KU LEUVEN

Statistical Analysis of Reliability and Survival Data

Group Members:
Shiqi Wang,
Hao Zhu,
Xiao Wang

May 2021

Introduction

Survival analysis which studies the time until an event of interest takes place is widely used in the fields of medicine, agriculture, sociology and engineering. In this report, we will apply the methods and software we learned in the course to analyze the dataset we selected. In part A, we will do a descriptive analysis of the variables in the dataset and mainly consider the variable which is censored. In part B, we will focus on the response variable and categorical variable. In part C, we will consider all variables in the dataset by using the COX model and parametric regression model. Through the research, we can have a better understanding of the course.

Part A - Exploratory Analysis

The dataset used in this analysis is the *melanom* dataset in the R package *ISwR*, which recorded information about survival conditions of patients after receiving an operation for malignant melanoma. The dataset has 205 rows and 6 columns, and the description of variables is shown in Table 1.

variable	description
no	a numeric vector,patient code;
status	a numeric vector code,survival status; 1: dead from melanoma, 2: alive, 3: dead from other causes;
days	a numeric vector, observation time;
ulc	a numeric vector code, ulceration; 1: present, 2: absent;
thick	a numeric vector, tumor thickness (1/100 mm);
sex	a numeric vector code; 1: female, 2: male.

Table 1 Variable and Description

To have an overview of the data, summary statistics were first computed. Figure 1 shows that all the variables are numeric. Since ulc and sex are actually categorical variable but are numerically coded in the dataset, we will convert them to categorical variable in the following model buiding process.

```
       no              status           days            ulc             thick            sex       
 Min.   :  2.0   Min.   :1.00   Min.   :  10   Min.   :1.000   Min.   :  10   Min.   :1.000  
 1st Qu.:222.0   1st Qu.:1.00   1st Qu.:1525   1st Qu.:1.000   1st Qu.:  97   1st Qu.:1.000  
 Median :469.0   Median :2.00   Median :2005   Median :2.000   Median : 194   Median :1.000  
 Mean   :463.9   Mean   :1.79   Mean   :2153   Mean   :1.561   Mean   : 292   Mean   :1.385  
 3rd Qu.:731.0   3rd Qu.:2.00   3rd Qu.:3042   3rd Qu.:2.000   3rd Qu.: 356   3rd Qu.:2.000  
 Max.   :992.0   Max.   :3.00   Max.   :5565   Max.   :2.000   Max.   :1742   Max.   :2.000  
```

	no	status	days	ulc	thick	sex
Min	2.0	1.00	10	1.000	10	1.000

1st Qu	222.0	1.00	1525	1.000	97	1.000
Median	469.0	2.00	2005	2.000	194	1.000
Mean	463.9	1.79	2153	1.561	292	1.385
3rd Qu	731.0	2.00	3042	2.000	356	2.000
Max	992.0	3.00	5565	2.000	1742	2.000

Figure XX Summary Statistics

From Figure XX we can see that more than a half of patients (134 out of 205) are still alive at the end of the study and Figure XX shows the observation time and survival status for each patient , which means the observations with status 2 and 3 as censored.

