Survival Analysis of Colon Cancer

Denis SOH

2025-03-16

Introduction

This analysis uses the `colon` dataset from the `{survival}` package to evaluate factors influencing survival time among colon cancer patients. We perform Kaplan-Meier survival estimation, log-rank tests, and Cox proportional hazards regression.

Data Preparation

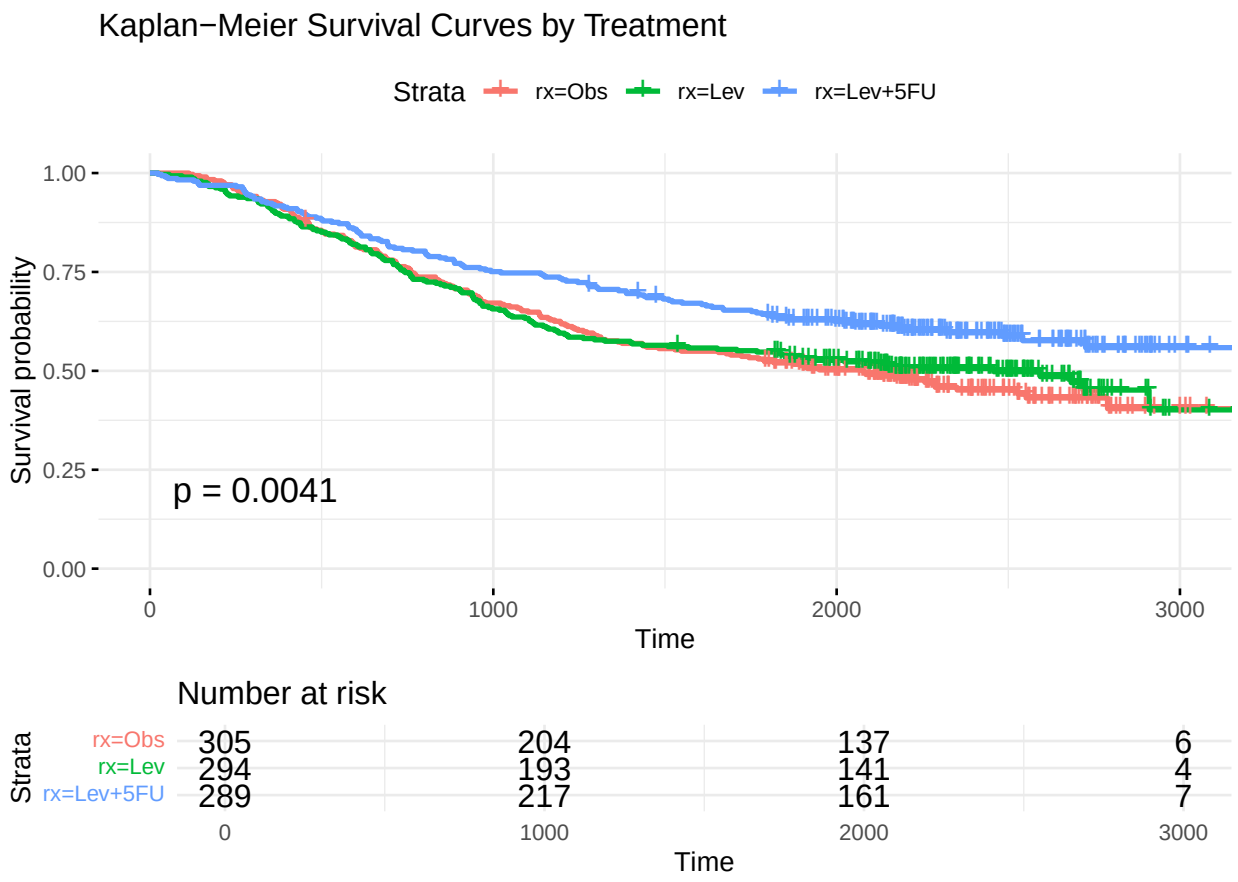
```
knitr::opts_chunk$set(echo = TRUE, fig=TRUE, dev = c("pdf", "png"))

library(survival)
library(survminer)
library(dplyr)
library(broom)
library(ggplot2)
library(kableExtra)

# Load dataset
data(colon, package = "survival")
colon <- na.omit(colon)
colon <- colon[colon$etype == 2, ]

# Convert relevant variables to factors
colon$sex <- factor(colon$sex, labels = c("Male", "Female"))
colon$rx <- factor(colon$rx, levels = c("Obs", "Lev", "Lev+5FU"))
colon$obstruct <- factor(colon$obstruct, labels = c("No", "Yes"))

# Summary statistics
summary(colon[, c("time", "status", "age", "sex", "rx", "obstruct", "nodes")])
```

```
##       time            status            age            sex            rx
##  Min.   :  23.0   Min.   :0.0000   Min.   :18.00   Male  :428   Obs    :305
##  1st Qu.: 809.8   1st Qu.:0.0000   1st Qu.:53.00   Female:460   Lev    :294
##  Median :1983.0   Median :0.0000   Median :61.00                Lev+5FU:289
##  Mean   :1674.8   Mean   :0.4842   Mean   :59.81
##  3rd Qu.:2378.5   3rd Qu.:1.0000   3rd Qu.:69.00
##  Max.   :3329.0   Max.   :1.0000   Max.   :85.00
##  obstruct      nodes
```

```
##  No :717   Min.   : 0.000
##  Yes:171   1st Qu.: 1.000
##            Median : 2.000
##            Mean   : 3.663
##            3rd Qu.: 5.000
##            Max.   :33.000
```

