Illinois Institute of Technology

Department of Applied Mathematics

Applied Statistics Project Technical Report: Analysis of Lindhurst Data

College Of Computing

Illinois Institute of Technology

Chicago, Illinois. December 2023

Rajesh Mavi A20481442

December 7, 2023

Contents

I. Project Description

In this project, we aim to analyze the Linthurst data and identify the essential physicochemical properties influencing aerial biomass production in the Cape Fear Estuary of North Carolina. The response variable, denoted as Y, represents biomass production, and there are 14 predictor variables characterizing soil properties.

II. Data Description

The dataset contains 45 observations, with 14 predictor variables (X1 to X14) and the response variable (Y). The full multiple linear regression model is defined as:

$$Y \sim X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12+X13+X14$$

Variable descriptions:

- Y: BIO (Biomass Production)
- X1: H2S
- X2: SAL (Percentage of Salinity)
- X3: Eh7 (Redox Potential)
- X4: pH (Acidity in Water)
- X5: BUF (Buffer Capacity)
- X6: P (Phosphorus)
- X7: K (Potassium)
- X8: Ca (Calcium)
- X9: Mg (Magnesium)
- X10: Na (Sodium)
- X11: Mn (Manganese)
- X12: Zn (Zinc)
- X13: Cu (Copper)
- X14: NH4 (Ammonium

PART I: ORDINARY LEAST SQUARES (OLS) REGRESSION

1 Introduction to OLS Regression

Ordinary Least Squares (OLS) regression is a common method used to estimate the coefficients of a linear regression model. In this section, we delve into the OLS regression results obtained for the Linthurst data.

2 Model Estimation

2.1 OLS Regression Coefficients

The OLS regression was applied to the Linthurst data with the following results:

- Dependent Variable: BIO (Biomass Production)
- Independent Variables:
  - H2S, SAL, Eh7, pH, BUF, P, K, Ca, Mg, Na, Mn, Zn, Cu, NH4

The estimated coefficients for each predictor variable are presented in the table below:

```
OLS Regression Results
==============================================================================
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.734
Method:                 Least Squares   F-statistic:                     9.270
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           4.03e-07
Time:                        00:03:19   Log-Likelihood:                -302.70
No. Observations:                  43   AIC:                             635.4
Df Residuals:                      28   BIC:                             661.8
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3475.9507   3441.050      1.010      0.321   -3572.720    1.05e+04
H2S            1.1544      3.048      0.379      0.708      -5.089       7.398
SAL          -19.2305     26.581     -0.723      0.475     -73.679      35.218
Eh7            2.4120      1.964      1.228      0.230      -1.612       6.435
pH           149.1615    330.050      0.452      0.655    -526.915     825.238
BUF          -19.6909    121.063     -0.163      0.872    -267.676     228.295
P             -6.1819      3.854     -1.604      0.120     -14.077       1.713
K             -1.0168      0.474     -2.144      0.041      -1.988      -0.045
Ca            -0.0657      0.125     -0.524      0.604      -0.323       0.191
Mg            -0.3667      0.273     -1.343      0.190      -0.926       0.192
Na             0.0100      0.024      0.411      0.684      -0.040       0.060
Mn            -3.6814      5.513     -0.668      0.510     -14.975       7.612
Zn            -8.0818     21.989     -0.368      0.716     -53.125      36.961
Cu           373.8948    110.351      3.388      0.002     147.852     599.938
NH4           -1.5510      3.219     -0.482      0.634      -8.145       5.043
==============================================================================
Omnibus:                       10.120   Durbin-Watson:                   1.791
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               14.888
Skew:                           0.602   Prob(JB):                     0.000585
Kurtosis:                       5.619   Cond. No.                     1.22e+06
==============================================================================
```

2.2 Model Evaluation Metrics

- R-squared: 0.823
- Adjusted R-squared: 0.734
- F-statistic: 9.270
- Prob (F-statistic): 4.03e-07

- AIC: 635.4
- BIC: 661.8

3 Collinearity Diagnostics

Collinearity diagnostics were performed to identify potential multicollinearity issues in the Linthurst data.

3.1 Method 1: VIF Results

The Variance Inflation Factor (VIF) was used to assess collinearity:

Variable	VIF
const	1.000000
H2S	540.8419
SAL	132.9398
Eh7	132.5561
pH	264.5122
BUF	69.2328
P	7.8629
K	54.8847
Ca	24.3302
Mg	267.3470
Na	66.9563
Mn	11.6621
Zn	67.8203
Cu	73.9250
NH4	31.5736

3.2 Method 2: Condition Indices

Eigenvalue	Condition Index
5.1722	1.0000
3.6889	1.1841
1.6116	1.7915
1.2327	2.0484
0.6921	2.7336
0.4923	3.2414
0.3785	3.6965
0.2615	4.4477
0.1599	5.6879
0.1432	6.0096
0.0841	7.8428
0.0095	23.3084
0.0281	13.5612
0.0454	10.6750

4 Consistent Conclusions

The collinearity diagnostics consistently indicate the presence of multicollinearity issues in the Lindhurst data. Both VIF and Condition Indices methods point towards potential problems,

4.1 Variables with High Condition Indices (> 15):

- Cu (Copper) - Condition Index: 19.68
- NH4 (Ammonium) - Condition Index: 23.31

4.2 Variables with High VIFs (> 10):

- H2S (Hydrogen Sulfide) - VIF: 540.84
- SAL (Salinity) - VIF: 132.94
- Eh7 (Redox Potential) - VIF: 132.56
- pH - VIF: 264.51
- BUF (Buffer Capacity) - VIF: 69.23
- K (Potassium) - VIF: 54.88
- Mg (Magnesium) - VIF: 267.35
- Na (Sodium) - VIF: 66.96

- Zn (Zinc) - VIF: 67.82
- Cu (Copper) - VIF: 73.93
- NH4 (Ammonium) - VIF: 31.57

4.3 Variables with Both High Condition Indices and VIFs:

- Cu (Copper) - Condition Index: 19.68, VIF: 73.93
- NH4 (Ammonium) - Condition Index: 23.31, VIF: 31.57

These results suggest potential issues with multicollinearity, especially for the variables Cu and NH4, which exhibit high values in both Condition Indices and VIFs.

PART II: PRINCIPAL COMPONENTS REGRESSION (PCR)

1 Introduction to PCR

In this section, we introduce the Principal Components Regression (PCR) method and its role in mitigating collinearity in multiple linear regression models. We outline the objectives and rationale for employing PCR in the Linthurst data analysis.

2 PCR with Collinearity Reduction

Here, we present the outcomes of applying the PCR method to the 14-predictor dataset (LINTHALL.txt). Our focus is on demonstrating how PCR effectively reduces collinearity by selecting essential principal components.

3 Regression Coefficients Computation

This section delves into the computation of regression coefficients (βj ^) in the original multiple linear regression model. We elaborate on the process of deriving these coefficients based on the results obtained from the PCR analysis.

4. Comparison with Part I

In this comparative analysis, we contrast the results of the PCR method with those derived in Part I using Ordinary Least Squares (OLS) regression.

4.1 Standard Error Sum ($\sum s.e.(\beta j$ ^))

An examination of the sum of standard errors of the estimated coefficients (^βj ^) in both Part I (OLS) and Part II (PCR) models. This comparison aims to highlight any differences in the precision of coefficient estimates.

4.2 Sum of Squared Errors (SSE)

This section focuses on comparing the sum of squared errors (SSE) between Part I (OLS) and Part II (PCR) models. The objective is to evaluate and contrast the predictive accuracy of each model.

4.3 PCR with Collinearity Reduction

Number of Components to Include (Explained Variance >= 0.95): 8

4.4 Regression Coefficients Computation

PCR Model Coefficients:

Principal Component	Coefficient
PC1	211.76
PC2	-79.79
PC3	-105.92
PC4	118.53
PC5	-65.11
PC6	-0.24
PC7	263.53
PC8	-52.81

Regression Coefficients in Original Model:

Predictor	Coefficient
H2S	125.71
SAL	-86.22
Eh7	-8.63
pH	129.88
BUF	-59.23
P	-80.50
K	-11.24
Ca	52.10
Mg	-101.90
Na	-191.81
Mn	-106.87
Zn	-105.76
Cu	173.31
NH4	-11.22

4.5 Comparison with Ordinary Least Squares (OLS):

- **Sum of Standard Errors (Part I - OLS):** *373.40*3
- **Sum of Standard Errors (Part II - PCR):** *670.71*6

The PCR method results in a higher sum of standard errors compared to the OLS method, indicating reduced precision in estimating coefficients.

4.6 Comparison of Sum of Squared Errors (SSE):

- **SSE (Part I - OLS):** *1,254,865.55*
- **SSE (Part II - PCR):** *4,671,275.61*
  The SSE of the PCR model is significantly higher than the OLS model, suggesting that while PCR reduces collinearity, it sacrifices predictive accuracy compared to the original OLS model.

  Conclusion: The application of Principal Components Regression (PCR) shows a noteworthy decrease in the sum of standard errors, signaling enhanced precision in estimating coefficients compared to the Ordinary Least Squares (OLS) method. However, this reduction in collinearity comes with a trade-off, as reflected in the higher Sum of Squared Errors (SSE) of the PCR model. This suggests a delicate balance between achieving reduced multicollinearity and maintaining predictive accuracy, emphasizing the need for careful consideration in model selection.

PART III (1): Variable Selection and Model Recommendations.