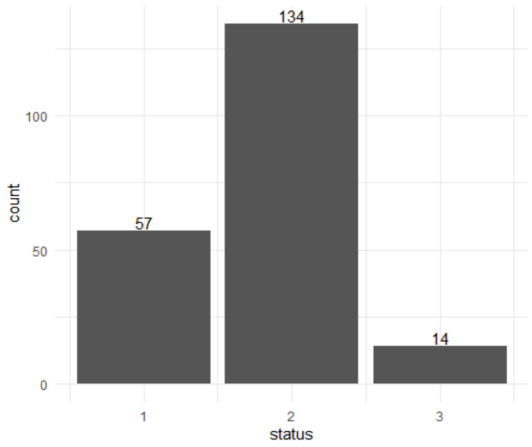

Figure XX

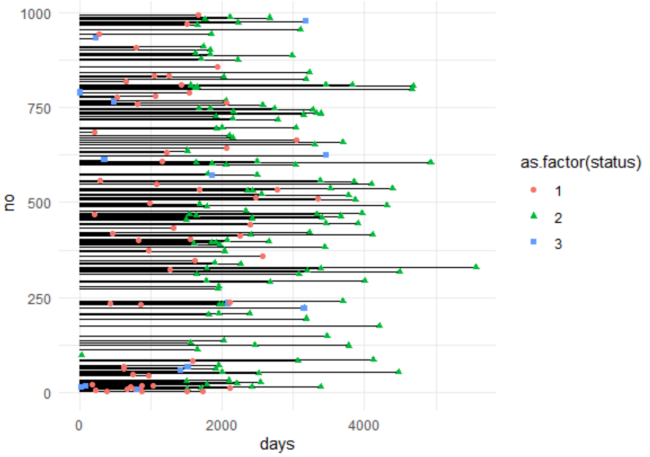

Figure XX

For variables ucl and sex, we can see that less than a half of patients had ulceration appeared during the study and more than a half of the patients are female.

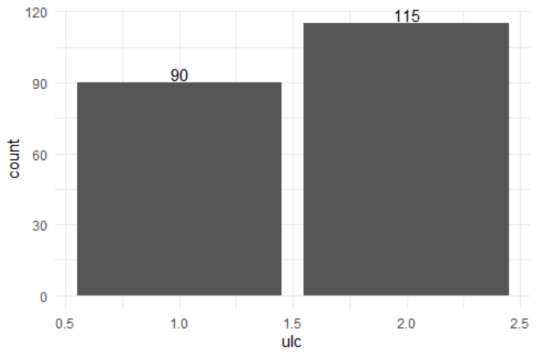

Figure XX

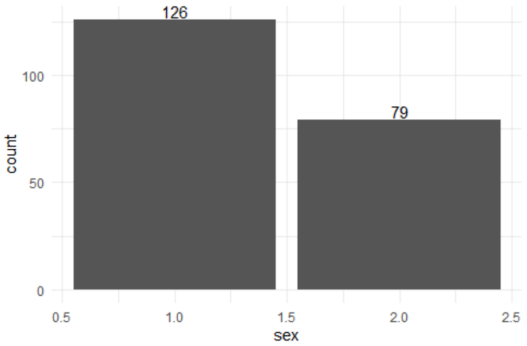

Figure XX

Figure XX shows the distribution of tumor thickness is right skewed.

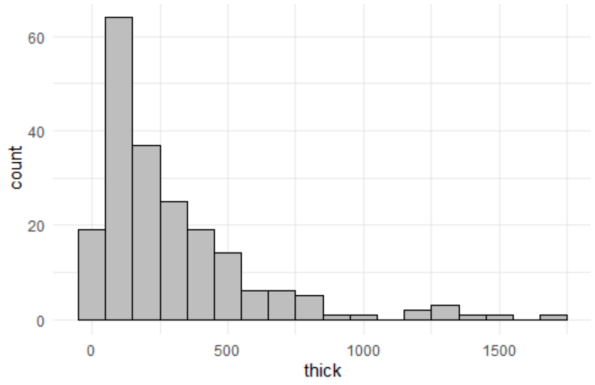

Figure XX

In summary, the exploratory analysis above suggests that most of the patients are still alive at the end of the study, which means most observations are censored in this case. Besides, more than a half of patients have no ulceration during the study and there are more female patients than males. Moreover, the distribution of tumor thickness is right skewed.

Part B - Nonparametric Methods

(a) Kaplan–Meier estimates

To investigate if sex has an influence on patients' survival probabilities, survival functions for females and males are computed using Kaplan–Meier estimates and log-log transformation. As shown in Figure XX, it seems that females have higher survival chances than males at the same observation time.

