Lab03

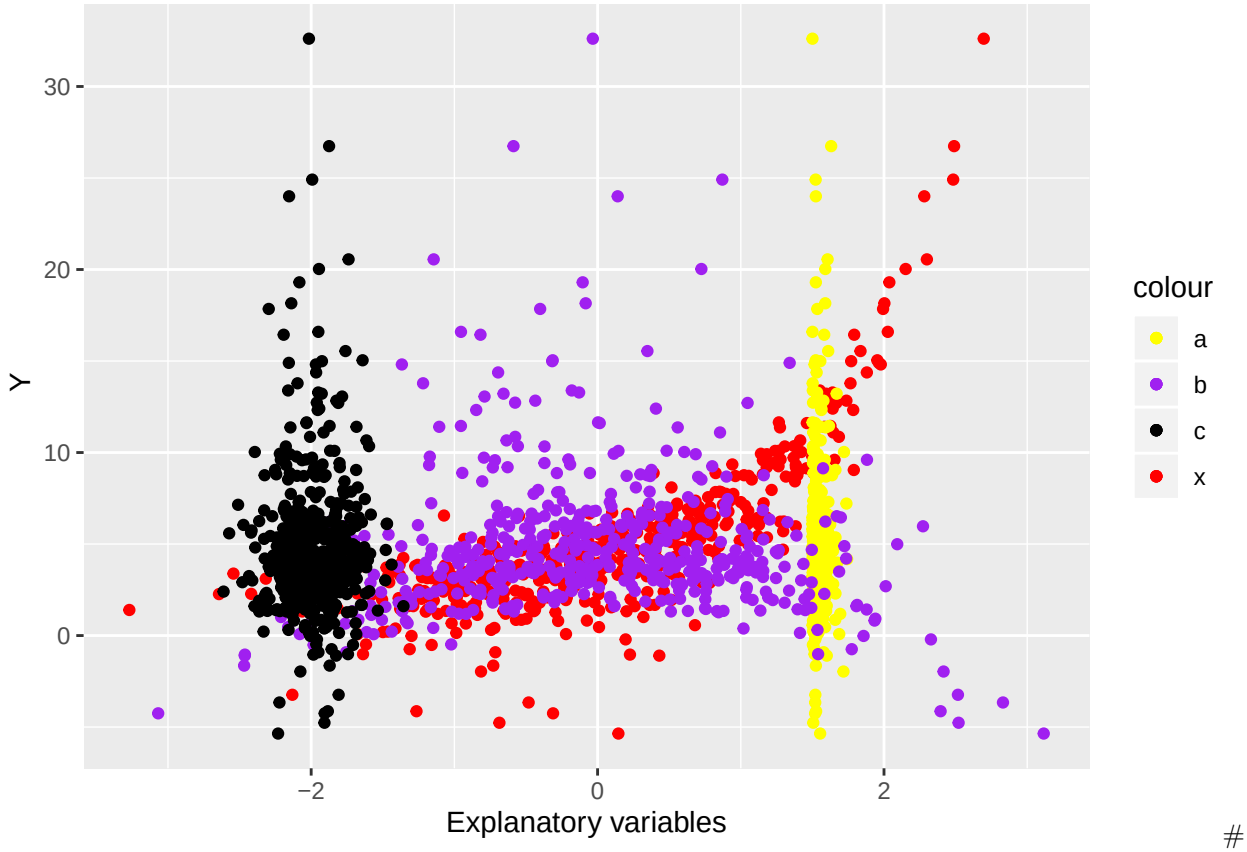
Minh Tam Hoang

10/7/2019

```
library(ggplot2)
lab_data <- read.csv("class11_LAB_dataFrame_20190930T2158.csv")
head(lab_data)
```

```
##   X        y        a          x           b         c
## 1 0 8.760203 1.578072  1.0220428  0.34599508 -1.898979
## 2 1 2.081936 1.508801 -0.2148006  1.03857945 -2.177976
## 3 2 6.046254 1.523975  1.0412572  0.16278736 -2.472784
## 4 3 5.636724 1.609808 -0.1726115 -0.02560988 -2.011834
## 5 4 1.643514 1.509789 -0.9514274 -1.58571591 -1.828793
## 6 5 6.034250 1.577474  0.8941980 -1.25455739 -2.125850
```

```
ggplot2::ggplot()+
  geom_point(mapping = aes(x = lab_data$x, y = lab_data$y,  col = "x"))+
  geom_point(mapping = aes(x = lab_data$a, y = lab_data$y,col = "a"))+
  geom_point(mapping = aes(x = lab_data$b, y = lab_data$y, col = "b"))+
  geom_point(mapping = aes(x = lab_data$c, y = lab_data$y,col = "c"))+
  scale_colour_manual(values = c("x" = "red", "a" = "yellow", "b" = "purple", "c" = "black"))+
                      labs(x = "Explanatory variables", y = " Y")
```



#

Exploratory Data Analysis

Summarize the relationships between y and the set of explanatory variables (x,a,b,c).

There is a strong, positive exponetial relationship between x and y, which means an increase in the value of x corresponds to an exponential increase in the value of y.

The plot of y vs b looks like a parabola with negative slope.

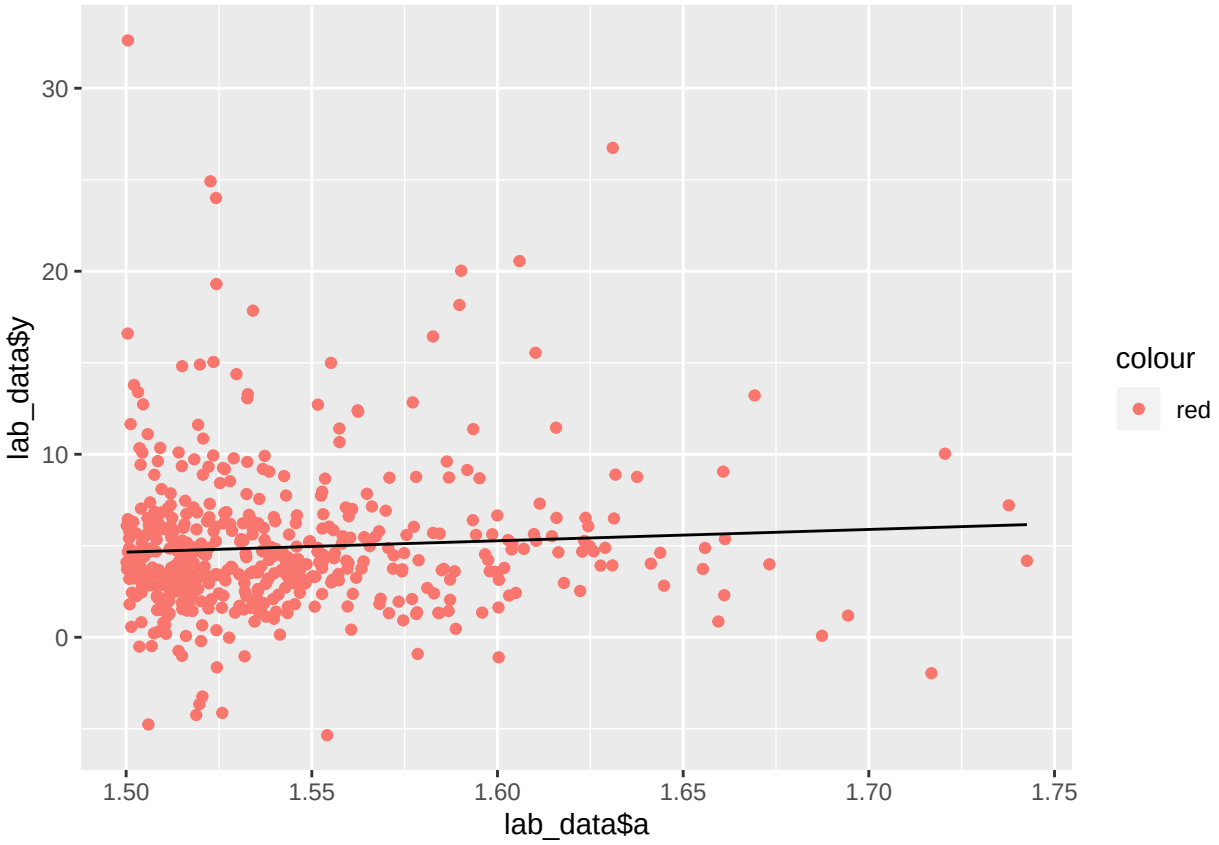
The relationship between y and a: There is an extremely dense cluster of points while x is between 1.5 and 1.55. From 1.55 onwards,points scatter all over the place. There does not seem to be any specific trend in this plot. and the one between y and c follow the same pattern which is the explanatory variable remaining almost unchanged(or slight difference among all the values of explanatory variable) corresponds to an increase in y.

Regression

```
# Fit a simple linear regression model between y and a

lin_reg <- lm(lab_data$y~lab_data$a)
print(lin_reg)
```

```
##
## Call:
## lm(formula = lab_data$y ~ lab_data$a)
##
## Coefficients:
## (Intercept)   lab_data$a
##      -4.641        6.195
```

```
ggplot()+
  geom_point(mapping = aes(x = lab_data$a, y = lab_data$y,  col = "red"))+
  geom_line(mapping = aes(x = lab_data$a , y =fitted(lin_reg)))
```

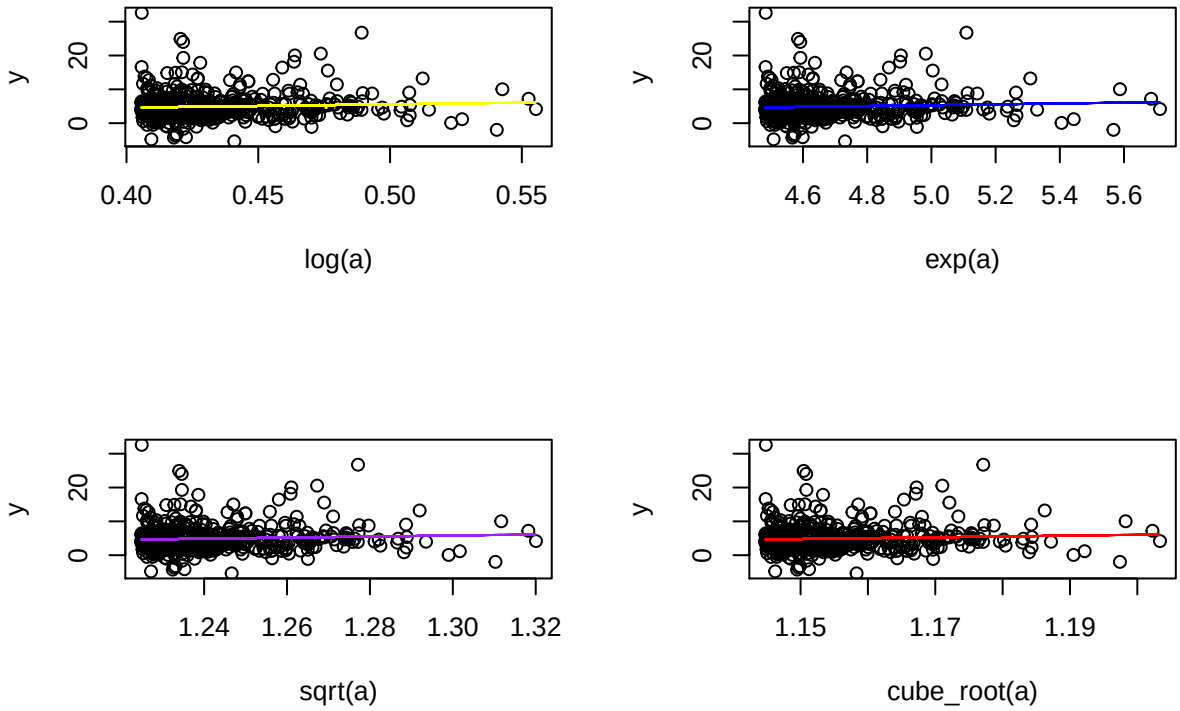


Comment on the predicted model versus ground truth:

The predicted model indicates that when x is below 1.65, y remains nearly constant and when x is above 1.65 y experience a slight increase. This trend does not reflect the behavior of y with respect to x accurately since the model over-estimates most of y-values when x values are greater than 1.65 and fails to capture the behavior of y for x less than 1.65.

