HAB Predictive Model Building

Table of Contents

# Variable Selection

From our full dataset, a best subset regression is used to compare all possible models.

With the best variables from the regression subset, a forward and backward step-wise regression will be preformed to further search the best fit model. First, variables will be backwardly selected based on Anova type II Sum of Square test. Type II SS tests each variable after accounting all the other variable.

Our best model will be chosen by the lowest Bayesian Information Criterion (BIC). The goal is to have a model that is simple and is robust.

The chosen model(s) will be further validated by k-fold repeated cross-validation.

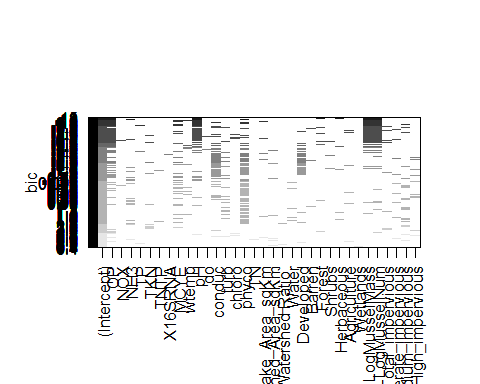
The MC congeners and categorical variables are not observed.

Subset analysis is done on an averaged dataset based on each lake. This results with observations. Microcystin response variable is assessed seperatly.

Here are all variable with their shorten name:

|  |  |
| --- | --- |
| Short.Name | Variable.Name./.Type |
| OP | Ortho-P (mg-P/L) |
| NOX | Nitrate/Nitrite (mg-N/L) |
| NH3 | Ammonia (mg-N/L) |
| TP | Total Phosphorus (mg-P/L) |
| TKN | Total Kjeldahl Nitrogen (mg-N/L) |
| TNTP | Total Nitrogen to Total Phosphorus Ratio |
| X16SRNA | Cyano 16s rRNA gene copies (cp/mL) |
| MCYE | mcyE gene copies (cp/mL) |
| SUM | Microcysin sum from all 12 MC congeners (ppb) |
| S\_SUM | Microcystin sum of 12 congeners from SPATTS (ng of MC/g of resin) |
| wtemp | Water Temprature ( C ) |
| pH | pH |
| do | Dissolved Oxygen (mg/L) |
| conduc | Conductance (uS/cm) |
| turb | Turbidity (NTU) |
| chloro | Chloraphyll-a (RFU) |
| phyco | Phycocyanin (RFU) |
| TN | Total Nitrogen to Total Phosphorus Ratio |
| Lake\_Area\_sqKm | Lake Area (sq Km) |
| Watershed\_Area\_sqKm | Watershed Area (sq Km) |
| Lake Area\_Watershed Ratio | Lake Area/Watershed Area Ratio |
| Water | Land-Use Percentage in Lakes Watershed |
| Developed | Land-Use Percentage in Lakes Watershed |
| Barren | Land-Use Percentage in Lakes Watershed |
| Forest | Land-Use Percentage in Lakes Watershed |
| Shrubs | Land-Use Percentage in Lakes Watershed |
| Herbaceous | Land-Use Percentage in Lakes Watershed |
| Agriculture | Land-Use Percentage in Lakes Watershed |
| Wetlands | Land-Use Percentage in Lakes Watershed |
| LogMusselMass | Zebra Mussel Mass (grams) |
| LogMusselNum | Zebra Mussel Counts |
| Total\_impervious | Area of all impervious surfaces |
| Moderate\_impervious | Area of moderate to high impervious surfaces |
| Medium\_impervious | Area of medium to high impervious surfaces |
| High\_impervious | Area of high impervious surfaces |

## MC Sum from grab sample as response

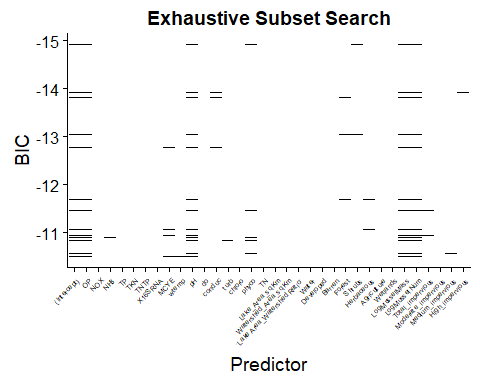


Subset Analysis: Exhaustive Search

Each row is a model, if a variable is included in the model it is represented as a shaded rectangle. The BIC is plotted on the y axis where the lowest value is higher up on the axis. The better the model, the lower the BIC, thus the top rows are the better model. The frequency of the variables being included can also be important. The presence of one variable in the model and the effect of another variables presence/absence can also be important.

In the full subset analysis, ortho-P, *mcyE*, turb, phyco, pH, conduc, forest, shrubs, mussle mass/number and medium impervious areas frequently included in the lowest BIC values (Higher up on the y-axis). They may signify some importance.

*There is a clearer plot in the next page*



Subset Analysis for Predicting MC Sum: Top 15 Models from Exhaustive Search

*A plot of a portion of the full subset. BIC range of -10 to -15 is shown here.*

The variables in the best 3 models includes ortho-P, *mcyE*, pH, mussel mass/number and area of impervious surfaces.

