Lung Cancer Survival Analysis

Jesús Veiga Morandeira

2024-10-10

Introduction

A database provided by the R package survival is going to be analysed in this document. The specific database is called lung. It contains 228 observations of patients with advanced lung cancer from the North Central Cancer Treatment Group.

This database contains the following columns:

- *inst*: Institution code
- *time*: Survival time in days
- *status*: censoring status 1=censored, 2=dead
- *age*: Age in years
- *sex*: Male=1 Female=2
- *ph.ecog*: ECOG performance score as rated by the physician
- *karno*: Karnofsky performance score (bad=0-good=100) rated by physician
- *pat.karno*: Karnofsky performance score as rated by patient
- *meal.cal*: Calories consumed at meals
- *wt.loss*: Weight loss in last six months (pounds)

It will start by making an EDA of the data. Exploratory Data Analysis (EDA) is a crucial step in any analysis project, as it helps to better understand the structure and characteristics of the dataset. In this analysis, we will focus on various aspects of the data, including patient status, sex, and age, to identify patterns that may influence subsequent analyses.

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in any analysis project, as it helps to better understand the structure and characteristics of the dataset. In this analysis, we will focus on various aspects of the data, including patient status, sex, and age, to identify patterns that may influence subsequent analyses.

Status Analysis

First, we will examine the status variable to understand its distribution in the dataset. This will give us an idea of how many patients are alive compared to those who have passed away.

```r
# Count of statuses
status_counts <- SQ_survival_data %>%
  count(status)
```

Table 1: Count of Survival Statuses

status	n
1	63
2	165

Table 2: Proportions of Survival Statuses

status	n	proportion
1	63	0.2763158
2	165	0.7236842

```r
# Proportions of status
status_proportions <- SQ_survival_data %>%
  count(status) %>%
  mutate(proportion = n / sum(n))

kable(status_counts, caption = "Count of Survival Statuses", escape = F) %>%
  kable_styling(full_width = F, position = "center") %>%
  row_spec(0, bold = TRUE, color = "black") %>%
  row_spec(1:nrow(status_counts), color = "black")
```

```r
kable(status_proportions, caption = "Proportions of Survival Statuses", escape = F) %>%
  kable_styling(full_width = F, position = "center") %>%
  row_spec(0, bold = TRUE, color = "black") %>%
  row_spec(1:nrow(status_proportions), color = "black")
```

Sex and Status Analysis

Next, we will explore the relationship between patients' sex and their status. This may help us identify any significant differences in survival based on gender.

```r
# Frequency table by sex and status
sex_table <- TQ_SurvivalData %>%
  count(sex, status) %>%
  pivot_wider(names_from = status, values_from = n, values_fill = list(n = 0))

kable(sex_table, caption = "Frequency table by sex and status", escape = F) %>%
  kable_styling(full_width = F, position = "center") %>%
  row_spec(0, bold = TRUE, color = "black") %>%
  row_spec(1:nrow(status_proportions), color = "black")
```

Table 3: Frequency table by sex and status

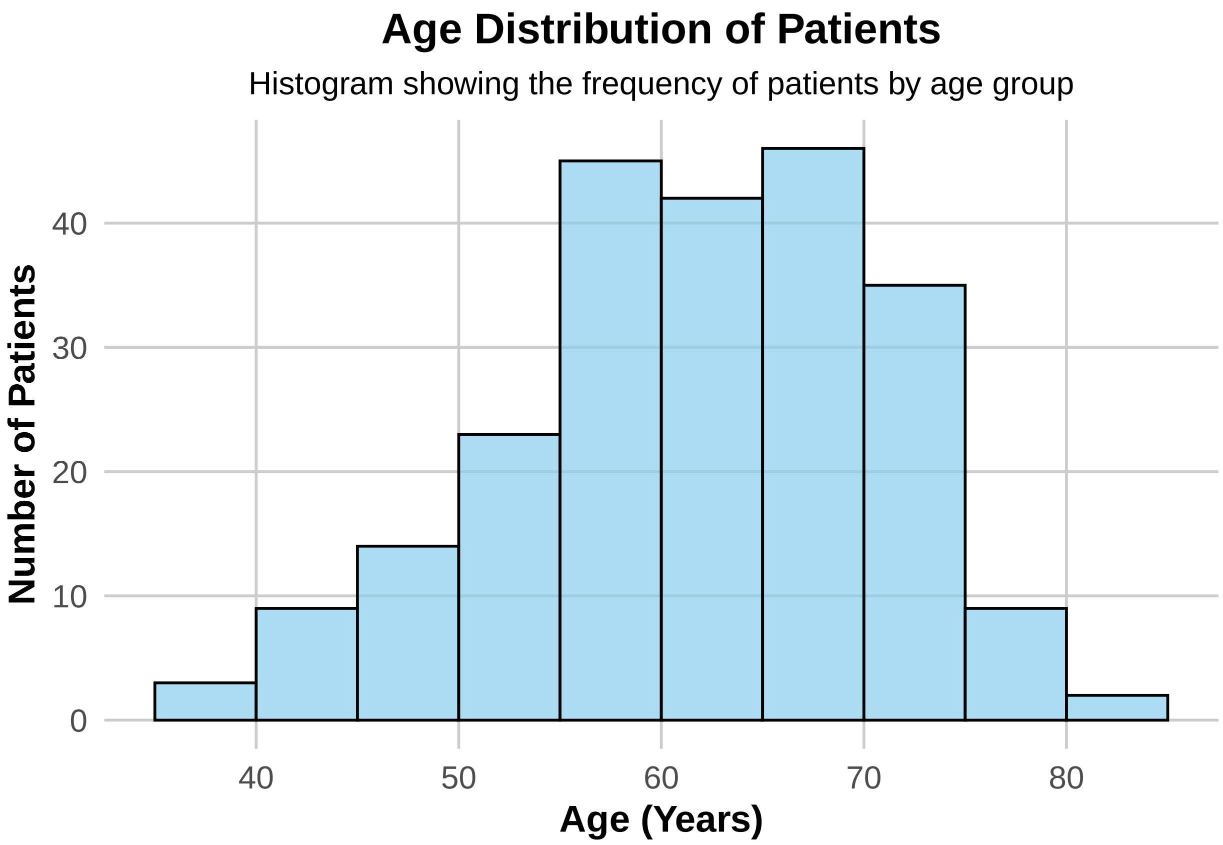
sex	1	2
1	26	112
2	37	53

Age Analysis

Age is a critical variable in survival studies. Next, we will visualize the distribution of age in the dataset using a histogram, which will help us identify the frequency of different age ranges.

```r
# Frequency table by sex and status
# Age histogram
ggplot(TQ_SurvivalData, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "#87CEEB", color = "black", boundary = 0, alpha = 0.7) +
  labs(title = "Age Distribution of Patients",
       subtitle = "Histogram showing the frequency of patients by age group",
       x = "Age (Years)",
       y = "Number of Patients") +
  theme_minimal(base_size = 14) +
  theme(panel.grid.major = element_line(color = "grey80", size = 0.5),
        panel.grid.minor = element_blank(),
        plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
        plot.subtitle = element_text(hjust = 0.5, size = 12),
        axis.title = element_text(face = "bold"),
        axis.text = element_text(size = 12),
        plot.background = element_rect(fill = "white", color = NA))
```

```
## Warning: The ‘size‘ argument of ‘element_line()‘ is deprecated as of ggplot2 3.4.0.
## i Please use the ‘linewidth‘ argument instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```



We will also create age groups to facilitate further analysis. Patients will be divided into age groups of 25 to 50 years, 50 to 70 years, and 70 to 90 years.

```
# Create age groups
TQ_SurvivalDataAgeGroups <- TQ_SurvivalData %>%
  mutate(age_group = cut(age, breaks = c(25, 50, 70, 90),
                         labels = c("25 to 50", "50 to 70", "70 to 90"),
                         right = TRUE))

# Frequency table for age groups
age_group_table <- TQ_SurvivalDataAgeGroups %>%
  count(age_group, status) %>%
  pivot_wider(names_from = status, values_from = n, values_fill = list(n = 0))


