PROJECT REPORT

Introduction

In this project I was required to analyze the Linthurst data and identify the important physicochemical properties of the substrate influencing the aerial biomass production in the Cape Far Estuary of North Carolina. I have used linear regression models to analyze this dataset.

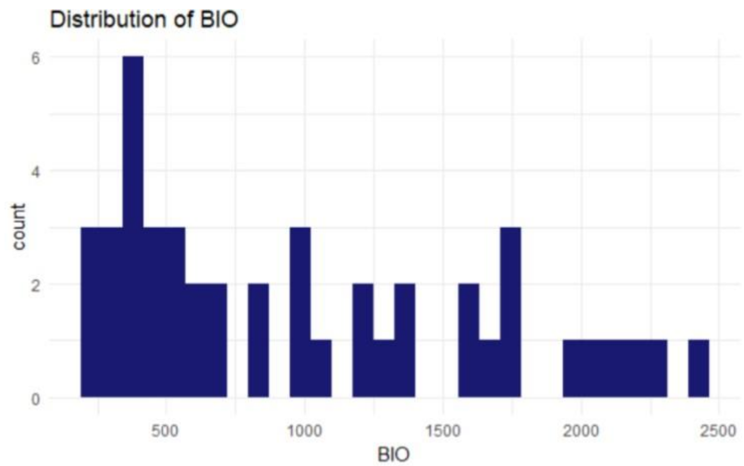
The full multiple linear regression model is

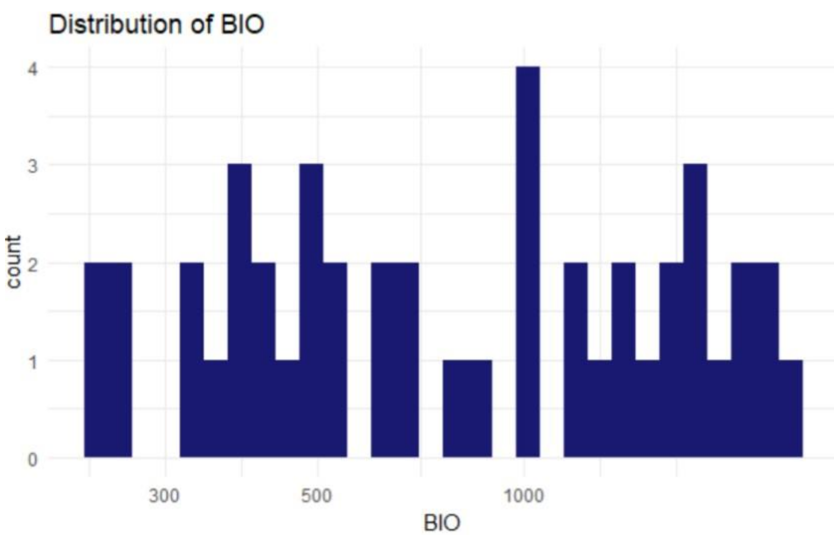
Y ~ X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+X11+X12+X13+X14 The dataset is shown here:

OBS	BIO	H2S	SAL	Eh7	pH	BUF	P	K	Ca	Mg	Na	Mn	Zn	Cu	NH4
1	676	-610	33	-290	5.00	2.34	20.238	1441.67	2150.00	5169.05	35184.5	14.2857	16.4524	5.02381	59.524
2	516	-570	35	-268	4.75	2.66	15.591	1299.19	1844.76	4358.03	28170.4	7.7285	13.9852	4.19019	51.378
3	1052	-610	32	-282	4.20	4.18	18.716	1154.27	1750.36	4041.27	26455.0	17.8066	15.3276	4.79221	68.788
4	868	-560	30	-232	4.40	3.60	22.821	1045.15	1674.36	3966.08	25072.9	49.1538	17.3128	4.09487	82.256
5	1008	-610	33	-318	5.55	1.90	37.843	521.62	3360.02	4609.39	31664.2	30.5229	22.3312	4.60131	70.904

Analysis

For the analysis the distribution of BIO had to be checked and we can see that it is rightly skewed as shown here.





