MATH 564 APPLIED STATISTICS PROJECT REPORT

Muhammad Usman Umer

Department of Mathematics, Illinois Institute of Technology

MATH 564: Applied Statistics

Professor Maggie Cheng

A20525892

Date: 11/29/2023

Abstract

This project examines advanced statistical techniques to analyze the Linthurst data to determine the primary physicochemical properties influencing aerial biomass production in the Cape Fear Estuary of North Carolina. The study uses the dataset given with 14 predictors and focuses on employing Principal Components Regression, Collinearity Diagnostics, and Ordinary Least Square Estimation to determine the correlations between different soil properties and biomass production. Furthermore, variable selection techniques such as stepwise regression, ridge regression, and subset selection methods are investigated on a reduced 5-predictor dataset. The project's objectives are to reduce collinearity, identify critical predictors, and improve knowledge of the intricate interactions between environmental conditions and aerial biomass production.

Methodology

For the methodology we will divide the section into three parts as the project also follows the same template. For each of the parts of the project we will be discussing the methodology employed and models used that lead us to the results.

PART I:

In the first part we use the ordinary least square estimation to estimate regression coefficients and also employ collinearity diagnostics to conclude the presence of multicollinearity among predictors.

The first step is to import the Linthurst dataset (LINTHALL.txt) into the python data frame, on observation we see that it comprises of 45 observations and 14 predictor variables. The response variable is Biomass production (BIO), while predictors (X1 to X14) represent diverse soil physicochemical properties. The model is given by the expression below:

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

Next up we deploy a multiple linear regression model to predict BIO using all 14 predictors and OLS (ordinary least squares) estimation is used to calculate the regression coefficients as well as standard error, t-stat, p-value estimation. Then to test the collinearity the two methods used are: Variance Inflation Factor (VIF) and calculating eigen values and condition index (K) for the predictors. For the VIF case if the VIF value is higher than 10 it can conclude the presence of collinearity. (Potters & Li, 2023) And for the case of eigen values we use the correlation matrix to calculate the eigen values, and using the eigen values we calculate the condition index K using Equation 1 given below. In general, Eigenvalues reveal how much of each principal component's variation is explained and the higher the eigenvalue the greater percentage of variance is present in the data. Whereas for the condition index (K) they quantify the degree to which an eigenvalue's variance deviates from its expected variance in the absence of multicollinearity. Higher values of the condition index signify rising multicollinearity, whereas a condition index approaching 1 suggests low multicollinearity.

$$K_j = \sqrt{\frac{\lambda_1}{\lambda_p}} \quad where\ j = 1,2, \dots \dots p. \qquad \text{Equation 1 (Chatterjee \& Hadi, 1938)}$$

The last thing we calculate in this part is the standard error sum $\Sigma_j\, s \cdot \mathrm{e}\left(\hat{\beta}_j\right)$ and SSE (sum of squared errors) for comparison with the results of part two of the project.

PART II:

Again, the first step is to import the Linthurst dataset (LINTHALL.txt) into the python data frame. On observation we see that it comprises of 45 observations and 14 predictor variables. We separate out the response variable is Biomass production (BIO) and predictors (X1 to X14).

In this part we are using the PCR (Principal Components Regression) with collinearity reduction to conclude which components should be included in the model. So first we standardize the predictor variables to ensure equal contribution during PCA.(Vinayedula, 2022) Predictor variables were standardized by subtracting their means and dividing by their standard deviations. PCA is employed using the scikit-learn library to transform the standardized predictor variables into principal components. The variance explained by each principal component was examined to inform the decision on the number of components to retain. The cumulative variance ratio was computed by summing the explained variance ratios of the principal components. The number of components to include was determined based on a threshold of 95% cumulative explained variance. This leads to selection of the principal components from the transformed data. A constant term is added to the principal components' matrix to account for the intercept in the PCR model. Then we employ the PCR model by fitting it with the response variable (Y) and the selected principal components using Ordinary Least Squares (OLS). Next, we extract the coefficients $\hat{\beta}_j$ from the PCR model. These coefficients transcribe the relationship between the original predictors and the response variable in the reduced model.

