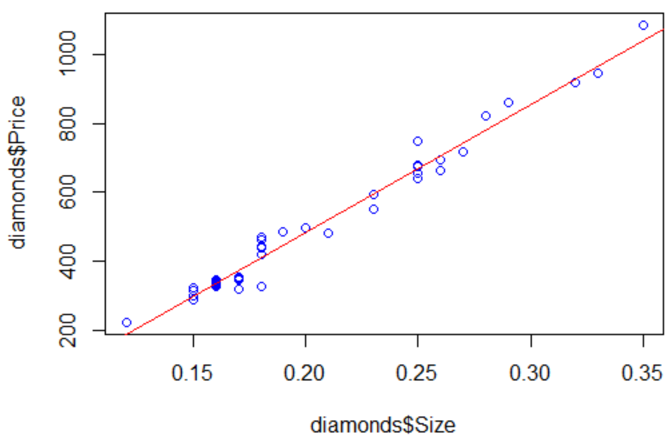
Statistics 101A
Homework Two

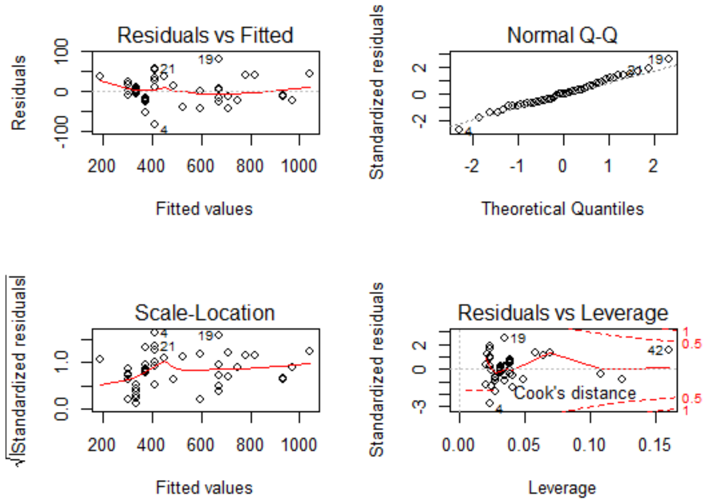
Question One: Problem one from chapter three 3.4 Exercises (The data file airfares.txt)

a) Obviously the model is not a valid one, even though the R^2 is very high. The standardized residuals are having a pattern which means they are not independent nor having a constant variance.
b) The line seems to be fitting the pattern well but it does because of the large scale of the y variable. This is not a valid model; we can use transformations on the x or the y variables to try to fix the violations.

Question Two: Problem eight from chapter three 3.4 exercises (The Diamond stones data file)
Part 1:
Part 1 a)



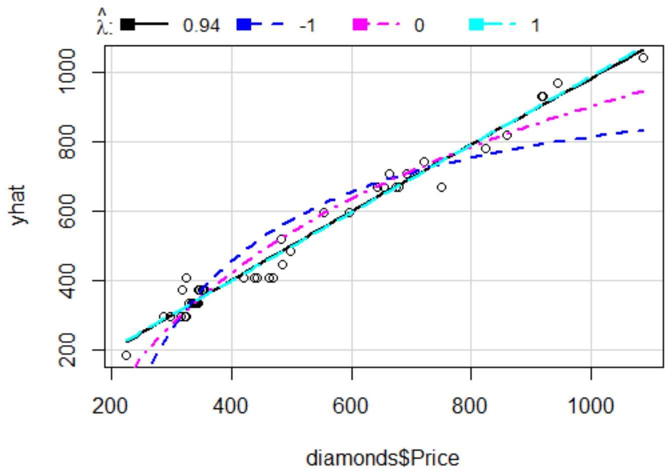

```
> Dmod1<-lm(diamonds$Price~diamonds$Size)
> summary(Dmod1)

Call:
lm(formula = diamonds$Price ~ diamonds$Size)

Residuals:
    Min      1Q  Median      3Q     Max 
-85.654 -21.503  -1.203  16.797  79.295 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -258.05      16.94  -15.23   <2e-16 ***
diamonds$Size  3715.02      80.41   46.20   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 31.6 on 47 degrees of freedom
Multiple R-squared:  0.9785,   Adjusted R-squared:  0.978 
F-statistic:  2135 on 1 and 47 DF,  p-value: < 2.2e-16
```



Part 1 B: Slight pattern in the residual plot, and seem to violate the non-constant variance assumption. R^2 is very high and significant, both the slope and the y-intercept estimates are significant

Part 2:

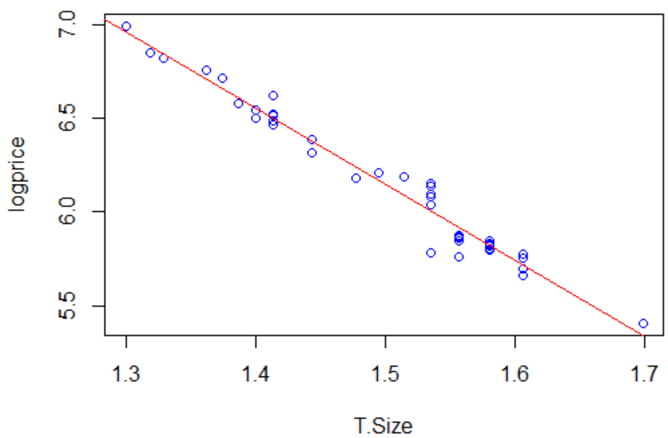
```
> summary(powerTransform(cbind(Size,Price)~1,data=diamonds))
bcPower Transformations to Multinormality
      Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Size    -0.2393           0      -1.0400       0.5615
Price   -0.0172           0      -0.6114       0.5771

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                             LRT df    pval
LR test, lambda = (0 0) 1.432924  2 0.48848

Likelihood ratio test that no transformations are needed
                             LRT df      pval
> inverseResponsePlot(Dmod1)
      lambda       RSS
1  0.9376257  45670.12
2 -1.0000000 272143.61
3  0.0000000 101071.53
4  1.0000000  45918.17
```

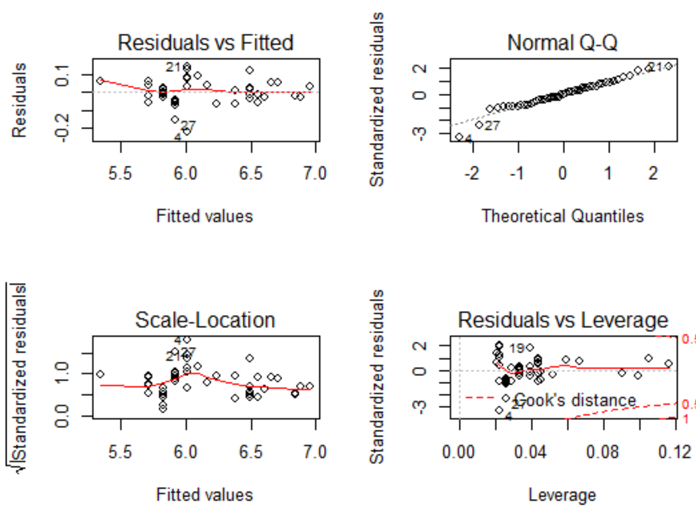


Inverse response plot is suggesting power of 1 as the best transformation of the response variable.

Power transformation suggests $log\ of\ the\ response\ and\ \frac{1}{Size^{0.25}}$ for the predictor.

```
> logprice<- log(diamonds$Price)
> T.Size<- diamonds$Size^(-0.25)
> Dmod2<-lm(logprice~T.Size)
> summary(Dmod2)

Call:
lm(formula = logprice ~ T.Size)

Residuals:
      Min         1Q    Median        3Q       Max 
-0.223411 -0.045628  0.001625  0.038482  0.141232 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  12.2252     0.1546   79.09   <2e-16 ***
T.Size       -4.0501     0.1025  -39.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06816 on 47 degrees of freedom
Multiple R-squared:  0.9708,   Adjusted R-squared:  0.9702 
F-statistic:  1563 on 1 and 47 DF,  p-value: < 2.2e-16
```





The diagnostics plots look better and less violation. More good leverage points. Almost a constant variance.

Question Three:

a) Using the stress echo UCLA data (see week four), fit a linear model to predict basal blood pressure from systolic blood pressure. Report the equation for the model. Report a residual plot and comment what it tells us about the assumption of linearity.

b) Report the ANOVA table. Show how you can find the F value reported in the ANOVA table using R^2. What is the null hypothesis that you are testing through ANOVA? Compare the F value that you calculate with value that you find from the F table and decide whether you are going to reject or fail to reject the null hypothesis). Check if this equation is true: $(Se)^2$ is approximately equal to $var(Y) * (1 - r^2)$.

c) Calculate R^2 adjusted and compare it to R^2. Comment on the difference.

