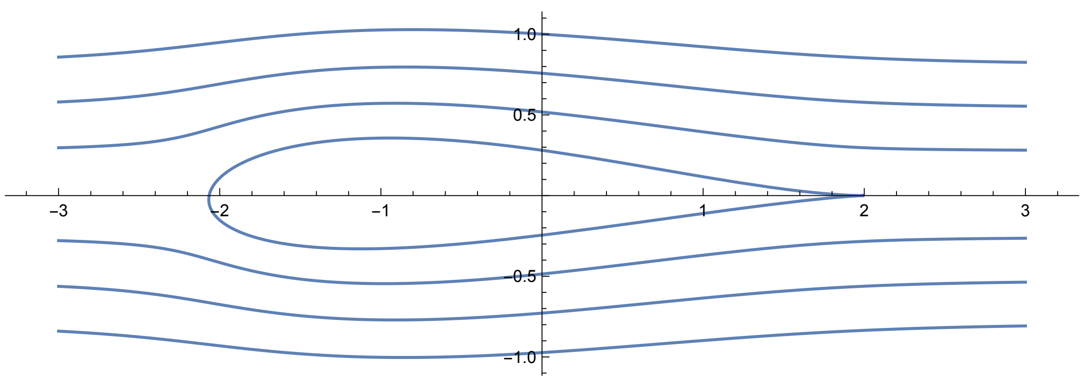
General Linear Model Project

Heran Song

Introduction

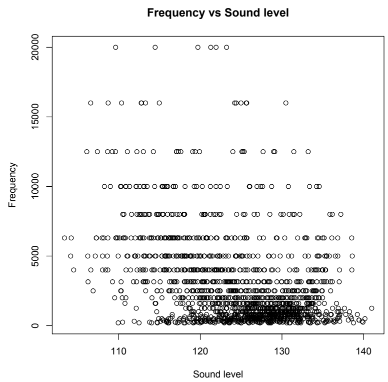
The airfoil noise data set is a 1503 by 6 data set provided by NASA comprises of 1503 measurements of noise generated by a NACA 0012 airfoil in a wind tunnel under 5 different parameters. The 5 parameters are Frequency in hertz(V1), angle of attack in degrees(V2), chord length in meters(V3), free-stream velocity in meters per second(V4), suction side displacement thickness in meters(V5), and the response variable is the scaled sound pressure level in decibels(V6). The goal of this paper will be trying to model our data using techniques of linear regression, and trying to predict the noise generated by the airfoil in flight using parameters described above.

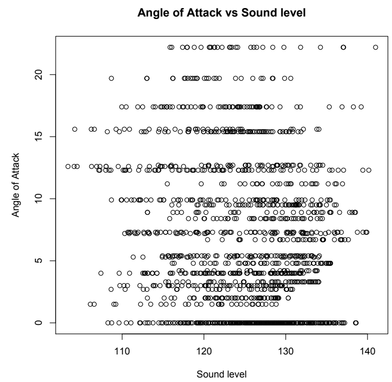
A little background about the NACA 0012 airfoil. First digit describing maximum camber as percentage of the chord. Second digit describing the distance of maximum camber from the airfoil leading edge in tens of percents of the chord. Last two digits describing maximum thickness of the airfoil as percent of the chord. Our case would be describing a symmetric airfoil with the maximum thickness 12% of the chord length.

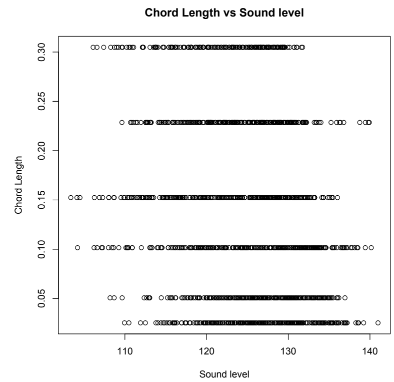


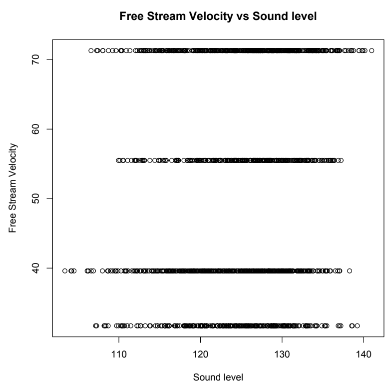
Analysis:

