S24 survival analysis

Yuting WU

2025-03-10

#This R script performs Survival Analysis with a focus on nonparametric comparison of 2 or more groups

#1. Load and process the data, check missing values, remove rows with missing values

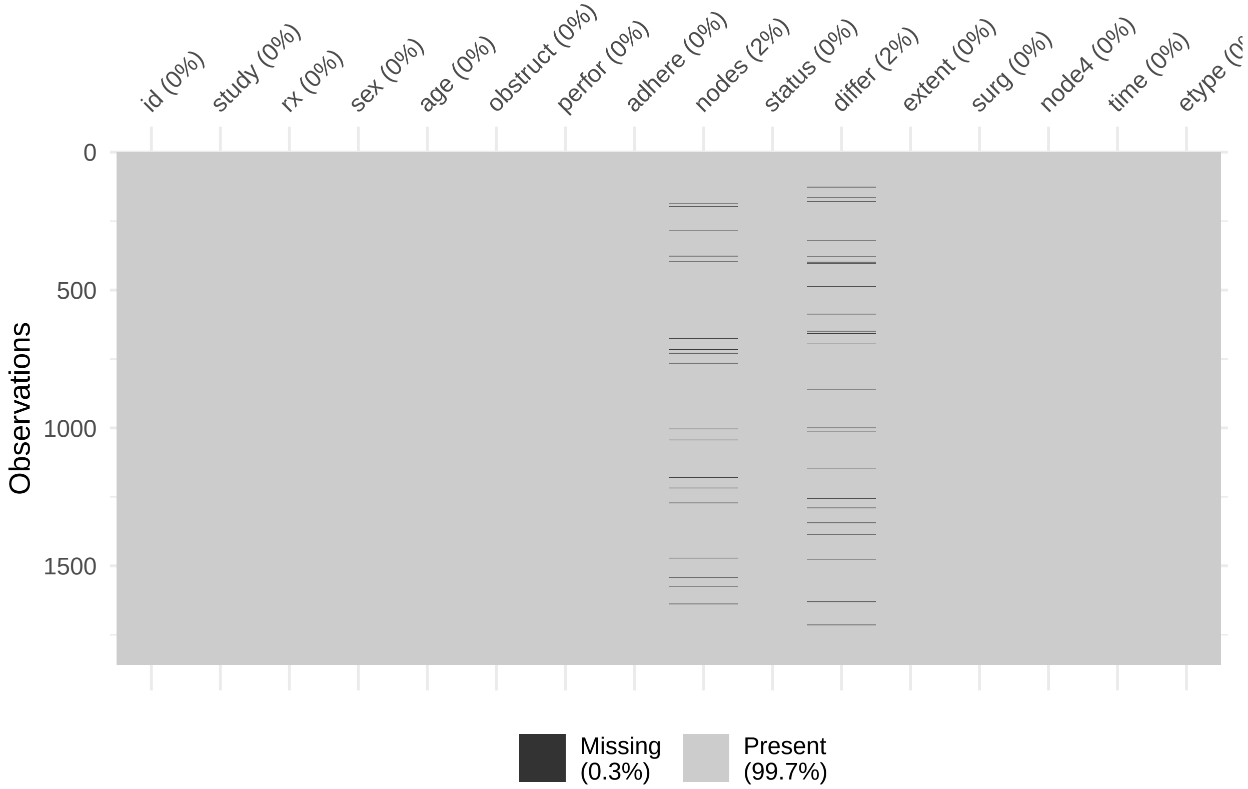
```
colon <- read.csv("C:/Users/ywu09/Downloads/dataset-42025.csv")
summary(colon)
```

```
##        id          study         rx                 sex              age
##  Min.   :  1   Min.   :1   Length:1858        Min.   :0.000   Min.   :18.00
##  1st Qu.:233   1st Qu.:1   Class :character   1st Qu.:0.000   1st Qu.:53.00
##  Median :465   Median :1   Mode  :character   Median :1.000   Median :61.00
##  Mean   :465   Mean   :1                      Mean   :0.521   Mean   :59.75
##  3rd Qu.:697   3rd Qu.:1                      3rd Qu.:1.000   3rd Qu.:69.00
##  Max.   :929   Max.   :1                      Max.   :1.000   Max.   :85.00
##
##     obstruct          perfor            adhere           nodes
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.0000   Min.   : 0.00
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.: 1.00
##  Median :0.0000   Median :0.00000   Median :0.0000   Median : 2.00
##  Mean   :0.1938   Mean   :0.02906   Mean   :0.1453   Mean   : 3.66
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.: 5.00
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.0000   Max.   :33.00
##                                                      NA’s   :36
##      status           differ          extent            surg
##  Min.   :0.0000   Min.   :1.000   Min.   :1.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:0.0000
##  Median :0.0000   Median :2.000   Median :3.000   Median :0.0000
##  Mean   :0.4952   Mean   :2.063   Mean   :2.887   Mean   :0.2659
##  3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :3.000   Max.   :4.000   Max.   :1.0000
##                   NA’s   :46
##      node4             time          etype
##  Min.   :0.0000   Min.   :   8   Min.   :1.0
##  1st Qu.:0.0000   1st Qu.: 566   1st Qu.:1.0
##  Median :0.0000   Median :1855   Median :1.5
##  Mean   :0.2745   Mean   :1538   Mean   :1.5
##  3rd Qu.:1.0000   3rd Qu.:2331   3rd Qu.:2.0
##  Max.   :1.0000   Max.   :3329   Max.   :2.0
##
```

