Project Report: Airfoil self-noise data analysis

Aditya Deshkar(CH23B006)
Ashwajit Tayade(CH23B008)

Dataset:- https://archive.ics.uci.edu/dataset/291/airfoil+self+noise

Research Questions

The study investigates the **factors influencing airfoil self-noise** using an experimental dataset derived from wind tunnel tests. The main research questions are:

1. Which features significantly affect the **Scaled Sound Pressure Level (SPL)** of an airfoil?
2. Can **multiple linear regression (MLR)** accurately predict SPL using aerodynamic and geometric parameters?
3. Do **feature interactions** and **non-linear effects** improve predictive performance compared to linear models?
4. Are there **dependencies (multicollinearity)** among predictors that could affect model stability?

Significance of the Problem

Understanding airfoil noise is crucial for designing quieter and more efficient aircraft. Aerodynamic noise contributes significantly to environmental sound pollution, particularly near airports. By modeling SPL as a function of key airfoil and flow parameters, this study provides statistical insights that can help in optimizing airfoil geometry and operating conditions to minimize noise.

Data Description

The dataset comprises **1,503 experimental observations** collected in a **wind tunnel** under controlled conditions, making it an **experimental dataset** rather than an observational one

Features

1. **Frequency (Hz):** Number of oscillations of airflow over the airfoil per second.
2. **Angle of Attack (°):** Angle between the oncoming flow and the chord line.
3. **Chord Length (m):** Distance from the leading edge to the trailing edge.
4. **Free-Stream Velocity (m/s):** Velocity of undisturbed airflow upstream of the airfoil.
5. **Suction Side Displacement Thickness (m):** Thickness of the boundary layer on the suction side, affecting flow separation.

Response Variable

- **Scaled Sound Pressure Level (SPL)** measured in decibels (dB), representing the acoustic energy radiated from the airfoil surface.

Preliminary Studies

Descriptive Statistics:

```
   Frequency      Angle_of_attack   Chord_length     Free_stream_velocity Suction_side_displacement_thickness Scaled_sound_pressure_level
 Min.   :  200    Min.   : 0.000   Min.   :0.0254   Min.   :31.70        Min.   :0.0004007                   Min.   :103.4
 1st Qu.:  800    1st Qu.: 2.000   1st Qu.:0.0508   1st Qu.:39.60        1st Qu.:0.0025351                   1st Qu.:120.2
 Median : 1600    Median : 5.400   Median :0.1016   Median :39.60        Median :0.0049574                   Median :125.7
 Mean   : 2886    Mean   : 6.782   Mean   :0.1365   Mean   :50.86        Mean   :0.0111399                   Mean   :124.8
 3rd Qu.: 4000    3rd Qu.: 9.900   3rd Qu.:0.2286   3rd Qu.:71.30        3rd Qu.:0.0155759                   3rd Qu.:130.0
 Max.   :20000    Max.   :22.200   Max.   :0.3048   Max.   :71.30        Max.   :0.0584113                   Max.   :141.0
```

Data Quality Checks

- **Missing Values:** None found.
- **Correlation Analysis:**

airfoil | arf | corr_matrix

Filter

	Frequency	Angle_of_attack	Chord_length	Free_stream_velocity	Suction_side_displacement_thickness	Scaled_sound_pressure_level
Frequency	1.000000000	-0.27276454	-0.003660639	0.133663831	-0.230107353	-0.3907114
Angle_of_attack	-0.272764536	1.00000000	-0.504868150	0.058759565	0.753393785	-0.1561075
Chord_length	-0.003660639	-0.50486815	1.000000000	0.003786629	-0.220842431	-0.2361615
Free_stream_velocity	0.133663831	0.05875957	0.003786629	1.000000000	-0.003974013	0.1251028
Suction_side_displacement_thickness	-0.230107353	0.75339378	-0.220842431	-0.003974013	1.000000000	-0.3126695
Scaled_sound_pressure_level	-0.390711412	-0.15610753	-0.236161512	0.125102801	-0.312669506	1.0000000

  - **Angle of Attack** and **Suction Side Displacement Thickness**: strong positive correlation (0.753).
  - **Angle of Attack** and **Chord Length**: moderate negative correlation (-0.505).
    These correlations indicate potential multicollinearity among predictors.

