Logarithmic Modelling

#Working on ln(Y) = b0 + b1X1 for simple linear model, and multiple linear model

```
#Transforming Response variable using ln for logarithmic Model of the Simple Linear Model
logarithmitic_dataset$L_time_in_hospital <- log(logarithmitic_dataset$time_in_hospital)
```

```
> logarithmitic_dataset
    time_in_hospital num_medications num_lab_procedures L_time_in_hospital
1                  7               11                 62           1.9459101
2                  7               15                 60           1.9459101
3                  4               17                 45           1.3862944
4                  9               16                 25           2.1972246
5                  4               14                 40           1.3862944
6                  2               11                 32           0.6931472
7                  7               22                 75           1.9459101
8                  3               21                 57           1.0986123
9                 14               19                 78           2.6390573
10                 4               19                 65           1.3862944
```

#Summary of original simple linear model without transformation.

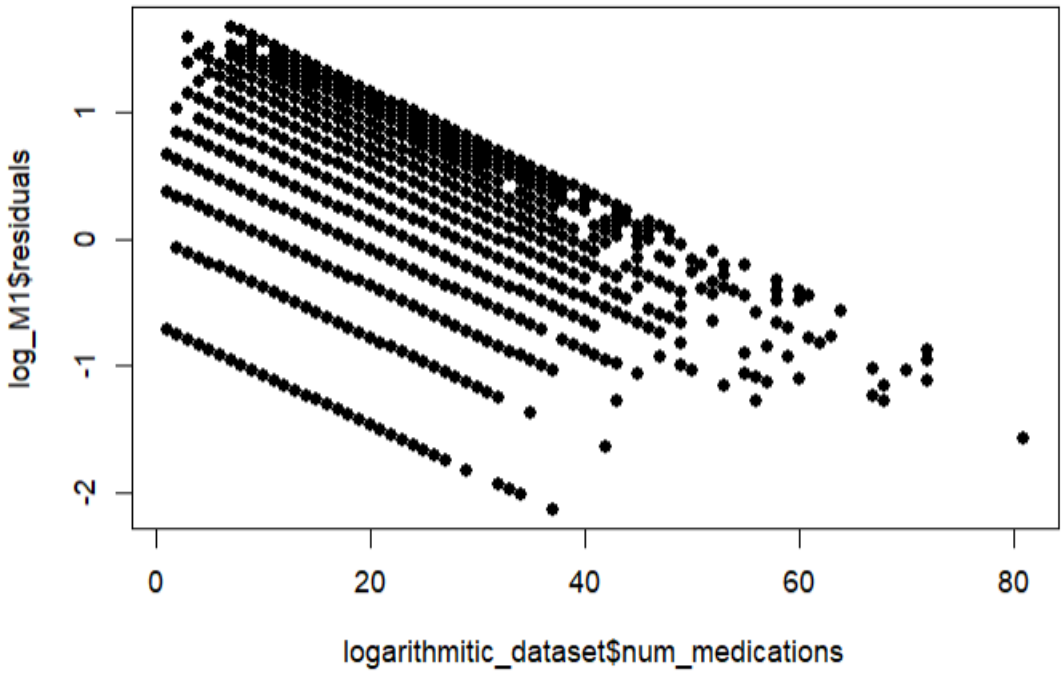
```
> summary(diabetic_less_M.1)

Call:
lm(formula = time_in_hospital ~ num_medications, data = diabetic_data.less)

Residuals:
    Min      1Q  Median      3Q     Max
-7.3714 -1.8892 -0.5306  1.4215 11.0073

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.73771    0.05773   30.10   <2e-16 ***
num_medications  0.17929    0.00308   58.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.658 on 11355 degrees of freedom
Multiple R-squared:  0.2298,    Adjusted R-squared:  0.2297
F-statistic:  3388 on 1 and 11355 DF,  p-value: < 2.2e-16
```

