Illinois Institute Of Technology Final - Project MATH-564 Fall 2023

Analysis of Linthurst data and identification of the important physicochemical properties of the substrate influencing the aerial biomass production in the Cape Fear Estuary of North Carolina.

Student's Detail:

Sanketkumar Patel : (A20523237)

I) Introduction:

In this project, our goal is to analyze the Linthurst data and identify the important physicochemical properties of the substrate influencing the aerial biomass production in the Cape Fear Estuary of North Carolina.

For Part I & Part II:
The response variable Y is BIO (the biomass production), and there are 14 predictor variables characterizing the soil. For instance, SAL is the percentage of salinity and pH is the acidity in the water, etc. We will use the all-predictor variable data set (LINTHALL.txt) to perform a variable selection procedure.

For Part III:

The data set only has 5 predictor variables and the same response variable as above, and yet it preserved some of the collinearity problem. We will use the 5-predictor data set (LINTH-5.txt) to perform a variable selection procedure.

II) Dataset:

Dataset in LINTHALL.txt has 43 number of observation and following variables:

- Y: BIO
- X1: H2S
- X2: SAL
- X3: Eh7
- X4:pH
- X5: BUF
- X6:P
- X7:K

- X8:Ca
- X9:Mg
- X10: Na
- X11: Mn
- X12: Zn
- X13: Cu
- X14: NH4

Dataset in LINTH-5.txt has 43 number of observation and following variables:

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

For both the data set:
The first column is the index of the observation, the second column ”Loc” and the third column ”Type” are not used in this project.

PART I

Problem Statement: Consider the 14-predictor data set (LINTHALL.txt). Use the ordinary least square estimation to estimate the regression coefficients. Run the collinearity diagnostics and identify if there is any collinearity. Try at least two collinearity diagnostics methods. What is the consistent conclusion you can draw from the two methods?

I) Introduction:

Our interest (or response) parameter is biomass production which is given as a BIO variable.

We aim to get a predicted biomass production using the given 14 predictor variables. To do it we can use OLS regression.

However, it is important to check the collinearity between these variables to get a better result of regression.

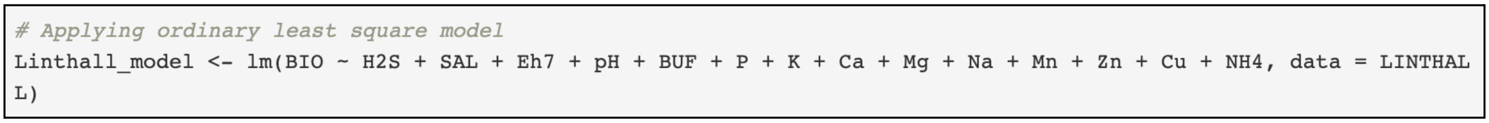
II) OLS Estimation:

As given BIO variable regressed on all other 14 variables. Using lm function in the R environment.

Following is the model which we are regressing.

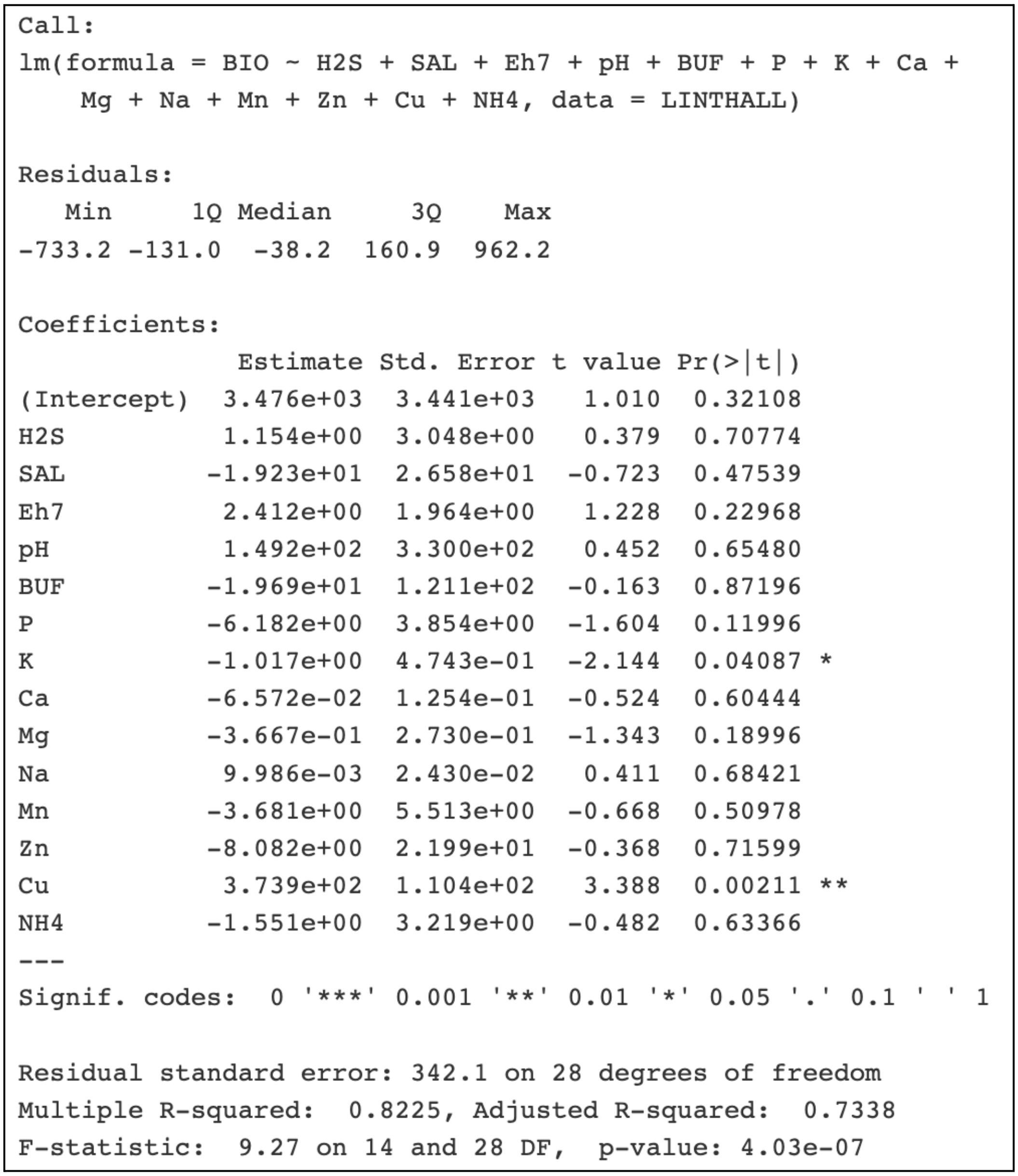
BIO ~ H2S + SAL + Eh7 + pH + BUF + P + K + Ca + Mg + Na + Mn + Zn + Cu + NH4

Snapshot of Code:



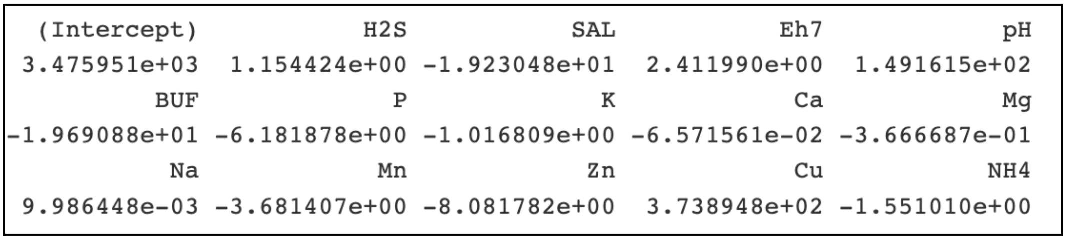
```
# Applying ordinary least square model
Linthall_model <- lm(BIO ~ H2S + SAL + Eh7 + pH + BUF + P + K + Ca + Mg + Na + Mn + Zn + Cu + NH4, data = LINTHAL
L)
```

Summary of OLS Model:



```
Call:
lm(formula = BIO ~ H2S + SAL + Eh7 + pH + BUF + P + K + Ca +
    Mg + Na + Mn + Zn + Cu + NH4, data = LINTHALL)

Residuals:
   Min     1Q Median     3Q    Max 
-733.2 -131.0  -38.2  160.9  962.2 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  3.476e+03  3.441e+03   1.010  0.32108   
H2S          1.154e+00  3.048e+00   0.379  0.70774   
SAL         -1.923e+01  2.658e+01  -0.723  0.47539   
Eh7          2.412e+00  1.964e+00   1.228  0.22968   
pH           1.492e+02  3.300e+02   0.452  0.65480   
BUF         -1.969e+01  1.211e+02  -0.163  0.87196   
P           -6.182e+00  3.854e+00  -1.604  0.11996   
K           -1.017e+00  4.743e-01  -2.144  0.04087 * 
Ca          -6.572e-02  1.254e-01  -0.524  0.60444   
Mg          -3.667e-01  2.730e-01  -1.343  0.18996   
Na           9.986e-03  2.430e-02   0.411  0.68421   
Mn          -3.681e+00  5.513e+00  -0.668  0.50978   
Zn          -8.082e+00  2.199e+01  -0.368  0.71599   
Cu           3.739e+02  1.104e+02   3.388  0.00211 **
NH4         -1.551e+00  3.219e+00  -0.482  0.63366   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 342.1 on 28 degrees of freedom
Multiple R-squared:  0.8225, Adjusted R-squared:  0.7338 
F-statistic:  9.27 on 14 and 28 DF,  p-value: 4.03e-07
```

Following are the values of the **regression coefficient** for the respective variables.



```
  (Intercept)           H2S           SAL           Eh7            pH 
 3.475951e+03  1.154424e+00 -1.923048e+01  2.411990e+00  1.491615e+02 
          BUF             P             K            Ca            Mg 
-1.969088e+01 -6.181878e+00 -1.016809e+00 -6.571561e-02 -3.666687e-01 
           Na            Mn            Zn            Cu           NH4 
 9.986448e-03 -3.681407e+00 -8.081782e+00  3.738948e+02 -1.551010e+00 
```

III) Collinearity Diagnostics:

Multicollinearity refers to a statistical phenomenon in which two or more independent variables in a regression model are highly correlated with each other. In other words, there is a linear relationship among the predictor variables. Multicollinearity can pose challenges and have several potential impacts on regression analysis such as:

- Unreliable Coefficient Estimates
- Inflated Standard Errors
- Misleading Significance Tests
- Difficulty in Variable Selection
- Decreased Precision in Prediction

Hence it is important to check and remove collinearity from data.

Following are the different collinearity methods:

1) Correlation matrix & Correlation matrix Plot: The correlation matrix of the predictor variable can if there is any collinearity between two variables.
2) VIF: Variance inflation factor: if VIF is higher than 10 then collinearity is present in the dataset.
3) Sum of the reciprocal of eigenvalues: if the sum of the reciprocal of eigenvalue is greater than five times the number of predictor variables then collinearity is present in the data set.
4) Condition index number: if the largest condition number is greater than 15 then collinearity is present in the system

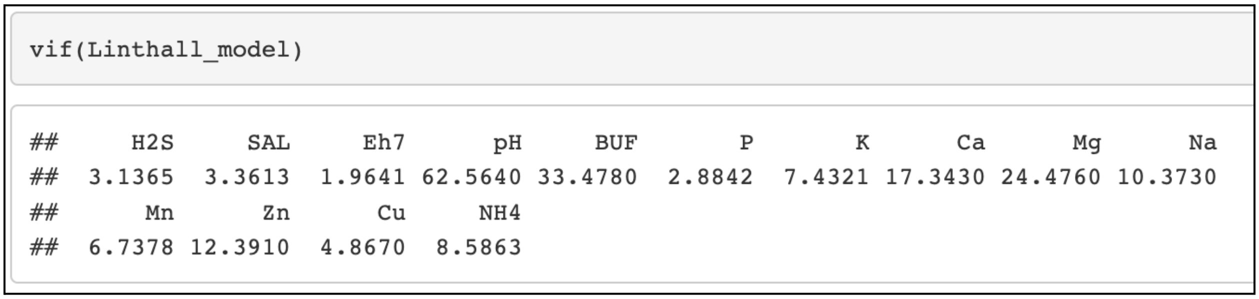
I have implemented the following methods:

1. VIF
2. Correlation matrix & Correlation matrix Plot

a) VIF:

The Variance Inflation Factor (VIF) values measures the extent to which the variance of an estimated regression coefficient increases due to collinearity among the predictor variables, the general rule is that a VIF greater than 10 is indicative of significant collinearity.

