Biostatistics 209

Homework #2

*In this homework assignment, Questions 1 and 2 require you to simply answer the questions using the provided tables. For Question 3, use the* pbc.dta *dataset available on the course website.*

1. [4 points] The follow Stata output gives the hazard ratio of a **1-year increase** in age (variable named agedx).

------------------------------------------------------------------------------

\_t | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

agedx | .9144948 .0264917 -3.09 0.002 .8640186 .9679198

------------------------------------------------------------------------------

Based on the above, with a **1-year decrease** in age, what are the corresponding

1. [1 point] Hazard ratio: Where HRdecrease = 1/HR increase, HR = 1/.9144948 = 1.0935
2. [2 points] Wald test statistic & p-value, and: χ² = 3.09, p = 0.002
3. [1 point] 95% CI?: Where 95% CI for 1 year increase = = = (1.033 - 1.157)
4. [4 points] The variable agedx was grouped as follows to create the variable “ages”

agedx: 0-2: ages =1

agedx: 2-6: ages =2

agedx: 6-12: ages =3

agedx: > 12: ages =4

This leads the following output from a Cox model

No. of subjects = 103 Number of obs = 103

No. of failures = 74

Time at risk = 1413.050005

LR chi2(3) = 28.55

Log likelihood = -266.66906 Prob > chi2 = 0.0000

------------------------------------------------------------------------------

\_t | Haz. Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

agedx |

2 | .2684887 .1685833 -2.09 0.036 .0784258 .9191639

3 | .0989741 .0632982 -3.62 0.000 .028258 .346658

4 | .0566843 .0391472 -4.16 0.000 .0146422 .2194414

------------------------------------------------------------------------------

1. [1 point] What is the hazard ratio for someone aged 0-2 (ages =1) compared to someone aged >12 (ages =4)? *i.e., now considering* (ages =4) *as the reference group.* Where the hazard ratio for ages = 1 using ages = 4 as reference = = = 17.638
2. [2 points] What are the Wald test statistic, p-value and 95% CI associated with the hazard ratio in (a)? χ² = 4.16, p = 0.00, Where 95% CI for someone aged 0-2 (ages =1) compared to someone aged >12 (ages =4)? *i.e., now considering* (ages =4) *as the reference group.* = = = (4.556 - 68.294)
3. [1 point] What is the hazard ratio for someone aged 2-6 (ages =2) compared to someone aged >12 (ages =4)? *i.e., considering* (ages =4) *as the reference group here.* (Hint: look at the definition of hazard ratio for 2 versus 1 and 4 versus 1)

HR (ages = 2) = 0.2684887 HR (ages = 4) = 0.0566843. Since both hazard ratios are relative to the same reference group, we can just divide them.

HR (ages =2) compared to someone aged >12 (ages = 4) = = = 4.736

Biostatistics 209 - Homework 2

Anya DeCarlo

April 25, 2025

Table of Contents

[Question 3 1](#_Toc196492567)

[Part A 1](#_Toc196492568)

[Part B 1](#_Toc196492569)

[Part C 1](#_Toc196492570)

[Part D 1](#_Toc196492571)

##### Week 3: Chapters 6.3-6.5, VGSM

# Question 3

#### 3. [5 points] For the variable age in the pbc dataset, check the proportional hazards assumption. Please find relevant R code in the Survival Lab #3 supplemental materials.

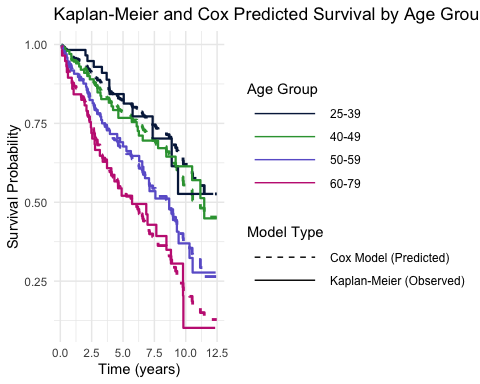
# Load the PBC dataset  
pbc <- read.dta13("~/biostats\_209/Data/pbchw.dta")  
colnames(pbc)

