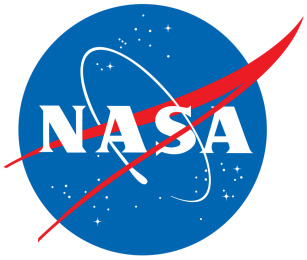
Airfoil Self-Noise Regression Analysis

March 2019

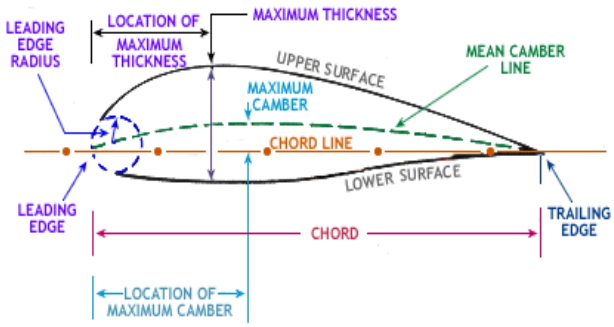


Joheun Kang (Friday 9am)
Catherine Miao (Friday 9am)
Katie Schmitzer (Friday 10am)

Abstract

We propose a regression model to represent airfoil self-noise while pursuing a rudimentary understanding of sound pressure mechanisms. In particular, we are interested in determining what the most influential factors of airfoil self-noise are and how it would best be minimized. Once we obtain a model that represents our data, these tasks are attainable. Using a dataset recorded by the NASA Langley Research Center, we identify five main elements of the self-noise mechanism that are important to our model. In addition, we find a strong negative correlation between frequency of noise, angle of attack, airfoil chord length, the natural logarithm of the suction-side displacement thickness and the scaled sound pressure level, as well as a strong positive correlation between the free-stream velocity and the scaled sound pressure level.

Introduction



Airfoil self-noise is the total noise produced when an airfoil encounters smooth non-turbulent inflow (Brooks 2). The noise is a result of the interaction between the airfoil blade and the turbulence produced by its boundary layer (Brooks 2). Our data, recorded by the NASA Langley Research Center, was the result of a series of aerodynamic and acoustic tests with two and three-dimensional airfoil blade sections conducted in an anechoic wind tunnel. This data set, comprised of 1503 observations, takes into account specified sizes of NACA 0012 airfoils at specified velocities and angles of attack (UCI Machine Learning Repository).

In this analysis, we propose a linear regression model for the dataset to develop a fundamental understanding of self-noise mechanisms. We identified chord length, frequency of noise, angle of attack, free-stream velocity, and suction-side displacement thickness as all being useful predictors of airfoil-self noise. Furthermore, we are interested in determining how each of these factors affects the resultant sound pressure level. Our regression model provides meaningful insight into decreasing airfoil self-noise for a variety of aerodynamic systems. For example, studying the self-noise of NACA 0012 airfoils will help us better understand the airfoil self-noise of helicopter rotors, high-lift devices, and wind turbines (Brooks 165).

Research Questions of Interest

Question 1: How much will airfoil self-noise change if frequency is increased by 1000 Hertz, and all other factors are held constant?
Hypothesis 1: We predict that as the level of frequency increases, the self-noise will also increase, after controlling for all other predictors.

Question 2: What is the predicted sound pressure level when frequency is 1600 Hertz, angle of attack is 6.778 degrees, chord length is 0.3408 meters, free-stream velocity is 71.3 meters per second, and suction side displacement thickness is 0.0153119 meters?

Hypothesis 2: After scrutinizing our dataset, we estimate that the predicted sound pressure for the independent variable values above is around 125 Hertz.

Data and Regression Methods

Our data was recorded at the NASA Langley Research Center during a series of aerodynamic and acoustic tests of airfoil blade sections. The dataset contains 6 variables: sound pressure level in decibels (**Sound**), frequency in Hertzs (**Frequency**), angle of attack in degrees (**Angle**), chord length in meters (**Chord**), free-stream velocity in meters-per-second (**Velocity**), and suction side displacement thickness in meters (**Displacement**). In our analysis, we treat sound pressure level (**Sound**) as our response variable, and the remaining five variables as our predictor variables.