Lastly, we calculate the standard error sum $\Sigma_j \, s \cdot \mathrm{e}(\hat{\beta}_j)$ and SSE (sum of squared errors) and compare the results with the results we calculated in part I. SSE reflects the overall goodness-of-fit of the model.

Part III:

In this part we use the smaller dataset (LINTH-5.txt) with five predictor variables is used for variable selection. The model is given as:

$$Y \sim X2 + X4 + X7 + X10 + X12$$

while the response variable is same (BIO), and predictor variables are as follows:

- X2: SAL
- X4: pH
- X7: K
- X10: Na
- X12: Zn

In the first step we do a stepwise regression analysis and the objective here is to identify the best model by adding or removing a predictor at each iteration based on significance levels. So, we start with an empty model and at each step, consider adding or removing a predictor based on the significance level (e.g., $\alpha_E = \alpha_R = 0.10$). The results of regression are calculated at each step, including which predictor enters or leaves the model and this process is continued until no change is observed. After arriving at the final model, we run collinearity diagnostics to ensure that collinearity does not exist in the model. (Hayes, 2022)

Next, we do a ridge regression with ridge trace and the goal is to apply ridge regression to eliminate collinearity and utilize the ridge trace to guide variable selection. So, we perform ridge regression on the five-predictor model while considering a range of ridge parameters. Then use the ridge trace to monitor the effect of changing the ridge parameter on the coefficients and monitor the variations on plot. Then we do variable selection and pick the model with the optimal set of remaining parameters and fit the model again with those variables. The variable selection is doing using the rules mentioned in the textbook in section 11.13 variable selection using ridge regression. (Chatterjee & Hadi, 1938). Lastly, we calculate the collinearity diagnostics again to verify that collinearity has been alleviated.

Finally, we have to conduct a subset selection to determine the best two-variable model using the Bayesian Information Criterion (BIC). So, we evaluate all possible variations of the two variables and calculate BIC results for each subset. Then we select the subset with the lowest value of BIC and VIF as the best two-variable model. Since in our case we did not encounter this problem but if there was a tie then the ideal strategy would be to use VIF to pick the model.

Results and Discussion

PART I:

In the first we performed an ordinary least square regression on the entire dataset with 14 predictors and the results of the regression are shown below. In this we see that R-squared value is 0.823 and also other statistics like F-stat, AIC and BIC are all shown below. Additionally, coefficients calculated for the 14 predictors are mentioned in the summary table in Figure 1 along with other statistics like t-stat, standard error, p-value etc.

1. **R-Squared: 0.823:**
   - This value means that approximately 82.3% of the variability in the dependent variable is explained by the independent variable in our model. A higher R-squared value indicates a better fit.

2. **Adjusted R-Squared: 0.734**
   - This value suggests that the adjusted R-squared is 73.4%, which is slightly lower than the R-squared. It provides a more conservative estimate of the model's goodness of fit.

3. **F-Statistic: 9.270**
   - The F-statistic tells us about the fit of the estimated model to a model with no predictors (null model). A higher value of F-stat suggests that at least one independent variable is contributing significantly to the model and reported value of 9.270 confirms that the model is significant.

4. **AIC: 635.5 and BIC: 661.8**
   - AIC and BIC values are a measure of relative quality in a statistical model. A lower AIC and BIC value suggests a better-fitting model and reported values are 635.5 and 661.8 respectively.

