Exploration of Proportional Hazards Model

Prepared by Tee Tian En

Table of Contents

1. Exploratory Data Analysis (EDA)

To start, we initialize our R environment by preparing the required libraries. The single generated survival dataset contains 42 observations, where durations can fall on any integer between 1 and 45, and is saved in **"simdata"**.

```
# Load library
library(coxed); library(survival); library(ggplot2)

# Set seed (57217)
set.seed(57217)

# Generate a survival dataset
simdata <- sim.survdata(N = 42, T = 45, num.data.frames = 1)$data
head(simdata, n = 5)
```

```
##           X1          X2          X3  y failed
## 1 -0.4191877  0.59973073  0.47023618 29   TRUE
## 2  1.2845013 -0.07866052 -0.06827143  4  FALSE
## 3 -0.2021333 -0.36243200 -0.24259505 29   TRUE
## 4  0.1955943  0.03445695  0.14416335  4   TRUE
## 5 -0.8798974  0.51863777  0.06074915 28   TRUE
```

The **"simdata"** consists of five columns, which are three covariates (X1, X2, X3), one survival time (y), and one event indicator (failed). In the event indicator (failed), FALSE means a censored observation. The summary of the "simdata" is provided below.

```
summary(simdata)
```

```
##        X1                  X2                  X3                  y         
##  Min.   :-1.19299   Min.   :-1.192959   Min.   :-1.160831   Min.   : 1.00  
##  1st Qu.:-0.24903   1st Qu.:-0.289067   1st Qu.:-0.237923   1st Qu.: 2.25  
##  Median : 0.02392   Median : 0.006619   Median :-0.029730   Median :21.00  
##  Mean   :-0.01614   Mean   :-0.013396   Mean   : 0.004191   Mean   :17.29  
##  3rd Qu.: 0.27319   3rd Qu.: 0.207220   3rd Qu.: 0.271781   3rd Qu.:31.00  
##  Max.   : 1.28450   Max.   : 1.365749   Max.   : 0.892033   Max.   :40.00  
##    failed
##  Mode :logical
##  FALSE:6
##  TRUE :36
##
##
##
```

Besides, there is no correlation between the covariates as all of the correlation's values are close to zero. The Kaplan-Meier survival curve without any covariates is plotted to provides a comprehensive visual representation of the survival probabilities over time.

```
cor(simdata[, c("X1", "X2", "X3")])

##            X1         X2         X3
## X1  1.0000000 -0.0733872 -0.1427242
## X2 -0.0733872  1.0000000  0.3229935
## X3 -0.1427242  0.3229935  1.0000000

km_fit <- survfit(Surv(y, failed) ~ 1, data=simdata)
plot(km_fit, xlab="Time", ylab="Survival Probability",
     main="Kaplan-Meier Survival Curve")
```

