ENV 710 Final Project

Emma Kaufman, Cara Kuuskvere, Jenn McNeill

2024-04-16

Contents

Introduction

In this report we examine the relationship of total phosphorus concentration on dissolved oxygen levels in Wisconsin lakes. Adequate levels of dissolved oxygen in lakes are crucial for the aquatic species that reside there. DO levels are at risk for becoming dangerously low when excess organic materials, such as algae, decompose, a process called eutrophication (US EPA, 2013). As a result, it is important to limit algae growth in lakes. When there is too much phosphorus in lakes conditions become suitable for algae blooms resulting in eutrophication. In addition, toxins found in algae blooms are harmful to human and animal health (US EPA, 2013).

Total phosphorus in lakes is influenced by rainfall, stream flow, overland runoff, groundwater discharge, decomposition of organic matter, and soil erosion. Humans contribute to phosphorus levels in lakes due to fertilization of agricultural fields and lawns, leaking waste water, and increased urbanization (Lee, 1973). Most phosphorus in lakes takes the form of dissolved or particulate phosphorus, and total phosphorus is a measure of all forms of phosphorus potentially available to algae (Total Nitrogen/Total Phosphorus and Chlorophyll a, n.d.).

Our research questions and associated hypotheses are three-fold:

- Is there a difference in the impact of total phosphorus on dissolved oxygen levels in the spring versus the summer?
  - H_0: There is no difference in the impact of phosphorus on DO in the spring versus the summer.
  - H_a: There is a difference in the impact of phosphorus on DO in spring versus summer.
- Is there a difference in the impact of total phosphorus on dissolved oxygen levels in shallow versus subsurface lake water?
  - H_0: There is no difference in the impact of phosphorus on DO in surface and subsurface lake water.
  - H_a: There is a difference in the impact of phosphorus on DO in surface and subsurface lake water.
- Do West Long, East Long, Peter, and Paul Lake all exhibit the same relationship of how total phosphorus impacts dissolved oxygen levels?
  - H_0: There is no difference in the impact of phosphorus on DO when accounting for different lakes.
  - H_a: There is a difference in the impact of phosphorus on DO when accounting for different lakes.

Methods

The data for this project are from lakes in the North Temperate Lakes District of Wisconsin. They were collected at the Long Term Ecological Research station located there. The physical and chemical variables were measured at a central station in the deepest point of each lake. Most measurements were made between 8 and 9 am. The nutrient measurements were made along vertical profiles at varied depth intervals. The dissolved oxygen chemistry data from 1991-1999 were obtained from a Lachat auto-analyser. The methods of data collection from 1991-1997 are described in more detail by Carpenter at al. (2001). More information on this data can be found here: https://lter.limnology.wisc.edu/core-datasets/.

These datasets contain daily nutrient and lake chemistry data at 26 different lakes within the North Temperate Lakes District in Wisconsin over 25 years. The variables between the two datasets include the lake depth, total nitrogen, total phosphorus, nh34, no23, and po4 concentrations, as well as lake temperature, dissolved oxygen, and irradiance. For the purposes of this project we focus on using lake depth, total phosphorus, and dissolved oxygen data across 4 lakes: East Long Lake, Paul Lake, Peter Lake, and West Long Lake from 1994 to 1999. A table of these variables is found below:

Table 1: Variable descriptions

Variable Analyzed	Unit of Measure	Type of Variable
Lake Name	Name Factor	Character/ Factor
Depth	Meters below surface	Numeric ; binary indicator
Season	Spring (May)/ Summer (August)	Binary indicator
Dissolved Oxygen (DO)	mg/ Liter	Numeric
Phosphorus	μg	Numeric

Our dependent variable is dissolved oxygen (mg/L) and our independent variables are total phosphorus (μg), season (May or August, coded as spring or summer), and depth (<6 ft as surface, deeper than 6 ft as subsurface).

Our initial data exploration of the range and measures of central tendency of each of our numeric variables is as follows:

Table 2: Measures of central tendency

Variable	Mean	Median	Range
Total Phosphorus (μg)	34.91	19.47	-3.04 - 284.03
Depth (Meters below surface)	3.67	2	0 - 12
Dissolved Oxygen (mg/Liter)	6.76	7.8	0 - 20

In this report we use linear models to estimate dissolved oxygen across these lakes based on total phosphorus content sampled within the lakes' water columns. We examine the difference in the relationship of DO and total phosphorus in spring and summer, as well as the difference in the relationship of DO and total phosphorus at the water surface versus deeper in the water column. Finally, we use a hierarchical linear model to examine the DO and total phosphorus relationship across East Long, West Long, Peter, and Paul lakes.

In order to check the assumptions of these models, we plot the residuals against a normal distribution (qq plot) as well as the residuals vs fitted values to check for constant variance. We also fit training models with 80% of the data and checked our root mean squared error (RMSE) values to see the accuracy of our model fit versus our observed data.

Results

MODEL FIT ONE: Dissolved Oxygen (DO) by Total Phosphorus (TP) with a Seasonal Binary Indicator Variable and a Seasonal Interaction Term

Table 3: Model 1 coefficients

Variable	Value	Interpretation
β_0 (intercept)	10.15 mg/L	Expected value of DO in spring with 0 TP

Variable	Value	Interpretation
β_1	-2.36 mg/L	Difference in DO between spring and summer with 0 TP
β_2	-0.067 (mg/L)/μg	Expected change in DO for each unit increase in TP in the spring
β_3	0.02 (mg/L)/μg	Adjustment to slope for interaction for each unit increase in TP during the summer

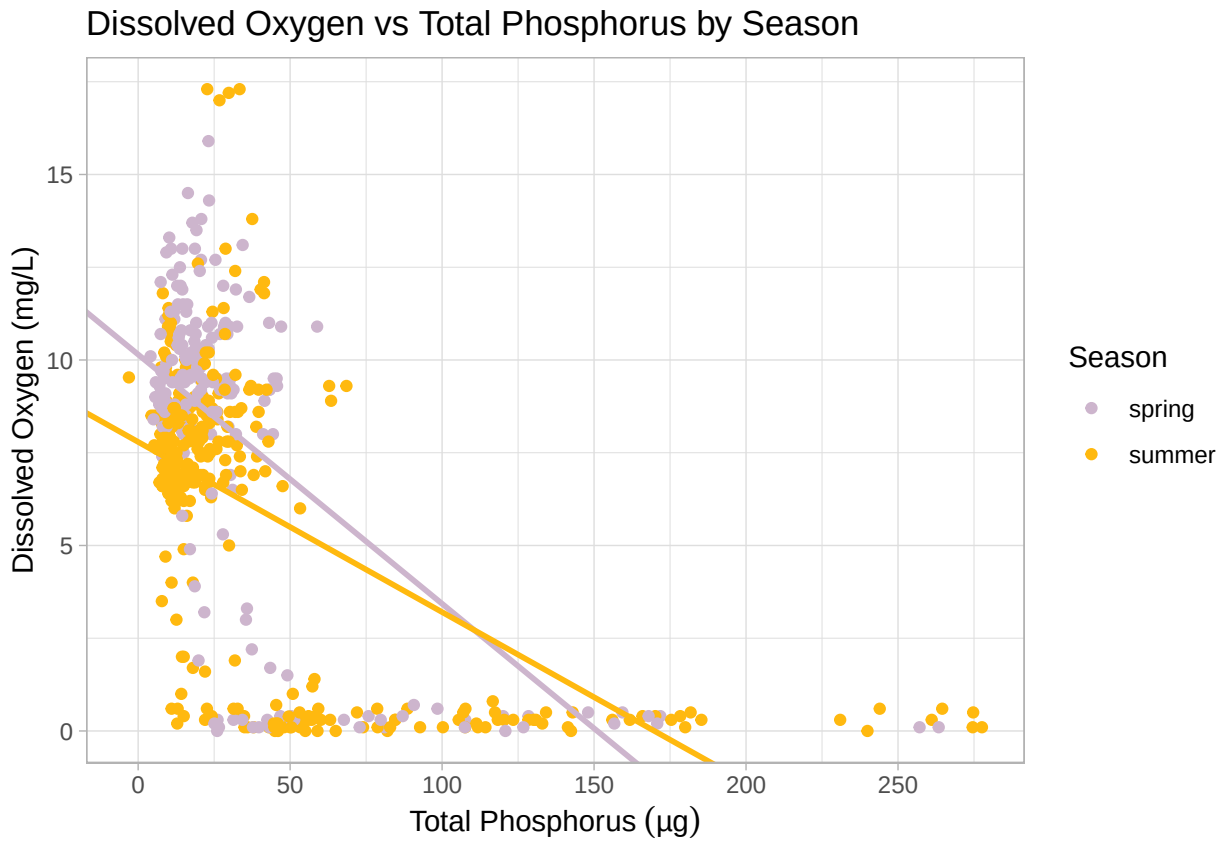


Figure 1: Dissolved Oxygen vs Total Phosphorus by Season

Results Model One

The results of this model are that the expected value of dissolved oxygen across all of the lakes with no phosphorus is 10.15 mg/L in the spring. In the summer, dissolved oxygen levels when there is no phosphorus decrease by 2.36 mg/L.

The expected change in DO for each unit increase in total phosphorus during the spring is -0.067 (mg/L)/μg . In the summer this change increases by 0.02.

These model results are displayed with the observed data in Figure 1. This linear model assumes that the residuals are normally distributed and have constant variance. The plots that examine these assumptions are found in the appendix plot 1. They show that these assumptions were not met, as the residuals do not

follow a normal distribution (as seen in the qq plot), and the variance is not constant (in the residuals v fitted plot).

We tried to improve our model by normalizing the data with a log transformation of total phosphorus. The results of this are in the appendix (Appendix plot log-transform). We still did not see normal distribution of residuals or constant variance of residuals. Additionally, we tried to use a subset of the DO data and fit our model to those data (Appendix plot subset DO), but still did not have promising results for our model assumptions. We chose to go forward with the original data because it is the most interpretable.

These results show that with increasing concentrations of phosphorus, there is a decrease in dissolved oxygen, and that the impact of phosphorus on DO levels is slightly stronger in the spring. These results are significant (all of the p-values are less than 0.05 for each coefficient), but that doesn't mean they are the best model of these data. The RMSE score of 2.99 shows that there is high error when you compare the fitted and observed data.