With all the best variables from regression subset analysis, variables will be tested by F-test and dropped when

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| OP | 0.8556 | 1 | 11.3316 | 0.0039 |
| MCYE | 0.5758 | 1 | 7.6254 | 0.0139 |
| turb | 0.0399 | 1 | 0.5278 | 0.4780 |
| Herbaceous | 0.0321 | 1 | 0.4249 | 0.5238 |
| Forest | 0.0876 | 1 | 1.1596 | 0.2975 |
| pH | 0.8419 | 1 | 11.1504 | 0.0042 |
| phyco | 0.1959 | 1 | 2.5939 | 0.1268 |
| LogMusselMass | 0.8955 | 1 | 11.8592 | 0.0033 |
| LogMusselNum | 0.7646 | 1 | 10.1263 | 0.0058 |
| Total\_impervious | 0.0527 | 1 | 0.6985 | 0.4156 |
| Medium\_impervious | 0.0453 | 1 | 0.6003 | 0.4498 |
| Shrubs | 0.0639 | 1 | 0.8456 | 0.3714 |
| Residuals | 1.2081 | 16 | NA | NA |

We will drop one variable with the highest value. This is repeated until the variables are all signifigant.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sum Sq Df F valu | e Pr(>F | ) |  |  |
| OP | 0.236 | 1 | 1.825 | 0.189 |
| MCYE | 0.604 | 1 | 4.667 | 0.041 |
| turb | 0.351 | 1 | 2.713 | 0.113 |
| Medium\_impervious | 0.207 | 1 | 1.601 | 0.218 |
| Residuals | 3.106 | 24 | NA | NA |

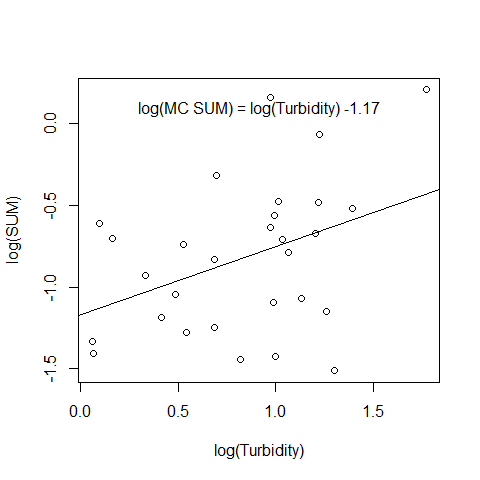
Our last four variables are ortho-P, *mcyE*, turbidity, and area of medium impervious surfaces.

Medium impervious surface gets dropped

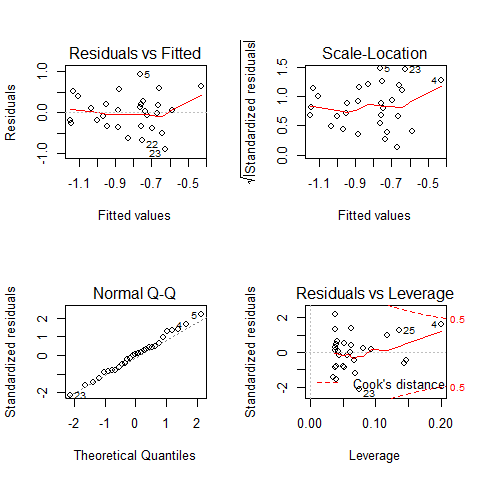
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| OP | 0.606 | 1 | 4.574 | 0.042 |
| MCYE | 1.190 | 1 | 8.977 | 0.006 |
| turb | 0.259 | 1 | 1.957 | 0.174 |
| Residuals | 3.313 | 25 | NA | NA |

Turbidity gets dropped

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| OP | 1.109 | 1 | 8.073 | 0.009 |
| MCYE | 1.287 | 1 | 9.369 | 0.005 |
| Residuals | 3.573 | 26 | NA | NA |



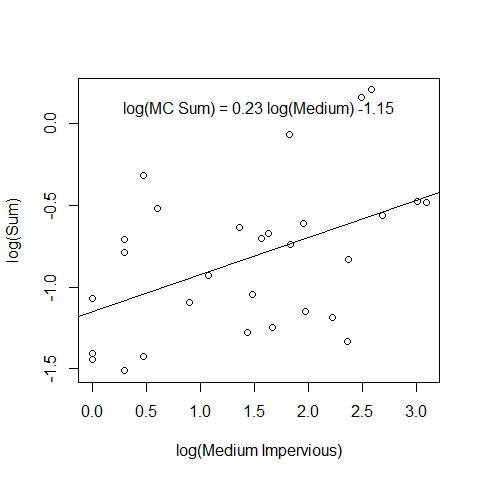
log(MC Sum) ~ log(Turbidity)



Residual Analysis of log(SUM) ~ log(Turbidity)

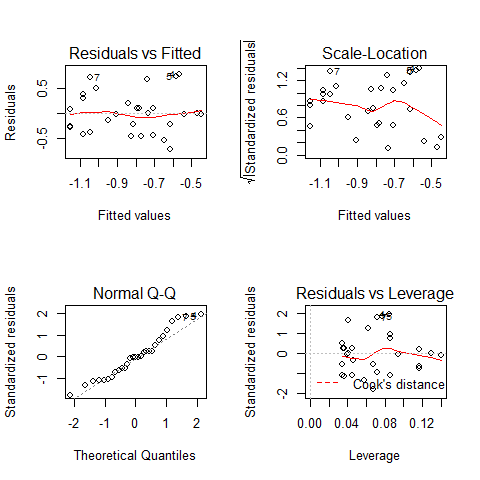
**The residuals is fairly normal. There is one data point that do seem to pull the correlation upward.**

log(MC Sum) ~ Area of Impervious surfaces



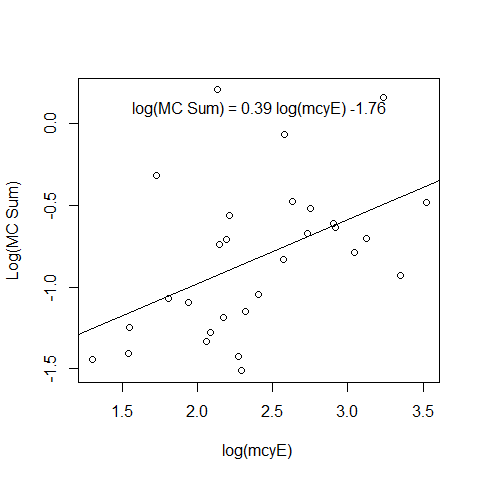
MC Sum as a Function of Impervious Surfaces