Following is the snapshot of the VIF number of the variables for the given data.



```
vif(Linthall_model)
```

```
##     H2S     SAL     Eh7      pH     BUF       P       K      Ca      Mg      Na 
##  3.1365  3.3613  1.9641 62.5640 33.4780  2.8842  7.4321 17.3430 24.4760 10.3730 
##      Mn      Zn      Cu     NH4 
##  6.7378 12.3910  4.8670  8.5863
```

Conclusion:

Variables with Low VIF (Below 10) are:
H2S: 3.14
SAL: 3.36
Eh7: 1.96
P: 2.88
K: 7.43
Mn: 6.74
Cu: 4.87
NH4: 8.59

These variables have low VIF values, suggesting that the collinearity for these variables are not significant.

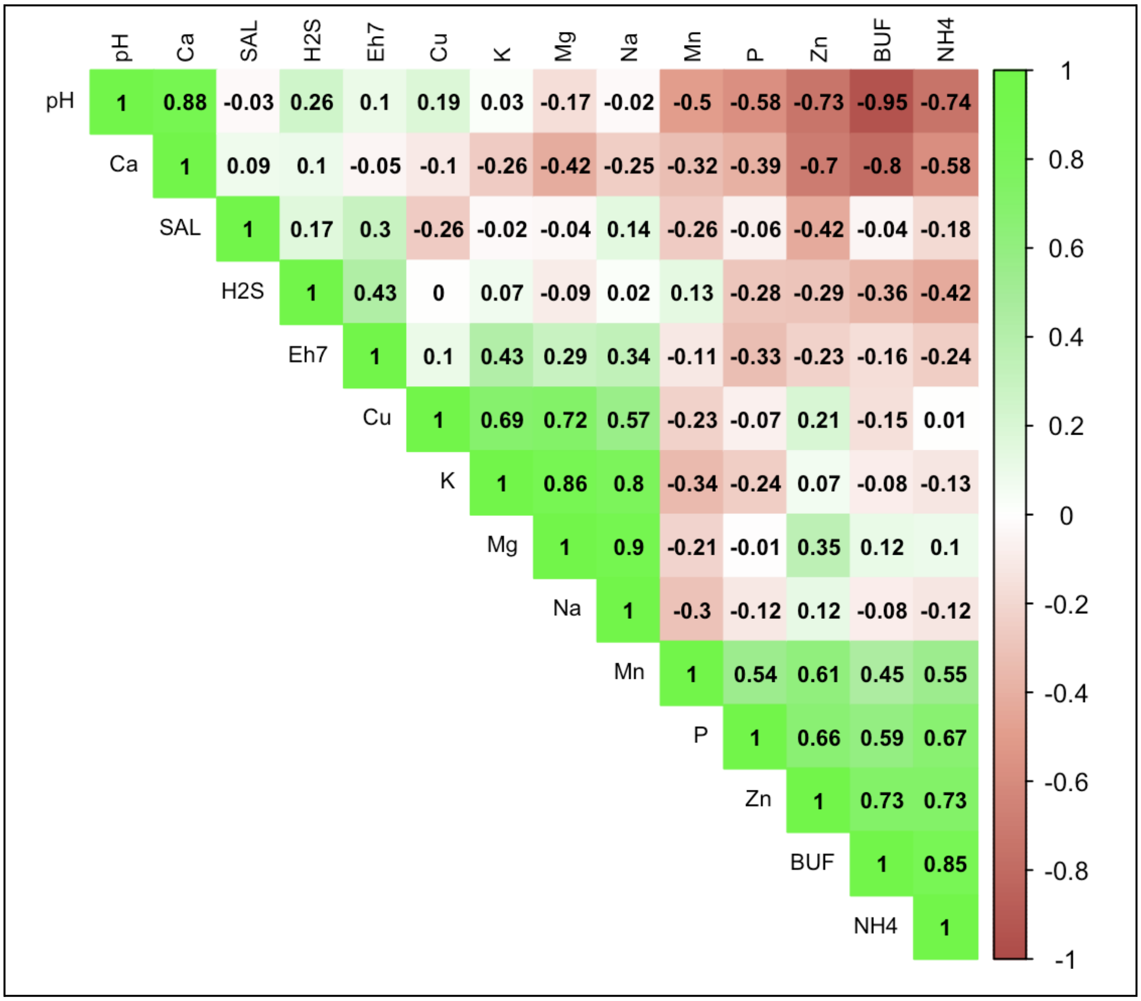
Variables with Moderate to High VIF (Above 10) are
pH: 62.56
BUF: 33.48
Ca: 17.34
Mg: 24.48
Na: 10.37
Zn: 12.39

The variable pH has an extremely high VIF (62.56), and indicating a strong correlation with the other predictors in the model. This high VIF for ‘pH’ suggests that there might be collinearity issues associated with this variable.

In Summary: As VIF for pH, BUF, Ca, Mg, Na & Zn are above 10, suggesting significant collinearity is present in the model.

b) Correlation matrix & Correlation matrix Plot:

The correlation matrix and its visual representation through a correlation matrix plot were utilized to detect collinearities among predictor variables, revealing potential instances of high linear associations that could impact the reliability of regression analysis.



	pH	Ca	SAL	H2S	Eh7	Cu	K	Mg	Na	Mn	P	Zn	BUF	NH4
pH	1	0.88	-0.03	0.26	0.1	0.19	0.03	-0.17	-0.02	-0.5	-0.58	-0.73	-0.95	-0.74
Ca		1	0.09	0.1	-0.05	-0.1	-0.26	-0.42	-0.25	-0.32	-0.39	-0.7	-0.8	-0.58
SAL			1	0.17	0.3	-0.26	-0.02	-0.04	0.14	-0.26	-0.06	-0.42	-0.04	-0.18
H2S				1	0.43	0	0.07	-0.09	0.02	0.13	-0.28	-0.29	-0.36	-0.42
Eh7					1	0.1	0.43	0.29	0.34	-0.11	-0.33	-0.23	-0.16	-0.24
Cu						1	0.69	0.72	0.57	-0.23	-0.07	0.21	-0.15	0.01
K							1	0.86	0.8	-0.34	-0.24	0.07	-0.08	-0.13
Mg								1	0.9	-0.21	-0.01	0.35	0.12	0.1
Na									1	-0.3	-0.12	0.12	-0.08	-0.12
Mn										1	0.54	0.61	0.45	0.55
P											1	0.66	0.59	0.67
Zn												1	0.73	0.73
BUF													1	0.85
NH4														1

Conclusion:

Considering the threshold value as 0.7 we can say that there are some collinearities are there in the data. Following are the sets of variable which shows significant collinearities.

Set 1: pH and Ca
Set 2: pH and Zn
Set 3: pH and BUF
Set 4: pH and NH4

Set 5: Ca and Zn
Set 6: Ca and BUF
Set 7: Cu and K
Set 8: Cu and Mg
Set 9: K and Mg
Set 10: K and Na
Set 11: Mg and Na
Set 12: Zn and BUF
Set 13: Zn and NH4
Set 14: BUF and NH4

Following are the different sets of collinearity.

1. pH, Zn, BUF, Ca || All the variables have high correlations with each other
2. Cu, K, Mg, Na || All the variables have high correlations with each other

following variables don't have significant correlations with any other variable.
H2S, SAL, Eh7, P, Mn, NH4

IV) Consistent Conclusion:

Consistent Conclusion from both methods can be given as follows:

1. Both methods suggest pH, BUF, Ca, Mg, Na, and Zn variables with high collinearity:
2. Result obtained from both methods align with each other ,i.e. Identified sets by the correlation method align with the high VIF variables: pH, Ca, Zn, BUF, Cu, K, Mg, and Na.
3. Variables H2S, SAL, Eh7, P, Mn, and NH4 have low VIF and also does not show strong correlations with any of the other variables in the correlation matrix.
4. High VIF for pH, BUF, Ca, Mg, Na, and Zn, suggests exploring a different method to remove collinearities. method such as variable selection, transformation, and regularization techniques can be beneficial.
5. Sets with high correlation identified by the correlation method (e.g., pH and Ca) reinforce the presence of collinearity and suggest potential variables that could be problematic in regression modeling.
6. The consistency between the two methods enhances the confidence in the findings, as both VIF and correlation analyses independently highlight similar sets of variables with collinearity issues.

PART II

Problem Statement: Consider the 14-predictor data set (LINTHALL.txt). Use the Principle Components Regression method with collinearity reduction to decide which principle components will be included in the model. From the results of Principle Component Regression on the reduced model, compute the regression coefficients βˆj in the original multiple linear regression model. Compare the standard error sum Pj s.e. (βˆj) and SSE with their counterparts in Part I.

I) Introduction:

In this section, we aim to apply Principle Components Regression (PCR) to the 14-predictor dataset (LINTHALL.txt) to mitigate collinearity issues. Collinearity reduction is essential for enhancing the stability of regression models. PCR, by transforming the original predictors into uncorrelated principal components, provides a method to address multicollinearity, allowing for more robust and reliable regression analysis.

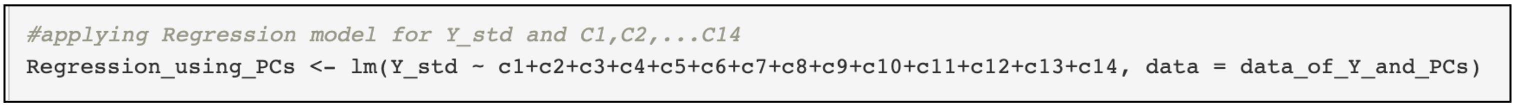
II) Principle Components Regression (PCR):

In this section, we delve into Principle Components Regression (PCR), a technique that transforms the original predictors into uncorrelated principal components. PCR is a powerful method designed to mitigate multicollinearity challenges in multiple linear regression models. By capturing the essence of the data through orthogonal components, PCR offers a valuable approach to enhance the reliability of regression analyses.

Principal components are generated with the help of Eigenvectors and standardized data of predictors.