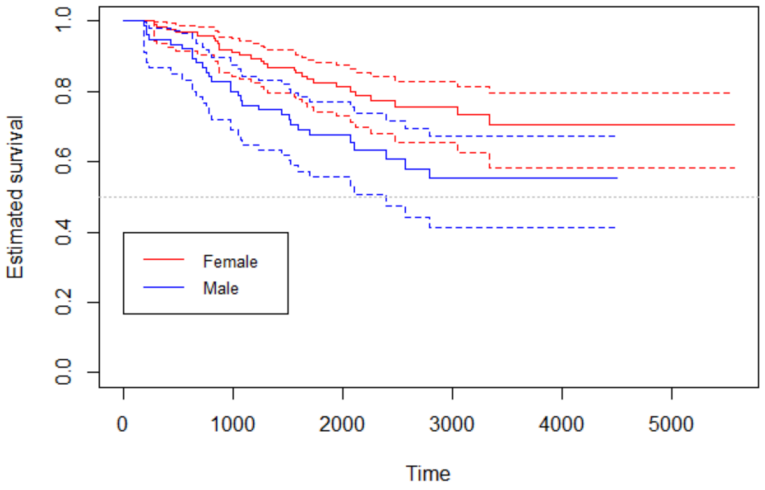

Figure XX

Figure XX shows that the estimates of the median of survival function is NA for survival functions for both males and females. This is because both of the survival curves did not reach the 50% line before the end of the study. The reasons for the NA in the estimates of confidence intervals are the same.

```
Call: survfit(formula = Surv(days, status == 1) ~ sex, data = mln,
    type = "kaplan-meier", conf.type = "log-log")

        n events median 0.95LCL 0.95UCL
sex=1 126     28     NA      NA      NA
sex=2  79     29     NA    2388      NA
```

Figure XX

(b) Quartiles of survival curves

Similarly, the NA in the first and third quartiles are due to the reason that the survival curves have not reached the 75% and 25% respectively at the end of the study, as shown in Figure XX. For interpretation, we can take the first quantile of males patients as an example. The first quantile is the time associated with the first survival probability in the survival curve that is less than or equal to 0.75. The point estimate of the first quartile for males is 1228, which means the corresponding observation time is expected to be 1228 days when the survival probability first reached 0.75. In other words, the probability of a male patient surviving just past 1228 days is expected to be 0.75. For the confidence interval, we can say that we have 95% confidence that the corresponding observation time that a male patient has 75% surviving probability is between 779 and 2103 days.

```
$quantile
        25 50 75
sex=1 3042 NA NA
sex=2 1228 NA NA

$lower
        25   50 75
sex=1 1726   NA NA
sex=2  779 2388 NA

$upper
        25 50 75
sex=1   NA NA NA
sex=2 2103 NA NA
```

Figure XX

(c) Log-rank test

From Figure XX (之前那个两个surviving curve的图) we inferred that there may be some difference for the surviving curves between males and female. In this part, we will conduct a hypothesis test to confirm if there is actually a significant difference between the two survival curves.

Since there are two populations (male and female) in this nonparametric setting, we can choose the weights in the test statistics as 1 for simplicity, which leads to the log-rank test for two samples. The result of the log-rank test in Figure XX suggests that we should reject the null hypothesis of the test and there is significant difference between the two survival curves.

```
Call:
survdiff(formula = Surv(days, status == 1) ~ sex, data = mln)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=1 126       28     37.1      2.25      6.47
sex=2  79       29     19.9      4.21      6.47

 Chisq= 6.5  on 1 degrees of freedom, p= 0.01
```

Figure XX

Similarly, we can investigate the other variable ulc in the same way. The test results in Figure XX suggests that the two survival curves for people who have ulceration present and absent are significantly different.

```
Call:
survdiff(formula = Surv(days, status == 1) ~ ulc, data = mln)

        N Observed Expected (O-E)^2/E (O-E)^2/V
ulc=1  90       41     21.2      18.5      29.6
ulc=2 115       16     35.8      10.9      29.6

 Chisq= 29.6  on 1 degrees of freedom, p= 5e-08
```

Figure XX

Part C - (a) Cox Proportional Hazard Model

(a1) Model Selection

As the model building and selection should be based on meaningful interpretation and biological knowledge, we will try out different models with interests and then use the AIC criteria to select the most appropriate model.

