Politechnika
Śląska

Survival Analysis Laboratory

Laboratory 6

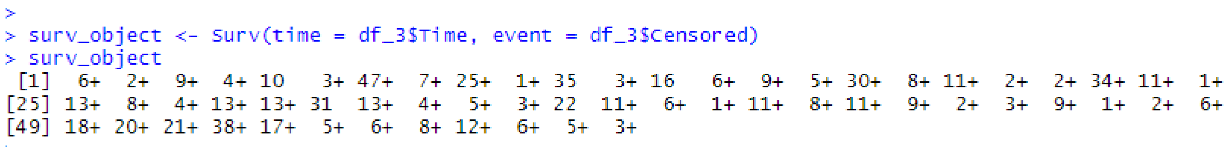
Ankit Rathi

Uwacu Jean Remy

We have used Dataset 16 to perform tasks in this laboratory.

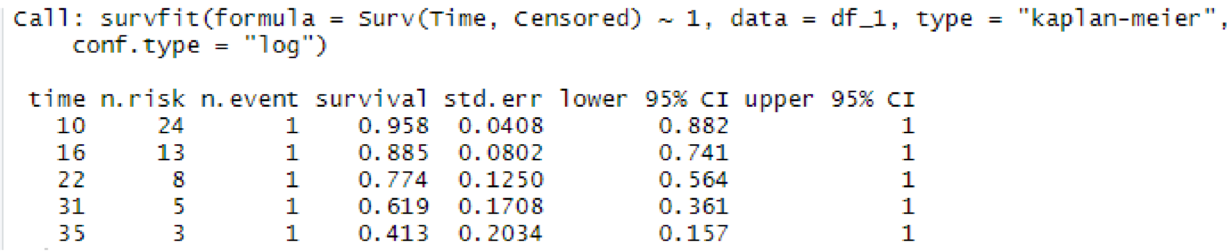
PART 1

1. For each set of data separately (censored and uncensored) compute the Kaplan-Meier estimator. In this part do not use implemented functions – perform all the computations yourselves according to the lecture. Present the results in the form of two tables (like at the lecture).



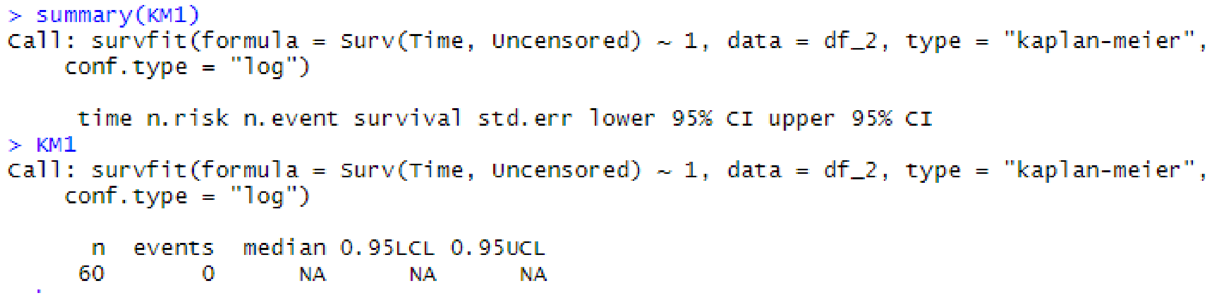
```
>
> surv_object <- Surv(time = df_3$Time, event = df_3$Censored)
> surv_object
 [1]  6+  2+  9+  4+ 10   3+ 47+  7+ 25+  1+ 35   3+ 16   6+  9+  5+ 30+  8+ 11+  2+  2+ 34+ 11+  1+
[25] 13+  8+  4+ 13+ 13+ 31  13+  4+  5+  3+ 22  11+  6+  1+ 11+  8+ 11+  9+  2+  3+  9+  1+  2+  6+
[49] 18+ 20+ 21+ 38+ 17+  5+  6+  8+ 12+  6+  5+  3+
```

censored Data



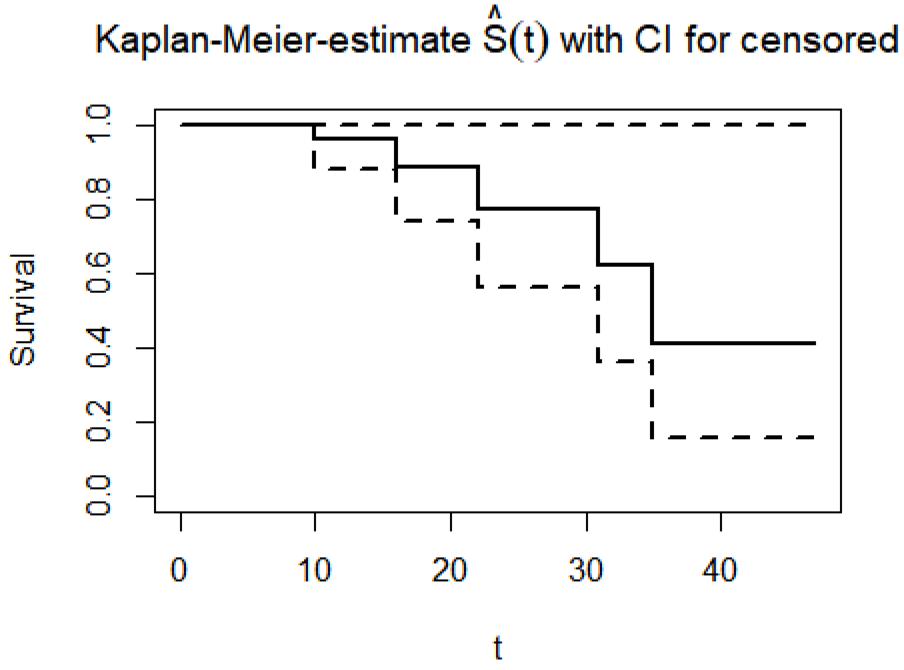
```
Call: survfit(formula = Surv(Time, Censored) ~ 1, data = df_1, type = "kaplan-meier",
    conf.type = "log")

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   10     24       1    0.958  0.0408        0.882            1
   16     13       1    0.885  0.0802        0.741            1
   22      8       1    0.774  0.1250        0.564            1
   31      5       1    0.619  0.1708        0.361            1
   35      3       1    0.413  0.2034        0.157            1
```