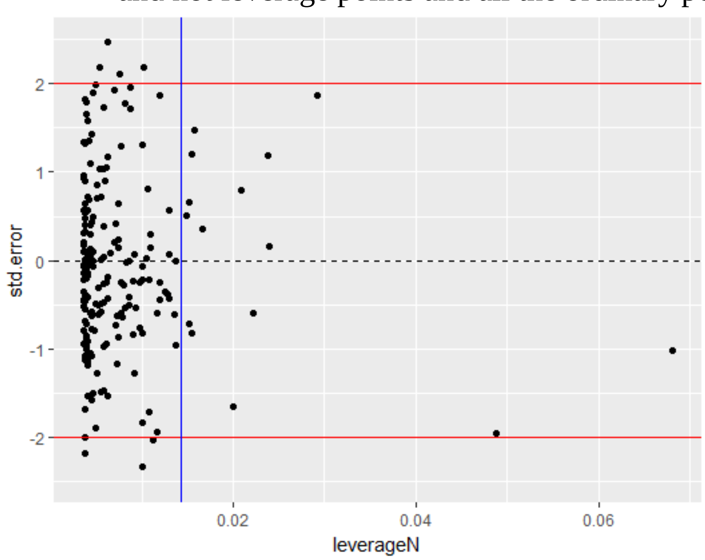
d) Check the diagnostic plots and comment on each one of them.

e) Create two new variables: one for the leverage of a point and one for the standardized residuals. Create a table from both variables to identify the following:

Leverage/Outliers	Yes	No
Yes	1	18
No	12	248

```
> table(LV)
LV
 No Yes
260  19
> OL<-ifelse(abs(std.error)>=2,"Yes","No")
> table(OL)
OL
 No Yes
266  13
> table(LV,OL)
     OL
LV     No Yes
  No  248  12
  Yes  18   1
```

f) Use ggplot2 library to create a plot of Leverage Vs Standardizes residuals divided into regions to help you identify bad and good leverage points, outliers and not leverage points and all the ordinary points.



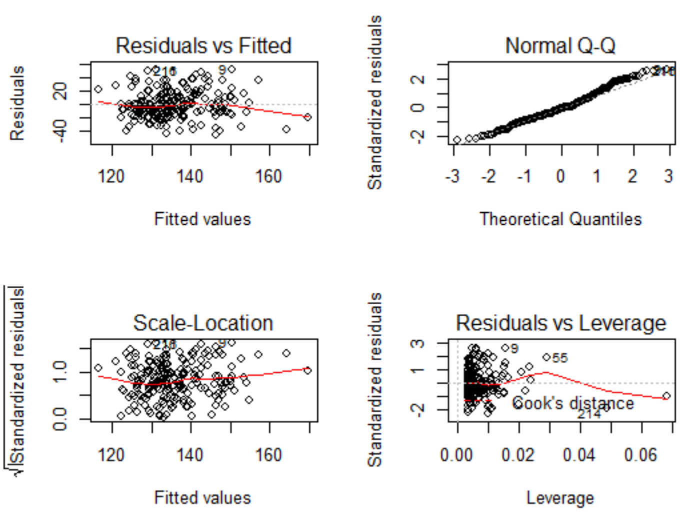
Question 5:
Use the Echo data from question three to transform the data and compare the results to the SLR created in question three:

a) Use the inverse response plot to find the best λ to transform the y variable to minimize the SSE. Construct a SLR of the transformed y variable and systolic blood pressure. Check diagnostics. Is this one better than the SLR in question three.

b) Use the power transform function to find the best λ(s) to transform both the y variable and the x variable to make the densities of these two variables as close as possible to normal. Construct a SLR of the transformed variables. Check diagnostics. Is this one better than the SLR in question three.

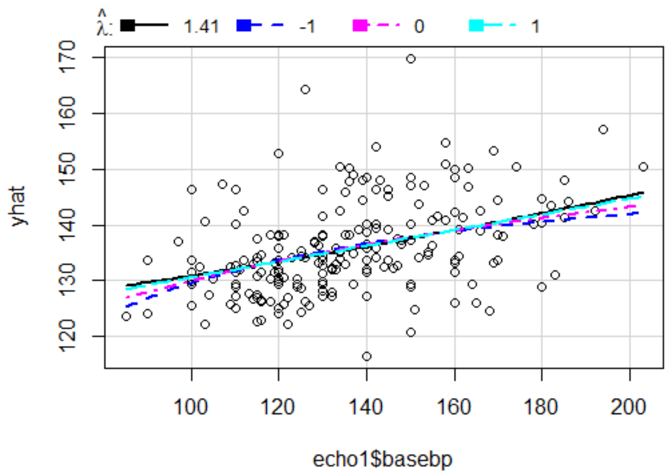
Q3 and Q5 Key:

```
> Em1<-lm(echo1$basebp~echo1$sbp)
> summary(Em1)

Call:
lm(formula = echo1$basebp ~ echo1$sbp)

Residuals:
    Min      1Q  Median      3Q     Max 
-46.449 -12.456  -1.273  11.444  52.490 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 103.70036    4.88943   21.21  < 2e-16 ***
echo1$sbp     0.21374    0.03176    6.73 9.71e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19.97 on 277 degrees of freedom
Multiple R-squared:  0.1405,	Adjusted R-squared:  0.1374 
F-statistic:  45.3 on 1 and 277 DF,  p-value: 9.705e-11

> par(mfrow=c(1,1))
> plot(echo1$sbp,echo1$basebp)
> abline(Em1)
> par(mfrow=c(2,2))
> plot(Em1)
```

