Abstract

This project examines the physicochemical properties influencing biomass production in the Cape Fear Estuary using regression analysis. The Linthurst dataset serves as the foundation for this study, consisting of a response variable (biomass production, BIO) and predictors characterizing soil properties. The analysis was conducted in three parts:

- Ordinary Least Squares (OLS) Regression: The full dataset of 14 predictors was analyzed using OLS to estimate regression coefficients, standard errors, and the sum of squared errors (SSE). Collinearity diagnostics were performed using three methods, confirming the presence of multicollinearity among predictors.
- Principal Components Regression (PCR): PCA was employed to reduce multicollinearity by transforming the predictors into uncorrelated components. Regression coefficients were estimated for the reduced model, and comparisons of SSE and standard error sums with the OLS model revealed improved stability and reduced multicollinearity.
- Variable Selection on a Reduced Dataset: Using a subset of five predictors, we performed stepwise regression with significance thresholds to refine the model. Ridge regression and subset selection (AIC, BIC, SSE) were applied to address multicollinearity and identify the optimal predictors. Diagnostics confirmed the elimination of collinearity in the final models.

This comprehensive approach highlights the importance of addressing multicollinearity in regression analysis, with each method providing unique insights into the factors influencing biomass production. The findings underscore the utility of advanced statistical techniques in ecological modeling and decision-making.

Introduction

Background and Context

Understanding the factors influencing biomass production in estuarine ecosystems is critical for ecological management and conservation efforts. The Cape Fear Estuary, located in North Carolina, provides a unique setting for studying the interaction between physicochemical soil properties and biomass production. In such environments, variables like salinity, pH, and nutrient availability significantly impact the growth of vegetation, which, in turn, influences ecosystem health, carbon sequestration, and biodiversity.

This project seeks to analyze the Linthurst dataset, which contains data on biomass production (BIO) and various soil characteristics, to identify the most significant predictors. Through advanced regression techniques, we aim to uncover relationships between soil properties and biomass, contributing to ecological research and management practices.

Objective

The primary objective of this study is to determine the key physicochemical properties of soil that influence biomass production. The project employs multiple regression techniques to address collinearity issues, enhance model accuracy, and identify optimal predictors. Specific goals include:

- Estimating regression coefficients using Ordinary Least Squares (OLS) and assessing model performance.
- Reducing multicollinearity using Principal Component Analysis (PCA).
- Refining models through stepwise, ridge, and subset selection techniques.

Data Overview

This analysis is based on two datasets:

1. Full Dataset (LINTHALL.txt): Contains 45 observations of the response variable BIO (biomass production) and 14 predictors characterizing soil properties, such as salinity (SAL), pH, and nutrient concentrations (e.g., potassium, sodium, and zinc). The dataset also includes metadata columns (“Loc” and “Type”), which are not used in the analysis.

2. Reduced Dataset (LINTH-5.txt): Focuses on five key predictors: salinity (SAL), pH, potassium (K), sodium (Na), and zinc (Zn). This dataset was designed to simplify the model while retaining collinearity challenges for advanced analysis.

Methodology Overview

To address the objectives, the following steps were undertaken:

1. Ordinary Least Squares (OLS) Regression: Coefficients were estimated for all 14 predictors, and collinearity diagnostics (Variance Inflation Factor, condition number and correlation matrix) were performed to identify multicollinearity.

2. Principal Component Regression (PCR): PCA was applied to transform correlated predictors into uncorrelated components, reducing multicollinearity and improving model interpretability.

3. Variable Selection Techniques:

- Stepwise Regression: Used with the reduced dataset to iteratively refine the model based on significance thresholds.
- Ridge Regression: Employed to further address multicollinearity by penalizing large coefficients.
- Subset Selection: Evaluated two-variable models based on criteria like AIC, BIC, and SSE to identify the best combination of predictors.