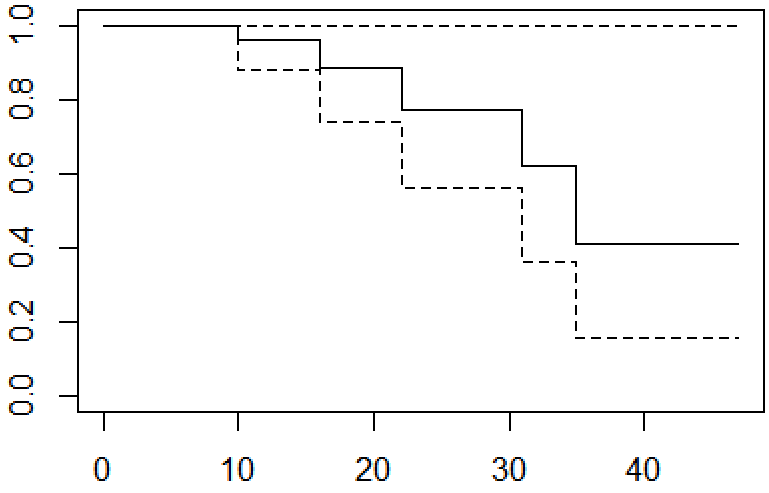
Uncensored Data



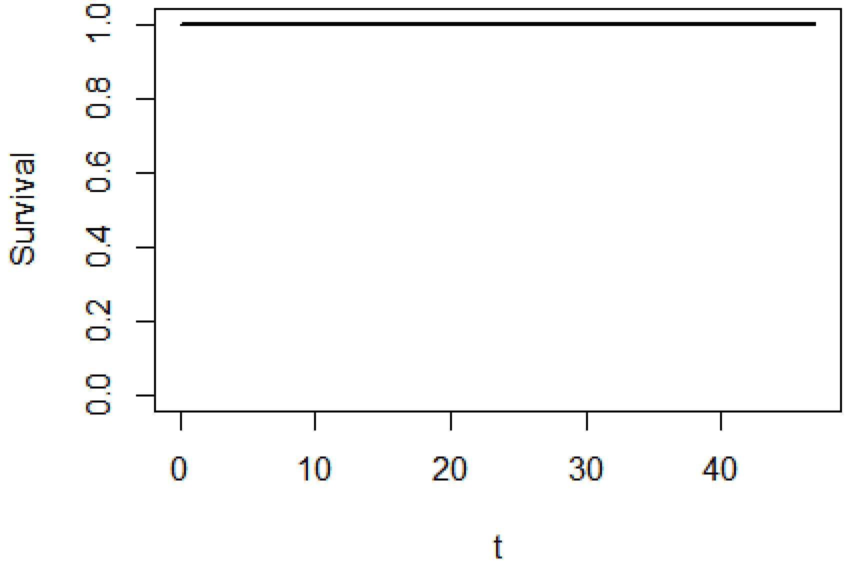
```
> summary(KM1)
Call: survfit(formula = Surv(Time, Uncensored) ~ 1, data = df_2, type = "kaplan-meier",
    conf.type = "log")

     time n.risk n.event survival std.err lower 95% CI upper 95% CI
> KM1
Call: survfit(formula = Surv(Time, Uncensored) ~ 1, data = df_2, type = "kaplan-meier",
    conf.type = "log")

      n  events  median 0.95LCL 0.95UCL
     60       0      NA      NA      NA
```

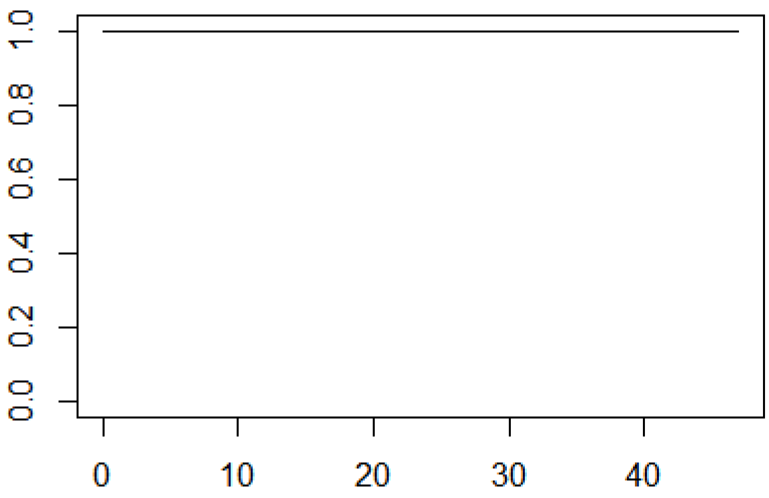
2. Using the results from point 1, draw the survival curves for uncensored and censored datasets at the same plot.





Kaplan-Meier-estimate $\hat{S}(t)$ with CI for uncensored



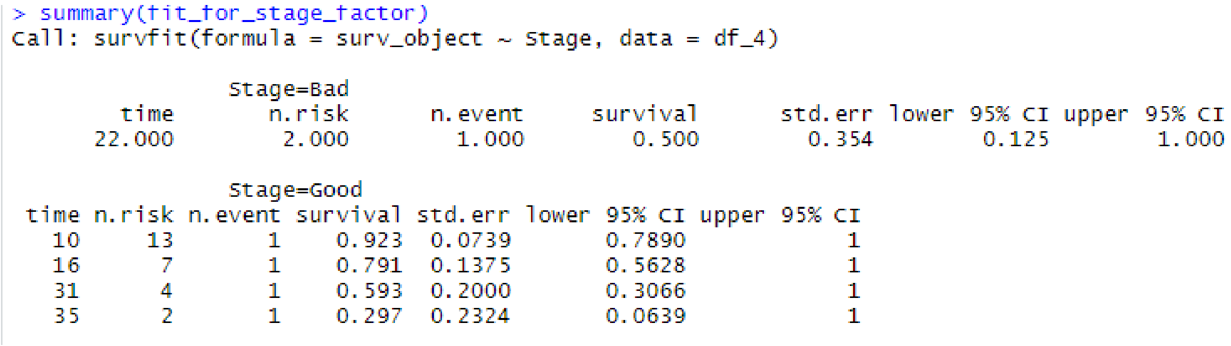


3. Based on the results from the previous two points answer the following questions: a. Does the censoring influence the values of the probability of survival? b. If so, what kind of changes in probability were made and why? c. Should censoring be used in the survival analysis? Comment on your results.

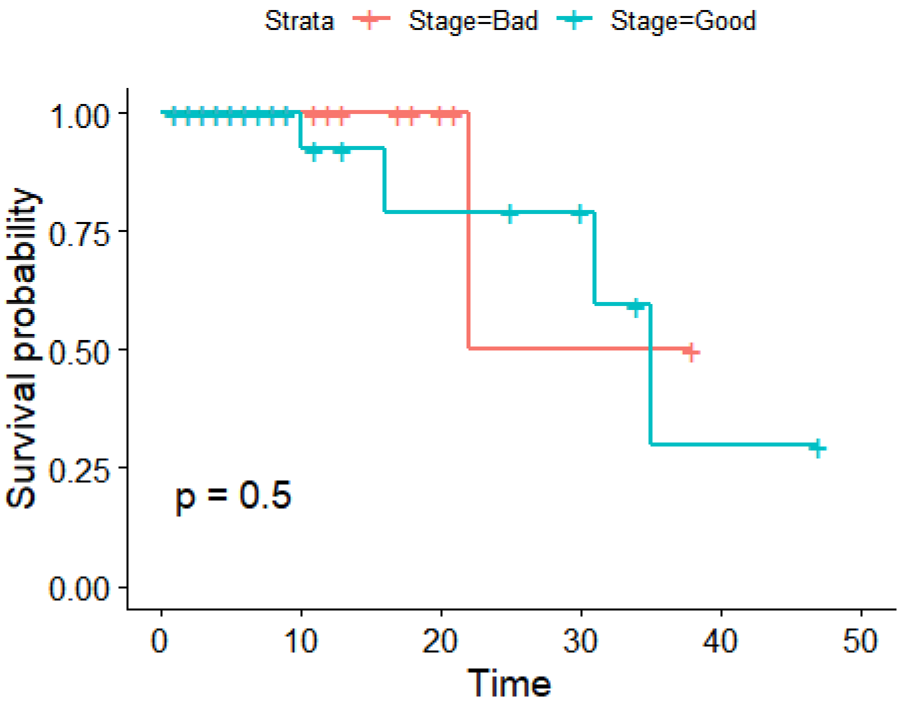
   As we can see from the data above censoring does affect the survival time, As we can see from the survival curves from point 2, with censored data the probability of survival changes as we move from time t0-t2, whereas with uncensored data there is no significant differences as if there is no changes and we have got a linear survival probability for t0-tn time.

Part 2: Comparison of two groups (log rank test)

1. Draw the survival curves for good and bad stages at the same plot with the confidence intervals. Comment on the plot: do the curves look different to you? Does the illness stage seem to influence the survival time?



```
> summary(fit_for_stage_factor)
Call: survfit(formula = surv_object ~ Stage, data = df_4)

                Stage=Bad
        time       n.risk       n.event      survival      std.err lower 95% CI upper 95% CI
      22.000        2.000         1.000         0.500        0.354        0.125        1.000

                Stage=Good
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   10     13       1    0.923  0.0739       0.7890            1
   16      7       1    0.791  0.1375       0.5628            1
   31      4       1    0.593  0.2000       0.3066            1
   35      2       1    0.297  0.2324       0.0639            1
```



Based on the Above curve the stage factor does not influence the survival time , The curves for bad stage and good stage doesn't look same, their survival probability vary from each other at different time taken in consideration.

2. Compare the survival curves with the use of the log-rank test. In this part do not use implemented functions – perform all the computations yourselves according to the lecture. Remember to write down hypotheses (null and alternative one), value of test statistics, p-value, decisions, and conclusions.

The log-rank p-value of 0.5 indicates a non-significant result if you consider $p < 0.05$ to indicate statistical significance.

