Final Project Part 3

The Effects of Temperature and Biological Organisms on the Dissolved Oxygen in the Water of Lakes

Justin Yu

2022-12-01

Contents

1 Introduction:

Contrary to popular beliefs, oxygen production comes mostly from the ocean and lakes and not trees. This production accounts for 50%-80% of Earth's oxygen and can be influenced by temperature and phytoplankton (microscopic marine algae) among other factors.[4] Previous studies have shown that increased levels of phytoplankton can lead to higher levels of dissolved oxygen in water.[2] Temperature can also affect the amount of dissolved oxygen in the water as cold water can hold more oxygen than warm water.[1] This study supports these findings and the importance of these bio-organisms in oxygen production for addressing climate change and sustainable development. We need to ask ourselves:

What are the effects of temperature and biological organisms like phytoplankton have on the dissolved oxygen in the water of lakes?

The purpose of this study is thus to explore these potential effects on dissolved oxygen and determine the strengths of the effect.

2 Method:

The study used data from the "Lake Simcoe Monitoring" program on the Ontario Government website, which provided measurements of various organisms and chemicals in Lake Simcoe's water and the water quality from 1980-2016. The analysis focused on the relationship between phytoplankton and temperature on dissolved oxygen levels. The relevant variables in the dataset include:

1. The dissolved oxygen levels in the lake water. ($\frac{mg}{L}$)
2. The levels of various phytoplanktons found in the water (Diatoms, Chrysophytes, Chlorophytes, Cryptophytes, Cyanobacteria, Dinoflagellates, Euglenophytes) ($\frac{mm^3}{L}$)
3. The temperature of the water. (C)

Our overall analysis will involve a linear regression to analyse the relationship of oxygen levels in Lake Simcoe and how it is affected by temperature and different types of phytoplankton found in the lake.

In building the linear model, various statistical tools are used to create the best model to answer the research question. The process includes combining data sets from the monitoring program, creating numerical summaries of the variables, and removing variables with insufficient observations. The data is then split into equal testing and training sets. The training set is used to test the model assumptions (linearity, homoscedasticity, independence, and normality) through residual plots and a normal qq plot, and to verify that the conditional mean response is a single function of a linear combination of the predictors and that the conditional mean of each predictor is a linear function with another predictor from the residual plots. These tests help determine whether the conditions for a linear regression model are met.

If the assumptions are not met but the two conditions hold, we can apply a common transformation on the data and then recheck the assumptions to see if the violations are fixed. The transformations applied to the data is determined using the Box-Cox method where the common transformation was chosen based on the proximity of the box-cox estimate. Once the model satisfy the assumptions we begin the model reduction step. We use the Anova test and partial F test to determine which predictors are insignificant and can be removed from the model and generate different models. We check for multicollinearity in these models using the variance inflation function in these models and if exists we return back to the reduction step. Problematic points were then found using cooks distance, dffitts and dfbetas. Outliers and leverage points were also highlighted based on their cut-offs calculated and the size of the dataset. We also determine the validity of these problematic points and determine if the removal is justified.

After developing multiple different models, we can compare each of then based on their adjusted rsquared, Akaike information criterion (AIC or AICC) and Bayesian information criterion (BIC) to test the goodness of fit for each model. The lowest AIC/AICC/BIC model is picked taking into account how low the adjusted r squared is (if too low compared to other models, choose a model with higher AIC/AICC/BIC). The model is then verifed by running the model on the test data and comparing the coefficients of both training and testing.

3 Results:

Table 1: Showing the Mean, Median and Standard Deviation of the Generalized Important Variables Used in the Study. The table also shows the number of observations for each variable, as well as the number of missing/unknown variables.