kable(age_group_table, caption = "Frequency table for age groups", escape = F,
      col.names = c("Age Group", "1", "2")) %>%
  kable_styling(full_width = F, position = "center") %>%
  row_spec(0, bold = TRUE, color = "black") %>%
  row_spec(1:nrow(status_proportions), color = "black")
```

This exploratory analysis has allowed us to better understand the demographic characteristics of our dataset. The distribution of status, the relationship between sex and status, as well as the age of patients, are factors that should be considered in subsequent analyses.

Table 4: Frequency table for age groups

Age Group	**1**	**2**
25 to 50	10	16
50 to 70	45	111
70 to 90	8	38

Survival Analysis: Introductory Theory and Kaplan-Meier Estimation

Survival analysis is a branch of statistics that deals with the analysis of time-to-event data. In many fields, particularly in medical research, it's crucial to analyze the time until a particular event occurs, such as death, disease recurrence, or failure of a treatment. The primary goal of survival analysis is to understand the survival function, which describes the probability that an event of interest has not occurred by a given time.

Survival Functions

Mathematically, the survival function, denoted as $S(t)$, represents the probability that the time to the event T is greater than some time t. This can be expressed as:

$$S(t) = P(T > t)$$

Conversely, the cumulative distribution function (CDF), $F(t)$, which provides the probability that the event has occurred by time t, is related to the survival function as follows:

$$F(t) = 1 - S(t)$$

In the context of survival analysis, we often encounter two important aspects:

1. **Censoring**: This occurs when we have incomplete information about the time to the event for some subjects. For instance, a patient may leave a study before the event occurs, or the study may end before the event is observed. In such cases, we only know that the event occurred after a certain time, but not exactly when.

2. **Hazard Function**: The hazard function, $h(t)$, describes the instantaneous risk of the event occurring at time t, given that it has not yet occurred. It can be defined as:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T \geq t)}{\Delta t}$$

This function is important as it helps to understand the underlying risk dynamics over time.

Non Parametric Survival Estimation. Kaplan-Meier Estimator

The Kaplan-Meier estimator is a non-parametric statistic used to estimate the survival function from observed survival data, particularly in the presence of censoring. The estimator is particularly useful when dealing with small sample sizes and when the exact distribution of the survival times is unknown.

Mathematically, the Kaplan-Meier estimator $\hat{S}(t)$ can be expressed as:

$$\hat{S}(t) = \prod_{i=1}^{k} \left(\frac{n_i - d_i}{n_i} \right)$$

where:

- k is the number of distinct time points where events occur,
- n_i is the number of individuals at risk just before time t_i,
- d_i is the number of events (e.g., deaths) that occur at time t_i.

The product continues until the last event time, and the Kaplan-Meier estimator provides a step function that allows for the visualization of survival probabilities over time.

Objectives of the Analysis

In this analysis, we aim to investigate how various factors, such as sex, age, and symptoms as measured by the Eastern Cooperative Oncology Group (ECOG) performance status, influence patient survival. By applying the Kaplan-Meier estimator, we can compare survival functions across these different groups and derive insights about the effects of these factors on survival rates.

Survival by Sex

A survival analysis stratified by sex was conducted, allowing us to observe differences in survival between men and women. The figure below shows the survival curve for each group. We are going to evaluate some different cases and for each of the cases, in order to test if there are significant differences between groups, we will use the graphical method and an analytical method, which in this case will be the Log-Rank.

The Log-Rank test is a statistical hypothesis test used to compare the survival distributions of two or more groups. It is particularly useful in survival analysis when analyzing the time until an event of interest occurs (e.g., death, relapse) and is applicable in the context of censored data.

This test compares the observed number of events (e.g., deaths) in each group to the expected number of events, assuming that the survival functions are equal across groups. The test is based on the following steps:

1. **Calculate the survival function**: For each group, the survival function $S(t)$ is estimated using the Kaplan-Meier method.

2. **Calculate the expected number of events**: For each time point where an event occurs, the expected number of events in each group is calculated based on the overall survival experience of all groups combined.

3. **Compute the test statistic**: The test statistic χ^2 for the Log-Rank test is calculated as follows:

   $$\chi^2 = \frac{(O - E)^2}{V}$$

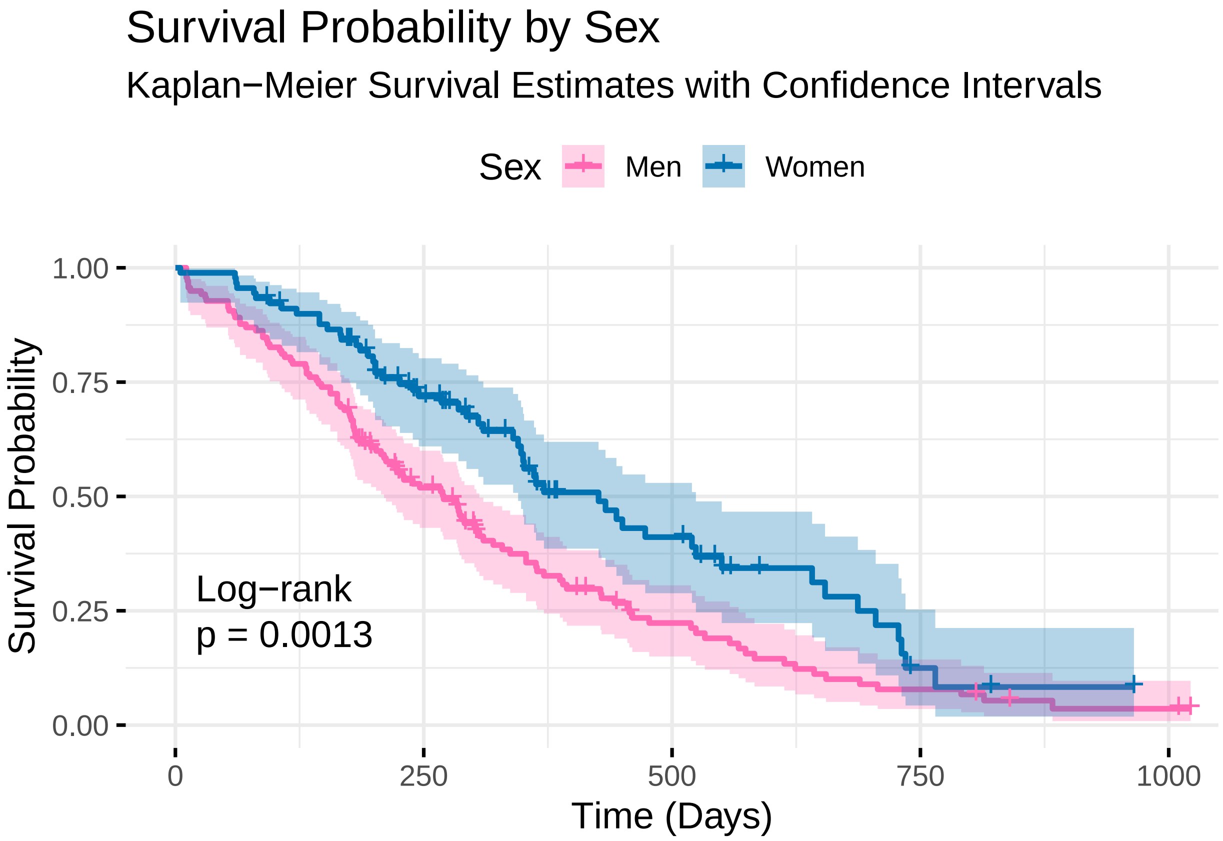
   Where:

   - O = observed number of events in the group,
   - E = expected number of events in the group,
   - V = variance of the observed number of events.

4. **Determine the p-value**: The test statistic is compared against a chi-squared distribution with degrees of freedom equal to the number of groups minus one. A small p-value (typically < 0.05) indicates a significant difference between the survival curves of the groups.

Graphical method

```
# Survival plot by sex

SurvPerSex <- survfit(Surv(time, status) ~ sex, TQ_SurvivalData, conf.type = "log-log")

ggsurvplot(SurvPerSex,
           title = "Survival Probability by Sex",
           subtitle = "Kaplan-Meier Survival Estimates with Confidence Intervals",
           xlab = "Time (Days)", ylab = "Survival Probability",
           conf.int = TRUE,
           legend.title = "Sex",
           legend.labs = c("Men", "Women"),
           palette = c("#FF69B4", "#0072B2"),
           pval = TRUE,
           pval.method = TRUE,
           ggtheme = theme_minimal(base_size = 14),
           theme = theme(plot.title = element_text(hjust = 0.5, face = "bold", size = 16),
                         plot.subtitle = element_text(hjust = 0.5, size = 12),
                         legend.position = "bottom",
                         legend.title = element_text(face = "bold"),
                         axis.title = element_text(face = "bold", size = 14),
                         axis.text = element_text(size = 12),
                         panel.grid.major = element_line(color = "grey90")))
```



### Analytical Method

```r
# Survival plot by sex

Test_0 <- survdiff(Surv(time, status) ~ sex, TQ_SurvivalData, rho = 1)
print(Test_0)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex, data = TQ_SurvivalData,
##     rho = 1)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=1 138     70.4     55.6      3.95      12.7
## sex=2  90     28.7     43.5      5.04      12.7
##
##  Chisq= 12.7  on 1 degrees of freedom, p= 4e-04
```

The Log-Rank test indicated a significant difference in survival between men and women, yielding a chi-squared statistic of 12.7 with 1 degree of freedom (df) and a p-value of 4e-04. The observed survival rates were 70.4% for men and 28.7% for women, with expected survival rates of 55.6% and 43.5%, respectively.

The (O-E)^2/E values, which represent the variance between observed and expected events, were 3.95 for men and 5.04 for women, further indicating the degree of deviation from what would be expected if there were no difference in survival between the groups. The significant p-value suggests strong evidence against the null hypothesis, leading us to conclude that sex is a critical factor influencing survival outcomes.

These findings underscore the necessity for personalized treatment strategies that account for sex differences and prompt further investigation into the biological and socio-economic factors contributing to these disparities.

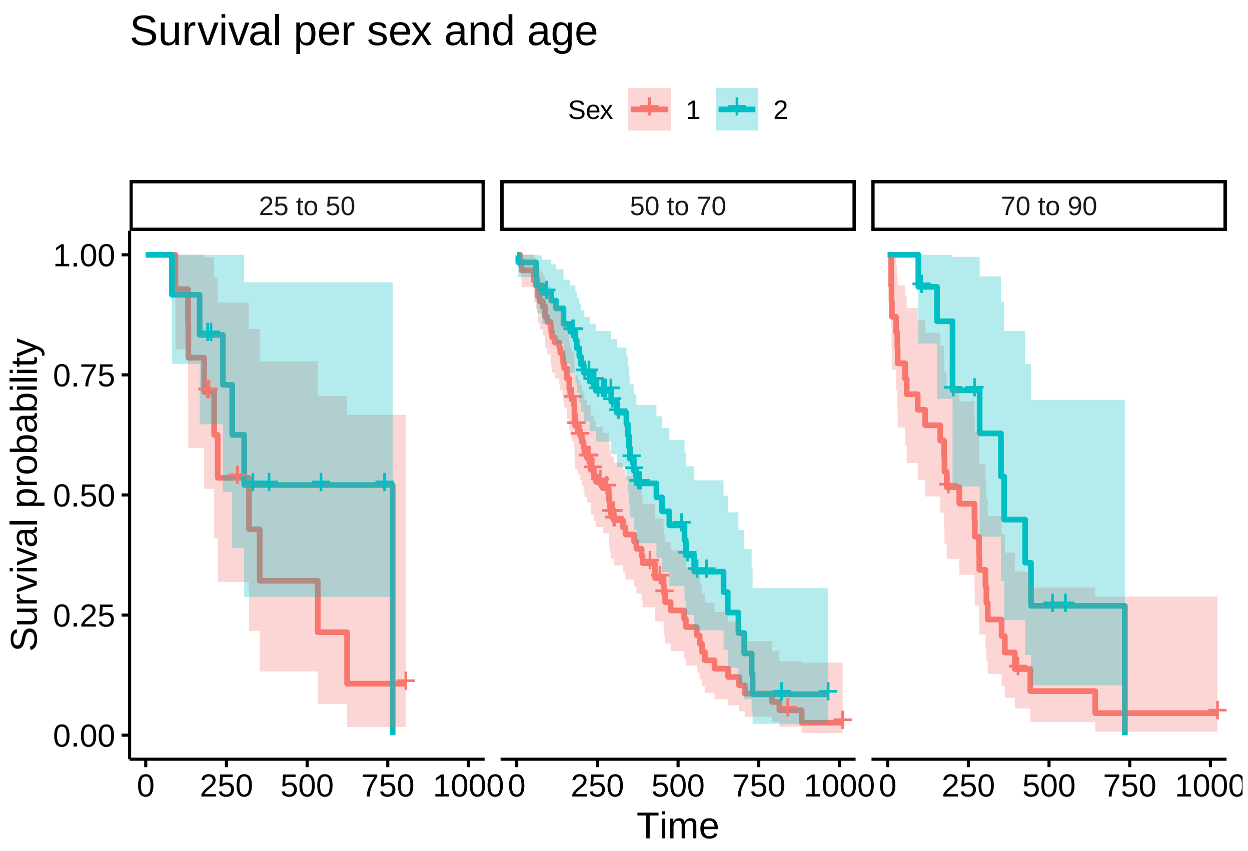
In conclusion, both methods indicates differences between genres.

We will continue to do more non-parametric studies with the Kapler-Meier function, although less detailed, following this first one as an example.

Survival by sex and age group

Graphical method