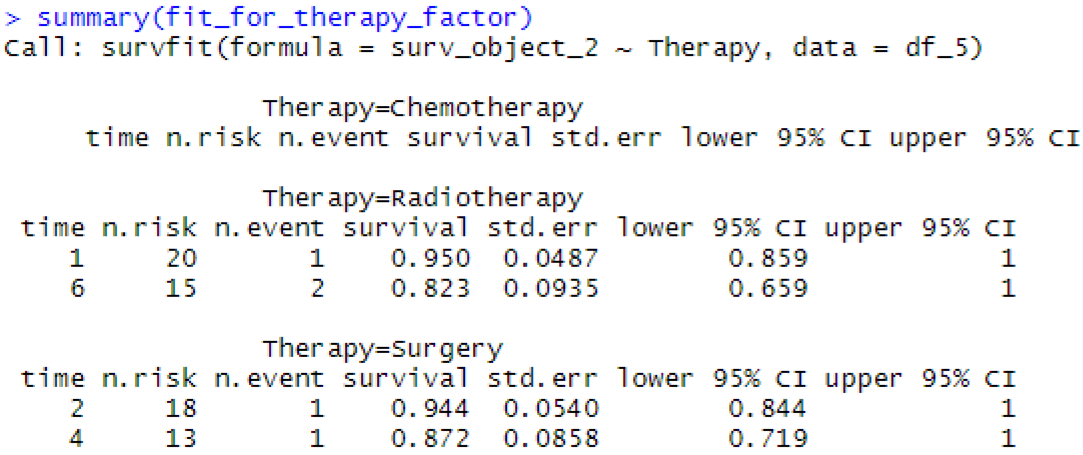
p-value(0.5) is less than 0.05, we reject the null hypothesis that there's no difference between the means and conclude that a significant difference does exist.

3. Does the survival time of patients diagnosed at good and bad stages differ significantly?

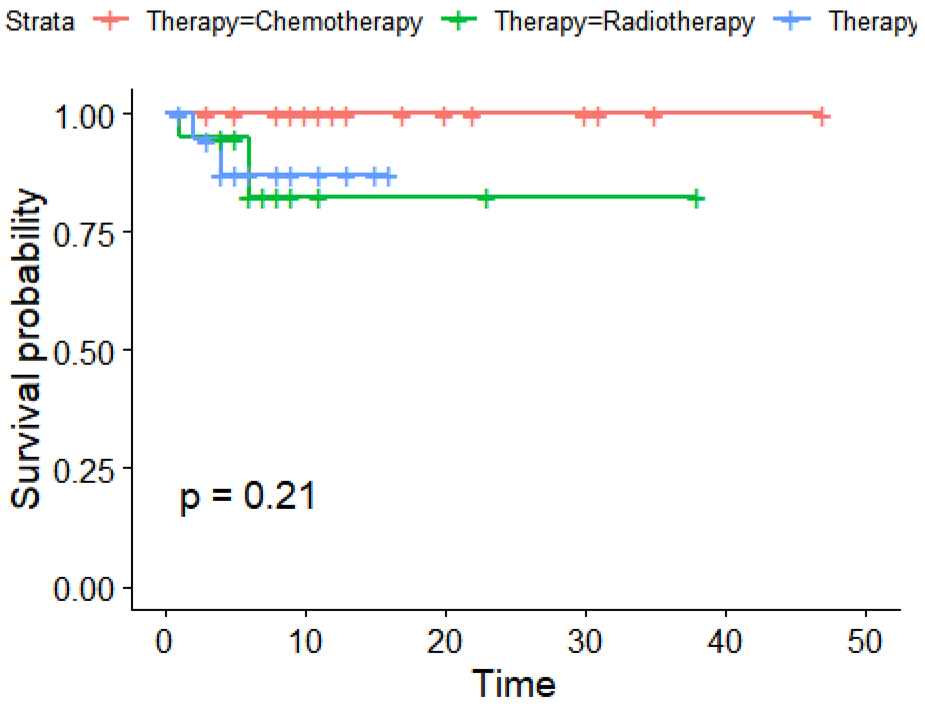
   Based on the log rank test the value of p-value is 0.5 which is less than 0.05 i.e. p<0.05 and hence it is insignificant
   i.e. stage doesnt effect survival time

Part 3 : Comparison of three or more group : log rank test

1. Draw the survival curves for each of the therapies at the same plot. Comment on the plot: do the curves look different to you? Does the treatment seem to influence the survival time?



```
> summary(fit_for_therapy_factor)
Call: survfit(formula = surv_object_2 ~ Therapy, data = df_5)

                Therapy=Chemotherapy
     time n.risk n.event survival std.err lower 95% CI upper 95% CI

                Therapy=Radiotherapy
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     20       1    0.950  0.0487        0.859            1
    6     15       2    0.823  0.0935        0.659            1

                Therapy=Surgery
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     18       1    0.944  0.0540        0.844            1
    4     13       1    0.872  0.0858        0.719            1
```



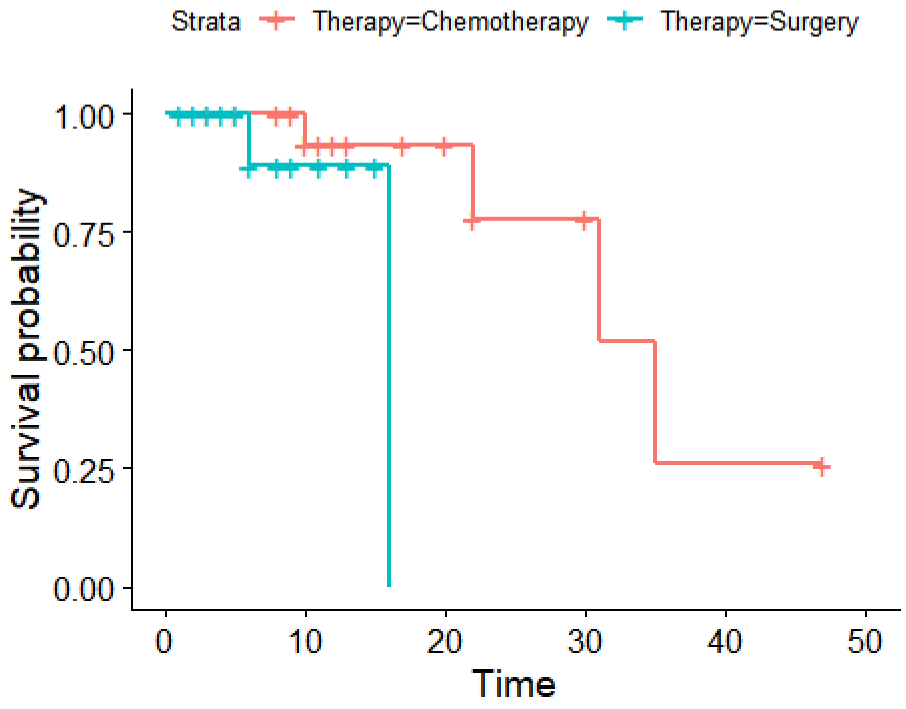
As we can see from the survival curves we can see that that the therapy is affecting the survival time, As people going under chemotherapy and Radiotherapy, tends to have a good survival time than the one going under the surgery.

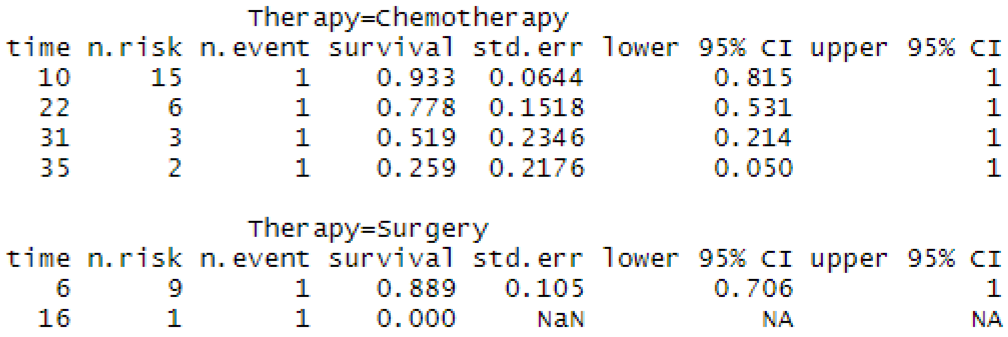
2. Compare all three survival curves with the use of the log-rank test. Remember to write down hypotheses (null and alternative one), value of test statistics, p-value, decisions, and conclusions.