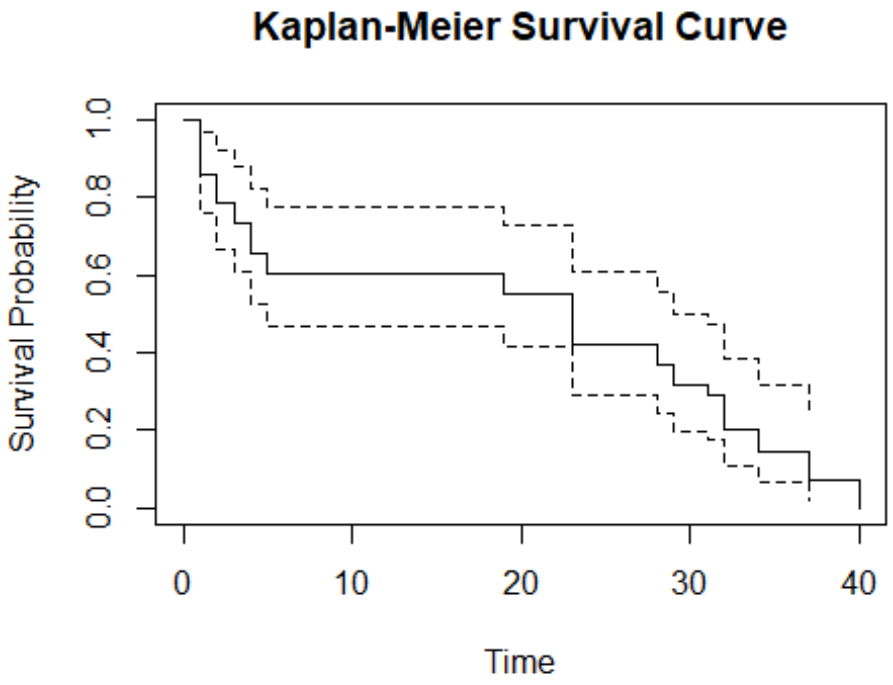

Figure 1: Kaplan-Meier survival curve.

At time 0, the survival probability is 1 (100%), indicating that all subjects are initially event-free. The first drop occurs at time 1, and half of the subjects experienced the event by time 21 (median survival time).

```r
# Histograms for each covariate
ggplot(simdata, aes(x=X1)) + geom_histogram(binwidth=1) +
  ggtitle("Distribution of X1") + theme_classic()
ggplot(simdata, aes(x=X2)) + geom_histogram(binwidth=1) +
  ggtitle("Distribution of X2") + theme_classic()
ggplot(simdata, aes(x=X3)) + geom_histogram(binwidth=1) +
  ggtitle("Distribution of X3") + theme_classic()

# Histogram for survival times
ggplot(simdata, aes(x=y)) + geom_histogram(binwidth=1) +
  ggtitle("Distribution of Survival Times") + theme_classic()
```

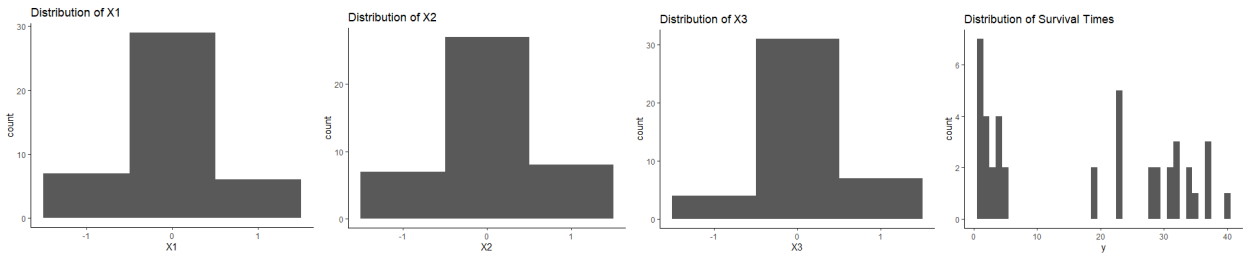

Figure 2: Distribution of covariates and survival times.

The histograms illustrate that the covariates X1, X2, and X3 have approximately normal distributions, while the survival times y has first peak near 0 suggests a right skewness, with a long tail extending towards higher survival times.

2. Fitting using Proportional Hazards Model

Fitted model: for $i = 1, \cdots, 42$,

$$h_i(t) = \exp(\beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i) h_0(t).$$

```
# Proportional Hazards Model
phm <- coxph(formula = Surv(y, failed) ~ X1 + X2 + X3,
             data = simdata, method = "breslow")
summary(phm)
```

```
## Call:
## coxph(formula = Surv(y, failed) ~ X1 + X2 + X3, data = simdata,
##     method = "breslow")
##
##   n= 42, number of events= 36
##
##       coef exp(coef) se(coef)      z Pr(>|z|)
## X1  0.4999    1.6485   0.3619  1.381    0.167
## X2  0.2897    1.3360   0.3684  0.786    0.432
## X3 -0.4164    0.6594   0.4259 -0.978    0.328
##
##    exp(coef) exp(-coef) lower .95 upper .95
## X1    1.6485     0.6066    0.8110     3.351
## X2    1.3360     0.7485    0.6490     2.750
## X3    0.6594     1.5165    0.2862     1.520
##
## Concordance= 0.605  (se = 0.059 )
## Likelihood ratio test= 3.33  on 3 df,   p=0.3
## Wald test            = 3.17  on 3 df,   p=0.4
## Score (logrank) test = 3.14  on 3 df,   p=0.4
```

