SA3 Individual Assignment

By:

Swapnil Vermani

2018-2019

Count Data Analysis

```
calves <- read.csv("/Users/swapnilvermani/Downloads/whale.csv")
```

1.Obtain summary statistics for the two whale calves, Hudson and Casey and make comparative statements.

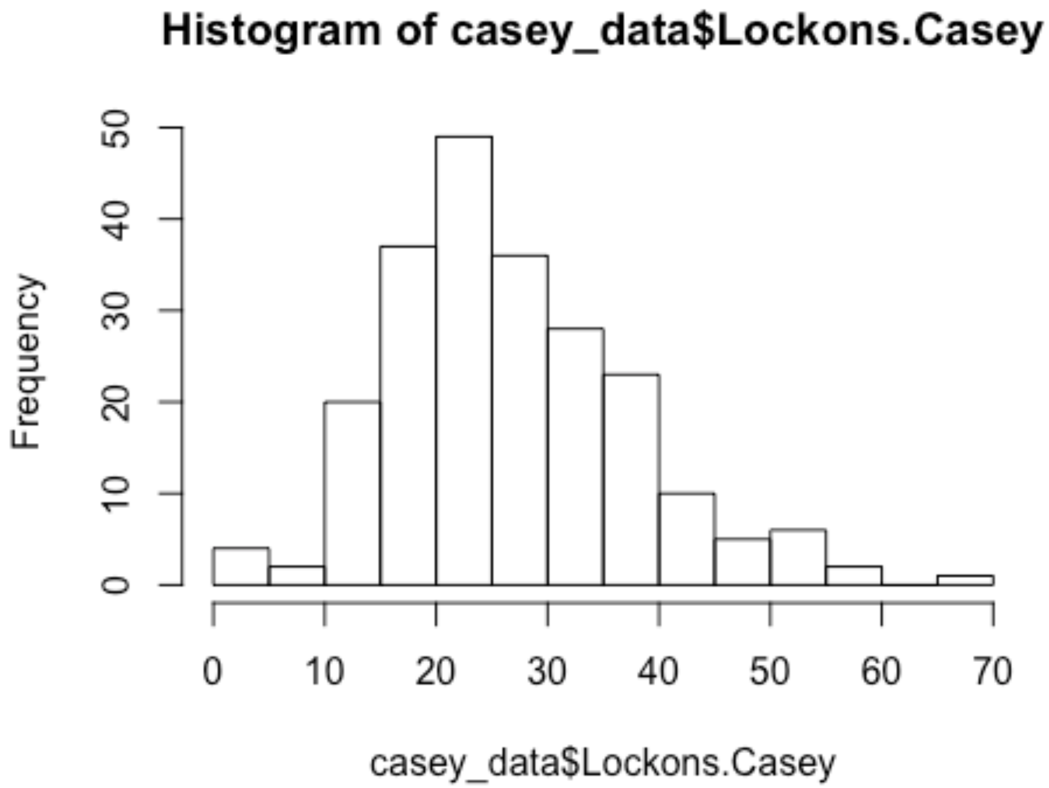
Separating the groups and changing the categorical variables into factors

```
head(calves)

##   Period.Hudson Bouts.Hudson Lockons.Hudson Daytime.Hudson Period.Casey
## 1             1            0              0              0            1
## 2             2            0              0              1            2
## 3             3            0              0              1            3
## 4             4            0              0              0            4
## 5             5            0              0              0            5
## 6             6            4              5              1            6
##   Bouts.Casey Lockons.Casey Daytime.Casey
## 1           0             0             0
## 2           0             0             1
## 3           0             0             1
## 4           9            12             0
## 5          10            28             0
## 6          13            35             1

hudson_data <- calves[,1:4]
casey_data <-calves[,5:8]
casey_data$Daytime.Casey <-as.factor(casey_data$Daytime.Casey)
casey_data$Daytime.Casey <-as.factor(relevel(casey_data$Daytime.Casey,ref="1"
))
hudson_data$Daytime.Hudson <- as.factor(hudson_data$Daytime.Hudson)
hudson_data$Daytime.Hudson <-as.factor(relevel(hudson_data$Daytime.Hudson,ref
="1"))
```

For the Casey calves

```
head(casey_data)

##   Period.Casey Bouts.Casey Lockons.Casey Daytime.Casey
## 1            1           0             0             0
## 2            2           0             0             1
## 3            3           0             0             1
## 4            4           9            12             0
## 5            5          10            28             0
## 6            6          13            35             1

summary(casey_data)
```

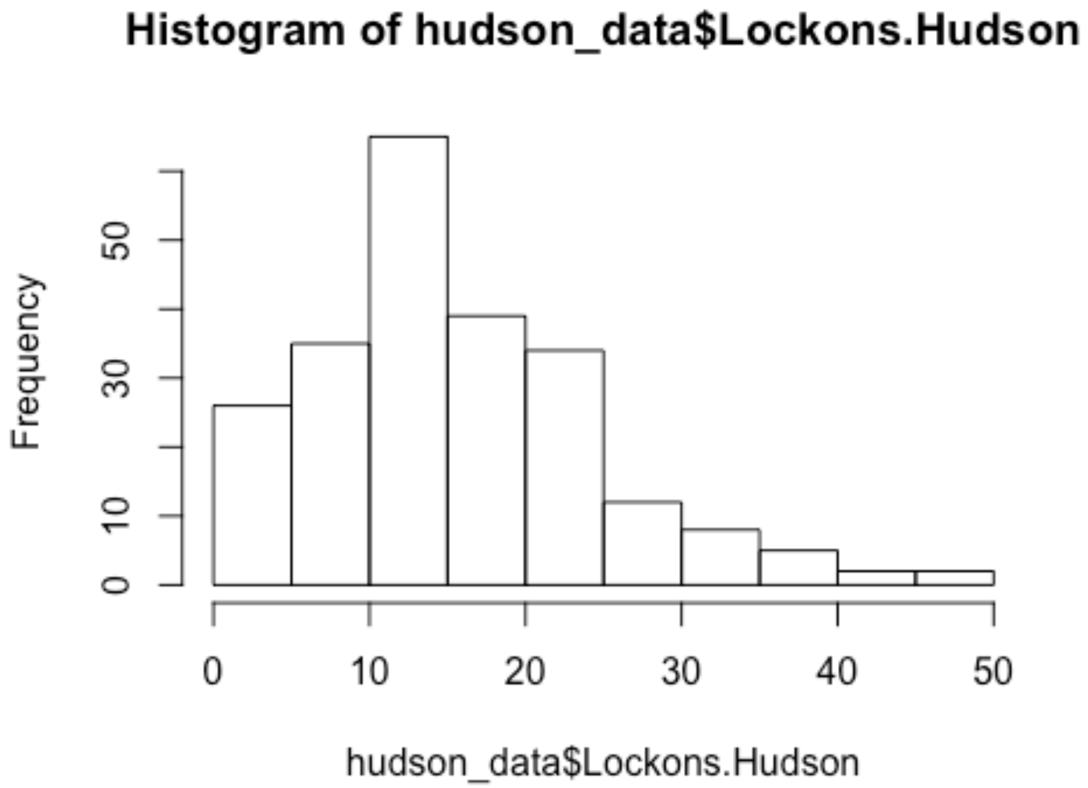
```
##   Period.Casey     Bouts.Casey     Lockons.Casey   Daytime.Casey
##  Min.   :  1.0   Min.   : 0.00   Min.   : 0.00   1   :112     
##  1st Qu.: 56.5   1st Qu.: 9.00   1st Qu.:20.00   0   :111     
##  Median :112.0   Median :11.00   Median :25.00   NA's:  5     
##  Mean   :112.0   Mean   :10.83   Mean   :27.13                
##  3rd Qu.:167.5   3rd Qu.:13.00   3rd Qu.:34.00                
##  Max.   :223.0   Max.   :24.00   Max.   :66.00                
##  NA's   :5       NA's   :5       NA's   :5                    

hist(casey_data$Lockons.Casey,breaks = 10)
```



For the Hudson calves

```
head(hudson_data)

##   Period.Hudson Bouts.Hudson Lockons.Hudson Daytime.Hudson
## 1             1            0              0              0
## 2             2            0              0              1
## 3             3            0              0              1
## 4             4            0              0              0
## 5             5            0              0              0
## 6             6            4              5              1

summary(hudson_data)
```

```
##   Period.Hudson      Bouts.Hudson     Lockons.Hudson Daytime.Hudson
##   Min.   :  1.00   Min.   : 0.000   Min.   : 0.0   1:114         
##   1st Qu.: 57.75   1st Qu.: 6.000   1st Qu.:10.0   0:114         
##   Median :114.50   Median : 9.000   Median :14.0                 
##   Mean   :114.50   Mean   : 8.741   Mean   :15.8                 
##   3rd Qu.:171.25   3rd Qu.:11.000   3rd Qu.:21.0                 
##   Max.   :228.00   Max.   :18.000   Max.   :49.0                 

hist(hudson_data$Lockons.Hudson,breaks = 10)
```



Studying the summary statistics and the distribution curves for the two calves we can clearly see some

a)The average number of bouts of the whales are almost the same with Casey(10.83) and Hudson(8.7) but there is a huge difference between the lockons of the whales with Casey cows able to achieve 27.13 lockons while Hudson having an average of 15.8

b)Also we see that there are a lot of NA values in the Casey whales data while for the Hudson whales there are no missing values.

c)The similarity we see in both of the subsets is that we dont see our intrest variable no of lockons following a normal curve in either of them.

2. Make comparative statements on the underlying probability distributions of the

number of lockons of the two calves. With justification, propose appropriate

regression model for number of lockons for each data set.

Both the probability distribution functions of Casey and hudson whales follows skewed distributions and are not exactly normal.

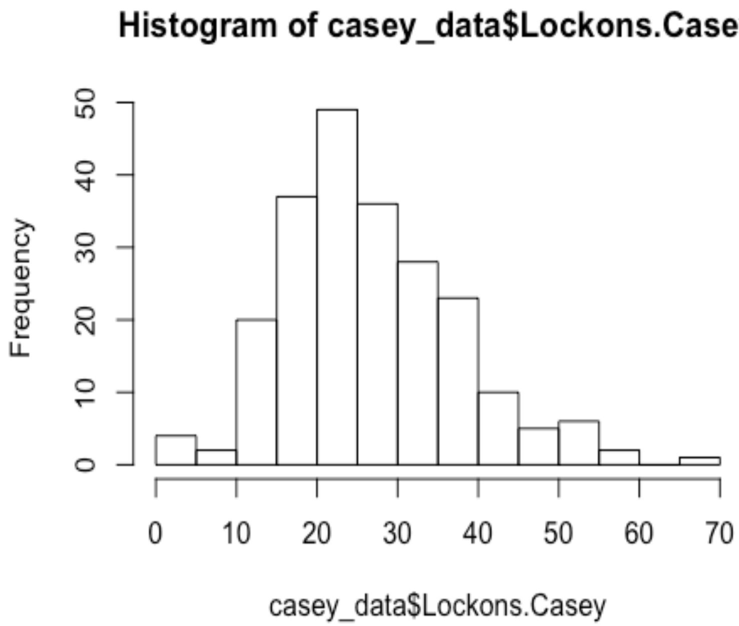
In order to predict them via linear regression but for that two assumptins must hold which is

1.Mean of the response variables has to be large enough to make it to normal distribution but it is not true in both cases of the whales.