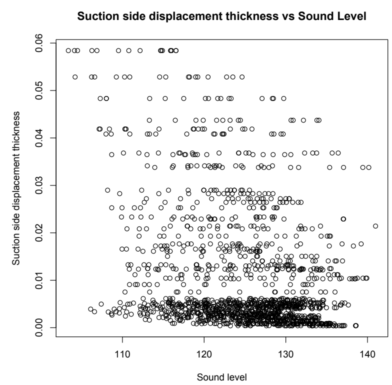
Before any analysis is done, we first try to find anything thing unusual with the data itself by plotting each of the predictors against the response.











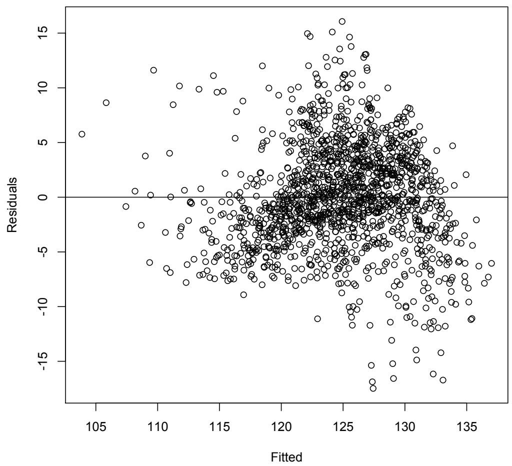
We see that some of the predictors might be categorical.
We will start out with the most simple model with all the predictors.

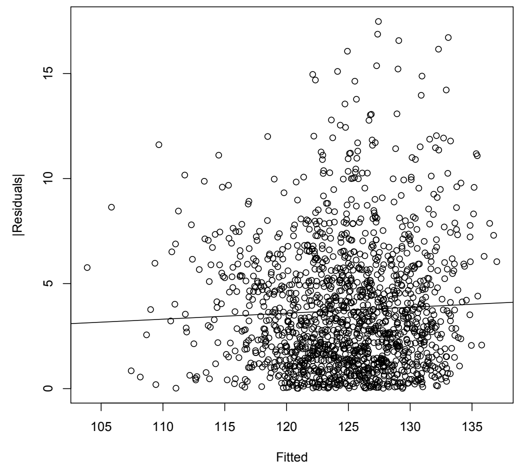
```
> lm1<-lm(V6~.,data=airfoil)
> summary(lm1)

Call:
lm(formula = V6 ~ ., data = airfoil)

Residuals:
Min      1Q  Median      3Q     Max
-17.480  -2.882  -0.209   3.152  16.064

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.328e+02  5.447e-01  243.87   <2e-16 ***
V1          -1.282e-03  4.211e-05  -30.45   <2e-16 ***
V2          -4.219e-01  3.890e-02  -10.85   <2e-16 ***
V3          -3.569e+01  1.630e+00  -21.89   <2e-16 ***
V4           9.985e-02  8.132e-03   12.28   <2e-16 ***
V5          -1.473e+02  1.501e+01   -9.81   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 4.809 on 1497 degrees of freedom
Multiple R-squared:  0.5157,Adjusted R-squared:  0.5141
F-statistic: 318.8 on 5 and 1497 DF,  p-value: < 2.2e-16
```

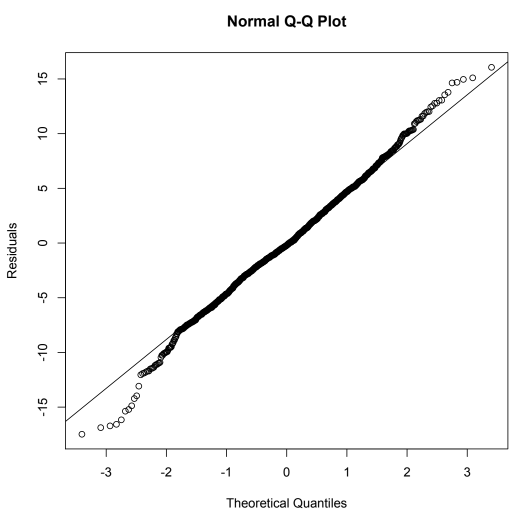
We see that all the predictors are significant with small p-values.

