ST525 HW 6

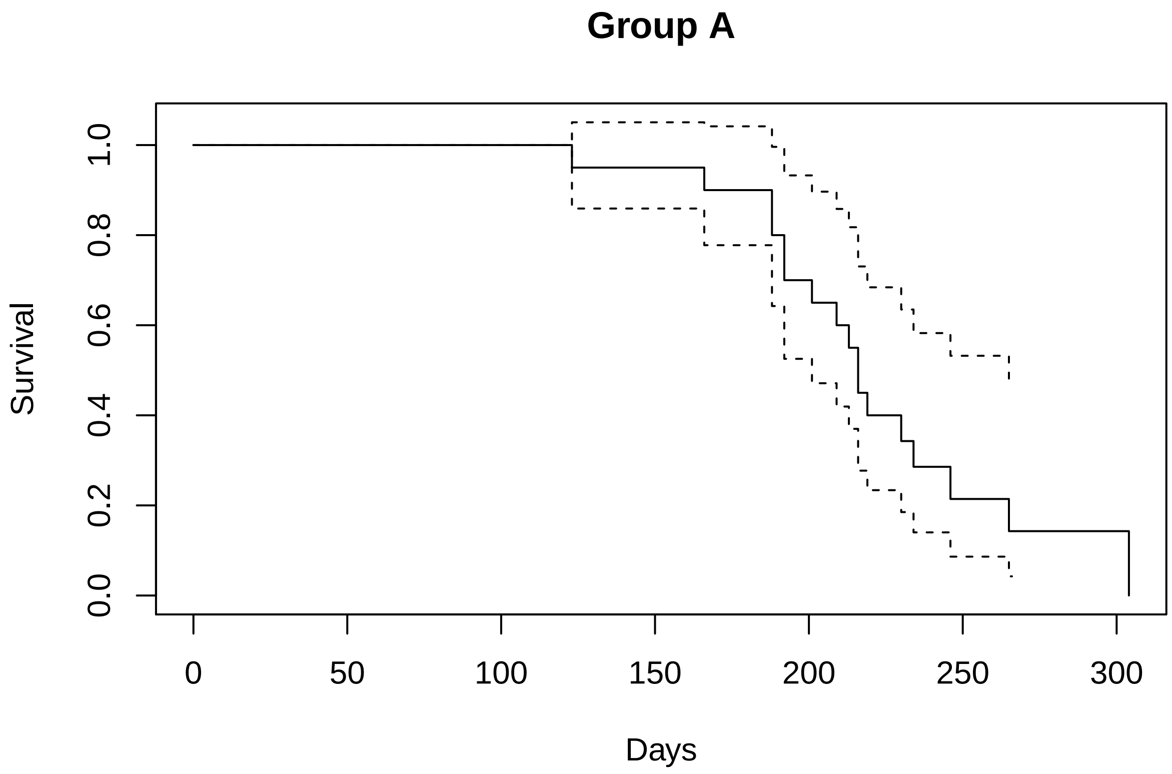
Nora Quick

Question 1

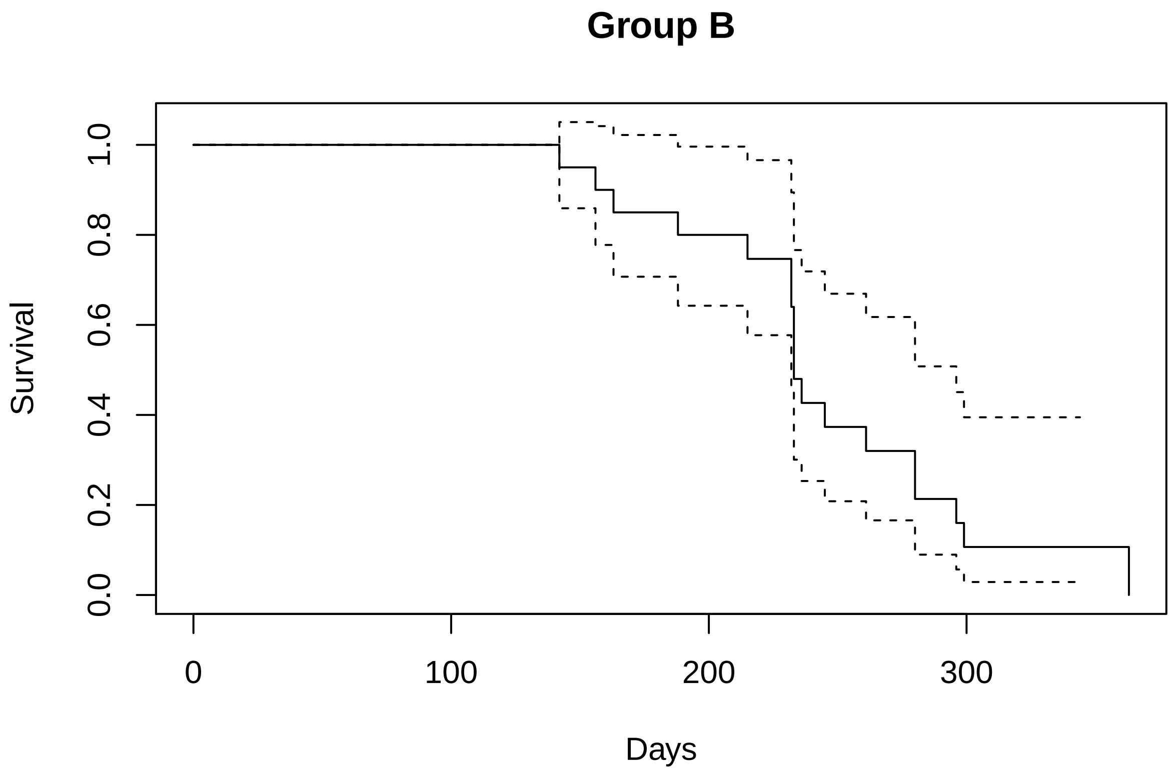
Part (a)

```
ratsA <- data.frame(days = c(123, 166, 188, 188, 192, 192, 201, 209, 213, 216, 216, 219, 230, 234, 246,
ratsB <- data.frame(days = c(142, 156, 163, 188, 215, 232, 232, 233, 233, 233, 236, 245, 261, 280, 280,
```

```
KM.A <- survfit(Surv(days, status) ~ 1, data = ratsA, conf.type="none")
summ.A <- summary(KM.A, times = c(0,ratsA$days))

plot(KM.A, main = 'Group A', xlab = 'Days',  ylab = 'Survival')
```