`



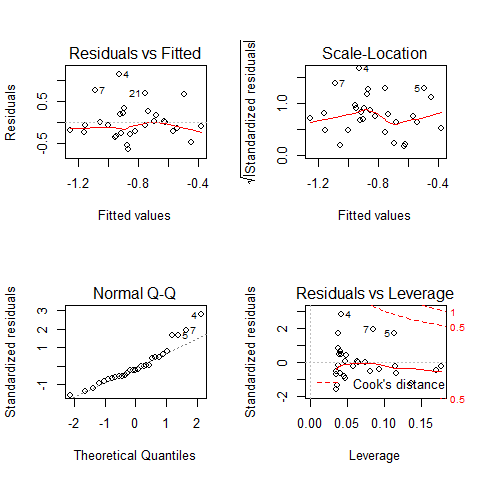
Residual Analysis of SUM ~ Medium Impervious surface

**Residuals are normally scattered for area of impervious surfaces. No leverage points.**



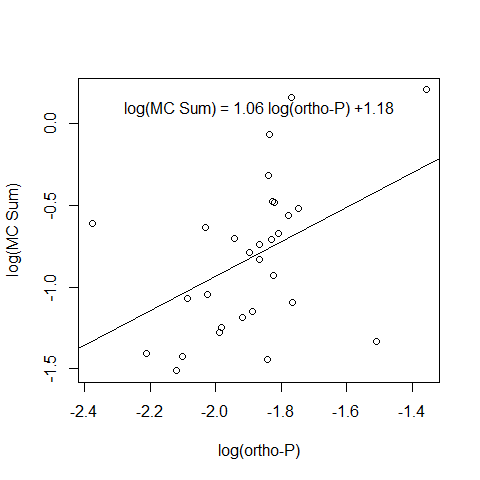
MC Sum as a function of mcyE

`



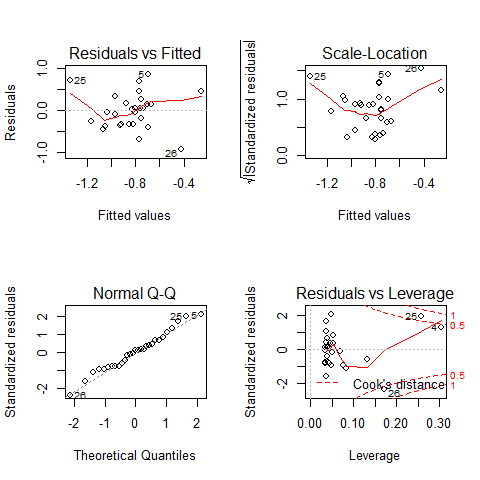
Residual Analysis of SUM ~ mcyE

**Residualis normal and seems fairly random. Variation is not uniformly scattered at the low and high end.**



MC Sum as a function of ortho-P

`

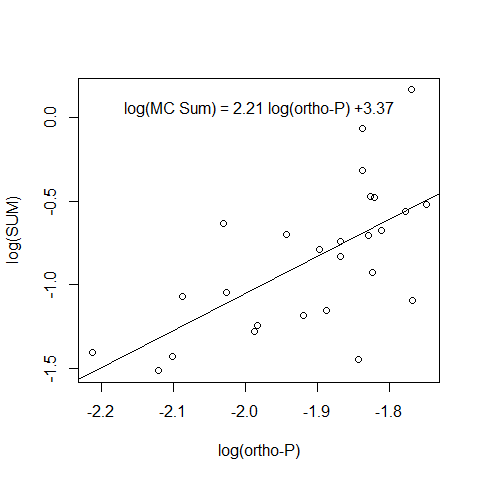


Residual Analysis of SUM~ortho-P

**ortho-P is driven by three signifigant outliers.**

*Outliers are ommitted*

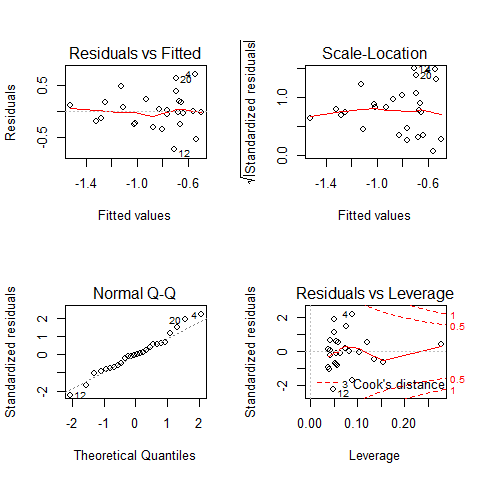
log(MC Sum) ~ log(ortho-P)



MC Sum as a function of ortho-P with omitted outliers

`

##   
## Call:  
## lm(formula = SUM ~ OP, data = omitted, na.action = na.omit)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.73774 -0.22009 -0.01207 0.17599 0.70645   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.373 1.044 3.230 0.003571 \*\*   
## OP 2.213 0.546 4.053 0.000461 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3394 on 24 degrees of freedom  
## Multiple R-squared: 0.4064, Adjusted R-squared: 0.3816   
## F-statistic: 16.43 on 1 and 24 DF, p-value: 0.0004606



Residual Analysis of MC Sum ~ ortho-P with outliers removed

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| OP | 1.8926 | 1 | 16.4286 | 5e-04 |
| Residuals | 2.7648 | 24 | NA | NA |

**With Brighton, Silver, and Stony creek ommited, we see ortho-P signifigant effect on MC sum. However, the residual does not have a uniform variation.**

Lets assess each variable and see which one is the most signifigant by value.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| MCYE | 1.3621 | 1 | 7.855 | 0.0093 |
| Residuals | 4.6820 | 27 | NA | NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| OP | 1.184 | 1 | 6.577 | 0.016 |
| Residuals | 4.860 | 27 | NA | NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| turb | 0.939 | 1 | 4.966 | 0.034 |
| Residuals | 5.105 | 27 | NA | NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| Medium\_impervious | 1.326 | 1 | 7.587 | 0.01 |
| Residuals | 4.718 | 27 | NA | NA |

