Combined Cycle Power Plants Regression

Becky Su (Thursday 2PM)
Helen Cui (Friday 12PM)
Sarah Hermann (Thursday 2PM)

ABSTRACT
Our project examined the Combined Cycle Power Plant dataset from the UCI Repository and the predictors that affect the net hourly electrical energy output for the combined cycle power plant. We tested the predictors to check their usefulness to the production of electricity as well as specifically ambient pressures' usefulness to the model and determined that all of the predictors are important to the model through graphs and tests detailed in the report. We created our fitted model to satisfy all the assumptions by transforming the response variable as well as one of the predictors and adding an interaction term. Additionally, we noted some potential data points that deviated from many of the others and checked for outliers and removed them to better fit the assumptions and the model. Our final model was responsible for a very high percentage of the variability in electrical energy output, meaning that it was a good predictor for the response.

INTRODUCTION
The Combined Cycle Power Plant dataset details the net hourly electrical energy output from a plant over the course of six years. There are 9568 instances in the dataset. With the importance of technology and electricity in society, having more electricity output is important and these combined cycle power plants produce 50% more electricity than a regular simple cycle power plant. The two turbines are combined in one cycle and transfer electricity between the two as they work and the vacuum captures the steam and turns it into more energy. The combined cycle plant utilizes fuel efficiently which leads to lower lifecycle costs. The dataset has four predictors which are temperature, which we called 'a', ambient pressure, 'c', relative humidity, 'd', and exhaust vacuum, 'b'. The temperature, ambient pressure and relative humidity affect the gas turbine, which generates electricity, and the vacuum affects the steam turbine which also contributes to the generation of electricity. The vacuum captures the steam that the gas turbine exhausts and turns it into more energy. Regular simple cycle power plants only have one turbine and cycle, which is why the combined cycle power plants produce more energy. The electrical efficiency of a simple cycle power plant is usually between 25-40% while the combined cycle power plant has an electrical efficiency of around 60% or higher, which explains why finding the best fit for the model could be beneficial to society as it would produce more electricity more efficiently.

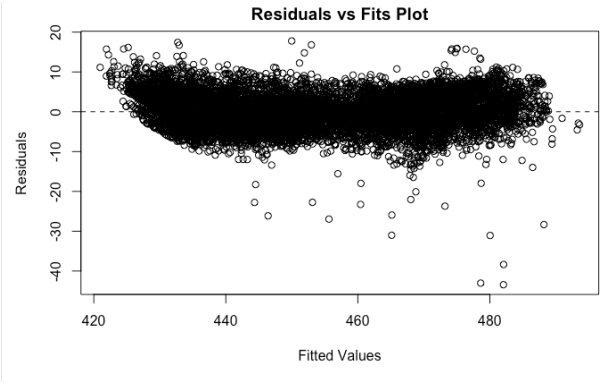
QUESTIONS OF INTEREST
1. Are all the predictors useful?
2. Is ambient pressure a useful predictor?
3. How much of the variability in electricity energy output is explained by our new transformed model?

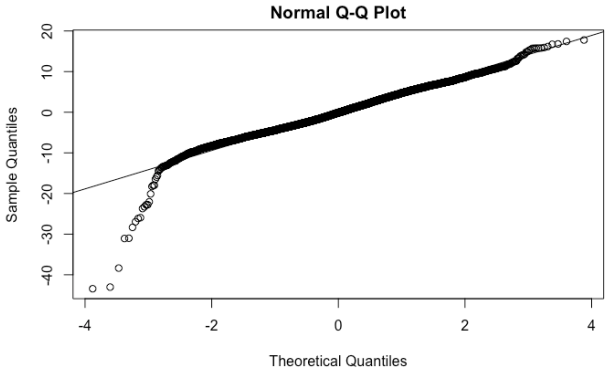
DATA AND REGRESSION METHODS
Our data set is Combined Cycle Power Plant and we are looking at the response, net hourly electrical energy output, with predictors of temperature, ambient pressure, relative humidity and exhaust volume. In our pairs plot with the initial simple model of Y=a+b+c+d, there were not

many relationships between predictors and the response but there was a relationship between a (temperature) and b (vacuum) which lead to us adding an interaction term (a*b). We used backward selection to check if all the predictors of the new model were important. Then, we used power transform on the model, and the estimated power rounded for b (vacuum) was 0 so we used the log function for b. By using box cox transformation, the lambda value is between -1 and -2 so we chose to use lambda=-2 since the confidence level is closer to -2 than -1, which lead to our transformation of the response variable (Y). Our final model is net hourly electrical energy output^(-2) = temperature + log(vacuum) + pressure + humidity +temperature*vacuum (Y^{-2}=a+log(b)+c+d+ab).

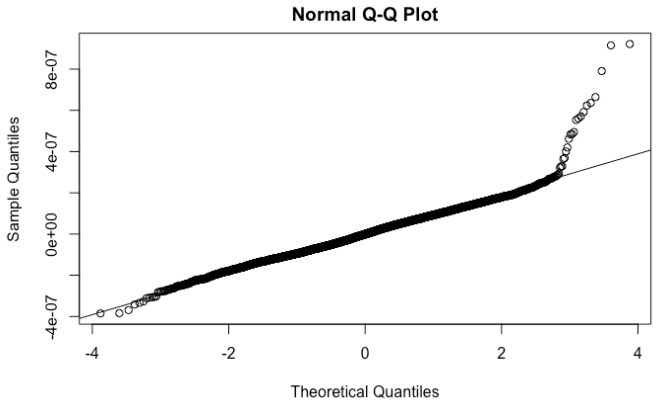
One of our questions is about the usefulness of all our predictors, which we will use a global F-test to check. To test the usefulness of ambient pressure, we will use a partial F-test. We used our new model and looked at the summary table to determine how much of the variability is due to our model.

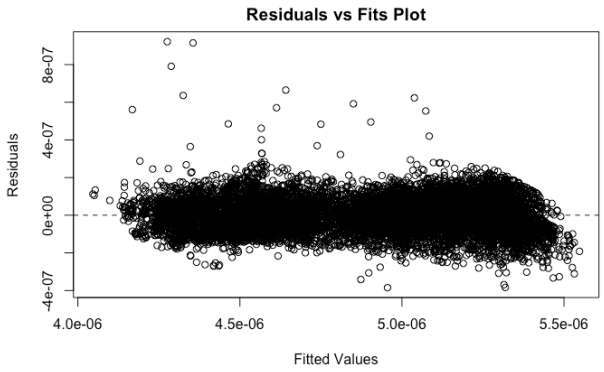
When we checked assumptions before any transformations, we determined that the normality, constant variance and linearity assumptions were violated based on the residual plot and normal Q-Q plot; they both had points that were deviated from the rest.



