Semi-parametric Cox regression model

Samuel Kong

2025-02-28

```r
# Load required libraries
library(survival)   # For Cox regression and survival analysis
library(survminer)  # For visualization of survival models
```

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

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

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

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

```r
library(dplyr)      # For data manipulation
```

```
##
## 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
```

```r
# Load the dataset from the survival package
data("colon", package = "survival")
```

```
## Warning in data("colon", package = "survival"): data set 'colon' not found
```

```r
# Check if the dataset is loaded
head(colon)  # View first few rows
```

```
##   id study      rx sex age obstruct perfor adhere nodes status differ extent
## 1  1     1 Lev+5FU   1  43        0      0      0     5      1      2      3
## 2  1     1 Lev+5FU   1  43        0      0      0     5      1      2      3
## 3  2     1 Lev+5FU   1  63        0      0      0     1      0      2      3
## 4  2     1 Lev+5FU   1  63        0      0      0     1      0      2      3
## 5  3     1     Obs   0  71        0      0      1     7      1      2      2
## 6  3     1     Obs   0  71        0      0      1     7      1      2      2
##   surg node4 time etype
## 1    0     1 1521     2
## 2    0     1  968     1
## 3    0     0 3087     2
## 4    0     0 3087     1
## 5    0     1  963     2
## 6    0     1  542     1
```

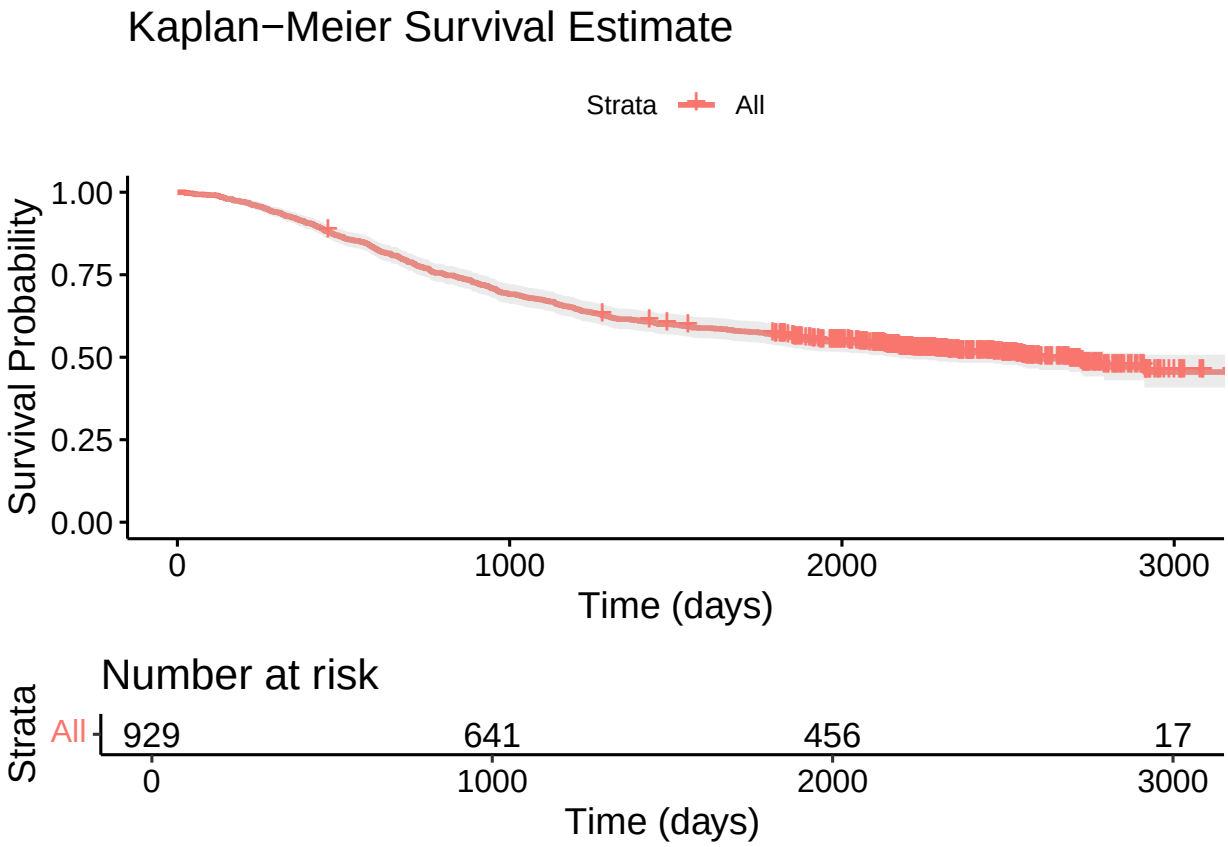
```r
# Subset relevant observations (e.g., patients from a clinical trial)
colon_data <- subset(colon, etype == 2)

# Convert categorical variables
colon_data$sex <- factor(colon_data$sex, labels = c("Male", "Female"))
colon_data$rx <- factor(colon_data$rx, labels = c("Obs", "Lev", "Lev+5FU"))
```

```r
# Install survminer if not installed
install.packages("survminer")
```

```
## Warning: package 'survminer' is in use and will not be installed
```

```r
# Load survminer library
library(survminer)
```

