Discussion 13 - Multiple Regression

Problem Statement Using R, build a multiple regression model for data that interests you. Include in this model at least one quadratic term, one dichotomous term, and one dichotomous vs. quantitative interaction term. Interpret all coefficients. Conduct residual analysis. Was the linear model appropriate? Why or why not?

Dataset overview This intermediate level data set has 155 rows and 20 columns and provides various attributes of a patient. Data can be download from below link:

https://www.kaggle.com/harinir/hepatitis

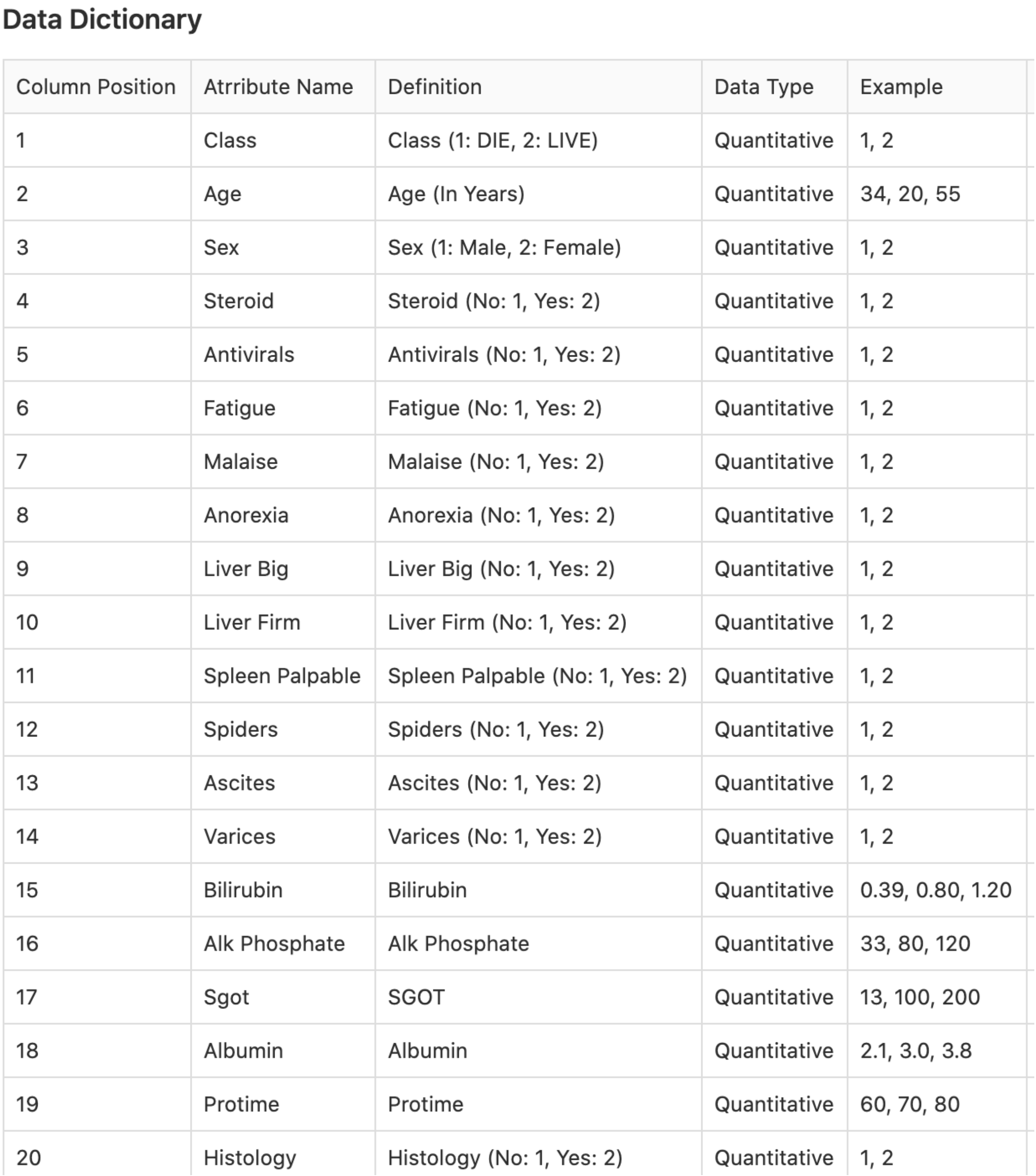
Below image describes variable name, description, levels, data type and values

