ST525 HW 7

Nora Quick

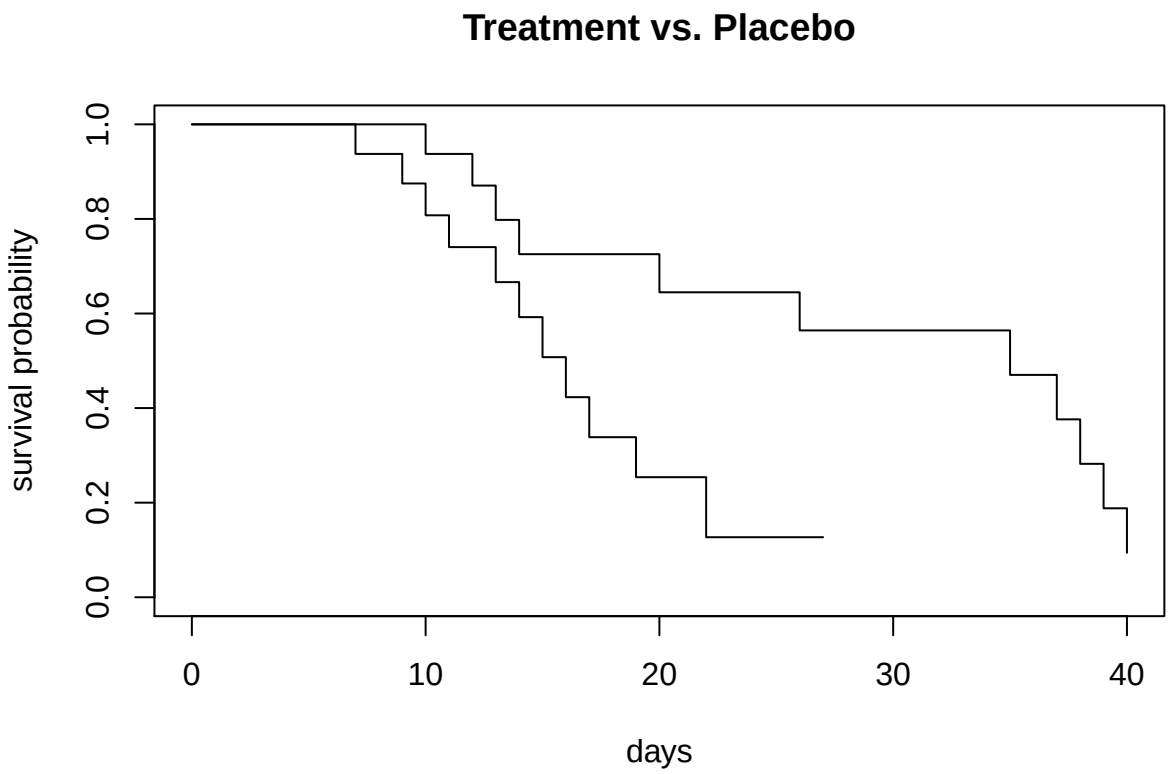
Question 1

```
treatment <- data.frame(days = c(10, 12, 13, 14, 20, 26, 35, 37, 38, 39, 40, 10, 12, 17, 34, 40), status
placebo <- data.frame(days = c(7, 9, 10, 11, 13, 14, 15, 16, 17, 19, 22, 9, 11, 14, 21, 27), status = r


q1data <- rbind(treatment, placebo)
head(q1data)
```

```
##   days status group
## 1   10      1     1
## 2   12      1     1
## 3   13      1     1
## 4   14      1     1
## 5   20      1     1
## 6   26      1     1
```

Part (a)

```
fit <- survfit(Surv(days, status) ~ group, data = q1data)

plot(fit, main = "Treatment vs. Placebo", xlab = "days", ylab = "survival probability")
```



Part (b)

```
logrank <- survdiff(Surv(days, status) ~ group, data = q1data)
logrank
```

```
## Call:
## survdiff(formula = Surv(days, status) ~ group, data = q1data)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## group=1 16       11    15.44      1.28      5.24
## group=2 16       11     6.56      3.01      5.24
##
##  Chisq= 5.2  on 1 degrees of freedom, p= 0.02
```

The test statistic is 15.44 and 6.56 respectively with a p-value of 0.02. Based on the log-rank test it appears that, yes, there is strong evidence to conclude that the treatment has an effect on the patient's survival time. Due to the p-value there is moderate evidence to reject the null hypothesis that they have the same survival.

