STAT645 - Homework 8

Salih Kilicli

10/1/2019

```
rm(list=ls())
knitr::opts_chunk$set(echo = TRUE)
```

Problem 1: Consider the dataset on 44 subjects given in the article https://www.bmj.com/content/317/7156/468.1 (https://www.bmj.com/content/317/7156/468.1). Consider *prednisolone* or *no prednisolone* as the binary treatment variable, and use it as the explanatory variable and fit lognormal, exponential, and Weibull model to the data. Then choose the best model and justify your choice.

```
library(survival)
a=c(2,6,12,54,56,68,89,96,96,125,128,131,140,141,143,145,146,148,162,168,173,181,
    2,3,4,7,10,22,28,29,32,37,40,41,54,61,63,71,127,140,146,158,167,182)
b=c(0*1:4+1,0,0*1:4+1,0*1:5,1,0,1,0,0,1,0,0,0*1:16+1,0*1:6)
data1 = data.frame("Treatment"=rep(c(1,0), each=22), "Time"=a, "Delta"=b)
logn=survreg(Surv(Time, Delta)~Treatment, data=data1, dist="lognormal")
exp=survreg(Surv(Time, Delta)~Treatment, data=data1, dist="exponential")
weib=survreg(Surv(Time, Delta)~Treatment, data=data1, dist="weibull")
AIC = c(extractAIC(logn)[2],extractAIC(exp)[2],extractAIC(weib)[2])
names(AIC)=c("Lognormal", "Exponential", "Weibull")
AIC
```

```
##   Lognormal Exponential     Weibull
##    319.4332    320.2051    320.0339
```

We will choose **lognormal** model since it gives the lowest AIC value.

(a) Obtain the analytical expression of the 25th, 50th, and 75th percentile of the time-to-event of the best fitted model for the two groups.

p^{th} percentile of T is given by: $inf(t : \hat{S}(t) \leq (1-p))$. Since we know the analytical value of $S(t)$, it can be found by solving the equality

$$S(t) = 1 - \Phi\left(\frac{log(t) - \beta_0 - x_0^T\beta_1}{\sigma}\right) = (1-p)$$

where Φ is the **cdf** of $Normal(0,1)$ distribution. Then, solving above equation for t yields,

$$t = exp\left(\Phi^{-1}(p)\sigma + \beta_0 + x_0^T\beta_1\right).$$

Notice, in the above equation $x_0 = 1$ for Group 1 (Treatment=1) and $x_0 = 0$ for Group 2 (Treatment=0) and $\Phi^{-1}(p) = qnorm(p)$ in R. Finally, the values of $Q1$, $Q2$ and $Q3$ can be given by setting $p = 0.25, 0.5, 0.75$ in the equation above, respectively.

(b) Estimate the above three percentiles and obtain the 95% CI for the percentiles for the two groups separately.