As we can see from the survival curve , By convention, vertical lines indicate censored data, their corresponding x values the time at which censoring occurred. The log-rank p-value of 0.21 indicates a non-significant result if you consider `p < 0.05` to indicate statistical significance.

3. Draw the survival curves for all pairs of therapies (Radiotherapy vs. Chemotherapy, Radiotherapy vs. Surgery, Surgery vs. Radiotherapy) with their confidence intervals. You should receive three plots with two curves at each of them. Comment on the plots

   Surg vs chemo

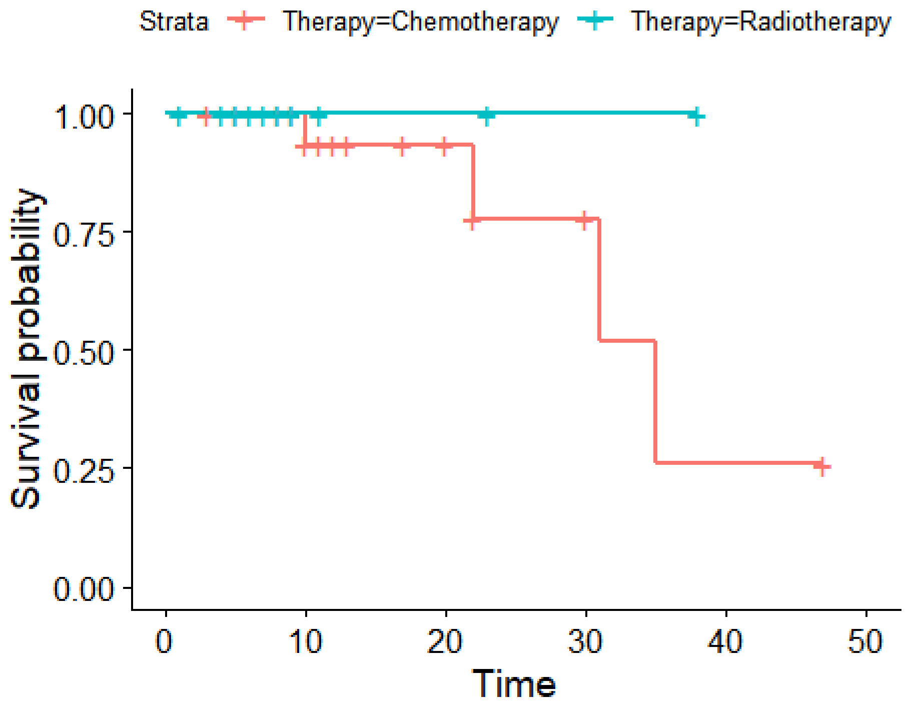


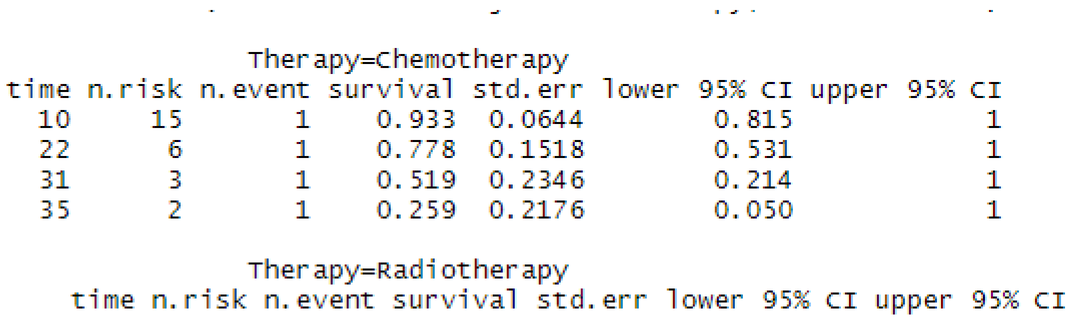


```
                Therapy=Chemotherapy
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  10     15       1    0.933  0.0644        0.815            1
  22      6       1    0.778  0.1518        0.531            1
  31      3       1    0.519  0.2346        0.214            1
  35      2       1    0.259  0.2176        0.050            1

                Therapy=Surgery
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   6      9       1    0.889   0.105        0.706            1
  16      1       1    0.000     NaN           NA           NA
```

As we can see from this table, in therapy treatment as chemotherapy, the lower 95%CI is gradually coming down where as upper 95CI is unchanged, where as in surgery after the time t=16 we don't have any confidence interval, so basically we can say surgery doesn't affect the survival probability.

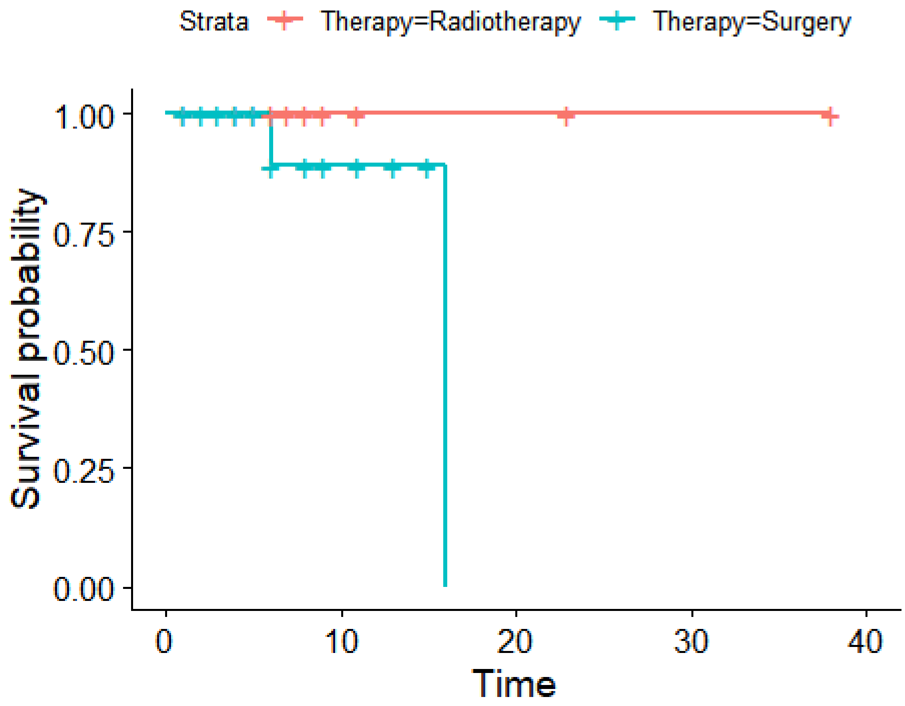
Chemo vs Radio



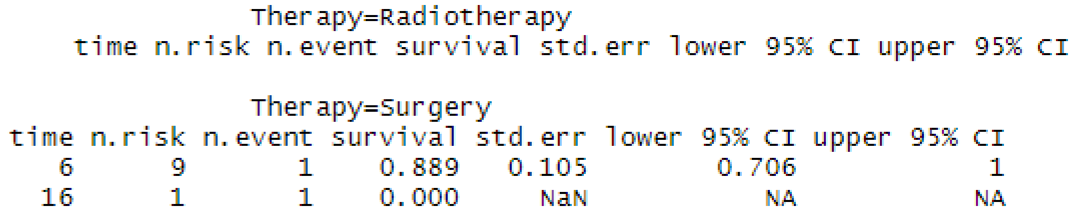


```
                Therapy=Chemotherapy
time n.risk n.event survival std.err lower 95% CI upper 95% CI
  10     15       1    0.933  0.0644        0.815            1
  22      6       1    0.778  0.1518        0.531            1
  31      3       1    0.519  0.2346        0.214            1
  35      2       1    0.259  0.2176        0.050            1

                Therapy=Radiotherapy
     time n.risk n.event survival std.err lower 95% CI upper 95% CI
```

Surgery vs Radio



As we can see from the survival curves surgery doesn't affect the survival probability.