## [1] "number" "status" "rx" "sex" "asictes" "hepatom"   
## [7] "spiders" "edema" "bilirubin" "cholest" "albumin" "copper"   
## [13] "alkphos" "sgot" "trigli" "platel" "prothrom" "histol"   
## [19] "age" "years" "logbili" "logalbu" "logprot"

## Part A

#### (a) [1 point] Generate log-log plot or plot Kaplan-Meier curves with predicted survival under Cox model. Note, for this question, you will need to create a grouped version of age. Use the cut command in R:

# Convert status to numeric if needed  
pbc$status\_numeric <- ifelse(pbc$status == "Dead", 1, 0)  
  
# Create age groups using the specified cutpoints from the instructions  
pbc$age\_group <- cut(pbc$age, breaks = c(25, 40, 50, 60, 80),   
 labels = c("25-39", "40-49", "50-59", "60-79"))  
  
# Fit Kaplan-Meier and Cox models  
km.fit <- survfit(Surv(years, status\_numeric) ~ age\_group, data = pbc)  
cox.fit <- coxph(Surv(years, status\_numeric) ~ age\_group, data = pbc)  
  
# Get Cox predicted survival  
newdata <- data.frame(age\_group = levels(pbc$age\_group))  
cox.pred <- survfit(cox.fit, newdata = newdata)  
  
# Tidy survival estimates  
km.df <- surv\_summary(km.fit, data = pbc) %>%  
 mutate(model = "KM")  
cox.df <- surv\_summary(cox.pred, data = newdata) %>%  
 mutate(model = "Cox")  
  
# Ensure consistent strata naming across both datasets  
km.df$strata <- gsub("age\_group=", "", km.df$strata)  
cox.df$strata <- levels(pbc$age\_group)[as.numeric(cox.df$strata)]  
  
# 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(aes(size = model)) +  
 scale\_linetype\_manual(  
 values = c("KM" = "solid", "Cox" = "dashed"),  
 labels = c("KM" = "Kaplan-Meier (Observed)", "Cox" = "Cox Model (Predicted)")  
 ) +  
 scale\_size\_manual(  
 values = c("KM" = 0.8, "Cox" = 1.0),  
 guide = "none" # Hide size from legend  
 ) +  
 scale\_color\_manual(  
 values = c("25-39" = "#052049", "40-49" = "#32A03E",   
 "50-59" = "#6C62D0", "60-79" = "#C42882"),  
 labels = function(x) paste0(x) # Remove any prefix  
 ) +  
 labs(  
 title = "Kaplan-Meier and Cox Predicted Survival by Age Group",  
 x = "Time (years)",  
 y = "Survival Probability",  
 color = "Age Group",  
 linetype = "Model Type"  
 ) +  
 theme\_minimal() +  
 theme(  
 legend.position = "right",  
 legend.box = "vertical",  
 legend.margin = margin(6, 6, 6, 6),  
 legend.key.width = unit(2, "cm"),  
 legend.text = element\_text(size = 9)  
 )



## Part B

#### (b) [1 point] Using Smoothed hazard ratios for the continuous variable age.

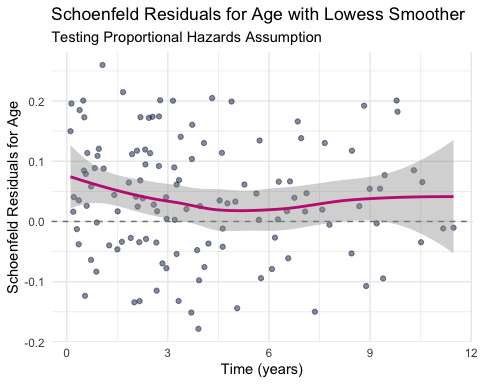
# Fit a Cox model with age as a continuous predictor  
cox.age <- coxph(Surv(years, status\_numeric) ~ age, data = pbc)  
summary(cox.age)

## Call:  
## coxph(formula = Surv(years, status\_numeric) ~ age, data = pbc)  
##   
## n= 312, number of events= 125   
##   
## coef exp(coef) se(coef) z Pr(>|z|)   
## age 0.039995 1.040806 0.008811 4.539 5.65e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## exp(coef) exp(-coef) lower .95 upper .95  
## age 1.041 0.9608 1.023 1.059  
##   
## Concordance= 0.625 (se = 0.027 )  
## Likelihood ratio test= 20.51 on 1 df, p=6e-06  
## Wald test = 20.6 on 1 df, p=6e-06  
## Score (logrank) test = 20.86 on 1 df, p=5e-06

