Lab Assignment 7 BIS 505b

Wenxin Xu

5/7/2021

Contents

Instructions

This Lab Assignment analyzes data from a high school athletic injury surveillance program. You may keep the sections on **Data Background** and the **Data Key** in your submission if you wish. Perform your work in the **Assignment** section below. In this assignment, report any p-values that are less than 0.001 as **<0.001** and round values reported in your narrative text to **3** decimal places. **Be sure to clearly state the reference category when interpreting the effects of categorical variables in any regression model.** Perform all hypothesis testing at the $\alpha = 0.05$-level.

Data Background

It is estimated that 7 million students participate in high school sports. To study factors related to returning to play after a sports-related injury, several high schools agreed to participate in an injury surveillance program. At each high school, a certified athletic trainer provided daily medical coverage at all scheduled practices and games. The athletic trainer was responsible for data collection for the injury surveillance program. Data elements collected included pain score [`pain`], previous history of injury [`prior`], cause of injury [`cause`] and biological sex [`sex`] of each student athlete. In addition, date of injury and the date the student athlete returned to play are recorded. The time to return-to-play (event of interest) or the end of the study (censoring) [`rtp`] is calculated [`t`]. Within a season, 137 student athletes experienced an injury and are included in this study. A CSV file [`sport.csv`] is provided which contains data from the student athletes in the study.

Data Key – `sport.csv`

Variable Name	Definition
`id`	Unique identifier for each subject
`cause`	Cause of injury. Injury due to… 1 = Contact with ground 2 = Contact with person 3 = Repetition (reference)

Variable Name	Definition
`pain`	Pain score (higher = greater pain)
`prior`	Prior injuries 0 = No (reference) 1 = Yes
`sex`	Sex 0 = Female (reference) 1 = Male
`t`	Time to return-to-play or censoring (days)
`rtp`	Return-to-play 0 = No 1 = Yes

Assignment

1. [5 points] Import the CSV file `sport.csv` in the third code chunk above. Name your data frame `sport` and create the factor variables `cause_factor`, `prior_factor`, and `sex_factor` using the reference categories indicated in the **Data Key**. After these steps, `sport` should contain 10 variables. [**Note:** No written response is required for this question. Display the code chunk(s) that perform the requested data management steps.]

```
# Creating factor variables in whas using mutate() function in "dplyr" package
sport <- mutate(sport,
               cause_factor = factor(cause,
                                   levels = 3:1,
                                   labels = c("Repetition", "Contact with person", "Contact with ground
               prior_factor = factor(prior,
                                   levels = 0:1,
                                   labels = c("No", "Yes")),
               sex_factor = factor(sex,
                                      levels = 0:1,
                                      labels = c("Female", "Male"))
               )

# after create 3 factor variables, sport has 10 variables
ncol(sport)
```

```
## [1] 10
```

2. The **research question** of this study is to identify factors associated with *return-to-play* (i.e., recovery) following a sports-related injury. We will begin our analysis with descriptive statistics and graphical summaries.

2a. [5 points] In the 137 injured athletes, report the number and percentage of student athletes who were *not* able to return-to-play before the study (season) ended.

```
tab = cbind(count = table(sport$rtp),
             percentage=prop.table(table(sport$rtp)))
tab
```

```
##   count percentage
## 0     9 0.06569343
## 1   128 0.93430657
```

In the 137 injured athletes, 9 (6.569%) of student athletes were not able to return-to-play before the study ended.