#Transform a. Four tranformations applied to a are: log(a), exp(a), a^(1/2), and a^(1/3)

```
log_a <- log(lab_data$a)
exp_a <- exp(lab_data$a)
sqrt_a <- sqrt(lab_data$a)
cube_root_a <- (lab_data$a^(1/3))

log_a_reg <- lm(lab_data$y~log_a)

exp_a_reg <- lm(lab_data$y~exp_a)

sqrt_a_reg <- lm(lab_data$y~sqrt_a)

cube_reg <- lm(lab_data$y~cube_root_a)


par(mfrow = c(2,2))
plot(log_a, lab_data$y, xlab= "log(a)", ylab = "y")
lines(log_a,predict.lm(log_a_reg,data.frame(x = lab_data$a)), col = "yellow")

plot(exp_a, lab_data$y, xlab= "exp(a)", ylab = "y")
lines(exp_a,predict.lm(exp_a_reg,data.frame(x = lab_data$a)), col = "blue")

plot(sqrt_a, lab_data$y, xlab= "sqrt(a)", ylab = "y")
lines(sqrt_a,predict.lm(sqrt_a_reg,data.frame(x = lab_data$a)), col = "purple")

plot(cube_root_a, lab_data$y, xlab= "cube_root(a)", ylab = "y")
lines(cube_root_a,predict.lm(cube_reg,data.frame(x = lab_data$a)), col = "red")
```

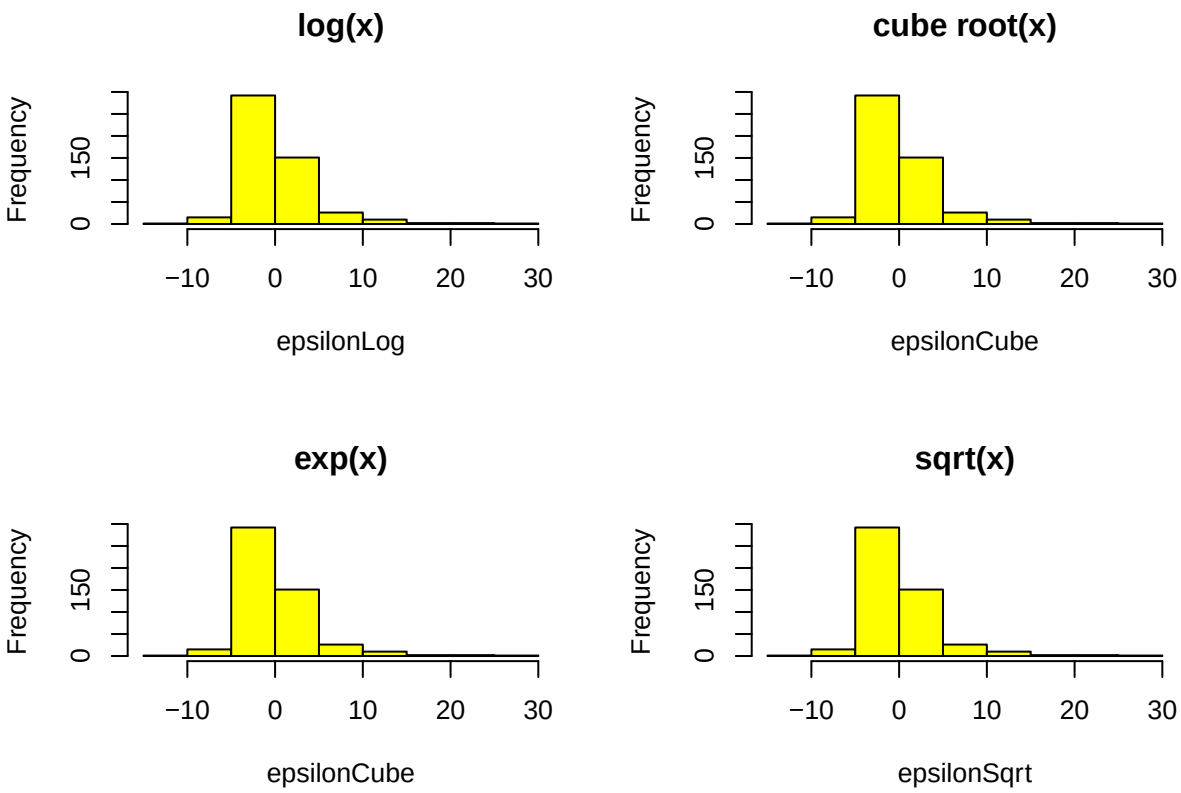


Comment on the predicted model versus ground truth:

All four transformations fail to represent the relationship between y and transformed a. Specifically, the predicted model indicates that when transformed a is greater than M(M ~0.5 for log(a), ~5.2 for exp(a),~1.28 for sqrt(a), and ~ 1.18 for a ^1/3), y remains nearly constant and when transformed a is greater than M, y experience a slight increase. This trend does not reflect the behavior of y with respect to x accurately since the model over-estimates most of y-values when transfomed a is greater than M and unaccurately fit for x smalller than M.

Check residuals:

```
par(mfrow = c(2,2))
epsilonLog <- residuals(log_a_reg)
hist(epsilonLog,col = "yellow", main = "log(x)")

epsilonCube <- residuals(cube_reg)
hist(epsilonCube,col = "yellow", main = "cube root(x)")

epsilonExp <- residuals(exp_a_reg)
hist(epsilonCube,col = "yellow", main = "exp(x)")

epsilonSqrt <- residuals(sqrt_a_reg)
hist(epsilonSqrt, col = "yellow", main = "sqrt(x)")
```

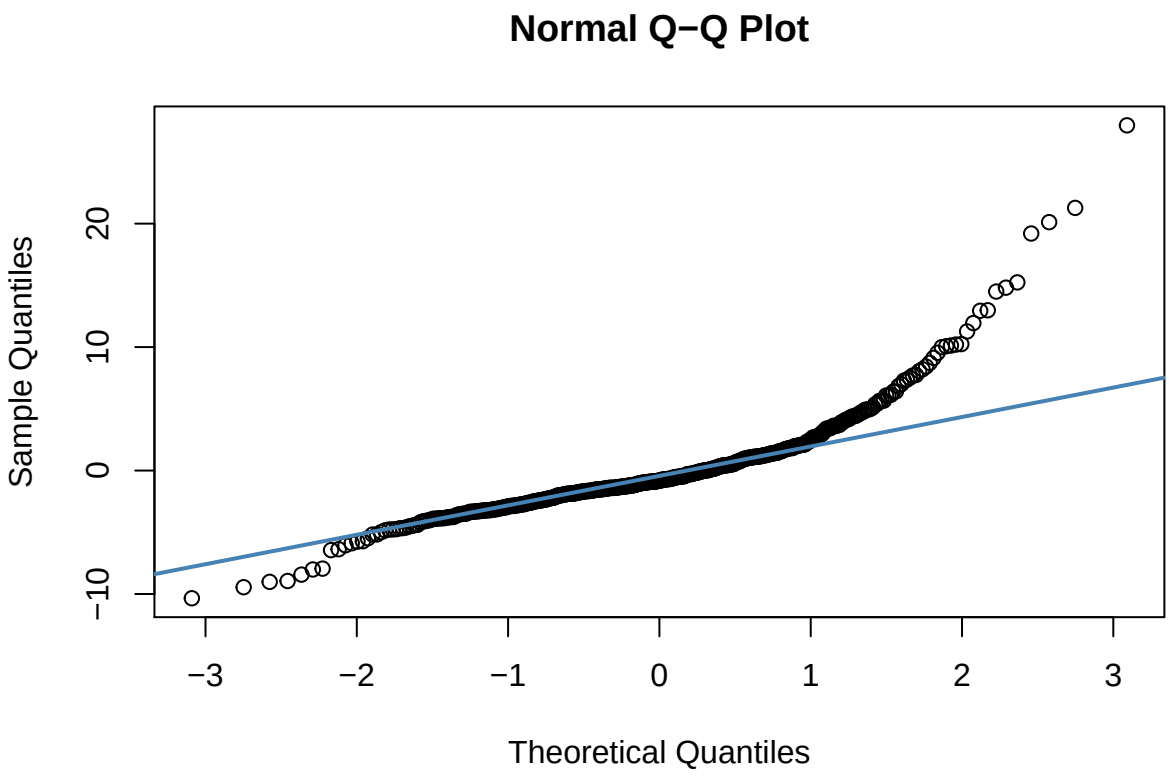


Check residuals:

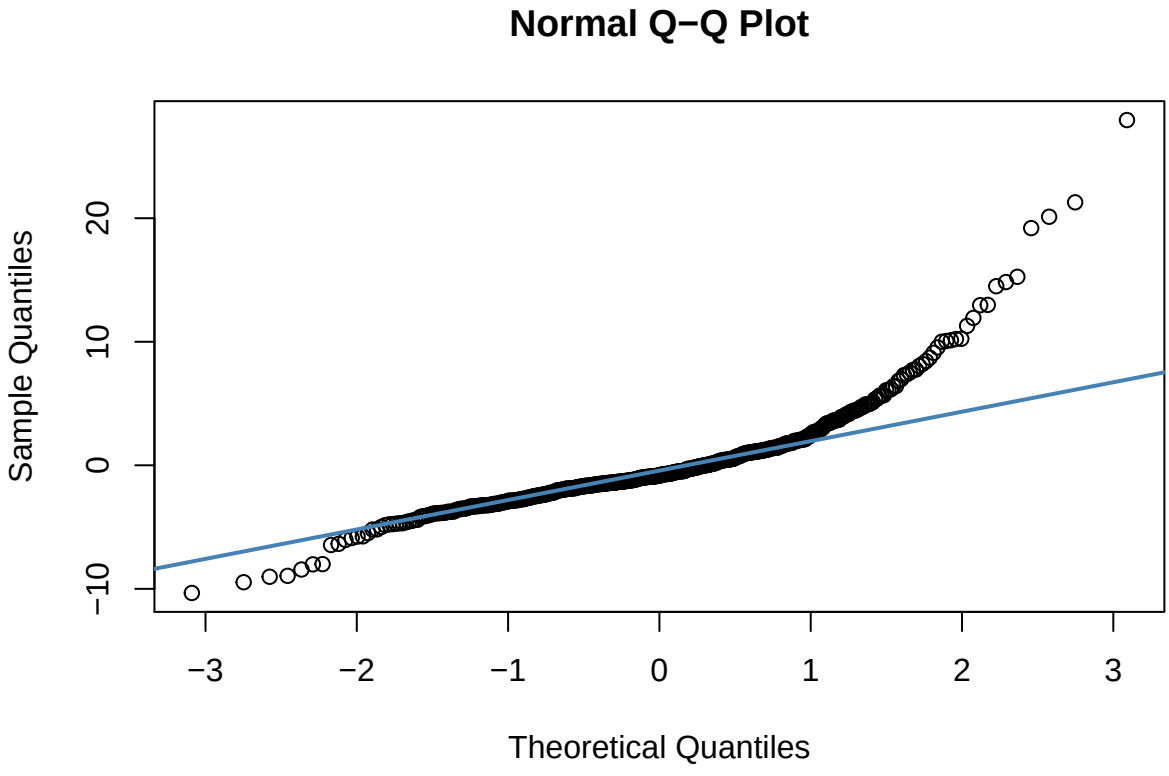
The error distribuition looks the same among four transformations. All transformation's errors are between -10 and -15 and are roughly symmetric around 0.