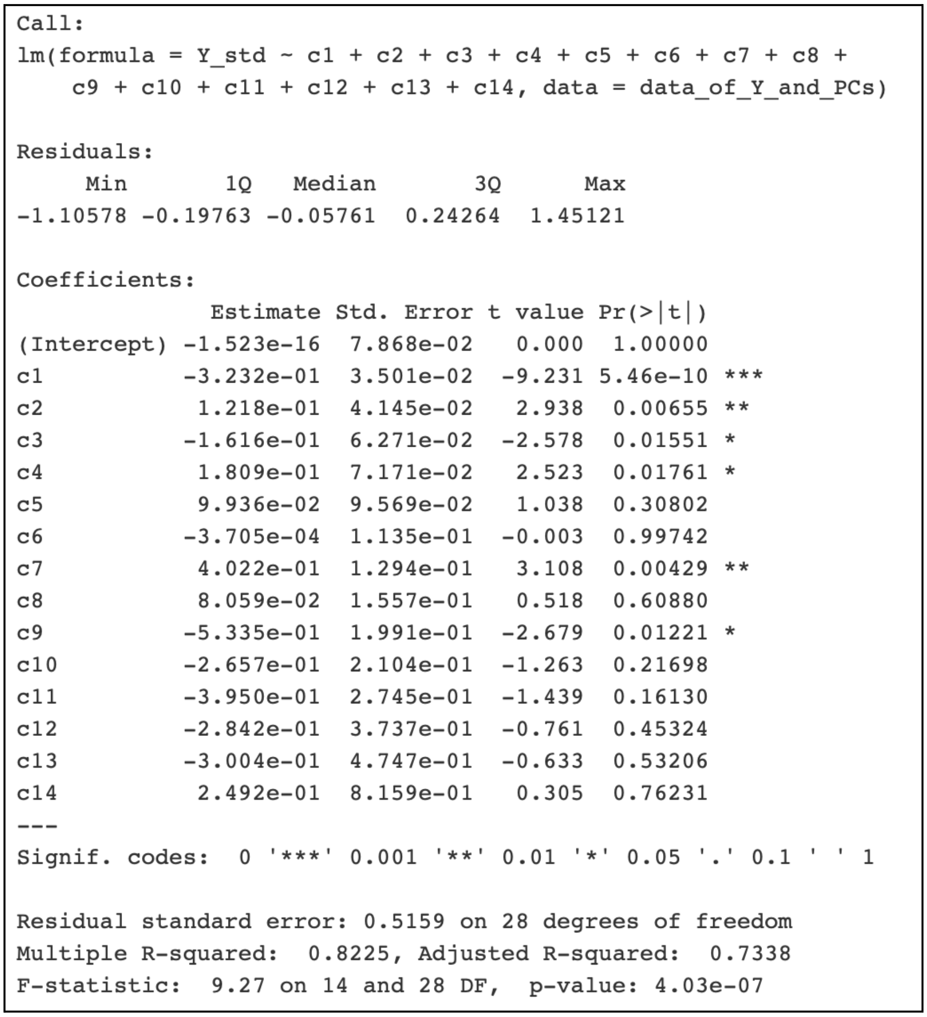
Then using standardized data of BIO (in here response variable) regressed on generated Principal components.

Snapshot of Code for PCR model:



```
#applying Regression model for Y_std and C1,C2,...C14
Regression_using_PCs <- lm(Y_std ~ c1+c2+c3+c4+c5+c6+c7+c8+c9+c10+c11+c12+c13+c14, data = data_of_Y_and_PCs)
```

Snapshot of result for PCR model:



```
Call:
lm(formula = Y_std ~ c1 + c2 + c3 + c4 + c5 + c6 + c7 + c8 +
    c9 + c10 + c11 + c12 + c13 + c14, data = data_of_Y_and_PCs)

Residuals:
     Min       1Q   Median       3Q      Max
-1.10578 -0.19763 -0.05761  0.24264  1.45121

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.523e-16  7.868e-02   0.000  1.00000
c1          -3.232e-01  3.501e-02  -9.231 5.46e-10 ***
c2           1.218e-01  4.145e-02   2.938  0.00655 **
c3          -1.616e-01  6.271e-02  -2.578  0.01551 *
c4           1.809e-01  7.171e-02   2.523  0.01761 *
c5           9.936e-02  9.569e-02   1.038  0.30802
c6          -3.705e-04  1.135e-01  -0.003  0.99742
c7           4.022e-01  1.294e-01   3.108  0.00429 **
c8           8.059e-02  1.557e-01   0.518  0.60880
c9          -5.335e-01  1.991e-01  -2.679  0.01221 *
c10         -2.657e-01  2.104e-01  -1.263  0.21698
c11         -3.950e-01  2.745e-01  -1.439  0.16130
c12         -2.842e-01  3.737e-01  -0.761  0.45324
c13         -3.004e-01  4.747e-01  -0.633  0.53206
c14          2.492e-01  8.159e-01   0.305  0.76231
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5159 on 28 degrees of freedom
Multiple R-squared:  0.8225, Adjusted R-squared:  0.7338
F-statistic:  9.27 on 14 and 28 DF,  p-value: 4.03e-07
```

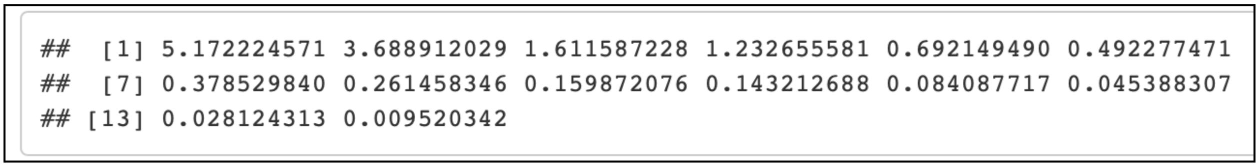
III) Collinearity Reduction:

For the collinearity reduction three methods were used.

1. Eigenvalues
2. Explained variance
3. Graph of RMSEP, MSEP, and R2.

1. **Eigenvalues:**
   lamda_j is the variance of the jth PC, if lamda_j is approximately zero, the corresponding PC, Cj, is approximately equal to a constant.

Following are the Eigenvalues of the data.



```
##  [1] 5.172224571 3.688912029 1.611587228 1.232655581 0.692149490 0.492277471
##  [7] 0.378529840 0.261458346 0.159872076 0.143212688 0.084087717 0.045388307
## [13] 0.028124313 0.009520342
```

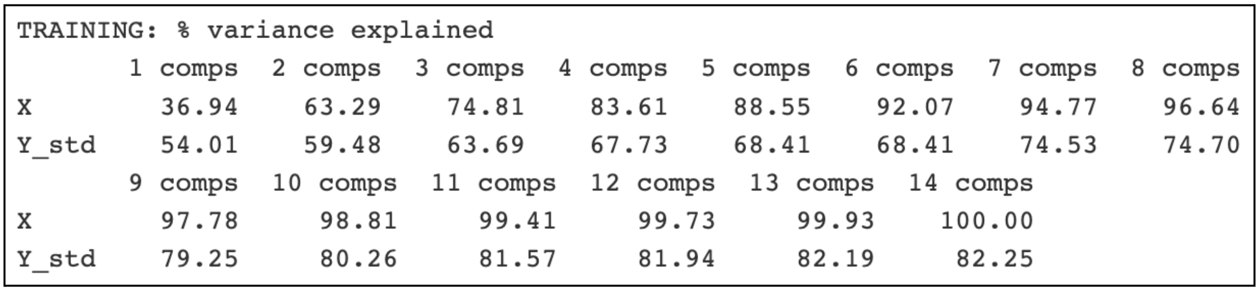
If we observe the above eigenvalue then we can observe that lamda12, lamda13 & lamda14 are very close to zero.

hence we can drop C12,C13 & C14 PCs.

2. **Explained variance:**

Fixing the cutoff value of explained variance as some value. The number of PCs can be decided.

Following is the snapshot of variance explained by the PCR model.

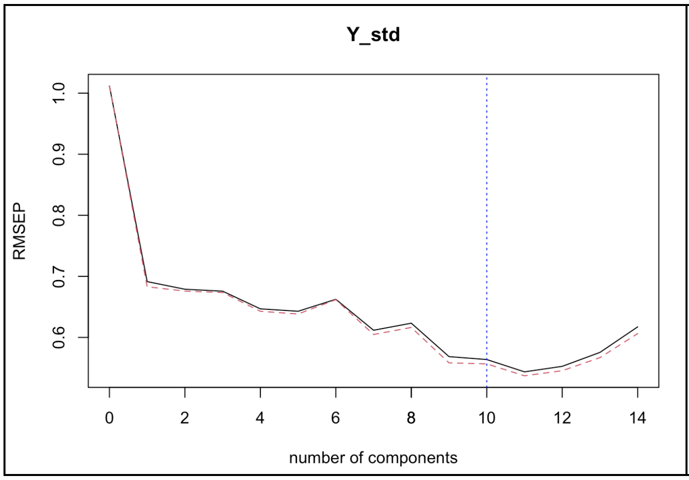


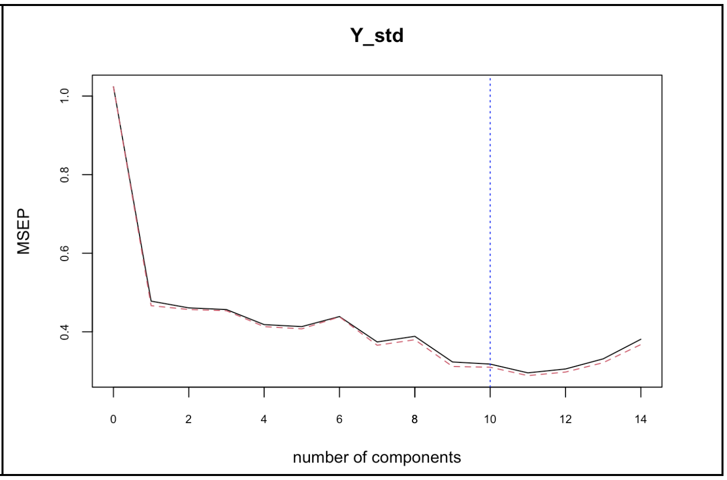
```
TRAINING: % variance explained
       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X        36.94    63.29    74.81    83.61    88.55    92.07    94.77    96.64
Y_std    54.01    59.48    63.69    67.73    68.41    68.41    74.53    74.70
       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps
X        97.78     98.81     99.41     99.73     99.93    100.00
Y_std    79.25     80.26     81.57     81.94     82.19     82.25
```

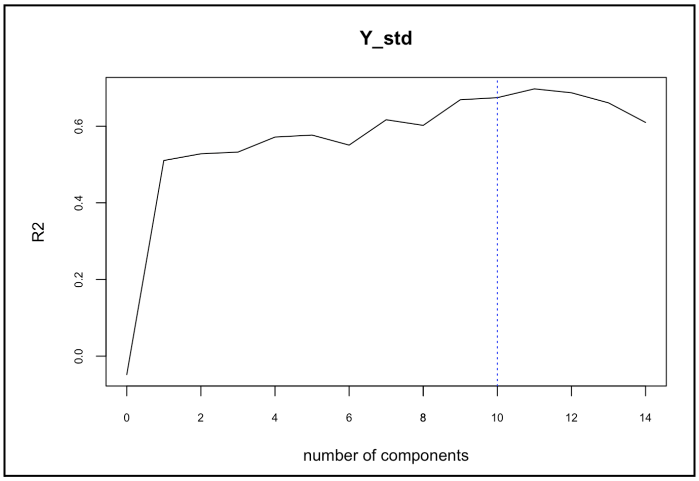
From the above, it can be concluded that as explained variance becomes 99.41 with the first 11 principal components which is above our desired limit of variance. Hence we can drop C12, C13 & C14

3. **Graph of RMSEP, MSEP, and R2:**

The following are the generated graphs







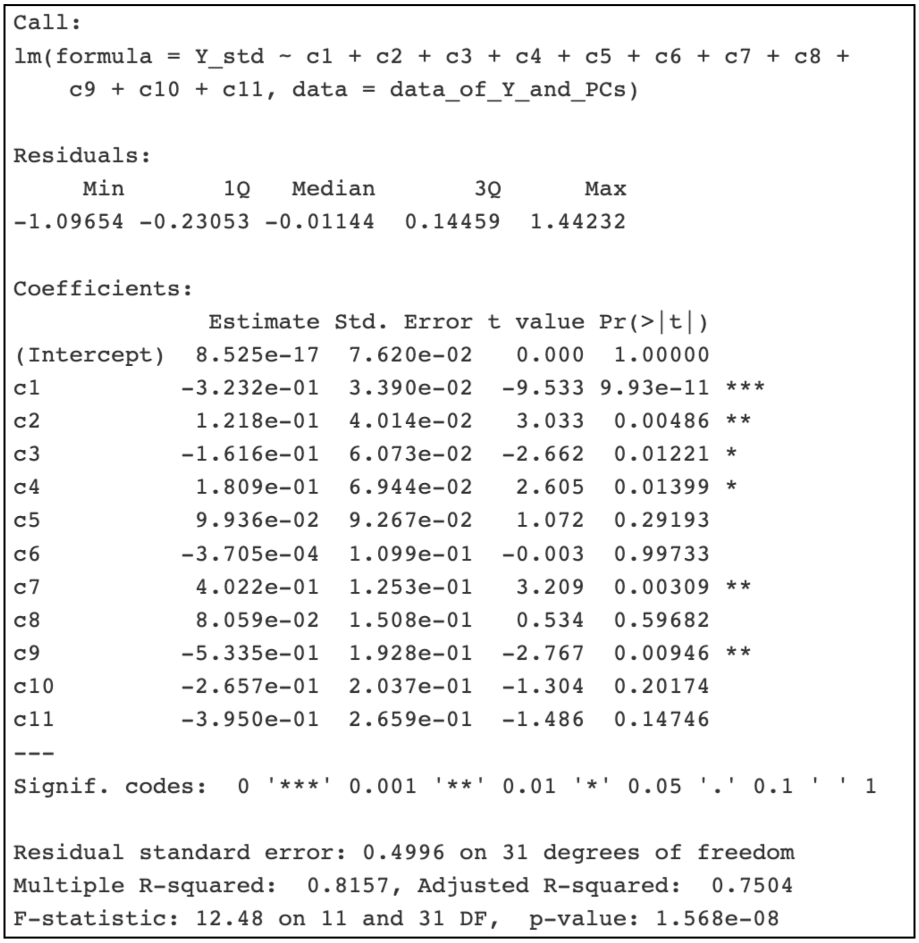
Hence we should proceed with 11 principal components as RMSEP & MSEP are optimal with 11 number of principal components and R^2 is at its peak with 11 principal components. all three parameters are changing in the desirable direction. we should remove 3 PCs.