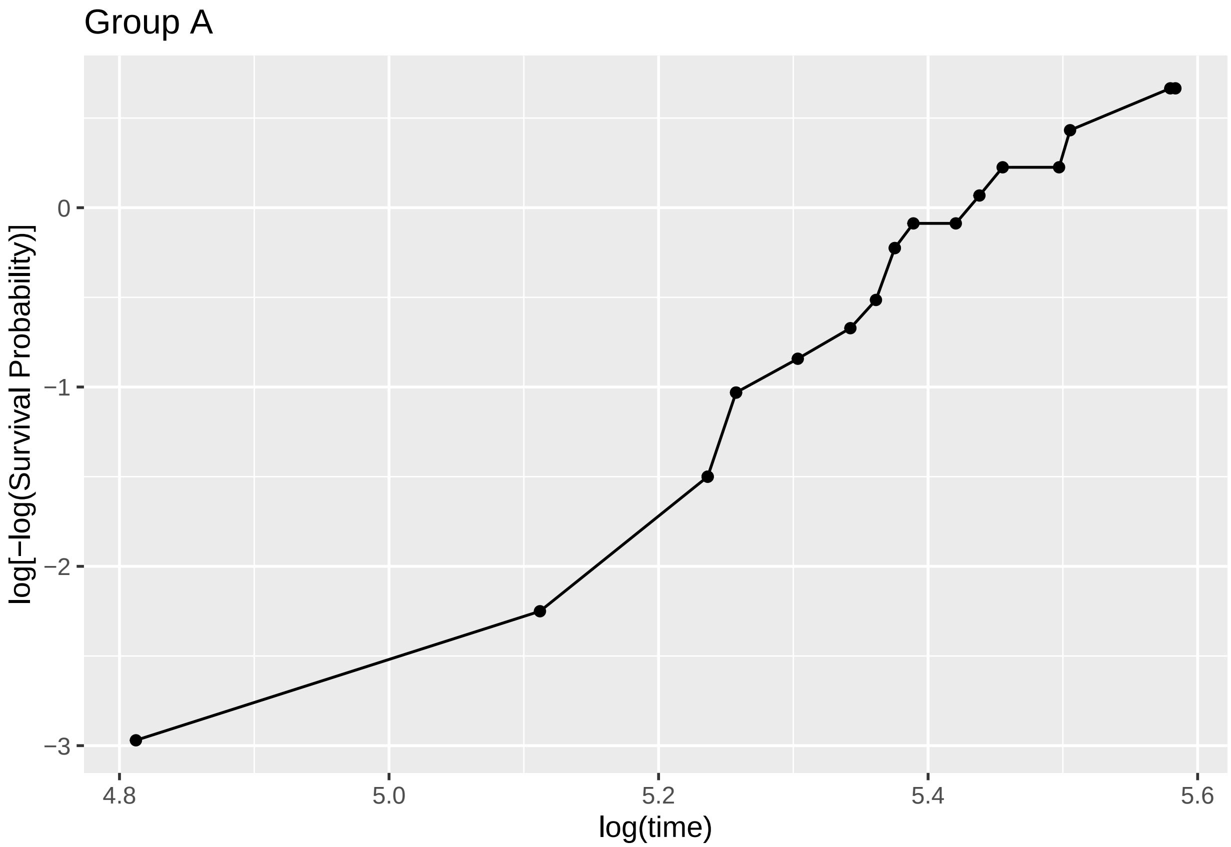


```
#----
KM.B <- survfit(Surv(days, status) ~ 1, data = ratsB, conf.type="none")
summ.B <- summary(KM.B, times = c(0,ratsB$days))

plot(KM.B, main = 'Group B', xlab = 'Days',  ylab = 'Survival')
```

