Data Analysis

Step:1 Data Preprocessing

The first step in any data analysis is to look at the dataset and find basic information such as the number of observations, variables, rows, columns, variable types, numeric summary, graphical summary, column and row names. It is performed using the following R commands

```
> data1 <- read.csv("Data_training.csv", header=TRUE)
> summary(data1)
    bldclot           prog            enzy          liverfunc
 Min.   : 2.600   Min.   : 8.00   Min.   : 23.00   Min.   :0.740
 1st Qu.: 5.075   1st Qu.:51.75   1st Qu.: 67.75   1st Qu.:1.995
 Median : 5.800   Median :62.00   Median : 77.50   Median :2.575
 Mean   : 5.842   Mean   :61.63   Mean   : 76.50   Mean   :2.677
 3rd Qu.: 6.525   3rd Qu.:76.00   3rd Qu.: 88.50   3rd Qu.:3.087
 Max.   :11.200   Max.   :96.00   Max.   :119.00   Max.   :6.400
      age            gender           alco          survival
 Min.   :30.00   Min.   :0.0000   Min.   :0.0000   Min.   : 181.0
 1st Qu.:40.75   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.: 480.8
 Median :50.50   Median :0.0000   Median :1.0000   Median : 605.5
 Mean   :51.30   Mean   :0.4833   Mean   :0.9167   Mean   : 685.5
 3rd Qu.:61.25   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.: 749.5
 Max.   :70.00   Max.   :1.0000   Max.   :2.0000   Max.   :2343.0
> str(data1)
'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   : int  0 0 0 0 0 1 1 1 0 0 ...
 $ alco     : int  1 0 0 0 2 1 2 1 1 1 ...
 $ survival : int  695 403 710 349 2343 348 518 749 1056 968 ...
```

The dataset is analyzed for missing values using the following command

```
> is.na(data1)
```

As the output is FALSE for all data points we shall proceed to the next step of analysis.

Encoding categorical variables: From the analysis it is identified that the variables 'gender' and 'alco' are categorical whereas they are represented as integers in the dataset. Those variables are encoded as follows

```
> data1$gender <- factor(data1$gender, levels=c(0,1), labels=c("male", "female"))
> data1$alco <- factor(data1$alco, levels=c(0,1,2), labels=c("none", "moderate", "severe"))
```

Multi collinearity analysis: Multi collinearity masks the real effect of a predictor variable on the response due to interaction. Thus, dataset should be analyzed for significant multicollinearity among predictor variables.

```
> cor(data1[, c(-6,-7,-8)])
              bldclot         prog        enzy  liverfunc         age
bldclot    1.000000000  0.006747791 -0.15191005  0.4321646 -0.04839277
prog       0.006747791  1.000000000 -0.06472451  0.3422187 -0.06970649
enzy      -0.151910052 -0.064724511  1.00000000  0.4438897 -0.01779729
liverfunc  0.432164627  0.342218686  0.44388971  1.0000000 -0.16417435
age       -0.048392770 -0.069706487 -0.01779729 -0.1641744  1.00000000
```

The result of cor() function implies that there is no significant correlation among predictor variables. The above output also indicates the direction of relationship between predictor variables (negative sign indicates inverse relation and positive sign indicates direct relation among each pair).

Step:2 Deciding the type of statistical learning method

- *Supervised or unsupervised:* As the dataset has a response/dependent variable to supervise the analysis 'supervised' learning approach will be a good fit.
- *Regression or Classification*: As the response variable 'survival' is a continuous numeric variable the given problem falls under 'Regression'.
- *Prediction vs Inference*: The reasons for estimating a statistical function for the given dataset are to predict the survival time of test set and also to understand the relationship among the predictors and response, identify the most significant predictors of survival time. Thus, the model is expected to predict and infer values from dataset.
- *Parametric vs Non-parametric*: Due to less number of observations in the dataset parametric method holds good.
- *Flexible vs Restrictive*: As it is necessary to figure out the relationship between each individual predictor and response variable and in order to avoid overfitting and minimize test MSE restrictive approach is used.

Thus, a supervised, parametric, restrictive, regression model capable of predicting and inferring relationships among variables from data is to be implemented. In addition to that linear, non-linear and data transformations are to be applied on various models and analyzed.