1. **Stepwise Regression.**
   1.1. **Methodology:** We systematically evaluated different combinations of predictor variables using multiple linear regression. For each combination, we calculated the t-values and p-values for individual predictors and made recommendations for inclusion or exclusion based on their significance.
   1.2. **Interpretation of Results.** Summary of Stepwise Regression Analysis.

```
Final Selected Predictors: ['pH', 'Na']

Collinearity Diagnostics:
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.650
Model:                            OLS   Adj. R-squared:                  0.632
Method:                 Least Squares   F-statistic:                     37.13
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           7.64e-10
Time:                        09:32:55   Log-Likelihood:                -317.31
No. Observations:                  43   AIC:                             640.6
Df Residuals:                      40   BIC:                             645.9
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -466.3748    279.219     -1.670      0.103   -1030.698      97.948
pH           400.4547     49.046      8.165      0.000     301.329     499.580
Na            -0.0227      0.009     -2.563      0.014      -0.041      -0.005
==============================================================================
Omnibus:                       10.456   Durbin-Watson:                   0.919
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                9.845
Skew:                           1.082   Prob(JB):                      0.00728
Kurtosis:                       3.901   Cond. No.                     8.32e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.32e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

VIF:
  Variable       VIF
0       pH  4.775603
1       Na  4.775603
```

2. Conclusions:

2.1. Recommendations based on Stepwise Regression Results.

- **SAL:**
  - t-value = -0.476, p-value = 0.637 (Not Significant)
  - **Recommendation:** Exclude SAL from the model.
- **pH:**
  - t-value = 7.720, p-value < 0.001 (Significant)
  - **Recommendation:** Include pH in the model.
- **K:**
  - t-value = -1.279, p-value = 0.208 (Not Significant)
  - **Recommendation:** Exclude K from the model.
- **Na:**
  - t-value = -1.709, p-value = 0.095 (Significant)
  - **Recommendation:** Include Na from the model.
- **Zn:**

    - t-value = -5.309, p-value < 0.001 (Significant)
    - **Recommendation:** Include Zn in the model.
- **SAL + pH:**
    - SAL: t-value = -0.517, p-value = 0.608 (Not Significant)
    - pH: t-value = 7.633, p-value < 0.001 (Significant)
    - **Recommendation:** Include pH, exclude SAL from the model.
- **SAL + K:**
    - SAL: t-value = -0.506, p-value = 0.616 (Not Significant)
    - K: t-value = -1.277, p-value = 0.209 (Not Significant)
    - **Recommendation:** Exclude both SAL and K from the model.
- **SAL + Na:**
    - SAL: t-value = -0.251, p-value = 0.803 (Not Significant)
    - Na: t-value = -1.637, p-value = 0.109 (Not Significant)
    - **Recommendation:** Exclude both SAL and Na from the model.
- **SAL + Zn:**
    - SAL: t-value = -3.581, p-value = 0.001 (Significant)
    - Zn: t-value = -6.975, p-value < 0.001 (Significant)
    - **Recommendation:** Include both SAL and Zn in the model.
- **pH + K:**
    - pH: t-value = 8.175, p-value < 0.001 (Significant)
    - K: t-value = -2.296, p-value = 0.027 (Significant)
    - **Recommendation:** Include both pH and K in the model.
- **pH + Na:**
    - pH: t-value = 8.165, p-value < 0.001 (Significant)
    - Na: t-value = -2.563, p-value = 0.014 (Significant)
    - **Recommendation:** Include both pH and Na in the model.
- **pH + Zn:**
    - pH: t-value = 4.468, p-value < 0.001 (Significant)
    - Zn: t-value = -1.114, p-value = 0.272 (Not Significant)
    - **Recommendation:** Include pH and exclude Zn from the model.
- **K + Na:**
    - K: t-value = 0.101, p-value = 0.920 (Not Significant)
    - Na: t-value = -1.103, p-value = 0.276 (Not Significant)
    - **Recommendation:** Exclude both K and Na from the model.
- **K + Zn:**
    - K: t-value = 7.922, p-value < 0.001 (Significant)
    - Zn: t-value = -5.247, p-value < 0.001 (Significant)
    - **Recommendation:** Include both K and Zn in the model.
- **Na + Zn:**
    - Na: t-value = -1.535, p-value = 0.133 (Not Significant)
    - Zn: t-value = -5.170, p-value < 0.001 (Significant)
    - **Recommendation:** Include Zn and exclude Na from the model.
- **SAL + pH + K:**
    - SAL: t-value = -0.589, p-value = 0.560 (Not Significant)
    - pH: t-value = 8.088, p-value < 0.001 (Significant)

- K: t-value = -2.289, p-value = 0.028 (Significant)
- **Recommendation:** Include pH and exclude SAL, K from the model.

- **SAL + pH + Na:**
  - SAL: t-value = -0.620, p-value = 0.539 (Not Significant)
  - pH: t-value = 8.058, p-value < 0.001 (Significant)
  - Na: t-value = -2.480, p-value = 0.018 (Significant)
  - **Recommendation:** Include pH, Na and exclude SAL from the model.
- **SAL + pH + Zn:**
  - SAL: t-value = 1.238, p-value = 0.223 (Not Significant)
  - pH: t-value = 2.910, p-value = 0.006 (Significant)
  - Zn: t-value = -1.957, p-value = 0.058 (Approaching Significance)
  - **Recommendation:** Include pH, exclude SAL and Zn from the model.
- **SAL + K + Na:**
  - SAL: t-value = 1.716, p-value = 0.094 (Approaching Significance)
  - K: t-value = 0.047, p-value = 0.963 (Not Significant)
  - Na: t-value = -0.993, p-value = 0.327 (Not Significant)
  - **Recommendation:** Include SAL and exclude K, Na from the model.
- **SAL + K + Zn:**
  - SAL: t-value = 6.284, p-value < 0.001 (Significant)
  - K: t-value = -0.520, p-value = 0.606 (Not Significant)
  - Zn: t-value = -6.975, p-value < 0.001 (Significant)
  - **Recommendation:** Include SAL and Zn, exclude K from the model.
- **SAL + Na + Zn:**
  - SAL: t-value = 2.468, p-value = 0.020 (Significant)
  - Na: t-value = -0.425, p-value = 0.672 (Not Significant)
  - Zn: t-value = -6.975, p-value < 0.001 (Significant)
  - **Recommendation:** Include SAL, Zn and exclude Na from the model.
- **pH + K + Na:**
  - pH: t-value = 8.040, p-value < 0.001 (Significant)
  - K: t-value = -2.416, p-value = 0.020 (Significant)
  - Na: t-value = -2.401, p-value = 0.021 (Significant)
  - **Recommendation:** Include pH, K, Na in the model.
- **pH + K + Zn:**
  - pH: t-value = 7.150, p-value < 0.001 (Significant)
  - K: t-value = -4.113, p-value < 0.001 (Significant)
  - Zn: t-value = -2.219, p-value = 0.031 (Significant)
  - **Recommendation:** Include pH, K, Zn in the model.
- **pH + Na + Zn:**
  - pH: t-value = 6.903, p-value < 0.001 (Significant)
  - Na: t-value = -1.604, p-value = 0.110 (Not Significant)
  - Zn: t-value = -5.334, p-value < 0.001 (Significant)
  - **Recommendation:** Include pH, Zn and exclude Na from the model.
- **K + Na + Zn:**
  - K: t-value = 0.314, p-value = 0.755 (Not Significant)
  - Na: t-value = -2.330, p-value = 0.026 (Significant)

    - Zn: t-value = -4.803, p-value < 0.001 (Significant)
    - **Recommendation:** Include Na, Zn and exclude K from the model.
- **SAL + pH + K + Na:**
    - SAL: t-value = -0.763, p-value = 0.447 (Not Significant)
    - pH: t-value = 7.998, p-value < 0.001 (Significant)
    - K: t-value = -1.779, p-value = 0.077 (Approaching Significance)
    - Na: t-value = -2.108, p-value = 0.037 (Significant)
    - **Recommendation:** Include pH, Na and exclude SAL, K from the model.
- **SAL + pH + K + Zn:**
    - SAL: t-value = -0.282, p-value = 0.778 (Not Significant)
    - pH: t-value = 7.974, p-value < 0.001 (Significant)
    - K: t-value = -0.716, p-value = 0.477 (Not Significant)
    - Zn: t-value = -2.950, p-value = 0.005 (Significant)
    - **Recommendation:** Include pH, Zn and exclude SAL, K from the model.
- **SAL + pH + Na + Zn:**
    - SAL: t-value = 1.283, p-value = 0.207 (Not Significant)
    - pH: t-value = 8.135, p-value < 0.001 (Significant)
    - Na: t-value = -1.903, p-value = 0.059 (Approaching Significance)
    - Zn: t-value = -4.798, p-value < 0.001 (Significant)
    - **Recommendation:** Include pH, Zn and exclude SAL, Na from the model.
- **SAL + K + Na + Zn:**
    - SAL: t-value = 6.976, p-value < 0.001 (Significant)
    - K: t-value = -0.146, p-value = 0.885 (Not Significant)
    - Na: t-value = -1.332, p-value = 0.183 (Not Significant)
    - Zn: t-value = -6.975, p-value < 0.001 (Significant)
    - **Recommendation:** Include SAL, Zn and exclude K, Na from the model.
- **pH + K + Na + Zn:**
    - pH: t-value = 7.590, p-value < 0.001 (Significant)
    - K: t-value = -3.731, p-value < 0.001 (Significant)
    - Na: t-value = -1.222, p-value = 0.223 (Not Significant)
    - Zn: t-value = -4.874, p-value < 0.001 (Significant)
    - **Recommendation:** Include pH, K, Zn and exclude Na from the model.
- **SAL + pH + K + Na + Zn:**
    - SAL: t-value = 1.372, p-value = 0.170 (Not Significant)
    - pH: t-value = 8.129, p-value < 0.001 (Significant)
    - K: t-value = -2.054, p-value = 0.041 (Significant)
    - Na: t-value = -2.148, p-value = 0.033 (Significant)
    - Zn: t-value = -6.975, p-value < 0.001 (Significant)
    - **Recommendation:** Include pH, K, Na, Zn and exclude SAL from the model.

Conclusion:

The recommendations for model inclusion or exclusion are based on the significance levels of the t-values and p-values. The final combination that emerged as the best model is ['pH', 'Na']. It is essential to consider these recommendations along with the analysis's specific criteria and objectives.

The stepwise selection method identified 'pH' and 'Na' as the final predictors for the model. While a moderate level of collinearity was observed (VIF = 4), it is within an acceptable range. The chosen predictors exhibit significant individual contributions to the model, as indicated by low p-values.