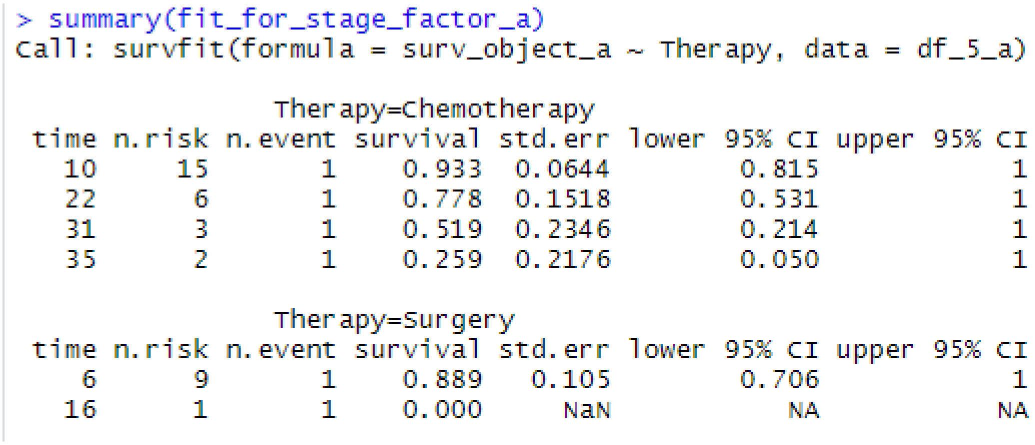


```
                Therapy=Radiotherapy
     time n.risk n.event survival std.err lower 95% CI upper 95% CI

                Therapy=Surgery
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6      9       1    0.889   0.105        0.706            1
   16      1       1    0.000     NaN           NA           NA
```

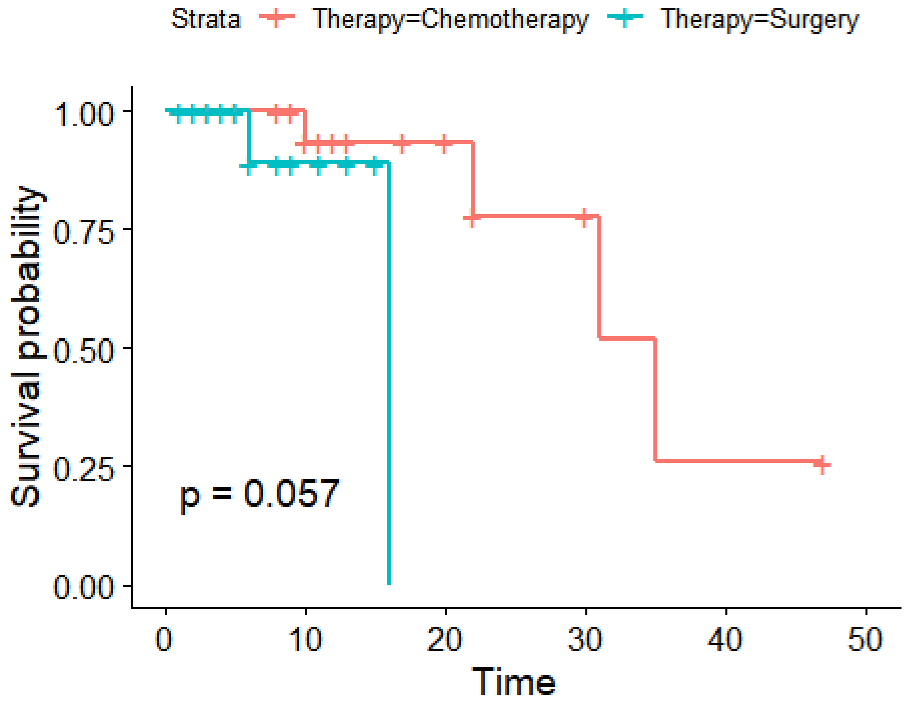
4. Compare the pairs of therapies (Radiotherapy vs. Chemotherapy, Radiotherapy vs. Surgery, Surgery vs. Radiotherapy) with the use of the log-rank test. Do the results of tests confirm what you have observed at the plots from the previous point? Remember to write down hypotheses (null and alternative one), values of test statistics, p-values, decisions, and conclusions.

   Plotting Log-Rank Test for Each

   Surg VS Chemo

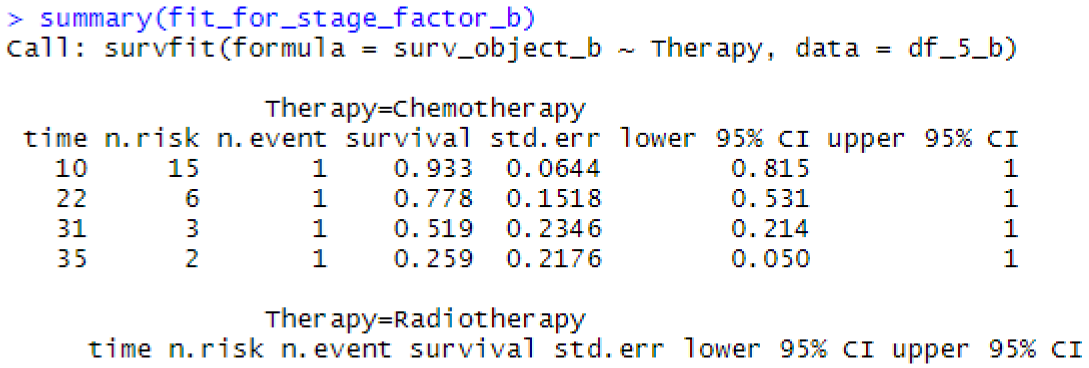


```
> summary(fit_for_stage_factor_a)
Call: survfit(formula = surv_object_a ~ Therapy, data = df_5_a)

                Therapy=Chemotherapy 
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   10     15       1    0.933  0.0644        0.815            1
   22      6       1    0.778  0.1518        0.531            1
   31      3       1    0.519  0.2346        0.214            1
   35      2       1    0.259  0.2176        0.050            1

                Therapy=Surgery 
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6      9       1    0.889   0.105        0.706            1
   16      1       1    0.000     NaN           NA           NA
```

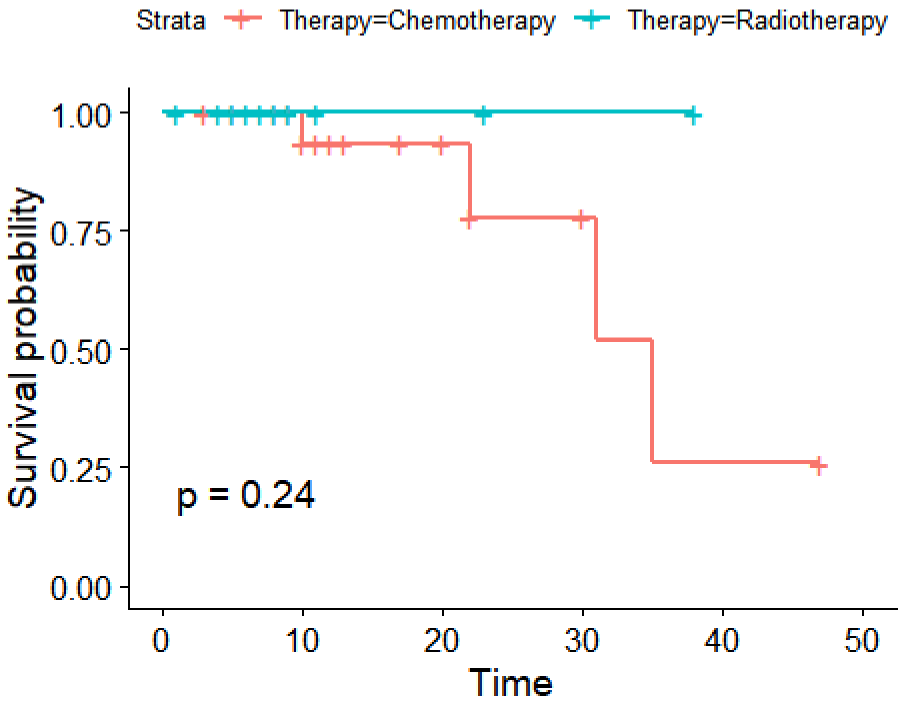


As we can see the p-value is 0.057 > 0.05 so we can accept the null hypothesis , A **p-value** higher than **0.05** (> **0.05**) is not statistically significant and indicates weak evidence against the null hypothesis. This means we fail to reject the null hypothesis and cannot accept the alternative hypothesis. Surgery doesn’t affect the survival probability.

