APPLIED STATISTICS

Project Report

By Sachin Janwalkar
A20479201

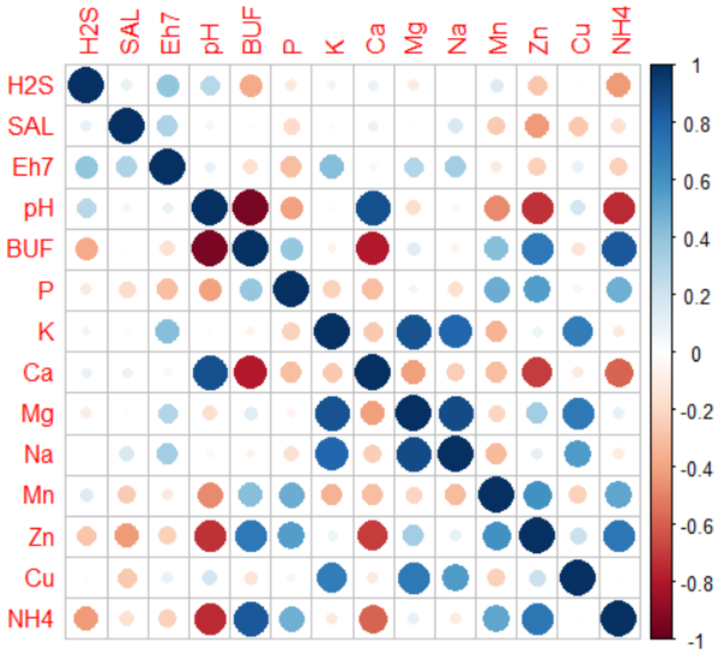
PART 1]

BIO ~ all predictor

Regression Co-efficient:

```
Residuals:
    Min      1Q  Median      3Q     Max 
-673.61 -148.68  -29.16  160.87 1012.18 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)  2.910e+03  3.413e+03   0.853  0.40062   
H2S          4.290e-01  2.998e+00   0.143  0.88717   
SAL         -2.398e+01  2.617e+01  -0.916  0.36678   
Eh7          2.553e+00  2.012e+00   1.269  0.21430   
pH           2.425e+02  3.342e+02   0.726  0.47361   
BUF         -6.902e+00  1.238e+02  -0.056  0.95592   
P           -1.702e+00  2.640e+00  -0.645  0.52409   
K           -1.047e+00  4.824e-01  -2.170  0.03808 * 
Ca          -1.161e-01  1.256e-01  -0.924  0.36293   
Mg          -2.802e-01  2.745e-01  -1.021  0.31540   
Na           4.451e-03  2.472e-02   0.180  0.85834   
Mn          -1.679e+00  5.373e+00  -0.312  0.75687   
Zn          -1.879e+01  2.178e+01  -0.863  0.39503   
Cu           3.452e+02  1.121e+02   3.080  0.00441 **
NH4         -2.705e+00  3.238e+00  -0.835  0.41007   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 350.8 on 30 degrees of freedom
Multiple R-squared:  0.8074,    Adjusted R-squared:  0.7175 
F-statistic: 8.983 on 14 and 30 DF,  p-value: 3.066e-07
```

Correlation Matrix:



As It can be seen from the correlation diagram features that have dark blue color relation are +ve correlated and feature that have red relation have –ve correlation.

Positive correlation: (PH, Ca), (NH4, BUF),(Mg, K),(Na, K),(Cu, K),(NH4, Zu)

Negative correlation: (BUF, pH),(Ca, BUF),(Zn, pH),(NH4, pH),(Zn, Ca)

Collinearity Diagnostic Test:

$$\sum_{j=1}^{p} \frac{1}{\lambda_j} \; . \; = 195.9633$$

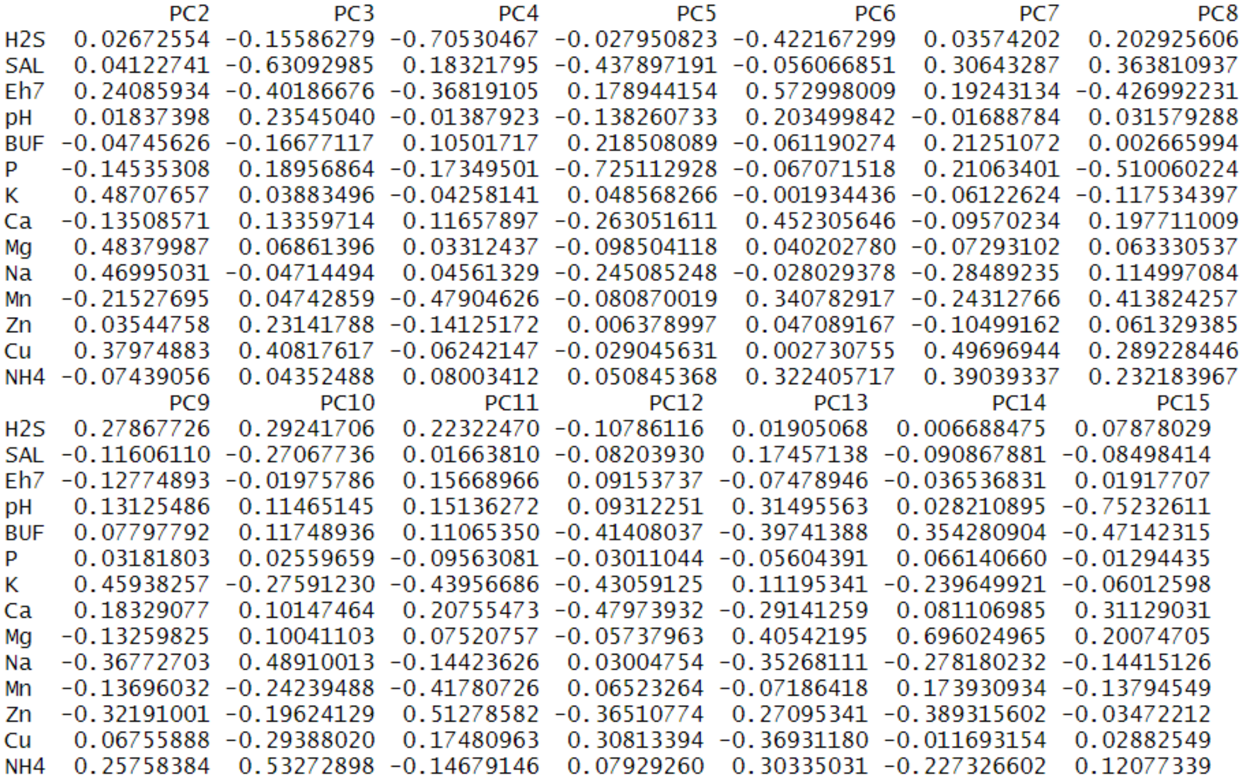
As the sum of reciprocals of all the Eigen values is 195.9633 > 5 x number of predictor, we can conclude there is collinearity.