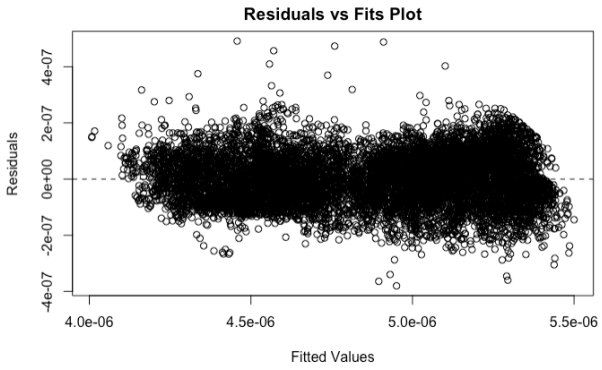


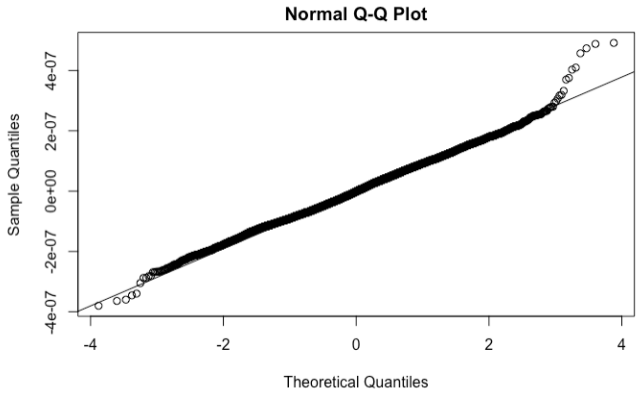
After examining the assumptions and fitting a new model, we tested again and noticed a difference in the plots. The residual plot and normal Q-Q plots still had some slight violations of the assumptions.





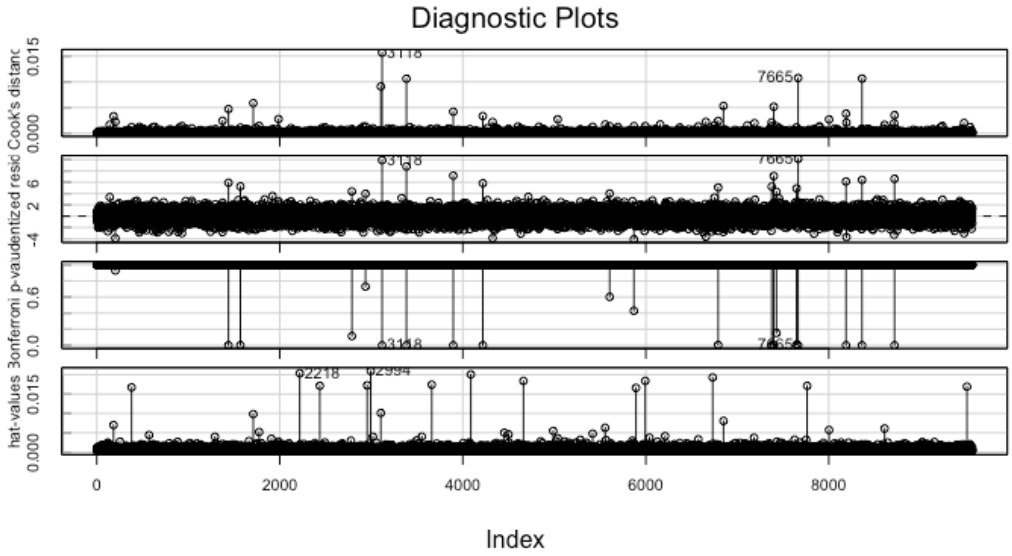
So, we performed a test for outliers and found that there were 10 outliers which we then removed from our model. When we ran the residual vs fitted and normal Q-Q plot with the new model without the 10 outliers, all of the assumptions were finally met.





Additionally, we tested the outliers to see if they were influential and determined that two points were more influential than the others but removed all 10 outliers as it had a greater impact on the fit of the model.

	rstudent <dbl>	unadjusted p-value <dbl>	Bonferonni p <dbl>
7665	10.146763	4.5347e-24	4.3388e-20
3118	9.898561	5.4466e-23	5.2113e-19
3384	8.790245	1.7501e-18	1.6744e-14
3896	7.153837	9.0585e-13	8.6672e-09
7399	7.090657	1.4294e-12	1.3676e-08
8718	6.609779	4.0545e-11	3.8793e-07
8363	6.416241	1.4630e-10	1.3998e-06
8188	6.134607	8.8755e-10	8.4920e-06
1439	5.908551	3.5692e-09	3.4150e-05
4219	5.815612	6.2351e-09	5.9658e-05



REGRESSION ANALYSIS, RESULTS AND INTERPRETATION

All Predictors Global F-Test

The null hypothesis is $\beta_1=\beta_2=\beta_3=\beta_4=\beta_{12}=0$ and the null hypothesis is at least one of the predictors is not equal to 0. We found a p-value of 2.2e-16 which means that we reject the null hypothesis and conclude that at least one of the predictors is not equal to 0 and therefore we include them in the model. Because the p-value is very small, we have strong evidence against the null hypothesis.

Partial F-Test for the Pressure

The null hypothesis is $\beta_3=0$ and the null hypothesis is that β_3 does not equal to 0. The partial F-test resulted in a p value of 2.2e-16, which is very small and smaller than alpha (.05) so we reject the null and conclude that ambient pressure (β_3) is a useful predictor and we cannot remove it from the model. Again, the small p-value tells us that the data is strong evidence against β_3 equalling 0.

Variability of Energy Output

The variability of energy output is tested by examining the summary of the fitted, transformed model and from that, we found a value of .9321 which means that 93.21% of variability in electrical energy output is explained by our new model. So our model is responsible for most of the variability in energy output. The high percentage further supports the results that all the predictors in our model are important because it shows that they result in a very high amount of the variation.