QQ plot

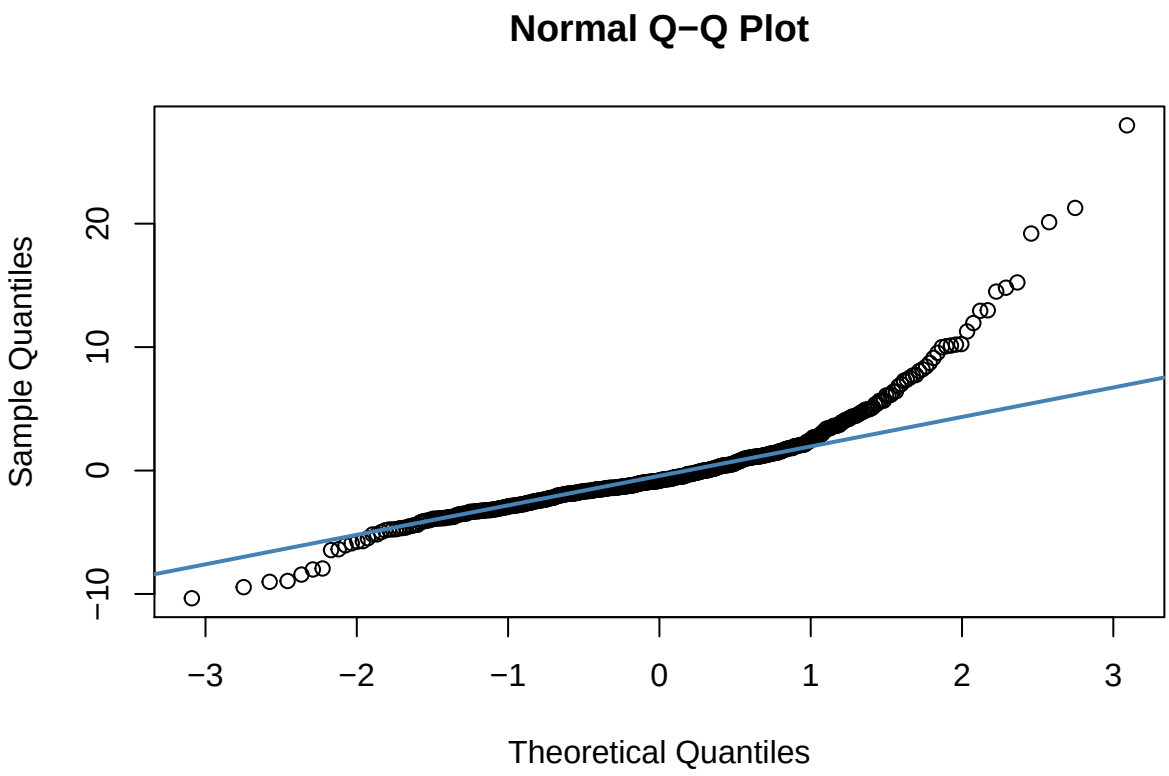
```
qqnorm(epsilonCube)
qqline(epsilonCube, col = "steelblue", lwd = 2)
```



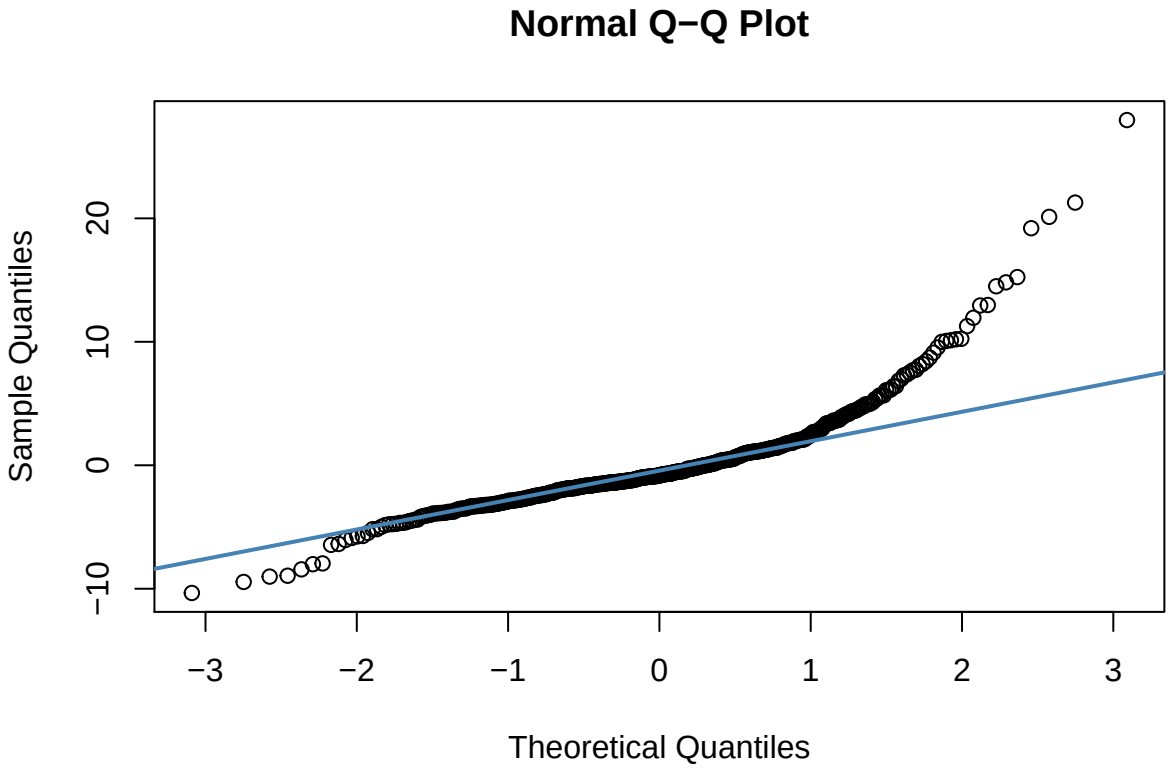

```r
qqnorm(epsilonExp)
qqline(epsilonExp, col = "steelblue", lwd = 2)
```




```r
qqnorm(epsilonLog)
qqline(epsilonLog, col = "steelblue", lwd = 2)
```




```
qqnorm(epsilonSqrt)
qqline(epsilonSqrt, col = "steelblue", lwd = 2)
```

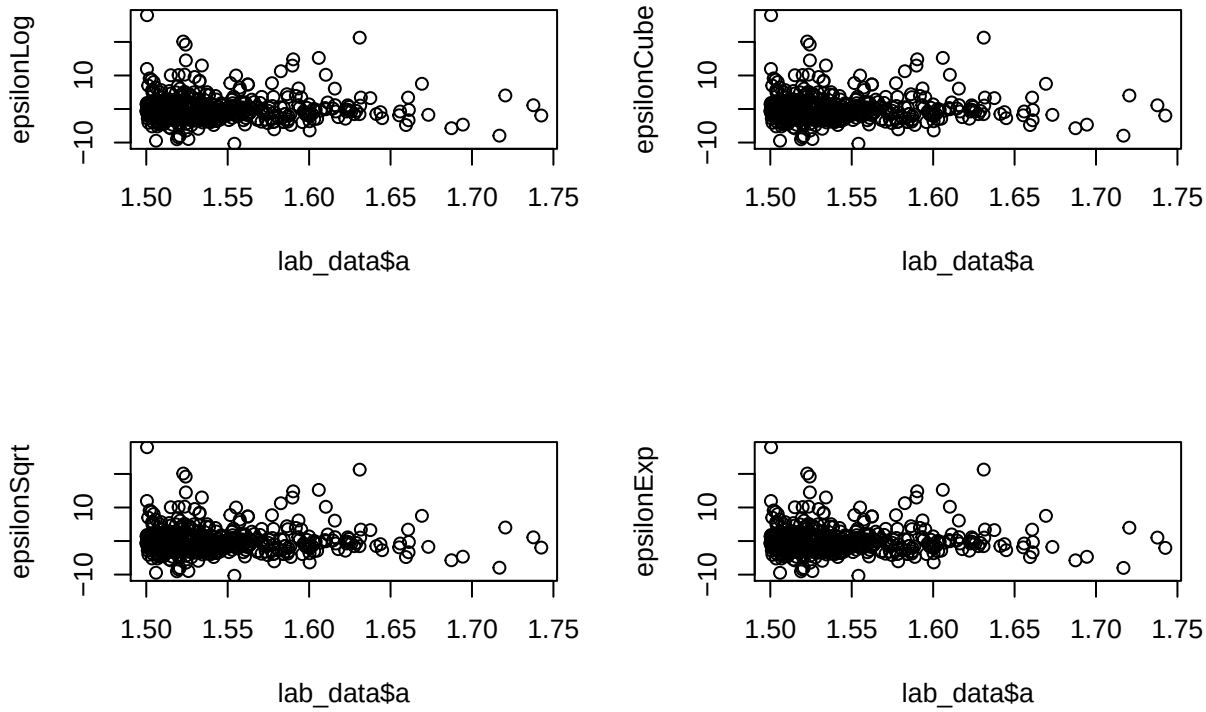


The QQ plot looks the same in all four transformations. The residuals fall along the line in the middle of the line and the left extremety but curve off in the right extremety(heavy right tail), suggesting an indication of

positive skew. All of the transformations violate the N assumption.

#Variance

```
par(mfrow = c(2,2))
plot(lab_data$a,epsilonLog)
plot(lab_data$a,epsilonCube)
plot(lab_data$a,epsilonSqrt)
plot(lab_data$a,epsilonExp)
```



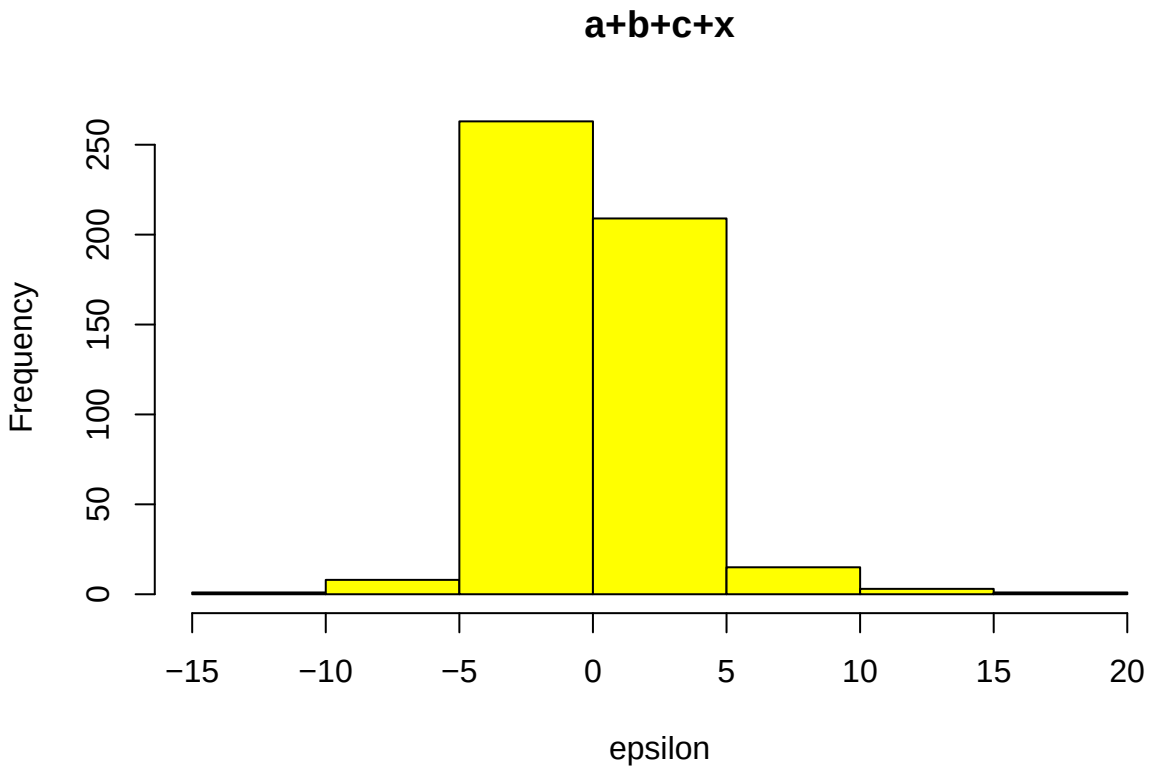
There is an inequal scatter of points around 0. Specifically, there is an extremely dense cluster of points for a less than 1.60 and the points begin to spread out after 1.65. ====> strong indication of non-linearity and violation of E assumption.