```r
# Survival plot by sex

SurvPerSexAndAge <- survfit(Surv(time, status) ~ sex + age_group, TQ_SurvivalDataAgeGroups, conf.type =
  ggsurvplot(title = "Survival per sex and age", conf.int = T,
             facet.by = "age_group", legend.title = "Sex", panel.labs = list(edadg = c("25 to 50", "50

print(SurvPerSexAndAge)
```



```
# Survival plot by sex

Test_1 <- survdiff(Surv(time, status) ~ sex + age_group, TQ_SurvivalDataAgeGroups, rho = 1)
print(Test_1)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex + age_group, data = TQ_SurvivalDataAgeGroups,
##     rho = 1)
##
##                               N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=1, age_group=25 to 50 14     5.96     6.08   0.00233   0.00346
## sex=1, age_group=50 to 70 93    44.70    39.13   0.79296   1.84233
## sex=1, age_group=70 to 90 31    19.71    10.36   8.45021  12.87699
## sex=2, age_group=25 to 50 12     3.48     5.97   1.03606   1.58708
## sex=2, age_group=50 to 70 63    19.94    30.44   3.61867   7.43569
## sex=2, age_group=70 to 90 15     5.30     7.13   0.46706   0.70805
##
##  Chisq= 20  on 5 degrees of freedom, p= 0.001
```

The Log-Rank test results for the combined analysis of sex and age groups indicate a significant difference in survival among the different subgroups. The test showed a chi-squared statistic of 20 with 5 degrees of freedom (df) and a p-value of 0.001, providing strong evidence against the null hypothesis.

The calculated (O-E)^2/E values highlight the degree of deviation between observed and expected survival rates, especially for men in the 70 to 90 age group, which showed the highest deviation (8.45). The significant p-value indicates that both sex and age group significantly impact survival outcomes, suggesting that survival analysis should consider these factors together

Survival by sex and Symptoms

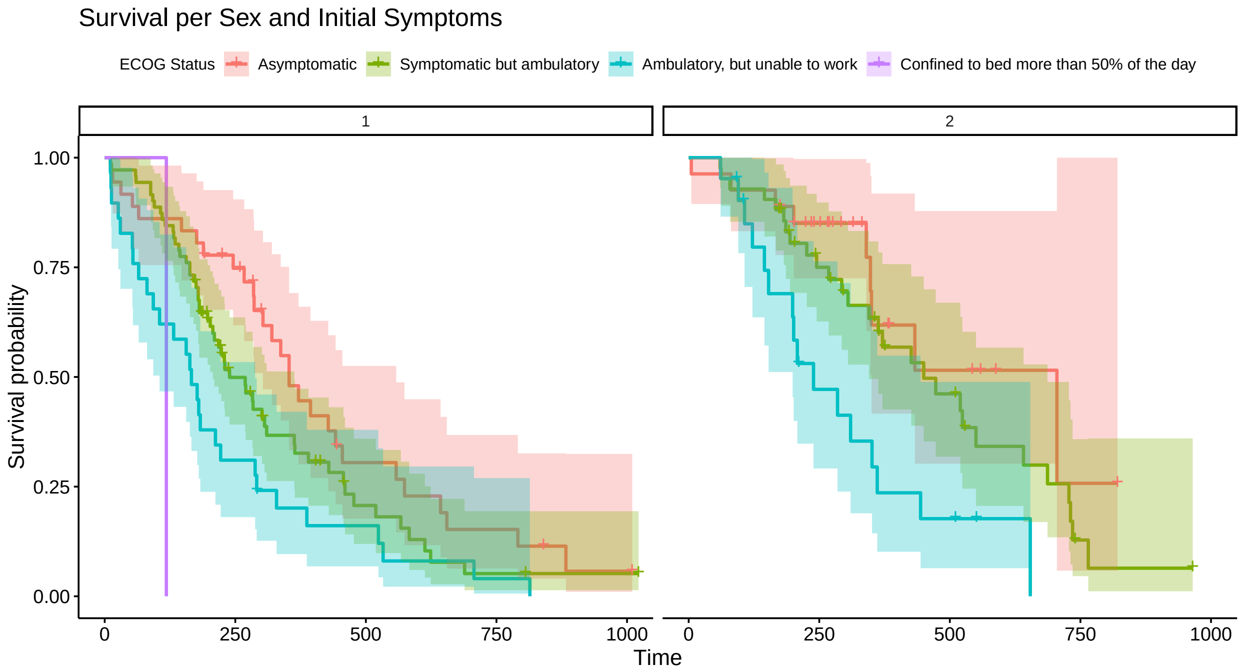
The Eastern Cooperative Oncology Group (ECOG) scale is a widely used system for assessing a patient's performance status in oncology. It helps determine how a patient's disease affects their daily living abilities. The ECOG scale ranges from 0 to 5, where:

- **0**: Asymptomatic
- **1**: Symptomatic but ambulatory
- **2**: Ambulatory and capable of all self-care but unable to work; up and about more than 50% of waking hours.
- **3**: Ambulatory, but unable to work
- **4**: Confined to bed more than 50% of the day
- **5**: Death

In this case, the information available to us indicates how the patient has arrived since the study began.

Graphical method

```
# Survival plot by sex

SurvPerSexAndSymptoms <- survfit(Surv(time, status) ~ ph.ecog + sex, TQ_SurvivalData, conf.type = "log-l
  ggsurvplot(title = "Survival per Sex and Initial Symptoms",
             conf.int = TRUE,
             facet.by = "sex",
             legend.title = "ECOG Status",
             legend.labs = c("Asymptomatic", "Symptomatic but ambulatory",
                             "Ambulatory, but unable to work",
                             "Confined to bed more than 50% of the day"),
             panel.labs = list(edadg = c("25 to 50", "More than 50")),
             short.panel.labs = TRUE)

print(SurvPerSexAndSymptoms)
```



```
# Survival plot by sex

Test_3 <- survdiff(Surv(time, status) ~ sex + ph.ecog, TQ_SurvivalData, rho = 1)
print(Test_3)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ sex + ph.ecog, data = TQ_SurvivalData,
##     rho = 1)
##
## n=227, 1 observation deleted due to missingness.
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## sex=1, ph.ecog=0 36   14.614   18.125     0.680      1.22
## sex=1, ph.ecog=1 71   34.497   27.985     1.516      2.93
## sex=1, ph.ecog=2 29   19.744    8.922    13.127     19.74
## sex=1, ph.ecog=3  1    0.846    0.159     2.967      3.24
## sex=2, ph.ecog=0 27    5.263   13.065     4.659      7.56
## sex=2, ph.ecog=1 42   13.317   22.362     3.658      6.93
## sex=2, ph.ecog=2 21   10.257    7.921     0.689      1.01
##
##  Chisq= 37.4  on 6 degrees of freedom, p= 1e-06
```

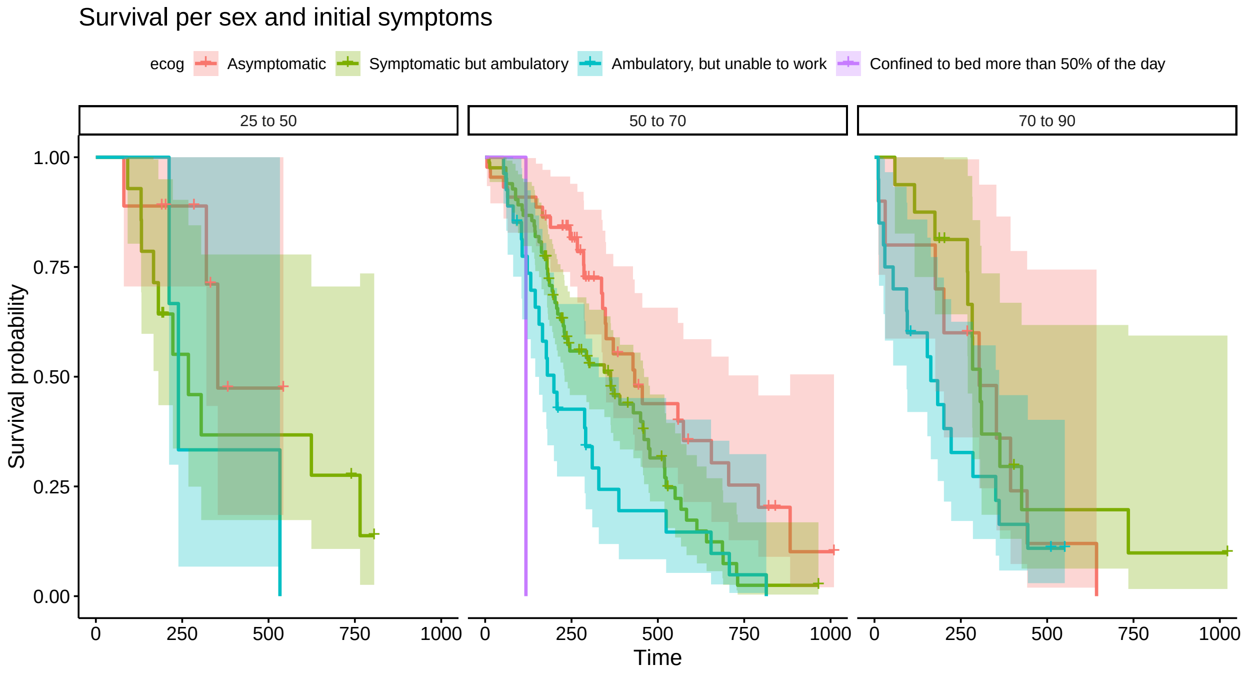
The chi-squared statistic of 37.4 indicates a substantial difference in survival rates among the groups when considering both sex and ECOG performance status. The p-value of 1e-06 is significantly lower than the conventional alpha level of 0.05, suggesting that we can reject the null hypothesis of equal survival functions across the defined groups.

Given the highly significant p-value, it is evident that there are marked differences in survival based on the combination of sex and ECOG status. Specifically, the results imply that male patients with higher ECOG scores (indicating more severe symptoms) tend to have worse survival outcomes compared to their female counterparts and those with lower ECOG scores. This analysis highlights the critical role that both sex and initial health status play in determining patient prognosis.

Survival by age and symptoms

Graphical method

