R code for Survival Lab #3

Spring 2025

Contents

Background

The purpose of this lab is to provide hands-on practice checking the proportional hazards assumption in Cox regression models. The R code in this document demonstrates various analyses for Survival Lab #3.

Data

Download the datasets lab3-actg019_a.dta and lab3-pbc_a.dta from the website (these are altered datasets from actg019.dta and pbc.dta that you used before!).

Checking the Cox Model for ZDV treatment in *lab3-actg019_a.dta*

Load all the R packages needed for Survival Lab #3

```r
packages_to_load <- c("haven", "survival", "survminer", "ggplot2", "pander", "dplyr")
lapply(packages_to_load, library, character.only = TRUE)
```

Load the ACTG dataset

```r
ACTG <- read_dta("lab3-actg019_a.dta")
ACTG$rx <- factor(ACTG$rx, levels=0:1, labels=c("ZDV", "Placebo"))
```

Generate the Cox-KM plot and the log-minus-log KM plot by treatment (*rx*)

- Generate the Cox-KM plot

```r
# Fit Kaplan-Meier and Cox models
km.fit <- survfit(Surv(days, cens) ~ rx, data = ACTG)
cox.fit <- coxph(Surv(days, cens) ~ rx, data = ACTG)

# Get Cox predicted survival
newdata <- data.frame(rx = c("ZDV", "Placebo"))
cox.pred <- survfit(cox.fit, newdata = newdata)

# Tidy survival estimates
km.df <- surv_summary(km.fit, data = ACTG) %>%
  mutate(model = "KM")
cox.df <- surv_summary(cox.pred, data = newdata) %>%
  mutate(model = "Cox")

# Standardize group labels
km.df$strata <- factor(km.df$strata, labels = c("ZDV", "Placebo"))
cox.df$strata <- factor(cox.df$strata, labels = c("ZDV", "Placebo"))

# Combine into one data frame
plot.df <- bind_rows(km.df, cox.df)
```

```
# Plot: solid = KM, dashed = Cox
ggplot(plot.df, aes(x = time, y = surv, color = strata, linetype = model)) +
  geom_step(size = 1) +
  scale_linetype_manual(values = c("KM" = "solid", "Cox" = "dashed")) +
  scale_color_manual(values = c("blue", "red")) +
  labs(
    title = "Kaplan-Meier and Cox Predicted Survival Curves",
    x = "Time (days)",
    y = "Survival Probability",
    color = "Treatment",
    linetype = "Model"
  ) +
  coord_cartesian(ylim = c(0.8, 1)) +
  theme_minimal()
```

- Generate log minus log plot

```
# Generate log-minus-log plot
ggsurvplot(
  km.fit,
  data = ACTG,
  censor = FALSE,
  fun = "cloglog",  # Complementary log-log = log(-log(S(t)))
  palette = c("blue", "red"),
  xlab = "Time (days)",
  ylab = "log(-log(Survival))",
  title = "Log-minus-Log Survival Curves by Treatment",
  legend.title = "Treatment",
  legend.labs = c("ZDV", "Placebo"),
  ggtheme = theme_minimal()
)
```

Question: Does the *rx* HR appear proportional?

Graph the log hazard ratio over time

Graph the log hazard ratio after fitting the Cox model and plot the scaled Schoenfeld residuals with a lowess smoother to approximate time-dependent coefficient for *rx*.

By default, cox.zph() applies a Kaplan-Meier transformation on time, but here we explicitly use the identity transformation (i.e., raw time) for better interpretability of the residual plot:

```
# Perform Schoenfeld residual test with identity time transformation
test.ph <- cox.zph(cox.fit, transform = "identity")

# Plot residuals with ggplot2-based diagnostic plot
ggcoxzph(test.ph)
```

Re-graph the log hazard ratio with a lowess smoother

- Compute a lowess-smoothed estimate of the log(HR) for *rx* to examine how the treatment effect may evolve over time.

```r
# Extract residuals and fit lowess smoother
smloghr <- data.frame(days = test.ph$time, rx = test.ph$y)
loess.fit <- loess(rx ~ days, smloghr, span = 0.8)
```

- Save the lowess values.

```r
# Predict log(HR) at selected time points and convert to HR scale
out <- data.frame(
  days = c(95, 181, 362, 540),
  logHR = predict(loess.fit, data.frame(days = c(95, 181, 362, 540)), se = FALSE)
)
out$HR <- exp(out$logHR)

# Display results
pander(out)
```

Question: Based on the results of the Schoenfeld test for the proportional hazards assumption, is there evidence of a violation? Do the plots support this conclusion?

```r
test.ph
```

Note that the result is different than Stata *estat phtest* output because different tests were used. As an alternative, we can apply Pearson's and Spearman's correlation between the residuals and time:

```r
test.ph
print(cor.test(test.ph$y, test.ph$time, method="pearson"))
print(cor.test(test.ph$y, test.ph$time, method="spearman"))
```

Question: How would you summarize the effect of ZDV on progression of HIV?