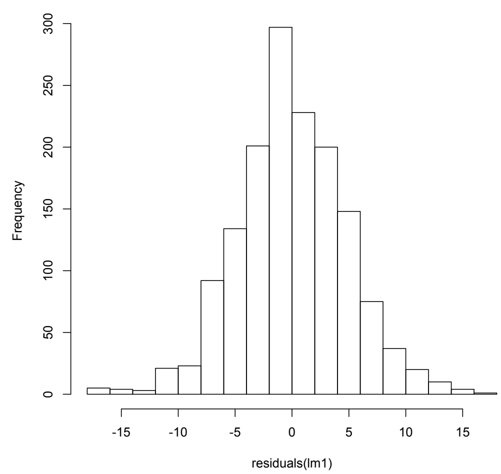
We will now perform diagnostics on the model. First we check for heteroscedasticity and linearity in the residuals.





On the left is a residuals vs the predicted value. On the right is absolute value of residuals vs the predicted value to increase the resolution for detecting nonconstant variance. We can see that are signs of mild nonconstant variance, but nothing indicates nonlinearity. Next we check for normality in the residuals.





The plots seems to show normality. But with a p-value 0.0002465, we reject our null hypothesis of normality using $\alpha = .05$.

```
> shapiro.test(residuals(lm1))

Shapiro-Wilk normality test

data:  residuals(lm1)
W = 0.9956, p-value = 0.0002465
```

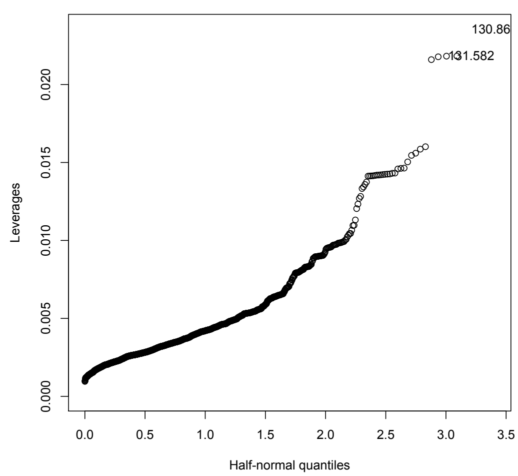
To check for correlation within data.

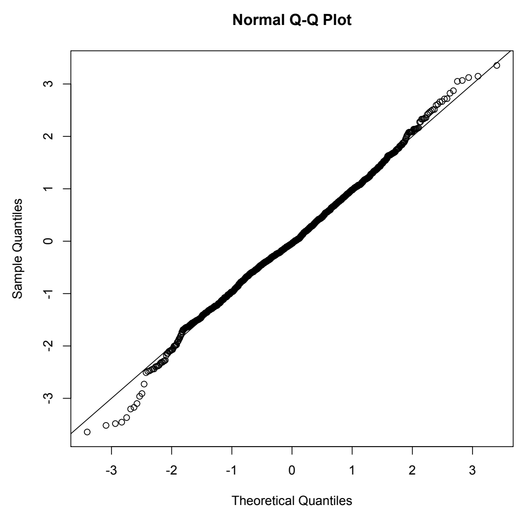
```
> dwtest(lm1)

Durbin-Watson test

data:  lm1
DW = 0.4474, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

We see that there are autocorrelation.

Next we look for leverage points, outliers and influential points. Though due to such large data set, 1 or 2 points will not change the result very much.

First we check for leverage points, unusuals in the predictor space which has the potential to influence the fit.





Next we check for possible outliers.

```
> lm1.jack<-rstudent(lm1)
> lm1.jack[which.max(abs(lm1.jack))]
1165
-3.657776
> qt(.05/(1503*2),1497)
[1] -4.162529
```

We see that observation 1165 is not an outlier. There doesn't seem to be any outliers.

Finally, we check for influential points.

