# A demonstration of the Airfoil Self Noise for Applied Statistical Methods MAT 565 University at Albany

Gabrielle Giacoppe

24, November 2021.

Statistical Methods learned in MAT 565 applied to the Airfoil Self Noise Dataset.

## Introduction

An airfoil is the cross-sectional shape of a wing on aircraft, that once in motion, produces lift based on its characteristics (Brooks et al.,1989). The sound of an airfoil is dependent upon several predictor variables. Statistical methods can be implemented accordingly.

### *Why it is important to do this statistical review?*

Airfoils positioned at greater angles of attack, have low frequencies placing emphasis on the amount noise, bringing correlations of higher magnitude to surface (Brooks et al.,1989).

### Data Extraction

The NASA data set in this statistical review contains experimental framework, testing both aerodynamic and acoustics of two and three-dimensional airfoil variability in the size NACA 0012 airfoils (n0012-il). In which, the span of the airfoil and observer position were the same in all experiments to retrieve metrics at various wind tunnel speeds and angles of attack (Kaggle). The data was retrieved from Kaggle which was obtained from UCI Machine Learning Repository.

The data set was then imported into the software (RStudio, 2021) and relevant code is implemented in RStudio markdown file, attached in a file folder to this review.

*Description of the variables*

1. f: Frequency in Hertzs, Hz.
2. alpha: Angle of attack (AoA, alpha), in degrees °.

   **Note:** A potential issue has been identified as 360 degrees = 0 degrees. The range of alpha is taken with cautious consideration.

   Hence, in the event where alpha gets close to 360, alpha will be transformed.
3. c: Chord length, in meters, m.

4. U_infinity: Free-stream velocity, in meters per second m/s.
5. delta: Suction side displacement thickness, in meters, m.

All 5 of them are numeric, are serving as explanatory variables to predict

the response y, SSPL: Scaled sound pressure level, in decibels, dB.

*Step 1: Analysis of the Pairs Scatterplots*

First, I used the pairs methods to visualize the scatterplots with all 5 predictors for SSPL. It is premeditated for relations to occur between each predictor and SSPL. However, Figure 1 shows f has a correlation with SSPL of 0.39, delta has a correlation with SSPL of 0.31, and so forth. The greatest correlation is seen between SSPL and alpha, this correlation equals 0.75. The last column of plots on the right has the y variable on the x axis, indicating there is a very weak relationship with SSPS.

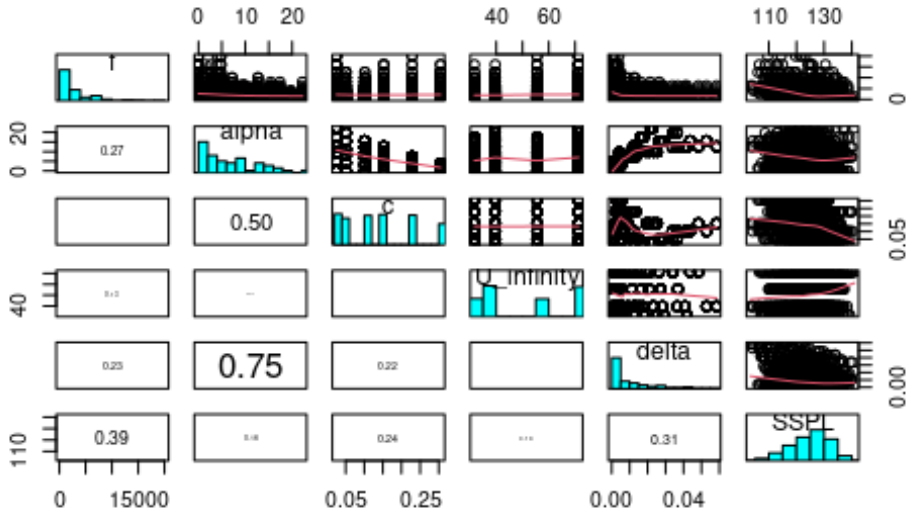



*Figure:1*

## Model Selection

*Step 2: Preliminary predictive modelling*

**Linear Regression**

Second, to provide a check of assumptions of linearity in regression, I used a multivariate regression model in (1). $R^2 = 0.5157$, Adj. $R^2 = 0.5141$, F-statistic = 318.8 on 5 and 1497 df, and p-value: < 2.2e-16.

$$SSPL = Beta0 + (f) + (alpha) + (c) + (U\_infinity) + (delta) \quad (1)$$

A linear model will not run efficiently in this case due to Figure: 1 displaying a curved fit in the plots, instead of a linear relation, as well as the multiple coefficient of determination, $R^2$ not being optimally close to 1.0.