After trying out several models, we found 3 models with one or multiple variables that show significant variable effects as well as significant global tests.

Model 1: coxph(Surv(days,status==1)~sex, data=mln)

From Figure XX we can see that, the test for the variable sex is statistcally significant, which suggest the estimate is significantly different from 1. Meanwhile, the three global tests are also all significant.

```
Call:
coxph(formula = Surv(days, status == 1) ~ sex, data = mln)

  n= 205, number of events= 57

       coef exp(coef) se(coef)     z Pr(>|z|)
sex2 0.6622    1.9390   0.2651 2.498   0.0125 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
sex2     1.939     0.5157     1.153      3.26

Concordance= 0.59  (se = 0.034 )
Likelihood ratio test= 6.15  on 1 df,   p=0.01
Wald test            = 6.24  on 1 df,   p=0.01
Score (logrank) test = 6.47  on 1 df,   p=0.01
```

Figure XX

Model 2: coxph(Surv(days,status==1)~ulc, data=mln)

From Figure XX we can see that the test for the variable ulc as well as the three global tests are all significant.

```
Call:
coxph(formula = Surv(days, status == 1) ~ ulc, data = mln)

  n= 205, number of events= 57

        coef exp(coef) se(coef)      z Pr(>|z|)
ulc2 -1.4717    0.2295   0.2954 -4.982 6.29e-07 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
ulc2    0.2295      4.357    0.1286    0.4095

Concordance= 0.689  (se = 0.029 )
Likelihood ratio test= 28.44  on 1 df,   p=1e-07
Wald test            = 24.82  on 1 df,   p=6e-07
Score (logrank) test = 29.56  on 1 df,   p=5e-08
```

Figure XX

Model 3: coxph(Surv(days,status==1)~ulc+log(thick), data=mln)

In Model 3, the variable ulc and thick are used. Since the variable *thick* is right skewed, the log transformation is applied to give more reliable results. From Figure XX we can see that the estimates of both variables as well as the three global tests.

```
Call:
coxph(formula = Surv(days, status == 1) ~ ulc + log(thick), data = mln)

  n= 205, number of events= 57

              coef exp(coef) se(coef)      z Pr(>|z|)
ulc2       -0.9712    0.3786   0.3209 -3.027  0.00247 **
log(thick)  0.6104    1.8411   0.1759  3.470  0.00052 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

           exp(coef) exp(-coef) lower .95 upper .95
ulc2          0.3786     2.6412    0.2019    0.7101
log(thick)    1.8411     0.5431    1.3043    2.5990

Concordance= 0.762  (se = 0.032 )
Likelihood ratio test= 40.68  on 2 df,   p=1e-09
Wald test            = 34.62  on 2 df,   p=3e-08
Score (logrank) test = 39.47  on 2 df,   p=3e-09
```

Figure XX

AIC criteria

The model selection will be conducted by inspecting the AIC criteria of the three models. Table XX shows the corresponding AIC criteria for the three models, in which smaller AIC indicates better performance.

Model	AIC
1	562.2479
2	539.9615
3	529.7198

Table XX

(a2) Relative Risk

In model 1, the relative risk (hazard ratio) for sex is 1.939 and the confidence interval is [1.153, 3.26].
In model 2, the relative risk for ulc is 0.2295 and the confidence interval is [0.1286, 0.4095].
In model 3, the relative risk for ulc is 0.3786 and the confidence interval is [0.2019, 0.7101].
The relative risk for log(thick) is 1.8411 and the confidence interval is [1.3043, 2.5990].

The formula of calculating the estimate of the hazard ratio and its confidence interval are shown below.

