A part

In this part, the 14-predictor data set is considered. Response variable is BIO and there are 14 predictor variables characterizing the soil. The Linthurst data is chose to identify the important physicochemical properties of the substrate influencing the aerial biomass production in the Cape Fear Estuary of North Carolina. Based on these, regression coefficients are obtained by ordinary least square estimation. The result is listed bellow.

Call:				
lm(formula = BIO ~ H2S + SAL + Eh7 + pH + BUF + P + K + Ca + Mg + Na + Mg + Zn + Cu + NH4, data = data) Mg + Na + Mg + Zn + Cu + NH4, data = data)				
Residuals:				
Min	1Q	Median	3Q	Max
-660.16	-147.2	-46.07	147.24	1020.86
Coefficients:				
Estimate	Std.	Error	t value	Pr(>|t|)
(Intercept)	2.18E+03	2.46E+03	0.887	0.38172
H2S	-1.20E-01	2.39E+00	-0.05	0.96025
SAL	-2.23E+01	2.53E+01	-0.884	0.38334
Eh7	2.41E+00	1.93E+00	1.248	0.22129
pH	2.98E+02	2.80E+02	1.063	0.29602
BUF	1.53E+01	9.98E+01	0.154	0.87894
P	-1.62E+00	2.59E+00	-0.624	0.53693
K	-1.04E+00	4.75E-01	-2.187	0.03639*
Ca	-1.34E-01	1.10E-01	-1.218	0.23238
Mg	-2.80E-01	2.70E-01	-1.036	0.30835
Na	5.67E-03	2.41E-02	0.236	0.81509
Zn	-2.18E+01	1.93E+01	-1.131	0.26656
Cu	3.48E+02	1.10E+02	3.165	0.00347**
NH4	-3.24E+00	2.70E+00	-1.2	0.23905
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 345.7 on 31 degrees of freedom				
Multiple R-squared: 0.8068, Adjusted R-squared: 0.7257				
F-statistic: 9.957 on 13 and 31 DF, p-value: 9.781e-08				

It is found that When the confidence level is 0.95, only the P values of K and Cu are less than 0.05, which pass the t test, and none of the other variables pass the test. The R-squared value of the entire model is 0.8068, and the p-value is 9.781e-08 less than 0.05, indicating that the entire model has passed the test.

Run the collinearity diagnostics and identify if there is any collinearity.

H2S	SAL	Eh7	pH	BUF	P	K
1.986493	3.250726	1.877174	44.859565	23.057453	1.875011	7.342187
Ca	Mg	Na	Zn	Cu	NH4	
13.167269	23.76416	10.091775	9.367063	4.793347	6.006372	

In the variance expansion factor method, the values of the variables PH, BUF, Ca, Mg, and Na are all greater than 10, indicating that there is serious collinearity between the independent variables, and this collinearity will defect the results of the data.

In the characteristic root discriminant method, the condition number k describes the extent of the characteristic roots. The obtained condition number is 552.2811. According to the discriminant principle, when the condition number k>100, there is very strong multicollinearity between variables.

In conclusion, both variance expansion factor method and characteristic root discriminant method show that there is collinearity.

