Date: 17/04/2025

REGRESSION ANALYSIS

Analysis of Parkinson's Disease

Guide: Prof. Monika Bhatacharjee

Parkinson's Disease:

- PD is a progressive neurological disorder that affects movement
- It occurs due to the loss of dopamine-producing neurons in the brain, leading to symptoms like tremors (involuntary, rhythmic shaking movements, most commonly affecting the hands and arms), stiffness, slow movement, and balance problems
- While there is no cure, treatments like medication and therapy can help manage symptoms

Motivation:

Speech-Based Assessment of Parkinson's Disease:

- Gaining popularity as an automatic, low-cost, and easy-to-administer method for identifying early stage Parkinson's disease
- Involves two key tasks:
  - Distinguishing people with Parkinson from healthy speakers
  - Identifying Parkinson's disease stage
- Reduces hospital visits, patient inconvenience, and healthcare costs

Dataset contains:

- 5,875 voice recording observation collected from 42 individuals with early-stage Parkinson’s disease
- 18 Predictor variables (age, sex and 16 Biomedical Voice Measures)
- 2 response variables, namely Motor_UPDRS, Total_UPDRS - we will only work with Total_UPDRS as the response variable
- Total_UPDRS represents the total score, combining both motor and non-motor symptoms
- Each patient contributed approximately 200 recordings over a 6-month telemonitoring trial

Biomedical Voice Measures:

- Jitter Series: These measures capture frequency variation in the voice, indicating pitch instability.
- Shimmer Series: These measures reflect amplitude variation in the voice, indicating loudness instability.
- NHR/HNR: The Noise-to-Harmonic and Harmonic-to-Noise Ratios provide insights into the clarity and breathiness of the voice.
- RPDE, DFA, PPE: These are nonlinear voice signal measures that assess the complexity, irregularity, and fractal patterns in speech, providing a deeper understanding of speech dynamics.

Objective:

Analyze the Parkinson’s disease dataset through regression analysis

Approach:

- Exploratory Data Analysis (EDA) with visualization
- Regression Modeling and Evaluation
- Residual Analysis
- Re-training Regression Model
- Regularization methods - Ridge and Lasso
- Random Forest Regressor Modeling

Response Variable

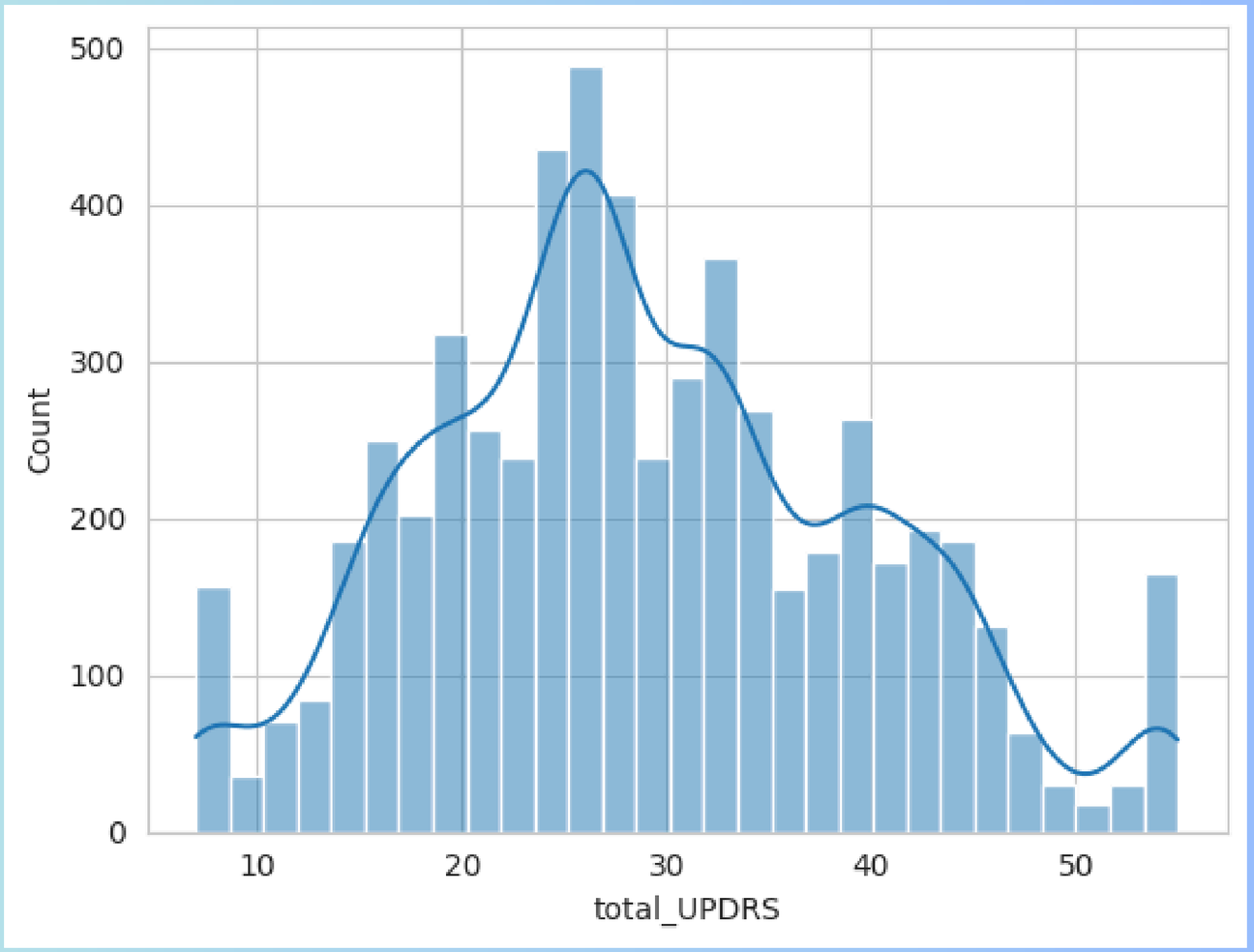
$$Y = \text{total\_UPDRS}$$

Predictor Variables

- X_1 = age
- X_2 = sex
- X_3 = Jitter(%)
- X_4 = Jitter(Abs)
- X_5 = Jitter:RAP
- X_6 = Jitter:PPQ5
- X_7 = Jitter:DDP
- X_8 = Shimmer
- X_9 = Shimmer(dB)
- X_{10} = Shimmer:APQ3
- X_{11} = Shimmer:APQ5
- X_{12} = Shimmer:APQ11
- X_{13} = Shimmer:DDA
- X_{14} = NHR
- X_{15} = HNR
- X_{16} = RPDE
- X_{17} = DFA
- X_{18} = PPE

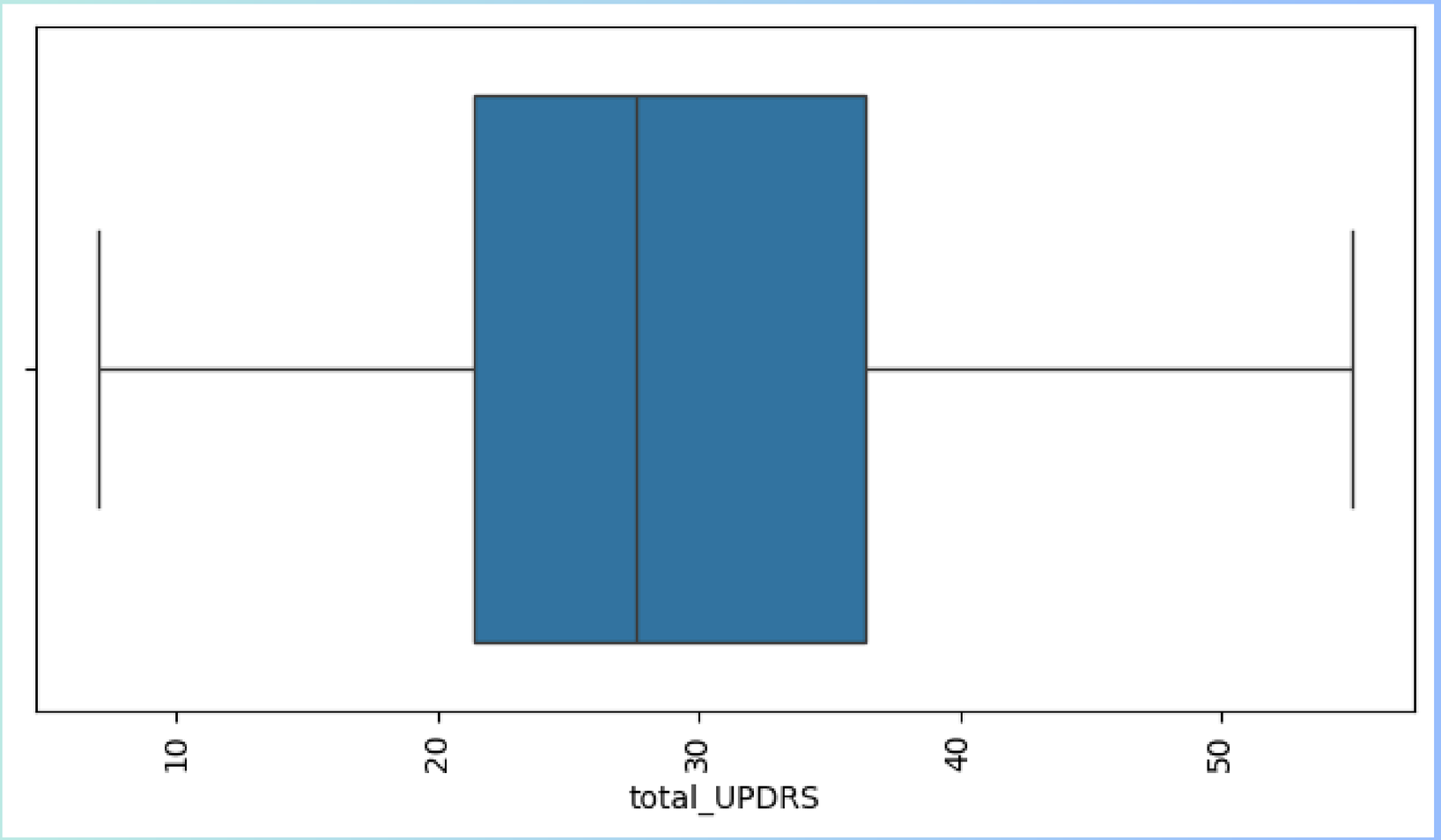
Distribution of Response (Y)