Load data from the csv file to R then perform EDA.

```
hepatitis <-
  read.csv("https://raw.githubusercontent.com/SubhalaxmiRout002/DATA-605/master/Week%2013/hepatitis.csv
head(hepatitis)
```

First 6 rows

```
##   class age sex steroid antivirals fatigue malaise anorexia liver_big
## 1     2  30   2       1          2       2       2        2         1
## 2     2  50   1       1          2       1       2        2         1
## 3     2  78   1       2          2       1       2        2         2
## 4     2  34   1       2          2       2       2        2         2
## 5     2  34   1       2          2       2       2        2         2
## 6     1  51   1       1          2       1       2        1         2
##   liver_firm spleen_palable spiders ascites varices bilirubin alk_phosphate
## 1          2              2       2       2       2      1.00            85
## 2          2              2       2       2       2      0.90           135
## 3          2              2       2       2       2      0.70            96
## 4          2              2       2       2       2      1.00           105
## 5          2              2       2       2       2      0.90            95
## 6          2              1       1       2       2      1.42           105
##   sgot albumin protime histology
## 1   18    4.00      61         1
## 2   42    3.50      61         1
## 3   32    4.00      61         1
## 4  200    4.00      61         1
## 5   28    4.00      75         1
## 6   85    3.81      61         1
```



Data Dictionary

Column Position	Atrribute Name	Definition	Data Type	Example
1	Class	Class (1: DIE, 2: LIVE)	Quantitative	1, 2
2	Age	Age (In Years)	Quantitative	34, 20, 55
3	Sex	Sex (1: Male, 2: Female)	Quantitative	1, 2
4	Steroid	Steroid (No: 1, Yes: 2)	Quantitative	1, 2
5	Antivirals	Antivirals (No: 1, Yes: 2)	Quantitative	1, 2
6	Fatigue	Fatigue (No: 1, Yes: 2)	Quantitative	1, 2
7	Malaise	Malaise (No: 1, Yes: 2)	Quantitative	1, 2
8	Anorexia	Anorexia (No: 1, Yes: 2)	Quantitative	1, 2
9	Liver Big	Liver Big (No: 1, Yes: 2)	Quantitative	1, 2
10	Liver Firm	Liver Firm (No: 1, Yes: 2)	Quantitative	1, 2
11	Spleen Palpable	Spleen Palpable (No: 1, Yes: 2)	Quantitative	1, 2
12	Spiders	Spiders (No: 1, Yes: 2)	Quantitative	1, 2
13	Ascites	Ascites (No: 1, Yes: 2)	Quantitative	1, 2
14	Varices	Varices (No: 1, Yes: 2)	Quantitative	1, 2
15	Bilirubin	Bilirubin	Quantitative	0.39, 0.80, 1.20
16	Alk Phosphate	Alk Phosphate	Quantitative	33, 80, 120
17	Sgot	SGOT	Quantitative	13, 100, 200
18	Albumin	Albumin	Quantitative	2.1, 3.0, 3.8
19	Protime	Protime	Quantitative	60, 70, 80
20	Histology	Histology (No: 1, Yes: 2)	Quantitative	1, 2

Figure 1: image of data description

```
dim(hepatitis)
```

Dimension of dataset

```
## [1] 142  20
```

```
hepatitis[!complete.cases(hepatitis),]
```

Check for null values

```
##  [1] class          age            sex            steroid        antivirals
##  [6] fatigue        malaise        anorexia       liver_big      liver_firm
## [11] spleen_palable spiders        ascites        varices        bilirubin
## [16] alk_phosphate  sgot           albumin        protime        histology
## <0 rows> (or 0-length row.names)
```

What is quadratic? In mathematics, the term quadratic describes something that pertains to squares, to the operation of squaring, to terms of the second degree, or equations or formulas that involve such terms.

