LINEAR REGRESSION: Homework #4

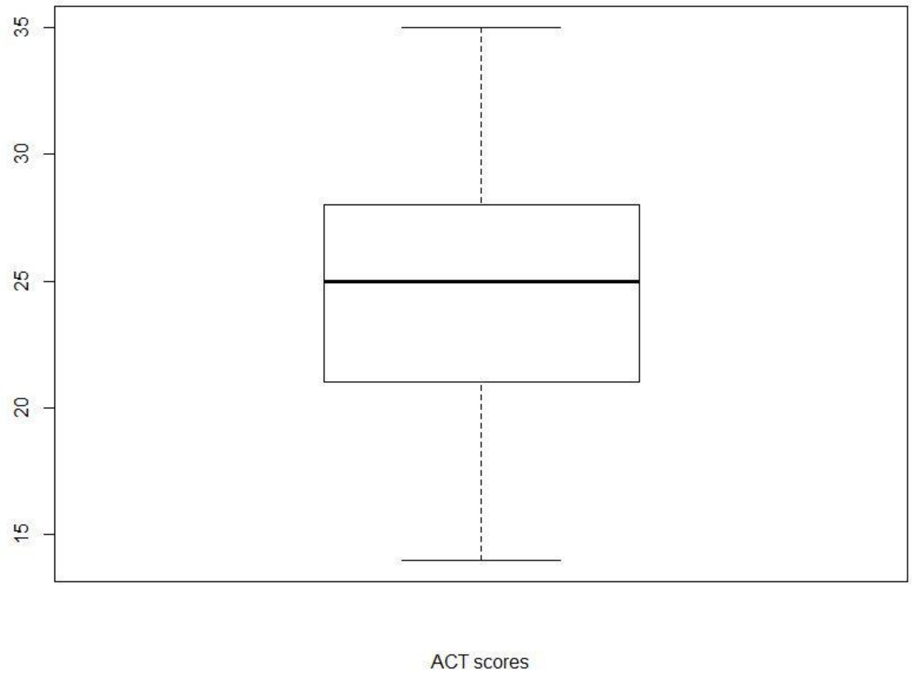
Professor Jingchen Liu

Fan Yang
UNI: fy2232

Problem 1 (3.3)

(a)

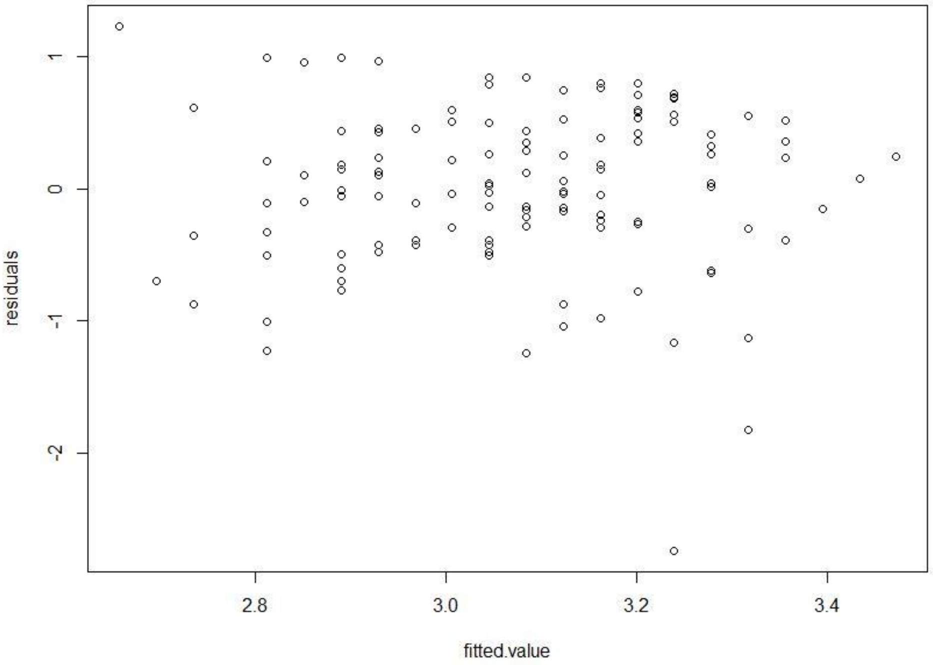
```
> boxplot(da$V2,xlab="ACT scores")
```



From the boxplot, we find most of the scores are between 20 and 30, with a median of around 25, which lie close to the middle of the range. This distribution looks very symmetrical.

(c)

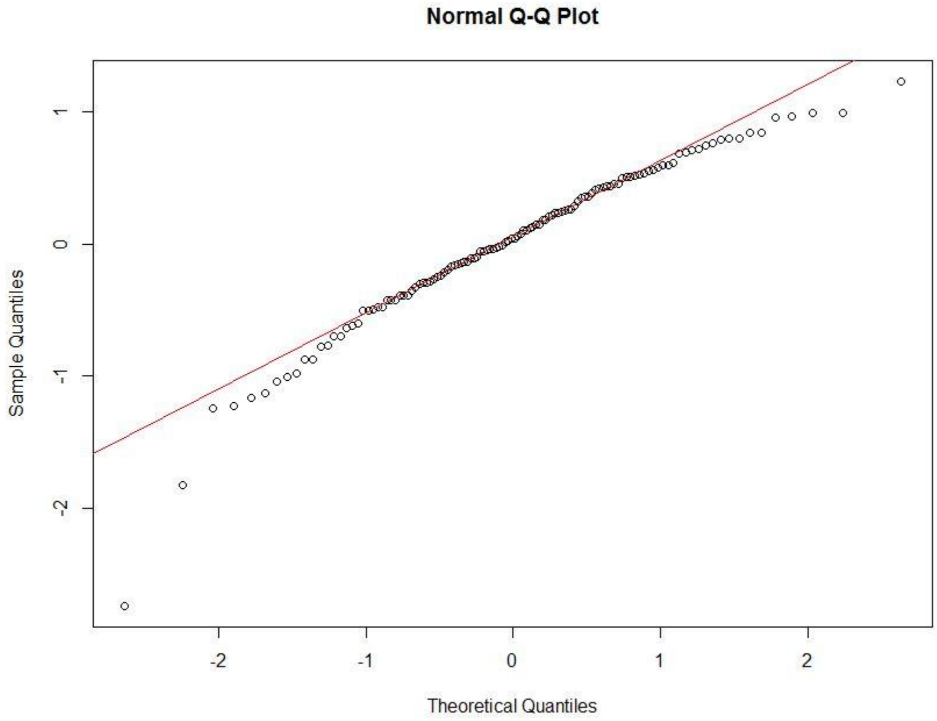
```
> plot(model19$fitted.values, model19$residuals,
+      xlab = "fitted.value", ylab = "residuals")
```



The variance of residuals seemed to be constant with the change of $\hat{Y}$, which indicates a constant variance. The plot indicated the presence of potential outliers. As we can see, most of the residuals were in the range of -2 and 1, however, there were two residuals beyond this range.

(d)

```
> qqnorm(model19$residuals)
> qqline(model19$residuals, col = 2)
>
> StdErr = sum((da$V1-mean(da$V1))^2)
> ExpVals = sapply(1:120, function(k) StdErr*qnorm((k-.375)/(120+.25)))
> cor(sort(ExpVals),sort(model19$residuals))
[1] 0.9737275
```



The coefficient of correlation between the ordered residuals and their expected values under normality is 0.9737275, with n=120. from Table B.6, the critical value for the coefficient of correlation between the ordered residuals and the expected values under normality when the distribution of error terms is normal using a 0.05 significance level is 0.987. Since 0.9737275 < 0.987, the assumption of normality appeared unreasonable.