Regression model between y and (x, a, b, c)

```
multiple_reg <- lm(lab_data$y~lab_data$b + lab_data$a+ lab_data$c + lab_data$x)
print(multiple_reg)
```

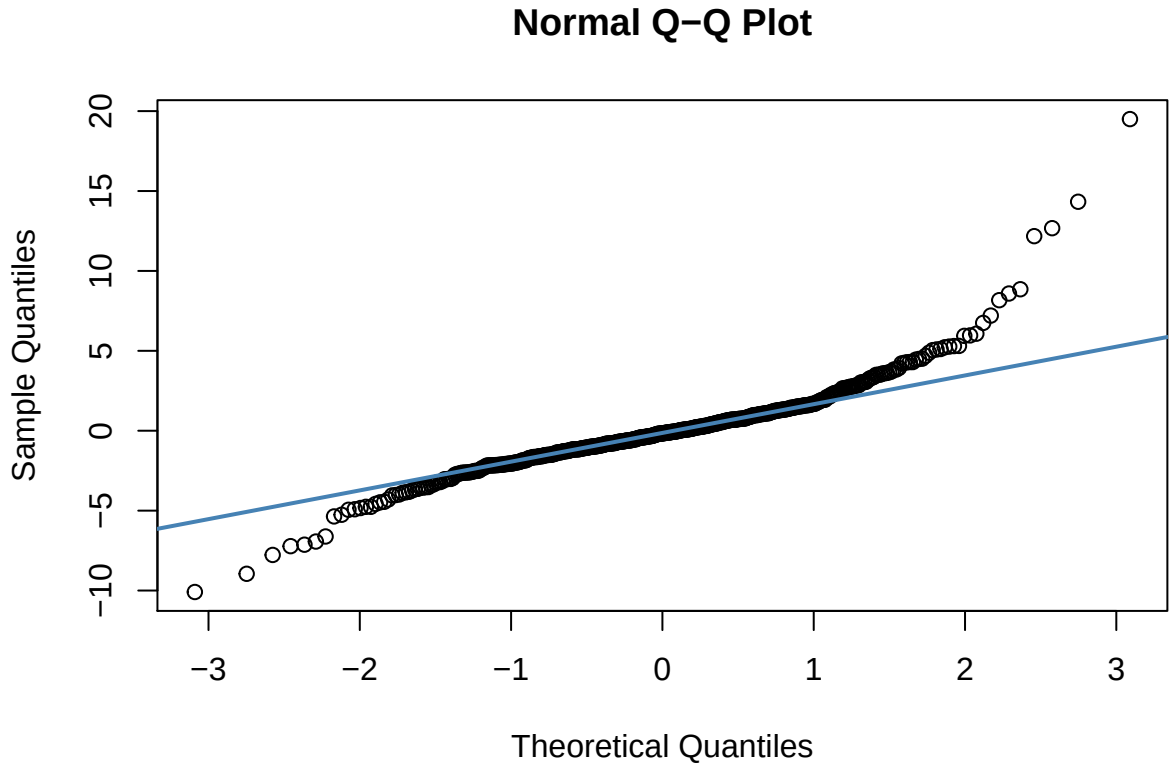
```
##
## Call:
## lm(formula = lab_data$y ~ lab_data$b + lab_data$a + lab_data$c +
##     lab_data$x)
##
## Coefficients:
## (Intercept)   lab_data$b   lab_data$a   lab_data$c   lab_data$x
##      6.5123      -0.2199      -1.2917      -0.2159       3.0016
```

```
epsilon <- residuals(multiple_reg)
hist(epsilon,col = "yellow", main = "a+b+c+x")
```



The distribution of errors has a bell-shaped curves ==> the distribution follows normal distribution(errors are from -15 to 15 and are symmetric around 0)

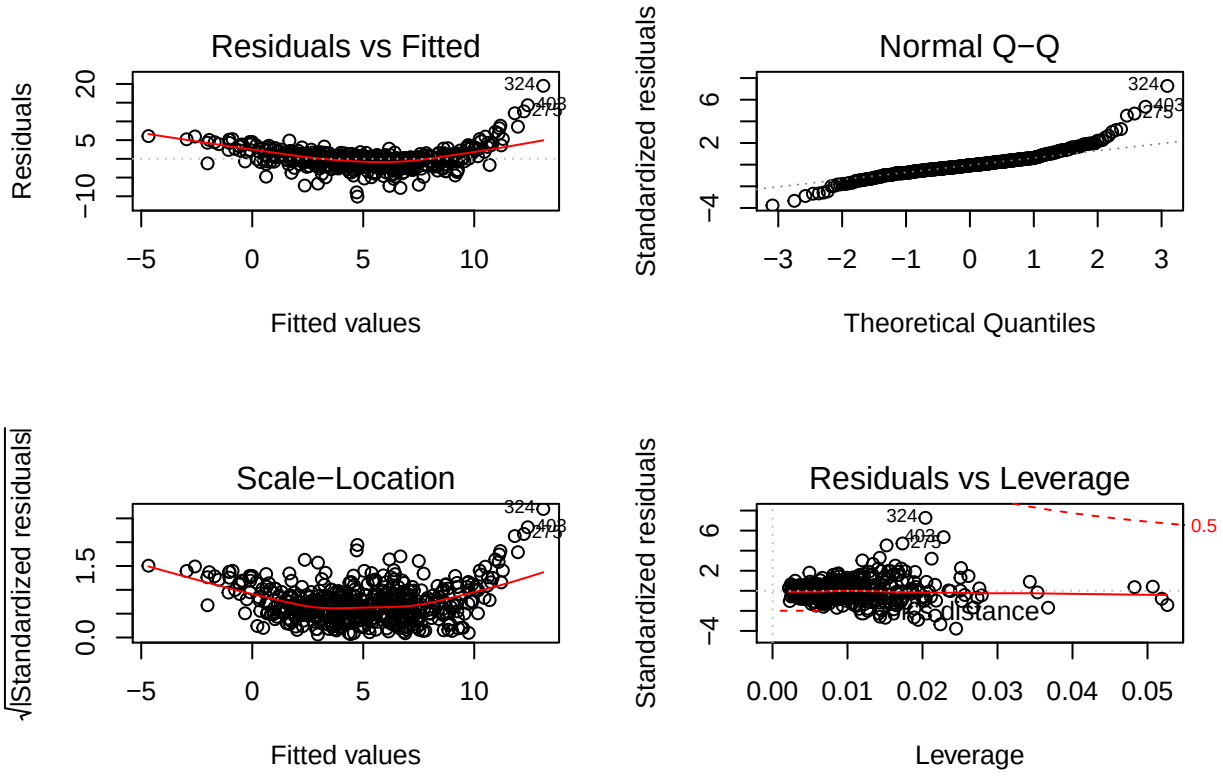
```
qqnorm(epsilon)
qqline(epsilon, col = "steelblue", lwd = 2)
```



The

points fall on the middle of the line but curves off in the extremeties, meaning there are more extreme data values than expected => the N assumption may be violated.

```
par(mfrow = c(2,2))
plot(multiple_reg)
```



According to the plot of the res vs fitted, there is an unequal scatter of points around 0 (the scatter of points follows U-shape)==> strong indication of non-linearity in the model

Also, the residual plot exhibits an U-shape => heteroscedasticity presents in the model => E assumption is violated

L: Violated(reason above)

I: Assume all the obserations/data points are independent

N: Violated(reason above)

E: Violated(reason above)

What transforms do you recommend be applied to each explanatory variable and why?

For x: exponential relationship ==> log transformation considered

For b: quadratic relationship ====> square root transformation considered

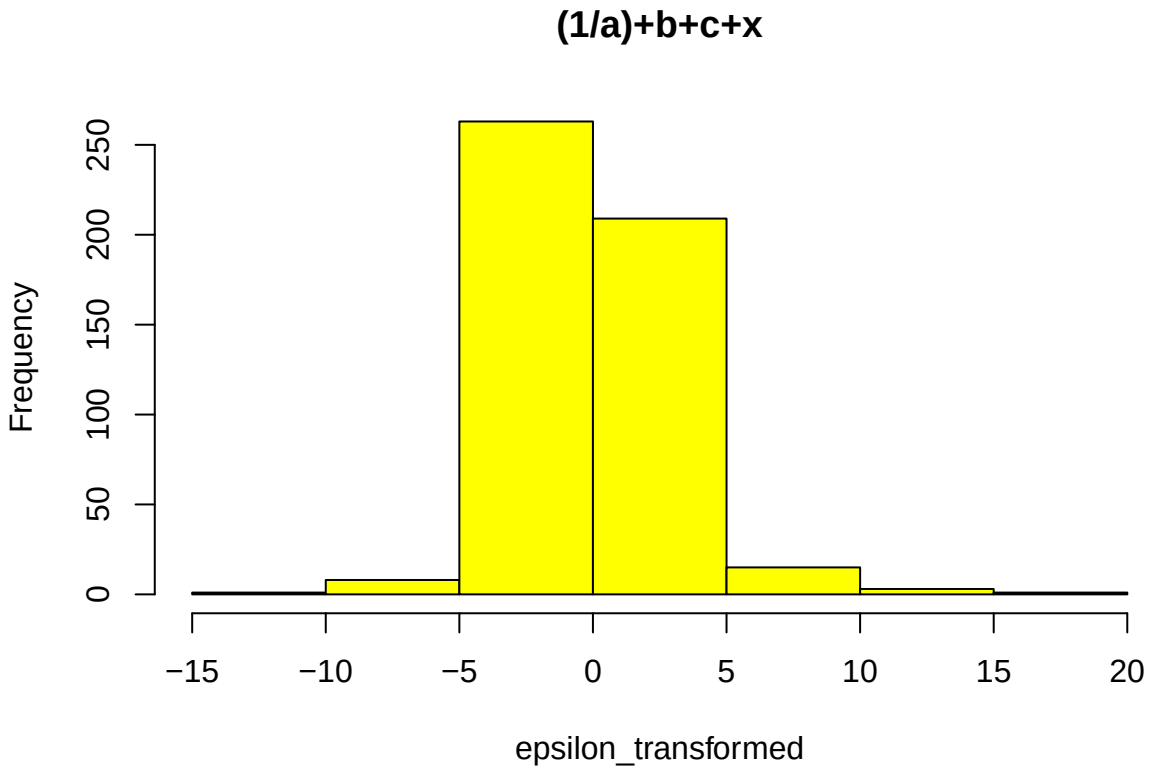
For a and c: minor changes in x values correspond to increase in y ===> reciprocal considered since this type of transformation drastically change the shape of the distribution.

Transform a: 1/a

```r
transform_a <- 1/lab_data$a
multiple_reg_transformed <- lm(lab_data$y~lab_data$b + transform_a+ lab_data$c + lab_data$x)
print(multiple_reg_transformed)
```

```
##
## Call:
## lm(formula = lab_data$y ~ lab_data$b + transform_a + lab_data$c +
##     lab_data$x)
##
## Coefficients:
## (Intercept)   lab_data$b  transform_a   lab_data$c   lab_data$x
##      2.4819      -0.2200       3.1421      -0.2157       3.0015
```