Hence all three methods suggest keeping 11 PCs.

IV) Principle Components Regression for reduced model:

Again reduced model considering 11 principal components regressed for BIO standardized data.

Snapshot of reduced model:

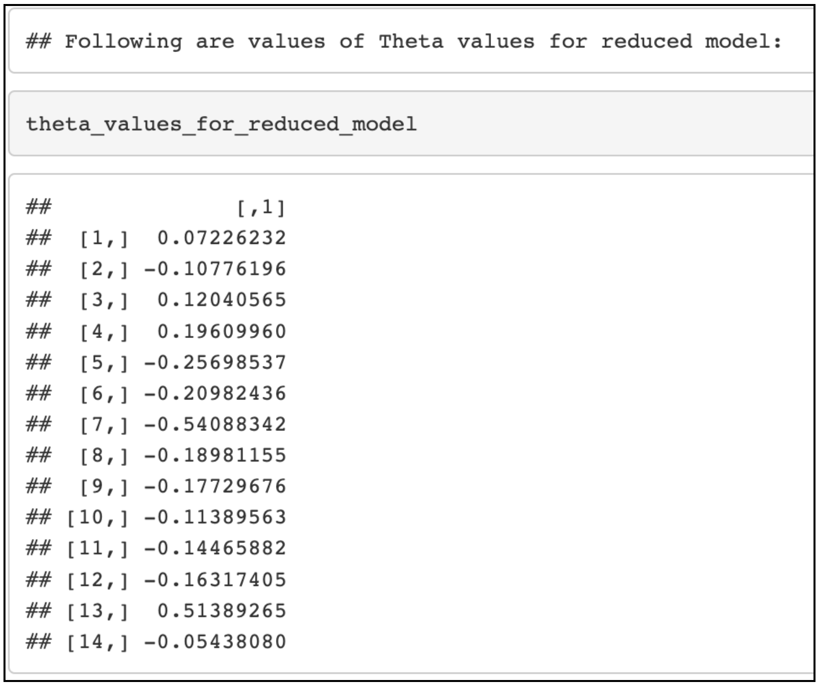


```
Call:
lm(formula = Y_std ~ c1 + c2 + c3 + c4 + c5 + c6 + c7 + c8 +
    c9 + c10 + c11, data = data_of_Y_and_PCs)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.09654 -0.23053 -0.01144  0.14459  1.44232 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.525e-17  7.620e-02   0.000  1.00000    
c1          -3.232e-01  3.390e-02  -9.533 9.93e-11 ***
c2           1.218e-01  4.014e-02   3.033  0.00486 ** 
c3          -1.616e-01  6.073e-02  -2.662  0.01221 *  
c4           1.809e-01  6.944e-02   2.605  0.01399 *  
c5           9.936e-02  9.267e-02   1.072  0.29193    
c6          -3.705e-04  1.099e-01  -0.003  0.99733    
c7           4.022e-01  1.253e-01   3.209  0.00309 ** 
c8           8.059e-02  1.508e-01   0.534  0.59682    
c9          -5.335e-01  1.928e-01  -2.767  0.00946 ** 
c10         -2.657e-01  2.037e-01  -1.304  0.20174    
c11         -3.950e-01  2.659e-01  -1.486  0.14746    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4996 on 31 degrees of freedom
Multiple R-squared:  0.8157, Adjusted R-squared:  0.7504 
F-statistic: 12.48 on 11 and 31 DF,  p-value: 1.568e-08
```

V) Methodology to calculate Regression Coefficients(beta_j) in Original Model:

1. From obtained coefficients of alpha and Standard error of alpha, Theta values and Standard error of theta calculated.

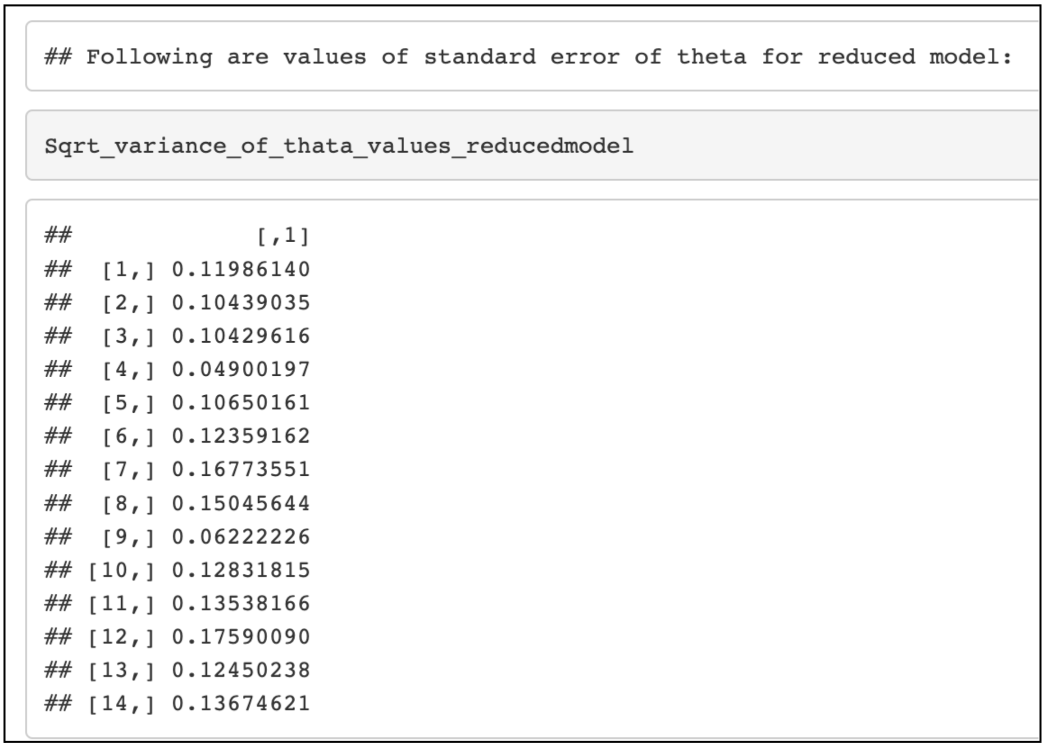
Following are the calculated values of theta values and the Standard error of theta.



```
## Following are values of Theta values for reduced model:
```

```
theta_values_for_reduced_model
```

```
##              [,1]
##  [1,]  0.07226232
##  [2,] -0.10776196
##  [3,]  0.12040565
##  [4,]  0.19609960
##  [5,] -0.25698537
##  [6,] -0.20982436
##  [7,] -0.54088342
##  [8,] -0.18981155
##  [9,] -0.17729676
## [10,] -0.11389563
## [11,] -0.14465882
## [12,] -0.16317405
## [13,]  0.51389265
## [14,] -0.05438080
```



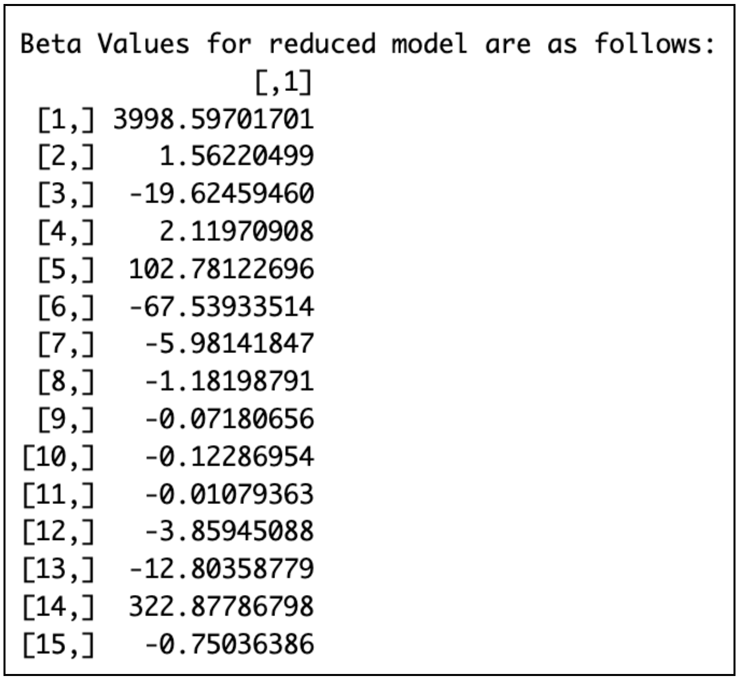
```
## Following are values of standard error of theta for reduced model:
```

```
Sqrt_variance_of_thata_values_reducedmodel
```

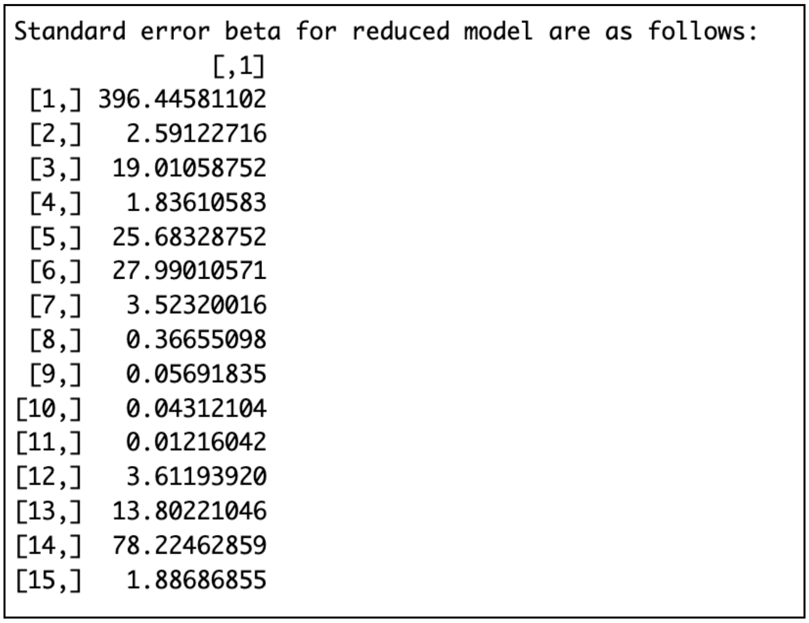
```
##             [,1]
##  [1,] 0.11986140
##  [2,] 0.10439035
##  [3,] 0.10429616
##  [4,] 0.04900197
##  [5,] 0.10650161
##  [6,] 0.12359162
##  [7,] 0.16773551
##  [8,] 0.15045644
##  [9,] 0.06222226
## [10,] 0.12831815
## [11,] 0.13538166
## [12,] 0.17590090
## [13,] 0.12450238
## [14,] 0.13674621
```

2. From Theta values and Standard error of theta, beta values and Standard error of beta are calculated.

Following are the calculated **values of beta values** and the Standard error of beta.