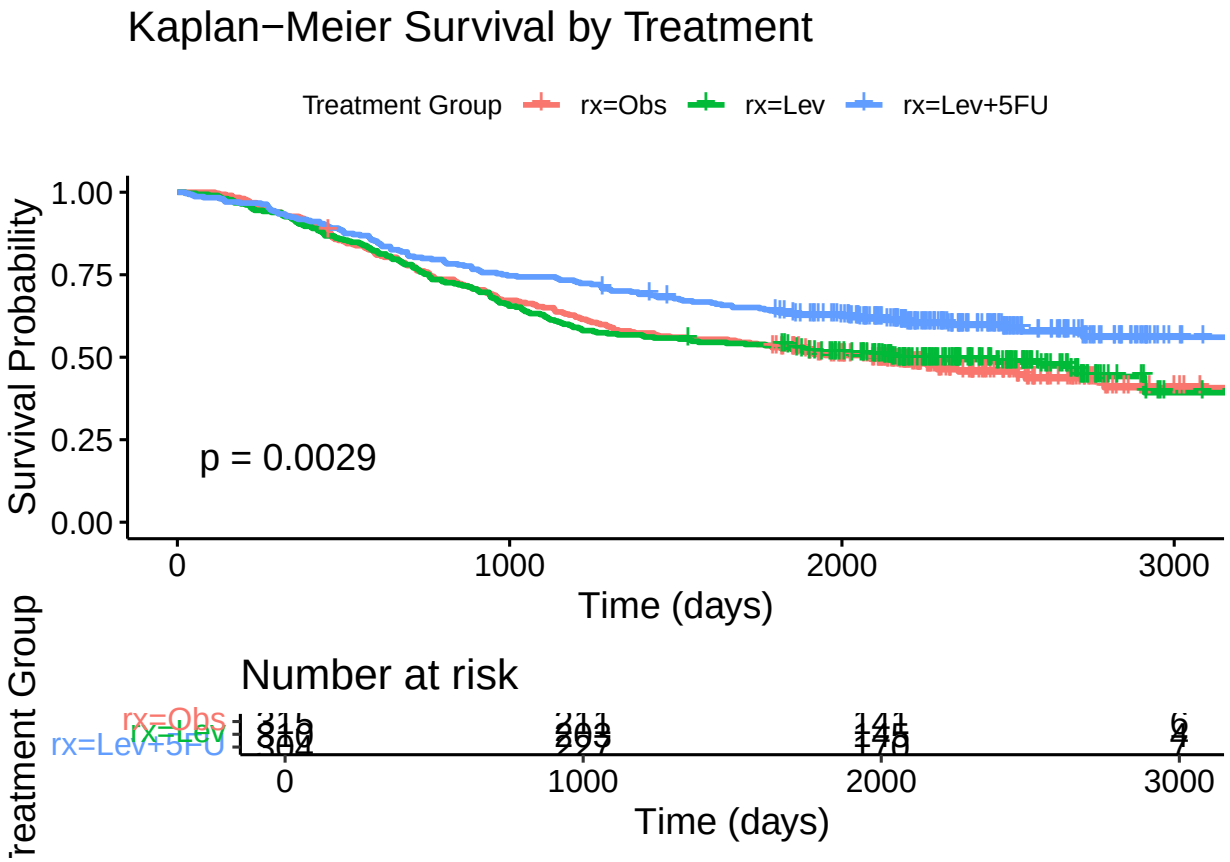
```r
# Kaplan-Meier Survival Estimate
km_fit <- survfit(Surv(time, status) ~ 1, data = colon_data)

# Plot Kaplan-Meier curve
ggsurvplot(km_fit, conf.int = TRUE, risk.table = TRUE,
           title = "Kaplan-Meier Survival Estimate",
           xlab = "Time (days)", ylab = "Survival Probability")
```




```r
# Kaplan-Meier estimate by treatment group
km_group_fit <- survfit(Surv(time, status) ~ rx, data = colon_data)

# Log-rank test for survival difference
survdiff(Surv(time, status) ~ rx, data = colon_data)
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ rx, data = colon_data)
##
##              N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=Obs     315      168      148      2.58      3.85
## rx=Lev     310      161      146      1.52      2.25
## rx=Lev+5FU 304      123      157      7.55     11.62
##
##  Chisq= 11.7  on 2 degrees of freedom, p= 0.003
```

```r
# Plot survival curves by treatment group
ggsurvplot(km_group_fit, pval = TRUE, risk.table = TRUE,
           title = "Kaplan-Meier Survival by Treatment",
           legend.title = "Treatment Group",
           xlab = "Time (days)", ylab = "Survival Probability")
```

