Oscar Huang
April 26, 2017

Statistics 516

Homework 04

Generalized Nonlinear Regression Models and Parametric Survival Models

Student: Oscar Huang

Date: 2017/04/26

Oscar Huang
April 26, 2017

Toxicity Study With Two Strains of Water Fleas

1.

a. Code and output:

```
options(digits=4)
m.nls <- nls(count ~ exp(b0 + b1 * concentration + b2 * (strain=="b")),
        start = c(b0 = 4, b1 = 0, b2=0), data = flea)

output_Nls<-summary(m.nls)$coefficients[,1]
output_Glm<-summary(m.glm)$coefficients[,1]
iteration=0

while(sum(round(output_Glm,digits =5)) != sum(round(output_Nls,digits =5))){
  flea$w <- 1/predict(m.nls)
  m.nls <- nls(count ~ exp(b0 + b1 * concentration + b2 * (strain=="b")),weights = w,
              start = c(b0 = 4, b1 = 0, b2=0), data = flea)
  output_Nls<-summary(m.nls)$coefficients[,1]
  iteration=iteration+1
    if(iteration > 1000){
    print("iteration more than 1,000 times")
    break
  }}

if(sum(round(output_Glm,digits =5)) == sum(round(output_Nls,digits =5)))
  {paste("The algorithms is converged in", iteration, "iterations" )}
```

```
[1] "The algorithms is converged in 4 iterations"
```

```
summary(m.nls)$coefficients
```

```
   Estimate Std. Error t value  Pr(>|t|)
b0    4.455    0.04272 104.273 7.167e-76
b1   -1.543    0.05087 -30.334 7.310e-41
b2   -0.275    0.05280  -5.208 1.988e-06
```

```
summary(m.glm)$coefficients
```

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)     4.455    0.04272 104.272 7.167e-76
concentration  -1.543    0.05087 -30.334 7.309e-41
strainb        -0.275    0.05280  -5.208 1.988e-06
```

b. Discussion:

The iteratively weighted least squares algorithm was converged in 4 iterations while the starting value of the parameters were: b0 = 4, b1 = 0, b2=0. The estimated parameters were: b0 = 4.455, b1 = -1.543, b2=-0.275 from the weighted least squares algorithm, which agreed with the parameters from the glm function.

2.

Oscar Huang
April 26, 2017

a. Code and output:

```
m.nls2 <- nls(count ~ exp(b0 + b1 * concentration ^ b3 + b2 * (strain=="b")),
              start = c(b0 = 4.5, b1 = -1.5, b2=0, b3 = 1), data = flea)
summary(m.nls2)$coefficients'
```

```
  Estimate Std. Error t value  Pr(>|t|)
b0   4.4846    0.02958 151.602 1.171e-85
b1  -1.5628    0.06284 -24.868 3.096e-35
b2  -0.3270    0.04533  -7.212 6.785e-10
b3   0.9614    0.08740  11.001 1.427e-16
```

b. Discussion:

The parameters estimated were: b0 = 4.4846, b1 = -1.5628, b2=-0.3270, b3 = 0.9614, while $E(Y_i)=\exp(\beta_0+\beta_1 c_i^{\beta_3}+\beta_2 d_i)$

3.

a. Code and output:

```
d<-expand.grid(strain=c("a","b"), concentration=seq(0,2,0.01))
d$predNls<-predict(m.nls, d, interval="confidence")
d$predNls2<-predict(m.nls2, d, interval="confidence")