(e)

```
> res1 <- model19$residuals[da$V2<26]
> res2 <- model19$residuals[da$V2>=26]
>
> d1 <- abs(res1-median(res1))
> d2 <- abs(res2-median(res2))
>
> s <- sqrt((sum((d1-mean(d1))^2) + sum((d2-mean(d2))^2))/(118))
> t <- (mean(d1)-mean(d2))/(s*(sqrt(1/length(res1) + 1/length(res2))))
> t
[1] -0.8967448
> qt(0.995,118)
[1] 2.618137
```

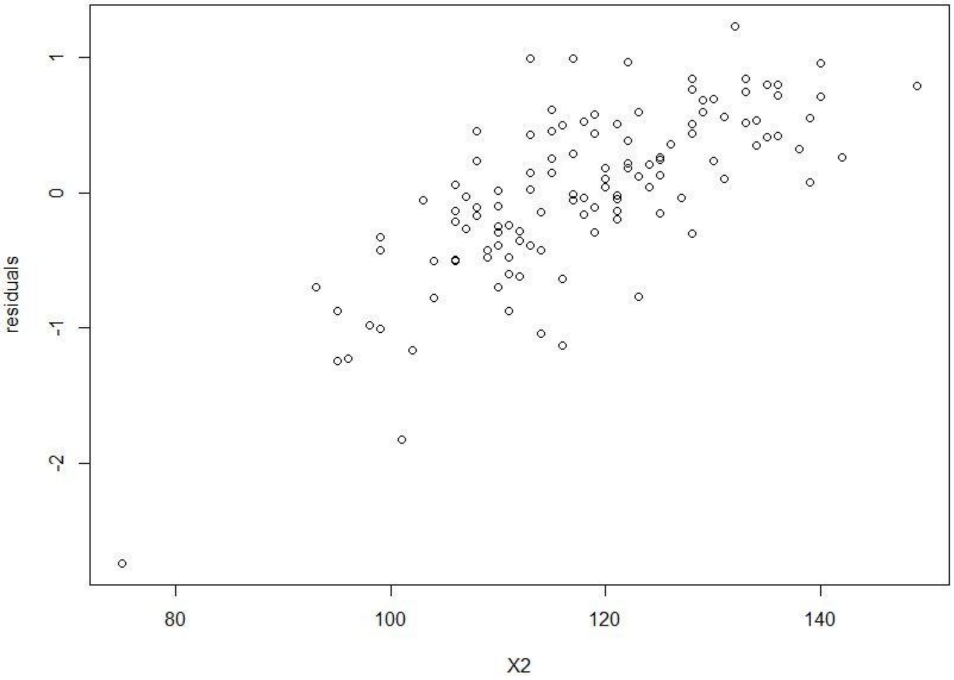
If $|t^*_{BF}| \leq 2.618137$, conclude the error variance is constant

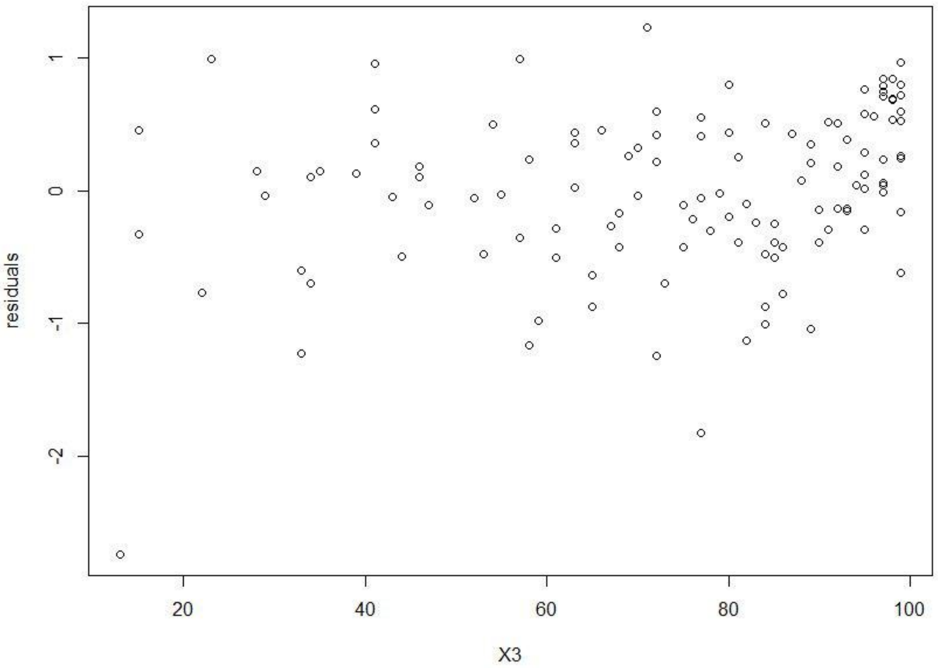
If $|t^*_{BF}| > 2.618137$, conclude the error variance is not constant

Now we know $|t^*_{BF}| = 0.8967448 < 2.618137$, then conclude the error variance is constant, which is consistent with the conclusion in (c).

(e)

```
> plot(d3$V3,model19$residuals,xlab="X2",ylab="residuals")
> plot(d3$V4,model19$residuals,xlab="X3",ylab="residuals")
```

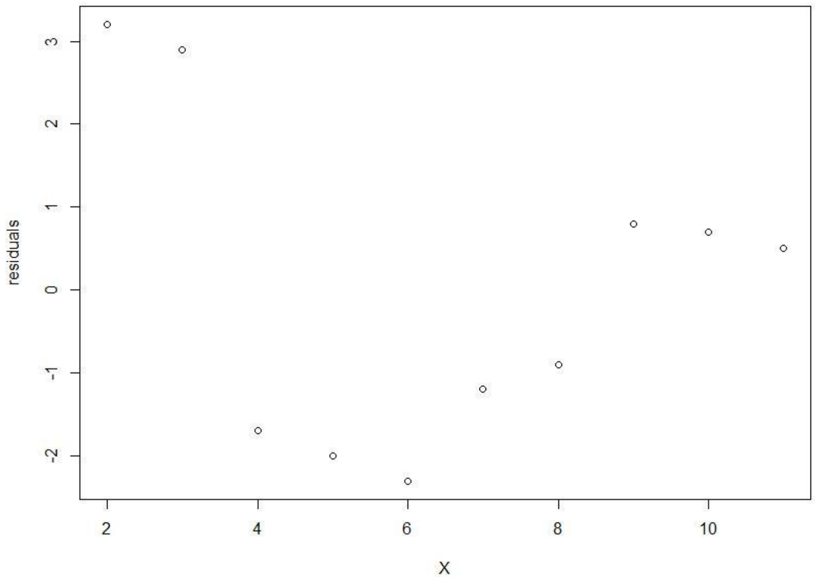




From the two graphs above, we find that X2 seems to have some linear relation with the residuals, so maybe include X2 in our model would improve it. But X3 shows little relation with the residuals.

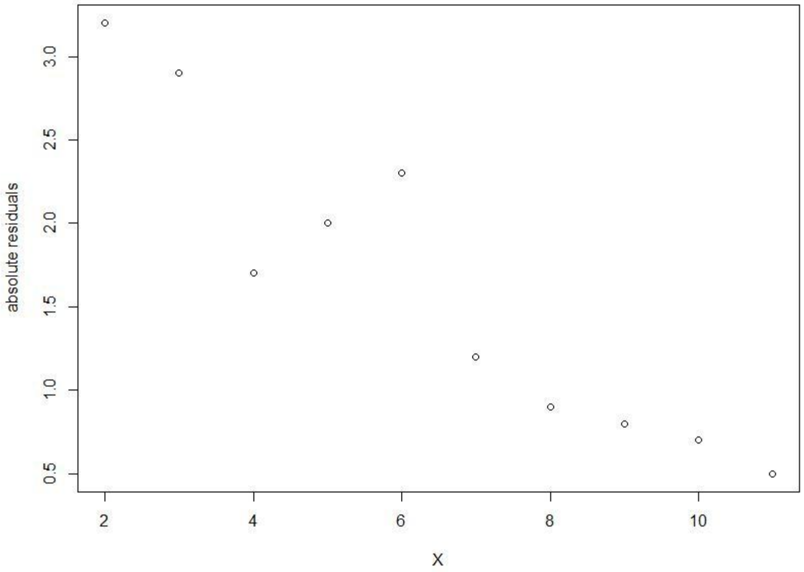
Problem 2 (3.9)

```
> plot(d9$V1,d9$V2,xlab="X",ylab="residuals")
```



It appears that there is no correlation between error terms that are near each other in the sequence. However, after plot the absolute residuals against *X*:

```
> plot(d9$V1,abs(d9$V2),xlab="X",ylab="absolute residuals")
```