From here, our starting variable for a forward selection is *mcyE*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sum Sq Df | F value | Pr | (>F) |  |
| MCYE | 1.287 | 1 | 9.369 | 0.005 |
| OP | 1.109 | 1 | 8.073 | 0.009 |
| Residuals | 3.573 | 26 | NA | NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| MCYE | 1.186 | 1 | 7.865 | 0.009 |
| turb | 0.763 | 1 | 5.058 | 0.033 |
| Residuals | 3.919 | 26 | NA | NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| MCYE | 0.593 | 1 | 3.735 | 0.064 |
| Medium\_impervious | 0.556 | 1 | 3.507 | 0.072 |
| Residuals | 4.126 | 26 | NA | NA |

ortho-P is the second most signifigant

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sum Sq Df F valu | e Pr(>F | ) |  |  |
| MCYE | 0.782 | 1 | 5.653 | 0.025 |
| OP | 0.668 | 1 | 4.833 | 0.037 |
| Medium\_impervious | 0.115 | 1 | 0.835 | 0.370 |
| Residuals | 3.457 | 25 | NA | NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| MCYE | 1.190 | 1 | 8.977 | 0.006 |
| OP | 0.606 | 1 | 4.574 | 0.042 |
| turb | 0.259 | 1 | 1.957 | 0.174 |
| Residuals | 3.313 | 25 | NA | NA |

Adding a third variable does not improve our model. SUM~OP+MCYE is most likely the best model.

### Best Model

Each model is compared by BIC. All possible combination of models are listed here:

* FullModel = SUM ~ MCYE + turb + Medium\_impervious + OP
* Model 1 = SUM ~ turb + Medium\_impervious + OP
* Model 2 = SUM ~ MCYE + Medium\_impervious + turb
* Model 3 = SUM ~ turb + MCYE + OP
* Model 4 = SUM ~ Medium\_impervious + MCYE + OP
* Model 5 = SUM ~ MCYE + OP
* Model 6 = SUM ~ MCYE + turb
* Model 7 = SUM ~ MCYE + Medium\_impervious
* Model 8 = SUM ~ Medium\_impervious + OP
* Model 9 = SUM ~ Medium\_impervious + turb
* Model 10 = SUM ~ Medium\_impervious
* Model 11 = SUM ~ OP
* Model 12 = SUM ~ turb
* Model 13 = SUM ~ MCYE
* Model 14 = SUM ~ OP + turb

|  |  |  |
| --- | --- | --- |
| * df | * B | * IC |
| * FullModel | * 6 | * 37.72 |
| * Model1 | * 5 | * 39.50 |
| * Model2 | * 5 | * 36.48 |
| * Model3 | * 5 | * 36.22 |
| * Model4 | * 5 | * 37.46 |
| * Model5 | * 4 | * 35.04 |
| * Model6 | * 4 | * 37.73 |
| * Model7 | * 4 | * 39.21 |
| * Model8 | * 4 | * 40.00 |
| * Model9 | * 4 | * 37.00 |
| * Model10 | * 3 | * 39.74 |
| * Model11 | * 3 | * 40.60 |
| * Model12 | * 3 | * 42.03 |
| * Model13 | * 3 | * 39.52 |
| * Model14 | * 4 | * 41.75 |

* SUM ~ *mcyE* + ortho-P is suggestably the best model.

##   
## Call:  
## lm(formula = SUM ~ MCYE + OP, data = MCSUMresponseAverage)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.77180 -0.23888 -0.06047 0.17799 0.70405   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.1971 0.7555 0.261 0.79623   
## MCYE 0.3819 0.1248 3.061 0.00507 \*\*  
## OP 1.0232 0.3601 2.841 0.00862 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3707 on 26 degrees of freedom  
## Multiple R-squared: 0.4089, Adjusted R-squared: 0.3634   
## F-statistic: 8.993 on 2 and 26 DF, p-value: 0.001076

* SUM ~ turbidity + ortho-P + *mcyE* is second best

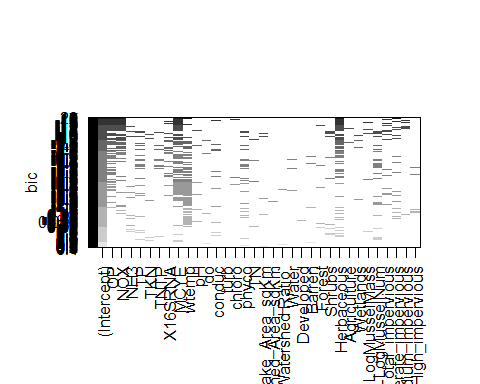
##   
## Call:  
## lm(formula = SUM ~ MCYE + turb + MCYE, data = MCSUMresponseAverage)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.82980 -0.21283 -0.05648 0.19541 0.80098   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.0210 0.3410 -5.926 2.98e-06 \*\*\*  
## MCYE 0.3677 0.1311 2.804 0.00941 \*\*   
## turb 0.3806 0.1692 2.249 0.03320 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3883 on 26 degrees of freedom  
## Multiple R-squared: 0.3515, Adjusted R-squared: 0.3016   
## F-statistic: 7.047 on 2 and 26 DF, p-value: 0.003586

* SUM ~ *mcyE* + Medium Impervious Area + turbidity is our third best.