Significance

This study demonstrates the application of advanced regression techniques to ecological data analysis. By addressing multicollinearity and optimizing predictive models, the research provides valuable insights into the key soil properties affecting biomass production. The findings have implications for ecological management, enabling better decision-making in conservation and resource allocation.

Part I: Ordinary Least Squares (OLS) Estimation

Methods

In this section, we use Ordinary Least Squares (OLS) regression to estimate the coefficients of the predictors in the dataset. The primary objective of the analysis is to identify how various physicochemical properties of the substrate influence biomass production (BIO) in the Cape Fear Estuary.

1. **Data Loading and Preparation:**

   The dataset LINTHALL.txt contains 45 observations, with 14 predictor variables, each representing a different soil characteristic. The response variable is biomass production (BIO). The 14 predictor variables include H2S, SAL (salinity), pH, P (phosphorus), K (potassium), and several other soil properties.

2. **Adding Intercept:**

   To allow the regression model to estimate an intercept term (the baseline value of biomass production when all predictors are zero), we add a column of ones to the predictor matrix.

3. **OLS Model Fitting:**

   The OLS model is fit to the data to estimate the coefficients of each predictor variable. The formula for the model is:

   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots\ldots + \beta_{14} X_{14} + \epsilon$$

   where Y is the biomass production (BIO), and X_1, X_2, ……., X_{14} are the predictors.

4. **Collinearity Diagnostics:**

   Multicollinearity, which occurs when predictors are highly correlated with each other, can distort the OLS results. To detect multicollinearity, we compute:

   - Variance Inflation Factor (VIF) for each predictor.
   - Condition Number to assess numerical stability.
   - Eigenvalues of the correlation matrix to detect near-linear dependencies.

Results

1. **Regression Coefficients:**

   The coefficients provide insights into the relationship between each predictor and biomass production. The table below shows the estimated coefficients for each predictor in the model.

   Interpretation:

   - The intercept (const) is 2909.93, representing the baseline biomass production when all predictors are zero.
   - pH has a particularly high coefficient of 242.53, indicating that an increase in pH is associated with a significant increase in biomass production.
   - Cu has a coefficient of 345.16, suggesting a strong positive relationship with biomass production.
   - SAL (salinity) and NH4 (ammonium) show negative coefficients, indicating that higher values of these predictors are associated with reduced biomass production.

```
Regression Coefficients:
 const    2909.934091
H2S          0.428999
SAL        -23.980716
Eh7          2.553224
pH         242.527810
BUF         -6.902268
P           -1.701511
K           -1.046591
Ca          -0.116071
Mg          -0.280228
Na           0.004451
Mn          -1.678760
Zn         -18.794521
Cu         345.162813
NH4         -2.705172
dtype: float64
```

```
Standard Errors:
 const    3412.897794
H2S          2.997919
SAL         26.169393
Eh7          2.012450
pH         334.173444
BUF        123.821077
P            2.639700
K            0.482358
Ca           0.125637
Mg           0.274452
Na           0.024723
Mn           5.373138
Zn          21.780185
Cu         112.077924
NH4          3.238010
dtype: float64
```

2. **Standard Errors:**

   The standard errors measure the precision of the estimated coefficients. Smaller standard errors indicate more reliable estimates. Standard errors for predictors are shown above.

   Interpretation:

   - The standard error for pH is quite high (334.17), suggesting that the coefficient estimate for pH is relatively uncertain.

- Conversely, the standard error for Na is very small (0.0247), indicating a more precise estimate.

