**LEUKEMIA**

#Loading the dataset

data2<-read.csv(file.choose())

colnames(data2)

"survival.times" "status" "sex" "logWBC" "Rx"

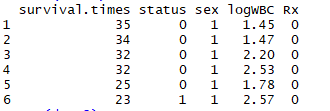
nrow(data2)

42

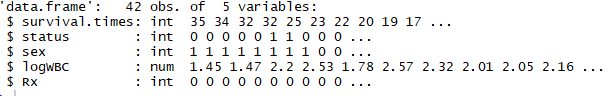
sum(is.na(data2))

0

Head(data2)

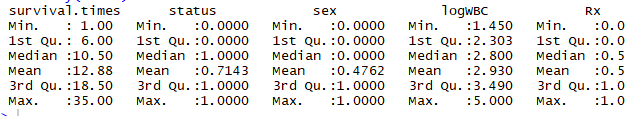


Str(data2)



***#Sex, Status and Rx are factor variables captured as incorrect types as shown above. We shall resolve that while exploring the attributes.***

summary(data2)



***Survival times is right skewed and log WBC is almost normally distributed.***

Quant\_colnames <- c ("survival.times","logWBC")

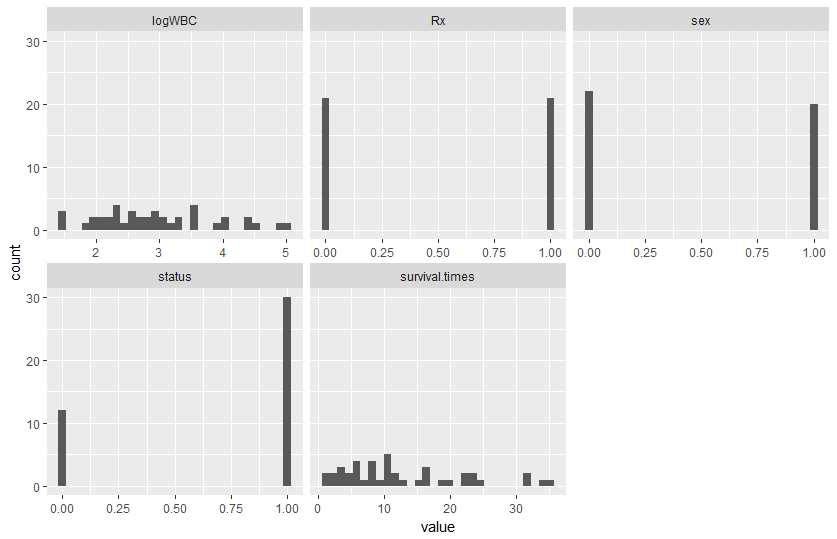
Qual\_colnames <- c("status", "sex", "Rx")

library(ggplot2)

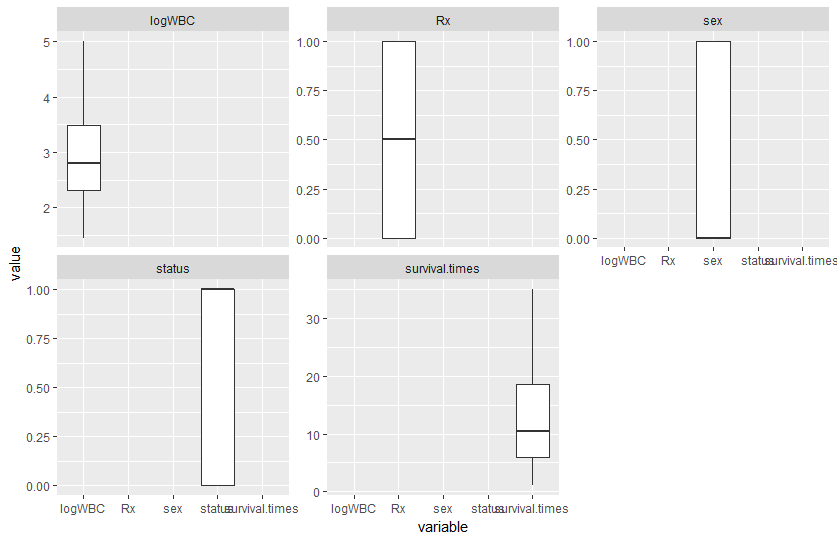
library(tidyverse)

data2 %>% gather(survival.times:Rx, key = "variable", value = "value") %>% ggplot(aes(x = value)) +

geom\_histogram(bins = 30) + facet\_wrap(~ variable, scales = 'free\_x')