Statistical Analysis

Methods

Several regression techniques were used to model SPL:

1. **Multiple Linear Regression (MLR)** – baseline model to establish linear relationships.
2. **Model Selection** – forward, backward, and stepwise approaches using **AIC** and **BIC**.
3. **Polynomial Regression** – added second-order and interaction terms to capture non-linearity.
4. **Robust Regression** – to handle deviations from normality.
5. **Generalized Least Squares (GLS)** – to address heteroscedasticity and correlated errors.

Multiple Linear Regression:

$$
\begin{aligned}
SPL = \beta_0 &+ \beta_1 \times \text{Frequency} \\
&+ \beta_2 \times \text{Angle of Attack} \\
&+ \beta_3 \times \text{Chord Length} \\
&+ \beta_4 \times \text{Free-Stream Velocity} \\
&+ \beta_5 \times \text{Suction Side Displacement Thickness}
\end{aligned}
$$

R-squared=0.5157, **p-value<2e-16** (statistically significant)

Summary of the model:

```
lm(formula = Scaled_sound_pressure_level ~ Frequency + Angle_of_attack +
    Chord_length + Free_stream_velocity + Suction_side_displacement_thickness,
    data = arf)

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 ***
Frequency                          -1.282e-03  4.211e-05  -30.45   <2e-16 ***
Angle_of_attack                    -4.219e-01  3.890e-02  -10.85   <2e-16 ***
Chord_length                       -3.569e+01  1.630e+00  -21.89   <2e-16 ***
Free_stream_velocity                9.985e-02  8.132e-03   12.28   <2e-16 ***
Suction_side_displacement_thickness -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
```

Diagnostic Analysis:

1)Multicollinearity Check:

As discussed in correlation part the, VIF(Variance Inflation Factor) was calculated

```
         Frequency                   Angle_of_attack                      Chord_length             Free_stream_velocity Suction_side_displacement_thickness
          1.144444                          3.441658                          1.510754                          1.041698                            2.532127
```

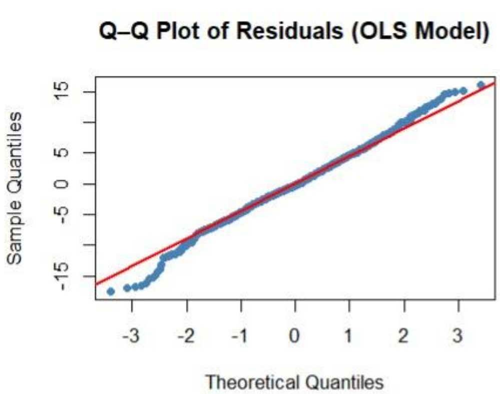
Although the values were not extremely high, there was some indication of multicollinearity.The highest VIF value was associated with Angle of Attack (VIF = 3.44).

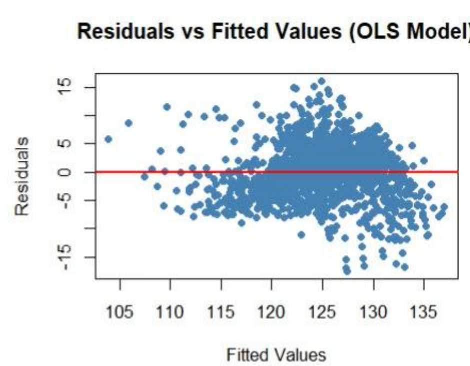
After removing AoA:

```
         Frequency                      Chord_length              Free_stream_velocity Suction_side_displacement_thickness
          1.079606                          1.054873                          1.019099                            1.114675
```

```
Call:
lm(formula = Scaled_sound_pressure_level ~ Frequency + Chord_length +
    Free_stream_velocity + Suction_side_displacement_thickness,
    data = arf)

Residuals:
    Min      1Q  Median      3Q     Max
-19.866  -3.109  -0.018   3.332  15.932

Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          1.304e+02  5.131e-01  254.03   <2e-16 ***
Frequency                           -1.174e-03  4.246e-05  -27.64   <2e-16 ***
Chord_length                        -2.597e+01  1.414e+00  -18.36   <2e-16 ***
Free_stream_velocity                 8.686e-02  8.351e-03   10.40   <2e-16 ***
Suction_side_displacement_thickness -2.692e+02  1.034e+01  -26.02   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.993 on 1498 degrees of freedom
Multiple R-squared:  0.4776,	Adjusted R-squared:  0.4763
F-statistic: 342.4 on 4 and 1498 DF,  p-value: < 2.2e-16
```