A polynomial term–a quadratic (squared) or cubic (cubed) term turns a linear regression model into a curve. Equation :

$$ax^2 + bx + c = 0$$

In out dataset we will create one new column called as `quardetic`

```
hepatitis$quardetic <- hepatitis$alk_phosphate ^ 2
```

What is dichotomous? Dichotomous variables are nominal variables which have only two categories or levels.

In this dataset we have many dichotomous variables. We will use `spiders` for this and named this column as dichotomous. Here we multiply quardetic variable with dichotomous variable.

```
hepatitis$dichotomous <- hepatitis$alk_phosphate * hepatitis$spiders
```

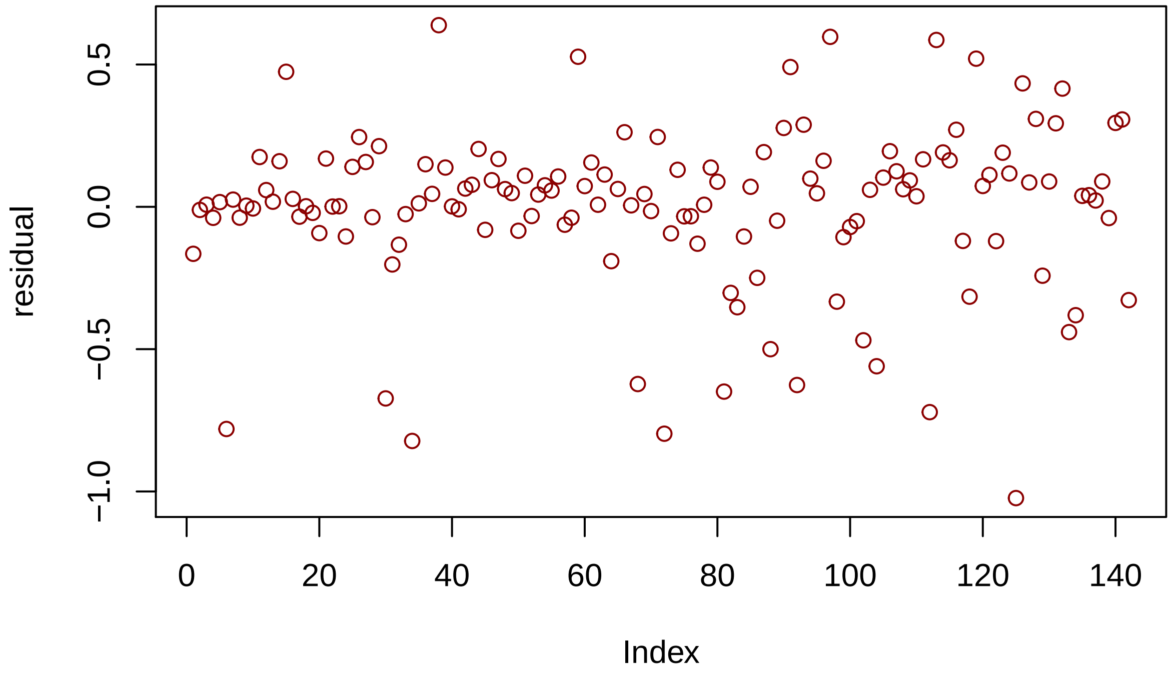
```
lm <- lm(class ~ age + sex + steroid + antivirals + fatigue + malaise + anorexia + liver_big + liver_fi
           spiders + ascites + varices + bilirubin + alk_phosphate + sgot + albumin + protime + histolog
summary(lm)
```