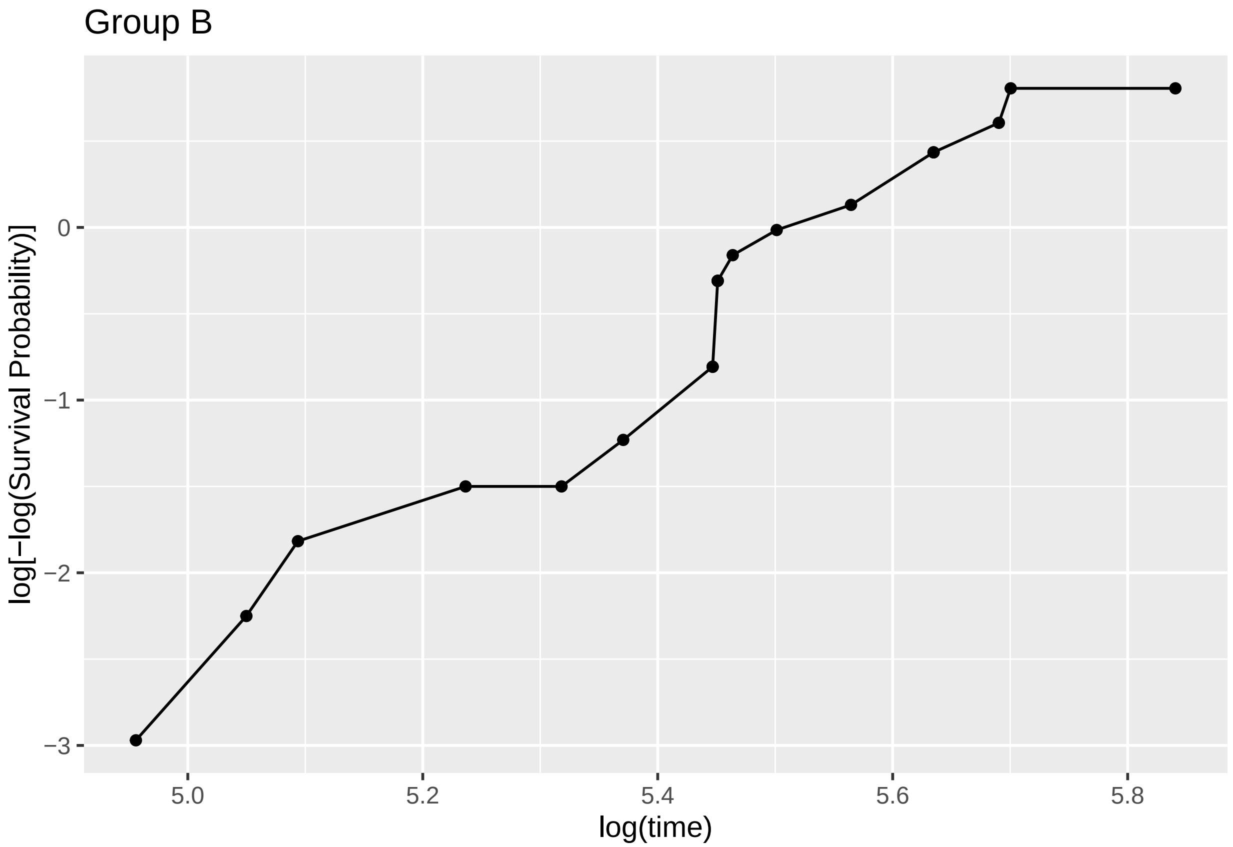


Part (b)

```
ltime <- log(summ.A$time)[-c(1,21)]
llsurv <- log(-log(summ.A$surv))[-c(1,21)]

wPlot.A <- data.frame(ltime, llsurv)

ggplot(data = wPlot.A) +
  geom_line(aes(x = ltime, y = llsurv)) +
  geom_point(aes(x = ltime, y = llsurv)) +
  labs(title = 'Group A') +
  ylab('log[-log(Survival Probability)]') +
  xlab('log(time)')
```



```
#----
ltime <- log(summ.B$time)[-c(1,21)]
llsurv <- log(-log(summ.B$surv))[-c(1,21)]

wPlot.B <- data.frame(ltime, llsurv)

ggplot(data = wPlot.B) +
  geom_line(aes(x = ltime, y = llsurv)) +
  geom_point(aes(x = ltime, y = llsurv)) +
  labs(title = ’Group B’) +
  ylab(’log[-log(Survival Probability)]’) +
  xlab(’log(time)’)
```



I believe based on the graphs above that the Kaplan-Meier estimators is a good fit for the survival fit of group A but less so for group B.

Part (c)

```
fitA <- survreg(formula = Surv(days, status) ~ ., data = ratsA, dist = ’weibull’)
summary(fitA)
```

```
##
## Call:
## survreg(formula = Surv(days, status) ~ ., data = ratsA, dist = "weibull")
##               Value Std. Error      z      p
## (Intercept)  5.4724     0.0434 126.21 <2e-16
## Log(scale)  -1.7378     0.1844  -9.43 <2e-16
##
## Scale= 0.176
##
## Weibull distribution
## Loglik(model)= -91.2   Loglik(intercept only)= -91.2
## Number of Newton-Raphson Iterations: 6
## n= 20
```

