Group Project Poster

Zain Qureshi

2024-10-14

```
# Define the path to the Cancer.xlsx file
file_path <- file.path("~", "Documents", "Spyder", "Cancer.xlsx")  # Use absolute path

# Load Cancer data from Excel file
Cancer <- read_excel(file_path)  # Ensure the path is correct

# Check if the data is loaded correctly
if (exists("Cancer")) {
  message("Cancer data loaded successfully.")
} else {
  stop("Cancer data could not be loaded.")
}
```

```
## Cancer data loaded successfully.
```

```
library(survival)
library(survminer)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Cancer$Age <- as.numeric(Cancer$Age)
Cancer$Survival <- as.numeric(Cancer$Survival)
Cancer$Type <- as.factor(Cancer$Type)
Cancer$Sex <- as.factor(Cancer$Sex)
```

```
str(Cancer)
```

```
## tibble [100 x 4] (S3: tbl_df/tbl/data.frame)
##  $ Type    : Factor w/ 22 levels "Bladder","Brain",..: 20 20 20 20 20 20 20 20 20 20 ...
##  $ Sex     : Factor w/ 2 levels "F","M": 1 2 1 1 2 2 2 2 2 1 ...
##  $ Age     : num [1:100] 61 69 62 66 42 79 76 54 62 69 ...
##  $ Survival: num [1:100] 121 12 9 18 258 43 142 36 149 182 ...
```

```
head(Cancer)
```

```
## # A tibble: 6 x 4
##   Type    Sex     Age Survival
##   <fct>   <fct> <dbl>    <dbl>
## 1 Stomach F        61      121
## 2 Stomach M        69       12
## 3 Stomach F        62        9
## 4 Stomach F        66       18
## 5 Stomach M        42      258
## 6 Stomach M        79       43
```

```
Cancer$Status <- rep(1, nrow(Cancer))
cox_model <- coxph(Surv(Survival, Status) ~ Age + Sex + Type, data = Cancer)

# Check the proportional hazards assumption
cox.zph(cox_model)
```

```
##           chisq df      p
## Age    1.43e-04  1 0.9904
## Sex    1.52e+00  1 0.2170
## Type   3.69e+01 21 0.0173
## GLOBAL 4.25e+01 23 0.0079
```

```
# Calculate AIC for model fit comparison
AIC(cox_model)
```

```
## [1] 743.5945
```

The p-values indicate whether the proportional hazards (PH) assumption holds for each variable. In the Cox Proportional Hazards model, the PH assumption means that the effect of a co-variate on the hazard is constant over time.

The P value for ‘Type’ is <0.05 violating the proportional hazards assumption for cancer type. This implies the model’s estimate for cancer type may be biased or misleading. The effects of cancer type may not be accurately reflected in the hazard ratio, because the impact of it changes over time.

Cox model probably not most suitable?

Instead let us try AFT models, as they do not require the proportional hazards assumption (which failed in the Cox model).

```
# Fit AFT models with different distributions
aft_model_weibull <- survreg(Surv(Survival) ~ Sex + Age + Type, data = Cancer, dist = "weibull")
aft_model_lognormal <- survreg(Surv(Survival) ~ Sex + Age + Type, data = Cancer, dist = "lognormal")
aft_model_exponential <- survreg(Surv(Survival) ~ Sex + Age + Type, data = Cancer, dist = "exponential"

# Compare AIC values
AIC(aft_model_weibull, aft_model_lognormal, aft_model_exponential)
```