Apply multiple regression model

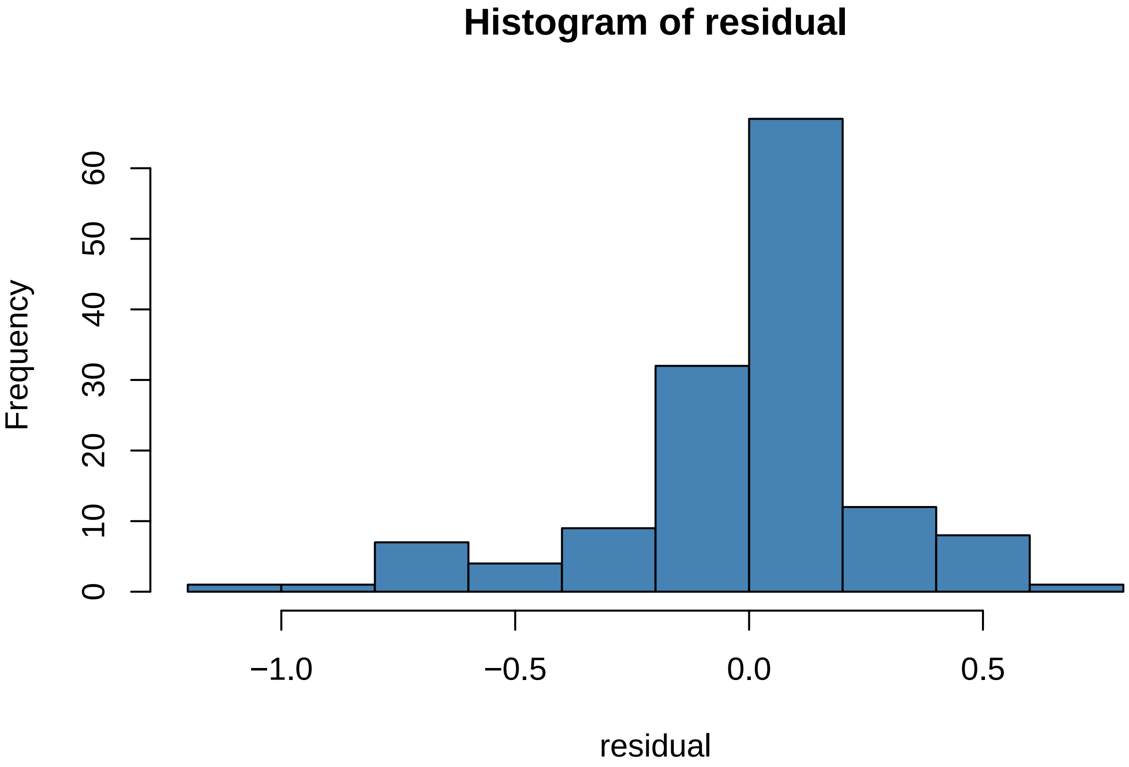
```
##
## Call:
## lm(formula = class ~ age + sex + steroid + antivirals + fatigue +
##     malaise + anorexia + liver_big + liver_firm + spleen_palable +
##     spiders + ascites + varices + bilirubin + alk_phosphate +
##     sgot + albumin + protime + histology, data = hepatitis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02318 -0.06890  0.03971  0.13978  0.63810
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     5.602e-01  4.570e-01   1.226 0.222673
## age             2.079e-05  2.387e-03   0.009 0.993063
## sex             1.605e-01  8.564e-02   1.875 0.063221 .
## steroid         4.204e-02  5.699e-02   0.738 0.462154
## antivirals      4.212e-02  7.608e-02   0.554 0.580903
## fatigue        -4.405e-03  7.281e-02  -0.060 0.951858
## malaise         1.234e-01  8.132e-02   1.518 0.131720
## anorexia       -1.385e-01  8.618e-02  -1.607 0.110678
## liver_big      -9.594e-02  8.126e-02  -1.181 0.240025
## liver_firm     -1.605e-02  6.630e-02  -0.242 0.809099
## spleen_palable  7.325e-02  7.195e-02   1.018 0.310611
## spiders         1.828e-01  6.972e-02   2.622 0.009846 **
## ascites         2.621e-01  1.082e-01   2.422 0.016910 *
## varices         4.558e-02  9.966e-02   0.457 0.648263
## bilirubin      -9.389e-02  2.783e-02  -3.373 0.000996 ***
## alk_phosphate   1.568e-04  6.680e-04   0.235 0.814843
## sgot            4.658e-04  3.507e-04   1.328 0.186611
## albumin         5.400e-02  5.824e-02   0.927 0.355644
## protime         1.218e-03  1.607e-03   0.758 0.449642
## histology      -2.173e-02  6.088e-02  -0.357 0.721768
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.3035 on 122 degrees of freedom
## Multiple R-squared:  0.4709, Adjusted R-squared:  0.3885
## F-statistic: 5.714 on 19 and 122 DF,  p-value: 6.319e-10
```

```r
residual <- resid(lm)
plot(residual, col = 'dark red')
```