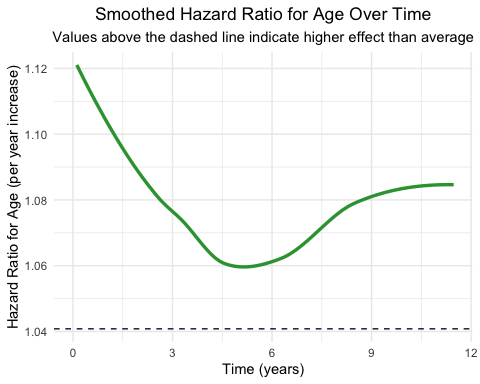
# Test the proportional hazards assumption using Schoenfeld residuals  
test.ph <- cox.zph(cox.age, transform = "identity")  
test.ph

## chisq df p  
## age 0.923 1 0.34  
## GLOBAL 0.923 1 0.34

# Extract the Schoenfeld residuals and time values  
schoen.age <- data.frame(  
 time = test.ph$time,  
 schoen.residuals = test.ph$y[,1] # Extract residuals for age  
)  
  
# Plot the Schoenfeld residuals with a lowess smoother  
ggplot(schoen.age, aes(x = time, y = schoen.residuals)) +  
 geom\_point(color = "#052049", alpha = 0.5) +  
 geom\_hline(yintercept = 0, linetype = "dashed", color = "gray50") +  
 geom\_smooth(method = "loess", span = 0.8, se = TRUE, color = "#C42882") +  
 labs(  
 title = "Schoenfeld Residuals for Age with Lowess Smoother",  
 x = "Time (years)",  
 y = "Schoenfeld Residuals for Age",  
 subtitle = "Testing Proportional Hazards Assumption"  
 ) +  
 theme\_minimal()



# Apply lowess smoother to approximate time-dependent coefficient for age  
loess.fit <- loess(schoen.residuals ~ time, data = schoen.age, span = 0.8)  
  
# Create a sequence of time points for prediction  
time.seq <- seq(0, max(schoen.age$time), length.out = 100)  
pred.df <- data.frame(  
 time = time.seq,  
 beta = predict(loess.fit, newdata = data.frame(time = time.seq)),  
 beta.coef = coef(cox.age)["age"]  
)  
  
# Calculate the time-varying log hazard ratio and hazard ratio  
pred.df$log.hr <- pred.df$beta + pred.df$beta.coef  
pred.df$hr <- exp(pred.df$log.hr)  
  
# Plot the smoothed hazard ratios over time  
ggplot(pred.df, aes(x = time, y = hr)) +  
 geom\_line(color = "#32A03E", size = 1.2) +  
 geom\_hline(yintercept = exp(coef(cox.age)["age"]),   
 linetype = "dashed", color = "#052049") +  
 labs(  
 title = "Smoothed Hazard Ratio for Age Over Time",  
 subtitle = "Values above the dashed line indicate higher effect than average",  
 x = "Time (years)",  
 y = "Hazard Ratio for Age (per year increase)"  
 ) +  
 theme\_minimal() +  
 theme(  
 plot.title = element\_text(hjust = 0.5),  
 plot.subtitle = element\_text(hjust = 0.5)  
 )



## Part C

#### (c) [1 point] Perform the Schoenfeld test for the continuous variable age.

# Fit Cox model with age as continuous predictor  
cox.age.test <- coxph(Surv(years, status\_numeric) ~ age, data = pbc)  
  
# Perform the Schoenfeld test with identity transformation of time  
test.ph <- cox.zph(cox.age.test, transform = "identity")  
print(test.ph)

## chisq df p  
## age 0.923 1 0.34  
## GLOBAL 0.923 1 0.34

# Alternative tests: Pearson and Spearman correlations between residuals and time  
print("Pearson correlation test:")

## [1] "Pearson correlation test:"

print(cor.test(test.ph$time, test.ph$y[,1], method = "pearson"))