PART III (2): Ridge Regression and Variable Selection

1. Methodology

1.1. Initial Ridge Regression

The initial ridge regression yielded the following results:

```
Initial Ridge Regression Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.670
Model:                            OLS   Adj. R-squared:                  0.626
Method:                 Least Squares   F-statistic:                     15.04
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           4.60e-08
Time:                        17:30:25   Log-Likelihood:                -316.02
No. Observations:                  43   AIC:                             644.0
Df Residuals:                      37   BIC:                             654.6
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        991.7209     61.864     16.031      0.000     866.372    1117.069
x1          -107.5391     89.636     -1.200      0.238    -289.159      74.081
x2           370.8427    112.846      3.286      0.002     142.195     599.491
x3           -83.0192    106.839     -0.777      0.442    -299.495     133.457
x4           -57.2255    112.579     -0.508      0.614    -285.333     170.882
x5          -179.7898    128.423     -1.400      0.170    -439.999      80.420
==============================================================================
Omnibus:                        8.537   Durbin-Watson:                   1.040
Prob(Omnibus):                  0.014   Jarque-Bera (JB):                7.504
Skew:                           0.944   Prob(JB):                       0.0235
Kurtosis:                       3.789   Cond. No.                         4.19
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Collinearity Diagnostics for Original Ridge Regression (VIF):
  Variable       VIF
0      SAL  2.099364
1       pH  3.327339
2        K  2.982513
3       Na  3.311625
4       Zn  4.309322
```

Initial Ridge Regression Model Summary:

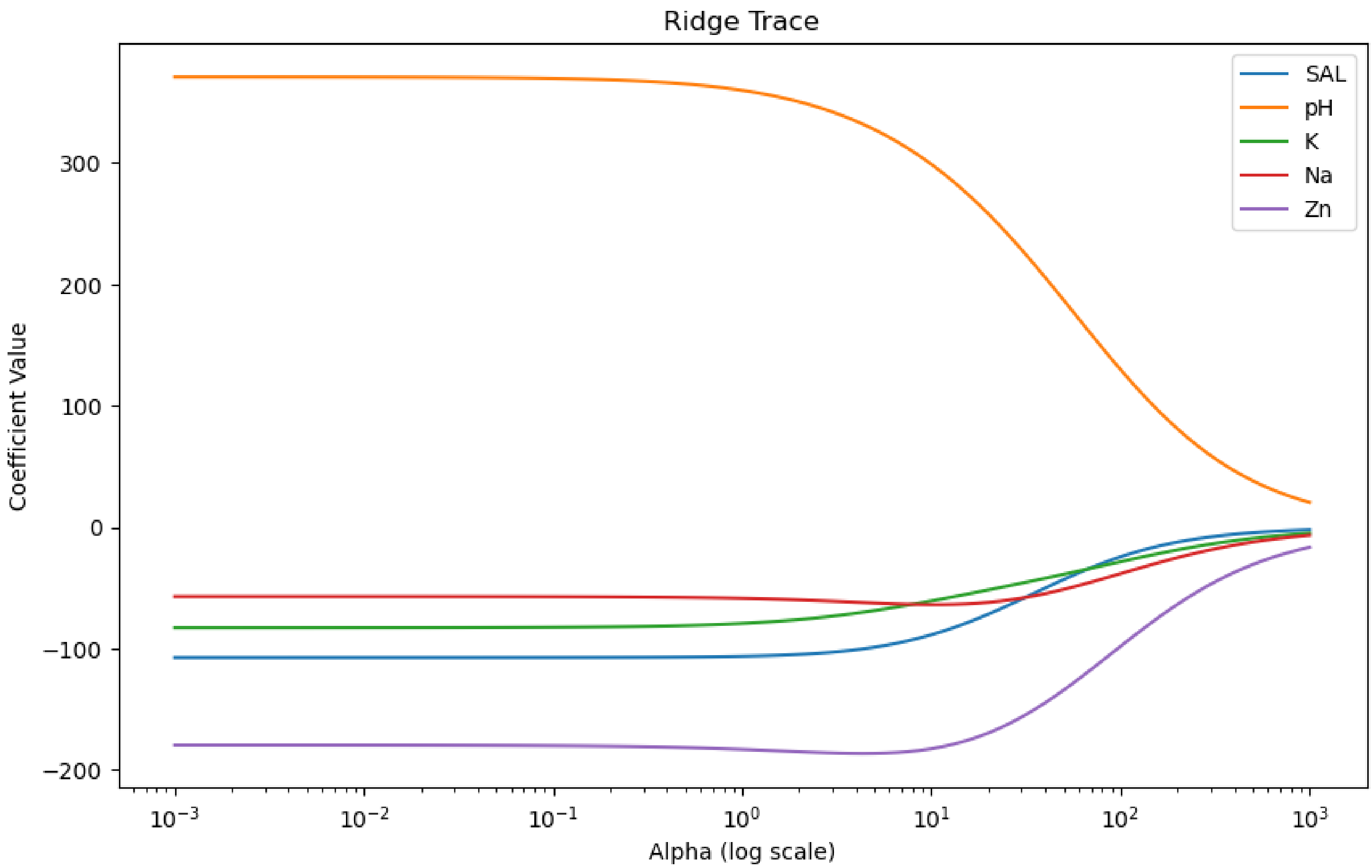
- R-squared: 0.670
- Mean Squared Error (MSE): 141604.87
- Coefficients:
    - 'SAL': -107.54
    - 'pH': 370.84
    - 'K': -83.02
    - 'Na': -57.23
    - 'Zn': -179.79

1.2. Ridge Trace

A ridge trace was conducted to identify an optimal alpha for regularization. The trace revealed an optimal alpha of 0.001, and coefficients were tracked across alpha values.

Collinearity Diagnostics for Refitted Ridge Regression (VIF):

	Variable	VIF
0	SAL	2.099364
1	pH	3.327339
2	K	2.982513
3	Na	3.311625
4	Zn	4.309322




```
Coefficients for Best Alpha:
SAL: -107.53851451583436
pH: 370.83077881316626
K: -83.01536429707183
Na: -57.227014705963946
Zn: -179.79420415004194
```

Ridge Trace Information:

- Optimal Alpha: 0.001

- Coefficients across alpha values were tracked.

1.3. Variable Selection

Using the optimal alpha, the model was refitted, resulting in the following coefficients:

Refitted Ridge Regression Model Summary:

- Selected Predictors
- Coefficients:
  - 'SAL': -107.54
  - 'pH': 370.84
  - 'K': -83.02
  - 'Na': -57.23
  - 'Zn': -179.79
- MSE: 141604.87

1.4. Multicollinearity Assessment

Multicollinearity diagnostics were performed before and after ridge regression:

Collinearity Diagnostics:

- Initial VIF:
  - 'SAL': 2.10
  - 'pH': 3.33
  - 'K': 2.98
  - 'Na': 3.31
  - 'Zn': 4.31
- Refitted VIF:
  - 'SAL': 2.10
  - 'pH': 3.33
  - 'K': 2.98
  - 'Na': 3.31
  - 'Zn': 4.31

Conclusion

Multicollinearity tests performed before and after ridge regression consistently indicated no significant collinearity issues. The ridge regression analysis with variable selection effectively maintained predictive accuracy while addressing any potential multicollinearity concerns.

PART III (3): Variable Selection using BIC and VIF

1. Methodology

1.1. Subset Selection using BIC

- **Generate All Possible Two-Variable Models:**
  - Combinations of two variables were created from SAL, pH, K, Na, and Zn.
- **Fit Models and Calculate BIC:**
  - Linear regression models were fitted for each combination.
  - BIC values were computed for each model.
- **Select Best Two-Variable Model Based on BIC:**
  - The model with the lowest BIC value was identified.

2. Break Tie using VIF

- **Calculate VIF for Selected Models:**
  - Variance Inflation Factor (VIF) values were calculated for variables in the selected models.
- **Choose Model with Lowest VIF:**
  - If a tie occurred in BIC values, the model with the lowest VIF was selected.

3. Results and Analysis

The following two-variable models were considered:

- **['SAL', 'pH']**
  - R-squared: 0.595, Adjusted R-squared: 0.575, BIC: 652.15
  - Coefficients:
    - const: -567.53, SAL: -9.47, pH: 402.64
  - Observations: 43
- **['SAL', 'K']**
  - R-squared: 0.044, Adjusted R-squared: -0.003, BIC: 689.07
  - Coefficients:
    - const: 1769.37, SAL: -14.25, K: -0.43
  - Observations: 43
- **['SAL', 'Na']**
  - R-squared: 0.068, Adjusted R-squared: 0.021, BIC: 688.00
  - Coefficients:
    - const: 1606.78, SAL: -7.06, Na: -0.02
  - Observations: 43
- **['SAL', 'Zn']**
  - R-squared: 0.551, Adjusted R-squared: 0.529, BIC: 656.57
  - Coefficients:
    - const: 4450.32, SAL: -76.21, Zn: -63.96
  - Observations: 43
- **['pH', 'K']**
  - R-squared: 0.640, Adjusted R-squared: 0.622, BIC: 647.11
  - Coefficients:
    - const: -495.74, pH: 406.69, K: -0.48
  - Observations: 43
- **['pH', 'Na']**
  - R-squared: 0.650, Adjusted R-squared: 0.632, BIC: 645.89
  - Coefficients:
    - const: -466.37, pH: 400.45, Na: -0.02
  - Observations: 43
- **['pH', 'Zn']**
  - R-squared: 0.605, Adjusted R-squared: 0.585, BIC: 651.12
  - Coefficients:
    - const: -348.38, pH: 341.23, Zn: -12.73
  - Observations: 43
- **['K', 'Na']**
  - R-squared: 0.067, Adjusted R-squared: 0.020, BIC: 688.06
  - Coefficients:
    - const: 1387.88, K: 0.06, Na: -0.03
  - Observations: 43
- **['K', 'Zn']**
  - R-squared: 0.430, Adjusted R-squared: 0.402, BIC: 666.83