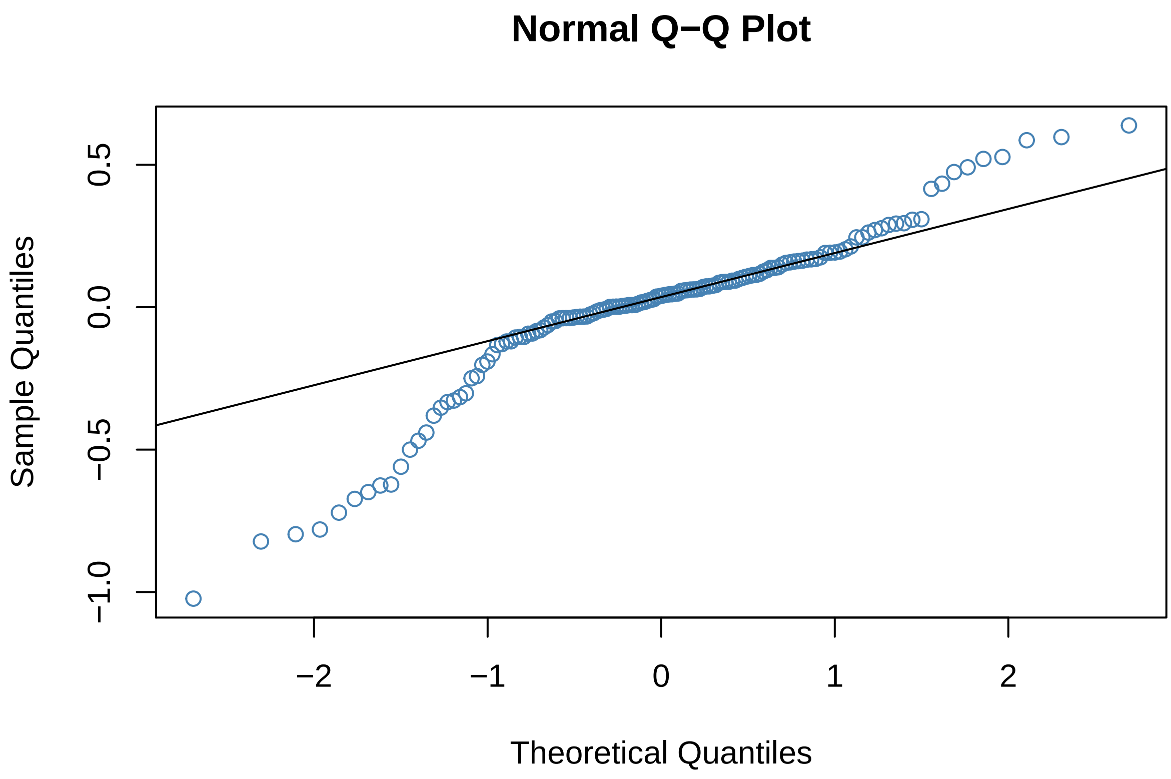
Residual Analysis of model 1



```r
hist(residual, col = "steelblue")
```



```r
qqnorm(residual, col = 'steelblue')
qqline(residual)
```