Variable	**N**	**N = 232**
Dissolved_Oxygen	232	9.61, 9.64, (0.83)
Unknown		0
Temperature_C	232	13.78, 13.83, (2.42)
Unknown		0
Diatoms	232	0.35, 0.27, (0.28)
Unknown		0
Chrysophytes	232	0.027, 0.020, (0.031)
Unknown		0
Chlorophytes	227	0.027, 0.013, (0.052)
Unknown		5
Cryptophytes	232	0.06, 0.05, (0.06)
Unknown		0
Cyanobacteria	231	0.026, 0.015, (0.035)
Unknown		1
Dinoflagellates	228	0.029, 0.023, (0.025)
Unknown		4
Euglenophytes	34	0.006, 0.002, (0.013)
Unknown		198

Three variables have missing values, so the observations with those values must be removed when creating the model. The variable “Euglenophytes” has 198 missing values, and removing the associated observations would leave only 34 observations to build the model with. Without imputation, the alternative is to exclude “Euglenophytes” from the model. Therefore, “Euglenophytes” is excluded from the model, resulting in 222 total observations to be split between the training and testing data.

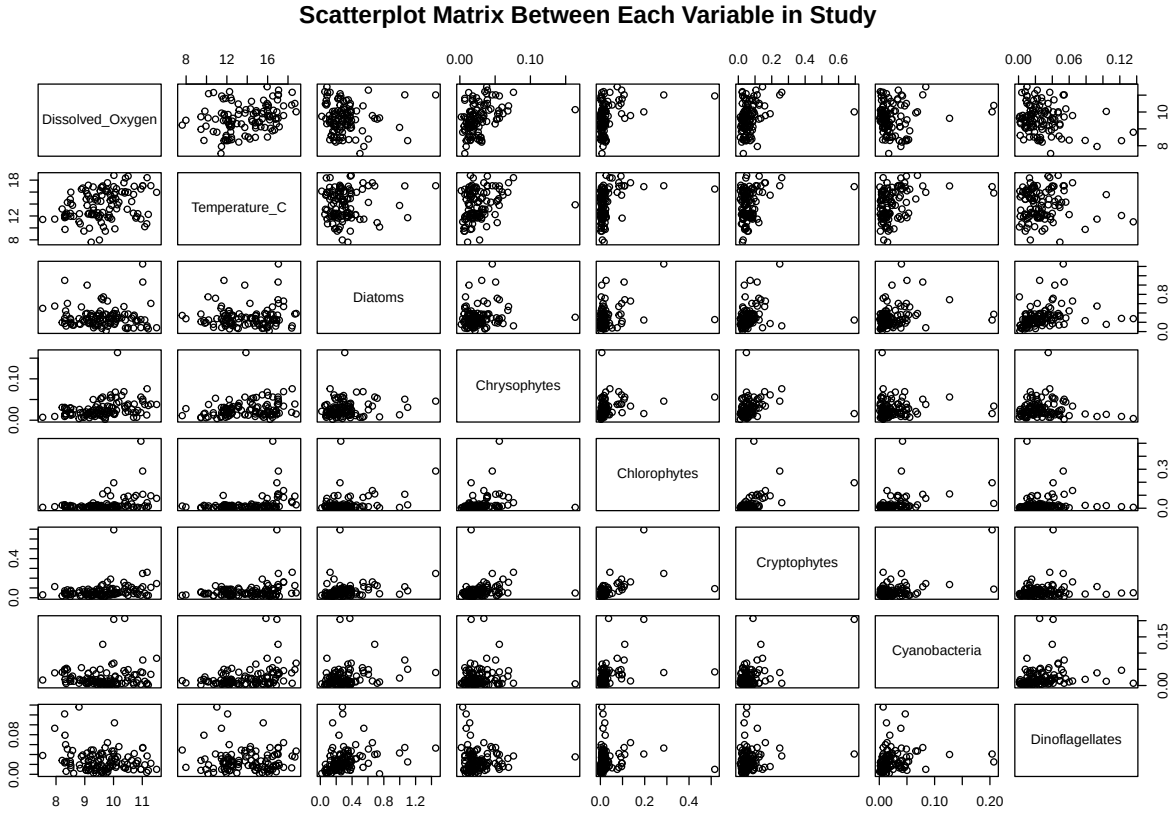

Figure 1: Plot Matrix of Scatterplots Showing the Relationship between Each Variables Being used in the Study. The response variable is Dissolved Oxygen and the remaining seven variables would be the predictor variables in the linear regression model.