Chemo vs Radio

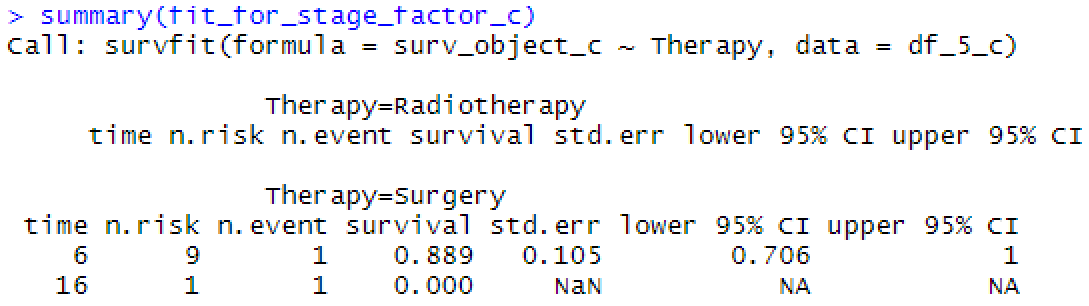


```
> summary(fit_for_stage_factor_b)
Call: survfit(formula = surv_object_b ~ Therapy, data = df_5_b)

                Therapy=Chemotherapy
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   10     15       1    0.933  0.0644        0.815            1
   22      6       1    0.778  0.1518        0.531            1
   31      3       1    0.519  0.2346        0.214            1
   35      2       1    0.259  0.2176        0.050            1

                Therapy=Radiotherapy
     time n.risk n.event survival std.err lower 95% CI upper 95% CI
```

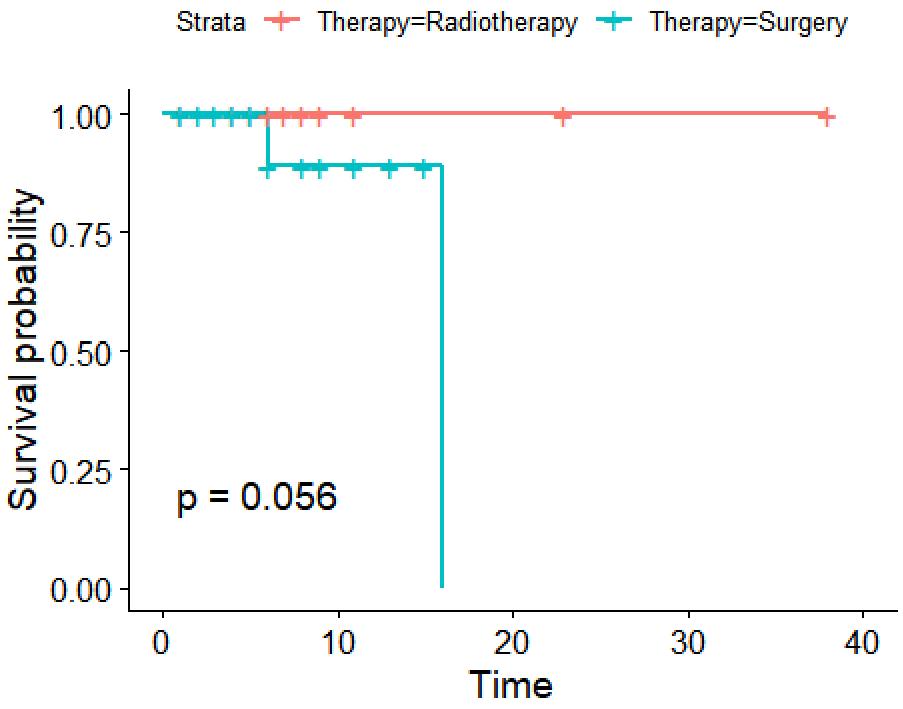


As we can see the p-value is 0.24 < 0.05 so we can reject the null hypothesis , A *p*-value less than 0.05 (typically ≤ 0.05) is statistically significant. It indicates strong evidence against the null hypothesis, as there is less than a 5% probability the null is correct (and the results are random). Therefore, we reject the null hypothesis, and accept the alternative hypothesis.

Surg vs Radio



```
> summary(fit_for_stage_factor_c)
Call: survfit(formula = surv_object_c ~ Therapy, data = df_5_c)

                Therapy=Radiotherapy
     time n.risk n.event survival std.err lower 95% CI upper 95% CI

                Therapy=Surgery
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6      9       1    0.889   0.105        0.706            1
   16      1       1    0.000     NaN           NA           NA
```



As we can see the p-value is 0.056 > 0.05 so we can accept the null hypothesis , A **p-value** higher than **0.05** (> **0.05**) is not statistically significant and indicates weak evidence against the null hypothesis. This means we fail to reject the null hypothesis and cannot accept the alternative hypothesis. Surgery doesn't affect the survival probability.

5. Does the survival time of patients treated with various methods differ significantly? Comment on the results of the comparison of all treatments and the pairs of them.

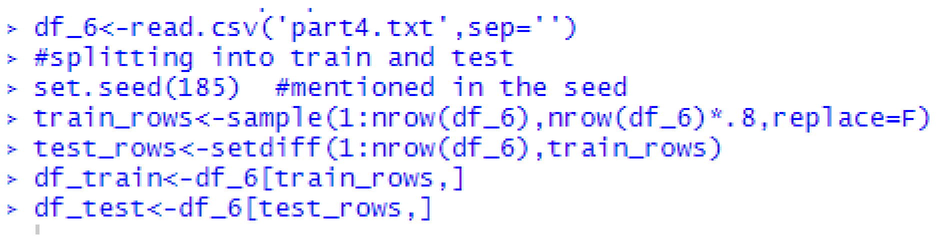
surgery does effect the survivial probabbility and time but the p_value is still >0.5 i.e. 0.056, As we can in all three different survival plots, the p-value for surg vs chemo is 0.057 which is >0.05, so we can accept the null hypothesis for that, and for the next group we have a chemo vs Radio where we have got a p-value of 0.24 and for the last survival curve we have surg vs radio where p-value(0.056)>0.05 so we can accept the null hypothesis.

Surgery doesn't affect the survival probabilities.

Part 4 :

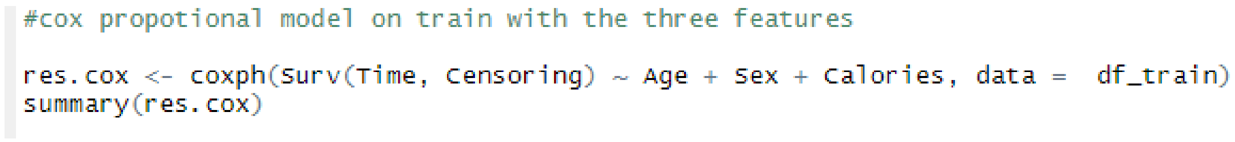
1. **type of columns**
   columns are in factor form so not converting yet into one hot encoding as mentioned in the assignment

2. **Here as asked In the task we have split the dataset into test and train set.**