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:	Mon, 20 Nov 2023	Prob (F-statistic):	4.03e-07
Time:	17:01:08	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

Figure 1 - OLS Regression with 14 Predictors

In the next part we are asked to determine collinearity diagnostics and for that we used the two methods: VIF and correlation matrix. The calculated VIF's values are shown in the table below:

S. No	Variable	VIF (value)
0	Const	4350.77
1	H2S	3.14
2	SAL	3.36
3	Eh7	1.96
4	pH	62.56
5	BUF	33.48
6	P	2.88
7	K	7.43

8	Ca	17.34
9	Mg	24.48
10	Na	10.37
11	Mn	6.73
12	Zn	12.39
13	Cu	4.86
14	NH4	8.58

Table 1 - VIF Values for OLS Regression

As described earlier a high VIF indicates that the associated independent variable is highly correlated with the other variables in the model, suggesting potential multicollinearity issues. For purposes of further analysis, we use the following threshold values:

- Variables with VIF values around 1 indicate low collinearity with other variables.
- Variables with VIF values between 5 and 10 are generally considered to have moderate collinearity.
- Variables with VIF values above 10 may have high collinearity.

As we see the values that are highlighted in the table 1 above are all greater than 10 and indicate a high collinearity with other variables. So, we conclude that variable pH, BUF, Ca, Mg, Na, and Zn all have high collinearity amongst them. While K, Mn, Cu (close to 5) and NH4 have moderate collinearity and the remaining have low collinearity.

Lambda	K (Condition Index)
5.17222	1
3.68891	1.1841
1.61159	1.79148
1.23266	2.04841
0.69215	2.73362
0.49228	3.24141
0.37853	3.69648
0.26146	4.44772
0.15987	5.68791
0.14321	6.00963
0.08409	7.84282
0.00952	23.3084
0.02812	13.56119
0.04539	10.67497

Table 2 - Eigen Values and Condition Index values

As we see the results in Table 2, we see a general trend that the eigenvalues are decreasing, as one would anticipate from a correlation matrix. Another important observation is that a small number of factors account for the majority of the variability in the data, as seen by the fact that the first few eigenvalues are significantly greater than the others.

In particular if we look at the first few rows in the table, the condition indices corresponding to the eigenvalues are typically low. This implies that there is not much collinearity between the variables.

If we look at the specific results the eigenvalues (lambda values) less than 1 (e.g., 0.69215, 0.492278, etc.) represent the proportion of variance explained by the respective components. Hence, components with smaller eigenvalues have a less effect on the variance of the entire dataset.

Additionally, condition indices greater than 20 are generally regarded as potential multicollinearity issues. In our case, some indices are relatively high, especially for the later eigenvalues. This could suggest increasing collinearity for those components but if we look at their eigenvalues, they are significantly less than 1 and hence their overall contribution to the overall variance of the dataset is very small. In our case there is only 1 value of condition index that is above 20 but its eigenvalue is the lowest amongst all.

Additionally, we can make the same conclusion using the correlation matrix as shown in Figure 2:

1. High Correlation (Absolute Value > 0.7):
    - The pair of variables "pH" and "BUF" exhibit a very high negative correlation (-0.946). This suggests a strong linear relationship, and it's consistent with the high VIF value observed for "pH" in the VIF analysis.
    - The variables "K" and "Mg" (0.865) have a high positive correlation, indicating potential collinearity.
    - The pair of variables "Na" and "Mg" (0.898) shows a strong positive correlation, although their VIF value was less than 10.
2. Moderate Correlation (Absolute Value > 0.5 and <= 0.7):
    - The pair of variables “NH4” and “Mn” (0.55) shows a moderate positive correlation.
    - The pair of variables “Ca” and “Na” (0.568) shows a moderate positive correlation.
3. Low Correlation (Absolute Value <= 0.5):
    - Many variables have low or no significant correlation based on the absolute values.