Condition Number = 22.7752

Sets of Collinearity =1, Since only 1 kappa value is greater than 15.

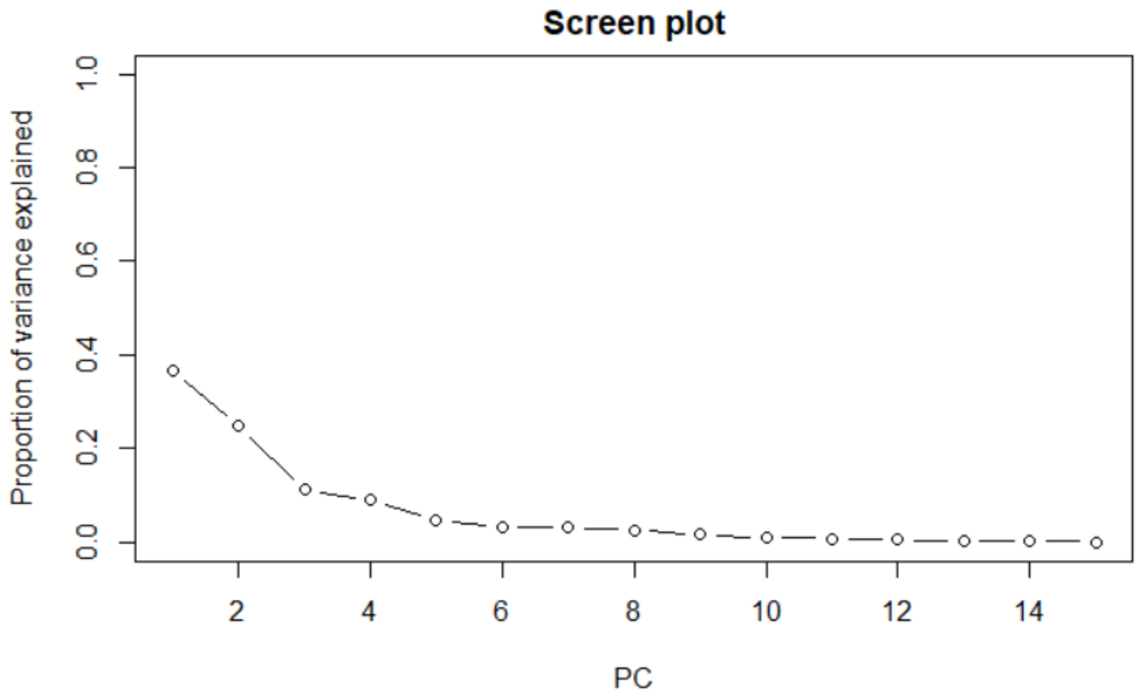
PART II]