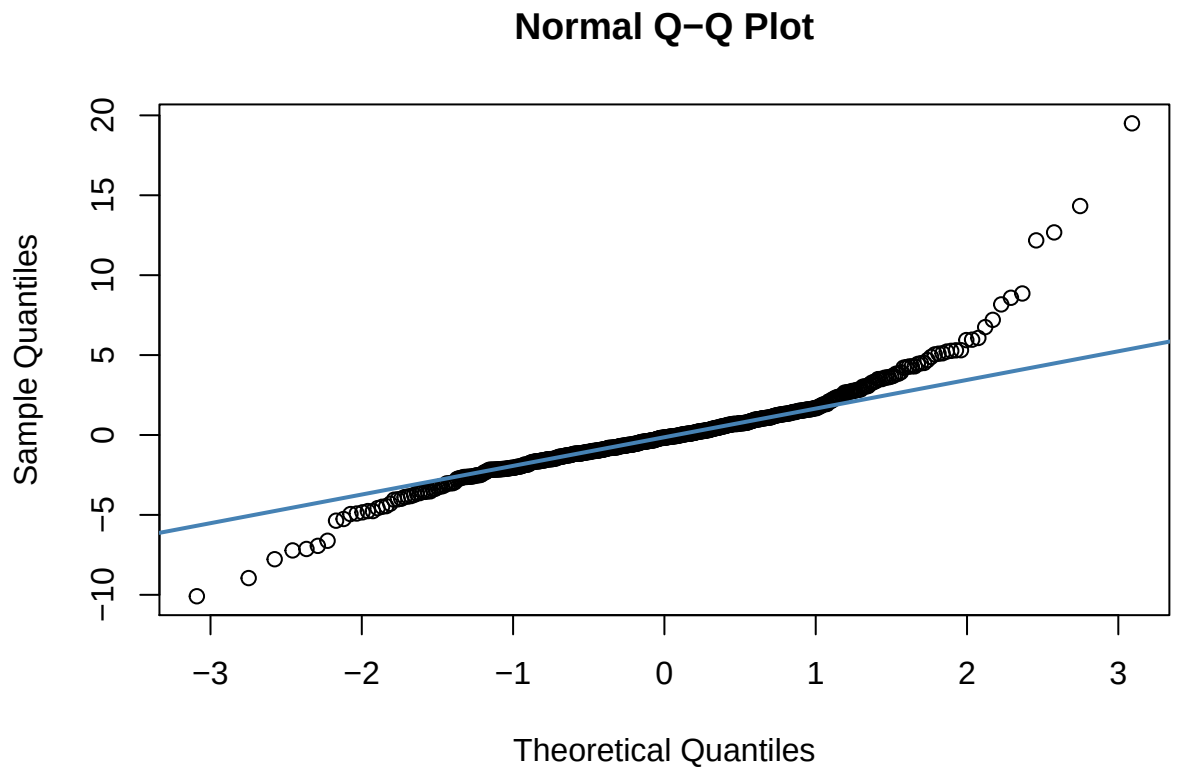
```r
epsilon_transformed <- residuals(multiple_reg_transformed)
hist(epsilon_transformed,col = "yellow", main = "(1/a)+b+c+x")
```



The distribution of errors has a bell-shaped curves ==> the distribution follows normal distribution(errors are from -15 to

2- and are roughly symmetric around 0)

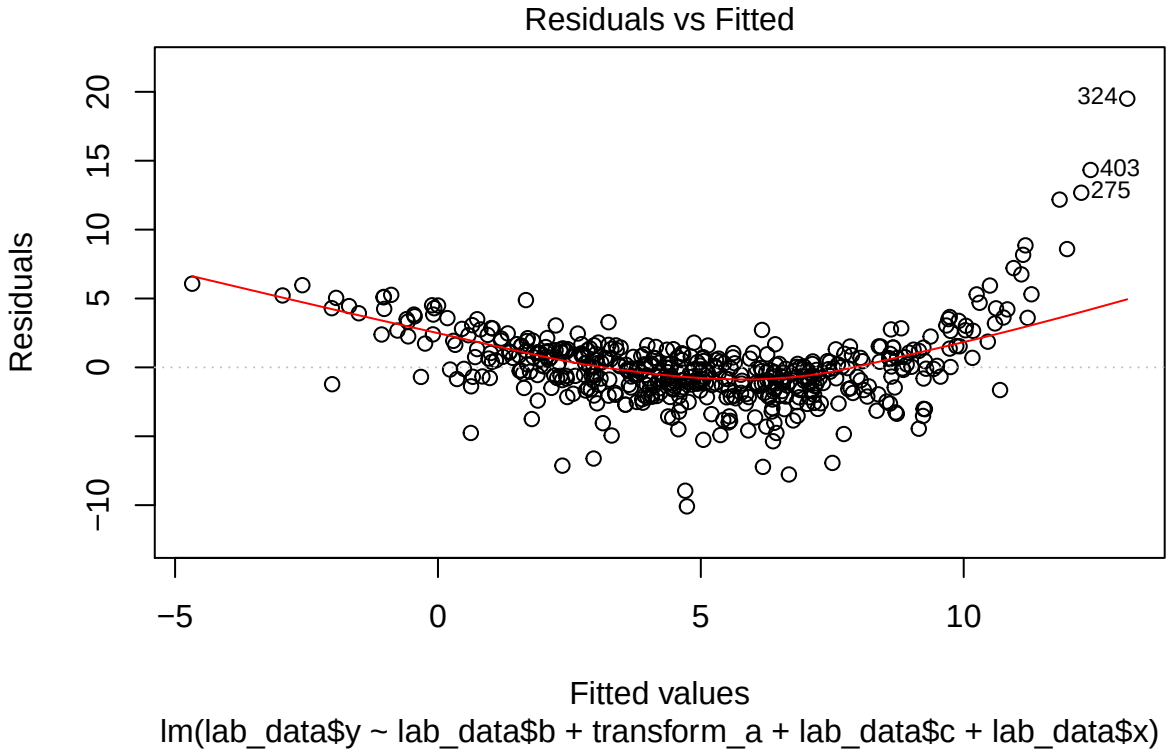
```r
qqnorm(epsilon_transformed)
qqline(epsilon_transformed, col = "steelblue", lwd = 2)
```

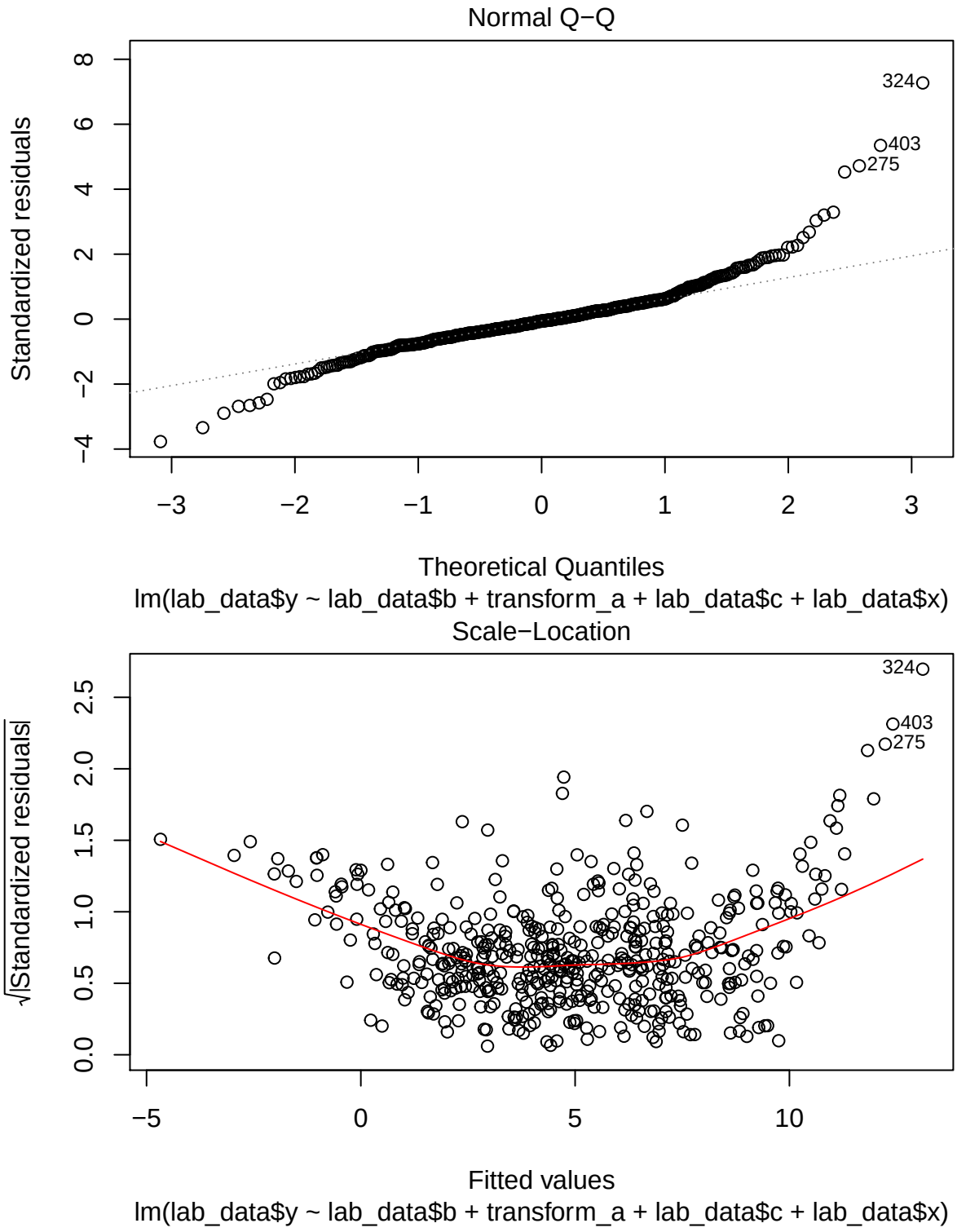


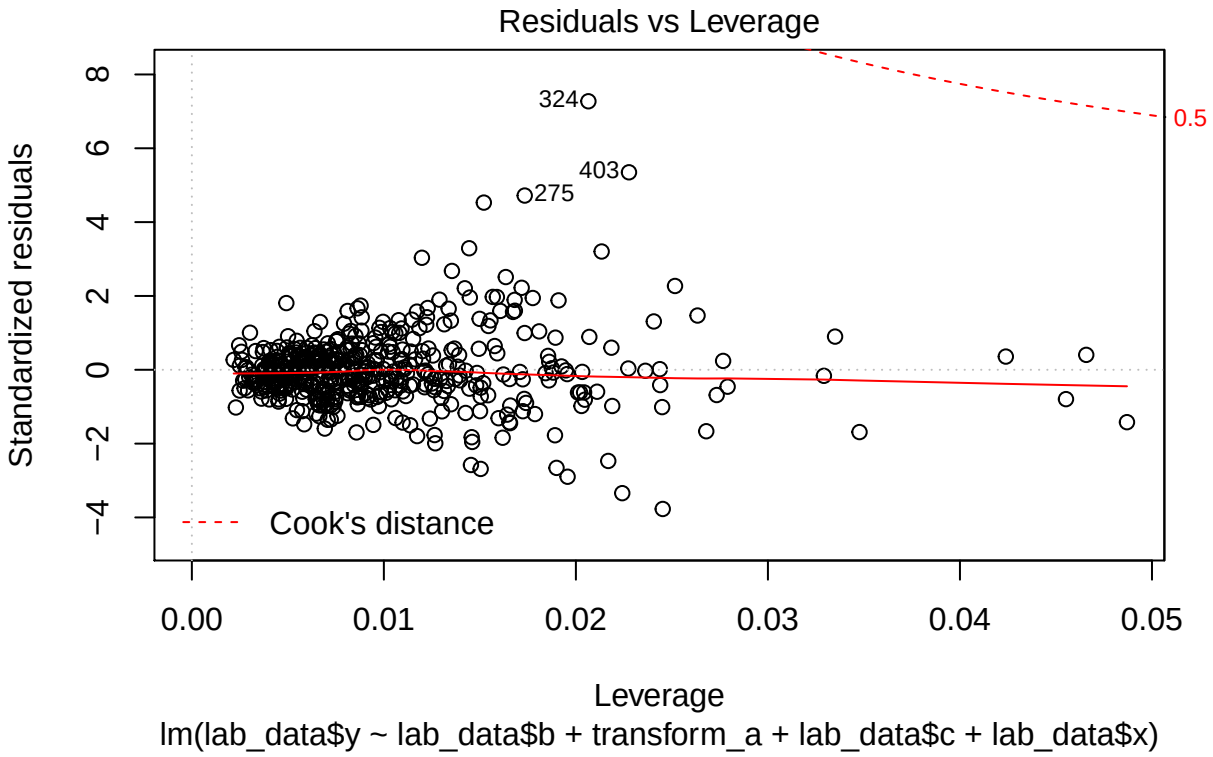
The points fall on the middle of the line but curves off in the extremeties, meaning there are more extreme data values than

