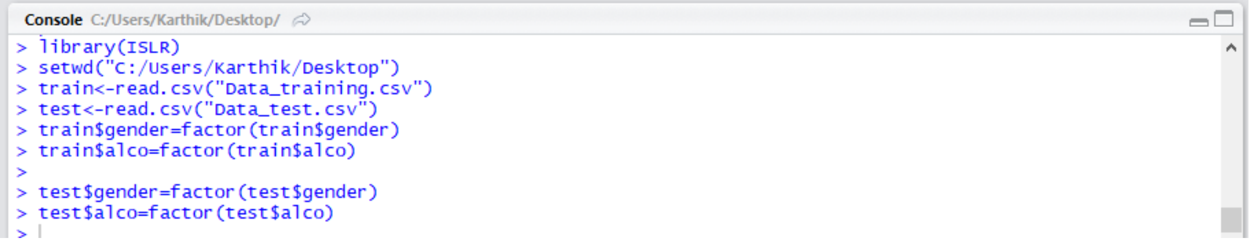
I. What analysis will you conduct?

Multiple Linear Regression Model is used in this problem as the response to be predicted is a continuous variable (Survival Time after a liver operation in ‘Days’).
Moreover as interpretability is asked for in the problem here (ie., to know which variables of patients affect their survival time and how), even though linear regression might be a bit restrictive, it gives a table of coefficient estimates and other entities useful for interpretation.

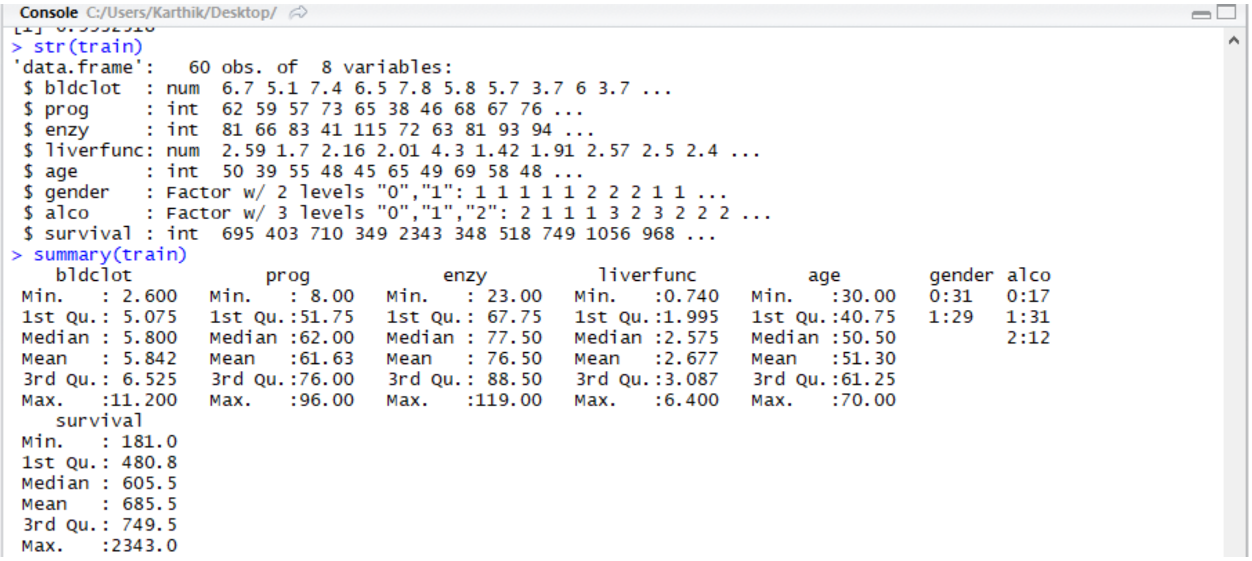
II. Find two good models for the training data.

- First, the directory is set to the folder where the data csv files are present.



```
> library(ISLR)
> setwd("C:/Users/Karthik/Desktop")
> train<-read.csv("Data_training.csv")
> test<-read.csv("Data_test.csv")
> train$gender=factor(train$gender)
> train$alco=factor(train$alco)
>
> test$gender=factor(test$gender)
> test$alco=factor(test$alco)
> |
```

- After loading the data the structure and the summary of the training and the test data are checked for.