MODEL FIT TWO: Dissolved Oxygen (DO) by Total Phosphorus (TP) with a Water Depth Binary Indicator Variable and a Water Depth Interaction Term

Table 4: Model 2 coefficients

Variable	Value	Interpretation
β_0 (Intercept)	8.52 mg/L	Expected value of DO in surface water with 0 TP
β_1	-7.67 mg/L	Difference in DO between subsurface and surface water with 0 TP
β_2	-0.0002 mg/L	Expected change in DO for each unit increase in TP in surface water
β_3	-0.003 mg/L	Adjustment to slope for interaction term between TP in subsurface water

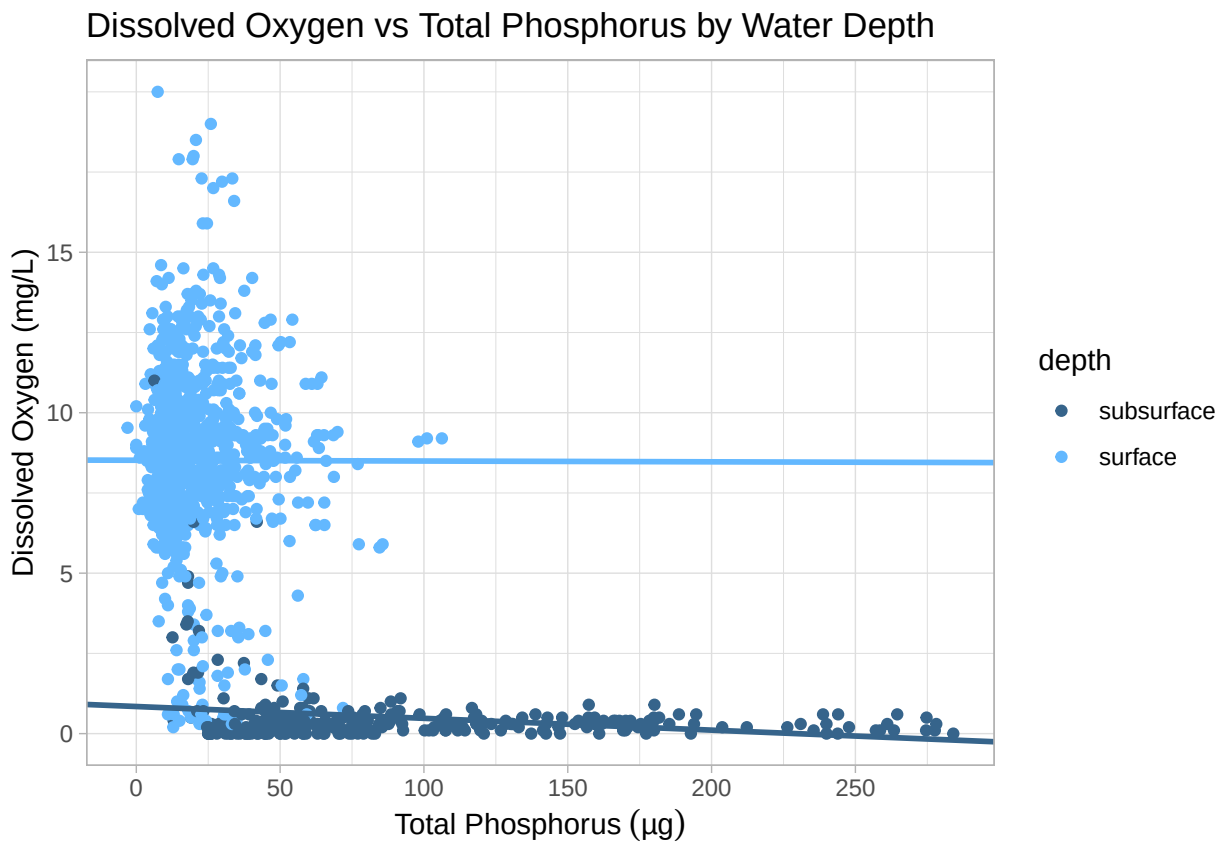


Figure 2: Dissolved Oxygen vs Total Phosphorus by Water Depth

Results Model Two

The results from this model tell us that the expected value for dissolved oxygen in surface waters with zero total phosphorus is 8.52 mg/L. In subsurface waters with zero total phosphorus, the dissolved oxygen level decreases by 7.67 mg/L. The rate of change of the dissolved oxygen concentration for each unit increase of micrograms of total phosphorus in the water is -0.0002 in surface waters and -0.003 for subsurface waters.

Figure 2 shows a graphical representation of this linear model. These data are consistent with our assumptions given what we know about the relationship between dissolved oxygen and total phosphorus. As total phosphorus levels increase, the phosphorus in the water stimulates the growth of algae, a process known as eutrophication. This algae consumes the oxygen in the water as it decomposes, and the level of dissolved oxygen decreases. Simultaneously, the algae growth in the water halts the production of oxygen because the water gets murkier and plants are not able to photosynthesize as efficiently. Knowing that this process occurs, it is logical that the levels of dissolved oxygen are lower in deeper waters because the sun penetration is weaker and plants are not photosynthesizing enough to overcome the depletion of oxygen caused by eutrophication. It follows that the downward slope of the line is also logical because the dissolved oxygen is decreasing as total phosphorus increases.

Although the results of this linear model are consistent with our predictions, the model still has a high degree of error. The QQ plot shown in Appendix Plot 2 does not form a straight line, which means that the residuals from this model do not follow a normal distribution. The clouds of data points in Figure 2 are consistent with this observation; the points are not evenly distributed about the best fit line. Additionally, the RMSE score of 2.01 is an improvement as compared to the first model we fit but still shows a high degree of error.

MODEL FIT THREE: Dissolved Oxygen by Total Phosphorus with a Hierarchical Linear Mixed Effects Lake Variable

Table 5: Model 3 coefficients

Variable	Value	Interpretation
β_0 (Intercept)	8.68 mg/L	Expected value of DO with zero phosphorus
β_1 (tp_ug)	-0.55 (mg/L)/μg	Change in dissolved oxygen for each unit increase in TP

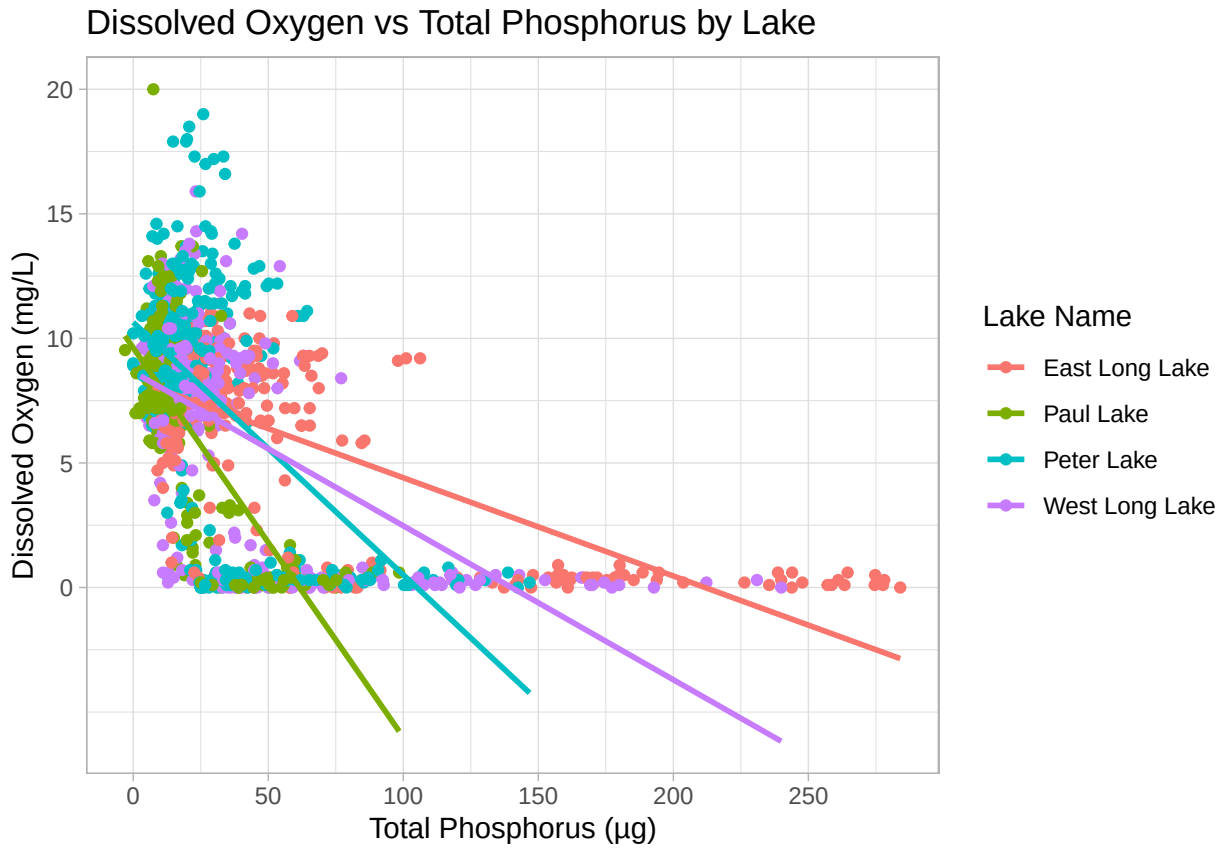


Figure 3: Dissolved Oxygen vs Total Phosphorus by Lake

Results Model Three

This model evaluates dissolved oxygen (dependent) based upon total phosphorus (independent) in each lake in our subset. The hierarchical model evaluates these two variables, with consideration that each lake may introduce differences in the relationship between the two, whose impacts should not be reflected in the pure randomness between the two variables but can be generalized between all lakes. Figure 3 shows a graphical representation of this model. Like the previous two models, linearity is assumed in the data, along with independence due to the nature of the data collection process. Normality of residuals and homoscedasticity

are assumptions made of successful models, however our diagnostic plots demonstrate that these assumptions were not met.

Based on the full dataset, the fixed estimate for the total phosphorus in micrograms is approximately -0.056, and for the training dataset it is approximately -0.054. Given the fixed estimate based on the full dataset, this suggests that for every one-unit increase in total phosphorus, dissolved oxygen decreases by approximately 0.056 units. The random effects, particularly the intercept variance at the lake level (0.6096), indicate that there are differences among lakes that contribute to the variation in dissolved oxygen beyond what can be explained by total phosphorus alone.

The model's assumptions regarding linearity and independence hold, as expected based on the nature of the hierarchical model. However, the diagnostic plots reveal departures from normality of residuals and homoscedasticity, suggesting that the model may not fully capture all nuances in the data. Despite these limitations, the model provides valuable insights into the relationship between total phosphorus and dissolved oxygen, accounting for lake-specific effects that might otherwise confound the analysis.

In terms of model performance, the Root Mean Squared Error (RMSE) score of 3.32 indicates a fairly high degree of error when comparing the fitted values to the observed data. This suggests that while the model may provides useful estimates, there is much room for improvement in capturing the variability in dissolved oxygen explained by total phosphorus across different lakes. However, the hypothesis test does indicate a statistically significant relationship between total phosporus and dissolved oxygen. Further refinements or alternative modeling approaches may be warranted to enhance predictive accuracy and better account for the complexities in the data.

Analysis and Conclusion

Model Performance: To assess the effectiveness of our models in predicting dissolved oxygen (DO) levels based on total phosphorus (TP) concentrations, we evaluated three different models: Depth Model, Seasonal Model, and Hierarchical Model by Lakename. Each model's performance was measured using the Root Mean Squared Error (RMSE) metric, providing insights into the accuracy of predictions compared to observed data.

Table 6: Model RMSE Comparission

Model	RMSE
Depth Model	2.42
Seasonal Model	3.02
Hierarchical Model by Lakename	3.31

The RMSE values indicate the percentage difference between predicted and observed DO values, with lower RMSE values indicating better model performance. Among the models evaluated, the Depth Model yielded the lowest RMSE, suggesting it provides the most accurate predictions compared to the other models evaluated

Residual Analysis and Model Assumptions

Despite the relatively lower RMSE of the Depth Model, all three models exhibited issues with the distribution of residuals and model assumptions. Residual analysis revealed non-normal distribution and heteroscedasticity, indicating that the models may not fully capture the variability in DO explained by TP concentrations.