PART – I

In this part we need to create an Ordinary Least Square (OLS) estimation to estimate the regression coefficients. The summary of the model is shown below.

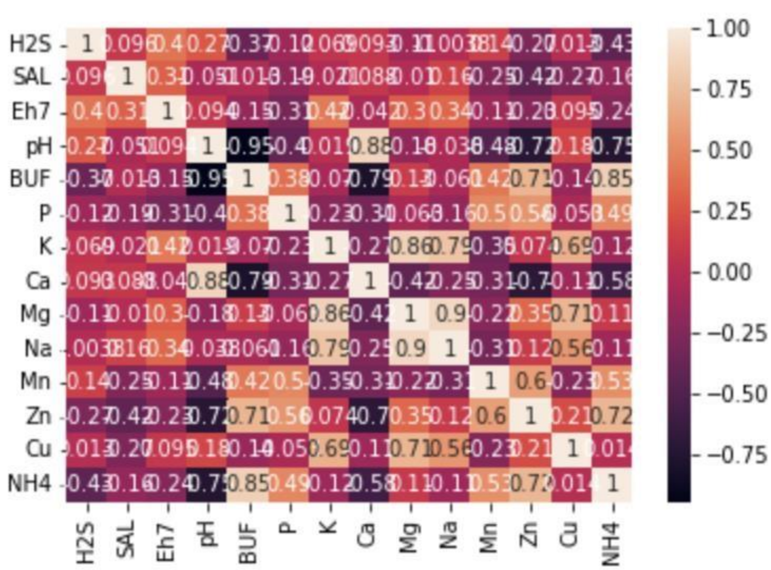
OLS Regression Results

Dep. Variable:	BIO	**R-squared:**	0.807
Model:	OLS	**Adj. R-squared:**	0.718
Method:	Least Squares	**F-statistic:**	8.983
Date:	Sun, 05 Dec 2021	**Prob (F-statistic):**	3.07e-07
Time:	18:42:53	**Log-Likelihood:**	-318.44
No. Observations:	45	**AIC:**	666.9
Df Residuals:	30	**BIC:**	694.0
Df Model:	14		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	2909.9341	3412.898	0.853	0.401	-4060.133	9880.001
H2S	0.4290	2.998	0.143	0.887	-5.694	6.552
SAL	-23.9807	26.169	-0.916	0.367	-77.426	29.464
Eh7	2.5532	2.012	1.269	0.214	-1.557	6.663
pH	242.5278	334.173	0.726	0.474	-439.945	925.001
BUF	-6.9023	123.821	-0.056	0.956	-259.779	245.974
P	-1.7015	2.640	-0.645	0.524	-7.092	3.689
K	-1.0466	0.482	-2.170	0.038	-2.032	-0.061
Ca	-0.1161	0.126	-0.924	0.363	-0.373	0.141
Mg	-0.2802	0.274	-1.021	0.315	-0.841	0.280
Na	0.0045	0.025	0.180	0.858	-0.046	0.055
Mn	-1.6788	5.373	-0.312	0.757	-12.652	9.295
Zn	-18.7945	21.780	-0.863	0.395	-63.276	25.687
Cu	345.1628	112.078	3.080	0.004	116.269	574.056
NH4	-2.7052	3.238	-0.835	0.410	-9.318	3.908

Omnibus:	10.874	**Durbin-Watson:**	1.907
Prob(Omnibus):	0.004	**Jarque-Bera (JB):**	14.037
Skew:	0.742	**Prob(JB):**	0.000895
Kurtosis:	5.299	**Cond. No.**	1.20e+06

I also created a correlation matrix to check the collinearity. The correlation matrix for the dataset:



	H2S	SAL	Eh7	pH	BUF	P	K	Ca	Mg	Na	Mn	Zn	Cu	NH4
H2S	1.000000	0.095809	0.399655	0.273529	-0.373831	-0.115394	0.068963	0.093307	-0.107822	-0.003763	0.141541	-0.272398	0.012719	-0.426213
SAL	0.095809	1.000000	0.309299	-0.051333	-0.012533	-0.185678	-0.020633	0.087978	-0.010043	0.162266	-0.253584	-0.420834	-0.266004	-0.156835
Eh7	0.399655	0.309299	1.000000	0.094018	-0.153083	-0.305431	0.422611	-0.042121	0.298503	0.342463	-0.111255	-0.232005	0.094544	-0.238966
pH	0.273529	-0.051333	0.094018	1.000000	-0.946372	-0.401372	0.019228	0.877978	-0.176148	-0.037720	-0.475143	-0.722167	0.181354	-0.745959
BUF	-0.373831	-0.012533	-0.153083	-0.946372	1.000000	0.382936	-0.070247	-0.791080	0.130459	-0.060714	0.420357	0.714683	-0.143153	0.849488
P	-0.115394	-0.185678	-0.305431	-0.401372	0.382936	1.000000	-0.226473	-0.306692	-0.063237	-0.163228	0.495410	0.557407	-0.053137	0.489739
K	0.068963	-0.020633	0.422611	0.019228	-0.070247	-0.226473	1.000000	-0.265206	0.862245	0.792096	-0.347455	0.073609	0.693051	-0.117581
Ca	0.093307	0.087978	-0.042121	0.877978	-0.791080	-0.306692	-0.265206	1.000000	-0.418446	-0.248187	-0.308985	-0.699866	-0.112247	-0.582609
Mg	-0.107822	-0.010043	0.298503	-0.176148	0.130459	-0.063237	0.862245	-0.418446	1.000000	0.899470	-0.219390	0.345217	0.712069	0.108226
Na	-0.003763	0.162266	0.342463	-0.037720	-0.060714	-0.163228	0.792096	-0.248187	0.899470	1.000000	-0.310061	0.117047	0.560069	-0.107024
Mn	0.141541	-0.253584	-0.111255	-0.475143	0.420357	0.495410	-0.347455	-0.308985	-0.219390	-0.310061	1.000000	0.603323	-0.233468	0.527021
Zn	-0.272398	-0.420834	-0.232005	-0.722167	0.714683	0.557407	0.073609	-0.699866	0.345217	0.117047	0.603323	1.000000	0.212102	0.720679
Cu	0.012719	-0.266004	0.094544	0.181354	-0.143153	-0.053137	0.693051	-0.112247	0.712069	0.560069	-0.233468	0.212102	1.000000	0.013657
NH4	-0.426213	-0.156835	-0.238966	-0.745959	0.849488	0.489739	-0.117581	-0.582609	0.108226	-0.107024	0.527021	0.720679	0.013657	1.000000

To find out the collinearity of the 14-predictor dataset, I applied some code to it. To confirm whether collinearity is present or not I applied a condition which is 22.7>15, and so collinearity exists. And to how that I created a correlation matrix to better understand the data. We can see high correlation between pH and Ca which can also be seen in the heat map above.

I then plotted a multiple linear regression model with BIO as the predictor variable and H2S, SAL, Eh7, pH, BUF, P, K, Ca, Mg, Na, Mn, Zn, CU and NH4 as the response variables. Only Eh7, K and Cu had a p-value of less than the significance value of 0.05. This means that they are the only elements that statistically have a significant effect on BIO. A unit increase in H2S results in an increase in BIO. A unit increase in SAL results in a decrease in BIO. A unit increase in Eh7 results in an increase in BIO. A unit increase in pH results in an increase in BIO. A unit increase in BUF results in a decrease in BIO. A unit increase in P results in a decrease in BIO. A unit increase in K results in a decrease in BIO. A unit increase in Ca results in a decrease in BIO. A unit increase in Mg results in a decrease in BIO. A unit increase in Na results in an increase in BI. A unit increase in Mn results in a decrease in BIO. A unit increase in Zn results in a decrease in BIO. A unit increase in Cu results in an increase in BIO. A unit increase in NH4 results in a decrease in BIO. I conducted a collinearity diagnostic, and the following variables have VIF scores of greater than 10; pH, BUF, Ca, Mg, Na and Zn. Collinearity exists is the variables mentioned above.