p<-ggplot(flea, aes(x= concentration, y=count,color=strain))
p<-p+geom_point()
p<-p + geom_line(aes(y=predNls), data=d)
p<-p + geom_line(aes(y=predNls2), data=d, linetype=2)
p<-p+labs(x="Concentraion of chemical fountd in jet fuel (%)", y="Number of water fleas")
plot(p)
```

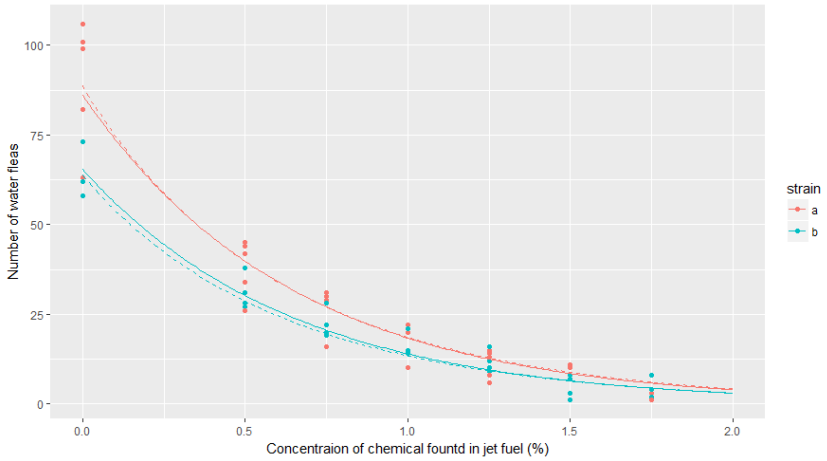

Figure 1 The number of water fleas with different concentration of chemical and different strain. The solid lines indicate the mean structure with no power transformation.

b. Discussion:

The model with and without the power transformation have no big difference since the β_3, the parameter for the power transformation, was 0.9614, which is close to the mean structure without power transformation ($\beta_3 = 1$).

Oscar Huang
April 26, 2017

Embryonic Duration of the Common Fruit Fly

1.

a. Code and output:

```
delta = 35
step =10
fly$x2<-1/(fly$temp-delta)
m<-glm(duration ~ temp + x2, data =fly,family=Gamma(link=log),weight =batch)
d<-m$deviance

while(abs(step)>0.0001){
    delta = delta + step
    fly$x2<-1/(fly$temp-delta)
    m<-glm(duration ~ temp + x2, data =fly,family=Gamma(link=log), weight = batch)
    if(m$deviance>d){
            step = step * -0.2}
    d<-m$deviance}

paste("delta:", delta)
```

```
[1] "delta: 58.638016"
```

```
summary(m)
```

```
Call:
glm(formula = duration ~ temp + x2, family = Gamma(link = log),
    data = fly, weights = batch)

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-0.23769  -0.09367  -0.01389   0.07567   0.21225

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.202e+00  4.392e-02   72.91   <2e-16 ***
temp        -2.648e-01  3.659e-03  -72.36   <2e-16 ***
x2          -2.170e+02  4.391e+00  -49.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.01591675)

    Null deviance: 617.86435  on 22  degrees of freedom
Residual deviance:   0.31823  on 20  degrees of freedom
AIC: 418.98

Number of Fisher Scoring iterations: 3
```

b. Discussion:

The estimated δ value that minimizes the residual deviance (0.318) was 58.638 .

Oscar Huang
April 26, 2017

2.

a. Code and output:

```
m.nls<-nls(duration~exp(b0 + b1 * temp + b2/(temp-58.638)),data=fly,
          start = c(b0=3.2, b1=-0.2, b2=-200))
fly$w <- fly$batch/predict(m.nls)^2

for(i in 1:10){
  m.nls<-nls(duration~exp(b0 + b1 * temp + b2/(temp-58.638)),weight = w,
data=fly,
           start = c(b0=3.2, b1=-0.2, b2=-200))
  fly$w <- fly$batch/predict(m.nls)^2
}
summary(m.nls)
```

```
Formula: duration ~ exp(b0 + b1 * temp + b2/(temp - 58.638))

Parameters:
   Estimate Std. Error t value Pr(>|t|)
b0  3.20e+00   4.39e-02    72.9   <2e-16 ***
b1 -2.65e-01   3.66e-03   -72.4   <2e-16 ***
b2 -2.17e+02   4.39e+00   -49.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.126 on 20 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 7.67e-06
```

b. Discussion:

While using δ=0.638, and starting vale: $b_0 = 3.2$, $b_1 = -0.2$, and $b_2 = -200$ and after 10 times iterations, the estimated parameters were $b_0 = 3.20$, $b_1 = -0.265$, and $b_2 = -217$, which agreed with the parameters estimated from last question.

3.

a. Code and output:

```
m.nls2 <- nls(duration~exp(b0 + b1 * temp + b2/(temp-delta)), data=fly,
        start = list(b0=3.2, b1=-0.2, b2=-216,delta=58.638))
fly$w <- fly$batch/predict(m.nls2)^2

for (i in 1:10) {
  m.nls2 <- nls(duration~exp(b0 + b1 * temp + b2/(temp-delta)), weight = w,
data=fly,
             start = list(b0=3.2, b1=-0.2, b2=-216,delta=58.638))
  fly$w <- fly$batch/predict(m.nls2)^2
}
summary(m.nls2)$coefficients
```

```
Formula: duration ~ exp(b0 + b1 * temp + b2/(temp - delta))
```

Oscar Huang
April 26, 2017

```
Parameters:
       Estimate Std. Error t value Pr(>|t|)
b0      3.20213    1.59485   2.008   0.0591 .
b1     -0.26480    0.03552  -7.454 4.71e-07 ***
b2   -216.97118  125.23342  -1.733   0.0994 .
delta  58.63835    6.48396   9.044 2.59e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1294 on 19 degrees of freedom

Number of iterations to convergence: 8
Achieved convergence tolerance: 5.99e-06
```

b. Discussion:

After 10 iteration with weighted least squares approach, the estimated parameters were: $b_0 = 3.202$, $b_1 = -0.265$, $b_2 = -216.971$, and $\delta = 58.64$.

4.

a. Code and output:

```
mydata<-expand.grid(temp=seq(15,33,0.1))
mydata$predict<-predict(m.nls2, mydata, interval="confidence")
p<-ggplot(fly,aes(x=temp,y=duration))
p<-p+geom_point(aes(size=batch),pch = 21) + theme_bw()
p<- p+scale_size(breaks = seq(25,250, by = 25))
p<- p+labs(x="Temperature (C)", size = "Egg Batch Size", y = "Mean Ebryonic Duration (hours)")
p<-p+geom_line(aes(y=predict), data = mydata)
plot(p)
```

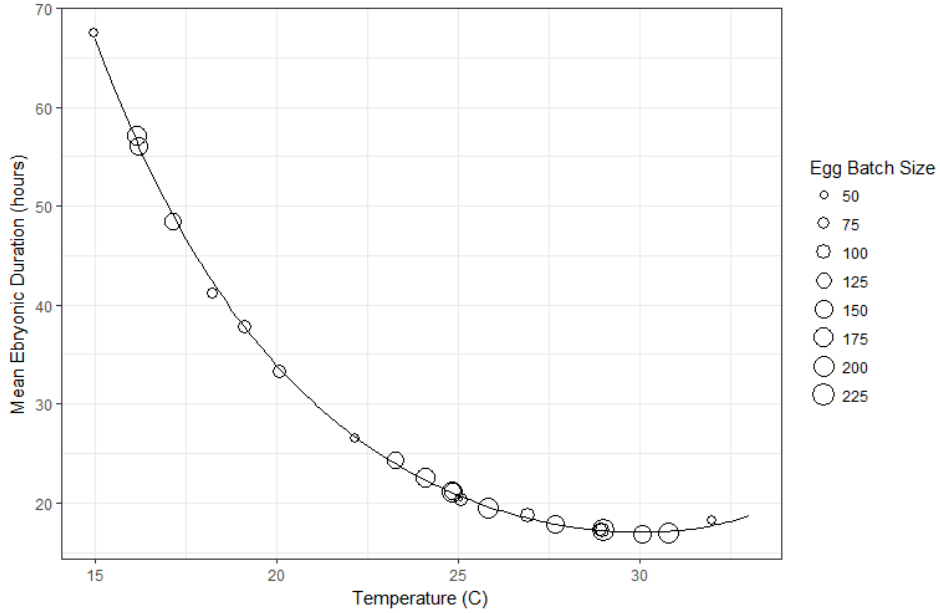

Figure 2 The observed (circles) and predicted (curve) mean embryonic duration with different

temperature.

Oscar Huang
April 26, 2017

Survival of Male Fruit Flies

1.

a. Code and output:

```
options(digits=4)
m.gamma<-glm(longevity ~ activity + thorax ,data = fruitfly, family = Gamma(link = log))
summary(m.gamma)$coefficients
```

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)   1.88722    0.19405   9.726 8.974e-17
activityone   0.05527    0.05337   1.036 3.024e-01
activitylow  -0.11646    0.05332  -2.184 3.091e-02
activitymany  0.08250    0.05413   1.524 1.302e-01
activityhigh -0.41466    0.05394  -7.687 4.935e-12
thorax        2.68778    0.22769  11.804 1.037e-21
```

```
m.flexSurv <- flexsurvreg(Surv(longevity) ~ activity + thorax, dist =
custom.gamma,
                  data = fruitfly)
m.flexSurv
```

```
Call:
flexsurvreg(formula = Surv(longevity) ~ activity + thorax, data = fruitfly,
   dist = custom.gamma)

Estimates:
             data mean  est      L95%     U95%     se      exp(est)  L95%     U95%
shape             NA    28.9025  37.0183  22.5660      NA      NA       NA      NA
rate              NA     4.3795   6.9007   2.7795      NA      NA       NA      NA
activityone    0.2016     0.0553  -0.0480   0.1585   0.0527   1.0568    0.9532   1.1718
activitylow    0.2016    -0.1165  -0.2197  -0.0134   0.0526   0.8900    0.8028   0.9867
activitymany   0.1935     0.0824  -0.0223   0.1871   0.0534   1.0859    0.9779   1.2057
activityhigh   0.2016    -0.4147  -0.5186  -0.3108   0.0530   0.6605    0.5953   0.7328
thorax         0.8224     2.6881   2.2393   3.1370   0.2290  14.7043    9.3864  23.0353

N = 124,  Events: 124,  Censored: 0
Total time at risk: 7145
Log-likelihood = -464.1, df = 7
AIC = 942.3
```

b. Discussion:

The estimated parameters from the flexsurvreg were $\beta_1 = 0.0553$, $\beta_2 = -0.1165$, $\beta_3 = 0.0824$, $\beta_1 = -0.4147$, and $\beta_1 = 2.6881$, which agreed with the parameters from the glm with family = Gamma(link = log).

2.

a. Code and output:

```
d.sur <- data.frame(activity = c("isolated","one","many","low","high"), thorax=0.82)
d.sur <- summary(m.flexSurv, newdata = d.sur, t = seq(0, 100, by = 1),
          type = "survival", tidy = TRUE)
```

Oscar Huang
April 26, 2017

```
d.haz <- data.frame(activity = c("isolated","one","many","low","high"),
thorax=0.82)
d.haz<-summary(m.flexSurv, newdata = d.haz, t = seq(0, 100, by = 1),
           type = "hazard", tidy = TRUE)

p.sur <- ggplot(d.sur, aes(x = time, y = est,color = activity))
p.sur <- p.sur + geom_line()
p.sur <- p.sur + labs(x = "Time (days)", y = "Probability",
           linetype = "Sexual Activity")
plot(p.sur)
```

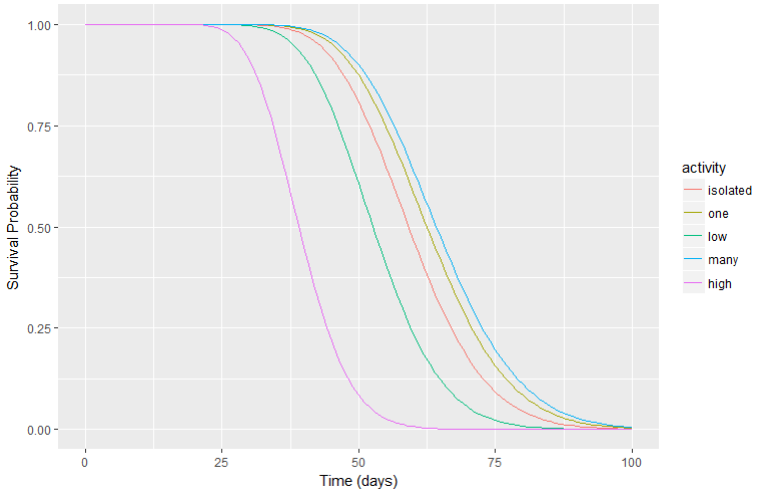

Figure 3 The survival function for each of the five sexual activity conditions.

```
p.haz <- ggplot(d.haz, aes(x = time, y = est, color = activity))
p.haz <- p.haz + geom_line()
p.haz <- p.haz + labs(x = "Time (days)", y = "Hazard Rate",
           linetype = "Sexual Activity")
plot(p.haz)
```

