Midwest Undergraduate Data Analytics Competition

Team 51

Yuchen Li, Ziqin Xiong, Keyu Hu, Changching Lan, Mengqi Huang

Problem 1(a): Simple Linear Regression

TSS

1.First do model selection by vif, then do further selection by Mallow Cp   OR do model selection by Cp on all predictors.

2. Transformation   **(1)** sqrt(response) **(2)** (response)^2   **(3)**degree 2 polynomial

3.for each transformation, perform model selection by Mallow Cp or AIC, etc.

4. Compare all the models with cross-validation mse, r^2, diagnostic plots.(Best subset selection with Mallow's Cp on sqrt(response) transformation is the optimal model)

5. Outlier only exists in model "3(1)" (sqrt(response))   observation#20 nemadji

Nitrate

1. The first two steps are the same as TSS.
2. Use Box-Cox transformation, the suggested transformation is natural logarithm
3. Do model selection with Mallow Cp
4. Compare all the models with cross-validation mse, r^2, diagnostic plots.(Best subset selection with Mallow’s Cp on log(response) transformation is the optimal model)
5. No outlier

TSS(with the optimal model)

Predictors: "Cropland", "Developed.Urban", "Shallow.bedrock.under.cropland", "Clay", "Lake.interception"

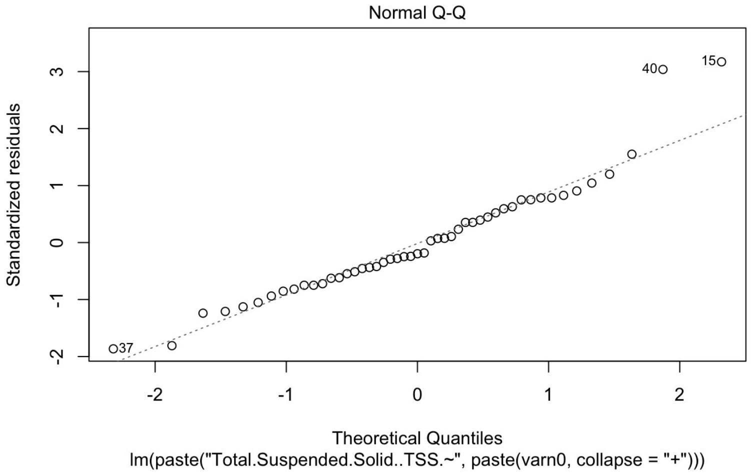
Cross-Validation MSE : 2736.756

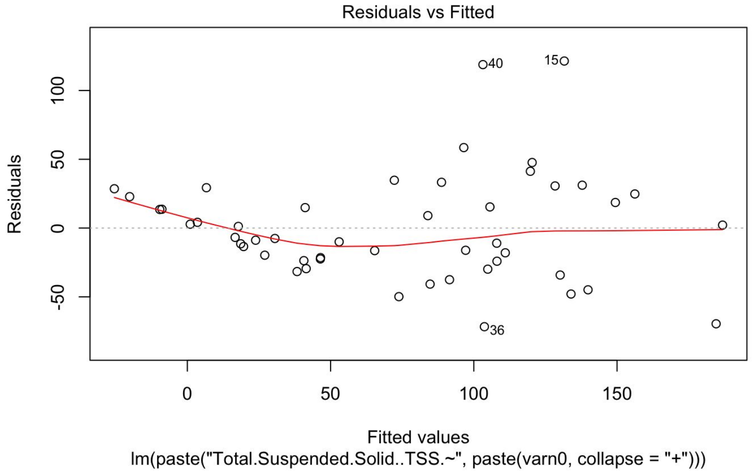
CV MSE for Full Model: 5231.847

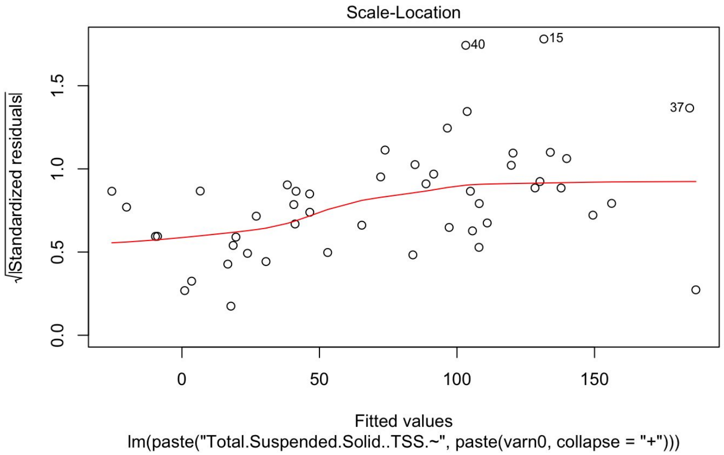
CV MSE for model with all predictors selected by Cp: 4083.401