We employ regression analysis to answer our research questions of interest. Firstly, we seek out an optimal model for our dataset. Once we have found our optimal model, we will be able to unravel the impact of **Frequency** on **Sound**, while holding all other predictors constant. Additionally, we can utilize our model to predict **Sound** at any given values of the independent variables, as well as construct the prediction interval.

All of the variables in the airfoil self-noise dataset are categorical, as the pairwise scatterplots are highly discrete. Since **Chord** and **Velocity** have 6 and 4 levels respectively, we treated them as categorical data points and used the as.factor() function to separate the values of each level. The other three predictors, **Frequency**, **Angle**, and **Displacement**, have numerous levels, so we treated them as if they were continuous. By looking at the pairs() plot, we find it difficult to tell if any of the predictors have strong correlation with one another, given their categorical nature. We do notice, in *Figure 1* however, that angle of attack and displacement have a seemingly quadratic relationship. This seems plausible when thinking about simple physics: displacement is quadratically related to the angle of release, while all other variables (**Chord**, **Frequency**, **Velocity**) are held constant.

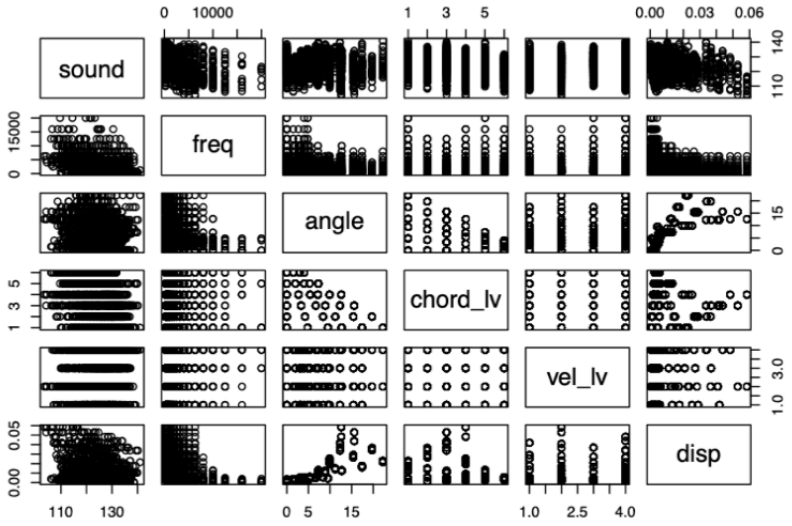


Figure 1: Pairwise Scatterplots

Next, we used the avPlots() function to further explore the nature of association between the response variable **Sound** and each individual predictor after controlling all other predictors. From the added variable plots, we can

see there is a strong negative relationship between **Sound** and **Frequency**. This finding is contradictory to our initial hypothesis that sound pressure and frequency are positively correlated. The relationship between **Sound** and **Angle** is slightly negative. As we factored out the values for each chord level and each velocity level, there is one separate added-variable plot for each chord and velocity value, except the baseline value. From the added-variable plots, we identified a negative correlation between **Sound** and **Chord**. In specific, as the chord level increases, the negative relationship becomes stronger. We also found that there is a relatively positive relationship between **Sound** and **Velocity** levels. Furthermore, the relationship becomes stronger as the velocity level increases. Lastly, the relationship between **Sound** and **Displacement** is slightly negative. However, none of the relationships above appear to be linear. It is our belief that certain transformations should be applied to the predictors and/or the response variable in order to meet the conditions necessary for linear regression.

Regression Analysis, Results, and Interpretation

Detailed Analysis and Diagnostic Checks

In order to determine which of the five predictor variables to include in the regression model, we conducted a variety of variable selection tests. First, we chose to use forward, backward, and stepwise selection methods on both AIC and BIC. We received the same result from all six tests: all five predictors are important to keep in our model. It is reasonable that this dataset contains only influential predictors. As drag, air-resistance, and aerodynamics have all been thoroughly studied, researchers are quite familiar with the significant factors of aerodynamic ability. Our interest, therefore, lies in how exactly each of these factors affects noise and aerodynamics.

