Test2 - Survival Prediction Model

Rick Galbo

November 20, 2015

Abstract

This research is to determine a proper model for predicting the rates of survival among patients who underwent this medical procedure. In order to predict the length of post-op patient's survival, we employ statistical methods to select from eight predictor variables and build the best predictive model from these. The independent variables are blood clotting score, prognostic index, enzyme function test score, Liver function test score, age, number of years employed, gender and alcohol use.

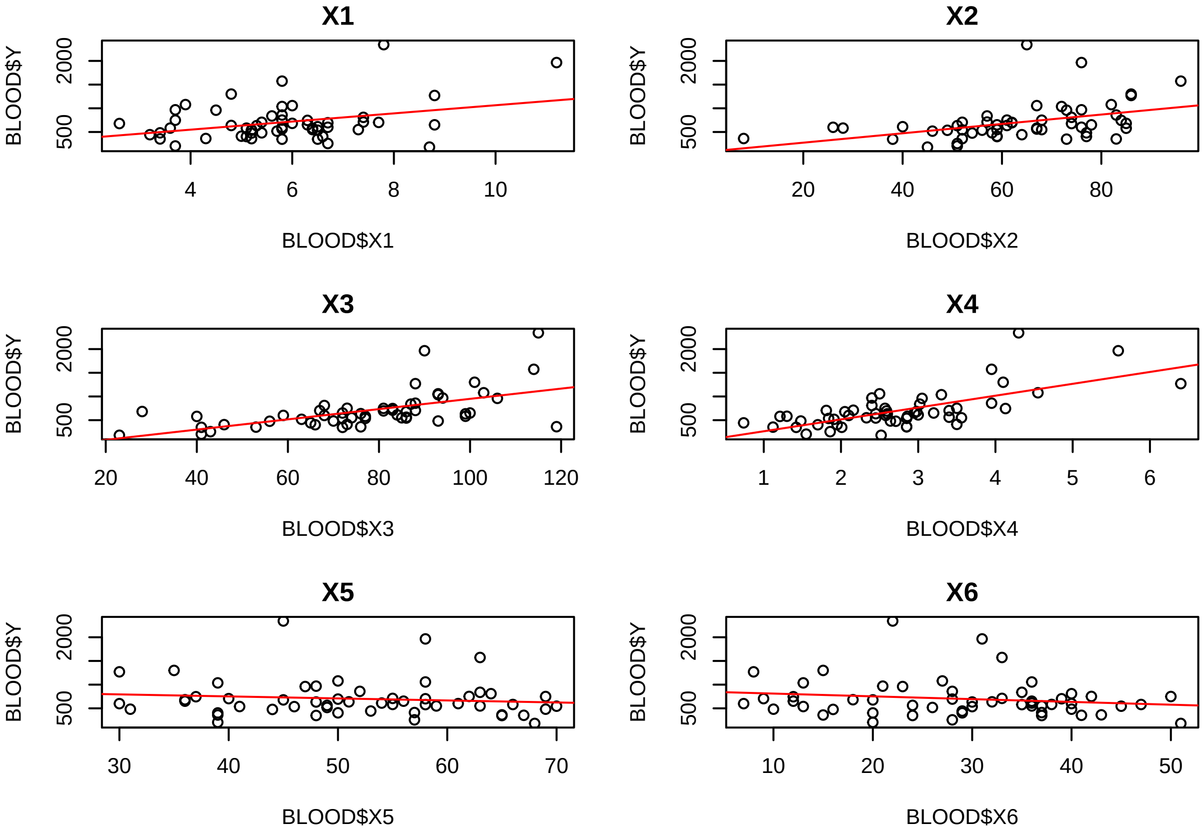
Intro

To build a statistically significant model for survival length of post-op patients, we will identify the most predictive variables which will be combined to produce a model to best describe our data.