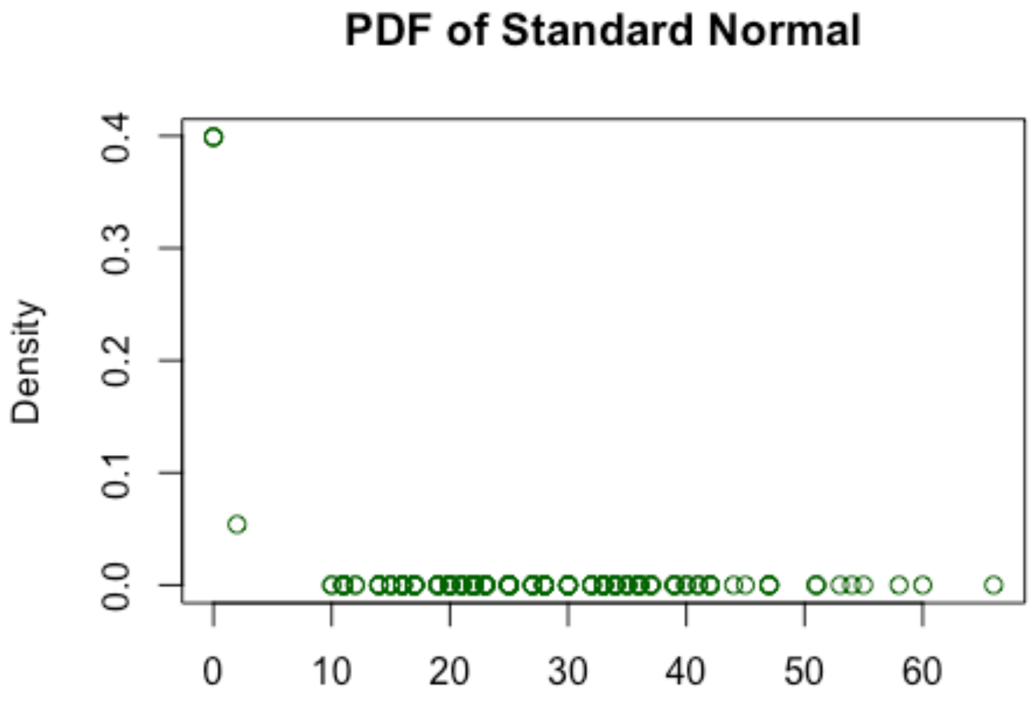
2.Also the pdf function of the response variable with transformation also doesnt follow normal ditribution hence we cant assume the distribution of lockons variable to follow linear regression

The attempts to see whether the probability mass function of lockons or any of its transformations follow normal distribution.

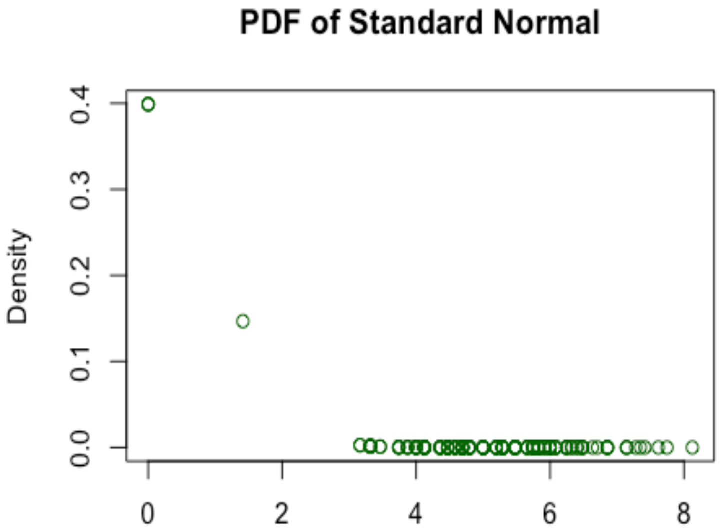
```
hist(casey_data$Lockons.Casey,breaks = 10)
```



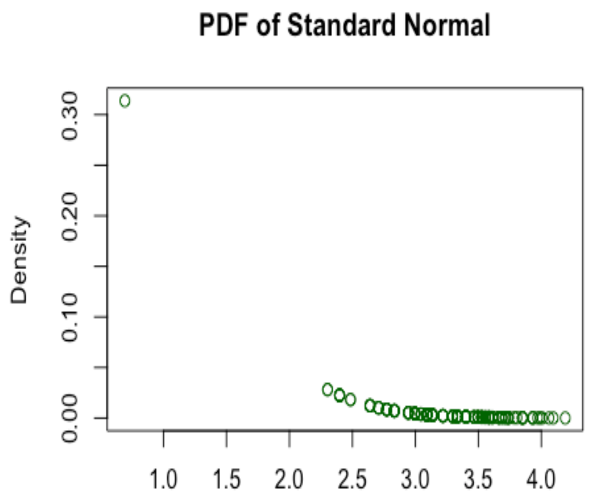
```
plot(casey_data$Lockons.Casey,dnorm(casey_data$Lockons.Casey),col="darkgreen"
,xlab="", ylab="Density", main="PDF of Standard Normal")
```



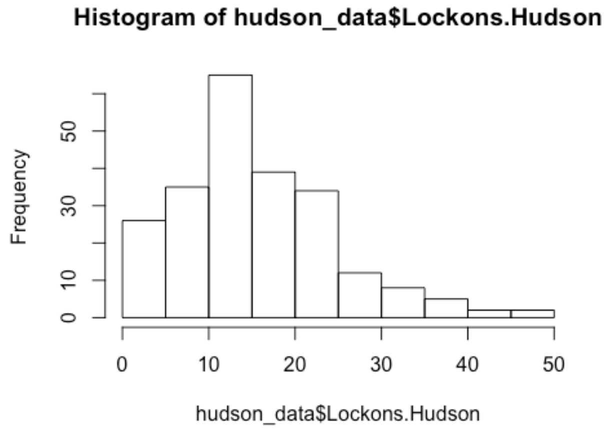
```
plot(sqrt(casey_data$Lockons.Casey),dnorm(sqrt(casey_data$Lockons.Casey)),col
="darkgreen",xlab="", ylab="Density", main="PDF of Standard Normal")
```



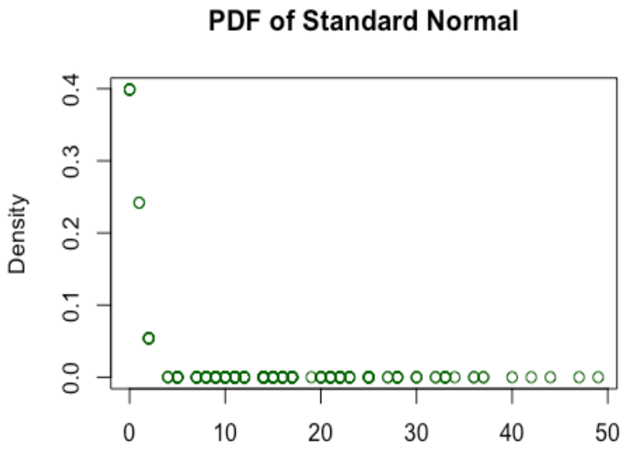
```r
plot(log(casey_data$Lockons.Casey),dnorm(log(casey_data$Lockons.Casey)),col="darkgreen",xlab="", ylab="Density", main="PDF of Standard Normal")
```



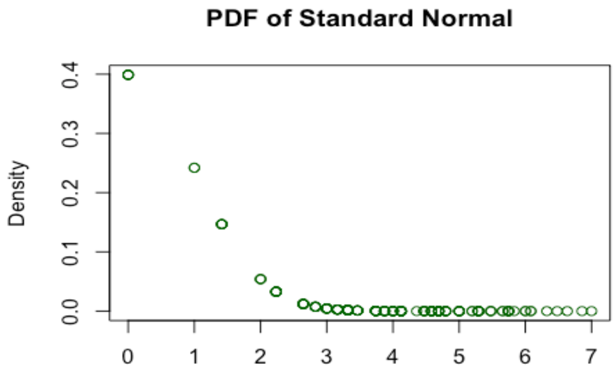
```r
hist(hudson_data$Lockons.Hudson)
```



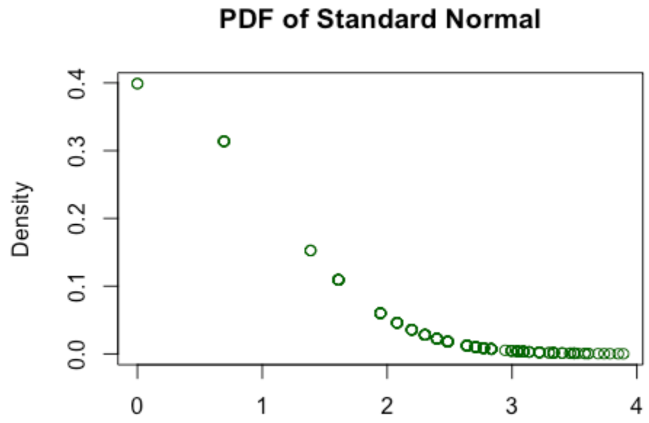
```r
plot(hudson_data$Lockons.Hudson,dnorm(hudson_data$Lockons.Hudson),col="darkgreen",xlab="", ylab="Density", main="PDF of Standard Normal")
```



```r
plot(sqrt(hudson_data$Lockons.Hudson),dnorm(sqrt(hudson_data$Lockons.Hudson)),col="darkgreen",xlab="", ylab="Density", main="PDF of Standard Normal")
```



```r
plot(log(hudson_data$Lockons.Hudson),dnorm(log(hudson_data$Lockons.Hudson)),col="darkgreen",xlab="", ylab="Density", main="PDF of Standard Normal")
```



Here we see that pdf of the lockons doesnt follow a normal distribution, hence linear model goes out of question.

Hence we can try Poisson model fitting for both the whales and if we see a case of overdispersion we may opt for Negative Binomial Distribution

3.For each data set, construct the regression model you have proposed in (2) abovefor the number of lockons in each period as a function of time, number of nursingbouts, and time of the day. Interpret your results.