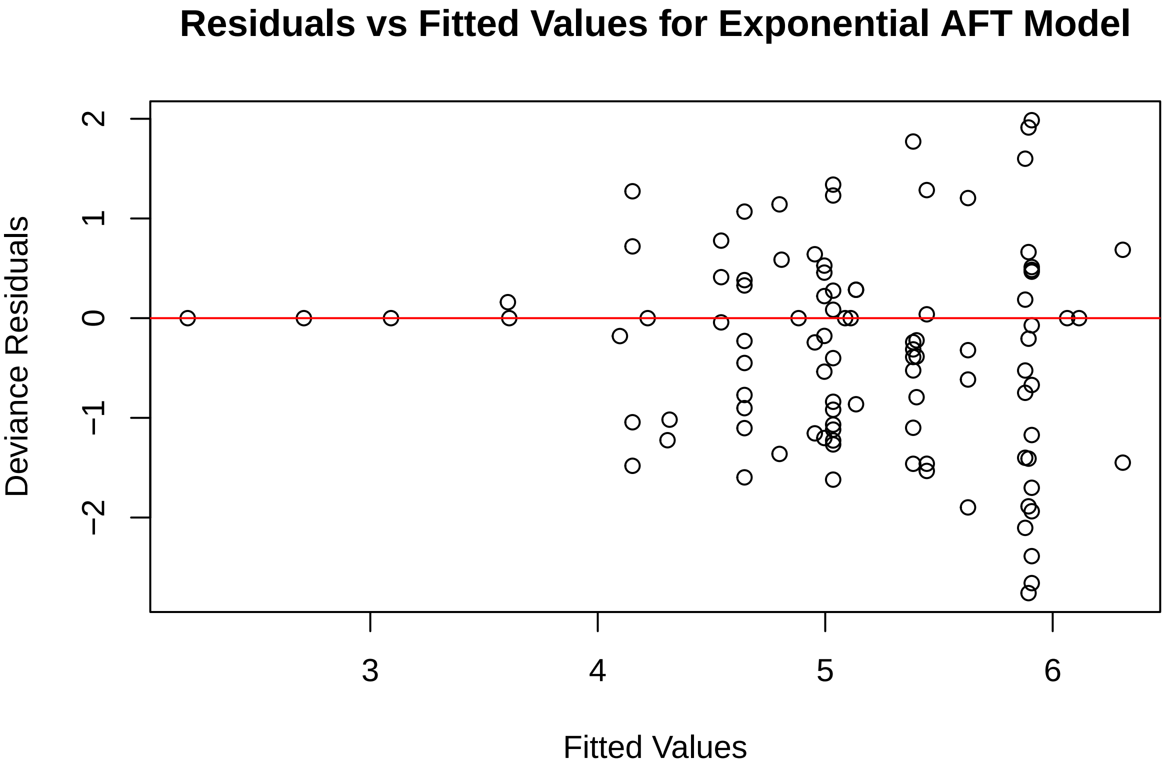
```
##                       df      AIC
## aft_model_weibull     25 1275.807
## aft_model_lognormal   25 1283.025
## aft_model_exponential 24 1273.861
```

The AFT model with exponential distribution has the lowest AIC value (1273.861), suggesting it provides the best fit for the cancer data among these three models. Gives the best balance between complexity and fit.

To confirm this choice, lets run some diagnostics.

```
# Obtain residuals from the Exponential AFT model
aft_model_exp <- survreg(Surv(Survival) ~ Sex + Type, data = Cancer, dist = "exponential")
aft_residuals_exp <- residuals(aft_model_exp, type = "deviance")

# Plot residuals against fitted values
plot(aft_model_exp$linear.predictors, aft_residuals_exp,
     xlab = "Fitted Values",
     ylab = "Deviance Residuals",
     main = "Residuals vs Fitted Values for Exponential AFT Model")
abline(h = 0, col = "red")
```