The hazard ratio: $\exp(\widehat{\beta})$

The confidence interval of the hazard ratio: $(\exp(\widehat{\beta_L}), \exp(\widehat{\beta_U}))$

(a3) Model Assumption

In the proportional hazard model, the hazard ratio between two subjects with different covariates is assumed to be constant over time. To check this assumption, the formal test is conducted. From Figure XX we can see that in model 1, the test for the covariate and the global test are both not statistically significant at 0.05 significance level. Therefore, the constant proportional hazards assumption is satisfied. For model 2, the p-value of test for the

covariate and the global test are around the borderline. For model 3, the model with smallest AIC, the result suggests the relative risk for ulc is constant but not for log(thick), and the global test is also significant.

	chisq	df	p
sex	1.5	1	0.22
GLOBAL	1.5	1	0.22

	chisq	df	p
ulc	3.95	1	0.047
GLOBAL	3.95	1	0.047

	chisq	df	p
ulc	3.73	1	0.0533
log(thick)	6.76	1	0.0093
GLOBAL	7.70	2	0.0213

Figure XX

In summary, although model 3 has the smallest AIC, it does not fulfill the constant relative risk assumption. As a result, we will select model 1 as the final model.

Part C - (b) Parametric Survival Model

(b1) Appropriate parametric model

Model 1 is the model selected for the previous model and we will build an appropriate parametric model based on four different distributions. As we can see, the log-normal model has the smallest AIC among the three models, so we chose the log-normal model as the appropriate model.

model	log-normal	weibull	exponential	log-logistic
AIC	1126.706	1134.075	1132.662	1130.805

Table XX

(b2) Point and interval estimates of two parametric models

Figure XX shows the summary of AFT model representation.

```
Call:
survreg(formula = Surv(days, status == 1) ~ sex, data = mln,
    dist = "lognormal")
             Value Std. Error     z      p
(Intercept)  8.929      0.228 39.11 <2e-16
sex2        -0.697      0.264 -2.64 0.0082
Log(scale)   0.362      0.107  3.39 0.0007

Scale= 1.44

Log Normal distribution
Loglik(model)= -560.4   Loglik(intercept only)= -563.9
        Chisq= 7.15 on 1 degrees of freedom, p= 0.0075
Number of Newton-Raphson Iterations: 4
n= 205
```

Figure XX

Since only the categorical variable *sex* has been chosen,we calculate the point and interval estimates from the summary,the point estimate of *sex* is -0.697,the 95% C.I. for this variable can be easily calculated.

```
> CI
     theta        LCL        UCL
-0.6968039 -1.2142439 -0.1793639
```

Figure XX

From the result,we can see that the estimated confidence interval estimate of sex is (-1.214,-0.179).

The log-normal distribution is most easily characterized by saying the lifetime T is log-normally distributed if Y = log(T) is normally distributed with mean and variance specified by μ and σ,and we can deduce that the vector of regression coefficients of AFT model is negative to the vector of regression coefficients of linear model.

So we can use these equations $\theta = -\gamma$to get the estimated parameter γ in the linear model representation.

```
> CI2
     gamma       LCL       UCL
0.6968039 0.1793639 1.2142439
```

Figure XX

In the linear model representation,the point estimate of *sex* is 0.697,and the 95% C.I. is (0.179,1.214).