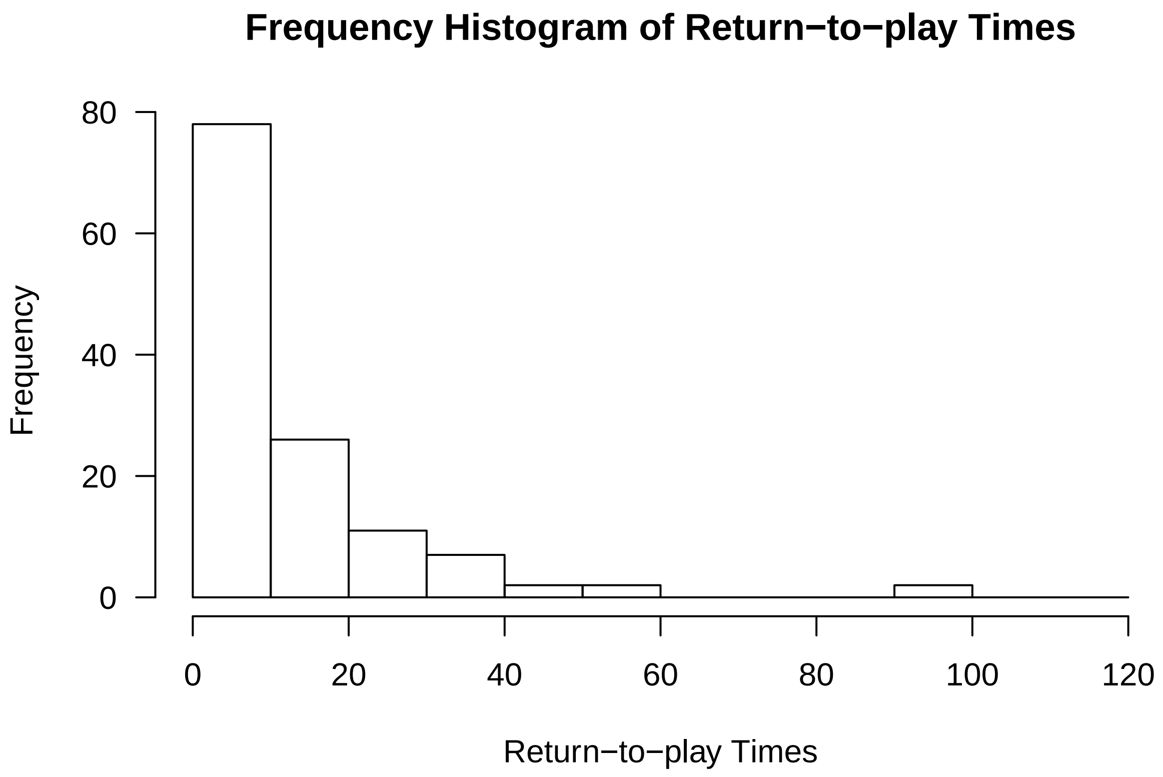
2b. [10 points] In the subset of student athletes who returned to play, report the mean, median, and range (min, max) of time to return-to-play and create a frequency histogram of return-to-play times. Comment on the shape of this distribution.

```
# report the mean, median, and range (min, max) of time to return-to-play

# subset student athletes who returned to play
sport_return = subset(sport, rtp == 1)

# calculate mean, median, and range (min, max) of time to return-to-play
tab2 = cbind(mean = mean(sport_return$t),
             median = median(sport_return$t),
             range = range(sport_return$t))

tab2
```

```
##          mean median range
## [1,] 12.2125    6.2   0.1
## [2,] 12.2125    6.2  99.9
```

Among the student athletes who returned to play,the mean of time to return-to-play is 12.213, the median is 6.2, and range is (0.1,99.9).

```
# create a frequency histogram of return-to-play times
hist(sport_return$t,
     breaks = seq(0, 120, by=10),
     right = FALSE, # left-closed intervals (preferred)
     main = "Frequency Histogram of Return-to-play Times",
     xlab = "Return-to-play Times",
     ylab = "Frequency",
     las = 1) # rotate y-axis text
```