    Note: an important observation is that for Ca the VIF value was 17.34 but it does not show high correlation values in the matrix especially with pH which suggests that VIF and correlation are complementary but not identical measures of collinearity.

	const	H2S	SAL	Eh7	pH	BUF	P
const	NaN	NaN	NaN	NaN	NaN	NaN	NaN
H2S	NaN	1.000000	0.169190	0.430436	0.259976	-0.360190	-0.284927
SAL	NaN	0.169190	1.000000	0.296771	-0.028847	-0.044753	-0.061205
Eh7	NaN	0.430436	0.296771	1.000000	0.101170	-0.163304	-0.326211
pH	NaN	0.259976	-0.028847	0.101170	1.000000	-0.946283	-0.579402
BUF	NaN	-0.360190	-0.044753	-0.163304	-0.946283	1.000000	0.590230
P	NaN	-0.284927	-0.061205	-0.326211	-0.579402	0.590230	1.000000
K	NaN	0.074175	-0.020650	0.427850	0.028564	-0.084963	-0.243725
Ca	NaN	0.097201	0.085460	-0.045807	0.882288	-0.797199	-0.391782
Mg	NaN	-0.090801	-0.035222	0.294838	-0.165704	0.115988	-0.008379
Na	NaN	0.020682	0.140143	0.338063	-0.023624	-0.081745	-0.118908
Mn	NaN	0.133887	-0.257924	-0.111838	-0.499060	0.453243	0.540701
Zn	NaN	-0.291303	-0.422449	-0.227293	-0.731131	0.728321	0.660655
Cu	NaN	0.002390	-0.259403	0.102979	0.187439	-0.152510	-0.070325
NH4	NaN	-0.423095	-0.181808	-0.244983	-0.742309	0.848178	0.670013

	K	Ca	Mg	Na	Mn	Zn	Cu	NH4
const	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
H2S	0.074175	0.097201	-0.090801	0.020682	0.133887	-0.291303	0.002390	-0.423095
SAL	-0.020650	0.085460	-0.035222	0.140143	-0.257924	-0.422449	-0.259403	-0.181808
Eh7	0.427850	-0.045807	0.294838	0.338063	-0.111838	-0.227293	0.102979	-0.244983
pH	0.028564	0.882288	-0.165704	-0.023624	-0.499060	-0.731131	0.187439	-0.742309
BUF	-0.084963	-0.797199	0.115988	-0.081745	0.453243	0.728321	-0.152510	0.848178
P	-0.243725	-0.391782	-0.008379	-0.118908	0.540701	0.660655	-0.070325	0.670013
K	1.000000	-0.259956	0.864980	0.795619	-0.340707	0.069921	0.692069	-0.127974
Ca	-0.259956	1.000000	-0.419053	-0.248135	-0.321979	-0.699351	-0.104983	-0.581496
Mg	0.864980	-0.419053	1.000000	0.898470	-0.212432	0.350283	0.720131	0.098942
Na	0.795619	-0.248135	0.898470	1.000000	-0.304095	0.121679	0.568729	-0.120313
Mn	-0.340707	-0.321979	-0.212432	-0.304095	1.000000	0.613772	-0.228238	0.549798
Zn	0.069921	-0.699351	0.350283	0.121679	0.613772	1.000000	0.206033	0.726463
Cu	0.692069	-0.104983	0.720131	0.568729	-0.228238	0.206033	1.000000	0.008608
NH4	-0.127974	-0.581496	0.098942	-0.120313	0.549798	0.726463	0.008608	1.000000

Figure 2 - Correlation Matrix for OLS regression

PART II:

In this part we have performed a principal component regression (PCR) on the entire dataset with 14 predictors and the results of the regression shown below in Figure 3. In this we see that R-squared value is 0.747 which is almost similar to what we obtained in Part 1 and if we compare the values of AIC and BIC, they are pretty much similar to what we obtained for ordinary least square regression. Additionally, coefficients calculated for the predictors are mentioned in the summary table in Figure 3 along with other statistics like t-stat, standard error, p-value etc.

1. **R-Squared: 0.747:**
   - This value means that approximately 74.7% of the variability in the dependent variable is explained by the independent variable in our model. A higher R-squared value indicates a better fit.

2. **Adjusted R-Squared: 0.687**
   - This value suggests that the adjusted R-squared is 68.7%, which is slightly lower than the R-squared. It provides a more conservative estimate of the model's goodness of fit.