Stratified Cox regression

Consider the "stratification" approach to dealing with non-proportional hazards: run the log-rank test to conclude that the effect of ZDV is statistically significant and present the K-M plot to show its effect on free of HIV progression. Note that this approach does not provide a summary estimate of the ZDV effect (the primary predictor of the study).

```
survdiff(Surv(days, cens) ~ rx, data=ACTG)
plot(km.fit, col=c("blue", "red"), xlab="Days", ylab="Survival Probability")
legend(10, 0.9, c("ZDV", "Placebo"),  col=c("blue", "red"), lty=c(1,1), bty="n")
```

Fit a Cox model with *rx* as a stratifying factor. Interpret the results.

```
coxph(Surv(days, cens) ~ strata(rx), data=ACTG)
```

Explore the time-dependent covariate approach

When the proportional hazards assumption is questionable, one way to address it is to allow a covariate's effect (e.g., treatment *rx*) to change over time. This can be done by creating time-dependent covariates through data splitting.

- Step 1: Split the follow-up time into intervals

  We use survSplit() to divide each subject's follow-up into two time intervals:

  - One before 365 days,
  - One after 365 days.

```
ACTG.td <- survSplit(Surv(days, cens) ~ ., data=ACTG,
                     cut=c(365), episode ="grp")
```

  The new variable *grp* indicates the time interval:

  - grp = 1 if time <= 365 days
  - grp = 2 if time > 365 days

  survSplit() also creates a tstart variable, which marks the start time of each interval (needed for time-dependent Cox models).

- Step 2: Define time-dependent covariates

  We now define two binary indicators:

  - rx01: Placebo group during interval 1 (<= 365 days)
  - rx1p: Placebo group during interval 2 (> 365 days)

  These variables allow us to estimate different hazard ratios in the two time periods.

```
ACTG.td$rx01 <- (ACTG.td$rx=="Placebo") * (ACTG.td$grp==1)
ACTG.td$rx1p <- (ACTG.td$rx=="Placebo") * (ACTG.td$grp==2)
```

- Step 3: Fit the time-dependent Cox model

  This model estimates two separate hazard ratios for the two time intervals.

```
cox.fit.td <-coxph(Surv(tstart, days, cens) ~rx01 + rx1p, data=ACTG.td)
pander(cox.fit.td)
```

Checking the Cox Model for Cholesterol in *lab3-pbc__a.dta*

Fit a Cox model to assess the effect of cholesterol. Then, perform the Schoenfeld residual test to evaluate the proportional hazards assumption. Is there evidence suggesting a violation of this assumption?

Note that the test result is different than what given by Stata "estat phtest" because different tests were used.

```
PBC <- read_dta("lab3-pbc_a.dta")
cox.fit <- coxph(Surv(years, status) ~ cholest, data=PBC)
test.ph <- cox.zph(cox.fit, transform="identity")
test.ph

print(cor.test(test.ph$y, test.ph$time, method="pearson"))
print(cor.test(test.ph$y, test.ph$time, method="spearman"))
```

Graph the log hazard ratio

What does the graph suggest? Do you have any concerns about the test?

```
#ggcoxzph(test.ph)
smloghr <- data.frame(years=test.ph$time, cholest=test.ph$y)
p <- ggplot(smloghr, aes(years, cholest)) +
  ylab("cholestrol log HR") +
  geom_point(color = "blue") +
  geom_smooth(method = "loess", span=0.8)
p
```

Delete some potential influential points and then re-run the plot and test for the proportional hazards assumption. What do you conclude?

Note that the test result is different than what given by Stata *estat phtest* because different tests were used.

```
cox.fit <- coxph(Surv(years, status) ~ cholest, data=PBC[PBC$years<=12,])
pander(cox.fit)
test.ph <- cox.zph(cox.fit, transform=identity)
ggcoxzph(test.ph)
```

```
test.ph
print(cor.test(test.ph$y, test.ph$time, method="pearson"))
print(cor.test(test.ph$y, test.ph$time, method="spearman"))
```

The bottom line is that the test of proportional hazards can be greatly affected by outlying values. It is important to always accompany the test by a graph so that you judge the directions, the magnitude of the violation and whether there appear to be points exerting a large influence on the test.