Estimated values of $Q1$, $Q2$ and $Q3$ can be found by using $\hat{\beta_0}, \hat{\beta_1}, \hat{\sigma}$ estimates from lognormal model and setting $p = 0.25, 0.5, 0.75$ in the equation above, respectively.

```
library(msm)

logn  =survreg(Surv(Time, Delta)~Treatment, data=data1, dist="lognormal")
bhat_0=logn$coefficients[1]  # gives the estimate for intercept - \hat{beta}_0
bhat_1=logn$coefficients[2]  # gives the estimate for treatment - \hat{beta}_1
sigmahat=logn$scale          # gives the estimate for sigmahat - \hat{sigma}
p=c(0.25,0.5,0.75)
treatment=c(1,0)
qp=lower=upper=integer()

for(j in 1:2){
  x=treatment[j]
  for(i in 1:3){
    ep=sigmahat*qnorm(p[i])+bhat_0+x*bhat_1 #First group - Treatment == 1, Second - Treatment == 0
    qp[i]=exp(ep)
    sestar= deltamethod(~(log(exp(ep))-x1-x*x2)/exp(x3),c(bhat_0,bhat_1,log(sigmahat)),logn$var) #s
  e for quartiles
    lower[i]=qp[i]-1.96*sestar
    upper[i]=qp[i]+1.96*sestar
  }
  cat("Confidence intervals for Group",j,"\n")
  cat(c("                 ","Q1" ,"      ", "Q2","     ", "Q3"),"\n")
  cat("Lower bounds: ", lower,"\n")
  cat("Estimates   : ", qp,"\n")
  cat("Upper bounds: ", upper,"\n \n")
}
```

```
## Confidence intervals for Group 1
##                    Q1        Q2       Q3
## Lower bounds:  49.15246 163.8267 543.5267
## Estimates   :  49.62121 164.3135 544.1004
## Upper bounds:  50.08996 164.8003 544.6741
##
## Confidence intervals for Group 2
##                    Q1        Q2       Q3
## Lower bounds:  13.77148 46.66035 155.4534
## Estimates   :  14.22303 47.09753 155.9567
## Upper bounds:  14.67458 47.5347 156.4599
##
```

(c) Using a nonparametric method obtain the estimate and 95% CI for the 25th, 50th, and 75th percentiles of the time-to-event for the two groups separately. Compare and comment on the differences between these nonparametric estimates and the parametric estimates obtained in step (b).

```
t = data1$Time[1:22]; delta1=data1$Delta[1:22]   #or you can use data1$Time[data1$Treatment==1]
c = data1$Time[23:44]; delta2=data1$Delta[23:44] #or you can use data1$Time[data1$Treatment==0]
one = survfit(Surv(t, delta1)~1)
two = survfit(Surv(c, delta2)~1)
quantile(one, prob=c(0.25,0.5,0.75), conf.int=TRUE)
```

```
## $quantile
##  25  50  75
##  89 146  NA
##
## $lower
##  25  50  75
##  12  96 168
##
## $upper
## 25 50 75
## NA NA NA
```

```
quantile(two, prob=c(0.25,0.5,0.75), conf.int=TRUE)
```

```
## $quantile
##   25   50   75
## 22.0 40.5   NA
##
## $lower
## 25 50 75
##  4 29 54
##
## $upper
## 25 50 75
## 41 NA NA
```

I believe the non-parametric method yields infintely big time-to-event values so that we don't have an upper bound for some of the confidence intervals. You can also see it is an increasing function of time since $Q1 < Q2 < Q3$ in each case. However, the parametric method yields very narrow confidence intervals which means delta method gives small standard error values for the problem.

Problem 2: Consider the colon data available in the survival package of R. Consider the subset where etype = 1 only (exclude the subjects who experienced death). You may find a descent description of the data at https://stat.ethz.ch/R-manual/R-devel/library/survival/html/colon.html (https://stat.ethz.ch/R-manual/R-devel/library/survival/html/colon.html). Consider the following seven explanatory variables, sex, age, perfor, adhere, nodes, differ, extent. Make sure to treat differ and extent as factor variables.