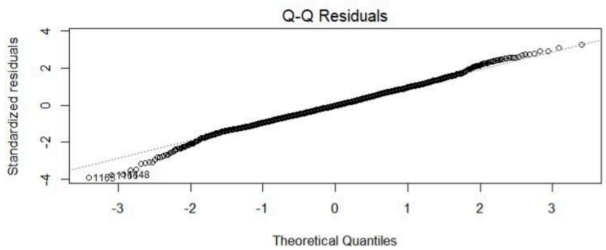
Q-Q Plot of Residuals and Residual vs Fitted::

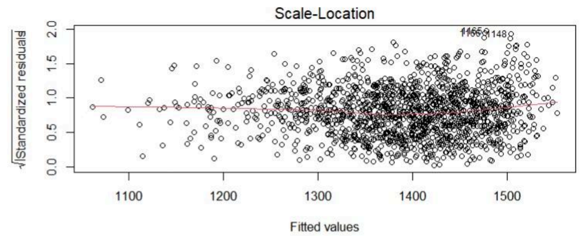




There are noticeable deviations at the tails, suggesting that the residuals depart from normality at both ends.The plot reveals that as the fitted values increase, the spread of the residuals also increases, which violates the assumption of homoscedasticity in the OLS model.

After using **box-cox transformation**:





Polynomial Regression

Second-order polynomial and interaction terms improved model fit:

$$
\begin{aligned}
SPL = \beta_0 &+ \beta_1 \times \text{Free\_Stream\_Velocity} \\
&+ \beta_2 \times (\text{Angle\_of\_Attack})^2 \\
&+ \beta_3 \times (\text{Chord\_Length})^2 \\
&+ \beta_4 \times (\text{Frequency} \times \text{Chord\_Length}) \\
&+ \beta_5 \times (\text{Frequency} \times \text{Suction\_Side\_Displacement\_Thickness}) \\
&+ \beta_6 \times (\text{Chord\_Length} \times \text{Angle\_of\_Attack})
\end{aligned}
$$

R^2 = 0.6192, **RSS = 4.266**
This model captured mild non-linear effects observed in residual plot

```
Residuals:
     Min       1Q   Median       3Q      Max
-17.4767  -2.5365  -0.2127   2.9646  16.1672

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                                 1.261e+02  4.097e-01 307.915  < 2e-16 ***
Free_stream_velocity                        9.575e-02  7.192e-03  13.314  < 2e-16 ***
I(Angle_of_attack^2)                       -1.074e-02  1.268e-03  -8.468  < 2e-16 ***
I(Chord_length^2)                          -2.141e+01  4.791e+00  -4.469 8.44e-06 ***
Frequency:Chord_length                     -5.818e-03  2.329e-04 -24.976  < 2e-16 ***
Frequency:Suction_side_displacement_thickness -7.638e-02  3.579e-03 -21.340  < 2e-16 ***
Chord_length:Angle_of_attack               -1.092e+00  2.412e-01  -4.528 6.43e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.266 on 1496 degrees of freedom
Multiple R-squared:  0.6192,	Adjusted R-squared:  0.6177
F-statistic: 405.4 on 6 and 1496 DF,  p-value: < 2.2e-16
```

.

Robust Regression

Applied to mitigate the influence of outliers:

- Improved residual distribution symmetry.
- Reduced sensitivity to heteroscedasticity, though R^2 slightly lower than polynomial regression.