CONCLUSION

Based on our final model, net hourly electrical energy output ^(-2) = temperature + log(vacuum) + ambient pressure + relative humidity + temperature*vacuum, we find the model that will predict the hourly energy output and therefore can figure out, for example, how to maximize the output or minimize the input, to more effectively use the combined cycle power plants to produce electricity. A main message from our model is that all of the predictors (the variables that contribute to the process of combined cycle power plants) are useful to predict the energy output. A problem that may have occurred had to do with the dataset. Because there were 9568 instances and the ranges of the predictors were not very large, many of the data points could have been very similar in predictors and responses. Also, with only four attributes, there could have been many other factors that influenced the results that were not tested, which could lead to further analysis of these other factors and whether or not they would have a significant effect on net hourly electrical energy output. The four attributes also did not allow us to make the model smaller but rather, we added terms to improve the model. This model could lead to further analysis of how to further improve the efficiency of the power plants in order to produce even more electricity more efficiently.

APPENDIX

```
library(readxl)
library(car)
```

```
## Loading required package: carData
```

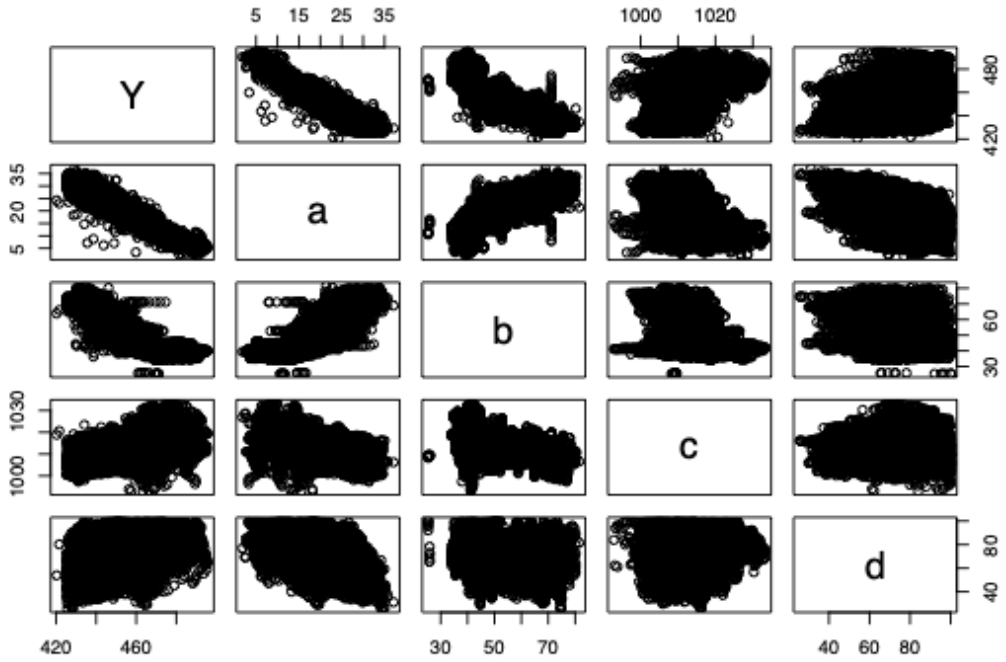
```
data.ccpp<-read_excel("/Users/sarahhermann/Desktop/Folds5x2_pp.xlsx")
data(data.ccpp)
```

```
## Warning in data(data.ccpp): data set 'data.ccpp' not found
```

```
attach(data.ccpp)

a<-data.ccpp$AT
b<-data.ccpp$V
c<-data.ccpp$AP
d<-data.ccpp$RH
Y<-data.ccpp$PE
```

```
pairs(~Y+a+b+c+d)
```

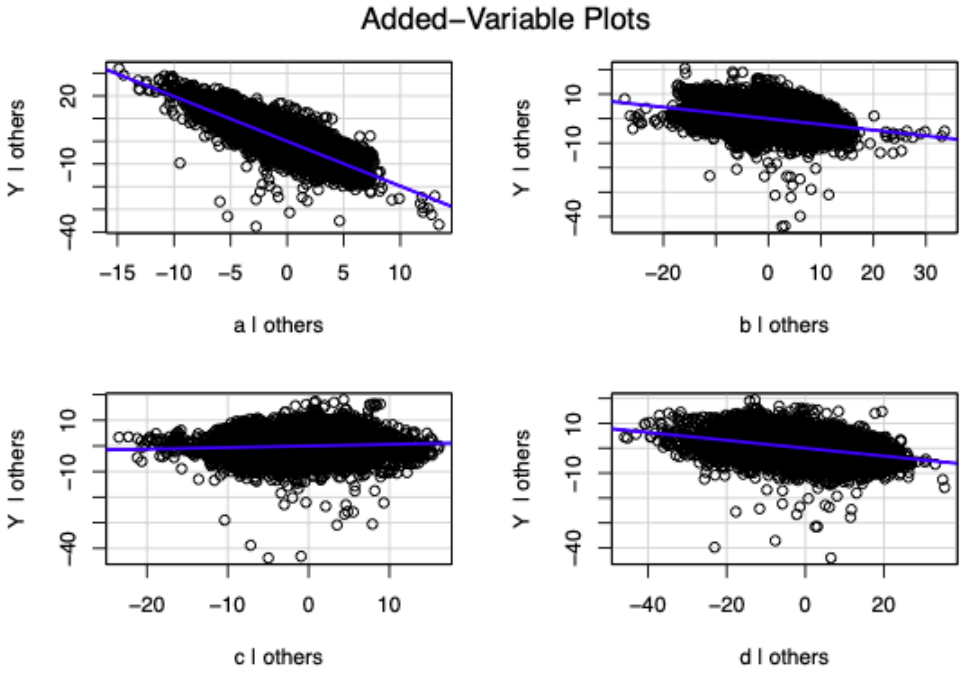



```
full.lm<-lm(Y~a+b+c+d)
summary(full.lm)
```