Selecting two best models for the training dataset

Two find two best models for the dataset, each and every model will be analyzed based on the following factors:

1. Adjusted R-squared – Indicates percentage variation of response explained due to predictors in the model. The model with high Adjusted R-squared is the best

2. RSE – Indicates the standard deviation of residual errors. The model with minimum RSE is the best

3. p-value – Indicates the significance/capability of the model is predicting the response variable. Small p-values (are preferred) indicate that the model has good predicting ability

4. Outliers – Values away from data in Standardized residuals vs Leverage indicates outliers, leverage points

5. Normality – If all the points lie on the straight diagonal line of QQ plot the residuals are normally distributed. *The following assumptions are applicable only for linear regression*

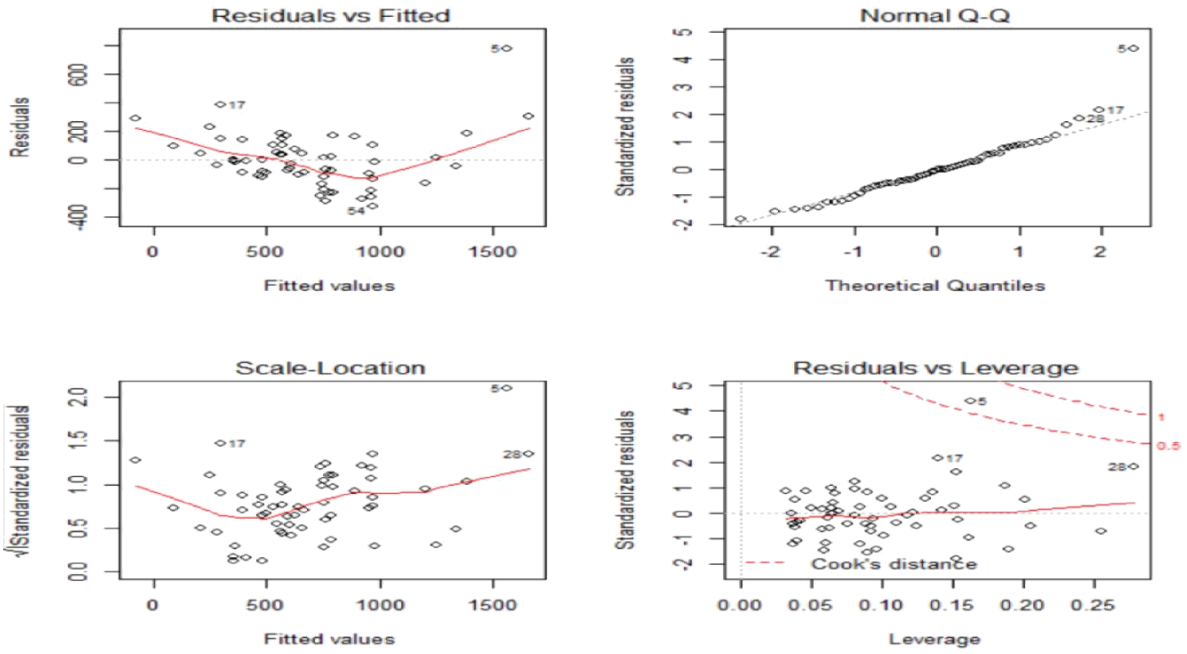
6. Linearity – No pattern in Residual vs Fitted plot indicates linearity.

7. Homogeneity of variance - No pattern in Standardized residuals vs Fitted plot indicates homogeneity. The model with homogenous variance will perform better for linear regression.

- **Model 1: Linear regression (without interaction terms)**

  To start from an initial model Linear regression is fitted with all given predictors. To reduce the complexity of the model and improve the goodness-of-fit, backward elimination is performed to remove insignificant predictors from the model.

```
> fit.lm4 <- lm(survival~bldclot+prog+enzy+alco, data1)
> summary(fit.lm4)

Call:
lm(formula = survival ~ bldclot + prog + enzy + alco, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-323.27 -105.84  -10.29  103.45  777.57

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1366.326    174.237  -7.842 1.78e-10 ***
bldclot         85.277     16.759   5.089 4.68e-06 ***
prog             9.471      1.379   6.867 6.77e-09 ***
enzy            11.825      1.219   9.697 2.01e-13 ***
alcomoderate    13.798     58.724   0.235 0.815121
alcosevere     291.110     74.811   3.891 0.000276 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 193.8 on 54 degrees of freedom
Multiple R-squared:  0.7689,    Adjusted R-squared:  0.7475
F-statistic: 35.93 on 5 and 54 DF,  p-value: 5.142e-16
```