Poisson Fitting for Casey

```
casey_poisson <- glm(casey_data$Lockons.Casey ~ casey_data$Bouts.Casey+casey_
data$Daytime.Casey+casey_data$Period.Casey , family = "poisson")
summary(casey_poisson)
```

```
##
## Call:
## glm(formula = casey_data$Lockons.Casey ~ casey_data$Bouts.Casey +
##     casey_data$Daytime.Casey + casey_data$Period.Casey, family = "poisson"
)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6903  -1.0117  -0.0307   0.9982   5.0066
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.3724816  0.0534981  44.347  < 2e-16 ***
## casey_data$Bouts.Casey     0.0919255  0.0038966  23.591  < 2e-16 ***
## casey_data$Daytime.Casey0  0.0265118  0.0259256   1.023    0.306
## casey_data$Period.Casey   -0.0011639  0.0002056  -5.661  1.5e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1061.86  on 222  degrees of freedom
## Residual deviance:  461.15  on 219  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 1586.4
##
## Number of Fisher Scoring iterations: 4
```

Interpretation

The regression coefficient of no of bouts means that the expected log(number of interlocks) for an increase in bouts by one is 0.0919255 and so the ratio of number of interlocks with number of bout x+1 and with number of bouts x is exp(0.0919255)= 1.09 i.e, 9%

The regression coefficient of daytime means that (keeping daytime as the baseline) in night time the number of interlocks will be higher by two percent than in daytime exp(0.0265118)= 1.02 i.e, 2%

The regression coefficient of period means that the expected log(number of interlocks) for an increase in period by one is -0.0011639 and so the ratio of number of interlocks with period x+1 and with period x is exp(-0.0011639)= 0.9988368.Hence explaining that as the period increases the number of interlocks decreases.

The residual deviance is much higher than the degrees of freedom hence proving that the model isnt a good fit and we might have some problem

Also we see that the daytime variable is not significant hence we need to modify our model by removing it.

```
casey_poisson_improved <- glm(casey_data$Lockons.Casey ~ casey_data$Bouts.Cas
ey+casey_data$Period.Casey, family = "poisson")
summary(casey_poisson_improved)
```

```
##
## Call:
## glm(formula = casey_data$Lockons.Casey ~ casey_data$Bouts.Casey +
##     casey_data$Period.Casey, family = "poisson")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.6469  -0.9800   0.0097   0.9391   4.9250
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.3804211  0.0529294  44.973  < 2e-16 ***
## casey_data$Bouts.Casey   0.0923997  0.0038692  23.881  < 2e-16 ***
## casey_data$Period.Casey -0.0011618  0.0002056  -5.651  1.6e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1061.9  on 222  degrees of freedom
## Residual deviance:  462.2  on 220  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 1585.4
##
## Number of Fisher Scoring iterations: 4
```

Although even removing it the problem of high Residual deviance is not solved also the model has the lower AIC but not by much.

Poisson Fitting for Hudson

```
hudson_poisson <- glm(hudson_data$Lockons.Hudson ~ hudson_data$Bouts.Hudson+h
udson_data$Daytime.Hudson+hudson_data$Period.Hudson ,   family = "poisson")
summary(hudson_poisson)
```

```
##
## Call:
## glm(formula = hudson_data$Lockons.Hudson ~ hudson_data$Bouts.Hudson +
##     hudson_data$Daytime.Hudson + hudson_data$Period.Hudson, family = "pois
son")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.0553  -1.2772  -0.3084   0.7820   5.0177
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.6680612  0.0607676  27.450  < 2e-16 ***
## hudson_data$Bouts.Hudson    0.1222453  0.0048531  25.189  < 2e-16 ***
## hudson_data$Daytime.Hudson0 0.0673002  0.0333983   2.015  0.04390 *
## hudson_data$Period.Hudson  -0.0008701  0.0002492  -3.492  0.00048 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1213.00  on 227  degrees of freedom
## Residual deviance:  571.65  on 224  degrees of freedom
## AIC: 1575.4
##
## Number of Fisher Scoring iterations: 4
```

Interpretation

The regression coefficient of no of bouts means that the expected log(number of interlocks) for an increase in bouts by one is 0.1222453 and so the ratio of number of interlocks with number of bout x+1 and with number of bouts x is exp(0.1222453)= 1.13 i.e, 13%

The regression coefficient of daytime means that (keeping daytime as the baseline) in night time the number of interlocks will be higher by two percent than in daytime exp(0.0265118)= 1.07 i.e, 7%

The regression coefficient of period means that the expected log(number of interlocks) for an increase in period by one is -0.0008701 and so the ratio of number of interlocks with period x+1 and with period x is exp(-0.0008701)= 0.9991303.Hence explaining that as the period increases the number of interlocks decreases.

The residual deviance is much higher than the degrees of freedom hence prving that the model isnt a good fit and we might have some problem

Also we see that all the variables are significant in the case of Hudson whales.

4.Specifically answer following questions for each data set.

1. Do the variables provide predictive power? Justify.

Yes the no of bouts and time period provides a predictive power but daytime categorical variable proves to be insignificant in the case of Hudson whale.On a general sense it gives us conclusions like:

1) Casey whale might have more interlocks in night time.
2) The chances of interlocking increases with number of bouts and decreases as time passes by(but not that much about 1%).

4.2. How would you interpret the coefficient for time period for the model?

For Casey whale,

The regression coefficient of period means that the expected log(number of interlocks) for an increase in period by one is -0.0011639 and so the ratio of number of interlocks with period x+1 and with period x is exp(-0.0011639)= 0.9988368.Hence explaining that as the period increases then the number of interlocks decreases but not by a significant amount.

For Hudson whale,

The regression coefficient of period means that the expected log(number of interlocks) for an increase in period by one is -0.0008701 and so the ratio of number of interlocks with period x+1 and with period x is exp(-0.0008701)= 0.9991303.Hence explaining that as the period increases the number of interlocks decreases but not by a significant amount.

And comparing the change between Hudson and Casey , Casey seems to have more effect than Hudson on the number of interlocks with passing time.

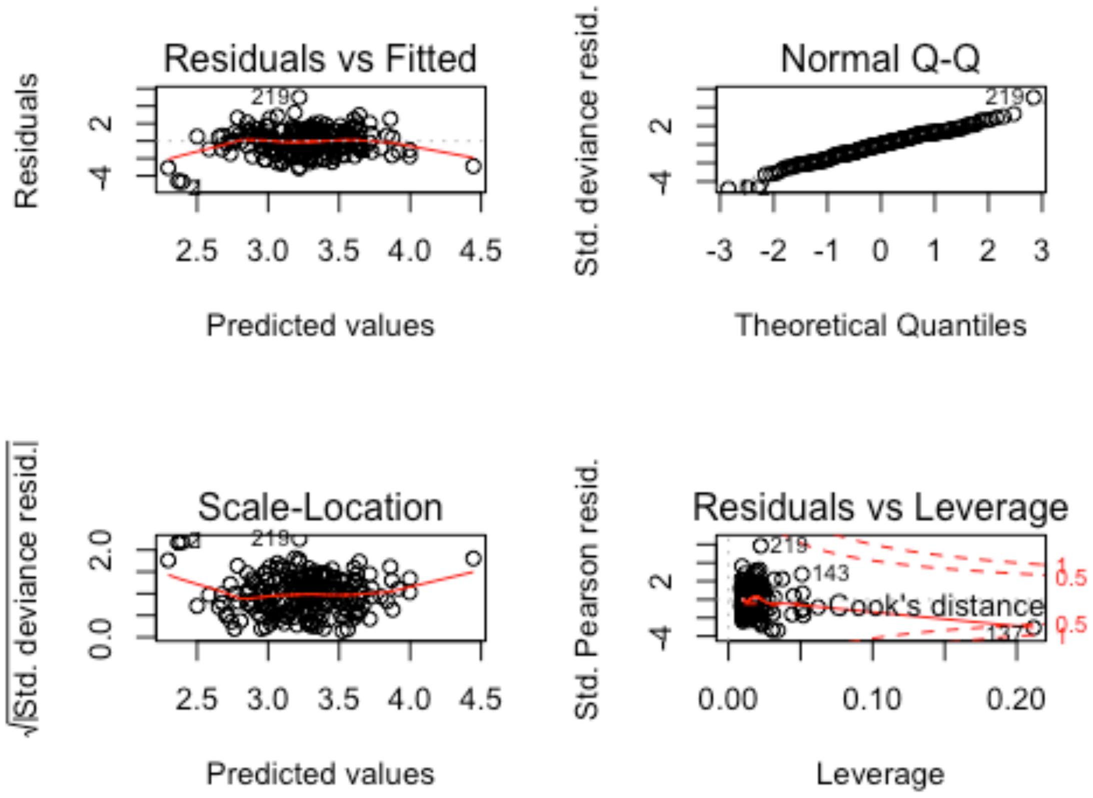
4.3 Does this model suffer from over dispersion? Justify with appropriate analysis.If the selected model suffers from over dispersion, propose and fit an alternativemodel which will take care of over dispersion.

CaseyWhales

In case of Casey whales the residual deviance(462) is higher almost double than the degrees of freedom(219) showing some level of overdispersion.

Also when we see the diagonstic plots we get the following results.

```
par(mfrow=c(2,2))
plot(casey_poisson)
```



```
par(mfrow=c(1,1))
```

By seeing the plots,

We can say that there is not much overdispersion but a little amount is present as,

-In the Residual vs Fitted we see them a little centric towards the middle, we wanted them to be random -The QQ Plot is quite a straight line except for some observations, especially

219 but we cant really say its normal -the Scale location plot & the ResidualvsLeverage plots is fine but with certain influential observations

In order to check the dispersion we can perform the overdispersion test.

```
#install.packages('AER')
library(AER)

## Loading required package: car

## Loading required package: lmtest

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival

dispersiontest(casey_poisson,trafo=2)

##
##  Overdispersion test
##
## data:  casey_poisson
## z = 4.1182, p-value = 1.909e-05
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##      alpha
## 0.03044849
```