From above plots and linear model Coefficients we found:

- P-value is small < 0.05, residual is normally distributed.
- R^2 is 0.4709 means model explains 47% variation in the response variable
- QQ-plot shows variations in tail

This model is not a good fit model, it needs more improvement

```
lm2 <- lm(class ~ age + sex + steroid + antivirals + fatigue + malaise + anorexia + liver_big + liver_fi
           spiders + ascites + varices + bilirubin + alk_phosphate + sgot + albumin + protime + histolog
          data = hepatitis)
summary(lm2)
```

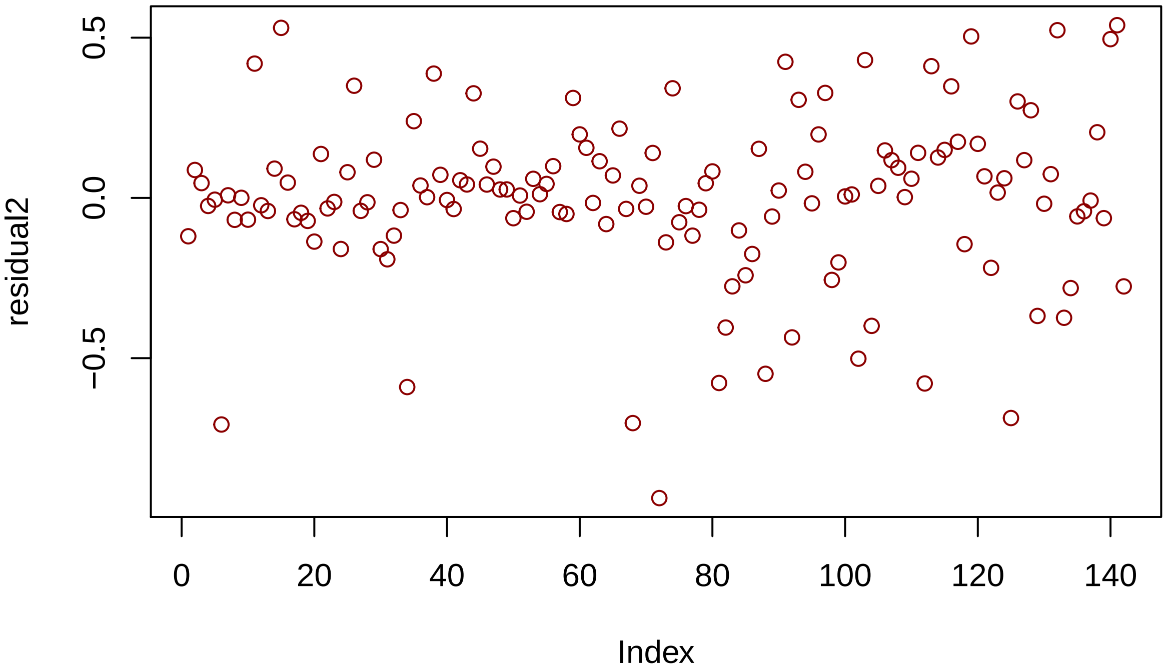
Apply multiple regression model with new variables

```
##
## Call:
## lm(formula = class ~ age + sex + steroid + antivirals + fatigue +
##     malaise + anorexia + liver_big + liver_firm + spleen_palable +
##     spiders + ascites + varices + bilirubin + alk_phosphate +
##     sgot + albumin + protime + histology + quardetic + dichotomous,
##     data = hepatitis)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93655 -0.06829  0.00775  0.11882  0.53896
```