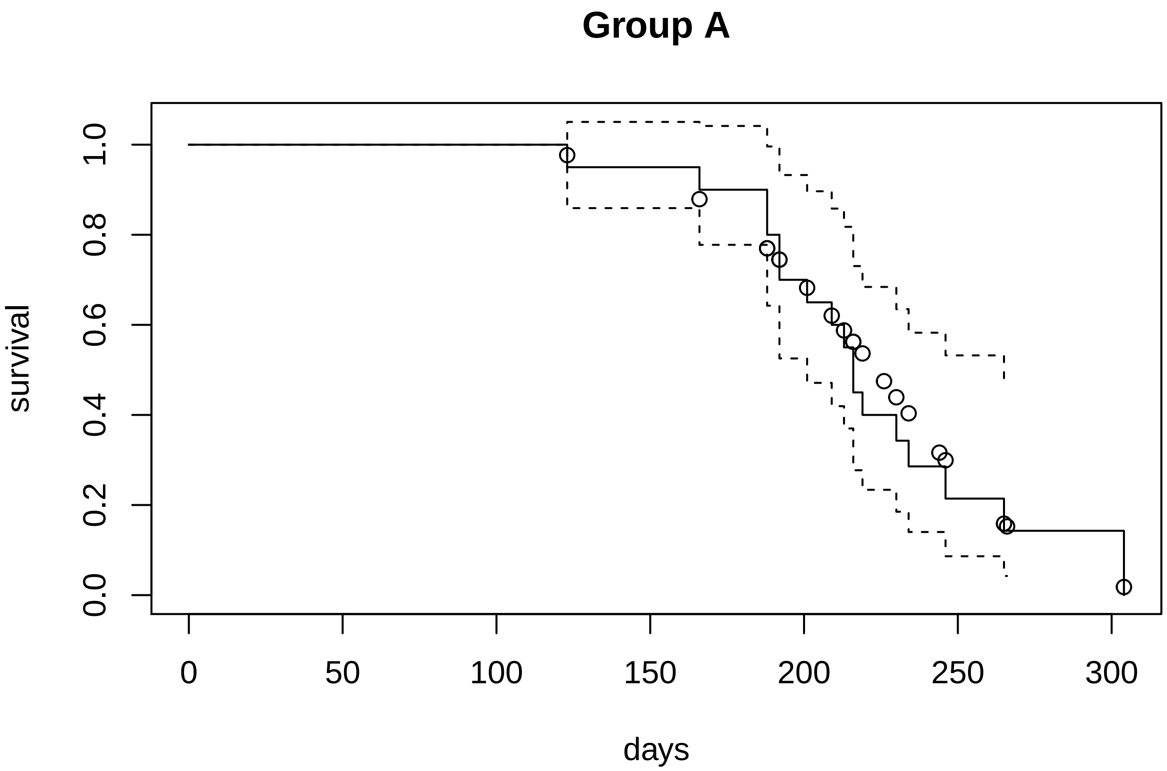
```r
#----
fitB <- survreg(formula = Surv(days, status) ~ ., data = ratsB, dist = ’weibull’)
summary(fitB)
```

```
##
## Call:
## survreg(formula = Surv(days, status) ~ ., data = ratsB, dist = "weibull")
##               Value Std. Error      z      p
## (Intercept)  5.5966     0.0544 102.81 <2e-16
## Log(scale)  -1.4960     0.1790  -8.36 <2e-16
##
## Scale= 0.224
##
## Weibull distribution
## Loglik(model)= -101.2   Loglik(intercept only)= -101.2
## Number of Newton-Raphson Iterations: 6
## n= 20
```