Part (c)

```
fit1 <- survreg(formula = Surv(days, status) ~ group,
                data = q1data, dist = 'exponential')
summary(fit1)
```

```
##
## Call:
## survreg(formula = Surv(days, status) ~ group, data = q1data,
##     dist = "exponential")
##              Value Std. Error     z       p
## (Intercept)  4.110      0.674  6.10 1.1e-09
## group       -0.524      0.426 -1.23    0.22
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -95.1   Loglik(intercept only)= -95.9
##  Chisq= 1.5 on 1 degrees of freedom, p= 0.22
## Number of Newton-Raphson Iterations: 4
## n= 32
```

Part (d)

To begin with the assumptions between the two approaches, I believe that it is okay to assume exponential survival times due to the longer a patient has leukemia the more likely it is to effect their survival. So I think that we could assume it for both.

There are discrepancies between the two approaches because one compares the treatment and placebo against each other while the other simply looks are the survival of each.

Question 2

```
kidney <- read.table('~/kidney.txt', header = TRUE)
head(kidney)
```

```
##   Time Delta Type
## 1  1.5     1    1
## 2  3.5     1    1
## 3  4.5     1    1
## 4  4.5     1    1
## 5  5.5     1    1
## 6  8.5     1    1
```

Part (a)

```
new_kidney <- kidney %>% filter(Type == 2)
head(new_kidney)
```

```
##   Time Delta Type
## 1  0.5     1    2
## 2  0.5     1    2
## 3  0.5     1    2
## 4  0.5     1    2
## 5  0.5     1    2
## 6  0.5     1    2
```

```
fit1 <- survreg(formula = Surv(Time, Delta) ~ 1,
                data = new_kidney, dist = 'weibull')
summary(fit1)
```

```
##
## Call:
## survreg(formula = Surv(Time, Delta) ~ 1, data = new_kidney, dist = "weibull")
##             Value Std. Error    z       p
## (Intercept) 5.411      1.024 5.29 1.3e-07
## Log(scale)  0.616      0.265 2.32    0.02
##
## Scale= 1.85
##
## Weibull distribution
## Loglik(model)= -51.6   Loglik(intercept only)= -51.6
## Number of Newton-Raphson Iterations: 7
## n= 76
```

Part (b)

The maximum likelihood estimates are 0.616 with a standard error of 0.265.

Part (c)

```
fit2 <- survreg(formula = Surv(Time, Delta) ~ 1,
                data = new_kidney, dist = 'exponential')
summary(fit2)
```

```
##
## Call:
## survreg(formula = Surv(Time, Delta) ~ 1, data = new_kidney, dist = "exponential")
##             Value Std. Error    z      p
## (Intercept) 4.011      0.302 13.3 <2e-16
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -55.1   Loglik(intercept only)= -55.1
## Number of Newton-Raphson Iterations: 6
## n= 76
```

```
2 * (fit1$loglik[2] - fit2$loglik[2])
```

```
## [1] 7.127063
```

From the likelihood ratio test I found a test statistic of 7.127 and from the Wald test I found a p-value of 0.02.

Part (d)

Based on the above information above (c) I would conclude that with moderate evidence we reject the null hypothesis of scale = 1.

Part (e)

```
fit3 <- survreg(formula = Surv(Time, Delta) ~ Type,
                data = kidney, dist = 'weibull')
summary(fit3)
```

```
##
## Call:
## survreg(formula = Surv(Time, Delta) ~ Type, data = kidney, dist = "weibull")
##             Value Std. Error    z       p
## (Intercept) 2.974      0.682 4.36 1.3e-05
## Type        0.623      0.469 1.33    0.18
## Log(scale)  0.129      0.167 0.77    0.44
##
## Scale= 1.14
##
## Weibull distribution
## Loglik(model)= -122   Loglik(intercept only)= -122.9
##  Chisq= 1.93 on 1 degrees of freedom, p= 0.16
## Number of Newton-Raphson Iterations: 7
## n= 119
```

Part (f)

The maximum likelihood estimates are 0.129 with a standard error of 0.167.

My interpritation is that with both groups in the mix and the numbers seen above we can conclude that this fit will likely provide information that scale = 1.

