

R code for Survival Lab #2

Spring 2025

Contents

1	Background	2
2	Data	2
3	Unadjusted and Adjusted Effect	2
4	Statistical Interaction	3
5	Adjusted Survival Curves	5

1 Background

This R code in this document demonstrates various analyses for the PMC dataset in the context of Survival Lab #2. The purpose of this lab is to give you practice in addressing interaction and confounding in Cox regression models.

2 Data

Load all the R packages needed for Survival Lab #2

```
packages_to_load <- c("haven", "survival", "survminer", "gtsummary", "ggplot2", "biostat3", "gridExtra")
lapply(packages_to_load, library, character.only = TRUE)
```

Load the PMC dataset (1=female, 0=male)

```
# Read the Stata data file "pbc.dta" into a data frame called PBC.
PBC <- read_dta("pbc.dta")

# Convert the numeric 'sex' variable into a factor with descriptive labels.
# - 'levels=0:1' tells R that the variable takes on values 0 and 1.
# - 'labels=c("male", "female")' assigns "male" to 0 and "female" to 1.
PBC$sex <- factor(PBC$sex, levels=0:1, labels=c("male", "female"))
```

3 Unadjusted and Adjusted Effect

Examine the effect of sex (1=female, 0=male) on survival both graphically and in a Cox model

```
# Note: We are using the default option in R (Efron) for handling ties.
# You may choose an alternative method (e.g., Breslow) to match results from STATA.
```

```
km.sex <- survfit(Surv(years, status) ~ sex, data = PBC)
```

```
# Generate Kaplan-Meier survival curves by sex without displaying censor ticks.
# You may play with different options.
```

```
ggsurvplot(km.sex,
  data = PBC,
  censor = FALSE,                # Do not display censoring tick marks
  xlab = "Years",                # Label for the x-axis
  ylab = "Survival Probability", # Label for the y-axis
  title = "Kaplan-Meier Curves by Sex", # Plot title
  legend.title = "Sex",          # Title for the legend
  legend.labs = c("Male", "Female"), # Labels for the two sex groups
  ggtheme = theme_bw()          # Apply a clean black-and-white theme
)
```

```
fit.cox.univ <- coxph(Surv(years, status) ~ sex, data = PBC)
summary(fit.cox.univ)

fit.cox.univ %>%
  tbl_regression(exp = TRUE)
```

Fit a model adjusting for the effect of sex on the following possible confounding variables: age, log bilirubin (logbili) and triglycerides (trigli)

```
fit.cox.multi <- coxph(Surv(years, status) ~ sex + age + logbili + trigli, data = PBC)
summary(fit.cox.multi)

fit.cox.multi %>%
  tbl_regression(exp = TRUE)
```

Which variable(s) has missing values?

Question 3.1: Interpret the results.

4 Statistical Interaction

Consider an interaction between sex and baseline level of copper in the urine (copper).

Fit the interaction model

```
fit.cox.int <- coxph(Surv(years, status) ~ sex*copper, data = PBC)
summary(fit.cox.int)

fit.cox.int %>%
  tbl_regression(exp = TRUE, digits = list(coef = 3, conf.int = 3))
```

Question 4.1: Does there appear to be a significant interaction?

Questions 4.2 & 4.3: Consider the use of the following tables to derive the lincom statements for *a 10 unit increase* in urine copper for female and male subjects

	Female Subjects			Male Subjects		
	sex	copper	interaction	sex	copper	interaction
copper = x+10	1	x+10	x+10	0	x+10	0
copper = x	1	x	x	0	x	0
difference	0	10	10	0	10	0

Use *lincom* to estimate the hazard ratio for a 10-unit increase in copper separately for males and females.

For male subjects:

```
lincom(fit.cox.int, c("10*copper" ), eform=TRUE)
```

For female subjects: ?

Interpret the results you obtain when you run the *lincom* commands above.