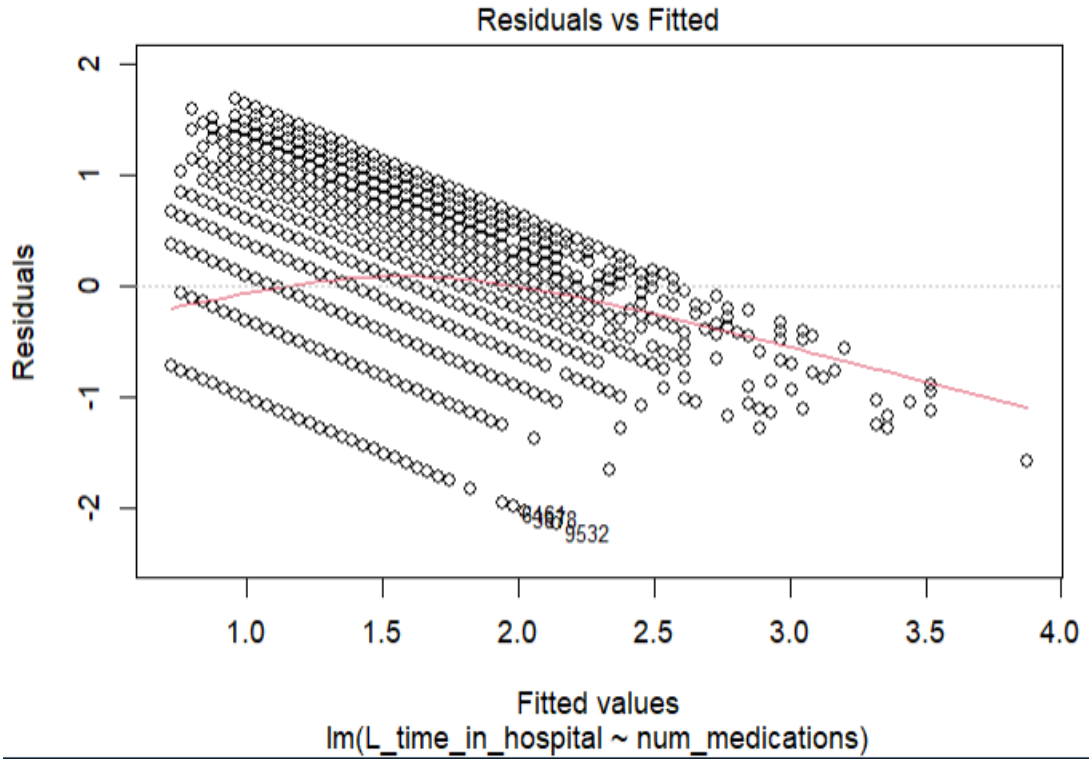
#Summary of the logarithmic transformation model with log (response variable)

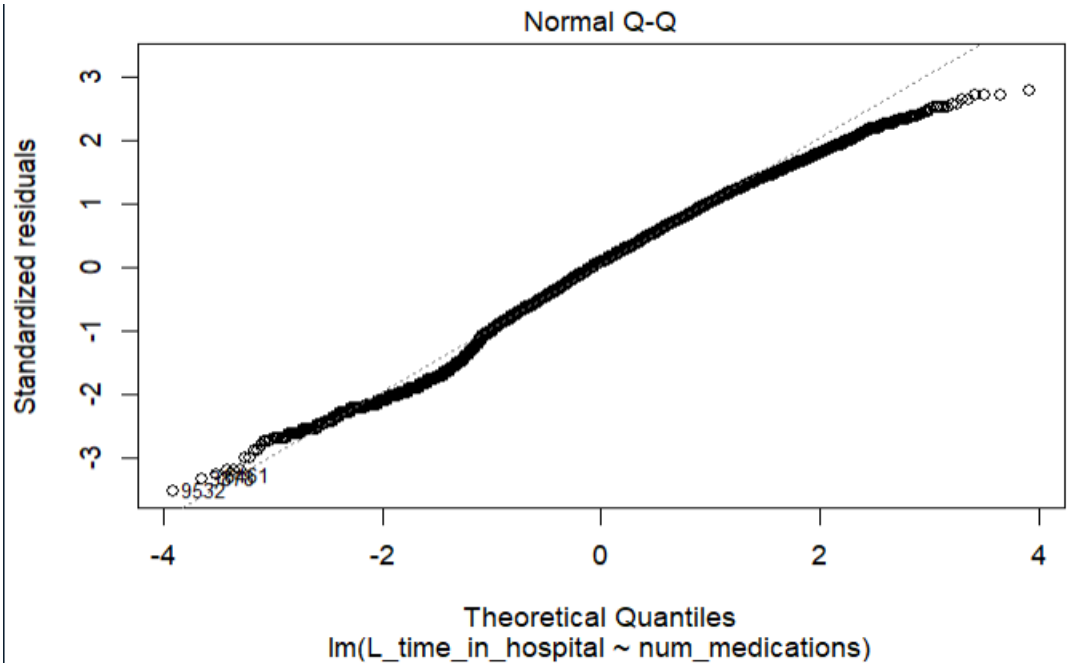
#Below are the outputs of the simple logarithmic transformation model.

```
> summary(log_M1)

Call:
lm(formula = L_time_in_hospital ~ num_medications, data = logarithmitic_dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-2.13997 -0.38221  0.05725  0.43682  1.68199

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.6810574  0.0132011   51.59   <2e-16 ***
num_medications 0.0394301  0.0007044   55.98   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6077 on 11355 degrees of freedom
Multiple R-squared:  0.2163,    Adjusted R-squared:  0.2162
F-statistic:  3134 on 1 and 11355 DF,  p-value: < 2.2e-16
```