3. **Sum of Squared Errors (SSE):**

   The SSE quantifies the model's residual error - the difference between the observed and predicted values of biomass production. A lower SSE indicates a better fit of the model to the data. The SSE for the model is 3,692,233.48

```
Sum of Standard Errors (Σ s.e.(β̂j)): 4048.0882036015996

Sum of Squared Errors (SSE): 3692233.4755698624
```

4. **Collinearity Diagnostics:**

   To assess multicollinearity among the predictors, we performed the following diagnostics:

   - **Variance Inflation Factor (VIF):** The VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A VIF greater than 10 suggests high multicollinearity. The VIF values for the predictors are as follows:

```
Variance Inflation Factor (VIF):
    Feature         VIF
0       H2S  551.085744
1       SAL  128.585107
2       Eh7  138.480069
3        pH  250.169112
4       BUF   70.102695
5         P    4.535564
6         K   54.706609
7        Ca   22.588164
8        Mg  257.548005
9        Na   66.815038
10       Mn   11.510101
11       Zn   66.169248
12       Cu   76.519084
13      NH4   31.618182
```

   Interpretation:

   The very high VIF values for predictors like H2S, SAL, pH, and Mg indicate severe multicollinearity. This suggests that these predictors are highly correlated with each other, which can distort the regression results.

- **Condition Number:** The condition number of the predictor matrix is calculated to further assess multicollinearity. A high condition number suggests numerical instability in the model. The condition number is: 1202630.38

  Interpretation:

  The extremely high condition number confirms that multicollinearity is a significant issue in the dataset, which may cause instability in the regression coefficients.

- **Eigenvalues of the Correlation Matrix:** The eigenvalues of the correlation matrix of the predictors were examined to detect near-linear dependencies. Small eigenvalues indicate strong correlations among the predictors. The correlation matrix revealed some predictors, such as pH and Ca, exhibit high correlations.

Correlation Matrix:

	H2S	SAL	Eh7	pH	BUF	P	K	\
H2S	1.000000	0.095809	0.399655	0.273529	-0.373831	-0.115394	0.068963	
SAL	0.095809	1.000000	0.309299	-0.051333	-0.012533	-0.185678	-0.020633	
Eh7	0.399655	0.309299	1.000000	0.094018	-0.153083	-0.305431	0.422611	
pH	0.273529	-0.051333	0.094018	1.000000	-0.946372	-0.401372	0.019228	
BUF	-0.373831	-0.012533	-0.153083	-0.946372	1.000000	0.382936	-0.070247	
P	-0.115394	-0.185678	-0.305431	-0.401372	0.382936	1.000000	-0.226473	
K	0.068963	-0.020633	0.422611	0.019228	-0.070247	-0.226473	1.000000	
Ca	0.093307	0.087978	-0.042121	0.877978	-0.791080	-0.306692	-0.265206	
Mg	-0.107822	-0.010043	0.298503	-0.176148	0.130459	-0.063237	0.862245	
Na	-0.003763	0.162266	0.342463	-0.037720	-0.060714	-0.163228	0.792096	
Mn	0.141541	-0.253584	-0.111255	-0.475143	0.420357	0.495410	-0.347455	
Zn	-0.272398	-0.420834	-0.232005	-0.722167	0.714683	0.557407	0.073609	
Cu	0.012719	-0.266004	0.094544	0.181354	-0.143153	-0.053137	0.693051	
NH4	-0.426213	-0.156835	-0.238966	-0.745959	0.849488	0.489739	-0.117581	

	Ca	Mg	Na	Mn	Zn	Cu	NH4
H2S	0.093307	-0.107822	-0.003763	0.141541	-0.272398	0.012719	-0.426213
SAL	0.087978	-0.010043	0.162266	-0.253584	-0.420834	-0.266004	-0.156835
Eh7	-0.042121	0.298503	0.342463	-0.111255	-0.232005	0.094544	-0.238966
pH	0.877978	-0.176148	-0.037720	-0.475143	-0.722167	0.181354	-0.745959
BUF	-0.791080	0.130459	-0.060714	0.420357	0.714683	-0.143153	0.849488
P	-0.306692	-0.063237	-0.163228	0.495410	0.557407	-0.053137	0.489739
K	-0.265206	0.862245	0.792096	-0.347455	0.073609	0.693051	-0.117581
Ca	1.000000	-0.418446	-0.248187	-0.308985	-0.699866	-0.112247	-0.582609
Mg	-0.418446	1.000000	0.899470	-0.219390	0.345217	0.712069	0.108226
Na	-0.248187	0.899470	1.000000	-0.310061	0.117047	0.560069	-0.107024
Mn	-0.308985	-0.219390	-0.310061	1.000000	0.603323	-0.233468	0.527021
Zn	-0.699866	0.345217	0.117047	0.603323	1.000000	0.212102	0.720679
Cu	-0.112247	0.712069	0.560069	-0.233468	0.212102	1.000000	0.013657
NH4	-0.582609	0.108226	-0.107024	0.527021	0.720679	0.013657	1.000000