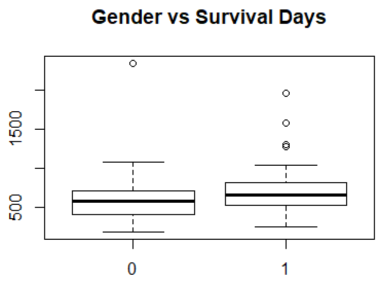
```
> str(train)
'data.frame':   60 obs. of  8 variables:
 $ bldclot  : num  6.7 5.1 7.4 6.5 7.8 5.8 5.7 3.7 6 3.7 ...
 $ prog     : int  62 59 57 73 65 38 46 68 67 76 ...
 $ enzy     : int  81 66 83 41 115 72 63 81 93 94 ...
 $ liverfunc: num  2.59 1.7 2.16 2.01 4.3 1.42 1.91 2.57 2.5 2.4 ...
 $ age      : int  50 39 55 48 45 65 49 69 58 48 ...
 $ gender   : Factor w/ 2 levels "0","1": 1 1 1 1 1 2 2 2 1 1 ...
 $ alco     : Factor w/ 3 levels "0","1","2": 2 1 1 1 3 2 3 2 2 2 ...
 $ survival : int  695 403 710 349 2343 348 518 749 1056 968 ...
> summary(train)
    bldclot            prog            enzy          liverfunc          age         gender  alco  
 Min.   : 2.600   Min.   : 8.00   Min.   : 23.00   Min.   :0.740   Min.   :30.00   0:31   0:17  
 1st Qu.: 5.075   1st Qu.:51.75   1st Qu.: 67.75   1st Qu.:1.995   1st Qu.:40.75   1:29   1:31  
 Median : 5.800   Median :62.00   Median : 77.50   Median :2.575   Median :50.50          2:12  
 Mean   : 5.842   Mean   :61.63   Mean   : 76.50   Mean   :2.677   Mean   :51.30                 
 3rd Qu.: 6.525   3rd Qu.:76.00   3rd Qu.: 88.50   3rd Qu.:3.087   3rd Qu.:61.25                 
 Max.   :11.200   Max.   :96.00   Max.   :119.00   Max.   :6.400   Max.   :70.00                 
    survival     
 Min.   : 181.0  
 1st Qu.: 480.8  
 Median : 605.5  
 Mean   : 685.5  
 3rd Qu.: 749.5  
 Max.   :2343.0  
```

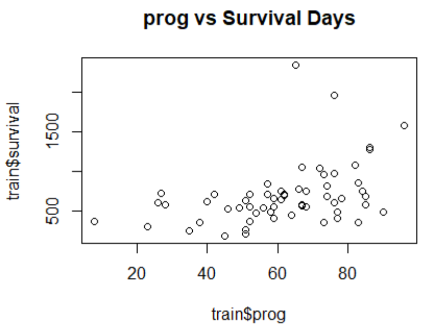
- Also some basic plots, correlations and scatter plots between survival and all other variables are checked for to get an idea of the data, to identify patterns and some inferences.

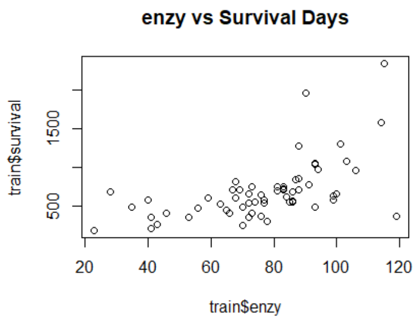
- All the categorical variables are converted into factors else for which they will be treated as continuous variables.

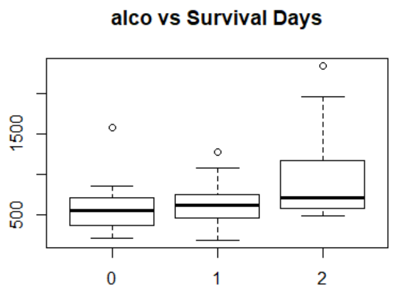
<u>PLOTS:</u>

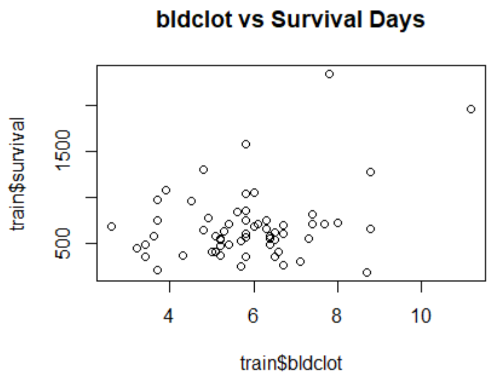
The ‘survival’ response variable is plotted against all the other predictor variables to see any relation between them.