#Residual Plot for Logarithmic model





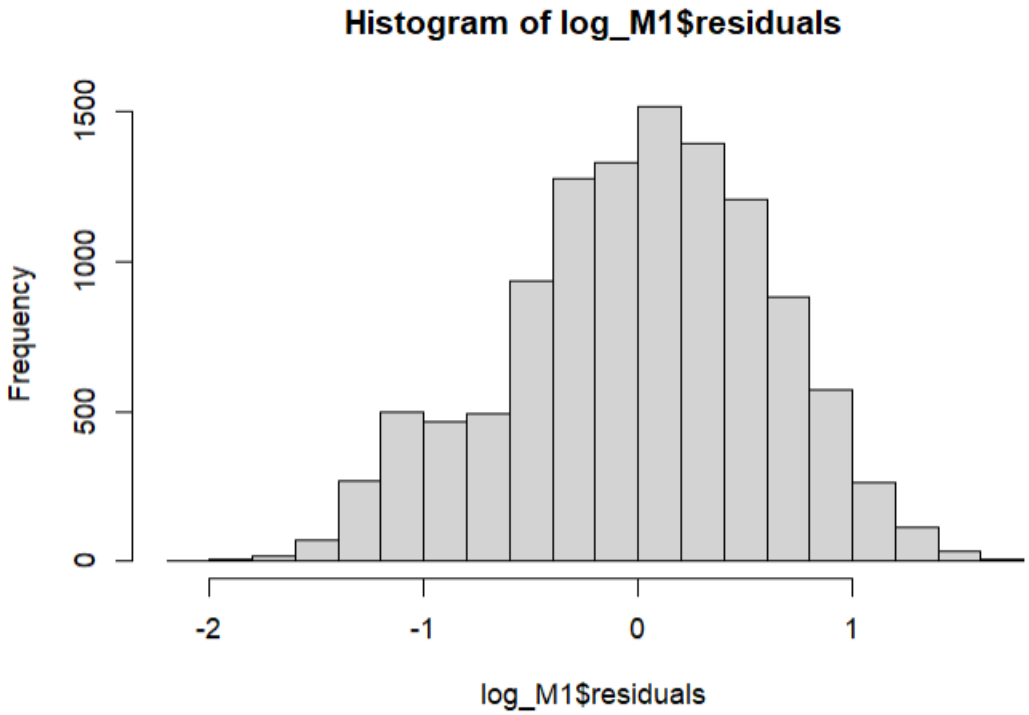



```
> mean(log_M1$residuals)
[1] -2.852719e-16
```

```
> #compute correlation using Durbin Watson test
> dwtest(log_M1)

        Durbin-Watson test

data:  log_M1
DW = 1.9589, p-value = 0.01425
alternative hypothesis: true autocorrelation is greater than 0
```

```
> #test for homoscedasticity using Breusch Pagan Test
> bptest(log_M1)

        studentized Breusch-Pagan test

data:  log_M1
BP = 87.653, df = 1, p-value < 2.2e-16
```

```
> shapiro.test(log_M1$residuals)
Error in shapiro.test(log_M1$residuals) :
  sample size must be between 3 and 5000
```