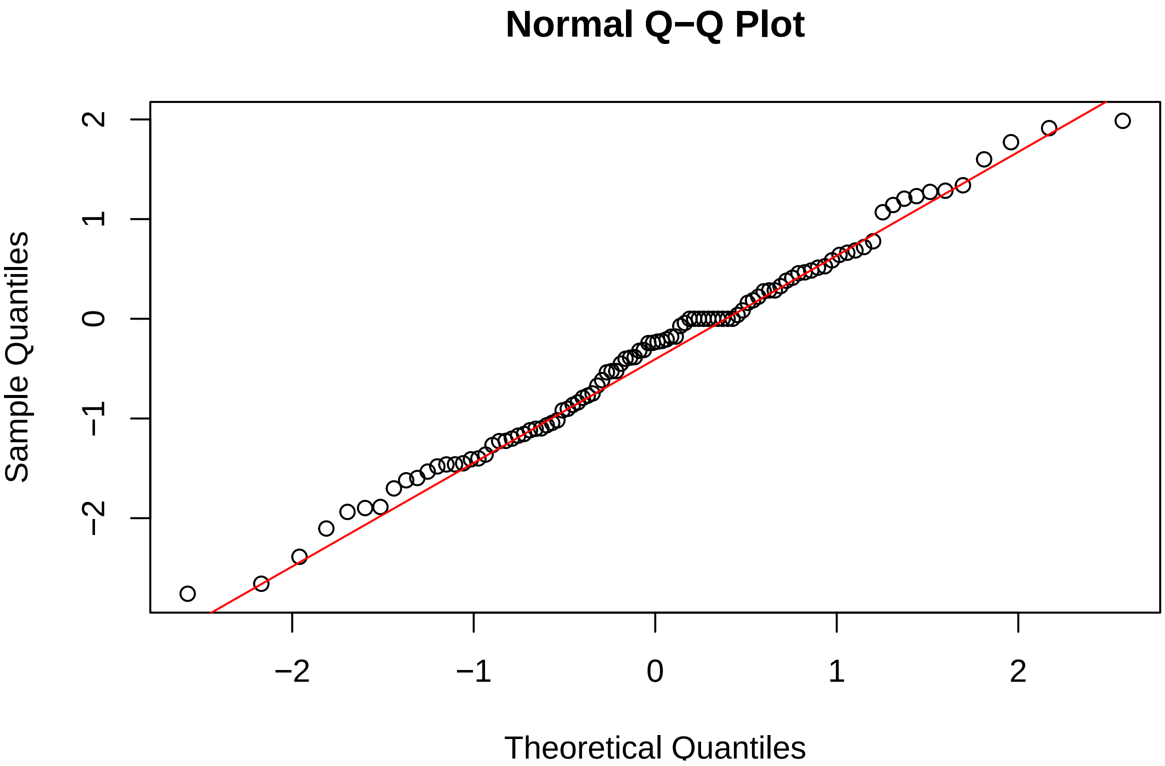


Analysis: - The plot shows deviance residuals against fitted values. - The residuals appear to be randomly scattered around zero. - There's no clear pattern or trend in the residuals. - The spread of residuals seems relatively constant across the range of fitted values.

Conclusion: The residuals vs fitted values plot does not show any major violations of model assumptions. The random scatter and lack of clear patterns suggest that the linearity assumption and the homoscedasticity assumption are reasonably met. This is a good indication for the model fit.

Let us consider a QQ Plot of Residuals to check normality of residuals.

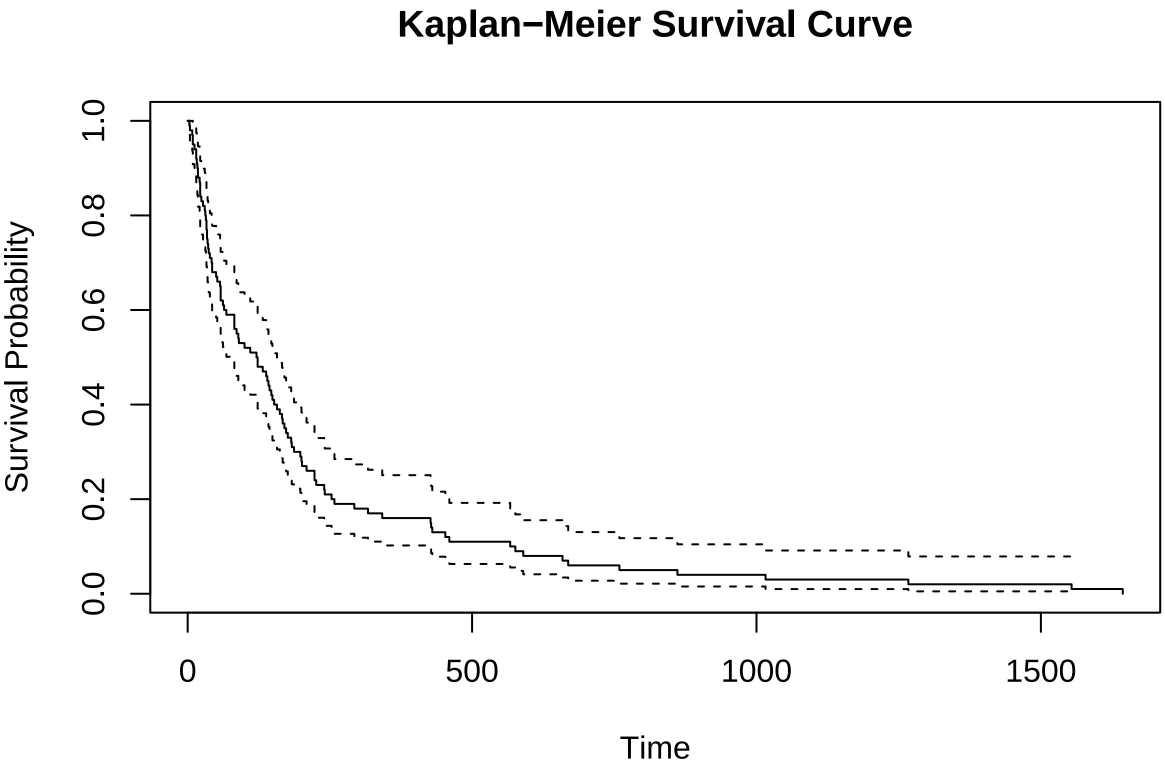
```
# QQ plot to check the normality of residuals
qqnorm(aft_residuals_exp)
qqline(aft_residuals_exp, col = "red")
```



Residuals follow a straight line in the QQ plot, this indicates that the residuals are normally distributed. This is also a postive sign.