Methodology

Dependent and Independent Variable Relationships

The first step Was to review the individual numerical and categorical variables to determine their statistical significance. Here are the scatter plots for the numerical variables with the underlying linear model overlay:



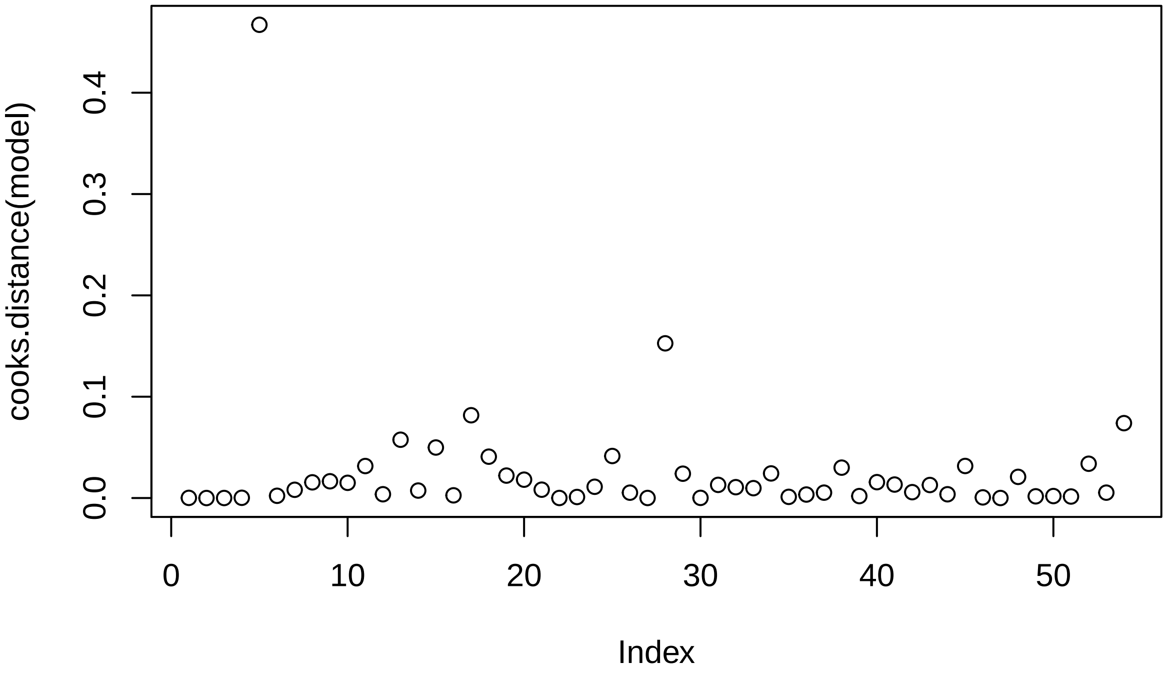
Full Model

We can see that the first four variables have decent predictability while the last two show little to no relation. Next we created linear models for the categorical variables to check their predictability.