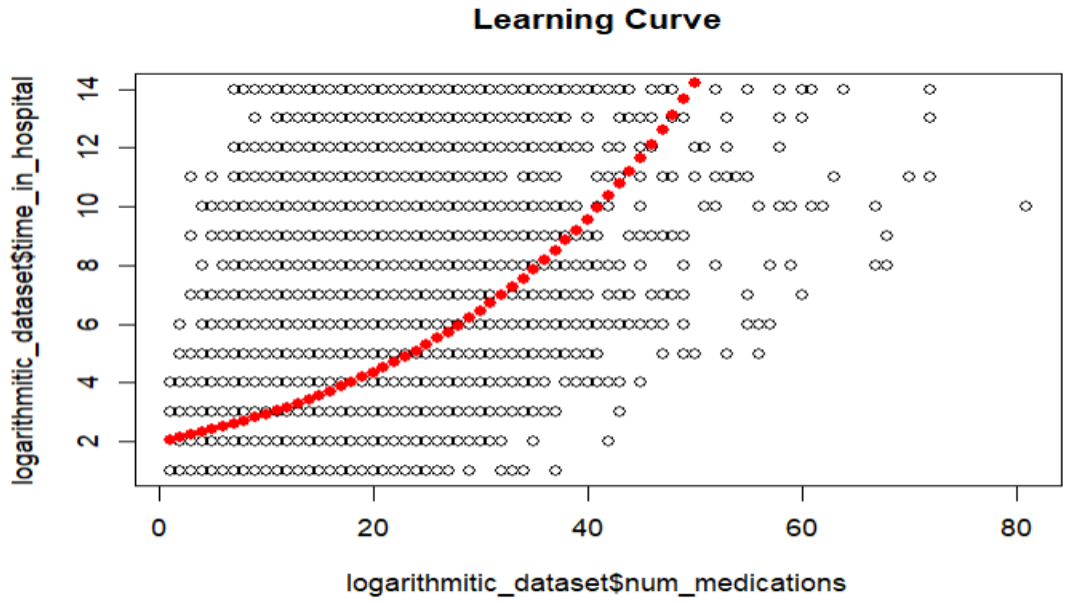
#Interpretation of Simple Logarithmic model output:

#Slope coefficients are significant.

ln(y) = 0.6810574 + 0.0394301(X1)

#Applying e to both sides,

Y = 1.975966 + 1.040218^num_medications



```
> #Residual Standard Error
> head(cbind(logarithmitic_dataset$time_in_hospital, exp(log_M1$fitted.values), logarithmitic_dataset$time_in_hospital-
exp(log_M1$fitted.values)))
  [,1]     [,2]       [,3]
1    7 3.048924  3.9510759
2    7 3.569799  3.4302011
3    4 3.862712  0.1372880
4    9 3.713368  5.2866315
5    4 3.431780  0.5682198
6    2 3.048924 -1.0489241
```

```
> # Learning rate
> 2^(log_M1$coefficients[2])
num_medications
       1.027708
> # Learning rate
> 2^(log_M1$coefficients[1])
(Intercept)
   1.603314
```

```
> sqrt(sum((logarithmitic_dataset$num_medications- exp(log_M1$fitted.values))^2)/log_M1$df.residual)
[1] 14.34492
```

```
> 1- M1_SSE/M1_TSE
[1] 0.08508087
```

```
> 1- log_M1_SSE/log_M1_TSE
[1] -21.16254
```