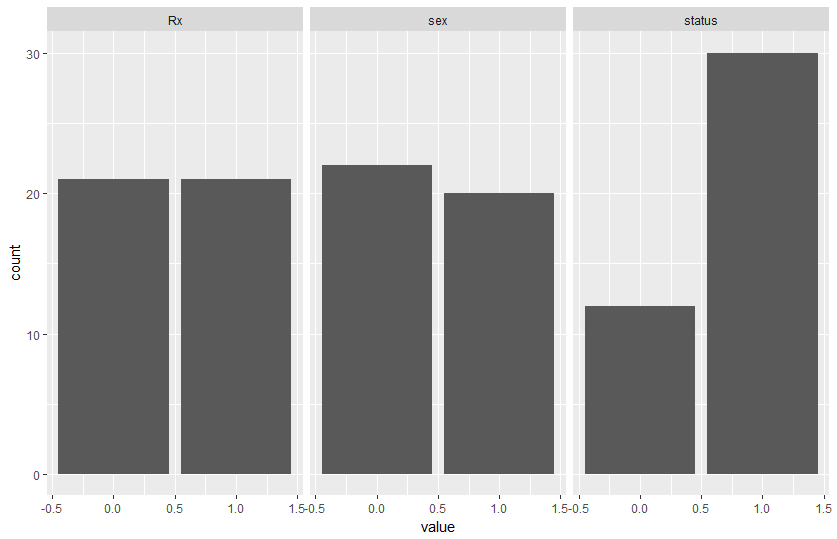


data2 %>% gather(survival.times:Rx, key = "variable", value = "value") %>% ggplot(aes(x=variable,y = value)) + geom\_boxplot() + facet\_wrap(~ variable, scales = 'free\_y')

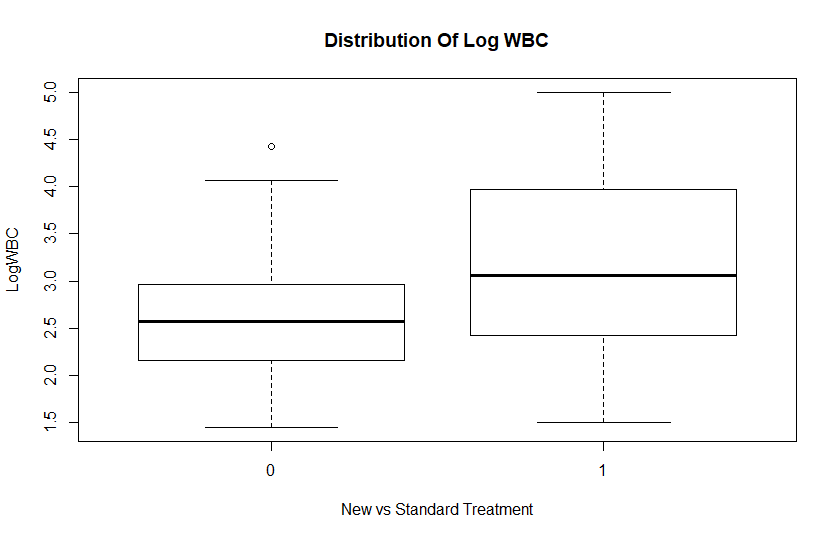


data2 %>% gather(status,sex,Rx, key = "variable", value = "value") %>% ggplot(aes(x = value)) +

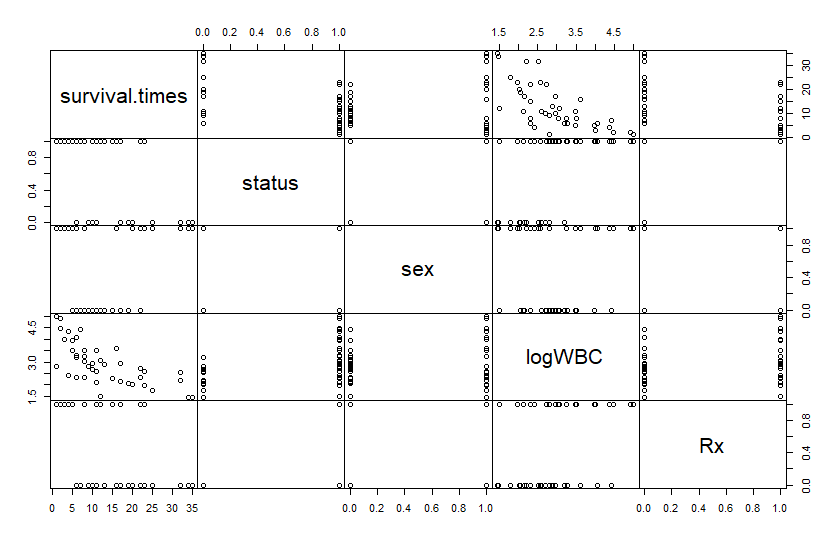
geom\_bar()+ facet\_wrap(~variable, scales = 'free\_x')