Part (g)

```
fit4 <- survreg(formula = Surv(Time, Delta) ~ Type,
                data = kidney, dist = 'exponential')
summary(fit4)
```

```
##
## Call:
## survreg(formula = Surv(Time, Delta) ~ Type, data = kidney, dist = "exponential")
##             Value Std. Error    z       p
## (Intercept) 2.944      0.598 4.92 8.5e-07
## Type        0.534      0.397 1.34    0.18
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -122.3   Loglik(intercept only)= -123.2
##  Chisq= 1.83 on 1 degrees of freedom, p= 0.18
## Number of Newton-Raphson Iterations: 6
## n= 119
```

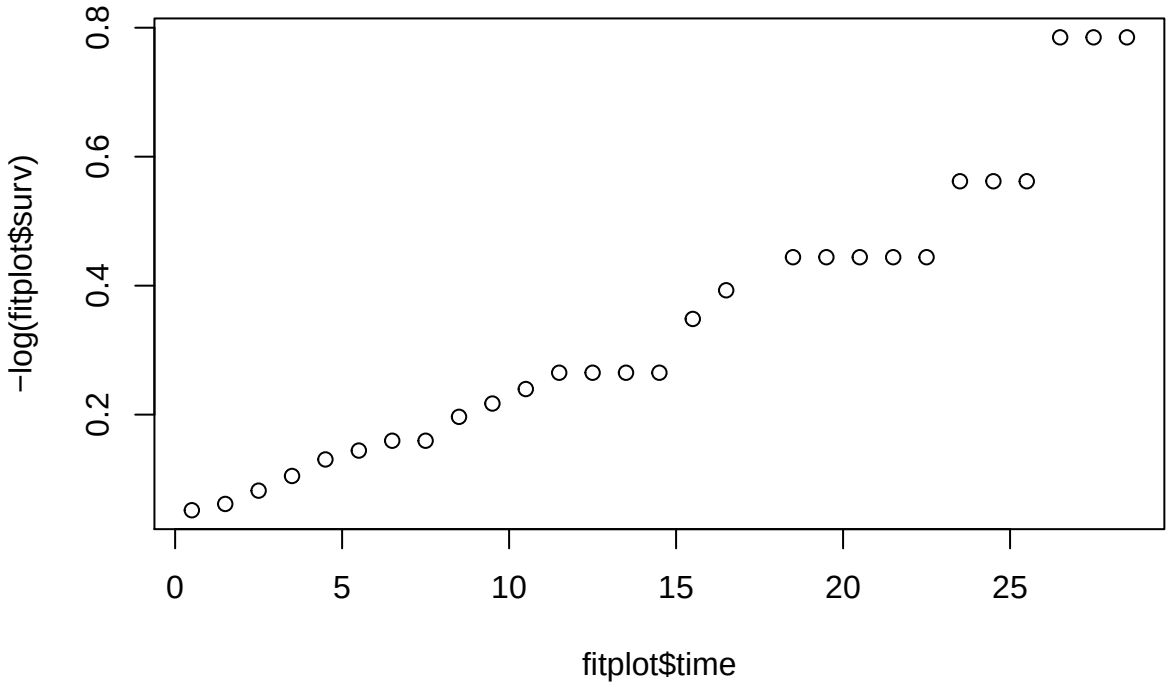
```
2 * (fit3$loglik[2] - fit4$loglik[2])
```

```
## [1] 0.628502
```

From the likelihood ratio test I found a test statistic of about 0.63 and from the Wald test I found a p-value of 0.44.

Part (h)

```
fitplot <- survfit(Surv(Time, Delta) ~ 1, data = kidney)

plot(fitplot$time,-log(fitplot$surv))
```



Based on the above graph and the data provided in fit1 and fit3 I would say that, no, the Weibull regression model is not a good fit for this data. As we know from lab we want a linear relationship within the graph above to prove a good fit which we do not see.

Question 3

```
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)

```
smoke$employment <- as.factor(smoke$employment)

fit5 <- survreg(formula = Surv(time, relapse) ~
                  age + gender + employment + yearsSmoking + priorAttempts,
                data = smoke, dist = 'weibull')
summary(fit5)
```