##   
## Pearson's product-moment correlation  
##   
## data: test.ph$time and test.ph$y[, 1]  
## t = -0.90258, df = 123, p-value = 0.3685  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.25311680 0.09585819  
## sample estimates:  
## cor   
## -0.08111477

print("Spearman correlation test:")

## [1] "Spearman correlation test:"

print(cor.test(test.ph$time, test.ph$y[,1], method = "spearman"))

##   
## Spearman's rank correlation rho  
##   
## data: test.ph$time and test.ph$y[, 1]  
## S = 359137, p-value = 0.2514  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## -0.10334

tbl\_regression(cox.age.test, exponentiate = TRUE)

| **Characteristic** | **HR** | **95% CI** | **p-value** |
| --- | --- | --- | --- |
| age | 1.04 | 1.02, 1.06 | <0.001 |
| Abbreviations: CI = Confidence Interval, HR = Hazard Ratio | | | |

# Fit a Cox model with age\_group as a categorical variable  
cox.age.group <- coxph(Surv(years, status\_numeric) ~ age\_group, data = pbc)  
  
# Display the results with hazard ratios and confidence intervals  
summary(cox.age.group)

## Call:  
## coxph(formula = Surv(years, status\_numeric) ~ age\_group, data = pbc)  
##   
## n= 312, number of events= 125   
##   
## coef exp(coef) se(coef) z Pr(>|z|)   
## age\_group40-49 0.2101 1.2338 0.3284 0.640 0.522388   
## age\_group50-59 0.7288 2.0726 0.3151 2.313 0.020746 \*   
## age\_group60-79 1.1630 3.1996 0.3267 3.560 0.000371 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## exp(coef) exp(-coef) lower .95 upper .95  
## age\_group40-49 1.234 0.8105 0.6482 2.348  
## age\_group50-59 2.073 0.4825 1.1175 3.844  
## age\_group60-79 3.200 0.3125 1.6866 6.070  
##   
## Concordance= 0.616 (se = 0.026 )  
## Likelihood ratio test= 20.72 on 3 df, p=1e-04  
## Wald test = 20.81 on 3 df, p=1e-04  
## Score (logrank) test = 22.19 on 3 df, p=6e-05

# For a nicely formatted table  
tbl\_regression(cox.age.group, exponentiate = TRUE)

| **Characteristic** | **HR** | **95% CI** | **p-value** |
| --- | --- | --- | --- |
| age\_group |  |  |  |
| 25-39 | — | — |  |
| 40-49 | 1.23 | 0.65, 2.35 | 0.5 |
| 50-59 | 2.07 | 1.12, 3.84 | 0.021 |
| 60-79 | 3.20 | 1.69, 6.07 | <0.001 |
| Abbreviations: CI = Confidence Interval, HR = Hazard Ratio | | | |

## Part D

#### (d) Based on (a)-(c), what do you conclude? Are hazards for the variable age proportional? How would you report the effect (in terms of HR and 95%CI) of age in a paper?

Based on part a-c, we conclude that the hazards for the variable age are proportional. Under the proportional hazards assumption, the hazard ratio (HR) does not vary with time. In part a, we plotted Kaplan-Meier curves with predicted survival under Cox model and from visual assessment, we can see that the Cox - predicted curves closely follow the Kaplan-Meier curves, supporting the proportional hazards assumption. In part b, we find minor evidence of a violation of the proportional hazards assumption for age, as we expect a smoother to be approximately flat, but instead we do not see this. We also find that the HR is not constant over time further supporting a violation of proportional hazards assumption. The Schoenfeld residuals for age: χ²=0.923, p=0.34 and Pearson/Spearman correlations: p=0.37 and p=0.25, which validates the proportional hazards assumption. In the Cox proportional hazards model, older age groups exhibited progressively higher risks of mortality compared to the 25-39 reference group. The 60-79 age group showed the strongest association (HR=3.20, 95% CI:1.69–6.07, p<0.001), followed by the 50-59 group (HR=2.07, 95% CI:1.12–3.84, p=0.021). No significant association was observed for the 40-49 group (HR=1.23, 95% CI:0.65–2.35, p=0.5). Proportional hazards assumptions were validated using Schoenfeld residuals (χ²=0.923, p=0.34) and correlation tests (Pearson p=0.37; Spearman p=0.25).