- Response variable shows a unimodal and slightly right-skewed distribution
- The variable has wide spread, indicating useful variation for modeling
- Potential outliers exist in the higher range (>50)



Box-Plot
Response Variable

- The response variable is slightly skewed



Correlation of Response with Predictors

age	0.310
sex	-0.097
Jitter(%)	0.074
Jitter(Abs)	0.067
Jitter:RAP	0.064
Jitter:PPQ5	0.063
Jitter:DDP	0.064
Shimmer	0.092
Shimmer(dB)	0.099
Shimmer:APQ3	0.079
Shimmer:APQ5	0.083
Shimmer:APQ11	0.121
Shimmer:DDA	0.079
NHR	0.061
HNR	-0.162
RPDE	0.157
DFA	-0.113
PPE	0.156

Summary Statistics

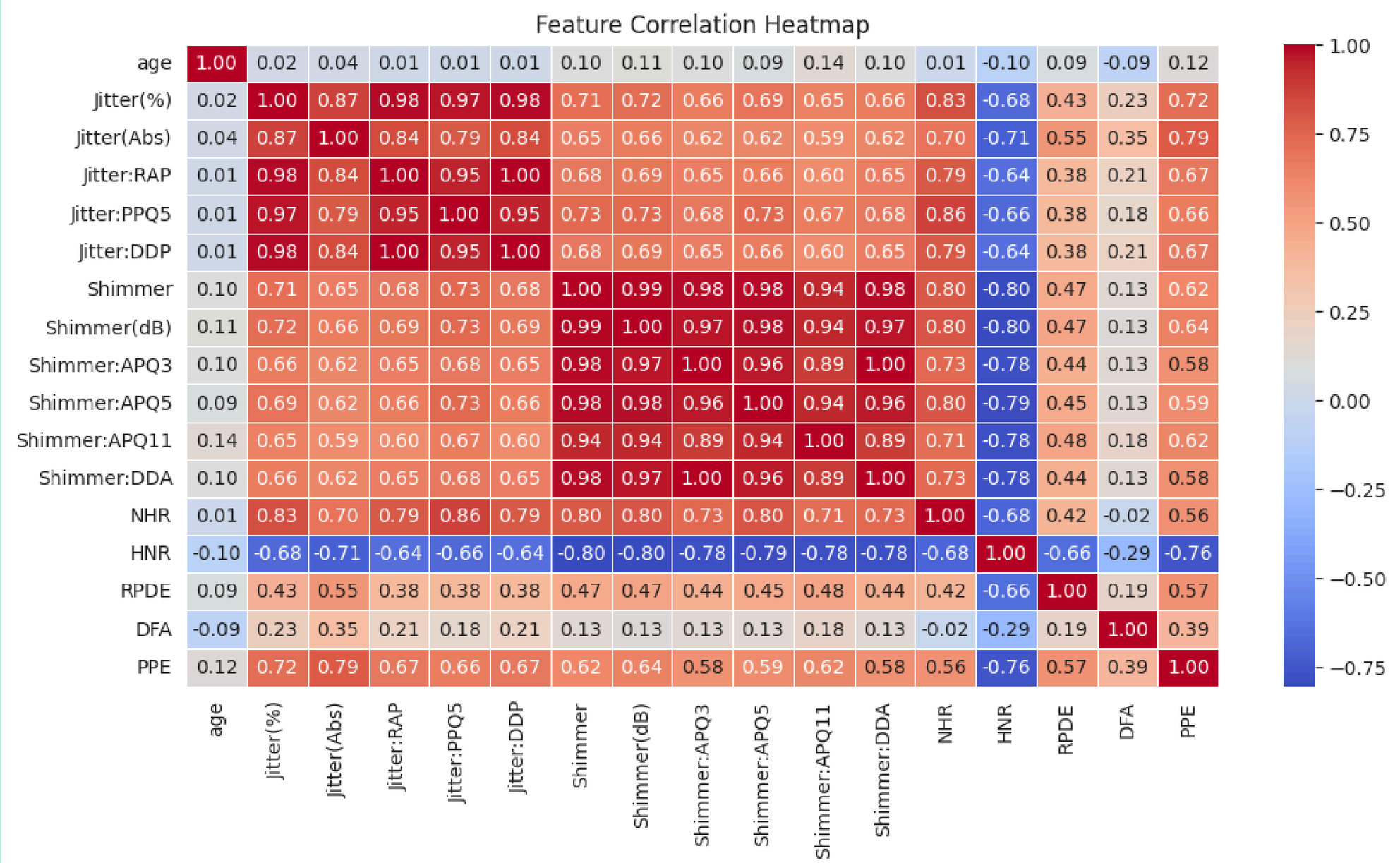
	count	mean	std	min	25%	50%	75%	max
age	5875.0	64.804936	8.821524	36.000000	58.000000	65.000000	72.000000	85.000000
sex	5875.0	0.317787	0.465656	0.000000	0.000000	0.000000	1.000000	1.000000
test_time	5875.0	92.863722	53.445602	-4.262500	46.847500	91.523000	138.445000	215.490000
motor_UPDRS	5875.0	21.296229	8.129282	5.037700	15.000000	20.871000	27.596500	39.511000
total_UPDRS	5875.0	29.018942	10.700283	7.000000	21.371000	27.576000	36.399000	54.992000
Jitter(%)	5875.0	0.006154	0.005624	0.000830	0.003580	0.004900	0.006800	0.099990
Jitter(Abs)	5875.0	0.000044	0.000036	0.000002	0.000022	0.000034	0.000053	0.000446
Jitter:RAP	5875.0	0.002987	0.003124	0.000330	0.001580	0.002250	0.003290	0.057540
Jitter:PPQ5	5875.0	0.003277	0.003732	0.000430	0.001820	0.002490	0.003460	0.069560
Jitter:DDP	5875.0	0.008962	0.009371	0.000980	0.004730	0.006750	0.009870	0.172630
Shimmer	5875.0	0.034035	0.025835	0.003060	0.019120	0.027510	0.039750	0.268630
Shimmer(dB)	5875.0	0.310960	0.230254	0.026000	0.175000	0.253000	0.365000	2.107000
Shimmer:APQ3	5875.0	0.017156	0.013237	0.001610	0.009280	0.013700	0.020575	0.162670
Shimmer:APQ5	5875.0	0.020144	0.016664	0.001940	0.010790	0.015940	0.023755	0.167020
Shimmer:APQ11	5875.0	0.027481	0.019986	0.002490	0.015665	0.022710	0.032715	0.275460
Shimmer:DDA	5875.0	0.051467	0.039711	0.004840	0.027830	0.041110	0.061735	0.488020
NHR	5875.0	0.032120	0.059692	0.000286	0.010955	0.018448	0.031463	0.748260
HNR	5875.0	21.679495	4.291096	1.659000	19.406000	21.920000	24.444000	37.875000
RPDE	5875.0	0.541473	0.100986	0.151020	0.469785	0.542250	0.614045	0.966080
DFA	5875.0	0.653240	0.070902	0.514040	0.596180	0.643600	0.711335	0.865600
PPE	5875.0	0.219589	0.091498	0.021983	0.156340	0.205500	0.264490	0.731730

MULTICOLLINEARITY

(Predictors are standardized)

$$Z_i = (X_i - \bar{x}_i) / s.d.(x_i)$$

HEATMAP



- Observing non-diagonal values suggests presence of high multicollinearity among the predictor variables

Variance Inflation Factor

- Features having VIF > 10 suggest serious multicollinearity

- Features responsible for multicollinearity are:
  - Jitter(%)
  - Jitter:RAP
  - Jitter:PPQ5
  - Jitter:DDP
  - Shimmer
  - Shimmer(dB)
  - Shimmer:APQ3
  - Shimmer:APQ5
  - Shimmer:APQ11
  - Shimmer:DDA

Feature	VIF	Feature	VIF
X_1	1.0958	X_{10}	23985721.3013
X_2	1.3487	X_{11}	52.5250
X_3	88.9910	X_{12}	15.2818
X_4	7.8520	X_{13}	23985640.3109
X_5	1323849.9985	X_{14}	8.5833
X_6	30.9966	X_{15}	5.4170
X_7	1324095.2665	X_{16}	2.0995
X_8	173.3748	X_{17}	1.6602
X_9	76.8452	X_{18}	4.4382

Condition Numbers and Indices

- Taking 100 to be the threshold for multicollinearity
- This approach suggests that last 8 principal components are responsible for Multicollinearity

Sorted Condition Indices		
1.00	6.53	7.66
1.12×10^{1}	1.41×10^{1}	1.57×10^{1}
3.79×10^{1}	5.45×10^{1}	6.68×10^{1}
7.39×10^{1}	1.14×10^{2}	2.75×10^{2}
5.57×10^{2}	8.10×10^{2}	1.30×10^{3}
2.53×10^{3}	2.98×10^{7}	5.41×10^{8}

LINEAR REGRESSION MODEL

Assumptions of Linear Regression:

- Relationship between predictor and response variables should be linear
- Errors are normally distributed with mean 0
- No autocorrelation - errors are independent and identically distributed
- Homoscedasticity - errors have constant variance
- Predictor variables are non-stochastic/determinastic

Linear Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_{18} X_{18}$$

where

Y and X_i's are as defined as earlier

Observations:

- R-squared = 0.170
- Adj. R-squared = 0.167

Some regression coefficients are

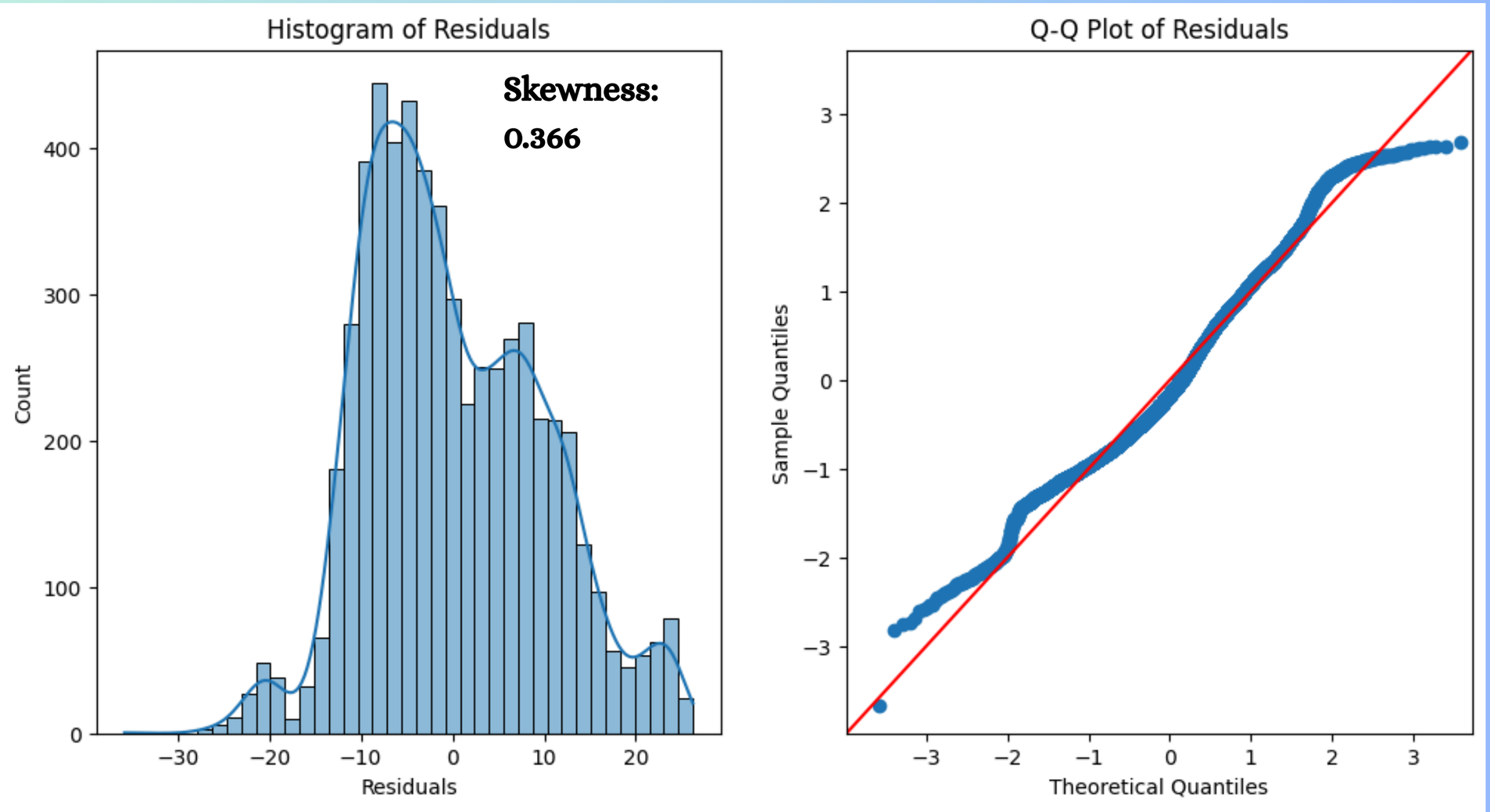
Variable	Coefficient	p-value
Intercept	29.0189	< 0.001
Age	2.6961	< 0.001
Sex	-1.2949	< 0.001
Jitter (Abs)	-2.2765	< 0.001
Shimmer	3.6130	0.031
Shimmer:APQ11	1.0529	0.035
NHR	-0.8554	0.022
HNR	-2.6368	< 0.001
RPDE	0.3742	0.043
DFA	-2.1968	< 0.001
PPE	1.6430	< 0.001

Model Performance

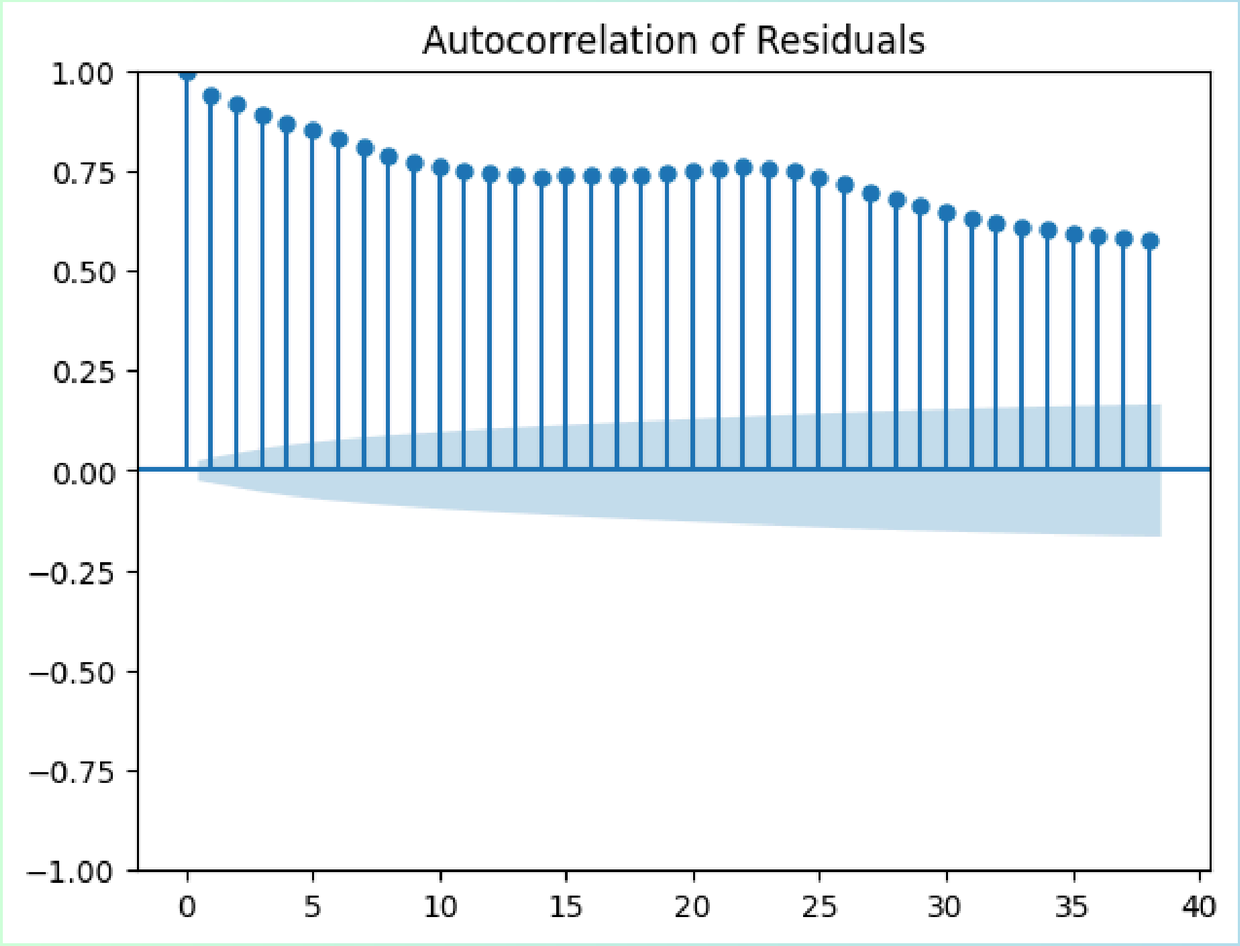
- R-squared = 0.170, Adj. R-squared = 0.167
- The model explains only 17% of the variability in the target variable
- Not very strong — there's a lot of variation unaccounted for
- F-statistic for significance of Linear Model is 66.53
- Statistically significant overall — at least one predictor is meaningfully associated with the outcome
- These variables are not statistically significant:
  - Jitter:RAP, Jitter:PPQ5, Jitter:DDP
  - Shimmer:APQ3, Shimmer:APQ5, Shimmer:DDA

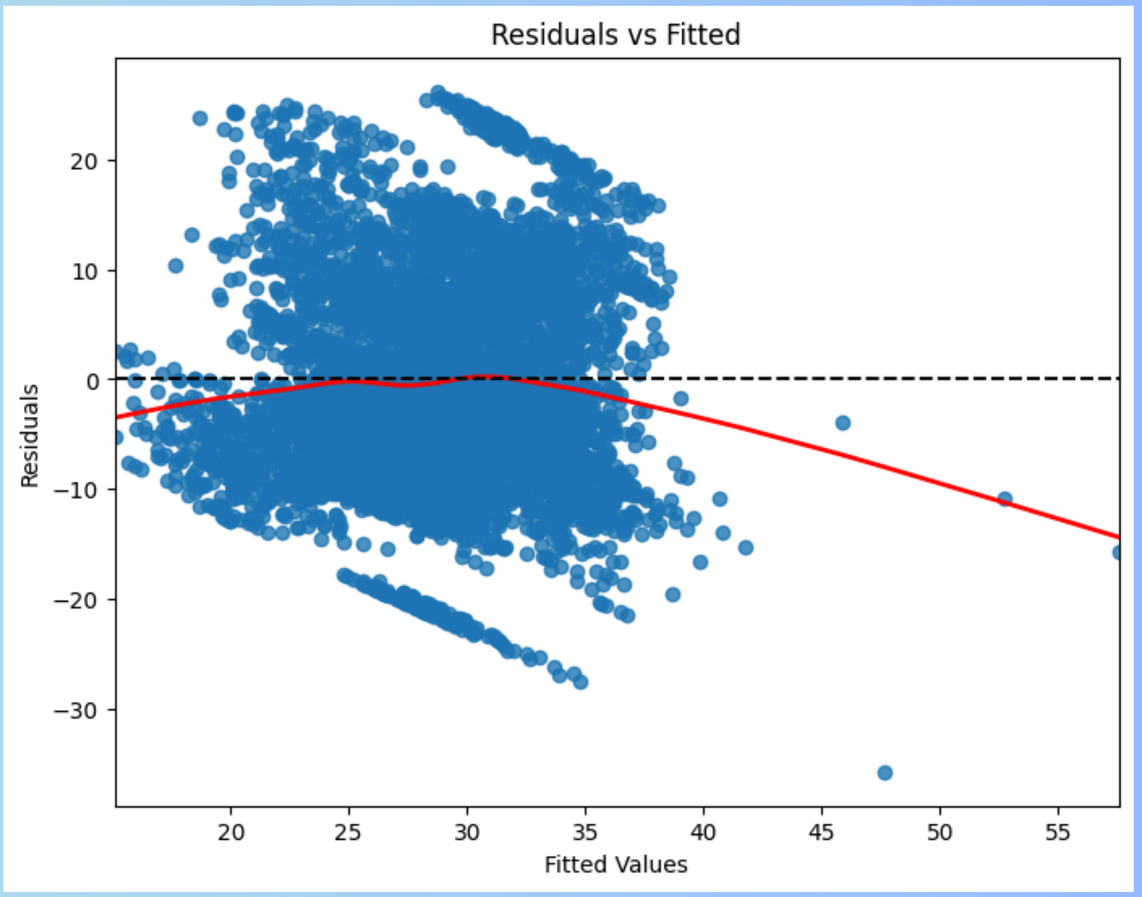
Residual Plots

- Residuals are approximately normal, but deviations in the tails indicate violations of the normality assumption.
- Transforming Y may improve normality