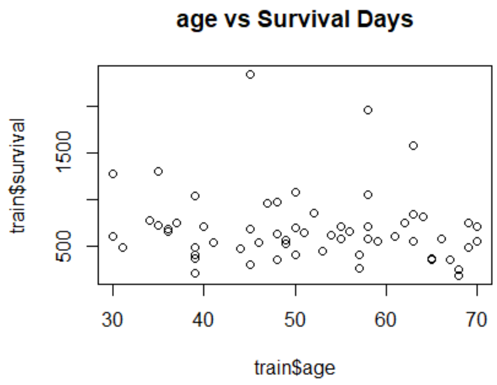


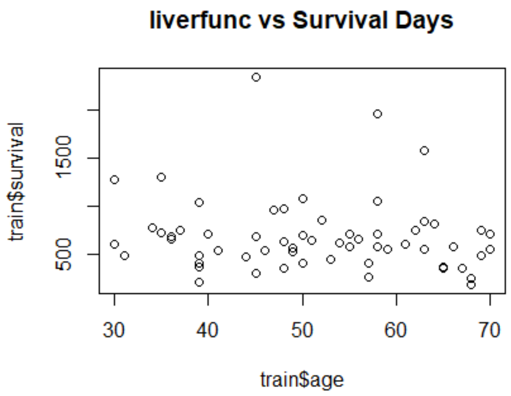






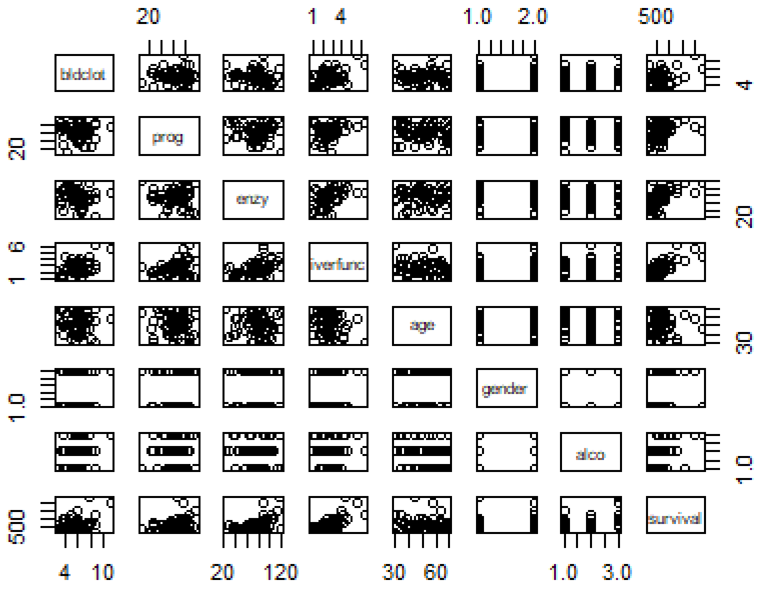






- ‘bldclot’, ‘age’ and ‘liverfunc’ don’t seem to have any relation to the response ‘survival’, but the other responses do, hence are included in the model.

Scatter Plot in Pairs[pairs(train)]:



- Then the model is fitted with all the basic predictor variables given in the question over the response “survival” for the training data using linear regression.

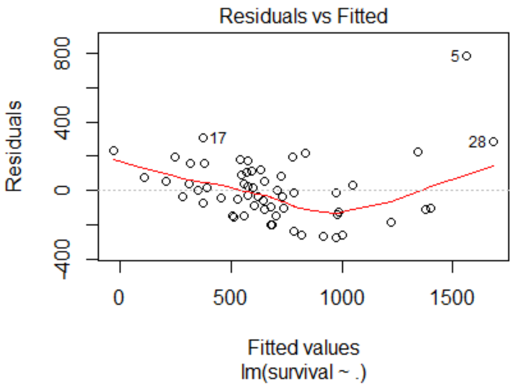
- The summary of the fitted linear model is got. Moreover the train MSE (between given response and predicted values for train data set) and test MSE(between test response and predicted values for test data set) are calculated of which test MSE is given high importance as it is the subject of interest here. Moreover the test data is also fitted using the same model as the train data and the test summary is also got.

- Next, the regression diagnostics is done.

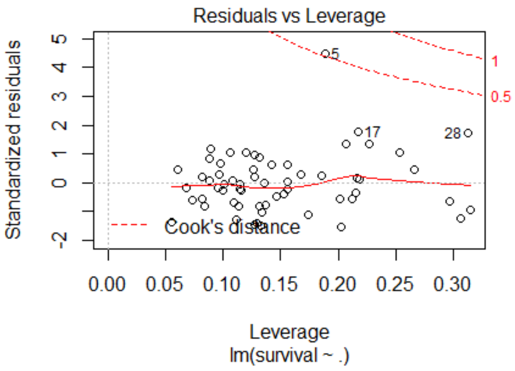
- First the Residual vs the fitted plot is plotted to check for non-linearity and heteroscedasticity.

- As was observed in the residual vs fitted value plot, there seems to be a funnel shape of increasing variance till about halfway mark of the maximum fitted value for the given data which contains around 90% of the data, hence it can be concluded that with the limited data highly concentrated for the first half of the fitted value axis, there is heteroscedasticity in the model.

