**PSTAT 126**

**Danming Wang T 5-5:50 pm**

**Xilin Huang W 10-10:50 am**

Final Project Analysis

# Regression Analysis of Scaled Sound Pressure Level

## 1. Introduction

This project will concentrate on studying the scaled sound pressure, given the attributes in the dataset “Airfoil Self-Noise” provided by the UC Irvine Machine Learning Repository. This dataset contains 1503 data points. We will determine the outliers in the datasets. Then, we will investigate whether the scaled sound pressure level can be predicted by the possible predictors and the relationship between the scaled sound pressure level and the predictors. Finally, we will try to predict the scaled sound pressure level using our model.

## 2. Questions of Interest

a. Are there any outliers in the dataset?
b. What is the best subset of predictor variables to predict scaled sound pressure level?
c. Predict the scaled sound pressure level when the values of predictors are at average level with an appropriate 95% confidence interval.

## 3. Regression Method

To examine if there are any outliers in the dataset, we will calculate the externally studentized residuals, and the outliers are those whose externally studentized residuals are greater than three. After removing the outliers from the dataset, we then need to build a model that meets all four LINE conditions before answering our questions of interest. We can obtain the model by applying the stepwise regression analysis procedure. And we will check if our model satisfies the LINE condition by applying residual analysis. Finally, we will calculate the mean of each predictor in our model in order to build a 95% confidence interval for scaled sound pressure level when the predictors are at average level.

## 4. Regression Analysis, Results, and Interpretation

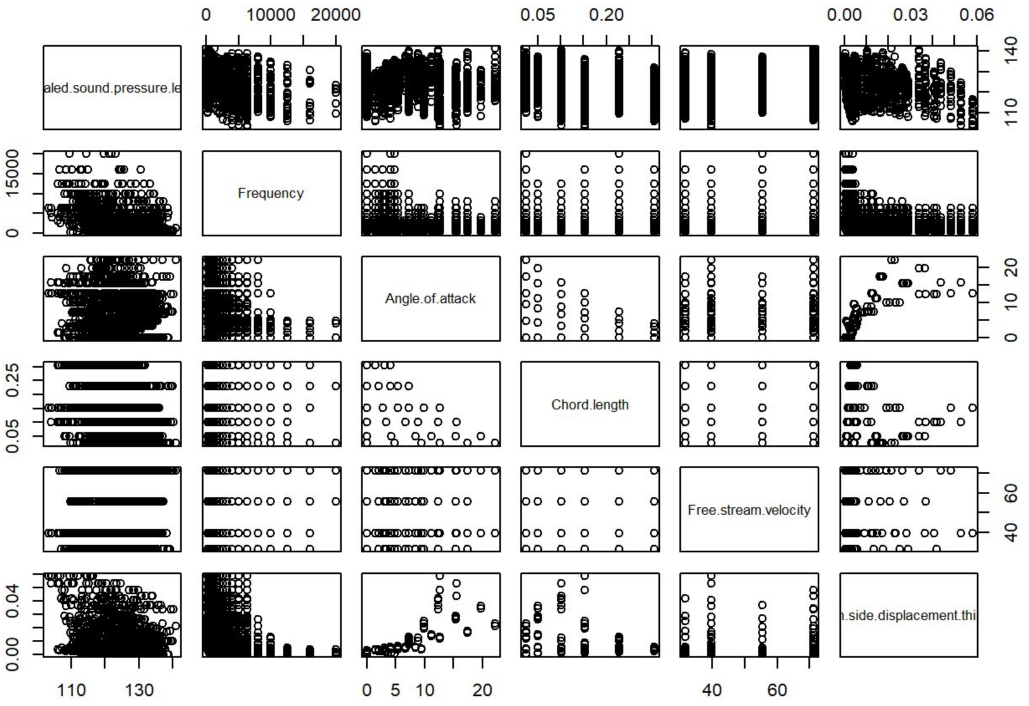
We start our analysis by examining the outliers in our dataset. We obtain the externally studentized residuals and find the observations whose externally studentized residuals are greater than three. Then we delete those observations from our dataset.