```
SurvPerAgeAndSymptoms <- survfit(Surv(time, status) ~ ph.ecog + age_group, TQ_SurvivalDataAgeGroups, co
  ggsurvplot(title = "Survival per sex and initial symptoms", conf.int = T,
             facet.by = "age_group", legend.title = "ecog", legend.labs = c("Asymptomatic", "Symptomatic
                                                                      "Ambulatory, but unable to work",
                                                                      "Confined to bed more than 50% of


print(SurvPerAgeAndSymptoms)
```



```
# Survival plot by sex

Test_4 <- survdiff(Surv(time, status) ~ age_group + ph.ecog, TQ_SurvivalDataAgeGroups, rho = 1)
print(Test_4)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ age_group + ph.ecog,
##     data = TQ_SurvivalDataAgeGroups, rho = 1)
##
## n=227, 1 observation deleted due to missingness.
##
##                                  N Observed Expected (O-E)^2/E (O-E)^2/V
## age_group=25 to 50, ph.ecog=0  9    1.849    4.105   1.24068   1.75806
## age_group=25 to 50, ph.ecog=1 14    6.099    6.400   0.01414   0.02214
## age_group=25 to 50, ph.ecog=2  3    1.539    1.484   0.00204   0.00286
## age_group=50 to 70, ph.ecog=0 44   12.747   22.877   4.48540   8.51900
## age_group=50 to 70, ph.ecog=1 83   34.913   36.467   0.06626   0.14769
## age_group=50 to 70, ph.ecog=2 27   15.461    9.648   3.50294   5.32321
## age_group=50 to 70, ph.ecog=3  1    0.846    0.159   2.96707   3.23565
## age_group=70 to 90, ph.ecog=0 10    5.282    4.208   0.27412   0.39939
## age_group=70 to 90, ph.ecog=1 16    6.802    7.479   0.06127   0.09331
## age_group=70 to 90, ph.ecog=2 20   13.002    5.712   9.30389  13.15663
##
##  Chisq= 29.4  on 9 degrees of freedom, p= 6e-04
```

The p-value of 0.0006 suggests strong evidence against the null hypothesis, indicating that there are significant differences in survival between the groups defined by both age and ECOG status.

The log-rank test reveals that survival varies significantly with age and performance status as measured by the ECOG scale. The test identifies critical disparities, particularly among individuals in the age group of 50 to 70 with different ECOG scores, which had the highest observed variance in survival times. The low p-value indicates that age and initial symptom severity, as quantified by the ECOG scale, are important factors influencing survival outcomes in this group. This emphasizes the necessity of considering both demographic and clinical variables in survival analyses to identify at-risk populations effectively.

Parametric Survival Estimation

Parametric survival estimation is a statistical approach used to analyze time-to-event data, commonly referred to as survival data. In this context, "survival" refers to the time until an event of interest occurs, such as death, failure, or relapse. Unlike non-parametric methods, which do not assume a specific distribution for the survival times, parametric methods model the survival function using predefined statistical distributions. This allows for more efficient estimation and hypothesis testing when the assumptions of these models are met.

Objectives

The primary objective of parametric survival estimation is to provide a mathematical framework for modeling the time until an event occurs. By fitting a specific distribution to the data, researchers can:

1. Estimate survival probabilities at different time points.
2. Predict the expected time to event for different groups based on covariates.
3. Assess the effect of covariates on the survival time using regression techniques.
4. Calculate the hazard function, which describes the instantaneous risk of the event occurring at any given time.

Common Parametric Distributions

Several distributions are commonly used in parametric survival analysis, each with its own characteristics:

1. **Weibull Distribution**:

    - The Weibull distribution is characterized by its shape parameter (β) and scale parameter (λ).
    - It can model increasing or decreasing hazard rates depending on the value of β.
    - For instance, if $\beta < 1$, the hazard decreases over time, suggesting that individuals are more likely to survive longer, while $\beta > 1$ indicates an increasing hazard rate.

$$S(t) = e^{-(t/\lambda)^{\beta}}$$

2. **Log-Normal Distribution**:

   - The log-normal distribution is used when the logarithm of the survival times follows a normal distribution.
   - It is particularly useful for modeling data that is positively skewed, as it can accommodate a wide range of shapes.

$$S(t) = 1 - \Phi\left(\frac{\ln(t) - \mu}{\sigma}\right)$$

3. **Log-Logistic Distribution**:

   - Similar to the log-normal distribution, the log-logistic model is often used for data that exhibits a “bathtub” shaped hazard function.
   - It can represent increasing hazard rates initially, followed by a decrease, making it suitable for certain biological processes.

$$S(t) = \frac{1}{1 + (t/\lambda)^{\alpha}}$$

4. **Gompertz Distribution**:

   - The Gompertz distribution is frequently applied in actuarial science and demography.
   - Its hazard function can model aging processes where the risk of an event increases exponentially over time.

$$S(t) = e^{-\frac{e^{\gamma t} - 1}{\gamma}}$$

5. **Exponential Distribution**:

   - The exponential distribution is a special case of the Weibull distribution with a shape parameter of $\beta = 1$.
   - It assumes a constant hazard rate, which means the risk of the event occurring is the same at any point in time. This property makes it particularly useful for modeling memoryless processes.

$$S(t) = e^{-\lambda t}$$

Rationale for Not Using the Exponential Distribution

While the exponential distribution is simple and mathematically tractable, it is often not appropriate for many real-world applications due to its assumption of a constant hazard rate. In most survival data, the risk of an event occurring changes over time. For instance, the risk of death may decrease after a certain age or increase as a disease progresses. In such cases, more flexible distributions like Weibull, log-normal, or log-logistic can better capture the varying hazard rates observed in the data. As a result, while the exponential distribution can serve as a useful starting point, it is often insufficient for accurately modeling the complexities inherent in survival data.

Advantages of Parametric Methods

1. **Efficiency**: Parametric methods can yield more efficient estimates than non-parametric methods, particularly when the underlying distribution is correctly specified.
2. **Predictive Power**: By incorporating covariates, parametric models can predict survival times and probabilities more accurately, providing valuable insights for decision-making.
3. **Interpretability**: The coefficients of parametric models can often be interpreted in terms of hazard ratios, making them straightforward to communicate.

Disadvantages of Parametric Methods

1. **Distribution Assumptions**: Parametric methods rely heavily on the assumption that the data follow a specific distribution. If this assumption is violated, the estimates can be biased or misleading.
2. **Model Complexity**: Some parametric models can become complex and difficult to fit, especially with limited data or when covariates interact in non-linear ways.
3. **Sensitivity to Outliers**: Parametric models can be sensitive to outliers, which may disproportionately influence the estimates.

In summary, parametric survival estimation provides a robust framework for analyzing time-to-event data through the lens of statistical distributions. The choice of distribution is crucial, as it directly affects the model's assumptions, efficiency, and interpretability. Researchers must weigh the advantages and disadvantages of parametric methods while considering the characteristics of their data to ensure accurate and meaningful survival analysis.

In this analysis, we will divide the patients into three age groups: 25-50, 50-70, and 70-90. For each group, we will fit several parametric survival models, including the Weibull, Log-Normal, Gompertz, and other relevant distributions. The goal is to compare these models to identify which one provides the best fit for the data within each age group.

Tools for Model Comparison

To determine the best-fitting model, two key tools will be employed:

1. **Comparison Plot:** This plot visually compares the estimated survival functions from both the parametric model and the non-parametric Kaplan-Meier estimator. By plotting both survival curves on the same graph, we can assess how closely the parametric model aligns with the non-parametric Kaplan-Meier curve, which serves as a baseline.The closer it is to a function y = x, the better the alignment.

2. **AIC, BIC, and Log-Likelihood:** These are numerical criteria that help assess the goodness of fit of the parametric models. Each criterion provides a different perspective on model performance:

   - **AIC (Akaike Information Criterion):** A lower AIC indicates a model that better balances fit and complexity. It penalizes models for having too many parameters but favors a good fit to the data.
   - **BIC (Bayesian Information Criterion):** Like AIC, BIC penalizes model complexity but applies a stricter penalty for additional parameters. A lower BIC suggests a model that better balances fit and complexity but with a greater emphasis on model simplicity.
   - **Log-Likelihood (LogLik):** This represents the likelihood of the data given the model. Higher log-likelihood values indicate that the model fits the data better. However, this criterion alone does not penalize complexity, which is why AIC and BIC are often preferred.

Interpretation of Results

- **Comparison Plot:** If the parametric survival curve closely follows the Kaplan-Meier curve, the parametric model is considered a good fit. Deviations between the two curves suggest that the parametric model may not fully capture the underlying survival dynamics.
- **AIC and BIC:** When comparing different parametric models, the model with the lowest AIC or BIC value is generally preferred. It indicates that the model strikes the best balance between complexity and accuracy. Since BIC applies a stricter penalty for additional parameters, it may favor simpler models more than AIC.
- **Log-Likelihood:** The model with the highest log-likelihood value fits the data better in terms of maximizing the likelihood of observing the given data. However, without penalizing for additional parameters, this criterion alone might favor overly complex models.

By using both visual and numerical tools, we can ensure that the chosen parametric model for each age group not only fits the data well but also avoids unnecessary complexity.

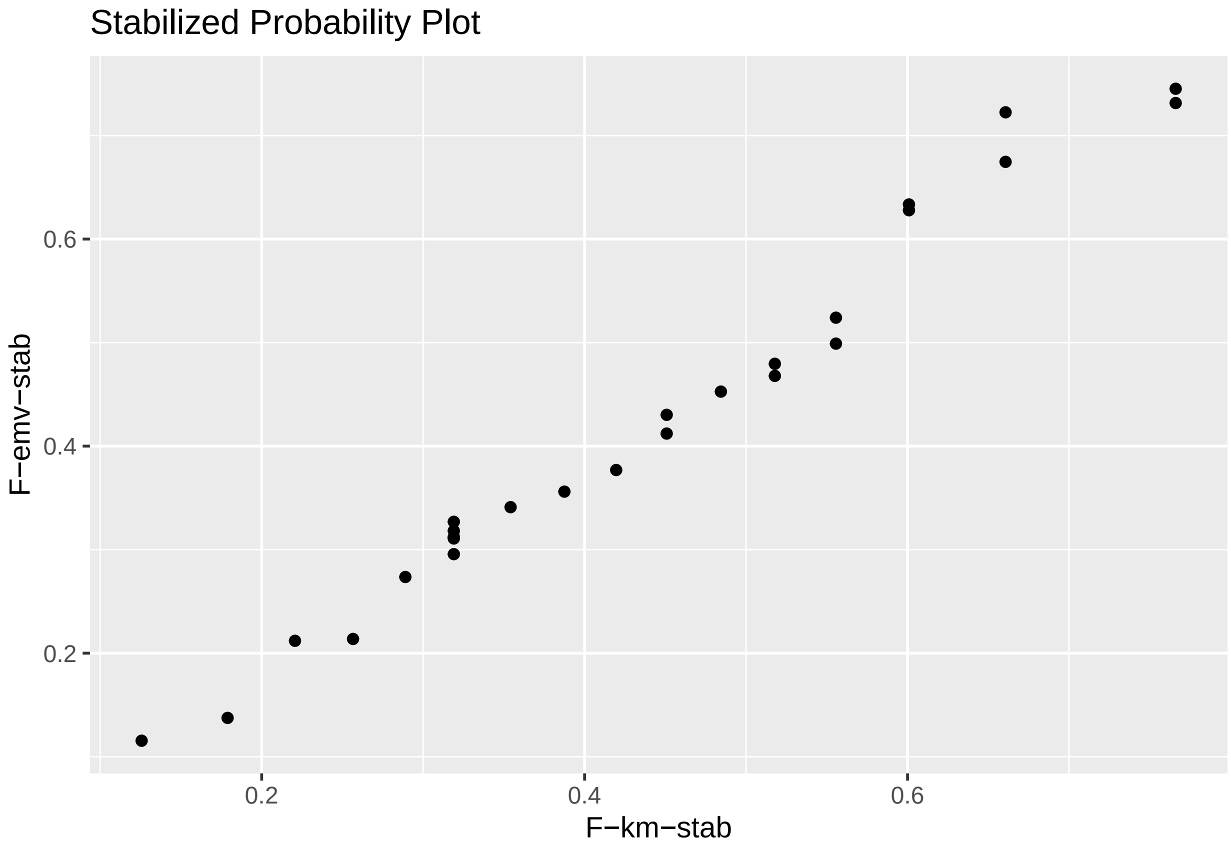
First Group (25 - 50)

