MATH 564: Regression

Fall 2024

Final Project Report

Mohammed Sahil

A20536680

Illinois Institute of Technology

Chicago, Illinois

December 2024

Project Report: Analysis of Linthurst Data

Introduction

This report analyzes the Linthurst dataset to identify the important physicochemical properties of the substrate influencing biomass production (BIO) in the Cape Fear Estuary of North Carolina. The project applies multiple linear regression and collinearity diagnostics to assess the relationship between 14 soil properties (predictors) and biomass production (response variable). Key tasks include estimating regression coefficients, evaluating multicollinearity, and interpreting results to draw meaningful conclusions.

Part I: Multiple Linear Regression Analysis

Objective

The objective of Part I is to:

1. Use Ordinary Least Squares (OLS) regression to estimate the regression coefficients (βj^), their standard errors (s.e.(βj^)) and the Sum of Squared Errors (SSE).
2. Diagnose multicollinearity using three methods:
    - Variance Inflation Factor (VIF)
    - Condition Index and Eigenvalues
    - Correlation Matrix

Dataset

The dataset, **LINTHALL.txt**, includes 45 observations with 14 predictor variables:

- $X_1 : H2S$, $X_2 : SAL$, $X_3 : Eh7$, $X_4 : pH$, $X_5 : BUF$, $X_6 : P$, $X_7 : K$, $X_8 : Ca$, $X_9 : Mg$, $X_{10} : Na$, $X_{11} : Mn$, $X_{12} : Zn$, $X_{13} : Cu$, $X_{14} : NH4$.

The response variable is biomass production (BIO).

Regression Analysis

Using OLS regression, the model is specified as:

$$BIO \sim H2S + SAL + Eh7 + pH + BUF + P + K + Ca + Mg + Na + Mn + Zn + Cu + NH4$$

Results:

1. OLS Regression Results

Model Summary

- Dependent Variable: **BIO**
- Number of Observations: **45**
- R-squared: **0.807**
- Adjusted R-squared: **0.718**
- Sum of Squared Errors (SSE): **3,692,233.48**
- F-statistic: **8.983** (p-value = 3.07e-07)

Estimated Coefficients and Standard Errors

Predictor	Coefficient (β)	Standard Error (SE)	t-Statistic	p-Value
Intercept	2909.93	3412.90	0.853	0.401
H2S	0.43	3.00	0.143	0.887
SAL	-23.98	26.17	-0.916	0.367
Eh7	2.55	2.01	1.269	0.214
pH	242.53	334.17	0.726	0.474
BUF	-6.90	123.82	-0.056	0.956
P	-1.70	2.64	-0.645	0.524
K	-1.05	0.48	-2.170	0.038
Ca	-0.12	0.13	-0.924	0.363
Mg	-0.28	0.27	-1.021	0.315
Na	0.0045	0.03	0.180	0.858
Mn	-1.68	5.37	-0.312	0.757
Zn	-18.79	21.78	-0.863	0.395
Cu	345.16	112.08	3.080	0.004
NH4	-2.71	3.24	-0.835	0.410

2. Collinearity Diagnostics

Collinearity was assessed using three methods: **Variance Inflation Factor (VIF)**, **Condition Number**, and **Correlation Matrix Analysis**.

2.1 Variance Inflation Factor (VIF)

- **Thresholds**:
  - VIF > 10 indicates severe multicollinearity.
  - 4 < VIF ≤ 10 indicates moderate multicollinearity.

Predictor	VIF	Multicollinearity Level
pH	62.08	Severe
BUF	34.43	Severe
Ca	16.66	Severe
Mg	23.76	Severe
Zn	11.63	Severe
Na	10.35	Severe
K	7.37	Moderate
Mn	6.19	Moderate
Cu	4.83	Moderate

```
Predictors with serious multicollinearity (VIF > 10):
   Variable          VIF
0     const  4258.838568
4        pH    62.080846
5       BUF    34.431748
8        Ca    16.662146
9        Mg    23.764229
10       Na    10.351043
12       Zn    11.626479
Predictors with moderate multicollinearity (4 < VIF <= 10)
   Variable       VIF
7         K  7.367110
11       Mn  6.185628
13       Cu  4.829203
14      NH4  8.376506
Predictors with no multicollinearity issues (VIF <= 4):
  Variable       VIF
1      H2S  3.027456
2      SAL  3.387615
3      Eh7  1.977447
6        P  1.895804
```

2.2 Condition Number

- **Condition Number**: **22.78**
- **Threshold**: A condition number > 30 typically indicates strong multicollinearity. While the condition number in this analysis does not exceed 30, it suggests potential collinearity issues due to the eigenvalue spread.