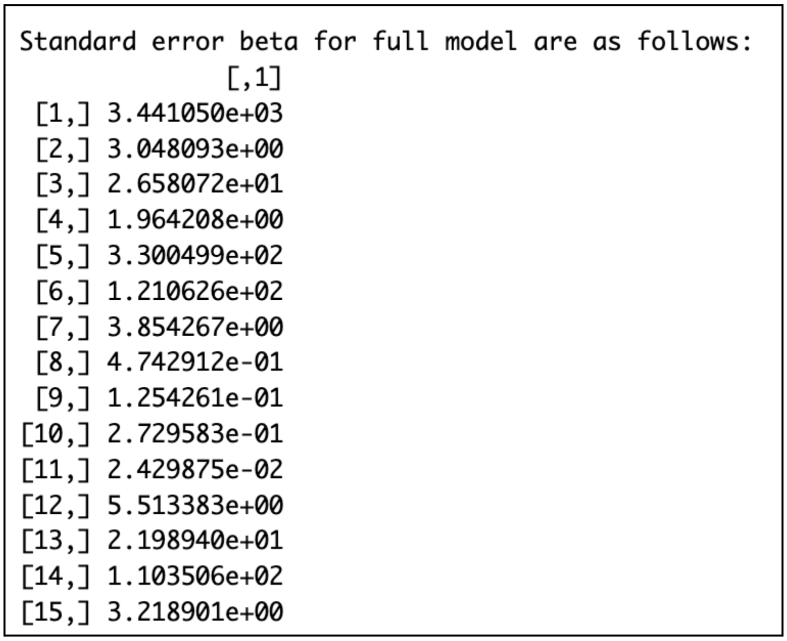


```
Beta Values for reduced model are as follows:
               [,1]
 [1,] 3998.59701701
 [2,]    1.56220499
 [3,]  -19.62459460
 [4,]    2.11970908
 [5,]  102.78122696
 [6,]  -67.53933514
 [7,]   -5.98141847
 [8,]   -1.18198791
 [9,]   -0.07180656
[10,]   -0.12286954
[11,]   -0.01079363
[12,]   -3.85945088
[13,]  -12.80358779
[14,]  322.87786798
[15,]   -0.75036386
```



```
Standard error beta for reduced model are as follows:
              [,1]
 [1,] 396.44581102
 [2,]   2.59122716
 [3,]  19.01058752
 [4,]   1.83610583
 [5,]  25.68328752
 [6,]  27.99010571
 [7,]   3.52320016
 [8,]   0.36655098
 [9,]   0.05691835
[10,]   0.04312104
[11,]   0.01216042
[12,]   3.61193920
[13,]  13.80221046
[14,]  78.22462859
[15,]   1.88686855
```

Following are the standard errors of beta for the full model.:



```
Standard error beta for full model are as follows:
              [,1]
 [1,] 3.441050e+03
 [2,] 3.048093e+00
 [3,] 2.658072e+01
 [4,] 1.964208e+00
 [5,] 3.300499e+02
 [6,] 1.210626e+02
 [7,] 3.854267e+00
 [8,] 4.742912e-01
 [9,] 1.254261e-01
[10,] 2.729583e-01
[11,] 2.429875e-02
[12,] 5.513383e+00
[13,] 2.198940e+01
[14,] 1.103506e+02
[15,] 3.218901e+00
```

VI) Results & Conclusions:

If we compare the standard error of beta for a reduced model with the full model then it can be concluded that values are decreased in the reduced model.

If Compare the standard error sum summation of {s.e. (βˆj) } of the reduced model with the full model then the following is the result:

Sum Standard error beta for the full model is: 4069.579
Sum Standard error beta for the reduced model is: 575.0847

From this it can be concluded that overall Standard error of beta values are significantly reduced using collinearity reduction through Principal component regression.

If Compare the standard error SSE of the reduced model with the full model then the following is the result:

SSE for the full model are as follows: 3276740
SSE for the reduced model are as follows: 3402208

From this it can be concluded that SSE is increased if we use collinearity reduction through Principal component regression.

To calculate the SSE of the reduced model matrix multiplication of the Predictor variable data matrix and obtained beta values from the reduced model were done to get residuals value. Then the square of residuals was done to get the SSE of the reduced model.

PART III

Problem 1:

Use the stepwise regression method to decide the best model. Use significance level αE = αR = 0.10. At each step, report the result of the regression, indicate which predictor variable enters or leaves the model, and how the decision is made. In the end, run the collinearity diagnostics again to verify that collinearity has disappeared.

I) Introduction:

Here we are Utilizing the stepwise regression method to determine the best model for the dataset. Specific significance level alpha is given for variable entry and removal.

II) Stepwise Regression Method:

The following are the steps of how the method works.

1. Forward Selection:

- In the initial step, we start with an empty model and evaluate each predictor's contribution using a t-test
- The predictor with the highest correlation with the response variable is added to the model.

2. Backward Elimination:

- After adding a predictor, the algorithm assesses the significance of all predictors in the model.
- If any predictor no longer meets the t-test, it gets eliminated from the model.

3. Iterative Process:

- we repeat the forward selection and backward elimination steps until no more variables meet the criteria for entry or removal.
- At each iteration, the model is updated, and the process continues until the optimal set of predictors is determined.

4.. Final Model:

- After following the above steps we get the final Model

III) Explanation of steps followed to get the final model and conclusion.

In the first step:

The forward step: pH entered the model, and stayed in the model due to statistical significance

Backward step: all variables in the model are statistically significant hence no variable was removed.

In the Second step:

The forward step: Zn entered the model, but left the model due to not being statistically significant

Backward step: all variables in the model are statistically significant hence no variable was removed.

In the Third step:

The forward step: Na entered the model, and stayed in the model due to statistical significance

Backward step: all variables in the model are statistically significant hence no variable was removed.

In the Fourth step:

The forward step: K entered the model, but left the model due to not being statistically significant

Backward step: all variables in the model are statistically significant hence no variable was removed.

In the Fifth step:

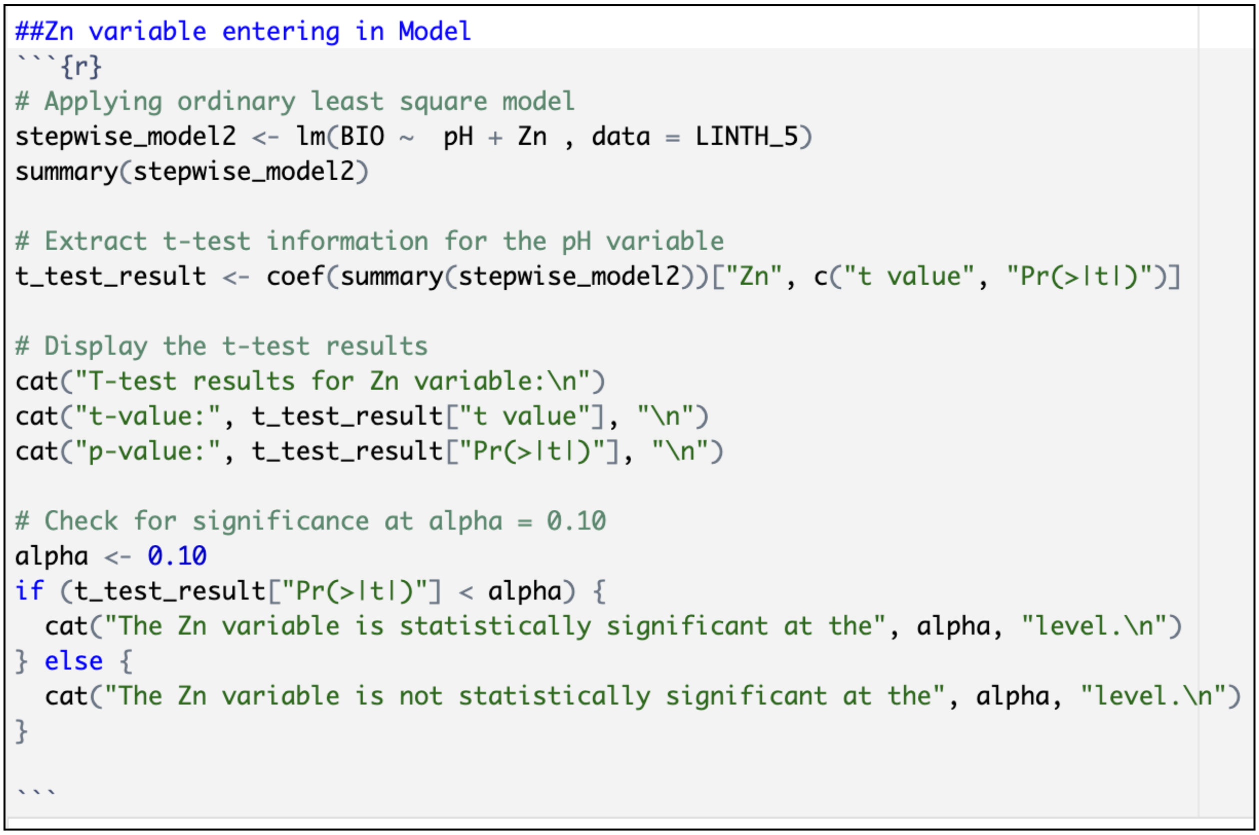
The forward step: SAL entered the model, but left the model due to not being statistically significant

Backward step: all variables in the model are statistically significant hence no variable was removed.

The final Model is achieved with variable pH and Na, without collinearity.

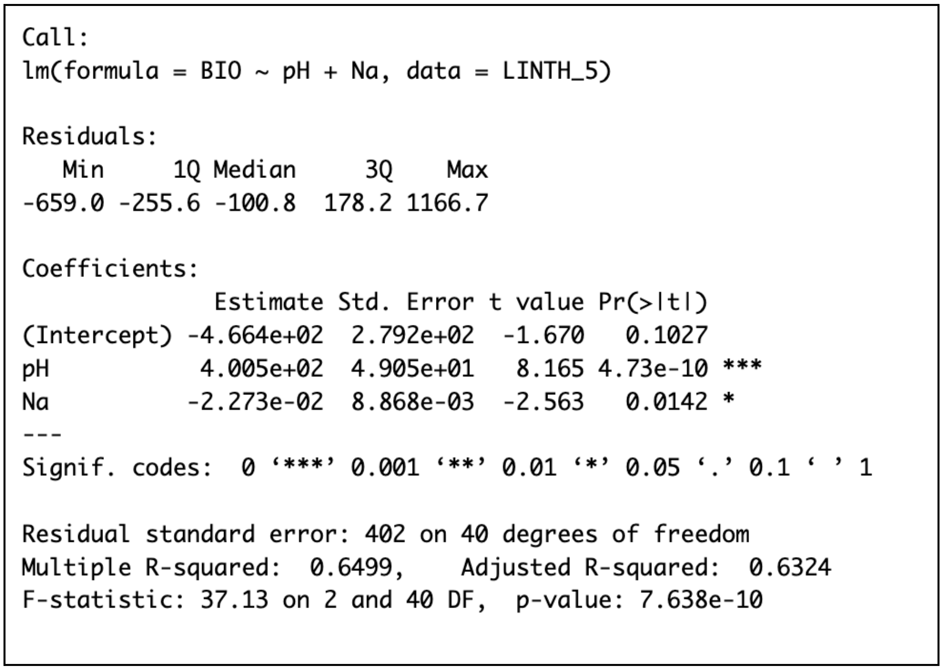
To keep the report simple adding just a few snapshots.