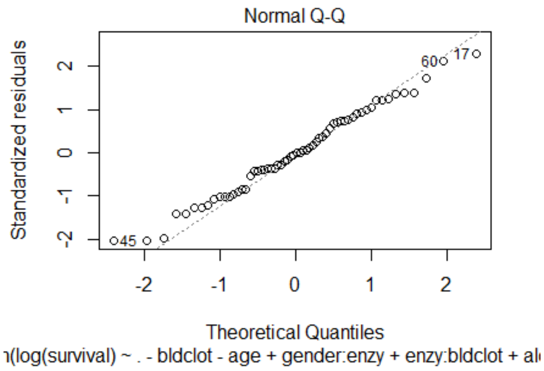
plot(lm.fit,which=1)



- Hence a 'log' transformation of the response variable('survival') is taken and the model is again modelled using linear regression to arrive at the Model 1.



plot(lm.fit,which=3)

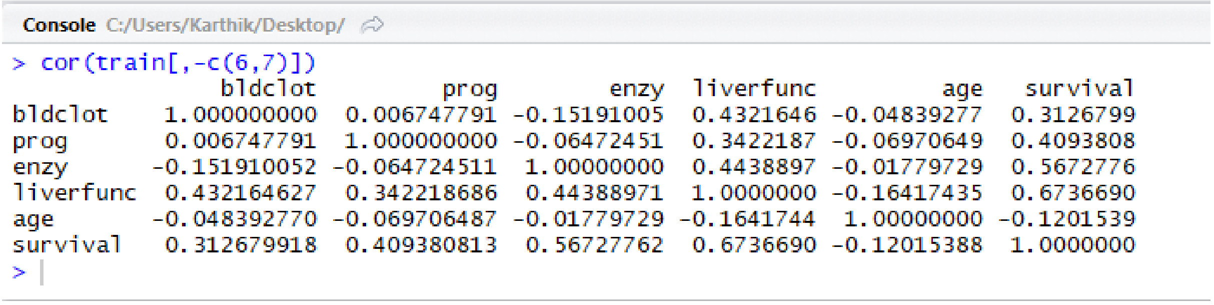


plot(lm.fit,which=2)

- The above plot says that the there is one possible outlier (point 5) which corresponds to survival days= 2343 =6.4 years which is practical, possible and hence cannot be removed just because it has a standardized residual greater than 3.
- Error terms are almost normally distributed

- The summary table for this model was derived to analyze the estimates, p,F,R^2 values, adjusted R^2 values, etc.

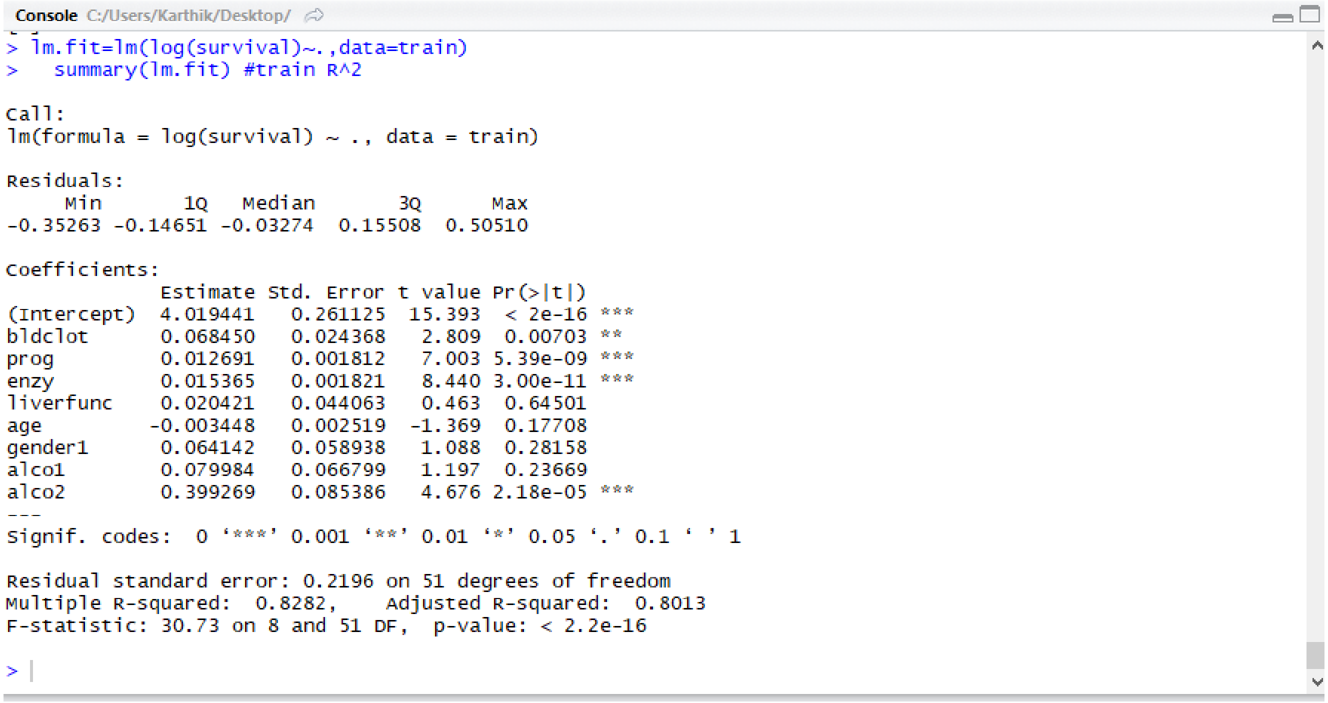
- To get a better model (Model 2), various interaction terms were tried into the model along with removing the terms that were not significant after some trial and error and according to the above scatter plots and correlation values. Hence the predictor 'liverfunc', 'bldclot' and 'age' are removed and some interaction terms were added recursively till the R^2 value reached a maximum and the MSE decreased as much as possible.