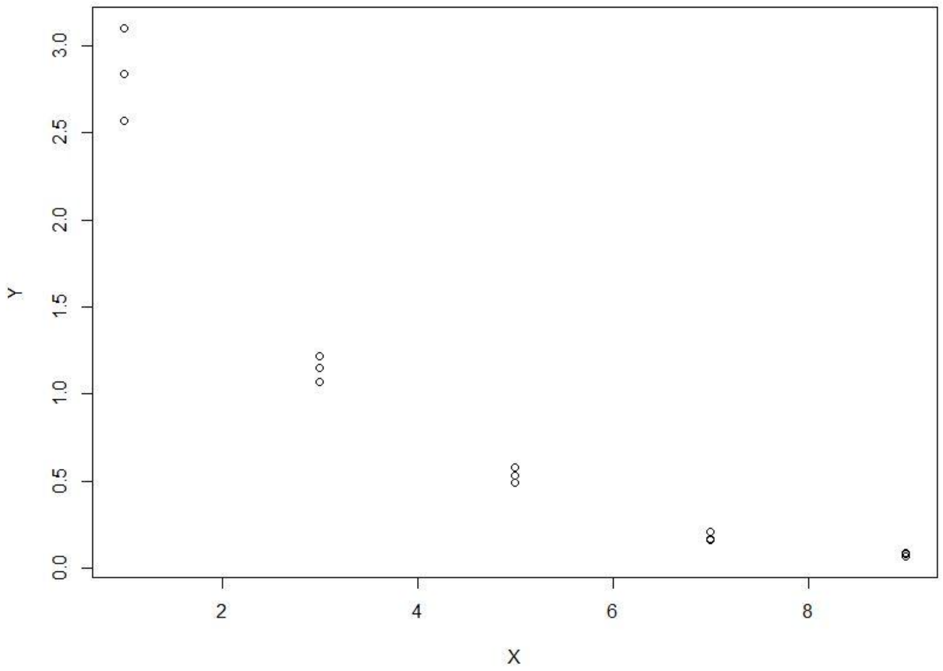


We can see that when X becomes larger, the absolute residuals tend to be smaller. So we the transformation of absolute seems to alleviate this problem.

Problem 3 (3.16)

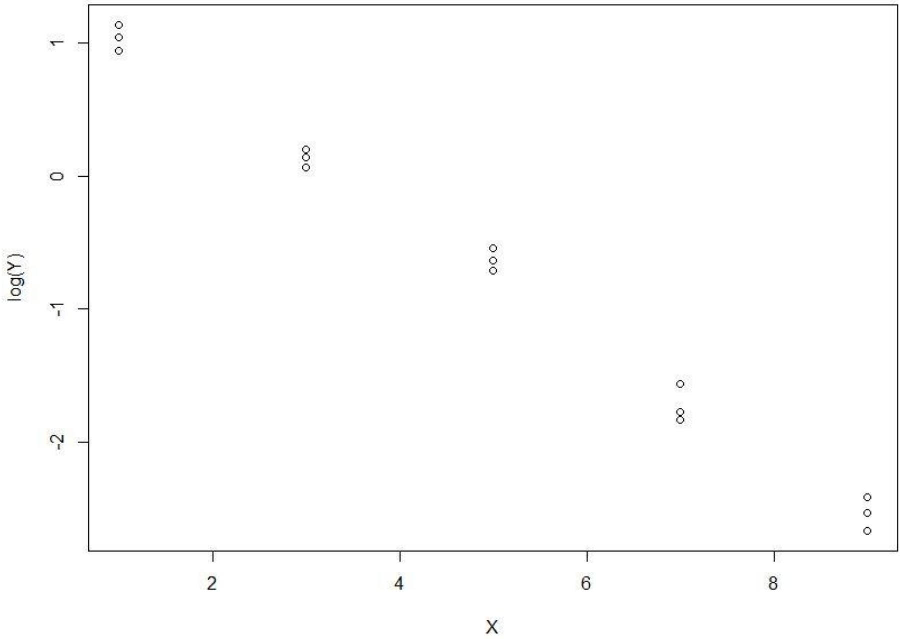
(a)

```
> plot(d15$V2,d15$V1,xlab="X",ylab="Y")
```



We see from the graph that X and Y shows approximate linear relation but the deviation might not be constant. So we apply log(Y) transformation and can get the following graph:

```
> plot(d15$V2,log(d15$V1),xlab="X",ylab="log(Y)")
```



After the log transformation we achieve constant variance and linearity.

(b)

```
> SSE=c()
> lambda=c(-0.2,-0.1,0.1,0.2)
> for (i in 1:4){
+   model.boxcox <- lm(d15$V1^lambda[i]~d15$V2)
+   a=anova(model.boxcox)
+   SSE[i]=a$‘Sum Sq‘[2]
+ }
> model.boxcox <- lm(log(d15$V1)~d15$V2)
> a=anova(model.boxcox)
> SSE[5]=a$‘Sum Sq‘[2]
> SSE
[1] 0.01065715 0.00332598 0.00167064 0.01065715 0.17179392
```

We can see from the 5 SEEs above, when we apply power of 0.1 we can get the smallest SSE. So we will choose $\lambda = 0.1$.

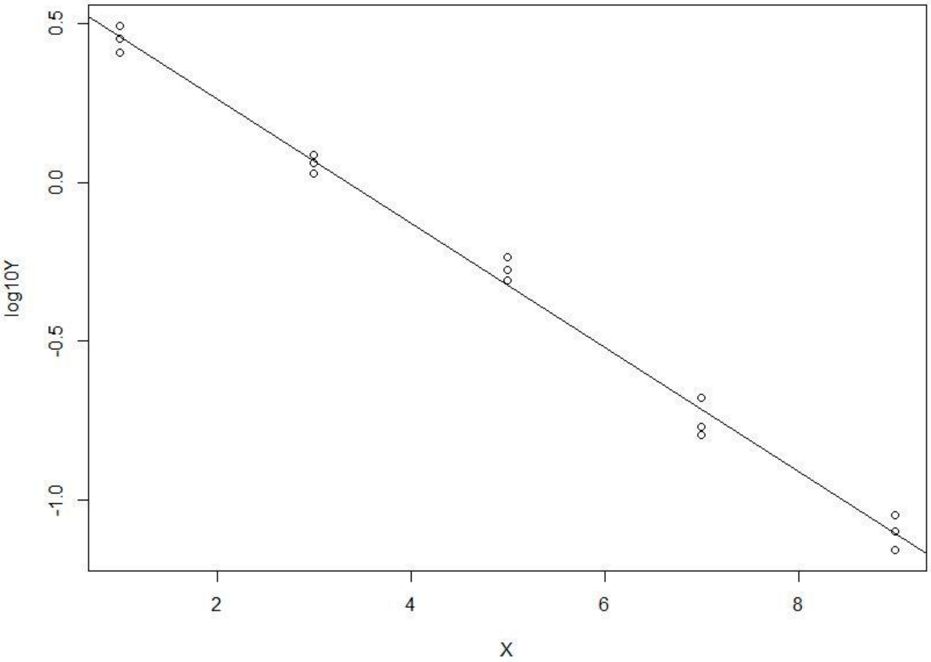
(c)

```
> model.boxcox <- lm(log(d15$V1)/log(10)~d15$V2)
> model.boxcox$coefficients
(Intercept)      d15$V2
  0.6548798  -0.1954003
```

So $\log_{10}\hat{Y} = 0.6548798 - 0.1954003X$

(d)

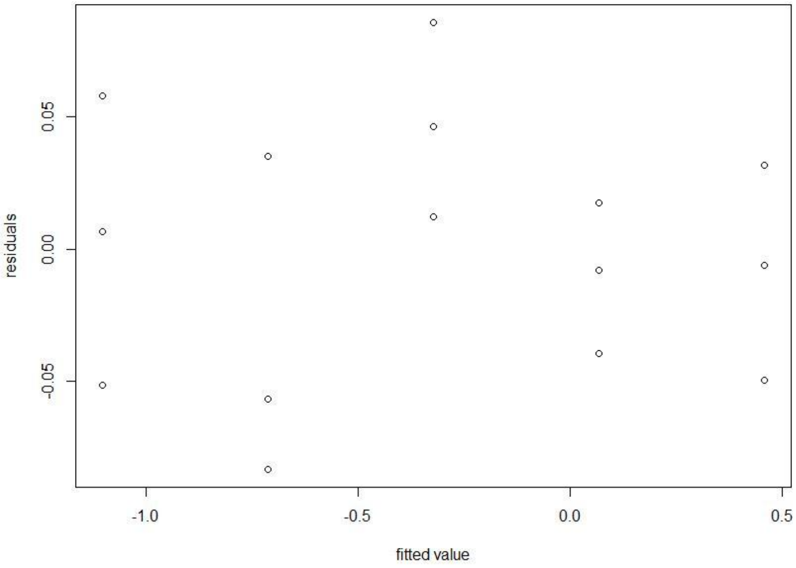
```
> plot(d15$V2,log(d15$V1)/log(10),xlab="X",ylab="log10Y")
> abline(model.boxcox$coefficients[1:2])
```