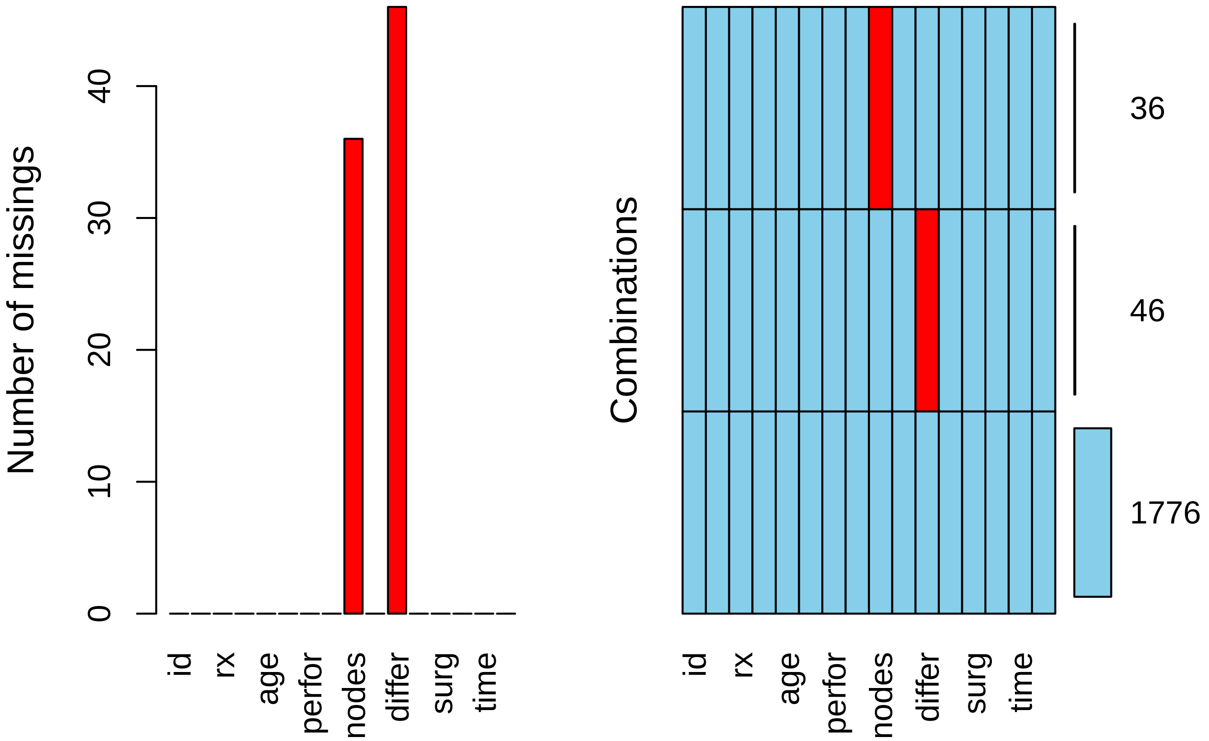
```
miss_var_summary(colon)
```

```
## # A tibble: 16 x 3
##    variable n_miss pct_miss
##    <chr>     <int>    <num>
##  1 differ       46     2.48
##  2 nodes        36     1.94
##  3 id            0     0
##  4 study         0     0
##  5 rx            0     0
##  6 sex           0     0
##  7 age           0     0
##  8 obstruct      0     0
##  9 perfor        0     0
## 10 adhere        0     0
## 11 status        0     0
## 12 extent        0     0
## 13 surg          0     0
## 14 node4         0     0
## 15 time          0     0
## 16 etype         0     0
```

```
vis_miss(colon)
```



```
aggr(colon, prop = FALSE, numbers = TRUE)
```



```
colon_death <- subset(colon, etype == 2)
sum(is.na(colon_death))
```

```
## [1] 41
```

```
#handling missing values: remove rows with any missing values
colon_complete <- na.omit(colon)
sum(is.na(colon_complete))
```

