Math 564 Project

Robert Mead

12/05/2021

Introduction

In the Cape Fear Estuary in North Carolina there is a substrate that is influencing the aerial biomass production. The data set has fourteen predictors that characterize the substrates physicochemical properties that are in the soil.Let the variable BIO represent the biomass production that influences by the presence of the substrate. The physicochemical properties include : H_2S, $Salinity$, EH_7,pH, BUF, P,K, Ca, Mg, Na, Mn, Zn, Cu, and NH_4. This project's objective is to identify the important physicochemical properties of the substrate that influence the aerial biomass production in the Cape Fear Estuary.

Collinearity of Full Regression Model

The value of coefficients in a ordinary least squares regression can be classified as unstable, and lead to erroneous inferences with the presence of collinearity within the predictor variables. In this section, an ordinary least squares regression will be completed on the full model of fourteen predictors on the response variable, BIO, to predict the amount of aerial biomass is produced due to the physicochemical substrate that is present. The full regression model will be analyzed for any collinearity using three techniques:

1. Standardized Residuals versus Fitted Values Plot
2. Pairwise Correlation Coefficient Matrix
3. Variance Inflation Factor (VIF)

To fully understand the extent of the collinearity between the predictor variables a regression of all the predictors is a necessary step in the process of determining the important physicochemical substrates that influence the aerial biomass production. The full regression model can be characterized by the coefficient values in the table. Completing the ordinary least squares regression on the full model the coefficients are below in Table 1. It is observed that the coefficient value are have a large range. The procedures, stated above, will assist in determining the predictors that influence the collinearity in the full regression model.

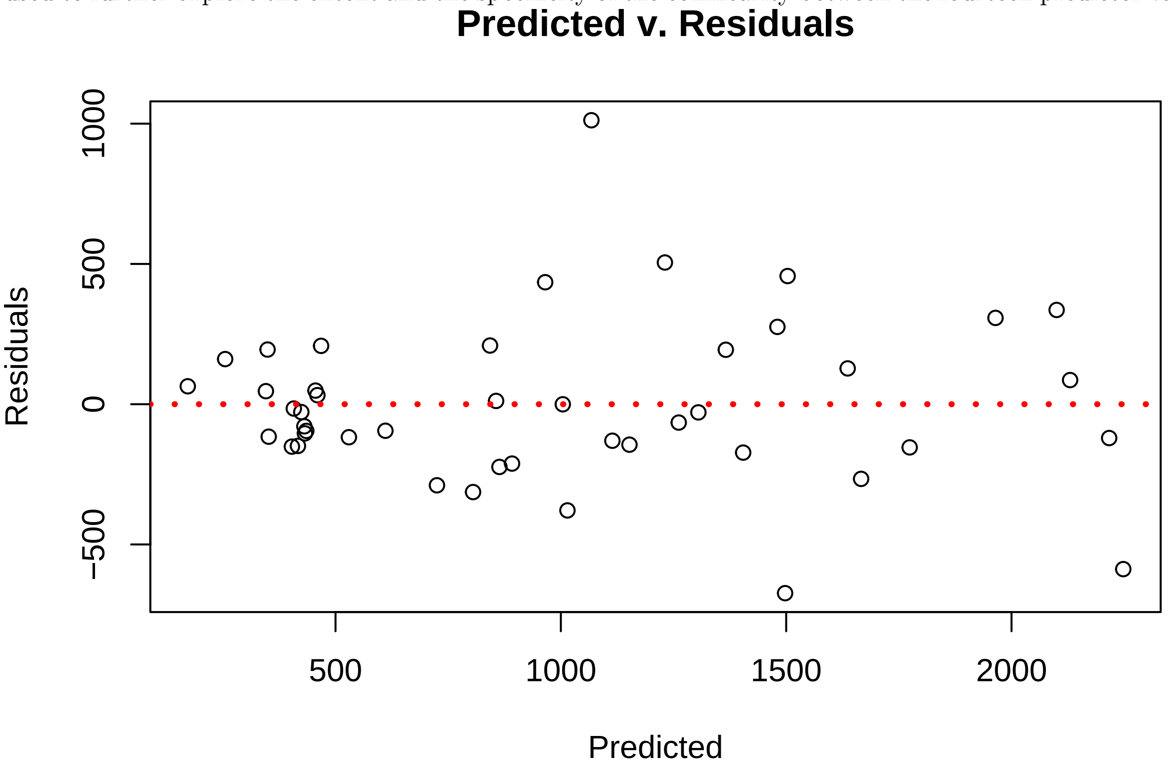
Table 1: Ordinary Least Squares Regression Coefficients