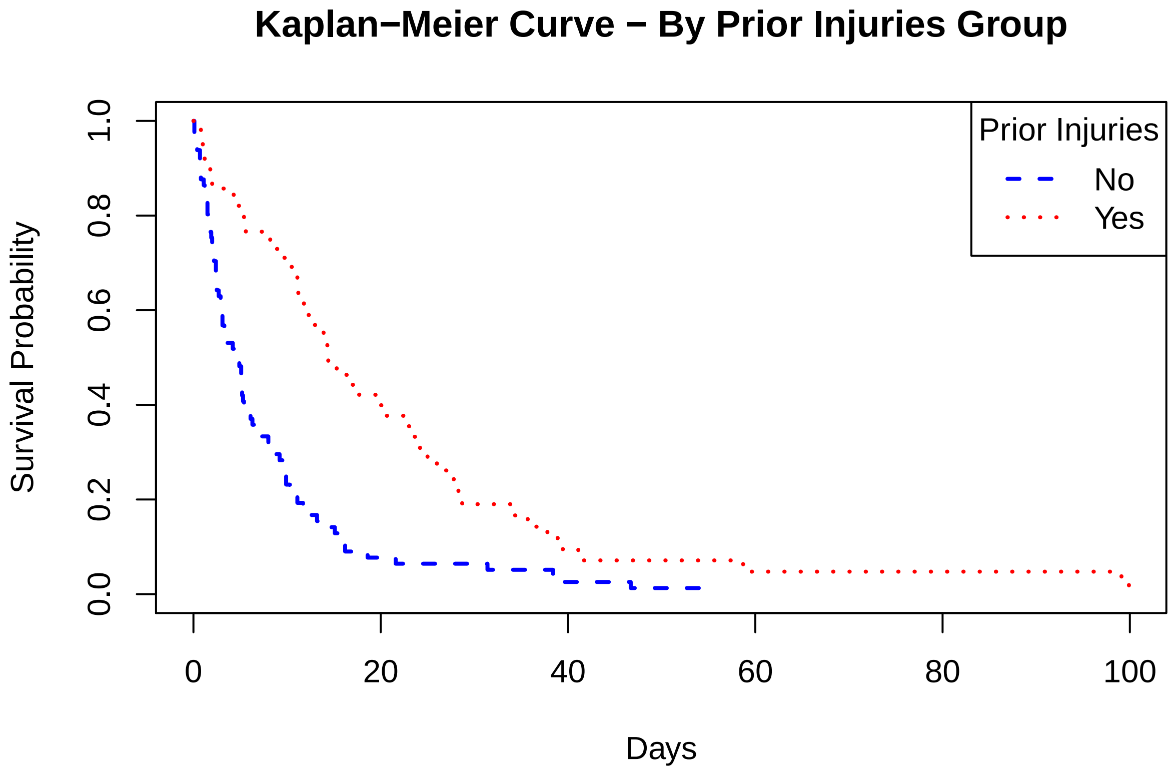


The shape of this distribution is highly right-skewed, most of the return-to-play times are small (0-10 days), and only a very samll amount of return-to-play times are large (90-100 days).

3. Student athletes who experienced past injuries [`prior`] are expected to have longer time to return to play than athletes who have not had a previous injury.

3a. [10 points] Plot the Kaplan-Meier survival curves comparing time to return-to-play in those with and without previous injuries. Be sure to include a figure legend. You should see that one curve is consistently above the other. How do we interpret survival probabilities in this context? Does the group with past injuries tend to have a longer time to return-to-play (recovery)? [*Note:* When thinking about survival analysis in general (i.e., modeling time-to-event such as time-to-death), you may find it helpful to think of a survival probability $S(t)$ as the probability of being event-free at time t.]

```
#Plot the Kaplan-Meier survival curves comparing time to return-to-play in those with and without previ
km.prior = survfit(Surv(t, rtp) ~ prior_factor, data = sport)

plot(km.prior,
     xlab = "Days",
     ylab = "Survival Probability",
     col = c("blue", "red"),
     lty = 2:4,
     lwd = 2)

# Be sure to include a figure legend
legend("topright",
       title = "Prior Injuries",
       legend = levels(sport$prior_factor),
       col = c("blue", "red"),
       lty = 2:4,
       lwd = 2)

title("Kaplan-Meier Curve - By Prior Injuries Group")
```



- How do we interpret survival probabilities in this context? The survival probability $S(t)$ is the probability of not being able to return-to-play or not being recovered at time t.
- Does the group with past injuries tend to have a longer time to return-to-play (recovery)?

Yes, the curve of the group with past injuries is consistently above the curve of the group without past injuries (ref).

3b. [5 points] Report median survival times for each group. Which group has the longer median survival time? In this case, is it better or worse to have a longer median survival time? Explain.

```
# median survival times by prior injuries group
quantile(km.prior)$quantile[,2]
```

```
##  prior_factor=No prior_factor=Yes
##              4.8              14.4
```

- Report median survival times for each group.

The median survival time in the group with past injuries is 14.4 days, while median survival time in the group without past injuries is 4.8 days.

- Which group has the longer median survival time?

The group with past injuries has the longer median survival time.

- In this case, is it better or worse to have a longer median survival time? Explain.

