Problem Set 8, Winter 2024

Michael Ghattas

2024-11-10

```
# Load necessary packages
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

```
library(mlbench)
library(haven)
library(MASS)
library(survival)
library(survminer)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

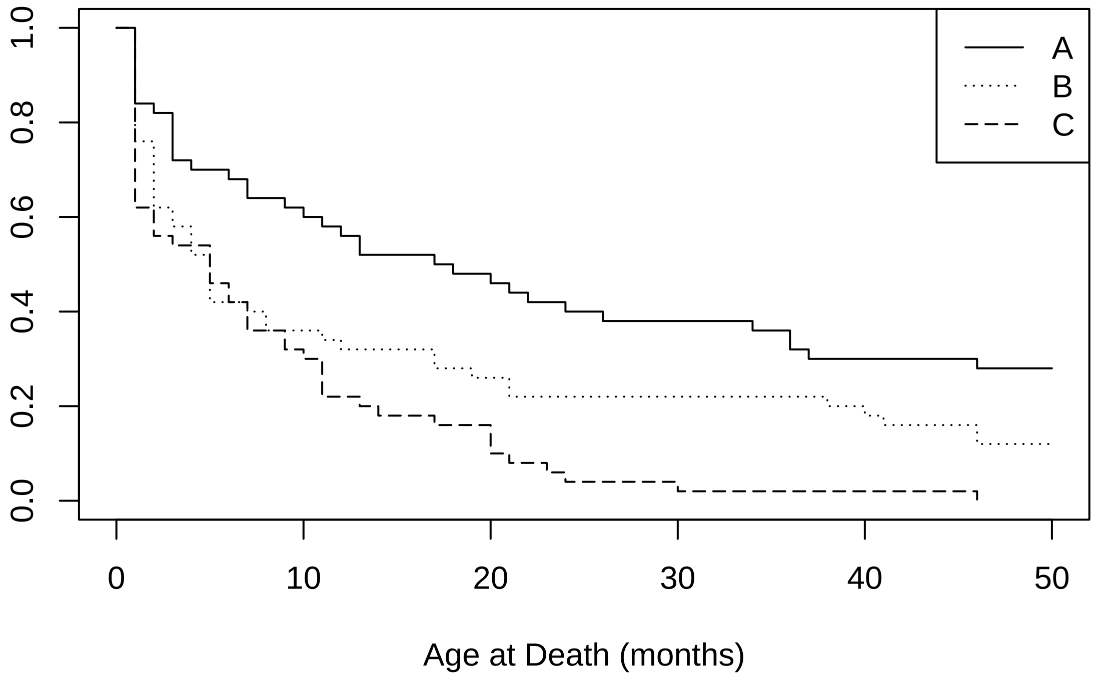
```
##
## Attaching package: 'survminer'
```

```
## The following object is masked from 'package:survival':
##
##     myeloma
```

Question 1 - 25 points

The following code is excerpted from the example shown in 9.1.3. The outcome of interest is time to death of sheep. Each sheep received some level of anti-parasite treatment; A and B contained actual anti-parasite ingredients and C was a placebo (i.e., no active ingredient in the treatment). Please run the three code chunks and examine their output. Once you've done that, answer the four questions below.

```
# Chunk 1
sheep <- read.csv("sheep.deaths.csv")
with(sheep, plot(survfit(Surv(death, status) ~ group), lty = c(1, 3, 5), xlab = "Age at Death (months)"
legend("topright", c("A", "B", "C"), lty = c(1, 3, 5))
```