To substantiate our belief that all the predictors should be incorporated in the model, we applied the regsubset() function and examined coefficient of determination R^2, Adjusted R^2, Mallows' C_p statistic, and BIC. The benefit of using this function, is the ability to see how each of these selection criteria varies for the model, at each level, when an additional predictor is added. In particular, the Mallows' C_p statistic has the greatest influence on our decision to include all five predictors. The Mallows' C_p statistic for the five predictors are as follows: 1135.33, 605.31, 233.87, 57.69, 6.00. There is a notable decrease in each two adjacent statistics. Even though the drop between the last two figures are not as considerable compared to the other pairs, we do believe that we should keep the last variable **Displacement** (from the "which" column) in our model. Our reasoning is that a Mallows' C_p statistic of 57.69 is still far off from the rule of p+1, which equals 6 for our dataset.

After determining the influential predictors, we move on to construct our model. First of all, we regress our response variable on all five predictors and obtain the summary statistics, residual-fits plots, and the normal Q-Q plots. The summary table shows that all predictors except one (out of 6) level of velocity are significant at the 0.1% level. This result again supports our decision to include all predictors in the model. Examining the residual - fits plot and the normal Q-Q plot, we discovered strong violations of linear assumptions. The residual - fits plot displays great concentration, conspicuous patterns, and a lack of well-scattered points around the 0 line, indicating that the linearity assumption is not satisfied. In addition, the plot has a fanning shape, which suggests that the equal variance assumption is violated. By looking at the normal Q-Q plot, we found that most data

points lie closely on the Q-Q line, but the plot appears slightly heavy-tailed, suggesting that the normality assumption is also somewhat violated.

Since there are strong violations of linear regression assumptions, we are convinced that transformations on the predictor variables and/or the response variable are necessary. To address the violations, we start with transformation on the predictors. Before doing so, we scrutinized the dataset and found multiple zero values for the predictor **Angle**. As the great number of zero values create obstacles for transformation, we decided on adding "noise" to the **Angle** variable to get rid of the zeros. After we tried adding various numbers without shifting the plots, we decided to add the median value (5.4) of the variable **Angle** to the original variable. From there, we were able to carry out transformations on the predictors. We then applied the powerTransform() function to our three non-categorical predictors and obtained the accurate lambas of 0.0169, 0.130, and 0.0483, respectively. As all three powers are close to zero, we decided to apply the log transformation on Frequency, Angle, and Displacement. Afterwards, we examined the residual-fits plot and found out that the fanning pattern was even stronger and the normal Q-Q plot was more off the Q-Q line. Therefore, we conclude that placing log transformations on all three variables are not appropriate. In order to determine which of the three predictors should be log-transformed, we conducted a Likelihood Ratio test on each value of lambda. From the result, we conclude that either one or two predictors (from **Frequency, Angle, Displacement**) should be log-transformed.

Next, we attempted the log transformations on all three variables individually as well as pairwisely, and reached the conclusion that log transforming the variable **Displacement** gives us the most satisfactory residual-fits plot, normal Q-Q Plot, and adjusted R-squared. Hence, we decided to apply the log transformation solely on **Displacement**. Below, in *Figure 2*, you can see how our residuals-fits plot is still patterned after transforming predictors.

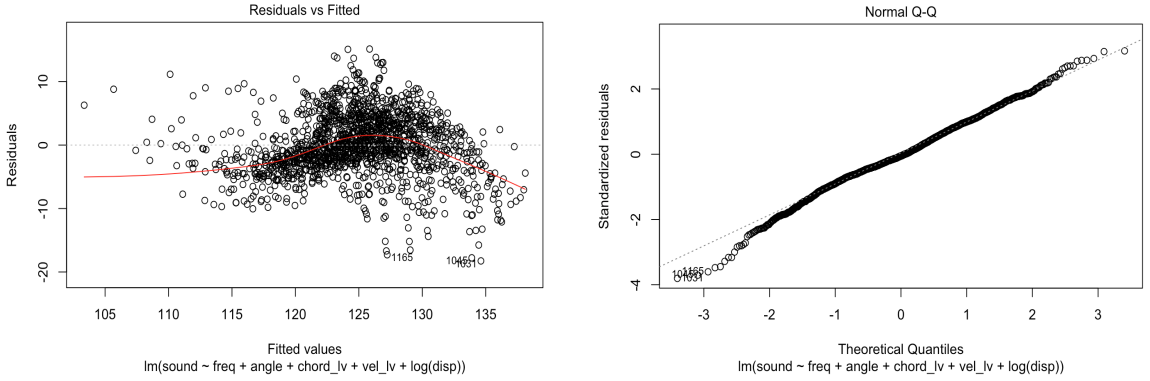


Figure 2: Residuals vs Fitted plot and Normal Q-Q plot for the transformed model

Afterwards, we conducted diagnostic checks and realized that the violations of linear assumptions are still not remedied. Then we examined the BoxCox transformation on the response **Sound**. The lambda value that we retrieved from the BoxCox transformation is approximately 2. We applied this transformation to the variable **Sound**, however, a quadratic response only complicates the model without resolving any of the linear violations. Thus, we decided to stick with our existing model.