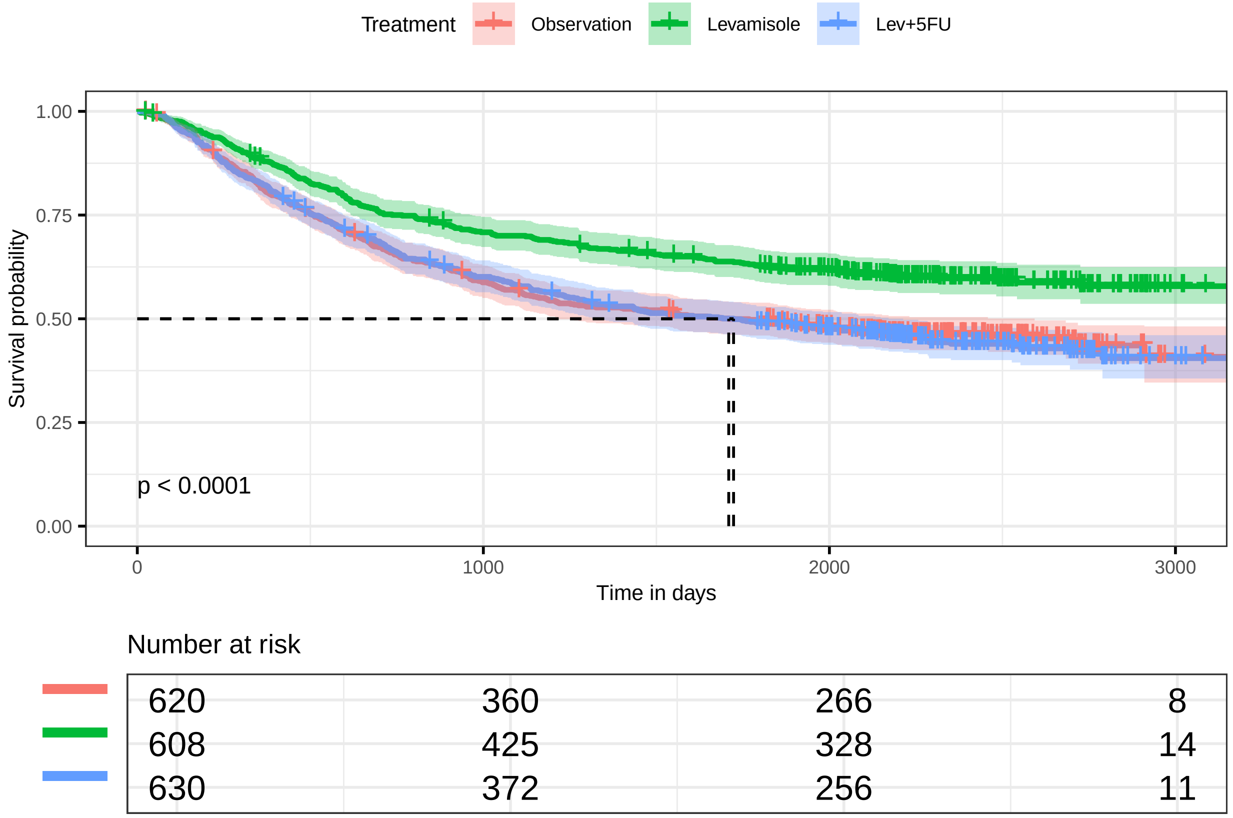
```
## [1] 0
```

#2. Nonparametric comparison of 2 or more groups_ treatment # Result: Significant difference in survival between treatment groups ##Lev+5FU vs. Observation: Significant survival benefit (adjusted p = 0.009). #all three observed groups has a survival rateat time 0 at 100% and decline overtime, but shows differences in rates. #“3000-day mark: #Levamisole group: ~60% survival probability #Observation group: ~43% survival probability #Lev+5FU group: ~40% survival probability

#conclusion: ##The Levamisole group shows a consistently better survival rate through the entire period, comparing with the 2 other groups.Observation vs Lev+5FU show relatively close result for the first 1500days, after 1500days Lev+5FU group shows slightly worse survival than observation group.

```
# Create a survival object
surv_obj <- Surv(time = colon$time, event = colon$status)
```

```r
fit_rx <- survfit(surv_obj ~ rx, data = colon)


smaller_theme <- theme(
  plot.title = element_text(size = 10),
  axis.title = element_text(size = 8),
  axis.text = element_text(size = 7),
  legend.title = element_text(size = 8),
  legend.text = element_text(size = 7),
  legend.position = "top"
)


km_plot <- ggsurvplot(fit_rx,
                     data = colon,
                     pval = TRUE,
                     risk.table = TRUE,
                     conf.int = TRUE,
                     xlab = "Time in days",
                     legend.title = "Treatment",
                     legend.labs = c("Observation", "Levamisole", "Lev+5FU"),
                     surv.median.line = "hv",

                     ggtheme = theme_bw() + smaller_theme,

                     tables.theme = theme_cleantable() + smaller_theme,
                     risk.table.y.text.col = TRUE,
                     risk.table.y.text = FALSE,
                     risk.table.height = 0.25,

                     pval.size = 3,
                     pval.coord = c(0, 0.1)
)


km_plot$table <- km_plot$table + theme(
  axis.text = element_text(size = 6),
  axis.title = element_text(size = 7)
)


print(km_plot)
```