```
> lm1.cook1<-lm(V6~.,data=airfoil,subset=(lm1.cook<max(lm1.cook)))
> summary(lm1.cook1)

Call:
lm(formula = V6 ~ ., data = airfoil, subset = (lm1.cook < max(lm1.cook)))

Residuals:
Min       1Q   Median       3Q      Max
-17.4870  -2.8699  -0.2291   3.1354  16.0752

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.329e+02  5.439e-01 244.278   <2e-16 ***
V1          -1.297e-03  4.245e-05 -30.547   <2e-16 ***
V2          -4.240e-01  3.884e-02 -10.916   <2e-16 ***
V3          -3.564e+01  1.628e+00 -21.895   <2e-16 ***
V4           1.001e-01  8.120e-03  12.333   <2e-16 ***
V5          -1.468e+02  1.499e+01  -9.796   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 4.801 on 1496 degrees of freedom
Multiple R-squared:  0.5176,Adjusted R-squared:  0.5159
F-statistic:   321 on 5 and 1496 DF,  p-value: < 2.2e-16
```

We see that leaving out the point with the largest cook's distance doesn't really change our model. Which is not a surprise due to our large data size.

Model Selection

Due to heteroscedasticity and nonnormality in the residuals, the new model seems to point out a generalized least square model.

```
> lm2<-gls(V6~V1+V2+V3+V4+V5,corr =corAR1(form=),data=airfoil)
> summary(lm2)
Generalized least squares fit by REML
Model: V6 ~ V1 + V2 + V3 + V4 + V5
Data: airfoil
AIC      BIC    logLik
7510.877 7553.367 -3747.439
```

```
Correlation Structure: AR(1)
Formula: ~1
Parameter estimate(s):
Phi
0.8224366

Coefficients:
Value Std.Error   t-value p-value
(Intercept) 131.46467   1.13029 116.31102  0.0000
V1           -0.00095   0.00003 -30.47885  0.0000
V2           -0.60235   0.09585  -6.28404  0.0000
V3          -40.46753   4.90815  -8.24496  0.0000
V4            0.11330   0.01149   9.86260  0.0000
V5           -5.68705  39.89805  -0.14254  0.8867

Correlation:
(Intr) V1     V2     V3     V4
V1 -0.080
V2 -0.447  0.094
V3 -0.714  0.035  0.437
V4 -0.447 -0.124 -0.165 -0.042
V5  0.149 -0.026 -0.785 -0.286  0.150

Standardized residuals:
Min           Q1           Med           Q3           Max
-3.121147004 -0.640185885  0.007920799  0.697461156  3.345565706

Residual standard error: 5.128825
Degrees of freedom: 1503 total; 1497 residual
```

Using $\alpha = .05$, we see that the predictor for V5 is statistically insignificant. Using backwards elimination, we construct a new model by taking out the insignificant predictor.

```
> lm3<-gls(V6~V1+V2+V3+V4,corr =corAR1(form=),data=airfoil)
> summary(lm3)
Generalized least squares fit by REML
Model: V6 ~ V1 + V2 + V3 + V4
Data: airfoil
AIC      BIC    logLik
7518.108 7555.291 -3752.054

Correlation Structure: AR(1)
Formula: ~1
Parameter estimate(s):
Phi
0.8226537
```

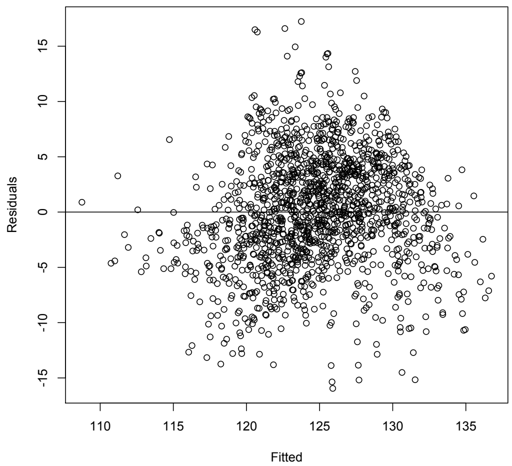
```
Coefficients:
Value Std.Error   t-value p-value
(Intercept) 131.48891  1.118007 117.61014       0
V1           -0.00095  0.000031 -30.50361       0
V2           -0.61301  0.059403 -10.31963       0
V3          -40.67479  4.706700  -8.64189       0
V4            0.11355  0.011354  10.00038       0

Correlation:
(Intr) V1     V2     V3
V1 -0.077
V2 -0.538  0.118
V3 -0.709  0.029  0.358
V4 -0.480 -0.122 -0.076  0.001

Standardized residuals:
Min           Q1          Med           Q3          Max
-3.109047571 -0.637493378  0.009263463  0.701079531  3.360047455