```
##
## Call:
## survreg(formula = Surv(time, relapse) ~ age + gender + employment +
##     yearsSmoking + priorAttempts, data = smoke, dist = "weibull")
##                   Value Std. Error     z       p
## (Intercept)    4.54e+00   2.65e-01 17.16 < 2e-16
## age           -7.74e-03   9.72e-03 -0.80    0.43
## gender        -1.31e-02   1.16e-01 -0.11    0.91
## employment2   -4.86e-02   1.27e-01 -0.38    0.70
## employment3   -4.28e-02   1.63e-01 -0.26    0.79
## yearsSmoking   1.19e-02   9.72e-03  1.22    0.22
## priorAttempts  8.01e-06   5.04e-04  0.02    0.99
## Log(scale)    -7.00e-01   8.89e-02 -7.87 3.5e-15
##
## Scale= 0.497
##
## Weibull distribution
## Loglik(model)= -468.1   Loglik(intercept only)= -469
##  Chisq= 1.78 on 6 degrees of freedom, p= 0.94
## Number of Newton-Raphson Iterations: 9
## n= 125
```

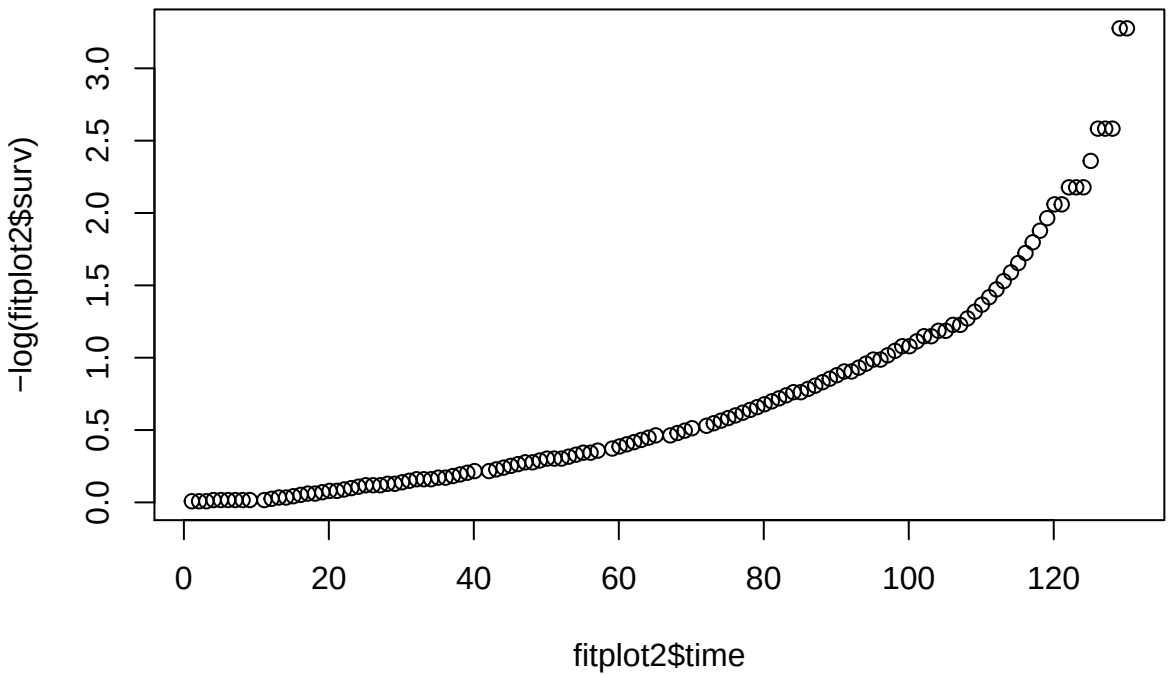
Part (b)

Age, yearsSmoking, and priorAttempts all have very small positive coefficients. On the other side of it they all have relatively large p-values; 0.43, 0.22, and 0.99 respectivly.

Gender and employement (2 and 3) also have very small coefficients but unlike the other variables they are negative. However, similar to the other variables they also have relatively large p-values; 0.91, 0.70, and 0.79 respecibly.

Part (c)

```
fitplot2 <- survfit(Surv(time, relapse) ~ 1, data = smoke)

plot(fitplot2$time,-log(fitplot2$surv))
```



Based on the graph above I would conclude that, no, the Weibull regression model is not appropriate for this data set. Again, as we know from lab we want to see a linear relationship but here we see an explonential model. Therefore, it appears that an exponential model would be better.

Part (d)

As we can see in part (b) out model and interpritation makes it look like Weibull is a good model for the data due to the p-values (indicating that we would fail to reject the null hypothesis) but looking at the graph it doesn't appear to be a good model. If we are not using an appropriate model we could conlude an incorrect conclusion such as failing to reject when we should be rejecting the null hypothesis or rejectin when we should be failing to reject the null hypothesis. Additionally, the estimates could be wrong so we may put undo importance on variables.