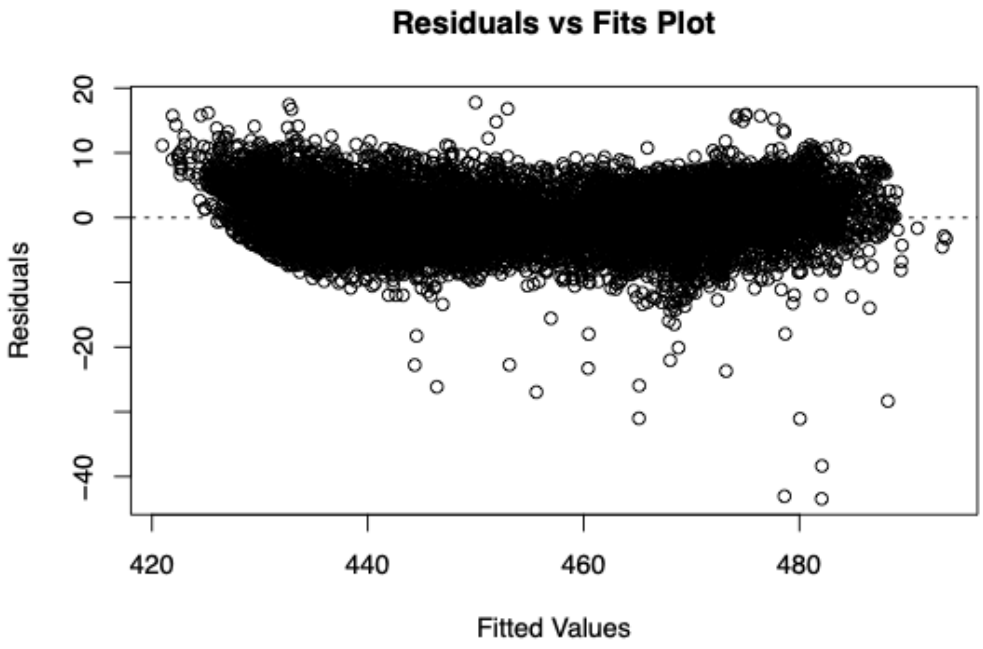
```
##
## Call:
## lm(formula = Y ~ a + b + c + d)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.435  -3.166  -0.118   3.201  17.778
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 454.609274   9.748512   46.634  < 2e-16 ***
## a            -1.977513   0.015289 -129.342  < 2e-16 ***
## b            -0.233916   0.007282  -32.122  < 2e-16 ***
## c             0.062083   0.009458    6.564 5.51e-11 ***
## d            -0.158054   0.004168  -37.918  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.558 on 9563 degrees of freedom
## Multiple R-squared:  0.9287, Adjusted R-squared:  0.9287
## F-statistic: 3.114e+04 on 4 and 9563 DF,  p-value: < 2.2e-16
```

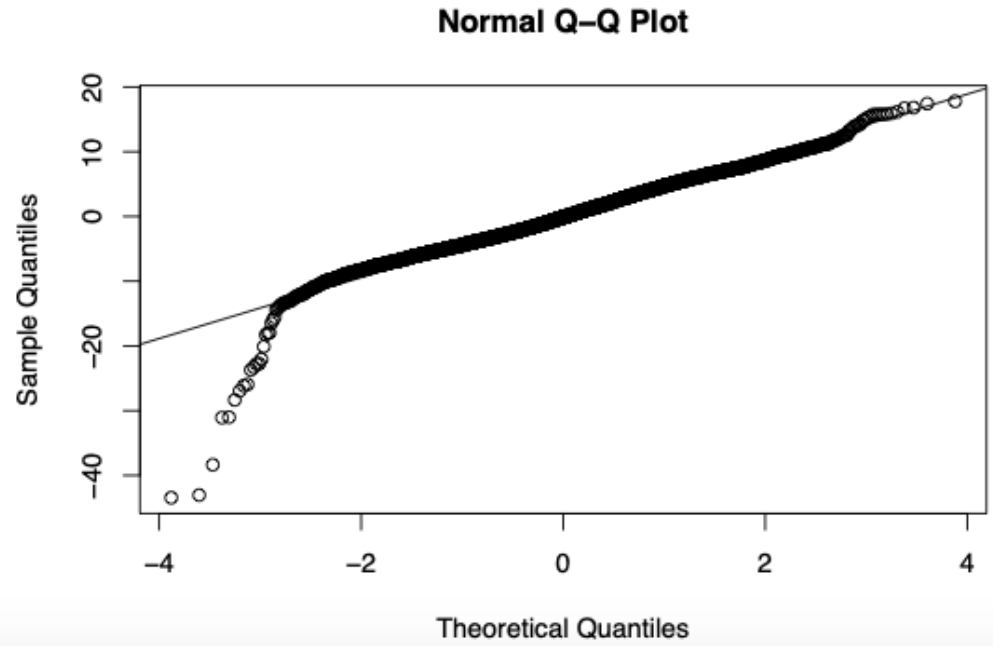
```
avPlots(full.lm, id=FALSE)
```




```
Residualtest<-full.lm$residuals
Fitted<-full.lm$fitted.values
plot(Residualtest-Fitted, xlab = 'Fitted Values', ylab = 'Residuals', main = 'Residuals vs Fits Plot')
abline(h = 0, lty = 2)
```




```
e<-Residualtest
qqnorm(e)
qqline(e)
```



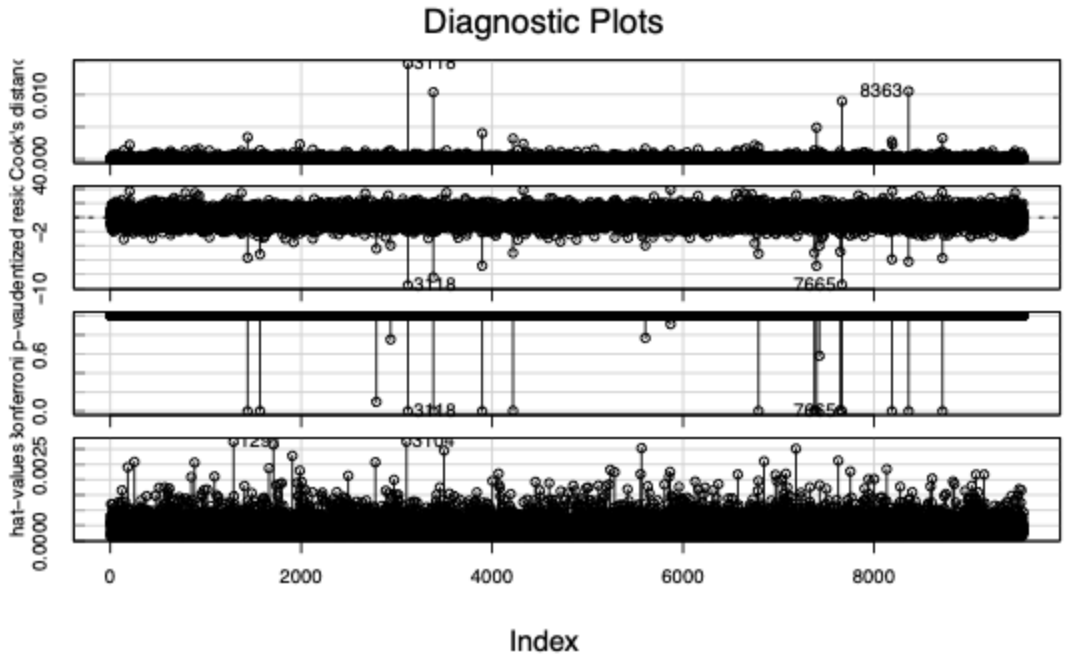
```r
ccpp.lm <- lm(Y-a+b+c+d)
s.resid <- rstudent(ccpp.lm)
as=abs(s.resid)
which(as==max(as))
```

```
## 3118
## 3118
```

```r
#test for outliers
outlierTest(ccpp.lm)
```

```
##       rstudent unadjusted p-value Bonferonni p
## 3118 -9.577798         1.2395e-21   1.1859e-17
## 7665 -9.487390         2.9376e-21   2.8107e-17
## 3384 -8.448690         3.3759e-17   3.2301e-13
## 7399 -6.835304         8.6831e-12   8.3080e-08
## 3896 -6.820705         9.6082e-12   9.1931e-08
## 8363 -6.233478         4.7546e-10   4.5492e-06
## 8188 -5.924763         3.2355e-09   3.0957e-05
## 8718 -5.745868         9.4249e-09   9.0177e-05
## 1439 -5.702510         1.2156e-08   1.1631e-04
## 1569 -5.207803         1.9507e-07   1.8664e-03
```

```r
influenceIndexPlot(full.lm)
```




```r
library(leaps)
#Full model
mod.0 <- lm(Y-1)
mod.full= -a + d + b + c

mod.1 <- update(mod.0, mod.full)
mod.backward <- step(mod.1, scope = c(lower = - 1, direction = 'backward'))
```