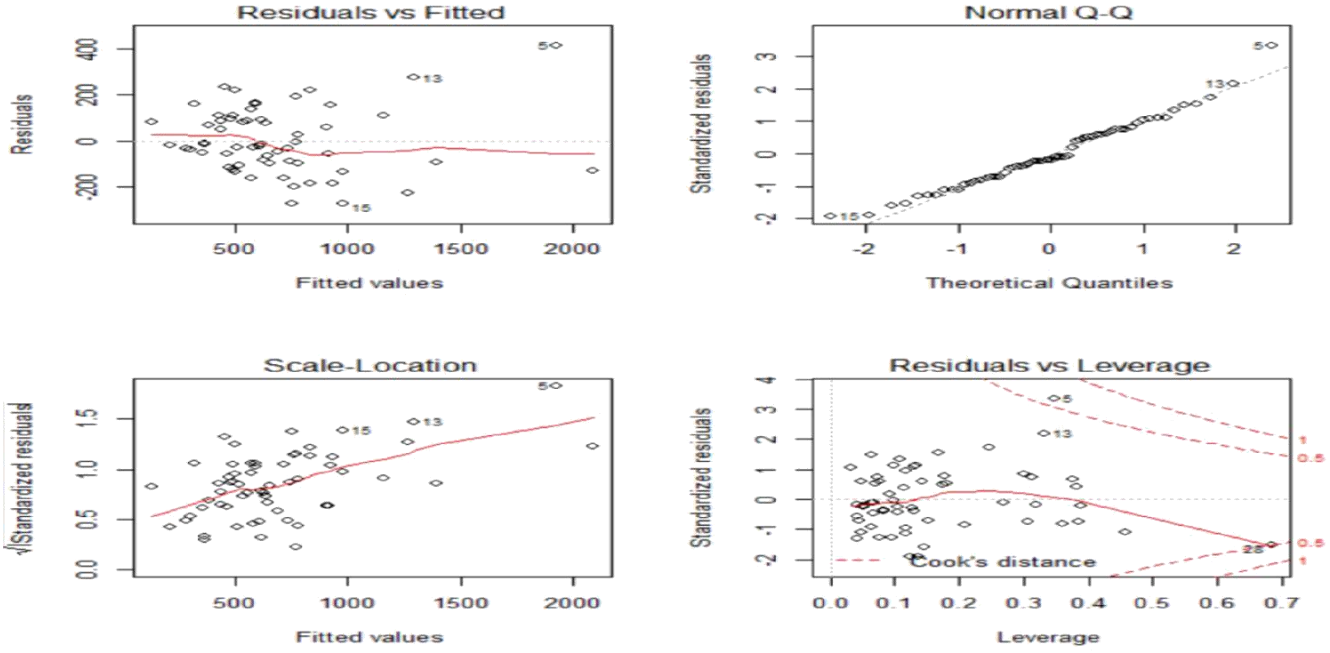


After removing all insignificant predictors bldclot, prog, enzy and alco are the significant predictors.The diagnostic plot indicates that the residuals are almost linear, has nearly homogenous variance with normal distribution and a few outliers. Adj R-squared = 0.7475, p-value=5.142e^-16 , RSE=193.8

- **Model2: Linear regression (with interaction terms)**

To further analyze for a better model, the interaction terms are included. Initially all interaction terms are included as shown below. (As for interaction among three variables the model becomes more complicated and flexible many observations are required by the approach to estimate the parameters. As this problem has few observations we are limiting our analysis to two-factor interactions). To reduce the complexity of the model and improve the goodness-of-fit, backward elimination is performed to remove insignificant predictors and interaction terms from the model.

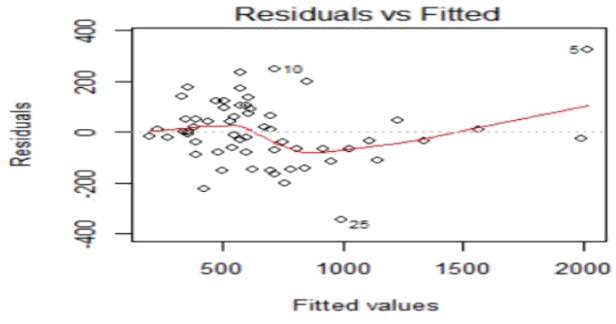
```
> fit.lmint3 <- lm(survival~(bldclot+prog+enzy+alco+bldclot:alco+enzy:alco), data1)
> summary(fit.lmint3)

Call:
lm(formula = survival ~ (bldclot + prog + enzy + alco + bldclot:alco +
    enzy:alco), data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-274.11  -99.83  -24.87   91.49  415.15

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         -812.515    223.396  -3.637 0.000652 ***
bldclot               10.442     27.280   0.383 0.703517
prog                   9.616      1.161   8.281 6.13e-11 ***
enzy                   9.901      1.852   5.345 2.23e-06 ***
alcomoderate        -201.062    289.762  -0.694 0.490965
alcosevere         -1284.925    312.227  -4.115 0.000144 ***
bldclot:alcomoderate   57.035     34.730   1.642 0.106814
bldclot:alcosevere    146.845     36.528   4.020 0.000196 ***
enzy:alcomoderate      -1.315      2.343  -0.561 0.577246
enzy:alcosevere         8.998      2.652   3.393 0.001361 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 153.8 on 50 degrees of freedom
Multiple R-squared:  0.8653,    Adjusted R-squared:  0.841
F-statistic: 35.68 on 9 and 50 DF,  p-value: < 2.2e-16
```