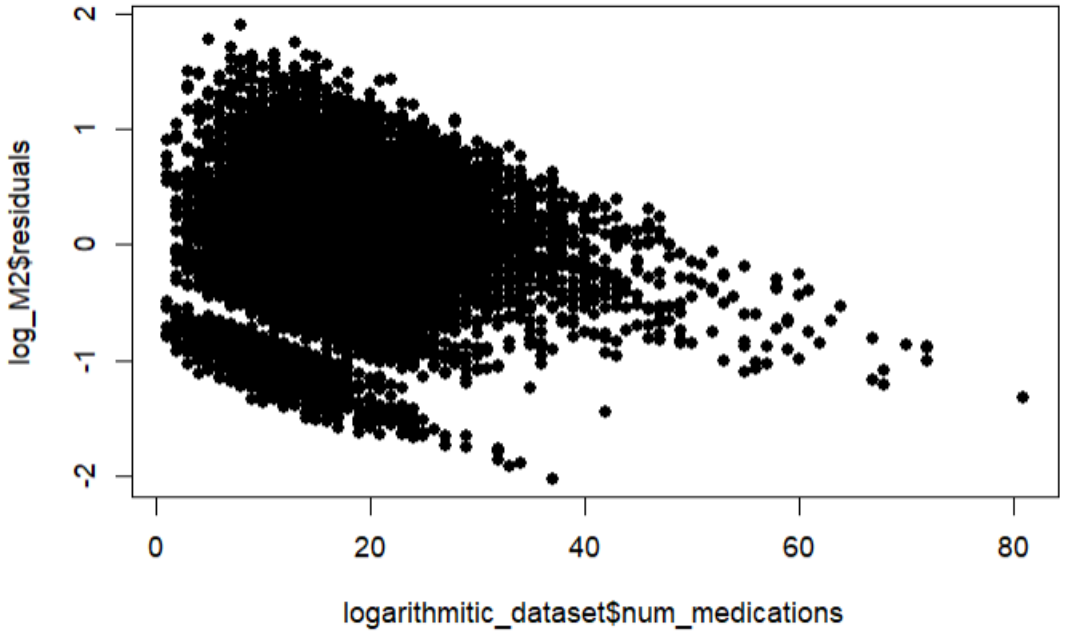
Summary of original Multiple Model

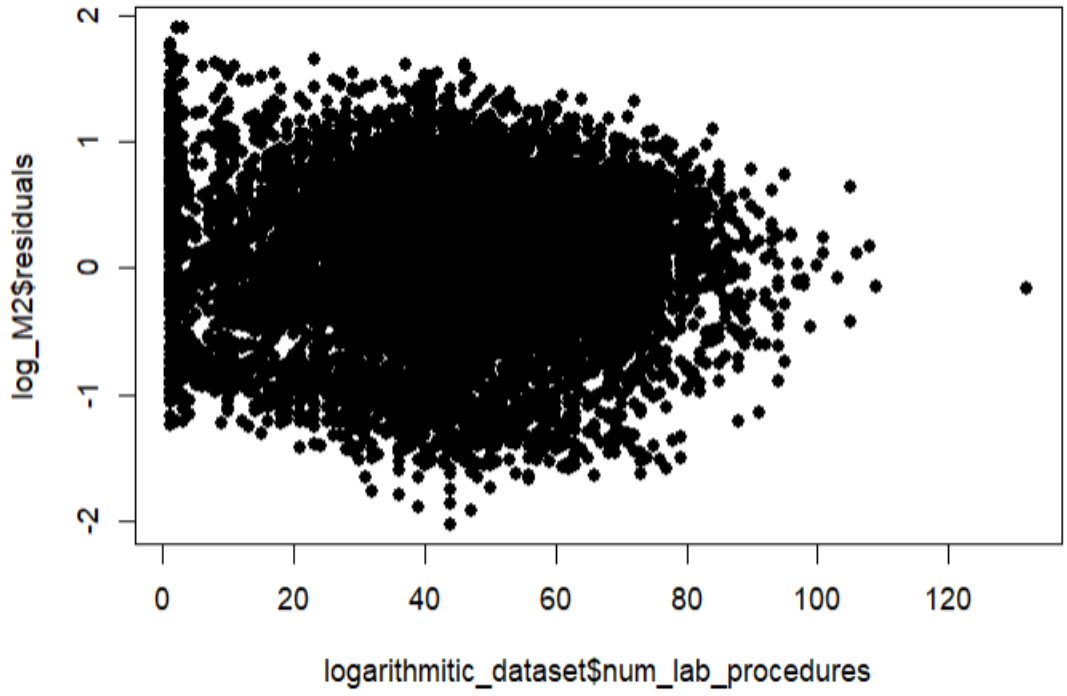
```
> summary(diabetic_less_M.2)

Call:
lm(formula = time_in_hospital ~ num_medications + num_lab_procedures,
    data = diabetic_multi)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0584 -1.7553 -0.5329  1.1955 12.5644

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.6189856  0.0226068   27.38   <2e-16 ***
num_medications    0.1506737  0.0010297  146.32   <2e-16 ***
num_lab_procedures 0.0316257  0.0004254   74.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.572 on 101763 degrees of freedom
Multiple R-squared:  0.2576,    Adjusted R-squared:  0.2576
F-statistic: 1.766e+04 on 2 and 101763 DF,  p-value: < 2.2e-16
```