(a) Build a Weibull model with the above explanatory variables and their two factor interactions. Then choose the best subset of the explanatory variables using the stepwise regression method.

```
library(MuMIn)
library(survival)

col= colon[colon$etype==1,]         #pick the rows with etype==1
data2a=col[complete.cases(col),]    #get rid of NA values, complete the cases

data2a$differ=as.factor(data2a$differ)
data2a$extent=as.factor(data2a$extent)

model2a=survreg(formula = Surv(time, status) ~ sex + age + perfor + adhere + nodes + differ + exten
  t
              + sex*age + sex*perfor + sex*adhere+ sex*nodes +  sex*differ + sex*extent
              + age*perfor + age*adhere +age*nodes + age*differ + age*extent
              + perfor*adhere + perfor*nodes + perfor*differ + perfor*extent
              + adhere*nodes + adhere*differ + adhere*extent
              + nodes*differ + nodes*extent
              + differ*extent, data=data2a, dist="weibull")

best = step(model2a, trace=0)
summary(best)
```

```
##
## Call:
## survreg(formula = Surv(time, status) ~ sex + age + perfor + adhere +
##     nodes + differ + extent + sex:age + sex:perfor + sex:nodes +
##     sex:extent + age:perfor + age:adhere + age:differ + perfor:nodes +
##     adhere:nodes + nodes:extent, data = data2a, dist = "weibull")
##                  Value Std. Error     z       p
## (Intercept)   12.0332     1.8798  6.40 1.5e-10
## sex            2.4207     1.7074  1.42 0.15625
## age           -0.0211     0.0176 -1.20 0.23079
## perfor         3.5815     2.4754  1.45 0.14795
## adhere        -2.2303     0.8713 -2.56 0.01047
## nodes         -0.6547     0.6126 -1.07 0.28515
## differ2       -2.2164     1.1558 -1.92 0.05516
## differ3       -4.4880     1.3013 -3.45 0.00056
## extent2       -1.5397     1.5587 -0.99 0.32324
## extent3       -1.9848     1.5296 -1.30 0.19441
## extent4       -2.0683     1.6462 -1.26 0.20897
## sex:age       -0.0300     0.0111 -2.71 0.00671
## sex:perfor     1.2641     0.7052  1.79 0.07306
## sex:nodes      0.0955     0.0284  3.36 0.00077
## sex:extent2    0.5332     1.6572  0.32 0.74765
## sex:extent3   -0.9545     1.5882 -0.60 0.54785
## sex:extent4   -1.7347     1.7042 -1.02 0.30874
## age:perfor    -0.0570     0.0347 -1.64 0.10007
## age:adhere     0.0275     0.0144  1.91 0.05612
## age:differ2    0.0370     0.0183  2.03 0.04281
## age:differ3    0.0687     0.0210  3.27 0.00107
## perfor:nodes  -0.2933     0.1326 -2.21 0.02694
## adhere:nodes   0.0613     0.0419  1.46 0.14350
## nodes:extent2  0.3518     0.6139  0.57 0.56656
## nodes:extent3  0.4790     0.6127  0.78 0.43430
## nodes:extent4  0.4778     0.6156  0.78 0.43769
## Log(scale)     0.2971     0.0414  7.17 7.4e-13
##
## Scale= 1.35
##
## Weibull distribution
## Loglik(model)= -3859.1   Loglik(intercept only)= -3938.1
##  Chisq= 158.1 on 25 degrees of freedom, p= 2.7e-21
## Number of Newton-Raphson Iterations: 5
## n= 888
```

Step function chooses a model with the lowest AIC in a stepwise algorithm, and the best model found to be:

$$Surv(time, status) = sex + age + perfor + adhere + nodes + differ + extent + sex : age + sex : perfor + sex : nodes + sex : extent + age : perfor + age : adhere + age : differ + perfor : nodes + adhere : nodes + nodes : extent$$

(b) For the best chosen model, obtain the estimate and 95% CI for the survival probability at time 365, 730, 1095, 1460, 1825 days and for the following set of covariates. Discuss the results.

```
zeros = 0*(1:8)
data2b=data.frame(
  sex = c(rep(1,4),rep(0,4)), age = zeros + 60 , perfor = rep(c(0,0,1,1),2),
  adhere = rep(c(0,1), 4), nodes = zeros + 2, differ = as.factor(zeros + 2), extent = as.factor(zer
 os + 3)
  )
names(data2b)=c("sex","age","perfor","adhere","nodes","differ","extent")
data2b
```