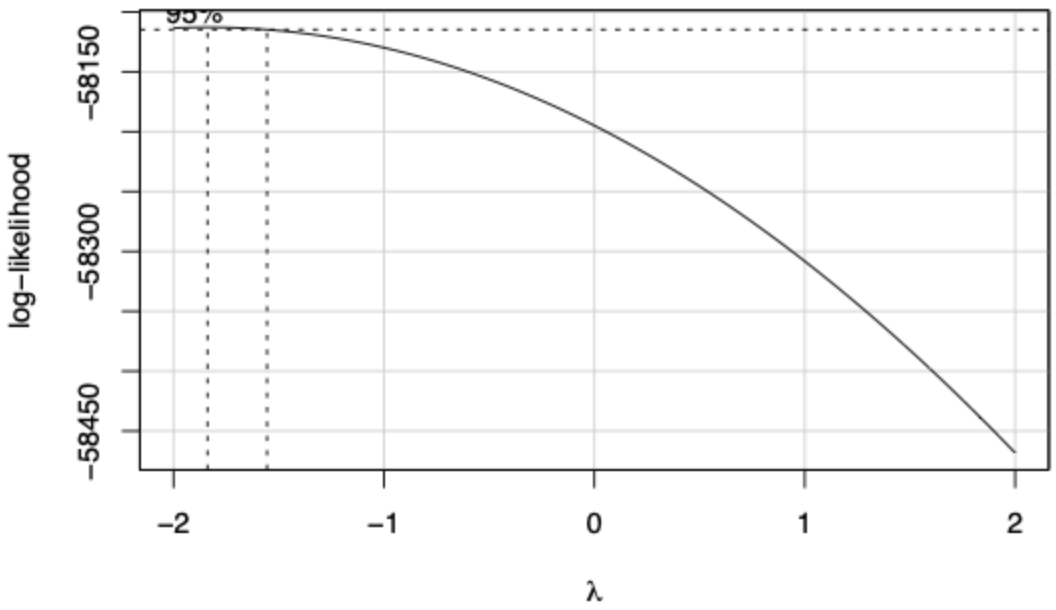
```
## Start:  AIC=29033.42
## Y ~ a + d + b + c
##
##        Df Sum of Sq    RSS   AIC
## <none>              198702 29033
## - c     1       895 199598 29074
## - b     1     21440 220142 30012
## - d     1     29875 228578 30372
## - a     1    347607 546309 38708
```

```
Trans.ccpp<-powerTransform(cbind(a, b, c, d)~1)
summary(Trans.ccpp)
```

```
## bcPower Transformations to Multinormality
##   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## a    1.2705        1.27       1.2387       1.3023
## b   -0.1366       -0.14      -0.2153      -0.0578
## c   -2.9843       -2.98      -2.9970      -2.9717
## d    1.6195        1.62       1.5374       1.7016
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                   LRT df       pval
## LR test, lambda = (0 0 0 0) 8554.261  4 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                   LRT df       pval
## LR test, lambda = (1 1 1 1) 1528.759  4 < 2.22e-16
```

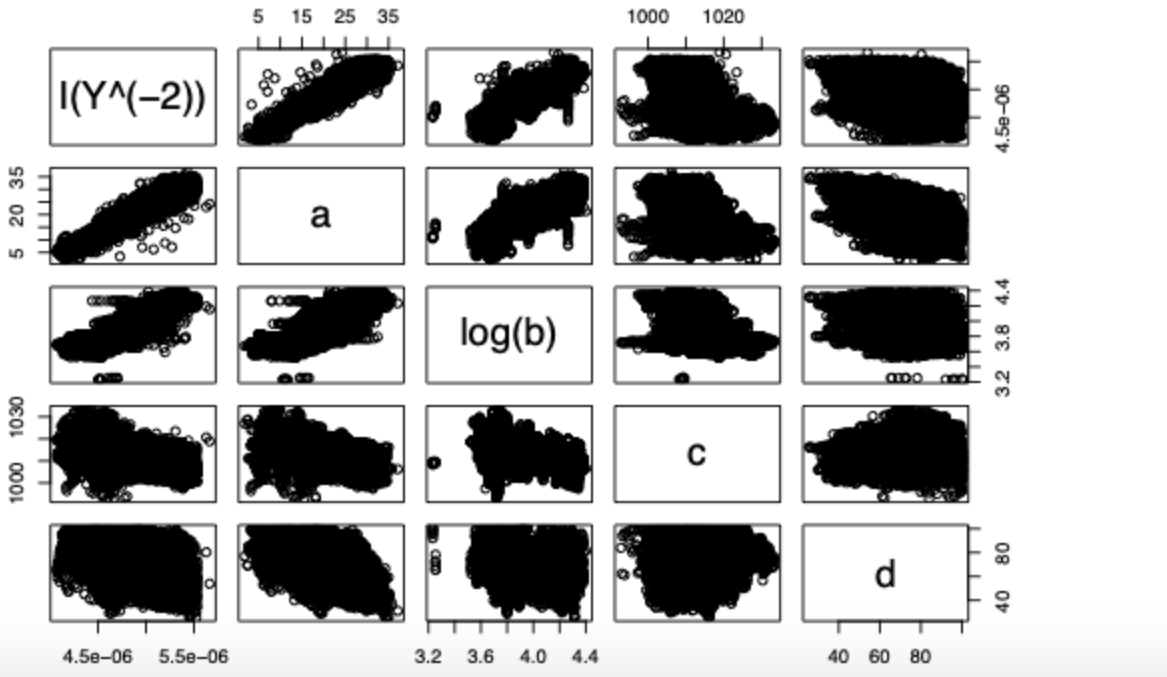
From using power transform, the likelihood ratio tests inidicate that using log transformations for all variables is not appropriate, neither should we use no transformations because both p-values (2.22e-16) are less than the default alpha level of 0.05. By rounding the estimated lambda powers, 1.2705 to 1, -0.1366 to 0, -2.9843 to -3, and 1.6195 to 1, we decided to log transform the predictor b and not transform the predictor c because it would over complicate our model.

```
ccpp.Trans <- with(data.ccpp, data.frame(Y, a, log(b), c, d))
ccpp.lm<-lm(Y~., data = ccpp.Trans)
boxCox(ccpp.lm)
```