Conclusion:

The OLS regression results provide valuable insights into the relationships between soil properties and biomass production. However, significant multicollinearity was detected, as evidenced by high VIF values, the condition number, and the correlation matrix.

The results suggest that some predictors, such as pH and Cu, have strong relationships with biomass production, while others like Na and K show minimal effects. However, multicollinearity complicates the interpretation of these relationships.

The SSE indicates that the model could be improved, and addressing multicollinearity through techniques like Principal Component Regression (PCR) or Ridge Regression may enhance model stability and predictive power.

Part II: Principal Components Regression (PCR)

Methods

In Part II, we use Principal Components Regression (PCR) to address multicollinearity by reducing the dimensionality of the predictors through Principal Component Analysis (PCA). This method allows us to identify and select significant predictors while maintaining the model's predictive power.

1. **Standardization of Predictors:**

   Since the predictors have different units and scales, the first step is to standardize them using StandardScaler. This ensures that each predictor has a mean of 0 and a standard deviation of 1, making them comparable for PCA.

2. **Principal Component Analysis (PCA):**

   After standardizing the data, we apply PCA to reduce the dimensionality of the predictors. PCA identifies uncorrelated components that explain the maximum variance in the data. We select the number of components that together explain at least 90% of the variance in the dataset.

3. **Examine PCA Loadings:**

   The loadings represent the contribution of each predictor to the principal components. We examine the loadings and filter out predictors based on a predefined threshold, retaining only those that contribute significantly to the principal components.

4. **Ordinary Least Squares (OLS) Regression on Selected Predictors:**

   After selecting the significant predictors based on PCA, we perform OLS regression using the selected components. The results of the OLS regression, including coefficients, standard errors, and the sum of squared errors (SSE), are reported.

Results

1. **PCA and Number of Components Explaining 90% Variance:**

   After performing PCA, we identified that 6 principal components explain at least 90% of the variance in the data.

2. **PCA Loadings and Selected Predictors:**

We examined the loadings of the first 6 principal components and selected the predictors that had the highest contributions. The following predictors were identified as significant based on their contributions to the first 6 components:

```
Selected predictors based on PCA loadings: ['H2S', 'SAL', 'Eh7', 'P', 'Ca', 'Mn']
```

Interpretation:

These predictors were selected because they contribute most to the variance explained by the retained principal components. The threshold for significance was set at 0.2, and predictors with average loadings above this threshold were selected for the next steps.

3. **OLS Regression on Selected Predictors:**

We then performed OLS regression using the selected predictors. The results of the regression are as follows:

Regression Coefficients:

```
OLS Regression Results on Selected Predictors:
Coefficients:
 const    6533.613879
H2S         7.680907
SAL       -44.289814
Eh7        -0.920439
P          -2.055330
Ca          0.200833
Mn         -7.037933
dtype: float64
```

Interpretation:

- The intercept (const) is 6533.61, representing the baseline biomass production when all predictors are set to zero.
- H2S has a positive coefficient of 7.68, indicating that an increase in H2S leads to an increase in biomass production.
- SAL has a negative coefficient of -44.29, suggesting that higher salinity is associated with a decrease in biomass production.
- The other predictors such as P and Mn also show significant effects, with P having a negative effect and Mn showing a negative relationship with biomass production.