boxplot(data2$logWBC ~ data2$Rx, xlab = "New vs Standard Treatment", ylab = "LogWBC", main = "Distribution Of Log WBC",data = data2)



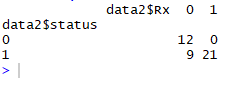
***The new treatment (at 0) shows lower values relatively lower values for log WBC.***

pairs(data2, gap = 0, pch = 21)

***Survival time and logWBC is negatively linearly related***

#Comparing the new and standard treatment wrt the status of failures

ftable(data2$Rx ~ data2$status)



***0 is new treatment; status - 1 is failures***

***For standard treatment we see that all 21 relapsed while for new treatment only 9 out of 21 relapsed***

***Thus, we see 100% relapse rate in standard treatment.***

ftable(data2$status ~ data2$sex)



***We see that the treatment/relapse is not related to gender; also across relapse variable gender is almost 50-50 proportionately distributed***

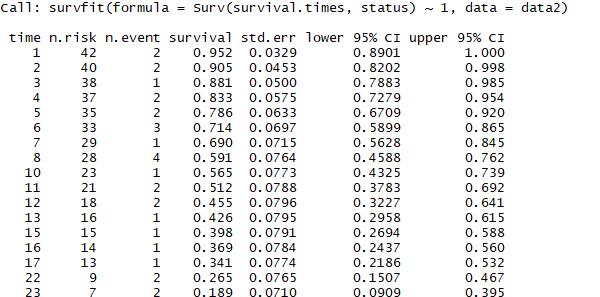
We shall study the survival on the overall data and deduce the survival probabilities of people or people dropping out of the study at different survival times

**Kaplan Meier estimate**

library(survival)

lukemia.survival = survfit(Surv(survival.times, status)~ 1, data=data2)

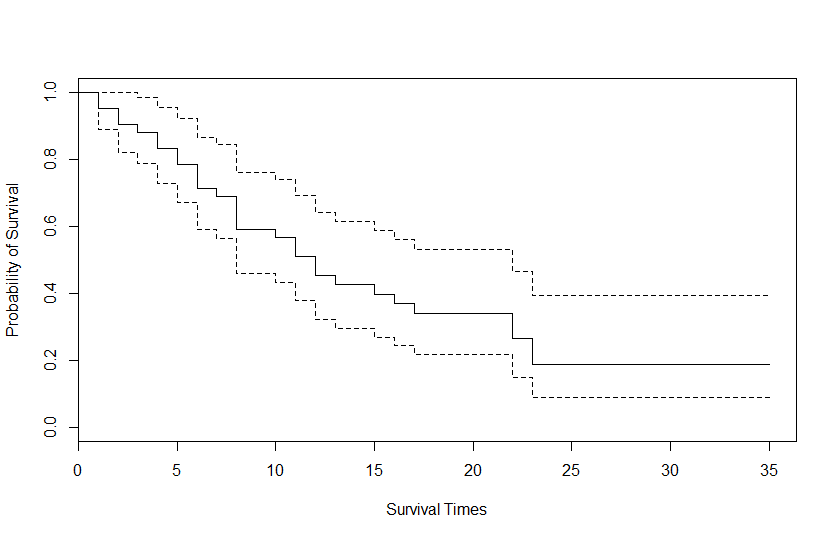
summary(lukemia.survival)



***The summary states how many people drop at each level ; n.event is failure#, we observe survival rate of 95% in the beginning because only 2 people dropped out of 42. By the end of the table, the survival rate decreases, confidence interval also goes down; the gap between lower and upper CI increases***

***that means confidence level is decreasing.***

plot(lukemia.survival, xlab="Survival Times", ylab="Probability of Survival")



***dotted line - conf. int.***

***bold line - survival probability***

***At survival time 8, we see 4 drops shown by the steep drop in graph***

**Nelson-Allen estimates**

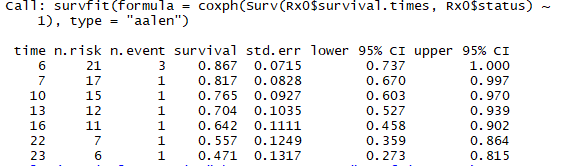
First we shall partition the data as per the new and old treatment. Nelson Allen gives a cumulative picture.