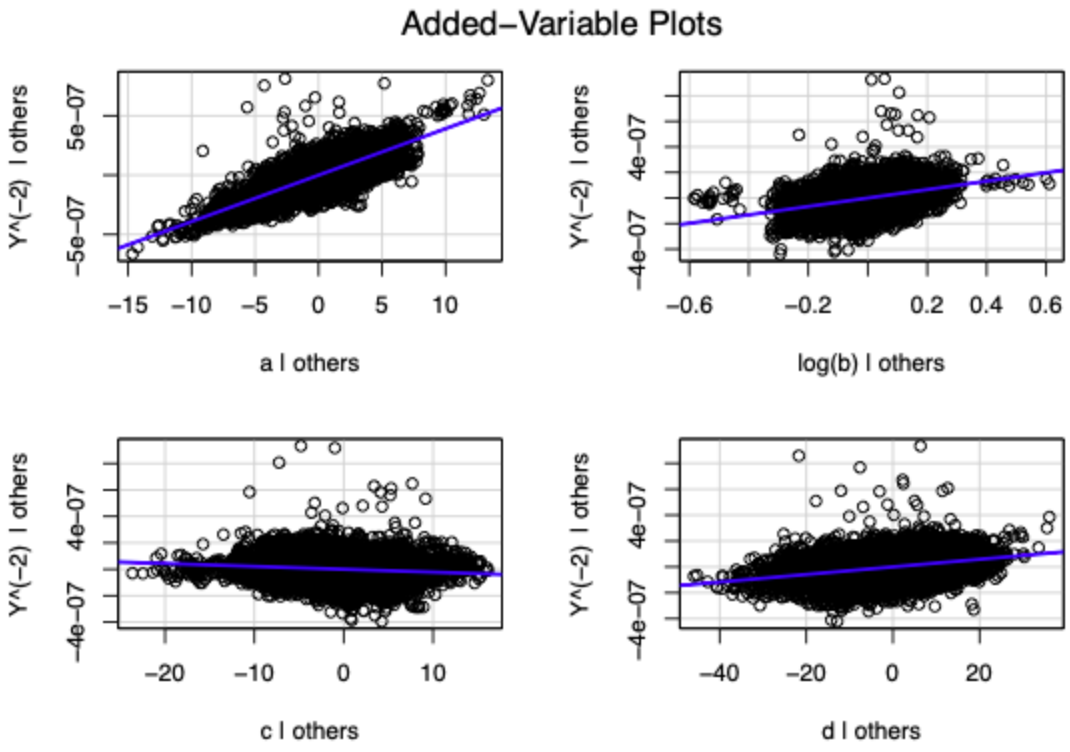


By using box cox transformation, the lambda value is between -1 and -2 so we chose to use lambda=-2 since the confidence level is closer to -2 than -1. Thus our final transformed model with the interaction terms is Y^(-2)~a+log(b)+c+d.

```
pairs(-I(Y^(-2))+a+log(b)+c+d)
```



```
full.fit<-lm(Y^(-2) ~ a + log(b) + c+ d)
avPlots(full.fit, id=FALSE)
```



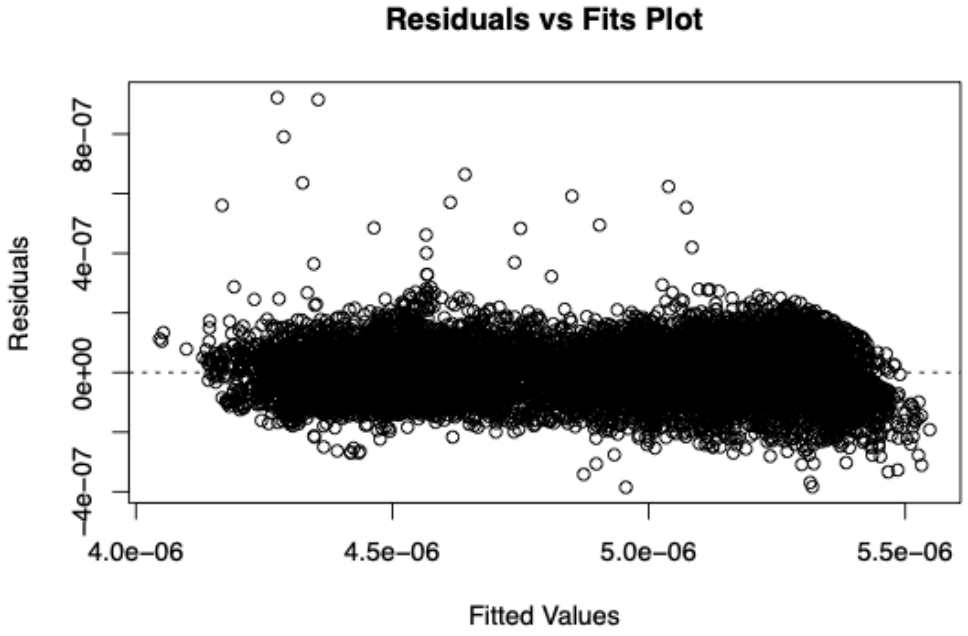

```
library(leaps)
#Full model
mod.full= ~a +log(b) + c + d + a*b

# Base model
mod.0 <- lm(Y^(-2)~1)
mod.1 <- update(mod.0, mod.full)
mod.backward <- step(mod.1, scope = c(lower = ~ 1, direction = 'backward'))
```