	Variable	Coefficients
(Intercept)	X0	2909.9340915
H2S	X1	0.4289992
SAL	X2	-23.9807157
Eh7	X3	2.5532238
pH	X4	242.5278101
BUF	X5	-6.9022678
P	X6	-1.7015107
K	X7	-1.0465910
Ca	X8	-0.1160706
Mg	X9	-0.2802284
Na	X10	0.0044510
Mn	X11	-1.6787598
Zn	X12	-18.7945212
Cu	X13	345.1628131
NH4	X14	-2.7051724

Standardized Residuals versus Fitted Values Plot

The Standardized Residuals versus Fitted Values Plot is a method of testing collinearity, because it tests for linearity of the data set, unequal variance and outliers.The presence of collinearity in a Standardized Residual versus Fitted Value Plot will be indicated by a trend in the data set. The plot, ideally, should have no trends with the points roughly forming a band of values close or near zero. the Standardized Residuals versus Fitted Values

Plot for the data set shows the there is an unequal variance due to the increase in the value of the residuals as the value of the predicted values increase. This increase in the residuals as the predictor values get larger shows that there is an existence of collinearity. Though the graph does not reveal which predictors have a collinear relationship the gradualy increase in the residuals as the predictor values get larger is an indication. Other tests will be used to further explore the extent and the specificity of the collinearity between the fourteen predictor variables.



Pairwise Correlation Coefficient Matrix

A *condtion indice*, or pairwise correlation coefficient matrix, is a method to further detect collinearity among the fourteen predictor variables. The correlation coefficient matrix takes the predictor variables and assess the correlation that two predictor variables have through the correlation coefficient. Correlation coefficients between two predictors that are closer to one indicate that two predictor variables are highly correlated.

The *condtion indice* shows that there is a correlation between multiple predictor values. The predictor variable pH has a high correlation with BUF, Ca,Zn, NH_4 with correlation coefficient values of -0.946, 0.877,-0.722 and -0.745 respectively. The predictor variable BUF has a high correlation coefficient with Ca, Zn and NH_4 with correlation coefficient values of -0.791, 0.714 and 0.849 respectively.The predictor variable K has a high correlation coefficient with Mg, Na and Cu with correlation coefficients as 0.862, 0.792 and 0.693 respectively.The predictor variable Ca has a high correlation coefficient with the predictor Zn with a correlation coefficient of -0.700.The predictor variable Mg has a high correlation coefficient with the following predictors Na and Cu with correlation coefficient values of 0.899 and 0.712. Lastly, Mn and Zn have a high correlation coefficient of 0.603, and Zn and NH_4 have a high correlation coefficient of 0.721.

Table 2: Pairwise Correlation Coefficient Matrix

	H2S	SAL	Eh7	pH	BUF	P	K	Ca	Mg	Na	Mn	Zn	Cu	NH4
H2S	1.00	0.10	0.40	0.27	-0.37	-0.12	0.07	0.09	-0.11	0.00	0.14	-0.27	0.01	-0.43
SAL	0.10	1.00	0.31	-0.05	-0.01	-0.19	-0.02	0.09	-0.01	0.16	-0.25	-0.42	-0.27	-0.16
Eh7	0.40	0.31	1.00	0.09	-0.15	-0.31	0.42	-0.04	0.30	0.34	-0.11	-0.23	0.09	-0.24
pH	0.27	-0.05	0.09	1.00	-0.95	-0.40	0.02	0.88	-0.18	-0.04	-0.48	-0.72	0.18	-0.75
BUF	-0.37	-0.01	-0.15	-0.95	1.00	0.38	-0.07	-0.79	0.13	-0.06	0.42	0.71	-0.14	0.85
P	-0.12	-0.19	-0.31	-0.40	0.38	1.00	-0.23	-0.31	-0.06	-0.16	0.50	0.56	-0.05	0.49
K	0.07	-0.02	0.42	0.02	-0.07	-0.23	1.00	-0.27	0.86	0.79	-0.35	0.07	0.69	-0.12
Ca	0.09	0.09	-0.04	0.88	-0.79	-0.31	-0.27	1.00	-0.42	-0.25	-0.31	-0.70	-0.11	-0.58
Mg	-0.11	-0.01	0.30	-0.18	0.13	-0.06	0.86	-0.42	1.00	0.90	-0.22	0.35	0.71	0.11
Na	0.00	0.16	0.34	-0.04	-0.06	-0.16	0.79	-0.25	0.90	1.00	-0.31	0.12	0.56	-0.11
Mn	0.14	-0.25	-0.11	-0.48	0.42	0.50	-0.35	-0.31	-0.22	-0.31	1.00	0.60	-0.23	0.53
Zn	-0.27	-0.42	-0.23	-0.72	0.71	0.56	0.07	-0.70	0.35	0.12	0.60	1.00	0.21	0.72
Cu	0.01	-0.27	0.09	0.18	-0.14	-0.05	0.69	-0.11	0.71	0.56	-0.23	0.21	1.00	0.01
NH4	-0.43	-0.16	-0.24	-0.75	0.85	0.49	-0.12	-0.58	0.11	-0.11	0.53	0.72	0.01	1.00