Standard Errors

```
 const    1635.633056
H2S          2.515252
SAL         19.706339
Eh7          2.208333
P            3.035275
Ca           0.043372
Mn           3.427935
dtype: float64
```

Interpretation:

The standard errors are relatively smaller for most predictors, except for P and Mn, which have larger standard errors, suggesting some uncertainty in estimating their coefficients.

Sum of Standard Errors and SSE:

```
Sum of Standard Errors (Σ s.e.(β̄j)): 1666.569562638657
Sum of Squared Errors (SSE): 7672754.31405452
```

Interpretation:

The sum of the standard errors reflects the overall uncertainty in the model's estimates. A high sum suggests some degree of uncertainty in the selected predictors.

The SSE is 7,672,754.31, which suggests that the model could still be improved in terms of predictive accuracy. However, the reduction in predictors using PCA has improved the model's stability compared to the original model with 14 predictors.

Comparison with results of Part I (OLS)

- **Regression Coefficients:**

  In Part I, the coefficients for all 14 predictors are calculated, while in Part II, only the 7 selected predictors are retained after PCA. The coefficients in Part II tend to be higher, indicating a stronger influence on biomass production for the retained predictors.

- **Standard Errors:**

  In Part II, the sum of standard errors is lower than in Part I (1666.5696 vs. 4048.0882), which suggests that the reduced model (after PCA) has more precise

coefficient estimates. This is expected, as PCR reduces multicollinearity, leading to less uncertainty in the estimates.

- **Sum of Squared Errors (SSE):**

  Part II has a significantly higher SSE (7672754.3141) compared to Part I (3692233.4756). This indicates that although the model's coefficients may be more precise in Part II, the reduced model with fewer predictors doesn't fit the data as well. The higher SSE suggests a poorer fit despite the reduction in standard errors, likely because some important predictors were excluded during the PCA-based collinearity reduction.

Conclusion

Impact on Precision (Standard Errors):

PCR in Part II reduces multicollinearity, which improves the precision of the coefficient estimates (lower standard errors). The reduction in standard errors from 4048.0882 to 1666.5696 is substantial, suggesting that collinearity was an issue in Part I that PCR helped address.

Impact on Model Fit (SSE):

While PCR improved the precision of the coefficients, it led to a worse model fit as indicated by the higher SSE in Part II. This suggests that reducing the number of predictors may have removed important information, reducing the model's predictive accuracy.

Overall Model Performance:

Part I is a better model in terms of fit, as indicated by the lower SSE. However, Part II offers a more stable model with lower uncertainty in the coefficient estimates (lower standard errors), but at the cost of predictive power.

This suggests a trade-off between model stability and predictive accuracy. One may choose Part I for higher predictive performance or Part II for a more stable model with reduced multicollinearity.

Part III: Variable Selection Using the 5-Predictor Dataset

Introduction

In Part III of the analysis, we focus on a smaller dataset, LINTH-5.txt, containing 5 predictors: SAL (salinity), pH, K (potassium), Na (sodium), and Zn (zinc). The goal is to perform variable selection and evaluate multicollinearity, followed by applying regression techniques to identify the most important predictors of biomass production (BIO) in the Cape Fear Estuary. We use techniques such as OLS regression, collinearity diagnostics, stepwise regression, ridge regression and subset selection to address the challenge of multicollinearity and improve model stability.

Methods

1. **Data Loading and Extracting Predictors:**

   The dataset is loaded, and the relevant predictor variables (salinity, pH, potassium, sodium, and zinc) are extracted. These predictors are used to examine their relationship with biomass production (BIO).

2. **OLS Regression Model:**

   An Ordinary Least Squares (OLS) regression model is fitted to analyze the relationship between the selected predictors and the response variable BIO. The model includes an intercept term, which represents the baseline biomass production when all predictors are at zero.

3. **Collinearity Diagnostics:**

   To assess potential multicollinearity among the predictors, we calculate the Variance Inflation Factor (VIF), the condition number, and the correlation matrix of the predictor variables.