It's worse to have a longer median survival time. The median survival time is estimated as the smallest survival time for which the survivor function $\hat{S}(t) \leq 0.5$. Because the event (return-to-play) is positive, a longer median survival time means a longer time to recovery.

3c. [7 points] Perform a log-rank test to compare the survivor functions for the two prior injury groups. (i) State the null and alternative hypotheses; (ii) From your **R** output, report the value of the test statistic and p-value; (iii) State your statistical conclusion and your conclusion in the context of the problem.

```r
# Perform a log-rank test to compare the survivor functions for the two prior injury groups
logrank.prior = survdiff(Surv(t, rtp) ~ prior_factor, data = sport)

logrank.prior
```

```
## Call:
## survdiff(formula = Surv(t, rtp) ~ prior_factor, data = sport)
##
##                    N Observed Expected (O-E)^2/E (O-E)^2/V
## prior_factor=No  81       80     53.7     12.92        24
## prior_factor=Yes 56       48     74.3      9.33        24
##
##  Chisq= 24  on 1 degrees of freedom, p= 1e-06
```

(i) State the null and alternative hypotheses;

$H_0 : S_1(t) = S_2(t)$ for all t vs. $H_1 : S_1(t) \neq S_2(t)$ for some t

(ii) From your **R** output, report the value of the test statistic and p-value;

The log-rank test statistic is $\chi^2 = 24.009$, and p-value $<.001$

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We have evidence to reject H_0 and conclude that the survival experience is significantly different in the two prior injury groups, which indicates that student athletes who experienced past injuries have a longer time to return to play than athletes who have not had a previous injury (ref).

3d. [8 points] Fit a simple Cox proportional hazards regression model to estimate the unadjusted hazard ratio for those with prior injuries vs. those without prior injuries. Report the equation of the fitted simple Cox PH regression model. Report and interpret the unadjusted hazard ratio associated with the prior injury indicator variable and report its 95% confidence interval. Does this point estimate support what you saw graphically in the Kaplan-Meier curves?

```r
# Fit a simple Cox proportional hazards regression model
cox.prior = coxph(Surv(t, rtp) ~ prior_factor, data = sport)
summary(cox.prior)
```

```
## Call:
## coxph(formula = Surv(t, rtp) ~ prior_factor, data = sport)
##
##   n= 137, number of events= 128
##
##                      coef exp(coef) se(coef)     z Pr(>|z|)
## prior_factorYes -0.9111    0.4021   0.1910 -4.77 1.84e-06
##
##                 exp(coef) exp(-coef) lower .95 upper .95
## prior_factorYes    0.4021      2.487    0.2765    0.5847
##
## Concordance= 0.628  (se = 0.023 )
## Likelihood ratio test= 23.89  on 1 df,   p=1e-06
## Wald test            = 22.75  on 1 df,   p=2e-06
## Score (logrank) test = 24.13  on 1 df,   p=9e-07
```

- Report the equation of the fitted simple Cox PH regression model

The fitted model is given by the equation, $\log(\hat{h}(t;x)) = \log(\hat{h}_0(t)) - 0.911$ Prior Injury

- Report and interpret the unadjusted hazard ratio associated with the prior injury indicator variable and eport its 95% confidence interval.

The unadjusted hazard ratio is given by the exponentiated slope $\hat{HR} = e^{b_1} = 0.402$ [95% CI (0.277, 0.585)], which means that student athletes who experienced past injuries have 0.402 times the hazard of returning to play compared to athletes who have not had a previous injury (ref).

- Does this point estimate support what you saw graphically in the Kaplan-Meier curves?

Yes, because a smaller hazard of return to play in student athletes who experienced past injuries means a longer time to recovery.

4. Next, we will examine the effect of an athlete's pain score [`pain`] on time to return-to-play.

4a. [10 points] Create a dichotomous variable `painhigh` that is equal to `1` if the pain score > 4 (high pain score) and `0` otherwise (low pain score, reference). How many individuals are in each group? Plot the Kaplan-Meier survival curves comparing time to return-to-play in the two pain score groups. Be sure to include a figure legend. Which group tends to return-to-play sooner? Can you think of an explanation for the direction of the effect observed here?

```r
# Create a dichotomous variable `painhigh` for quantitative variable `pain`
# 1: pain score > 4 (high pain score) ; 0: pain score <= 4 (low pain score, ref).
sport$painhigh = ifelse(sport$pain > 4, 1, 0)

sport = mutate(sport,
               painhigh_factor=factor(painhigh,
                                      levels = 0:1,
                                      labels = c("Low pain score","High pain score")))
```