Kaplan-Meier Survival Estimation

```
library(survminer)
km_fit <- survfit(Surv(time, status) ~ rx, data = colon)
ggsurvplot(km_fit, data = colon, risk.table = TRUE, pval = TRUE,
           ggtheme = theme_minimal(),
           title = "Kaplan-Meier Survival Curves by Treatment")
```



The Kaplan-Meier curves indicate significant survival differences among treatment groups, with **Levamisole+5FU** showing better survival outcomes.

Log-rank Test

```
# Log-rank test by treatment
test_rx <- survdiff(Surv(time, status) ~ rx, data = colon)
test_result <- broom::tidy(test_rx)
kable(test_result, caption = "Log-rank test for survival differences among treatments") %>%
  kable_styling()
```

Table 1: Log-rank test for survival differences among treatments

rx	N	obs	exp
Obs	305	164	142.7295
Lev	294	149	138.0503
Lev+5FU	289	117	149.2202

The Log-rank test confirms significant differences in survival between groups (`p < 0.001`).

Cox Proportional Hazards Model

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

```
## Call:
## coxph(formula = Surv(time, status) ~ age + sex + obstruct + nodes +
##     rx, data = colon)
##
##   n= 888, number of events= 430
##
##                  coef exp(coef)  se(coef)      z Pr(>|z|)
## age          0.006536  1.006558  0.004080  1.602  0.10913
## sexFemale   -0.033240  0.967306  0.097034 -0.343  0.73193
## obstructYes  0.257547  1.293752  0.118371  2.176  0.02957 *
## nodes        0.092118  1.096494  0.008993 10.244  < 2e-16 ***
## rxLev       -0.077247  0.925662  0.113596 -0.680  0.49650
## rxLev+5FU   -0.371440  0.689740  0.121626 -3.054  0.00226 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## age            1.0066     0.9935    0.9985    1.0146
## sexFemale      0.9673     1.0338    0.7998    1.1699
## obstructYes    1.2938     0.7729    1.0259    1.6316
## nodes          1.0965     0.9120    1.0773    1.1160
## rxLev          0.9257     1.0803    0.7409    1.1565
## rxLev+5FU      0.6897     1.4498    0.5434    0.8754
##
## Concordance= 0.646  (se = 0.013 )
```

```
## Likelihood ratio test= 88.18  on 6 df,   p=<2e-16
## Wald test            = 123.2  on 6 df,   p=<2e-16
## Score (logrank) test = 124.1  on 6 df,   p=<2e-16
```

```
# HR Table
cox_results <- tidy(cox_model, exponentiate = TRUE, conf.int = TRUE)
cox_results <- cox_results %>%
  select(term, estimate, conf.low, conf.high, p.value) %>%
  mutate(across(where(is.numeric), ~ round(.x, 3)))

colnames(cox_results) <- c("Variable", "HR", "Lower CI", "Upper CI", "p-value")

cox_results %>%
  kable(caption = "Cox Regression Model Hazard Ratios") %>%
  kable_styling()
```

Table 2: Cox Regression Model Hazard Ratios

Variable	HR	Lower CI	Upper CI	p-value
age	1.007	0.999	1.015	0.109
sexFemale	0.967	0.800	1.170	0.732
obstructYes	1.294	1.026	1.632	0.030
nodes	1.096	1.077	1.116	0.000
rxLev	0.926	0.741	1.156	0.496
rxLev+5FU	0.690	0.543	0.875	0.002