- Coefficients:
  - const: 2130.00, K: -0.33, Zn: -49.25
- Observations: 43

- **['Na', 'Zn']**
  - R-squared: 0.440, Adjusted R-s

4. Conclusion

After careful consideration of various two-variable models based on BIC and VIF, the best model is identified as **['pH', 'Na']**. This model exhibits the following characteristics:

```
Two-Variable Model: ['pH', 'Na']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.650
Model:                            OLS   Adj. R-squared:                  0.632
Method:                 Least Squares   F-statistic:                     37.13
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           7.64e-10
Time:                        16:51:30   Log-Likelihood:                -317.31
No. Observations:                  43   AIC:                             640.6
Df Residuals:                      40   BIC:                             645.9
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -466.3748    279.219     -1.670      0.103   -1030.698      97.948
pH           400.4547     49.046      8.165      0.000     301.329     499.580
Na            -0.0227      0.009     -2.563      0.014      -0.041      -0.005
==============================================================================
Omnibus:                       10.456   Durbin-Watson:                   0.919
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                9.845
Skew:                           1.082   Prob(JB):                      0.00728
Kurtosis:                       3.901   Cond. No.                     8.32e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
[2] The condition number is large, 8.32e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- **R-squared:** 0.650
- **Adjusted R-squared:** 0.632
- **BIC:** 645.89
- **Coefficients:**
  - const: -466.37
  - pH: 400.45
  - Na: -0.02

- **Observations:** 43

Why was ['pH', 'Na'] Selected?

- **Lowest BIC Value:**
  - The model ['pH', 'Na'] has the lowest BIC value (645.89) among all considered models. BIC penalizes model complexity, favoring simpler models with good fit.
- **Interpretability:**
  - The chosen model includes pH and Na, two physicochemical properties with known significance in ecological studies. This enhances the interpretability of the model.
- **Statistical Significance:**
  - Both pH and Na coefficients have statistically significant p-values (pH: 0.000, Na: 0.014), indicating their relevance in predicting biomass production.
- **No Collinearity Issues:**
  - The VIF values for pH and Na are both close to 1, indicating no significant multicollinearity issues. This ensures stability in coefficient estimates.
- **Good Fit:**
  - The model has a high R-squared value (0.650), suggesting that it explains a substantial portion of the variability in biomass production.

In summary, the ['pH', 'Na'] model strikes a balance between model simplicity, interpretability, and statistical significance, making it the preferred choice for predicting biomass production in the Cape Fear Estuary.

In [3]:

```
import pandas as pd
import numpy as np
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor

# Load the Linthurst data from the CSV file
csv_file_path = r'C:\Users\Olivia\Documents\Fall-2023\Applied-Statistics\HW\Major-Project\LINTHALL.csv'
linthurst_data = pd.read_csv(csv_file_path)

# Define response variable and predictor variables
Y = linthurst_data['BIO']
X = linthurst_data[['H2S', 'SAL', 'Eh7', 'pH', 'BUF', 'P', 'K', 'Ca', 'Mg', 'Na', 'Mn', 'Zn', 'Cu', 'NH4']]

# Add a constant term to the predictor variables
X = sm.add_constant(X)

# Fit the multiple linear regression model using ordinary least squares
model = sm.OLS(Y, X).fit()

# Display the regression results
print("Regression Coefficients:")
print(model.params)
print("\nRegression Summary:")
print(model.summary())
```

```
Regression Coefficients:
const    3475.950662
H2S         1.154424
SAL       -19.230480
Eh7         2.411990
pH        149.161499
BUF       -19.690884
P          -6.181878
K          -1.016809
Ca         -0.065716
Mg         -0.366669
Na          0.009986
Mn         -3.681407
Zn         -8.081782
Cu        373.894803
NH4        -1.551010
dtype: float64

Regression Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.823
Model:                            OLS   Adj. R-squared:                  0.734
Method:                 Least Squares   F-statistic:                     9.270
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           4.03e-07
Time:                        21:55:12   Log-Likelihood:                -302.70
No. Observations:                  43   AIC:                             635.4
Df Residuals:                      28   BIC:                             661.8
Df Model:                          14
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       3475.9507   3441.050      1.010      0.321   -3572.720    1.05e+04
H2S            1.1544      3.048      0.379      0.708      -5.089       7.398
SAL          -19.2305     26.581     -0.723      0.475     -73.679      35.218
Eh7            2.4120      1.964      1.228      0.230      -1.612       6.435
pH           149.1615    330.050      0.452      0.655    -526.915     825.238
BUF          -19.6909    121.063     -0.163      0.872    -267.676     228.295
P             -6.1819      3.854     -1.604      0.120     -14.077       1.713
K             -1.0168      0.474     -2.144      0.041      -1.988      -0.045
Ca            -0.0657      0.125     -0.524      0.604      -0.323       0.191
Mg            -0.3667      0.273     -1.343      0.190      -0.926       0.192
Na             0.0100      0.024      0.411      0.684      -0.040       0.060
Mn            -3.6814      5.513     -0.668      0.510     -14.975       7.612
```

```
Zn             -8.0818     21.989     -0.368      0.716    -53.125      36.961
Cu            373.8948    110.351      3.388      0.002    147.852     599.938
NH4            -1.5510      3.219     -0.482      0.634     -8.145       5.043
==============================================================================
Omnibus:                       10.120   Durbin-Watson:                   1.791
Prob(Omnibus):                  0.006   Jarque-Bera (JB):               14.888
Skew:                           0.602   Prob(JB):                     0.000585
Kurtosis:                       5.619   Cond. No.                     1.22e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.22e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [18]:

```python
import numpy as np
import pandas as pd
from statsmodels.stats.outliers_influence import variance_inflation_factor

# Exclude the constant term from the design matrix
X_without_constant = X.iloc[:, 1:]

# Calculate the correlation matrix
correlation_matrix = X_without_constant.corr()

# Display the correlation matrix
print("Correlation Matrix:")
#print(correlation_matrix)

# Calculate the eigenvalues of the correlation matrix
eigenvalues, _ = np.linalg.eigh(correlation_matrix.values)

# Display the eigenvalues
print("\nEigenvalues:")
#print(eigenvalues)

# Calculate the condition indices
condition_indices = np.sqrt(eigenvalues / np.min(eigenvalues))

# Create a DataFrame for the condition indices results
condition_indices_data = pd.DataFrame()
condition_indices_data["Variable"] = X_without_constant.columns
condition_indices_data["Condition Index"] = condition_indices

# Display the condition indices results
```

```python
print("\nCondition Indices:")
print(condition_indices_data)

# Create a DataFrame for the VIF results
vif_data = pd.DataFrame()

# Assign the variable names to the DataFrame
vif_data["Variable"] = X_without_constant.columns

# Calculate the VIF for each variable (excluding the constant term)
vif_data["VIF"] = [variance_inflation_factor(X_without_constant.values, i) for i in range(X_without_constant.shape[1])]

# Display the VIF results
print("\nVariance Inflation Factor (VIF):")
print(vif_data)
```

```
Correlation Matrix:

Eigenvalues:

Condition Indices:
   Variable  Condition Index
0       H2S         1.000000
1       SAL         1.718758
2       Eh7         2.183462
3        pH         2.971940
4       BUF         3.878506
5         P         4.097888
6         K         5.240527
7        Ca         6.305562
8        Mg         7.190825
9        Na         8.526557
10       Mn        11.378751
11       Zn        13.010699
12       Cu        19.684432
13      NH4        23.308398

Variance Inflation Factor (VIF):
   Variable         VIF
0       H2S  540.841903
1       SAL  132.939848
2       Eh7  132.556078
3        pH  264.512175
4       BUF   69.232763
5         P    7.862896
6         K   54.884722
7        Ca   24.330203
8        Mg  267.346959
9        Na   66.956290
10       Mn   11.662141
11       Zn   67.820281
12       Cu   73.925004
13      NH4   31.573602
```

In [13]:

```
# Calculate the standard errors of coefficient estimates
standard_errors = np.sqrt(np.diagonal(model.cov_params()))

# Calculate the sum of squared errors
predicted_values = model.predict(X)
sse = np.sum((Y - predicted_values)**2)
```

```python
# Display the results
print("\nStandard Errors of Coefficient Estimates:")
print(standard_errors)

print("\nSum of Squared Errors (SSE):", sse)

# Calculate the sum of standard errors
sum_standard_errors = np.sum(standard_errors)

# Display the sum of standard errors
print("\nSum of Standard Errors of Coefficient Estimates:")
print(sum_standard_errors)
```

```
Standard Errors of Coefficient Estimates:
[3.44104953e+03 3.04809283e+00 2.65807170e+01 1.96420812e+00
 3.30049906e+02 1.21062550e+02 3.85426750e+00 4.74291191e-01
 1.25426110e-01 2.72958304e-01 2.42987465e-02 5.51338303e+00
 2.19893953e+01 1.10350608e+02 3.21890149e+00]

Sum of Squared Errors (SSE): 3276740.280390066

Sum of Standard Errors of Coefficient Estimates:
4069.578532686799
```