**Polynomial**

Third, is the check of assumptions in regression fitting a polynomial; this requires the transformation of the response variable into a polynomial and is justified by looking at the pairs scatter plots of SSPL vs numerical variables. I considered changing the base model by including polynomial terms for the numerical variables. Starting with polynomials of degree 5, $R^2$ =0.5732, Adj. $R^2$ =0.5666, F-statistic = 86.38 on 23 and 1479 df with p-value: < 2.2e-16.

While interactions are usually performed with categorical variables, given the additive nature of the predictors x and y relationship, I considered the additive effect of the variable c: Chord length. Chord length in an airfoil is the measurement between its leading edge and the trailing edge. This interaction term was decided to have the most influence, because the polynomial base model changed by including polynomial interaction term of c on the other variables: $R^2$ =0.596, Adj. $R^2$ =0.5928, F-statistic = 183.2 on 12 and 1490 df with p-value: < 2.2e-16.

**Logarithm Polynomial**

Based on the polynomial model's increase in the Multiple coefficient of determination, $R^2$ , it is advised to do a logarithm transformation. Looking back, in reference to Figure 1, the fourth column of scatter plots there is an emphasis on a logarithmic shape. On the quintic polynomial, with c: Chord length, as the interaction induced on the remaining terms the output is: $R^2$ =0.6317, Adj. $R^2$ =0.6267, F-statistic = 127.1 on 20 and 1482 df with p-value: < 2.2e-16. This Adj. $R^2$ shows there is more power to using a polynomial log model rather than just a polynomial model without a transformation.

In the extent of this study, I have decided to limit the polynomial's degree intentionally. There is evidence that increasing the degree of the polynomial while fitting the model on a logarithmic transformation will increase the coefficient of determination.

## Model Assessment

### Examining Normality

The Normal QQ plot displays standardized residuals that are linear towards the middle of the quantile, but there is note of the curving the data is doing in the lower quantile and curving on the upper portion of the highest quantiles (Figure 2). The points of interest are located on the bottom left of the graph (1148 and 1165).

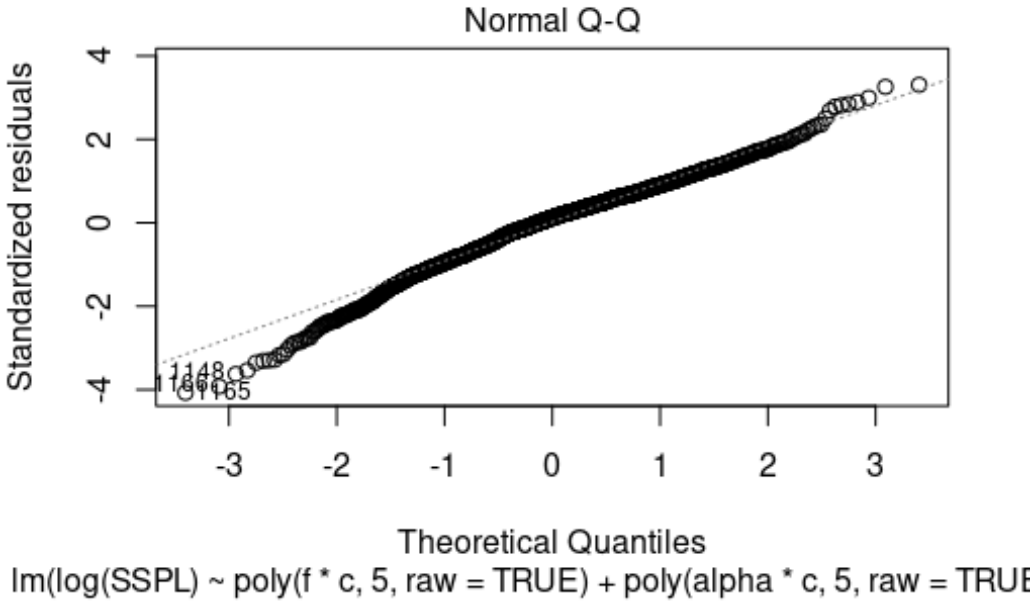



*Figure:2*

To test if normality is satisfied, The Shapiro-Wilk test proposed in 1965, is run to validate whether data in the sample comes from a normal distribution. Here, W =0.98222, p-value = 1.128e-12. It is commonly addressed that small values of W give evidence of departure from normality and the correlation of the percentiles of the normal vs the percentiles of the data are present as uncorrelated; In this case, W is very close to 1.0 and thus, there is high correlation. W alone is not enough to gather this test. The p-value is so small at a significance level less than 0.01, meaning we reject the null and we are to accept the alternative hypothesis H1, it does not come from normal distribution.