##   
## Call:  
## lm(formula = SUM ~ MCYE + turb + Medium\_impervious, data = MCSUMresponseAverage)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.65566 -0.18601 -0.03566 0.17403 0.88606   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.99387 0.32144 -6.203 1.73e-06 \*\*\*  
## MCYE 0.25539 0.13477 1.895 0.0697 .   
## turb 0.38582 0.15940 2.420 0.0231 \*   
## Medium\_impervious 0.16446 0.07916 2.078 0.0482 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3656 on 25 degrees of freedom  
## Multiple R-squared: 0.447, Adjusted R-squared: 0.3806   
## F-statistic: 6.736 on 3 and 25 DF, p-value: 0.001744

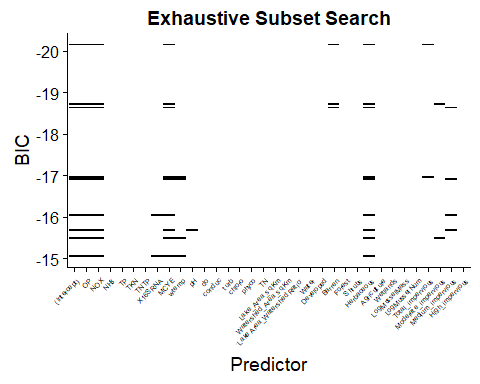
These models will be validated later on with K-fold cross validation.

## MC Sum from SPATTs as Response Variable



Subset Analysis for Predicting MC from SPATT: Exhaustive Search

The notable variables frequently chosen are ortho-P, nitrate+nitrite-N, *16s rRNA*, *mcyE*, water temprature, herbaceous, mussel mass, and area of medium impervious surface.



Subset Analysis for Predicting MC from SPATT:Top 15 Models Exhaustive Search

The best subset model is with variables including ortho-P, nitrate+nitrite-N, *mcyE*, barren, herbaceous, and impervious surfaces. They are selected as candidate variables.

`

model1 <- lm(S\_SUM~ #Starting out all variables, then backward selection  
 OP +  
 NOX +  
 X16SRNA +  
 MCYE +  
 wtemp +  
 Herbaceous +  
 LogMusselMass +  
 LogMusselNum +  
 Medium\_impervious ,  
 data=SPATTSresponseAverage)  
kable(Anova(model1), digits = 3)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| OP | 0.512 | 1 | 1.763 | 0.200 |
| NOX | 0.004 | 1 | 0.015 | 0.905 |
| X16SRNA | 0.433 | 1 | 1.490 | 0.237 |
| MCYE | 1.029 | 1 | 3.545 | 0.075 |
| wtemp | 0.121 | 1 | 0.417 | 0.526 |
| Herbaceous | 0.539 | 1 | 1.858 | 0.189 |
| LogMusselMass | 0.003 | 1 | 0.011 | 0.919 |
| LogMusselNum | 0.068 | 1 | 0.233 | 0.635 |
| Medium\_impervious | 0.005 | 1 | 0.018 | 0.893 |
| Residuals | 5.516 | 19 | NA | NA |

One variables with > 0.05 will be dropped and reanalyzed with ANOVA. This is repeated until the last few variables.

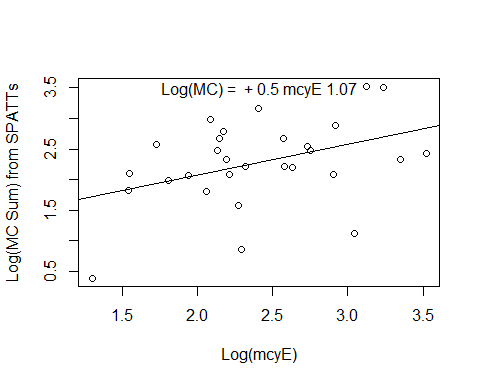
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| OP | 0.765 | 1 | 3.087 | 0.092 |
| X16SRNA | 0.550 | 1 | 2.222 | 0.149 |
| MCYE | 1.928 | 1 | 7.782 | 0.010 |
| Herbaceous | 5.291 | 1 | 21.356 | 0.000 |
| Residuals | 5.946 | 24 | NA | NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| OP | 0.987 | 1 | 3.797 | 0.063 |
| MCYE | 2.018 | 1 | 7.764 | 0.010 |
| Herbaceous | 4.751 | 1 | 18.282 | 0.000 |
| Residuals | 6.497 | 25 | NA | NA |

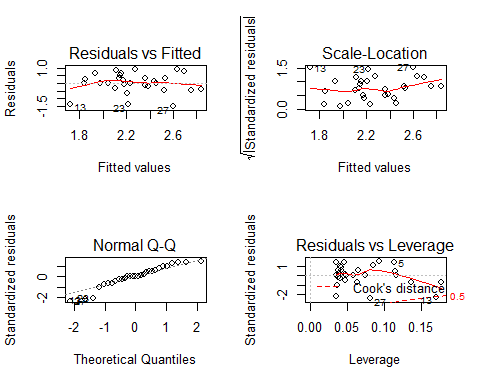
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| MCYE | 2.125 | 1 | 7.384 | 0.012 |
| Herbaceous | 3.938 | 1 | 13.683 | 0.001 |
| Residuals | 7.483 | 26 | NA | NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sum Sq Df | F value | Pr( | >F) |  |
| Herbaceous | 4.048 | 1 | 11.374 | 0.002 |
| Residuals | 9.609 | 27 | NA | NA |

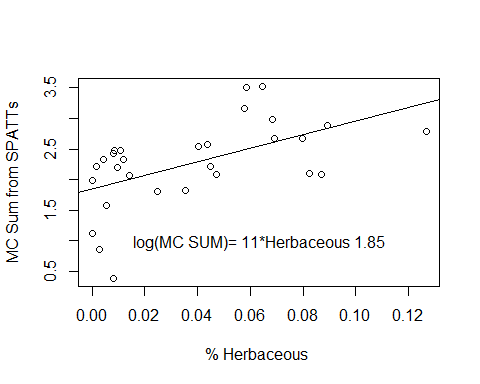
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Sum Sq | Df | F value | Pr(>F) |
| MCYE | 2.235 | 1 | 5.282 | 0.03 |
| Residuals | 11.422 | 27 | NA | NA |