Hypothesis Testing Results

Based on the hypothesis tests conducted for each model given the p-values of our statistical evaluations, all coefficients associated with TP were found to be statistically significant ($p < 0.05$). This indicates a significant relationship between TP concentrations and DO levels across the lakes studied.

Model	Hypothesis Test Outcome
Depth Model	p-value: < 2.2e-16; Reject Null
Seasonal Model	p-value: < 2.2e-16; Reject Null
Hierarchical Model by Lakename	p-value: < 2.2e-16; Reject Null

Interpretation and Insights

Our analysis suggests a negative relationship between TP concentrations and DO levels, consistent with the phenomenon of eutrophication where increased TP leads to decreased DO due to algae growth and oxygen consumption. However, the presence of distinct data clusters, as observed in residual plots, raises questions about the exact nature of this relationship.

The two clusters of data, one showing high DO and low TP and the other showing consistently low DO across a range of TP values, imply potential thresholds or critical points where TP concentrations drastically impact DO levels. Beyond these thresholds, DO levels may drop significantly, possibly indicating eutrophication events leading to near-zero DO levels.

Limitations and Future Directions

While our models provide valuable insights into the TP-DO relationship, several limitations exist. The non-normality of residuals and heteroscedasticity indicate that our models may not fully capture all nuances in the data. Future research could explore alternative modeling approaches or data transformations to improve model accuracy and address these limitations.

Additionally, our study focused on TP as the sole predictor of DO levels. Including other variables such as temperature, pH, and nutrient concentrations could enhance the predictive power of our models and provide a more comprehensive understanding of factors influencing DO in lakes.

In conclusion, while our models demonstrate a negative relationship between TP concentrations and DO levels, further refinement and exploration are needed to better predict and understand the complexities of this relationship, especially regarding critical thresholds and non-linear effects.

Description of project roles

Emma wrote the intro, methods, model 1 results, fit model 1, and annotated the r script. Jenn wrangled the datasets for analysis, completed the testing/training of all models, wrote model 2 results, and fit model 2. Cara wrote the analysis and conclusion, model 3 results, and fit model 3. All group members contributed in fine-tuning the final submission and helped one-another out when necessary.

Bibliography

Carpenter, S. R., Cole, J. J., Hodgson, J. R., Kitchell, J. F., Pace, M. L., Bade, D., ... & Schindler, D. E. (2001). Trophic cascades, nutrients, and lake productivity: whole-lake experiments. Ecological monographs, 71(2), 163-186.

Lee, G. F. (1973). Role of phosphorus in eutrophication and diffuse source control. In Phosphorus in Fresh Water and the Marine Environment (pp. 111-128). Pergamon.

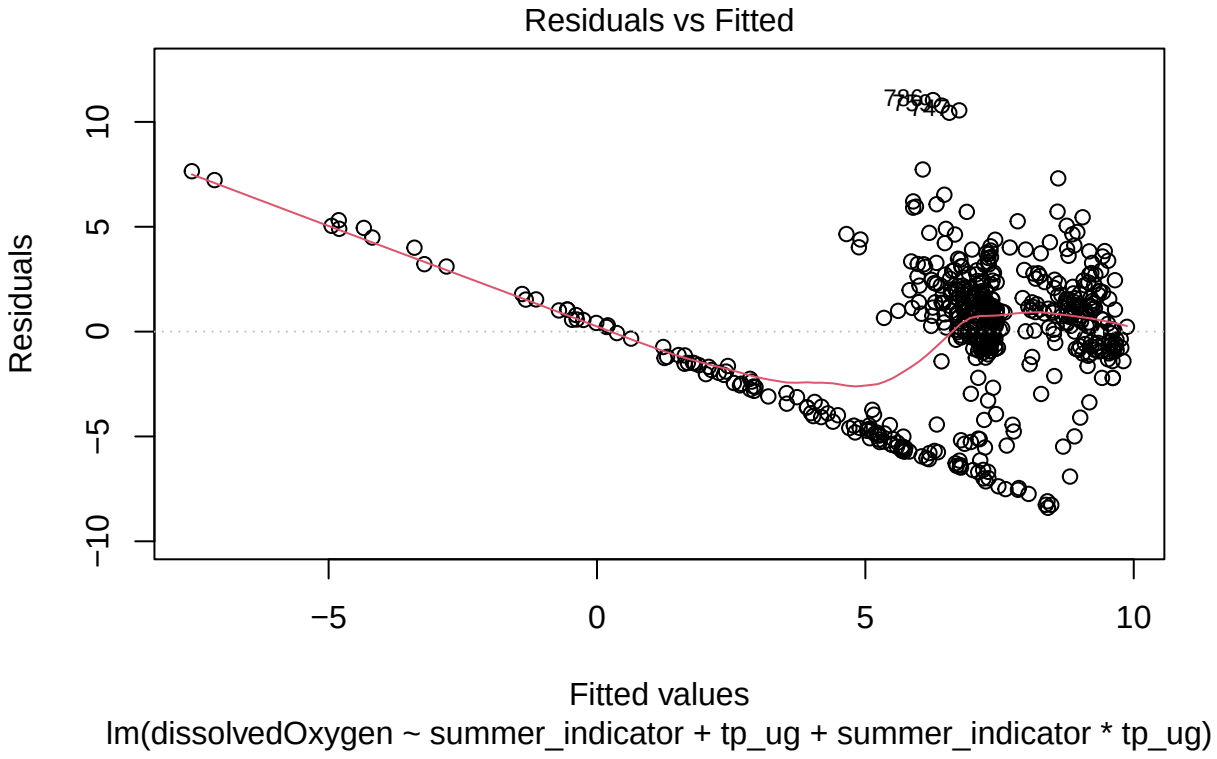
Total nitrogen/total phosphorus and chlorophyll a: Volunteer data entry and information: indiana clean lakes program: indiana university. (n.d.). Indiana Clean Lakes Program. Retrieved April 15, 2024, from https://clp.indiana.edu/volunteer-data/phosphorus-chlorophyll.html

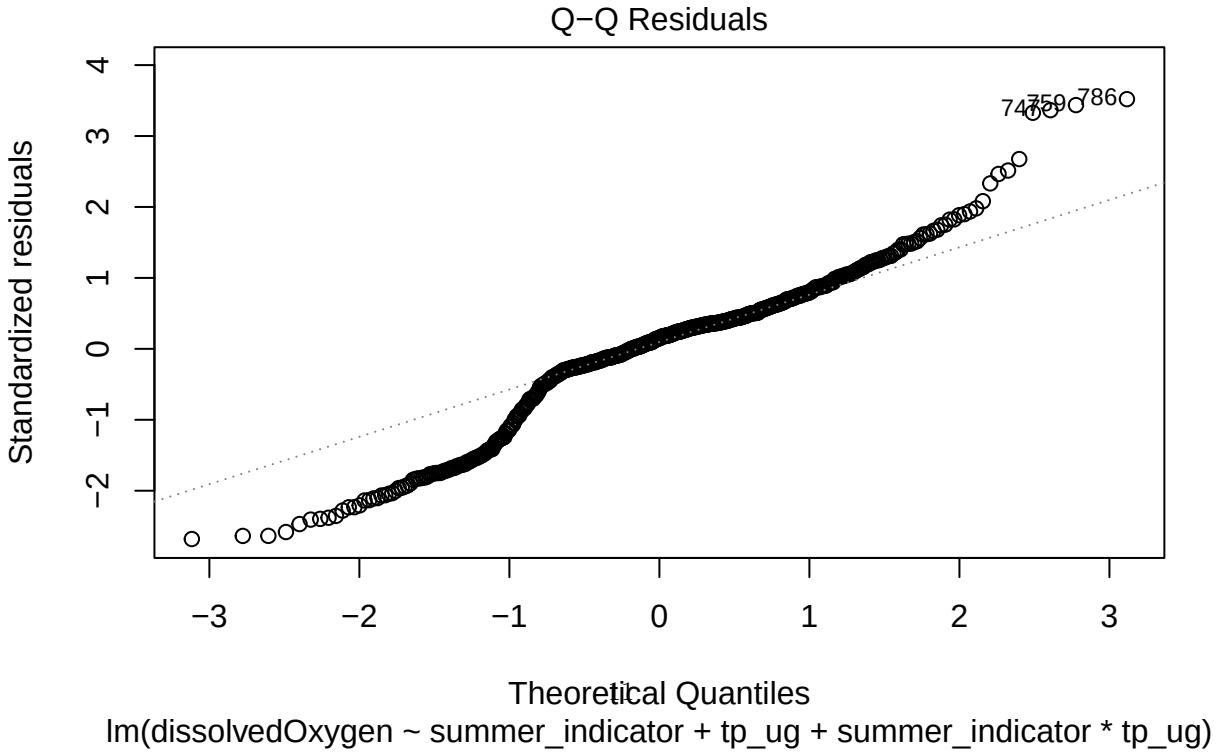
US EPA, O. (2013, November 20). Indicators: Dissolved oxygen [Overviews and Factsheets]. https://www.epa.gov/national-aquatic-resource-surveys/indicators-dissolved-oxygen

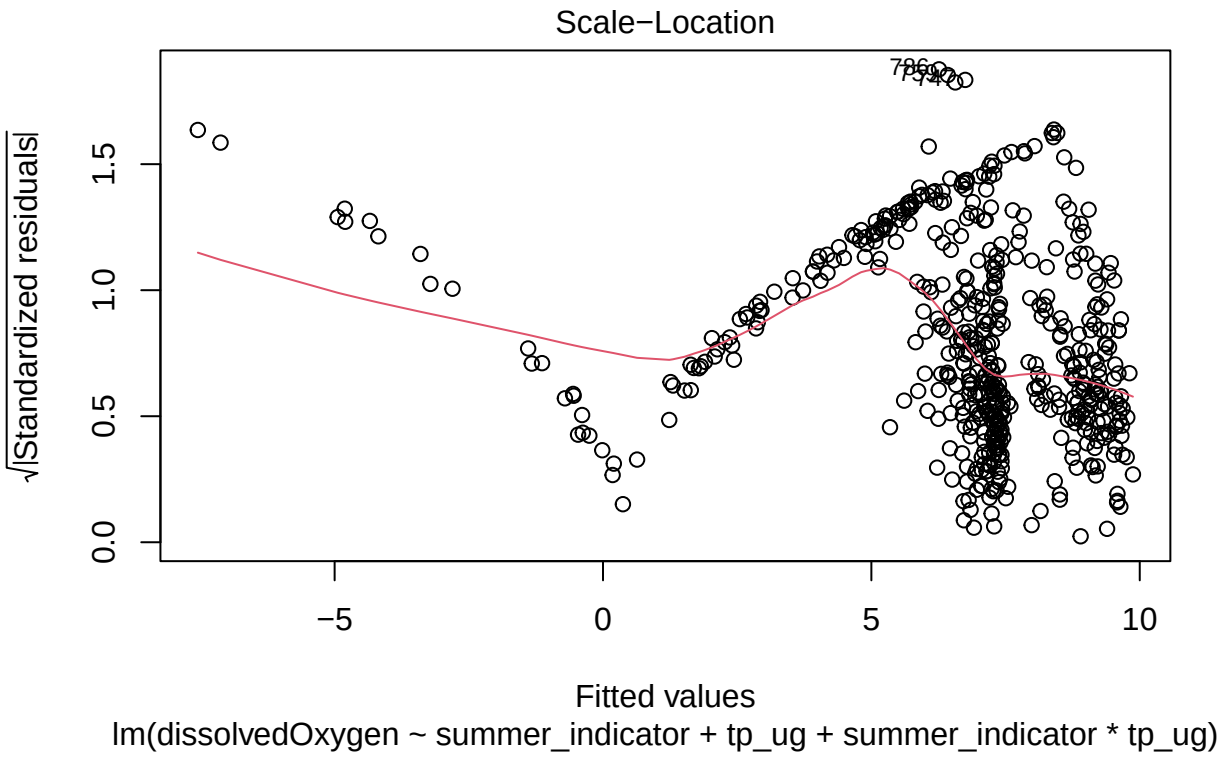
US EPA, O. (2013, November 27). Indicators: Phosphorus [Overviews and Factsheets]. https://www.epa.gov/national-aquatic-resource-surveys/indicators-phosphorus