Snapshot of the Sample code how variables were entered model and significance were checked:



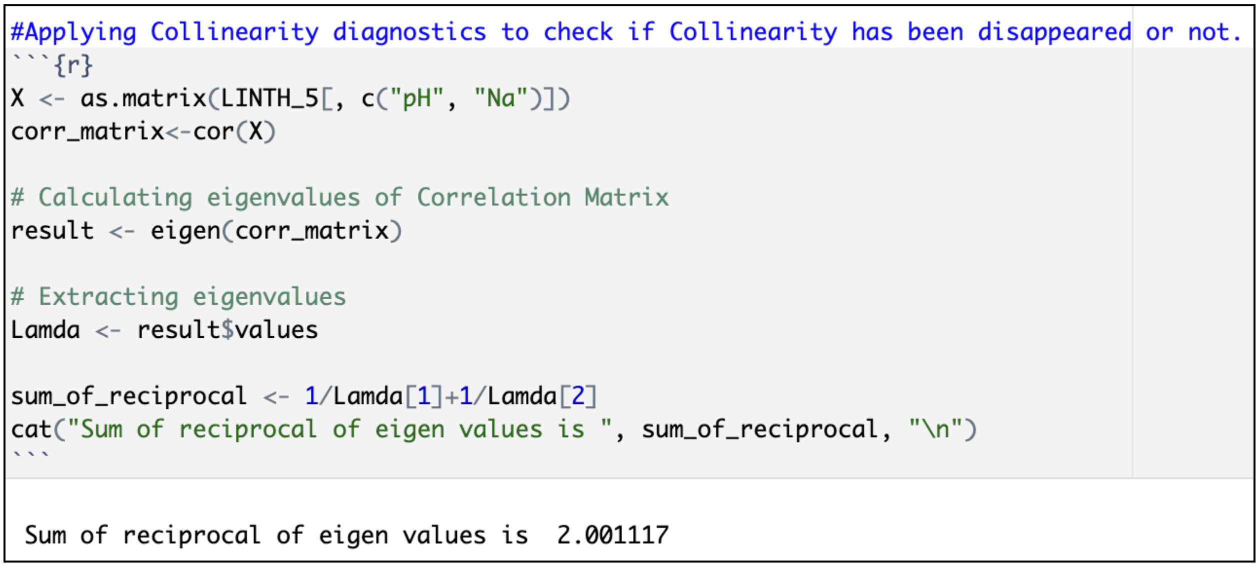
~~~
##Zn variable entering in Model
```{r}
# Applying ordinary least square model
stepwise_model2 <- lm(BIO ~  pH + Zn , data = LINTH_5)
summary(stepwise_model2)

# Extract t-test information for the pH variable
t_test_result <- coef(summary(stepwise_model2))["Zn", c("t value", "Pr(>|t|)")]

# Display the t-test results
cat("T-test results for Zn variable:\n")
cat("t-value:", t_test_result["t value"], "\n")
cat("p-value:", t_test_result["Pr(>|t|)"], "\n")

# Check for significance at alpha = 0.10
alpha <- 0.10
if (t_test_result["Pr(>|t|)"] < alpha) {
  cat("The Zn variable is statistically significant at the", alpha, "level.\n")
} else {
  cat("The Zn variable is not statistically significant at the", alpha, "level.\n")
}

```
~~~

Snapshot of final model:



```
Call:
lm(formula = BIO ~ pH + Na, data = LINTH_5)

Residuals:
   Min     1Q Median     3Q    Max 
-659.0 -255.6 -100.8  178.2 1166.7 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.664e+02  2.792e+02  -1.670   0.1027    
pH           4.005e+02  4.905e+01   8.165 4.73e-10 ***
Na          -2.273e-02  8.868e-03  -2.563   0.0142 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 402 on 40 degrees of freedom
Multiple R-squared:  0.6499,	Adjusted R-squared:  0.6324 
F-statistic: 37.13 on 2 and 40 DF,  p-value: 7.638e-10
```

Snapshot of Collinearity check



~~~
#Applying Collinearity diagnostics to check if Collinearity has been disappeared or not.
```{r}
X <- as.matrix(LINTH_5[, c("pH", "Na")])
corr_matrix<-cor(X)

# Calculating eigenvalues of Correlation Matrix
result <- eigen(corr_matrix)

# Extracting eigenvalues
Lamda <- result$values

sum_of_reciprocal <- 1/Lamda[1]+1/Lamda[2]
cat("Sum of reciprocal of eigen values is ", sum_of_reciprocal, "\n")
```

 Sum of reciprocal of eigen values is  2.001117
~~~

Conclusion for collinearity: Since the reciprocal of eigenvalue is less than 5 times the number of variables (i.e. 10). Hence it can be concluded that collinearity has disappeared

PART III

Problem 2:

Use ridge regression on the 5-predictor model, and use ridge trace to do variable selection. Refit the model that includes the remaining variables and then run the collinearity diagnostics again to verify that collinearity has disappeared.

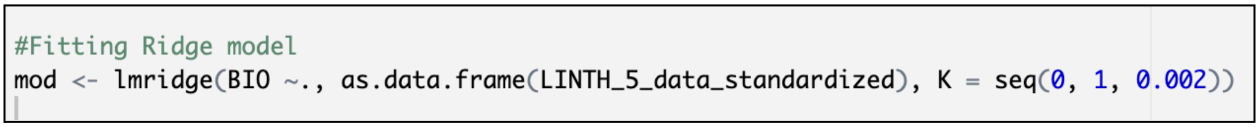
I) Introduction:

In this section, ridge regression is applied to a 5-predictor model, leveraging the ridge trace technique for variable selection. The ridge trace offers insights into the optimal regularization parameter, guiding the identification of important predictors. Following variable selection, the model is refitted, and collinearity diagnostics are conducted to ensure the mitigation of multicollinearity.

II) Ridge regression method:

Ridge regression is a regularization technique applied to linear regression, introducing a penalty term to address multicollinearity by shrinking coefficients. It provides a balance between model complexity and variance, offering improved stability in the presence of highly correlated predictors.

Hence first we fit the model in Ridge regression. Following is the snapshot code of the fitted model:



```
#Fitting Ridge model
mod <- lmridge(BIO ~., as.data.frame(LINTH_5_data_standardized), K = seq(0, 1, 0.002))
```

III)Ridge Trace and Variable Selection methodology:

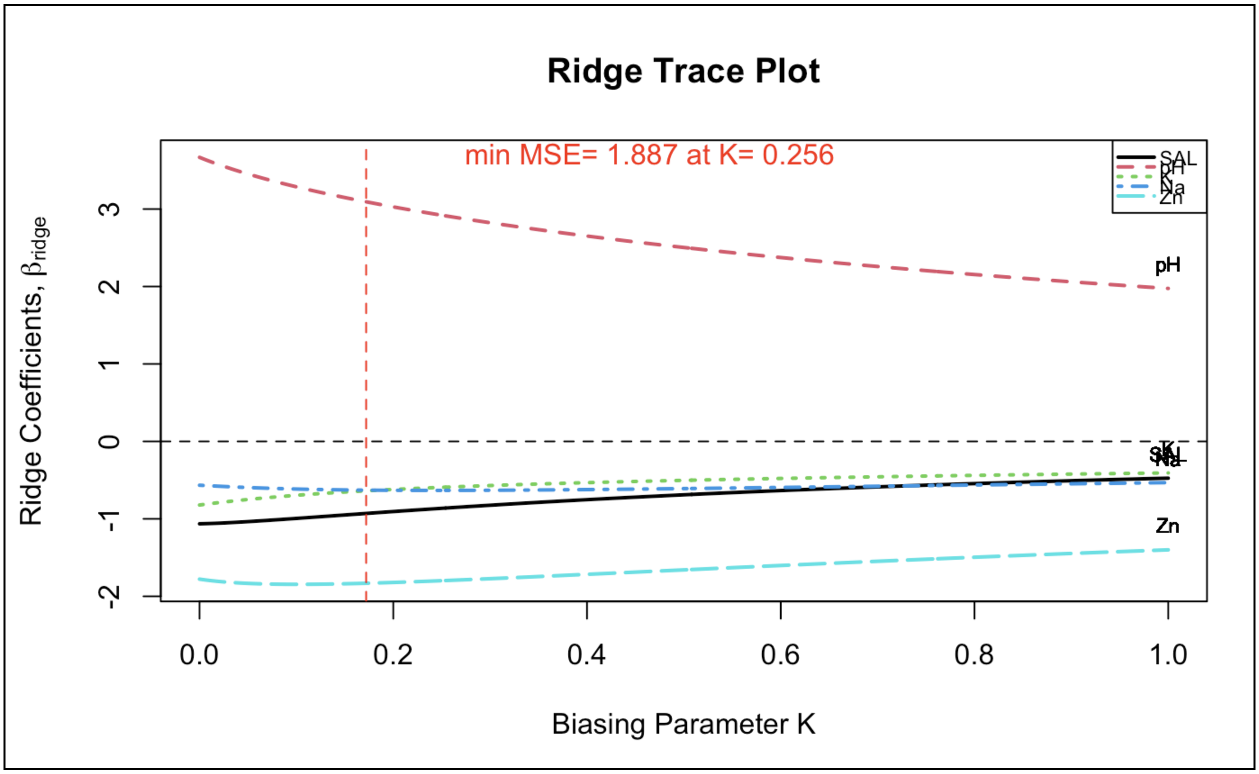
The ridge trace is used to eliminate variables from the equation. The guidelines for the elimination of variables as given in the book are as follows.

- Eliminate variables whose coefficients are stable but small. Since ridge regression is applied to standardized data, the magnitude of the various coefficients is directly comparable.
- Eliminate variables with unstable coefficients that do not hold their predicting power, that is, unstable coefficients that tend to zero.
- Eliminate one or more variables with unstable coefficients. The variables remaining from the original set, say p in number, are used to form the regression equation.

At the end of each of the above steps, we refit the model that includes the remaining variables before we proceed to the next.

