Project Report

(MATH 584 Applied Statistics)

Kyung Jin Kwak

A20497336

Illinois Institute of Technology

12.07.2023

Part 1

1) Use OLS to estimate Regression coefficients.

- Based on the summary output using OLS, Regression coefficients of each of predictors are as follows in the Red box:

OLS Regression Results

Dep. Variable:	BIO	R-squared:	0.823
Model:	OLS	Adj. R-squared:	0.734
Method:	Least Squares	F-statistic:	9.270
Date:	Thu, 07 Dec 2023	Prob (F-statistic):	4.03e-07
Time:	11:46:30	Log-Likelihood:	-302.70
No. Observations:	43	AIC:	635.4
Df Residuals:	28	BIC:	661.8
Df Model:	14		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	3475.9507	3441.050	1.010	0.321	-3572.720	1.05e+04
H2S	1.1544	3.048	0.379	0.708	-5.089	7.398
SAL	-19.2305	26.581	-0.723	0.475	-73.679	35.218
Eh7	2.4120	1.964	1.228	0.230	-1.612	6.435
pH	149.1615	330.050	0.452	0.655	-526.915	825.238
BUF	-19.6909	121.063	-0.163	0.872	-267.676	228.295
P	-6.1819	3.854	-1.604	0.120	-14.077	1.713
K	-1.0168	0.474	-2.144	0.041	-1.988	-0.045
Ca	-0.0657	0.125	-0.524	0.604	-0.323	0.191
Mg	-0.3667	0.273	-1.343	0.190	-0.926	0.192
Na	0.0100	0.024	0.411	0.684	-0.040	0.060
Mn	-3.6814	5.513	-0.668	0.510	-14.975	7.612
Zn	-8.0818	21.989	-0.368	0.716	-53.125	36.961
Cu	373.8948	110.351	3.388	0.002	147.852	599.938
NH4	-1.5510	3.219	-0.482	0.634	-8.145	5.043

Omnibus:	10.120	Durbin-Watson:	1.791
Prob(Omnibus):	0.006	Jarque-Bera (JB):	14.888
Skew:	0.602	Prob(JB):	0.000585
Kurtosis:	5.619	Cond. No.	1.22e+06

2) Run Collinearity diagnostics (VIF, Condition Index)

- Looking at the result from the two Collinearity diagnostics method, it is observed from VIF that there are 6 predictors that exceeds 10, meaning there presents serious multicollinearity (pH, BUF, Ca, Mg, Na, Zn). From Condition Index, since there are no predictors that exceeds 30, we see no serious collinearity, but it suggests that there exists collinearity from the predictor NH4 for it is exceeding 15.
- Below is the code output of 'VIF' and 'Condition Index' to check collinearity of the predictors :

VIF	Condition Index
Predictors VIF const 4350.771896 H2S 3.136506 SAL 3.361283 Eh7 1.964076 pH 62.564383 BUF 33.478422 P 2.884226 K 7.432133 Ca 17.343432 Mg 24.476419 Na 10.372624 Mn 6.737786 Zn 12.391033 Cu 4.866983 NH4 8.586275	Predictors Condition Index H2S 1.000000 SAL 1.184103 Eh7 1.791479 pH 2.048414 BUF 2.733624 P 3.241408 K 3.696482 Ca 4.447720 Mg 5.687906 Na 6.009633 Mn 7.842823 Zn 10.674971 Cu 13.561188 NH4 23.308398
Thresholds: - VIF = 1 : Best case - VIF between 4 & 10 : Needs further investigation - VIF > 10 : Serious Multicollinearity	Thresholds (Where, $k = \sqrt{\frac{\lambda_1}{\lambda_j}}$) : - k ≥ 15 : Collinearity exists - k ≥ 30 : Serious Multicollinearity

Part 2

1) Use Principal Components Regression (PCR) method with collinearity reduction.

- The collinearity reduction was conducted based on value of Coefficients, Explained Variance, and Condition Index.
    i. If the Coefficient is close to zero, it means there is no significance in the model, so it is better to be excluded.
    ii. Explained Variance is the proportion of the total variability in a dataset that is accounted for by the statistical model, similar concept as R-squared. If the Explained Variance is low, it can be also considered to be excluded from the model.
    iii. If Condition Index is greater than 15, it means there exists collinearity, hence the predictor will be excluded from the model. (Explained Variance here is used only to double check but solely used to select predictor)
- Following is the code output :

Coefficients

```
Coefficients θj in the Standardized Model:
         H2S        SAL         Eh7          pH        BUF          P  \
0  211.75609 -79.789789 -105.921327  118.530564 -65.106255 -0.242776

           K         Ca          Mg          Na          Mn          Zn  \
0  263.530008 -52.807876  349.583378  174.118541  258.838493  186.25233

          Cu         NH4
0  196.818913  163.294501
```