- Heteroscedasity detected
- The spread of residuals increases slightly as fitted values increase





- Strong positive autocorrelation
- Violation of the residual independence assumption
- It is suggested to use tree-based models

Detecting Outliers

There are mainly two types of outliers:

- Leverage points - a sample point unusual in predictors but not unusual response
- Influential points - a sample point unusual in both predictors and response

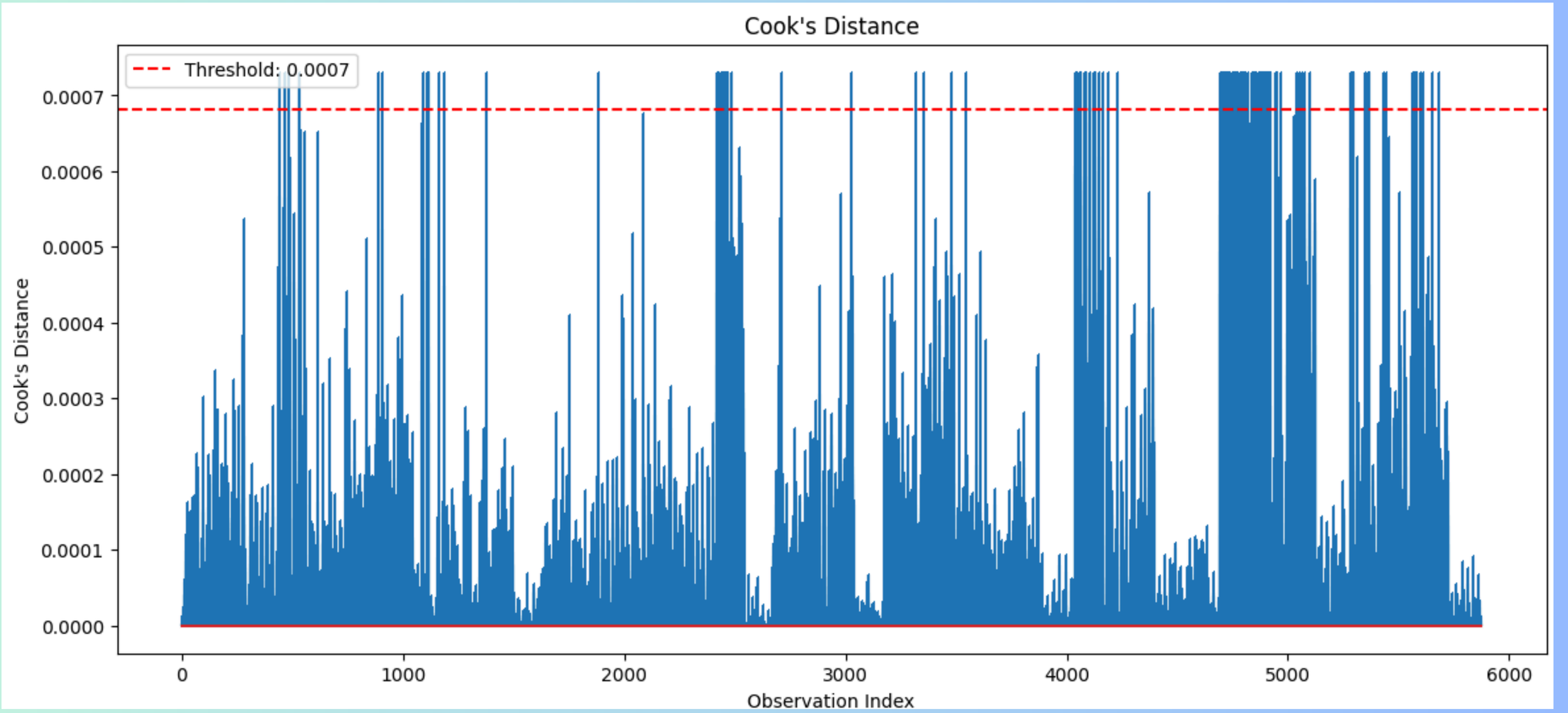
Various approaches to identify influential observations:

1. Cooks' distance method
2. DFFITS method
3. DFBETAS method
4. COVRATIO method

Cooks' Distance Approach

- Used 0.0007 (4/n) as cutoff for influential observation
- Notable high-influence spikes near indices - 1900 (very prominent), 4100, 5300

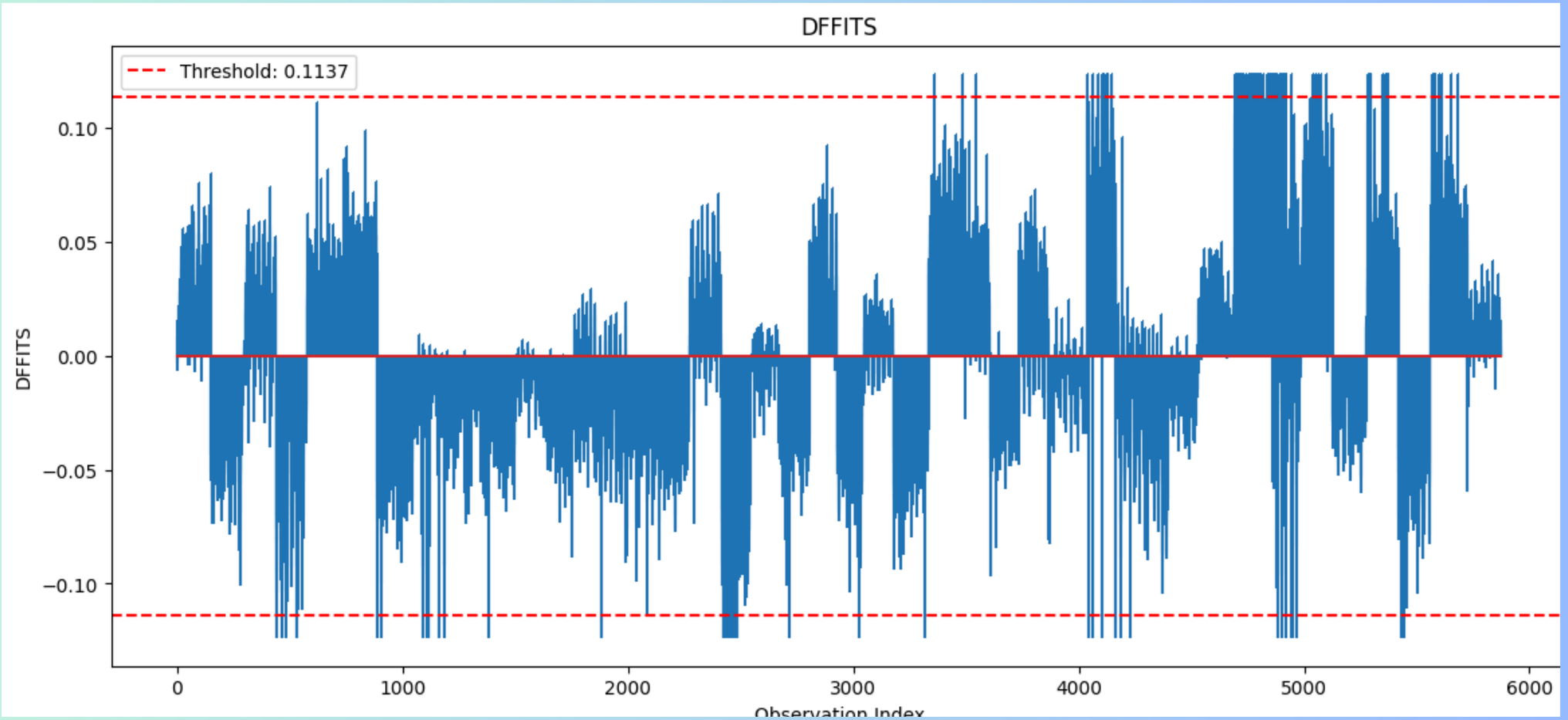
Observations are capped for better visualization



DFFITS Approach

- Used ±0.1137 as cutoff according to conventional formula $2 * \sqrt{(p/n)}$
- Notable high-influence spikes positive and negative near indices - 1900 (very prominent), 4100, 5300