Rx0 = subset(data2, Rx == 0)

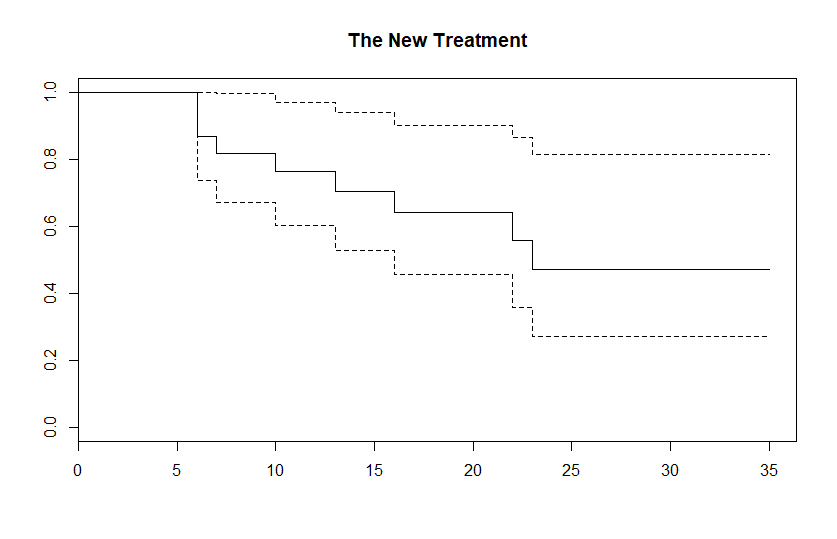
Rx1 = subset(data2, Rx == 1)

survivalRx0 = survfit(coxph(Surv(Rx0$survival.times, Rx0$status) ~ 1), type="aalen")

summary(survivalRx0)

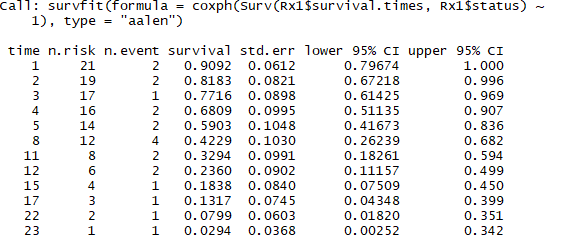


plot(survivalRx0, main="The New Treatment",conf.int=TRUE)

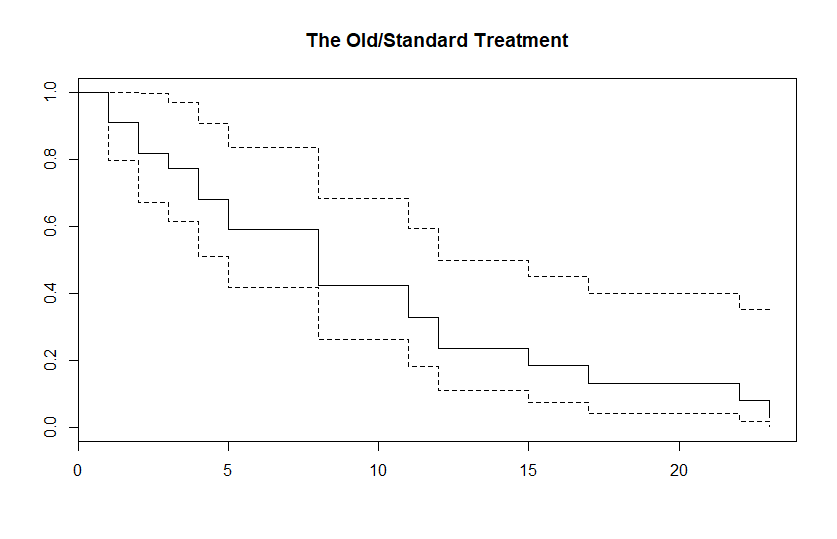


survivalRx1 = survfit(coxph(Surv(Rx1$survival.times, Rx1$status) ~ 1), type="aalen")

summary(survivalRx1)