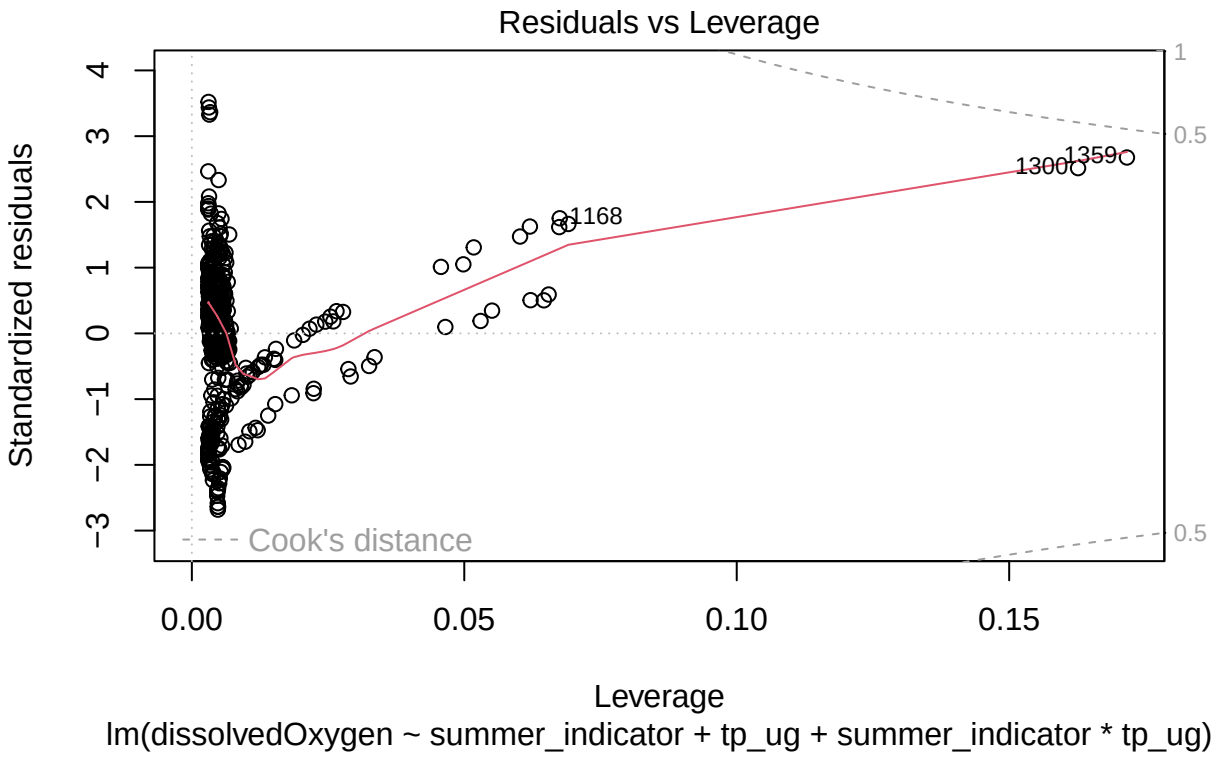
Appendix

Model One Residuals

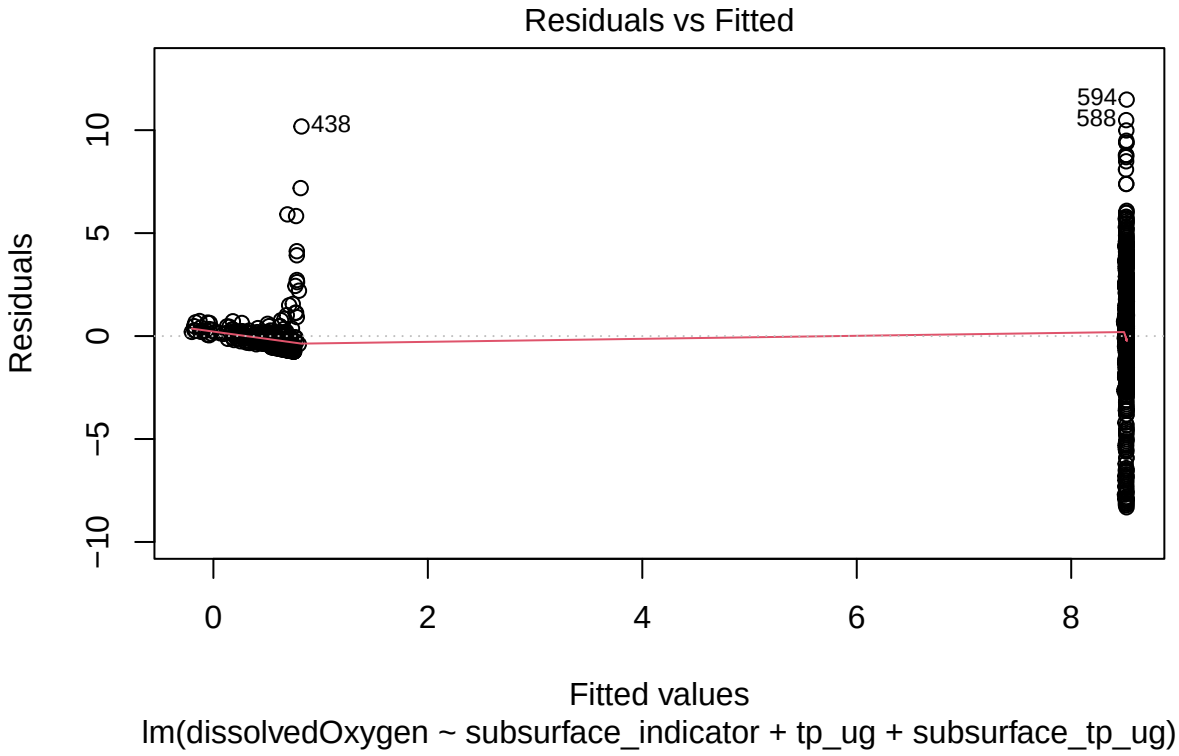


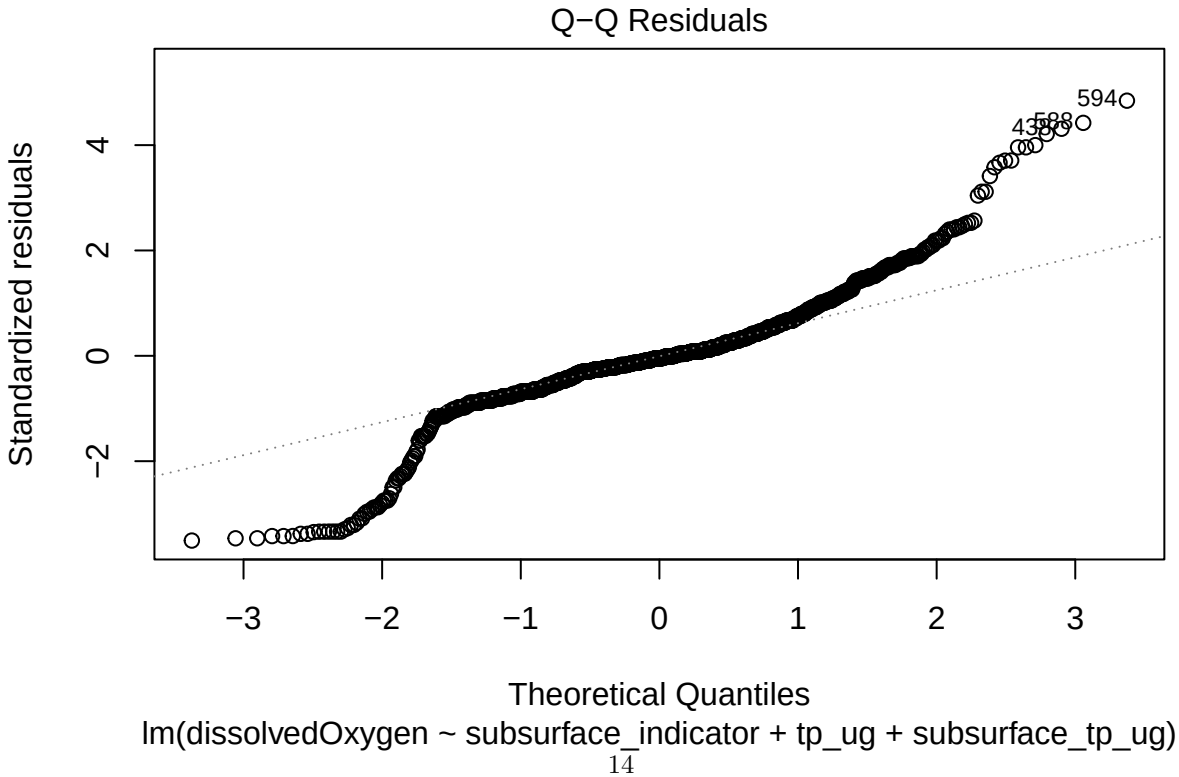


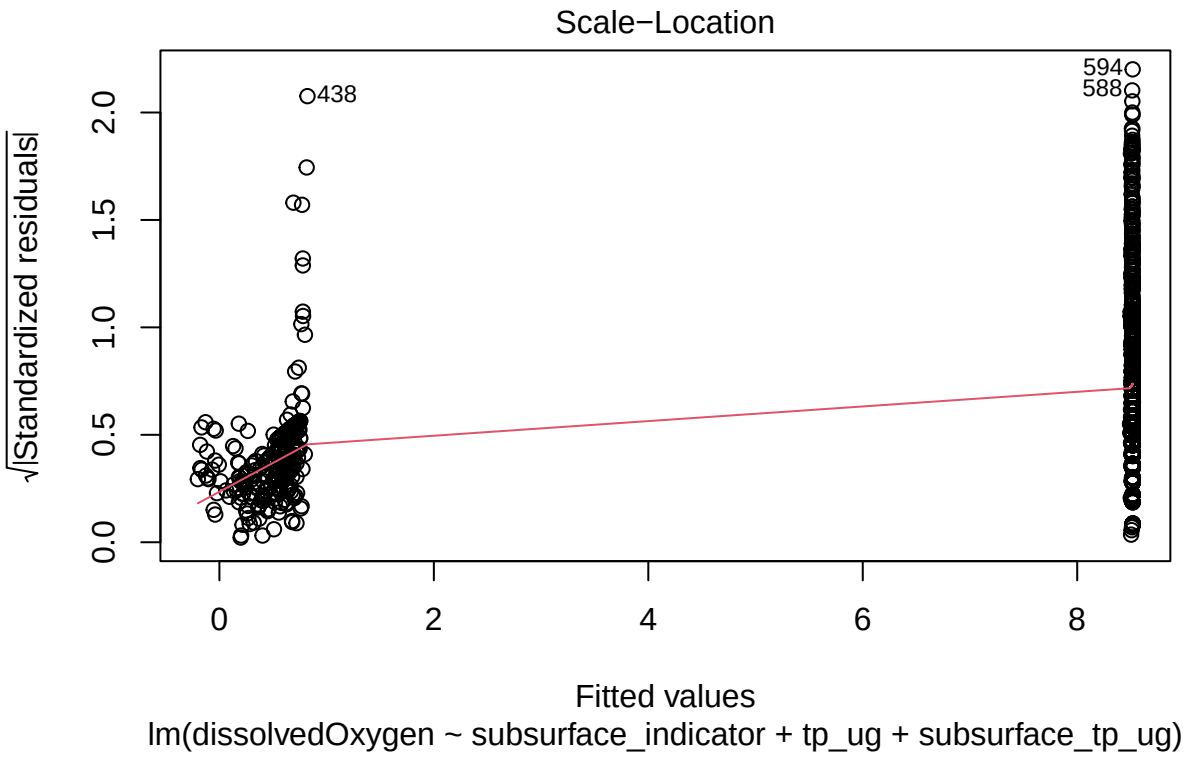

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−5
0
5
10
Fitted values
lm(dissolvedOxygen ~ summer_indicator + tp_ug + summer_indicator * tp_ug)