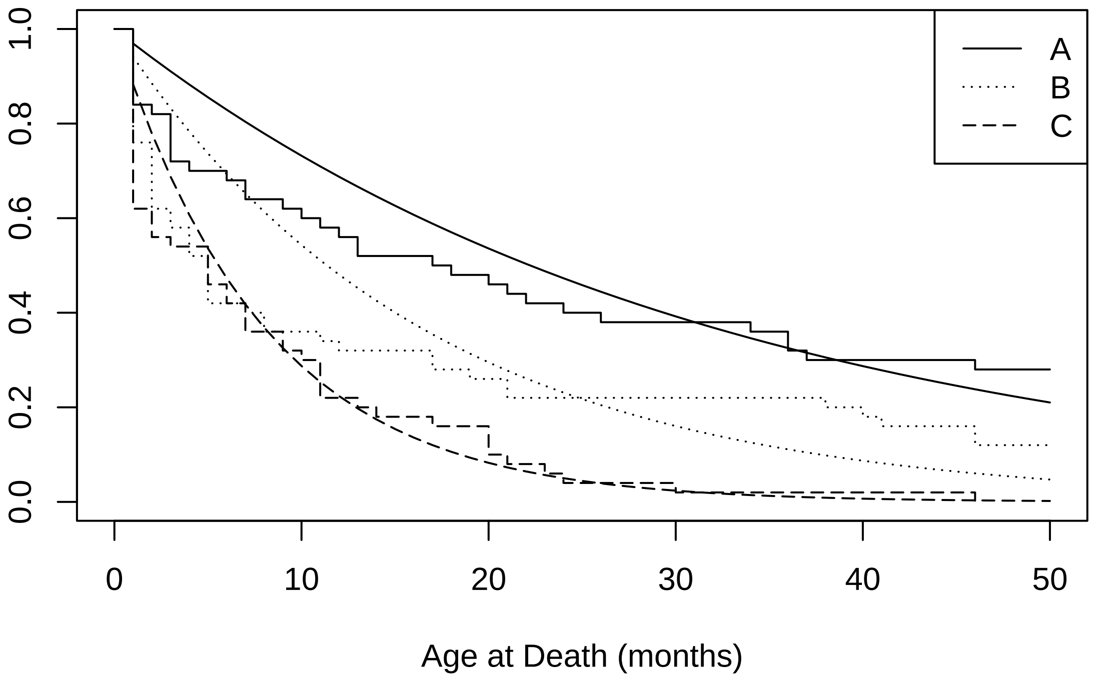


```
# Chunk 2
model <- survreg(Surv(death, status) ~ group, dist = "exponential", data = sheep)
summary(model)
```

```
##
## Call:
## survreg(formula = Surv(death, status) ~ group, data = sheep,
##     dist = "exponential")
##              Value Std. Error     z       p
## (Intercept)  3.467      0.167 20.80 < 2e-16
## groupB      -0.671      0.225 -2.99  0.0028
## groupC      -1.386      0.219 -6.34 2.3e-10
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -482   Loglik(intercept only)= -502.1
##  Chisq= 40.35 on 2 degrees of freedom, p= 1.7e-09
## Number of Newton-Raphson Iterations: 5
## n= 150
```

```
# Chunk 3
plot(survfit(Surv(sheep$death, sheep$status) ~ sheep$group), lty = c(1, 3, 5), xlab = "Age at Death (mo
legend("topright", c("A", "B", "C"), lty = c(1, 3, 5))
points(1:50, 1 - pexp(1:50, rate = 1 / exp(model$coefficients[1])), type = "l", lty = 1)
```

```r
points(1:50, 1 - pexp(1:50, rate = 1 / exp(sum(model$coefficients[c(1, 2)]))), type = "l", lty = 3)
points(1:50, 1 - pexp(1:50, rate = 1 / exp(sum(model$coefficients[c(1, 3)]))), type = "l", lty = 5)
```



Question about Chunk 1

A) What kind of plot is this? It has a specific name.

- Your answer here: This plot is a Kaplan-Meier survival plot, showing the estimated survival function over time for each treatment group (A, B, C).

B) Which group had the most number of sheep whose outcomes were censored?

- Your answer here: The Kaplan-Meier plot indicates that Group C (the placebo) likely has the most censored outcomes, as its survival curve extends higher without reaching many endpoint events, suggesting more censored cases compared to Groups A and B.

C) In the context of this data, what does it mean if a sheep's outcome was censored?

- Your answer here: In this context, a censored outcome means that for a particular sheep, the exact time of death was not observed within the study period. This could occur if the sheep were still alive at the study's end or were lost to follow-up.

Questions about Chunk 2

D) What kind of survival model is being fitted in this code? Be specific.

- Your answer here: The model fitted here is an exponential parametric survival model with treatment group as a covariate, specifically using a constant hazard assumption (exponential distribution).

E) Looking at the p-values, is Group A significantly different from Group B?

- Your answer here: Looking at the summary, the p-value for Group B relative to Group A is 0.0028, which is less than 0.05. Thus, Group A is significantly different from Group B in terms of survival time.