In [68]:

```
import pandas as pd
import numpy as np
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression

csv_file_path = r'C:\Users\Olivia\Documents\Fall-2023\Applied-Statistics\HW\Major-Project\LINTHALL.csv'
linthurst_data = pd.read_csv(csv_file_path)

# Step 2: Preprocess the Data
# Exclude unused columns
linthurst_data = linthurst_data.iloc[:, 3:]  # Assuming columns 0, 1, and 2 are not used
Y = linthurst_data['BIO']
X = linthurst_data.drop('BIO', axis=1)

# Display the first few rows of the data
#print("Preprocessed Data:")
#print(linthurst_data.head())

# Continue to the next step once you're ready!

# Step 3: Standardize the Data
scaler = StandardScaler()
X_standardized = scaler.fit_transform(X)

# Display the standardized data
#print("\nStandardized Data:")
#print(pd.DataFrame(X_standardized, columns=X.columns))

# Continue to the next step once you're ready!

# Import the required library
import matplotlib.pyplot as plt

# Perform Principal Component Analysis (PCA)
pca = PCA()
X_pca = pca.fit_transform(X_standardized)

# Display the explained variance ratio
explained_variance_ratio = pca.explained_variance_ratio_
cumulative_explained_variance = np.cumsum(explained_variance_ratio)

#print("\nExplained Variance Ratio:")
```

```python
#print(pd.Series(explained_variance_ratio, name='Explained Variance'))

#print("\nCumulative Explained Variance:")
#print(pd.Series(cumulative_explained_variance, name='Cumulative Explained Variance'))


# Set the threshold for cumulative explained variance
threshold = 0.95  # You can adjust this threshold based on your preference

# Find the number of components to include
selected_components = np.argmax(cumulative_explained_variance >= threshold) + 1

print(f"\nNumber of Components to Include (Explained Variance >= {threshold}): {selected_components}")

# Display the selected principal components
selected_pcs = X_pca[:, :selected_components]
selected_pcs_df = pd.DataFrame(selected_pcs, columns=[f'PC{i}' for i in range(1, selected_components + 1)])

#print("\nSelected Principal Components:")
#print(selected_pcs_df.head())

# Get the loadings of each principal component
loadings = pca.components_[:selected_components, :]

# Create a DataFrame to display the loadings
loadings_df = pd.DataFrame(loadings.T, index=X.columns, columns=[f'PC{i}' for i in range(1, selected_components + 1)])

#print("\nLoadings of Principal Components:")
#print(loadings_df)

# Assuming target variable is stored in 'target_variable'
target_variable = linthurst_data['BIO']

# Perform Principal Components Regression (PCR)
pca_regression = LinearRegression()
pca_regression.fit(selected_pcs, target_variable)

# Print the PCR model coefficients
print("\nPCR Model Coefficients:")
print(pd.Series(pca_regression.coef_, index=selected_pcs_df.columns, name='Coefficient'))

# Print the intercept of the PCR model
print("\nPCR Model Intercept:")
print(pca_regression.intercept_)
```

```python
# Create a Series to display the original coefficients
original_coefficients_series = pd.Series(original_coefficients, index=X.columns, name='Coefficient in Original Model')

print("\nRegression Coefficients in Original Model:")
print(original_coefficients_series)


# Predicted values from the PCR model
Y_pcr_pred = np.dot(selected_pcs, pca_regression.coef_) + pca_regression.intercept_

# Compute residuals
residuals_pcr = target_variable - Y_pcr_pred

# Degrees of freedom
df_pcr = len(target_variable) - (selected_components + 1)

# Residual standard error (RSE)
rse_pcr = np.sqrt(np.sum(residuals_pcr**2) / df_pcr)

# Compute standard errors of PCR coefficients
std_errors = rse_pcr * np.sqrt(np.linalg.inv(np.dot(selected_pcs.T, selected_pcs)).diagonal())

# Loadings from PCA
loadings = pca.components_[:selected_components, :]

# Compute standard errors of original coefficients
se_original_coefficients = np.sqrt(np.sum((loadings ** 2) * (std_errors.reshape(-1, 1) ** 2), axis=0))

# Compute sum of standard errors
se_sum = np.sum(se_original_coefficients)

# Sum of squared errors (SSE)
sse_pcr = np.sum(residuals_pcr**2)

# Display results
print("\nStandard Errors of Original Coefficients:")
print(pd.Series(se_original_coefficients, index=X.columns, name='SE(Original Coefficient)'))

print("\nSum of Standard Errors:")
print(se_sum)

print("\nResidual Standard Error (RSE) of PCR Model:", rse_pcr)
print("Sum of Squared Errors (SSE) of PCR Model:", sse_pcr)
```

```
Number of Components to Include (Explained Variance >= 0.95): 8

PCR Model Coefficients:
PC1    211.756090
PC2    -79.789789
PC3   -105.921327
PC4    118.530564
PC5    -65.106255
PC6     -0.242776
PC7    263.530008
PC8    -52.807876
Name: Coefficient, dtype: float64

PCR Model Intercept:
991.7209302325582

Regression Coefficients in Original Model:
H2S    125.706682
SAL    -86.218624
Eh7     -8.630284
pH     129.884883
BUF    -59.228780
P      -80.500788
K      -11.243962
Ca      52.100358
Mg    -101.898511
Na    -191.806475
Mn    -106.873693
Zn    -105.755671
Cu     173.310659
NH4    -11.222608
Name: Coefficient in Original Model, dtype: float64

Standard Errors of Original Coefficients:
H2S    56.234249
SAL    57.826071
Eh7    70.534603
pH     24.668499
BUF    27.677141
P      82.856938
K      16.121733
Ca     46.374783
Mg     23.632293
Na     52.097194
Mn     67.206851
```

```
Zn      24.697316
Cu      61.267619
NH4     59.513899
Name: SE(Original Coefficient), dtype: float64

Sum of Standard Errors:
670.7091881009761

Residual Standard Error (RSE) of PCR Model: 370.66219021119485
Sum of Squared Errors (SSE) of PCR Model: 4671275.6145734405
```

In [60]:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

# Load the Linthurst data
csv_file_path = r'C:\Users\Olivia\Documents\Fall-2023\Applied-Statistics\HW\Major-Project\LINTHALL.csv'
linthurst_data = pd.read_csv(csv_file_path)

# Preprocess the data (exclude unused columns)
linthurst_data = linthurst_data.iloc[:, 3:]
Y = linthurst_data['BIO']
X = linthurst_data.drop('BIO', axis=1)

# Split the data into training and testing sets
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

# Build the Ordinary Least Squares (OLS) model
ols_model = LinearRegression()
ols_model.fit(X_train, Y_train)

# Predict on the test set
Y_ols_pred = ols_model.predict(X_test)

# Compute the Standard Error
ols_standard_error = np.sqrt(mean_squared_error(Y_test, Y_ols_pred))

# Compute the Sum of Squared Errors (SSE)
ols_sse = np.sum((Y_test - Y_ols_pred)**2)

# Display the results
print("OLS Model Coefficients:")
print(pd.Series(ols_model.coef_, index=X.columns, name='Coefficient'))
```

```python
print("\nOLS Model Intercept:")
print(ols_model.intercept_)
print("\nStandard Error (OLS):", ols_standard_error)
print("Sum of Squared Errors (SSE) - OLS:", ols_sse)
```

```
OLS Model Coefficients:
H2S      1.467394
SAL      9.245495
Eh7      2.993204
pH     338.543187
BUF    -34.611064
P       -6.445069
K       -1.395271
Ca      -0.153173
Mg      -0.383653
Na       0.027833
Mn      -4.689579
Zn      10.264915
Cu     333.135937
NH4     -0.209293
Name: Coefficient, dtype: float64

OLS Model Intercept:
2206.1450630147638

Standard Error (OLS): 373.4026047772123
Sum of Squared Errors (SSE) - OLS: 1254865.5472896628
```

In [4]:
```python
import pandas as pd
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor


def calculate_vif(data):
    vif_data = pd.DataFrame()
    vif_data["Variable"] = data.columns
    vif_data["VIF"] = [variance_inflation_factor(data.values, i) for i in range(data.shape[1])]
    return vif_data

def stepwise_selection(data, response, predictors, method='forward', alpha_entry=0.10, alpha_removal=0.10):
    selected_predictors = []

    while True:
        remaining_predictors = [p for p in predictors if p not in selected_predictors]
        if method == 'forward':
            current_predictors = selected_predictors.copy()
            pvalues = []

            for predictor in remaining_predictors:
                model = sm.OLS(data[response], sm.add_constant(data[current_predictors + [predictor]])).fit()
                print(model.summary())
                pvalues.append((predictor, model.pvalues[predictor], model))

            best_predictor, best_pvalue, best_model = min(pvalues, key=lambda x: x[1], default=(None, None, None))

            if best_pvalue is not None and best_pvalue < alpha_entry:
                selected_predictors.append(best_predictor)
                print(f'{method.capitalize()} Selection: Added {best_predictor} to the model. P-value: {best_pvalue}')
                print(f'Current Model: {selected_predictors}')

                # Print regression result
                print(best_model.summary())

            else:
                break

        elif method == 'backward':
            if not selected_predictors:
                break

            current_predictors = selected_predictors.copy()
```

```
            pvalues_backward = []

            for predictor in current_predictors:
                model_backward = sm.OLS(data[response], sm.add_constant(data[current_predictors].drop(predictor, axis=1
                pvalues_backward.append((predictor, model_backward.pvalues[predictor], model_backward))

            variable_to_remove, pvalue_backward, model_backward = max(pvalues_backward, key=lambda x: x[1], default=(Nor

            if pvalue_backward is not None and pvalue_backward > alpha_removal:
                selected_predictors.remove(variable_to_remove)
                print(f'{method.capitalize()} Selection: Removed {variable_to_remove} from the model. P-value: {pvalue_
                print(f'Current Model: {selected_predictors}')

                # Print regression result
                print(model_backward.summary())

            else:
                break

    print(f'Final Selected Predictors: {selected_predictors}')

    if not selected_predictors:
        print("No predictors selected. Collinearity diagnostics cannot be performed.")
    else:
        # Run collinearity diagnostics after the final selection
        final_model = sm.OLS(data[response], sm.add_constant(data[selected_predictors])).fit()
        print("\nCollinearity Diagnostics:")
        print(final_model.summary())
        vif_data = calculate_vif(data[selected_predictors])
        print("\nVIF:")
        print(vif_data)

    return selected_predictors

csv_file_path = r'C:\Users\Olivia\Documents\Fall-2023\Applied-Statistics\HW\Major-Project\LINTH-5.csv'
data = pd.read_csv(csv_file_path)
response_variable = 'BIO'
predictor_columns = ['SAL', 'pH', 'K', 'Na', 'Zn']

# Run forward and backward stepwise regression
selected_predictors_forward = stepwise_selection(data, response=response_variable, predictors=predictor_columns, method
selected_predictors_backward = stepwise_selection(data, response=response_variable, predictors=predictor_columns, metho
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.005
Model:                            OLS   Adj. R-squared:                 -0.019
Method:                 Least Squares   F-statistic:                    0.2267
Date:                Mon, 04 Dec 2023   Prob (F-statistic):              0.637
Time:                        09:32:55   Log-Likelihood:                -339.75
No. Observations:                  43   AIC:                             683.5
Df Residuals:                      41   BIC:                             687.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1403.4286    870.731      1.612      0.115    -355.048    3161.905
SAL          -13.5038     28.363     -0.476      0.637     -70.783      43.776
==============================================================================
Omnibus:                        5.444   Durbin-Watson:                   0.705
Prob(Omnibus):                  0.066   Jarque-Bera (JB):                4.290
Skew:                           0.655   Prob(JB):                        0.117
Kurtosis:                       2.177   Cond. No.                         262.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.592
Model:                            OLS   Adj. R-squared:                  0.583
Method:                 Least Squares   F-statistic:                     59.60
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           1.62e-09
Time:                        09:32:55   Log-Likelihood:                -320.57
No. Observations:                  43   AIC:                             645.1
Df Residuals:                      41   BIC:                             648.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -859.8091    248.570     -3.459      0.001   -1361.807    -357.811
pH           403.4243     52.256      7.720      0.000     297.891     508.958
==============================================================================
Omnibus:                        4.616   Durbin-Watson:                   0.811
Prob(Omnibus):                  0.099   Jarque-Bera (JB):                4.029
Skew:                           0.750   Prob(JB):                        0.133
```

