -.25304418 | .559309059 | .313 |
| G3PDH\_570 | 272 | -1.951243000000000 | 1.897414000000000 | -.025153838235294 | .626507299118175 | .393 |
| Valid N (listwise) | 272 |  |  |  |  |  |

Comment:

* For the age variable, the mean age of people in the data set is 44.05, on average , the people in the data set have an age of 44.05. as for the standard deviation, a 5.465 indicated a large variance which means that the age values are more spread out from the mean. The maximum age in the data set is 53 and minimum age is 26.
* For the time recurrence variable, the mean is 7.25 which means that on average, individuals experience recurrence after approximately 7.25 unit in time . as for the standard deviation, 4.177 which means that the recurrence time are more spread out from the mean . the maximum recurrence time is 18.3 units in time and the minimum recurrence time is 0.271 units in time .
* For esr1 variable , it has a mean of -0.253 which means that on average the data set has ESR1 level of -0.253 . and as for the standard deviation 0.559 which is a small value indicates less variability or more tightly clustered distribution of ESR1 level values around the mean . the max ESR1 value in the data set is 0.5961 and the minimum value is -1.5.
* For the G3PDH\_570 variable , a mean of 0.025 indicates that on average the data set has a G3PDH\_570 level of 0.025. as for standard deviation 0.626 which is a small value indicates less variability or more tightly clustered distribution of G3PDH\_570 level values around the mean. The maximum value of G3PDH\_570 is 1.89 and the minimum value is -1.95.

***2) survival analysis***

**First Caplan-Meier:**

**Graph 1**

A graph with a line going up

Description automatically generated

* We can see from the above graph that people with cumulative survival rate higher than 0.5 are most likely to live, also we can notice that the curve is decreasing slightly which means that people are dying in small rate until they reached a cumulative survive ratio of 0.7, they became fixed.
* A graph with a line graph

  Description automatically generated with medium confidence

Graph 2

* we can see from the graph above that both curves are above the median which means that people had high survival rate for both groups (chemo and no chemo ) but we can notice that the group of chemo had slightly higher survival rates . and to make sure whether this difference is significant or not we will look at the log rank test.

Table 4

|  |  |  |  |
| --- | --- | --- | --- |
| **Overall Comparisons** | | | |
|  | Chi-Square | df | Sig. |
| Log Rank (Mantel-Cox) | 1.349 | 1 | .245 |
| Test of equality of survival distributions for the different levels of chemo. | | | |

Comment:

We can notice from the log rank test that there is no significant difference between both groups.

***3)cox regression***

Table 5

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | | |
|  | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
| Lower | Upper |
| age | -.046 | .020 | 5.137 | 1 | .023 | .955 | .918 | .994 |
| chemo | -.228 | .244 | .878 | 1 | .349 | .796 | .493 | 1.283 |
| esr1 | -.957 | .190 | 25.238 | 1 | .000 | .384 | .264 | .558 |
| G3PDH\_570 | .053 | .191 | .078 | 1 | .780 | 1.055 | .726 | 1.532 |

We can see in this table that we have 2 insignificant variables (chemo and G3PDH\_570) and 2 significant ( age and esr1) so we will remove the 2 insignificant variables and re-regress the model.

Table 6

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation** | | | | | | | | |
|  | B | SE | Wald | df | Sig. | Exp(B) | 95.0% CI for Exp(B) | |
| Lower | Upper |
| age | -.046 | .020 | 5.134 | 1 | .023 | .955 | .919 | .994 |
| esr1 | -.975 | .190 | 26.307 | 1 | .000 | .377 | .260 | .548 |

Comment:

* After removing the 2 insignificant variables we computed the model in table 6.
* For the age variable, the hazard ratio is 0.955 which is <1 and the confidence interval is entirely less than 1, this indicates for every 1 unit increase in age, the hazard of death decreases by approximately 4.5% (HR=0.955).
* For esr1 variable, the hazard ratio is 0.377 which is <1 and the confidence interval is entirely less than 1, this indicates for every 1 unit increase in esr1, the hazard of death decreases by approximately 62.3% (HR=0.955).

**Checking the PH assumption (proportional hazard rates) :**

In order for the model to be reliable we need to check the PH assumption and we will use SPSS for that . we will check it for each variable alone.

For the esr 1 variable :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation table 7** | | | | | | |
|  | B | SE | Wald | df | Sig. | Exp(B) |
| T\_COV\_ | .042 | .078 | .293 | 1 | .588 | 1.043 |
| esr1 | -1.139 | .304 | 14.058 | 1 | .000 | .320 |

* For the T\_COV significance is above 0.05 then the proportion hazard ratio is satisfied in the esr1 case.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation table 8** | | | | | | |
|  | B | SE | Wald | df | Sig. | Exp(B) |
| T\_COV\_ | .017 | .009 | 4.159 | 1 | .041 | 1.018 |
| age | -.109 | .032 | 11.554 | 1 | .001 | .897 |

* Here we can see that the significance of age is less than 0.05 so the hazard ratio assumption isn’t satisfied here.

So we will regress time only on esr1 level , and our final model will be :

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Variables in the Equation table 9** | | | | | | |
|  | B | SE | Wald | df | Sig. | Exp(B) |
| esr1 | -1.009 | .187 | 28.992 | 1 | .000 | .365 |

* For esr1 variable, the hazard ratio is 0.365 which is <1 and the confidence interval is entirely less than 1, this indicates for every 1 unit increase in esr1, the hazard of death decreases by approximately 63.5% (HR=0.365).