##Log-Rank Test and Pairwise #The pairwise log-rank tests compare each pair of treatment groups to determine which specific group differences are driving the overall significant result. Result shows significant differences in survival bewteen treatment groups, and significant pairwise differences between certain treatment groups. #bonferroni : P-adjusted values < 0.05 indicate there's statistically significant differences even after taking into account multiple comparisons.

```
# Perform log-rank test
log_rank_test <- survdiff(surv_obj ~ rx, data = colon)
print(log_rank_test)
```

```
## Call:
## survdiff(formula = surv_obj ~ rx, data = colon)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## rx=Lev       620      333      295      4.93      7.26
## rx=Lev+5FU 608      242      326     21.61     33.54
## rx=Obs       630      345      299      7.01     10.40
##
##  Chisq= 33.6  on 2 degrees of freedom, p= 5e-08
```

```
# Create a function for pairwise log-rank tests
pairwise_logrank <- function(formula, data, ...) {
  groups <- unique(eval(formula[[3]], data))
  combinations <- combn(groups, 2)
  results <- data.frame(group1 = character(),
                        group2 = character(),
                        p.value = numeric(),
                        stringsAsFactors = FALSE)

  for(i in 1:ncol(combinations)) {
    g1 <- combinations[1, i]
    g2 <- combinations[2, i]
    subset_data <- data[eval(formula[[3]], data) %in% c(g1, g2), ]
    test <- survdiff(formula, data = subset_data, ...)
    p.val <- 1 - pchisq(test$chisq, length(test$n) - 1)
    results <- rbind(results, data.frame(group1 = g1, group2 = g2, p.value = p.val))
  }

  return(results)
}

#Create a survival object
surv_obj <- Surv(time = colon$time, event = colon$status)

#pairwise comparisons
pairwise_results <- pairwise_logrank(Surv(time, status) ~ rx, data = colon)
print(pairwise_results)
```

```
##    group1 group2      p.value
## 1 Lev+5FU    Obs 1.095864e-07
## 2 Lev+5FU    Lev 6.302887e-07
## 3     Obs    Lev 7.842255e-01
```

```
# Apply Bonferroni
pairwise_results$p.adjusted <- p.adjust(pairwise_results$p.value, method = "bonferroni")
print(pairwise_results)
```

```
##    group1 group2      p.value   p.adjusted
## 1 Lev+5FU    Obs 1.095864e-07 3.287592e-07
## 2 Lev+5FU    Lev 6.302887e-07 1.890866e-06
## 3     Obs    Lev 7.842255e-01 1.000000e+00
```