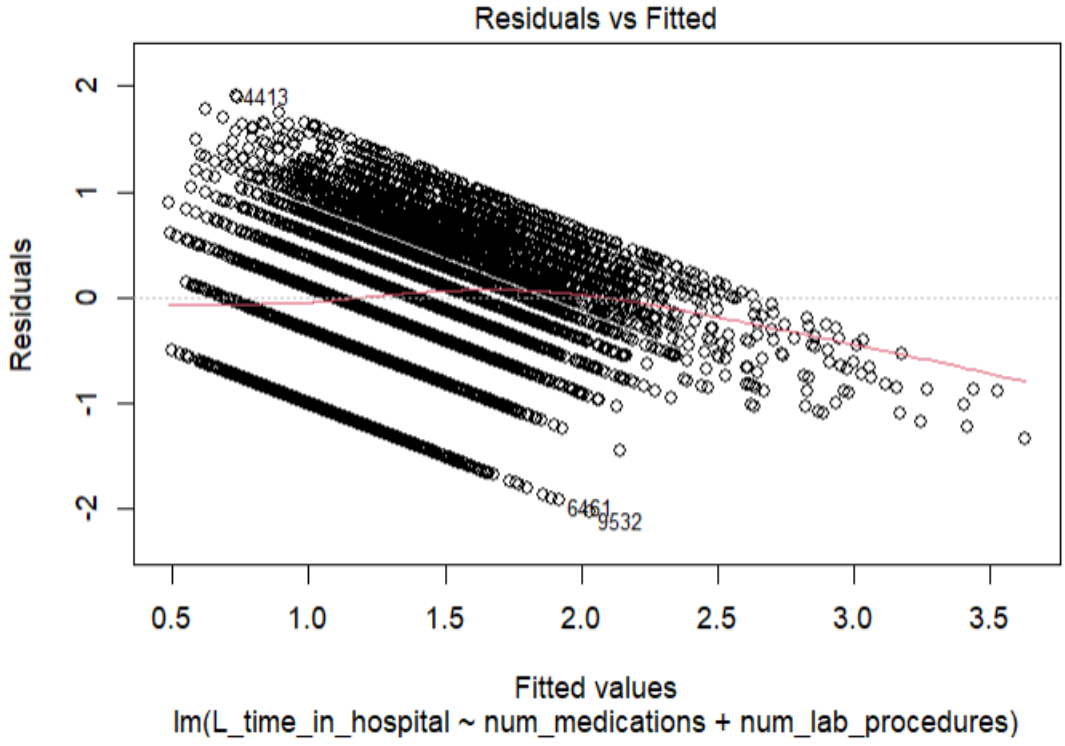
#Below are the results and outputs of the Multiple logarithmic transformation model.

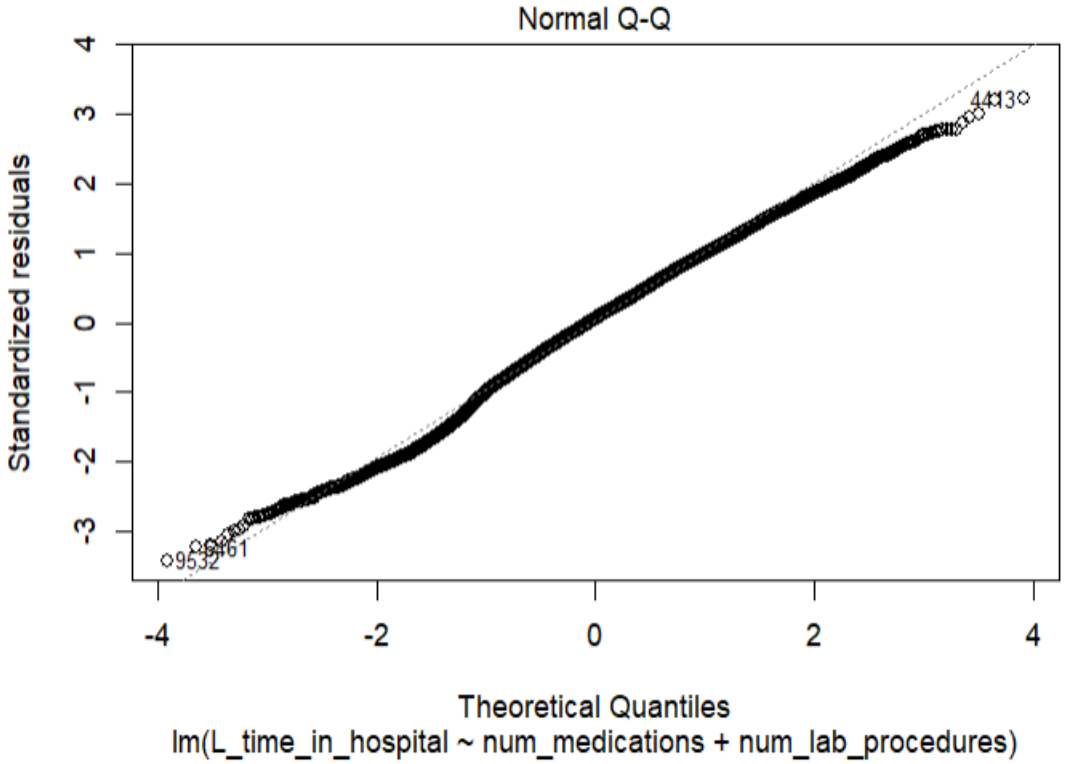
```
> summary(log_M2)

Call:
lm(formula = L_time_in_hospital ~ num_medications + num_lab_procedures,
    data = logarithmitic_dataset)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0315 -0.3738  0.0348  0.4168  1.9084

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.4428923  0.0161182   27.48   <2e-16 ***
num_medications    0.0341168  0.0007198   47.40   <2e-16 ***
num_lab_procedures 0.0074159  0.0003023   24.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5923 on 11354 degrees of freedom
Multiple R-squared:  0.2557,    Adjusted R-squared:  0.2556
F-statistic:  1951 on 2 and 11354 DF,  p-value: < 2.2e-16
```