I also decided to find out the collinearity using eign values of the correlation matrix and here is the result I got:

```
[1.0,
 1.1543421199134034,
 1.7503706145221423,
 1.9205711388302702,
 2.6682605881301633,
 3.1363506616759906,
 3.574190384264785,
 3.5960198810762334,
 5.44680089907441,
 5.868093621736142,
 7.528834952129797,
 22.77520296800637,
 10.42697614897929,
 12.84292260061667]
```

This means that there is collinearity present between pH, BUF, Ca, Mg, Na, Mn, and NH4, and affects our ability to use SLR on the base set of variables. Instead, we must look at other options to reduce the collinearity present in this data.

PART - II

In the second part I used Principal Component Analysis (PCA) to check which components need to be added in the model. This is the result I got from the analysis:

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
0	1.112151	4.275722	-0.186982	0.612804	0.818981	0.024611	-0.433643	-0.733046	0.363693	-0.014452	0.053981	-0.054142	0.049663	0.053001
1	1.599943	3.041830	-1.603856	-0.094490	0.444074	0.509136	-0.245642	-0.009583	0.305317	-0.152675	-0.318981	-0.235367	0.181633	-0.065849
2	0.445421	2.568878	-0.643423	0.569739	-0.007717	-0.010850	-0.226922	0.068273	-0.035248	0.046137	0.151764	0.428004	0.015452	-0.057882
3	0.342600	2.064535	-1.452285	-1.770957	-0.363058	-0.614552	-0.480617	-0.113295	0.170582	-0.517821	0.130711	0.119400	-0.084190	-0.096411
4	0.708068	1.595353	0.462733	0.407358	1.645849	-0.249627	-0.117282	-1.471000	-1.073394	-1.396635	0.103036	-0.196953	-0.306649	0.038709

Using PCA I regressed Y on both sides to check my code.

$$
\begin{aligned}
\hat{\theta}_1 &= 0.532\hat{\alpha}_1 - 0.024\hat{\alpha}_2 - 0.668\hat{\alpha}_3 + 0.074\hat{\alpha}_4 - 0.514\hat{\alpha}_5, \\
\hat{\theta}_2 &= -0.232\hat{\alpha}_1 + 0.825\hat{\alpha}_2 + 0.158\hat{\alpha}_3 - 0.037\hat{\alpha}_4 - 0.489\hat{\alpha}_5, \\
\hat{\theta}_3 &= -0.389\hat{\alpha}_1 - 0.022\hat{\alpha}_2 - 0.217\hat{\alpha}_3 + 0.895\hat{\alpha}_4 + 0.010\hat{\alpha}_5, \\
\hat{\theta}_4 &= 0.395\hat{\alpha}_1 - 0.260\hat{\alpha}_2 + 0.692\hat{\alpha}_3 + 0.338\hat{\alpha}_4 - 0.428\hat{\alpha}_5, \\
\hat{\theta}_5 &= -0.595\hat{\alpha}_1 - 0.501\hat{\alpha}_2 - 0.057\hat{\alpha}_3 - 0.279\hat{\alpha}_4 - 0.559\hat{\alpha}_5.
\end{aligned}
$$