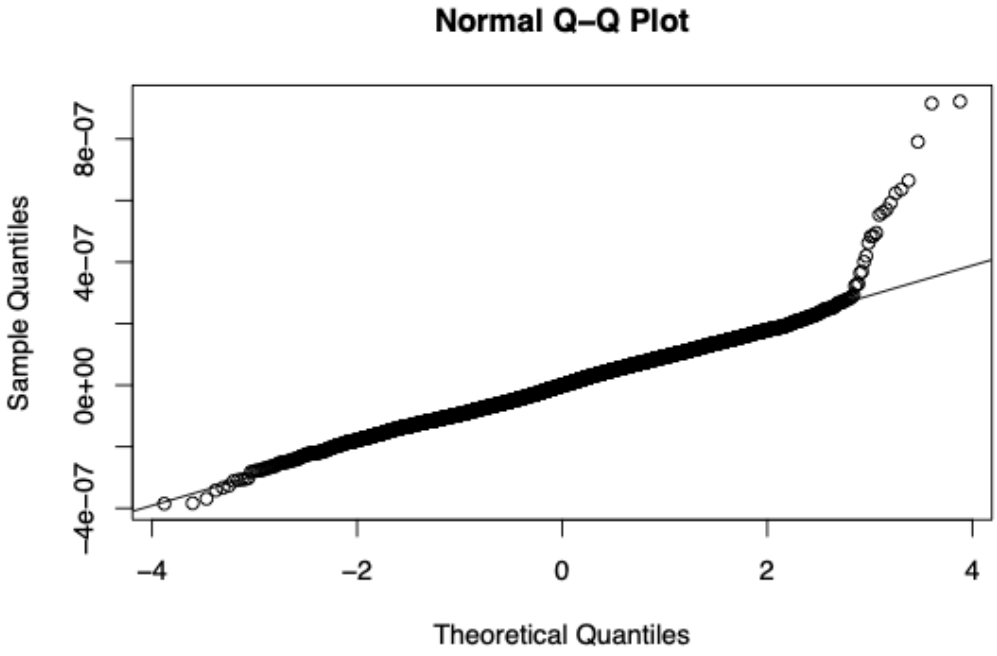
```
## Start:  AIC=-309757.8
## Y^(-2) ~ a + log(b) + c + d + b + a:b
##
##          Df  Sum of Sq        RSS     AIC
## <none>                 8.3211e-11 -309758
## - log(b)  1 7.6990e-13 8.3981e-11 -309672
## - c       1 9.7090e-13 8.4182e-11 -309649
## - a:b     1 2.8788e-12 8.6090e-11 -309434
## - d       1 7.7052e-12 9.0916e-11 -308913
```

By using backward selection, we found out that all predictors are important predictors. Thus, the final model is Y^(-2) ~ a + log(b) + c + d + a*b.

```
Residual<-full.fit$residuals
Fitted<-full.fit$fitted.values
plot(Residual~Fitted, xlab = 'Fitted Values', ylab = 'Residuals', main = 'Residuals vs Fits Plot')
abline(h = 0, lty = 2)
```




```
e<-Residual
qqnorm(e)
qqline(e)
```



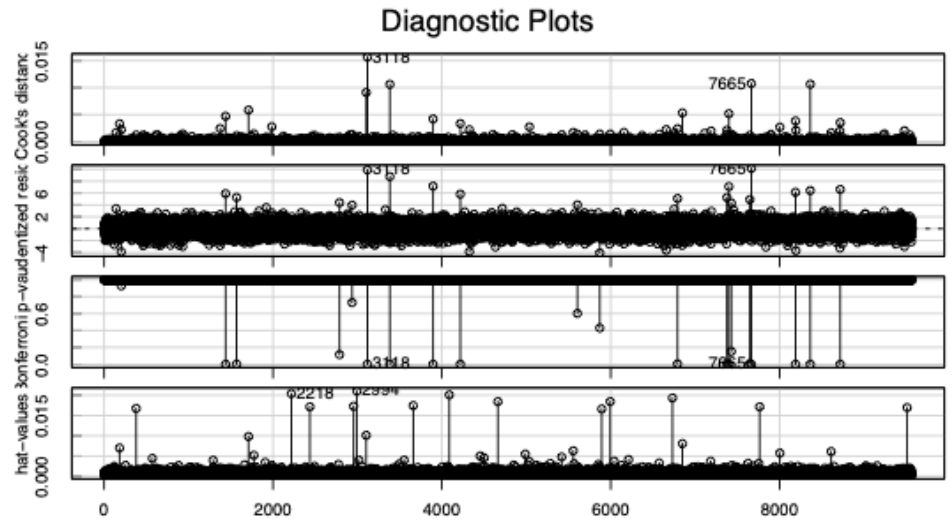

```
ccpp.lm <- lm(Y^(-2)~a+log(b)+c+d+a*b)
s.resid <- rstudent(ccpp.lm)
as=abs(s.resid)
which(as==max(as))
```

```
## 7665
## 7665
```

```
#test for outliers
outlierTest(ccpp.lm)
```

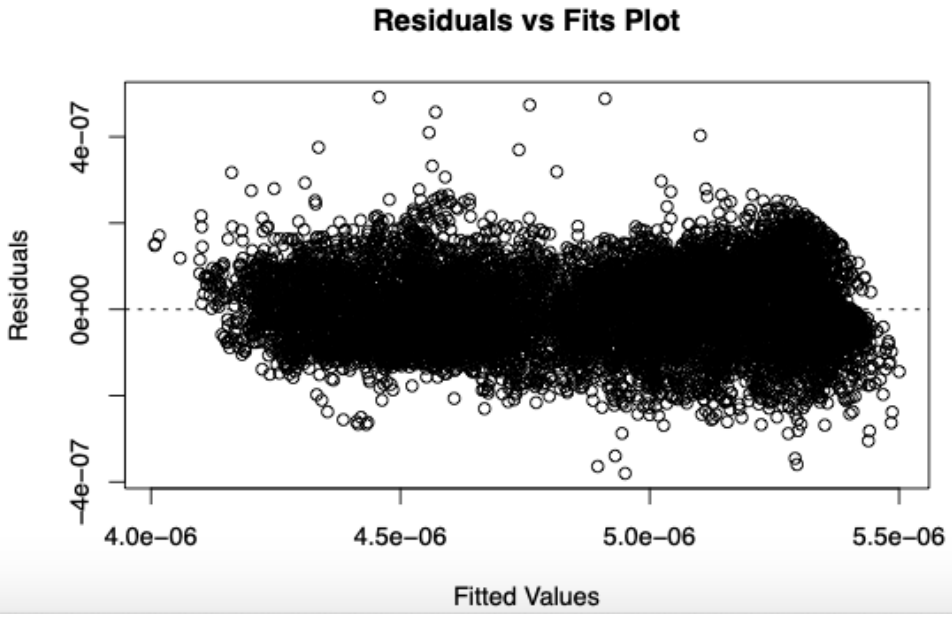
```
##       rstudent unadjusted p-value Bonferonni p
## 7665 10.146763         4.5347e-24   4.3388e-20
## 3118  9.898561         5.4466e-23   5.2113e-19
## 3384  8.790245         1.7501e-18   1.6744e-14
## 3896  7.153837         9.0585e-13   8.6672e-09
## 7399  7.090657         1.4294e-12   1.3676e-08
## 8718  6.609779         4.0545e-11   3.8793e-07
## 8363  6.416241         1.4630e-10   1.3998e-06
## 8188  6.134607         8.8755e-10   8.4920e-06
## 1439  5.908551         3.5692e-09   3.4150e-05
## 4219  5.815612         6.2351e-09   5.9658e-05
```

```
full.fit<-lm(Y^(-2) ~ a + log(b) + c + d+a*b)
influenceIndexPlot(full.fit)
```