The plot of Residuals vs. Fitted Values, it can be inferred that the residuals start to thicken at the fitted value 4.81 (Figure 3) with the same outlying observation points of interest recalled in the Normal Q-Q plot.

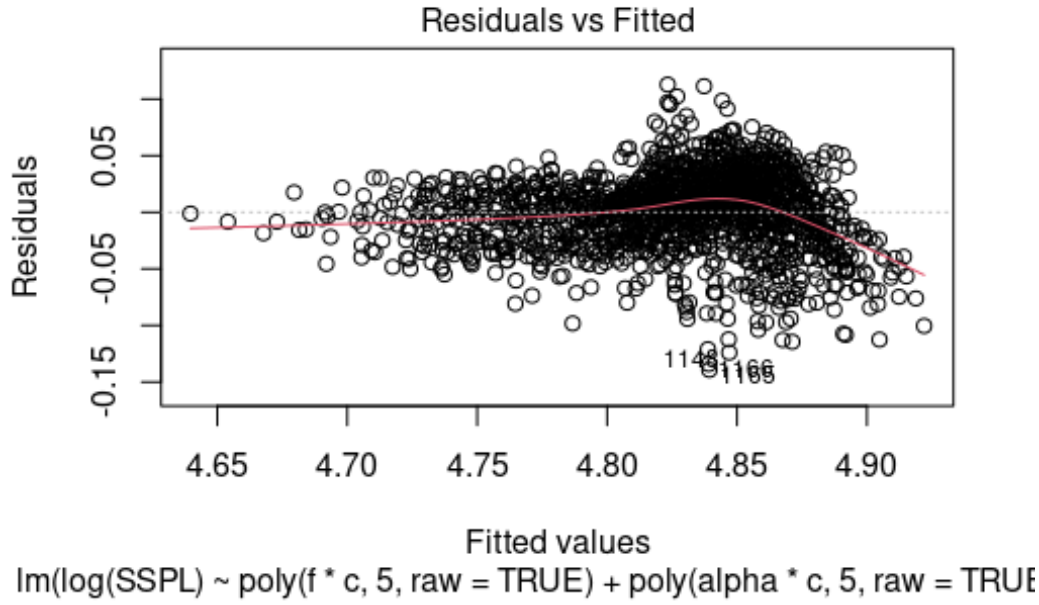



*Figure:3*

The observation point (1165) is the highest in residual leverage and is at a standardized residuals value of 2.0 (Figure 4).