Interpretation:

1. X1: $\psi = \exp(0.4999) = 1.6485$. For the same X2 and X3, the hazard for someone with unit $k + 1$ of X1 is approximately 65% higher compared to someone with unit k of X1. However, the p-value for X1 is 0.167, which is greater than 0.05, indicating that this effect is not statistically significant at the 5% level.
2. X2: $\psi = \exp(0.2897) = 1.3360$. For the same X1 and X3, the hazard for someone with unit $k + 1$ of X2 is approximately 34% higher compared to someone with unit k of X2. The p-value for X2 is 0.432, which is also greater than 0.05, indicating that this effect is not statistically significant.
3. X3: $\psi = \exp(-0.4164) = 0.6594$. For the same X1 and X2, the hazard for someone with unit $k + 1$ of X3 is approximately 34% lower compared to someone with unit k of X3. The p-value for X3 is 0.328, which is again greater than 0.05, indicating that this effect is not statistically significant.

The 95% confidence interval for the hazard ratio of X1, X2, and X3 are (0.8110, 3.351), (0.6490, 2.750), and (0.2862, 1.520), respectively. These wide intervals include 1, indicating no statistically significant effect.

This initial full model included all covariates, but none significantly contributed to explaining the variation in survival times. The wide confidence intervals and the non-significant p-values imply that these associations are not statistically significant.

The model fit statistics (likelihood ratio test, Wald test, score test) all have high p-values, suggesting that the model does not fit the data well.

3. Reduced Model and Interpretation

To find the reduced model, we start with the initial full model and compare the AIC for each covariate removed or added into the model. We stop the process if the AIC is the smallest when we did not add or remove any covariate. Three directions (both, backward, forward) have been considered to confirm our findings.

```
# Reduced model (both direction)
reduced_model <- step(phm, direction = "both")
```

```
## Start:  AIC=212.41
## Surv(y, failed) ~ X1 + X2 + X3
##
##        Df    AIC
## - X2    1 211.03
## - X3    1 211.37
## - X1    1 212.34
## <none>    212.41
##
## Step:  AIC=211.03
## Surv(y, failed) ~ X1 + X3
##
##        Df    AIC
## - X3    1 209.58
## - X1    1 210.75
## <none>    211.03
## + X2    1 212.41
##
## Step:  AIC=209.58
## Surv(y, failed) ~ X1
##
##        Df    AIC
## <none>    209.58
## - X1    1 209.75
## + X3    1 211.03
## + X2    1 211.37
```

```
# Reduced model (backward direction)
reduced_model2 <- step(phm, direction = "backward")
```

```
## Start:  AIC=212.41
## Surv(y, failed) ~ X1 + X2 + X3
##
##        Df    AIC
## - X2    1 211.03
## - X3    1 211.37
## - X1    1 212.34
## <none>    212.41
##
```

```
## Step:  AIC=211.03
## Surv(y, failed) ~ X1 + X3
##
##        Df    AIC
## - X3    1 209.58
## - X1    1 210.75
## <none>    211.03
##
## Step:  AIC=209.58
## Surv(y, failed) ~ X1
##
##        Df    AIC
## <none>    209.58
## - X1    1 209.75
```

```r
# Reduced model (forward direction)
nullphm <- coxph(formula = Surv(y, failed) ~ 1,
                 data = simdata, method = "breslow")
reduced_model3 <- step(nullphm, direction = "forward",
                       scope = ~ X1 + X2 + X3)
```

```
## Start:  AIC=209.75
## Surv(y, failed) ~ 1
##
##        Df    AIC
## + X1    1 209.58
## <none>    209.75
## + X3    1 210.75
## + X2    1 211.68
##
## Step:  AIC=209.58
## Surv(y, failed) ~ X1
##
##        Df    AIC
## <none>    209.58
## + X3    1 211.03
## + X2    1 211.37
```