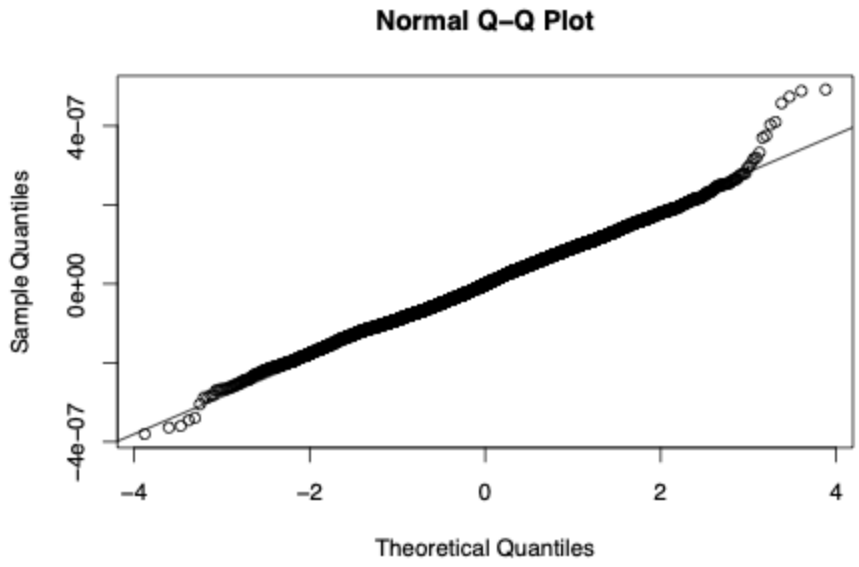



```
ccpp.lm<-lm((Y^(-2) ~ a + log(b) + c + d+a*b), subset=-c(3118,7665,3384,3896,7399,8718,8363,8188,1439,
```

```
Residualt<-ccpp.lm$residuals
Fittedt<-ccpp.lm$fitted.values
plot(Residualt~Fittedt, xlab = 'Fitted Values', ylab = 'Residuals', main = 'Residuals vs Fits Plot')
abline(h = 0, lty = 2)
```



```
Residual<-ccpp.lm$residuals
e<-Residual
qqnorm(e)
qqline(e)
```




```
#PROBLEM 3B:
ccpp.full.lm<-lm(Y^(-2)~a+log(b)+c+d+a*b)
summary(ccpp.full.lm)
```

```
##
## Call:
## lm(formula = Y^(-2) ~ a + log(b) + c + d + a * b)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -3.804e-07 -6.587e-08 -2.120e-09  6.237e-08  9.412e-07
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.276e-06  2.629e-07  27.676   <2e-16 ***
## a            5.776e-08  1.072e-09  53.877   <2e-16 ***
## log(b)      -6.695e-07  7.118e-08  -9.405   <2e-16 ***
## c           -2.073e-09  1.962e-10 -10.562   <2e-16 ***
## d            2.614e-09  8.785e-11  29.755   <2e-16 ***
## b            2.648e-08  1.650e-09  16.050   <2e-16 ***
## a:b         -3.672e-10  2.019e-11 -18.187   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.329e-08 on 9561 degrees of freedom
```

```
## Multiple R-squared:  0.9329, Adjusted R-squared:  0.9329
## F-statistic: 2.215e+04 on 6 and 9561 DF,  p-value: < 2.2e-16
```

The null hypothesis for the global F-test for this model using log(Life) as the response is:

H_0: β_1=β_2=β_3=β_4=$\beta_{(12)}$=0 and the alternative hypothesis is H_1: at least one these predictors is not 0. Since the p-value is 2.2e-16 which is smaller than the default alpha level of 0.05, we reject the null hypothesis and conclude that at least one predictor is useful.

```
#PROBLEM 3D:
ccpp.red.lm<-lm(Y^(-2) ~ a + log(b) + d +a*b)
summary(ccpp.red.lm)
```

```
##
## Call:
## lm(formula = Y^(-2) ~ a + log(b) + d + a * b)
##
## Residuals:
##        Min          1Q      Median          3Q         Max
## -3.873e-07 -6.533e-08 -1.800e-09  6.305e-08  9.435e-07
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.421e-06  1.968e-07   27.55   <2e-16 ***
## a            5.946e-08  1.066e-09   55.79   <2e-16 ***
## log(b)      -7.653e-07  7.101e-08  -10.78   <2e-16 ***
## d            2.866e-09  8.503e-11   33.70   <2e-16 ***
## b            2.810e-08  1.652e-09   17.01   <2e-16 ***
## a:b         -3.715e-10  2.030e-11  -18.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.383e-08 on 9562 degrees of freedom
## Multiple R-squared:  0.9321, Adjusted R-squared:  0.9321
## F-statistic: 2.626e+04 on 5 and 9562 DF,  p-value: < 2.2e-16
```

The null hypothesis for the partial F-test to test if ambient pressure is an useful predictor is:

H_0: β_3=0 and the alternative hypothesis is H_1: β_3 is not 0. Since the p-value is 2.2e-16 which is smaller than the default alpha level of 0.05, we reject the null hypothesis and conclude that ambient pressure is an useful predictor.

Adjusted R^2: 93.21% of the variability of energy output is explained by the model.