Next, we are going to build our model. We define our variables as follows:

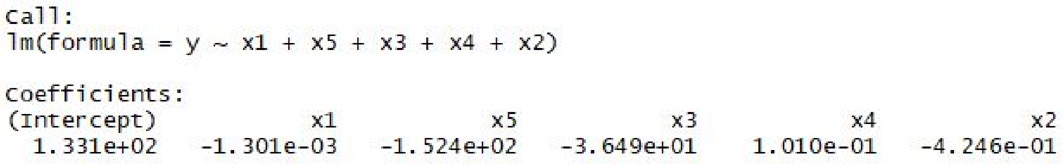
a. Y = Scaled sound pressure level, in decibel.
b. x1 = Frequency, in Hertz
c. x2 = Angle of attack, in Hertz
d. x3 = Chord length, in meter
e. x4 = Free-stream velocity, in meter per second
f. x5 = Suction side displacement thickness, in meters

The Y = Scaled sound pressure level is the response variable with potential predictors x1,x2,...,x5.

To begin with, we plot each potential predictor (x1,...,x5) against the response variable (Scaled sound pressure level) using a pairs() function in R. And the results are as follows:

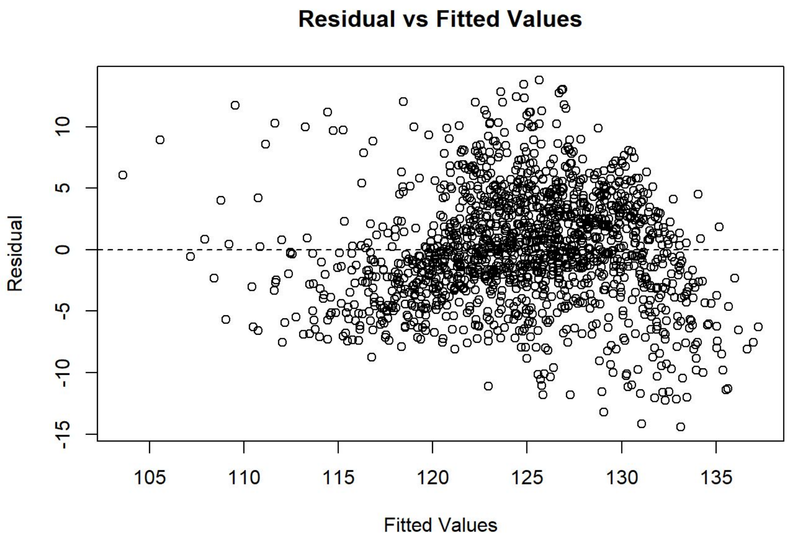


Performing a stepwise regression will help us select variables to predict the values of Y. By following the AIC criterion and using the step() procedure in R, we get that our resulting best fitting model uses frequency, suction side displacement thickness, chord length, free-stream velocity, and angle of attack as predictors.

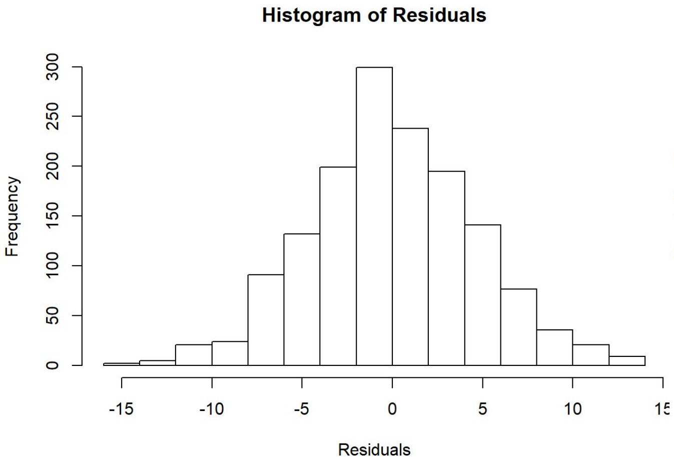


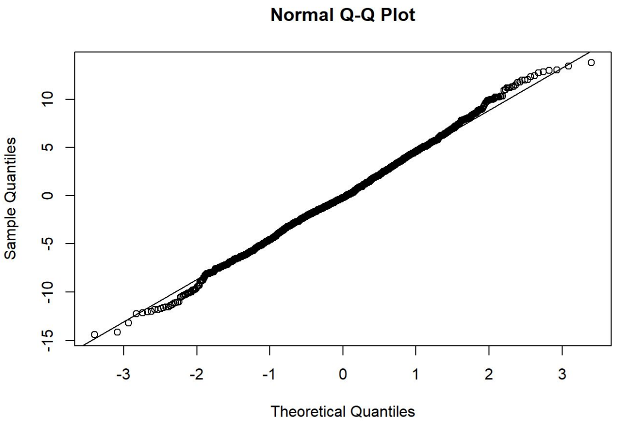
```
Call:
lm(formula = y ~ x1 + x5 + x3 + x4 + x2)

Coefficients:
(Intercept)           x1           x5           x3           x4           x2  
  1.331e+02   -1.301e-03   -1.524e+02   -3.649e+01    1.010e-01   -4.246e-01  
```

Next, we check if our model satisfies the L.I.N.E. assumptions. We first observe the plot of residuals against the fitted values as shown below to see if our model meets the linearity and equal variance assumptions. Since the data points are quite randomly scattered and do not show any pattern, our model meets the linearity and equal variance assumptions.