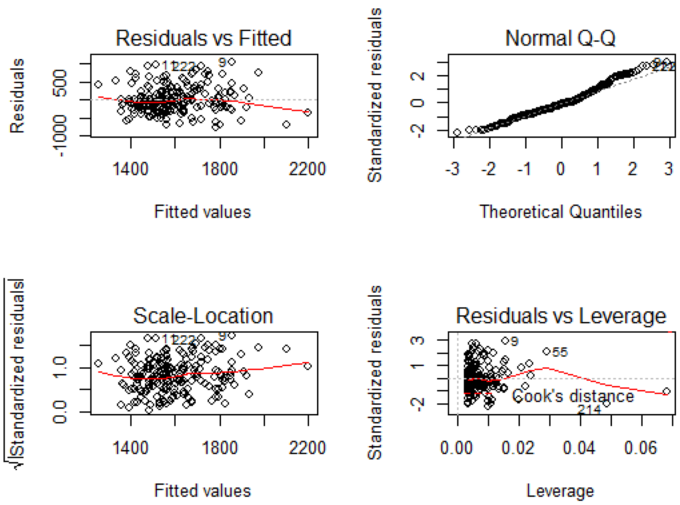



```
> anova(Em1)
Analysis of Variance Table

Response: echo1$basebp
           Df Sum Sq Mean Sq F value    Pr(>F)
```

```
echo1$sbp    1  18065 18064.7   45.298 9.705e-11 ***
Residuals 277 110466   398.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> library(car)
> par(mfrow=c(1,1))
> inverseResponsePlot(Em1)
     lambda      RSS
1  1.413234 15521.43
2 -1.000000 15677.80
3  0.000000 15574.03
4  1.000000 15525.77
```

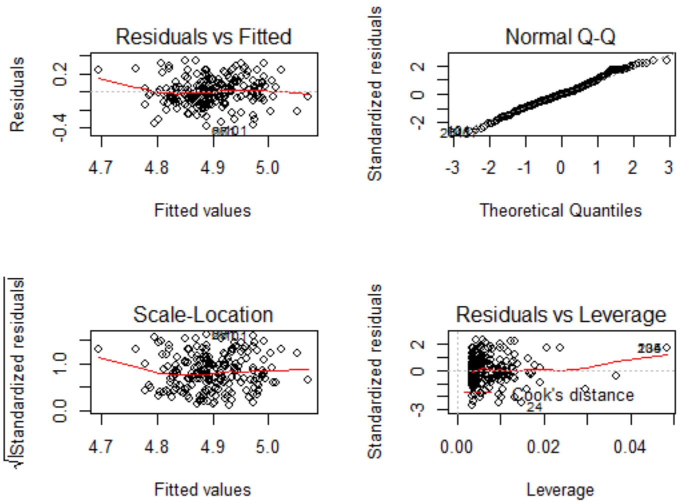



```
> Em2<-lm(echo1$basebp^(3/2)~echo1$sbp)
> summary(Em2)

Call:
lm(formula = echo1$basebp^(3/2) ~ echo1$sbp)

Residuals:
    Min      1Q  Median      3Q     Max
-786.35 -225.06  -34.72  199.56 1033.86

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1027.4596    86.7166  11.848  < 2e-16 ***
echo1$sbp      3.7944     0.5632   6.737 9.35e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 354.2 on 277 degrees of freedom
Multiple R-squared:  0.1408,   Adjusted R-squared:  0.1377
F-statistic: 45.38 on 1 and 277 DF,  p-value: 9.345e-11

> par(mfrow=c(2,2))
> plot(Em2)
```




```
> summary(powerTransform(cbind(echo1$basebp,echo1$sbp)~1,data=echo1))
bcPower Transformations to Multinormality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.0105           0      -0.6074       0.6283
Y2    0.1356           0      -0.2080       0.4792

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                              LRT df    pval
LR test, lambda = (0 0) 0.6038973  2 0.73938

Likelihood ratio test that no transformations are needed
                             LRT df        pval
LR test, lambda = (1 1) 32.62618  2 8.2284e-08
> anova(Em2)
Analysis of Variance Table

Response: echo1$basebp^(3/2)
           Df   Sum Sq Mean Sq F value    Pr(>F)
echo1$sbp   1  5693057 5693057  45.385 9.345e-11 ***
Residuals 277 34746894  125440
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


> Em3<-lm(log(echo1$basebp)~log(echo1$sbp))
> summary(Em3)

Call:
lm(formula = log(echo1$basebp) ~ log(echo1$sbp))

Residuals:
     Min       1Q   Median       3Q      Max
-0.38990 -0.08984 -0.00312  0.09195  0.34268

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.74996    0.17598  21.309  < 2e-16 ***
log(echo1$sbp)  0.23064    0.03533   6.528 3.16e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.147 on 277 degrees of freedom
```

```
Multiple R-squared:  0.1333,   Adjusted R-squared:  0.1302
F-statistic: 42.62 on 1 and 277 DF,  p-value: 3.163e-10

> par(mfrow=c(2,2))
> plot(Em3)
```




```
> anova(Em3)
Analysis of Variance Table

Response: log(echo1$basebp)
                Df Sum Sq Mean Sq F value    Pr(>F)    
log(echo1$sbp)   1 0.9209 0.92089  42.619 3.163e-10 ***
Residuals      277 5.9852 0.02161                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Question Four:

Consider the following R output predicting Marine water growth from Freshwater growth in Salmon:

```
> SL1<- lm(salmon$Marine~salmon$Freshwater)
> summary(SL1)
Call:
lm(formula = salmon$Marine ~ salmon$Freshwater)

Residuals:
    Min      1Q  Median      3Q     Max
-88.222 -27.382  -3.406  24.784  89.977
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       511.3656    18.2547   28.01  < 2e-16 ***
salmon$Freshwater  -0.9602     0.1512   -6.35 6.75e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 39.12 on 98 degrees of freedom
Multiple R-squared:  0.2915,   Adjusted R-squared:  0.2843
F-statistic: 40.32 on 1 and 98 DF,  p-value: 6.747e-09
> summary(salmon$Freshwater)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   53.0    99.0   117.5   117.9   140.0   179.0
> var(salmon$Freshwater)
[1] 676.0541
> summary(salmon$Marine)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  301.0   367.0   396.5   398.1   428.2   511.0
> var(salmon$Marine)
[1] 2138.142
```

a) Construct ANOVA table based on the given output.

```
> anova(SL1)
Analysis of Variance Table

Response: salmon$Marine
                  Df Sum Sq Mean Sq F value    Pr(>F)
salmon$Freshwater  1  61706   61706  40.323 6.747e-09 ***
Residuals         98 149970    1530
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Consider the three observations: 4, 41 and 53

```
       Observation SalmonOrigin      Freshwater Marine
   1           4      Alaska               86     506 (outlier)
   2          41      Alaska               84     511 (Outlier)
   3          53      Canada              179     407 (Good Leverage)
```

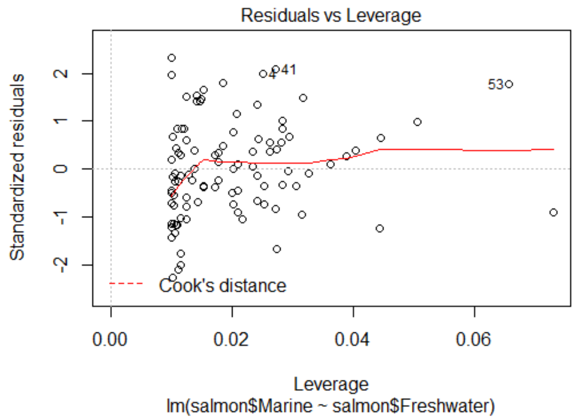
b) Which of these three points is a leverage point?

c)Which of these three points is an outlier?

d) Based on your answers of part b and c, classify these points as one of the following:

i) A bad leverage point   ii) An outlier but Not a leverage point.

iii) A good leverage point   iv) Not a leverage point nor an outlier (ordinary)



Residuals vs Leverage
Standardized residuals
-2
-1
0
1
2
41
53
Cook's distance
0.00
0.02
0.04
0.06
Leverage
lm(salmon$Marine ~ salmon$Freshwater)