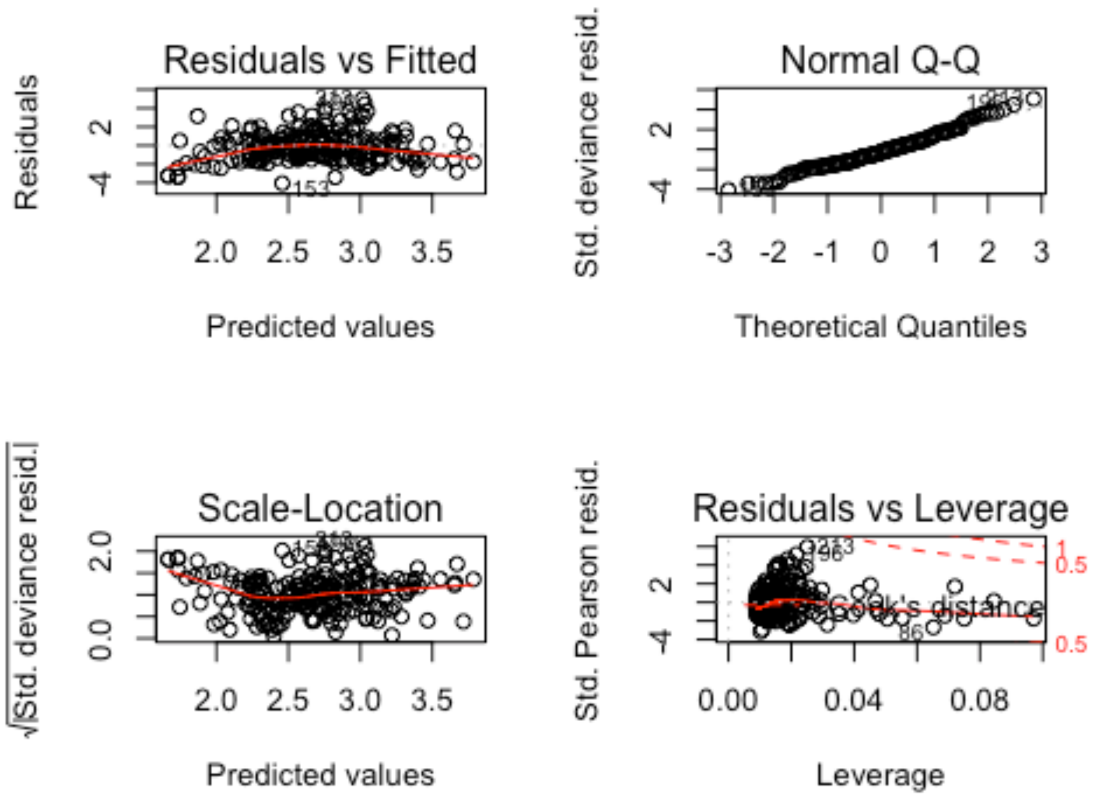
Hence we can say that the data suffers from overdispersion.

Hudson Whales

In case of Hudson whales the residual deviance(551) is higher almost double than the degrees of freedom(224) showing some level of overdispersion.

Also when we see the diagonstic plots we get the following results.

```
par(mfrow=c(2,2))
plot(hudson_poisson)
```



```
par(mfrow=c(1,1))
```

The Hudson whales also have the similar results but the dispersion is more in this case.

In order to check the dispersion we can perform the overdispersion test.

```
dispersiontest(hudson_poisson,trafo=2)
```

```
##
##  Overdispersion test
##
## data:  hudson_poisson
## z = 5.4108, p-value = 3.138e-08
## alternative hypothesis: true alpha is greater than 0
## sample estimates:
##      alpha
## 0.08667107
```

Both the datasets suffer from overdispersion and a Negative Binomial model would fit better in this case.

4.4. Does this alternative model have different implications than the initial model youselected in (3) above? Is it a better fit to the data? Justify your answer with proper analysis

```
#install.packages('MASS')
library(MASS)
```

Casey Whales

```
casey_Neg_Bin <- glm.nb(casey_data$Lockons.Casey ~ casey_data$Bouts.Casey+cas
ey_data$Daytime.Casey+casey_data$Period.Casey)
summary(casey_Neg_Bin)
```

```
##
## Call:
## glm.nb(formula = casey_data$Lockons.Casey ~ casey_data$Bouts.Casey +
##     casey_data$Daytime.Casey + casey_data$Period.Casey, init.theta = 29.16
910176,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1688  -0.7408   0.0216   0.7229   3.3550
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.2862734  0.0761289  30.032  < 2e-16 ***
## casey_data$Bouts.Casey     0.0980853  0.0057166  17.158  < 2e-16 ***
## casey_data$Daytime.Casey0  0.0295272  0.0364269   0.811 0.417601
## casey_data$Period.Casey   -0.0010278  0.0002861  -3.593 0.000327 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(29.1691) family taken to be 1)
##
##     Null deviance: 583.74  on 222  degrees of freedom
## Residual deviance: 262.33  on 219  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 1532.5
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  29.17
##           Std. Err.:  6.11
##
##  2 x log-likelihood:  -1522.462
```

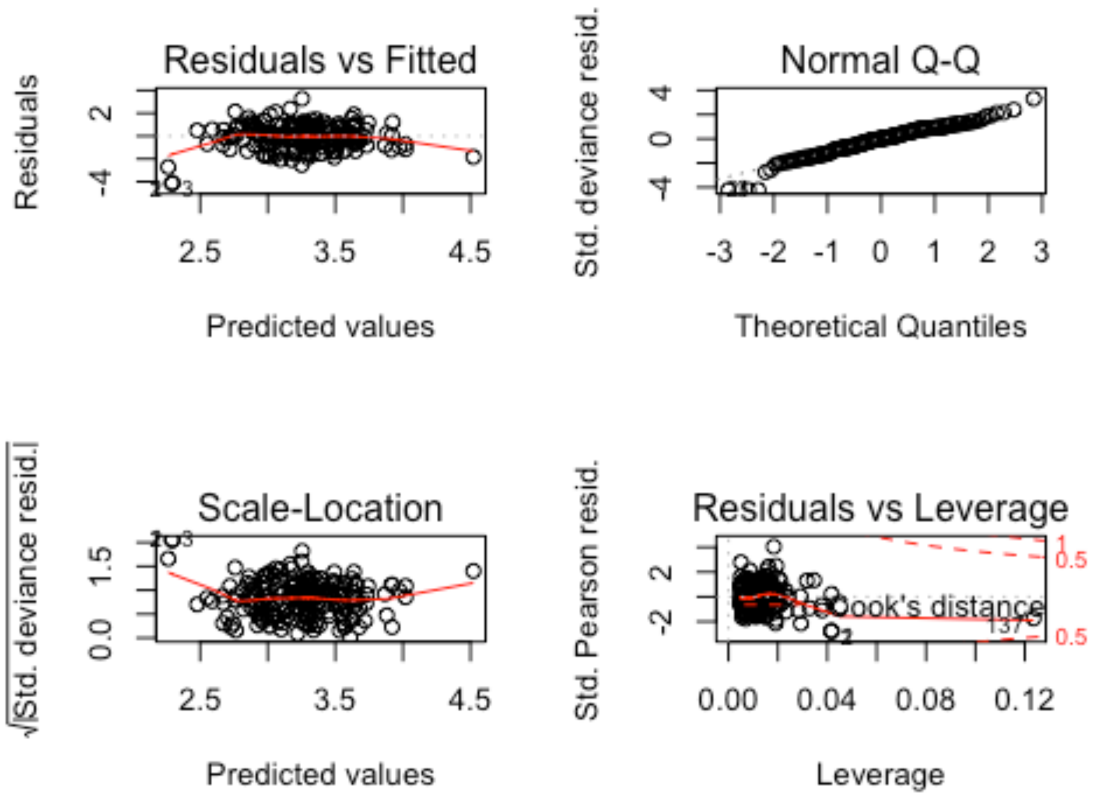
Since daytime is insignificant for Casey,we try to remove it and improve the model

```
casey_Neg_Bin <- glm.nb(casey_data$Lockons.Casey ~ casey_data$Bouts.Casey+cas
ey_data$Period.Casey)
summary(casey_Neg_Bin)

##
## Call:
## glm.nb(formula = casey_data$Lockons.Casey ~ casey_data$Bouts.Casey +
##     casey_data$Period.Casey, init.theta = 28.98066847, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.1285  -0.7325  -0.0003   0.7477   3.2760
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             2.2943353  0.0754379  30.414  < 2e-16 ***
## casey_data$Bouts.Casey   0.0986444  0.0056875  17.344  < 2e-16 ***
## casey_data$Period.Casey -0.0010215  0.0002865  -3.565 0.000364 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(28.9807) family taken to be 1)
##
##     Null deviance: 582.11  on 222  degrees of freedom
## Residual deviance: 262.31  on 220  degrees of freedom
##   (5 observations deleted due to missingness)
## AIC: 1531.1
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  28.98
##           Std. Err.:  6.05
##
##  2 x log-likelihood:  -1523.118
```

AIC doesnt change much but we can say it is bettr than the previous.

Now checking for the plots,

```
par(mfrow=c(2,2))
plot(casey_Neg_Bin)
```



```
par(mfrow=c(1,1))
```

We can clearly see that the residual deviance has reduced significantly and is clearly closer to the degrees of freedom and the diagonostc plots are more satisfactory hence it is better than following a Poisson distribution

Hudson Whales

```
hudson_Neg_Bin <- glm.nb(hudson_data$Lockons.Hudson ~ hudson_data$Bouts.Hudso
n+hudson_data$Daytime.Hudson+hudson_data$Period.Hudson)
summary(hudson_Neg_Bin)

##
## Call:
## glm.nb(formula = hudson_data$Lockons.Hudson ~ hudson_data$Bouts.Hudson +
##     hudson_data$Daytime.Hudson + hudson_data$Period.Hudson, init.theta = 1
1.04845616,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1784  -0.8407  -0.2124   0.5024   2.5333
##
## Coefficients:
```

```
##                               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)                  1.5301521  0.0931356  16.429   <2e-16 ***
## hudson_data$Bouts.Hudson     0.1350092  0.0078914  17.108   <2e-16 ***
## hudson_data$Daytime.Hudson0  0.0629402  0.0532722   1.181   0.2374    
## hudson_data$Period.Hudson   -0.0006999  0.0004032  -1.736   0.0826 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(11.0485) family taken to be 1)
## 
##     Null deviance: 540.62  on 227  degrees of freedom
## Residual deviance: 255.69  on 224  degrees of freedom
## AIC: 1453.1
## 
## Number of Fisher Scoring iterations: 1
## 
## 
##               Theta:  11.05 
##           Std. Err.:  1.85 
## 
##  2 x log-likelihood:  -1443.135
```

For Hudson whale, the daytime and Period variable are insignificant, hence we can remove it to get better results.

```
hudson_Neg_Bin <- glm.nb(hudson_data$Lockons.Hudson ~ hudson_data$Bouts.Hudso
n)
summary(hudson_Neg_Bin)
```