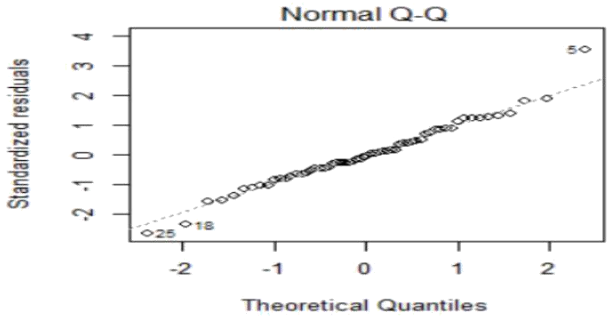


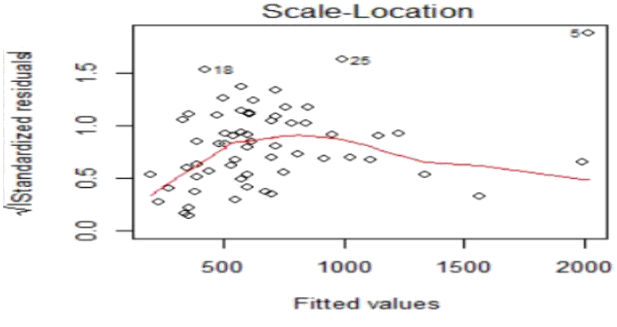
After removing all insignificant predictors, the diagnostic plot indicates that the residuals are almost linear, has nearly homogenous variance with normal distribution and a few outliers. Adj R-squared = 0.841, p-value=2.2e^-16, RSE=153.8. The residual vs fitted plot indicates that the linear model fits the data well. But, there are some outliers and it has nearly homogenous variance.