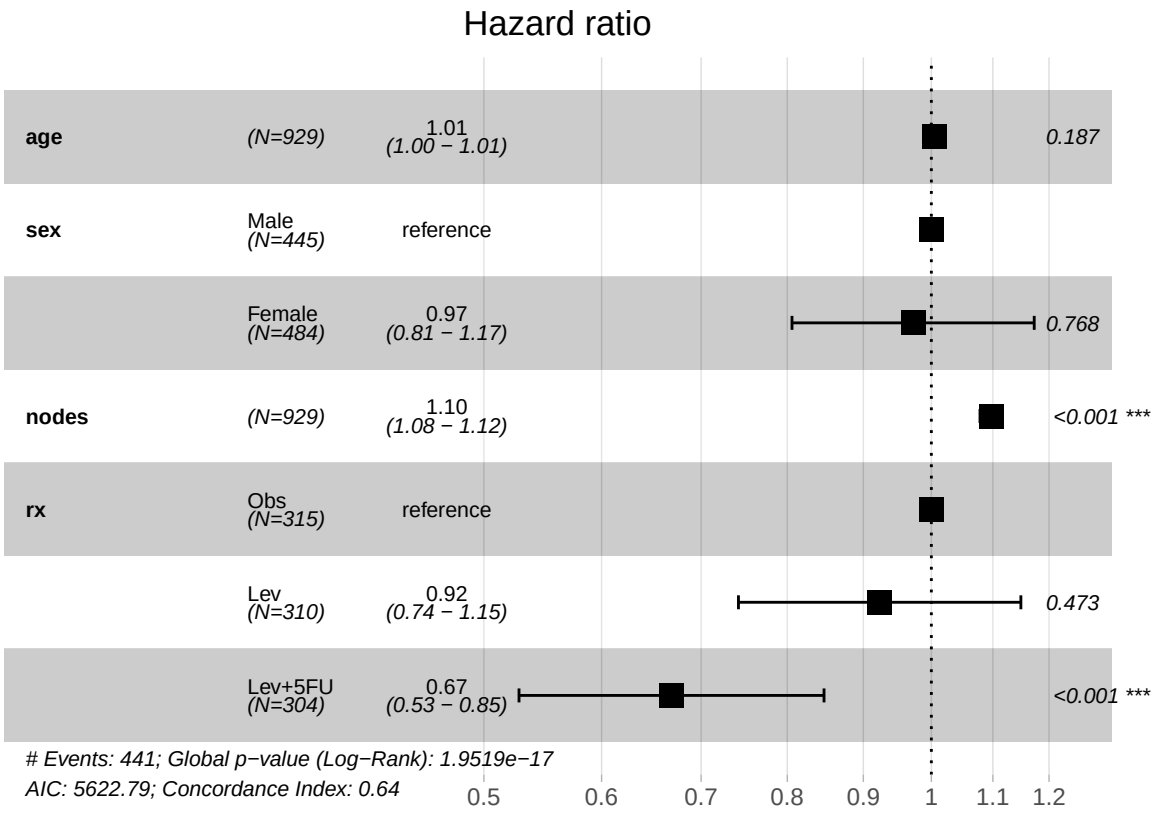



```
# Fit Cox model
cox_model <- coxph(Surv(time, status) ~ age + sex + nodes + rx, data = colon_data)

# Model summary
summary(cox_model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ age + sex + nodes + rx,
##     data = colon_data)
##
##   n= 911, number of events= 441
##    (18 observations deleted due to missingness)
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## age         0.005333  1.005347  0.004045  1.318  0.18739
## sexFemale -0.028257  0.972138  0.095728 -0.295  0.76786
## nodes       0.092755  1.097193  0.008871 10.456  < 2e-16 ***
## rxLev      -0.080072  0.923049  0.111613 -0.717  0.47312
## rxLev+5FU -0.402527  0.668628  0.120539 -3.339  0.00084 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          1.0053     0.9947    0.9974    1.0134
## sexFemale    0.9721     1.0287    0.8058    1.1728
## nodes        1.0972     0.9114    1.0783    1.1164
```

```
## rxLev        0.9230     1.0834    0.7417    1.1488
## rxLev+5FU    0.6686     1.4956    0.5279    0.8468
##
## Concordance= 0.638  (se = 0.014 )
## Likelihood ratio test= 87.79  on 5 df,   p=<2e-16
## Wald test            = 123  on 5 df,   p=<2e-16
## Score (logrank) test = 125  on 5 df,   p=<2e-16
```

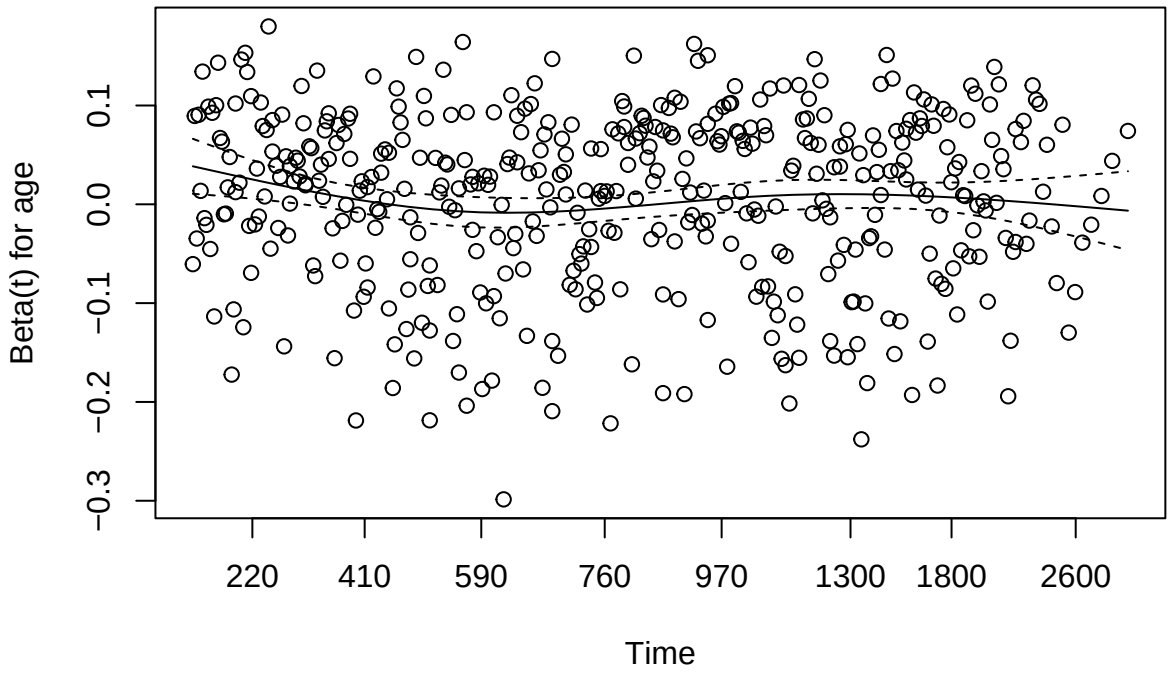
```r
# Visualize Cox model results
ggforest(cox_model, data = colon_data)
```