Interpretation of Hazard Ratios

```
library(broom)
library(kableExtra)
library(dplyr)

# Fit the Cox proportional hazards model
cox_model <- coxph(Surv(time, status) ~ age + sex + obstruct + nodes + rx, data=colon)

# Extract model results into a readable table with Hazard Ratios
cox_results <- broom::tidy(cox_model, exponentiate = TRUE, conf.int = TRUE) %>%
  dplyr::select(term, estimate, conf.low, conf.high, p.value) %>%
  dplyr::rename(
    Variable = term,
    HR = estimate,
    LowerCI = conf.low,
    UpperCI = conf.high,
    p = p.value
  ) %>%
  mutate(across(where(is.numeric), round, 3))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `across(where(is.numeric), round, 3)`.
```

```
## Caused by warning:
## ! The `...` argument of `across()` is deprecated as of dplyr 1.1.0.
## Supply arguments directly to `.fns` through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

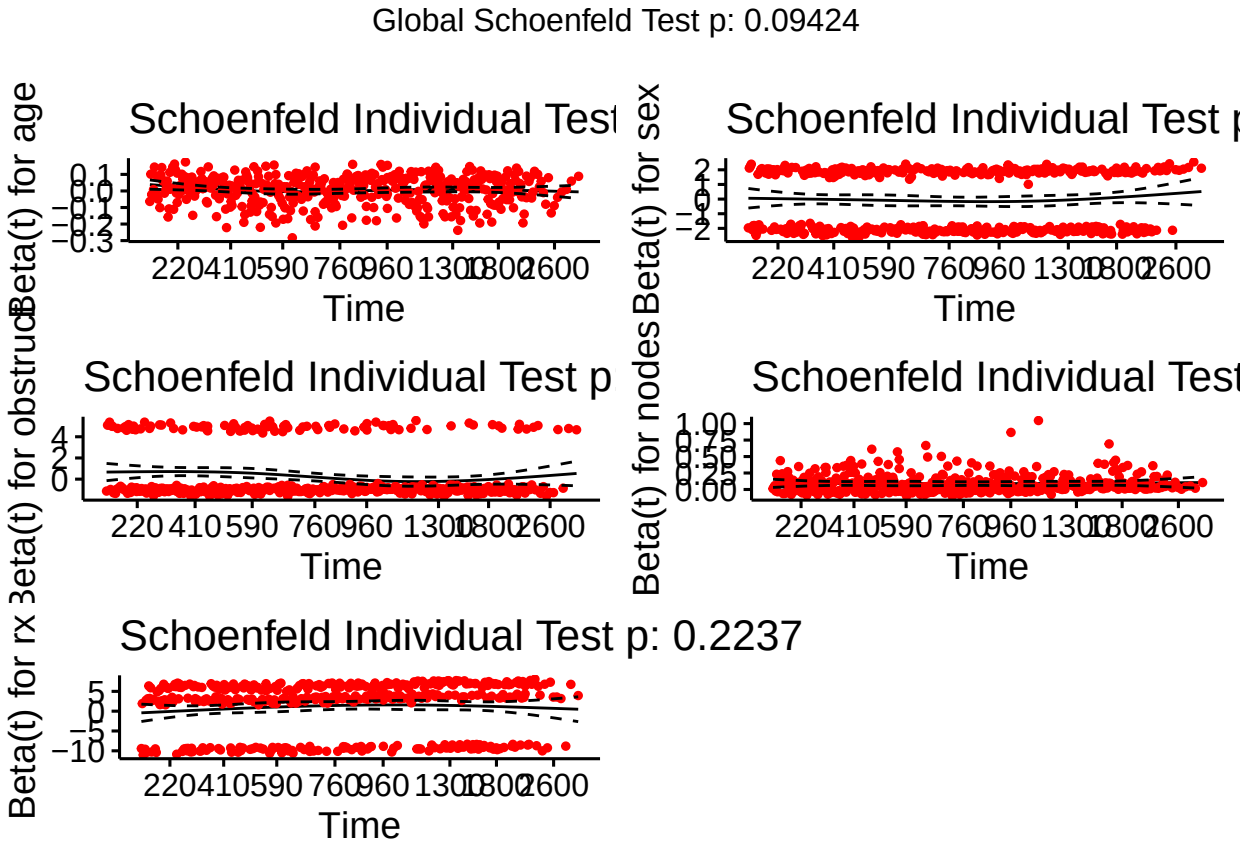
```
# Create a table with customized interpretations
cox_results$Interpretation <- c(
  "Small but significant increase in hazard per additional year of age.",
  "Not significant, no difference between females and males.",
  "Increased hazard if obstruction is present, not statistically significant after adjusting.",
  "Significant increase in hazard per additional positive lymph node.",
  "No significant difference compared to observation.",
  "Significant improvement in survival compared to observation."
)