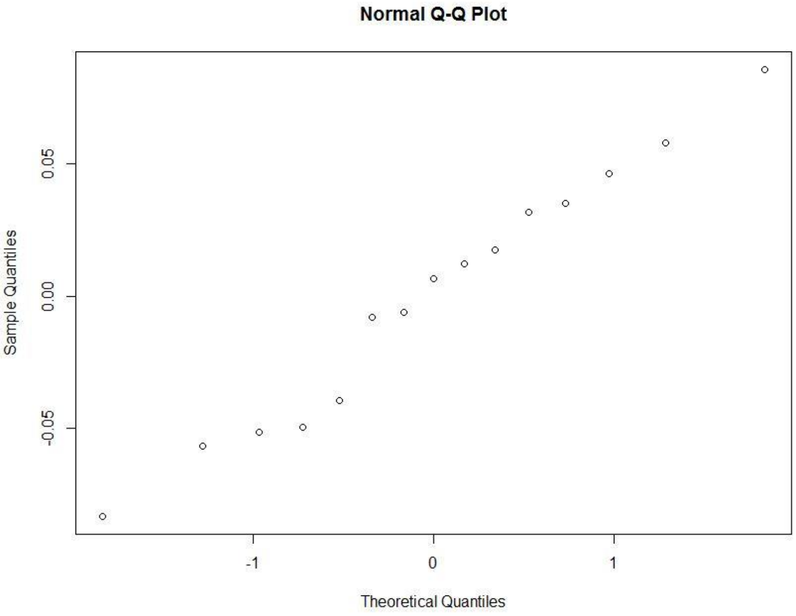
The regression line seems to be fitted to the transformed data, and this line is a good fit to the data.

(e)

```
> plot(model.boxcox$fitted.values,model.boxcox$residuals,
+       xlab="fitted value", ylab="residuals")
```



```
> qqnorm(model.boxcox$residuals)
```



Although the residual against fitted value plot shows that the error variance appears to be more stable and the points in the normal probability plot fall roughly on a straight line, the residual plot now suggests that Y is nonlinearly related to X.

(f)

$\log_{10}\hat{Y} = 0.6548798 - 0.1954003X$
So

$$\begin{aligned}\hat{Y} &= 10^{0.6548798-0.1954003X}\\ &= 4.517309 \times 10^{-0.1954003X}\end{aligned}$$

Problem 4 (3.23)

Our full model is:

$$Y_{ij} = \mu_j + \epsilon_{ij}, \quad i = 1, ..., n_j; \quad j = 1, ..., 10$$

where c is the number of different levels of X, n_j is the number of observations at level X_j, and μ_j is the expected value of Y_{ij}.
There are $n - c = 10$ degrees of freedom in the full model.

Our reduced model is a simple linear regression:

$$Y_{ij} = \beta_1 X_i + \epsilon_{ij}, \quad i = 1, ..., n_j; \quad j = 1, ..., 10$$

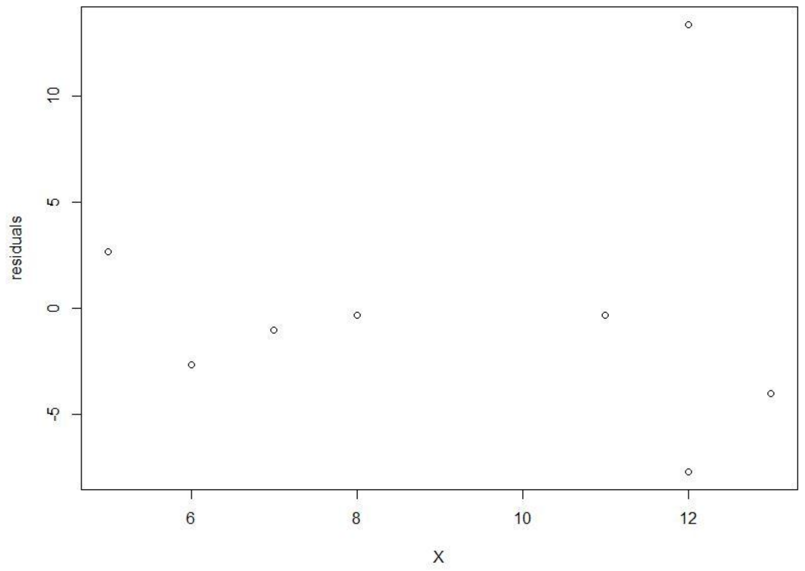
There are $n - 1 = 9$ degrees of freedom in the reduced model.

Problem 5 (3.24)

(a)

```
> colnames(d24)<-c("Y","X")
> model24<-lm(Y~X,data=d24)
> model24$coefficients
(Intercept)      d24$V2
  48.666667    2.333333
> plot(d24$V2,model24$residuals,xlab="X",ylab="residuals")
```

The fitted regression function is $\hat{Y} = 48.666667 + 2.333333X$



This plot tells us that most residuals are below 0, and except for residual when *i*=7 (x=12), all other points have the same tendency, likely a linearity relation between X and residuals.

(b)

```
> d24=d24[-7,]
> model24<-lm(Y~X,data=d24)
> model24$coefficients
(Intercept)      d24$V2
  53.067961    1.621359
```

The fitted regression function is $\hat{Y} = 53.067961 + 1.621359X$

Compared with the regression function in part (a), we get a smaller estimator of β_1, which means case 7 has a strong effect on our model which influences the slope. So, case 7 must be an outlier.

(c)

```
> newdata24=data.frame(X=12)
> Ynew=predict(model24, newdata24,level=0.99)
> MSE=sum((d24$Y-model24$fitted.values)^2)/model24$df.residual
> s2=MSE*(1+1/7+(12-mean(d24$X))^2/sum((d24$X-mean(d24$X))^2))
> Ynew+sqrt(s2)*qt(0.995,5)
       1
84.73588
> Ynew-sqrt(s2)*qt(0.995,5)
       1
60.31266
```

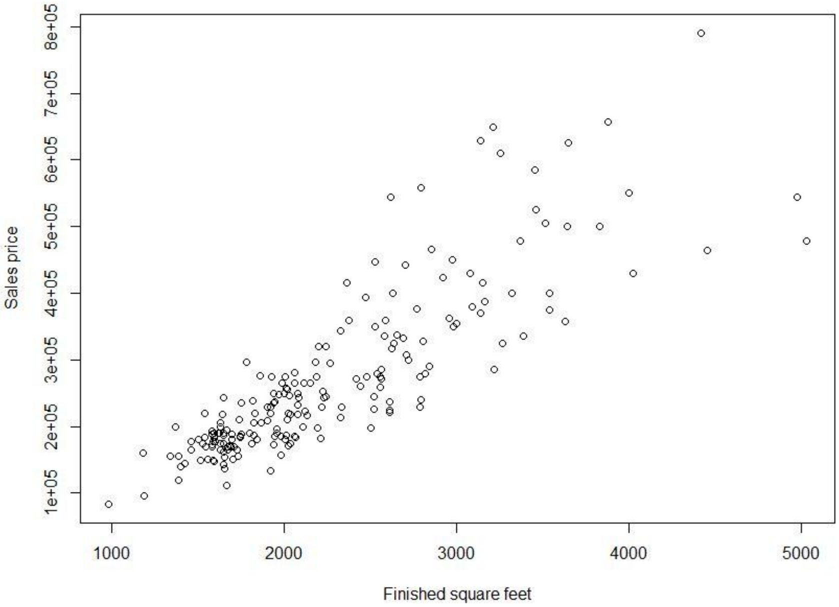
The .99 percent prediction interval for X=12 is $[60.31266,\ 84.73588]$.