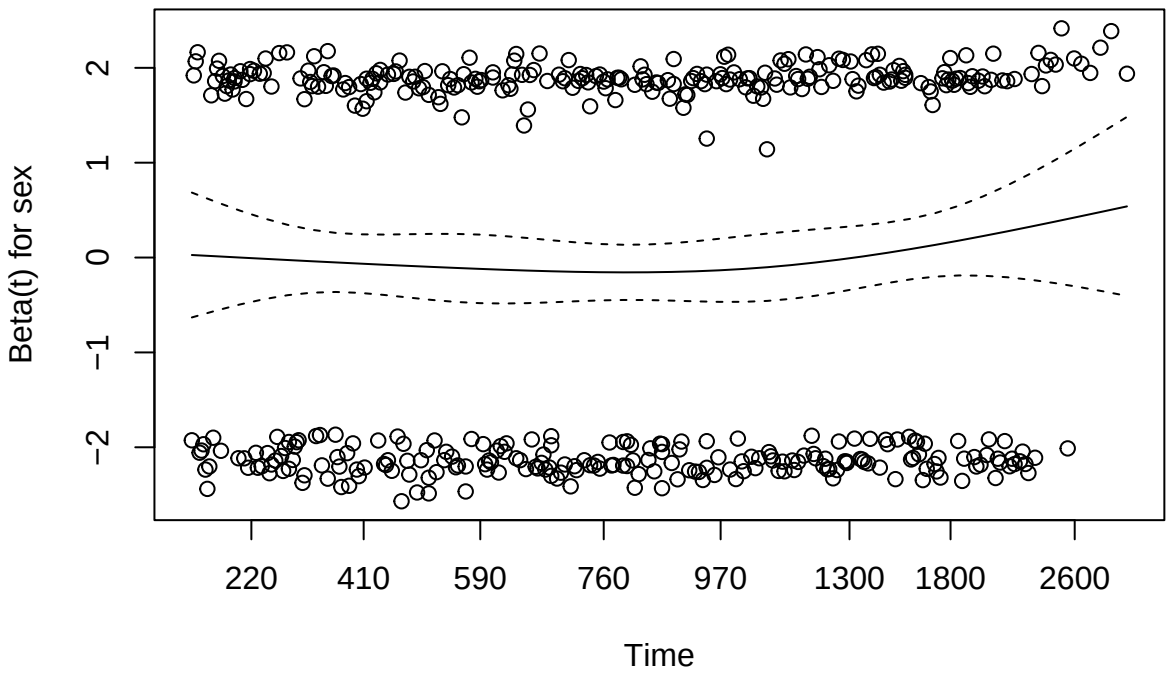


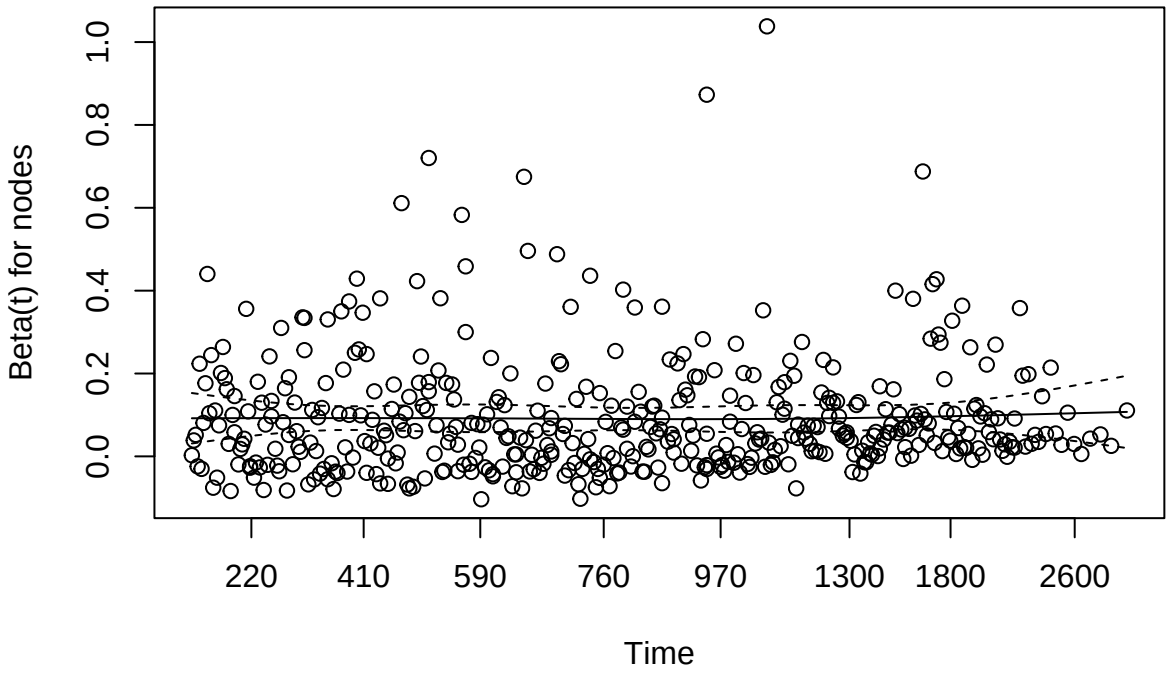

```r
# Test for proportional hazards assumption
cox_zph <- cox.zph(cox_model)

# Print results
print(cox_zph)
```