CV MSE for model with vif selected predictors(further selected by Cp): 4343.982

Diagnostic Plots:







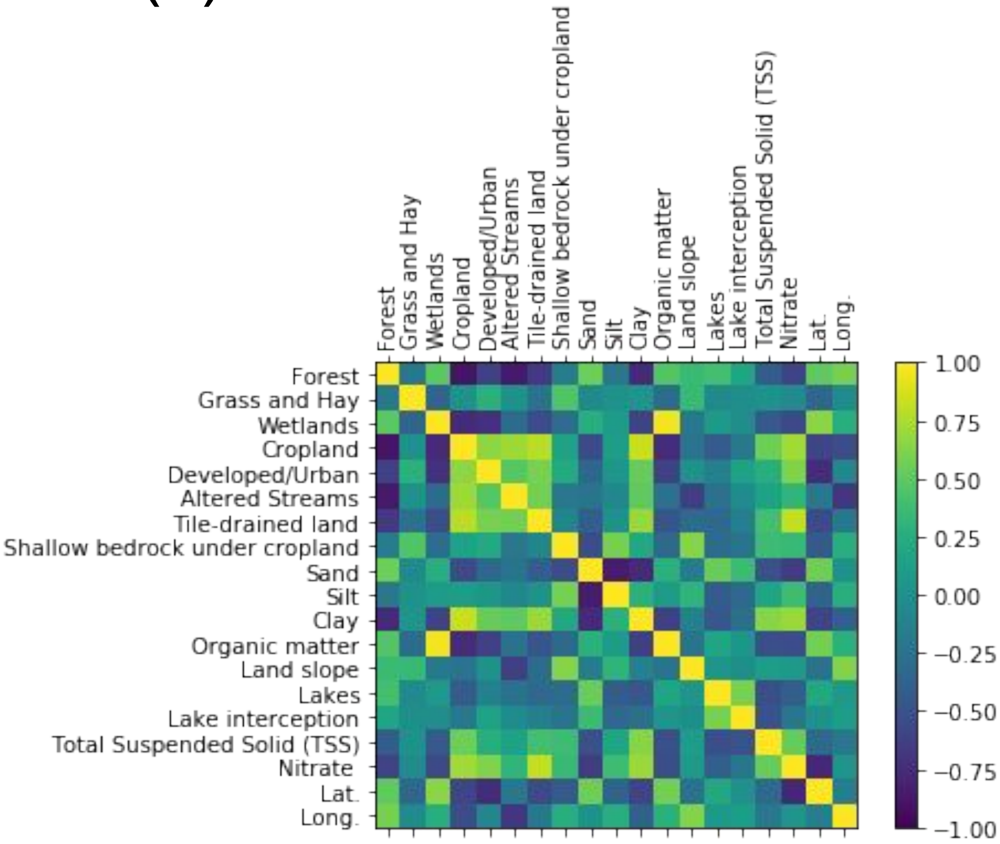
Nitrate(with the optimal model)

1. Predictors: "Forest", "Wetlands", "Tile.drained.land" "Sand", "Clay", "Land.slope", "Lakes".
   a. Lowest CV MSE
   b. Highest Adjusted R-squared
   c. Smoothest Diagnostic Plot