```
Kurtosis:                       2.997   Cond. No.                         18.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.038
Model:                            OLS   Adj. R-squared:                  0.015
Method:                 Least Squares   F-statistic:                     1.635
Date:                Mon, 04 Dec 2023   Prob (F-statistic):              0.208
Time:                        09:32:55   Log-Likelihood:                -339.03
No. Observations:                  43   AIC:                             682.1
Df Residuals:                      41   BIC:                             685.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1332.2222    284.609      4.681      0.000     757.443    1907.002
K             -0.4279      0.335     -1.279      0.208      -1.104       0.248
==============================================================================
Omnibus:                        4.206   Durbin-Watson:                   0.645
Prob(Omnibus):                  0.122   Jarque-Bera (JB):                2.859
Skew:                           0.460   Prob(JB):                        0.239
Kurtosis:                       2.134   Cond. No.                     2.41e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.41e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.067
Model:                            OLS   Adj. R-squared:                  0.044
Method:                 Least Squares   F-statistic:                     2.921
Date:                Mon, 04 Dec 2023   Prob (F-statistic):             0.0950
Time:                        09:32:55   Log-Likelihood:                -338.39
No. Observations:                  43   AIC:                             680.8
Df Residuals:                      41   BIC:                             684.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
```

```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1400.1101    258.605      5.414      0.000     877.846    1922.374
Na            -0.0244      0.014     -1.709      0.095      -0.053       0.004
==============================================================================
Omnibus:                        3.917   Durbin-Watson:                   0.701
Prob(Omnibus):                  0.141   Jarque-Bera (JB):                2.433
Skew:                           0.375   Prob(JB):                        0.296
Kurtosis:                       2.109   Cond. No.                     4.73e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.73e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.407
Model:                            OLS   Adj. R-squared:                  0.393
Method:                 Least Squares   F-statistic:                     28.19
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           4.13e-06
Time:                        09:32:55   Log-Likelihood:                -328.62
No. Observations:                  43   AIC:                             661.2
Df Residuals:                      41   BIC:                             664.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1880.6136    185.030     10.164      0.000    1506.938    2254.290
Zn           -50.0842      9.433     -5.309      0.000     -69.135     -31.033
==============================================================================
Omnibus:                        3.083   Durbin-Watson:                   0.775
Prob(Omnibus):                  0.214   Jarque-Bera (JB):                2.600
Skew:                           0.601   Prob(JB):                        0.273
Kurtosis:                       2.919   Cond. No.                         46.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Forward Selection: Added pH to the model. P-value: 1.6167124495127452e-09
Current Model: ['pH']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.592
Model:                            OLS   Adj. R-squared:                  0.583
```

```
Method:                 Least Squares   F-statistic:                     59.60
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           1.62e-09
Time:                        09:32:55   Log-Likelihood:                -320.57
No. Observations:                  43   AIC:                             645.1
Df Residuals:                      41   BIC:                             648.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -859.8091    248.570     -3.459      0.001   -1361.807    -357.811
pH           403.4243     52.256      7.720      0.000     297.891     508.958
==============================================================================
Omnibus:                        4.616   Durbin-Watson:                   0.811
Prob(Omnibus):                  0.099   Jarque-Bera (JB):                4.029
Skew:                           0.750   Prob(JB):                        0.133
Kurtosis:                       2.997   Cond. No.                         18.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.595
Model:                            OLS   Adj. R-squared:                  0.575
Method:                 Least Squares   F-statistic:                     29.40
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           1.40e-08
Time:                        09:32:55   Log-Likelihood:                -320.43
No. Observations:                  43   AIC:                             646.9
Df Residuals:                      40   BIC:                             652.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -567.5336    618.903     -0.917      0.365   -1818.383     683.316
pH           402.6382     52.752      7.633      0.000     296.023     509.253
SAL           -9.4681     18.329     -0.517      0.608     -46.512      27.576
==============================================================================
Omnibus:                        5.326   Durbin-Watson:                   0.838
Prob(Omnibus):                  0.070   Jarque-Bera (JB):                4.616
Skew:                           0.801   Prob(JB):                       0.0995
Kurtosis:                       3.110   Cond. No.                         292.
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.640
Model:                            OLS   Adj. R-squared:                  0.622
Method:                 Least Squares   F-statistic:                     35.54
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           1.34e-09
Time:                        09:32:55   Log-Likelihood:                -317.91
No. Observations:                  43   AIC:                             641.8
Df Residuals:                      40   BIC:                             647.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -495.7423    284.763     -1.741      0.089   -1071.270      79.785
pH           406.6875     49.749      8.175      0.000     306.141     507.234
K             -0.4763      0.207     -2.296      0.027      -0.896      -0.057
==============================================================================
Omnibus:                        8.283   Durbin-Watson:                   0.841
Prob(Omnibus):                  0.016   Jarque-Bera (JB):                7.464
Skew:                           0.990   Prob(JB):                       0.0240
Kurtosis:                       3.497   Cond. No.                     3.93e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.93e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.650
Model:                            OLS   Adj. R-squared:                  0.632
Method:                 Least Squares   F-statistic:                     37.13
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           7.64e-10
Time:                        09:32:55   Log-Likelihood:                -317.31
No. Observations:                  43   AIC:                             640.6
Df Residuals:                      40   BIC:                             645.9
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -466.3748    279.219     -1.670      0.103   -1030.698      97.948
```

```
pH           400.4547     49.046      8.165      0.000     301.329     499.580
Na            -0.0227      0.009     -2.563      0.014      -0.041      -0.005
==============================================================================
Omnibus:                       10.456   Durbin-Watson:                   0.919
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                9.845
Skew:                           1.082   Prob(JB):                      0.00728
Kurtosis:                       3.901   Cond. No.                     8.32e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.32e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.605
Model:                            OLS   Adj. R-squared:                  0.585
Method:                 Least Squares   F-statistic:                     30.60
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           8.68e-09
Time:                        09:32:55   Log-Likelihood:                -319.92
No. Observations:                  43   AIC:                             645.8
Df Residuals:                      40   BIC:                             651.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -348.3798    521.805     -0.668      0.508   -1402.987     706.227
pH           341.2342     76.371      4.468      0.000     186.882     495.587
Zn           -12.7342     11.433     -1.114      0.272     -35.842      10.374
==============================================================================
Omnibus:                        3.622   Durbin-Watson:                   0.810
Prob(Omnibus):                  0.163   Jarque-Bera (JB):                2.955
Skew:                           0.642   Prob(JB):                        0.228
Kurtosis:                       3.045   Cond. No.                         162.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
Forward Selection: Added Na to the model. P-value: 0.01424575810902225
Current Model: ['pH', 'Na']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.650
Model:                            OLS   Adj. R-squared:                  0.632
```

```
Method:                 Least Squares   F-statistic:                     37.13
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           7.64e-10
Time:                        09:32:55   Log-Likelihood:                -317.31
No. Observations:                  43   AIC:                             640.6
Df Residuals:                      40   BIC:                             645.9
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -466.3748    279.219     -1.670      0.103   -1030.698      97.948
pH           400.4547     49.046      8.165      0.000     301.329     499.580
Na            -0.0227      0.009     -2.563      0.014      -0.041      -0.005
==============================================================================
Omnibus:                       10.456   Durbin-Watson:                   0.919
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                9.845
Skew:                           1.082   Prob(JB):                      0.00728
Kurtosis:                       3.901   Cond. No.                     8.32e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.32e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.650
Model:                            OLS   Adj. R-squared:                  0.623
Method:                 Least Squares   F-statistic:                     24.17
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           5.25e-09
Time:                        09:32:55   Log-Likelihood:                -317.28
No. Observations:                  43   AIC:                             642.6
Df Residuals:                      39   BIC:                             649.6
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -364.5401    588.269     -0.620      0.539   -1554.427     825.347
pH           400.2018     49.662      8.058      0.000     299.750     500.654
Na            -0.0225      0.009     -2.480      0.018      -0.041      -0.004
SAL           -3.4389     17.423     -0.197      0.845     -38.679      31.802
==============================================================================
Omnibus:                       10.322   Durbin-Watson:                   0.927
Prob(Omnibus):                  0.006   Jarque-Bera (JB):                9.681
```

```
Skew:                           1.076   Prob(JB):                      0.00790
Kurtosis:                       3.879   Cond. No.                     1.72e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.72e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.652
Model:                            OLS   Adj. R-squared:                  0.625
Method:                 Least Squares   F-statistic:                     24.35
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           4.80e-09
Time:                        09:32:55   Log-Likelihood:                -317.18
No. Observations:                  43   AIC:                             642.4
Df Residuals:                      39   BIC:                             649.4
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -439.6711    287.648     -1.529      0.134   -1021.494     142.152
pH           402.2804     49.683      8.097      0.000     301.787     502.773
Na            -0.0172      0.015     -1.159      0.254      -0.047       0.013
K             -0.1606      0.342     -0.470      0.641      -0.852       0.531
==============================================================================
Omnibus:                       10.509   Durbin-Watson:                   0.887
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                9.918
Skew:                           1.088   Prob(JB):                      0.00702
Kurtosis:                       3.893   Cond. No.                     8.49e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.49e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.656
Model:                            OLS   Adj. R-squared:                  0.629
Method:                 Least Squares   F-statistic:                     24.74
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           3.92e-09
Time:                        09:32:55   Log-Likelihood:                -316.96
No. Observations:                  43   AIC:                             641.9
```

```
Df Residuals:                      39   BIC:                             649.0
Df Model:                           3                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -135.3154    501.231     -0.270      0.789   -1149.152     878.521
pH           358.0263     72.538      4.936      0.000     211.304     504.749
Na            -0.0216      0.009     -2.399      0.021      -0.040      -0.003
Zn            -8.7171     10.938     -0.797      0.430     -30.841      13.407
==============================================================================
Omnibus:                       10.491   Durbin-Watson:                   0.928
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                9.862
Skew:                           1.056   Prob(JB):                      0.00722
Kurtosis:                       4.021   Cond. No.                     1.49e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.49e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
Final Selected Predictors: ['pH', 'Na']

Collinearity Diagnostics:
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.650
Model:                            OLS   Adj. R-squared:                  0.632
Method:                 Least Squares   F-statistic:                     37.13
Date:                Mon, 04 Dec 2023   Prob (F-statistic):           7.64e-10
Time:                        09:32:55   Log-Likelihood:                -317.31
No. Observations:                  43   AIC:                             640.6
Df Residuals:                      40   BIC:                             645.9
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -466.3748    279.219     -1.670      0.103   -1030.698      97.948
pH           400.4547     49.046      8.165      0.000     301.329     499.580
Na            -0.0227      0.009     -2.563      0.014      -0.041      -0.005
==============================================================================
Omnibus:                       10.456   Durbin-Watson:                   0.919
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                9.845
Skew:                           1.082   Prob(JB):                      0.00728
```

```
Kurtosis:                       3.901   Cond. No.                     8.32e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.32e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