Principal components:



```
            PC2         PC3         PC4          PC5          PC6         PC7          PC8
H2S  0.02672554 -0.15586279 -0.70530467 -0.027950823 -0.422167299  0.03574202  0.202925606
SAL  0.04122741 -0.63092985  0.18321795 -0.437897191 -0.056066851  0.30643287  0.363810937
Eh7  0.24085934 -0.40186676 -0.36819105  0.178944154  0.572998009  0.19243134 -0.426992231
pH   0.01837398  0.23545040 -0.01387923 -0.138260733  0.203499842 -0.01688784  0.031579288
BUF -0.04745626 -0.16677117  0.10501717  0.218508089 -0.061190274  0.21251072  0.002665994
P   -0.14535308  0.18956864 -0.17349501 -0.725112928 -0.067071518  0.21063401 -0.510060224
K    0.48707657  0.03883496 -0.04258141  0.048568266 -0.001934436 -0.06122624 -0.117534397
Ca  -0.13508571  0.13359714  0.11657897 -0.263051611  0.452305646 -0.09570234  0.197711009
Mg   0.48379987  0.06861396  0.03312437 -0.098504118  0.040202780 -0.07293102  0.063330537
Na   0.46995031 -0.04714494  0.04561329 -0.245085248 -0.028029378 -0.28489235  0.114997084
Mn  -0.21527695  0.04742859 -0.47904626 -0.080870019  0.340782917 -0.24312766  0.413824257
Zn   0.03544758  0.23141788 -0.14125172  0.006378997  0.047089167 -0.10499162  0.061329385
Cu   0.37974883  0.40817617 -0.06242147 -0.029045631  0.002730755  0.49696944  0.289228446
NH4 -0.07439056  0.04352488  0.08003412  0.050845368  0.322405717  0.39039337  0.232183967
            PC9        PC10        PC11        PC12        PC13         PC14        PC15
H2S  0.27867726  0.29241706  0.22322470 -0.10786116  0.01905068  0.006688475  0.07878029
SAL -0.11606110 -0.27067736  0.01663810 -0.08203930  0.17457138 -0.090867881 -0.08498414
Eh7 -0.12774893 -0.01975786  0.15668966  0.09153737 -0.07478946 -0.036536831  0.01917707
pH   0.13125486  0.11465145  0.15136272  0.09312251  0.31495563  0.028210895 -0.75232611
BUF  0.07797792  0.11748936  0.11065350 -0.41408037 -0.39741388  0.354280904 -0.47142315
P    0.03181803  0.02559659 -0.09563081 -0.03011044 -0.05604391  0.066140660 -0.01294435
K    0.45938257 -0.27591230 -0.43956686 -0.43059125  0.11195341 -0.239649921 -0.06012598
Ca   0.18329077  0.10147464  0.20755473 -0.47973932 -0.29141259  0.081106985  0.31129031
Mg  -0.13259825  0.10041103  0.07520757 -0.05737963  0.40542195  0.696024965  0.20074705
Na  -0.36772703  0.48910013 -0.14423626  0.03004754 -0.35268111 -0.278180232 -0.14415126
Mn  -0.13696032 -0.24239488 -0.41780726  0.06523264 -0.07186418  0.173930934 -0.13794549
Zn  -0.32191001 -0.19624129  0.51278582 -0.36510774  0.27095341 -0.389315602 -0.03472212
Cu   0.06755888 -0.29388020  0.17480963  0.30813394 -0.36931180 -0.011693154  0.02882549
NH4  0.25758384  0.53272898 -0.14679146  0.07929260  0.30335031 -0.227326602  0.12077339
```

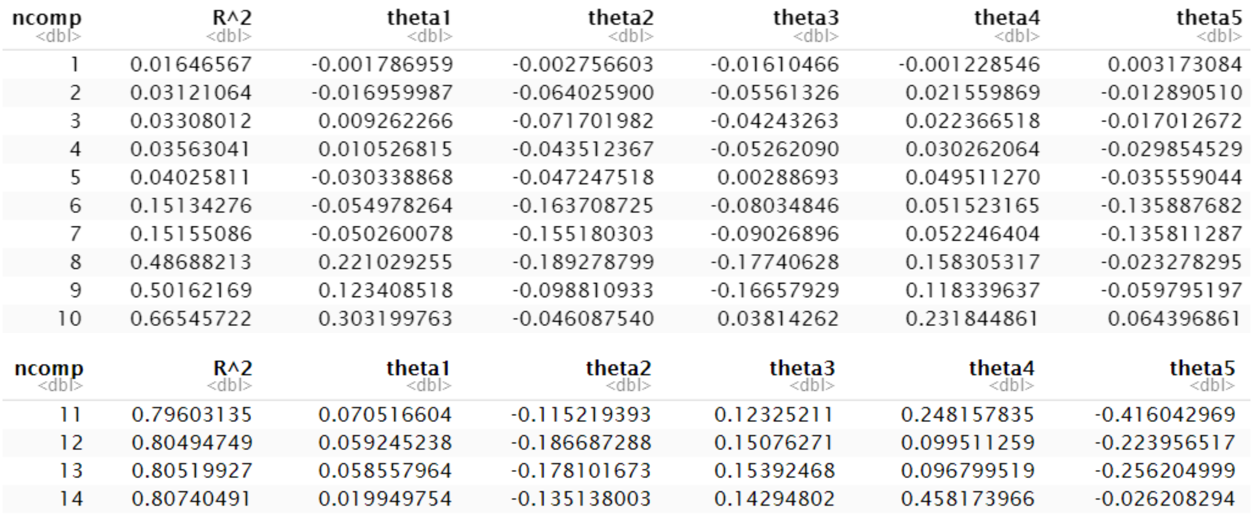
Screen Plot explaining variance explained by each PC.



we will only consider first 11 PC's as the proportion of variance explained from the 12th PC is very less.

Same calculation for considering number of PC is done with calculation of R^2.

Calculation of R^2 for theta:



ncomp	R^2	theta1	theta2	theta3	theta4	theta5
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0.01646567	-0.001786959	-0.002756603	-0.01610466	-0.001228546	0.003173084
2	0.03121064	-0.016959987	-0.064025900	-0.05561326	0.021559869	-0.012890510
3	0.03308012	0.009262266	-0.071701982	-0.04243263	0.022366518	-0.017012672
4	0.03563041	0.010526815	-0.043512367	-0.05262090	0.030262064	-0.029854529
5	0.04025811	-0.030338868	-0.047247518	0.00288693	0.049511270	-0.035559044
6	0.15134276	-0.054978264	-0.163708725	-0.08034846	0.051523165	-0.135887682
7	0.15155086	-0.050260078	-0.155180303	-0.09026896	0.052246404	-0.135811287
8	0.48688213	0.221029255	-0.189278799	-0.17740628	0.158305317	-0.023278295
9	0.50162169	0.123408518	-0.098810933	-0.16657929	0.118339637	-0.059795197
10	0.66545722	0.303199763	-0.046087540	0.03814262	0.231844861	0.064396861
11	0.79603135	0.070516604	-0.115219393	0.12325211	0.248157835	-0.416042969
12	0.80494749	0.059245238	-0.186687288	0.15076271	0.099511259	-0.223956517
13	0.80519927	0.058557964	-0.178101673	0.15392468	0.096799519	-0.256204999
14	0.80740491	0.019949754	-0.135138003	0.14294802	0.458173966	-0.026208294

We will consider first 11 PC as after that there is barely any change/ increase in R^2 value by adding another Principal component.