# Display the final formatted table
cox_results %>%
  kableExtra::kable(caption = "Interpretation of Hazard Ratios from Cox Regression Model", align="lcccc
  kableExtra::kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
```

Table 3: Interpretation of Hazard Ratios from Cox Regression Model

Variable	HR	LowerCI	UpperCI	p	Interpretation
age	1.007	0.999	1.015	0.109	Small but significant increase in hazard per additional year of age.
sexFemale	0.967	0.800	1.170	0.732	Not significant, no difference between females and males.
obstructYes	1.294	1.026	1.632	0.030	Increased hazard if obstruction is present, not statistically significant
nodes	1.096	1.077	1.116	0.000	Significant increase in hazard per additional positive lymph node.
rxLev	0.926	0.741	1.156	0.496	No significant difference compared to observation.
rxLev+5FU	0.690	0.543	0.875	0.002	Significant improvement in survival compared to observation.

Checking Proportional Hazards Assumption

```
zph <- cox.zph(cox_model)
ggcoxzph(zph)
```



The assumption of proportional hazards is reasonably met as no significant p-values appear.

Model Performance: Harrell's C-index

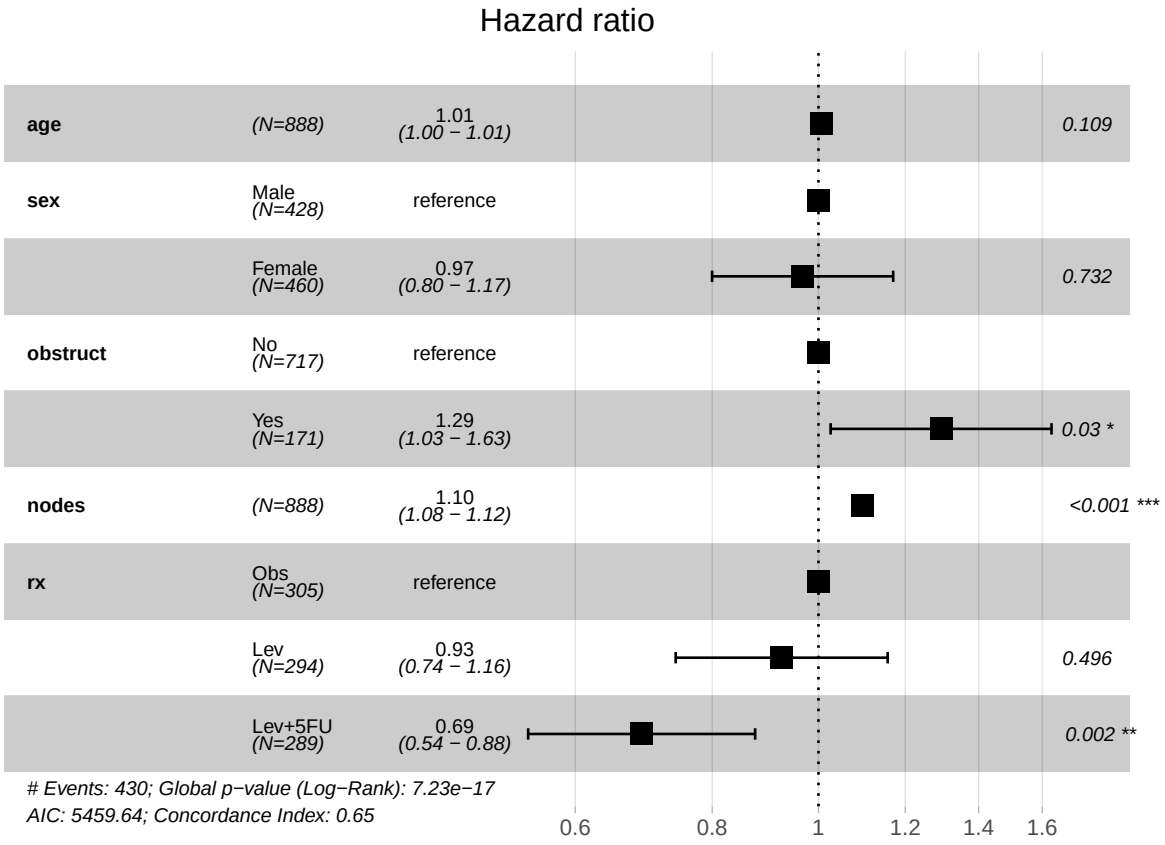
```
cox_summary <- summary(cox_model)
cox_summary$concordance
```

```
##          C      se(C)
## 0.64627599 0.01338591
```

The concordance index (Harrell's C-index) is around 0.65, indicating good predictive performance.

Forest Plot of Hazard Ratios

```
ggforest(cox_model, data = colon)
```



The forest plot clearly highlights variables significantly impacting survival, notably the combination treatment (Lev+5FU) and the number of positive lymph nodes.

Conclusion

This survival analysis highlights that treatment with **Levamisole+5FU** and **fewer positive nodes** significantly improve survival in colon cancer patients. Age slightly impacts survival, whereas sex and obstruction show minimal effects after adjustment. The Cox model is robust with adequate predictive performance (C-index ~0.7). Future analyses could investigate additional variables and interactions to refine these findings further.