3. **F-Statistic: 12.55**
   - The F-statistic tells us about the fit of the estimated model to a model with no predictors (null model). A higher value of F-stat suggests that at least one independent variable is contributing significantly to the model and reported value of 12.55 confirms that the model is significant.

4. **AIC: 638.6 and BIC: 654.5**
   - AIC and BIC values are a measure of relative quality in a statistical model. A lower AIC and BIC value suggests a better-fitting model and reported values are 638.6 and 654.5 respectively.

If we look closely at the results obtained from PCR regression, they are pretty much similar to what we have obtained for OLS regression in part 1. The difference between these results arises from the fact that we have a smaller number of components in the PCR regression. Despite the reduced number of components, we still have similar results compare to Part 1 which suggests overall good fit of the model.

Summary for PCR Regression:

OLS Regression Results

Dep. Variable:	BIO	R-squared:	0.747
Model:	OLS	Adj. R-squared:	0.687
Method:	Least Squares	F-statistic:	12.55
Date:	Sun, 03 Dec 2023	Prob (F-statistic):	3.58e-08
Time:	20:20:39	Log-Likelihood:	-310.32
No. Observations:	43	AIC:	638.6
Df Residuals:	34	BIC:	654.5
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	991.7209	56.525	17.545	0.000	876.847	1106.594
x1	214.2622	25.149	8.520	0.000	163.154	265.370
x2	-80.7341	29.779	-2.711	0.010	-141.252	-20.217
x3	-107.1749	45.053	-2.379	0.023	-198.734	-15.616
x4	119.9333	51.515	2.328	0.026	15.242	224.624
x5	-65.8768	68.747	-0.958	0.345	-205.587	73.834
x6	-0.2456	81.517	-0.003	0.998	-165.908	165.417
x7	266.6488	92.962	2.868	0.007	77.728	455.570
x8	-53.4328	111.854	-0.478	0.636	-280.748	173.882

Omnibus:	10.353	Durbin-Watson:	1.319
Prob(Omnibus):	0.006	Jarque-Bera (JB):	9.712
Skew:	1.017	Prob(JB):	0.00778
Kurtosis:	4.134	Cond. No.	4.45

Figure 3 - PCR Model Regression Results

Furthermore, to determine which components of PCR are included in the model we have plotted cumulative explained variance vs the number of components in Figure 4. As you can see, with 8 components we have above 95% explained variance and hence those components have been included in the model.

The method of eigenvalue decomposition method to calculate 95% explained variance from the total variance of all the predictors. If we sum up the eigenvalues calculated in the first part and add them to 95% of total variance, we will have the answer of 8 components. Hence the number of components included in the model is 8 principal components.