Problem 1(b): Correlation in Responses

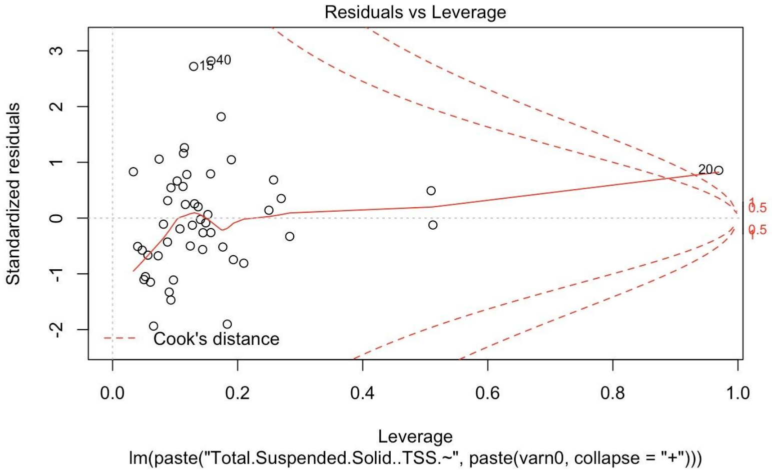
```
Pearson's product-moment correlation

data:  p1$Total.Suspended.Solid..TSS. and p1$Nitrate

t = 4.3599, df = 48, p-value = 6.847e-05

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

 0.2985244 0.7062585

sample estimates:

      cor

0.5326152
```

Problem 1(c): Correlation in Predictors



Problem 1(d): Outliers

The diagnostic plot of the best subset selection with Mallow’s Cp on sqrt(response) transformation implies that the 20th observation(nemadji) is an influential point(it has Cook’s Distance larger than 1).



Problem 1: Common Predictors for TSS and Nitrate

- Predictors: Clay
- When wet, clay soils are easily compacted, which also increases runoff that contain contaminants and could affect the quality of surface and ground-water by holding water. Clay-sized sediment is of small size that it can greatly increase water treatment costs.
- Insights:As soil managers we can help build soil aggregates by growing green manure cover crops or adding animal manure. If farmers can reduce tillage operations, it may be possible to reduce the disturbance to the soil biota that are essential for building aggregation. Feeding the soil food web with cover crops or other organic materials also increases the numbers of these organisms. Then bacteria and fungi work to help make aggregation happen.

Problem 2(a): Annual Precipitation Intensity Index

1. Number of days Mar-Nov where at least 1/2 the precipitation zones in a watershed had more than 1 inch of precipitation.

2. Number of days Mar-Nov where at least 1/2 the precipitation zones in a watershed had more than 2 inch of precipitation.

3. Days: identify all days Mar-Nov when at 1/3 of the precipitation zones in a watershed had > 1 inch of precipitation. Hours: from those dates, the number of hours when at 1/3 of the precipitation zones in a watershed had > 0.25 inch of precipitation.

Problem 2(a): APII 4-6 (April to June)

❏ The above three measures, but with the months changed to April-June.

❏ Motivation: as pointed out in the problem, (April-June) may have the greatest impact on water quality because vegetative cover on the landscape has not fully developed.

Problem 2(a): APII 7-8 (May to September)

7. Number of days May-Sep where at least 1/3 the precipitation zones in a watershed had more than 0.5 inch of precipitation.

8. From those dates, the number of hours when at 1/3 of the precipitation zones in a watershed had > 0.25 inch of precipitation.

Motivations:

- ❏ December, January, and February should be excluded as precipitation in these month is likely to be in the form of snow.
- ❏ Take it further to consider warmer months only.

Problem 2(a): APII 9-10 (0.75 Coverage)

9. Number of days Apr-Jun where at least 3/4 the precipitation zones in a watershed had more than 0.25 inch of precipitation.

10. From those dates, the number of hours when at 3/4 of the precipitation zones in a watershed had > 0.1 inch of precipitation.

Motivations:

- From the previous analyses of correlations, we see that April-June are the most important months, consistent with the provided knowledge.
- A large area of precipitation tends to have a large effect on water quality.

Problem 2(a): APII 11-12 (0.67 Coverage)

11. Number of days Apr-Jun where at least 2/3 the precipitation zones in a watershed had more than 0.4 inch of precipitation.

12. From those dates, the number of hours when at 2/3 of the precipitation zones in a watershed had > 0.1 inch of precipitation.