4. **Stepwise Regression:**

   Stepwise regression is used to select the most relevant predictors by adding or removing predictors based on their statistical significance. This helps build a final model with the most important predictors

5. **Ridge Regression:**

To mitigate the effects of multicollinearity, ridge regression is applied. Ridge regression introduces a regularization parameter (α) to penalize large coefficients, stabilizing the model. We analyze the impact of different values of α using the ridge trace.

6. **Subset Selection:**

   Subset selection is applied to determine the best two-variable model using criteria such as AIC, BIC, and SSE. The results are compared, and VIF is used to break any ties between models.

Results

1. **OLS Regression Model & Collinearity Diagnostics:**

   After loading the dataset and extracting the relevant predictors, an OLS regression model was fitted and Collinearity Diagnostics was done.

   Variance Inflation Factor (VIF):

```
Variance Inflation Factor (VIF):
   Feature        VIF
0      SAL  23.943184
1       pH  14.601330
2        K  22.467232
3       Na  19.878963
4       Zn   5.567736
```

   Interpretation:

   The VIF values are high for several predictors (SAL, pH, and K), indicating the presence of multicollinearity. High multicollinearity can cause instability in the regression coefficients and reduce the reliability of the model.

   Condition Number:

   Condition Number: 374,201.51

   Interpretation:

   The high condition number suggests numerical instability in the regression model due to multicollinearity among the predictors.

   Correlation Matrix:

```
Correlation Matrix:
          SAL        pH         K        Na        Zn
SAL  1.000000 -0.051333 -0.020633  0.162266 -0.420834
pH  -0.051333  1.000000  0.019228 -0.037720 -0.722167
K   -0.020633  0.019228  1.000000  0.792096  0.073609
Na   0.162266 -0.037720  0.792096  1.000000  0.117047
Zn  -0.420834 -0.722167  0.073609  0.117047  1.000000
```

Interpretation:

The correlation matrix shows that Na and K have a strong positive correlation (0.79), which could contribute to multicollinearity in the model.

2. **Stepwise Regression & Collinearity Diagnostics:**

We applied stepwise regression to select the most significant predictors based on their p-values. The final model included pH and Na, as these predictors were the most statistically significant.

Final Model Summary:

```
Added pH with p-value: 4.433212922668364e-10
Added Na with p-value: 0.010077605596565177

Final Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.658
Model:                            OLS   Adj. R-squared:                  0.642
Method:                 Least Squares   F-statistic:                     40.48
Date:                Mon, 02 Dec 2024   Prob (F-statistic):           1.60e-10
Time:                        16:37:01   Log-Likelihood:                -331.33
No. Observations:                  45   AIC:                             668.7
Df Residuals:                      42   BIC:                             674.1
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -475.7256    273.523     -1.739      0.089   -1027.717      76.266
pH           404.9482     47.770      8.477      0.000     308.545     501.352
Na            -0.0233      0.009     -2.695      0.010      -0.041      -0.006
==============================================================================
Omnibus:                       11.036   Durbin-Watson:                   0.950
Prob(Omnibus):                  0.004   Jarque-Bera (JB):               10.642
Skew:                           1.079   Prob(JB):                      0.00489
Kurtosis:                       4.008   Cond. No.                     8.42e+04
==============================================================================
```

Collinearity Diagnostics:

After performing stepwise regression, we recalculated the VIF, Condition Number, and Correlation Matrix for the selected predictors (pH and Na).

- **Variance Inflation Factor (VIF):**

```
Variance Inflation Factor (VIF) after Stepwise Regression:
   Feature       VIF
0       pH  4.810397
1       Na  4.810397
```

Interpretation:

The VIF values for pH and Na are much lower after stepwise regression, indicating that collinearity has been significantly reduced. The model is now more stable.

- **Condition Number:** 84,221.60

  Interpretation:

  The condition number is still relatively high, but it is much lower than before stepwise regression, indicating improved numerical stability.