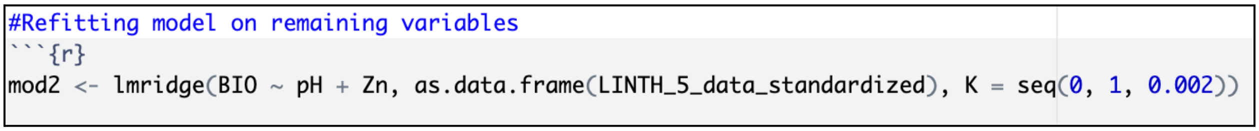
IV) Process followed to get results:

1. Ridge trace Graph for all variables:



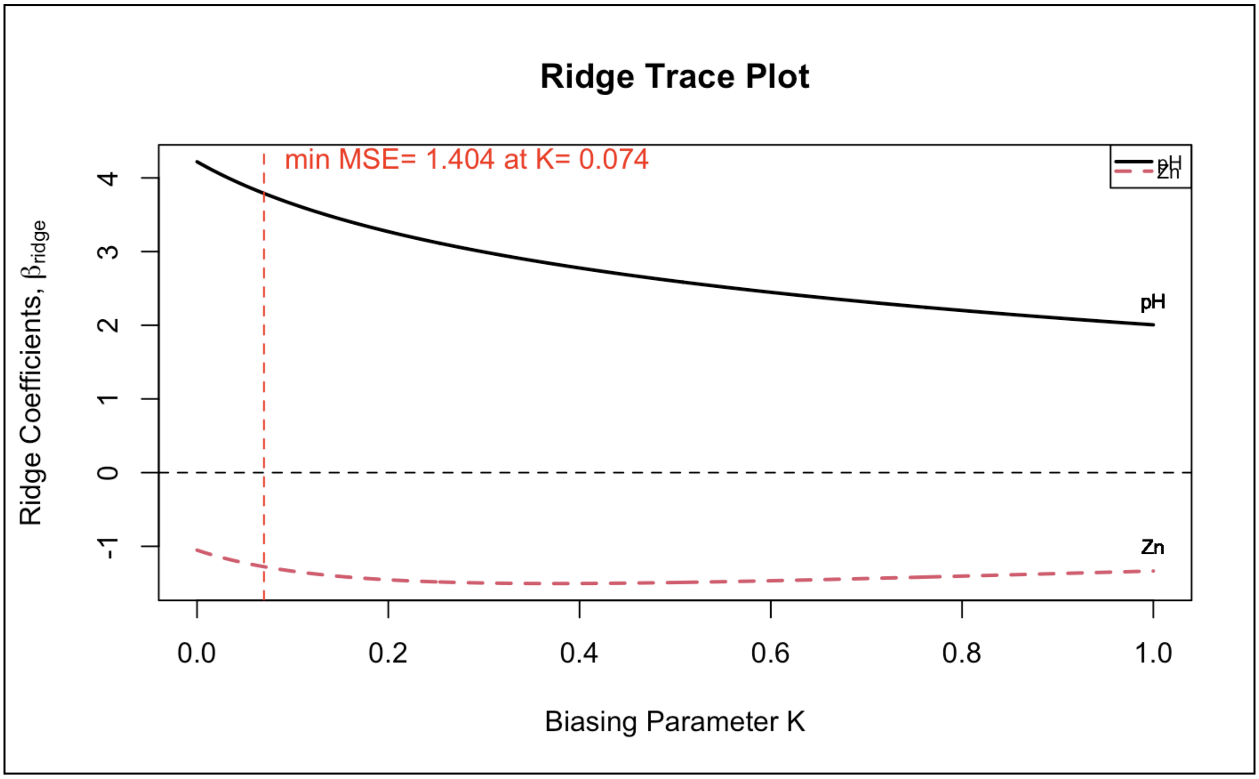
1st step: We can eliminate SAL, K, and Na variables because the coefficients of those are stable but small. Since ridge regression is applied to standardized data, the magnitude of the various coefficients is directly comparable.

2. Refitting model after first step for remaining variables:



~~~
#Refitting model on remaining variables
```{r}
mod2 <- lmridge(BIO ~ pH + Zn, as.data.frame(LINTH_5_data_standardized), K = seq(0, 1, 0.002))
~~~

3. Ridge trace Graph for refitted variable:

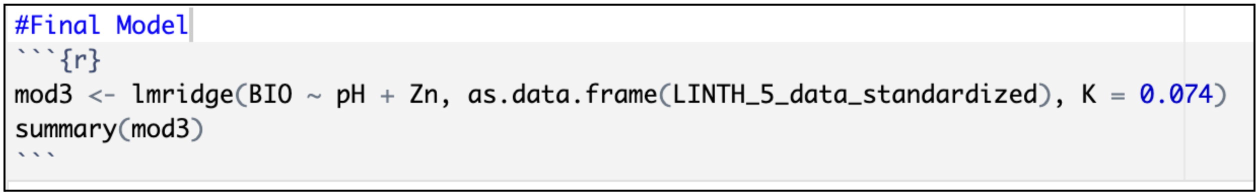


No variable can be removed from step 2 and step 3 as no conditions are being met to remove the variable.

IV) Final Model:

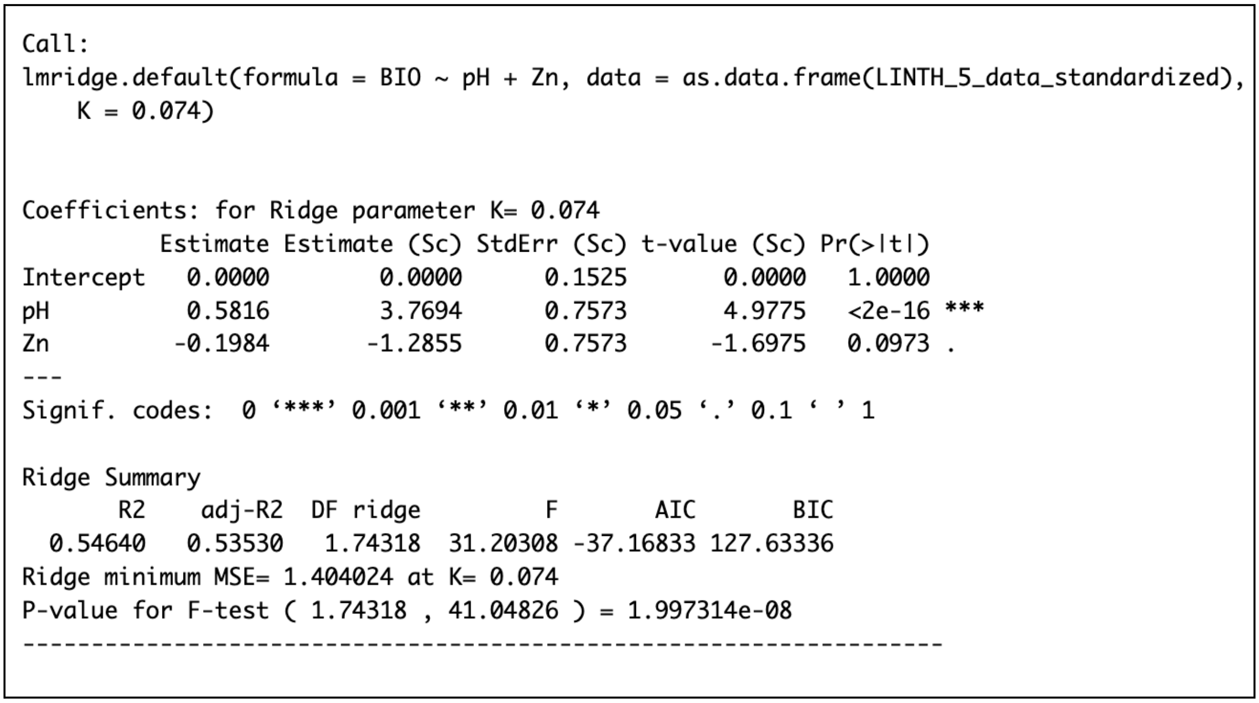
The final model with reduced variables obtained by following the above methodology and from the above graph, the optimal value of K=0.074 with min MSE hence our final model can be given as follows.

Snapshot of final model fitting in the ridge regression:



~~~
#Final Model
```{r}
mod3 <- lmridge(BIO ~ pH + Zn, as.data.frame(LINTH_5_data_standardized), K = 0.074)
summary(mod3)
```
~~~

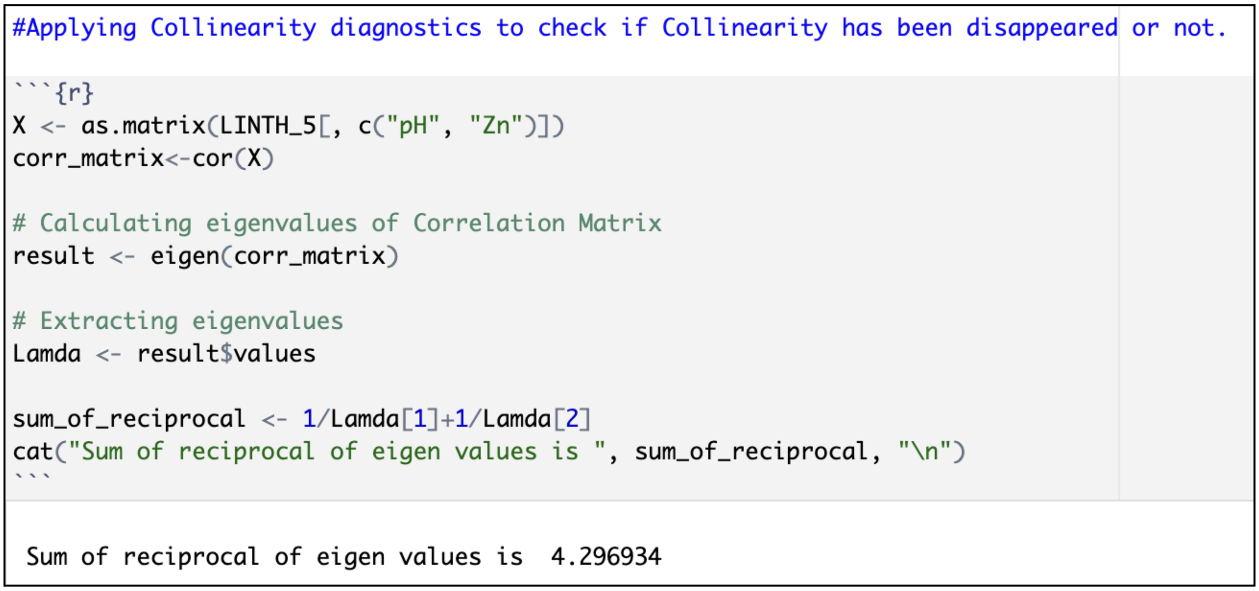
Summary of Final model:



```
Call:
lmridge.default(formula = BIO ~ pH + Zn, data = as.data.frame(LINTH_5_data_standardized),
    K = 0.074)


Coefficients: for Ridge parameter K= 0.074
          Estimate Estimate (Sc) StdErr (Sc) t-value (Sc) Pr(>|t|)
Intercept   0.0000        0.0000      0.1525       0.0000   1.0000
pH          0.5816        3.7694      0.7573       4.9775   <2e-16 ***
Zn         -0.1984       -1.2855      0.7573      -1.6975   0.0973 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Ridge Summary
       R2     adj-R2   DF ridge          F        AIC       BIC
  0.54640    0.53530    1.74318   31.20308 -37.16833 127.63336
Ridge minimum MSE= 1.404024 at K= 0.074
P-value for F-test ( 1.74318 , 41.04826 ) = 1.997314e-08
-------------------------------------------------------------------
```

V)Collinearity checks:

Snapshot of Collinearity check:



~~~
#Applying Collinearity diagnostics to check if Collinearity has been disappeared or not.

```{r}
X <- as.matrix(LINTH_5[, c("pH", "Zn")])
corr_matrix<-cor(X)

# Calculating eigenvalues of Correlation Matrix
result <- eigen(corr_matrix)

# Extracting eigenvalues
Lamda <- result$values

sum_of_reciprocal <- 1/Lamda[1]+1/Lamda[2]
cat("Sum of reciprocal of eigen values is ", sum_of_reciprocal, "\n")
```