Appendix

```
####################################
# Part A - Descriptive Analysis
####################################
# load packages & import data
library(ggplot2)
library(survival)
library(ISwR)
data(melanom)
summary(melanom)

# status
table(melanom$status)
ggplot(melanom, aes(status)) +
geom_bar()+theme_minimal()+geom_text(stat='count', aes(label =..count..,
vjust = -0.2))
# status & days
ggplot(melanom, aes(x = no, y = days)) +
  geom_linerange(aes(ymin = 0, ymax = days)) +
  geom_point(aes(color = as.factor(status), shape = as.factor(status)))
+ coord_flip()+theme_minimal()
# ulc
table(melanom$ulc)
ggplot(melanom, aes(ulc)) + geom_bar()+geom_text(stat='count', aes(label
=..count.., vjust = -0.2))+theme_minimal()
# thick
ggplot(melanom, aes(x=thick)) +
geom_histogram(binwidth=100,color="black", fill="grey")+theme_minimal()
# sex
table(melanom$sex)
ggplot(melanom, aes(sex)) + geom_bar()+geom_text(stat='count', aes(label
=..count.., vjust = -0.2))+theme_minimal()

####################################
# Part B -  the response variable, the censoring indicator and the
categorical variable
####################################
# convert categorical variables to factor type
mln <- melanom
mln$ulc <- as.factor(mln$ulc)
mln$sex <- as.factor(mln$sex)
summary(mln)

# (a) For each of the levels of the categorical variable, compute the
```

```r
survival distribution. Plot them on the same graph. What do the graphs
suggest ?
# survival functions by sex
fit_sex <- survfit(Surv(days, status==1)~sex,
data=mln,type='kaplan-meier',conf.type='log-log')
fit_sex
summary(fit_sex)
plot(fit_sex,
     conf.int=TRUE,
     col=c('red','blue'),
     xlab='Time',
     ylab='Estimated survival')
legend(0,0.4, legend=c('Female', 'Male'),
       col=c('red','blue'), lty=1:1, cex=0.8)
abline(h=0.5,col='gray',lty=3)

# (b) For each level obtain an appropriate estimator and confidence
interval for the 3 quartiles of the survival curves. Interpret the
results.
# Q3 is the time associated with the first survival probability in the
table less than or equal to 0.25
# from the result, Q3 (75) is NA
quantile(fit_sex)

# (c) Conduct a single test of differences between the survival curves.
Justify your choice of test.
# log-rank test: test whether the two survival curves are identical
# p-value<0.05 reject H0, suggest significant difference
survdiff(Surv(days, status==1)~sex, data=mln)
survdiff(Surv(days, status==1)~ulc, data=mln)

#####################################
# Part C -  the response variable, the censoring indicator and the
categorical variable
#####################################
# (a) proportional hazard model / Cox model
# a.1 Build an appropriate model for these data, using the model
building procedures seen in class.
# There are three models that the significant tests for all variables as
well as global tests are significant
cox1 <- coxph(Surv(days,status==1)~sex, data=mln)
cox2 <- coxph(Surv(days,status==1)~ulc, data=mln)
cox3 <- coxph(Surv(days,status==1)~ulc+log(thick), data=mln)

# AIC
AIC(cox1)
```

```r
AIC(cox2)
AIC(cox3)

# a.2 Find an estimator and confidence interval for these relative risks
under this model.
summary(cox1)
summary(cox2)
summary(cox3)

# a.3 Is there evidence that these risks are indeed constant over time ?
cox.zph(cox1)
cox.zph(cox2)
cox.zph(cox3)

# (b) parametric regression models / AFT
# b.1 Build an appropriate parametric model for the data, taking (for
simplicity) the same variables as the ones selected for the previous
model.
logn=survreg(Surv(days,status==1)~sex, data=mln,dist="lognormal")
weib=survreg(Surv(days,status==1)~sex, data=mln, dist="weibull")
expon=survreg(Surv(days,status==1)~sex, data=mln,dist="exponential")
loglogist=survreg(Surv(days,status==1)~sex,data=mln,dist="loglogistic")
AIC=c(extractAIC(logn)[2], extractAIC(weib)[2], extractAIC(expon)[2],
extractAIC(loglogist)[2])
names(AIC)=c( "log(normal)", "weibull", "exponential", "log(logistic)")
AIC
summary(logn)
# b.2 Give point and interval estimates of the coefficient of the
categorical variable both in the AFT model and in the linear model
representation.
#AFT estimates
theta<-logn$coeff[2]
theta
stdev1<-0.264
CI<-c(theta,theta-1.96*stdev1,
      theta+1.96*stdev1)
names(CI) <- c("theta","LCL","UCL")
CI
#linear estimates
#The vector of regression coefficients of AFT model (theta) is negative
to the vector of regression coefficients of linear model (gamma).
gamma<-(-theta)
stdev2<-0.264
CI2<-c(gamma,gamma-1.96*stdev2,
       gamma+1.96*stdev2)
names(CI2)<-c("gamma","LCL","UCL")
```

```
CI2
```