expected => the N assumption may be violated.

```
plot(multiple_reg_transformed)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(lab_data$y ~ lab_data$b + transform_a + lab_data$c + lab_data$x)
Scale−Location
√|Standardized residuals|
Fitted values
lm(lab_data$y ~ lab_data$b + transform_a + lab_data$c + lab_data$x)
324
403
275



According to the plot of the res vs fitted, there is an unequal scatter of points around 0(the scatter of points follows U-shape)==> E assumption is violated

L: unable to determine if linearity is violated(do not have plot to observe pattern, t-test needed)

I: Assume all the obserations/data points are independent

N: Violated(reason above)

E: Violated(reason above)

Does transforming this variable help? why or why not? No, it does not help. The model after transforming a still violates LINE assumption. This is because the reciprocal transformation of variable "a" itselt fails to satisfy the LINE assumptions and therefore, is not the good choice in this case. Moreover, the influence of other variables in the regression play a big part in the violation of the assumptions and the transformation of only one variable is unable to get rid of the effects of other variables, which does not result in any improvement in the model.

Perform transformations of x and b, respectively.

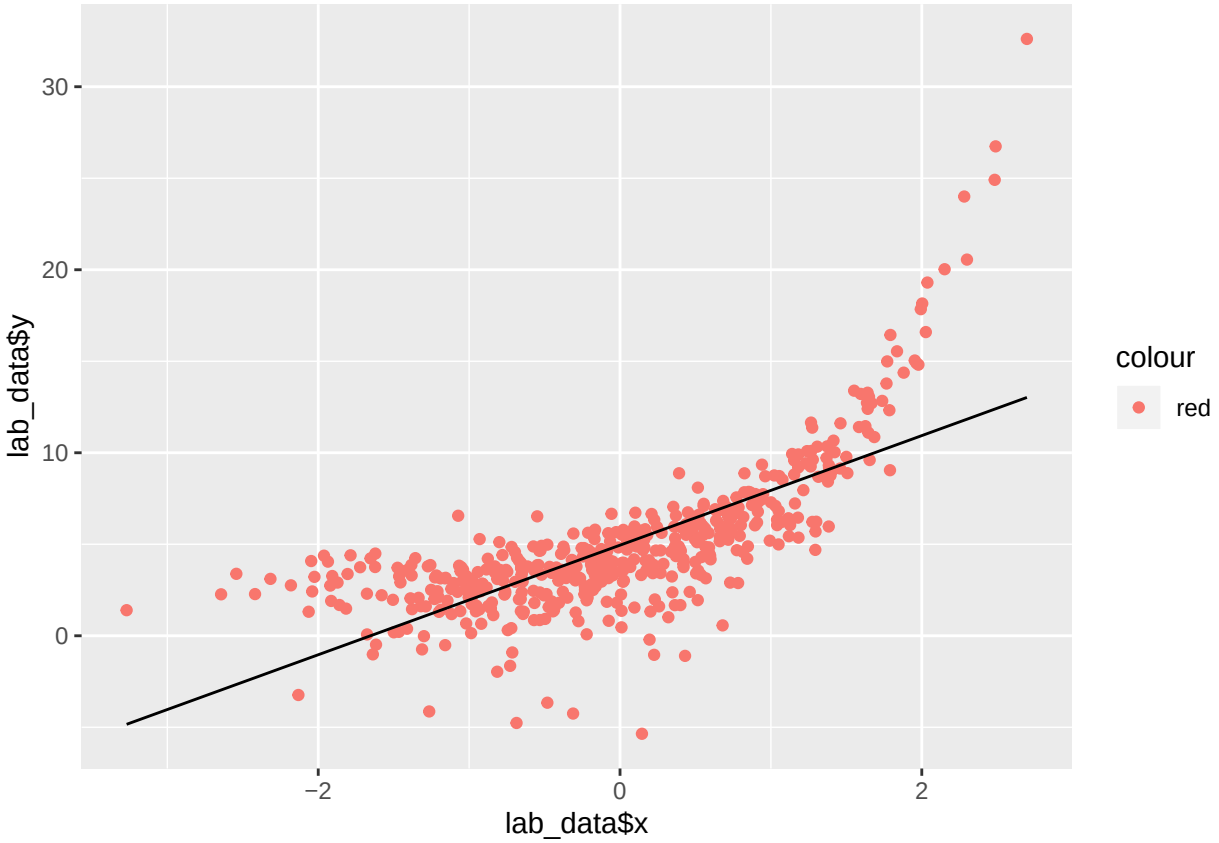
Note: Since there are negative values in x and b, log transformation or sqrt transformation might not be a good option; therefore, the model obtained from there transformations might give some inaccuracies in the analysis of assumptions(include analysis of residual plot, QQ plot due to insufficient transformed data.)

```
lin_reg_x <- lm(lab_data$y~lab_data$x)
print(lin_reg_x)
```

```
##
## Call:
## lm(formula = lab_data$y ~ lab_data$x)
##
```

```
## Coefficients:
## (Intercept)   lab_data$x
##       4.950        2.994
```

```
ggplot()+
  geom_point(mapping = aes(x = lab_data$x, y = lab_data$y,  col = "red"))+
  geom_line(mapping = aes(x = lab_data$x , y =fitted(lin_reg_x)))
```

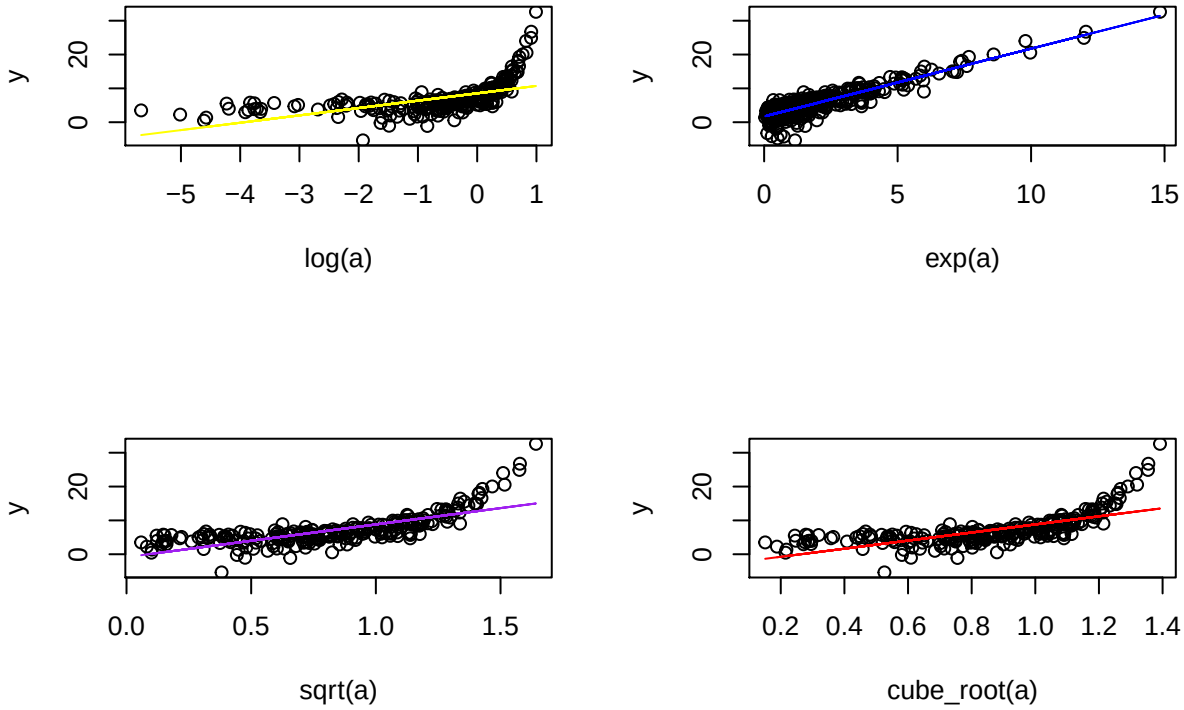



```
# Comment on the predicted model versus ground truth:
# The predicted model indicates that for each (unit) increase in x , y increases by 2.994. This trend r
# behavior of y for x in [-2,1.8] but under-estimate for x greater than 2. The relationship between x a
# plot is not linear and the predicted model is unable to capture the exponetial growth in y when x inc
# is violated.


###Transform x. Four tranformations applied to x are: log(x), exp(x), x^(1/2), and x^(1/3)
log_a_x <- log(lab_data$x)
```

```
## Warning in log(lab_data$x): NaNs produced
```

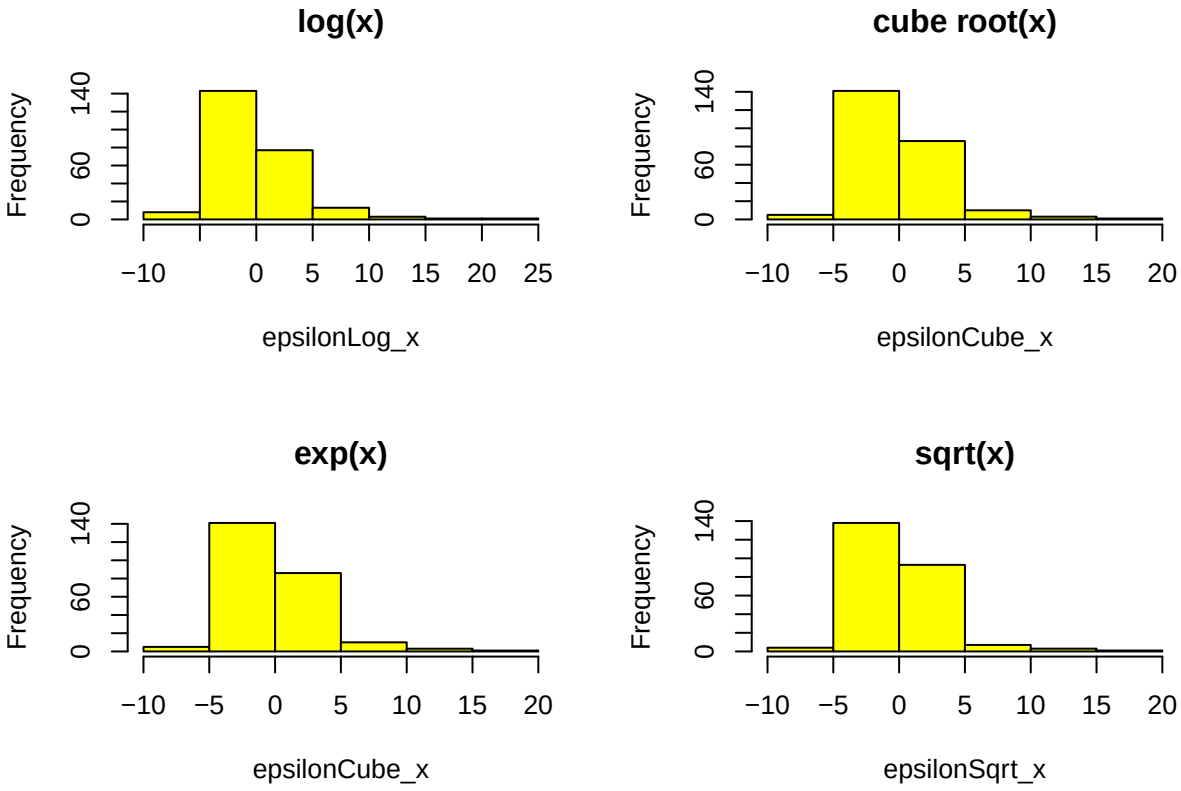
```
exp_a_x <- exp(lab_data$x)
sqrt_a_x <- (lab_data$x^(1/2))
cube_root_a_x <- (lab_data$x^(1/3))

par(mfrow = c(2,2))
log_a_reg_x <- lm(lab_data$y~log_a_x)

exp_a_reg_x <- lm(lab_data$y~exp_a_x)
```

```r
sqrt_a_reg_x <- lm(lab_data$y~sqrt_a_x)

cube_reg_x <- lm(lab_data$y~cube_root_a_x)


par(mfrow = c(2,2))
plot(log_a_x, lab_data$y, xlab= "log(a)", ylab = "y")
lines(log_a_x,predict.lm(log_a_reg_x,data.frame(x = lab_data$x)), col = "yellow")

plot(exp_a_x, lab_data$y, xlab= "exp(a)", ylab = "y")
lines(exp_a_x,predict.lm(exp_a_reg_x,data.frame(x = lab_data$x)), col = "blue")

plot(sqrt_a_x, lab_data$y, xlab= "sqrt(a)", ylab = "y")
lines(sqrt_a_x,predict.lm(sqrt_a_reg_x,data.frame(x = lab_data$x)), col = "purple")

plot(cube_root_a_x, lab_data$y, xlab= "cube_root(a)", ylab = "y")
lines(cube_root_a_x,predict.lm(cube_reg_x,data.frame(x = lab_data$x)), col = "red")
```

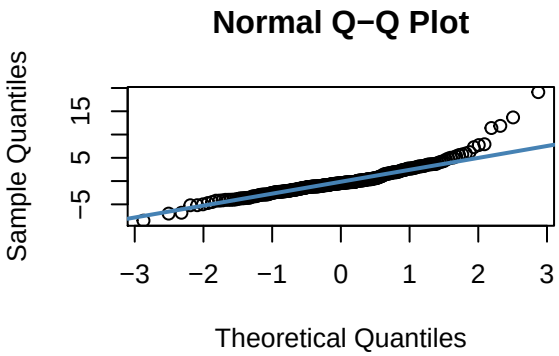


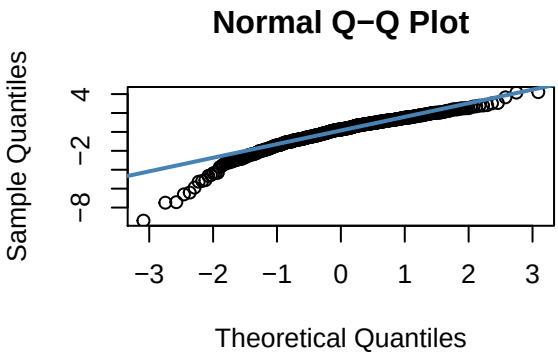