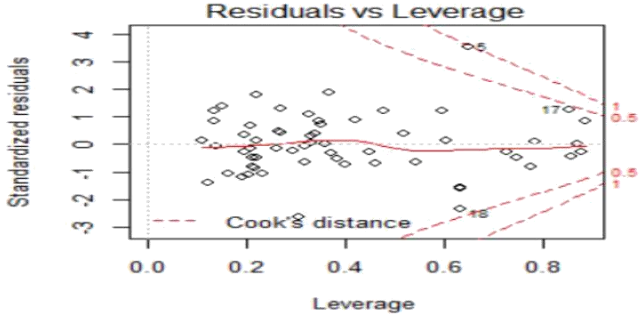
- **Model 3: Polynomial Regression**
  Fitting polynomial regression to the dataset

```
> poly2 <- lm(survival~polym(bldclot,prog,enzy,liverfunc,age,degree=2)+gender+alco,data=data1)
> summary(poly2)

Call:
lm(formula = survival ~ (bldclot + prog + enzy + liverfunc +
    age + gender + alco)^2, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-265.69  -59.90   10.62   58.01  214.62

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.481e+02  1.159e+03   0.559   0.5812
bldclot                  1.748e+02  1.598e+02   1.094   0.2849
prog                     8.132e+00  1.814e+01   0.448   0.6579
enzy                    -5.305e+00  1.090e+01  -0.487   0.6310
liverfunc               -9.236e+02  3.779e+02  -2.444   0.0222 *
age                     -1.267e+01  2.367e+01  -0.535   0.5973
genderfemale             1.135e+03  5.767e+02   1.968   0.0607 .
alcomoderate            -4.742e+02  5.445e+02  -0.871   0.3925
alcosevere              -1.835e+03  9.513e+02  -1.929   0.0657 .
bldclot:prog            -1.184e+00  1.890e+00  -0.626   0.5371
bldclot:enzy             2.043e-01  1.143e+00   0.179   0.8596
bldclot:liverfunc       -6.446e+00  2.145e+01  -0.300   0.7664
bldclot:age             -6.528e-01  2.608e+00  -0.250   0.8045
bldclot:genderfemale    -6.202e+01  4.728e+01  -1.312   0.2020
bldclot:alcomoderate     3.815e+01  5.007e+01   0.762   0.4535
bldclot:alcosevere       1.837e+02  1.027e+02   1.789   0.0862 .
prog:enzy                1.031e-01  9.209e-02   1.120   0.2739
prog:liverfunc           2.385e+00  2.396e+00   0.995   0.3296
prog:age                -9.487e-02  1.755e-01  -0.540   0.5939
prog:genderfemale       -5.815e+00  3.242e+00  -1.794   0.0855 .
prog:alcomoderate        2.147e+00  5.051e+00   0.425   0.6746
prog:alcosevere          1.810e+00  6.389e+00   0.283   0.7794
enzy:liverfunc           4.582e+00  1.924e+00   2.382   0.0255 *
enzy:age                 3.826e-02  1.724e-01   0.222   0.8263
enzy:genderfemale       -9.219e+00  4.128e+00  -2.233   0.0351 *
enzy:alcomoderate       -4.195e+00  3.386e+00  -1.239   0.2273
enzy:alcosevere          1.339e+01  7.253e+00   1.847   0.0772 .
liverfunc:age            6.071e+00  4.461e+00   1.361   0.1862
liverfunc:genderfemale   1.774e+02  7.965e+01   2.228   0.0355 *
liverfunc:alcomoderate   6.642e+01  1.123e+02   0.591   0.5598
liverfunc:alcosevere    -1.238e+02  2.035e+02  -0.608   0.5486
age:genderfemale        -2.414e+00  5.354e+00  -0.451   0.6562
age:alcomoderate         5.680e+00  5.733e+00   0.991   0.3317
age:alcosevere           4.753e+00  7.834e+00   0.607   0.5498

Residual standard error: 156 on 36 degrees of freedom
Multiple R-squared:  0.9001,    Adjusted R-squared:  0.8363
F-statistic:  14.1 on 23 and 36 DF,  p-value: 7.427e-12
```