OLS Regression Results

Dep. Variable:	y	**R-squared (uncentered):**	0.807
Model:	OLS	**Adj. R-squared (uncentered):**	0.720
Method:	Least Squares	**F-statistic:**	9.283
Date:	Sun, 05 Dec 2021	**Prob (F-statistic):**	1.58e-07
Time:	18:42:55	**Log-Likelihood:**	-26.791
No. Observations:	45	**AIC:**	81.58
Df Residuals:	31	**BIC:**	106.9
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
C1	0.3275	0.036	9.219	0.000	0.255	0.400
C2	-0.1195	0.041	-2.915	0.007	-0.203	-0.036
C3	0.2018	0.062	3.245	0.003	0.075	0.329
C4	-0.1392	0.068	-2.040	0.050	-0.278	-4.29e-05
C5	-0.1276	0.095	-1.346	0.188	-0.321	0.066
C6	0.0311	0.111	0.279	0.782	-0.196	0.258
C7	0.2059	0.127	1.622	0.115	-0.053	0.465
C8	0.2946	0.128	2.306	0.028	0.034	0.555
C9	-0.5013	0.193	-2.591	0.014	-0.896	-0.107
C10	-0.2028	0.208	-0.973	0.338	-0.628	0.222
C11	0.5096	0.267	1.906	0.066	-0.036	1.055
C12	0.2171	0.370	0.586	0.562	-0.538	0.972
C13	0.0610	0.456	0.134	0.895	-0.869	0.991
C14	0.4358	0.809	0.539	0.594	-1.214	2.086

Omnibus:	10.874	**Durbin-Watson:**	1.907
Prob(Omnibus):	0.004	**Jarque-Bera (JB):**	14.037
Skew:	0.742	**Prob(JB):**	0.000895
Kurtosis:	5.299	**Cond. No.**	22.8

I used the following eleven variables for prediction; H2S, SAL, Eh7, P, K, Ca, Na, Mn, Zn, CU and NH4. For this model, SAL, K, Cu and Zn had a p-value of less than 0.05 implying that they statistically have a significant effect on BIO. A unit increase in H2S results in an increase in BIO. A unit increase in SAL results in a decrease in BIO. A unit increase in Eh7 results in an increase in BIO. A unit increase in P results in a decrease in BIO. A unit increase in K results in a decrease in BIO. A unit increase in Ca results in an increase in BIO. A unit increase in Na results in a decrease in BIO. A unit increase in Mn results in a decrease in BIO. A unit increase in Zn results in a decrease in BIO. A unit increase in Cu results in an increase in BIO. A unit increase in NH4 results in a decrease in BIO. The summary of the model is shown above.

```
model with regression up to C1 R^2= 0.528
model with regression up to C2 R^2= 0.5808
model with regression up to C3 R^2= 0.6462
model with regression up to C4 R^2= 0.6721
model with regression up to C5 R^2= 0.6833
model with regression up to C6 R^2= 0.6838
model with regression up to C7 R^2= 0.7002
model with regression up to C8 R^2= 0.7332
model with regression up to C9 R^2= 0.7749
model with regression up to C10 R^2= 0.7808
model with regression up to C11 R^2= 0.8034
model with regression up to C12 R^2= 0.8055
model with regression up to C13 R^2= 0.8056
model with regression up to C14 R^2= 0.8074
```

we can use OLS regression and achieve the following estimation model:

$$Y\sim_{est}=0.328C_1-0.120C_2+0.202C_3-0.139C_4-0.128C_5+0.031C_6+0.206C_7+0.295C_8-0.501C_9-0.203C_{10}+0.510C_{11}+0.217C_{12}+0.061C_{13}+0.436C_{14}$$

Then, I calculate the $theta_i$ values, the coefficients for the standardized_y ~ standardized_X regression, by computing V.Alpha, with each excluded Ci's alpha set equal to 0. The values for all the thetas and betas are shown below for all the models along with the value of beta0.