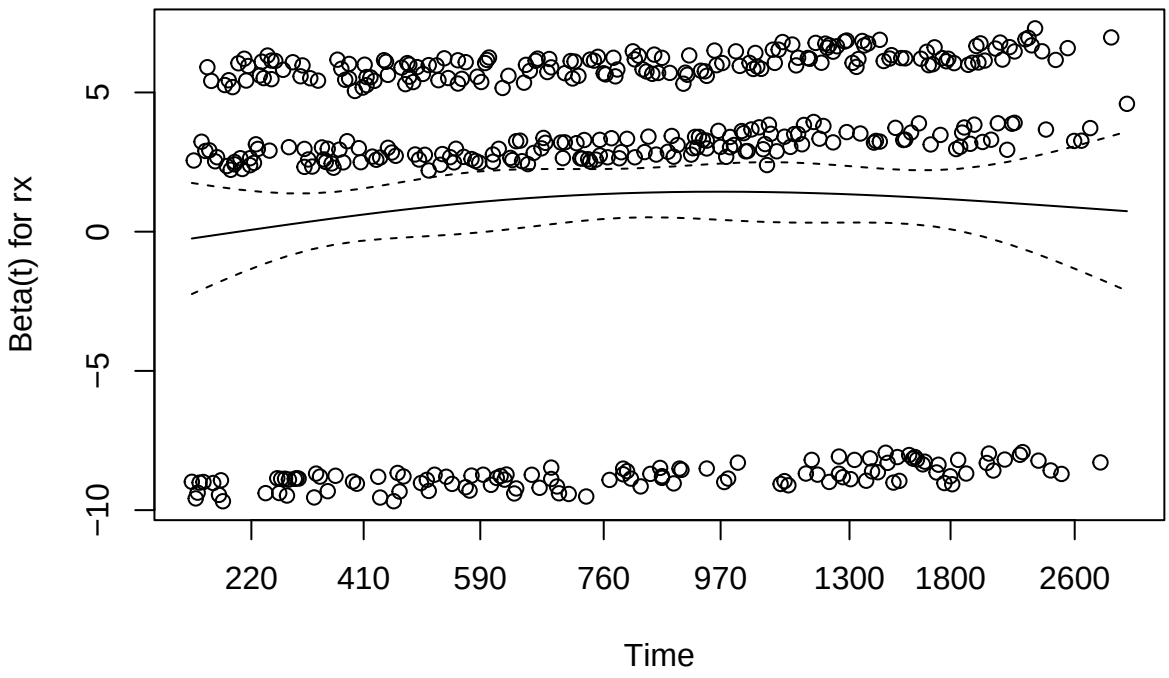
```
##         chisq df    p
## age     0.546  1 0.46
## sex     0.719  1 0.40
## nodes   0.156  1 0.69
## rx      2.332  2 0.31
## GLOBAL 3.687  5 0.60
```

```r
# Plot Schoenfeld residuals
plot(cox_zph)
```