F) Looking at the p-values, is Group A significantly different from Group C?

- Your answer here: The p-value for Group C relative to Group A is 2.3e-10, which is highly significant ($p < 0.05$). Therefore, Group A is significantly different from Group C as well.

G) Looking at the coefficient estimates, which group - B or C - had the lowest predicted survival time?

- Your answer here: Based on the coefficients, Group C has the lowest survival time because it has the most negative coefficient (-1.386), indicating shorter survival compared to both Group A and Group B.

Question about Chunk 3

H) The jagged lines on this plot are the same as those from the plot shown in Chunk 1. What is being visualized by the the *smooth, curved lines* in this plot? Again, be specific.

- Your answer here: The smooth curves represent the fitted exponential survival functions for each group based on the model in Chunk 2. Unlike the Kaplan-Meier estimates, which are non-parametric and empirical, these curves are derived from a parametric model, assuming an exponential distribution with group-specific hazard rates.

Question 2 (25 pts).

Survival in patients with advanced lung cancer from the North Central Cancer Treatment Group. Performance scores rate how well the patient can perform usual daily activities.

```
?lung
```

```
## Help on topic ’lung’ was found in the following packages:
##
##   Package               Library
##   survival              /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library
##   KMsurv                /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library
##
##
## Using the first match ...
```

```r
head(lung)
```

```
##   inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
## 1    3  306      2  74   1       1       90       100     1175      NA
## 2    3  455      2  68   1       0       90        90     1225      15
## 3    3 1010      1  56   1       0       90        90       NA      15
## 4    5  210      2  57   1       1       90        60     1150      11
## 5    1  883      2  60   1       0      100        90       NA       0
## 6   12 1022      1  74   1       1       50        80      513       0
```

We would like to create a model to describe the survival time of lung cancer patients based on their sex and ph.karno score effects. Create a Cox proportional hazard model, an exponential model, and weibull models.

```r
# Cox proportional hazards model
cox_fit <- coxph(Surv(time, status) ~ sex + ph.karno, data = lung)
summary(cox_fit)
```

```
## Call:
## coxph(formula = Surv(time, status) ~ sex + ph.karno, data = lung)
##
##   n= 227, number of events= 164
##    (1 observation deleted due to missingness)
##
##               coef exp(coef)  se(coef)      z Pr(>|z|)
## sex      -0.504210  0.603982  0.167723 -3.006  0.00265 **
## ph.karno -0.015155  0.984959  0.005727 -2.646  0.00814 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## sex          0.604      1.656    0.4348    0.8391
## ph.karno     0.985      1.015    0.9740    0.9961
##
## Concordance= 0.638  (se = 0.026 )
## Likelihood ratio test= 17.05  on 2 df,   p=2e-04
## Wald test            = 17.18  on 2 df,   p=2e-04
## Score (logrank) test = 17.5  on 2 df,   p=2e-04
```

```r
# Exponential
exp.model <- survreg(Surv(time, status) ~ sex + ph.karno, dist = "exponential", data = lung)
summary(exp.model)
```

```
##
## Call:
## survreg(formula = Surv(time, status) ~ sex + ph.karno, data = lung,
##     dist = "exponential")
##               Value Std. Error    z       p
## (Intercept) 4.20530    0.51373 8.19 2.7e-16
## sex         0.48028    0.16717 2.87  0.0041
## ph.karno    0.01443    0.00581 2.48  0.0131
##
## Scale fixed at 1
```

```
##
## Exponential distribution
## Loglik(model)= -1148.5   Loglik(intercept only)= -1156
##  Chisq= 15.13 on 2 degrees of freedom, p= 0.00052
## Number of Newton-Raphson Iterations: 4
## n=227 (1 observation deleted due to missingness)
```

```r
# Weibull
exp.model2 <- survreg(Surv(time, status) ~ ph.karno + sex + ph.ecog, dist = "weibull", data = lung)
summary(exp.model2)
```

```
##
## Call:
## survreg(formula = Surv(time, status) ~ ph.karno + sex + ph.ecog,
##     data = lung, dist = "weibull")
##                 Value Std. Error     z       p
## (Intercept)  6.66434    0.65785 10.13 < 2e-16
## ph.karno    -0.00906    0.00675 -1.34 0.17983
## sex          0.41141    0.12287  3.35 0.00081
## ph.ecog     -0.47729    0.12618 -3.78 0.00016
## Log(scale)  -0.32452    0.06207 -5.23 1.7e-07
##
## Scale= 0.723
##
## Weibull distribution
## Loglik(model)= -1126.3   Loglik(intercept only)= -1141.1
##  Chisq= 29.49 on 3 degrees of freedom, p= 1.8e-06
## Number of Newton-Raphson Iterations: 5
## n=226 (2 observations deleted due to missingness)
```