The scatterplot matrix is used to informally check for potential violations of the model assumptions. The response appears to be normally distributed, but the predictors are skewed and may affect the linearity assumption or model fit. Homoscedasticity also appears to be violated, as the response's variability increases across values of each predictor. Formal testing is required, but a common transformation using Box Cox estimation may be applied to the variables to address these model violations.

Appendix A tests the assumptions formally. We can see that the residual graphs are interpretable since the two conditions (noted in the method) are not violated. We can see a violation in the Homoscedasticity assumption but the other assumptions seem to hold (linearity - no observable curves, uncorrelated errors - no separate clusters, normality - straight QQ plot). We can now apply a common transformation to the variables suggested by the Box Cox method.

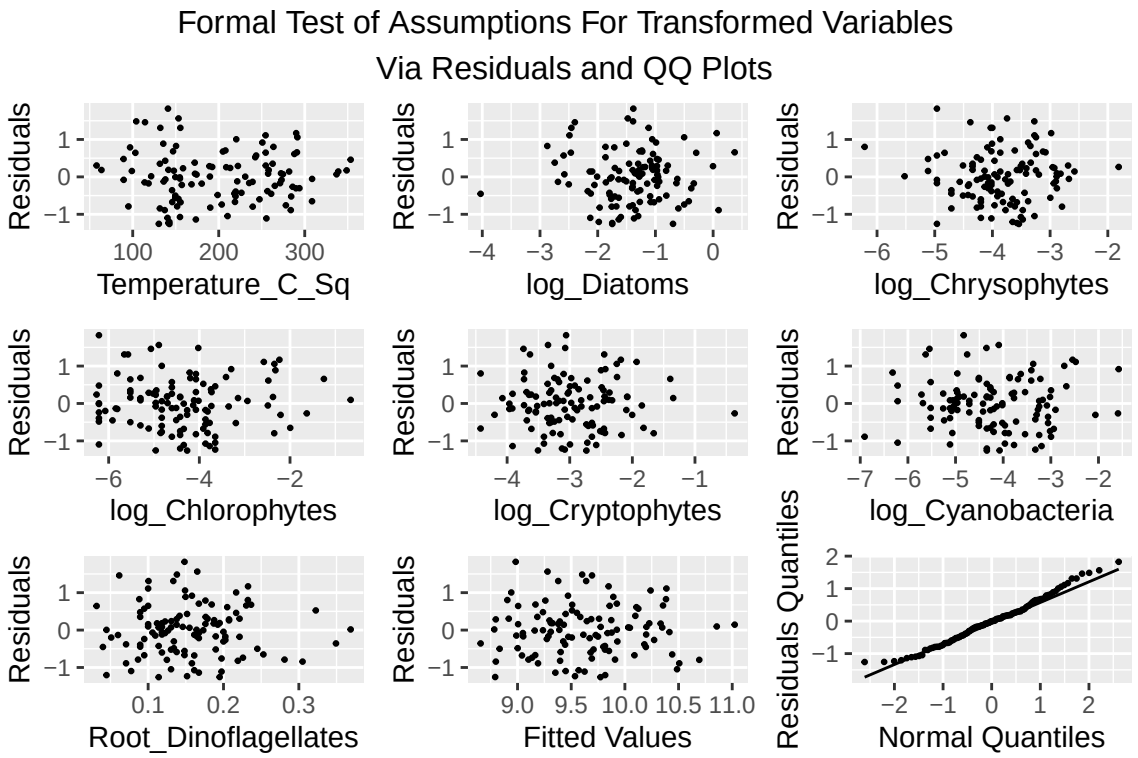

Figure 2: The test of assumptions after the variables were transformed using common transformations determined by the Box-Cox Method. This figure shows the differences seen after applying the transformations as opposed to Appendix A

The variables appear to satisfy the assumptions needed to model and interpret a linear model mentioned in the method. Thus we can start to build the linear model and proceed with the reduction process via Anova and Partial F Test to test and compare different models.