```r
# Define distributions
Distributions <- c("weibull", "llogis", "gompertz", "lnorm")

# 1st group 25 to 50
data1stgroup <- TQ_SurvivalDataAgeGroups %>%
  filter(age_group == "25 to 50")

# Fit parametric models and get AIC, BIC, LogLik
AIC.model <- fit_parametric_models(data1stgroup, Distributions)

# Fit a LogNormal model
flex_lnorm <- fit_lognormal_model(data1stgroup)

# Obtain the parametric estimates from the fitted model
F_est_values <- summary(flex_lnorm) %>%
  as.data.frame() %>%
  pull(est)

F_est <- 1 - F_est_values

# Fit a non-parametric Kaplan-Meier survival estimate
KM_est <- survfit(Surv(time, status) ~ 1, data = data1stgroup)
KM_est <- 1 - KM_est$surv

# Create stabilized probability plot
stabilized_plot <- create_stabilized_probability_plot(KM_est, F_est)

kable(AIC.model, caption = "Parametric Model Comparison Results", escape = F, col.names = c("Model", "A
  kable_styling(full_width = F, position = "center") %>%
  row_spec(0, bold = TRUE, color = "black") %>%
  row_spec(1:nrow(AIC.model), color = "black")
```

Table 5: Parametric Model Comparison Results

Model	AIC	BIC	LogLik
lnorm	230.5883	233.1045	-113.2942
llogis	231.4998	234.0160	-113.7499
weibull	233.1531	235.6693	-114.5765
gompertz	235.3213	237.8375	-115.6606

```r
print(stabilized_plot)
```

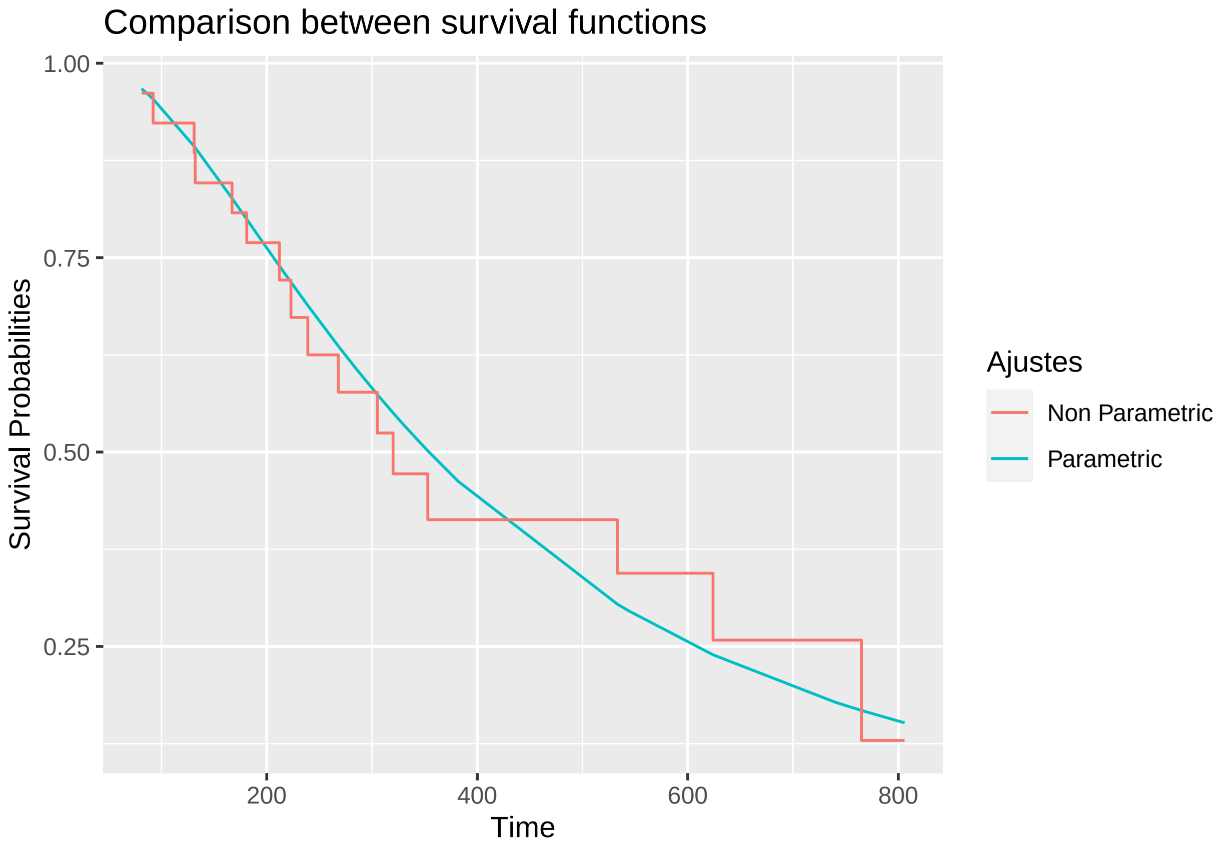


Based on the AIC, BIC, and Log-Likelihood values, the **Log-Normal (lnorm)** model is the best-fitting model for the data. It minimizes both AIC and BIC, and has the highest Log-Likelihood, indicating it provides the best balance between model fit and complexity. The **Log-Logistic (llogis)** model is a close second, while the **Weibull** and **Gompertz** models are less preferred due to their higher AIC, BIC, and lower Log-Likelihood values.

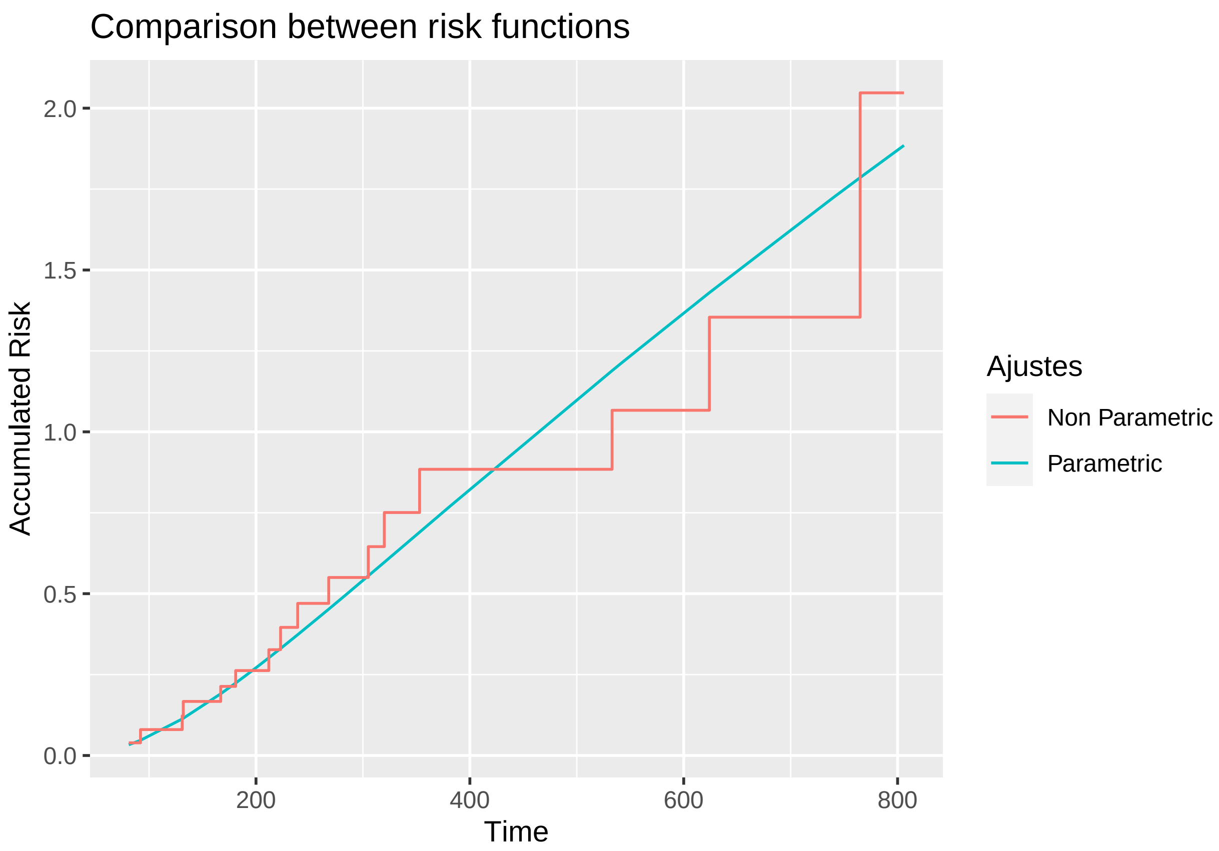
Regarding the stabilized plot , there is a straight line tendency, however, as there is a large percentage of censored observations, vertical lines appear.