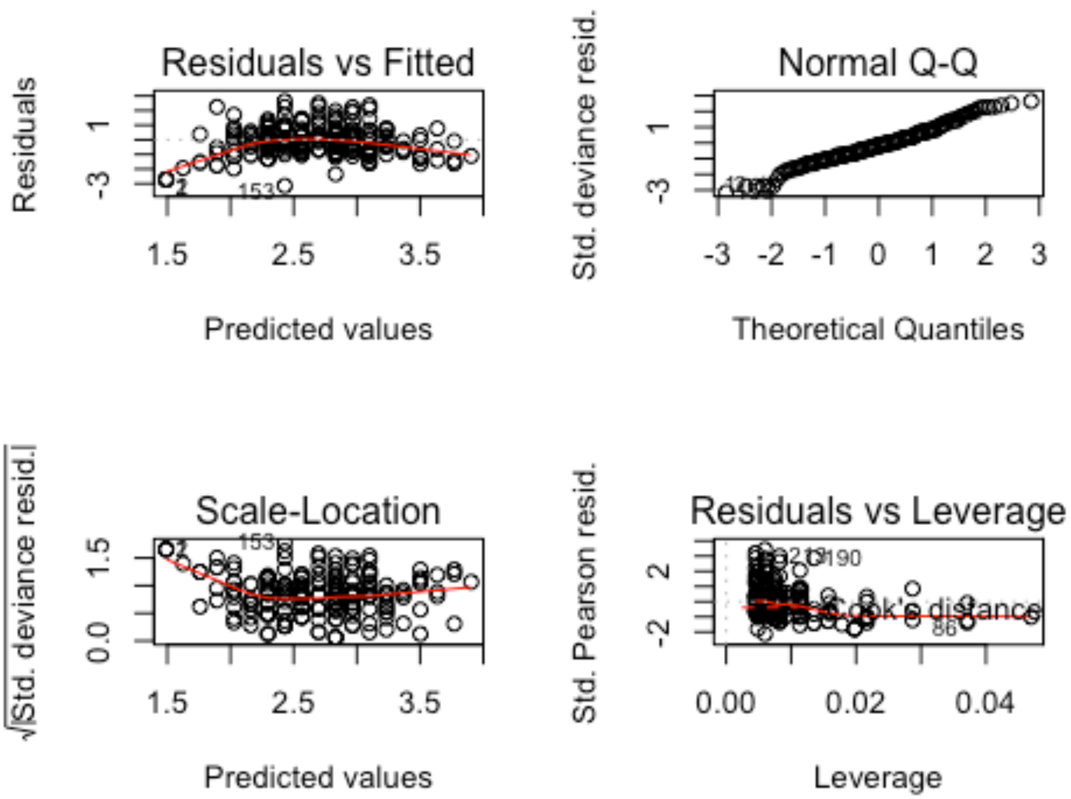
```
## 
## Call:
## glm.nb(formula = hudson_data$Lockons.Hudson ~ hudson_data$Bouts.Hudson, 
##     init.theta = 10.51771505, link = log)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -3.1488  -0.7930  -0.1810   0.4516   2.6361  
## 
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)    
## (Intercept)              1.491511   0.079567   18.75   <2e-16 ***
## hudson_data$Bouts.Hudson 0.134028   0.007961   16.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Negative Binomial(10.5177) family taken to be 1)
## 
##     Null deviance: 526.90  on 227  degrees of freedom
## Residual deviance: 253.91  on 226  degrees of freedom
## AIC: 1453.6
```

```
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  10.52
##           Std. Err.:  1.72
##
##  2 x log-likelihood:  -1447.607
```

AIC doesnt change much but we can say it is bettr than the previous.

The plots,

```
par(mfrow=c(2,2))
plot(hudson_Neg_Bin)
```



```
par(mfrow=c(1,1))
```

We can clearly see that the residual deviance has reduced significantly and is clearly closer to the degrees of freedom and the diagonostc plots are more satisfactory hence it is better than following a Poisson distribution

5. Using the two models for two calves you have finally selected, make comparative statements on how the predictors are affecting the number of lockons.

For the Casey whale, the predictors that are significant are the number of bouts and time period. The number of interlocks increases with number of bouts and slightly decreases/remains almost similar as time passes by.

For the Hudson whale, the predictors that is significant is just the number of bouts. The number of interlocks increases with the number of bouts with Hudson.

Survival Analytics

loading the neccesary libraries and the csv

```
library (survival)
luekemia_data <- read.csv("/Users/swapnilvermani/Downloads/leukemia.csv")
library(survminer)

## Loading required package: ggplot2

## Loading required package: ggpubr

## Loading required package: magrittr
```

Studying the summary and changing categorical variables

```
head(luekemia_data)

##   survival.times status sex logWBC Rx
## 1             35      0   1   1.45  0
## 2             34      0   1   1.47  0
## 3             32      0   1   2.20  0
## 4             32      0   1   2.53  0
## 5             25      0   1   1.78  0
## 6             23      1   1   2.57  0

summary(luekemia_data)

##  survival.times      status            sex              logWBC     
##  Min.   : 1.00   Min.   :0.0000   Min.   :0.0000   Min.   :1.450  
##  1st Qu.: 6.00   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:2.303  
##  Median :10.50   Median :1.0000   Median :0.0000   Median :2.800  
##  Mean   :12.88   Mean   :0.7143   Mean   :0.4762   Mean   :2.930  
##  3rd Qu.:18.50   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3.490  
##  Max.   :35.00   Max.   :1.0000   Max.   :1.0000   Max.   :5.000  
##        Rx     
##  Min.   :0.0  
##  1st Qu.:0.0  
##  Median :0.5  
##  Mean   :0.5  
##  3rd Qu.:1.0  
##  Max.   :1.0
```

```
luekemia_data$sex<-as.factor(luekemia_data$sex)
luekemia_data$Rx<-as.factor(luekemia_data$Rx)
```

1. Explore the data. What is the basic difference you are noticing between the two groups?

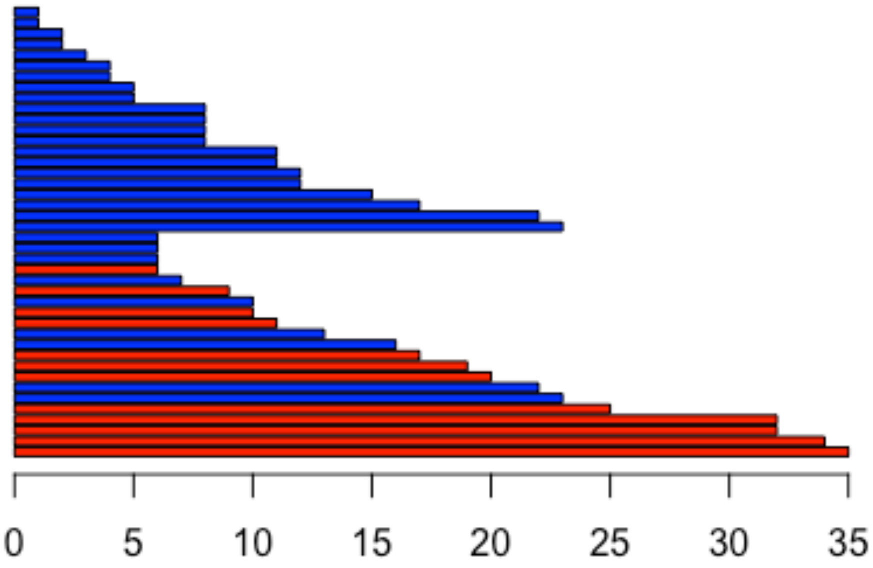
Let us try to plot the data with both censored and uncensored observations

```
luekemia_data
```

```
##    survival.times status sex logWBC Rx
## 1              35      0   1   1.45  0
## 2              34      0   1   1.47  0
## 3              32      0   1   2.20  0
## 4              32      0   1   2.53  0
## 5              25      0   1   1.78  0
## 6              23      1   1   2.57  0
## 7              22      1   1   2.32  0
## 8              20      0   1   2.01  0
## 9              19      0   0   2.05  0
## 10             17      0   0   2.16  0
## 11             16      1   1   3.60  0
## 12             13      1   0   2.88  0
## 13             11      0   0   2.60  0
## 14             10      0   0   2.70  0
## 15             10      1   0   2.96  0
## 16              9      0   0   2.80  0
## 17              7      1   0   4.43  0
## 18              6      0   0   3.20  0
## 19              6      1   0   2.31  0
## 20              6      1   1   4.06  0
## 21              6      1   0   3.28  0
## 22             23      1   1   1.97  1
## 23             22      1   0   2.73  1
## 24             17      1   0   2.95  1
## 25             15      1   0   2.30  1
## 26             12      1   0   1.50  1
## 27             12      1   0   3.06  1
## 28             11      1   0   3.49  1
## 29             11      1   0   2.12  1
## 30              8      1   0   3.52  1
## 31              8      1   0   3.05  1
## 32              8      1   0   2.32  1
## 33              8      1   1   3.26  1
## 34              5      1   1   3.49  1
## 35              5      1   0   3.97  1
## 36              4      1   1   4.36  1
## 37              4      1   1   2.42  1
## 38              3      1   1   4.01  1
## 39              2      1   1   4.91  1
## 40              2      1   1   4.48  1
## 41              1      1   1   2.80  1
## 42              1      1   1   5.00  1
```

```
barplot(luekemia_data$survival.times,type="h",horiz=TRUE,col=ifelse(luekemia_
data$status==0,"red","blue"))
```

```
## Warning in plot.window(xlim, ylim, log = log, ...): graphical parameter
## "type" is obsolete

## Warning in title(main = main, sub = sub, xlab = xlab, ylab = ylab, ...):
## graphical parameter "type" is obsolete

## Warning in axis(if (horiz) 1 else 2, cex.axis = cex.axis, ...): graphical
## parameter "type" is obsolete
```



Lets divide the data into the groups of new treatment and standard treatment.

```
luekemia_data_new <- luekemia_data[luekemia_data$Rx==0,]
luekemia_data_standard <- luekemia_data[luekemia_data$Rx==1,]

summary(luekemia_data_new)

##  survival.times     status        sex        logWBC       Rx  
##  Min.   : 6.0   Min.   :0.0000   0:11   Min.   :1.450   0:21  
##  1st Qu.: 9.0   1st Qu.:0.0000   1:10   1st Qu.:2.160   1: 0  
##  Median :16.0   Median :0.0000          Median :2.570         
##  Mean   :17.1   Mean   :0.4286          Mean   :2.636         
##  3rd Qu.:23.0   3rd Qu.:1.0000          3rd Qu.:2.960         
##  Max.   :35.0   Max.   :1.0000          Max.   :4.430         

summary(luekemia_data_standard)

##  survival.times       status  sex        logWBC       Rx  
##  Min.   : 1.000   Min.   :1   0:11   Min.   :1.500   0: 0  
##  1st Qu.: 4.000   1st Qu.:1   1:10   1st Qu.:2.420   1:21  
##  Median : 8.000   Median :1          Median :3.060         
##  Mean   : 8.667   Mean   :1          Mean   :3.224         
##  3rd Qu.:12.000   3rd Qu.:1          3rd Qu.:3.970         
##  Max.   :23.000   Max.   :1          Max.   :5.000         
```

Seeing the summary of the two groups the major differences we can figure out is as follows:

a) All the cases of standard treatment result in a failure and there is no censored data.
b) On an average with the new treatment, the remission period for the patient lies for around 17 months as compared to only 8 weeks in case of standard treatment. c)But also we have to keep in mind that the group which is given the new treatment has avg logWBC's 3.6 while the average logWBC's for the other is high.