Residual Plot for both the predictors of the logarithmic transformation Multiple Model





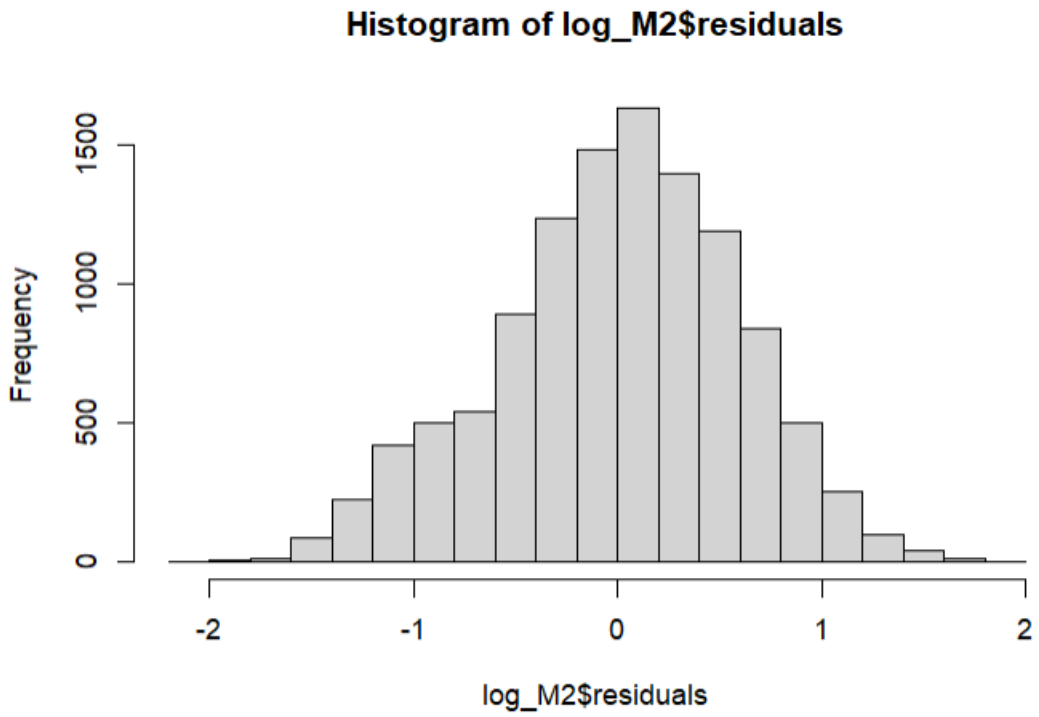

Residuals vs Fitted
Residuals
Fitted values
lm(L_time_in_hospital ~ num_medications + num_lab_procedures)
4413
6461
9532


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(L_time_in_hospital ~ num_medications + num_lab_procedures)
4413
6461
9532

```
> mean(log_M2$residuals)
[1] 4.349063e-17
```