Residual standard error: 5.129987
Degrees of freedom: 1503 total; 1498 residual
```

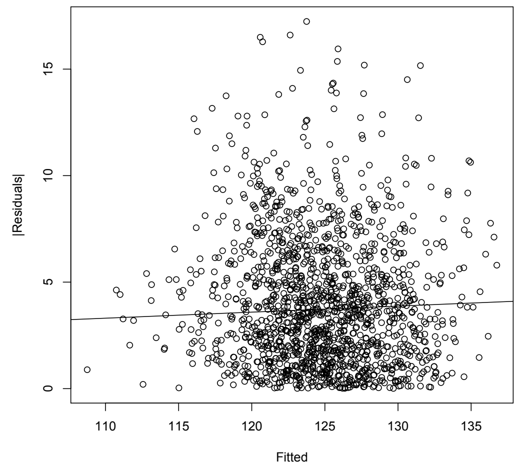
Does taking out the predictor actually makes our model better?

```
> anova(lm2)
Denom. DF: 1497
numDF  F-value p-value
(Intercept)     1 87209.04  <.0001
V1              1   822.31  <.0001
V2              1    48.21  <.0001
V3              1    74.92  <.0001
V4              1    99.95  <.0001
V5              1     0.02  0.8867
> anova(lm2,lm3)
Model df      AIC      BIC    logLik   Test  L.Ratio p-value
lm2     1  8 7510.877 7553.367 -3747.439
lm3     2  7 7518.108 7555.291 -3752.054 1 vs 2 9.230325  0.0024

> rmse(fitted(lm2), airfoil[,6])
[1] 5.058705
> rmse(fitted(lm3), airfoil[,6])
[1] 5.070249
```

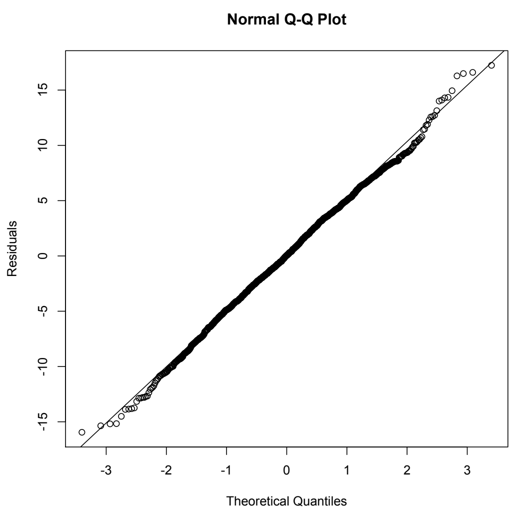
Looking at the log likelihood ratios of the two models, we see that the model with predictor of V5 taking out is a better model judging by the small p-value of 0.0024.

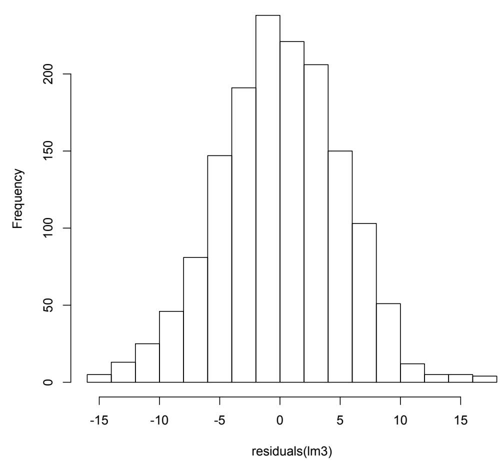
Repeating our previous diagnostics above. Though some of the functions used above does not work for gls models in R.





We see that the signs of mild nonconstant variance has disappeared.