Variance Inflation Factor (VIF)

By identifying the predictor variables that have high correlation coefficients, the Variance Inflation Factor utilizes the correlation coefficient, R^2, to further analyze the predictors that are influenced by collinearity by regressing each predictor variables against the other predictor variables. The Variance Inflation Factor, or VIF, is calculated by finding the correlation coefficient of a specific predictor variables regressed on all the other predictor variables. Then, using the formula, $VIF = \frac{1}{1-R_j^2}$ where each R_j^2 represents the correlation coefficient for each predictor variable X_j from $j = 1, ..., 13, 14$. In the absence of a linear relationship between any predictor variables, the VIF will be 1, and the presence of a linear relationship between any predictor variables will have a VIF value that is large. VIF shows that the predictor variables pH, BUF, Ca, Mg, Na, Z are heavily affected by collinearity, because the VIF values exceed 10. The predictor variables H_2S, SAL, K, Mn, CU and NH_4 show that they are also effected by collinearity, since there VIF is between the interval of 3 and 10. The predictor values of Eh_7 and P are indicating no effect of collinearity.

Table 3: VIF Values for Each Predictor

	VIF
H2S	3.027456
SAL	3.387615
Eh7	1.977447
pH	62.080846
BUF	34.431748
P	1.895804
K	7.367110
Ca	16.662146
Mg	23.764229
Na	10.351043
Mn	6.185628
Zn	11.626479
Cu	4.829203
NH4	8.376506

Conclusion

Through the analysis of the full regression model it is apparent there is collinearity. This effect of collinearity will prove the full regression model to be inaccurate and cause erroneous predictions of the aerial biomass production. In the next section a Principle Component Regression will be executed to identify the important physicochemical properties that influence the aerial biomass production in the Cape Fear Estuary.

Principle Component Regression

It is evident that there are a number of predictors that are collinear. In order to complete a multiple linear regression with fidelity, a Principle Component Regression is in order, because with the existence of collinearity among the predictor variables the coefficient estimates of the model may prove to be unreliable and have high variance. The steps used to complete the Principle Component Regression, known as PCR, are the following:

1. Standardize the data
2. Construct Principle Component Matrix
3. Complete Principle Component Regression of the Full Model
4. Complete Principle Component Regression of Reduced Model
5. Compute Coefficient Values of Reduced Model

The data set is standardized by doing the following $z_{ij} = \frac{x_{ij} - \bar{x_{ij}}}{\sigma_{ij}}$ where i is the row value for $i = 1, ..., 13, 14$ and j is the column value for $j = 1, ..., 13, 14$. The standardized data will provide the necessary information to then compute the Principle components of each of the predictors. The full Principle Component Regression reveals that there are a considerable amount of Principle Components that are not significant. the following Principle Components that can be excluded from the Principle Component Regression model are the following: PC4, PC5, PC6, PC10, PC11, PC12, PC13, PC14. The table shows that the Principle Components are significant, because there p-values are smaller than the alpha level $\alpha = 0.05$, and there are Principle Component values whose $\alpha > 0.05$, thus making the Princple Component not significant.