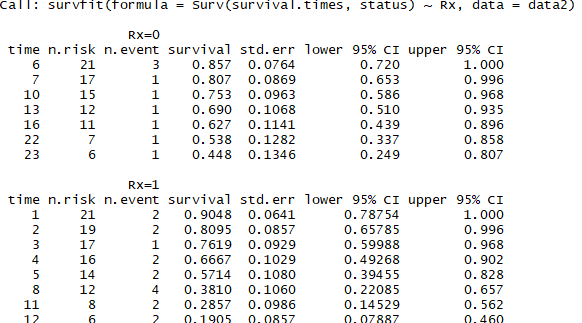
plot(survivalRx1, main="The Old/Standard Treatment",conf.int=TRUE)



**KM estimate for the two groups with confidence intervals**

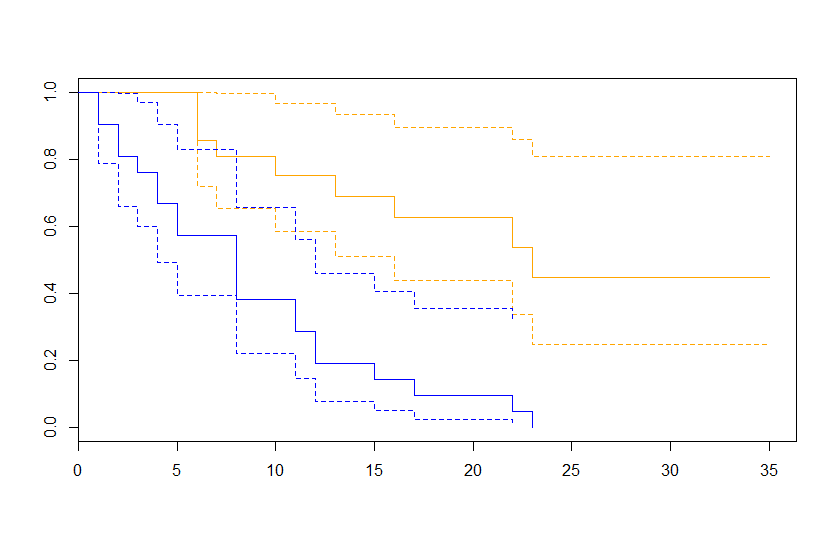
lukemia\_survival = survfit(Surv(survival.times, status)~ Rx, data=data2)

summary(lukemia\_survival)





plot(lukemia\_survival, conf.int=TRUE,col=c("orange","blue"))



***We can see from the table and the graph above that for the new treatment, the survival ends at 0.448, all patients survive for rx 0 - new treatment***

***While for rx-1, it reaches 0 by 23rd time period; hence depicting all failures***

***The new- treatment(shown in orange color) is much more successful and is better as it shows a lower failure rate, but the width of confidence levels is high ; depicting high variance or risk.***

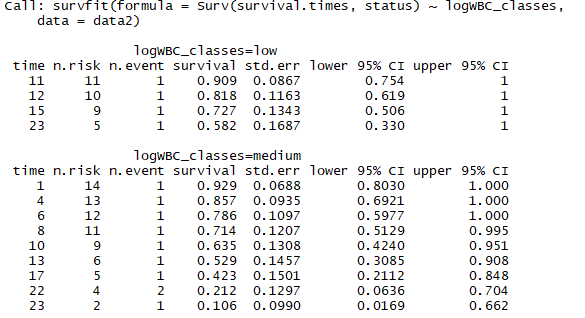
***The old or the standard treatment shows a higher failure rate and the graph shows that it drops early.***

**Categorizing logWBC into 3 classes and plotting 3 KM graphs**

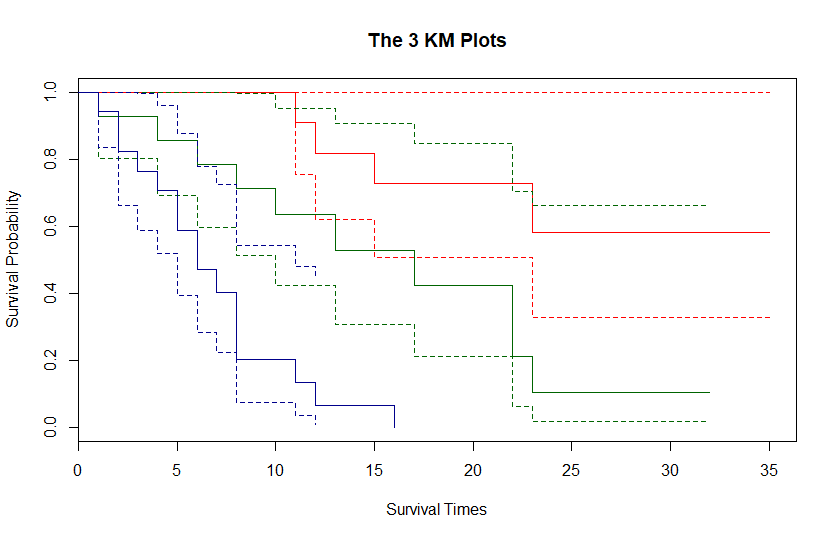
data2$logWBC\_classes = cut(data2$logWBC, breaks = c(0,2.3,3,1000), labels = ("low","medium","high"))