Computed Regression co-efficient from result of PC regression:

Calculate alpha: we need to get PC's coefficients i.e. alpha by regression Y ~ PC1+…..PC14

Calculate theta: multiplying the eigen vectors with alphas

Calculate betas : theta * (Sy/Sj)

Where Sy = standard deviation of Response variable

Sj = standard deviation of all the predictor variables where j=1,2….p

For Beta_0 Intercept co-effcicient

Beta_0 = 2909.934

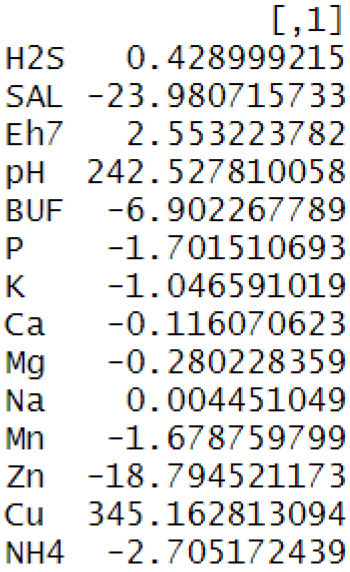
Beta_0 = y_bar – summation(betas * x_bar)

Where y_bar is mean of Y

Beats are the coefficients that we got from above

X_bar is the mean of each predictor variable

Beta Coefficients obtained:



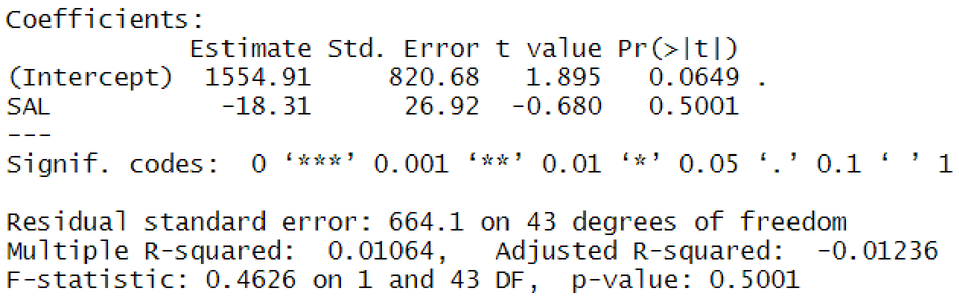
```
            [,1]
H2S    0.428999215
SAL  -23.980715733
Eh7    2.553223782
pH   242.527810058
BUF   -6.902267789
P     -1.701510693
K     -1.046591019
Ca    -0.116070623
Mg    -0.280228359
Na     0.004451049
Mn    -1.678759799
Zn   -18.794521173
Cu   345.162813094
NH4   -2.705172439
```

PART III]

Stepwise Regression

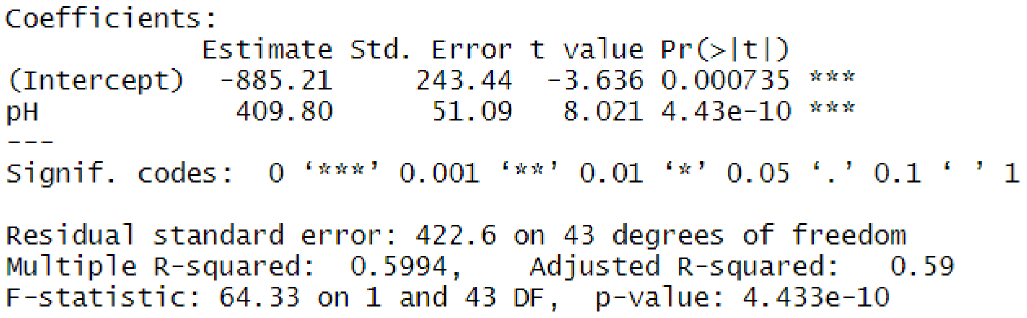
Model with 1 predictor:

BIO ~SAL



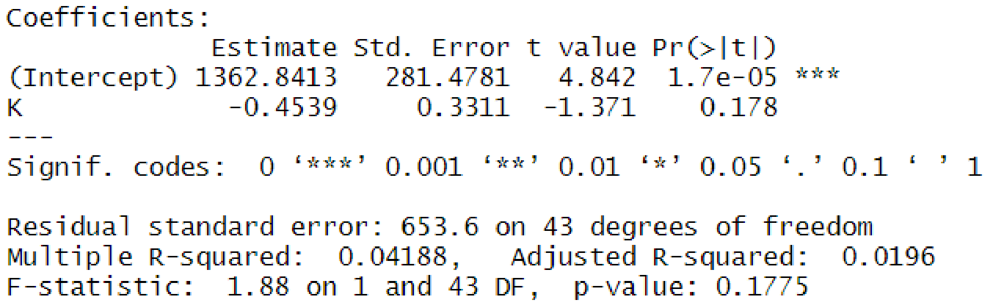
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept)  1554.91     820.68   1.895   0.0649 .
SAL           -18.31      26.92  -0.680   0.5001  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 664.1 on 43 degrees of freedom
Multiple R-squared:  0.01064,	Adjusted R-squared:  -0.01236 
F-statistic: 0.4626 on 1 and 43 DF,  p-value: 0.5001
```

BIO ~ pH



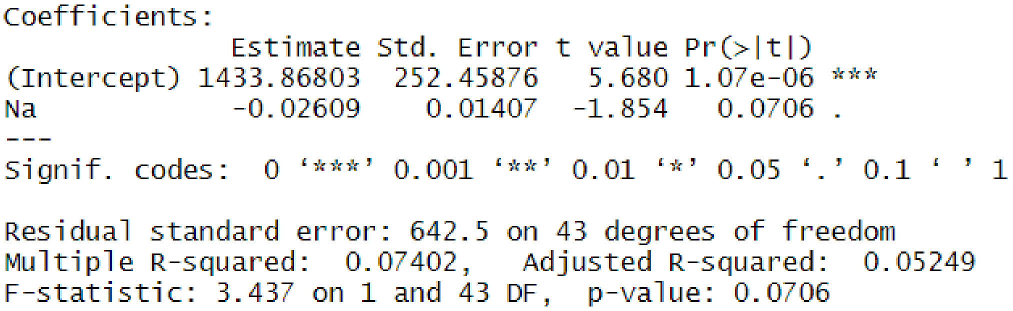
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -885.21     243.44  -3.636 0.000735 ***
pH            409.80      51.09   8.021 4.43e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 422.6 on 43 degrees of freedom
Multiple R-squared:  0.5994,	Adjusted R-squared:   0.59 
F-statistic: 64.33 on 1 and 43 DF,  p-value: 4.433e-10
```

BIO ~ K



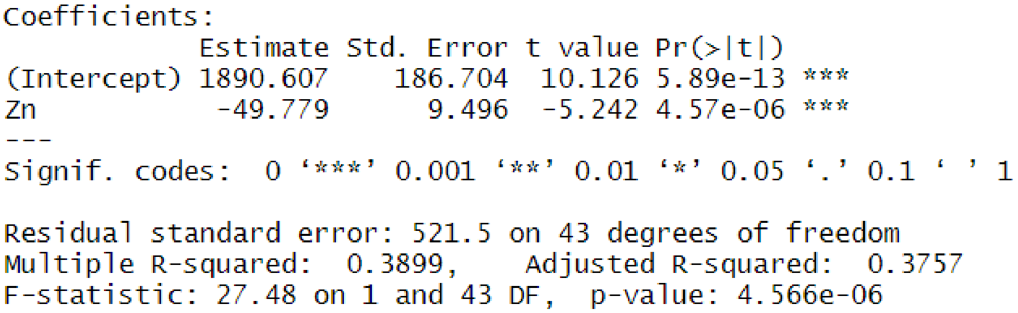
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1362.8413   281.4781   4.842  1.7e-05 ***
K             -0.4539     0.3311  -1.371    0.178    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 653.6 on 43 degrees of freedom
Multiple R-squared:  0.04188,	Adjusted R-squared:  0.0196 
F-statistic:  1.88 on 1 and 43 DF,  p-value: 0.1775
```

BIO ~ Na



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1433.86803  252.45876   5.680 1.07e-06 ***
Na            -0.02609    0.01407  -1.854   0.0706 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 642.5 on 43 degrees of freedom
Multiple R-squared:  0.07402,	Adjusted R-squared:  0.05249 
F-statistic: 3.437 on 1 and 43 DF,  p-value: 0.0706
```