Table 4: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PC1	1	10122651	10122651	82.25	4.232e-10
PC2	1	1011910	1011910	8.222	0.007502
PC3	1	1254161	1254161	10.19	0.003304
PC4	1	495734	495734	4.028	0.05384
PC5	1	215738	215738	1.753	0.1955
PC6	1	9289	9289	0.07547	0.7854
PC7	1	313159	313159	2.544	0.1212
PC8	1	633449	633449	5.147	0.03065
PC9	1	799741	799741	6.498	0.01615
PC10	1	112793	112793	0.9165	0.3461
PC11	1	432512	432512	3.514	0.07061
PC12	1	40906	40906	0.3324	0.5686
PC13	1	2128	2128	0.01729	0.8963
PC14	1	34561	34561	0.2808	0.6001
Residuals	30	3692233	123074	NA	NA

After removing the insignificant Principle Components from the regression model, the reduced Principle Component Regression can determine the model that identifies the important physicochemical properties of the substrate that influence the aerial biomass production.The reduced Principle Regression Model can be used to find the coefficient values for an ordinary least squares regression model. The Principle Components that are removed from the full Principle Component Regression model are the values that have a p-value that is larger than $\alpha > 0.05$. It is permissible for the Principle Components whose p-value exceeds α due to the fact that the Principle Components show no collinearity. This means that the Principle Component Regression is unaffected by the relationship predictor variables have with each other, so by removing Principle Components allows for a more accurate model.

Table 6: Beta Values of Ordinary Least Squares Regression

	Beta Values
H2S	13.16840
SAL	-89.20170
EH7	94.35692
pH	302.43081
BUF	-17.29953
P	-46.93905
K	-311.46793
Ca	-199.44732
Mg	-263.24839
Na	30.63400
Mn	-41.09700
Zn	-155.61499
Cu	357.93097
NH4	-127.88094

Table 5: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
PC1	1	10122651	10122651	73.8	1.569e-10
PC2	1	1011910	1011910	7.378	0.009792
PC3	1	1254161	1254161	9.144	0.004397
PC8	1	633449	633449	4.618	0.03789
PC9	1	799741	799741	5.831	0.02054
Residuals	39	5349052	137155	NA	NA

The coefficient values for the full regression model are shown below in the table. These coefficient values are calculated through using the Principle Component matrix, constructed earlier, multiplied by the coefficients from the model on the Principle Component Regression. These coefficient values reflect a model that is not effected by the collinearity of the predictor variables. Completing the Principle Component Regression is a methodology that can eliminate the collinearity of predictor variables, and prevent erroneous predictions from the regression.

Conclusion

The Principle Component Regression enables a regression on collinear data. The physicochemical properties of the substrate that influence the aerial biomass production in the Cape Fear Estuary were found due to first, identifying that there indeed exists a collinear relationship among the fourteen predictor variables. After identifying the collinearity, a Principle Component Regression enables the coefficients of the ordinary least squares regression to be found by standardizing the data, creating a principle component matrix, regressing BIO on the principle component matrix to find the princple components that are significant. From there, the reduced model can be computed by removing the principle components that are not significant due to there larger p-values. Through the reduced principle component regression model it shows that there are a large amount of principle components that were removed.

Best Subset Selection

When building a regression model for the data, removing the insignificant variables makes the model more accurate, easier to interpret, and less susceptible to overfit the data. The best subset selection process provides an exhaustive search for the best subset of predictor variables by considering all possible combination. The best subset selection process will be produced for a smaller regression where $BIO \sim SAL + pH + K + Na + Zn$. The five predictor variables will be used for the selection procedure. The stepwise selection procedure is a combination of forward selection and backward selection, because variables that enter into the regression model can be removed later in the process, much like backward selection. The p-values are utilized throughout the process to determine which predictor variables are allowed to enter the regression model, and which predictors are removed from the model.

Best Model

To select the best model, a stepwise regression method will be utilized. Throughout the stepwise process, alpha values of $\alpha_{Enter} = 0.15$ and $\alpha_{Remove} = 0.15$ will be used to access the significance of the p-values.

Step 1

In the first step, the respinse variable, BIO, will be regressed five times on the predictor variables, individually, so there are five regression models. The regression model with the predictor that has the smallest p-value will be entered into the model.Since pH has the smallest p-value of the predictors, pH is now entered into the regression model.