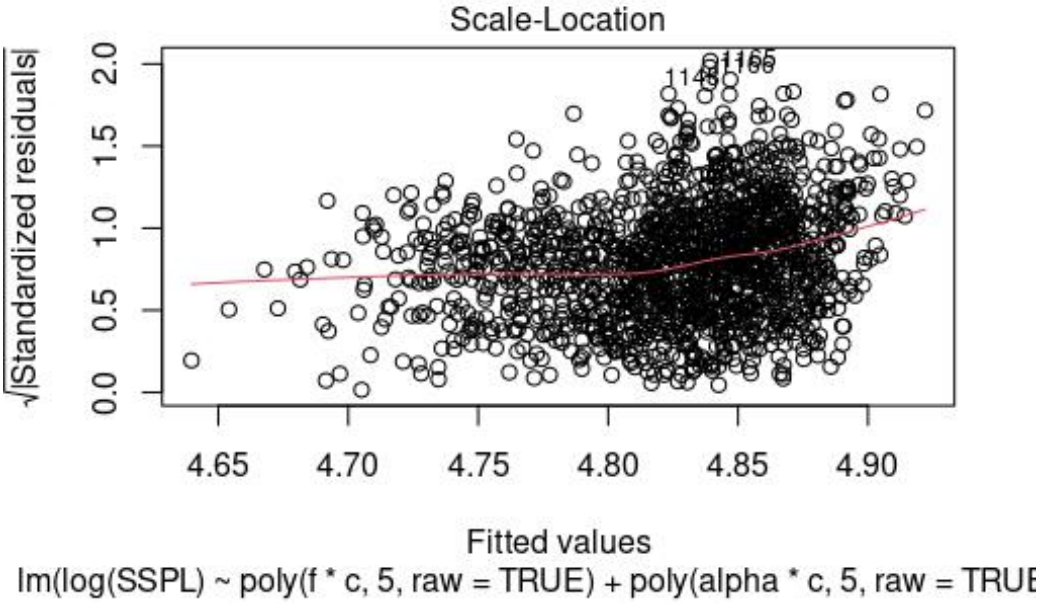



*Figure: 4*

## Examining Outlying data

## Cook's Distance

The observation point with Cook's distance greater than 1 is (201). (201, 707, 80) are points classified as being influential. When Cook's distance is plotted against leverage, (80) lies in the upper right, and has the highest leverage (Figure 5).

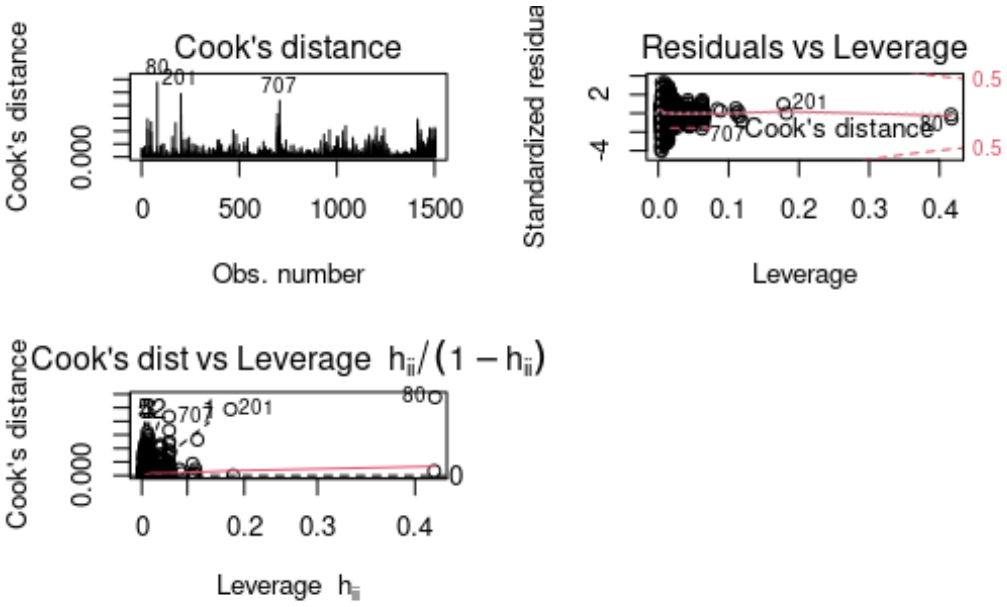
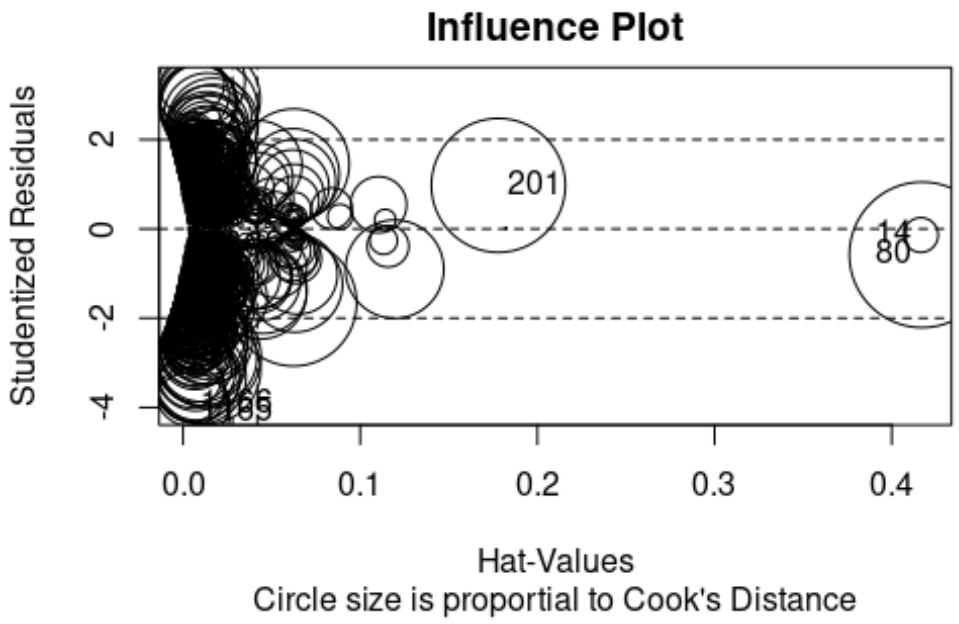