Beta(t) for sex
2
1
0
−1
−2
220
410
590
760
970
1300
1800
2600
Time


Beta(t) for nodes
0.0
0.2
0.4
0.6
0.8
1.0
220
410
590
760
970
1300
1800
2600
Time




```r
# Concordance index (C-index)
cat("Concordance Index:", summary(cox_model)$concordance[1])
```

```
## Concordance Index: 0.6382539
```

```r
# Log-Likelihood Test
anova(cox_model)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(time, status)
## Terms added sequentially (first to last)
##
##        loglik   Chisq Df Pr(>|Chi|)
## NULL  -2850.3
## age   -2850.2  0.2522  1   0.615507
## sex   -2850.2  0.0018  1   0.966312
## nodes -2812.6 75.1026  1  < 2.2e-16 ***
## rx    -2806.4 12.4378  2   0.001991 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Load the dataset
data(colon, package = "survival")
```

```
## Warning in data(colon, package = "survival"): data set 'colon' not found
```

```r
# Inspect the first few rows of the dataset
head(colon)
```

```
##   id study      rx sex age obstruct perfor adhere nodes status differ extent
## 1  1     1 Lev+5FU   1  43        0      0      0     5      1      2      3
## 2  1     1 Lev+5FU   1  43        0      0      0     5      1      2      3
## 3  2     1 Lev+5FU   1  63        0      0      0     1      0      2      3
## 4  2     1 Lev+5FU   1  63        0      0      0     1      0      2      3
## 5  3     1     Obs   0  71        0      0      1     7      1      2      2
## 6  3     1     Obs   0  71        0      0      1     7      1      2      2
##   surg node4 time etype
## 1    0     1 1521     2
## 2    0     1  968     1
## 3    0     0 3087     2
## 4    0     0 3087     1
## 5    0     1  963     2
## 6    0     1  542     1
```

```r
# Subset the dataset to remove missing values and exclude non-useful variables
colon_clean <- colon %>%
  filter(etype == 1) %>%  # Select only recurrence or death as event of interest
  select(time, status, age, sex, obstruct, perfor, nodes, rx)

# Convert categorical variables to factors
colon_clean$sex <- factor(colon_clean$sex, labels = c("Male", "Female"))
colon_clean$rx <- factor(colon_clean$rx, labels = c("Obs", "Lev", "Lev+5FU"))
colon_clean$obstruct <- as.factor(colon_clean$obstruct)
colon_clean$perfor <- as.factor(colon_clean$perfor)

# Inspect the cleaned data
str(colon_clean)
```

```
## 'data.frame':    929 obs. of  8 variables:
##  $ time    : num  968 3087 542 245 523 ...
##  $ status  : num  1 0 1 1 1 1 1 0 0 0 ...
##  $ age     : num  43 63 71 66 69 57 77 54 46 68 ...
##  $ sex     : Factor w/ 2 levels "Male","Female": 2 2 1 1 2 1 2 2 2 1 ...
##  $ obstruct: Factor w/ 2 levels "0","1": 1 1 1 2 1 1 1 1 1 1 ...
##  $ perfor  : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
##  $ nodes   : num  5 1 7 6 22 9 5 1 2 1 ...
##  $ rx      : Factor w/ 3 levels "Obs","Lev","Lev+5FU": 3 3 1 3 1 3 2 1 2 3 ...
```

```r
# Fit a Cox Proportional Hazards model
cox_model <- coxph(Surv(time, status) ~ age + sex + obstruct + perfor + nodes + rx, data = colon_clean)

# Display model summary
summary(cox_model)
```

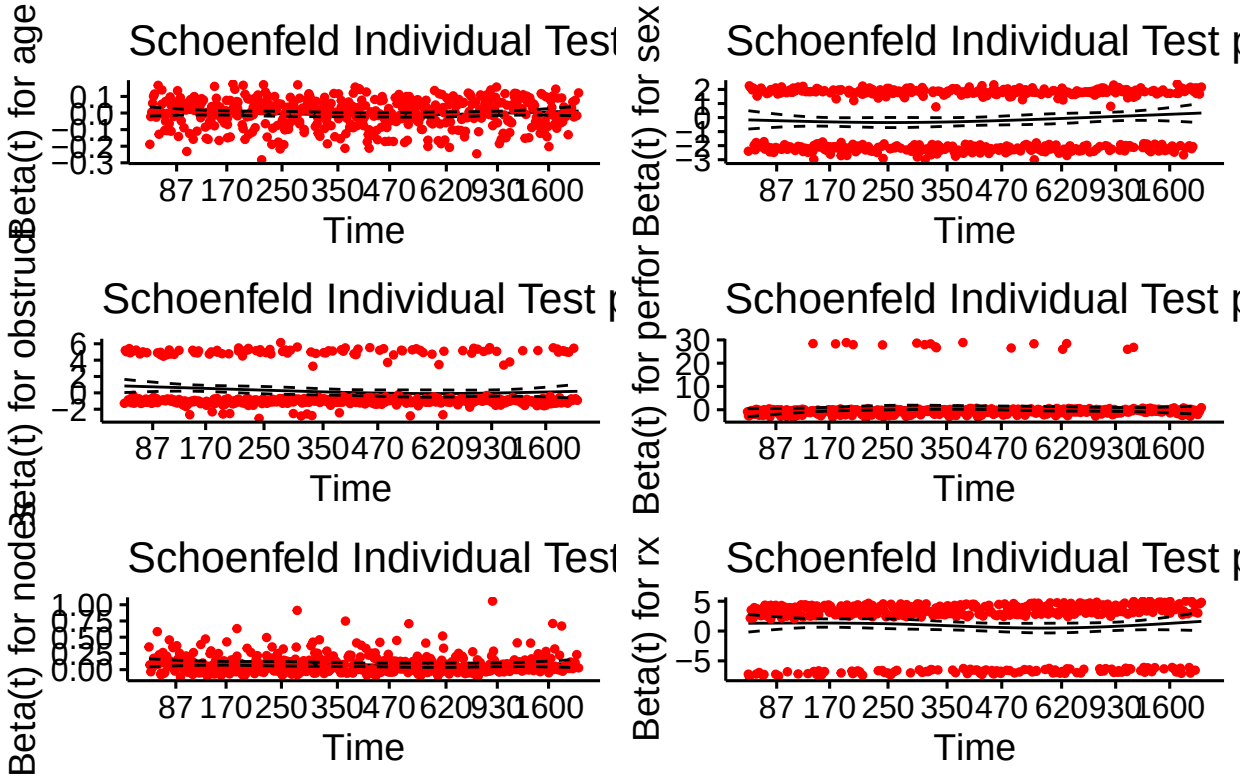
```
## Call:
## coxph(formula = Surv(time, status) ~ age + sex + obstruct + perfor +
##     nodes + rx, data = colon_clean)
##
```

```
##   n= 911, number of events= 456
##    (18 observations deleted due to missingness)
##
##                coef exp(coef)  se(coef)      z Pr(>|z|)
## age       -0.002657  0.997346  0.003962 -0.671   0.5024
## sexFemale -0.149270  0.861336  0.094463 -1.580   0.1141
## obstruct1  0.218168  1.243796  0.116523  1.872   0.0612 .
## perfor1    0.323247  1.381607  0.250331  1.291   0.1966
## nodes      0.084077  1.087713  0.008905  9.442  < 2e-16 ***
## rxLev     -0.063643  0.938340  0.108741 -0.585   0.5584
## rxLev+5FU -0.535193  0.585556  0.120638 -4.436 9.15e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##           exp(coef) exp(-coef) lower .95 upper .95
## age          0.9973     1.0027    0.9896    1.0051
## sexFemale    0.8613     1.1610    0.7158    1.0365
## obstruct1    1.2438     0.8040    0.9898    1.5629
## perfor1      1.3816     0.7238    0.8459    2.2567
## nodes        1.0877     0.9194    1.0689    1.1069
## rxLev        0.9383     1.0657    0.7582    1.1612
## rxLev+5FU    0.5856     1.7078    0.4623    0.7417
##
## Concordance= 0.642  (se = 0.013 )
## Likelihood ratio test= 95.72  on 7 df,   p=<2e-16
## Wald test            = 121.4  on 7 df,   p=<2e-16
## Score (logrank) test = 123.1  on 7 df,   p=<2e-16
```

```
# Test the Proportional Hazards assumption
cox.zph(cox_model)
```