BIO ~ Zn



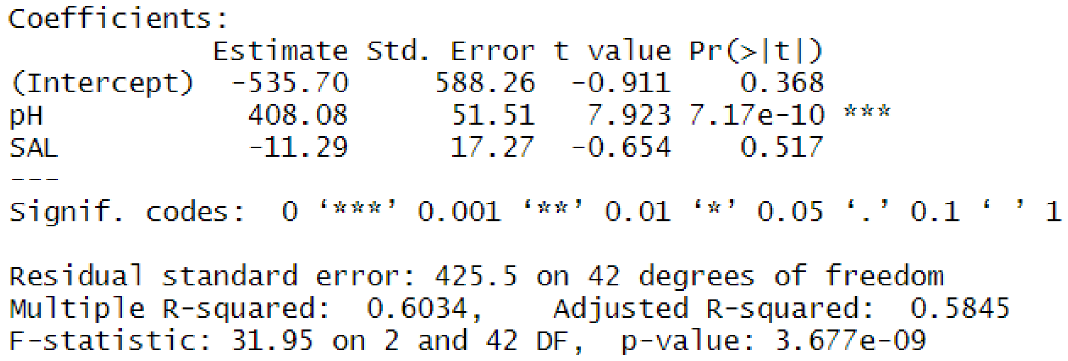
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1890.607    186.704  10.126 5.89e-13 ***
Zn           -49.779      9.496  -5.242 4.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 521.5 on 43 degrees of freedom
Multiple R-squared:  0.3899,	Adjusted R-squared:  0.3757 
F-statistic: 27.48 on 1 and 43 DF,  p-value: 4.566e-06
```

From all the models with 1 Predictor the model BIO ~ PH has the lowest p-value, means it’s the most statistically significant, Also it’s R^2 value is highest among all the significant models with 1 predictors. So we will include PH into our model.

Model with 2 predictors:

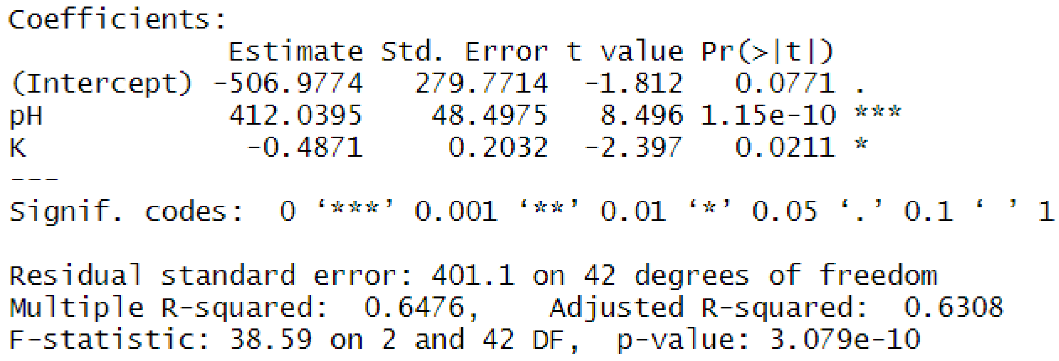
BIO ~ PH + 2ND Feature

BIO ~ pH + SAL



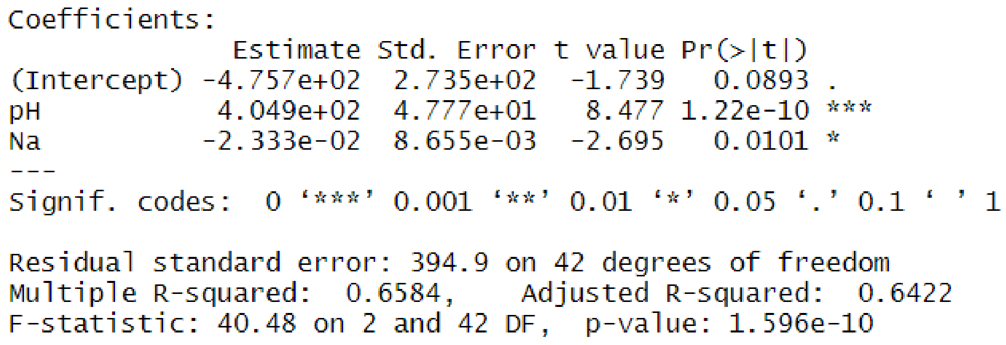
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -535.70     588.26  -0.911    0.368    
pH            408.08      51.51   7.923 7.17e-10 ***
SAL           -11.29      17.27  -0.654    0.517    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 425.5 on 42 degrees of freedom
Multiple R-squared:  0.6034,	Adjusted R-squared:  0.5845 
F-statistic: 31.95 on 2 and 42 DF,  p-value: 3.677e-09
```

BIO ~ pH + K



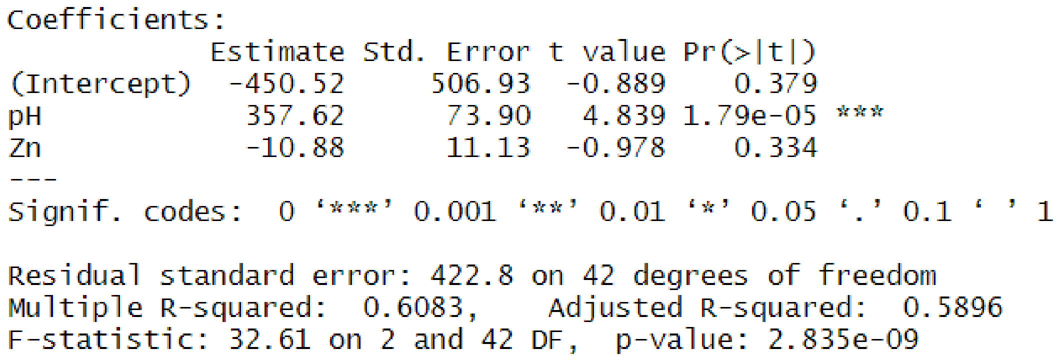
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -506.9774   279.7714  -1.812   0.0771 .  
pH           412.0395    48.4975   8.496 1.15e-10 ***
K             -0.4871     0.2032  -2.397   0.0211 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 401.1 on 42 degrees of freedom
Multiple R-squared:  0.6476,	Adjusted R-squared:  0.6308 
F-statistic: 38.59 on 2 and 42 DF,  p-value: 3.079e-10
```

BIO ~ Ph + Na



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.757e+02  2.735e+02  -1.739   0.0893 .  
pH           4.049e+02  4.777e+01   8.477 1.22e-10 ***
Na          -2.333e-02  8.655e-03  -2.695   0.0101 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 394.9 on 42 degrees of freedom
Multiple R-squared:  0.6584,	Adjusted R-squared:  0.6422 
F-statistic: 40.48 on 2 and 42 DF,  p-value: 1.596e-10
```