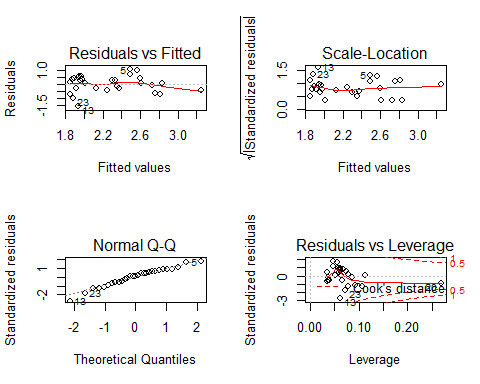


MC from Spatts ~ mcyE



Residual Analysis of MC Sum ~ mcyE





Residual analysis of MC Sum ~ Herbaceous

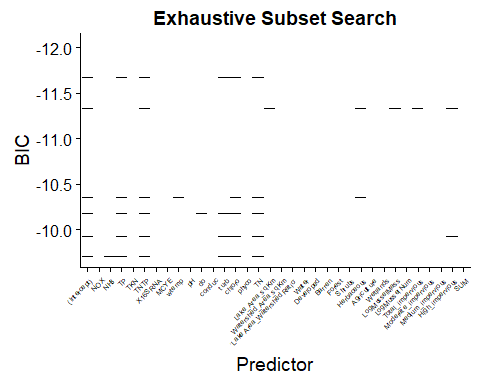
### Best Model

* FullModel = S\_SUM ~ MCYE + Herbaceous
* Model 1 = S\_SUM ~ MCYE
* Model 2 = S\_SUM ~ Herbaceous
* Model 3 = S\_SUM ~ Agriculture
* Model 4 = S\_SUM ~ Forest
* Model 5 = S\_SUM ~ Shrubs

|  |  |  |
| --- | --- | --- |
| * df | * B | * IC |
| * FullModel | * 4 | * 56.48 |
| * Model1 | * 3 | * 65.38 |
| * Model2 | * 3 | * 60.37 |
| * Model3 | * 3 | * 68.77 |
| * Model4 | * 3 | * 68.09 |
| * Model5 | * 3 | * 66.52 |

The best predictor model for MC from SPATTs is *mcyE* and Herbaceous land use.

## mcyE gene as Response



Subset Analysis for Predicting mcyE:Top 15 Models Exhaustive Search

* Here we can select total phosphorus, TN:TP ratio, turbidity, chloraphyl-a, phyocyanin, herbaceous, and area of impervious surfaces.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * Sum | * Sq Df | * F v | * alue Pr( | * >F) |
| * NOX | * 0.134 | * 1 | * 1.034 | * 0.325 |
| * wtemp | * 1.239 | * 1 | * 9.595 | * 0.007 |
| * do | * 1.465 | * 1 | * 11.342 | * 0.004 |
| * conduc | * 1.554 | * 1 | * 12.030 | * 0.003 |
| * Herbaceous | * 0.085 | * 1 | * 0.656 | * 0.431 |
| * LogMusselNum | * 1.606 | * 1 | * 12.433 | * 0.003 |
| * High\_impervious | * 1.707 | * 1 | * 13.215 | * 0.002 |
| * Total\_impervious | * 0.223 | * 1 | * 1.728 | * 0.208 |
| * Medium\_impervious | * 0.164 | * 1 | * 1.273 | * 0.277 |
| * Residuals | * 1.937 | * 15 | * NA | * NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * Sum | * Sq Df | * F v | * alue Pr( | * >F) |
| * wtemp | * 0.259 | * 1 | * 0.919 | * 0.347 |
| * LogMusselNum | * 0.236 | * 1 | * 0.838 | * 0.369 |
| * Medium\_impervious | * 1.093 | * 1 | * 3.888 | * 0.060 |
| * Residuals | * 7.032 | * 25 | * NA | * NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * Sum | * Sq Df | * F v | * alue Pr( | * >F) |
| * LogMusselNum | * 0.122 | * 1 | * 0.437 | * 0.514 |
| * Medium\_impervious | * 1.284 | * 1 | * 4.580 | * 0.042 |
| * Residuals | * 7.290 | * 26 | * NA | * NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * Sum | * Sq Df | * F v | * alue Pr( | * >F) |
| * Medium\_impervious | * 1.422 | * 1 | * 5.178 | * 0.031 |
| * Residuals | * 7.413 | * 27 | * NA | * NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * Sum | * Sq Df | * F v | * alue Pr( | * >F) |
| * LogMusselNum | * 0.260 | * 1 | * 0.818 | * 0.374 |
| * Residuals | * 8.574 | * 27 | * NA | * NA |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| * Sum | * Sq Df | * F v | * alue Pr( | * >F) |
| * wtemp | * 0.252 | * 1 | * 0.791 | * 0.382 |
| * Residuals | * 8.583 | * 27 | * NA | * NA |

The only model best to predict *mcyE* is area of medium impervious surfaces. Water temprature and mussel data did not have signifigant relationship with MC Sum.

# K-fold Cross-validation

Here we will verify each model by K-fold cross-validation. First, the full dataset will be split into two. Roughly 66% of the full set is for training and the remaining 33% for testing.

On our training set, it is further divided into “K” folds. One fold is held away for validation, while the remaining are used to train the model. Next the trained model is validated against the validation fold. Each validation, a skill score is recorded. Here the RMSE is assessed along with its standard deviation. We can compare the RMSE between our models. The lowest RMSE is the best model.

Ultimately, the trained model will further be tested again against the 33%. If the prediction is signifigant with the actual value of the training dataset it is deemed successful model.

Using the library(caret) package.

Here each model that was found to be effective from previous section is tested. If the trained model fitted against the untrained dataset has a signifigant slope, then the model does have potential predictablility power, proving to be robust.