Question 2.1

A) Interpret the regression coefficients for the Cox model:

- Sex: The coefficient for sex is -0.5042, with an exp(coef) of 0.604, indicating that males (sex=1) have approximately 40% lower hazard of death compared to females, controlling for performance score.
- ph.karno: The coefficient for performance score (ph.karno) is -0.0152, with an exp(coef) of 0.985. This suggests that for each additional point in performance score, the hazard decreases by approximately 1.5%, meaning higher performance scores are associated with better survival.

B) Interpret the regression coefficients for the Exponential model:

- Sex: The positive coefficient for sex (0.4803) suggests that males (sex=1) have higher survival times than females.
- ph.karno: The coefficient for ph.karno (0.0144) indicates that higher performance scores correspond to longer survival times in this exponential model.

C) Are all the variables significant in all the Weibull model. If not which variables are not significant?

- sex: p=0.00081 – This p-value is less than 0.05, indicating sex is statistically significant.
- ph.karno: p=0.17983 – This p-value is greater than 0.05, meaning ph.karno is not statistically significant in this Weibull model.
- ph.ecog: p=0.00016 – This p-value is also less than 0.05, making ph.ecog statistically significant.

Question 2.2

A) Create a model an Exponential model with the variables ph.ecog, sex and ph.karno score as main effects, and compare the model to the Exponential model we generated earlier with sex and ph.karno score using a p-value to compare the models.

- When comparing the model with ph.ecog, sex, and ph.karno against the simpler model (only sex and ph.karno), a likelihood ratio test reveals if the additional variable (ph.ecog) significantly improves the model. The p-value (2.7e-05) for the addition of ph.ecog is less than 0.05, indicating that adding ph.ecog significantly improves model fit. (See output below)

B) Compare the Weibull model with the Exponential model using a p-value and state which model you should use.

- The likelihood ratio test comparing the Weibull model to the Exponential model shows a p-value of 2.24e-05, suggesting that the Weibull model provides a significantly better fit. Therefore, the Weibull model is preferable. (See output below)

```
# Exponential model comparison
anova(exp.model, exp.model2, test = "Chisq")
```

```
##                      Terms Resid. Df    -2*LL     Test Df Deviance     Pr(>Chi)
## 1           sex + ph.karno       224 2296.957          NA       NA           NA
## 2 ph.karno + sex + ph.ecog       221 2252.632 +ph.ecog  3 44.32513 1.287275e-09
```

```
# Fit Weibull model and compare
weibull_model <- survreg(Surv(time, status) ~ sex + ph.karno, dist = "weibull", data = lung)
anova(exp.model, weibull_model, test = "Chisq")
```

```
##            Terms Resid. Df    -2*LL Test Df Deviance     Pr(>Chi)
## 1 sex + ph.karno       224 2296.957      NA       NA           NA
## 2 sex + ph.karno       223 2278.987    =  1 17.96995 2.244204e-05
```

Question 2.3

A) Assess the assumption of proportional hazards for the Cox model.

- The p-value for ph.karno is 0.0053, which is less than 0.05, suggesting that ph.karno may violate the proportional hazards assumption. (See output below)
- The global test also shows a significant result (p=0.0060), indicating that the proportional hazards assumption may not hold overall. (See output below)

B) Explain why it would not be ideal to use the Kaplan Meier model with the variables sex and ph.karno score.

- The Kaplan-Meier model is non-parametric and does not adjust for multiple covariates (like sex and ph.karno). Therefore, it cannot account for these covariates simultaneously, which may lead to oversimplified results when there are multiple predictors with different impacts on survival. (See output below)

```
# Checking proportional hazards assumption
cox.zph(cox_fit)
```

```
##            chisq df      p
## sex         3.32  1 0.0686
## ph.karno    7.76  1 0.0053
## GLOBAL     10.23  2 0.0060
```