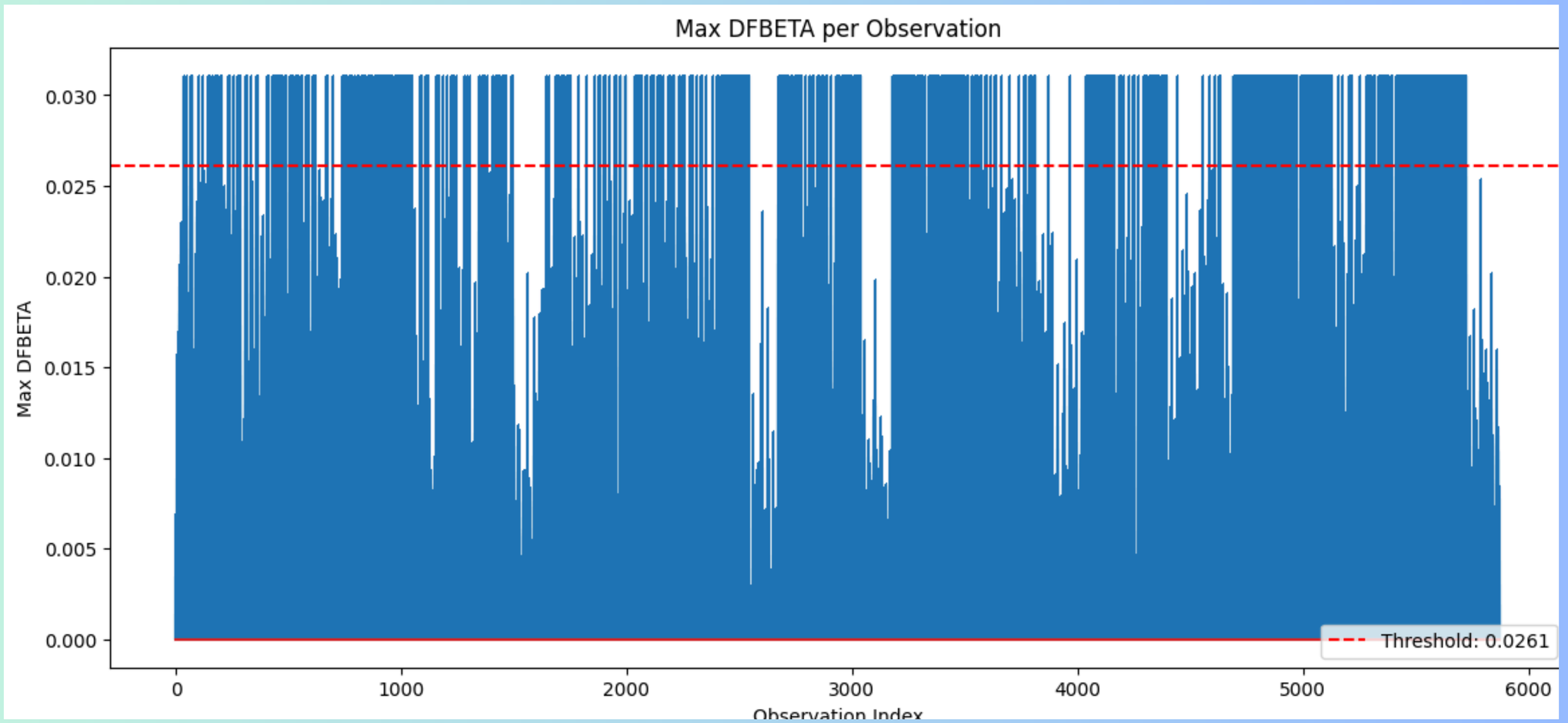
Observations are capped for better visualization



DFBETAS Approach

- Threshold used: 0.0261
- A few observations exceed the threshold, notably at indecies: 1900 (very high, >1.4), 4100, 5300

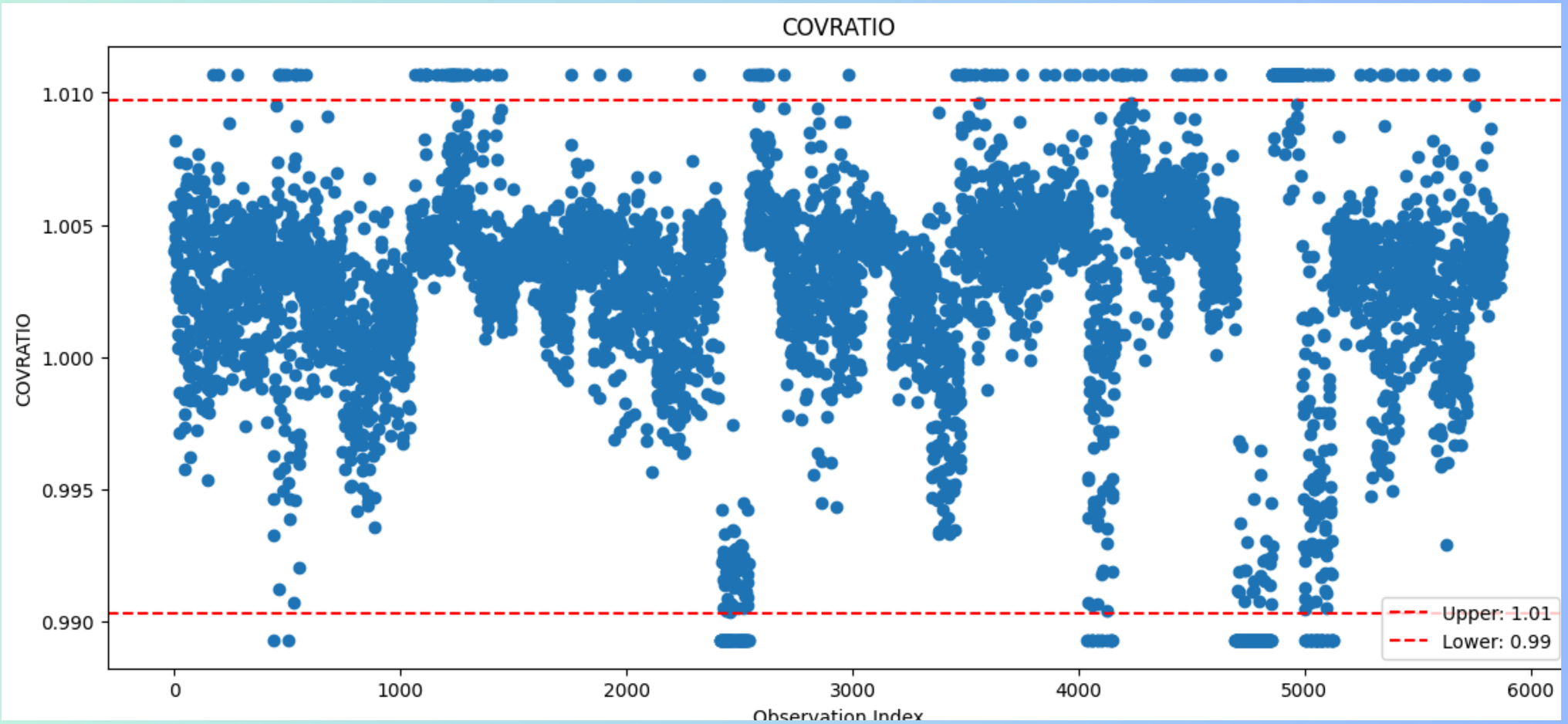
Observations are capped for better visualization



COVRATIO Approach

- Thresholds used: Upper = 1.01, Lower = 0.99
- A few data points lie outside this band, particularly around indices ~1800, ~4100, and ~5000

Observations are capped for better visualization



Number of outliers identified by each approach:

Method	#Outliers
Cook's Distance	146
DFFITS	147
DFBETAS	1492
COVRATIO	528

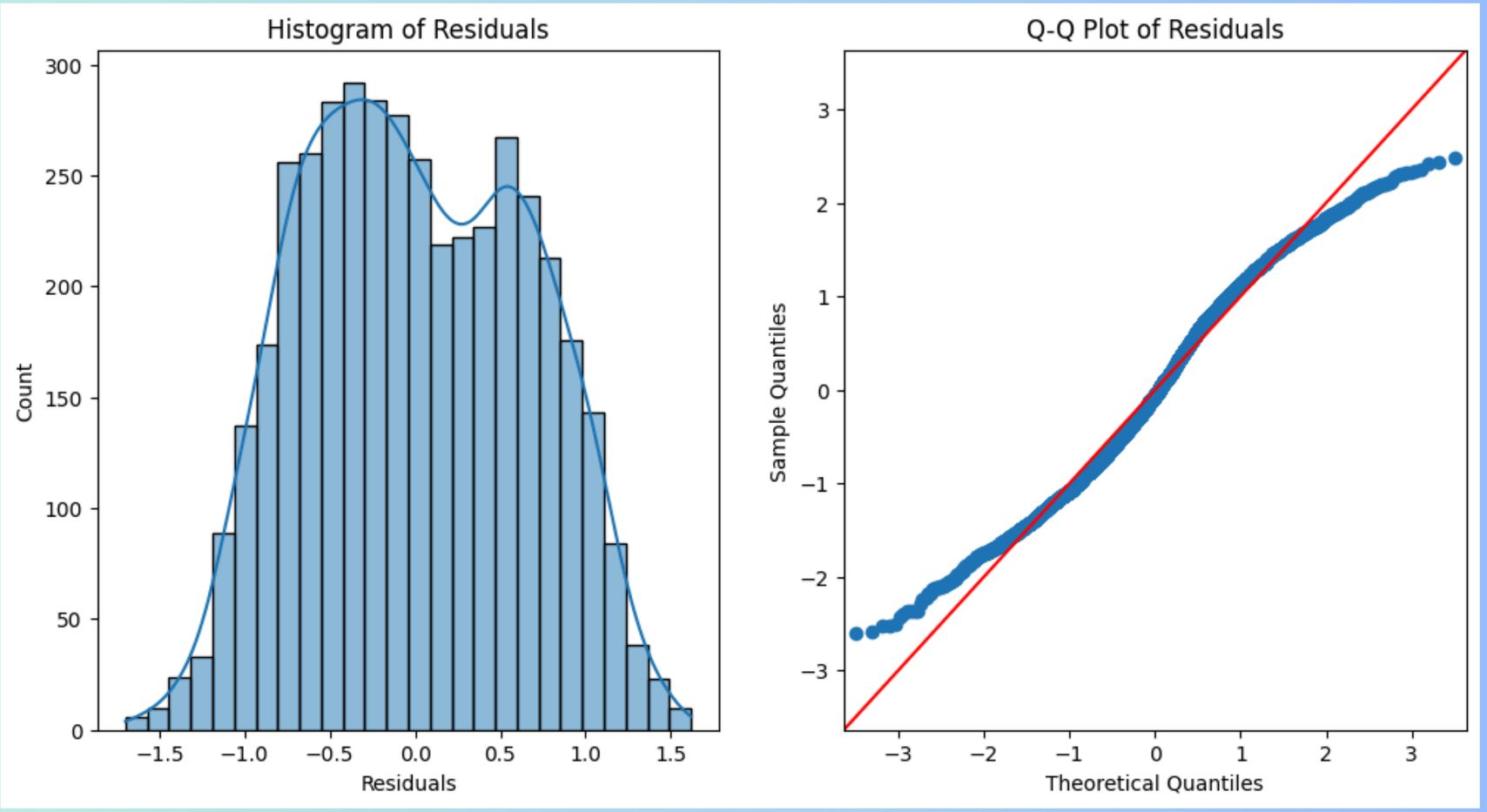
- Considering all the three methods, total of 1630 outliers have been detected
- We will remove these data points
- Then re-train the regression model
- Further perform residual analysis