BIO ~ Ph + Zn



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -450.52     506.93  -0.889    0.379    
pH            357.62      73.90   4.839 1.79e-05 ***
Zn            -10.88      11.13  -0.978    0.334    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 422.8 on 42 degrees of freedom
Multiple R-squared:  0.6083,	Adjusted R-squared:  0.5896 
F-statistic: 32.61 on 2 and 42 DF,  p-value: 2.835e-09
```

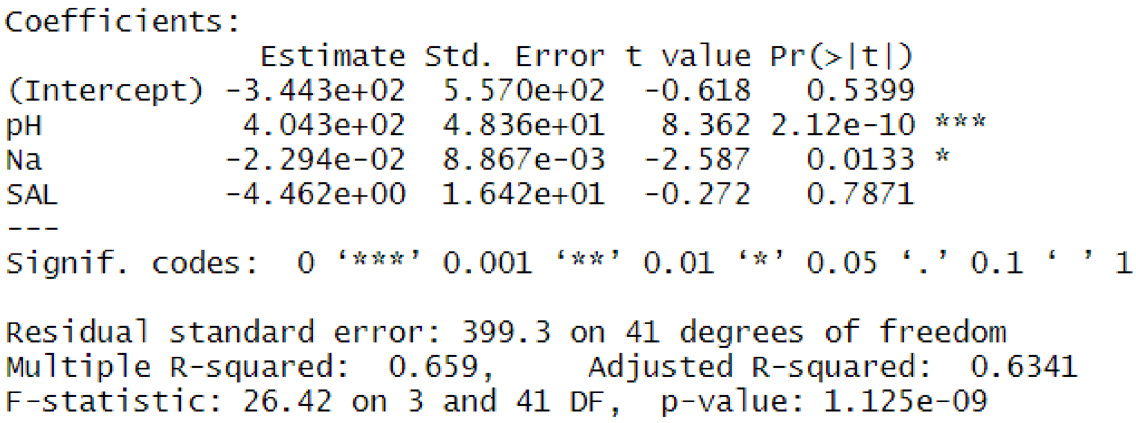
The model with PH + Na and PH+ K are both statistically significant, but the p-value of predictor Na is the smallest and it has the largest R^2 (means this model can capture the variation in the i/p data better). So we will consider model with 2nd feature as Na

BIO ~ PH + Na

Model 3 predictor:

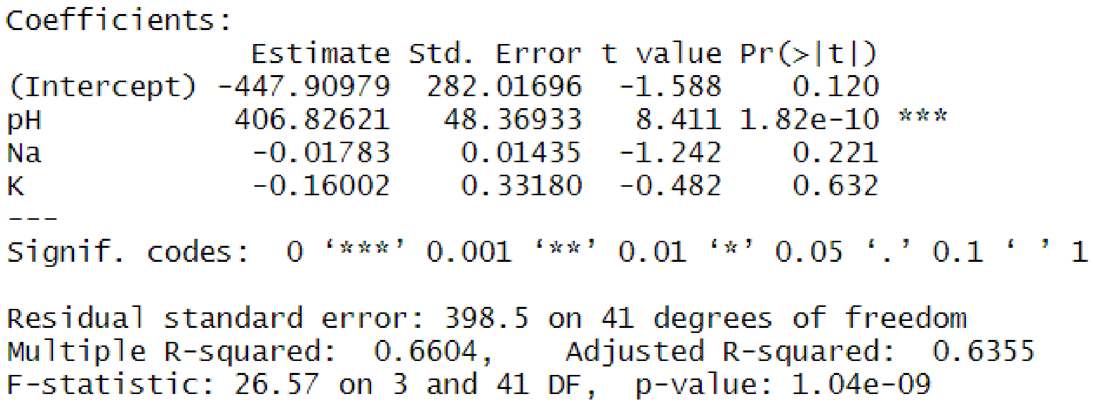
BIO ~ PH + Na + 3rd feature

BIO ~ PH + Na +SAL



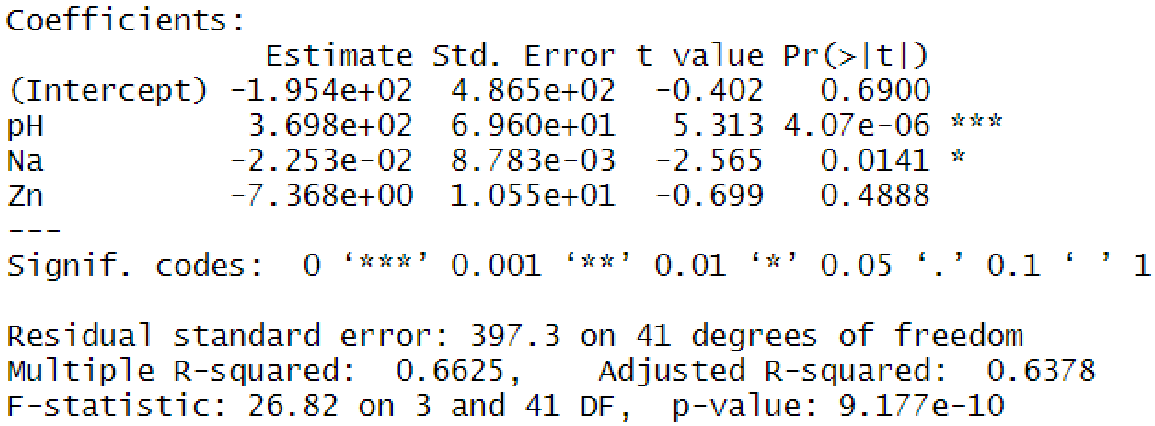
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.443e+02  5.570e+02  -0.618   0.5399    
pH           4.043e+02  4.836e+01   8.362 2.12e-10 ***
Na          -2.294e-02  8.867e-03  -2.587   0.0133 *  
SAL         -4.462e+00  1.642e+01  -0.272   0.7871    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 399.3 on 41 degrees of freedom
Multiple R-squared:  0.659,	Adjusted R-squared:  0.6341 
F-statistic: 26.42 on 3 and 41 DF,  p-value: 1.125e-09
```