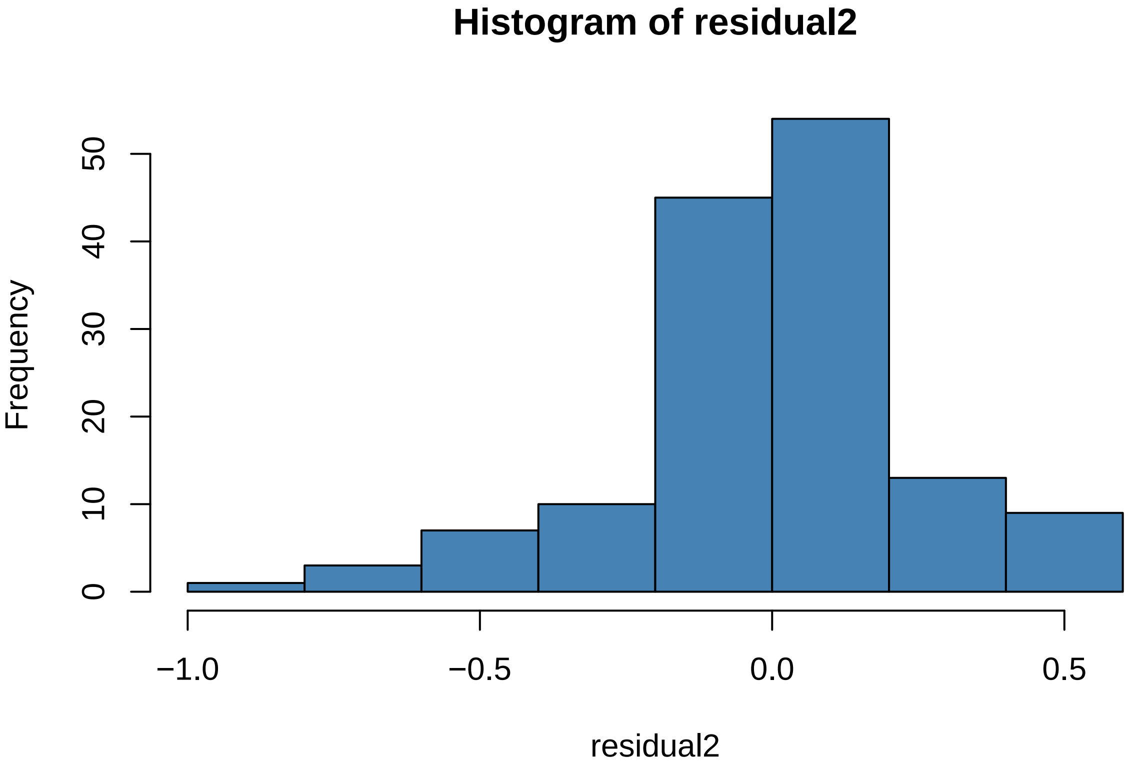
```
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.141e+00  5.342e-01  -2.135  0.03476 *
## age            -1.246e-03  2.222e-03  -0.561  0.57590
## sex             1.238e-01  7.858e-02   1.575  0.11793
## steroid         1.964e-02  5.244e-02   0.375  0.70863
## antivirals     -7.316e-03  7.131e-02  -0.103  0.91845
## fatigue        -3.959e-03  6.677e-02  -0.059  0.95281
## malaise         1.099e-01  7.434e-02   1.478  0.14194
## anorexia       -7.679e-02  7.965e-02  -0.964  0.33694
## liver_big      -3.141e-02  7.531e-02  -0.417  0.67740
## liver_firm     -3.904e-02  6.077e-02  -0.642  0.52183
## spleen_palable  6.650e-02  6.612e-02   1.006  0.31652
## spiders         8.276e-01  1.424e-01   5.811 5.22e-08 ***
## ascites         3.009e-01  9.940e-02   3.028  0.00302 **
## varices         1.526e-01  9.415e-02   1.621  0.10770
## bilirubin      -8.013e-02  2.559e-02  -3.132  0.00218 **
## alk_phosphate   1.435e-02  3.233e-03   4.439 2.02e-05 ***
## sgot            2.949e-04  3.243e-04   0.910  0.36489
## albumin         6.834e-02  5.329e-02   1.283  0.20211
## protime         1.264e-03  1.468e-03   0.861  0.39083
## histology       1.780e-02  5.626e-02   0.316  0.75224
## quardetic      -1.576e-05  7.261e-06  -2.170  0.03197 *
## dichotomous    -5.873e-03  1.169e-03  -5.025 1.77e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2773 on 120 degrees of freedom
## Multiple R-squared:  0.5657, Adjusted R-squared:  0.4897
## F-statistic: 7.443 on 21 and 120 DF,  p-value: 1.42e-13
```

```
residual2 <- resid(lm2)
plot(residual2, col = 'dark red')
```