#Nonparametric comparison of 2 or more group____age

#conclusion: age significantly impacts recurrence-free survival (log-rank p=0.002). Patients <30 years old demonstrate the best outcomes (median survival 3021 days), while those >60 years have the poorest prognosis (median 1622 days). These differences persist after pairwise comparisons (p<0.05).

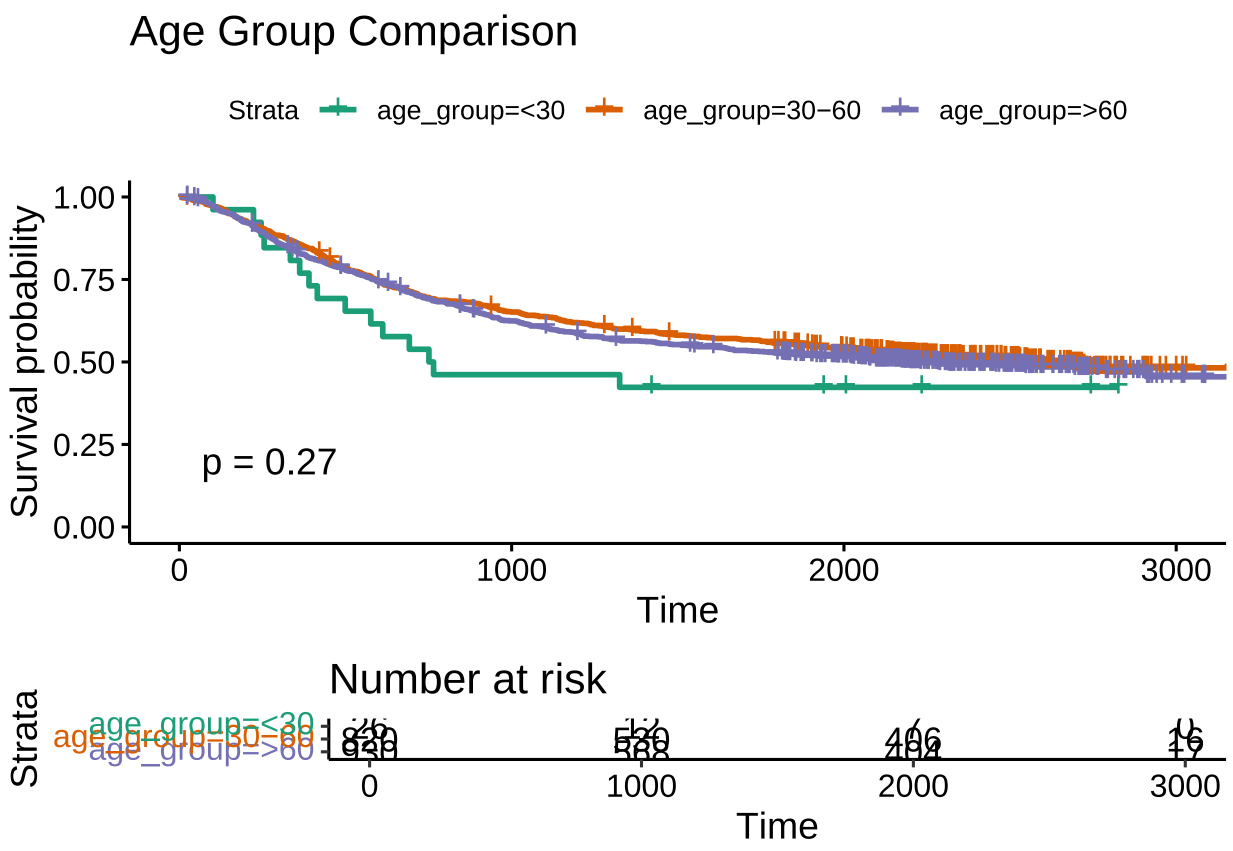
```
colon_complete <- colon_complete |>
  mutate(
    age_group = cut(age,
      breaks = c(0, 30, 60, 100),
      labels = c("<30", "30-60", ">60")
    )
)

# Verify group distribution
table(colon_complete$age_group)
```

```
##
##   <30 30-60   >60
##    26   820   930
```

```
# Log-Rank Test for age groups
survdiff(Surv(time, status) ~ age_group, data = colon_complete) |>
  print()
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ age_group, data = colon_complete)
##
##                     N Observed Expected (O-E)^2/E (O-E)^2/V
## age_group=<30      26       15     10.8     1.662     1.684
## age_group=30-60   820      397    413.4     0.650     1.231
## age_group=>60     930      464    451.8     0.327     0.676
##
##  Chisq= 2.6  on 2 degrees of freedom, p= 0.3
```

```
# Kaplan-Meier Plot
ggsurvplot(
  survfit(Surv(time, status) ~ age_group, data = colon_complete),
  pval = TRUE, risk.table = TRUE, palette = "Dark2",
  title = "Age Group Comparison"
)
```