From the directions of both and backward, the covariates have been removed in the same sequence, which is X2, and then X3. This leads to the model with only covariate X1 being the model with lowest AIC. From the direction of forward, the AIC slightly reduces by adding X1, hence it suggests that the model with covariate X1 is the best model. Reduced model: for $i = 1, \cdots, 42$,

$$h_i(t) = \exp(\beta_1 X1_i) h_0(t).$$

```
summary(reduced_model)
```

```
## Call:
## coxph(formula = Surv(y, failed) ~ X1, data = simdata, method = "breslow")
##
##   n= 42, number of events= 36
##
##       coef exp(coef) se(coef)     z Pr(>|z|)
## X1 0.5163    1.6758   0.3540 1.458    0.145
##
##    exp(coef) exp(-coef) lower .95 upper .95
## X1     1.676     0.5967    0.8373     3.354
##
## Concordance= 0.585  (se = 0.056 )
## Likelihood ratio test= 2.17  on 1 df,   p=0.1
## Wald test            = 2.13  on 1 df,   p=0.1
## Score (logrank) test = 2.13  on 1 df,   p=0.1
```

In short, we get the same reduced model from all of the directions, which is the model with covariate X1 (hazards ratio: 1.6758). The reduced model suggests that removing the covariates X2 and X3 does not worsen the model fit according to the AIC criterion, the same as adding the covariates X2 and X3. Note that the p-value of X1 is still 0.145 > 0.05, indicating no strong statistically significant.

4. Adequacy of the Reduced Model

For the reduced model, we still need to ensure that it's a reasonable fit given the data. The plot of martingale residuals against covariate X1 has been checked and it shows a linear relationship, suggesting no transformation of variable needed. Two other approaches are considered as follows.

4.1. Proportional Hazards Assumption:

The proportional hazards assumption for the reduced model is checked to ensure it is valid. We conducted the Schoenfeld residuals test to assess the proportional hazards assumption for each predictor variable and the overall model. The results are as follows:

```
# par(pty = "s")
plot(cox.zph(reduced_model))
```

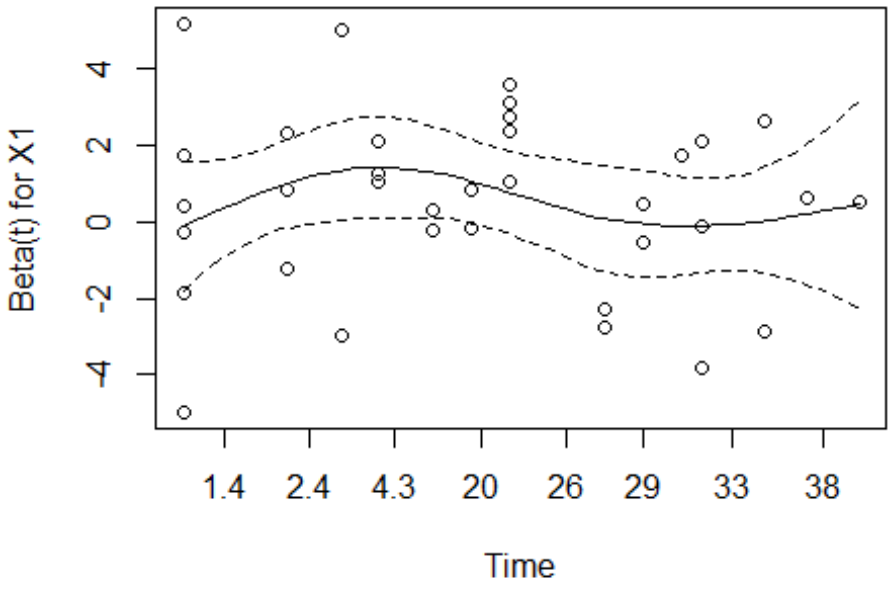

Figure 3: Schoenfeld residuals plot for X1, X2, X3.

```
cox.zph(reduced_model)

##        chisq df    p
## X1     0.198  1 0.66
## GLOBAL 0.198  1 0.66
```

The proportional hazards assumption holds for the model (the p-values 0.66 > 0.05). The Schoenfeld residuals plot for X1 showed an almost horizontal line, suggesting not significant violation of the proportional hazards assumption.