Explained Variance	Condition Index

Explained Variance Ratio of Each Principal Component:

	Explained Variance
H2S	0.369445
SAL	0.263494
Eh7	0.115113
pH	0.088047
BUF	0.049439
P	0.035163
K	0.027038
Ca	0.018676
Mg	0.011419
Na	0.010229
Mn	0.006006
Zn	0.003242
Cu	0.002009
NH4	0.000680

Condition Index of Each Principal Component:

	Condition Index
H2S	1.000000
SAL	1.184103
Eh7	1.791479
pH	2.048414
BUF	2.733624
P	3.241408
K	3.696482
Ca	4.447720
Mg	5.687906
Na	6.009633
Mn	7.842823
Zn	10.674971
Cu	13.561188
NH4	23.308398

- Looking at the result, it is observed that predictor 'P' has insignificant coefficient and its Explained Variance is also low. There are other predictors whose Explained Variance are also very low, but their coefficients are not considered as insignificant. Also, predictor 'NH4' condition index is 23 which is greater than 15, there can be collinearity exists. so will exclude.
- Excluded predictors: P, NH4

  Remaining predictors: H2S, SAL, Eh7, pH, BUF, K, Ca, Mg, Na, Mn, Zn, Cu

2) Compare the standard error sum $\sum_j s.e(\hat{\beta}_j)$ and SSE with their counterparts in Part I

- From Part 1, we have $\sum_j s.e(\hat{\beta}_j)$ and SSE for Full model as follows :

```
Standard Error Sum (SSE): 3276740.2803900684
Sum of Standard Errors of Coefficients: 628.529003575566
```

- If we apply the previous result to exclude predictor to make Reduced model, we can get the value of $\hat{\beta}_j$, and SSE as follows:

```
Sum of Squared Errors (SSE): 3287657.542488552
Standard Errors of the Coefficients:
s.e.(β̂_1): 18.541127672701244
s.e.(β̂_2): 21.954607722598457
s.e.(β̂_3): 33.21604517681828
s.e.(β̂_4): 37.979912316268276
s.e.(β̂_5): 50.68446649294327
s.e.(β̂_6): 68.53694945916295
s.e.(β̂_7): 82.46574301595605
s.e.(β̂_8): 105.46018376016357
s.e.(β̂_9): 111.42536934798409
s.e.(β̂_10): 145.4147795921936
s.e.(β̂_11): 197.92599540395935
s.e.(β̂_12): 251.4397114448181
```

- Let's compare standard error sum $\sum_j s.e(\widehat{\beta_j})$ and SSE each both from Full model and Reduced model :

```
Full Model - SSE: 3276740.280390065 , Sum of Standard Errors: 1614.6207173853195
Reduced Model - SSE: 3287657.542488552 , Sum of Standard Errors: 1125.0448914055673
```

- We can conclude that SSE from Full model is smaller than that of Reduced model after excluding predictor 'P' and 'NH4'. On the other hand, $\sum_j s.e(\widehat{\beta_j})$ from Full model is larger than that of Reduced model.

Part 3

Part 3.1)

3.1.1) Build stepwise regression method (using $\alpha_E = \alpha_R = 0.1$) and report each step explicitly.

- The process of Stepwise regression method:
  i. Start from empty model, with no predictors included.
  ii. Forward selection: Evaluate all available predictors (SAL, pH, K, Na, Zn) by adding each one individually to the model and calculating the p-value of its coefficient, and then decide which predictor enters the model based on the p-value.
  iii. Backward selection: Evaluate all included variables to ensure they still have p-values below the threshold, 0.1.
  iv. Iterate until no more variables meet the criteria for inclusion or exclusion.

- The stepwise regression process output:

Iterations	Notes

1st Forward step
`Add pH with p-value 1.61671e-09`
`Compared to other variables:`
`pH 1.616712e-09`
`Zn 4.126330e-06`
`Na 9.500137e-02`
`K 2.082624e-01`
`SAL 6.365223e-01`
	- Models:
	Y ~ pH
	Y ~ Zn
	Y ~ Na
	Y ~ K
	Y ~ SAL
	
	- 'pH' enters the model with lowest p-value of 1.61671e-09.

1st Backward step
`No variable dropped. Current model p-values:`
`pH 1.616712e-09`
	- No change, p-value of pH < 0.1.

2nd Forward step
`Add Na with p-value 0.0142458`
`Compared to other variables:`
`Na 0.014246`
`K 0.026971`
`Zn 0.272026`
`SAL 0.608300`
	- Models:
	Y ~ pH + Na
	Y ~ pH + K
	Y ~ pH + Zn
	Y ~ pH + SAL
	
	- 'Na' enters the model with lowest p-value of 0.0142458.