Console C:/Users/Karthik/Desktop/

```
> cor(train[,-c(6,7)])
              bldclot         prog        enzy  liverfunc         age   survival
bldclot   1.000000000  0.006747791 -0.15191005  0.4321646 -0.04839277  0.3126799
prog      0.006747791  1.000000000 -0.06472451  0.3422187 -0.06970649  0.4093808
enzy     -0.151910052 -0.064724511  1.00000000  0.4438897 -0.01779729  0.5672776
liverfunc 0.432164627  0.342218686  0.44388971  1.0000000 -0.16417435  0.6736690
age      -0.048392770 -0.069706487 -0.01779729 -0.1641744  1.00000000 -0.1201539
survival  0.312679918  0.409380813  0.56727762  0.6736690 -0.12015388  1.0000000
> |
```

MODEL 1:

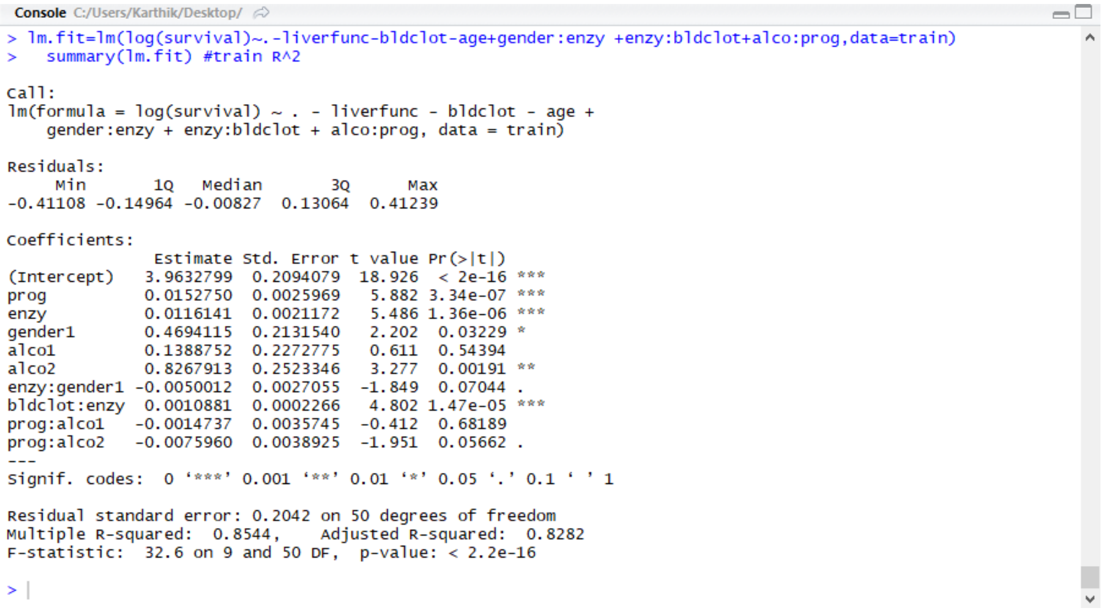


Console C:/Users/Karthik/Desktop/

```
> lm.fit=lm(log(survival)~.,data=train)
>   summary(lm.fit) #train R^2

Call:
lm(formula = log(survival) ~ ., data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.35263 -0.14651 -0.03274  0.15508  0.50510 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.019441   0.261125  15.393  < 2e-16 ***
bldclot      0.068450   0.024368   2.809  0.00703 ** 
prog         0.012691   0.001812   7.003 5.39e-09 ***
enzy         0.015365   0.001821   8.440 3.00e-11 ***
liverfunc    0.020421   0.044063   0.463  0.64501    
age         -0.003448   0.002519  -1.369  0.17708    
gender1      0.064142   0.058938   1.088  0.28158    
alco1        0.079984   0.066799   1.197  0.23669    
alco2        0.399269   0.085386   4.676 2.18e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2196 on 51 degrees of freedom
Multiple R-squared:  0.8282,	Adjusted R-squared:  0.8013 
F-statistic: 30.73 on 8 and 51 DF,  p-value: < 2.2e-16

> |
```

Log(survival)=b0 +b1(bldclot)+ b2(prog)+ b3(enzy)+ b4(liverfunc)+ b5(age)+ b6(gender)+ b7(alco)

MODEL 2:

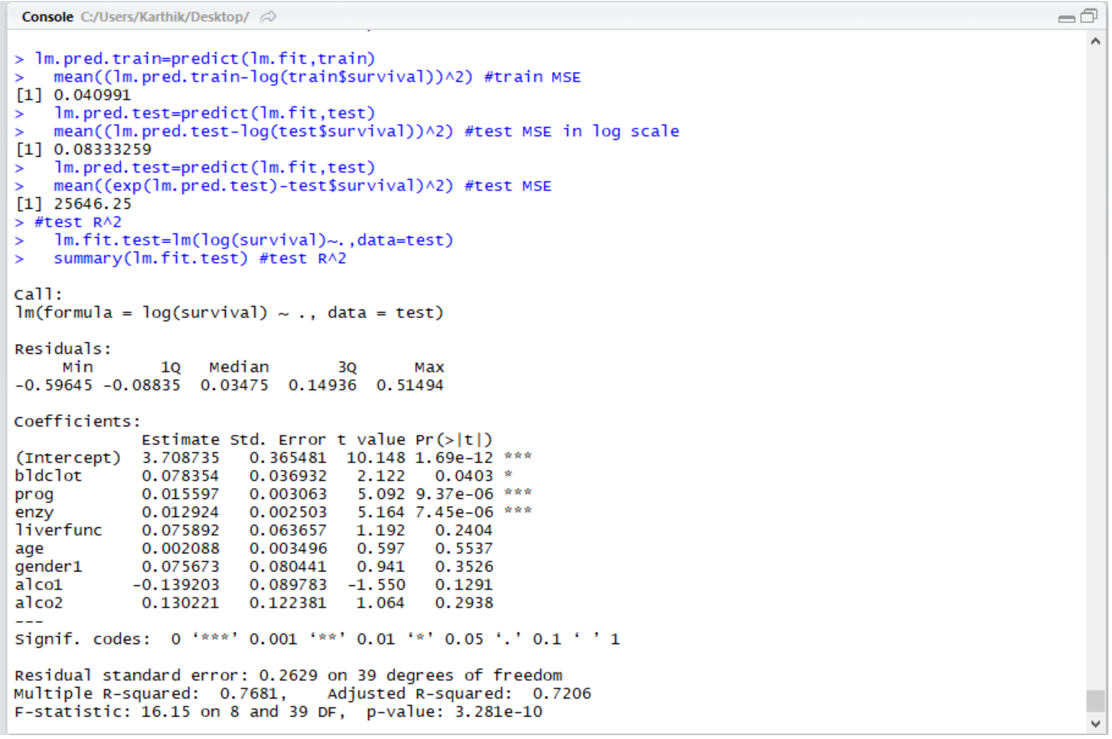


```
> lm.fit=lm(log(survival)~.-liverfunc-bldclot-age+gender:enzy +enzy:bldclot+alco:prog,data=train)
>   summary(lm.fit) #train R^2

Call:
lm(formula = log(survival) ~ . - liverfunc - bldclot - age +
    gender:enzy + enzy:bldclot + alco:prog, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.41108 -0.14964 -0.00827  0.13064  0.41239 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.9632799  0.2094079  18.926  < 2e-16 ***
prog          0.0152750  0.0025969   5.882 3.34e-07 ***
enzy          0.0116141  0.0021172   5.486 1.36e-06 ***
gender1       0.4694115  0.2131540   2.202  0.03229 *  
alco1         0.1388752  0.2272775   0.611  0.54394    
alco2         0.8267913  0.2523346   3.277  0.00191 ** 
enzy:gender1 -0.0050012  0.0027055  -1.849  0.07044 .  
bldclot:enzy  0.0010881  0.0002266   4.802 1.47e-05 ***
prog:alco1   -0.0014737  0.0035745  -0.412  0.68189    
prog:alco2   -0.0075960  0.0038925  -1.951  0.05662 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2042 on 50 degrees of freedom
Multiple R-squared:  0.8544,	Adjusted R-squared:  0.8282 
F-statistic:  32.6 on 9 and 50 DF,  p-value: < 2.2e-16

> |
```

Log(survival)= b0 + b1(prog)+ b2(enzy)+ b3(gender)+ b4(alco)+ b5(enzy*gender)+ b6(enzy*bldclot)+ b7(prog*alco)