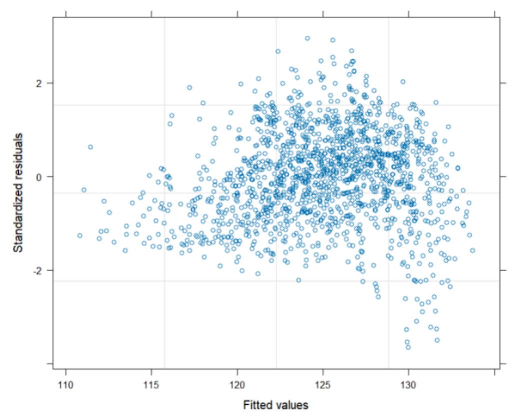
```
Call: rlm(formula = Scaled_sound_pressure_level ~ Frequency + Chord_length +
    Free_stream_velocity + Suction_side_displacement_thickness,
    data = arf)
Residuals:
      Min        1Q    Median        3Q       Max
-20.37126  -3.02783  -0.02526   3.21111  15.86943

Coefficients:
                                    Value     Std. Error t value
(Intercept)                          131.1444    0.4891   268.1483
Frequency                             -0.0013    0.0000   -32.5707
Chord_length                         -27.7069    1.3482   -20.5516
Free_stream_velocity                   0.0879    0.0080    11.0445
Suction_side_displacement_thickness -285.6688    9.8579   -28.9788

Residual standard error: 4.634 on 1498 degrees of freedom
```

Generalized Least Squares (GLS)

Used to correct for non-constant variance and correlated residuals:

- Provided stable coefficient estimates.
- Improved model diagnostics with more homoscedastic residuals

```
Generalized least squares fit by REML
  Model: Scaled_sound_pressure_level ~ Frequency + Chord_length + Free_stream_vel
ction_side_displacement_thickness
  Data: arf
       AIC      BIC    logLik
  7533.388 7575.883 -3758.694

Correlation Structure: AR(1)
 Formula: ~1
 Parameter estimate(s):
      Phi
0.8214131
Variance function:
 Structure: Power of variance covariate
 Formula: ~Suction_side_displacement_thickness
 Parameter estimates:
     power
0.05803285

Coefficients:
                                        Value Std.Error   t-value p-value
(Intercept)                          128.61172  0.985538 130.49898       0
Frequency                             -0.00086  0.000030 -28.85955       0
Chord_length                         -29.55854  4.335925  -6.81712       0
Free_stream_velocity                   0.10178  0.011397   8.93017       0
Suction_side_displacement_thickness -203.14591 26.688630  -7.61170       0

 Correlation:
                                    (Intr) Frqncy Chrd_l Fr_st_
Frequency                           -0.044
Chord_length                        -0.631  0.001
Free_stream_velocity                -0.604 -0.118  0.026
Suction_side_displacement_thickness -0.313  0.074  0.059  0.039

Standardized residuals:
         Min           Q1          Med           Q3          Max
-3.648500486 -0.696654735  0.006126964  0.665812342  2.947649013
```



Conclusion

- **Best Model:** The **Polynomial Regression Model** demonstrated the best accuracy, capturing non-linear relationships with improved R^2 (0.6192).
- **Interpretability:** The **transformed MLR model** (with multicollinearity correction and Box–Cox transformation) remains better for **interpretation** and **statistical inference**.
- **Broader Implications:**
  The results suggest that statistical regression techniques can effectively model aerodynamic noise. This approach can inform future airfoil design and noise prediction frameworks in aeronautical engineering.
- **Future Work**: In future we can explore the use of **Ridge** and **Lasso regression** for regularization, along with **tree-based models** like **Random Forests** and **XGBoost** to capture complex relationships and improve predictive accuracy.

1. Frequency:Higher frequencies cause acoustic energy to spread over a wider range, reducing intensity per band.Additionally, viscous damping is stronger at high frequencies, lowering overall SPL.
2. Angle of Attack:A moderate increase in angle of attack smoothens airflow and reduces turbulence near the surface.This minimizes flow separation and weakens noise-producing pressure fluctuations.
3. Chord Length: A longer chord length promotes smoother airflow and reduces vortex shedding at the trailing edge.This decreases broadband noise generation, leading to lower SPL.
4. Free-Stream Velocity: As velocity increases, aerodynamic forces and unsteady pressure fluctuations intensify.Since noise power scales sharply with velocity, SPL rises significantly at higher speeds.
5. Suction Side Displacement Thickness:A thicker boundary layer reduces the velocity gradient and suppresses surface turbulence.This lowers high-frequency noise production, thereby decreasing SPL.

Answers to Research Questions:

1) Almost all features are significantly important.
2) Yes, MLR can accurately predict SPL, after removing multicollinearity and using transformations.
3) Yes, the Polynomial regression model gives the best result as it captures the interactions better..
4) Yes, there was a multicollinearity issue due to the feature Angle of Attack which was highly correlated with Suction Side Displacement Thickness and moderately correlated with Chord Length. This variable was removed to address the issue.