The observation $Y_7 = 90$ fall outside this prediction interval.

Since the prediction interval is of 0.99 percent confidence, the significance is 1-0.99=0.01, which is 1%.

Problem 6 (3.31)

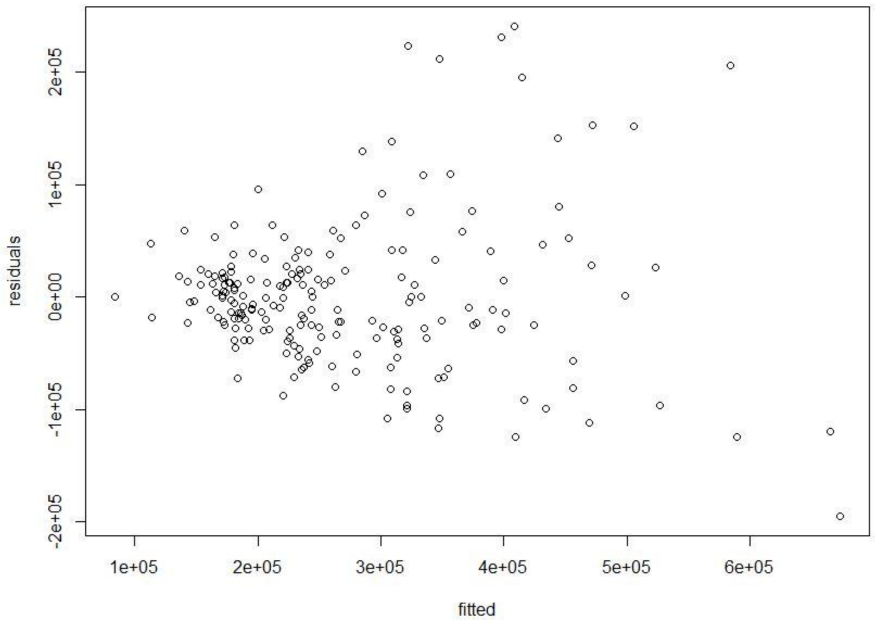
We first plot the scatter plot for original data:

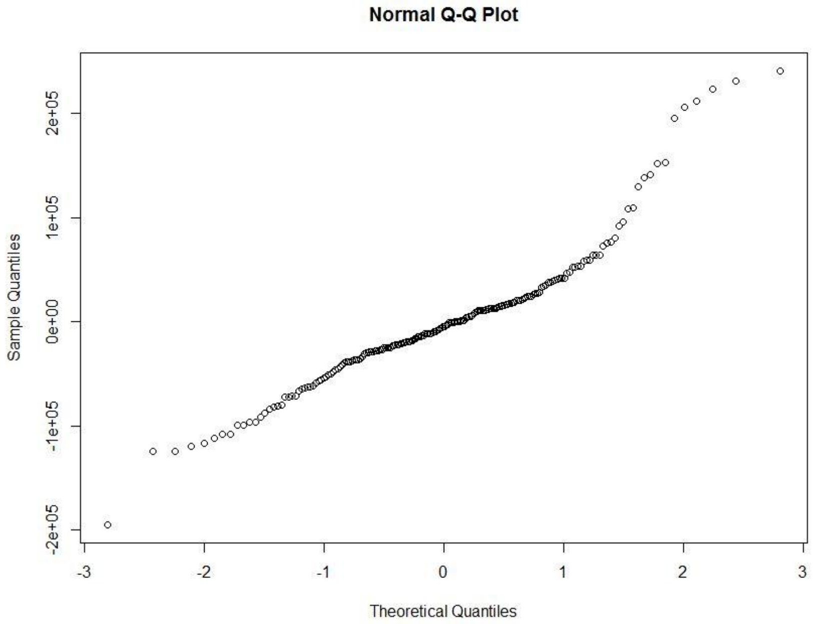


```
> summary(model31)
Call:
lm(formula = Y ~ X, data = d31.fit)
Residuals:
    Min      1Q  Median      3Q     Max 
-194676  -31606   -4721   22358  240584 
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -58917.437  15228.399  -3.869 0.000148 ***
X              145.587      6.418  22.683  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 65330 on 198 degrees of freedom
Multiple R-squared:  0.7221,Adjusted R-squared:  0.7207 
F-statistic: 514.5 on 1 and 198 DF,  p-value: < 2.2e-16
```

We see from the regression output that the slope of the regression line is not zero (F* = 514.5, P-value < 2.2e-16) so that a regression relationship exists.

```
> plot(model31$fitted.values,model31$residuals
+      ,xlab="fitted", ylab="residuals")
> qqnorm(model31$residuals)
```



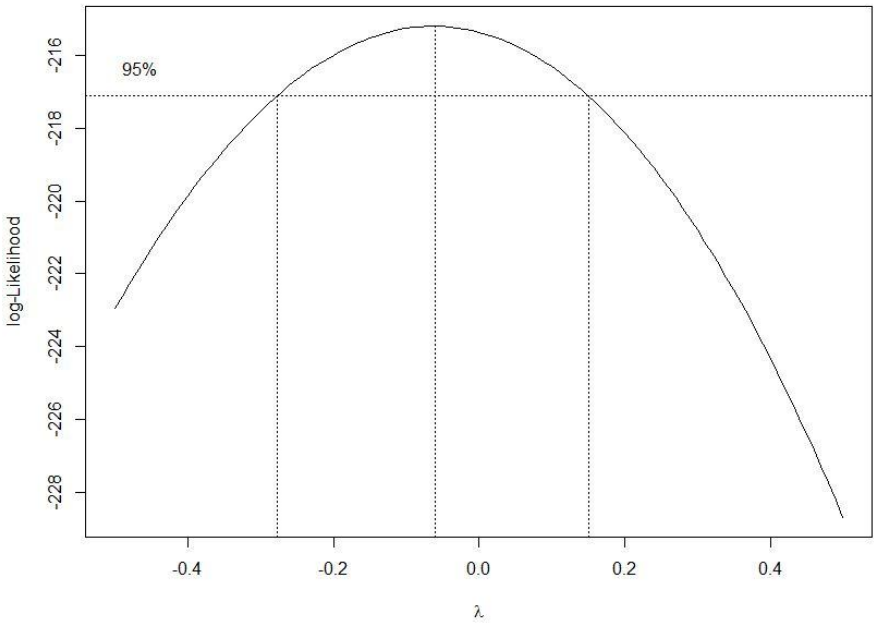


But we also see from the residual plot that the error variance appears to be increasing with the level of finished squared feet.

The normal probability plot suggests nonnormality (heavy tails), but the nonlinearity of the plot is likely to be related to the unequal error variances.

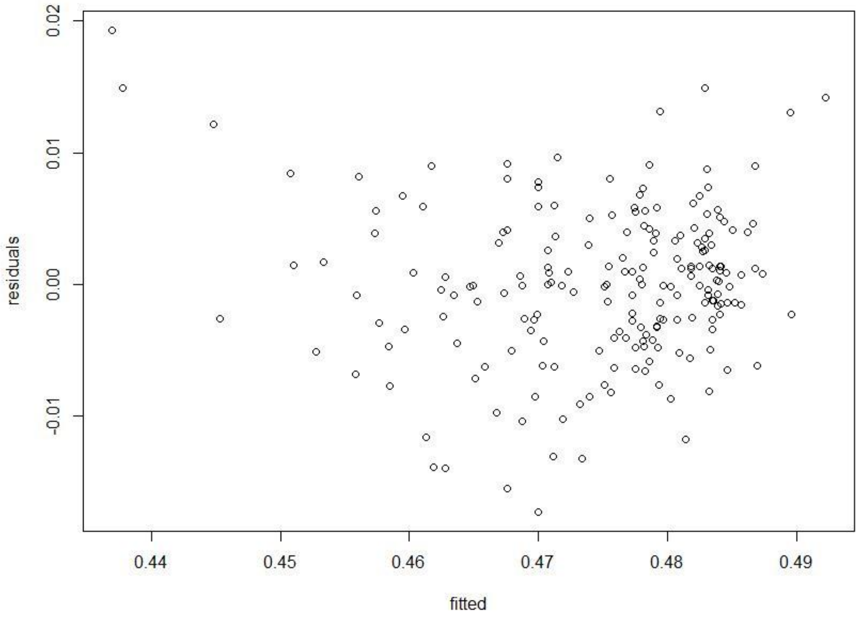
The presence of nonconstant variance clearly requires remediation. We shall use the Box-Cox procedure to suggest an appropriate power transformation.