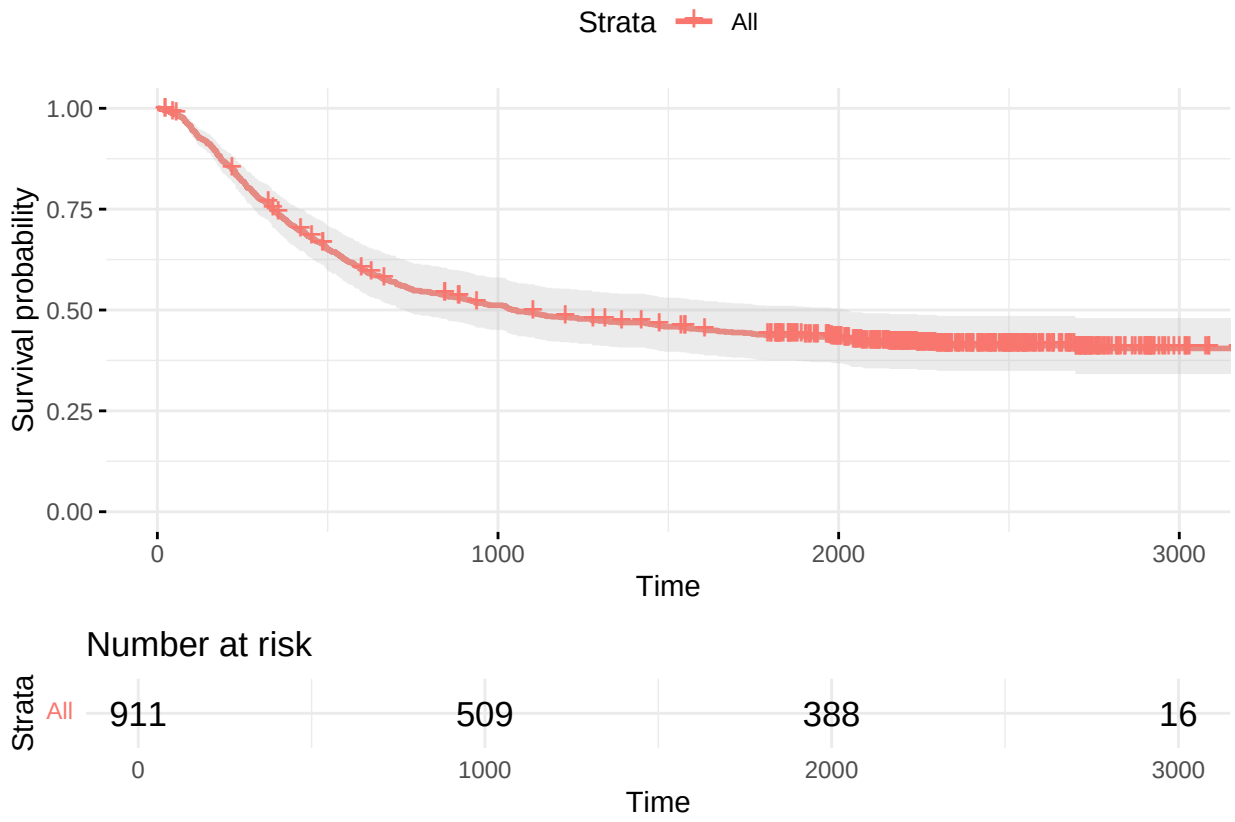
```
##            chisq df     p
## age      0.00123  1 0.972
## sex      3.01978  1 0.082
## obstruct 4.13539  1 0.042
## perfor   0.02894  1 0.865
## nodes    0.53520  1 0.464
## rx       0.46977  2 0.791
## GLOBAL   8.35473  7 0.302
```

```
# Visualize Schoenfeld residuals to check PH assumption
ggcoxzph(cox.zph(cox_model))
```