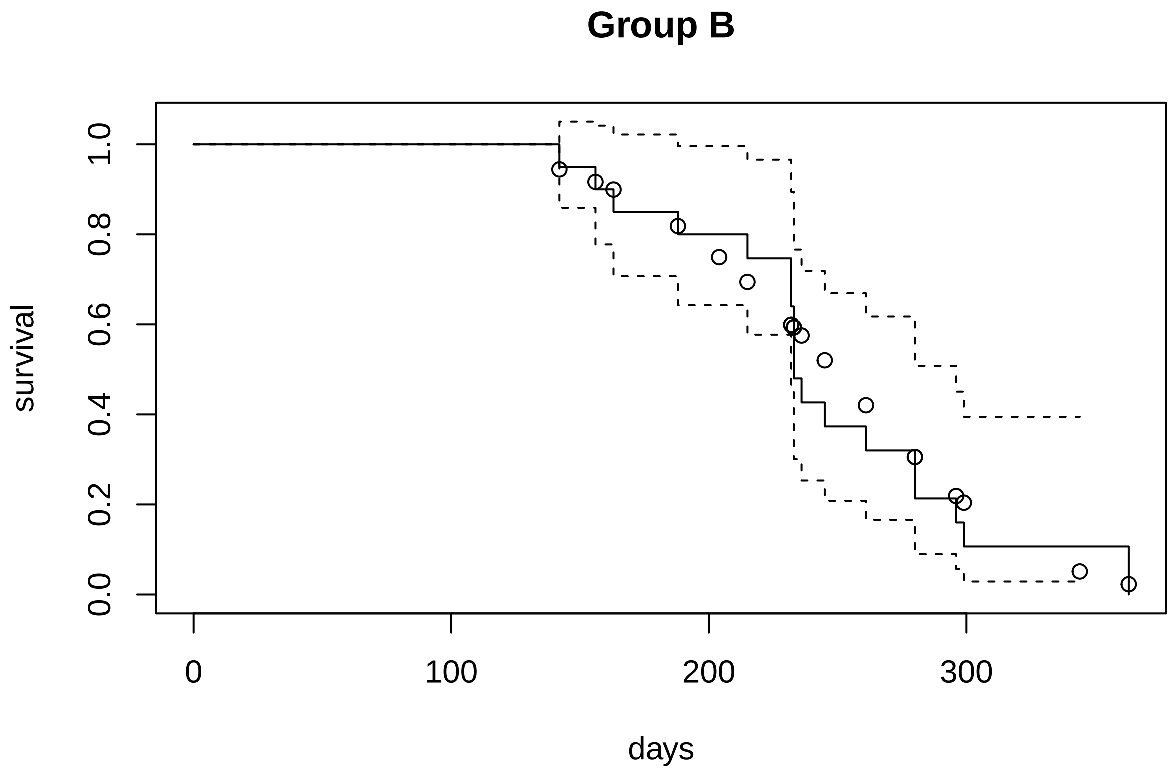
The parameter estimates for the Weibull distributions indicate significance with both having a p-value of <2e-16.

Part (d)

```r
shapeA <- 1/fitA$scale
scaleA <- exp(fitA$coefficients)

plot(KM.A, main = ’Group A’, xlab = ’days’, ylab = ’survival’)
points(ratsA$days, pweibull(ratsA$days, shapeA, scaleA, lower.tail=F),type=’p’)
```



```
#----
shapeB <- 1/fitB$scale    #shape of Weibull
scaleB <- exp(fitB$coefficients) #scale of Weibull

plot(KM.B, main = 'Group B', xlab = 'days', ylab = 'survival')
points(ratsB$days, pweibull(ratsB$days, shapeB, scaleB, lower.tail=F),type='p')
```



Assessing again if Weibull distirbution is a good fit for the data I would conclude that, yes, it is a good fit for both group A and group B.