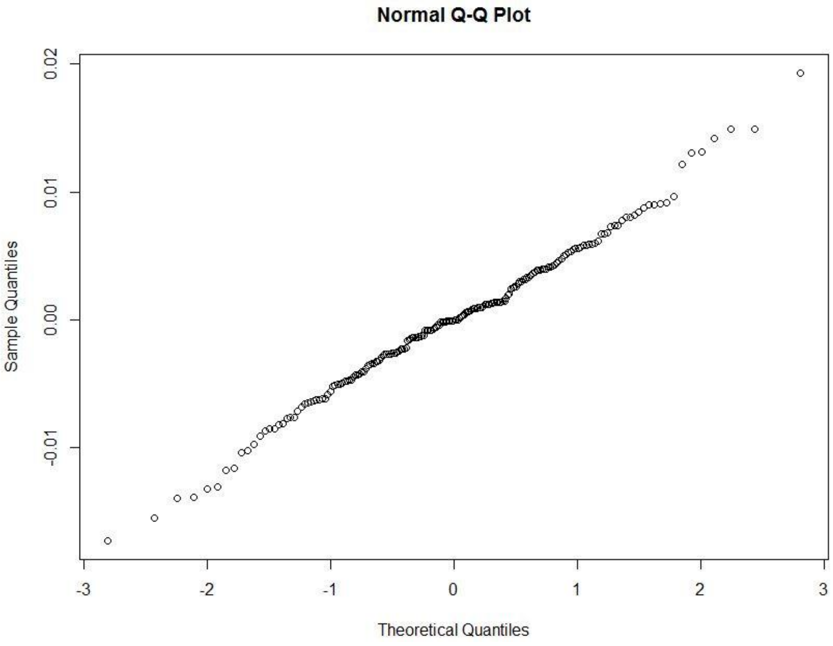
```
> boxcox(model31,lambda = seq(-0.5, 0.5, by=0.01))
```



We find the maximum likelihood estimate of λ to be $\lambda = -0.06$.

```
> lambda <- -0.06
> d31.fit <- cbind(d31.fit,Ylam<-d31.fit$Y^lambda)
> model31.lam <- lm(Ylam~X, data=d31.fit)
>
> plot(model31.lam$fitted.values,model31.lam$residuals
+      ,xlab="fitted", ylab="residuals")
> qqnorm(model31.lam$residuals)
```





The error variance appears to be more stable and the points in the normal probability plot fall likely on a straight line.

Now let's compute the lack of fit test:

```
> anova(lm(Ylam~X,data=d31.fit), lm(Ylam~factor(X),data=d31.fit))
Analysis of Variance Table
Model 1: Ylam ~ X
Model 2: Ylam ~ factor(X)
  Res.Df       RSS  Df Sum of Sq      F Pr(>F)
1    198 0.0069756
2     24 0.0009108 174 0.0060649 0.9185 0.6391
```

The lack of fit statistic is $F^* = 0.9185$ with p-value$= 0.6391$, supporting the linearity of the regression model.

```
> summary(model31.lam)
Call:
lm(formula = Ylam ~ X, data = d31.fit)
Residuals:
      Min        1Q    Median        3Q       Max
-0.017242 -0.003415 -0.000050  0.003865  0.019276
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.056e-01  1.384e-03  365.46   <2e-16 ***
X           -1.365e-05  5.832e-07  -23.41   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.005936 on 198 degrees of freedom
Multiple R-squared:  0.7346,Adjusted R-squared:  0.7333
F-statistic: 548.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

We now get the fitted regression function

$$\hat{Y}' = 0.5056 - 1.365 \times 10^{-5} X$$

where $\hat{Y}' = \hat{Y}^{-0.06}$.

Now let's make the prediction:

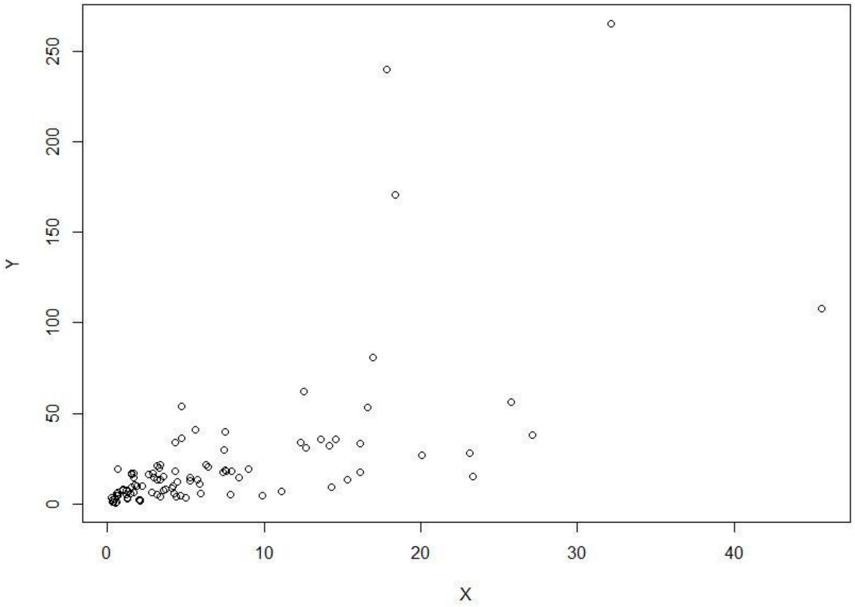
when X=1100, $\hat{Y}' = 0.5056 - 1.365 \times 10^{-5} \times 1100 = 0.490585$, so $\hat{Y} = 142811.2$

when X=4900, $\hat{Y}' = 0.5056 - 1.365 \times 10^{-5} \times 4900 = 0.438715$, so $\hat{Y} = 919655.7$

Comparing with the original model, this final model shows more validation for the constant error variance, and tend to follow normal distribution. The lack of fit static $F^* = 0.9185$ p-value$= 0.6391$ supports our linearity conclusion. But after transformation, this model still has the problem that Y and X might be nonlinear. And the transformation λ =-0.06 makes the data change scale a lot which may lead to large error even with a small change.

Problem 7 (3.32)

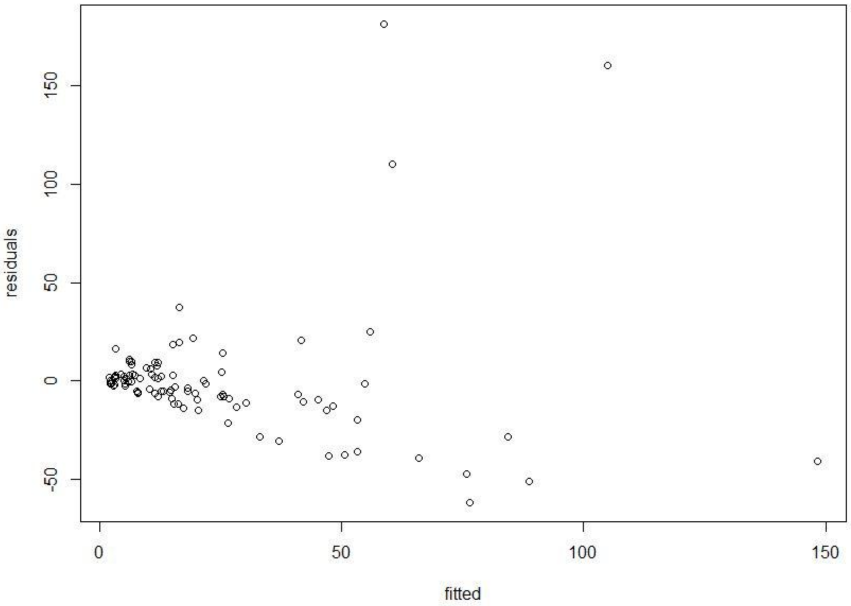
We first plot the scatter plot for original data:

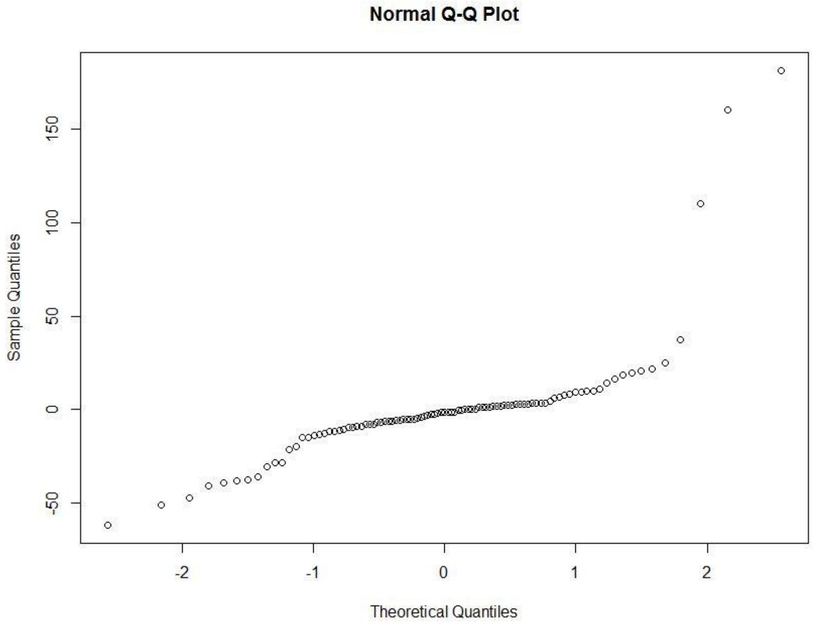