2nd Backward step
`No variable dropped. Current model p-values:`
`pH 4.731149e-10`
`Na 1.424576e-02`
	- No change, p-value of pH and Na < 0.1.

3rd Forward step No additional predictors enter the model because their p-v Best p-value among excluded variables: 0.4302895702889037 Excluded variables and their p-values: Zn 0.430290 K 0.641029 SAL 0.844553	- The iteration stops here, since there is no p-value from predictors below 0.1. **- Resulting predictors: 'pH' and 'Na'**

- After Including the selected predictors chosen from stepwise function to the model, the p-value of both 'pH' and 'Na' is below the threshold 0.1, hence these two are the final model.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-466.3748	279.219	-1.670	0.103	-1030.698	97.948
pH	400.4547	49.046	8.165	0.000	301.329	499.580
Na	-0.0227	0.009	-2.563	0.014	-0.041	-0.005

Therefore, the model we get from this stepwise function is as follows (X_1= pH, X_2= Na) :

$$\hat{Y} = -466.3748 + 400.4547X_1 - 0.0227X_2$$

3.1.2) Check collinearity diagnostics (VIF)

- Comparison of VIF from Full model vs. Reduced model (stepwise regression)

Full model

Predictors	VIF
const	420.277700
SAL	2.099364
pH	3.327339
K	2.982513
Na	3.311625
Zn	4.309322

Reduced model (Stepwise regression)

Predictors	VIF
const	20.746465
pH	1.000558
Na	1.000558

- Looking at the result above, the VIF after selecting predictors with stepwise regression shows much smaller value compared to that of predictors from the Full model.

Part 3.2)

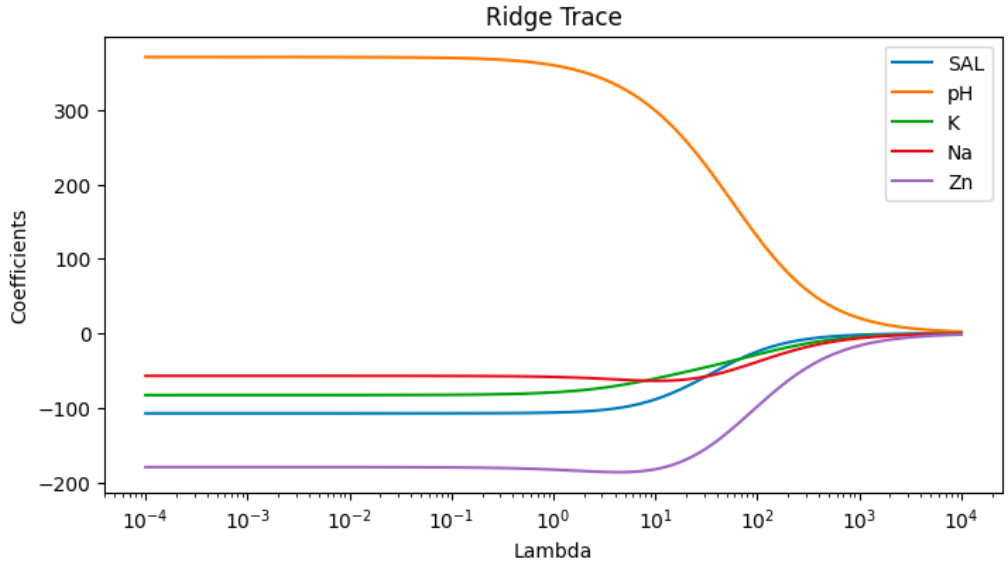
3.2.1) Conduct Ridge Trace and Cross-validate calculating MSE to get optimal λ value.