```r
# Comparison between survival functions (K-M and lnorm)
flex_gg <- flex_lnorm %>%
  summary(type = "survival") %>%
  data.frame() %>%
  fortify()
```

```r
km_gg <- survfit(Surv(time, status) ~ 1, data = data1stgroup) %>%
  tidy() %>%
  fortify()

survival_comparison_plot <- create_survival_comparison_plot(flex_gg, km_gg)

# Comparison between accumulated risk functions
risk_comparison_plot <- create_risk_comparison_plot(flex_gg, km_gg)

print(survival_comparison_plot)
```



```r
print(risk_comparison_plot)
```



The plot shows that both the non-parametric (Kaplan-Meier) and parametric (likely Log-Normal) survival estimates align well over time. This indicates that:

- **Good fit**: The parametric model fits the data well, as seen from the close alignment of the curves.
- **Minor deviations**: There are small differences at later time points (after ~600 days), suggesting slight limitations in capturing the full survival behavior.
- **General adequacy**: Overall, the parametric model provides a reasonable approximation, and can be used for further analysis.

Despite the slight deviations, the parametric approach is suitable for modeling these survival data.

Second Group (50 - 70)

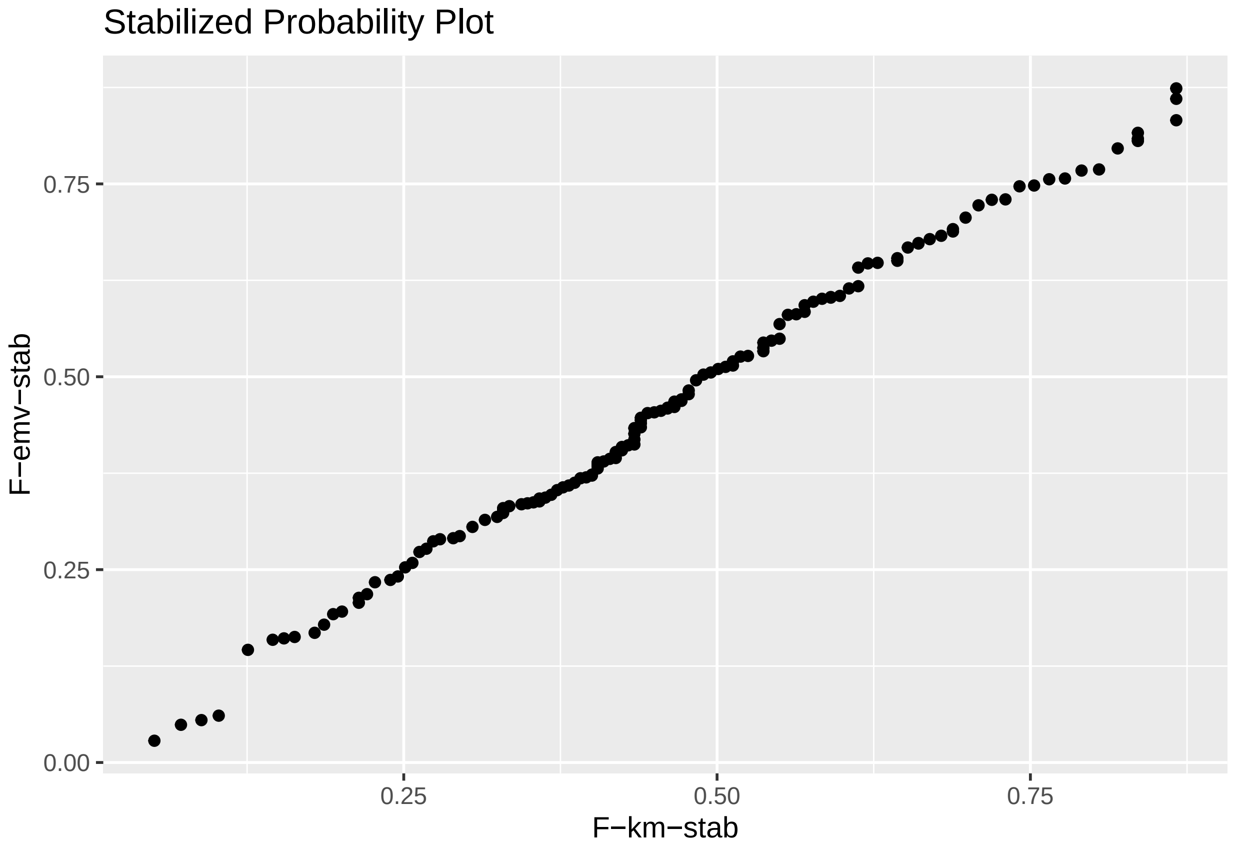
```r
# 2nd group 50 to 70
data2ndgroup <- TQ_SurvivalDataAgeGroups %>%
  filter(age_group == "50 to 70")

# Fit parametric models and get AIC, BIC, LogLik
AIC.model2 <- fit_parametric_models(data2ndgroup, Distributions)

# Fit a Weibull model (AIC indicates it)
flex_weibull <- fit_weibull_model(data2ndgroup)

# Obtain the parametric estimates from the fitted model
```

Table 6: Parametric Model Comparison Results

Model	AIC	BIC	LogLik
weibull	1561.320	1567.420	-778.6600
gompertz	1563.955	1570.055	-779.9776
llogis	1570.946	1577.045	-783.4728
lnorm	1581.573	1587.673	-788.7866

```r
F_est_values2 <- summary(flex_weibull) %>%
  as.data.frame() %>%
  pull(est)

F_est2 <- 1 - F_est_values2

# Fit a non-parametric Kaplan-Meier survival estimate
KM_est2 <- survfit(Surv(time, status) ~ 1, data = data2ndgroup)
KM_est2 <- 1 - KM_est2$surv

# Create stabilized probability plot
stabilized_plot2 <- create_stabilized_probability_plot(KM_est2, F_est2)

kable(AIC.model2, caption = "Parametric Model Comparison Results", escape = F, col.names = c("Model", "A
  kable_styling(full_width = F, position = "center") %>%
  row_spec(0, bold = TRUE, color = "black") %>%
  row_spec(1:nrow(AIC.model2), color = "black")
```

```r
print(stabilized_plot2)
```



Interpretation of the Stabilized Probability Plot

The stabilized probability plot compares the empirical Kaplan-Meier survival function (F-km-stab) with the parametric survival function (F-emy-stab). In this plot:

- **Good alignment**: The points closely follow a diagonal pattern, indicating a strong agreement between the parametric and non-parametric models.
- **Model validation**: This confirms that the parametric model fits the data well, particularly in the middle range of survival probabilities.

The plot shows that the parametric distribution is a good approximation for the survival data in this context.

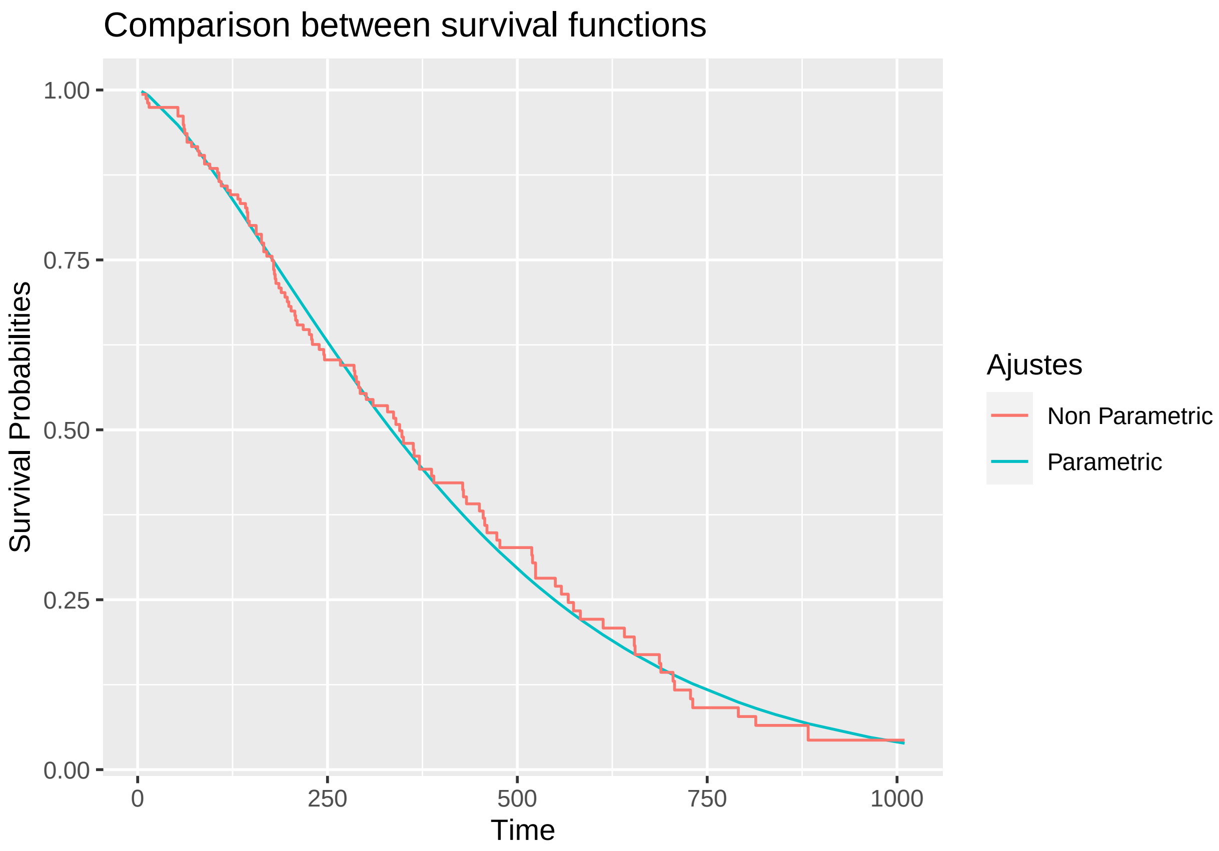
Parametric Model Comparison Interpretation

Based on the AIC, BIC, and LogLik values:

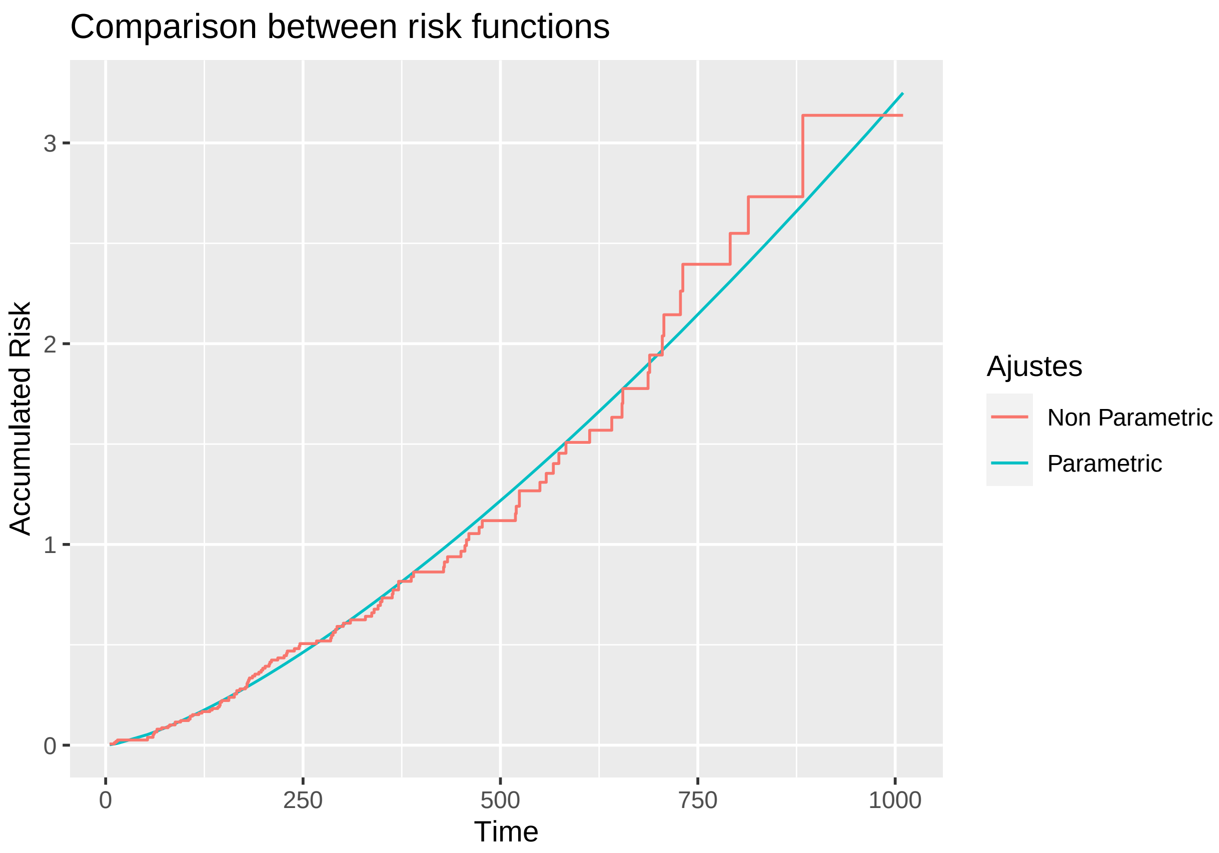
- **Weibull model** has the lowest AIC (1561.320) and BIC (1567.420), which suggests it is the best fitting model among those tested.
- **Gompertz model** comes in second, with slightly higher AIC and BIC values.
- **Log-logistic and Log-normal models** perform worse in terms of both AIC and BIC, indicating they do not fit the data as well.

In summary, the **Weibull distribution** is the preferred parametric model, as it provides the best fit according to AIC and BIC, and this is further supported by the stability shown in the probability plot.