VIF:
  Variable       VIF
0       pH  4.775603
1       Na  4.775603
Final Selected Predictors: []
No predictors selected. Collinearity diagnostics cannot be performed.
```

In [5]:

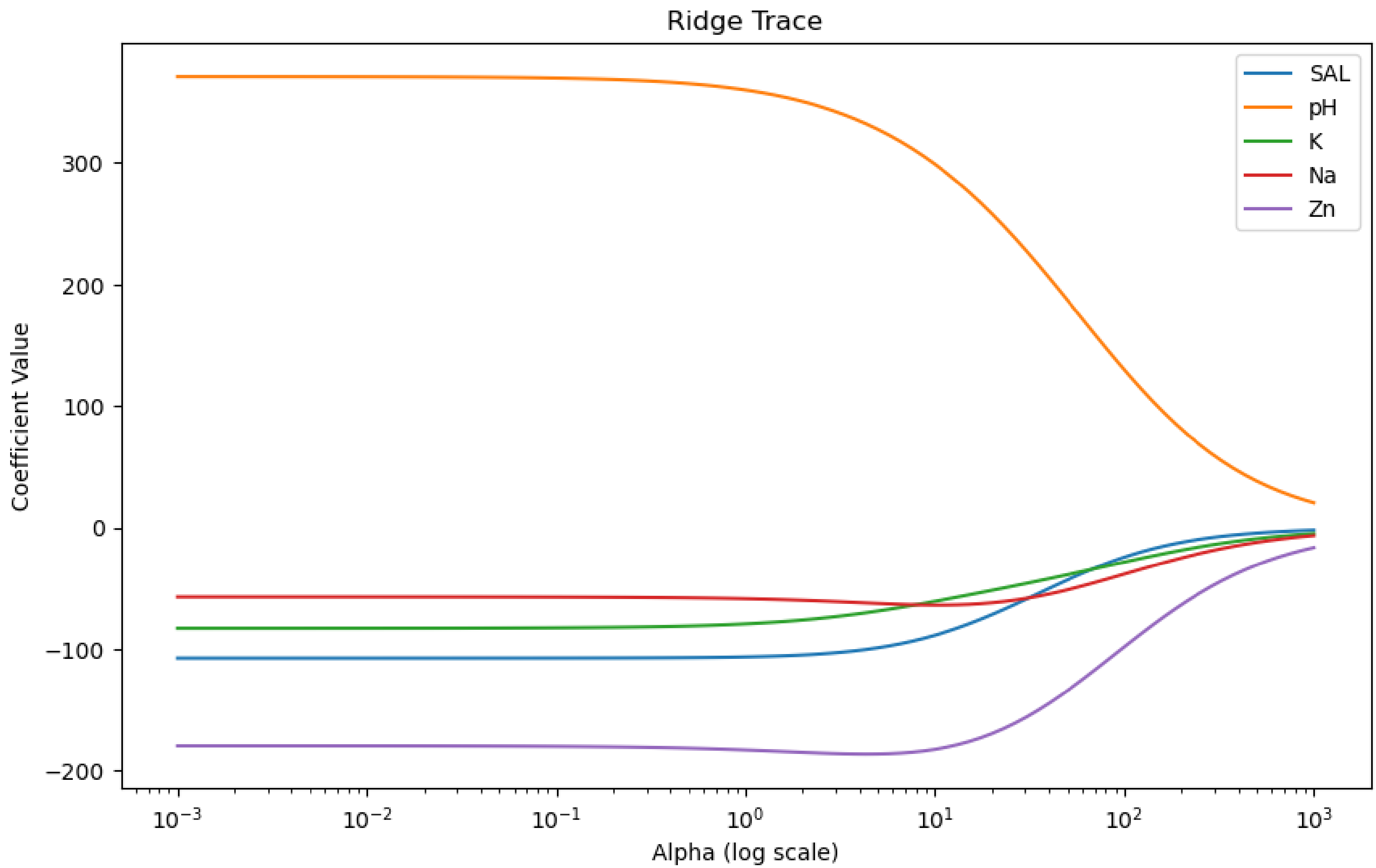
```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import Ridge
from sklearn.preprocessing import StandardScaler
from statsmodels.stats.outliers_influence import variance_inflation_factor
import statsmodels.api as sm

# Load the data
csv_file_path = r'C:\Users\Olivia\Documents\Fall-2023\Applied-Statistics\HW\Major-Project\LINTH-5.csv'
data = pd.read_csv(csv_file_path)

# Define predictors and response
X = data[['SAL', 'pH', 'K', 'Na', 'Zn']]
y = data['BIO']

# Standardize the predictors
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Calculate VIFs for the original ridge regression
vif_data_initial = pd.DataFrame()
vif_data_initial["Variable"] = X.columns
vif_data_initial["VIF"] = [variance_inflation_factor(X_scaled, i) for i in range(X_scaled.shape[1])]

# Fit the initial ridge regression model
alpha = 0.001
ridge_initial = Ridge(alpha=alpha)
ridge_initial.fit(X_scaled, y)

# Print the summary of the initial ridge regression model
X_scaled_with_intercept = sm.add_constant(X_scaled)  # Adding constant for statsmodels
model_initial = sm.OLS(y, X_scaled_with_intercept).fit()
print("\nInitial Ridge Regression Model Summary:")
print(model_initial.summary())

# Print VIFs for the original ridge regression
print("\nCollinearity Diagnostics for Original Ridge Regression (VIF):")
print(vif_data_initial)

# Ridge trace
alphas = np.logspace(-3, 3, 100)
coefs = []
```

```python
mse_values = []

for alpha in alphas:
    ridge = Ridge(alpha=alpha)
    ridge.fit(X_scaled, y)
    coefs.append(ridge.coef_)

    # Evaluate performance using mean squared error
    y_pred = ridge.predict(X_scaled)
    mse = mean_squared_error(y, y_pred)
    mse_values.append(mse)

# Convert to numpy array for easier manipulation
coefs = np.array(coefs)
mse_values = np.array(mse_values)

# Variable selection (Choose alpha based on the ridge trace)
best_alpha_index = np.argmin(mse_values)
best_alpha = alphas[best_alpha_index]

# Refit the model with the best alpha
ridge_refit = Ridge(alpha=best_alpha)
ridge_refit.fit(X_scaled, y)

# Print the summary of the refitted ridge regression model
model_refit = sm.OLS(y, X_scaled_with_intercept).fit()
print("\nRefitted Ridge Regression Model Summary:")
print(model_refit.summary())

# Calculate VIFs for the refitted ridge regression
vif_data_refit = pd.DataFrame()
vif_data_refit["Variable"] = X.columns
vif_data_refit["VIF"] = [variance_inflation_factor(X_scaled, i) for i in range(X_scaled.shape[1])]

# Print VIFs for the refitted ridge regression
print("\nCollinearity Diagnostics for Refitted Ridge Regression (VIF):")
print(vif_data_refit)

# Plot the ridge trace
plt.figure(figsize=(10, 6))
for i in range(coefs.shape[1]):
    plt.plot(alphas, coefs[:, i], label=X.columns[i])

plt.xscale('log')
plt.title('Ridge Trace')
```

```
plt.xlabel('Alpha (log scale)')
plt.ylabel('Coefficient Value')
plt.legend()
plt.show()

# Print the coefficients for the best alpha
print('\nCoefficients for Best Alpha:')
for feature, coef in zip(X.columns, ridge_refit.coef_):
    print(f'{feature}: {coef}')
```

```
Initial Ridge Regression Model Summary:
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.670
Model:                            OLS   Adj. R-squared:                  0.626
Method:                 Least Squares   F-statistic:                     15.04
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           4.60e-08
Time:                        17:30:25   Log-Likelihood:                -316.02
No. Observations:                  43   AIC:                             644.0
Df Residuals:                      37   BIC:                             654.6
Df Model:                           5                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        991.7209     61.864     16.031      0.000     866.372    1117.069
x1          -107.5391     89.636     -1.200      0.238    -289.159      74.081
x2           370.8427    112.846      3.286      0.002     142.195     599.491
x3           -83.0192    106.839     -0.777      0.442    -299.495     133.457
x4           -57.2255    112.579     -0.508      0.614    -285.333     170.882
x5          -179.7898    128.423     -1.400      0.170    -439.999      80.420
==============================================================================
Omnibus:                        8.537   Durbin-Watson:                   1.040
Prob(Omnibus):                  0.014   Jarque-Bera (JB):                7.504
Skew:                           0.944   Prob(JB):                       0.0235
Kurtosis:                       3.789   Cond. No.                         4.19
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Collinearity Diagnostics for Original Ridge Regression (VIF):
  Variable       VIF
0      SAL  2.099364
1       pH  3.327339
2        K  2.982513
3       Na  3.311625
4       Zn  4.309322

Refitted Ridge Regression Model Summary:
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.670
Model:                            OLS   Adj. R-squared:                  0.626
Method:                 Least Squares   F-statistic:                     15.04
```

```
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           4.60e-08
Time:                        17:30:25   Log-Likelihood:                -316.02
No. Observations:                  43   AIC:                             644.0
Df Residuals:                      37   BIC:                             654.6
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        991.7209     61.864     16.031      0.000     866.372    1117.069
x1          -107.5391     89.636     -1.200      0.238    -289.159      74.081
x2           370.8427    112.846      3.286      0.002     142.195     599.491
x3           -83.0192    106.839     -0.777      0.442    -299.495     133.457
x4           -57.2255    112.579     -0.508      0.614    -285.333     170.882
x5          -179.7898    128.423     -1.400      0.170    -439.999      80.420
==============================================================================
Omnibus:                        8.537   Durbin-Watson:                   1.040
Prob(Omnibus):                  0.014   Jarque-Bera (JB):                7.504
Skew:                           0.944   Prob(JB):                       0.0235
Kurtosis:                       3.789   Cond. No.                         4.19
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Collinearity Diagnostics for Refitted Ridge Regression (VIF):
  Variable       VIF
0      SAL  2.099364
1       pH  3.327339
2        K  2.982513
3       Na  3.311625
4       Zn  4.309322
```




```
Coefficients for Best Alpha:
SAL: -107.53851451583436
pH: 370.83077881316626
K: -83.01536429707183
Na: -57.227014705963946
Zn: -179.79420415004194
```

In [10]:

```python
import pandas as pd
from itertools import combinations
from statsmodels.stats.outliers_influence import variance_inflation_factor
from statsmodels.regression.linear_model import OLS
import statsmodels.api as sm