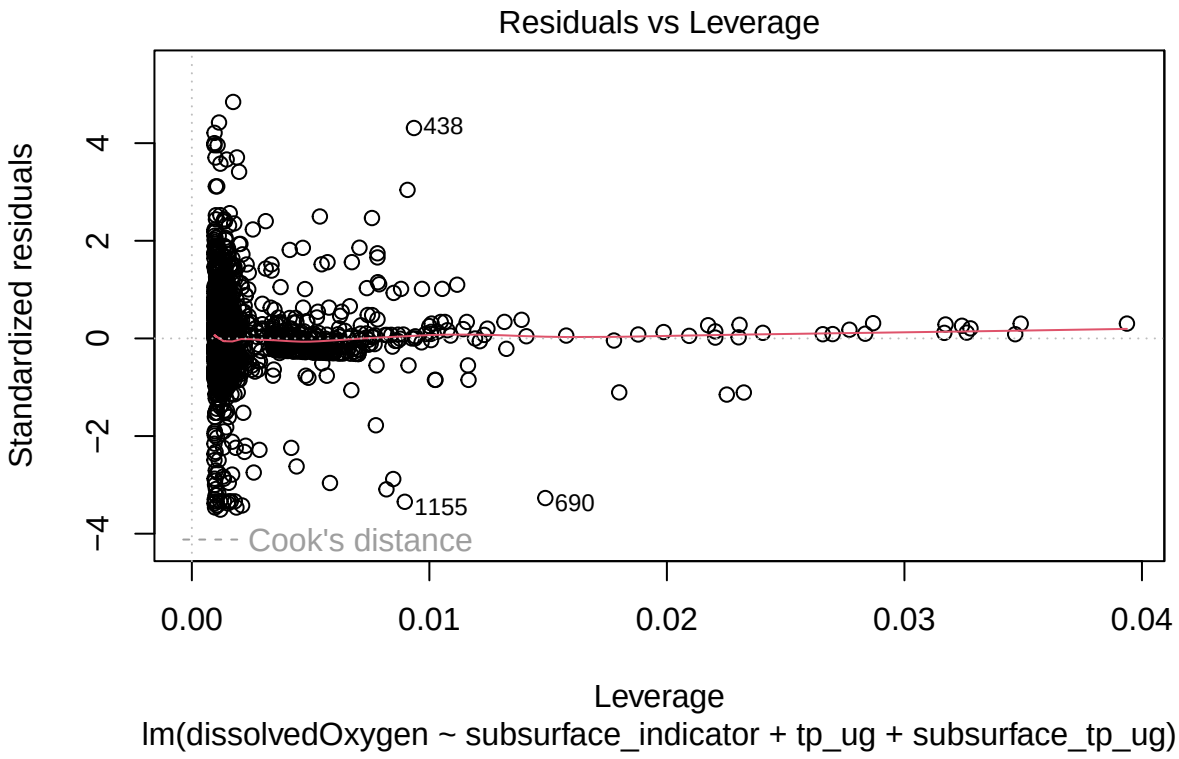

Residuals vs Leverage
Standardized residuals
4
3
2
1
0
−1
−2
−3
1
0.5
0.5
1359
1300
1168
Cook's distance
0.00
0.05
0.10
0.15
Leverage
lm(dissolvedOxygen ~ summer_indicator + tp_ug + summer_indicator * tp_ug)

Model Two Residuals

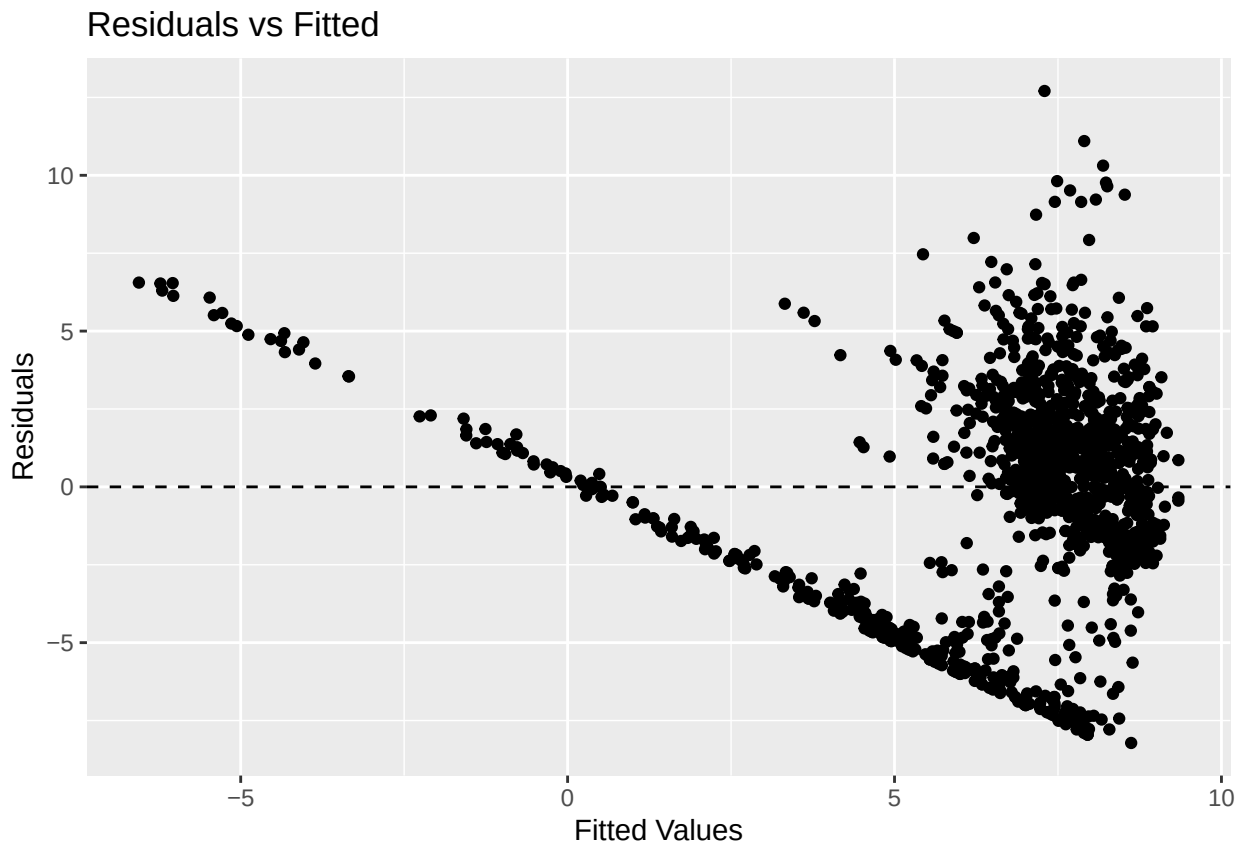


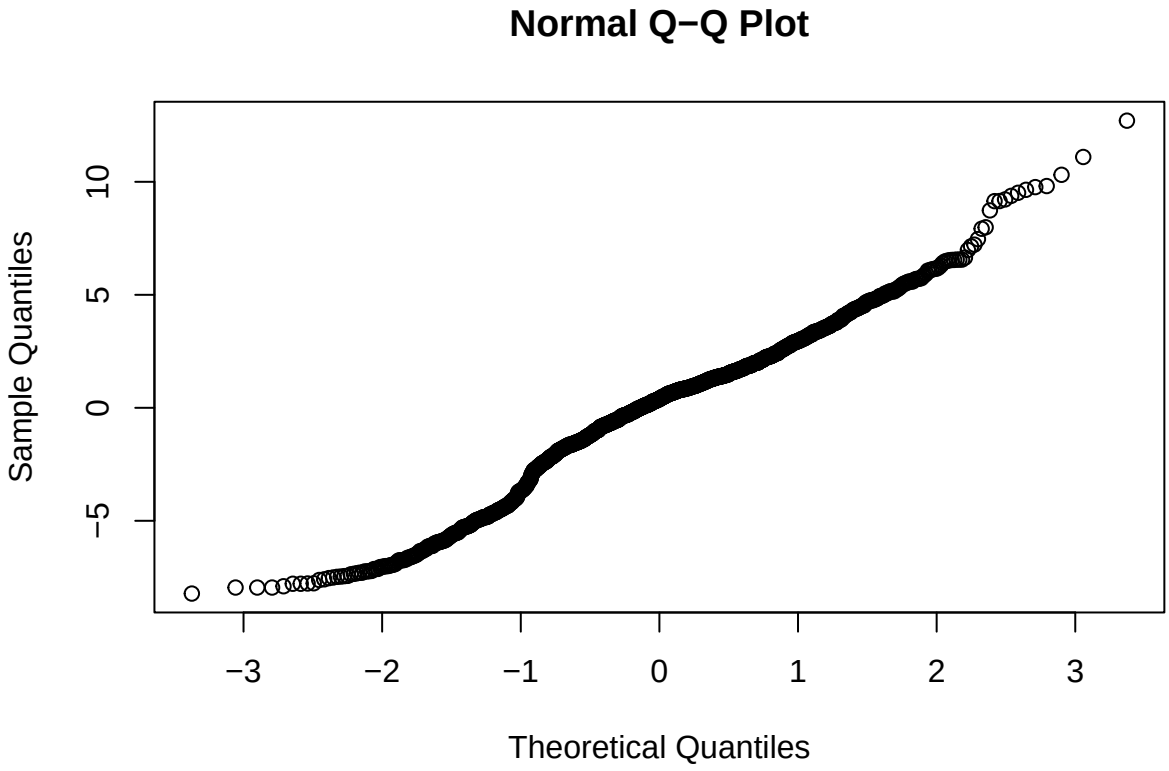



Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
438
594
588
0
2
4
6
8
Fitted values
lm(dissolvedOxygen ~ subsurface_indicator + tp_ug + subsurface_tp_ug)