Table 2: Showing Different Linear Regressions Ran, the Lowest and Highest Estimates for the Variance Inflation Function are also included in the table. The full model is the original model with untransformed variables which violate the assumptions. All the models has "Dissolved_Oxygen" as the response variable. Model 2 uses all the variables which have been transformed, Model 3 is a reduced version of Model 2 (without "Root_Dinoflagellates"). Model 4 is a reduced version of Model 2 (without "log_Chrysophytes') and Model 5 and 6 further removes"log_Diatoms" and then "log_Cryptophytes" from Model 4. Model 7 is where only "log_Diatoms" is removed and model 8 is where only "log_Cryptophytes" is removed from model 2.

Model	Adjusted R^2	AIC	AICC	BIC	Lowest VIF	Highest_VIF
Full	0.229	257.634	259.381	282.02	1.128	1.67
2	0.303	246.525	248.272	270.91	1.288	2.416
3	0.294	246.932	248.316	268.608	1.241	2.335
4	0.299	246.133	247.517	267.809	1.275	2.071
5	0.292	246.256	247.323	265.223	1.196	2.069
6	0.287	246.127	246.919	262.384	1.192	1.576
7	0.293	247.08	248.465	268.757	1.232	2.404
8	0.294	246.982	248.367	268.659	1.285	2.052

From Table 2, we can judge which model is a better fit on the variables, than the other by comparing AICC, AIC, R^2 and BIC. All the models built do not show a multi-collinearity issue based on the range of their Variance Inflation Function estimates (< 5). The models have a relatively similar AIC, AICC and BIC despite having different amounts of predictors. Model 4, which is the model that does not include log_Chrysophytes, has a lower AIC, AICC and BIC relative to a higher adjusted R^2 compared to the other models. Based on these notations (lower AIC, etc being better) and that the model satisfies the assumptions needed for a linear model, we choose model 4 as the final model.

Using model 4 as the chosen model, We identified 8 leverage points in the data that are distant from the rest of the observations in the predictor space. We also identified 4 outlier observations when considering the dataset as "small", but none when considering it as "large". No observations were identified as being influential on the entire regression surface, but we identified 5 observations who influenced their own fitted values and between 5-11 observations being influential on at least one estimated coefficient. Observing the values of these problematic points (to check if any measurement errors occurred) however, it was determined that the points reasonable in context and should not be removed from the model data. Now that the training model has been developed, we compare it to the testing model.

Table 3: Summary of characteristics of Model 4 between the training and test datasets. Model 4 uses Temperature_C_Sq, log_Diatoms, log_Chlorophytes, log_Cryptophytes, log_Cyanobacteria and Root_Dinoflagellates as predictors. The Response is Dissolved_Oxygen in both models. Coefficients are presented as estimate $\pm$ SE.

Characteristic	Model 4 (Train)	Model 4 (Test)
Largest VIF value	2.070864	1.9487494
Cook's D	0	0
DFFITS	5	7
Leverage Points	8	8
Outliers (Small)	4	7
Outliers (Large)	0	0
Violations	none	none
Intercept	10.026 $\pm$ 0.644	8.906 $\pm$ 0.609
Temperature_C_Sq	0.003 $\pm$ 0.001	0.005 $\pm$ 0.001
log_Diatoms	-0.168 $\pm$ 0.119	-0.197 $\pm$ 0.133
log_Chlorophytes	0.299 $\pm$ 0.083	0.016 $\pm$ 0.1
log_Cryptophytes	0.215 $\pm$ 0.127	0.252 $\pm$ 0.155
log_Cyanobacteria	-0.269 $\pm$ 0.084	-0.165 $\pm$ 0.094
Root_Dinoflagellates	-2.23 $\pm$ 1.253	-1.687 $\pm$ 1.227
Adjusted R^2	0.299	0.16

Based on Appendix B and C, both the training and testing models seem to satisfy the conditions and assumptions of a linear regression model and both had similar residual and QQ plots. In comparing model 4 between the training data and testing data, we see some similarities in both the amount of problematic points and the VIF values. The testing data set had more outliers and points who influenced their fitted values but a lower max VIF value. The signs of coefficients remained the same for both sets, however, there is a noticeable difference in size of the coefficient in "log_Chlorophytes" and also Adjusted R^2. The difference seen from the Adjusted R^2 may suggest that the training model was over-fitted.