2. Compute Kaplan-Meier estimate of survival function and Nelson-Allen estimates of cumulative hazard rate.

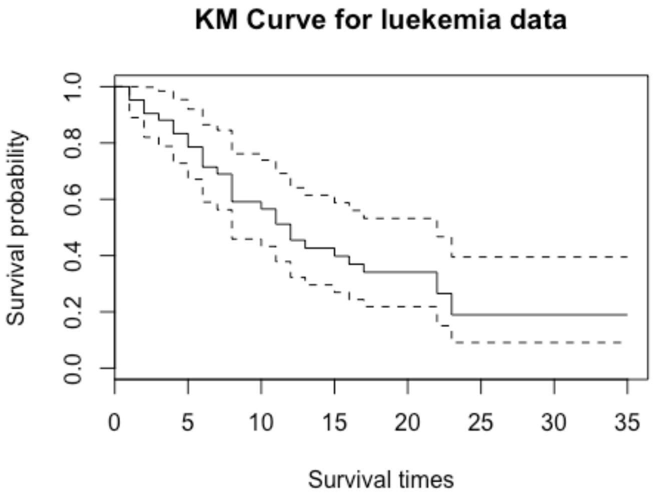
Kaplein Mier Estimates

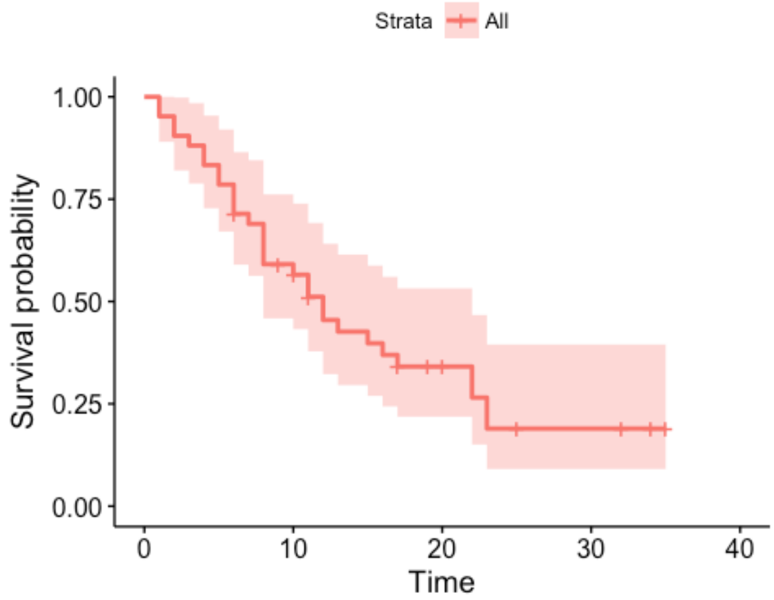
```
luekemia_data

##    survival.times status sex logWBC Rx
## 1              35      0   1   1.45  0
## 2              34      0   1   1.47  0
## 3              32      0   1   2.20  0
## 4              32      0   1   2.53  0
## 5              25      0   1   1.78  0
## 6              23      1   1   2.57  0
## 7              22      1   1   2.32  0
## 8              20      0   1   2.01  0
## 9              19      0   0   2.05  0
## 10             17      0   0   2.16  0
## 11             16      1   1   3.60  0
## 12             13      1   0   2.88  0
## 13             11      0   0   2.60  0
## 14             10      0   0   2.70  0
## 15             10      1   0   2.96  0
## 16              9      0   0   2.80  0
## 17              7      1   0   4.43  0
## 18              6      0   0   3.20  0
## 19              6      1   0   2.31  0
## 20              6      1   1   4.06  0
## 21              6      1   0   3.28  0
## 22             23      1   1   1.97  1
## 23             22      1   0   2.73  1
## 24             17      1   0   2.95  1
## 25             15      1   0   2.30  1
## 26             12      1   0   1.50  1
## 27             12      1   0   3.06  1
## 28             11      1   0   3.49  1
## 29             11      1   0   2.12  1
## 30              8      1   0   3.52  1
## 31              8      1   0   3.05  1
## 32              8      1   0   2.32  1
## 33              8      1   1   3.26  1
## 34              5      1   1   3.49  1
## 35              5      1   0   3.97  1
## 36              4      1   1   4.36  1
## 37              4      1   1   2.42  1
## 38              3      1   1   4.01  1
## 39              2      1   1   4.91  1
## 40              2      1   1   4.48  1
## 41              1      1   1   2.80  1
## 42              1      1   1   5.00  1
```

```
fit = survfit(Surv(luekemia_data$survival.times,luekemia_data$status)~1 )
summary(fit)
```

```
## Call: survfit(formula = Surv(luekemia_data$survival.times, luekemia_data$s
tatus) ~
##     1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     42       2    0.952  0.0329       0.8901        1.000
##     2     40       2    0.905  0.0453       0.8202        0.998
##     3     38       1    0.881  0.0500       0.7883        0.985
##     4     37       2    0.833  0.0575       0.7279        0.954
##     5     35       2    0.786  0.0633       0.6709        0.920
##     6     33       3    0.714  0.0697       0.5899        0.865
##     7     29       1    0.690  0.0715       0.5628        0.845
##     8     28       4    0.591  0.0764       0.4588        0.762
##    10     23       1    0.565  0.0773       0.4325        0.739
##    11     21       2    0.512  0.0788       0.3783        0.692
##     12      18         2     0.455  0.0796        0.3227         0.641
##     13      16         1     0.426  0.0795        0.2958         0.615
##     15      15         1     0.398  0.0791        0.2694         0.588
##     16      14         1     0.369  0.0784        0.2437         0.560
##     17      13         1     0.341  0.0774        0.2186         0.532
##     22       9         2     0.265  0.0765        0.1507         0.467
##     23       7         2     0.189  0.0710        0.0909         0.395
```

```
plot(fit, xlab="Survival times",ylab = "Survival probability",main="KM Curve
for luekemia data")
ggsurvplot(fit, data = luekemia_data)
```



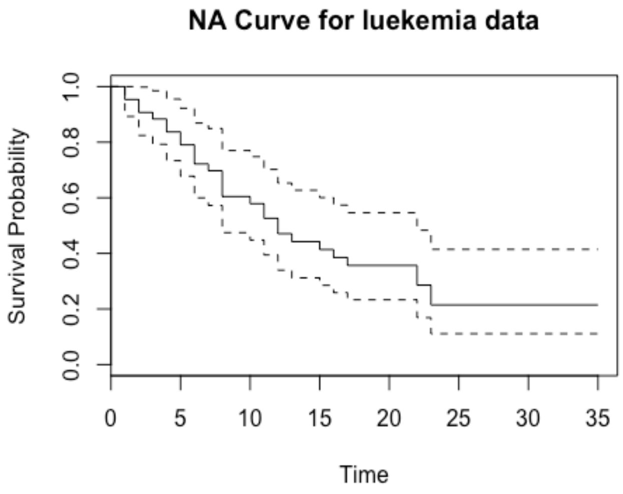


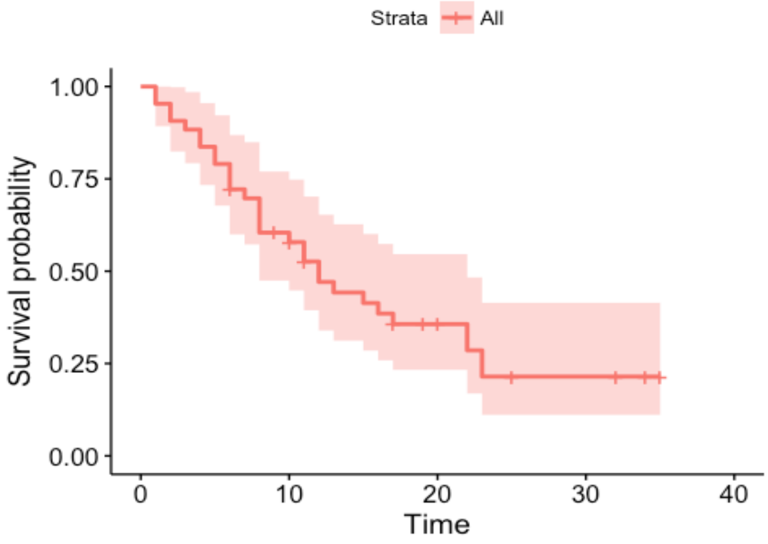
Nelson Aelen Estimates

```
NA_surv <- survfit(coxph(Surv(luekemia_data$survival.times,luekemia_data$stat
us)~1), type="aalen")
summary(NA_surv)
```

```
## Call: survfit(formula = coxph(Surv(luekemia_data$survival.times, luekemia_
data$status) ~
##     1), type = "aalen")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     42       2    0.953  0.0321        0.893        1.000
##     2     40       2    0.907  0.0443        0.824        0.998
##     3     38       1    0.883  0.0490        0.792        0.985
##     4     37       2    0.837  0.0564        0.733        0.955
##     5     35       2    0.790  0.0621        0.678        0.922
##     6     33       3    0.722  0.0682        0.600        0.869
##     7     29       1    0.697  0.0701        0.573        0.849
##     8     28       4    0.604  0.0746        0.475        0.770
##    10     23       1    0.579  0.0757        0.448        0.748
##    11     21       2    0.526  0.0774        0.394        0.702
##    12     18       2    0.471  0.0785        0.340        0.653
##    13     16       1    0.442  0.0788        0.312        0.627
##    15     15       1    0.414  0.0787        0.285        0.601
##    16     14       1    0.385  0.0783        0.259        0.574
##    17     13       1    0.357  0.0775        0.233        0.546
##    22      9       2    0.286  0.0766        0.169        0.483
##    23      7       2    0.215  0.0721        0.111        0.414
```

```
plot(NA_surv, xlab="Time", ylab="Survival Probability",main="NA Curve for lue
kemia data")
ggsurvplot(NA_surv, data = luekemia_data)
```