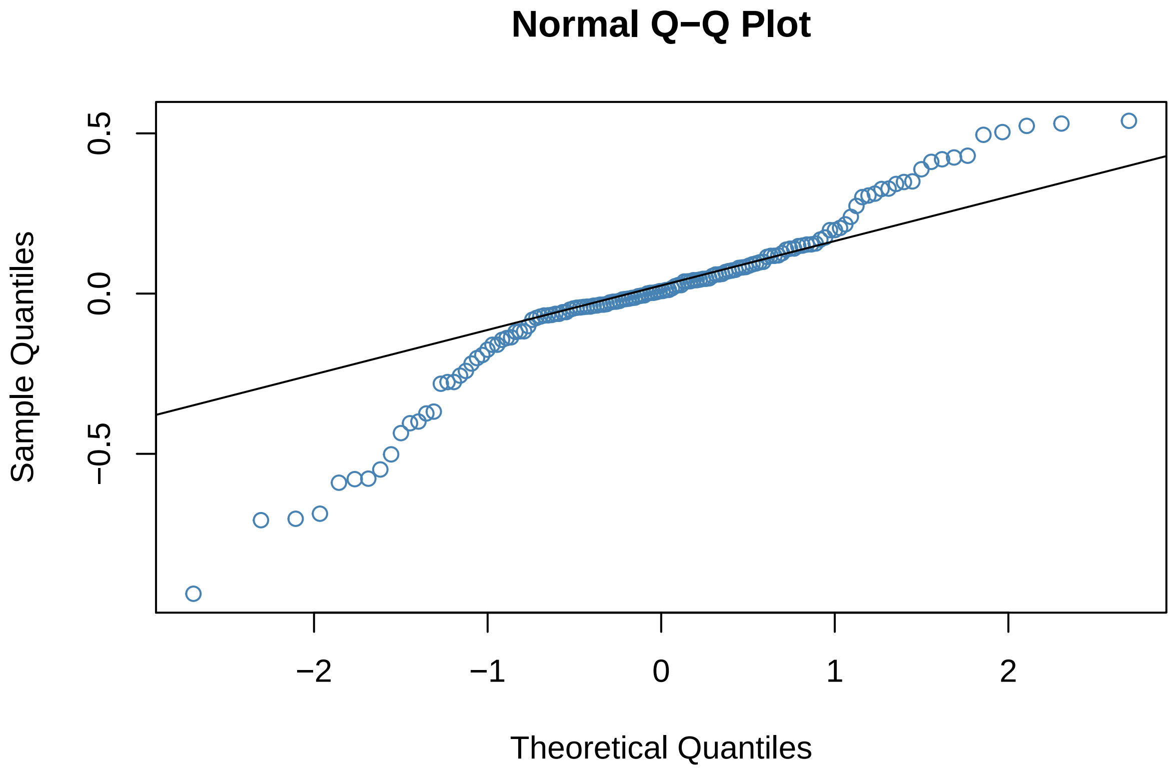
Residual Analysis of model 2



```r
hist(residual2, col = "steelblue")
```



```r
qqnorm(residual2, col = "steelblue")
qqline(residual2)
```



After apply quadratic and dichotomous the model has little improved.

- Residual standard error getting improved, get high R^2 means model expalins 57% of variation in the response variable
- Residual ditribution is unimodel and symmetric.
- QQ-plot shows variations

Model 2 is beter than model 1 but model2 is also not a good fit, improvement required for this model to be a good fit. We can apply backword elimination, transformation or different machine learning algorithms such as KNN, random forest to make model good fit.