960-542– Life Data Analysis– Spring, 2020

Homework 5 Solutions,

1.
The data set at `http://lib.stat.cmu.edu/datasets/csb/ch15.dat` from contains results compiled by the Cooperative Early Lung Cancer Detection Program. A full description of the data set can be found at `http://lib.stat.cmu.edu/datasets/csb/ch15.txt`, and SAS code to read the data can be found at `http://lib.stat.cmu.edu/datasets/csb/ch15.sas`. Focus attention on three variables: cancer cell type (variable 5), stage (variable 6), survival time in days (variable 11), and status (variable 12). Status is coded as 0 if alive, 1 if dead from lung cancer, and 2 if dead from other causes. Consider individuals with status either 0 or 2 as censored.

a. Observations numbered 608 and 922 have a trait which violates the assumptions under which we developed the accelerated life model. What is this trait? Delete these observations and all other observations which share this violation.
These observations (and others) have a survival time of zero.

b. Fit an accelerated life model to these data, using a generalized gamma error distribution, and indicator functions for the various cell types and stages. Do either the log normal or Weibull distributions appear to fit these data? Does any simpler model fit these data?
R commands are

```
fle<-as.data.frame(scan("ch15.dat",what=list(number=0,inst=0,gr=0,
  det=0,ct=0,st=0,T=0,N=0,M=0,op=0,surv=0,status=0),na.strings="."))
fle$delta<-(fle$status==1)*1
library(survival)
library(flexsurv)
cat("Generalized Gamma Fit")
ggamfit<-flexsurvreg(Surv(surv,delta)~as.factor(ct)+as.factor(st),
  data=fle, dist="gengamma.orig",subset=surv>0)
#Don't print the summary here.  flexsurvreg summary is quite long.
print(ggamfit$res.t)
```

The important part of the output from the first run is: The `flexsurvreg` *parameterization of the generalized gamma on the original scale has* $k = 1$ *giving the Weibull, and gives the confidence interval on the log scale, and so the Weibull fits if the given value is close to zero. Here's the important line:*

```
              est     L95%     U95%      se
k          0.201834068 -0.3106770  0.71434515 0.2614900
```

The closeness of the shape estimate to 1 (or 0 on the log scale) makes the Weibull model seem reasonable, and the test of shape=0, corresponding to the log normal model, is clearly rejected. Running in R gives

```
cat("Weibull Fit")
wfit<-summary(survreg(Surv(surv,delta)~as.factor(ct)+as.factor(st),
  data=fle, dist="weibull",subset=surv>0))
print(wfit)
```

gives output

```
            Value Std. Error    z     p
(Intercept)    8.6096    0.1172  73.44 < 2e-16
as.factor(ct)1 -0.1860    0.1072  -1.73 0.08286
as.factor(ct)2 -0.4234    0.1224  -3.46 0.00054
as.factor(ct)3 -0.5520    0.1149  -4.81 1.5e-06
as.factor(ct)4 -0.0110    0.3473  -0.03 0.97464
as.factor(st)2 -0.9108    0.1544  -5.90 3.6e-09
as.factor(st)3 -1.4362    0.1141 -12.59 < 2e-16
Log(scale)     0.0132    0.0315   0.42 0.67530
Scale= 1.01
```

The hypothesis that the extreme value scale parameter is 1 is not rejected. Probably more importantly, the scale parameter is close enough to one to make the exponential adequate.

c. Interpret the parameter estimates for stage from part (b). Comment on both the direction and magnitude of the effects. Use the simplest error distribution that appears to fit reasonably well.

Fit the exponential model in R using

```
cat("Exponential Fit")
expfit<-survreg(Surv(surv,delta)~as.factor(ct) + as.factor(st),
   data = fle, subset = surv > 0, dist = "exponential")
print(expfit)
```

The output from survreg is

```
            Value Std. Error    z     p
(Intercept)    8.6096    0.1172  73.44 < 2e-16
as.factor(ct)1 -0.1860    0.1072  -1.73 0.08286
as.factor(ct)2 -0.4234    0.1224  -3.46 0.00054
as.factor(ct)3 -0.5520    0.1149  -4.81 1.5e-06
as.factor(ct)4 -0.0110    0.3473  -0.03 0.97464
as.factor(st)2 -0.9108    0.1544  -5.90 3.6e-09
as.factor(st)3 -1.4362    0.1141 -12.59 < 2e-16
Log(scale)     0.0132    0.0315   0.42 0.67530
Scale= 1.01
```

Hence the individuals in stage 1 have a time scale that is approximately $\exp(1.43)$ *times slower than those in stage 3, and the individuals in stage 2 have a time scale that is approximately* $\exp(0.52)$ *times slower than those in stage 3.*

d. Give confidence intervals for the median survival times for the different cell type groups.