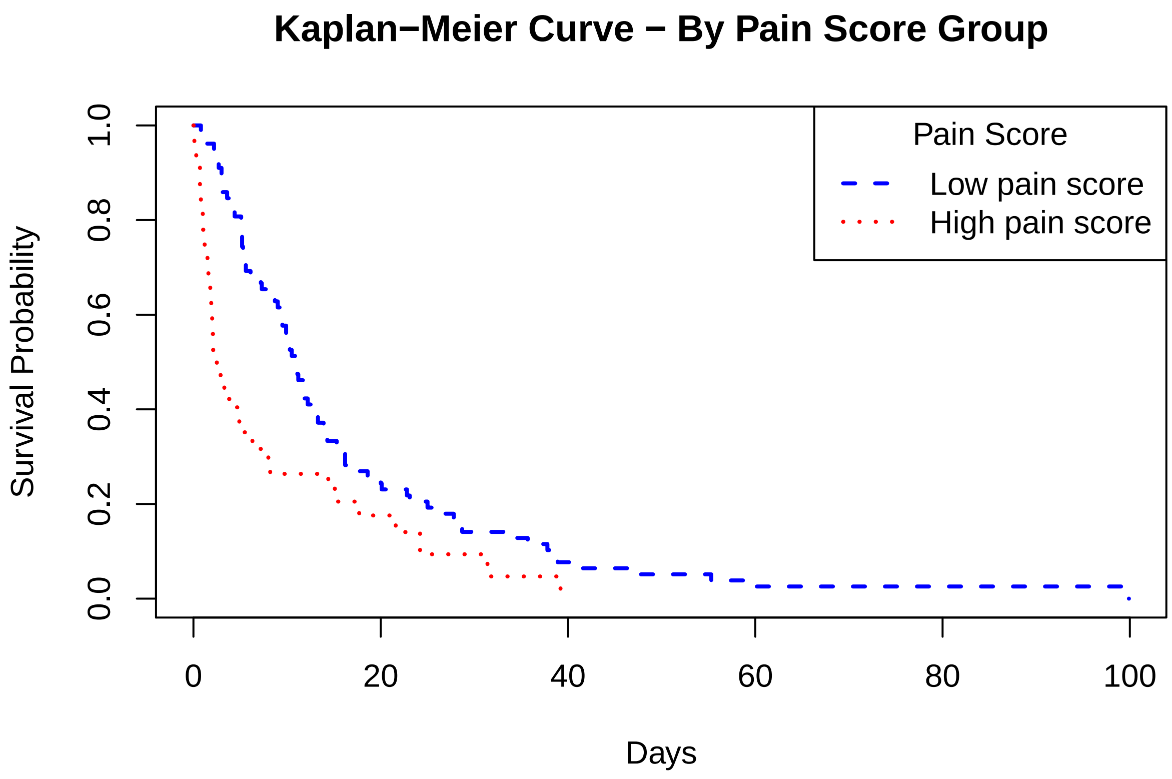
- How many individuals are in each group?

```r
table(sport$painhigh_factor)
```

```
##
##  Low pain score High pain score
##              78              59
```

There are 59 student athletes who have high pain score and 78 student athletes who have low pain score (ref).

```r
# Plot the Kaplan-Meier survival curves comparing time to return-to-play in the two pain score groups.
km.prior = survfit(Surv(t, rtp) ~ painhigh_factor, data = sport)

plot(km.prior,
     xlab = "Days",
     ylab = "Survival Probability",
     col = c("blue", "red"),
     lty = 2:4,
     lwd = 2)

# Be sure to include a figure legend
legend("topright",
       title = "Pain Score",
       legend = levels(sport$painhigh_factor),
       col = c("blue", "red"),
       lty = 2:4,
       lwd = 2)

title("Kaplan-Meier Curve - By Pain Score Group")
```



- Which group tends to return-to-play sooner?

Student athletes group who have high pain score tends to return-to-play sooner.

- Can you think of an explanation for the direction of the effect observed here?

Student athletes who experienced high level of pain may get more intense treatment than student athletes who experienced low level of pain, which make them return-to-play sooner.