Re-trained Regression Model:

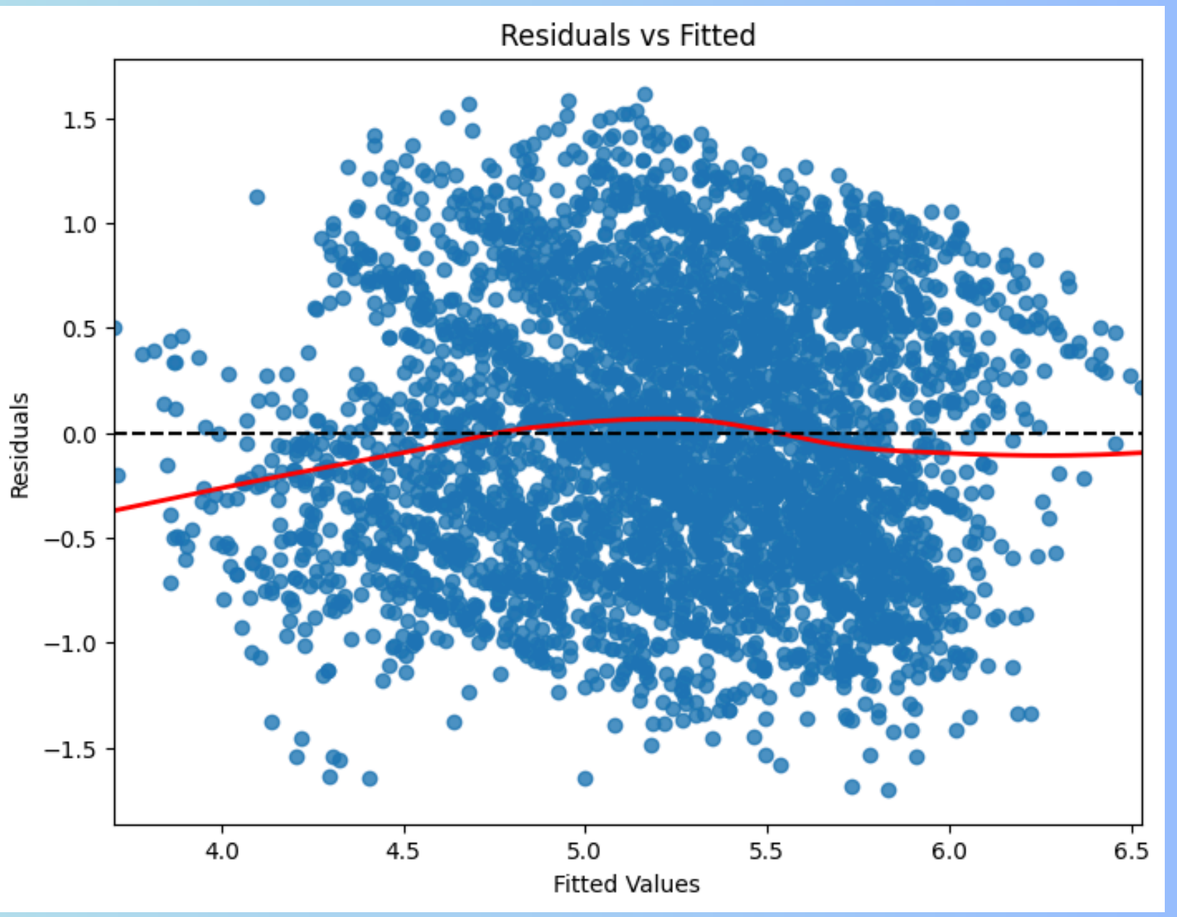
- We applied sqrt transformation on the response variable Y
- Also dropped the outliers identified by all the methods
- After fitting the regression model, we get:
  - R-squared = 0.356
  - Adj. R-squared = 0.353
- For the significance of linear model:
  - F-statistic = 129.5
  - Thus the overall model is still statistically significant

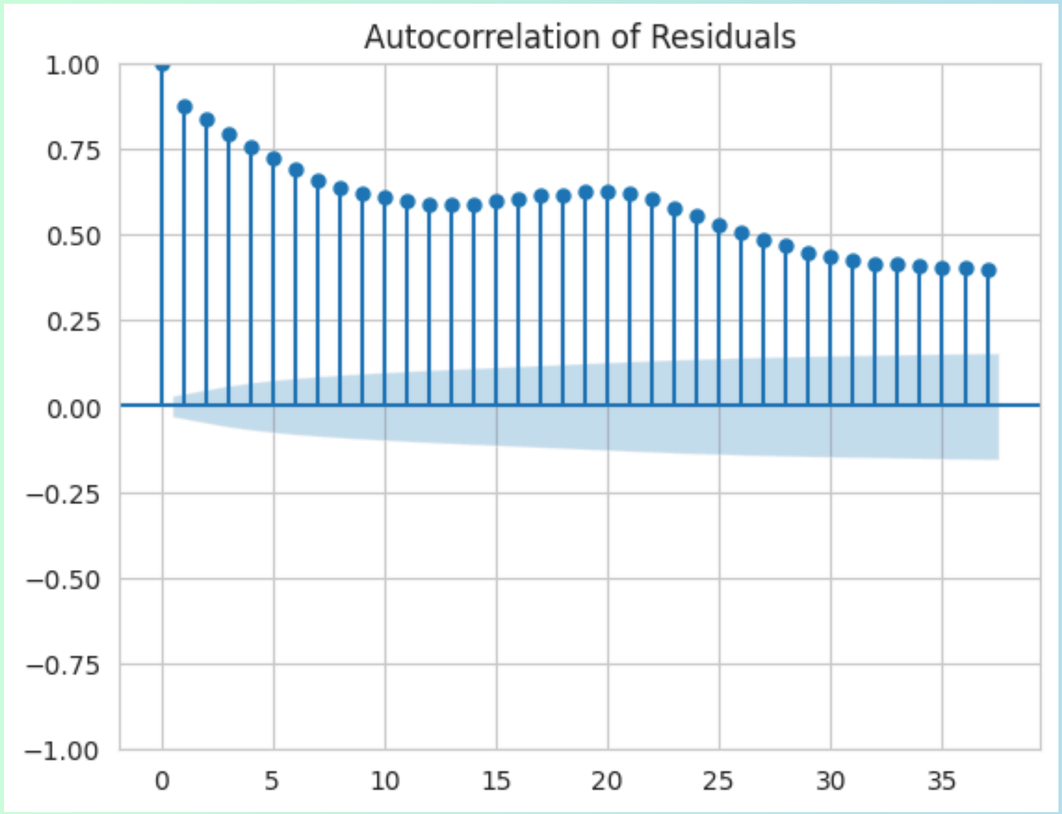
Residual Analysis on Re-trained model

Residual Distribution



Residual vs Fitted values





Auto-correlation function

Model Performance

- Although the Linear Model is significant, R-squared = 0.356 and adjusted R-squared = 0.353
- Also residuals are approximately normally distributed
  - Skewness : 0.069
  - Suggests residuals to be mildly skewed and light-tailed
- The residuals are not independent of fitted values
- Residuals have strong positive auto-correlation - violating the assumption of Linear Regression
- Results obtained are still not satisfactory. Thus we want to see whether results can be improved using Regularization techniques

REGULARIZATION REGRESSION

1. Ridge Regression
2. Lasso Regression

Ridge Regression

Objective:

$$\min_{\beta}\{\sum_{i=1}^{n}\left(Y_i - X_i^T\beta\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\}$$

where,

- Y_i: Actual response
- X_i: Feature vector for i^{th} observation
- β: Coefficient vector
- λ: Regularization parameter

Interpretation:

- Shrinks coefficients toward zero, but never exactly zero
- Helps in case of multicollinearity or high-dimensional data
- The tuning parameter λ controls the bias–variance trade-off
  - Small λ: behaves like OLS
  - Large λ: heavy shrinkage (increased bias, lower variance)

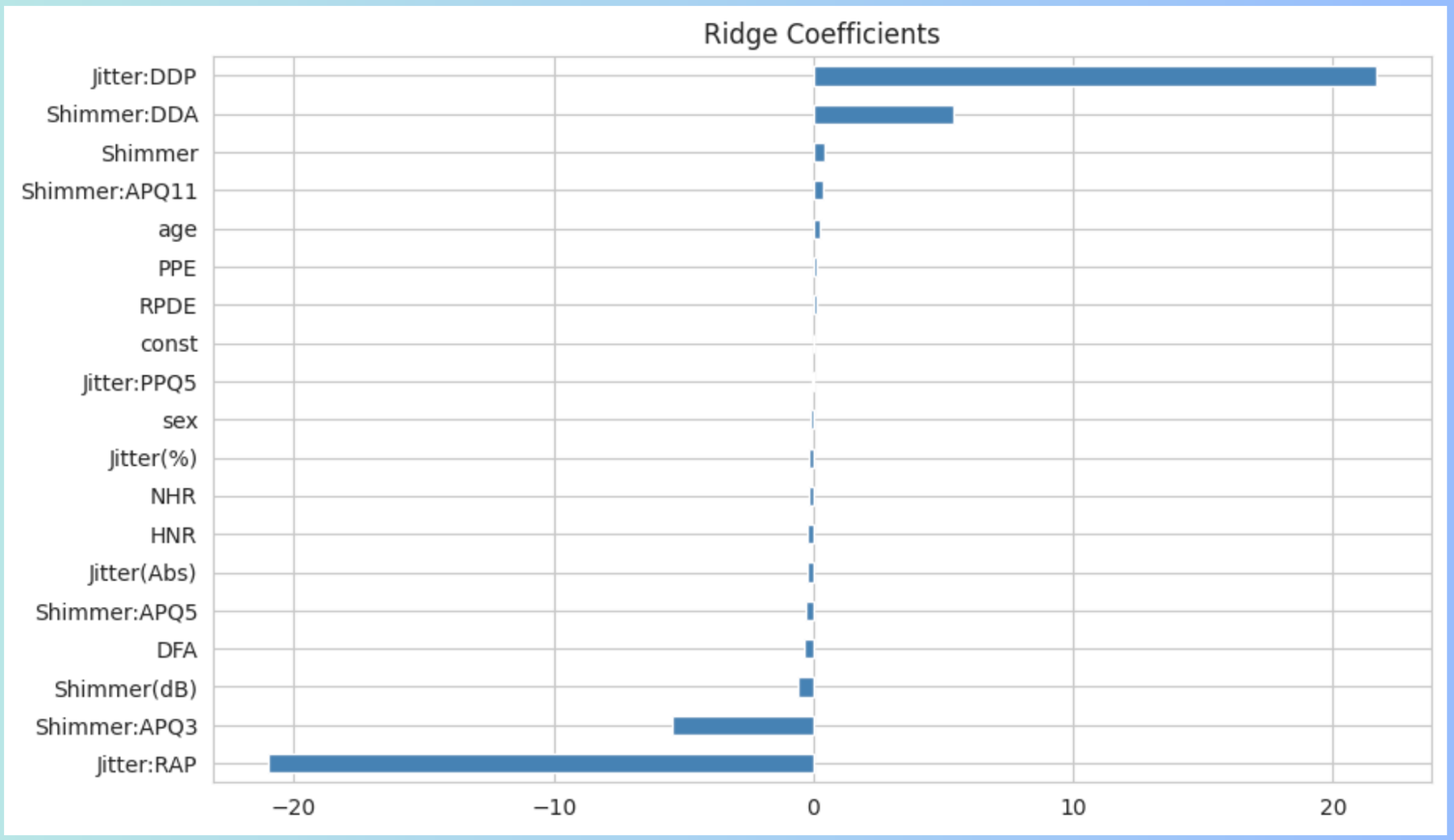
Obtained:

- Ridge_R^2: 0.3326
- Best λ: 0.001

Insights:

- Low value of best λ can be explained by low R-square
- The most significant features affecting response are Jitter:DDP and Jitter:RAP

Feature Coefficient Plot



Lasso Regression

Objective:

$$\min_{\beta}\{\sum_{i=1}^{n}\left(Y_i - X_i^T\beta\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\}$$

where,

- Y_i: Actual response
- X_i: Feature vector for i^{th} observation
- β: Coefficient vector
- λ: Regularization parameter

Interpretaion:

- Adds an L1 penalty to shrink coefficients
- Performs feature selection by setting some coefficients exactly to zero
- Useful in high-dimensional datasets
- Helps reduce overfitting and improves model interpretability
- Controlled by regularization parameter λ

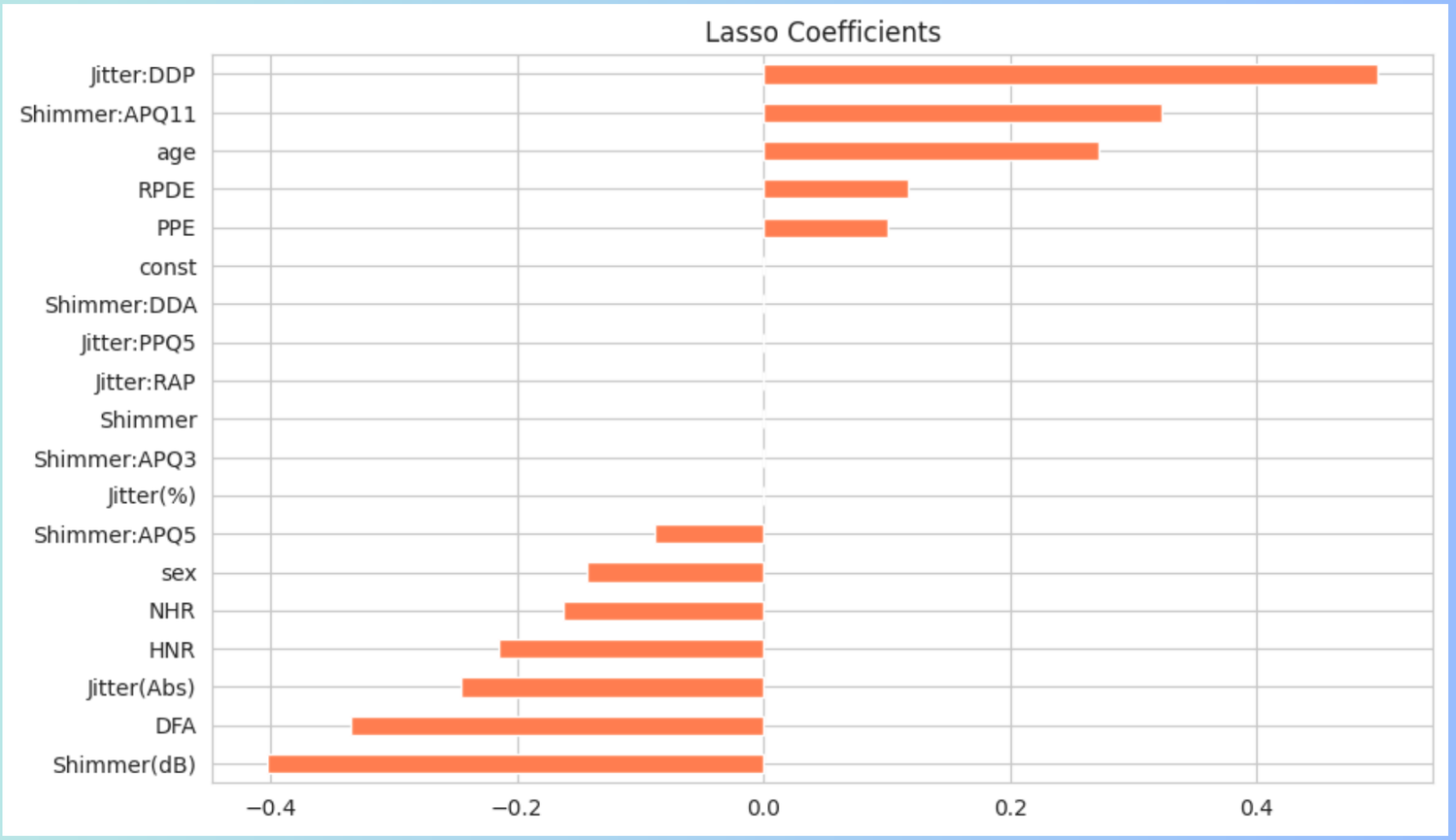
Obtained:

- Lasso_R^2: 0.3319
- Best λ: 0.001

Insights:

- Low value of best λ can be explained by low R-square
- The most significant features affecting response are Jitter:DDP and Shimmer(dB)

Feature Coefficient Plot



Conclusions:

- Jitter:DDP emerges as a dominant feature in both models
- Lasso Regression sets many coefficients exactly to zero, selecting only key predictors
- Ridge Regression retains all predictors but shrinks their impact
- Ridge is preferred when all variables may contribute
- Lasso is preferred when feature selection is desired
- Lasso aids interpretability, highlighting dominant voice measures

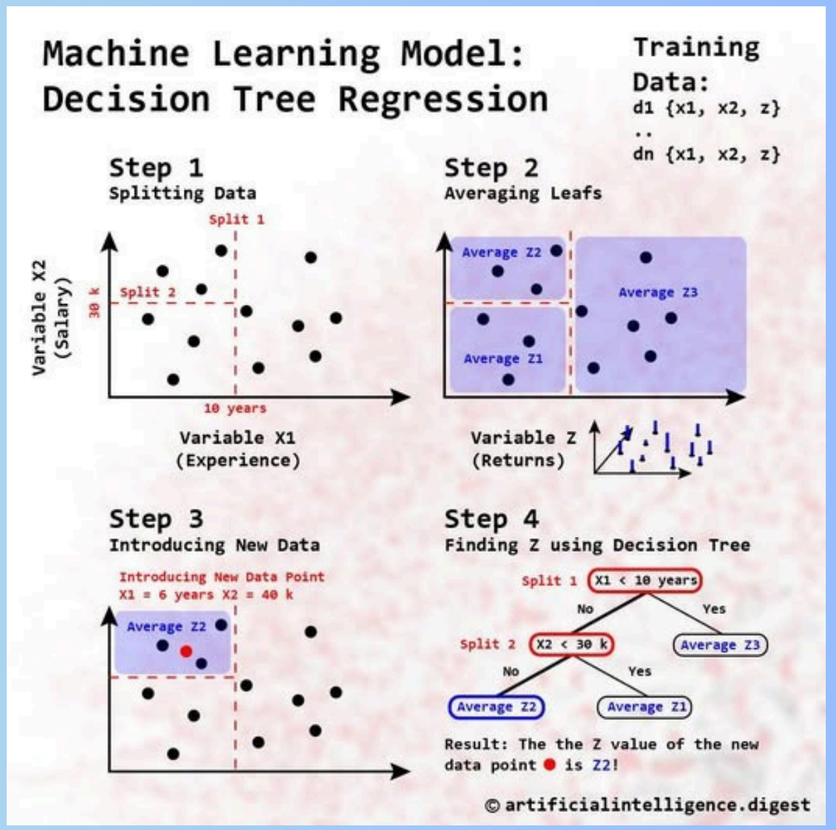
Thinking:

- None of the regression models implemented gave satisfactory results
- Even the data did not satisfy the assumptions of Linear Regression
- Thus, we take the dataset not fit for implementing Linear Regression
- Hence we now reside to Tree-Based-Algorithms
    - We will now implement Random Forest Regressor algorithm

RANDOM FOREST REGRESSOR

Decision Tree

- A decision tree is a supervised learning method used for classification and regression tasks.
- It splits the data based on features into binary decisions recursively.
- Each internal node represents a feature, each branch a decision, and each leaf a predicted outcome.