As all five predictors are strongly associated with the sound pressure, we hypothesized that some pairs of predictors, when interacting with one another, may effectively influence the response and resolve some of our violation issues. Therefore, we tried adding possible interaction terms to arrive at a potentially better model. Through careful exploration, we discovered that the most important interaction term of all - the interaction between Chord and log(Displacement), i.e. **Chord*log(Displacement)** only increases the *Adjusted* R^2 by about 1%, and it hardly improves the residual-fits plot or normal Q-Q plot. Evaluating the tradeoff between goodness of fit and model complexity, we decided to not include any interaction terms in our model.

Our next step was to inspect whether potential influential points play a significant role in determining the regression line. We used the infuenceIndexPlot() function to examine the Cook's distance plots and found that observation 1030 and 1217 had the greatest Cook's distances, suggesting that these two points might be influential to the regression. Using intuition, we believed that with a dataset containing over 1500 observations, the influence of individual observations is trivial. However, removing observations 1030 and 1217 from our dataset and refitting the regression model, we noticed that the significant code for **Angle** decreased from 0.01 to 0.001. In contrast to our original thought, we discovered that these two points were indeed "influential" and, thus, we determined that it was crucial to exclude them from our dataset.

Up until this point, the greatest issue confronting our model was non-constant variance, which is a violation of one of the linear regression assumptions. We believed using a Weighted Least Squares (WLS) model would be the most practical approach to address this issue. We attempted to use fitted values and different variables as "weight" to adjust our model. Finally, we identified the predictor **Frequency** as the most effective weight parameter. Although applying this weight parameter increased the p-value of the non-constant variance test from 1.20e-07 to 0.02 (improving our constant variance violation), this model is still insufficient in resolving all of our violations. Specifically, as seen in *Figure 3*, this weighted model made our residuals better-scattered, but it severely and negatively affected the normality in the Q-Q plot. Moreover, the weighted regression model largely shrinks the goodness of fit (adjusted R-squared). Hence, the cons of employing a complicated model using Weighted Least Squares in our model overweights its pros.

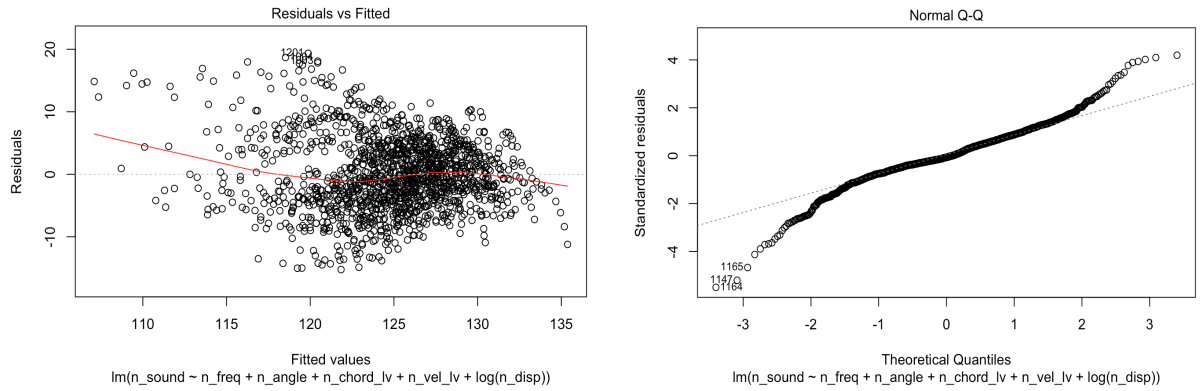


Figure 3: Residuals vs Fitted plot and Normal Q-Q plot for the WLS model

Therefore, we determined that the optimal regression model based on our analysis is:

$$Sound_i = \beta_0 + \beta_1 freq_i + \beta_2 angle_i + \beta_3 chord\_lv_i + \beta_4 vel\_lv_i + \beta_5 log(disp_i) + \varepsilon_i$$

It is important to notice that this model contains 11 predictors (since chord_lv and vel_lv are categorical variables with 6 and 4 levels, respectively) and provides an *Adjusted* R^2 of 0.5158.