sex	age	perfor	adhere	nodes	differ	extent
1	60	0	0	2	2	3
1	60	0	1	2	2	3
1	60	1	0	2	2	3
1	60	1	1	2	2	3
0	60	0	0	2	2	3
0	60	0	1	2	2	3
0	60	1	0	2	2	3
0	60	1	1	2	2	3

```
time=365*(1:5)

for(i in 1:5){

  tm=time[i] ; lb=ub=si=integer()

  for(j in 1:nrow(data2b)){

  d=data2b[j,]; a1=d$sex; a2=d$age; a3=d$perfor; a4=d$adhere; a5=d$nodes; a6=1; a7=1;
  avec=c(1,a1,a2,a3,a4,a5,a6,0,0,a7,0,a1*a2,a1*a3,a1*a5,0,a1*a7,0,a2*a3, a2*a4,a2*a6,0,a3*a5,a4*a5,
  0,a5*a7,0);
  pred=sum(avec*as.vector(best$coefficients));

  estm=(tm*exp(-pred))^(1/best$scale)
  est=exp(-estm)
  si=c(si,est)

  sestar= deltamethod(~(tm*exp(-x1-a1*x2-a2*x3-a3*x4-a4*x5-a5*x6-a6*x7-0*x8-0*x9-a7*x10-
                              0*x11-(a1*a2)*x12-(a1*a3)*x13-(a1*a5)*x14-0*x15-(a1*a7)*x16-
                              0*x17-(a2*a3)*x18-(a2*a4)*x19-(a2*a6)*x20-0*x21-(a3*a5)*x22-
                              (a4*a5)*x23-0*x24-(a5*a7)*x25-0*x26))^(1/exp(x27)),
                        c(as.vector(best$coefficients), log(best$scale)),best$var)
  lb=c(lb,exp(-estm-1.96*sestar))
  ub=c(ub,min(1,exp(-estm+1.96*sestar)))
  }
CI=data.frame("time"=tm,"new_data"=(1:nrow(data2b)),"lower_bound"=lb,"estimate"=si,"upper_bound"=u
  b)
print(CI)
}
```

```
##   time new_data lower_bound  estimate upper_bound
## 1  365        1   0.8166418 0.8440539   0.8723862
## 2  365        2   0.7315171 0.7878112   0.8484374
## 3  365        3   0.8441729 0.9129425   0.9873143
## 4  365        4   0.7881268 0.8797398   0.9820020
## 5  365        5   0.8315179 0.8588060   0.8869896
## 6  365        6   0.7537759 0.8072493   0.8645161
## 7  365        7   0.6822331 0.8112670   0.9647057
## 8  365        8   0.5868709 0.7451047   0.9460019
##   time new_data lower_bound  estimate upper_bound
## 1  730        1   0.7159255 0.7529627   0.7919160
## 2  730        2   0.5947286 0.6708919   0.7568090
## 3  730        3   0.7537349 0.8586146   0.9780880
## 4  730        4   0.6720184 0.8069963   0.9690851
## 5  730        5   0.7373006 0.7751167   0.8148723
## 6  730        6   0.6250829 0.6988251   0.7812667
## 7  730        7   0.5279530 0.7046557   0.9404998
## 8  730        8   0.4104656 0.6111438   0.9099344
##   time new_data lower_bound  estimate upper_bound
## 1 1095        1   0.6380258 0.6814831   0.7279003
## 2 1095        2   0.4961387 0.5830623   0.6852150
## 3 1095        3   0.6827219 0.8138157   0.9700817
## 4 1095        4   0.5846920 0.7483998   0.9579441
## 5 1095        5   0.6636641 0.7087218   0.7568385
## 6 1095        6   0.5306000 0.6161105   0.7154016
## 7 1095        7   0.4219798 0.6230682   0.9199825
## 8 1095        8   0.3002961 0.5140028   0.8797946
##   time new_data lower_bound  estimate upper_bound
## 1 1460        1   0.5738313 0.6219688   0.6741446
## 2 1460        2   0.4200124 0.5127274   0.6259087
## 3 1460        3   0.6235379 0.7748276   0.9628248
## 4 1460        4   0.5146610 0.6984585   0.9478943
## 5 1460        5   0.6024360 0.6528978   0.7075866
## 6 1460        6   0.4564292 0.5489538   0.6602344
## 7 1460        7   0.3436191 0.5566407   0.9017219
## 8 1460        8   0.2254683 0.4386222   0.8532882
##   time new_data lower_bound  estimate upper_bound
## 1 1825        1   0.5192866 0.5709278   0.6277046
## 2 1825        2   0.3591448 0.4545415   0.5752777
## 3 1825        3   0.5727283 0.7399914   0.9561031
## 4 1825        4   0.4566471 0.6546907   0.9386239
## 5 1825        5   0.5499769 0.6045863   0.6646181
## 6 1825        6   0.3962058 0.4926848   0.6126570
## 7 1825        7   0.2834112 0.5008380   0.8850698
## 8 1825        8   0.1723278 0.3780531   0.8293736
```