#Cox model

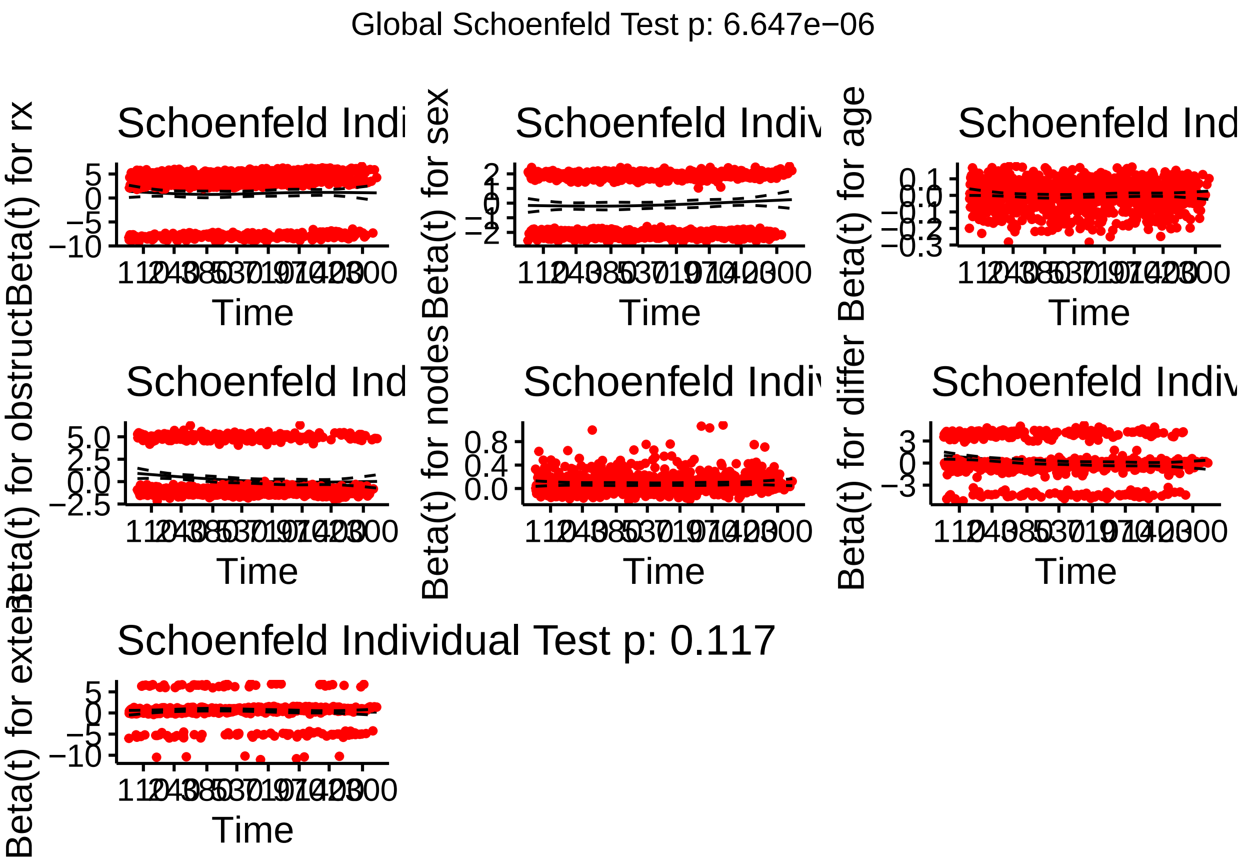
```r
cox_model <- coxph(
  Surv(time, status) ~ rx + sex + age + obstruct + nodes + differ + extent,
  data = colon_complete
)
summary(cox_model)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ rx + sex + age + obstruct + 
##     nodes + differ + extent, data = colon_complete)
## 
##   n= 1776, number of events= 876 
## 
##                coef exp(coef)  se(coef)      z Pr(>|z|)    
## rxLev+5FU -0.377435  0.685618  0.087672 -4.305 1.67e-05 ***
## rxObs      0.071061  1.073647  0.079202  0.897   0.3696    
## sex       -0.087680  0.916054  0.068014 -1.289   0.1973    
## age        0.002056  1.002058  0.002871  0.716   0.4739    
## obstruct   0.210999  1.234912  0.083699  2.521   0.0117 *  
## nodes      0.080964  1.084332  0.006687 12.108  < 2e-16 ***
## differ     0.148493  1.160085  0.070075  2.119   0.0341 *  
## extent     0.471209  1.601930  0.081610  5.774 7.75e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
##           exp(coef) exp(-coef) lower .95 upper .95
## rxLev+5FU    0.6856     1.4585    0.5774    0.8142
## rxObs        1.0736     0.9314    0.9193    1.2539
## sex          0.9161     1.0916    