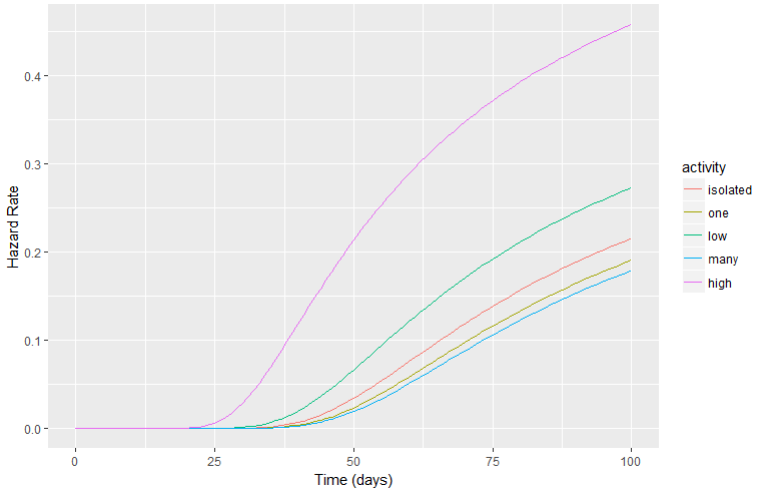

Figure 4 The hazard function for each of the five sexual activity conditions.

3.

a. Code and output:

```
fruitfly$activity<-relevel(fruitfly$activity,ref="many")
m.flexSurv <- flexsurvreg(Surv(longevity) ~ activity + thorax, dist =
custom.gamma,
                          data = fruitfly)
m.flexSurv$coefficients
```

shape	rate	activityhigh	activitylow	activityone
-3.36408	-1.39431	-0.49711	-0.1989	-0.02718
activityisolated	thorax			
-0.08253	2.68770			

```
exp(m.flexSurv$coefficients)
```

shape	rate	activityhigh	activitylow	activityone
0.03459	0.24800	0.60829	0.81959	0.97318
activityisolated	thorax			
0.92078	14.69787			

```
fruitfly$activity<-relevel(fruitfly$activity,ref="one")
m.flexSurv <- flexsurvreg(Surv(longevity) ~ activity + thorax, dist =
custom.gamma,
                          data = fruitfly)
m.flexSurv$coefficients
```

shape	rate	activitymany	activityhigh	activitylow
-3.36354	-1.42059	0.02714	-0.46983	-0.17181
activityisolated	thorax			
-0.05537	2.68728			

```
exp(m.flexSurv$coefficients)
```

shape	rate	activitymany	activityhigh	activitylow
0.03461	0.24157	1.02752	0.62511	0.84214
activityisolated	thorax			
0.94614	14.69171			

b. Discussion:

When using the sexual activity = “many” as reference group, the β_1 = -0.4971. (x_{i1} = 1 if the i-th fruit fly was in group “high”.) The survival functions will be:

$$S_{many}(t) = P\left(e^{\beta_0} e^{\sigma\epsilon_i} \geq t\right)$$

$$S_{high}(t) = P\left(e^{\beta_0} e^{\beta_1} e^{\sigma\epsilon_i} \geq t\right)$$

Since $\exp(\beta_1)$= 0.6083<1, the sexual activity will decelerate time. Which also means that the estimated survival probability of the fruit flies in group “high” will lower than the fruit flies in group “many”, which agreed with Figure 3; the estimated hazard rate of the fruit flies in group “high” will higher than the fruit flies in group “many”, which agreed with Figure 4.