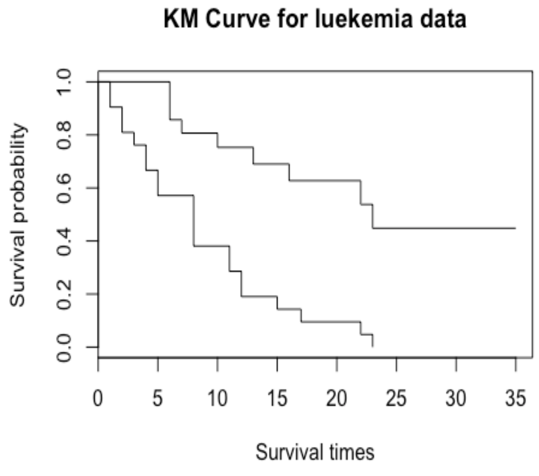


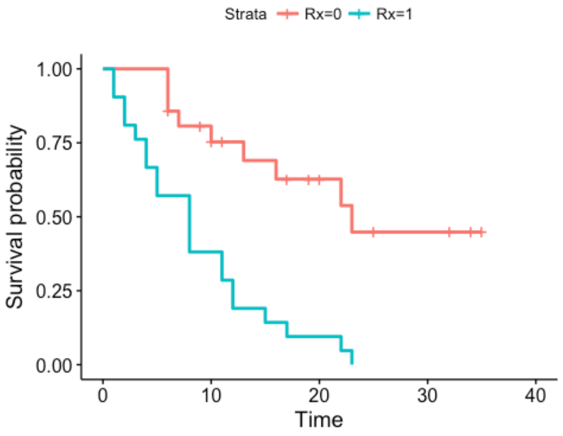
Here we can see that both the estimates are almost same.And can be interpreted as “the chances of the remission time to be say 10 weeks is approximately 50% for the leukemia patients”

3)Plot the KM estimate for the two groups with confidence intervals. Can you notice any major differences between two groups? Specify.

```
group_fit = survfit(Surv(luekemia_data$survival.times,luekemia_data$status)~l
uekemia_data$Rx )
summary(group_fit)

## Call: survfit(formula = Surv(luekemia_data$survival.times, luekemia_data$s
tatus) ~
##     luekemia_data$Rx)
##
##                 luekemia_data$Rx=0
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##      6      21        3     0.857  0.0764        0.720        1.000
##      7      17        1     0.807  0.0869        0.653        0.996
##     10      15        1     0.753  0.0963        0.586        0.968
##     13      12        1     0.690  0.1068        0.510        0.935
##     16      11        1     0.627  0.1141        0.439        0.896
##     22       7        1     0.538  0.1282        0.337        0.858
##     23       6        1     0.448  0.1346        0.249        0.807
##
##                 luekemia_data$Rx=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##      1      21        2   0.9048  0.0641      0.78754        1.000
##      2      19        2   0.8095  0.0857      0.65785        0.996
##      3      17        1   0.7619  0.0929      0.59988        0.968
##      4      16        2   0.6667  0.1029      0.49268        0.902
##      5      14        2   0.5714  0.1080      0.39455        0.828
##      8      12        4   0.3810  0.1060      0.22085        0.657
##     11       8        2   0.2857  0.0986      0.14529        0.562
##     12       6        2   0.1905  0.0857      0.07887        0.460
##     15       4        1   0.1429  0.0764      0.05011        0.407
##     17       3        1   0.0952  0.0641      0.02549        0.356
##     22       2        1   0.0476  0.0465      0.00703        0.322
##     23       1        1   0.0000     NaN           NA           NA
```

```
plot(group_fit, xlab="Survival times",ylab = "Survival probability",main="KM
Curve for luekemia data")
ggsurvplot(group_fit, data = luekemia_data)
```





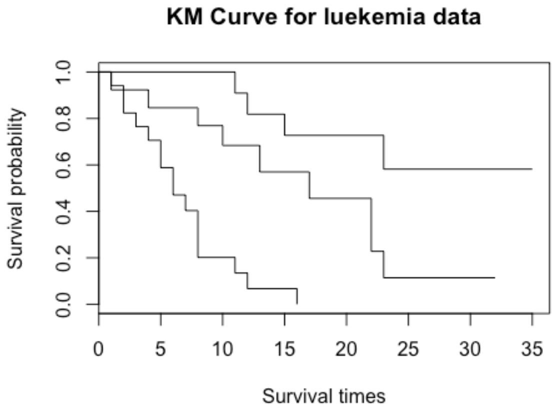
We can say that the survival probability for a larger time like 30 weeks is more with the new treatment as compared to the standard treatment proving new treatment to be effective.

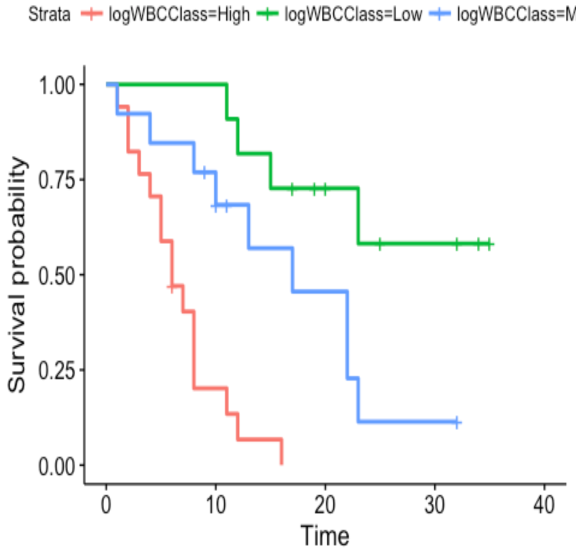
d)Suppose we wish to compare KM estimates given the variable logWBC, for which we categorize logWBC into 3 classes—low, medium, high—as follows:

```
#luekemia_data$logWBCClass <- factor(ordered = TRUE,levels=c("Low,Medium,High
"))
luekemia_data$logWBCClass[luekemia_data$logWBC<=2.30 & luekemia_data$logWBC>0
]<-'Low'
luekemia_data$logWBCClass[luekemia_data$logWBC>2.31 & luekemia_data$logWBC<=3
.0]<-'Medium'
luekemia_data$logWBCClass[luekemia_data$logWBC>3.0]<-'High'


group_fit = survfit(Surv(luekemia_data$survival.times,luekemia_data$status)~l
uekemia_data$logWBCClass )
summary(group_fit)
```

```
## Call: survfit(formula = Surv(luekemia_data$survival.times, luekemia_data$s
tatus) ~
##     luekemia_data$logWBCClass)
##
## 1 observation deleted due to missingness
##                 luekemia_data$logWBCClass=High
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     17       1   0.9412  0.0571       0.8357        1.000
##     2     16       2   0.8235  0.0925       0.6609        1.000
##     3     14       1   0.7647  0.1029       0.5875        0.995
##     4     13       1   0.7059  0.1105       0.5194        0.959
##     5     12       2   0.5882  0.1194       0.3952        0.876
##     6     10       2   0.4706  0.1211       0.2842        0.779
##     7      7       1   0.4034  0.1210       0.2241        0.726
##     8      6       3   0.2017  0.1022       0.0747        0.544
##    11      3       1   0.1345  0.0875       0.0376        0.481
##    12      2       1   0.0672  0.0646       0.0102        0.442
##    16      1       1   0.0000     NaN           NA           NA
##
##                 luekemia_data$logWBCClass=Low
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    11     11       1    0.909  0.0867        0.754            1
##    12     10       1    0.818  0.1163        0.619            1
##    15      9       1    0.727  0.1343        0.506            1
##    23      5       1    0.582  0.1687        0.330            1
##
##                 luekemia_data$logWBCClass=Medium
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     13       1    0.923  0.0739       0.7890        1.000
##     4     12       1    0.846  0.1001       0.6711        1.000
##     8     11       1    0.769  0.1169       0.5711        1.000
##    10      9       1    0.684  0.1315       0.4691        0.997
##    13      6       1    0.570  0.1511       0.3389        0.958
##    17      5       1    0.456  0.1581       0.2310        0.900
##     22       4        2     0.228  0.1387       0.0692         0.751
##     23       2        1     0.114  0.1063       0.0183         0.709
```

```
plot(group_fit, xlab="Survival times",ylab = "Survival probability",main="KM
Curve for luekemia data")
ggsurvplot(group_fit, data = luekemia_data)
```



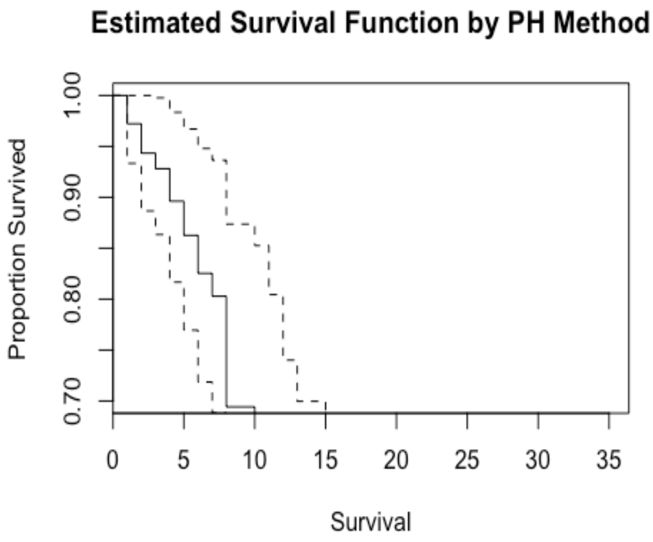


We can say that the survival probability for a larger time like 30 weeks is more with the people with low logWBC's and to be least with high logWBC's.

e)Fit the Cox PH model that can be used to assess the relationship of interest, which considers the potential confounders Sex and logWBC.

```
CPH <- coxph(Surv(luekemia_data$survival.times,luekemia_data$status) ~ luekem
ia_data$sex + luekemia_data$logWBCClass)
summary(CPH)
```

```
## Call:
## coxph(formula = Surv(luekemia_data$survival.times, luekemia_data$status) ~ 
##     luekemia_data$sex + luekemia_data$logWBCClass)
## 
##   n= 41, number of events= 29 
##    (1 observation deleted due to missingness)
## 
##                                   coef exp(coef) se(coef)      z
## luekemia_data$sex1            -0.16129   0.85104  0.41814 -0.386
## luekemia_data$logWBCClassLow    -2.81398   0.05997  0.64363 -4.372
## luekemia_data$logWBCClassMedium -1.62066   0.19777  0.48789 -3.322
##                               Pr(>|z|)    
## luekemia_data$sex1            0.699693    
## luekemia_data$logWBCClassLow    1.23e-05 ***
## luekemia_data$logWBCClassMedium 0.000894 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                 exp(coef) exp(-coef) lower .95 upper .95
## luekemia_data$sex1                0.85104      1.175   0.37500    1.9314
## luekemia_data$logWBCClassLow      0.05997     16.676   0.01698    0.2117
## luekemia_data$logWBCClassMedium   0.19777      5.056   0.07601    0.5146
##
## Concordance= 0.738  (se = 0.062 )
## Rsquare= 0.474   (max possible= 0.987 )
## Likelihood ratio test= 26.36  on 3 df,   p=7.998e-06
## Wald test            = 21.62  on 3 df,   p=7.842e-05
## Score (logrank) test = 28.67  on 3 df,   p=2.625e-06

plot(survfit(CPH), ylim=c(0.7, 1), xlab="Survival", ylab="Proportion Survived
",
main="Estimated Survival Function by PH Method")
```



```
#ggforest(CPH, data = luekemia_data$survival.times)
```

6. Compare the fits using KM and Cox PH plots for estimated survival function.

The KM and the Cox PH curves does not seem to be parallel and hence violates the PH assumption.We can say that the covariates are not staying the same in due course of time.