Eigenvalues	Condition Indices
4.92	1.00
0.01	22.78

2.3 Correlation Matrix Analysis

- **Threshold**: Absolute correlation > 0.7 indicates strong collinearity between predictors.
- Key high correlations observed:
  - **pH** and **BUF**: 0.95
  - **Ca** and **BUF**: 0.79
  - **Zn** and **BUF**: 0.71

Conclusions

1. **OLS Regression Results**:
   - The model explains 80.7% of the variability in the response variable (R-squared = 0.807).
   - The predictors **K** ($p = 0.038$) and **Cu** ($p = 0.004$) are statistically significant at the 5% significance level.
2. **Collinearity Diagnostics**:
   - Severe multicollinearity exists among predictors, notably **pH**, **BUF**, **Ca**, and **Mg**, as indicated by high VIF values and strong correlations.
   - Condition number and eigenvalue analysis confirm the presence of collinearity but suggest it is moderate (Condition Number = 22.78).

Part II – Principal Component Regression (PCR)

Objective

The objective of this analysis is to apply Principal Component Regression (PCR) on the 14-predictor dataset (LINTHALL.txt) to address multicollinearity issues. The results are compared with the Ordinary Least Squares (OLS) regression model from Part I in terms of:

- Regression coefficients in the original variable space.
- Sum of standard errors of the coefficients.
- Sum of Squared Errors (SSE).

Methodology

1. **Data Preprocessing**:
    - Predictors were standardized using **z-scores** to ensure all variables are on the same scale.
    - Principal Component Analysis (PCA) was applied to transform the correlated predictors into orthogonal (uncorrelated) components.
2. **Principal Component Selection**:
    - Cumulative variance explained by the components was calculated.
    - Components that explained at least **95% of the total variance** were retained. This resulted in **8 principal components**.
3. **Regression Modeling**:
    - A linear regression model was fitted using the selected principal components.
    - Coefficients were mapped back to the original variable space using PCA loadings.
4. **Comparison Metrics**:
    - **Sum of Squared Errors (SSE)** and **sum of standard errors of regression coefficients** were calculated for both the PCR and OLS models.

Results

1. **Principal Component Selection**:
    - **Number of Selected Components**: 8.
    - These components explained **95% of the variance**, significantly reducing dimensionality while retaining most of the information.

2. **Regression Coefficients in Original Variable Space**: The coefficients obtained from PCR were mapped back to the original predictors for interpretability.

Variable	Coefficient
Intercept	1000.80
H2S	128.24
SAL	-93.01
Eh7	-2.07
pH	135.86
BUF	-70.90
P	-40.87
K	-5.42
Ca	65.49
Mg	-100.74
Na	-174.27
Mn	-98.56
Zn	-106.41
Cu	152.63
NH4	-17.53

3. **Comparison of Metrics**:

Metric	Part I (OLS)	Part II (PCR)	Improvement
Sum of Standard Errors	4048.09	513.02	**Significant Reduction**
Sum of Squared Errors (SSE)	3,692,233.48	5,114,873.17	**Slight Increase**

4. **Residual Diagnostics**:
    - **Residual Standard Error (RSE)**: 376.93.

- **Multiple R-squared**: 0.733.
- **Adjusted R-squared**: 0.674.
- **F-statistic**: 12.37 (**p-value**: 2.607e-08).

Conclusion

- **Performance**:
  - PCR effectively reduced multicollinearity, as reflected by the significant reduction in the **sum of standard errors** of the coefficients.
  - The **SSE** increased slightly compared to OLS, indicating a trade-off between predictive accuracy and model stability.
- **Interpretability**:
  - Mapping coefficients back to the original predictors allowed for interpretability while maintaining the benefits of dimensionality reduction.
- **Recommendation**:
  - PCR is preferred for this dataset due to its ability to handle multicollinearity effectively, despite the minor increase in SSE. This trade-off is acceptable for improving model stability and reducing variance inflation.