A 10 fold cross-validation with 20 repeated cycles will be done on a training dataset. K=10

## Successful Models

### SUM ~ MCYE + OP + Medium\_impervious

Model Validation Performance:

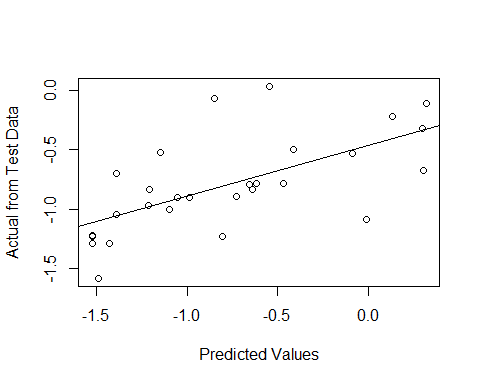
Our model: SUM ~ MCYE + OP + Medium\_impervious

Across all 10-fold cross-validation, the average of the root mean square deviation:

(RMSE) = 0.57 with = 0.19.

Average = 0.35 with = 0.28

Evaluate the trained model and further validate this model on 33% of rest of the data.



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -0.019 | 0.199 | -0.095 | 0.925 |
| Prediction | 0.952 | 0.224 | 4.253 | 0.000 |

### SUM ~ MCYE

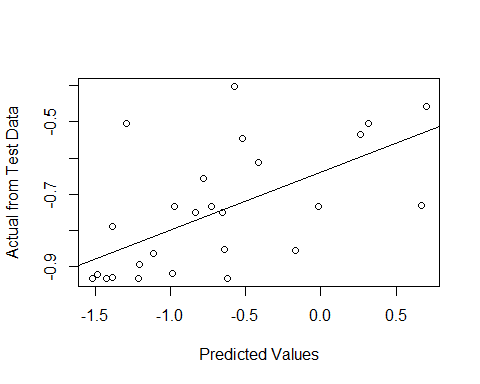
Model Validation Performance:

Our model: SUM ~ MCYE

(RMSE) = 0.59 with = 0.16.

Average = 0.25 with = 0.26

Final validation



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 1.083 | 0.466 | 2.324 | 0.028 |
| Prediction | 2.416 | 0.599 | 4.032 | 0.000 |

### SUM ~ OP + Turbidity

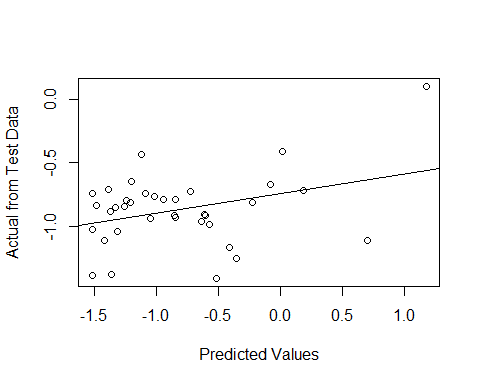
Model Validation Performance:

Our model: SUM ~ OP + turb

(RMSE) = 0.54 with = 0.12.

Average = 0.24 with = 0.21

Final validation



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -0.143 | 0.33 | -0.433 | 0.668 |
| Prediction | 0.783 | 0.36 | 2.173 | 0.037 |

### SUM ~ OP

## Linear Regression   
##   
## 113 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 101, 103, 101, 102, 101, 102, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.5777739 0.1807362 0.4746605  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Model Validation Performance:

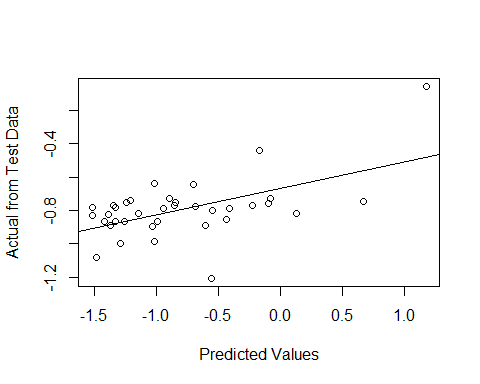
## Linear Regression   
##   
## 113 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 101, 103, 101, 102, 101, 102, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.5777739 0.1807362 0.4746605  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Our model: SUM ~ OP

(RMSE) = 0.58 with = 0.11.

Average = 0.18 with = 0.19

Final validation



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 0.730 | 0.406 | 1.796 | 0.081 |
| Prediction | 1.938 | 0.500 | 3.878 | 0.000 |

### SPATTS SUM ~ MCYE + Herbaceous

Model Validation Performance:

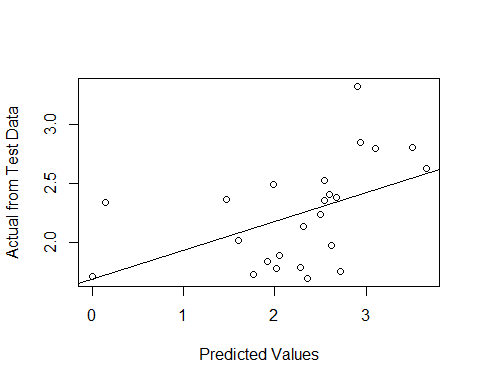
## Linear Regression   
##   
## 55 samples  
## 2 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 49, 51, 49, 50, 50, 49, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.6846797 0.4042762 0.5672053  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Our model: S\_SUM ~ MCYE + Herbaceous

(RMSE) = 0.68 with = 0.22.

Average = 0.4 with = 0.26

Final validation



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 0.142 | 0.837 | 0.170 | 0.867 |
| Prediction | 0.945 | 0.367 | 2.575 | 0.017 |

### SPATTS SUM ~ Herbaceous

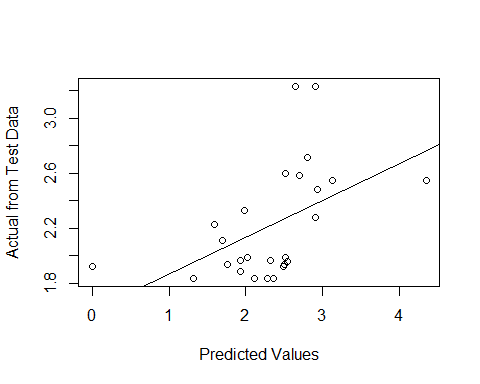
Model Validation Performance:

Our model: S\_SUM ~ Herbaceous

(RMSE) = 0.73 with = 0.24.