Oscar Huang
April 26, 2017

When using the sexual activity = “one” as reference group, the $\beta_3 = -0.1718$. ($x_{i3} = 1$ if the i-th fruit fly was in group “low”.) The survival functions will be:

$$S_{one}(t) = P\left(e^{\beta_0} e^{\sigma\epsilon_i} \geq t\right)$$

$$S_{low}(t) = P\left(e^{\beta_0} e^{\beta_3} e^{\sigma\epsilon_i} \geq t\right)$$

Since $\exp(\beta_3) = 0.8421 < 1$, the sexual activity will decelerate time. Which also means that the estimated survival probability of the fruit flies in group “low” will lower than the fruit flies in group “one”, which agreed with Figure 3; the estimated hazard rate of the fruit flies in group “low” will higher than the fruit flies in group “one”, which agreed with Figure 4.

Oscar Huang
April 26, 2017

Breast Cancer Survival Data

1.

a. Code and output:

```
bc$group<-relevel(bc$group, ref = "Medium")
m <- flexsurvreg(Surv(recyrs, censored == "yes") ~ group, dist = "weibull", data = bc)
cbind(coef(m), confint(m))
                  2.5 % 97.5 %
shape       0.86342  0.78465 0.9422
scale       1.53541  1.47329 1.5975
groupMedium  0.06360 -0.03299 0.1602
groupPoor   -0.05298 -0.16355 0.0576
```

2.

a. Discussion:

$$T_i = e^{\beta_1 x_{i1}} e^{\beta_2 x_{i2}} e^{\sigma \epsilon_i}$$

$\hat{\beta}_1 \approx 0.06360\,;\ \hat{\beta}_2 \approx -0.05298\,;\ \sigma : scale\ parameter \approx 1.53541\,;$

$x_{i1} = 1\ if\ the\ i-th\ subject\ was \in \text{medium}\ group\,;\ x_{1i} = 0\ other\ wise$

$x_{i1} = 1\ if\ the\ i-th\ subject\ was \in \text{ poor }\ group\,;\ x_{1i} = 0\ other\ wise$

3.

a. Code and output:

```
dPlot.sur <- data.frame(group = c("Good", "Medium", "Poor"))
dPlot.sur <- summary(m, newdata = dPlot.sur, t = seq(0.1, 8, by = 0.1),
                     type = "survival", tidy = TRUE)
p.sur <- ggplot(dPlot.sur, aes(x = time, y = est, group=group, color = group))
p.sur<- p.sur + geom_line()
p.sur <- p.sur + labs(x = "Time Till Death (years)", y = "Probability",
                      linetype = "Prognostic group")
plot(p.sur)
```

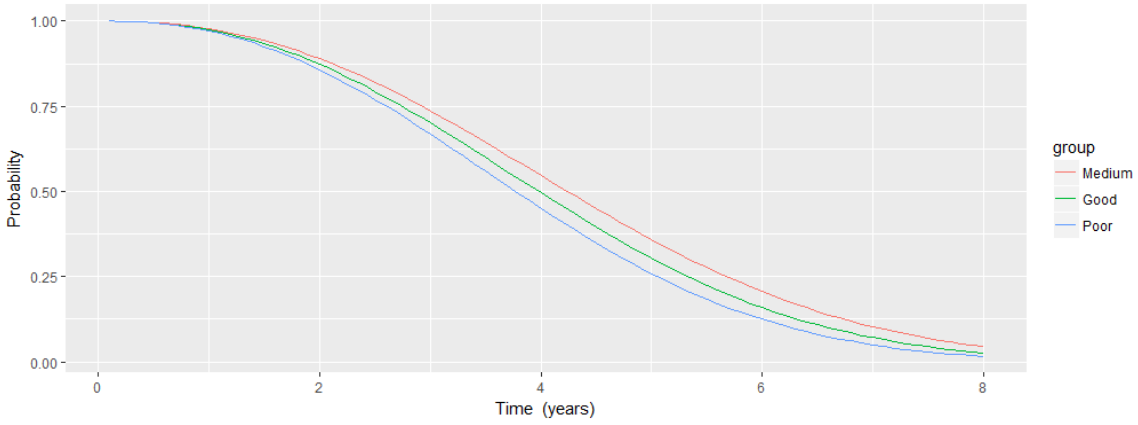

Figure 5 The survival function for each of the three prognostic groups.