Interpretation

With the above model, we were able to answer our research questions of interest:

Question 1: As the estimated slope for the **Frequency** predictor is -1.349e-03, we know that a 1000 Hertz increase in frequency results in a 1.349 decibels decrease in sound pressure level, assuming all other predictors are held constant.

Question 2: With the given data values, we obtained a fitted value of 122.18 and a 95% prediction interval of (112.74, 131.63). This is our estimate of predicted sound pressure level when frequency is 1600 Hertz, angle of attack is 6.778 degrees, chord length is 0.3408 meters, free-stream velocity is 71.3 meters per second, and suction side displacement thickness is 0.0153119 meters is 122.1829 decibels. This means that we are 95% confident that the predicted sound pressure is within an interval between 112.74 decibels and 131.63 decibels at these given conditions.

Conclusion

Analyzing this dataset, we determine that when frequency is increased, airfoil self-noise is decreased, assuming all other predictors are held constant. This same phenomenon occurs when the angle of attack, chord length, and suction side displacement thickness are increased separately. Conversely, when free-stream velocity is increased, the airfoil self-noise is also increased as all other predictors are held constant. This finding is instrumental to self-noise prediction and minimization for a variety of aerodynamic systems in real-life applications. While seeking the optimal regression model for the dataset, we consistently weighed the pros and cons between making our model more linear and making our model overly complicated. We finally landed on a model that struck a balance between the complexity of the model and the goodness of fit. From our analysis, we perceive several violations of the linear regression assumptions as irreconcilable within the scope of our study. In trying many methods to fix non-constant variances, the normality was decremented, and vice versa. In addition, we realize that the five major attributes in our dataset cannot unravel the self-noise mechanism to its entirety. Therefore, it is our belief that applying more advanced statistical methodology as well as incorporating additional critical predictors to the regression model are the next logical steps for a more in-depth research and analysis.

References

Airfoil. (2019, January 31). Retrieved from https://en.wikipedia.org/wiki/Airfoil

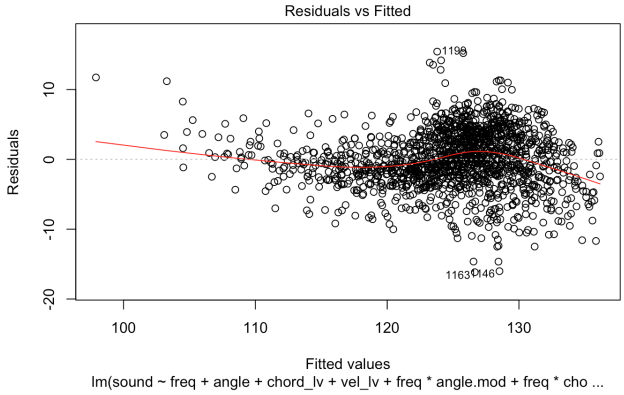
Brooks, Thomas F., D. Stuart Pope, and Michael A. Marcolini. "Airfoil self-noise and prediction." (1989).

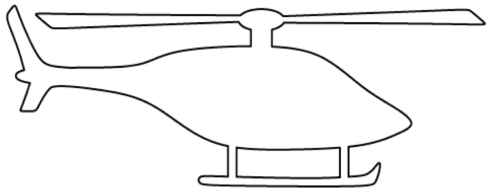
Appendices

Appendix I - Interesting Findings

While testing to see if any of our interaction terms were significant, we came across a model whose residuals-fits plot resembled the shape of a helicopter. The model that produced this plot was:

$$Sound_i = \beta_0 + \beta_1 freq_i + \beta_2 angle_i + \beta_3 chord\_lv_i + \beta_4 vel\_lv_i + \beta_5 log(disp_i) + \beta_{12} freq_i * angle_i + \beta_{15} freq_i * log(disp_i) + \varepsilon_i$$





Appendix II - R Code