The results are as in the following table:

```
1 thetas
```

	model 1	model 2	model 3	model 4	model 5	model 6	model 7	model 8	model 9	model 10	model 11	model 12	model 13	model 14
0	0.054	0.053	0.006	0.102	0.100	0.113	0.175	0.196	0.112	0.172	0.058	0.055	0.055	0.020
1	0.035	0.033	-0.089	-0.127	-0.191	-0.192	-0.113	-0.143	-0.055	-0.101	-0.146	-0.180	-0.174	-0.135
2	0.041	0.014	-0.079	-0.037	-0.016	-0.034	-0.095	-0.003	0.110	0.127	0.139	0.151	0.153	0.143
3	0.134	0.137	0.194	0.205	0.194	0.188	0.199	0.208	0.196	0.226	0.205	0.133	0.131	0.458
4	-0.135	-0.135	-0.176	-0.199	-0.178	-0.178	-0.154	-0.109	-0.156	-0.136	-0.312	-0.213	-0.234	-0.026
5	-0.089	-0.076	-0.044	-0.016	-0.111	-0.111	-0.180	-0.063	-0.102	-0.105	-0.088	-0.074	-0.078	-0.071
6	0.011	-0.047	-0.043	-0.037	-0.029	-0.028	-0.042	-0.008	-0.289	-0.401	-0.512	-0.519	-0.504	-0.472
7	0.117	0.139	0.181	0.173	0.147	0.133	0.155	0.100	0.007	0.022	-0.239	-0.164	-0.168	-0.302
8	-0.026	-0.085	-0.075	-0.081	-0.094	-0.095	-0.104	-0.154	-0.149	-0.126	-0.187	-0.273	-0.315	-0.399
9	0.006	-0.051	-0.061	-0.068	-0.099	-0.097	-0.134	-0.267	-0.311	-0.222	-0.112	-0.033	-0.016	0.046
10	-0.091	-0.069	-0.073	-0.006	-0.011	-0.020	0.006	-0.151	-0.194	-0.268	-0.130	-0.113	-0.124	-0.062
11	-0.132	-0.143	-0.107	-0.087	-0.086	-0.087	-0.102	-0.163	0.057	0.060	-0.230	-0.278	-0.254	-0.236
12	0.004	-0.043	0.033	0.047	0.039	0.036	0.152	0.234	0.423	0.396	0.495	0.561	0.561	0.542
13	-0.131	-0.127	-0.125	-0.140	-0.139	-0.151	-0.070	-0.027	-0.237	-0.158	-0.091	-0.157	-0.143	-0.194

```
1 beta
```

```
[1.2472310858941984,
 -25.908214059222704,
 2.4827073093744088,
 108.5138005018635,
 -82.1689295233718,
 -2.105613323151283,
 -1.1356100516405887,
 -0.09180949765663837,
 -0.131396430933868,
 -0.01074168079257986,
 -3.505237217386276,
 -18.335944629105118,
 315.08339883248823,
 -1.270650546358688]
```

```
1 beta0
```

```
4222.976532878156
```

Component variables are required for an accurate model and proceed to convert the thetas for model to our original betas. Thus, we arrive at an estimate, in terms of the original variables, of:

$$y_{est}=4222.98+1.25X_1-25.91X_2+2.48X_3+108.51X_4-82.17X_5-2.11X_6-1.14X_7-0.09X_8-0.13X_9-0.01X_{10}-3.51X_{11}-18.34X_{12}+315.08X_{13}-1.27X_{14}$$

In comparison with our original results, the coefficients have changed drastically, and with the use of PCR technique we can see that the new estimation is better than the original.

PART – III

Y ~X2+X4+X7+X10+X12

For this part I considered only few data BIO, SAL, pH, K, Na and Zn as the predictor variables yet it preserved some collinearity problem. Here stepwise regression is used to decide which is the best fit model. These variables selected are those that have a significant effect on BIO. The summary of the model is shown below.

OLS Regression Results

Dep. Variable:	BIO	**R-squared:**	0.677
Model:	OLS	**Adj. R-squared:**	0.636
Method:	Least Squares	**F-statistic:**	16.37
Date:	Sun, 05 Dec 2021	**Prob (F-statistic):**	1.08e-08
Time:	20:29:41	**Log-Likelihood:**	-330.05
No. Observations:	45	**AIC:**	672.1
Df Residuals:	39	**BIC:**	682.9
Df Model:	5		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	1252.2794	1234.750	1.014	0.317	-1245.239	3749.798
SAL	-30.2851	24.031	-1.260	0.215	-78.893	18.323
pH	305.5254	87.879	3.477	0.001	127.774	483.277
K	-0.2851	0.348	-0.818	0.418	-0.990	0.420
Na	-0.0087	0.016	-0.544	0.589	-0.041	0.024
Zn	-20.6764	15.055	-1.373	0.177	-51.127	9.774