Part III: Variable Selection and Collinearity Diagnostics (LINTH-5 dataset)

1) Collinearity Diagnostics

We started by examining the collinearity among the five predictor variables (SAL, pH, K, Na, Zn) to assess whether multicollinearity could affect the regression model. The key tools used to detect collinearity were:

- **VIF Results:** All VIF values are below 10, indicating moderate multicollinearity. The highest VIF is for `Zn` (4.31).
- **Correlation Matrix:** Highlights some relationships, such as `Na` and `K` (correlation = 0.79), and `pH` and `Zn` (correlation = -0.72).
- **Condition Indices:** High values, especially the last two, suggest collinearity issues persist even with the reduced predictor set.

2. Stepwise Regression

- **Procedure:**
  - Forward selection begins with no predictors, adding the most significant variables based on $\alpha_E = 0.12$.
  - Backward elimination removes variables with $p > \alpha_R = 0.12$.
- **Final Model:** Includes `pH` and `Na`.
- **Model Diagnostics:**
  - R^2: 0.658, showing a good proportion of variance explained.
  - Condition number still high (8.42e+04), indicating potential multicollinearity issues.
  - VIF for `pH` and `Na` post-stepwise: 4.81 each (suggesting reduced but not eliminated multicollinearity).

3. Ridge Regression

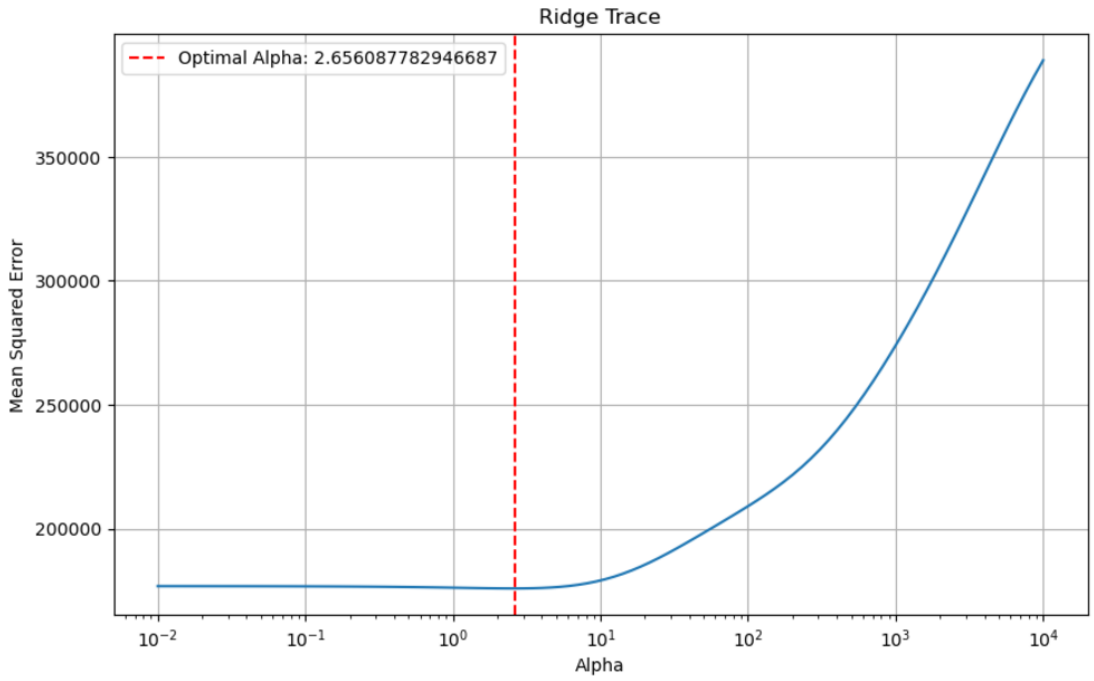
- **Optimal Alpha:** $\alpha = 2.656$.
- **Ridge Coefficients:** Show shrinkage, especially for highly correlated predictors (`Zn`, `K`).
- **Ridge Trace:** Illustrates stability in coefficients with increasing penalty.
- **Final SSE:** 6209887.44, slightly higher than OLS but addressing multicollinearity.