BIO ~ PH + Na + K



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -447.90979  282.01696  -1.588    0.120    
pH           406.82621   48.36933   8.411 1.82e-10 ***
Na            -0.01783    0.01435  -1.242    0.221    
K             -0.16002    0.33180  -0.482    0.632    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 398.5 on 41 degrees of freedom
Multiple R-squared:  0.6604,	Adjusted R-squared:  0.6355 
F-statistic: 26.57 on 3 and 41 DF,  p-value: 1.04e-09
```

BIO ~ PH + Na + Zn

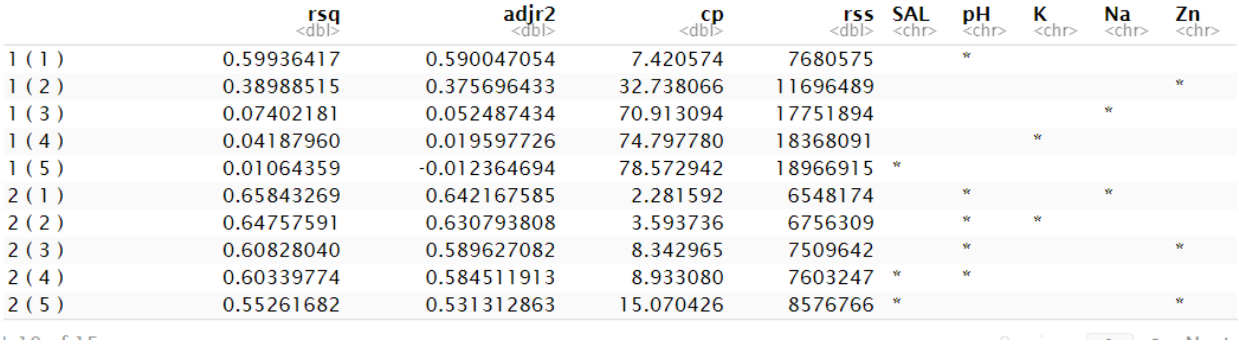


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.954e+02  4.865e+02  -0.402   0.6900    
pH           3.698e+02  6.960e+01   5.313 4.07e-06 ***
Na          -2.253e-02  8.783e-03  -2.565   0.0141 *  
Zn          -7.368e+00  1.055e+01  -0.699   0.4888    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 397.3 on 41 degrees of freedom
Multiple R-squared:  0.6625,	Adjusted R-squared:  0.6378 
F-statistic: 26.82 on 3 and 41 DF,  p-value: 9.177e-10
```

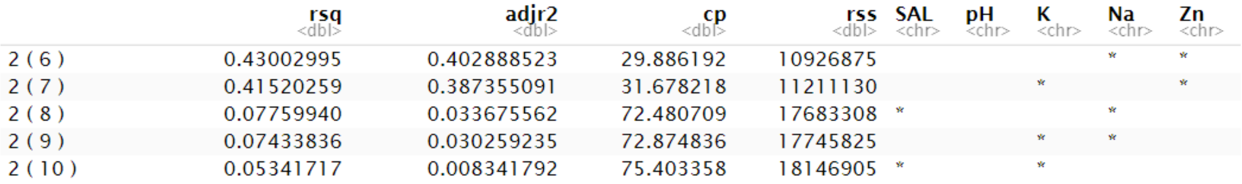
In all the models with 3 features the 3^{rd} feature is not statistically significant, Also the R^2 of the model doesn’t increases by adding 3^{rd} predictor. Hence we will stop the stepwise regression with addition of 2 predictors as our final model

Final model: BIO ~ pH + Na

Subset Selection:



	rsq	adjr2	cp	rss	SAL	pH	K	Na	Zn
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>
1 (1)	0.59936417	0.590047054	7.420574	7680575		*			
1 (2)	0.38988515	0.375696433	32.738066	11696489					*
1 (3)	0.07402181	0.052487434	70.913094	17751894				*	
1 (4)	0.04187960	0.019597726	74.797780	18368091			*		
1 (5)	0.01064359	-0.012364694	78.572942	18966915	*				
2 (1)	0.65843269	0.642167585	2.281592	6548174		*		*	
2 (2)	0.64757591	0.630793808	3.593736	6756309		*	*		
2 (3)	0.60828040	0.589627082	8.342965	7509642		*			*
2 (4)	0.60339774	0.584511913	8.933080	7603247	*	*			
2 (5)	0.55261682	0.531312863	15.070426	8576766	*				*



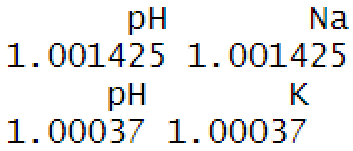
	rsq	adjr2	cp	rss	SAL	pH	K	Na	Zn
	<dbl>	<dbl>	<dbl>	<dbl>	<chr>	<chr>	<chr>	<chr>	<chr>
2 (6)	0.43002995	0.402888523	29.886192	10926875				*	*
2 (7)	0.41520259	0.387355091	31.678218	11211130			*		*
2 (8)	0.07759940	0.033675562	72.480709	17683308	*			*	
2 (9)	0.07433836	0.030259235	72.874836	17745825			*	*	
2 (10)	0.05341717	0.008341792	75.403358	18146905	*		*		

By looking at the summary table Adj_R^2 and Cp values, we can narrow it down to 2 models

BIO~ pH + Na   & BIO ~ pH + K as their Cp values are close to the the number of features+1 i.e. (p+1) and they have highest Adj_R^2

To break the tie, we use VIF

VIF:



```
      pH       Na 
1.001425 1.001425 
     pH       K 
1.00037 1.00037 
```

Since VIF values of features pH and Na are higher that pH+K, higher VIF values indicate collinearity also Cp value of model with pH+ K is equal to 3 that equals p+1 but on the other hand Cp value of pH and Na is closer to 2.

Hence we can proceed with model with 2 features pH +k as there is barely any difference in the R^2 value.

Final model: BIO~ PH + k

The model selected in Stepwise and best subset regression are different as in stepwise there is a possibility that we may not get the most optimum model as it does not verify every possibility of subset as subset selection method does. Also the order in which the features are added also matters in computing the final result in step wise regression.