```
> shapiro.test(residuals(lm3))

Shapiro-Wilk normality test

data:  residuals(lm3)
W = 0.9981, p-value = 0.07665
```

The signs of nonnormality has also disappeared.

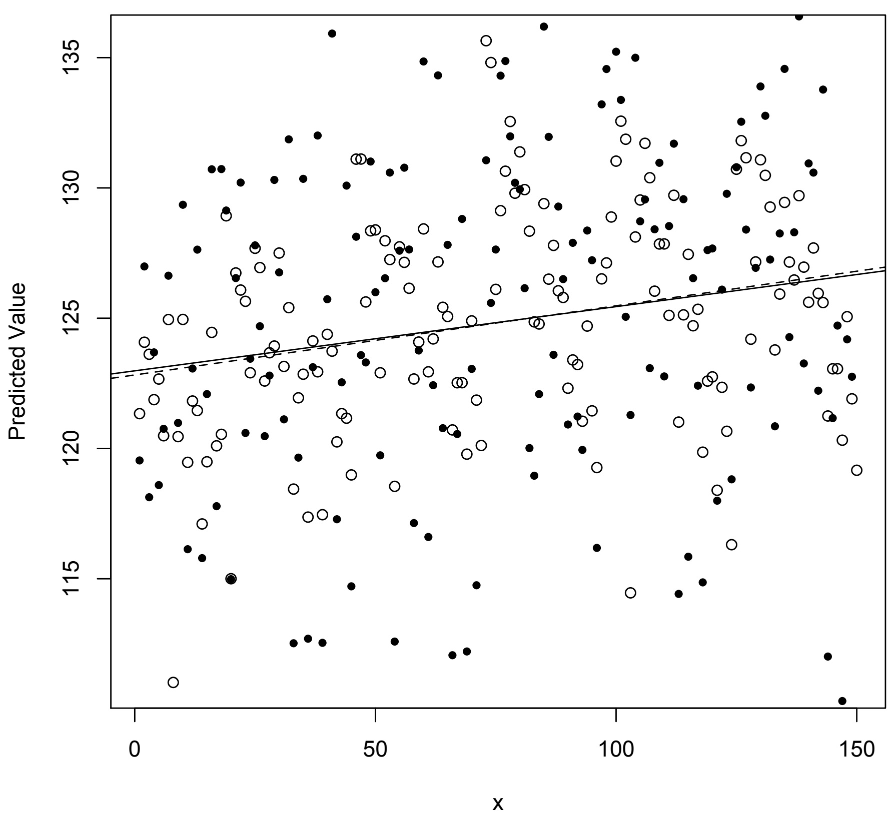
Here are the 95% confidence interval for our estimated predictors. Since 0 is not included in any of the intervals, we can consider our predictors to be significant.

```
> round(confint(lm3),4)
2.5 %   97.5 %
(Intercept) 129.2977 133.6802
V1           -0.0010  -0.0009
V2           -0.7294  -0.4966
V3          -49.8997 -31.4498
V4            0.0913   0.1358
```

Predictions

To test the prediction power of our model, we took out every 10th observation from the original data set and use it as test samples to verify our model.

```
> airfoil1<-Nth.row.delete(airfoil,10)
> test<-Nth.row.get(airfoil,10)
> lm4<-gls(V6~V1+V2+V3+V4,corr =corAR1(form=),data=airfoil1)
```



In the above plot, the empty points represent the predicted values using our model found above. The solid points represent the actual data. The dashed line is the line of best fit for the predicted value. The solid line is the line of best fir for the actual data. We can see the two lines matches very well.

How do we interpret our model?

- With 95% confidence, an increase in one unit in Frequency(V1) will cause a 0.0010 to 0.0009 times decrease in the sound pressure level while holding all other parameters constant.
- With 95% confidence, an increase in one unit in Angle of attack(V2) will cause a 0.7294 to 0.4966 times decrease in sound pressure level while holding all other parameters constant.
- With 95% confidence, an increase in one unit in Chord Length(V3) will cause a 49.8997 to 31.4498 times decrease in sound pressure level while holding all other parameters constant.
- With 95% confidence, an increase in one unit in Free Stream Velocity(V4) will cause a 0.0913 to 0.1358 times increase in sound pressure level while holding all other parameters constant.

Conclusion

From our analysis, we conclude that best way to decrease noise generated by a NACA 0012 airfoil is to increase the chord length. Increase the frequency has almost no effect on sound level. Changing the angle of attack has some minor affect on sound level. Increase air speed will also increase the noise. Suction side displacement thickness doesn't affect the noise generated by the airfoil, this is due to the characteristic of our airfoil being symmetric or 0 camber.

Appendix

```
library(gdata);library(MASS);library(faraway);library(pls);library(psych)