Omnibus:	8.971	**Durbin-Watson:**	1.069
Prob(Omnibus):	0.011	**Jarque-Bera (JB):**	8.050
Skew:	0.938	**Prob(JB):**	0.0179
Kurtosis:	3.878	**Cond. No.**	3.74e+05

Stepwise regression method results:

```
const     0.064876
SAL       0.500058
dtype: float64
const     7.349989e-04
pH        4.433213e-10
dtype: float64
const     0.000017
K         0.177500
dtype: float64
const     0.000001
Na        0.070604
dtype: float64
const     5.890805e-13
Zn        4.566092e-06
dtype: float64
```

```
const     3.676763e-01
pH        7.171988e-10
SAL       5.169516e-01
dtype: float64
const     7.712215e-02
pH        1.148936e-10
K         2.105660e-02
dtype: float64
const     8.931536e-02
pH        1.220284e-10
Na        1.007761e-02
dtype: float64
const     0.379211
pH        0.000018
Zn        0.333797
dtype: float64
```

```
const     5.398634e-01
pH        2.122008e-10
Na        1.333140e-02
SAL       7.871337e-01
dtype: float64
const     1.199157e-01
pH        1.820207e-10
Na        2.212038e-01
K         6.321722e-01
dtype: float64
const     0.690023
pH        0.000004
Na        0.014071
Zn        0.488756
dtype: float64
```

To begin, a regression on each of the 5 variables, individually, results in the following p-values:

	SAL	pH	K	Na	Zn
p – value	0.500058	4.433213e-10	0.177500	0.070604	4.566092e-06

Here, pH, Na, and Zn are less than the provided $\alpha_E = \alpha_R = 0.15$ value. Since pH is the smallest value, we will begin by entering pH first.

$$Y \sim pH + \{SAL, K, Na, Zn\} Individually$$

These regressions result in the following p-values:

	SAL	K	Na	Zn
p – value	5.169516e-01	2.105660e-02	1.007761e-02	0.333797

Since pH is the smallest value, we will begin by entering pH first.
$Y \sim pH + \{SAL, K, Na, Zn\}$-Individual

These regressions result in the following p-values:

	SAL	K	Zn
p – value	7.871337e-01	6.321722e-01	0.488756

Here, all p-values are greater than the provided $\alpha_E = \alpha_R = 0.15$ value. Thus we stop here and accept the variables {pH,Na} in the final model.

Thus, our model with reduced collinearity will be:

$$Y \sim pH + K$$

Subset :

$yest$=1252.28−30.29X1+305.53X2−0.29X3−0.009X4−20.68X5

These models, and their corresponding C_P values are in the following table, using the formula from the text:

$$C_p = \frac{\mathrm{SSE}_p}{\hat{\sigma}^2} + (2p - n),$$

I have used the subset selection method to select the best-two variable model. The final variables were pH and Na. The summary of the sub-set selection is shown below and I have used VIF to break the tie.

```
8.93307991706591
75.40335753168574
72.48070912729082
15.070426388832018
3.5937361149892624
2.2815921539121646
8.34296479530905
72.87483611257235
31.67821760934585
29.886192091557888
```

The VIFs are simple to calculate as, and here we find:

$$\mathrm{VIF}_j = \frac{1}{1 - R_j^2}$$

```
  feature       VIF
0   const  21.89570
1      pH   1.00037
2       K   1.00037
  feature        VIF
0   const  21.593762
1      pH   1.001425
2      Na   1.001425
```

Thus, we see that the model using {pH, K} is closer to one and therefore can be chosen as the best model, or the reduced model eliminating collinearity. Therefore, we have analysed methods to detect, identify, and reduce collinearity in two sets of the Linthall Data.