4b. [10 points] Fit a simple Cox proportional hazards regression model to estimate the unadjusted hazard ratio associated with *quantitative* pain score [`pain`]. Report the equation of the fitted simple Cox PH regression model. Report and interpret the unadjusted hazard ratio associated with the pain score and report its 95% confidence interval. Does this point estimate support what you saw graphically in the Kaplan-Meier curves? Finally, using your Cox PH model, perform a hypothesis test to determine if there is a significant association between the hazard of returning to play and the athlete's pain score (i) State the null and alternative hypotheses; (ii) From your **R** output, report the value of the test statistic and p-value; (iii) State your statistical conclusion and your conclusion in the context of the problem.

```
# Fit a simple Cox proportional hazards regression model with quantitative variable pain score `pain`
cox.pain = coxph(Surv(t, rtp) ~ pain, data = sport)
summary(cox.pain)
```

```
## Call:
## coxph(formula = Surv(t, rtp) ~ pain, data = sport)
##
##   n= 137, number of events= 128
##
##         coef exp(coef) se(coef)     z Pr(>|z|)
## pain 0.14862   1.16024  0.04087 3.636 0.000277
##
##      exp(coef) exp(-coef) lower .95 upper .95
## pain      1.16     0.8619     1.071     1.257
```

```
##
## Concordance= 0.658  (se = 0.026 )
## Likelihood ratio test= 12.38  on 1 df,   p=4e-04
## Wald test            = 13.22  on 1 df,   p=3e-04
## Score (logrank) test = 13.43  on 1 df,   p=2e-04
```

- Report the equation of the fitted simple Cox PH regression model.

The fitted model is given by the equation, $\log(\hat{h}(t;x)) = \log(\hat{h}_0(t)) + 0.149 \text{ Pain}$

- Report and interpret the unadjusted hazard ratio associated with the pain score and report its 95% confidence interval.

The unadjusted hazard ratio is given by the exponentiated slope $\hat{HR} = e^{b_1} = 1.16$ [95% CI (1.071, 1.257)], which means that a 1-unit increase in pain score increases the hazard of return-to-play by 16%.

- Does this point estimate support what you saw graphically in the Kaplan-Meier curves?

Yes, because a larger hazard of return to play in student athletes who experienced high level of pain means a shorter time to recovery.

- Finally, using your Cox PH model, perform a hypothesis test to determine if there is a significant association between the hazard of returning to play and the athlete's pain score

(i) State the null and alternative hypotheses;

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$

(ii) From your **R** output, report the value of the test statistic and p-value;

The z-statistic $z = 3.636$, p-value $<.001$

(iii) State your statistical conclusion and your conclusion in the context of the problem.

We have evidence to reject H_0 and conclude that the hazard of returning to play is significantly associated with the athletes' pain score.

5. In this question, we would like to estimate the adjusted effect of pain score and the adjusted effect of prior injury status. We will estimate these effects using a multiple Cox proportional hazards model that contains quantitative pain score, prior injury status, cause of injury, and athlete's sex.

5a. [10 points] Build the multiple Cox proportional hazards model described in the previous sentence. Report the equation of the fitted Cox PH regression model. Report and interpret the adjusted hazard ratio associated with pain score and report its 95% confidence interval. Based on the p-value of the coefficient of the quantitative pain variable in this model, is there a significant association between the hazard of returning to play and the athlete's pain score?

```r
# Build the multiple Cox proportional hazards model with quantitative pain score, prior injury status,
cox.mul = coxph(Surv(t, rtp) ~ pain + prior_factor + cause_factor + sex_factor, data = sport)
summary(cox.mul)
```

```
## Call:
## coxph(formula = Surv(t, rtp) ~ pain + prior_factor + cause_factor +
##     sex_factor, data = sport)
##
##   n= 137, number of events= 128
##
##                                    coef exp(coef) se(coef)      z Pr(>|z|)
## pain                            0.13407   1.14348  0.04191  3.199 0.001378
## prior_factorYes                -0.70881   0.49223  0.21461 -3.303 0.000957
## cause_factorContact with person  0.15170   1.16381  0.28906  0.525 0.599714
## cause_factorContact with ground  0.67421   1.96248  0.27122  2.486 0.012925
## sex_factorMale                    0.27711   1.31932  0.19042  1.455 0.145591
##
##                                 exp(coef) exp(-coef) lower .95 upper .95
## pain                               1.1435     0.8745    1.0533    1.2414
## prior_factorYes                    0.4922     2.0316    0.3232    0.7496
## cause_factorContact with person    1.1638     0.8592    0.6604    2.0508
## cause_factorContact with ground    1.9625     0.5096    1.1533    3.3394
## sex_factorMale                     1.3193     0.7580    0.9084    1.9162
##
## Concordance= 0.714  (se = 0.023 )
## Likelihood ratio test= 45.63  on 5 df,   p=1e-08
## Wald test            = 44.88  on 5 df,   p=2e-08
## Score (logrank) test = 47.29  on 5 df,   p=5e-09
```