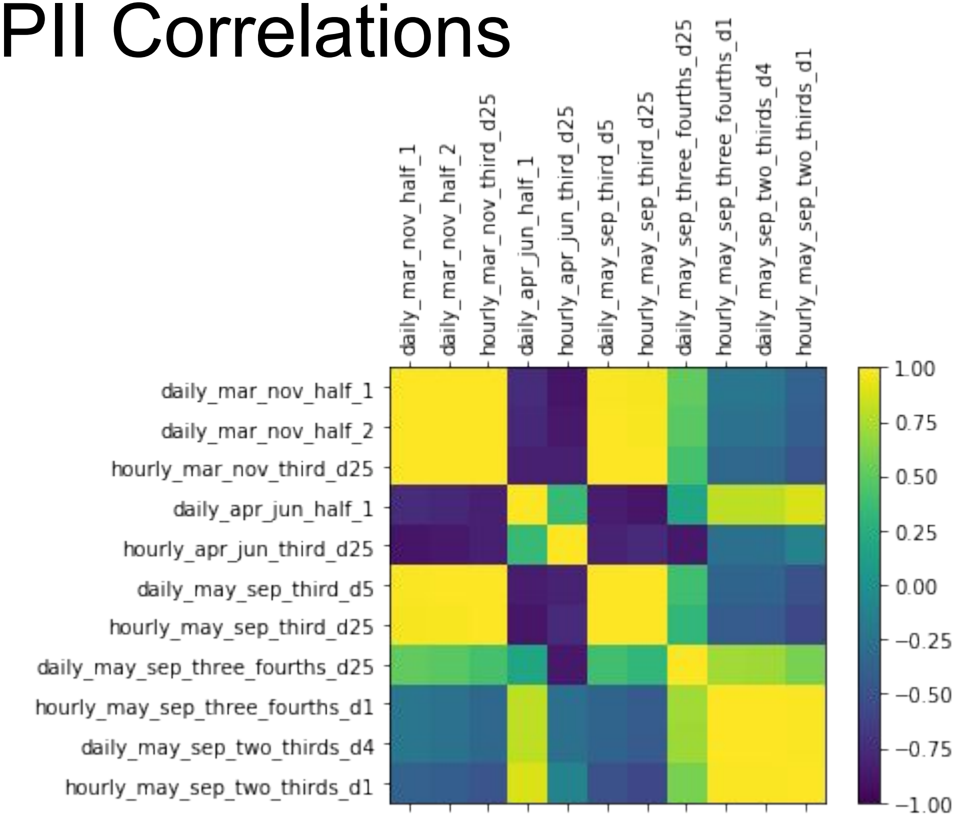
Motivations:

- ❏ The previous index defined on date+hour resulted in sparse indices in some cases.
- ❏ We loosen the requirements to reduce sparsity.

Problem 2(a): APII Example

2014-otter	daily_mar_nov_half_1	daily_mar_nov_half_2	hourly_mar_nov_third_d25	daily_apr_jun_half_1	daily_apr_jun_half_2	hourly_apr_jun_third_d25
type	daily	daily	daily +hourly	daily	daily	daily +hourly
Month range	Mar-Nov	Mar-Nov	Mar-Nov	April-June	April-June	April-June
Threshold	>= ½ the precipitation zones had >1 inch of precipitation	>= ½ the precipitation zones had >1 inch of precipitation	1.days: ⅓ the precipitation zones had >1 inch of precipitation 2.hours:⅓ 0.25 inch	>= ½ the precipitation zones had >1 inch of precipitation	>= ½ the precipitation zones had >2 inch of precipitation	1.days: ⅓ the precipitation zones had >1 inch of precipitation 2.hours:⅓ 0.25 inch
	6	0	15	6	0	11

Problem 2(b): APII Correlations



Problem 2(b): Relation between APII and Quality

- Using single linear model (only one APII)

Le Sueur		Otter		Root		St.Louis	
Nitrate	TSS	Nitrate	TSS	Nitrate	TSS	Nitrate	TSS
hourly, May-Sep, 3/4, 0.1	hourly, May-Sep, 1/3, 0.25	daily, May, Sep, 1/5, 0.5	hourly, May-Sep, 1/3, 0.25	daily, Mar-Nov, 1/2, 1	daily, Mar-Nov, 1/2, 1	daily, Mar-Nov, 1/2, 1	daily, Apr-Jun, 1/2, 1

Problem 2(b): Cont.

- Using multivariate linear model:
    - We only have 5 to 8 observations here, so it’s very easy to go overfitting when we include more than one predictor.
    - With higher-frequency data, we can try more complex model.

Problem 2(c): Applying APII to Previous Models

- In spite of adding APII as an extra predictors in previous model, interaction between APII and other spacial variables may also be useful.
- Essentially, we lose data when creating the APII. We can create better time-series model using higher frequency time-based water quality data.