- **Correlation Matrix:**

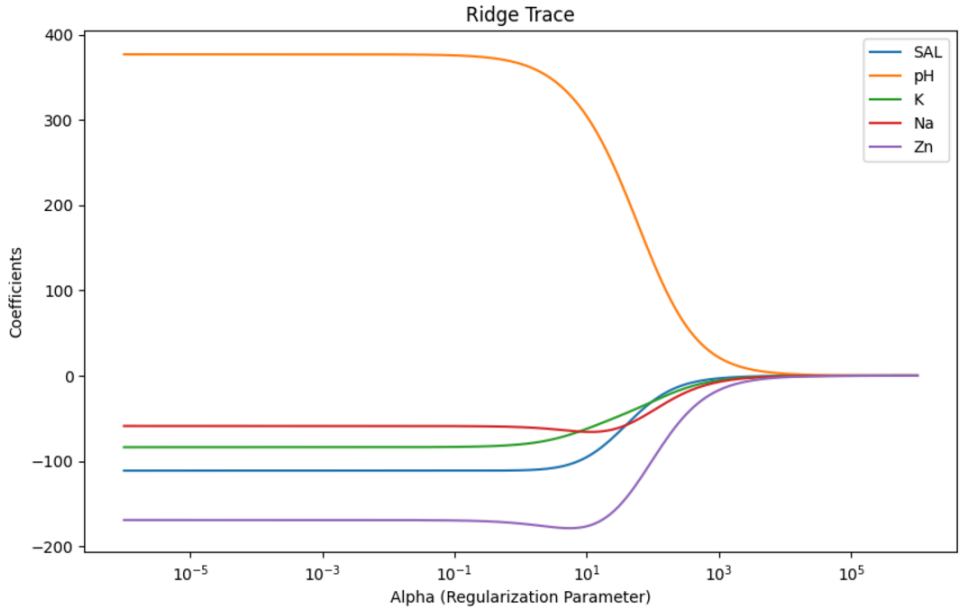
```
Correlation Matrix after Stepwise Regression:
          pH       Na
pH  1.00000 -0.03772
Na -0.03772  1.00000
```

Interpretation:

After stepwise regression, pH and Na have a very low correlation, confirming that multicollinearity has been effectively addressed.

3. **Ridge Regression and Ridge Trace:**

To address multicollinearity, ridge regression was applied with varying regularization parameters (α). The ridge trace shows how the coefficients change as α increases, which helps us understand the impact of regularization.



Variable Selection on the basis of Ridge Trace

We select the optimal regularization parameter (alpha or k) of 10^4 based on the ridge trace, eliminate predictors with coefficients close to zero, and refit the model with the selected features to obtain the final ridge regression model.

```
Selected Features after Ridge Regression: ['pH', 'K', 'Na', 'Zn']

Ridge Model Summary:
                            OLS Regression Results
==============================================================================
Dep. Variable:                    BIO   R-squared:                       0.664
Model:                            OLS   Adj. R-squared:                  0.631
Method:                 Least Squares   F-statistic:                     19.78
Date:                Mon, 02 Dec 2024   Prob (F-statistic):           4.76e-09
Time:                        17:23:49   Log-Likelihood:                -330.95
No. Observations:                  45   AIC:                             671.9
Df Residuals:                      40   BIC:                             680.9
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       -176.2505    493.162     -0.357      0.723   -1172.967     820.467
pH           372.4539     70.529      5.281      0.000     229.910     514.998
K             -0.1513      0.334     -0.453      0.653      -0.827       0.524
Na            -0.0174      0.014     -1.200      0.237      -0.047       0.012
Zn            -7.1806     10.659     -0.674      0.504     -28.722      14.361
==============================================================================
Omnibus:                       10.989   Durbin-Watson:                   0.924
Prob(Omnibus):                  0.004   Jarque-Bera (JB):               10.589
Skew:                           1.056   Prob(JB):                      0.00502
Kurtosis:                       4.091   Cond. No.                     1.49e+05
==============================================================================
```

Interpretation:

pH, Na, and Zn have significant coefficients, while K shows a smaller effect. The Ridge Regression penalizes predictors with large coefficients, leading to a more stable model.