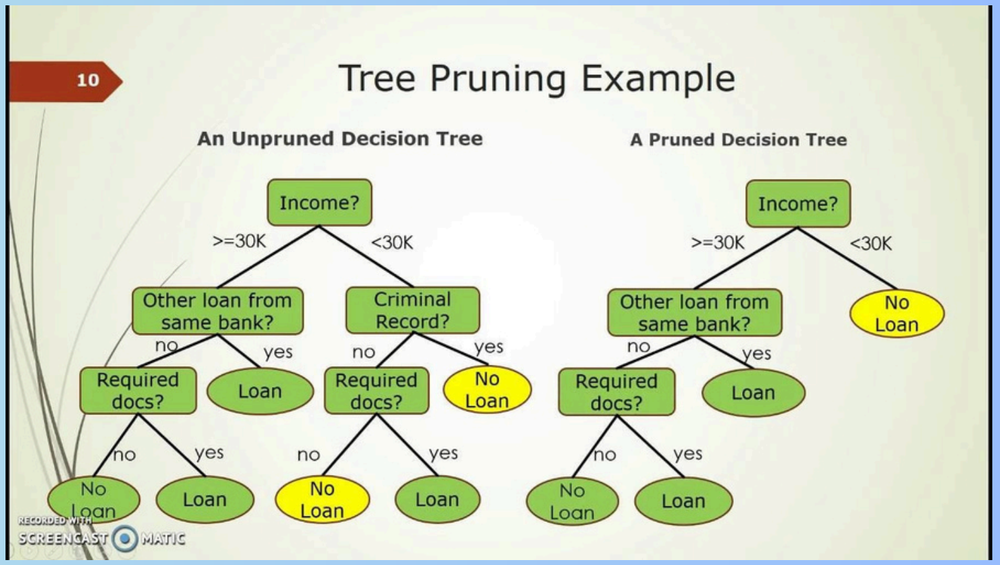


Cost Complexity Pruning

- Used to avoid overfitting by pruning unnecessary branches.
- Cost Function:

$$\sum_{m=1}^{|T|} \sum_{x_i \in R_m} \left(Y_i - \overset{\Lambda}{Y} R_m \right)^2 + \alpha |T|$$

- $|T|$: number of terminal nodes
- R_m : region of terminal node
- $\overset{\Lambda}{Y} R_m$: predicted response in region
- α : complexity tuning parameter



Recursive Binary Splitting

- Tree-building process uses recursive binary splitting to partition data:
  - Find the best feature & split point that minimizes MSE.
  - Split the data and repeat the process on each branch.
- It results in a tree structure that adapts to the data complexity.

Bias-Variance Tradeoff

- Total Error: $E[(Y - \hat{f}(x))^2] = Bias^2 + \sigma^2$, (where σ^2 is variance)
- Random Forest reduces variance while keeping bias low.

Decision Tree Regression

- Choose feature x_j and threshold s at each node.
- Split: $R_1(j,s) = \{x_i \mid x_{ij} \leqslant s\}$, $R_2(j,s) = \{x_i \mid x_{ij} > s\}$
- Minimize total variance (MSE):
- $L(j,s) = \sum_m \sum_{x_i \in R_m} (Y_i - \bar{Y}_{R_m})^2$

Problem Setup

- Feature Matrix: $X = [x_1, x_2, ..., x\_p] \in \mathbb{R}^{nxp}$
- Response Vector: $Y = [y_1, y_2, ..., y_n]^T \in \mathbb{R}^n$
- Goal: Estimate $Y = f(x) + \varepsilon$ using trees.

Building the Forest

- For m = 1 to M:
  1. Draw bootstrap sample.
  2. Train regression tree on subset.
  3. At each split: randomly select $k < p$ features.
- Final prediction: $\hat{f}(x) = (1/M) \sum_m f_m(x)$

Random Forest on Parkinson's Dataset

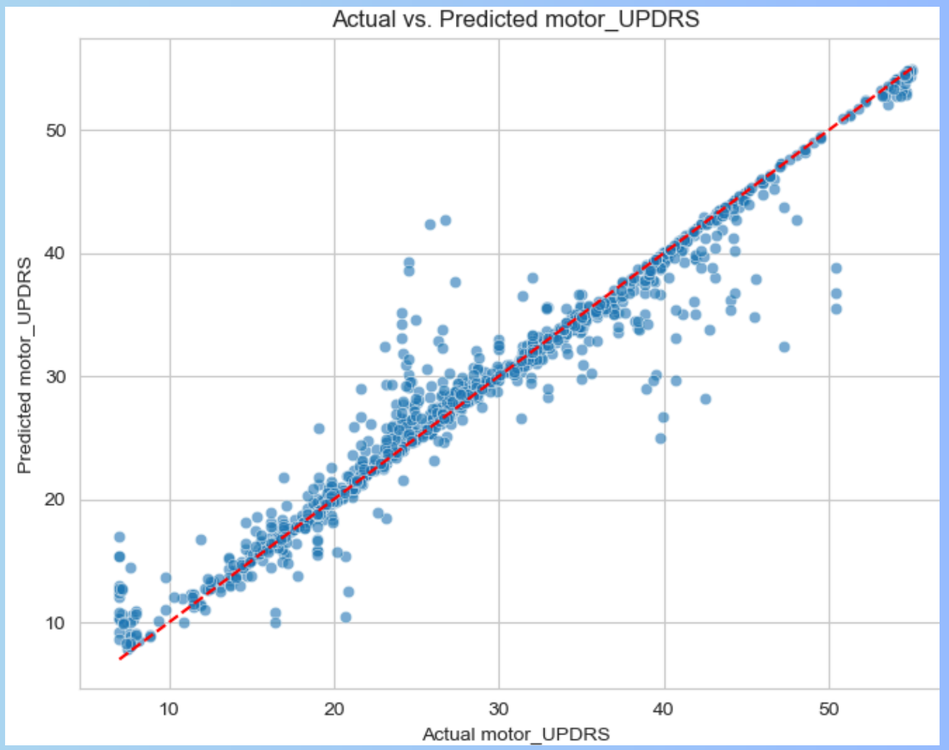
- Used RandomForestRegressor to predict response Y score
- Achieved:

  R^2 Score : 0.9476
  (High prediction accuracy)

  R^2 Score : 0.9465
  (after 5-fold cv)

  Mean Squared Error : 5.95

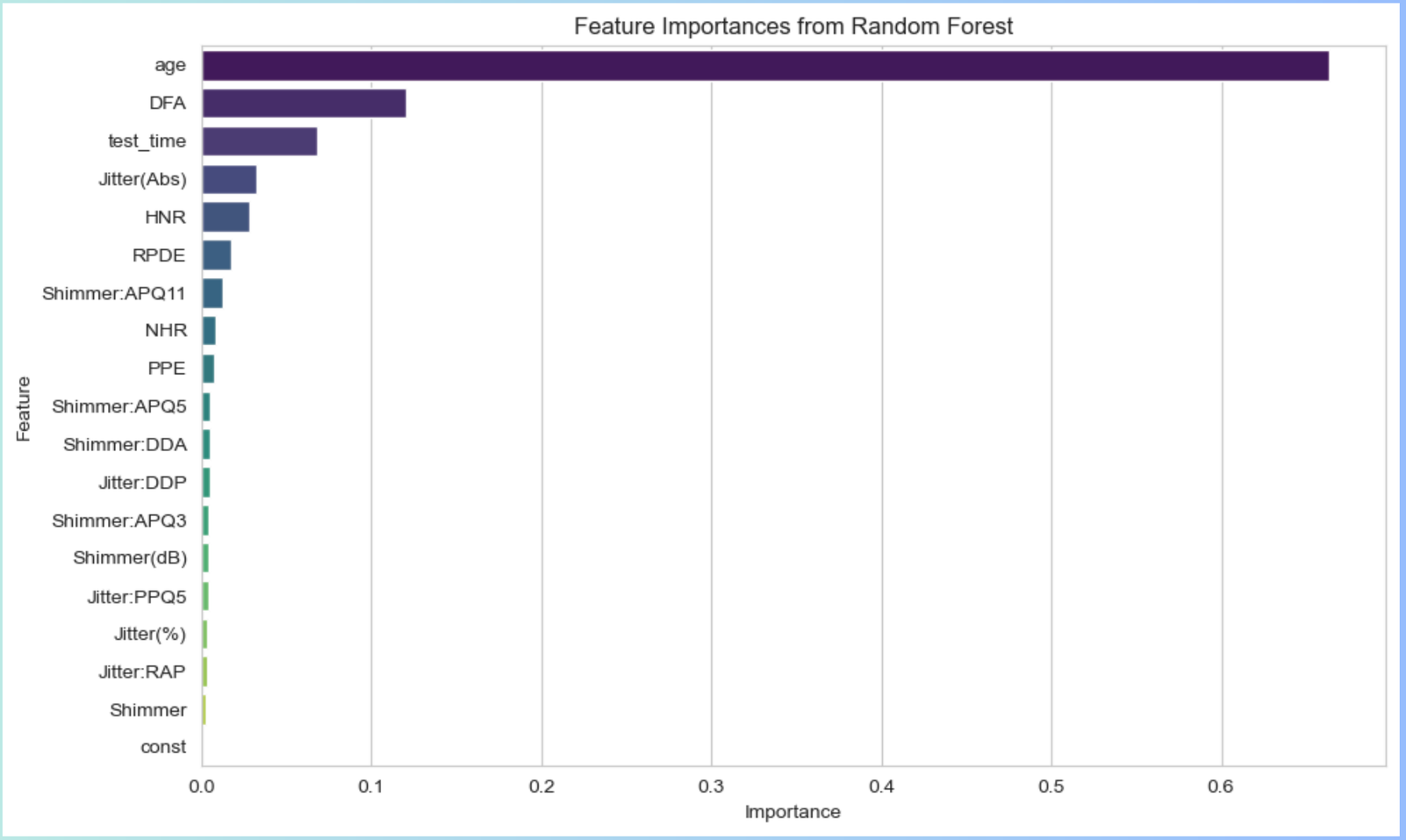
- Model output closely follows actual values, as seen in scatter plot.



Feature Importance plot

Most significant features:

- age
- DFA
- Jitter(Abs)
- HNR
- RDPE
- Shimmer:APQ11
- NHR
- Shimmer:APQ5



Trees vs Linear Models

Criteria	Linear Modely text	Decision Tree
Interpretation	Requires understanding coefficients	Visually easy to interpret
Handles Non-linearity	No	Yes
Qualitative Predictors	Needs dummy variables	Handles natively
Robustness	Stable	Sensitive to small data changes
Predictive Accuracy	Good for linear data	Better for non-linear relationships

Project Credits:

- Aditya Pandey : 24NOO64
- Anup Kumar Singh : 24N0060
- Lekhraj Meena : 24N0049
- Mansi Singh : 24N0061
- Ratnesh Pati Tripathi : 24N0065
- Sachin Dimri : 24N0083

THANK YOU!