```
> summary(model32)
Call:
lm(formula = Y ~ X, data = d32.lm)
Residuals:
    Min      1Q  Median      3Q     Max
-61.619  -9.023  -1.586   3.151 181.183
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.1249     4.3596   0.258    0.797
X             3.2299     0.4148   7.786 8.47e-12 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 32.03 on 95 degrees of freedom
Multiple R-squared:  0.3896,Adjusted R-squared:  0.3831
F-statistic: 60.63 on 1 and 95 DF,  p-value: 8.468e-12
```

We see from the regression output that the slope of the regression line is not zero (F* = 60.63, P-value = 8.568e-12) so that a regression relationship exists.

```
> plot(model32$fitted.values,model32$residuals
+      ,xlab="fitted", ylab="residuals")
> qqnorm(model32$residuals)
```



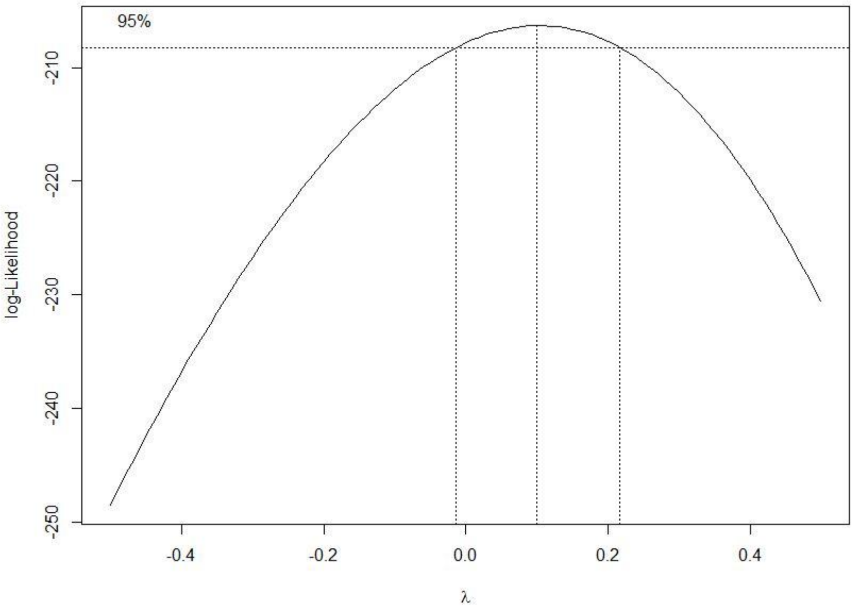


But we also see from the residual plot that the error variance appears to be increasing with the level of cancer volume, which is likely to be related to the unequal error variances.

Except for 3 points in the normal probability plot, other points fall roughly on a straight line. We may suggest these three points might be outliers.

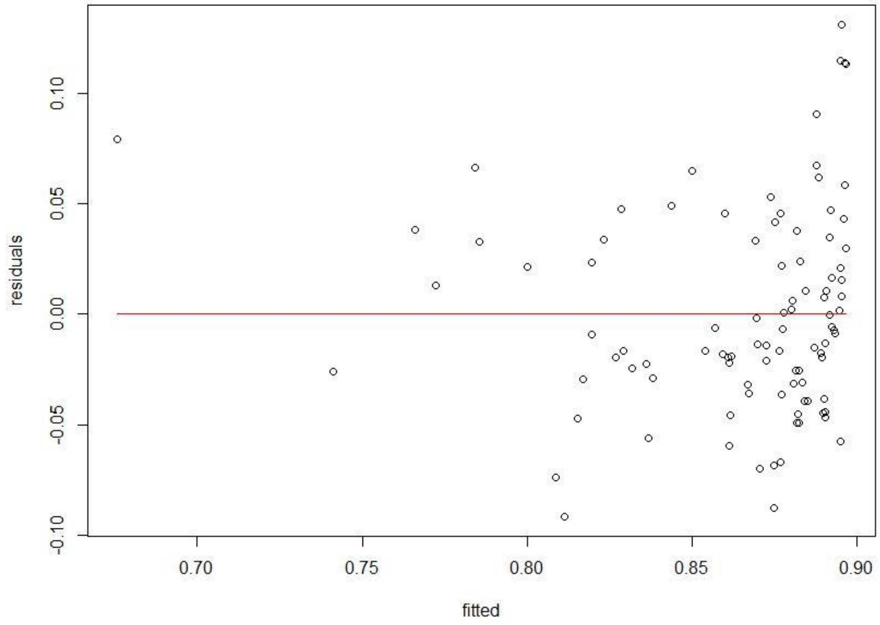
The presence of nonconstant variance clearly requires remediation. We shall use the Box-Cox procedure to suggest an appropriate power transformation.

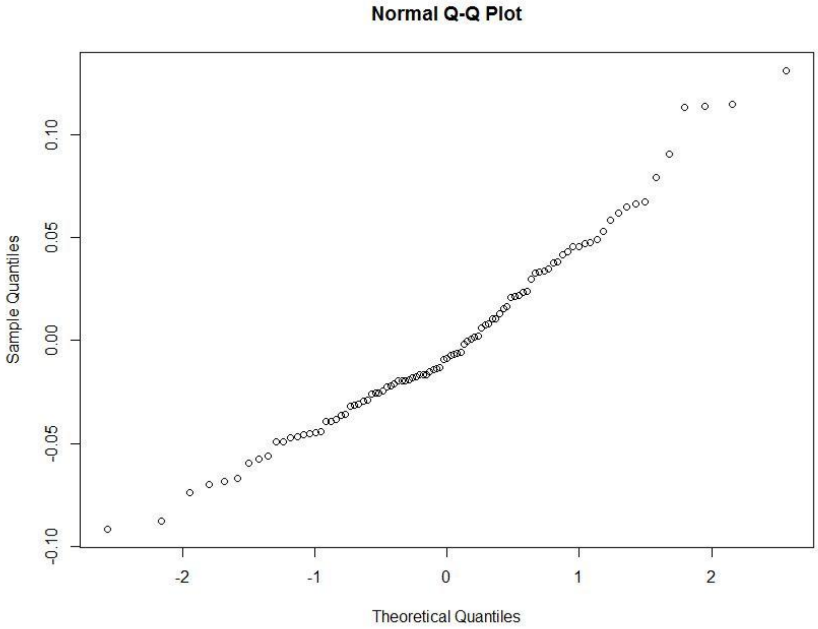
```
> boxcox(model32,lambda = seq(-0.5, 0.5, by=0.01))
```



We find the maximum likelihood estimate of λ to be $\lambda = 0.1$.

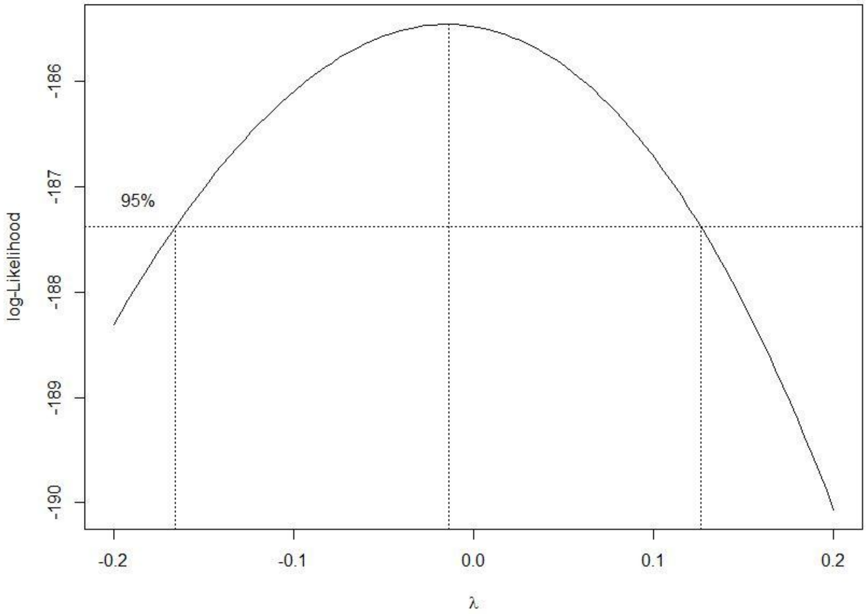
```
> lambda <- 0.1
> d32.lm <- cbind(d32.lm,Ylam<-d32.lm$Y^lambda)
> model32.lam <- lm(Ylam~X, data=d32.lm)
> plot(model32.lam$fitted.values,model32.lam$residuals
+      ,xlab="fitted", ylab="residuals")
> qqnorm(model32.lam$residuals)
> lines(model32.lam$fitted.values,
+       rep(0,length(model32.lam$fitted.values)),col=’red’)
```





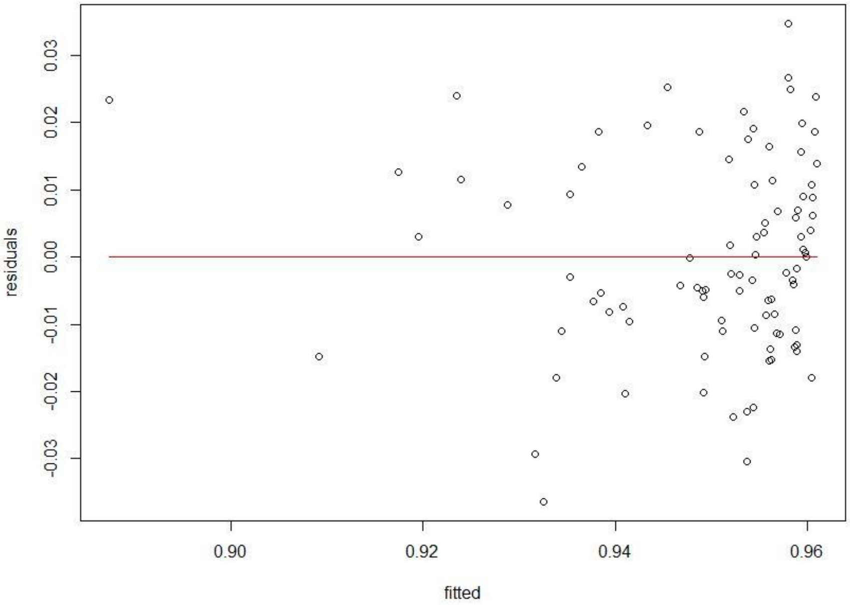
In the residual plot, all points fall within the range [-0.1, 0.1] except for 4 points, which also cause a nonlinearity in the normal probability plot. So we intuitively consider these 4 points as outliers. Now we rebuild the model exclude these points.

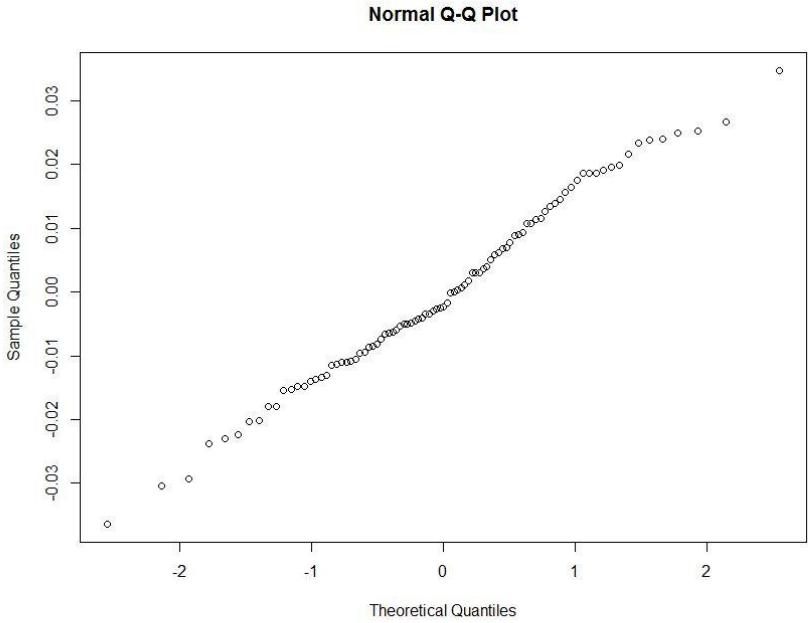
```
> boxcox(model32,lambda = seq(-0.2, 0.2, by=0.01))
```



We get the new $\lambda = -0.02$