```r
# For log transformation, the regression line does not fit the relationship between y and log(x). The l
# y for x greater than 0. The relationship appears non-linear(exponential relationship)=> violates the

# For exponential transformation, the line seems to fit the data much better in comparison with other t
# However, the dense cluster of y values under 7 is quite concerning. The L assumption may be violated.

# For sqrt and cube root transformation, the line does not capture well the relationship between the tr
# a litte bit underestimates y for x in [0.5,1.5] and fails to reflect the exponential growth in y for


#Check residuals:
par(mfrow = c(2,2))
epsilonLog_x <- residuals(log_a_reg_x)
```

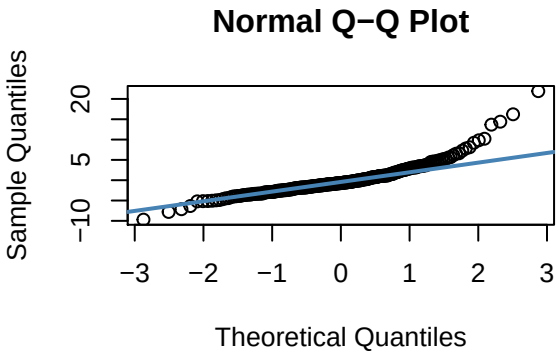
```r
hist(epsilonLog_x,col = "yellow", main = "log(x)")

epsilonCube_x <- residuals(cube_reg_x)
hist(epsilonCube_x,col = "yellow", main = "cube root(x)")

epsilonExp_x <- residuals(exp_a_reg_x)
hist(epsilonCube_x,col = "yellow", main = "exp(x)")

epsilonSqrt_x <- residuals(sqrt_a_reg_x)
hist(epsilonSqrt_x, col = "yellow", main = "sqrt(x)")
```

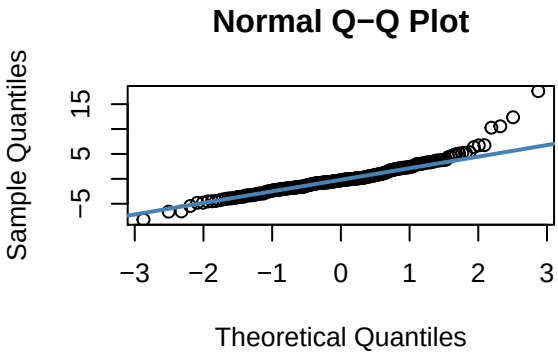



```r
#The error distribution looks similar among four transformations. All transfomation's errors are betwee
#a little bit skewed to the right.

#QQ plot
par(mfrow = c(2,2))
qqnorm(epsilonCube_x)
qqline(epsilonCube_x, col = "steelblue", lwd = 2)
qqnorm(epsilonExp_x)
qqline(epsilonExp_x, col = "steelblue", lwd = 2)
qqnorm(epsilonLog_x)
qqline(epsilonLog_x, col = "steelblue", lwd = 2)
qqnorm(epsilonSqrt_x)
qqline(epsilonSqrt_x, col = "steelblue", lwd = 2)
```

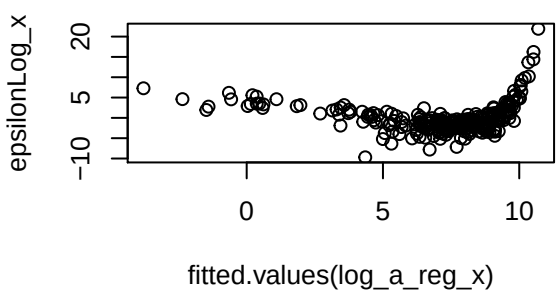


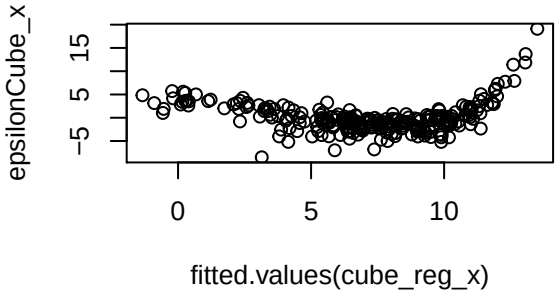


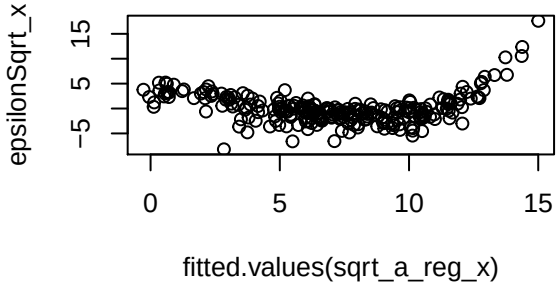


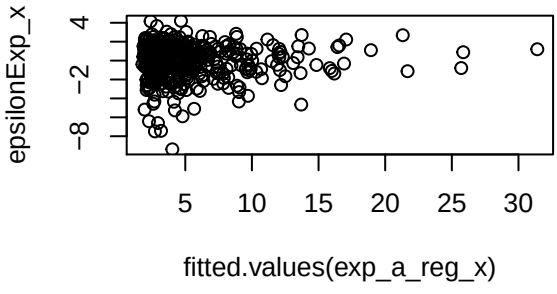



```
# For log transformation, sqrt transformation, and cube root transformation, the residuals fall along t
# the line and the left extremety but curve off in the right extremety.
# However, for exponential transformation, the residuals curve of in the left extremety but fall along
# All of the transformations violate the N assumption.

##Variance
par(mfrow = c(2,2))
plot(fitted.values(log_a_reg_x),epsilonLog_x)
plot(fitted.values(cube_reg_x),epsilonCube_x)
plot(fitted.values(sqrt_a_reg_x),epsilonSqrt_x)
plot(fitted.values(exp_a_reg_x),epsilonExp_x)
```









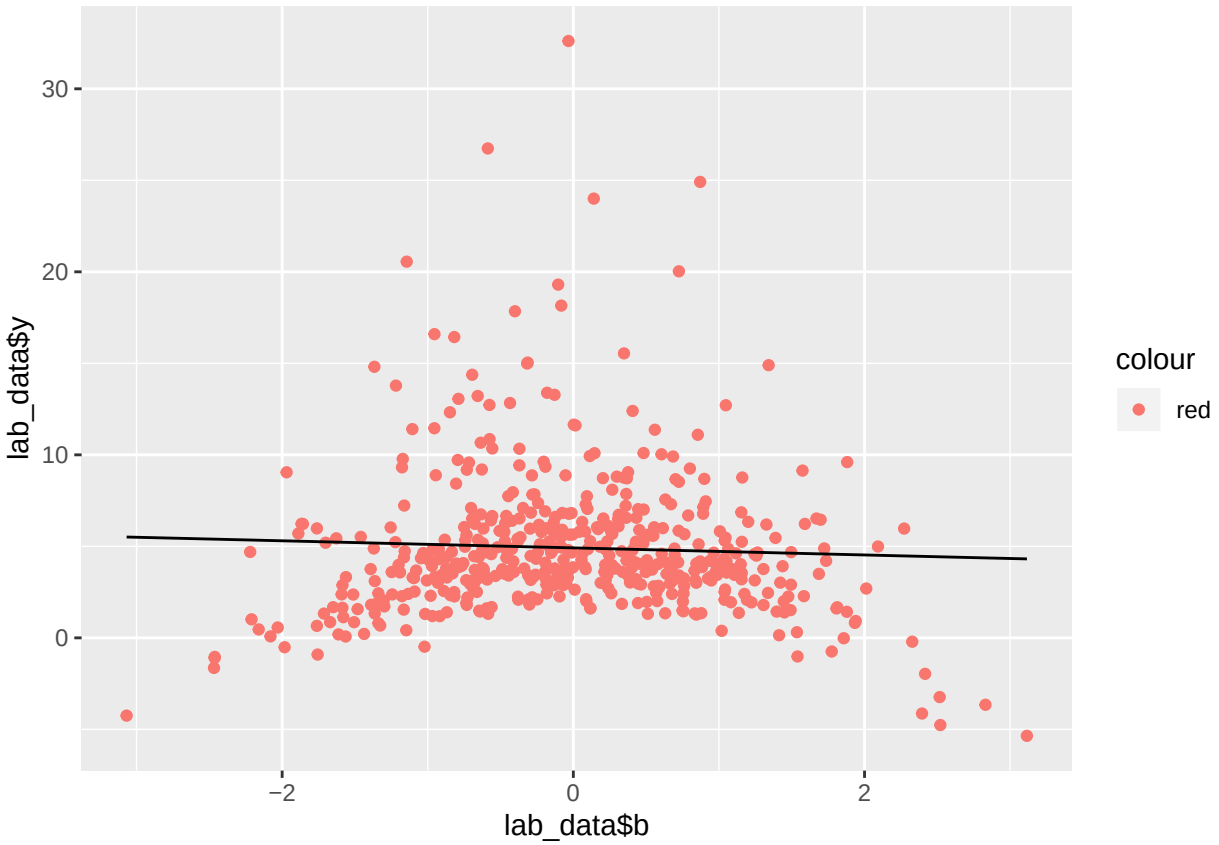

```
##Unequal scatter of points around 0 in all models.
## U-shaped is exhibited in three models(log trans, cube trans, and sqrt) ===> inequality of variance
```

Regression between y and b

```
lin_reg_b <- lm(lab_data$y~lab_data$b)
print(lin_reg_b)
```

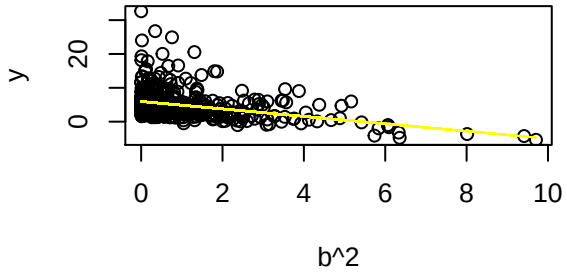
```
##
## Call:
## lm(formula = lab_data$y ~ lab_data$b)
##
## Coefficients:
## (Intercept)   lab_data$b
##      4.9163      -0.1935
```

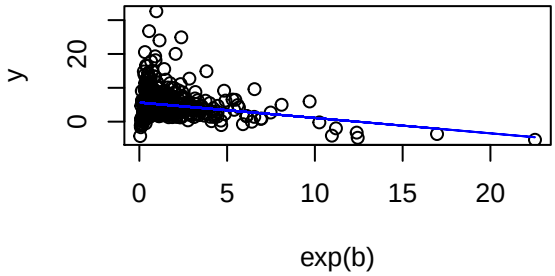
```
ggplot()+
  geom_point(mapping = aes(x = lab_data$b, y = lab_data$y,  col = "red"))+
  geom_line(mapping = aes(x = lab_data$b , y =fitted(lin_reg_b)))
```

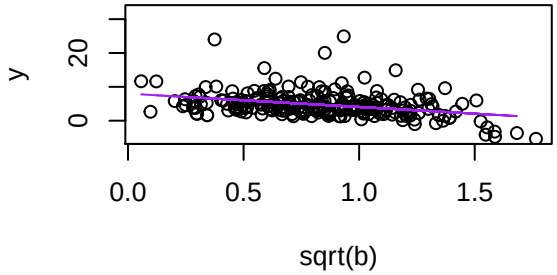


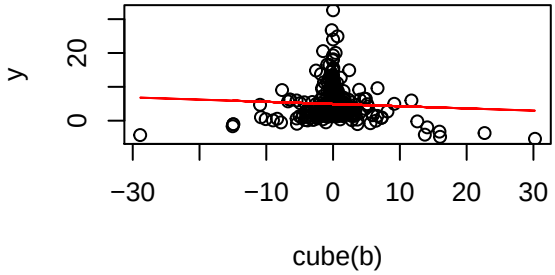