```
# Comparison between survival functions (K-M and lnorm)
flex_gg2 <- flex_weibull %>%
  summary(type = "survival") %>%
  data.frame() %>%
  fortify()

km_gg2 <- survfit(Surv(time, status) ~ 1, data = data2ndgroup) %>%
  tidy() %>%
  fortify()

survival_comparison_plot2 <- create_survival_comparison_plot(flex_gg2, km_gg2)

# Comparison between accumulated risk functions
risk_comparison_plot2 <- create_risk_comparison_plot(flex_gg2, km_gg2)

print(survival_comparison_plot2)
```



```
print(risk_comparison_plot2)
```



Interpretation of the Comparison Between Survival Functions Plot

This plot compares the non-parametric Kaplan-Meier survival function (in red) with the parametric survival function (in green/blue). Key observations:

- **Close alignment**: The two survival curves are closely aligned throughout the observed time, indicating that the parametric model is a good fit for the data.
- **Deviation in tail**: There is a slight divergence in the tail of the distribution (at higher time values), which could suggest that the parametric model slightly underestimates or overestimates the survival probability in the long-term. However, the difference is minimal.

We have to take into account that this is the best fit also partly conditioned by the fact that this is the age group with the highest number of subjects.

In summary, the parametric model provides a reliable approximation of the survival function, closely following the non-parametric Kaplan-Meier estimate.

Third Group (70 - 90)

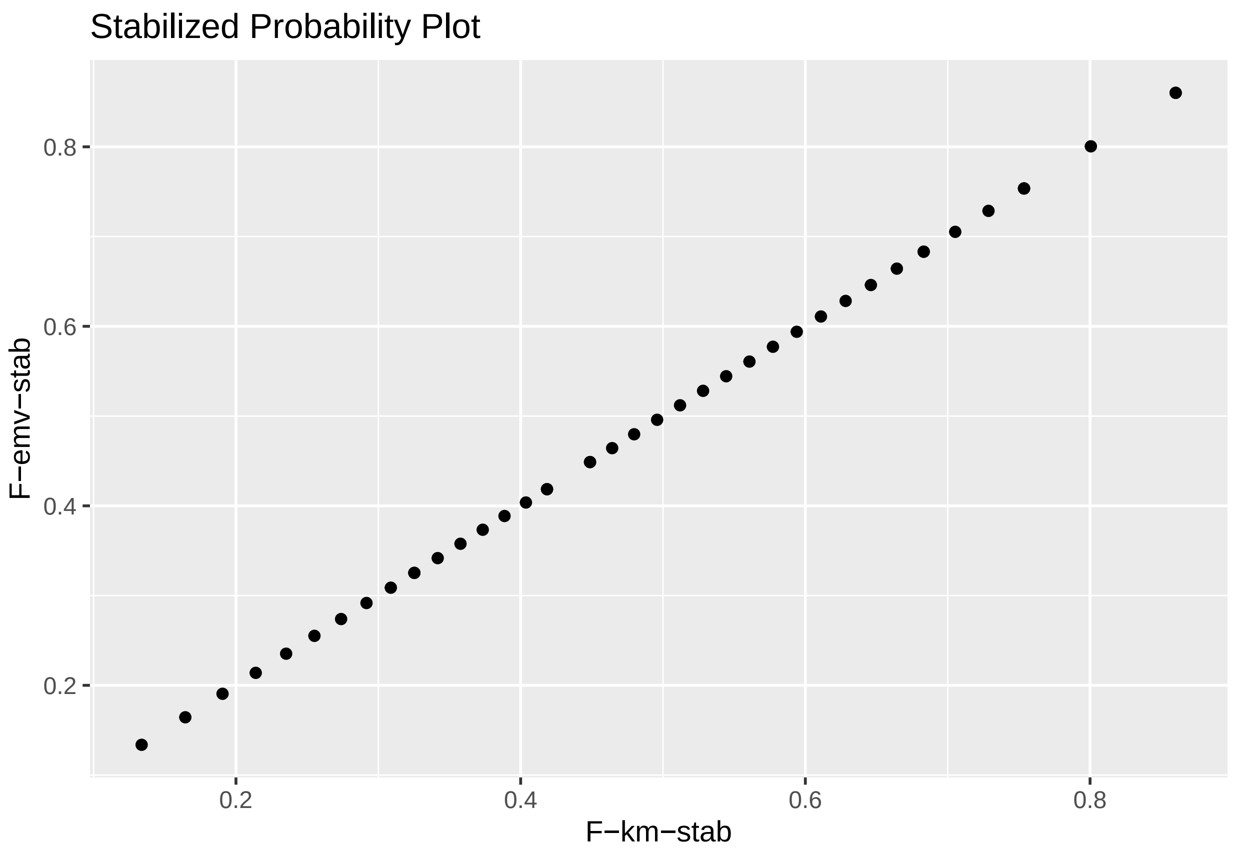
```
data3rdgroup <- TQ_SurvivalDataAgeGroups %>%
  filter(age_group == "70 to 90")

# Fit parametric models and get AIC, BIC, LogLik
```

Table 7: Parametric Model Comparison Results

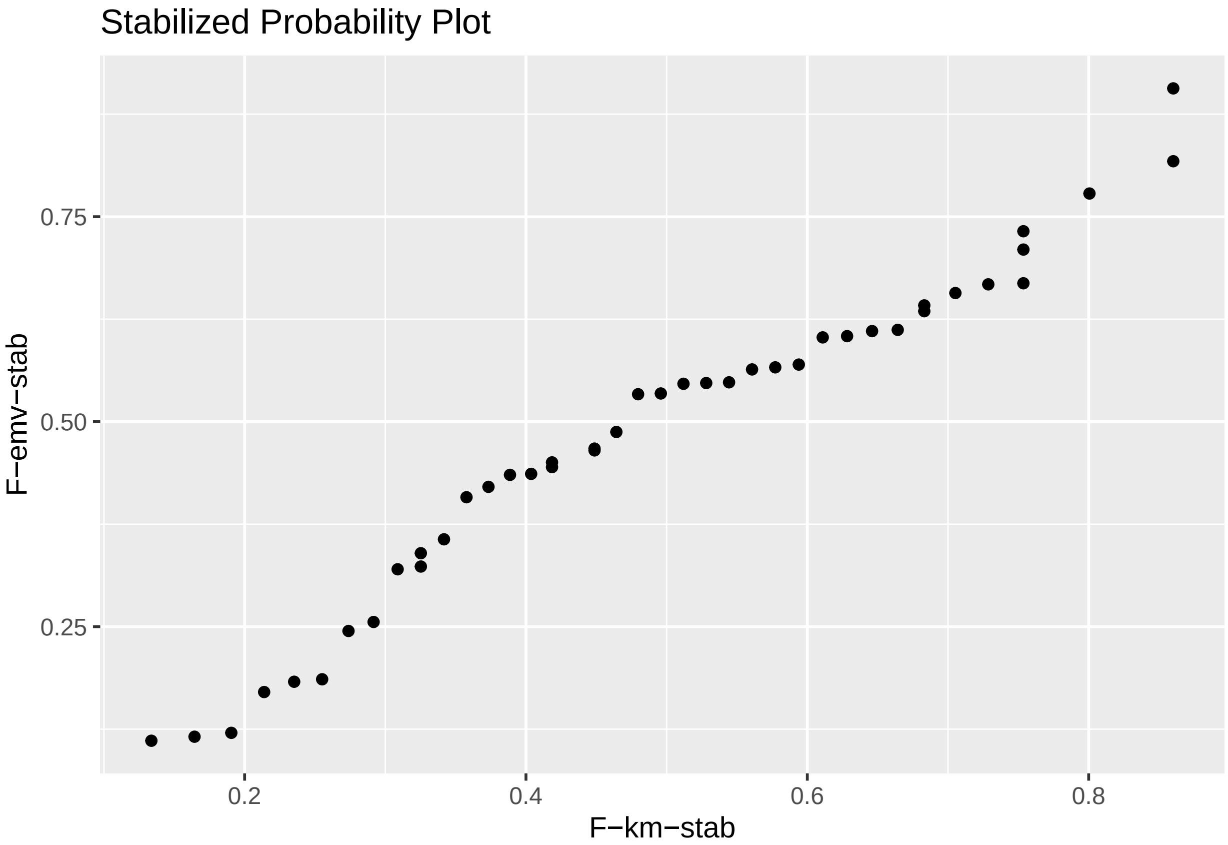
Model	AIC	BIC	LogLik
gompertz	517.5310	521.1883	-256.7655
weibull	517.6061	521.2634	-256.8031
llogis	522.8604	526.5177	-259.4302
lnorm	524.3460	528.0033	-260.1730