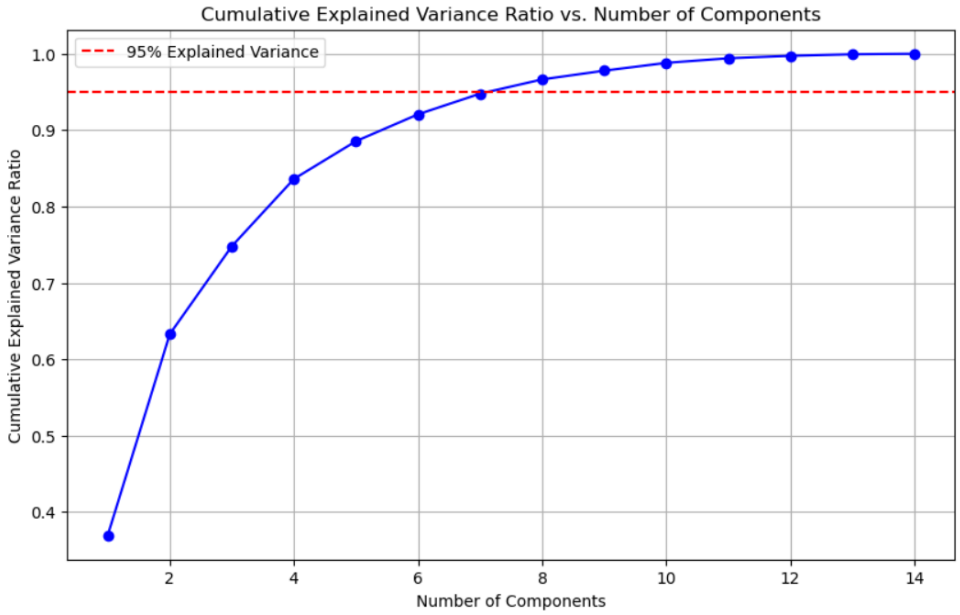

Figure 4 - 95% explained variance Vs No. of Components of PCR

The last result to be discussed in this part is the standard error sum $\Sigma_j\, s \cdot \mathrm{e}(\hat{\beta}_j)$ and SSE (sum of squared errors) of both the models: OLS regression and PCR regression. These results are summarized in the table below. The difference between the results is due to the reduction in the number of components between the OLS and PCR models.

	Standard Error Sum $\Sigma_j\, s \cdot \mathrm{e}(\hat{\beta}_j)$	SSE
OLS Regression (Part1)	4,069.58	3,276,740.28
PCR Regression (Part 2)	563.10	4,671,275.61

Table 3 - Regression Results comparison between OLS and PCR Models

PART III:

Stepwise Regression Results:

In this part we have used the smaller dataset (LINTH-5.txt) with five predictor variables and we are required to perform a variable selection procedure using three different methods.

The first method is using step wise regression with significance level., $\alpha_E = \alpha_R = 0.10$ and we do an iteration by adding or removing the predictor and then finding the best model using the

iteration results calculated at each step. From the results we have selected the variables “Na” and “pH” as the best model and the results are as summarized in Figure 5 below.

```
Selected variables using stepwise regression: ['pH', 'Na']

Summary for Stepwise Regression:
                            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:                Sat, 02 Dec 2023   Prob (F-statistic):           7.64e-10
Time:                        01:15:45   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
==============================================================================
```

Figure 5 - Stepwise Regression Results

- R-Squared: 0.650
- Adjusted R-Squared: 0.632
- F-Statistic: 37.13
- AIC: 640.6
- BIC: 645.9

From the above results we can conclude that almost 65.0% percent of data variability is explained in our model and similarly with the slightly lower adjusted R-Squared value. The higher F-statistic values confirms that our model is significant. By general rule of rule of thumb, the AIC and BIC values should be lower as that ensures a good fit of the model and BIC value in this case is lower than the one in previous parts.

For the variables selected in the above stepwise regression model we have calculated the VIF values to determine the collinearity in the model and the results is that both “pH” and “Na” have VIF values close to 1 as shown by the table 4 below which suggest that collinearity has been almost eliminated.

	Variable	VIF
0	Constant	20.746
1	pH	1.0006
2	Na	1.0006

Table 4 - VIF values for selected variables with Stepwise regression

Ridge Regression Results:

The second method is using the ridge regression using the ridge trace to do the variable selection. So first we fit the ridge regression model and use the ridge trace plot the variation for the ridge coefficients versus the log alpha variations as shown in the figure below:

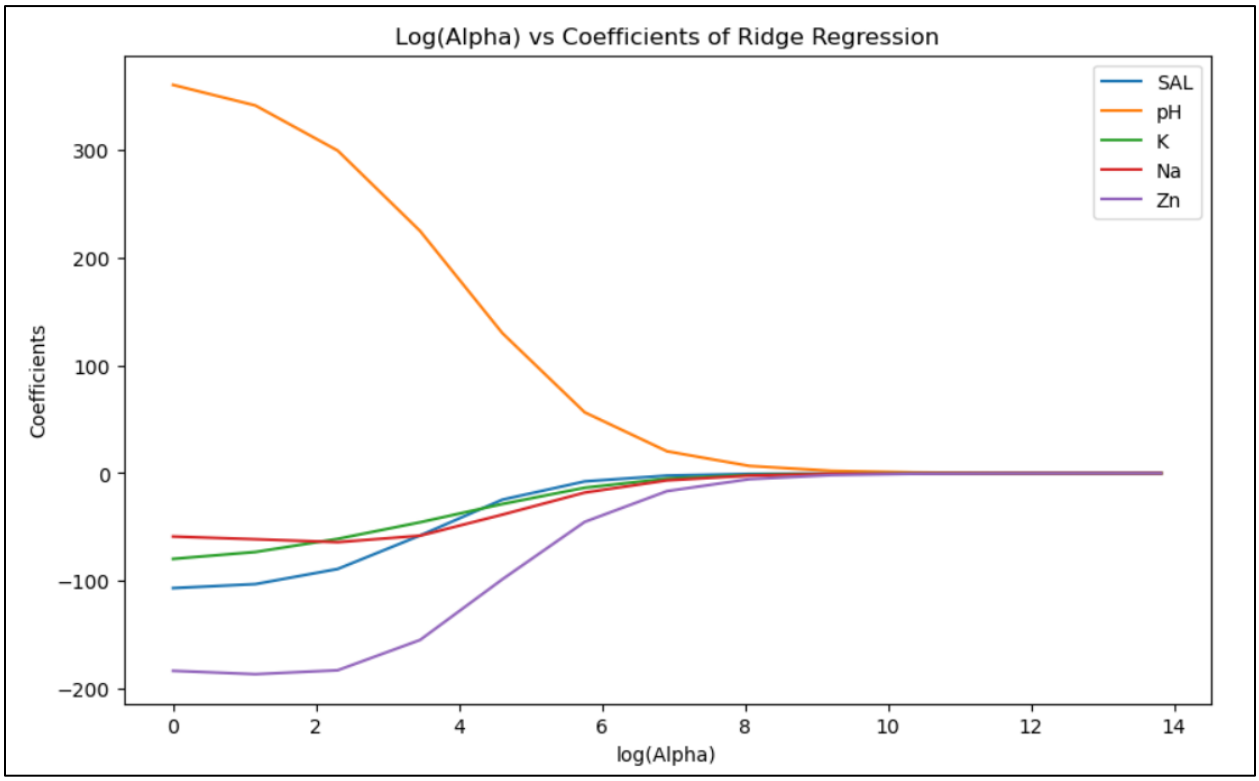

Figure 6 - Log(alpha) vs Ridge Coefficients

As we see that the variables 'pH' and 'Zn' are most dominant in the graph and have the largest coefficient amongst other variables. All the variables have a decreasing trend and almost all of them are reduced to zero as we approach the higher values of log alpha except 'pH' and 'Zn'.

Next, we now look to eliminate some of these variables and for this we take reference from the textbook (Chatterjee & Hadi, 1938). The rules stated in the textbook state that we can eliminate those variables coefficients are stable but small and also those coefficients whose values tend to zero. As discussed earlier from the graph of the ridge trace only 'pH' and 'Zn' variables are most dominant and hence we can use the rules stated in the textbook to eliminate the other variables that tend to zero using a specific log alpha value.

Hence in the code we have used the log alpha value of 8.0 and then used the threshold value of ridge coefficients to eliminate the other remaining variables. Then the variables selected are 'pH' and 'Zn' and the results of the ridge regression are summarized below:

```
Remaining variables using Ridge regression: Index(['pH', 'Zn'], dtype='object')

Summary for Ridge Regression:
                            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:                Wed, 06 Dec 2023   Prob (F-statistic):           8.68e-09
Time:                        16:47:18   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        991.7209     65.142     15.224      0.000     860.065    1123.377
x1           426.6230     95.482      4.468      0.000     233.646     619.600
x2          -106.3453     95.482     -1.114      0.272    -299.322      86.632
==============================================================================
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.                         2.54
==============================================================================
```

Figure 7 - Ridge Regression Results

The final step in this variable selection technique is to calculate the VIF values for the selected variables and the results are summarized in the table below. The results conclude that ‘pH’ and ‘Zn’ are the best models based on the VIF values as they are close to 1 which suggest that collinearity has been almost eliminated.