 Sum of reciprocal of eigen values is  4.296934
~~~

Since the reciprocal of eigenvalue is less than 5 times the number of variables (i.e. 10). Hence it can be concluded that collinearity has disappeared.

PART III

Problem 3:

Use the subset selection method to decide the best two-variable model on the basis of BIC. If there is a tie, use VIF to break the tie.

I) Introduction:

Subset selection methods are used to choose the best subset of predictor variables for inclusion in a regression model.

II) Subset Selection method:

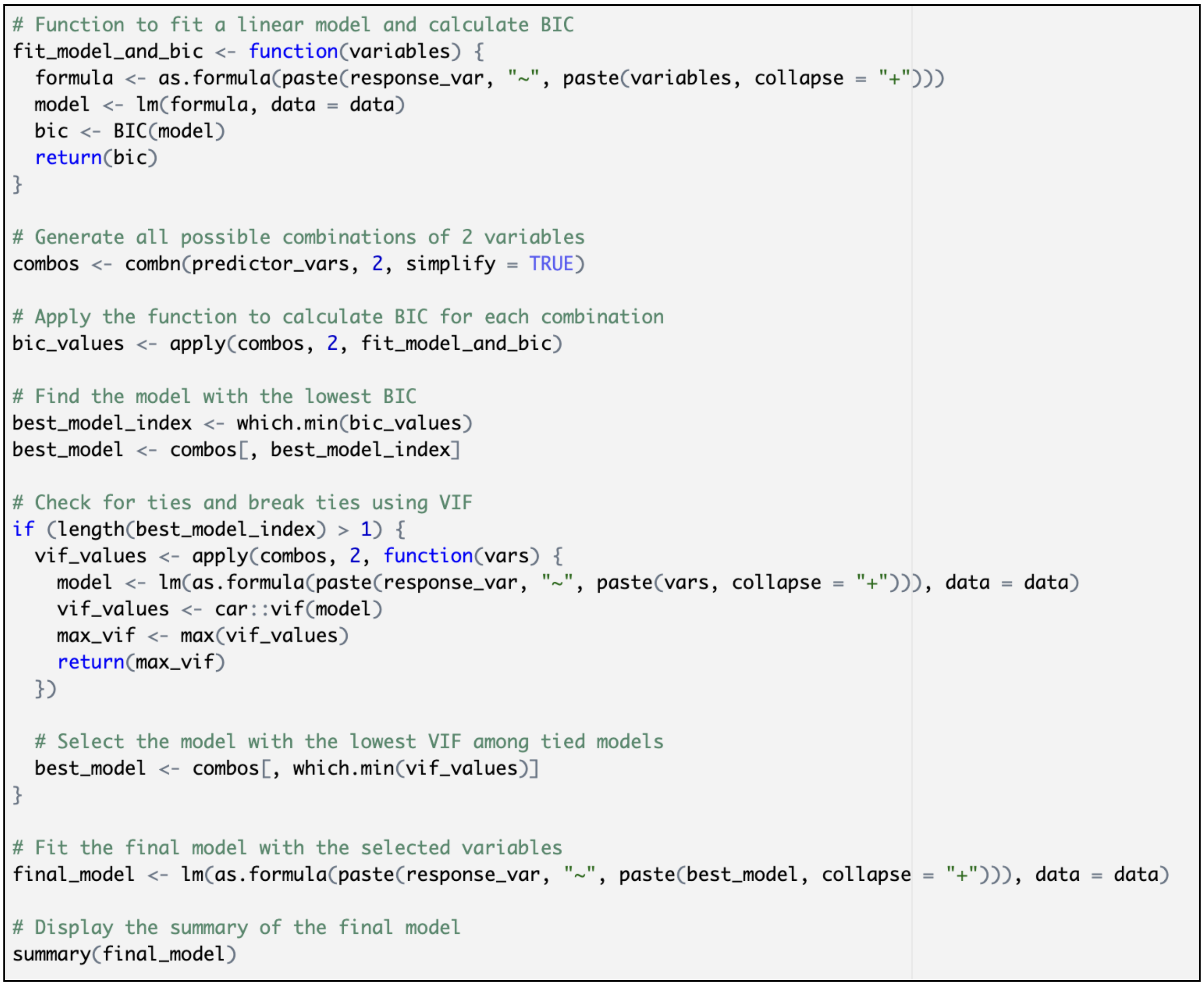
Subset variable selection methods involve systematically considering different subsets of predictors to identify the optimal set that yields the best model fit. These methods, such as forward selection, backward elimination, and stepwise regression, evaluate models based on criteria like AIC or BIC, progressively adding or removing predictors to enhance model interpretability and performance.

In our case, we are comparing the BIC of the model to get the best model.

III) Procedure to remove variable by subset selection method:

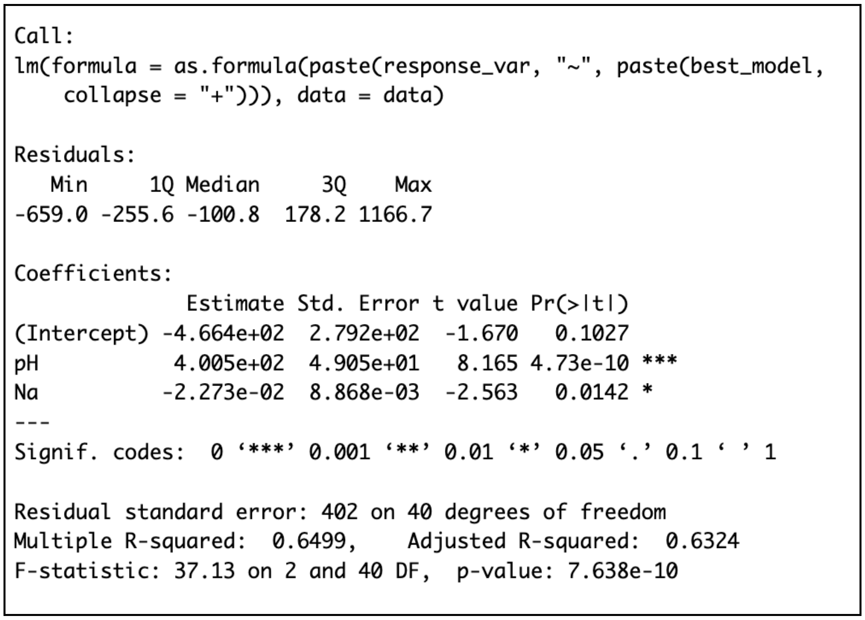
- Fit All Possible Models: Consider all possible combinations of two-variable models from your set of predictors.
- Calculate BIC: For each model, calculate the BIC. The BIC is a measure of the goodness of fit adjusted for the number of parameters in the model. Lower BIC values indicate better-fitting models.
- Identify Best Models: Identify the models with the lowest BIC values. If there is a tie (i.e., multiple models have the same low BIC), proceed to the next step.
- Use VIF to Break Ties: For the tied models, calculate the Variance Inflation Factor (VIF) for each predictor variable in each model. VIF measures the extent of multicollinearity in a regression analysis. If one of the tied models has lower average VIF values, it may be considered more suitable, as lower VIF values suggest less collinearity among the predictors.
- Select the Best Model: Based on the combined consideration of BIC and VIF, choose the model that strikes the best balance between model fit and multicollinearity.

IV) Developed code with the above procedure:



```r
# Function to fit a linear model and calculate BIC
fit_model_and_bic <- function(variables) {
  formula <- as.formula(paste(response_var, "~", paste(variables, collapse = "+")))
  model <- lm(formula, data = data)
  bic <- BIC(model)
  return(bic)
}

# Generate all possible combinations of 2 variables
combos <- combn(predictor_vars, 2, simplify = TRUE)

# Apply the function to calculate BIC for each combination
bic_values <- apply(combos, 2, fit_model_and_bic)

# Find the model with the lowest BIC
best_model_index <- which.min(bic_values)
best_model <- combos[, best_model_index]

# Check for ties and break ties using VIF
if (length(best_model_index) > 1) {
  vif_values <- apply(combos, 2, function(vars) {
    model <- lm(as.formula(paste(response_var, "~", paste(vars, collapse = "+"))), data = data)
    vif_values <- car::vif(model)
    max_vif <- max(vif_values)
    return(max_vif)
  })

  # Select the model with the lowest VIF among tied models
  best_model <- combos[, which.min(vif_values)]
}

# Fit the final model with the selected variables
final_model <- lm(as.formula(paste(response_var, "~", paste(best_model, collapse = "+"))), data = data)

# Display the summary of the final model
summary(final_model)
```

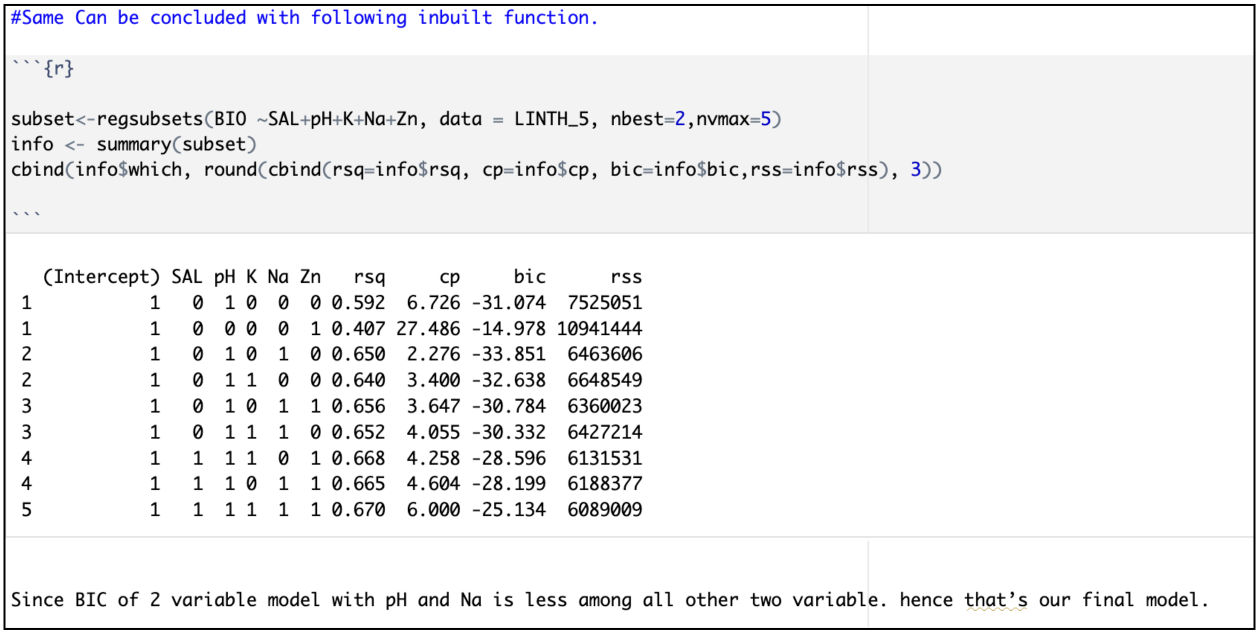
V) Final Model



```
Call:
lm(formula = as.formula(paste(response_var, "~", paste(best_model,
    collapse = "+"))), data = data)

Residuals:
   Min     1Q Median     3Q    Max
-659.0 -255.6 -100.8  178.2 1166.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.664e+02  2.792e+02  -1.670   0.1027
pH           4.005e+02  4.905e+01   8.165 4.73e-10 ***
Na          -2.273e-02  8.868e-03  -2.563   0.0142 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 402 on 40 degrees of freedom
Multiple R-squared:  0.6499,	Adjusted R-squared:  0.6324
F-statistic: 37.13 on 2 and 40 DF,  p-value: 7.638e-10
```

Hence Final Model is with variable pH and Na by following the subset selection method to decide the best two-variable model based on BIC. If there is a tie, use VIF to break the tie.

The same results were obtained with inbuilt function:



~~~
#Same Can be concluded with following inbuilt function.

```{r}

subset<-regsubsets(BIO ~SAL+pH+K+Na+Zn, data = LINTH_5, nbest=2,nvmax=5)
info <- summary(subset)
cbind(info$which, round(cbind(rsq=info$rsq, cp=info$cp, bic=info$bic,rss=info$rss), 3))

```
~~~

```
   (Intercept) SAL pH K Na Zn   rsq     cp     bic      rss
 1           1   0  1 0  0  0 0.592  6.726 -31.074  7525051
 1           1   0  0 0  0  1 0.407 27.486 -14.978 10941444
 2           1   0  1 0  1  0 0.650  2.276 -33.851  6463606
 2           1   0  1 1  0  0 0.640  3.400 -32.638  6648549
 3           1   0  1 0  1  1 0.656  3.647 -30.784  6360023
 3           1   0  1 1  1  0 0.652  4.055 -30.332  6427214
 4           1   1  1 1  0  1 0.668  4.258 -28.596  6131531
 4           1   1  1 0  1  1 0.665  4.604 -28.199  6188377
 5           1   1  1 1  1  1 0.670  6.000 -25.134  6089009
```

Since BIC of 2 variable model with pH and Na is less among all other two variable. hence that’s our final model.

References

- Book: “Regression Analysis By Example by Samprit Chatterjee Ali S. Hadi
- Class Notes
- Lectures