Residuals vs Leverage
Standardized residuals
4
2
0
−2
−4
438
1155
690
Cook's distance
0.00
0.01
0.02
0.03
0.04
Leverage
lm(dissolvedOxygen ~ subsurface_indicator + tp_ug + subsurface_tp_ug)

Model Three Residuals





Attempts to improve normalization of data

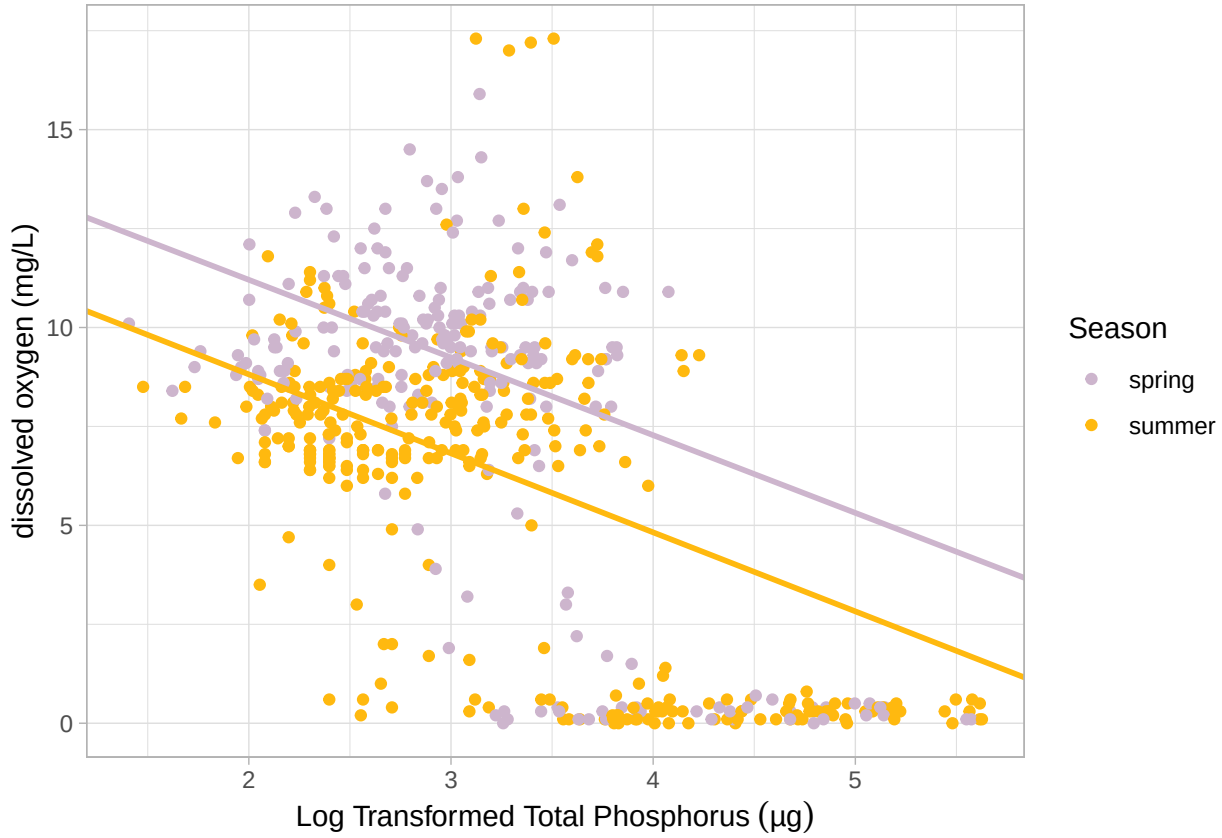


Figure 4: Appendix plot log-transform

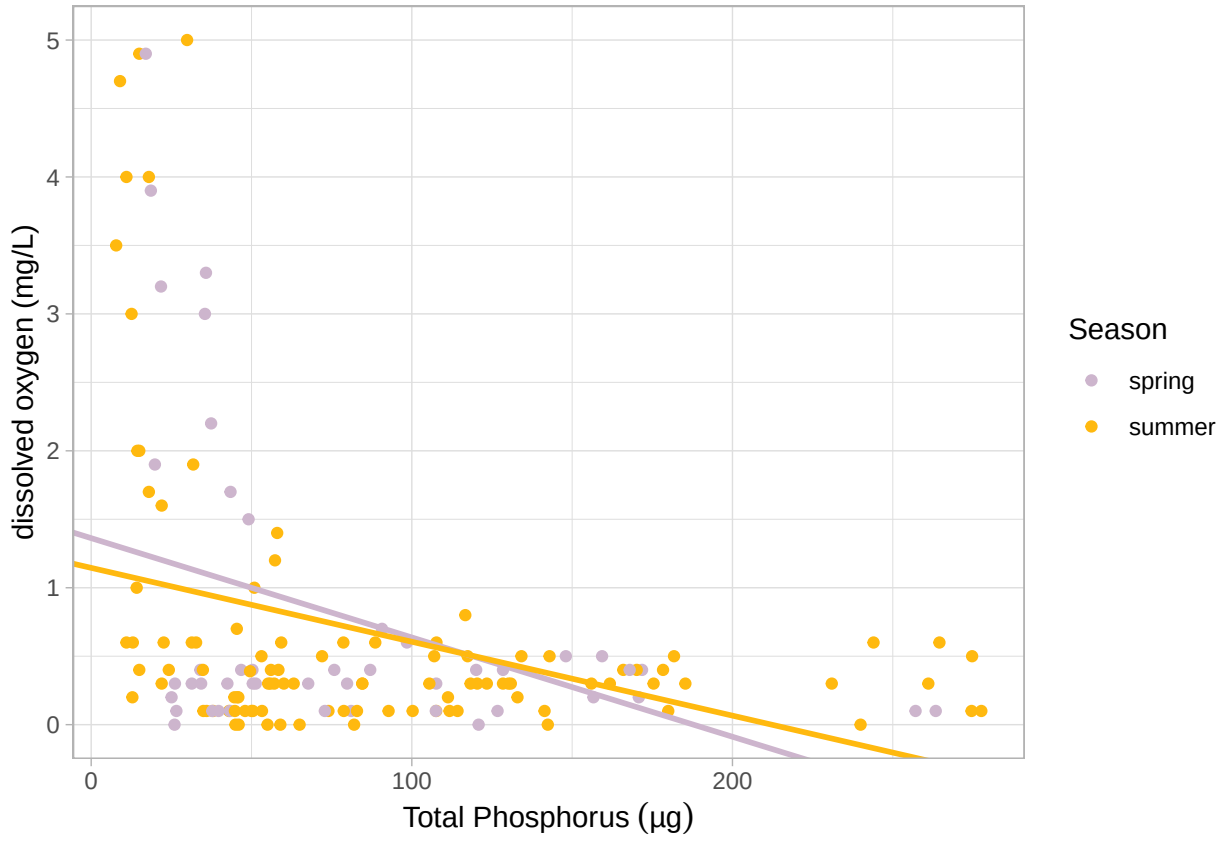
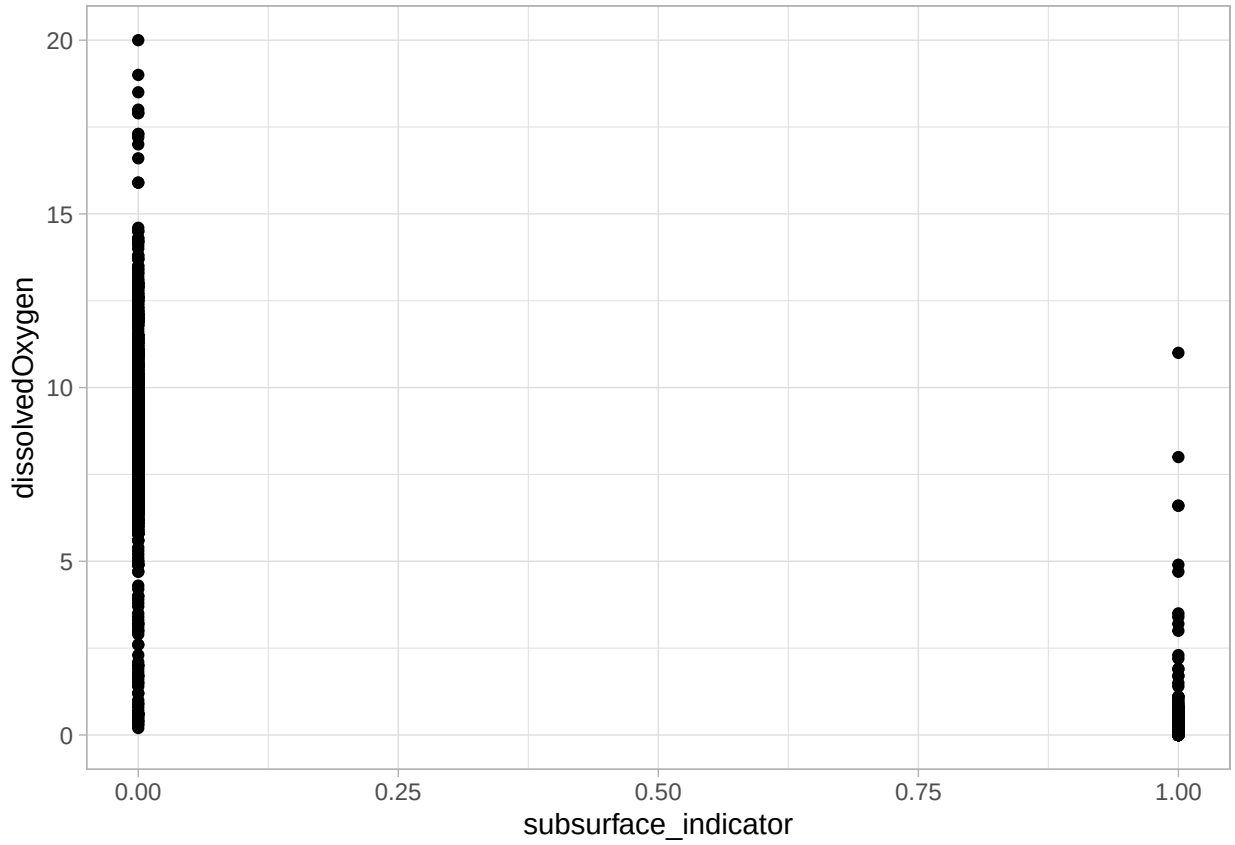
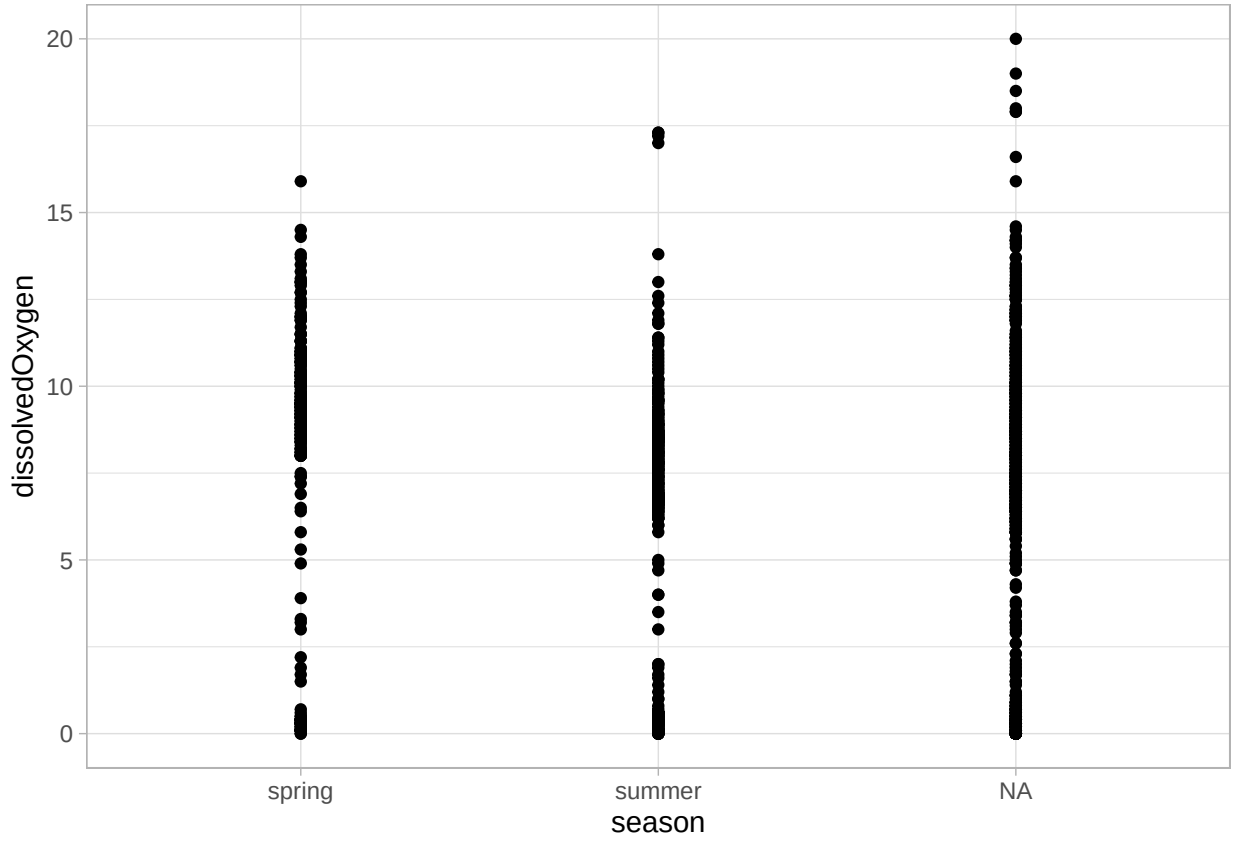