Missing Data Question

1.Carry out summary statistics and identify the missing values, if any.

```
class_data<-read.csv("/Users/swapnilvermani/Downloads/class_data.csv")

class_data
```

```
##    Age Weight Sex Height
## 1   19    180   0     61
## 2   19    160   0     70
## 3   19    135   0     70
## 4   19    195   0     71
## 5   19    130   1     64
## 6   19    120   1     64
## 7   21    135   1     69
## 8   19    125   0     67
## 9   19    120   1     62
## 10  20    145   0     66
## 11  19    155   0     65
## 12  19    135   1     69
## 13  19    140   0     66
## 14  NA    120   1     63
## 15  19    140   0     69
## 16  18    113   1     66
## 17  18    180   0     68
## 18  19    175   0     72
## 19  19    169   0     70
## 20  19    210   0     74
## 21  20    104   1     66
## 22  20    105   1     64
## 23  20    125   1     65
## 24  20    120   1     71
## 25  19    119   1     69
## 26  NA    140   1     64
## 27  20    185   1     67
## 28  19    110   1     60
## 29  20    120   1     66
## 30  19    175   0     71
## 31  19    135   1     65
## 32  19    120   0     70
## 33  21     NA   0     69
## 34  20    108   1     63
## 35  19    118   1     63
## 36  20    135   0     72
## 37  19    169   0     73
## 38  19    145   0     69
## 39  27    130   1     69
## 40  18    135   0     64
## 41  20    115   1     61
## 42  19    140   0     68
## 43  21    152   0     70
## 44  19    118   1     64
## 45  19    112   1     62
## 46  19    100   1     64
## 47  20    135   1     67
## 48  20    110   1     63
## 49  20     NA   0     68
## 50  18    115   1     63
## 51  19    145   0     68
## 52  19    115   1     65
## 53  19    128   1     63
## 54  20     NA   1     68
## 55  19    130   0     69
## 56  19    165   0     69
## 57  19    130   0     69
## 58  20    180   0     70
## 59  28    110   1     65
## 60  19    155   0     55
```

```
summary(class_data)
```

```
##       Age            Weight            Sex              Height     
##  Min.   :18.00   Min.   :100.0   Min.   :0.0000   Min.   :55.00  
##  1st Qu.:19.00   1st Qu.:119.0   1st Qu.:0.0000   1st Qu.:64.00  
##  Median :19.00   Median :135.0   Median :1.0000   Median :67.00  
##  Mean   :19.59   Mean   :137.5   Mean   :0.5167   Mean   :66.62  
##  3rd Qu.:20.00   3rd Qu.:152.0   3rd Qu.:1.0000   3rd Qu.:69.00  
##  Max.   :28.00   Max.   :210.0   Max.   :1.0000   Max.   :74.00  
##  NA's   :2       NA's   :3
```

```
library(mice)
```

```
## Loading required package: lattice
```

```
md.pattern(class_data)
```

```
##    Sex Height Age Weight  
## 55   1      1   1      1 0
##  2   1      1   0      1 1
##  3   1      1   1      0 1
##      0      0   2      3 5
```

As we can clearly see in the summary statistics the no of NA are less 2 in age and 3 in weight which is the response variable but it could affect our model.

2.Fit a linear regression model to the data with missing values. Interpret your results.

-Changing into categorical variable

```
class_data$Sex<-as.factor(class_data$Sex)
class_data
```

```
##    Age Weight Sex Height
## 1   19    180   0     61
## 2   19    160   0     70
## 3   19    135   0     70
## 4   19    195   0     71
## 5   19    130   1     64
## 6   19    120   1     64
## 7   21    135   1     69
## 8   19    125   0     67
## 9   19    120   1     62
## 10  20    145   0     66
## 11  19    155   0     65
## 12  19    135   1     69
## 13  19    140   0     66
## 14  NA    120   1     63
## 15  19    140   0     69
## 16  18    113   1     66
## 17  18    180   0     68
## 18  19    175   0     72
## 19  19    169   0     70
## 20  19    210   0     74
## 21  20    104   1     66
## 22  20    105   1     64
## 23  20    125   1     65
## 24  20    120   1     71
## 25  19    119   1     69
## 26  NA    140   1     64
## 27  20    185   1     67
## 28  19    110   1     60
## 29  20    120   1     66
## 30  19    175   0     71
## 31  19    135   1     65
## 32  19    120   0     70
## 33  21     NA   0     69
## 34  20    108   1     63
## 35  19    118   1     63
## 36  20    135   0     72
## 37  19    169   0     73
## 38  19    145   0     69
## 39  27    130   1     69
## 40  18    135   0     64
## 41  20    115   1     61
## 42  19    140   0     68
## 43  21    152   0     70
## 44  19    118   1     64
## 45  19    112   1     62
## 46  19    100   1     64
## 47  20    135   1     67
## 48  20    110   1     63
## 49  20     NA   0     68
## 50  18    115   1     63
## 51  19    145   0     68
## 52  19    115   1     65
## 53  19    128   1     63
## 54  20     NA   1     68
## 55  19    130   0     69
## 56  19    165   0     69
## 57  19    130   0     69
## 58  20    180   0     70
## 59  28    110   1     65
## 60  19    155   0     55
```

Fitting the linear model on the data,

```
class_model<-lm(Weight~Age+Sex+Height, data=class_data)
summary(class_model)

##
## Call:
## lm(formula = Weight ~ Age + Sex + Height, data = class_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.741 -11.038  -2.756   9.271  60.600
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  52.8982    58.1504   0.910   0.3673
## Age          -0.5416     1.6309  -0.332   0.7412
## Sex1        -27.8643     6.1462  -4.534 3.54e-05 ***
## Height        1.6448     0.8093   2.032   0.0474 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.18 on 51 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.4804, Adjusted R-squared:  0.4498
## F-statistic: 15.72 on 3 and 51 DF,  p-value: 2.297e-07
```

Here age is seeming to be an insignificant variable.

3.Impute the missing values by multiple imputation techniques using MICE.

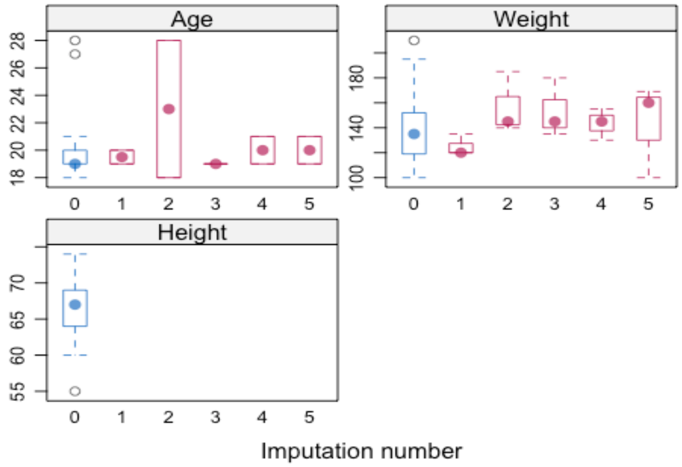
Give summary of the imputed data.

```
class_data_imp<-mice(class_data,m=5,maxit=2,meth="pmm",seed=500)

##
##  iter imp variable
##   1   1  Age  Weight
##   1   2  Age  Weight
##   1   3  Age  Weight
##   1   4  Age  Weight
##   1   5  Age  Weight
##   2   1  Age  Weight
##   2   2  Age  Weight
##   2   3  Age  Weight
##   2   4  Age  Weight
##   2   5  Age  Weight
```

```
summary(class_data_imp)
```

```
## Multiply imputed data set
## Call:
## mice(data = class_data, m = 5, method = "pmm", maxit = 2, seed = 500)
## Number of multiple imputations:  5
## Missing cells per column:
##     Age Weight    Sex Height
##       2      3      0      0
## Imputation methods:
##     Age Weight    Sex Height
##   "pmm"  "pmm"  "pmm"  "pmm"
## VisitSequence:
##     Age Weight
##       1      2
## PredictorMatrix:
##         Age Weight Sex Height
## Age       0      1   1      1
## Weight    1      0   1      1
## Sex       0      0   0      0
## Height    0      0   0      0
## Random generator seed value:  500
```

```
bwplot(class_data_imp)
```



```
class_data_imp$imp$Age
```

```
##     1  2  3  4  5
## 14 19 28 19 21 19
## 26 20 18 19 19 21
```

```
class_data_imp$imp$Weight
```

```
##      1   2   3   4   5
## 33 120 140 145 130 169
## 49 120 145 180 145 160
## 54 135 185 135 155 100

class_data_imputed <- complete(class_data_imp)
```

Here we can clearly see that the five sets of imputed data for 2 observations of age and 3 observations of weight and the exact values.We can create and combine these decks using complete function

4.Fit liner regression models to the imputed data decks and get the combined result.

Now we can use the five decks and fit a linear model modelfit1 to all the five imputed datasets and then combine them using pool and have the summary.

```
modelFit1 <- with(class_data_imp,lm(Weight~Age+Sex+Height))
summary(pool(modelFit1))

##                      est         se          t       df     Pr(>|t|)
## (Intercept)  53.1239776 58.4955076  0.9081719 52.32982 0.3679509447
## Age          -0.7474853  1.6161652 -0.4625055 46.88025 0.6458584397
## Sex2        -25.6672959  6.6450362 -3.8626270 29.58357 0.0005659108
## Height        1.6901432  0.8247588  2.0492577 49.80708 0.0457225958
##                   lo 95      hi 95 nmis        fmi     lambda
## (Intercept) -64.23823889 170.486194   NA 0.06011219 0.02486407
## Age          -3.99900967   2.504039    2 0.11114753 0.07401938
## Sex2        -39.24628527 -12.088306   NA 0.25412754 0.20535152
## Height        0.03340749   3.346879    0 0.08589392 0.04991051
```

5. Briefly compare the two models, one with missing data and second with imputed data.

As we can see from the two models the coefficients have changed a bit although the significant predictors have remained the same sex and height.

The coefficient for sex which was -27.86 which is now -25.66 which suggests that the weight is reduced by 27.86 centimeter (if female) but after imputation it changed to 25.66

Height coefficeint has almost remained the same.

Age has changed from -0.54 to -0.74. But the predictor stays insignificant because of its p value.