Question 2

```r
smoke <- read.csv('pharmacoSmoking-old.csv')
head(smoke)
```

```
##    id ttr relapse grp age gender race employment yearsSmoking levelSmoking
## 1  21  41       0   2  36      1    4          1           26            1
## 2 113  14       1   2  41      1    4          2           27            1
## 3  39   5       1   1  25      0    4          2           12            1
## 4  80  16       1   1  54      1    4          1           39            1
## 5  87   0       1   1  45      1    4          2           30            1
## 6  29 157       0   1  43      1    2          1           30            1
##    admitdate      fdate priorAttempts longestNoSmoke
## 1 11/20/2005 12/31/2005             0              0
## 2  6/16/2005  6/30/2005             3             90
## 3   5/9/2005  5/14/2005             3             21
## 4 10/26/2005 11/11/2005             0              0
## 5  9/27/2005  9/27/2005             0              0
## 6   7/6/2005 12/10/2005             2           1825
```

```
smoke$admitdate <- smoke$admitdate %>%
  sapply(function(x) x[1]) %>% as.Date(format = c("%m/%d/%y"))

smoke$fdate <- smoke$fdate %>%
  sapply(function(x) x[1]) %>% as.Date(format = c("%m/%d/%y"))

smoke$time <- difftime(smoke$fdate , smoke$admitdate, units = ’days’) %>% as.numeric()

head(smoke)
```

```
##    id ttr relapse grp age gender race employment yearsSmoking levelSmoking
## 1  21  41       0   2  36      1    4          1           26            1
## 2 113  14       1   2  41      1    4          2           27            1
## 3  39   5       1   1  25      0    4          2           12            1
## 4  80  16       1   1  54      1    4          1           39            1
## 5  87   0       1   1  45      1    4          2           30            1
## 6  29 157       0   1  43      1    2          1           30            1
##    admitdate      fdate priorAttempts longestNoSmoke time
## 1 2020-11-20 2020-12-31             0              0   41
## 2 2020-06-16 2020-06-30             3             90   14
## 3 2020-05-09 2020-05-14             3             21    5
## 4 2020-10-26 2020-11-11             0              0   16
## 5 2020-09-27 2020-09-27             0              0    0
## 6 2020-07-06 2020-12-10             2           1825  157
```

```
smoke$time <- smoke[,1] + 0.1
head(smoke)
```

```
##    id ttr relapse grp age gender race employment yearsSmoking levelSmoking
## 1  21  41       0   2  36      1    4          1           26            1
## 2 113  14       1   2  41      1    4          2           27            1
## 3  39   5       1   1  25      0    4          2           12            1
## 4  80  16       1   1  54      1    4          1           39            1
## 5  87   0       1   1  45      1    4          2           30            1
## 6  29 157       0   1  43      1    2          1           30            1
##    admitdate      fdate priorAttempts longestNoSmoke  time
## 1 2020-11-20 2020-12-31             0              0  21.1
## 2 2020-06-16 2020-06-30             3             90 113.1
## 3 2020-05-09 2020-05-14             3             21  39.1
## 4 2020-10-26 2020-11-11             0              0  80.1
## 5 2020-09-27 2020-09-27             0              0  87.1
## 6 2020-07-06 2020-12-10             2           1825  29.1
```

Part (a)

```
fitSmoke <- survreg(formula = Surv(time, relapse) ~
                  gender + age + grp,
                data = smoke, dist = ’weibull’)
summary(fitSmoke)
```

```
##
## Call:
## survreg(formula = Surv(time, relapse) ~ gender + age + grp, data = smoke,
##     dist = "weibull")
##                 Value Std. Error     z      p
## (Intercept)  4.681913   0.340858 13.74 <2e-16
## gender      -0.013984   0.114511 -0.12   0.90
## age          0.000627   0.004948  0.13   0.90
## grp         -0.130668   0.113332 -1.15   0.25
## Log(scale)  -0.691637   0.088632 -7.80  6e-15
##
## Scale= 0.501
##
## Weibull distribution
## Loglik(model)= -468.2   Loglik(intercept only)= -469
##  Chisq= 1.61 on 3 degrees of freedom, p= 0.66
## Number of Newton-Raphson Iterations: 8
## n= 125
```

Part (b)

The coefficient of age (0.0006) indicates being older is better after ajusting for gender and grp as well. However, this improvement is not statistically significant with a p-value of 0.90.

Part (c)

The coefficient of gender (-0.014) indicates being a female is better than being a male, however, with a p-value of 0.90 it is not statistically significant.