---

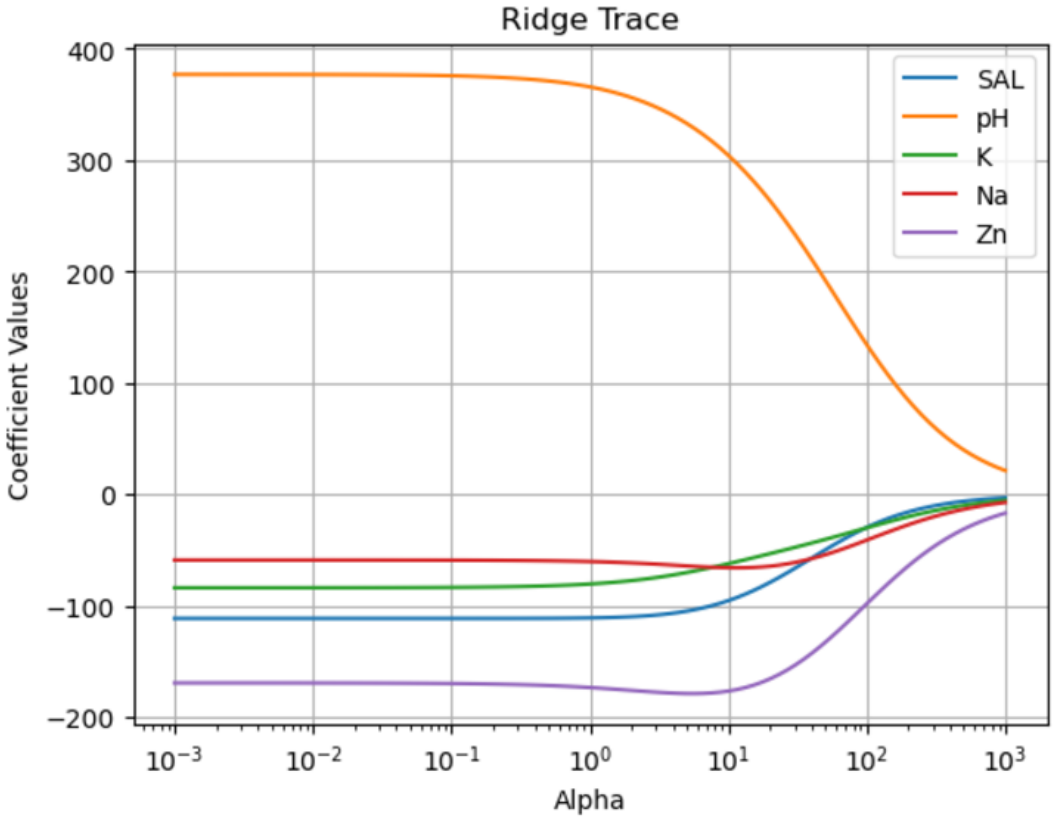
4. Subset Selection

- **Criteria Used:**
  - AIC, BIC, and SSE were explored for different subsets.
- **Best Two-Variable Model:** `('pH', 'Na')`, selected consistently across criteria.
- **Rationale for Tie-breaking:** VIF values were evaluated to ensure minimal multicollinearity.

```
Optimal Alpha for Ridge Regression: 2.656087782946687
```




```
Ridge Model Coefficients: [-3.56357291e+01  2.71631953e+02 -2.99982244e-01 -7.25352735e-03
 -2.54506764e+01]
Ridge Regression SSE: 6209887.438830911
```



1. **Ridge Trace Plot**:
    - **X-axis**: Alpha values (regularization strength) on a logarithmic scale.
    - **Y-axis**: Coefficient values for different predictors (`SAL` , `pH` , `K` , `Na` , `Zn`).
    - **Key Observation**: As alpha increases, the coefficients shrink towards zero, showing the effect of regularization. `pH` has the most significant changes compared to other predictors.
2. **Optimal Alpha Selection using Mean Squared Error (MSE)**:
    - **X-axis**: Alpha values (log scale).
    - **Y-axis**: Mean Squared Error (MSE).
    - **Key Observation**: The red dashed line indicates the optimal alpha value (2.656), minimizing the MSE. Beyond this value, the error increases as the model becomes overly regularized.