Average = 0.34 with = 0.25

Final validation



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 0.248 | 0.745 | 0.333 | 0.742 |
| Prediction | 0.931 | 0.330 | 2.820 | 0.009 |

## Failed Models

### SUM ~ Medium\_impervious

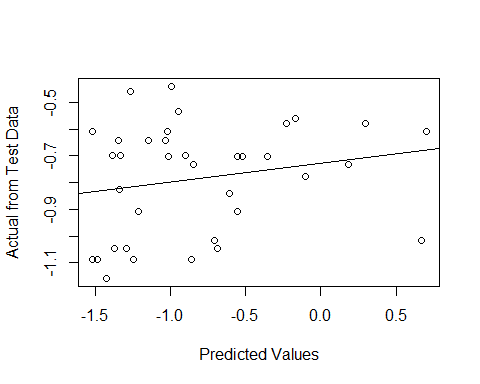
Model Validation Performance:

Our model: SUM ~ Medium\_impervious

(RMSE) = 0.55 with = 0.1.

Average = 0.2 with = 0.18

Final validation



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -0.329 | 0.399 | -0.824 | 0.416 |
| Prediction | 0.613 | 0.493 | 1.244 | 0.222 |

### SUM ~ Turbidity

Model Validation Performance:

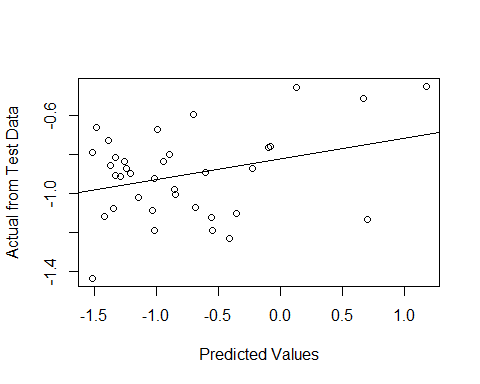
## Linear Regression   
##   
## 77 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 69, 69, 70, 70, 69, 69, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.5265933 0.2144668 0.4372665  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Our model: SUM ~ turb

(RMSE) = 0.53 with = 0.1.

Average = 0.21 with = 0.21

Final validation



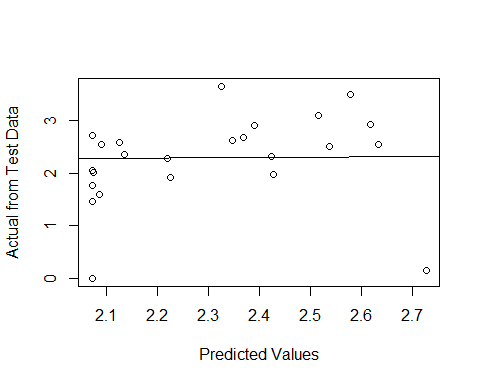
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 0.080 | 0.451 | 0.176 | 0.861 |
| Prediction | 0.952 | 0.484 | 1.966 | 0.058 |

### SPATTS SUM ~ MCYE

## Linear Regression   
##   
## 79 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 71, 72, 71, 71, 71, 71, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 0.774129 0.2286159 0.5840406  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Our model: S\_SUM ~ MCYE

Across all 10-fold cross-validation, the average of the root mean square deviation (RMSE) = 0.77 = 0.24 with an average of = 0.23 with = 0.2



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 0.270 | 1.905 | 0.142 | 0.888 |
| Prediction | 0.864 | 0.825 | 1.048 | 0.306 |

### MCYE ~ Medium\_impervious

## Linear Regression   
##   
## 91 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 81, 82, 82, 81, 83, 82, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 1.051635 0.1865997 0.8855092  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Model Validation Performance:

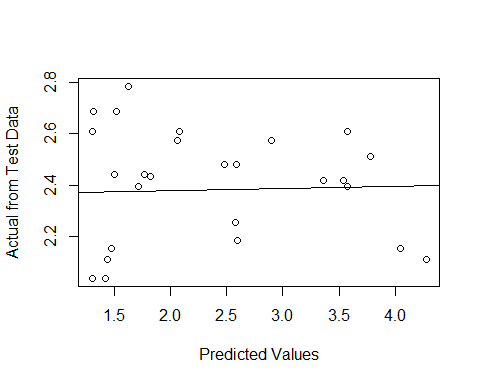
## Linear Regression   
##   
## 91 samples  
## 1 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 20 times)   
## Summary of sample sizes: 81, 82, 82, 81, 83, 82, ...   
## Resampling results:  
##   
## RMSE Rsquared MAE   
## 1.051635 0.1865997 0.8855092  
##   
## Tuning parameter 'intercept' was held constant at a value of TRUE

Our model: MCYE ~ Medium\_impervious

(RMSE) = 1.05 with = 0.2.

Average = 0.19 with = 0.2

Final validation



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | 1.915 | 1.989 | 0.963 | 0.345 |
| Prediction | 0.159 | 0.832 | 0.191 | 0.850 |

# Conclusion

So far, we found three predictor models for MC Sum which has a signifigant predictive power. Observing the K-fold validation resampling, the best model with the lowest RMSE is the model including ortho-P and turbidity with RMSE=0.54 and a = 0.12. The second best model had *mcyE*, ortho-P, and area of medium impervious surfaces included with RMSE=0.57. Although this model does have a larger variation of RMSE with = 0.19. With ortho-P alone as a predictor preformed well with testing against the untrained dataset. In the cross-validation, the RMSE=0.57 with = 0.18. MC from SPATTs are best predicted with a model including *mcyE* and herbaceous. Cross-validation average RMSE=0.68. This was the lowest RMSE compared to models with only one of the two as predictors.

# R Packages

#install.packages("tidyverse")  
#install.packages("lme4")  
#install.packages("leaps")  
#install.packages("caret")  
  
  
  
library(lme4)  
library(leaps)  
library(leaps)  
library(lavaan)  
library(car)  
library(caret)  
library(reshape2)