Fit the exponential model with just cell type:

```
smexpfit<-survreg(formula = Surv(surv, delta) ~ as.factor(ct),
   data = fle, subset = surv > 0, dist = "exponential")
newdata<-data.frame(ct=factor(0:4,levels=0:4))
out<-predict(smexpfit,type="quantile",p=0.5,newdata=newdata,se=TRUE)
cimat<-cbind(out$fit,out$fit)+outer(out$se.fit,c(-1,1))*1.96
dimnames(cimat)<-list(0:4,c("Lower","Upper"))
print(cimat)
```

to give

```
    Lower    Upper
0 1475.2814 2016.3089
1 1217.0431 1603.0989
2  691.9338  984.6902
3  519.5328  698.4670
4  495.9607 2365.3509
```

2. Greenberg and White (1963) obtained data on the length of time between successive births of children to certain families.Cox and Snell (1981) report that these times appear to be lognormally distributed; that is, conditional on an additional birth being observed, the logs of the months between successive live births were found to be approximately normally distributed.Means of times between births under various scenarios were reported, but no standard deviations were reported. Assume that the mean of the normal distribution is 3.5, with a standard deviation of 2 units.

a. Calculate the probability of more than 4 years passing between successive births.

Let X represent the interval between births, in years. Then

$$\mathrm{P}\left[X \geq 4\right] = \mathrm{P}\left[(\log(X) - 3.5)/2 \geq (\log(48) - 3.5)/2\right] = .43.$$

b. Calculate the hazard rate of a birth three years after the first.

The density and and survival function are

$$\phi((\log(36) - 3.5)/2)/(2 \times 36) \text{ and } \Phi(-(\log(36) - 3.5)/2))$$

respectively, and the hazard is 0.0107.

c. One might examine intervals between subsequent births on the log scale, since intervals between births can take any positive number, and so potential values of the log are the same as the support of the normal distribution. Assess this rational.

Except for children who are born of the same pregnancy, the interval between births can not be smaller than the minimal pregnancy length; hence there is an interval in the support of the log normal distribution during which no observations can occur. Hence the rational is flawed.

3.

The Center for Analysis and Management of Multicenter AIDS Cohort Study (2002) reported on a the health history of a cohort of individuals. A subset of these individuals were followed up for as many as 31 visits, approximately six months apart. All individuals were HIV-positive at the beginning of the study. The visit at which they were last found to

be without AIDS (or in one case, the visit at which an individual was last followed) was recorded, along with an indicator of educational level. Individuals who did not report an educational level, or who missed visits during the interval in which they developed AIDS, were omitted. Furthermore, some individuals had additional screenings between scheduled screenings; results from these visits were ignored. Data may be found at `http://stat.rutgers.edu/home/kolassa/960-542/seroconvert.dat` The entries are educational level (1= 8th grade or less, 2= 9,10,11th grade, 3= 12th grade, 4= At least one year college but no degree, 5= Four years college/got a degree, 6= Some graduate work, 7= Post-graduate degree) , last visit before seroconversion or AIDS, and status indicator (1 for lost to followup prior to AIDS, 4 for seroconversion, 6 for AIDS prevalence; treat 1 as censored and the others as having the event), and the count of individuals with this pattern. Treat these individuals as though the times to AIDS have a continuous distribution with the hazards for the various educational groups proportional, and fit a model measuring the effect of education. Carefully account for interval censoring.

In R do

```
sero<-as.data.frame(
  scan('seroconvert.dat',what=list(ed=0,last=0, status=0, count=0)))
sero$event<-1
sero$event[sero$status==1] <-0
sero<-sero[ rep(seq(length(sero$count)),sero$count),]
library(icenReg)
test<-sero
test$begin<-test$last
test$end<-ifelse(test$status!=1,test$last+1,10000)
print(ic_sp(Surv(begin,end,type="interval2")~as.factor(ed), bs_samples = 9999,data=test))
```

to obtain

```
                Estimate Exp(Est) Std.Error   z-value    p
as.factor(ed)2 -0.62890  0.5332    280.9 -0.0022390 0.9982
as.factor(ed)3 -0.16140  0.8509    280.8 -0.0005748 0.9995
as.factor(ed)4 -0.09204  0.9121    280.8 -0.0003278 0.9997
as.factor(ed)5 -0.20560  0.8141    280.8 -0.0007322 0.9994
as.factor(ed)6 -0.05411  0.9473    280.8 -0.0001927 0.9998
as.factor(ed)7  0.21190  1.2360    280.8  0.0007546 0.9994
```