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
##     data = BLOOD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -288.53 -133.68   -9.18   89.64  788.76
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1132.818    269.581  -4.202 0.000127 ***
## X1             63.041     25.159   2.506 0.015994 *
## X2              9.055      1.981   4.571 3.92e-05 ***
## X3              9.976      1.866   5.347 3.04e-06 ***
## X4             48.645     47.123   1.032 0.307570
## X5             -2.131      8.715  -0.245 0.807921
## X6              1.181      8.299   0.142 0.887463
## X7             16.724     59.423   0.281 0.779685
## X8              7.503     65.692   0.114 0.909582
## X9            320.283     86.061   3.722 0.000559 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 203.7 on 44 degrees of freedom
## Multiple R-squared:  0.7819, Adjusted R-squared:  0.7373
## F-statistic: 17.53 on 9 and 44 DF,  p-value: 7.465e-12
```

Influential Values

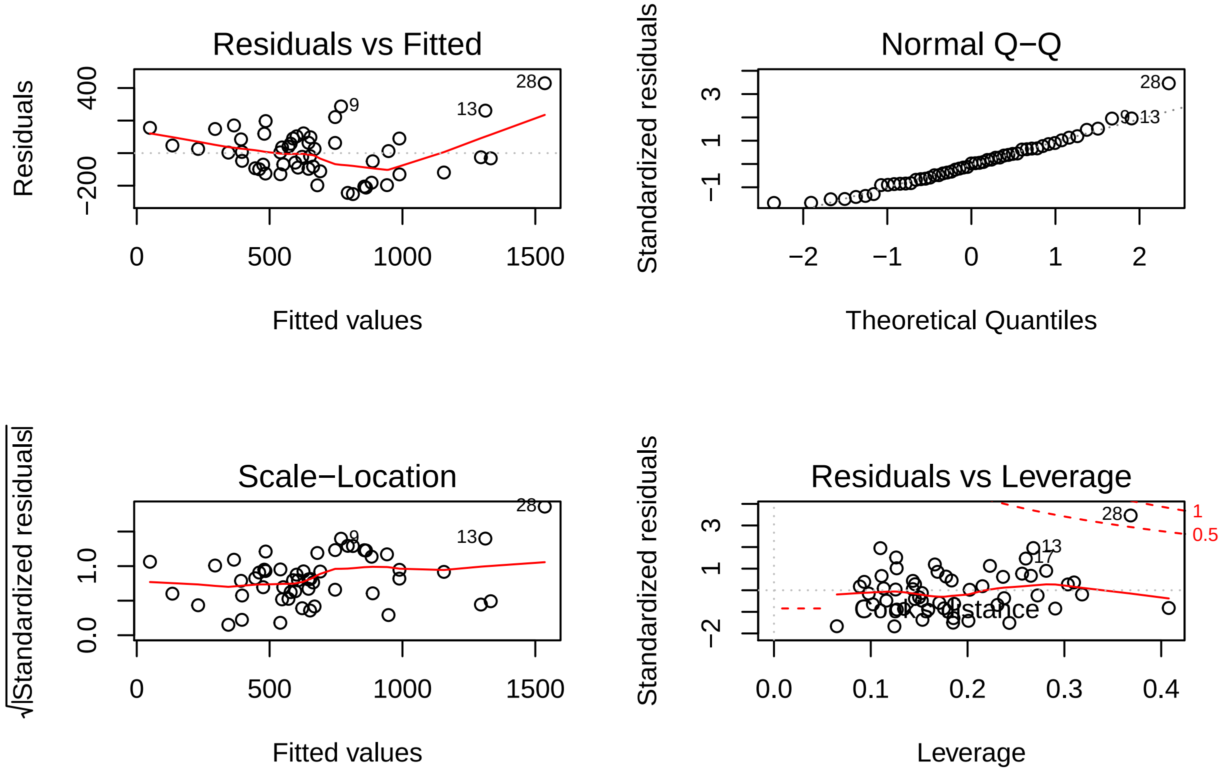
Using the Cook's D statistic we will identify influential points. These points will apear to be outlierswhen the Cook's D statistic is plotted. We do identify one significantly influential poin after plotting Cook's D here:



Checking for Outliers