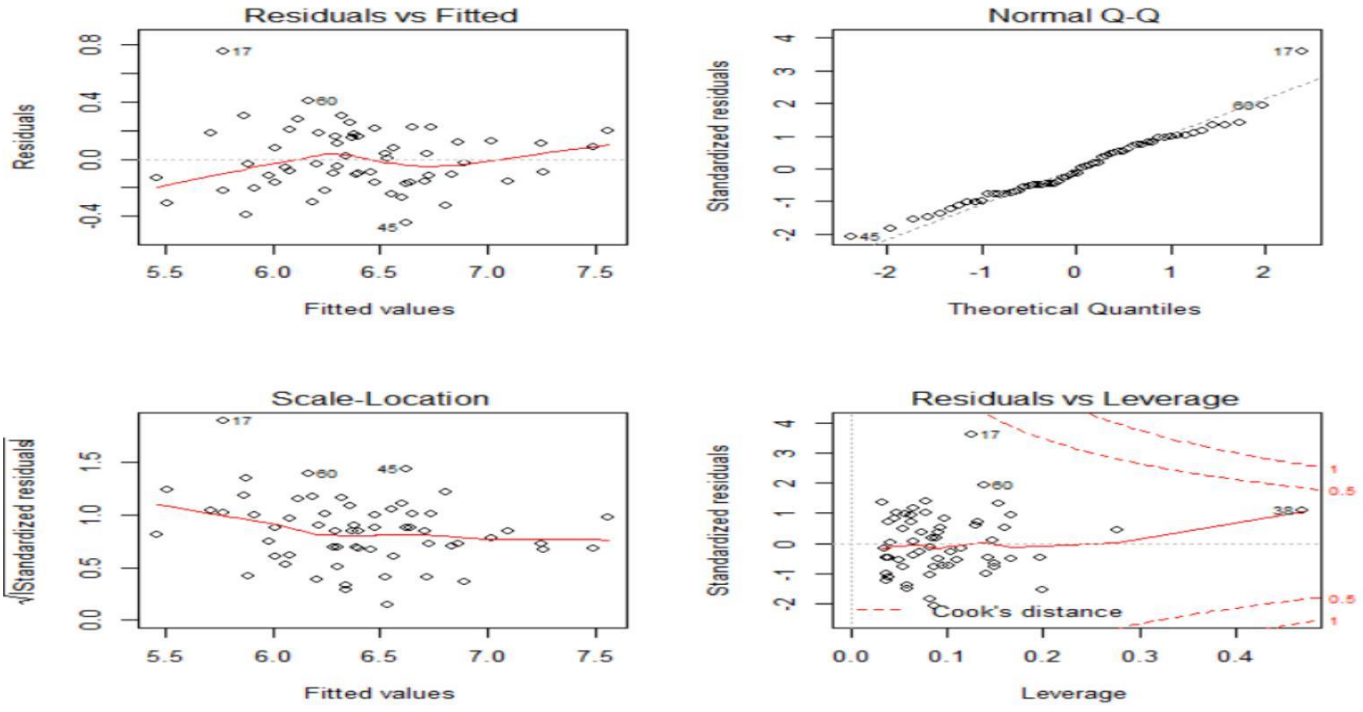




The plot indicates that the residuals have a curved pattern indicating non-linearity, variance is not homogenous, has normal distribution and a few outliers. Adj R-squared = 0.8363, p-value=7.427e^-12 , RSE=156

- **Model 4: Regression with transformed variables**
  Applying log transformation to the response and predictor variable 'prog'

```
> fit.lm6 <- lm(log(survival)~((bldclot)+log(prog)+(enzy)+(alco)), datal)
> summary(fit.lm6)

Call:
lm(formula = log(survival) ~ ((bldclot) + log(prog) + (enzy) +
    (alco)), data = datal)

Residuals:
     Min       1Q   Median       3Q      Max
-0.44387 -0.15361 -0.03108  0.15522  0.75381

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.085662   0.349828   5.962 1.96e-07 ***
bldclot      0.077006   0.019390   3.971 0.000213 ***
log(prog)    0.607121   0.072273   8.400 2.24e-11 ***
enzy         0.017090   0.001422  12.018  < 2e-16 ***
alcomoderate 0.047785   0.068139   0.701 0.486138
alcosevere   0.379279   0.086586   4.380 5.49e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2242 on 54 degrees of freedom
Multiple R-squared:  0.8104,    Adjusted R-squared:  0.7928
F-statistic: 46.15 on 5 and 54 DF,  p-value: < 2.2e-16
```



The plot indicates that the residuals have a slightly curved pattern indicating non-linearity, variance is not homogenous, distribution is deviating from normal towards the end and a few outliers. Adj R-squared = 0.7928, p-value=2.2e^-16 , RSE=0.2242

From the above-mentioned models, I would select Model2(with significant interaction terms) and Model3(Polynomial model) based on the specified criteria discussed above. In addition to that transformed data

are difficult to re-transform and hence it makes the inference complicated. Thus, model4 is not considered as the best model.

Finding prediction performance of two best models on test data

Model with significant interaction terms:

```
> predint <- predict(fit.lmint3,test.set)
> mse <- mean((test.set$survival-predint)^2)
> mse
[1] 27781.38
```

The test data results are predicted using parameters estimated from training set and the MSE is calculated as 27781.38

Polynomial model:

```
> predpoly <- predict(poly2,test.set)
> mse <- mean((test.set$survival-predpoly)^2)
> mse
[1] 70793.75
```

The test data results are predicted using parameters estimated from training set and the MSE is calculated as 70793.75

Choosing the final model

The quality of fit of a model is indicated by its Test MSE and Adjusted R-squared values. Among the two models the one with significant interaction terms has the minimum MSE of 27781.38 and highest adjusted R2 of 0.841. Thus, the model with significant predictors and interaction terms is the best among all the models in predicting the test set values with high accuracy and minimum mean square error.

Conclusion

The best model chosen above indicates bldclot, prog, enzy. alcosevere and the interactions between alcosevere and enzy and alcosevere and bldclot as significant. Among these bldclot and alcosevere has negative slope. Based on this, I conclude that

Having high blood clotting score and severe alcohol consumption history significantly reduces the survival time of patients after the liver operation. Whereas high prognostic index and enzyme scores significantly increases the survival time of patients after liver operation.

Thus, the four variables,

a. Blood clotting score
b. prognostic index
c. enzyme function test score
d. alcohol consumption – severe level

should be taken care of while treating patients. Of those variables, severe alcohol consumption history has significant interaction with enzyme function as well as blood clotting score. Hence, it is the most important variable to be monitored and act upon.