The proportional hazards assumption was tested, revealing no significant violations.

4.2. Goodness-of-Fit:

We can assess the goodness-of-fit by examining the deviance residuals.

```
# Deviance Residuals Plot (Reduced Model)
deviance_residuals <- resid(reduced_model, type="deviance")
plot(deviance_residuals, main="Deviance Residuals",
     ylab="Residuals", xlab="Index")
abline(h=0, col="red")
```

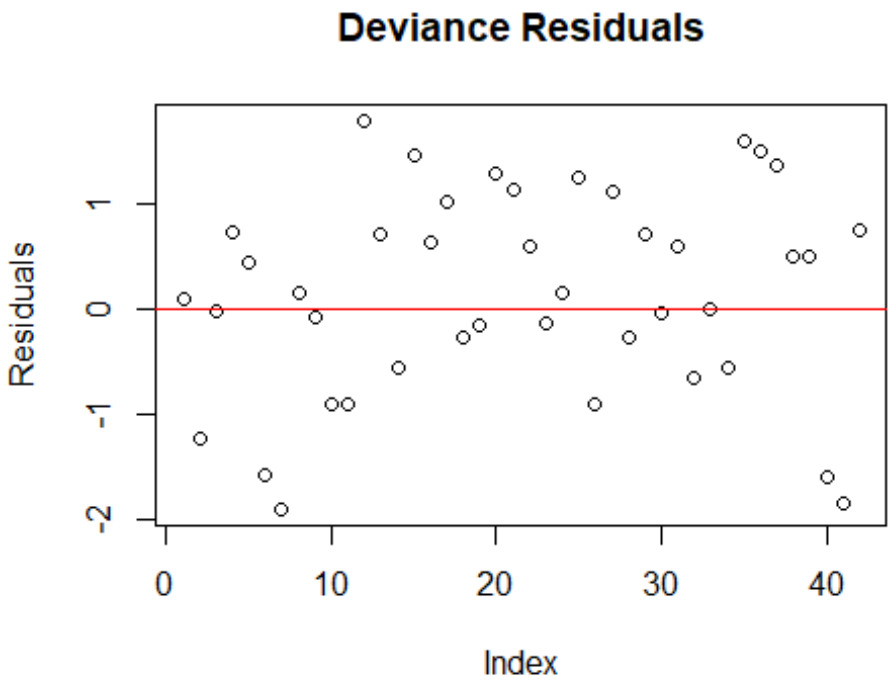

Figure 4: Deviance Residuals plot for reduced model.

The plot should show residuals scattered randomly around zero, indicating a reasonable fit for the reduced model.

5. Conclusion

In our analysis of the simulated dataset, the results indicated that only one of the covariates (X1) are statistically significant. This finding suggests that the covariates X2 and X3 do not significantly improve the model fit over the X1 only model.

The proportional hazards model (full model) is $h_i(t) = \exp(\beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i)h_0(t)$ for $i = 1, \cdots, 42$, while the reduced model is $h_i(t) = \exp(\beta_1 X1_i)h_0(t)$ for $i = 1, \cdots, 42$.

Although the p-value for X1 in the reduced model (AIC: 209.58) is not small enough to indicate strong statistical significance, the AIC value does not improve much from the null model (AIC: 209.75). Nevertheless, we checked the proportional hazards assumption and examined the deviance residuals plot, both of which suggested that the reduced model may be a good fit.

In conclusion, while the reduced model with the covariate X1 appears to be a reasonable fit, further exploration is necessary to determine whether it is practically significant to include this covariate in the model. Additional data or covariates may be needed to improve the model's explanatory power and practical relevance.

6. References

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370. https://doi.org/10.1007/BF02294361

Fine, J., & Gray, R. (1999). A Proportional Hazards Model for the Subdistribution of a Competing Risk. *Journal of the American Statistical Association, 94*, 496-509. https://doi.org/10.1080/01621459.1999.10474144