Part (d)

The coeggicient of grp (-0.13) indicates that being in combination is better than being in patch-only, however, with a p-value of 0.25 we can, again, conclude that it is not statistically significant.

Question 3

The censoring indicator is indeed switched in this dataset, so you will have to create a new variable that converts it to the expected 1=observed, 0=censored format.

```
color <- read.csv('color.csv')
head(color)
```

```
##   group time status DVAL FVAL
## 1 Green   19      0   43   85
## 2  Blue   88      0   33   63
## 3 Green   23      0   45   77
## 4  Blue   89      0   38   41
## 5  Blue   24      0   45   51
## 6 Green   91      0   49   77
```

```
color$status <- ifelse(color$status==1, 0, ifelse(color$status==0, 1, color$status))
head(color)
```

```
##   group time status DVAL FVAL
## 1 Green   19      1   43   85
## 2  Blue   88      1   33   63
## 3 Green   23      1   45   77
## 4  Blue   89      1   38   41
## 5  Blue   24      1   45   51
## 6 Green   91      1   49   77
```

Part (a)

```
fitColor <- survreg(formula = Surv(time, status) ~
                  group + DVAL + FVAL,
                data = color, dist = 'weibull')
summary(fitColor)
```

```
##
## Call:
## survreg(formula = Surv(time, status) ~ group + DVAL + FVAL, data = color,
##     dist = "weibull")
##               Value Std. Error      z       p
## (Intercept)  5.01120    0.20122  24.90 < 2e-16
## groupGreen   0.26720    0.07693   3.47 0.00051
## DVAL        -0.00520    0.00403  -1.29 0.19768
## FVAL        -0.00810    0.00296  -2.73 0.00630
## Log(scale)  -1.23798    0.10059 -12.31 < 2e-16
##
## Scale= 0.29
##
## Weibull distribution
## Loglik(model)= -316.8   Loglik(intercept only)= -324.4
##  Chisq= 15.29 on 3 degrees of freedom, p= 0.0016
## Number of Newton-Raphson Iterations: 6
## n=80 (1 observation deleted due to missingness)
```

Part (b)

Yes, the failure times do depend on the group with a p-value of 2e-16 in favor of the green group. It also appears that FVAL also statistically matters (p-value = 0.006) for faulure times but DVAL does not (p-value = 0.198).

Part (c)

Group green has a longer expected survival time. FVAL is also significant and has a 0.99 times (exp(-0.008)) shorter survival time. The larger this value is the shorter the survival time would be and the smaller it get the longer the survival time would be.

Part (d)

```
fitexColor <- survreg(formula = Surv(time, status) ~
                  group + DVAL + FVAL,
                data = color, dist = 'exponential')
summary(fitexColor)
```

```
##
## Call:
## survreg(formula = Surv(time, status) ~ group + DVAL + FVAL, data = color,
##     dist = "exponential")
##                 Value Std. Error     z       p
## (Intercept)  4.97238    0.72996  6.81 9.6e-12
## groupGreen   0.25205    0.25926  0.97    0.33
## DVAL        -0.00258    0.01281 -0.20    0.84
## FVAL        -0.00817    0.01051 -0.78    0.44
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -362.4   Loglik(intercept only)= -363.1
##  Chisq= 1.43 on 3 degrees of freedom, p= 0.7
## Number of Newton-Raphson Iterations: 3
## n=80 (1 observation deleted due to missingness)
```

Part (e)

No, there is no statistically significant data to indicate failure times depend on group, DVAL, or FVAL with p-values of 0.33, 0.84, and 0.44.

Part (f)

Yes, two of my conclusions differ. In particular, assuming exponenetial instead of Weibell decreases the significance of the covariates.

Part (g)

```
2*(-316.8-(-362.4))
```

```
## [1] 91.2
```

The outputted p-value is extremely small (2e-16) indicating that there is a difference between the two fits. We can find that the Weibull distribution is the better fit for this data based on the log likelihood values.