# Load the data
csv_file_path = r'C:\Users\Olivia\Documents\Fall-2023\Applied-Statistics\HW\Major-Project\LINTH-5.csv'
data = pd.read_csv(csv_file_path)

# Define predictors and response
X = data[['SAL', 'pH', 'K', 'Na', 'Zn']]
y = data['BIO']

# Generate all possible combinations of two variables
predictor_combinations = list(combinations(X.columns, 2))

# Initialize variables to store best model information
best_bic = float('inf')
best_model = None

# Iterate through all two-variable combinations
for combo in predictor_combinations:
    # Select the two variables
    current_predictors = list(combo)

    # Fit the linear regression model
    X_current = X[current_predictors]
    X_current = sm.add_constant(X_current)  # Add constant term for intercept
    model = OLS(y, X_current).fit()

    print("\nTwo-Variable Model:", current_predictors)
    print(model.summary())

    # Calculate BIC
    bic = model.bic

    # Check if the current model has the lowest BIC
    if bic < best_bic:
        best_bic = bic
        best_model = current_predictors
```

```python
# Display the best two-variable model
print("\nBest Two-Variable Model:", best_model)
```

```
Two-Variable Model: ['SAL', 'pH']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.595
Model:                            OLS   Adj. R-squared:                  0.575
Method:                 Least Squares   F-statistic:                     29.40
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           1.40e-08
Time:                        16:51:30   Log-Likelihood:                -320.43
No. Observations:                  43   AIC:                             646.9
Df Residuals:                      40   BIC:                             652.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -567.5336    618.903     -0.917      0.365   -1818.383     683.316
SAL           -9.4681     18.329     -0.517      0.608     -46.512      27.576
pH           402.6382     52.752      7.633      0.000     296.023     509.253
==============================================================================
Omnibus:                        5.326   Durbin-Watson:                   0.838
Prob(Omnibus):                  0.070   Jarque-Bera (JB):                4.616
Skew:                           0.801   Prob(JB):                       0.0995
Kurtosis:                       3.110   Cond. No.                         292.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Two-Variable Model: ['SAL', 'K']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.044
Model:                            OLS   Adj. R-squared:                 -0.003
Method:                 Least Squares   F-statistic:                    0.9305
Date:                Thu, 07 Dec 2023   Prob (F-statistic):              0.403
Time:                        16:51:30   Log-Likelihood:                -338.89
No. Observations:                  43   AIC:                             683.8
Df Residuals:                      40   BIC:                             689.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1769.3738    910.385      1.944      0.059     -70.583    3609.330
SAL          -14.2462     28.153     -0.506      0.616     -71.145      42.653
```

```
K             -0.4314      0.338     -1.277      0.209      -1.114       0.251
==============================================================================
Omnibus:                        4.214   Durbin-Watson:                   0.659
Prob(Omnibus):                  0.122   Jarque-Bera (JB):                2.744
Skew:                           0.433   Prob(JB):                        0.254
Kurtosis:                       2.116   Cond. No.                     7.65e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.65e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Two-Variable Model: ['SAL', 'Na']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.068
Model:                            OLS   Adj. R-squared:                  0.021
Method:                 Least Squares   F-statistic:                     1.459
Date:                Thu, 07 Dec 2023   Prob (F-statistic):              0.245
Time:                        16:51:30   Log-Likelihood:                -338.36
No. Observations:                  43   AIC:                             682.7
Df Residuals:                      40   BIC:                             688.0
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1606.7805    862.396      1.863      0.070    -136.186    3349.747
SAL           -7.0609     28.075     -0.251      0.803     -63.803      49.682
Na            -0.0239      0.015     -1.637      0.109      -0.053       0.006
==============================================================================
Omnibus:                        4.165   Durbin-Watson:                   0.707
Prob(Omnibus):                  0.125   Jarque-Bera (JB):                2.469
Skew:                           0.367   Prob(JB):                        0.291
Kurtosis:                       2.084   Cond. No.                     1.56e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.56e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Two-Variable Model: ['SAL', 'Zn']
                            OLS Regression Results
```

```
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.551
Model:                            OLS   Adj. R-squared:                  0.529
Method:                 Least Squares   F-statistic:                     24.57
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           1.09e-07
Time:                        16:51:30   Log-Likelihood:                -322.64
No. Observations:                  43   AIC:                             651.3
Df Residuals:                      40   BIC:                             656.6
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       4450.3222    735.825      6.048      0.000    2963.165    5937.480
SAL          -76.2098     21.280     -3.581      0.001    -119.219     -33.201
Zn           -63.9559      9.169     -6.975      0.000     -82.487     -45.425
==============================================================================
Omnibus:                        1.044   Durbin-Watson:                   1.524
Prob(Omnibus):                  0.593   Jarque-Bera (JB):                1.085
Skew:                           0.315   Prob(JB):                        0.581
Kurtosis:                       2.543   Cond. No.                         377.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Two-Variable Model: ['pH', 'K']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.640
Model:                            OLS   Adj. R-squared:                  0.622
Method:                 Least Squares   F-statistic:                     35.54
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           1.34e-09
Time:                        16:51:30   Log-Likelihood:                -317.91
No. Observations:                  43   AIC:                             641.8
Df Residuals:                      40   BIC:                             647.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -495.7423    284.763     -1.741      0.089   -1071.270      79.785
pH           406.6875     49.749      8.175      0.000     306.141     507.234
K             -0.4763      0.207     -2.296      0.027      -0.896      -0.057
==============================================================================
```

```
Omnibus:                        8.283   Durbin-Watson:                   0.841
Prob(Omnibus):                  0.016   Jarque-Bera (JB):                7.464
Skew:                           0.990   Prob(JB):                       0.0240
Kurtosis:                       3.497   Cond. No.                     3.93e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.93e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Two-Variable Model: ['pH', 'Na']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.650
Model:                            OLS   Adj. R-squared:                  0.632
Method:                 Least Squares   F-statistic:                     37.13
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           7.64e-10
Time:                        16:51:30   Log-Likelihood:                -317.31
No. Observations:                  43   AIC:                             640.6
Df Residuals:                      40   BIC:                             645.9
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -466.3748    279.219     -1.670      0.103   -1030.698      97.948
pH           400.4547     49.046      8.165      0.000     301.329     499.580
Na            -0.0227      0.009     -2.563      0.014      -0.041      -0.005
==============================================================================
Omnibus:                       10.456   Durbin-Watson:                   0.919
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                9.845
Skew:                           1.082   Prob(JB):                      0.00728
Kurtosis:                       3.901   Cond. No.                     8.32e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.32e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Two-Variable Model: ['pH', 'Zn']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.605
```

```
Model:                            OLS   Adj. R-squared:                  0.585
Method:                 Least Squares   F-statistic:                     30.60
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           8.68e-09
Time:                        16:51:30   Log-Likelihood:                -319.92
No. Observations:                  43   AIC:                             645.8
Df Residuals:                      40   BIC:                             651.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -348.3798    521.805     -0.668      0.508   -1402.987     706.227
pH           341.2342     76.371      4.468      0.000     186.882     495.587
Zn           -12.7342     11.433     -1.114      0.272     -35.842      10.374
==============================================================================
Omnibus:                        3.622   Durbin-Watson:                   0.810
Prob(Omnibus):                  0.163   Jarque-Bera (JB):                2.955
Skew:                           0.642   Prob(JB):                        0.228
Kurtosis:                       3.045   Cond. No.                         162.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Two-Variable Model: ['K', 'Na']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.067
Model:                            OLS   Adj. R-squared:                  0.020
Method:                 Least Squares   F-statistic:                     1.430
Date:                Thu, 07 Dec 2023   Prob (F-statistic):              0.251
Time:                        16:51:30   Log-Likelihood:                -338.39
No. Observations:                  43   AIC:                             682.8
Df Residuals:                      40   BIC:                             688.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1387.8770    288.304      4.814      0.000     805.192    1970.562
K              0.0558      0.551      0.101      0.920      -1.058       1.169
Na            -0.0264      0.024     -1.103      0.276      -0.075       0.022
==============================================================================
Omnibus:                        4.043   Durbin-Watson:                   0.712
Prob(Omnibus):                  0.132   Jarque-Bera (JB):                2.490
```

```
Skew:                           0.382   Prob(JB):                        0.288
Kurtosis:                       2.101   Cond. No.                     5.21e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.21e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Two-Variable Model: ['K', 'Zn']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.430
Model:                            OLS   Adj. R-squared:                  0.402
Method:                 Least Squares   F-statistic:                     15.11
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           1.29e-05
Time:                        16:51:30   Log-Likelihood:                -327.77
No. Observations:                  43   AIC:                             661.5
Df Residuals:                      40   BIC:                             666.8
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2130.0023    268.882      7.922      0.000    1586.572    2673.433
K             -0.3320      0.261     -1.270      0.211      -0.860       0.196
Zn           -49.2507      9.387     -5.247      0.000     -68.222     -30.280
==============================================================================
Omnibus:                        6.307   Durbin-Watson:                   0.715
Prob(Omnibus):                  0.043   Jarque-Bera (JB):                5.316
Skew:                           0.841   Prob(JB):                       0.0701
Kurtosis:                       3.373   Cond. No.                     2.93e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.93e+03. This might indicate that there are
strong multicollinearity or other numerical problems.

Two-Variable Model: ['Na', 'Zn']
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.440
Model:                            OLS   Adj. R-squared:                  0.412
Method:                 Least Squares   F-statistic:                     15.74
```

```
Date:                Thu, 07 Dec 2023   Prob (F-statistic):           9.07e-06
Time:                        16:51:30   Log-Likelihood:                -327.39
No. Observations:                  43   AIC:                             660.8
Df Residuals:                      40   BIC:                             666.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2139.2998    248.072      8.624      0.000    1637.927    2640.673
Na            -0.0173      0.011     -1.535      0.133      -0.040       0.005
Zn           -48.3377      9.351     -5.170      0.000     -67.236     -29.440
==============================================================================
Omnibus:                        5.749   Durbin-Watson:                   0.844
Prob(Omnibus):                  0.056   Jarque-Bera (JB):                4.759
Skew:                           0.798   Prob(JB):                       0.0926
Kurtosis:                       3.331   Cond. No.                     5.79e+04
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.79e+04. This might indicate that there are
strong multicollinearity or other numerical problems.

Best Two-Variable Model: ['pH', 'Na']
```