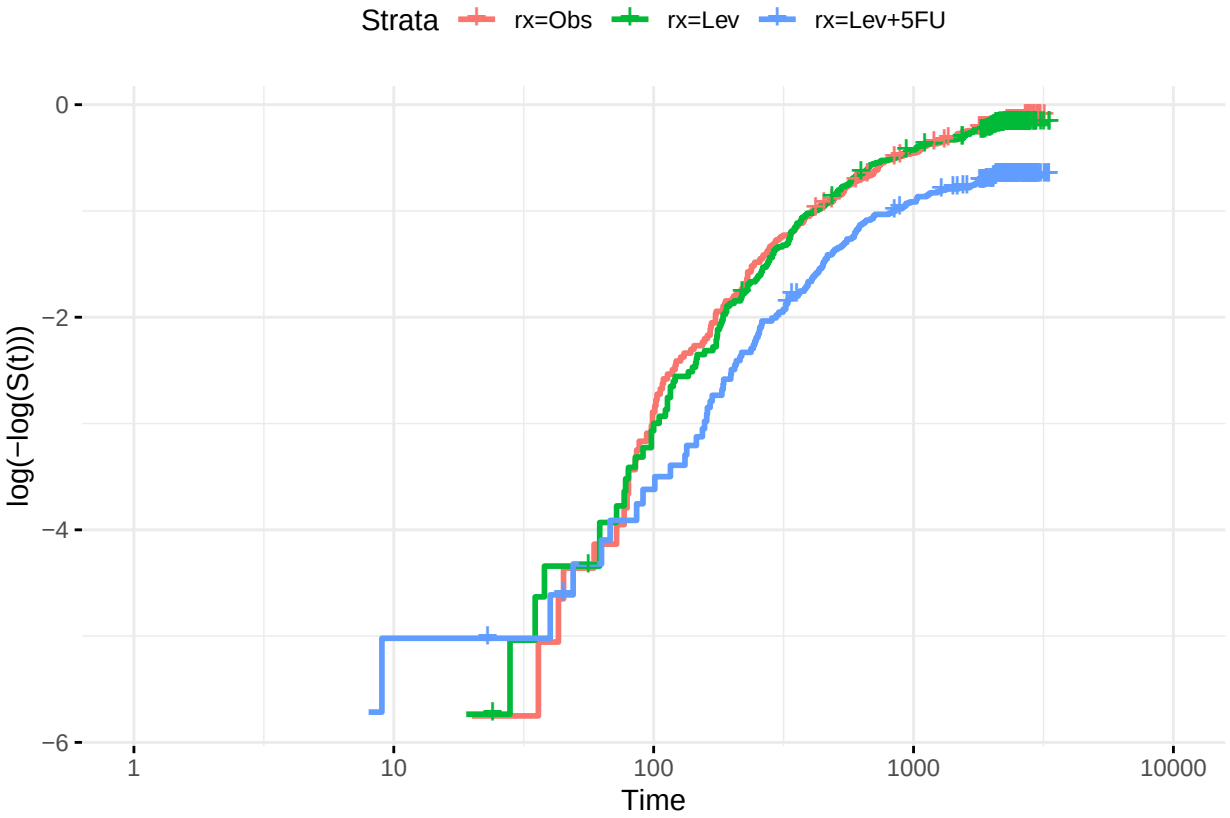



```
# Plot the survival curves for different treatment groups
ggsurvplot(survfit(cox_model), data = colon_clean, risk.table = TRUE, pval = TRUE,
           ggtheme = theme_minimal(), conf.int = TRUE)
```

```
## Warning in .pvalue(fit, data = data, method = method, pval = pval, pval.coord = pval.coord, : There a
##  This is a null model.
```




```
# Plot the log-log survival curves to assess proportionality visually
ggsurvplot(survfit(Surv(time, status) ~ rx, data = colon_clean), fun = "cloglog",
           ggtheme = theme_minimal(), conf.int = FALSE)
```




```
# Print Hazard Ratios (HR) with 95% confidence intervals
exp(coef(cox_model))
```

```
##       age sexFemale obstruct1   perfor1     nodes     rxLev rxLev+5FU
## 0.9973461 0.8613362 1.2437959 1.3816072 1.0877129 0.9383402 0.5855563
```

```
exp(confint(cox_model))
```

```
##                2.5 %    97.5 %
## age        0.9896315 1.0051207
## sexFemale  0.7157565 1.0365259
## obstruct1  0.9898391 1.5629088
## perfor1    0.8458684 2.2566610
## nodes      1.0688941 1.1068631
## rxLev      0.7582285 1.1612360
## rxLev+5FU  0.4622547 0.7417471
```