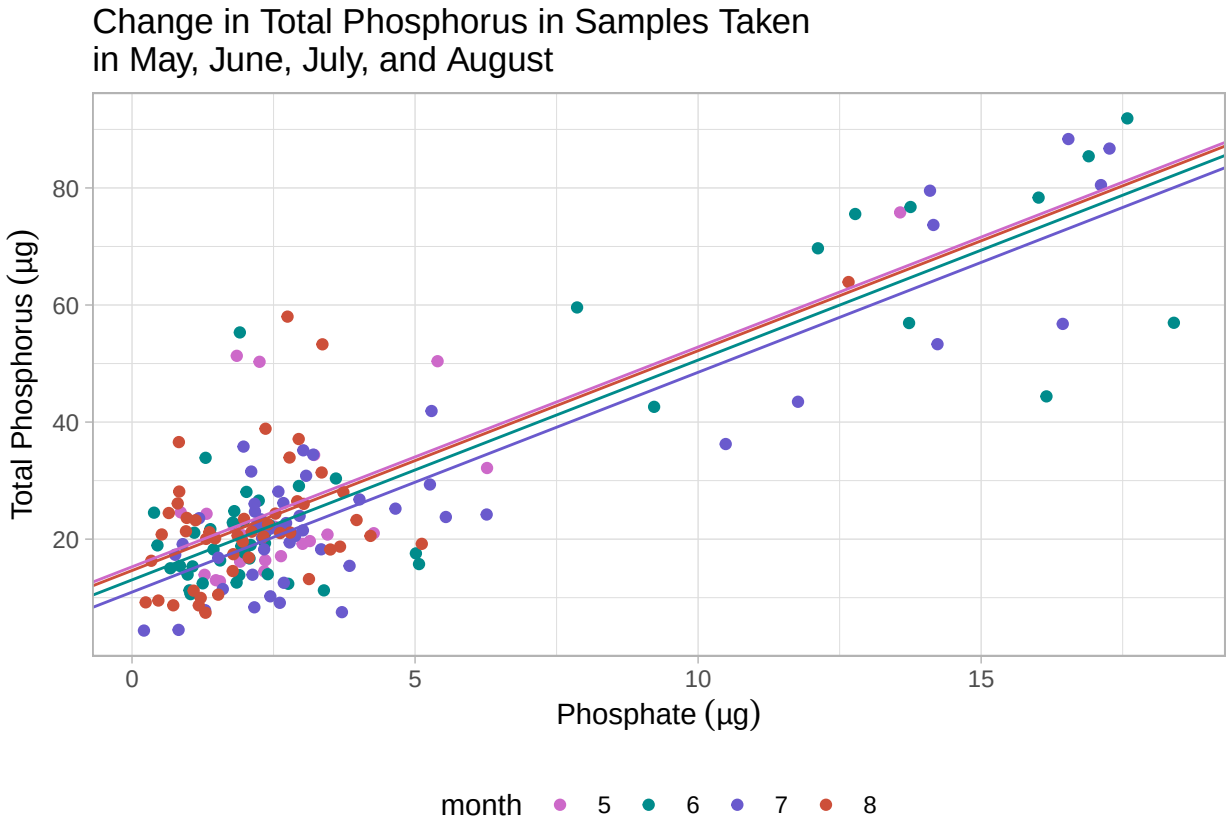
Figure 5: Appendix plot subset DO

Iterative plots and data exploration

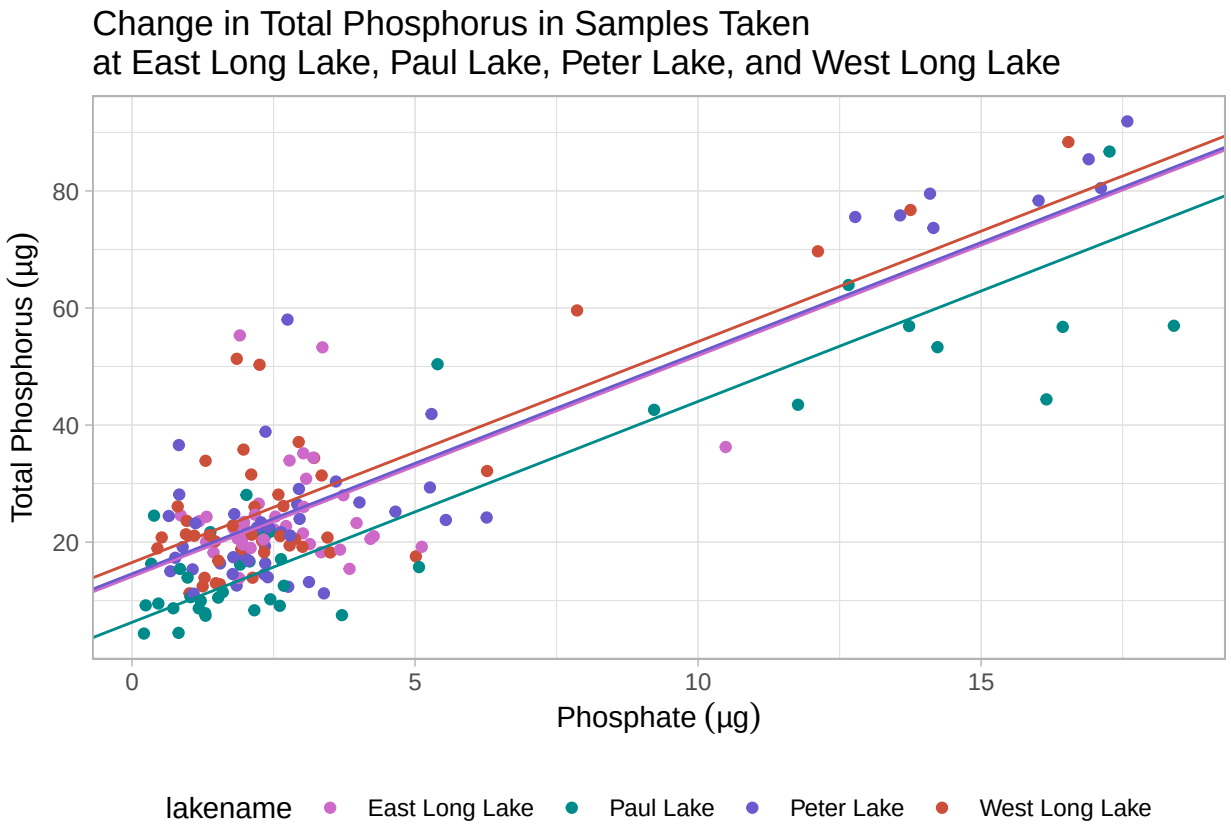


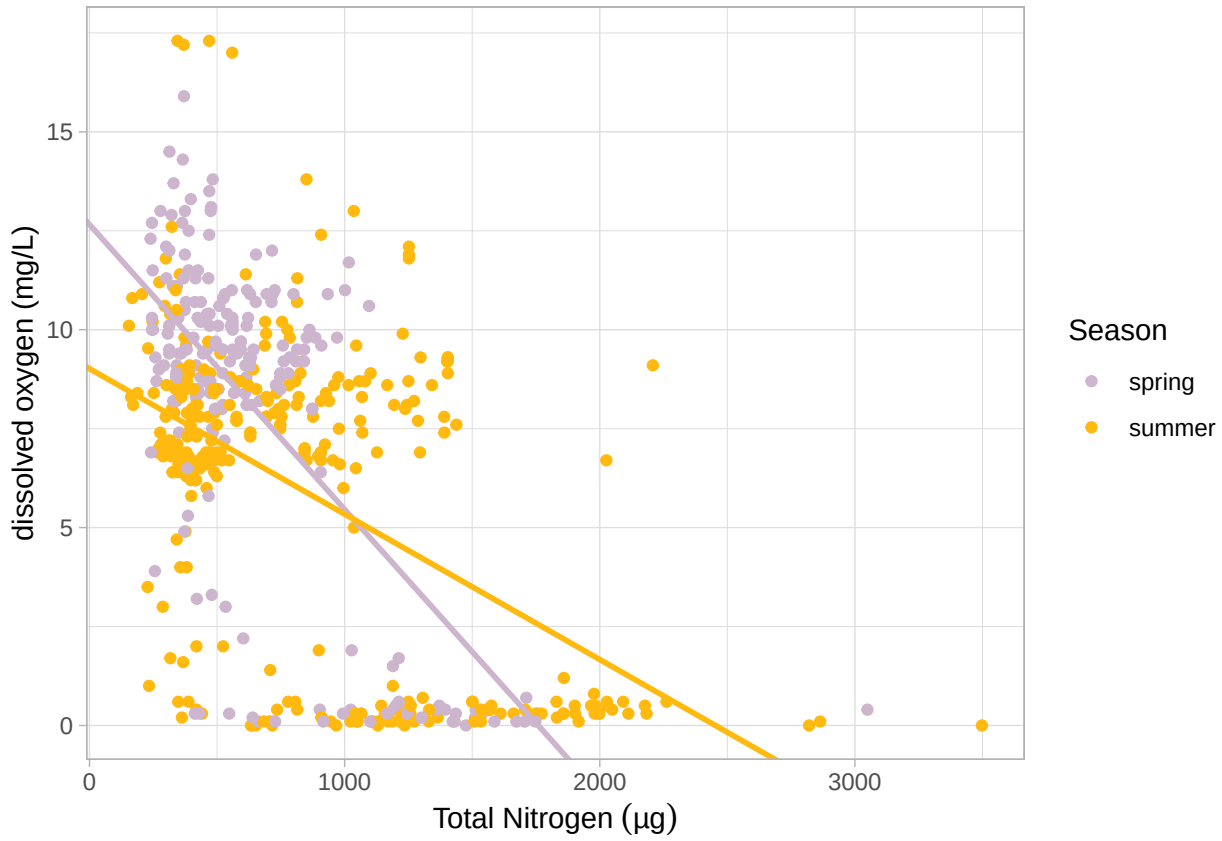


```
##
## Call:
## lm(formula = tp_ug ~ month + po4, data = lake_1994)
##
## Coefficients:
## (Intercept)       month6       month7       month8          po4
##     15.2690      -2.2260      -4.3199      -0.6307       3.7553
```