Oscar Huang
April 26, 2017

```
dPlot.haz <- data.frame(group = c("Good", "Medium", "Poor"))
dPlot.haz <- summary(m, newdata = dPlot.haz, t = seq(0.1, 8, by = 0.1),
                     type = "hazard", tidy = TRUE)
p.haz <- ggplot(dPlot.haz, aes(x = time, y = est, group=group, color = group))
p.haz<- p.haz + geom_line()
p.haz <- p.haz + labs(x = "Time Till Death (years)", y = "Hazard Rate",
                      linetype = "Prognostic group")
plot(p.haz)
```

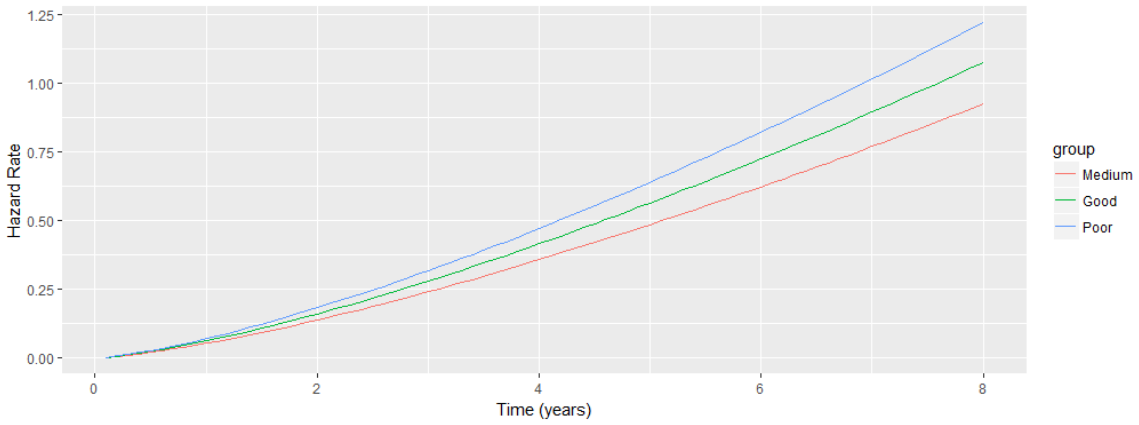

Figure 6 The hazard function for each of the three prognostic groups.

b. Discussion:

From the Figure 5, we can see at the same time, the subject in the medium prognostic group has highest probability to survive, the subject in the poor prognostic group has lowest probability to survive. From the Figure 6, we can also see at the same time, the subject in the medium prognostic group has lowest hazard rate to dead, the subject in the poor prognostic group has highest hazard rate to dead.

4.

a. Code and output:

exp(cbind(coef(m), confint(m)))

```
              2.5 % 97.5 %
shape     2.3713 2.1916  2.566
scale     4.9481 4.5964  5.327
groupGood 0.9384 0.8520  1.034
groupPoor 0.8900 0.7918  1.000
```

b. Discussion:

The subject in the “good” prognostic group has 16.16% shorter estimated time till death compared with the subject in the “medium” prognostic group (decreased by a factor of 0.9384). The subject in the “poor” prognostic group has 11.00% shorter estimated time till death compared with the subject in the “medium” prognostic group (decreased by a factor of 0.8900).

Oscar Huang
April 26, 2017

5.

c. Code and output:

m <- survreg(Surv(recyrs, censored == "yes") ~ group, dist = "weibull", data = bc)
exp(-m$coefficients[-1]/m$scale)

```
groupGood groupPoor
    1.163     1.318
```

d. Discussion:

When the subject was in the “good” prognostic group, the hazard function increase by a factor of 1.163 (increase 16.3%); When the subject was in the “poor” prognostic group, the hazard function increase by a factor of 1.318 (increase 31.8%). These results agreed with the Figure 6.