It is important after discovering an influential value to check if it is considered an outlier. This value may be affecting the accuracy of the model that has been built thus far. Upon inspection of the graphical output from the previous graph as well ass the one dimentional scatter plots which are created using the package 'mvoutliers', we can see that there are indeed a few extreme values present in the data. We will test the improvement of the model with that extreme value excluded.

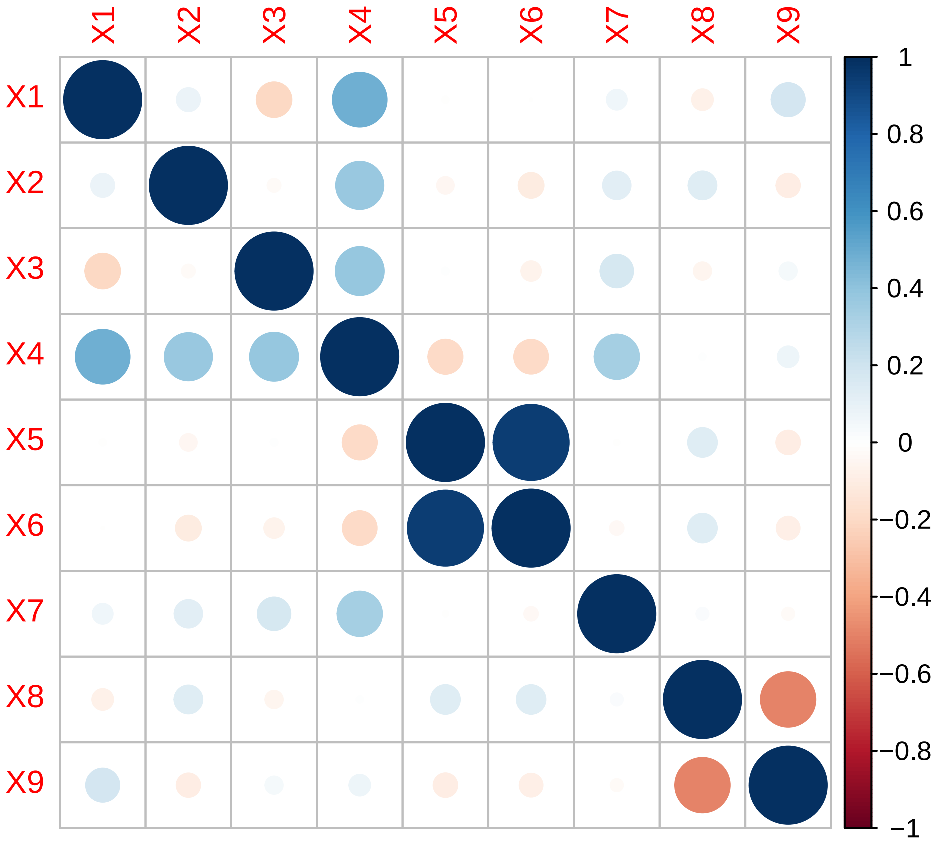
```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9,
##     data = BLOOD[-c(5), ])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -251.53  -96.47    3.17   89.66  429.42
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -977.9332   208.4420  -4.692 2.76e-05 ***
## X1            47.6265    19.4761   2.445 0.018640 *
## X2             8.6045     1.5205   5.659 1.14e-06 ***
## X3             8.1149     1.4675   5.530 1.76e-06 ***
## X4            51.5911    36.1226   1.428 0.160452
## X5             0.3586     6.6943   0.054 0.957524
## X6            -0.7038     6.3699  -0.110 0.912530
## X7            59.3557    46.1679   1.286 0.205447
## X8             5.4013    50.3531   0.107 0.915075
## X9           249.0801    67.1583   3.709 0.000592 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 156.1 on 43 degrees of freedom
## Multiple R-squared:  0.8138, Adjusted R-squared:  0.7748
## F-statistic: 20.88 on 9 and 43 DF,  p-value: 5.717e-13
```



With the removal of the largest outlier we see a great improvement in the fit statistics of the model before the predictor variables have even been screened.