Analysis of Subset Selection Models (Based on BIC)

Observations:

1. **Single Variable Models:**
    - The model using `pH` as a single predictor has the **lowest BIC (677.46)** among single-variable models.
    - This suggests that `pH` alone is the most significant predictor compared to `SAL` , `K` , `Na` , or `Zn` .
2. **Two-Variable Models:**
    - The combination of `pH` and `Na` yields the lowest BIC (674.09) among all models, indicating this pair is the best fit.
    - However, the model combining `SAL` and `pH` has a close BIC of 680.81, making it a competitive choice if we prioritize including `SAL` .
3. **Three-Variable Models:**
    - The best three-variable combination is `pH` , `K` , and `Na` (BIC: 677.64).
    - It outperforms combinations involving `SAL` , further emphasizing `pH` as a core feature.

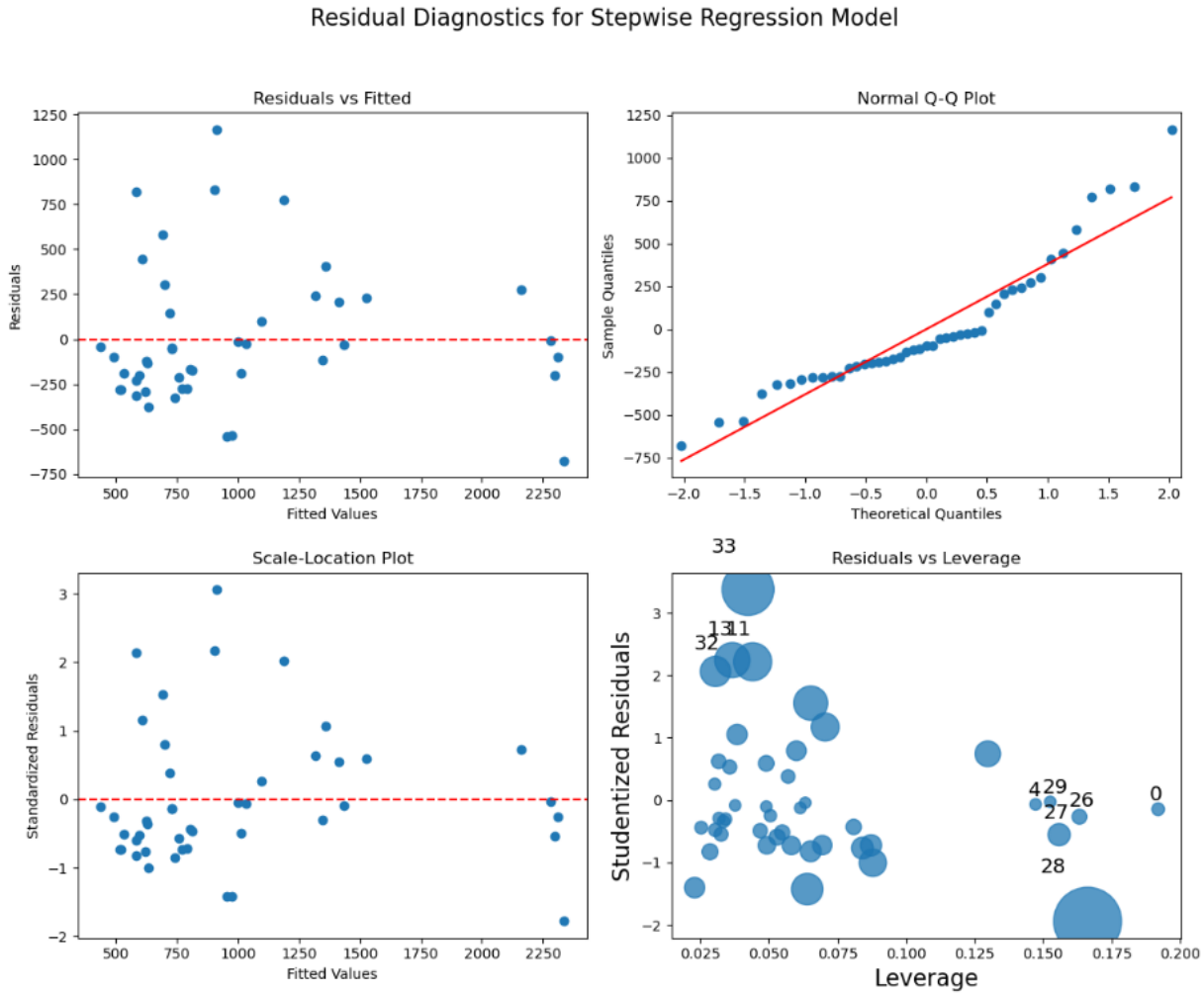
4. **Four and Five-Variable Models:**
    - Adding more variables beyond three slightly increases BIC, indicating overfitting or diminishing returns for model complexity.