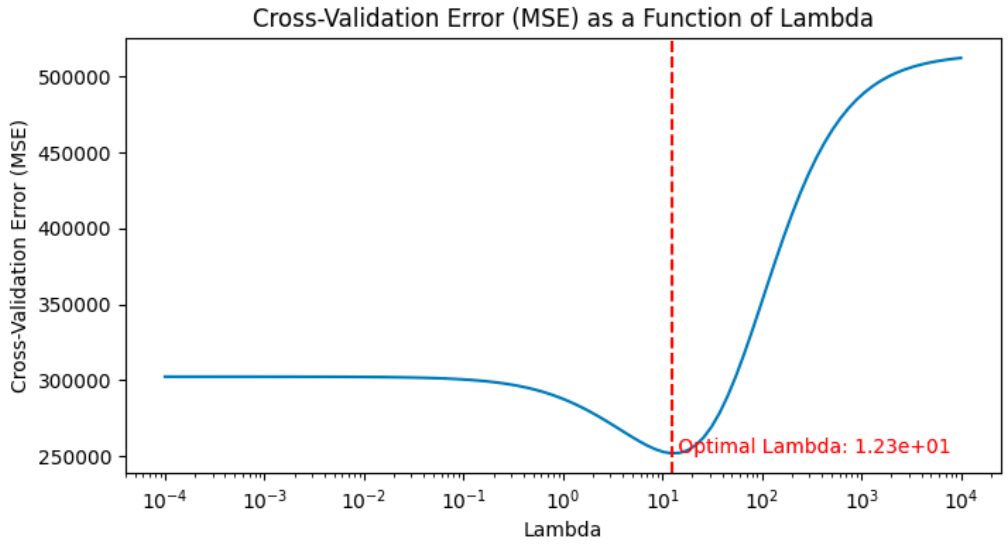
- The process of selecting variable based on Ridge Trace:
  i. Standardize the predictor variables (X)
  ii. Compute Ridge Regression for various λ values
  iii. Plot the Ridge Trace

iv. Select an appropriate λ based on cross-validation error which is Mean Squared Error(MSE) for each λ values that minimizes the MSE

- Ridge Trace and optimal λ value output:

  As seen in the two plots below, we can see that the optimal λ is approximately 12.33. This value represents the best balance between bias and variance for the model, according to the MSE metric used in cross-validation.





3.2.2) Variable Selection:

- Variables can be selected by comparing Ridge coefficients and OLS coefficients. If coefficients both from Ridge and OLS shows noticeable difference, those predictors could be excluded from the model.
- Looking at the result below after applying chosen λ value and calculating back for OLS coefficients, it is observed that predictor 'Na' and 'Zn' shows less difference each other compared to other predictors, hence these two predictors are selected to be included in the model.

Predictors	Ridge Coefficients	OLS Coefficients
SAL	-84.358184	-107.539104
pH	287.982203	370.842749
K	-58.204827	-83.019174
Na	-63.851492	-57.225485
Zn	-179.922429	-179.789802

3.2.3) Check Collinearity Diagnostics (VIF) for selected model:

- Looking at the result above, the VIF after selecting predictors Ridge Trace shows smaller value compared to that of predictors from the Full model.

Full model

Predictors	VIF
const	420.277700
SAL	2.099364
pH	3.327339
K	2.982513
Na	3.311625
Zn	4.309322

Reduced model (Ridge Trace)

Predictors	VIF
Constant	1.000000
Na	1.015028
Zn	1.015028

Part 3.3)

3.3.1) Build Subset Selection method (using BIC and VIF) and report each step explicitly.

- The process of Subset Selection method:
  i. For a two-variable model, generate all possible pairs of predictors.
  ii. For each pair of predictors, fit a OLS regression model and calculate BIC and VIF.
  iii. Identify the model with the lowest BIC value as it suggests a good balance between model complexity and fit. (In case of tie BIC, use VIF)

- The Subset Selection process result:
  There are total of 10 subsets (5 combination 2) as possible pairs of predictors. Looking at the result below, we can conclude that the best two variable subset model is ‘pH + Na’ with the lowest BIC value of 645.8937.

```
------------------------------
The best two-variable model based on the lowest BIC is: pH + Na
With BIC: 645.8937195289844 and Max VIF: 1.0005584029303853
```

- More details of the Subset Selection process from the code are shown below output:

BIC and VIF value of all possible pairs of predictors

```
Model: SAL + pH
BIC: 652.1459756701684
Max VIF: 1.0008328341478487
--------------------------------
Model: SAL + K
BIC: 689.0726713443028
Max VIF: 1.0004265994130863
--------------------------------
Model: SAL + Na
BIC: 688.0009853893404
Max VIF: 1.0200335426235634
--------------------------------
Model: SAL + Zn
BIC: 656.5687438839859
Max VIF: 1.2172307994771843
--------------------------------
Model: pH + K
BIC: 647.106807265751
Max VIF: 1.0008165409299896
--------------------------------
```

```
Model: pH + Na
BIC: 645.8937195289844
Max VIF: 1.0005584029303853
--------------------------------
Model: pH + Zn
BIC: 651.1186233056428
Max VIF: 2.1484671152888484
--------------------------------
Model: K + Na
BIC: 688.0579006245046
Max VIF: 2.7248709843797716
--------------------------------
Model: K + Zn
BIC: 666.8280215308667
Max VIF: 1.0049129920618014
--------------------------------
Model: Na + Zn
BIC: 666.0664388199644
Max VIF: 1.0150282539083788
--------------------------------
```