Checking for Multicollinearity

With all of the variables included in this initial model it is important to check for multicollinearity. Using the package 'corrplot' we are able to create a graphical correlation matrix to visually identify highly correlated pairs of predictor variables.



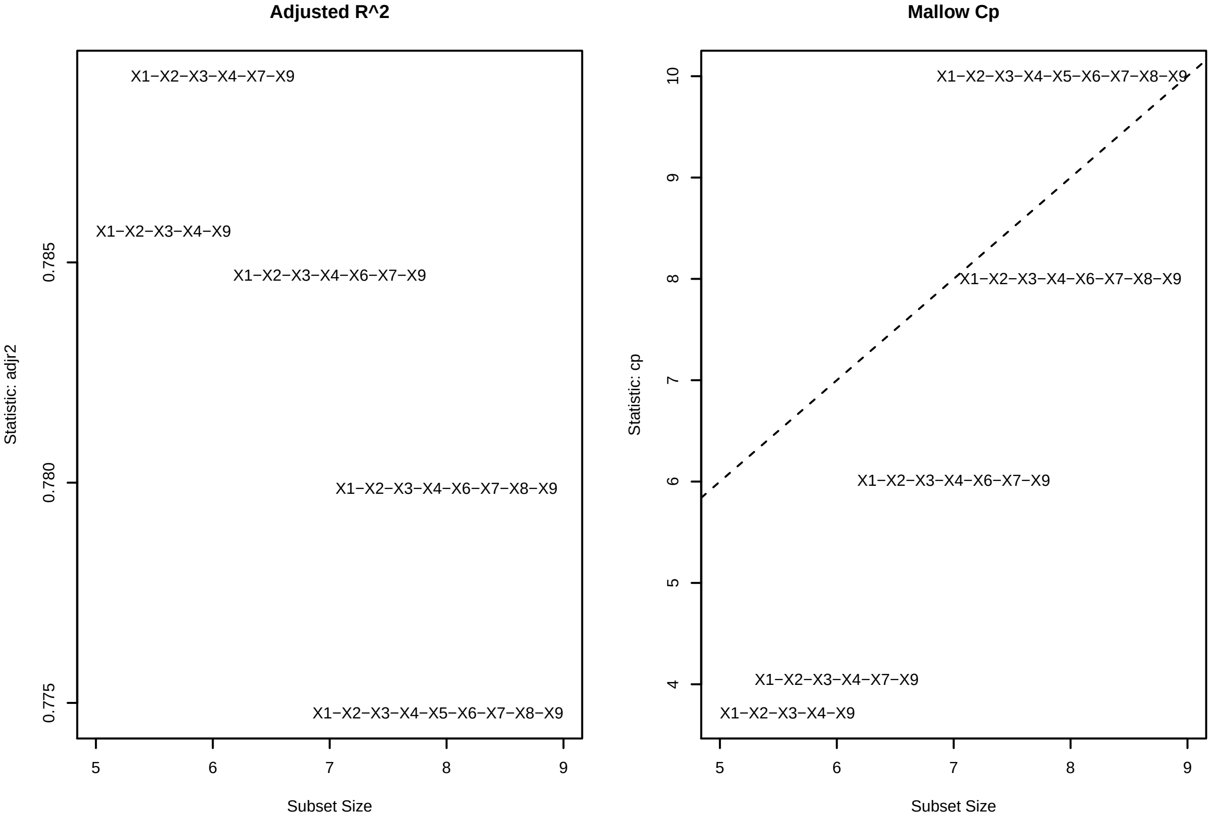
As we would expect the variables X5: Age and X6: Years employed are very highly correlated. The variable X4 is also considerably correlated to X1, X2 and X3. Removing the two variables X6 and X4 we get a new model:

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X5 + X7 + X8 + X9, data = BLOOD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -261.94  -97.66  -12.09   97.58  448.34
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1055.190    181.120  -5.826 5.66e-07 ***
## X1             66.447     14.330   4.637 3.05e-05 ***
## X2              9.597      1.302   7.374 2.85e-09 ***
## X3              9.404      1.092   8.613 4.50e-11 ***
## X5             -1.247      1.962  -0.636 0.528237
## X7             77.679     44.356   1.751 0.086715 .
## X8             10.071     50.297   0.200 0.842210
## X9           246.521     67.206   3.668 0.000644 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 156.3 on 45 degrees of freedom
## Multiple R-squared:  0.8046, Adjusted R-squared:  0.7742
## F-statistic: 26.47 on 7 and 45 DF,  p-value: 5.773e-14
```