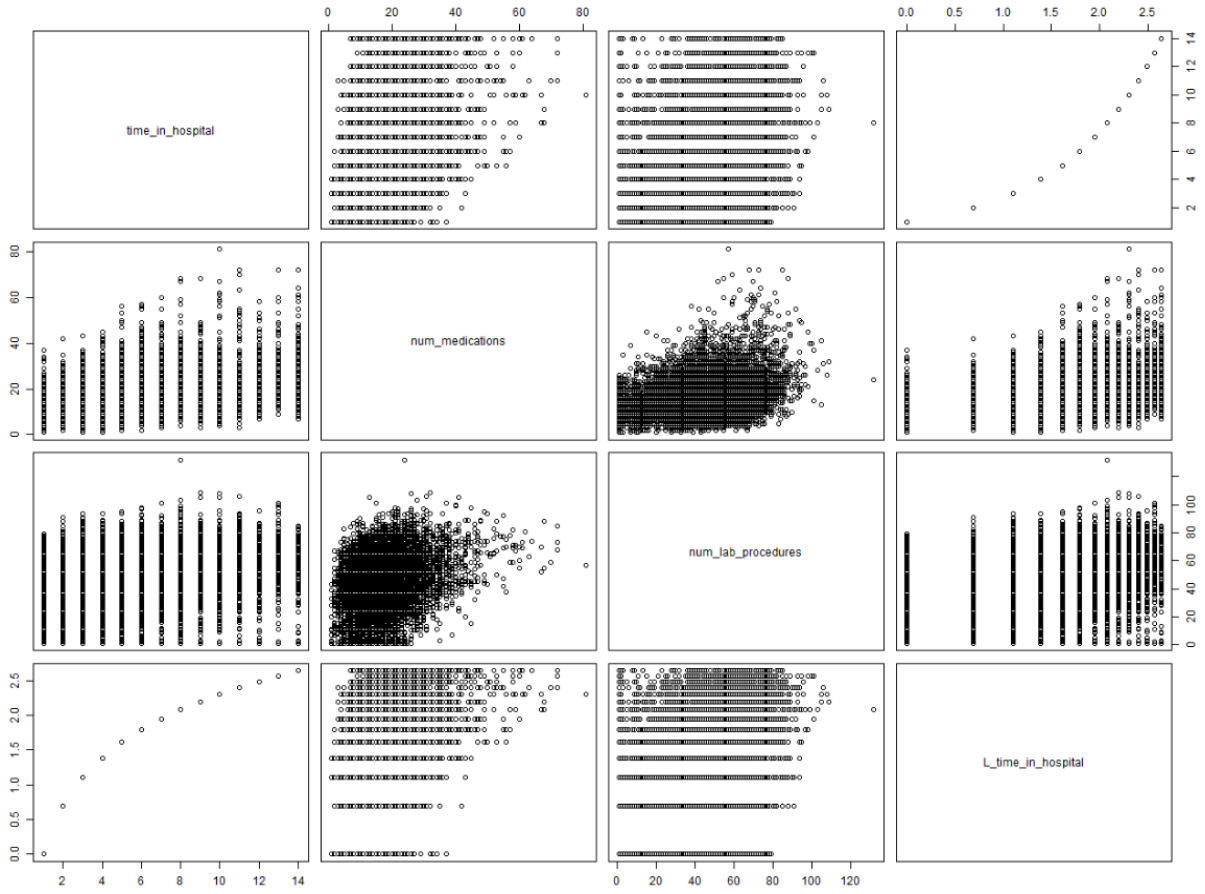
```
> dwtest(log_M2)

        Durbin-Watson test

data:  log_M2
DW = 1.9673, p-value = 0.04044
alternative hypothesis: true autocorrelation is greater than 0
```

```
> bptest(log_M2)

        studentized Breusch-Pagan test

data:  log_M2
BP = 244.84, df = 2, p-value < 2.2e-16
```



```
> shapiro.test(log_M2$residuals)
Error in shapiro.test(log_M2$residuals) :
  sample size must be between 3 and 5000
```

#Checking for Multicollinearity




```
> round(cor(logarithmitic_dataset),4)
                  time_in_hospital num_medications num_lab_procedures L_time_in_hospital
time_in_hospital            1.0000          0.4794             0.3177             0.9315
num_medications             0.4794          1.0000             0.3009             0.4651
num_lab_procedures          0.3177          0.3009             1.0000             0.3293
L_time_in_hospital          0.9315          0.4651             0.3293             1.0000
```

```
> library(car)
Loading required package: carData
> vif(log_M2)
   num_medications num_lab_procedures
          1.099588           1.099588
```

#Both the logarithmic transformation residual graphs show CONCAVE plots.

#Both the QQ plots are a lot better, as only a few points are outside the confidence bands when compared to the original

#Therefore, in both the tests mean of residual is close to zero, Histogram indicates normality,

#But still Durbin Watson, Breusch pagan test and Shapiro Watson tests failed.

#There also exists multi-collinearity between variables.