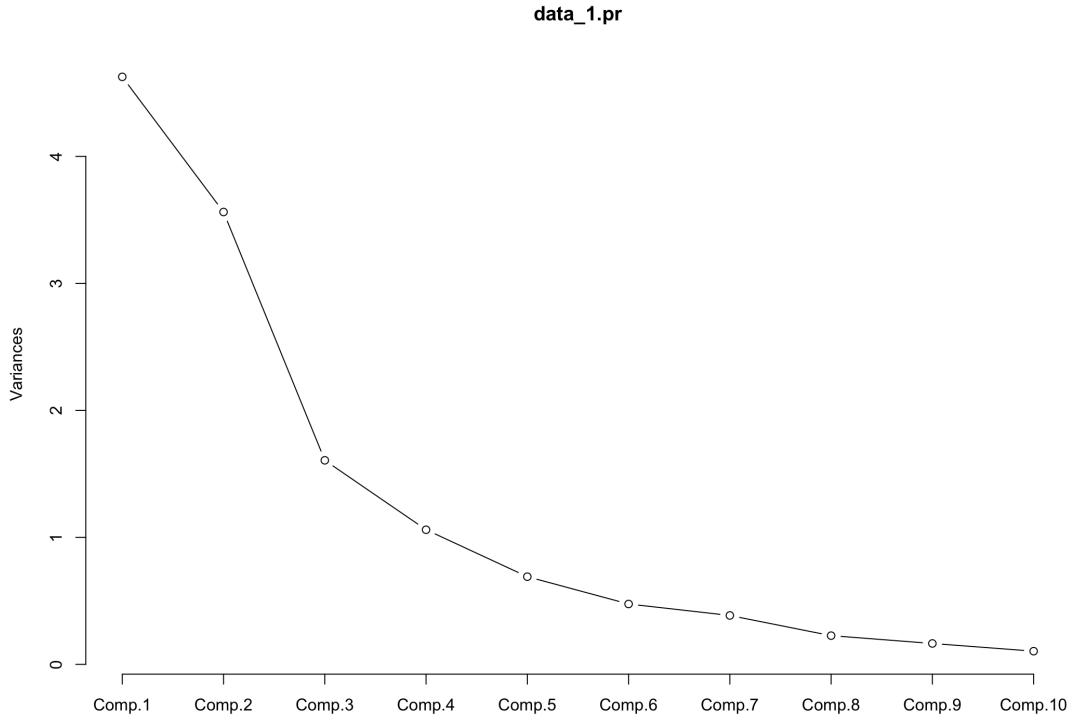
B part

In this part, the 14-predictor data set is considered. Use the Principle Components Regression method with collinearity reduction to decide which principle components will be included in the model. First, perform principal component analysis on the data, and each characteristic root obtained is shown in the following table.

Importance of components:					
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	2.1509034	1.8873635	1.2676074	1.02970302	0.83101723
Proportion of Variance	0.3558758	0.2740109	0.1236022	0.08156064	0.05312228
Cumulative Proportion	0.3558758	0.6298867	0.7534888	0.83504948	0.88817176
	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10
Standard deviation	0.68956449	0.62064086	0.47565084	0.40573791	0.32237596
Proportion of Variance	0.03657686	0.02963039	0.01740336	0.01266333	0.00799433
Cumulative Proportion	0.92474862	0.95437901	0.97178237	0.9844457	0.99244003
	Comp.11	Comp.12	Comp.13		
Standard deviation	0.2186159	0.192896	0.11522952		

Proportion of Variance	0.00367638	0.00286222	0.00102137		
Cumulative Proportion	0.99611641	0.99897863	1		

Observing the characteristic roots, it is found that the first 7 characteristic roots are all greater than 0.5, and the latter characteristic roots are relatively small.



The gravel diagram shows that when selecting the principal components, choosing 3 principal components is a better choice. The first 3 characteristic roots are obviously larger than the latter characteristic roots, and the cumulative proportion is guaranteed to reach 0.7534888. After selecting 3 principal components, the composition coefficient of each variable on the first three principal components are solved.

```
Loadings:
    Comp.1 Comp.2 Comp.3
H2S  0.190          0.216
SAL                 0.611
Eh7  0.103  0.268  0.450
pH   0.421         -0.286
BUF -0.425          0.210
P   -0.254 -0.159 -0.163
K            0.496
Ca   0.394 -0.107 -0.207
Mg  -0.159  0.482
Na           0.474
Zn  -0.415         -0.179
Cu           0.397 -0.380
NH4 -0.401 -0.108
```

Perform principal component regression and draw the model

$$BIO = \beta_1 + \alpha_1 * z_1 + \alpha_2 * z_2 + \alpha_3 * z_3$$

Among them, z_1, z_2, and z_3 are the selected three principal component values. The calculation formula of the principal component z_1 can be written as,

$$Z_1 = 0.19 * H2S + 0.103 * Eh7 + 0.421 * pH - 0.425 * BUF - 0.254 * P + 0.394 * Ca - 0.159 * Mg - 0.415 * Zn - 0.401 * NH4$$

The other two principal component formulas can be derived similarly. The model results obtained by principal component regression are shown in the following table.