- Report the equation of the fitted Cox PH regression model.

The fitted model is given by the equation, $\log(\hat{h}(t;x)) = \log(\hat{h}_0(t)) + 0.134$ Pain $-$ 0.709 Prior Injury + 0.152 Contact with ground + 0.674 Contact with person + 0.277 Male

- Report and interpret the adjusted hazard ratio associated with pain score and report its 95% confidence interval.

Controlling for all of the other variables in the model, as pain score increases, the hazard of returning to play increases. A 1-unit increase in pain score increases the hazard of returning to play by 14%; adjusted $\hat{HR} = e^{b_1} = 1.143$ [95% CI (1.053, 1.241)].

- Based on the p-value of the coefficient of the quantitative pain variable in this model, is there a significant association between the hazard of returning to play and the athlete's pain score?

P-value = 0.001, we have evidence to reject H_0 and conclude that the hazard of returning to play is significantly associated with the athletes' pain score when controlling for all of the other variables in the model.

5b. [10 points] Report and interpret the adjusted hazard ratio associated with prior injury status and report its 95% confidence interval. Based on the p-value of the coefficient of the prior injury indicator in this model, is there a significant difference in the hazard of returning to play in those with and without prior injuries? How does the unadjusted hazard ratio for the effect of prior injury computed in question **3d** compare to the adjusted hazard ratio seen here? Has the adjusted effect moved closer to or away from the null?

- Report and interpret the adjusted hazard ratio associated with prior injury status and report its 95% confidence interval.

Controlling for all of the other variables in the model, the adjusted hazard ratio is given by the exponentiated slope $\hat{HR} = e^{b_2} = 0.492$ [95% CI (0.323, 0.75)], which means that student athletes who experienced past injuries have 0.492 times the hazard of returning to play compared to athletes who have not had a previous injury (ref).

- Based on the p-value of the coefficient of the prior injury indicator in this model, is there a significant difference in the hazard of returning to play in those with and without prior injuries?

P-value $<.001$, we have evidence to reject $H_0 : \beta_2 = 0$ and conclude that there is a significant difference in the hazard of returning to play in those with and without prior injuries when controlling for all of the other variables in the model.

- How does the unadjusted hazard ratio for the effect of prior injury computed in question **3d** compare to the adjusted hazard ratio seen here?

The unadjusted hazard ratio for the effect of prior injury is smaller than the adjusted hazard ratio: 0.402 vs. 0.492.

- Has the adjusted effect moved closer to or away from the null?

The adjusted effect moved closer to the null because under null hypothesis, the hazard ratio is $\hat{HR} = e^0 = 1$.

5c. [10 points] Using your multiple Cox PH model, plot the adjusted survival curves for those with and without prior injury. Assume the pain score is equal to the average pain score in the sample, the injury was due to repetition, and the athlete sex is male. Report median survival time for the two levels of prior injury.