Nth.row.delete<-function(dataframe, n)dataframe[-(seq(n,to=nrow(dataframe),by=n)),]
Nth.row.get<-function(dataframe, n)dataframe[(seq(n,to=nrow(dataframe),by=n)),]
rmse<-function(x, y){sqrt(mean((x-y)^2))}

airfoil<-read.table(file.choose(),header=FALSE)

plot(airfoil[,6],airfoil[,1],ylab="Frequency",xlab="Sound level"
,main="Frequency vs Sound level")
plot(airfoil[,6],airfoil[,2],ylab="Angle of Attack",xlab="Sound level"
,main="Angle of Attack vs Sound level")
plot(airfoil[,6],airfoil[,3],ylab="Chord Length",xlab="Sound level"
,main="Chord Length vs Sound level")
plot(airfoil[,6],airfoil[,4],ylab="Free Stream Velocity",xlab="Sound level"
,main="Free Stream Velocity vs Sound level")
plot(airfoil[,6],airfoil[,5],ylab="Suction side displacement thickness",xlab="Sound level"
,main="Suction side displacement thickness vs Sound Level")

lm1<-lm(V6~.,data=airfoil)
summary(lm1)
plot(fitted(lm1),residuals(lm1),xlab="Fitted",ylab="Residuals")
abline(h=0)
plot(fitted(lm1),abs(residuals(lm1)),xlab="Fitted",ylab="|Residuals|")
a<-summary(lm(abs(residuals(lm1))~fitted(lm1)))
abline(a)
qqnorm(residuals(lm1),ylab="Residuals")
qqline(residuals(lm1))
hist(residuals(lm1),main="")
shapiro.test(residuals(lm1))

dwtest(lm1)

halfnorm(lm1.inf$hat,labs=airfoil1$V6,ylab="Leverages")

m1.inf<-influence(lm1)
lm1.sum<-summary(lm1)
stud<-residuals(lm1)/(lm1.sum$sig*sqrt(1-lm1.inf$hat))
qqnorm(stud)
abline(0,1)

lm1.jack<-rstudent(lm1)
lm1.jack[which.max(abs(lm1.jack))]
qt(.05/(1503*2),1497)
```

```
lm1.cook <- cooks.distance(lm1)
halfnorm(lm1.cook,3,labs=airfoil1$V6,ylab="Cooks Distances")
lm1.cook1<-lm(V6~.,data=airfoil,subset=(lm1.cook<max(lm1.cook)))
summary(lm1.cook1)

lm2<-gls(V6~V1+V2+V3+V4+V5,corr =corAR1(form=),data=airfoil)
lm3<-gls(V6~V1+V2+V3+V4,corr =corAR1(form=),data=airfoil)
anova(lm2,lm3)

plot(fitted(lm3),residuals(lm3),xlab="Fitted",ylab="Residuals")
abline(h=0)
plot(fitted(lm3),abs(residuals(lm3)),xlab="Fitted",ylab="|Residuals|")
a<-summary(lm(abs(residuals(lm1))~fitted(lm1)))
abline(a)

qqnorm(residuals(lm3),ylab="Residuals")
qqline(residuals(lm3))
hist(residuals(lm3),main="")
shapiro.test(residuals(lm3))

dwtest(lm3)

halfnorm(lm1.inf$hat,labs=airfoil1$V6,ylab="Leverages")

lm3.inf<-influence(lm3)
lm3.sum<-summary(lm3)
stud<-residuals(lm3)/(lm3.sum$sig*sqrt(1-lm3.inf$hat))
qqnorm(stud)
abline(0,1)

lm3.jack<-rstudent(lm3)
lm3.jack[which.max(abs(lm3.jack))]
qt(.05/(1503*2),1497)

lm3.cook <- cooks.distance(lm3)
halfnorm(lm3.cook,3,labs=airfoil$V6,ylab="Cooks Distances")
lm3.cook1<-lm(V6~.,data=airfoil,subset=(lm1.cook<max(lm1.cook)))
summary(lm1.cook3)
plot(lm3.inf$coef[,5], ylab="Change in Expend coef")

lm4<-gls(V6~V1+V2+V3+V4,corr =corAR1(form=),data=airfoil1)

x<-c(1:150)
plot(x,predict(lm4,test[,1:5]),ylab="Predicted Value")
```

```
points(x,test[,6],pch=20)
abline(lm(predict(lm4,test)~x),lty=2)
abline(lm(test[,6]~x),type)
abline(lm4)
```