*Figure:5*

**Influence Plot**

To further analyze Cook's distance findings, (Figure 5) demonstrates the observation point (80, and 201) are largest in proportion to Cook's distance.




## Cross Validation

Since there is the potential to choose from two or more possible models, previously it was stated to look at adjusted R-squared values. While there are other measurements to find the best model such as Mallow's Cp, Akaike Information Criterion (AIC), and PRESSp Prediction Sums of Squares criterion. The reason why these measurements are of mention is for their quality:

$$\text{Mallow's Cp} = (\text{SSEp}/\text{MSEfull}) - n + 2(p + 1) \qquad (3)$$

An optimal answer to (3) will be a number closest to the number of independent variables in the model. AIC/n is an unbiased estimate of the expected log probability can derive a new data point (Reinhold, 2021). PRESS is predictive assisting in the prevention of over-fitting because it is calculated using ith observations, not included in the model estimation (Reinhold, 2021).

Next, there are methods listed below to ensure that these measurements can be retrieved. The null model is one with none of the variables contrasting the full model with all the variables.

*Methods*

I proceeded with backwards selection method, following "both" directions were ran, starting with the full model, it can be compared to the backsel model; both methods produced the same model.

Obtaining best subsets models, by randomly dividing the data into 2 parts, testing and training, I was able to create the models for each selected best subset. The model that had closest Cp to variables was cpm. The model that won is the one with lowest AIC which was

bsel. The model with better predictive power under PRESS was bsel since it had the least value for PRESS and the largest adjusted R-squared (Table 1).

| | p+1 | R2adj | Cp | AIC | PRESS |
|---|---|---|---|---|---|
| bsel | 8 | 0.7563 | 5.87 | -2310.81 | 0.1973157 |
| bicm | 6 | 0.7526 | 8.54 | -2307.99 | 0.1989558 |
| aicm | 7 | 0.7543 | 7.39 | -2309.19 | 0.1984300 |
| adjr2m | 8 | 0.7549 | 7.61 | -2309.02 | 0.1978929 |
| cpm | 9 | 0.7544 | 9.21 | -2307.42 | 0.1992304 |

*Table: 1*

### Strengths & Weakness in Results

There are limitations in this review involving Cross Validation. Cross Validation should've been a primary step before model selection. However, this was not possible to yield result, being there ended up being more data for testing than training, when it is favorable to have the training set to have at least 6 times as many observations than predictors in the testing set. The cause of this error was random generator, being that the line of code used in the cvindex as the variables was not classifying the data into testing and training. There was also difficulty in setting the appropriate random seed as dividing into 80-20 was not outputting proper Cp ranges.

Strength demonstrated in this review was recognition that the polynomial model was not the overall preferable fit on the NASA data, it happened that a transformation to the log model was most acceptable. Despite the CV setback, the adjusted R-squared value after doing CV had reached its highest with respect to all model discussions.

### Future Directions

The timeline of this paper was mid to end of a 14-week semester; there being, I would like to emphasize how important it is to fix any possible discrepancies in asserting a better train test split method. It is most likely that the results will become different. I would also like to continue in the application of more statistical methods that will bring the model beyond the results stated.

## References

Brooks, T.F., Pope, D.S., Marcolini, and A.M. (1989). Airfoil self-noise and prediction. Technical report, NASA RP-1218. https://ntrs.nasa.gov/api/citations/19890016302/downloads/19890016302.pdf

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository http://archive.ics.uci.edu/ml%5D. Irvine, CA: University of California, School of Information and Computer Science.

Fedesoriano. (2021). NASA Airfoil Self-Noise Dataset. A series of aerodynamic and acoustic

tests of two and three-dimensional airfoils. https://www.kaggle.com/fedesoriano/airfoil-selfnoise-dataset

Reinhold-L, Karin. (2021). Course Lecture Notes from MAT 565. University at Albany. Blackboard.

RStudio Team (2021). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/