0.8017    1.0467
## age          1.0021     0.9979    0.9964    1.0077
## obstruct     1.2349     0.8098    1.0481    1.4551
## nodes        1.0843     0.9222    1.0702    1.0986
## differ       1.1601     0.8620    1.0112    1.3309
## extent       1.6019     0.6242    1.3651    1.8798
## 
## Concordance= 0.654  (se = 0.009 )
## Likelihood ratio test= 212  on 8 df,   p=<2e-16
## Wald test            = 252.4  on 8 df,   p=<2e-16
## Score (logrank) test = 258.2  on 8 df,   p=<2e-16
```

```r
test_ph <- cox.zph(cox_model)
print(test_ph)
```

```
##            chisq df       p
## rx        1.6160  2  0.4457
## sex       2.9303  1  0.0869
## age       0.0486  1  0.8255
## obstruct  9.2802  1  0.0023
## nodes     0.0140  1  0.9058
## differ   20.6827  1 5.4e-06
## extent    2.4576  1  0.1170
## GLOBAL   38.2924  8 6.6e-06
```

```r
ggcoxzph(test_ph)
```


Global Schoenfeld Test p: 6.647e−06
Schoenfeld Indi
Beta(t) for rx
Time
Schoenfeld Indiv
Beta(t) for sex
Time
Schoenfeld Ind
Beta(t) for age
Time
Schoenfeld Ind
Beta(t) for obstruct
Time
Schoenfeld Indiv
Beta(t) for nodes
Time
Schoenfeld Indiv
Beta(t) for differ
Time
Schoenfeld Individual Test p: 0.117
Beta(t) for extent
Time