Call:				
lm(formula = BIO ~ z1 + z2 + z3, data = data_1)				
Residuals:				
Min	1Q	Median	3Q	Max
-669.7	-212.7	-88.4	153.2	1135.7
Coefficients:				
Estimate	Std.	Error	t value	Pr(>|t|)
(Intercept)	1.00E+03	6.02E+01	16.618	< 2e-16 ***
z1	2.30E+02	2.80E+01	8.211	3.41e-10 ***
z2	-3.34E+01	3.19E+01	-1.046	0.30151
z3	-1.34E+02	4.75E+01	-2.819	0.00738 **
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 404 on 41 degrees of freedom				
Multiple R-squared: 0.6509, Adjusted R-squared: 0.6254				
F-statistic: 25.49 on 3 and 41 DF, p-value: 1.81e-09				

In the model, the R-square of the overall model is 0.6509, indicating that the model can explain the predicted data to a degree of 65.09%, and the p-value is 1.81e-09 and less than 0.05, so the model passes the F test. The independent variable coefficients were t-tested, and the P values of z_1 and z_3 were all less than 0.05. The null hypothesis was rejected and the test passed. The P values of z_2 were both greater than 0.05. There was no reason to reject the null hypothesis and failed the test.

(Intercept)	H2S	SAL	Eh7	pH
2.22E+03	-5.70E-02	-1.79E-01	-1.64E-01	-1.10E+00
BUF	P	K	Ca	Mg
8.60E-01	1.31E-01	-3.76E-02	1.37E-03	-1.16E-02
Na	Zn	Cu	NH4	
-1.55E-03	-4.34E-02	-8.65E+00	5.29E-02	

Using the principal component equation, bring it into the model, restore the coefficients of these 13 independent variables, and get the coefficients as shown in the table above.

C part

Use stepwise regression to determine the best model. First, the backward method is adopted, and the results obtained are shown in the following table.

Start: AIC=544.4				
BIO ~ SAL + pH + K + Na + Zn				
	Df	Sum of Sq	RSS	AIC
-Na	1	47011	6233274	542.74
-K	1	106211	6292475	543.17
-SAL	1	251921	6438184	544.2
<none>			6186263	544.4
-Zn	1	299209	6485473	544.53
-pH	1	1917306	8103569	554.55
Step: AIC=542.74				
BIO ~ SAL + pH + K + Zn				
	Df	Sum of Sq	RSS	AIC
<none>			6233274	542.74
-Zn	1	434796	6668070	543.78
-SAL	1	436496	6669770	543.79
-K	1	732606	6965880	545.74
-pH	1	1885805	8119079	552.64

The complete model is

$$BIO = \beta_1 + \alpha_1 * SAL + \alpha_2 * pH + \alpha_3 * K + \alpha_4 * Na + \alpha_5 * Zn$$

In the first step, we compare the AIC value of the model after deleting each variable (5 variables) and find that when Na is subtracted, the minimum AIC value is 542.74, so the following model is obtained.

$$BIO = \beta_1 + \alpha_1 * SAL + \alpha_2 * pH + \alpha_3 * K + \alpha_4 * Zn$$

The second step is to compare the AIC value of the model after deleting the remaining 4 variables. It is found that no matter which variable is deleted, the AIC value of the obtained model is greater than the current model, so stop and get the optimal model, that is, the current model.

Then use the stepwise method, and the results obtained are shown in the following table.

Start: AIC=544.4				
BIO ~ SAL + pH + K + Na + Zn				
	Df	Sum of Sq	RSS	AIC
-Na	1	47011	6233274	542.74
-K	1	106211	6292475	543.17
-SAL	1	251921	6438184	544.2
<none>			6186263	544.4
-Zn	1	299209	6485473	544.53
-pH	1	1917306	8103569	554.55
Step: AIC=542.74				
BIO ~ SAL + pH + K + Zn				
	Df	Sum of Sq	RSS	AIC
<none>			6233274	542.74
-Zn	1	434796	6668070	543.78
-SAL	1	436496	6669770	543.79
+Na	1	47011	6186263	544.4
-K	1	732606	6965880	545.74
-pH	1	1885805	8119079	552.64

The complete model is

$$BIO = \beta_1 + \alpha_1 * SAL + \alpha_2 * pH + \alpha_3 * K + \alpha_4 * Na + \alpha_5 * Zn$$

In the first step, we compare the AIC value of the model after deleting each variable (5 variables) and find that when Na is subtracted, the minimum AIC value is 542.74, so the following model is obtained.

$$BIO = \beta_1 + \alpha_1 * SAL + \alpha_2 * pH + \alpha_3 * K + \alpha_4 * Zn$$

The second step is to compare the AIC value of the model after deleting the remaining Zn, SAL, K, and pH with the AIC value of the model with Na added. It is found that no matter whether it is deleted or increased, the AIC value of the obtained model is greater than the current model, so stop, The optimal model is the current model.