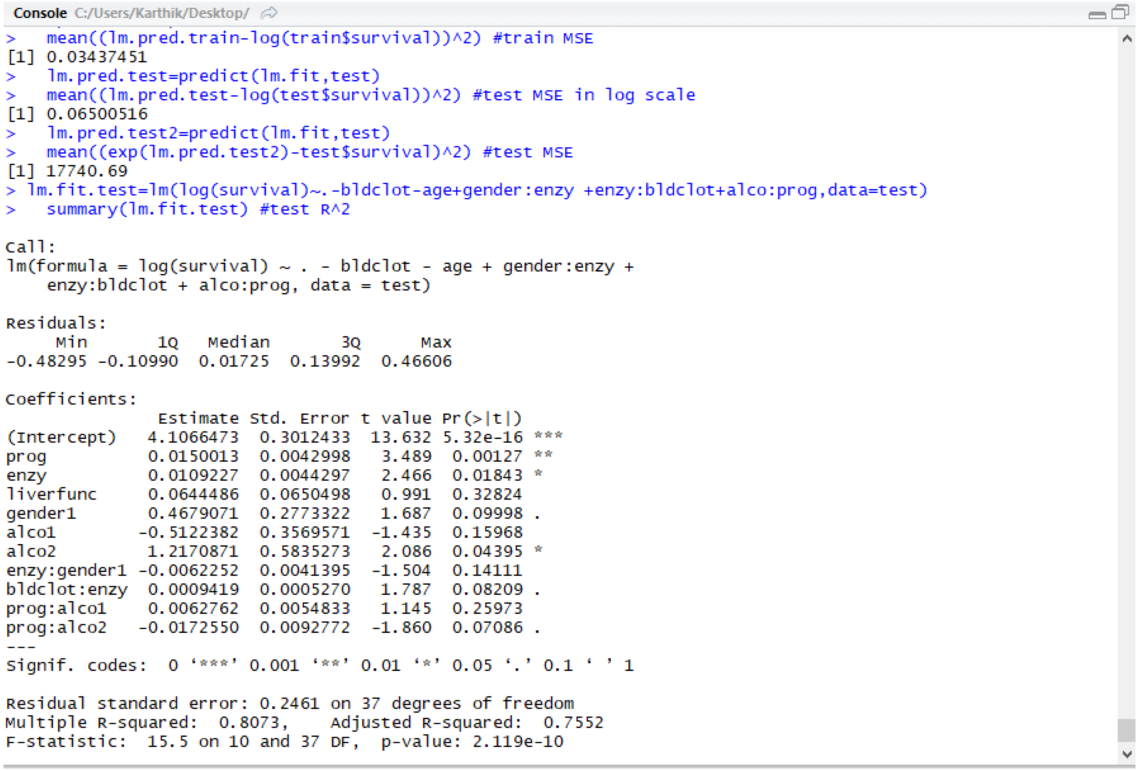
III. Find the prediction performance of the two models on the test data

MODEL 1:

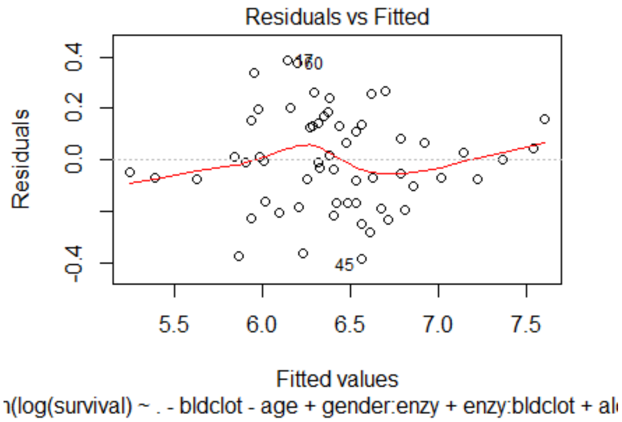


```
> lm.pred.train=predict(lm.fit,train)
>   mean((lm.pred.train-log(train$survival))^2) #train MSE
[1] 0.040991
>   lm.pred.test=predict(lm.fit,test)
>   mean((lm.pred.test-log(test$survival))^2) #test MSE in log scale
[1] 0.08333259
>   lm.pred.test=predict(lm.fit,test)
>   mean((exp(lm.pred.test)-test$survival)^2) #test MSE
[1] 25646.25
> #test R^2
>   lm.fit.test=lm(log(survival)~.,data=test)
>   summary(lm.fit.test) #test R^2

Call:
lm(formula = log(survival) ~ ., data = test)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.59645 -0.08835  0.03475  0.14936  0.51494 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.708735   0.365481  10.148 1.69e-12 ***
bldclot      0.078354   0.036932   2.122   0.0403 *  
prog         0.015597   0.003063   5.092 9.37e-06 ***
enzy         0.012924   0.002503   5.164 7.45e-06 ***
liverfunc    0.075892   0.063657   1.192   0.2404    
age          0.002088   0.003496   0.597   0.5537    
gender1      0.075673   0.080441   0.941   0.3526    
alco1       -0.139203   0.089783  -1.550   0.1291    
alco2        0.130221   0.122381   1.064   0.2938    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2629 on 39 degrees of freedom
Multiple R-squared:  0.7681,	Adjusted R-squared:  0.7206 
F-statistic: 16.15 on 8 and 39 DF,  p-value: 3.281e-10
```