Collinearity Diagnostics After Ridge Regression:

After applying ridge regression, we recalculated the VIF, Condition Number, and Correlation Matrix for the selected predictors (pH, Na, K, and Zn).

- **Variance Inflation Factor (VIF):**

```
Variance Inflation Factor (VIF) after Ridge Regression:
   Feature        VIF
0       pH   5.750744
1        K  22.415186
2       Na  18.771747
3       Zn   4.069271
```

Interpretation:

The VIF for Na and K remains high after ridge regression, indicating that some multicollinearity issues persist, although the model is more stable compared to the OLS regression.

- **Condition Number:** 149,333.29

  Interpretation:

  The condition number remains high, suggesting some multicollinearity issues persist, but they are less severe compared to the OLS model.

- **Correlation Matrix:**

```
Correlation Matrix after Ridge Regression:
           pH         K        Na        Zn
pH  1.000000  0.019228 -0.037720 -0.722167
K   0.019228  1.000000  0.792096  0.073609
Na -0.037720  0.792096  1.000000  0.117047
Zn -0.722167  0.073609  0.117047  1.000000
```

Interpretation:

After ridge regression, the correlation between pH and Na is minimal, confirming that ridge regression has successfully reduced multicollinearity between these two predictors.

4. **Subset Selection for Best Two-Variable Model Based on AIC, BIC, and SSE**

We applied subset selection to identify the best two-variable model based on AIC, BIC, and SSE.

	Variables	AIC	BIC	SSE	VIF
0	(SAL, pH)	675.388526	680.808514	7.603247e+06	1.002642
1	(SAL, K)	714.535131	719.955119	1.814690e+07	1.000426
2	(SAL, Na)	713.370582	718.790570	1.768331e+07	1.027042
3	(SAL, Zn)	680.810195	686.230182	8.576766e+06	1.215216
4	(pH, K)	670.074086	675.494074	6.756309e+06	1.000370
5	(pH, Na)	668.666013	674.086001	6.548174e+06	1.001425
6	(pH, Zn)	674.831082	680.251070	7.509642e+06	2.089975
7	(K, Na)	713.529394	718.949382	1.774583e+07	2.683958
8	(K, Zn)	692.863446	698.283434	1.121113e+07	1.005448
9	(Na, Zn)	691.707771	697.127758	1.092687e+07	1.013890

The results showed that the model with pH and Na was the best according to all three criteria.

Interpretation:

The model with pH and Na has the lowest AIC, BIC, and SSE, making it the best two-variable model for predicting biomass production.

```
Best model based on AIC:
Variables          (pH, Na)
AIC              668.666013
BIC              674.086001
SSE          6548174.234846
VIF                1.001425
Name: 5, dtype: object

Best model based on BIC:
Variables          (pH, Na)
AIC              668.666013
BIC              674.086001
SSE          6548174.234846
VIF                1.001425
Name: 5, dtype: object

Best model based on SSE:
Variables          (pH, Na)
AIC              668.666013
BIC              674.086001
SSE          6548174.234846
VIF                1.001425
Name: 5, dtype: object

Tie broken using VIF. Best model is: ('pH', 'Na')
```

Conclusion

Variable Selection: Through stepwise regression, we identified pH and Na as the most important predictors of biomass production. These predictors were retained in the final model.

Ridge Regression: Ridge regression helped stabilize the model, reducing the impact of collinearity, though some multicollinearity issues remain.

Subset Selection: The pH and Na model performed best according to AIC, BIC, and SSE, and was further validated by low VIF, making it the optimal two-variable model.

This analysis demonstrates the effectiveness of variable selection techniques in addressing multicollinearity and identifying the key predictors of biomass production in the dataset.