Without the inclusion of the highly correlated terms we can see that the adjusted R-squared value has went up, meaning that the model better explains the variance of the independent variable. To see if we have built the model that gives the highest adjusted R-Squared value, a good method to use is a step wise regression. This will build the best first-order model out of our given predictor variables. Just to make sure that nothing was missed when removing the correlated variables, they have been included in the pool of choices for the step wise.

R-Squared and Mallow CP

A way to evaluate the number of predictor variables used in a first order linear model is to look at the R-Squared and the Mallow CP statistic as functions of the number of predictor variables included in the model.



We can see that the model containing six variables yeilds the maximum R-Squared value while still maintaining a relatively small Mallow CP value. A Mallow CP value small and close to the number of variables in the model is indicative of a precise model.

Stepwise Regression

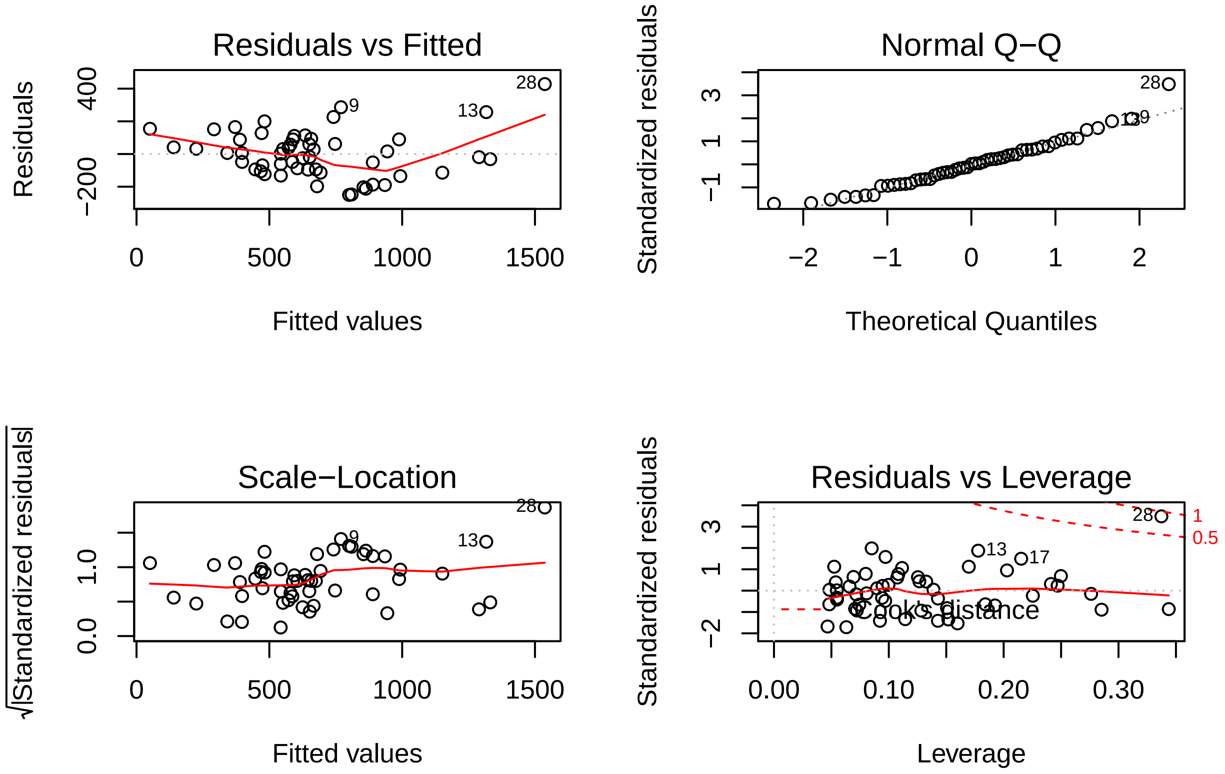
The best first order model that was chosen by the step wise regression consists of the variables X4: Liver function test score, X9: Severe alcohol use (1 = yes, 0 = no), X1: Blood clotting score, X2: Prognostic index and X3: Enzyme function test score. the summary from this model is given here:

```
##
## Call:
## lm(formula = Y ~ X4 + X3 + X2 + X9 + X1 + X7, data = BLOOD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -250.34  -95.61    2.27   89.09  428.10
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -979.280    165.624  -5.913 3.91e-07 ***
## X4            52.388     32.114   1.631   0.1097
## X3             8.118      1.310   6.195 1.48e-07 ***
## X2             8.644      1.398   6.182 1.54e-07 ***
## X9           246.640     56.867   4.337 7.81e-05 ***
## X1            47.316     18.075   2.618   0.0119 *
## X7            59.222     44.349   1.335   0.1883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 151 on 46 degrees of freedom
## Multiple R-squared:  0.8135, Adjusted R-squared:  0.7892
## F-statistic: 33.45 on 6 and 46 DF,  p-value: 3.383e-15
```

This model has a higher adjusted R-squared value than the previous model and does show good fit.