MODEL 2:



```
>   mean((lm.pred.train-log(train$survival))^2) #train MSE
[1] 0.03437451
>   lm.pred.test=predict(lm.fit,test)
>   mean((lm.pred.test-log(test$survival))^2) #test MSE in log scale
[1] 0.06500516
>   lm.pred.test2=predict(lm.fit,test)
>   mean((exp(lm.pred.test2)-test$survival)^2) #test MSE
[1] 17740.69
> lm.fit.test=lm(log(survival)~.-bldclot-age+gender:enzy +enzy:bldclot+alco:prog,data=test)
>   summary(lm.fit.test) #test R^2

Call:
lm(formula = log(survival) ~ . - bldclot - age + gender:enzy +
    enzy:bldclot + alco:prog, data = test)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48295 -0.10990  0.01725  0.13992  0.46606

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.1066473  0.3012433  13.632 5.32e-16 ***
prog          0.0150013  0.0042998   3.489  0.00127 **
enzy          0.0109227  0.0044297   2.466  0.01843 *
liverfunc     0.0644486  0.0650498   0.991  0.32824
gender1       0.4679071  0.2773322   1.687  0.09998 .
alco1        -0.5122382  0.3569571  -1.435  0.15968
alco2         1.2170871  0.5835273   2.086  0.04395 *
enzy:gender1 -0.0062252  0.0041395  -1.504  0.14111
bldclot:enzy  0.0009419  0.0005270   1.787  0.08209 .
prog:alco1    0.0062762  0.0054833   1.145  0.25973
prog:alco2   -0.0172550  0.0092772  -1.860  0.07086 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2461 on 37 degrees of freedom
Multiple R-squared:  0.8073,	Adjusted R-squared:  0.7552
F-statistic:  15.5 on 10 and 37 DF,  p-value: 2.119e-10
```



There is no pattern and heteroscedasticity in this final model. Hence the model has less bias(no non- linearity) and error terms have a constant variance

IV. Which one of the two models will you choose? Why?

Model 2 will be chosen as it has a lower test MSE and a higher train R^2.
This means that the prediction is as accurate as possible and a higher percentage of variability of the response is explained by the predictors. Also, the test MSE is significant than the train MSE as the prediction of the test data set is the point of interest. Also efforts were taken to remove as many insignificant predictors (data reduction) out from the model which is true in Model 2.

Initial Model:

survival=b0 +b1(bldclot)+ b2(prog)+ b3(enzy)+ b4(liverfunc)+ b5(age)+ b6(gender)+ b7(alco)

Model 1:

Log(survival)=b0 +b1(bldclot)+ b2(prog)+ b3(enzy)+ b4(liverfunc)+ b5(age)+ b6(gender)+ b7(alco)

Model 2:

Log(survival)= b0 + b1(prog)+ b2(enzy)+ b3(gender)+ b4(alco)+ b5(enzy*gender)+ b6(enzy*bldclot)+ b7(prog*alco)

	Train(R^2) %	Train(Adjusted R^2) %	Test (R^2) %	Test(Adjusted R^2) %	Train MSE	Test MSE
Initial Model	78.07	75.12	83.55	80.67	32067.77	26577.8

Model 1	82.82	80.13	76.81	72.06	0.04(log scale)	0.083(log scale)
						25646(normal scale)

Model 2	85.44	82.82	80.73	75.52	0.034(log scale)	0.065(log scale)
						17741(normal scale)

V. Based on the chosen model, what will you tell Doctor White?

The following points need to be made to Dr.White about which predictor variables are useful in predicting the survival days after a liver operation and how they affect it.

[All the interpretations are at 0.05 or lower significance levels]

- There is a relationship which is significant between the predictors and the survival time after a liver operation from the available data of the 108 patients.
- The lifetime of the patients is exponentially distributed as is expected for the lifetime distributions.
- The predictors ‘prognostic index’ and ‘enzyme function test score’ are highly useful in predicting the survival after a liver operation in days. As the response variables are log transformed, the relation between the predictors and the log of the transform need to be analyzed.
- For 1 unit increase of the ‘prognostic index’, survival on an average increases by [(e^(0.01527)-1]*100%=1.54% [Positive Relationship]
- For 1 unit increase ‘enzyme function test score’, survival on an average increases by [(e^(0.0116+0.001*bldclot)-1]*100% [Positive Relationship]

- Female tend to survive for more days ((e^0.467)=1.6 times male) than males after the liver operation as the variable Gender1 corresponding to Female is significant and the coefficient estimate is positive.
- Having a severe history of alcohol leads to a survival time after liver operation of ((e^0.827)=2.286 times than having no history of alcohol use.

  Hence it can be concluded that ‘prognostic index’, ‘enzyme function test score’, ‘gender’, ‘alco’ are the variables that significantly affect the survival of the patient in days after a liver operation.
- Moreover, the interaction term between ‘bldclot’ and ‘enzy’ suggests that the effect of the blood clotting score on the survival days after liver operation is dependent on the enzyme function test score.
- Predictors ‘liver function test score’ and ‘age’ are not useful for predicting the survival days after a liver operation.