4 Discussion:

The final model displayed, helps us analyse the various effects of temperature and different types of phytoplankton on the dissolved oxygen levels found in Lake Water, thus answering our research question. By looking at the coefficients of the model, we are able to interpret what effect an increase of that variable can have on the level of dissolved oxygen in the water. For example, looking at the predictor variable of Temperature_C_Sq, we can see that it has a small and positive impact on the levels of dissolved oxygen in the lake water (with a small standard error of 0.01). Comparing this to log_Cyanobacteria, we see an increase in log_Cyanobacteria causes the dissolved oxygen levels to fall by -0.269 (with a small standard error of 0.084), it is important to not interpret this as an increase of Cyanobacteria leads to decreases in the oxygen level in water.

The model has some limitations, including a low adjusted r-squared of 0.299, indicating that the variability of dissolved oxygen levels is weakly explained by the predictor variables. This may be due to insufficient observations/predictors or a smaller correlation between phytoplankton and temperature and dissolved oxygen levels than initially thought. Additionally, data collected from the natural environment may be affected by other unobservable factors, such as water flow rate, that have a greater impact on dissolved oxygen levels.

The chosen model may be overfitted due to the difference in adjusted r-squared values between the training and testing data sets. This may be due to a small data size with insufficient samples to accurately represent all levels of input, or to the transformations applied. An overfitted model would produce biased estimates of the coefficients of the predictor variables and not accurately measure the effect on dissolved oxygen levels. To address these limitations, a larger study with more observations and possibly additional predictors would be necessary.

5 Bibliography:

Scientific Articles Used:

1. Khan, U. T., & Valeo, C. (2015). A new fuzzy linear regression approach for dissolved oxygen prediction. Hydrological Sciences Journal, 60(6), 1096–1119. https://doi.org/10.1080/02626667.2014.900558

2. Smith, D. W., & Piedrahita, R. H. (1988). The relation between phytoplankton and dissolved oxygen in fish ponds. Aquaculture, 68(3), 249–265. https://doi.org/10.1016/0044-8486(88)90357-2

3. Wang, J., & Zhang, Z. (2020). Phytoplankton, dissolved oxygen and nutrient patterns along a eutrophic river-estuary continuum: Observation and modeling. Journal of Environmental Management, 261, 110233. https://doi.org/10.1016/j.jenvman.2020.110233

Websites used:

4. How much oxygen comes from the ocean? NOAA's National Ocean Service. (n.d.). Retrieved October 15, 2022, from https://oceanservice.noaa.gov/facts/ocean-oxygen.html#:~:text=Scientists%20estimate%20that%2050%2D80,smallest%20photosynthetic%20organism%20on%20Earth

5. Government of Ontario. (n.d.). Lake Simcoe Monitoring - Ontario Data Catalogue. Lake Simcoe Monitoring - Datasets - Ontario Data Catalogue. Retrieved October 22, 2022, from https://data.ontario.ca/en/dataset/lake-simcoe-monitoring

6. United Nations. (n.d.). Global Population Growth and sustainable development | population division. United Nations. Retrieved October 15, 2022, from https://www.un.org/development/desa/pd/content/global-population-growth