	Variable	VIF
1	pH	2.15
2	Zn	2.15

Table 5 - VIF values based on ridge regression.

Best Subset Selection Results Using BIC:

The third method is using Subset selection using BIC and VIF and selected variables in this model are “pH” and “Na” as the best subset selection amongst the 5 predictor variables. In this technique we first define a function to calculate the BIC results for each predictor and then another method to select the best model based on the results of BIC and VIF. The function iterates over the entire set of variables and compares the BIC and VIF results to give us the set of variables with the lowest BIC as the best model which in this case is ‘pH’ and ‘Na’ as shown in figure below.

```
Selected variables using subset selection: ('pH', 'Na')

Best BIC for the selected model using subset selection: 523.8650056733826
```

Figure 8 - Subset Selection Using BIC and VIF

For the variables selected in the above best subset selection model we have calculated the VIF values to determine the collinearity in the model and the results is that both “pH” and “Na” have VIF values close to 1 as shown by the table 5 below which suggest that collinearity has been almost eliminated.

	Variable	VIF
1	pH	1.0006
2	Na	1.0006

Table 6 - VIF Values for Best Subset Selection Model

Conclusion

To conclude all of our working above in the project we will go back to our objective of the project that was to reduce collinearity, identify critical predictors, and improve knowledge of the intricate interactions between environmental conditions and aerial biomass production. We will go to the detailed conclusion part by part in this section.

In the first part, we used simple Ordinary least square regression to fit the dependent variable ‘BIO’ and used all 14 predictor variables to get our results. And then we calculated the VIF values for all the predictor variables. On analyzing those values, we saw that almost 6 of those variables’ values were above the threshold value of 10 which is indicative of the fact that collinearity is present in our data.

So, in order to reduce this collinearity amongst the dataset we employed the principal component analysis to identify the best component amongst the dataset and then employed PCR regression. On analysis of the results, we saw that our model was reduced to 8 principal components by virtue of the logic (i.e., explained by 95% of the cumulative variance). Hence using PCR and reducing the model to a specific number of principal components we are able to reduce the collinearity from our model.

Lastly, in the third part of the project we employ a reduced dataset with only 5 predictor variables and then employ multiple techniques to do variable selection. We employ the techniques: stepwise regression, ridge regression and best subset selection using BIC and VIF and we get results “pH” and “Na” from stepwise regression, “pH” and “Zn” from ridge regression and “pH” and “Na” from the best subset selection. After performing variable selection in all three of these techniques when we calculate the VIF values we see that the VIF values have reduced significantly and are close to 1 which suggests that collinearity has been reduced significantly from our model after performing variable selection.

So overall the work we have accomplished in this project involves fitting of regression model with OLS, then with PCR to reduce the collinearity in the dataset and then using a reduced dataset with to employ variable selection to get the best pair of variables. This indeed was the

goal of the project stated earlier in the abstract to reduce collinearity and identify critical predictors for Biomass production.

References

- Chatterjee, S., & Hadi, A. S. (1938). *Regression Analysis by Example FIFTH EDITION.* NewYork: Library of Congress Cataloging-in-Publication.
- Hayes, A. (2022, January 10). *Stepwise Regression: Definition, Uses, Example, and Limitations*. Retrieved from Investopedia: https://www.investopedia.com/terms/s/stepwise-regression.asp#:~:text=Stepwise%20regression%20is%20a%20method,incrementally%2C%20testing%20for%20statistical%20significance.
- Potters, C., & Li, T. (2023, September 30). *Variance Inflation Factor (VIF)*. Retrieved from Investopedia: https://www.investopedia.com/terms/v/variance-inflation-factor.asp
- Vinayedula. (2022, December 30). *Principal Component Regression (PCR)*. Retrieved from geeksforgeeks: https://www.geeksforgeeks.org/principal-component-regression-pcr/