Conclusion:

- **Best Overall Model:** The two-variable model with `pH` and `Na` is the optimal choice based on BIC (674.09).
- **Alternative Two-Variable Model:** The combination of `SAL` and `pH` is a solid alternative, especially if including `SAL` aligns with domain knowledge or interpretability goals.

```
Subset Selection Models (Based on BIC):
Model 1: ('SAL',), BIC: 718.1372831495836
Model 2: ('pH',), BIC: 677.4572036445928
Model 3: ('K',), BIC: 716.6936298908786
Model 4: ('Na',), BIC: 715.1581051817125
Model 5: ('Zn',), BIC: 696.3839493417659
Model 6: ('SAL', 'pH'), BIC: 680.8085136741549
Model 7: ('SAL', 'K'), BIC: 719.9551187756848
Model 8: ('SAL', 'Na'), BIC: 718.7905698980386
Model 9: ('SAL', 'Zn'), BIC: 686.2301820504596
Model 10: ('pH', 'K'), BIC: 675.4940736660518
Model 11: ('pH', 'Na'), BIC: 674.0860005873418
Model 12: ('pH', 'Zn'), BIC: 680.2510698772227
Model 13: ('K', 'Na'), BIC: 718.9493815606426
Model 14: ('K', 'Zn'), BIC: 698.2834338573199
Model 15: ('Na', 'Zn'), BIC: 697.1277584146529
Model 16: ('SAL', 'pH', 'K'), BIC: 678.7091509405648
Model 17: ('SAL', 'pH', 'Na'), BIC: 677.8116486504473
Model 18: ('SAL', 'pH', 'Zn'), BIC: 680.6753563450171
Model 19: ('SAL', 'K', 'Na'), BIC: 722.5967711393795
Model 20: ('SAL', 'K', 'Zn'), BIC: 687.5690359137295
Model 21: ('SAL', 'Na', 'Zn'), BIC: 688.832712984556
Model 22: ('pH', 'K', 'Na'), BIC: 677.6380987803094
Model 23: ('pH', 'K', 'Zn'), BIC: 678.720621298591
Model 24: ('pH', 'Na', 'Zn'), BIC: 677.3601939890351
Model 25: ('K', 'Na', 'Zn'), BIC: 700.9343401576419
Model 26: ('SAL', 'pH', 'K', 'Na'), BIC: 681.2663539335373
Model 27: ('SAL', 'pH', 'K', 'Zn'), BIC: 679.4815215599119
Model 28: ('SAL', 'pH', 'Na', 'Zn'), BIC: 679.9068914749192
Model 29: ('SAL', 'K', 'Na', 'Zn'), BIC: 691.2896504639349
Model 30: ('pH', 'K', 'Na', 'Zn'), BIC: 680.9370395214553
Model 31: ('SAL', 'pH', 'K', 'Na', 'Zn'), BIC: 682.94751108026

Best Two-Variable Model (Based on BIC):
('SAL', 'pH')
```



1. Residuals vs. Fitted Values (Top Left)

- **Purpose:** Checks for non-linearity and equal variance (homoscedasticity).
- **Observation:**
  - Ideally, residuals should scatter randomly around the red line (mean=0).
  - If there's a pattern (e.g., curve or funnel shape), it may indicate model misfit or heteroscedasticity.

2. Normal Q-Q Plot (Top Right)

- **Purpose:** Evaluates if residuals follow a normal distribution.
- **Observation:**

- Points should align closely with the red diagonal line. Deviations at the tails may indicate non-normality.

3. Scale-Location Plot (Bottom Left)

- **Purpose:** Checks for constant variance (homoscedasticity).
- **Observation:**
  - Points should form a horizontal band without a pattern. Increasing or decreasing spread may suggest heteroscedasticity.

4. Residuals vs. Leverage Plot (Bottom Right)

- **Purpose:** Identifies influential data points that disproportionately affect the model.
- **Observation:**
  - Look for points with high leverage (far from others on the x-axis) and high studentized residuals (y-axis). These are potential outliers or influential observations.

GitHub Link: https://github.com/Sahil-337/MATH564_Project/tree/main