Question 4.4: Fill out the following tables to determine the hazard ratio (female vs. male) at copper levels of 17, 73, and 256 (i.e. the 5th, 50th, and 95th percentiles of copper values).

	Copper of 17			Copper of 73		
	sex	copper	interaction	sex	copper	interaction
Female						
Male						
difference						

	Copper of 256		
	sex	copper	interaction
Female			
Male			
difference			

Use *lincom* function to estimate the hazard ratios

- HR (female vs male) at copper = 17
- HR (female vs male) at copper = 73
- HR (female vs male) at copper = 256

Interpret the results you obtain when you run the *lincom* commands above.

It can be helpful to graph the interaction. It is most easily visualized on the log hazard scale (i.e., on the scale of the coefficients).

The R code below performs the following steps:

- Create a new column `xb` in the PBC data frame and fill it with NA values.
- For rows in the PBC data frame where the copper value is not NA, calculate the linear predictor (XB) using the fitted Cox model (`fit.cox.int`) and store the result in the `xb` column.

```
PBC$xb <- NA
PBC$xb[!is.na(PBC$copper)] <- predict(fit.cox.int)
```

Create a line plot using `ggplot2`:

- Set the x aesthetic to `copper`, the y aesthetic to `xb`, and the color aesthetic to the factor version of the sex variable.
- Add lines for each group (males and females) using `geom_line()`.
- Add vertical dashed green lines at x-intercepts 17, 73, 153, and 256 using `geom_vline()`.
- Set the y-axis label to “Log Hazard Ratio” using the `labs()` function.
- Apply a minimal theme to the plot with `theme_minimal()`.

```
ggplot(PBC, aes(x = copper, y = xb, color = factor(sex))) +
  geom_line() +
  geom_vline(xintercept = c(17, 73, 153, 256), color="green", linetype = "dashed") +
  labs(y = "Log Hazard Ratio") +
  theme_minimal()
```

Feel free to experiment with other options to further customize the graph.

5 Adjusted Survival Curves

Return to the issue we considered in question 3, the effect of sex on mortality after adjusting for log bilirubin, age and triglycerides.

Re-run the adjusted regression

```
# fit Cox model
cox.reg <- coxph(Surv(years, status) ~ sex + age + logbili + trigli, data=PBC)
```

Graph the adjusted survival curve under the multivariate Cox model

```
# get the mean of each covariate.
means1 <- aggregate(cbind(age, logbili, trigli) ~ 1, data = PBC, FUN = mean)

# Create a new data frame with two rows, one for each sex category,
# and set the copper level to the overall means of other covariates.
newdata1 <- data.frame(sex = c("male", "female"), means1)
newdata1
# Fit the survival curves for the two gender groups at the overall means
```

```

surv1 <- survfit(cox.reg, newdata = newdata1)

# Generate a Kaplan-Meier plot using ggsurvplot:
# - 'palette' sets the curve colors (red for male, blue for female)
# - 'legend.labs' and 'legend.title' label the groups
# - 'title', 'xlab', and 'ylab' add appropriate plot labels
ggsurvplot(surv1,
            data = newdata1,
            censor=FALSE,
            conf.int = FALSE,
            legend.labs = c("Male", "Female"),
            legend.title = "Sex",
            title = "Adjusted Survival Curves",
            xlab = "Time",
            ylab = "Survival Probability",
            ggtheme = theme_minimal())

```

Graph the unadjusted Kaplan-Meiers for comparison

```

# Fit a Kaplan-Meier model
km_sex <- survfit(Surv(years, status) ~ sex, data = PBC)
# Plot the Kaplan-Meier estimates
ggsurvplot(km_sex, data = PBC,
            censor = FALSE,
            title = "Kaplan-Meier estimates",
            legend.title = "Sex", legend = "bottom",
            legend.labs = c("Male", "Female"),
            ylab = "Proportion Surviving",
            ggtheme = theme_bw())

```