Lets now check the distributional assumption.

```r
library(survival)
fit <- survfit(Surv(Survival) ~ 1, data = Cancer)
plot(fit, xlab = "Time", ylab = "Survival Probability", main = "Kaplan-Meier Survival Curve")
```



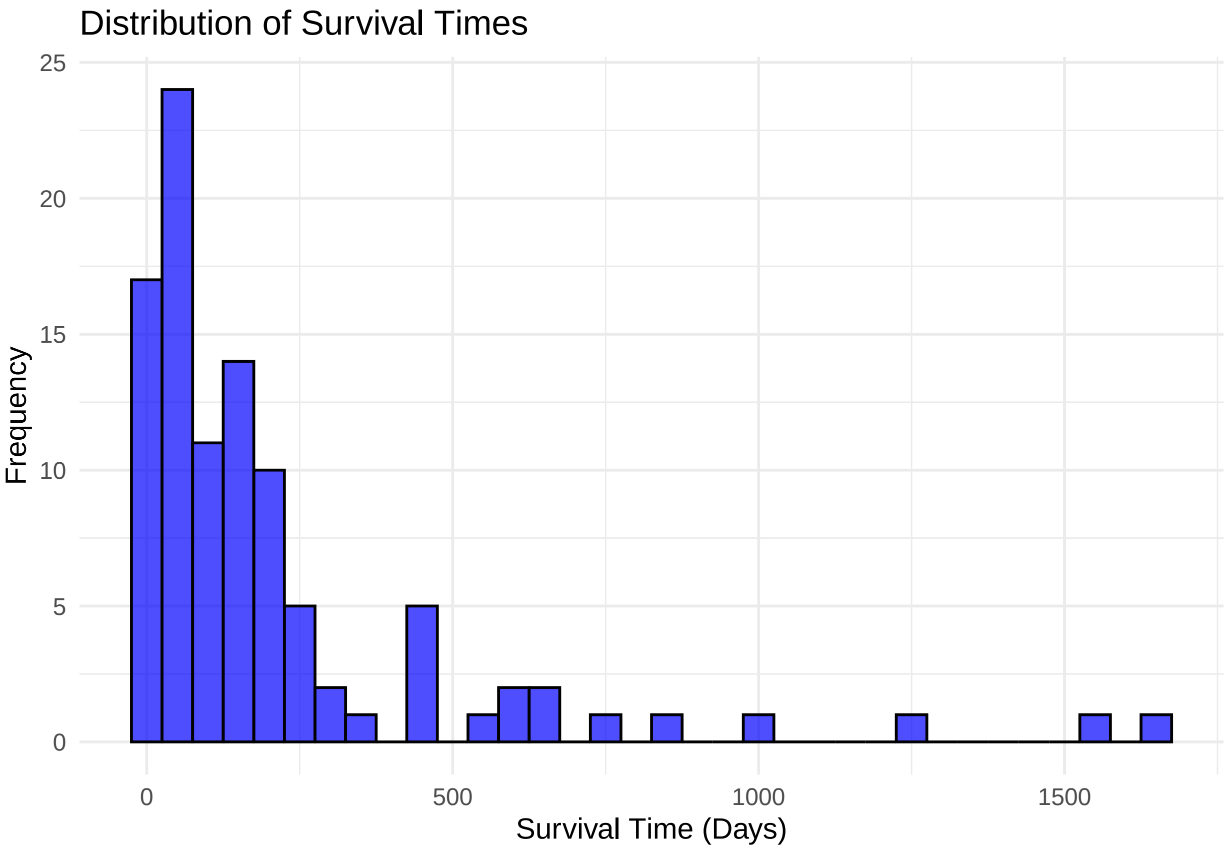
The observed Kaplan-Meier survival curves are consistent with the predictions made by the Exponential AFT model, so the distribution assumption checks out.

Lastly, let us do a Goodness of Fit Test.

The p-value is very significantly <0.05. we can therefore conclude: - The full model (including Sex, Age, and Type) provides a statistically significant improvement in fit compared to the null model. - The combination of all variables significantly contributes to explaining the survival times in the cancer data. - The result supports the overall validity of the AFT exponential model, demonstrating that the included variables collectively improve the model's fit.

Overall, the AFT exponential model appears to be valid and well-fitted to the cancer data. It passes the key diagnostic checks and shows significant improvement over a simpler model.

```r
# 1. Histogram of Survival Times
ggplot(Cancer, aes(x = Survival)) +
  geom_histogram(binwidth = 50, fill = "blue", alpha = 0.7, color = "black") +
  labs(title = "Distribution of Survival Times", x = "Survival Time (Days)", y = "Frequency") +
  theme_minimal()
```


Distribution of Survival Times
Frequency
25
20
15
10
5
0
0
500
1000
1500
Survival Time (Days)