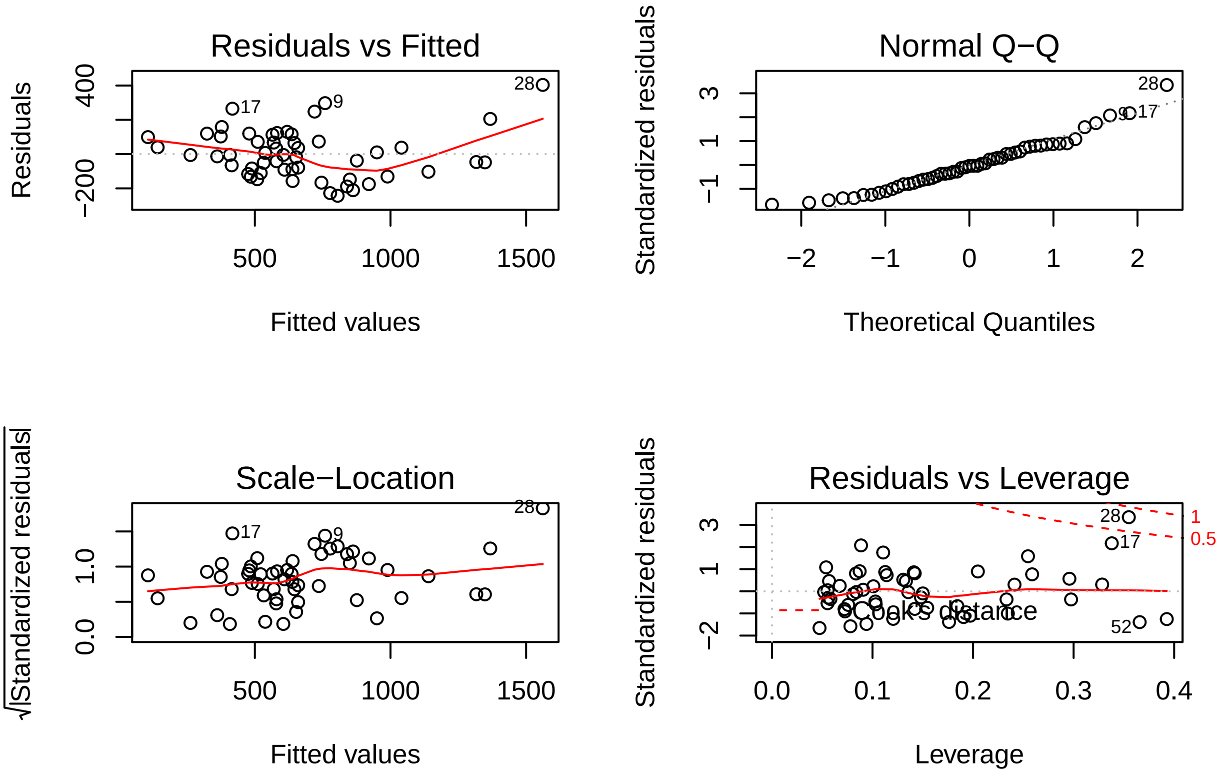
Residule Analysis

Now lets take a look at the residual plot of the last model that was selected by the step wise regression:



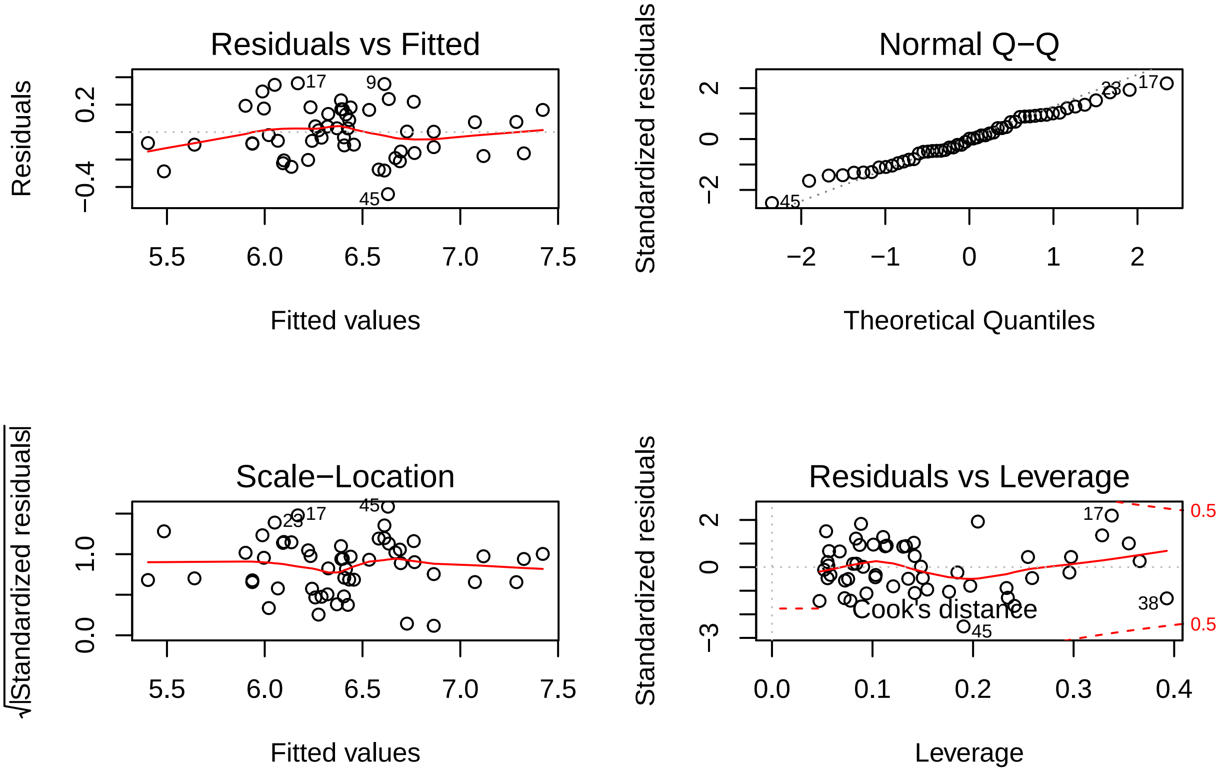
Although the probability plot looks good for this model there is a clear bias in the residue plot. Using a higher order term created by multiplying the two most significant variables, we now create a new model to account for this bias in the data.

```
##
## Call:
## lm(formula = Y ~ X4 + X3 + X2 + X9 + X1 + X7 + X3 * X4, data = BLOOD)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -243.19 -102.92   -5.51   99.13  403.57
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -633.250    325.612  -1.945 0.058064 .
## X4           -80.313    112.352  -0.715 0.478407
## X3             4.049      3.551   1.140 0.260185
## X2             8.093      1.460   5.542 1.48e-06 ***
## X9           234.126     57.455   4.075 0.000185 ***
## X1            48.839     18.016   2.711 0.009467 **
## X7            58.667     44.104   1.330 0.190151
## X4:X3          1.638      1.330   1.232 0.224377
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 150.2 on 45 degrees of freedom
## Multiple R-squared:  0.8196, Adjusted R-squared:  0.7916
## F-statistic: 29.21 on 7 and 45 DF,  p-value: 9.948e-15
```



No only has the adjusted R-squared went up again but the residual plot looks much better, however it still is not perfect and does appear to exhibit some multiplicative error. To remedy this we will use a logarithmic transformation of the dependent variable. This is the proper transformation for dealing with multiplicative error.

```
##
## Call:
## lm(formula = log(Y) ~ X4 + X3 + X2 + X9 + X1 + X7 + X3 * X4,
##     data = BLOOD)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45218 -0.14172  0.00262  0.16174  0.35581
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.351252   0.433623   7.728 8.59e-10 ***
## X4          0.283779   0.149622   1.897   0.0643 .
## X3          0.021669   0.004729   4.582 3.64e-05 ***
## X2          0.014312   0.001945   7.360 2.99e-09 ***
## X9           0.358231   0.076513   4.682 2.63e-05 ***
## X1           0.052928   0.023993   2.206   0.0325 *
## X7           0.097178   0.058733   1.655   0.1050
## X4:X3       -0.003184   0.001771  -1.798   0.0789 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2 on 45 degrees of freedom
## Multiple R-squared:  0.8365, Adjusted R-squared:  0.811
## F-statistic: 32.88 on 7 and 45 DF,  p-value: 1.149e-15
```



The model's residual graph now shows no signs of bias and looks relatively homoscedastic. The adjusted R-squared value has also went up once more and explains eighty one percent of the variance of the dependent variable.

Conclusion

After testing multiple models, a fairly complex model utilizing a logarithmic transformation of the dependent variable and a second order interaction term were determined to create the best fitting model for predicting the survival length of post operation patients. Using this model we can account for eighty percent of the variance in the survival rates of the post opperative patients.