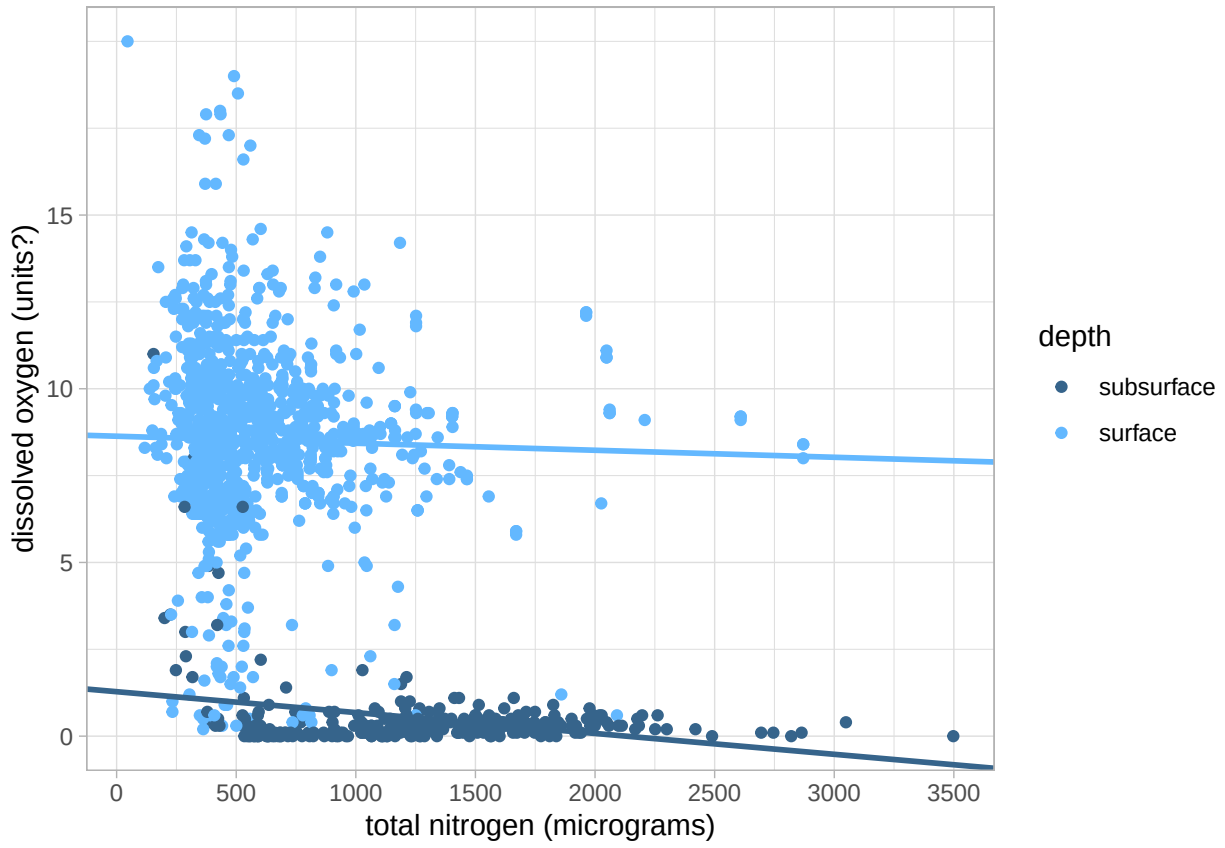



```
##
## Call:
## lm(formula = tp_ug ~ lakename + po4, data = lake_1994)
##
## Coefficients:
##            (Intercept)       lakenamePaul Lake      lakenamePeter Lake
##                14.1293                 -7.8305                  0.4614
## lakenameWest Long Lake                     po4
##                 2.3909                  3.7737
```


Change in Total Phosphorus in Samples Taken
at East Long Lake, Paul Lake, Peter Lake, and West Long Lake
Total Phosphorus (µg)
80
60
40
20
0
5
10
15
Phosphate (µg)
lakename
East Long Lake
Paul Lake
Peter Lake
West Long Lake



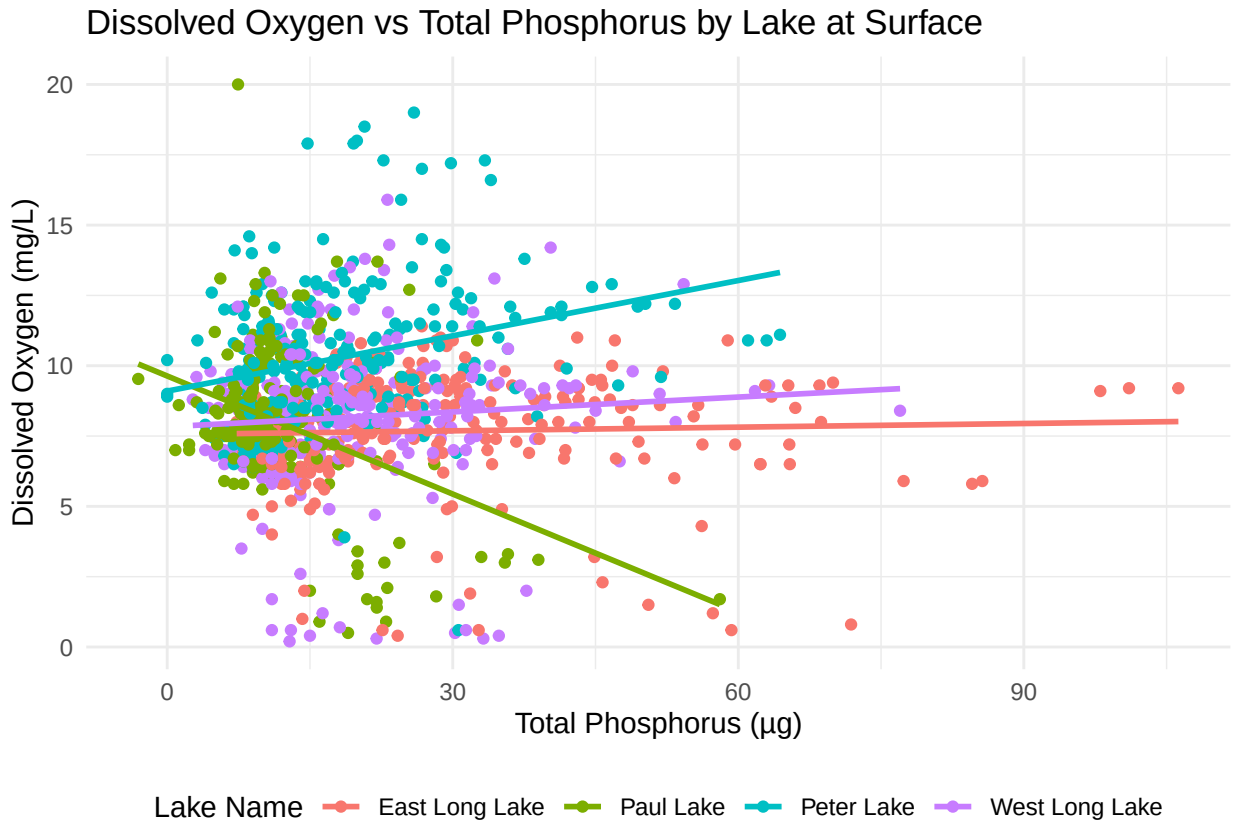

```
##
## Call:
## lm(formula = dissolvedOxygen ~ subsurface_indicator + tn_ug +
##     subsurface_tn_ug, data = jenn)
##
## Coefficients:
##          (Intercept)  subsurface_indicator                 tn_ug
##            8.6334713            -7.3533011            -0.0002020
##     subsurface_tn_ug
##           -0.0003988
```




```
## Linear mixed model fit by REML ['lmerMod']
## Formula: dissolvedOxygen ~ tp_ug + (1 | lakename)
##    Data: lakes_processed_surface
##
## REML criterion at convergence: 4851.2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.0558 -0.5070  0.0015  0.5606  5.0058
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  lakename (Intercept) 1.441    1.200
##  Residual             5.819    2.412
## Number of obs: 1050, groups:  lakename, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept) 8.330860   0.616156  13.521
## tp_ug       0.009972   0.006026   1.655
##
## Correlation of Fixed Effects:
##       (Intr)
## tp_ug -0.191

## Linear mixed model fit by REML ['lmerMod']
```

```
## Formula: dissolvedOxygen ~ tp_ug + (1 | lakename)
##    Data: lakes_processed_surface
## REML criterion at convergence: 4851.247
## Random effects:
##  Groups   Name        Std.Dev.
##  lakename (Intercept) 1.200
##  Residual             2.412
## Number of obs: 1050, groups:  lakename, 4
## Fixed Effects:
## (Intercept)        tp_ug
##    8.330860     0.009972
```




```
## Linear mixed model fit by REML ['lmerMod']
## Formula: dissolvedOxygen ~ tp_ug + (1 | lakename)
##    Data: lakes_processed_summer
##
## REML criterion at convergence: 1703.4
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.2227 -0.3910  0.1498  0.5243  3.3339
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  lakename (Intercept) 0.4257   0.6524
##  Residual             9.4989   3.0820
```

```
## Number of obs: 332, groups:  lakename, 4
##
## Fixed effects:
##              Estimate Std. Error t value
## (Intercept)  7.750824   0.391567   19.79
## tp_ug       -0.045821   0.003436  -13.34
##
## Correlation of Fixed Effects:
##       (Intr)
## tp_ug -0.342
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: dissolvedOxygen ~ tp_ug + (1 | lakename)
##    Data: lakes_processed_summer
## REML criterion at convergence: 1703.426
## Random effects:
##  Groups   Name        Std.Dev.
##  lakename (Intercept) 0.6524
##  Residual             3.0820
## Number of obs: 332, groups:  lakename, 4
## Fixed Effects:
## (Intercept)        tp_ug
##     7.75082     -0.04582
```