```
# Comment on the predicted model versus ground truth:
# The predicted model indicates that for each (unit) increase in x , y decreases by 0.1935. This trend
# quadratic relationship between y and transformed b( with the negative slope). ==> L assumption is vio


###Transform b. Four tranformations applied to b are: log(b), exp(b), b^(1/2), and b^(1/3)
squared_b <- (lab_data$b)^2
exp_b <- exp(lab_data$b)
sqrt_b <- (lab_data$b^(1/2))
cubed_b <- (lab_data$b^(3))

par(mfrow = c(2,2))
squared_reg_b <- lm(lab_data$y~squared_b)

exp_reg_b <- lm(lab_data$y~exp_b)

sqrt_reg_b <- lm(lab_data$y~sqrt_b)

cubed_reg_b <- lm(lab_data$y~cubed_b)


par(mfrow = c(2,2))
plot(squared_b, lab_data$y, xlab= "b^2", ylab = "y")
lines(squared_b,predict.lm(squared_reg_b,data.frame(x = lab_data$b)), col = "yellow")

plot(exp_b, lab_data$y, xlab= "exp(b)", ylab = "y")
lines(exp_b,predict.lm(exp_reg_b,data.frame(x = lab_data$b)), col = "blue")
```

```r
plot(sqrt_b, lab_data$y, xlab= "sqrt(b)", ylab = "y")
lines(sqrt_b,predict.lm(sqrt_reg_b,data.frame(x = lab_data$b)), col = "purple")

plot(cubed_b, lab_data$y, xlab= "cube(b)", ylab = "y")
lines(cubed_b,predict.lm(cubed_reg_b,data.frame(x = lab_data$b)), col = "red")
```

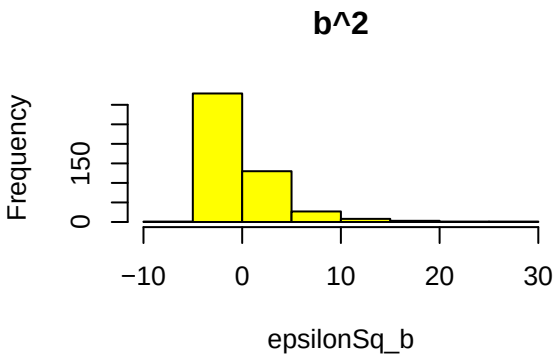


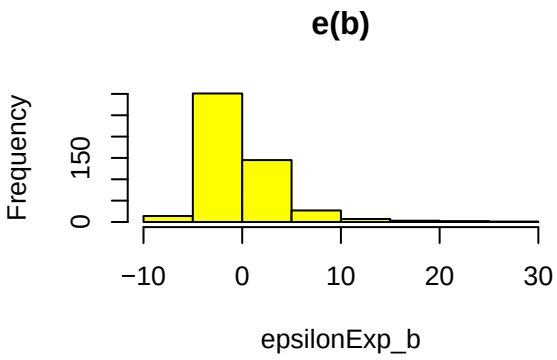


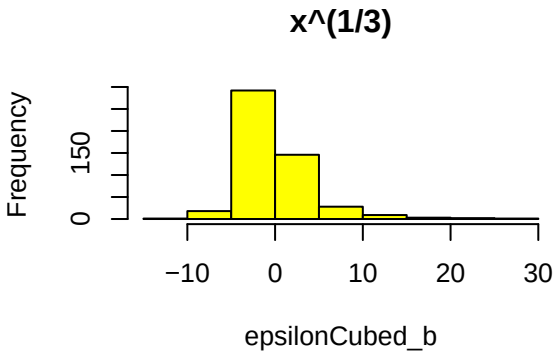


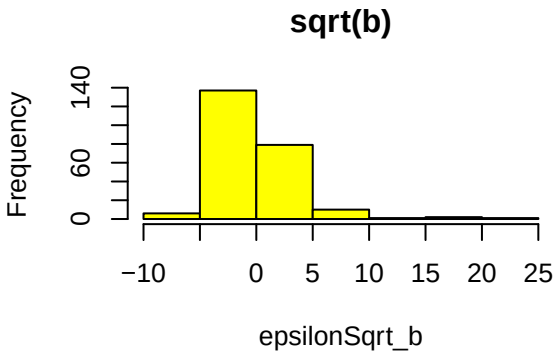



```r
# The sqrt transformation gives a better fit for the relationship between y and transformed b in compar
#transformations ==> L assumption is not violated in this case. For other three transformations, the li
# the relationship/trend between two variables.


#Check residuals:
par(mfrow = c(2,2))
epsilonSq_b <- residuals(squared_reg_b)
hist(epsilonSq_b,col = "yellow", main = "b^2")

epsilonExp_b <- residuals(exp_reg_b)
hist(epsilonExp_b,col = "yellow", main = "e(b)")

epsilonCubed_b <- residuals(cubed_reg_b)
hist(epsilonCubed_b,col = "yellow", main = "x^(1/3)")

epsilonSqrt_b <- residuals(sqrt_reg_b)
hist(epsilonSqrt_b, col = "yellow", main = "sqrt(b)")
```

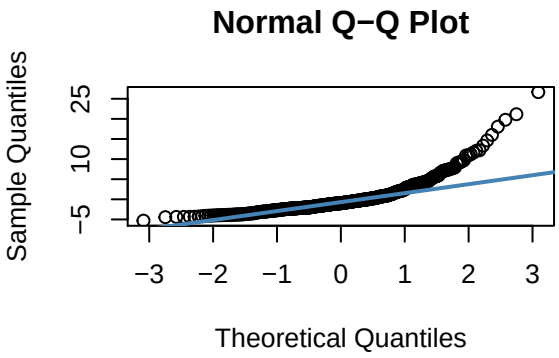


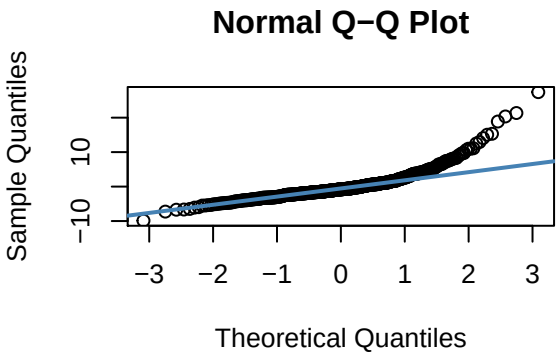


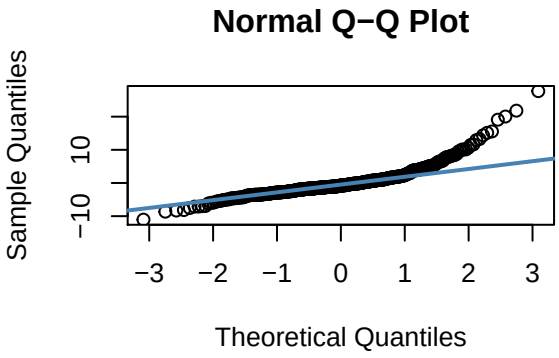


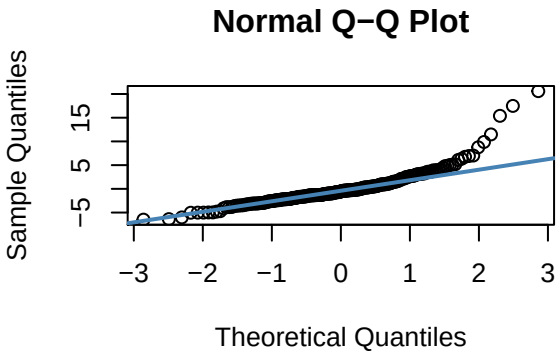



```
#The error distribution looks similar among four transformations. All transfomation's errors are betwee
#roughly symmetric around 0

#QQ plot
par(mfrow = c(2,2))
qqnorm(epsilonSq_b)
qqline(epsilonSq_b, col = "steelblue", lwd = 2)
qqnorm(epsilonExp_b)
qqline(epsilonExp_b, col = "steelblue", lwd = 2)
qqnorm(epsilonCubed_b)
qqline(epsilonCubed_b, col = "steelblue", lwd = 2)
qqnorm(epsilonSqrt_b)
qqline(epsilonSqrt_b, col = "steelblue", lwd = 2)
```

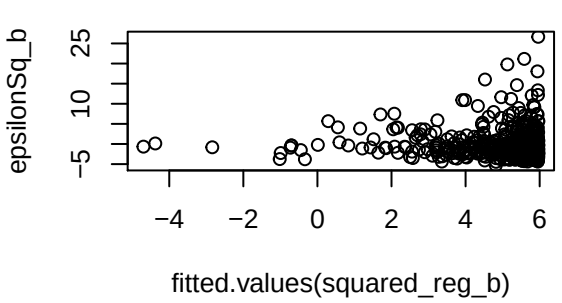


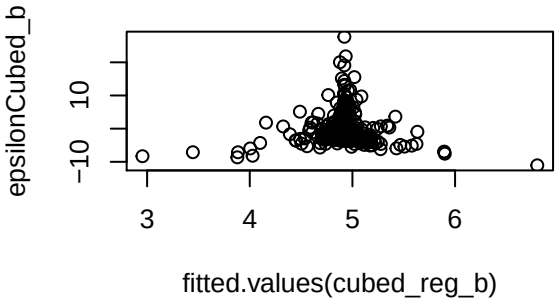


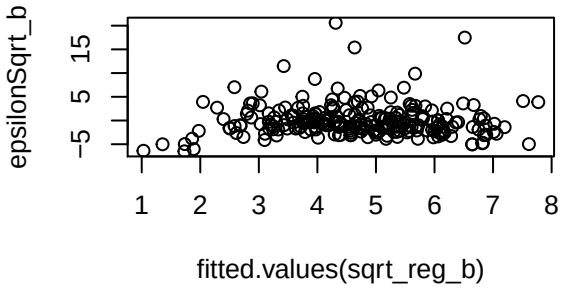


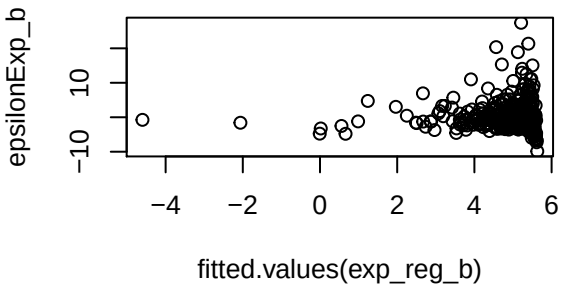



```
# For all transformations, the residuals curve off in the left extremety and fall along the rest of the
# ====> All of the transformations violate the N assumption.

##Variance
par(mfrow = c(2,2))
plot(fitted.values(squared_reg_b),epsilonSq_b)
plot(fitted.values(cubed_reg_b),epsilonCubed_b)
plot(fitted.values(sqrt_reg_b),epsilonSqrt_b)
plot(fitted.values(exp_reg_b),epsilonExp_b)
```










```
##Unequal scatter of points around 0 in all models.
## ===> inequality of variance
```