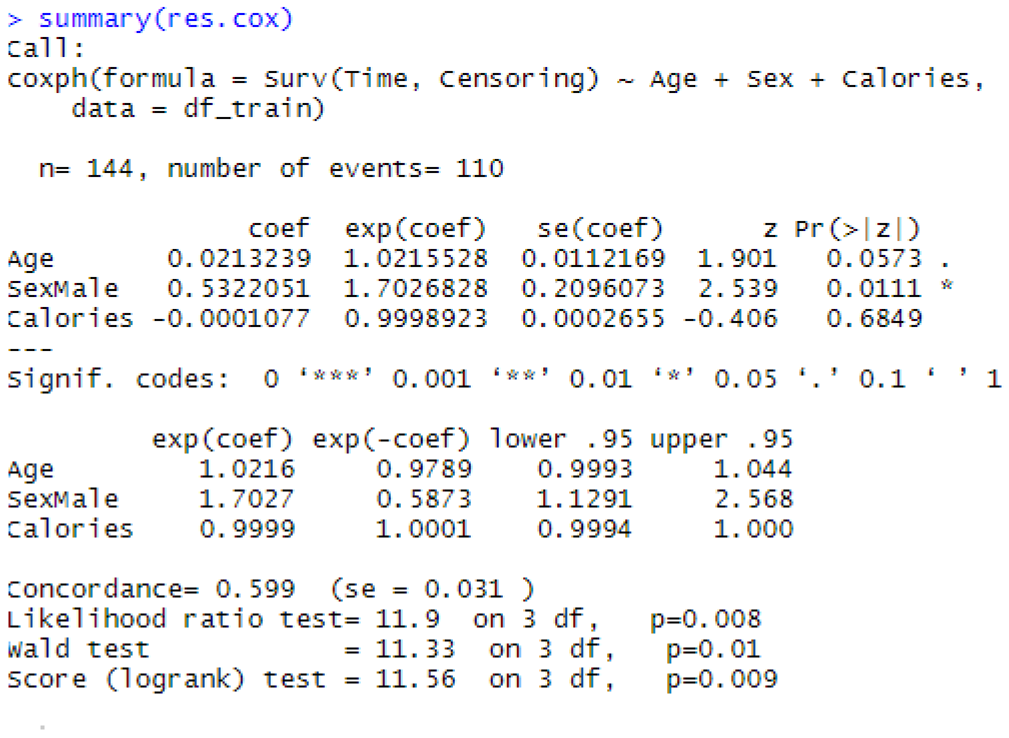


```
> df_6<-read.csv('part4.txt',sep='')
> #splitting into train and test
> set.seed(185)  #mentioned in the seed
> train_rows<-sample(1:nrow(df_6),nrow(df_6)*.8,replace=F)
> test_rows<-setdiff(1:nrow(df_6),train_rows)
> df_train<-df_6[train_rows,]
> df_test<-df_6[test_rows,]
```

3.



```
#cox propotional model on train with the three features

res.cox <- coxph(Surv(Time, Censoring) ~ Age + Sex + Calories, data =  df_train)
summary(res.cox)
```



```
> summary(res.cox)
Call:
coxph(formula = Surv(Time, Censoring) ~ Age + Sex + Calories,
    data = df_train)

  n= 144, number of events= 110

               coef  exp(coef)   se(coef)      z Pr(>|z|)
Age       0.0213239  1.0215528  0.0112169  1.901   0.0573 .
SexMale   0.5322051  1.7026828  0.2096073  2.539   0.0111 *
Calories -0.0001077  0.9998923  0.0002655 -0.406   0.6849
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
Age         1.0216     0.9789    0.9993     1.044
SexMale     1.7027     0.5873    1.1291     2.568
Calories    0.9999     1.0001    0.9994     1.000

Concordance= 0.599  (se = 0.031 )
Likelihood ratio test= 11.9  on 3 df,   p=0.008
Wald test            = 11.33  on 3 df,   p=0.01
Score (logrank) test = 11.56  on 3 df,   p=0.009
```

Here we have built a cox-propotional model based on train set with three feature. *Cox's proportional hazards regression model* ***(also called* Cox regression *or* Cox's model*) builds a survival function which tells you probability a certain event (e.g. death) happens at a particular time* t*. Once you've built the model from observed values, it can then be used to make predictions for new inputs.***

A hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

The column marked "z" gives the Wald statistic value. It corresponds to the ratio of each regression coefficient to its standard error (z = coef/se(coef)). The wald statistic evaluates, whether the beta (ββ) coefficient of a given variable is statistically significantly different from 0. From the output above, we can conclude that the variable age, sex have highly statistically significant coefficients.

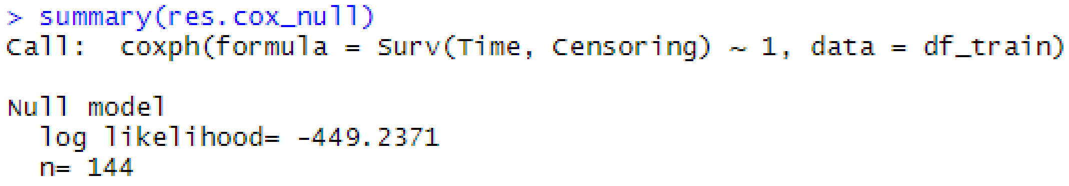
The second feature to note in the Cox model results is the the sign of the regression coefficients (coef). A positive sign means that the hazard (risk of death) is higher, and thus the prognosis worse, for subjects with higher values of that variable. The R summary for the Cox model gives the hazard ratio (HR) for the second group relative to the first group, that is, female versus male. The beta coefficient for sex = 0.53 indicates higher risk, whereas for age 0.021 is till positive, unless calories which have -0.0001.

The exponentiated coefficients (exp(coef) = exp(0.021) = 1.021), (exp(coef) = exp(-0.00001) = 0.9999), (exp(coef) = exp(0.532) = 1.702), also known as* hazard ratios*, give the effect size of covariates.

The summary output also gives upper and lower 95% confidence intervals for the hazard ratio (exp(coef)), lower 95% bound = 0.993, upper 95% bound =1.044(for age), lower 95% bound = 1.1291, upper 95% bound =2.568(for sex), lower 95% bound = 0.994, upper 95% bound =1.000(for age), .

Finally, the output gives p-values for three alternative tests for overall significance of the model: The likelihood-ratio test, Wald test, and score logrank statistics. These three methods are asymptotically equivalent. For large enough N, they will give similar results. For small N, they may differ somewhat. The Likelihood ratio test has better behavior for small sample sizes, so it is generally preferred.

4.

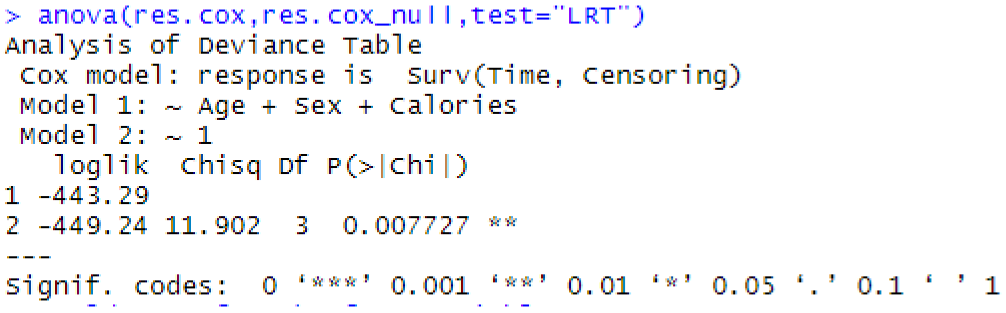


```
> summary(res.cox_null)
Call:  coxph(formula = Surv(Time, Censoring) ~ 1, data = df_train)

Null model
  log likelihood= -449.2371
  n= 144
```

The log-likelihood is the expression that Minitab maximizes to determine optimal values of the estimated coefficients (β).

Log-likelihood values cannot be used alone as an index of fit because they are a function of sample size but can be used to compare the fit of different coefficients. Because you want to maximize the log-likelihood, the higher value is better.