```
> lambda <- -0.02
> d32.fit <- cbind(d32.fit,Ylam<-d32.fit$Y^lambda)
> model32.fitlam <- lm(Ylam~X, data=d32.fit)
> plot(model32.fitlam$fitted.values,model32.fitlam$residuals
+       ,xlab="fitted", ylab="residuals")
> qqnorm(model32.fitlam$residuals)
> lines(model32.fitlam$fitted.values,
+        rep(0,length(model32.fitlam$fitted.values)),col='red')
```





The error variance appears to be more stable and the points in the normal probability plot fall likely on a straight line. And the points in residual plot seems be more equally placed.

Now let's compute the lack of fit test:

```
> anova(lm(Ylam~X,data=d32.fit), lm(Ylam~factor(X),data=d32.fit))
Analysis of Variance Table
Model 1: Ylam ~ X
Model 2: Ylam ~ factor(X)
  Res.Df       RSS Df Sum of Sq      F  Pr(>F)
1     91 0.0194604
2     10 0.0010197 81  0.018441 2.2326 0.08209 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The lack of fit statistic is $F^* = 2.2326$ with p-value$= 0.08209$, supporting the linearity of the regression model.

```
> summary(model32.fitlam)
Call:
lm(formula = Ylam ~ X, data = d32.fit)
Residuals:
      Min        1Q    Median        3Q       Max 
-0.036312 -0.010613 -0.002409  0.010697  0.034615 
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.9614485  0.0020637  465.88  < 2e-16 ***
X           -0.0016247  0.0001923   -8.45 4.46e-13 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.01462 on 91 degrees of freedom
Multiple R-squared:  0.4396,Adjusted R-squared:  0.4335 
F-statistic:  71.4 on 1 and 91 DF,  p-value: 4.458e-13
```

We now get the fitted regression function

$$\hat{Y}' = 0.9614485 - 0.0016247 \times X$$

where $\hat{Y}' = \hat{Y}^{-0.02}$.

Now let's make the prediction:

if X=20, $E(\hat{Y}') = 0.9614485 - 0.0016247 \times 20 = 0.9289545$,so $E(\hat{Y}) = 39.8362$

Comparing with the original model, this final model shows more validation for the constant error variance. After we successfully exclude 4 outliers, the model seems be more fitted and the error variance be more likely to be normal distribution. The lack of fit static $F^* = 2.2326$ with p-value$= 0.08209$ supports our linearity conclusion. But after transformation, this model still has the problem that *Y* and *X* might be nonlinear. And the transformation λ=-0.02 makes the data change scale a lot which may lead to large error even with a small change.