It is found that the optimal model selected by the backward method and the stepwise method is the same. Check the results of the model.

Call:				
lm(formula = BIO ~ SAL + pH + K + Zn, data = mydata_1)				
Residuals:				
Min	1Q	Median	3Q	Max
-749.1	-229.2	-94.2	127.2	1037.4
Coefficients:				
Estimate	Std.	Error	t value	Pr(>|t|)
(Intercept)	1.51E+03	1.13E+03	1.328	0.19172
SAL	-3.59E+01	2.15E+01	-1.674	0.10201
pH	2.94E+02	8.45E+01	3.479	0.00123**
K	-4.39E-01	2.02E-01	-2.168	0.03615*
Zn	-2.35E+01	1.40E+01	-1.67	0.10265
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 394.8 on 40 degrees of freedom				
Multiple R-squared: 0.6749, Adjusted R-squared: 0.6423				
F-statistic: 20.76 on 4 and 40 DF, p-value: 2.528e-09				

In the model, the R-square of the overall model is 0.6749, indicating that the model can explain the predicted data to a degree of 67.499%, and the p-value is 2.528e-09 less than 0.05, so the model passes the F test. Carry out t test on the independent variable coefficients. Based on the significance level of 0.15, the P values of SAL, pH, K, and Zn are all less than 0.15. The null hypothesis is rejected and the test is passed.

Using subset selection method, the best two-variable model is determined on the basis of analysis. The result of selecting a subset is as follows.

```
Selection Algorithm: exhaustive
         SAL pH  K   Na  Zn
1  ( 1 ) " " "*" " " " " " "
2  ( 1 ) " " "*" " " "*" " "
3  ( 1 ) " " "*" " " "*" "*"
4  ( 1 ) "*" "*" "*" " " "*"
5  ( 1 ) "*" "*" "*" "*" "*"
```

Compare the AIC value, Cp value and BIC value of these 5 models,

	AIC	Cp	BIC
1	0.5900471	7.420574	-33.54828
2	0.6421676	2.281592	-36.91949
3	0.6377518	3.796	-33.64529
4	0.6423444	4.29637	-31.52397
5	0.6359404	6	-28.05798

The results show that the Cp value and BIC value of Model 2 are the smallest among the 5 models, and its AIC value is not high, so we think that Model 2 is the best model. Two independent variables are selected in Model 2 as pH and Na. Next, we construct all the subsets of the two variables and find the AIC value and BIC value of the corresponding model.

	AIC	BIC
SAL + pH	677.3885	684.6152
SAL + K	716.5351	723.7618
SAL + Na	715.3706	722.5972
SAL + Zn	682.8102	690.0368
pH + K	672.0741	679.3007
pH + Na	670.666	677.8927
pH + Zn	676.8311	684.0577
K + Na	715.5294	722.756
K + Zn	694.8634	702.0901
Na + Zn	693.7078	700.9344

The results show that the AIC value and BIC value corresponding to the model of choosing pH and Na are both the smallest, which verifies that the choice we made in the previous step is indeed correct. Since there is no tie, there is no need to use VIF to break the tie. The best two-variable model is as follows:

$$BIO = -475.7 + 404.9 * pH - 233.3 * Na$$

The specific parameters of the model are as follows,

Call:				
lm(formula = BIO ~ pH + Na, data = mydata_1)				
Residuals:				
Min	1Q	Median	3Q	Max
-677.93	-229.76	-97.47	207.51	1168.4
Coefficients:				
Estimate	Std.	Error	t value	Pr(>|t|)
(Intercept)	-4.76E+02	2.74E+02	-1.739	0.0893.
pH	4.05E+02	4.78E+01	8.477	1.22e-10 ***
Na	-2.33E-02	8.66E-03	-2.695	0.01018
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				

In the model, the R-square of the overall model is 0.6584, indicating that the model can explain the predicted data to a degree of 65.84%, and the p-value is 1.596e-10 less than 0.05, so the model passes the F test. The independent variable coefficients were

tested by t. On the basis of the significance level of 0.05, the P values of pH and Na were both less than 0.05. The null hypothesis was rejected and the test passed.