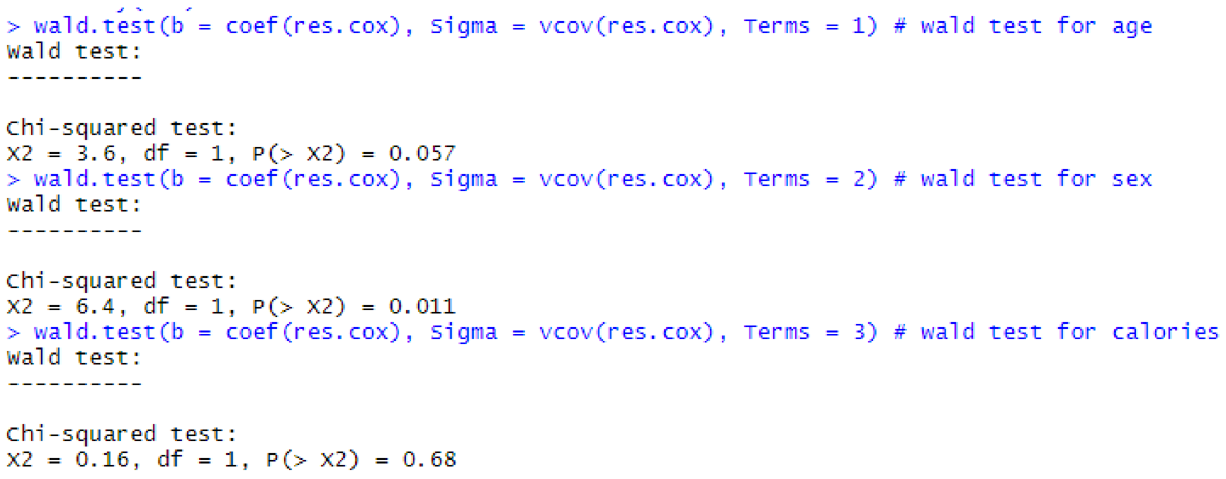
LR test to compare with the null model



```
> anova(res.cox,res.cox_null,test="LRT")
Analysis of Deviance Table
 Cox model: response is  Surv(Time, Censoring)
 Model 1: ~ Age + Sex + Calories
 Model 2: ~ 1
   loglik  Chisq Df P(>|Chi|)
1 -443.29
2 -449.24 11.902  3  0.007727 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

we reject the null hypothesis i.e. model doesn't depend on age+sex+calories as p value is 0.007245 <0.05
p-value is 0.007727 < 0.05 so we don’t have enough evidence to accept null hypothesis, and the log likelihood values is almost similar for both of them.

5. **Wald test**

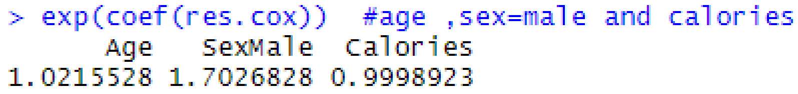


```
> wald.test(b = coef(res.cox), Sigma = vcov(res.cox), Terms = 1) # wald test for age
Wald test:
----------

Chi-squared test:
X2 = 3.6, df = 1, P(> X2) = 0.057
> wald.test(b = coef(res.cox), Sigma = vcov(res.cox), Terms = 2) # wald test for sex
Wald test:
----------

Chi-squared test:
X2 = 6.4, df = 1, P(> X2) = 0.011
> wald.test(b = coef(res.cox), Sigma = vcov(res.cox), Terms = 3) # wald test for calories
Wald test:
----------

Chi-squared test:
X2 = 0.16, df = 1, P(> X2) = 0.68
```

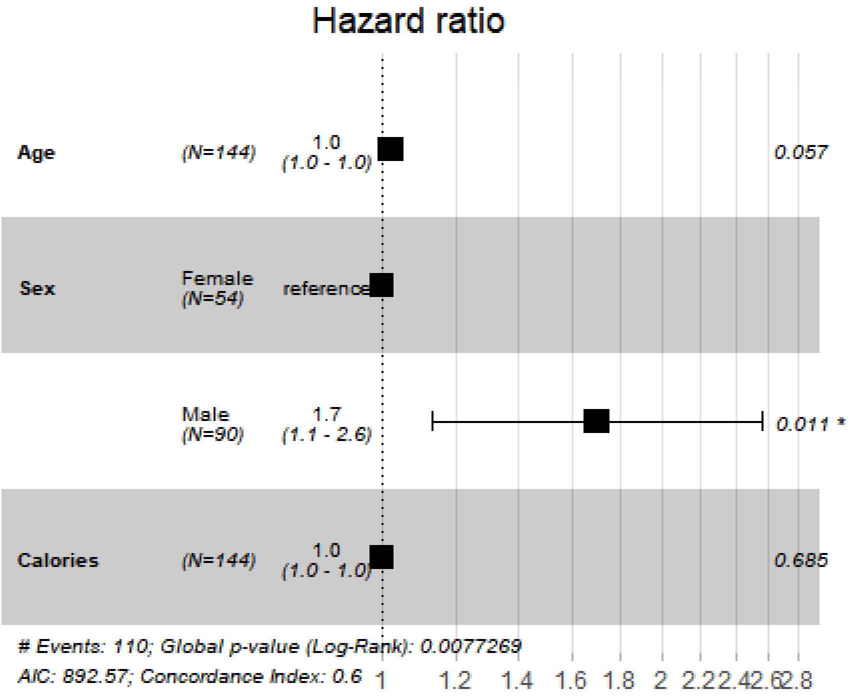
The **Wald test** (also called the Wald Chi-Squared Test) is a way to find out if explanatory variables in a model are significant. “Significant” means that they add something to the model; variables that add nothing can be deleted without affecting the model in any meaningful way.

The p-value for age is 0.057 >0.05, so we accept the null hypothesis, but in case of sex p-value0.011<0.05, so we don’t have enough evidence to accept null hypothesis, but in case of calories again p-value 0.68>0.05 so we can accept the null hypothesis.

6, Hazard Ratio

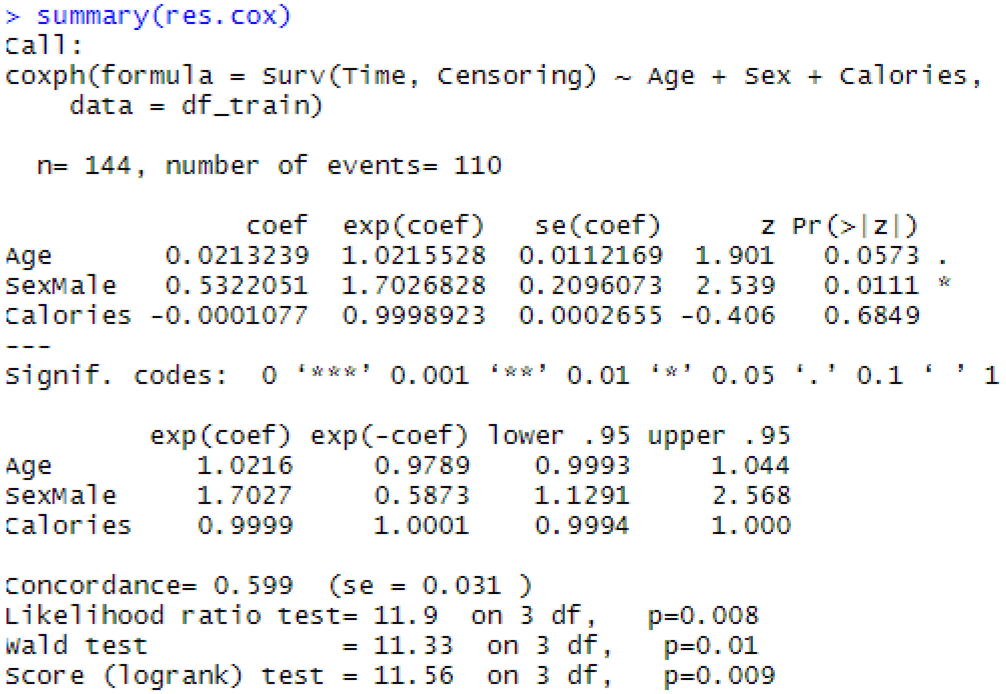


```
> exp(coef(res.cox))  #age ,sex=male and calories
      Age   SexMale  Calories 
1.0215528 1.7026828 0.9998923 
```



Every HR represents a relative risk of death that compares one instance of a binary feature to the other instance. For example, a hazard ratio of 1.7 for treatment groups tells you that male patients reduced risk of dying compared to Female patients (which served as a reference to calculate the hazard ratio).

7, Cox propotional Hazard Model



```
> summary(res.cox)
Call:
coxph(formula = Surv(Time, Censoring) ~ Age + Sex + Calories, 
    data = df_train)

  n= 144, number of events= 110 

               coef  exp(coef)   se(coef)      z Pr(>|z|)  
Age       0.0213239  1.0215528  0.0112169  1.901   0.0573 .
SexMale   0.5322051  1.7026828  0.2096073  2.539   0.0111 *
Calories -0.0001077  0.9998923  0.0002655 -0.406   0.6849  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
Age         1.0216     0.9789    0.9993     1.044
SexMale     1.7027     0.5873    1.1291     2.568
Calories    0.9999     1.0001    0.9994     1.000

Concordance= 0.599  (se = 0.031 )
Likelihood ratio test= 11.9  on 3 df,   p=0.008
Wald test            = 11.33  on 3 df,   p=0.01
Score (logrank) test = 11.56  on 3 df,   p=0.009
```