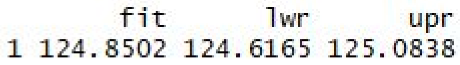


In addition, we look at the histogram of residuals and the normal Q-Q plot to check for normality. Since the histogram of residuals is normally distributed and the relationship in the normal Q-Q plot is approximately linear, our model satisfies the normality assumption.





Finally, we try to obtain the 95% confidence interval of the scaled sound pressure level when frequency, suction side displacement thickness, chord length, free-stream velocity, and angle of attack are at average level. We can see from the output shown below that we are 95% confident that when the frequency, suction side displacement thickness, chord length, free-stream velocity, and angle of attack are at average level, the mean scaled sound pressure level will be between 124.6165 and 125.0838 decibel.


```
       fit      lwr      upr
1 124.8502 124.6165 125.0838
```

## 5. **Conclusion**

After removing the outliers from the dataset, we get a model of scaled sound pressure level containing frequency, suction side displacement thickness, chord length, free-stream velocity, and angle of attack as predictors by using the stepwise regression. The exact model is Scaled sound pressure level = 133.1 - $1.301\times10^{-3}$ frequency - 152.4 suction side displacement thickness - 36.49 chord length + 0.101 free-stream velocity - 0.4246 angle of attack. Using this model, we are 95% confident that when the frequency, suction side displacement thickness, chord length, free-stream velocity, and angle of attack are at average, the mean scaled sound pressure level will be between 124.6165 and 125.0838 decibel. We may improve the model by taking the interaction between the predictors and some higher order terms into account.

## 6. **Appendix**

```
# set a working directory for our project
setwd("D:/UC Santa Barbara/Junior/Spring/Pstat 126/project")

# reading data into R
airfoil_self_noice <- read.csv("airfoil_self_noise.csv")
# check variable names
names(airfoil_self_noice)

# response
y = airfoil_self_noice$Scaled.sound.pressure.level
# predictors
x1 = airfoil_self_noice$Frequency
x2 = airfoil_self_noice$Angle.of.attack
x3 = airfoil_self_noice$Chord.length
x4 = airfoil_self_noice$Free.stream.velocity
x5 = airfoil_self_noice$Suction.side.displacement.thickness

# examine outliers
fit.all = lm(y~x1+x2+x3+x4+x5)
# studentized deleted residuals
rs=rstudent(fit.all)
outlier.pt = which(abs(rs)>3) #outliers
# remove outliers
y=y[-outlier.pt]
x1=x1[-outlier.pt]
x2=x2[-outlier.pt]
```

```
x3=x3[-outlier.pt]
x4=x4[-outlier.pt]
x5=x5[-outlier.pt]

#running a scatter plot matrix for all variables to test preliminary association
pairs(airfoil_self_noice[-outlier.pt])

# Running stepwise regression
mod0 = lm(y~1)
mod.upper = lm(y~x1+x2+x3+x4+x5)
step(mod0,scope=list(lower=mod0,upper=mod.upper))
# The resulting model is one using frequency, angle of attack, chord length, Free stream velocity, and Suction.side.displacement.thickness as the best predictors of scaled sound pressure level

# our model
mod = lm(y~x1+x2+x3+x4+x5)
# verify LINE conditions
e = residuals(mod)
yhat = fitted(mod)
plot(yhat, e, xlab = 'Fitted Values', ylab = 'Residual', main = 'Residual vs Fitted Values')
abline(h = 0, lty = 2)
hist(e, xlab = 'Residuals', main = 'Histogram of Residuals')
#Q-Q plot
qqnorm(e)
qqline(e)

# find the confidence interval
x <- model.matrix(mod)
x0 <- apply(x,2,mean) # mean amount
ci = predict(mod, new=data.frame(t(x0)), interval = "confidence", level = 0.95)
ci
```