As time increases (1 year, 2 years, …, 5 years) the survival probability estimates are decreasing and the confidence intervals are getting narrower however se values are increasing, simultaneously. Now, fixing a time period (year 1), and comparing affects of perfor and adhere on Males and Females we see that:

1. Males with perforation of colon and no adherence to nearby organs have the highest survival probability, whereas males without perforation of colon and adherence to nearby organs have the lowest survival probability.
2. Females without perforation of colon and no adherence to nearby organs have the highest survival probability, whereas males with perforation of colon and adherence to nearby organs have the lowest survival probability.

In addition, looking at the first 4 and last for rows of each year we see that gender has no clear effect. If we compare row 1, row 3, row 7 (perfor effect) we can see that perforation of colon in males increases the surival probability while it decreases in females. Similary, if we compare row 1, row 2, and row 6 (adhere affect) we see that adherence to nearby organs clearly decreases survival probability for each gender.

(c) Consider the Weibull model with age, sex, treatment, and nodes and their two factor interactions as the explanatory variables. Then conduct a likelihood ratio test using the anova function if age has a statistically significant effect on the model. Full points will be given only for properly writing the hypotheses, test statistics, p-value, and conclusions.

```
library(survival)
col= colon[colon$etype==1,]        #pick the rows with etype==1
data2c=col[complete.cases(col),]   #get rid of NA values, complete the cases
treatment = as.factor(data2c$rx)
model2c1=survreg(formula = Surv(time, status) ~ age + sex + treatment + nodes
                 + age:sex + age:treatment + age:nodes
                 + sex:treatment + sex:nodes
                 + treatment:nodes, data = data2c, dist = "weibull")

model2c2=survreg(formula = Surv(time, status) ~ sex + treatment + nodes
                 + sex:treatment + sex:nodes
                 + treatment:nodes, data = data2c, dist = "weibull")
anova(model2c2, model2c1)
```

Terms	Resid. Df	-2*LL	Test	Df	Deviance	Pr(>Chi)
sex + treatment + nodes + sex:treatment + sex:nodes + treatment:nodes	877	7764.350		NA	NA	NA
age + sex + treatment + nodes + age:sex + age:treatment + age:nodes + sex:treatment + sex:nodes + treatment:nodes	872	7751.877	=	5	12.47309	0.0288499

```
# To make sure my p-value is correct, I have calculated it by hand as well

LRtest = as.numeric(-2*(logLik(model2c2)-logLik(model2c1)))
print(LRtest)
```

```
## [1] 12.47309
```

```
p = 1 - pchisq(12.47309, 5)
print(p)
```

```
## [1] 0.02884999
```

The hypothesis for the problem can be written as below:

$$H_0 : Age = 0 \text{ (Age has no effect on the time to recurrence),}$$

$$H_a : Age \neq 0 \text{ (Age has a statistically significant effect on the time to recurrence)}$$

.

The test statistic is 12.47309 with a corresponding $p - value = 0.02884999$. Therefore, we reject H_0 at the 5% level, and conclude that age has statistically significant effect on the time to recurrence.