Table 7: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
SAL	1	204048	204048	0.4626	0.5001
Residuals	43	18966915	441091	NA	NA

Table 8: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pH	1	11490388	11490388	64.33	4.433e-10
Residuals	43	7680575	178618	NA	NA

Table 9: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
K	1	802872	802872	1.88	0.1775
Residuals	43	18368091	427165	NA	NA

Table 10: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Na	1	1419069	1419069	3.437	0.0706
Residuals	43	17751894	412835	NA	NA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Zn	1	7474474	7474474	27.48	4.566e-06
Residuals	43	11696489	272011	NA	NA

Table: Analysis of Variance Table

Step 2

The regression model will regress BIO on four predictor variables added with pH. In order for another predictor to be entered into the model the p-value associated has to be the smallest. The predictor with the smallest p-value is entered into the model. Then, the p-value of the previously added predictor, pH is verified to see if it is still a significant predictor by testing it's p-value. Since the p-values remained accepted at the alpha value of $\alpha = 0.15$ for both pH and Na,we can accept Na into the regression model. The p-values for the regression of Na were the smallest, while not effecting the p-value of pH. The addition of Na to the model did not disrupt the predictor value pH.

Table 12: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pH	1	11490388	11490388	63.47	6.223e-10
SAL	1	77327	77327	0.4272	0.517
Residuals	42	7603247	181030	NA	NA

Table 13: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pH	1	11490388	11490388	71.43	1.323e-10
K	1	924266	924266	5.746	0.02106
Residuals	42	6756309	160865	NA	NA

Table 14: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pH	1	11490388	11490388	73.7	8.679e-11
Na	1	1132401	1132401	7.263	0.01008
Residuals	42	6548174	155909	NA	NA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pH	1	11490388	11490388	64.26	5.307e-10
Zn	1	170933	170933	0.956	0.3338
Residuals	42	7509642	178801	NA	NA

Table: Analysis of Variance Table

Step 3

The regression model will regress BIO on three predictor variables added with pH and Na. In order for another predictor to be entered into the model, the remaining predictor has to have the smallest p-value that is significant. If there exist a predictor with a significant p-value, then it will be determined if the existence of the predictor in the model changes the p-value of the predictors in the current model. Since none of the predictors produced a significant p-value that was less than 0.15, then the remaining predictors cannot be included in the regression model. Thus, the regression model only has two predictors of pH and Na. Therefore, the best subset regression model is $BIO = -475.7 + 404.9 \cdot pH - 233.3 \cdot Na$.

Table 16: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pH	1	11490388	11490388	72.07	1.423e-10
Na	1	1132401	1132401	7.103	0.01096
SAL	1	11778	11778	0.07388	0.7871
Residuals	41	6536396	159424	NA	NA

Table 17: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pH	1	11490388	11490388	72.35	1.352e-10
Na	1	1132401	1132401	7.131	0.01081
K	1	36938	36938	0.2326	0.6322
Residuals	41	6511236	158811	NA	NA

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
pH	1	11490388	11490388	72.8	1.246e-10
Na	1	1132401	1132401	7.175	0.01059
Zn	1	77026	77026	0.488	0.4888
Residuals	41	6471149	157833	NA	NA

Table: Analysis of Variance Table

Subset Selection

Though assessing the best subset for each of the models that have 2 predictors, the best model there is when *pH* and Na are in the model because the C_p value, Cooks Distance, is smaller than that of the other model *pH* and K because the C_p value is larger,however, only by small margins. Interestingly, the VIF values are switched where the VIF of the second best 2 predictor model is smaller than that of the best 2 predictor model, but again, only by small margins. This subset selection process reinforces the idea that was concluded in the previous subsection, that the best model to use to determine which physicochemical properties that influence the arerial biomass production in Cape Fear are *pH* and Na. Both the Best Model procedure and the Subset Selection process show that the best two-predictor model include *pH* and Na. These conclusions have been derived through different means.

Table 19: Table of Subset Selection

	SAL	pH	K	Na	Zn	Cp	VIF
1 (1)		*				7.4205743520875	2.49603234631321
1 (2)					*	32.738065776256	1.63903566682368
2 (1)		*		*		2.28159215391216	2.92768068051418
2 (2)		*	*			3.59373611498924	2.83749045830573
3 (1)		*		*	*	3.79600025536262	2.96252865607793
3 (2)		*	*	*		4.0487221242631	2.94428945787905
4 (1)	*	*	*		*	4.29637032868621	3.07558481353295
4 (2)	*	*		*	*	4.66958691102062	3.0466493155314
5 (1)	*	*	*	*	*	6	3.0989569182019

References

1. Chatterjee, Samprit, and Ali S. Hadi. Regression Analysis by Example. 5 ed., Wiley, 2013.

2. “Lecture54 (Data2Decision) Principle Components in R.” Performance by Chris Mack, Youtube, 4 Nov. 2016, Accessed 18 Nov. 2021.