```r
AIC.model3 <- fit_parametric_models(data3rdgroup, Distributions)

# Fit a Weibull model (AIC indicates it)
flex_weibull3 <- fit_weibull_model(data3rdgroup)
flex_gompertz <- fit_gompertz_model(data3rdgroup)


# Obtain the parametric estimates from the fitted model
F_est_values3.1 <- summary(flex_weibull3) %>%
  as.data.frame() %>%
  pull(est)

F_est3.1 <- 1 - F_est_values3.1


F_est_values3.2 <- summary(flex_gompertz) %>%
  as.data.frame() %>%
  pull(est)

F_est3.2 <- 1 - F_est_values3.2

# Fit a non-parametric Kaplan-Meier survival estimate
KM_est3.1 <- survfit(Surv(time, status) ~ 1, data = data3rdgroup)
KM_est3.1 <- 1 - KM_est3.1$surv

# Create stabilized probability plot
stabilized_plot3.1 <- create_stabilized_probability_plot(KM_est3.1, KM_est3.1)

KM_est3.2 <- survfit(Surv(time, status) ~ 1, data = data3rdgroup)
KM_est3.2 <- 1 - KM_est3.2$surv

# Create stabilized probability plot
stabilized_plot3.2 <- create_stabilized_probability_plot(KM_est3.2, F_est3.2)

kable(AIC.model3, caption = "Parametric Model Comparison Results", escape = F, col.names = c("Model", "A
  kable_styling(full_width = F, position = "center") %>%
  row_spec(0, bold = TRUE, color = "black") %>%
  row_spec(1:nrow(AIC.model3), color = "black")
```

```r
print(stabilized_plot3.1)
```



```
print(stabilized_plot3.2)
```



```
#Better stabilized plot weibull so we go with it
```

Interpretation of Parametric Model Results for the Age Group 70-90

1. **AIC and BIC Comparison**: The Gompertz model has the lowest AIC (517.5310) and BIC (521.1883), indicating it provides the best fit compared to the other models. The Weibull model is extremely close in terms of AIC (517.6061), suggesting it is also a strong candidate. In comparison, the Log-Logistic and Log-Normal models have higher AIC and BIC values, which makes them less favorable for this dataset.

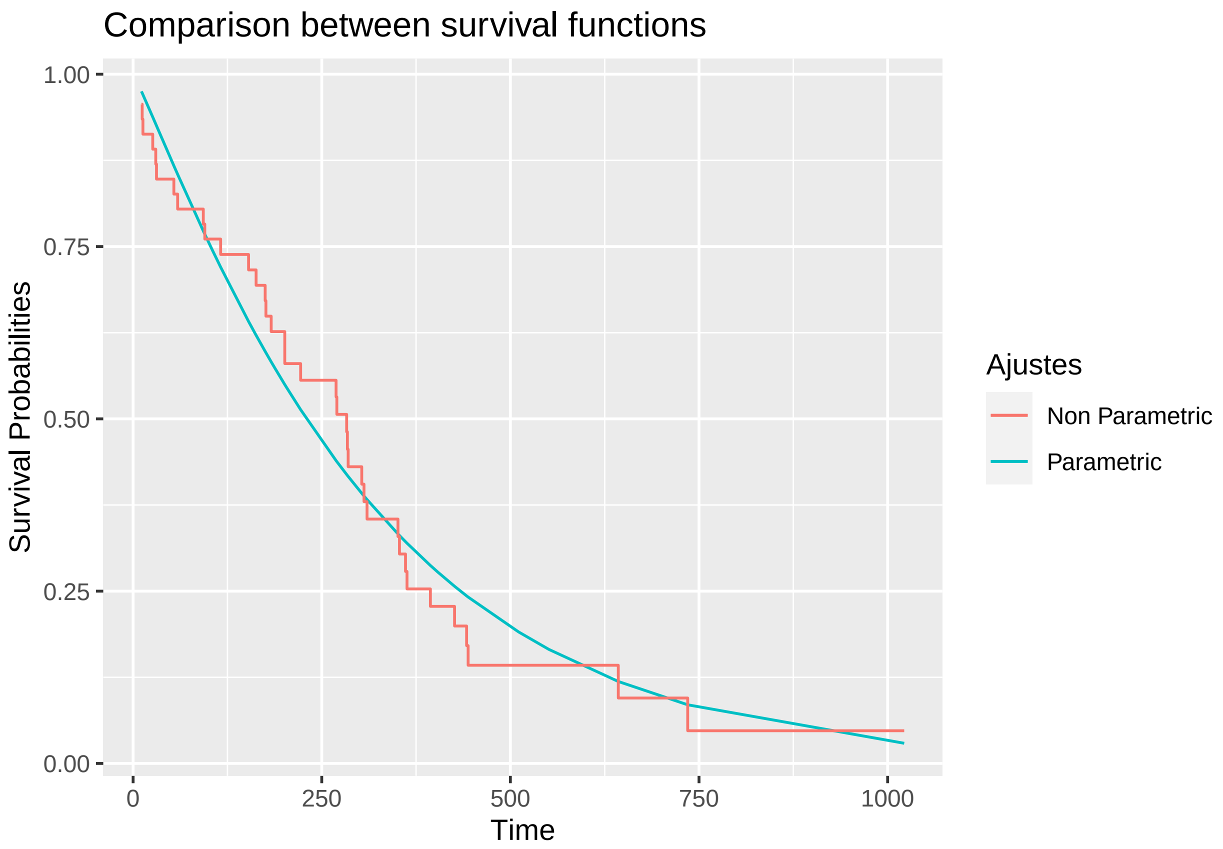
2. **Log-Likelihood (LogLik)**: The Log-Likelihood values are consistent with the AIC and BIC results. The Gompertz model has the highest LogLik (-256.7655), closely followed by the Weibull model (-256.8031), further indicating these two models offer better fits to the data.

3. **Stabilized Probability Plot Comparison**:

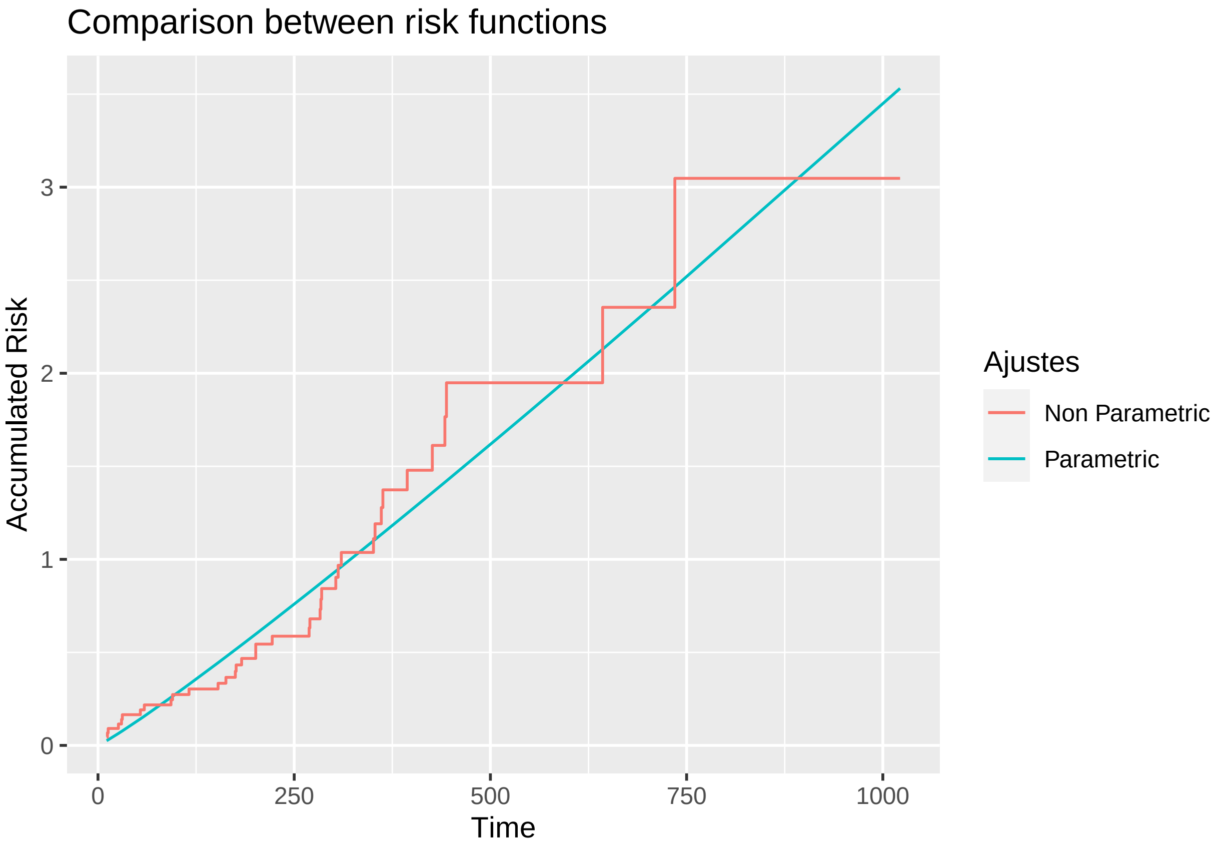
   - The **Weibull model's** stabilized probability plot shows a reasonable match with the non-parametric Kaplan-Meier estimates, suggesting a good fit.
   - The **Gompertz model's** plot also aligns well, but based on visual inspection and statistical comparison, the Weibull model slightly outperforms in terms of model stability.

In conclusion, while both the Gompertz and Weibull models offer strong fits for the 70-90 age group, we proceed with the **Weibull model** for further analysis based on the stabilized probability plot and close AIC/BIC values.

```
# Comparison between survival functions (K-M and lnorm)
flex_gg3 <- flex_weibull3 %>%
  summary(type = "survival") %>%
  data.frame() %>%
  fortify()

km_gg3 <- survfit(Surv(time, status) ~ 1, data = data3rdgroup) %>%
  tidy() %>%
  fortify()

survival_comparison_plot3 <- create_survival_comparison_plot(flex_gg3, km_gg3)

# Comparison between accumulated risk functions
risk_comparison_plot3 <- create_risk_comparison_plot(flex_gg3, km_gg3)

print(survival_comparison_plot3)
```



```
print(risk_comparison_plot3)
```


Comparison between risk functions
Accumulated Risk
0
1
2
3
0
250
500
750
1000
Time
Ajustes
Non Parametric
Parametric