new\_survival = survfit(Surv(survival.times, status)~logWBC\_classes, data=data2)

summary(new\_survival)



plot(new\_survival, xlab="Survival Times", ylab="Survival Probability",main ="The 3 KM Plots",conf.int=TRUE,col=c("red","darkgreen","darkblue"))



Red – indicates low WBC and higher survival ; higher variability and a wider confidence interval

DarkGreen – shows medium response

DarkBlue-indicates high WBC and lower survival while a narrow confidence interval

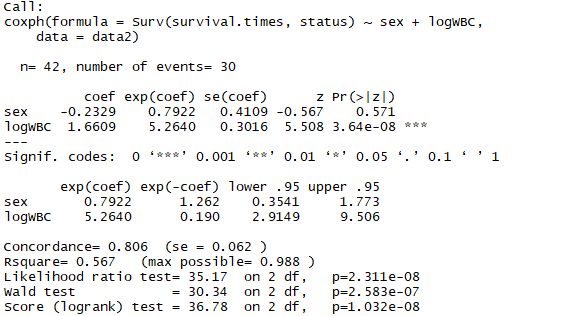
**COX PH**

It is used to show which attribute/factor affects the Survival or which factor is significant

The tests signifies if the two attributes are significant on survival or not

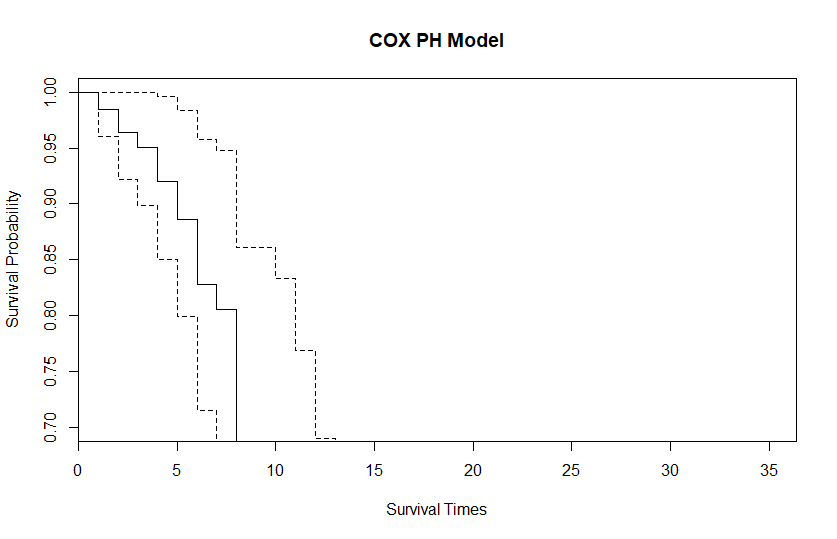
cox\_survival = coxph(Surv(survival.times, status) ~ sex+logWBC, data=data2)

summary(cox\_survival)



***Sex shows neutral effect on the survival time if all else is held constant as we can see from the summary. While the summary also shows that LogWBC is a significant factor that has an impact on Survival times.***

plot(survfit(cox\_survival), ylim=c(0.7, 1), xlab="Survival Times", ylab="Survival Probability", main="COX PH Model")



***p-value- A low value in Likelihood ratio test shows that the potential confounders Sex and logWBC are both significant together for the model .***

***We get the hazard ratio - exp(beta) using the Cox model. It shows the % increase/decrease in the hazard, whereas the KM gives us an estimate of the survival probability.***

***However, both the tests show the same outputs i.e. Gender does not affect the Survival times and logWBC has an inverse relationship with the survival times i.e. higher the logWBC, lower is the survival probability***