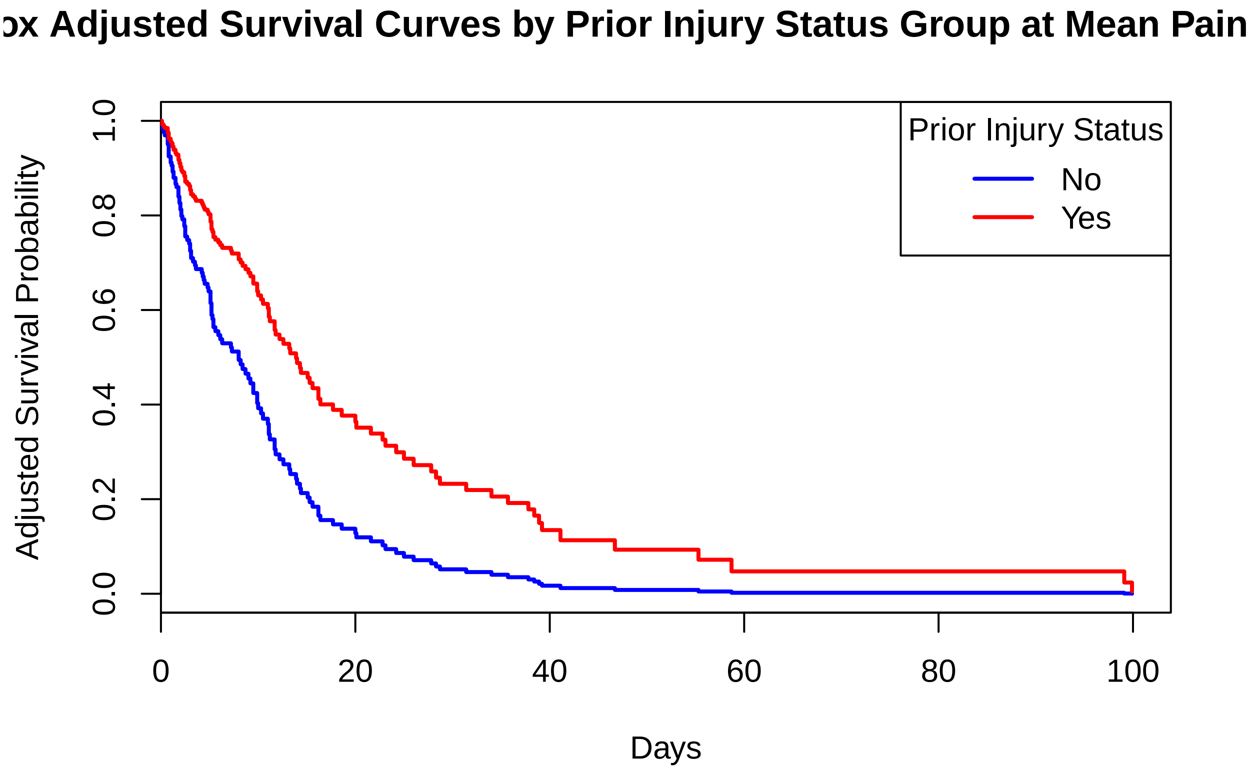
```r
## plot the adjusted survival curves by prior injury status group

# pain score = average pain score, cause= repetition, sex=male
pred.x = data.frame(pain = mean(sport$pain, na.rm=TRUE),
                    expand.grid(prior_factor = levels(sport$prior_factor),
                                cause_factor = levels(sport$cause_factor)[1],
                                sex_factor = levels(sport$sex_factor)[2])
                    )

# adjusted survival probabilities S(t)
Shat = survfit(cox.mul, newdata = pred.x, data=sport)

# Report median survival time for the two levels of prior injury
cbind(pred.x, quantile(Shat)$quantile)
```

```
##       pain prior_factor cause_factor sex_factor  25   50   75
## 1 4.470803           No   Repetition       Male 2.7  8.0 13.9
## 2 4.470803          Yes   Repetition       Male 5.6 13.9 28.3
```

- Report median survival time for the two levels of prior injury.

The adjusted median survival time for student athletes who experienced past injuries when pain score is equal to the average pain score in the sample, the injury was due to repetition, and the athlete sex is male is 13.9 days, while for student athletes who have not experienced past injuries is 8 days.

```r
# Plot of adjusted survival curve at fixed values of x
plot(Shat, xlab = "Days", ylab = "Adjusted Survival Probability",
     col = rep(c("blue", "red"), 2),
     lwd = 2,
     lty = c(rep(1,3), rep(2,3)),
     xaxs= "S")

legend("topright", title = "Prior Injury Status",
       legend = levels(sport$prior_factor),
       col = c("blue", "red"), lwd = 2)

title("Cox Adjusted Survival Curves by Prior Injury Status Group at Mean Pain Score")
```



ox Adjusted Survival Curves by Prior Injury Status Group at Mean Pain