Packages Used

7. tidyverse https://www.tidyverse.org/packages/

8. grid https://CRAN.R-project.org/package=grid

9. gridExtra https://cran.r-project.org/web/packages/gridExtra/index.html

10. knitr::kable https://rmarkdown.rstudio.com/lesson-7.html

11. gtsummary https://cran.r-project.org/web/packages/gtsummary/index.html

6 Appendix:

6.1 A

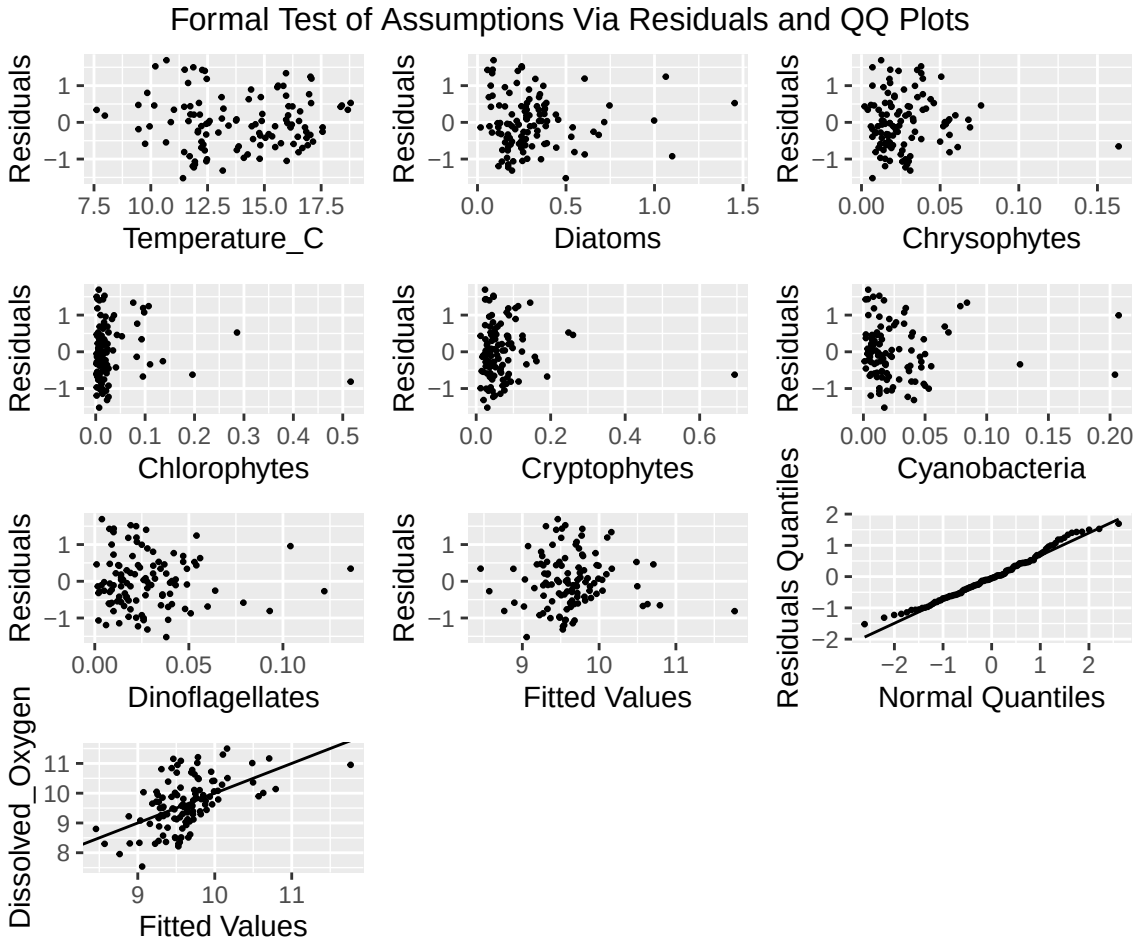

Figure 3: Formal Test Assumptions for the Untransformed Training Data Set. The QQ Plot represents the normality assumption and the Residual Vs Predictor Graphs can show violations of the Constant Variance Assumption. The Response vs Fitted plot is used to verify a condition of using the residual plots to interpret the assumption validity

6.2 B

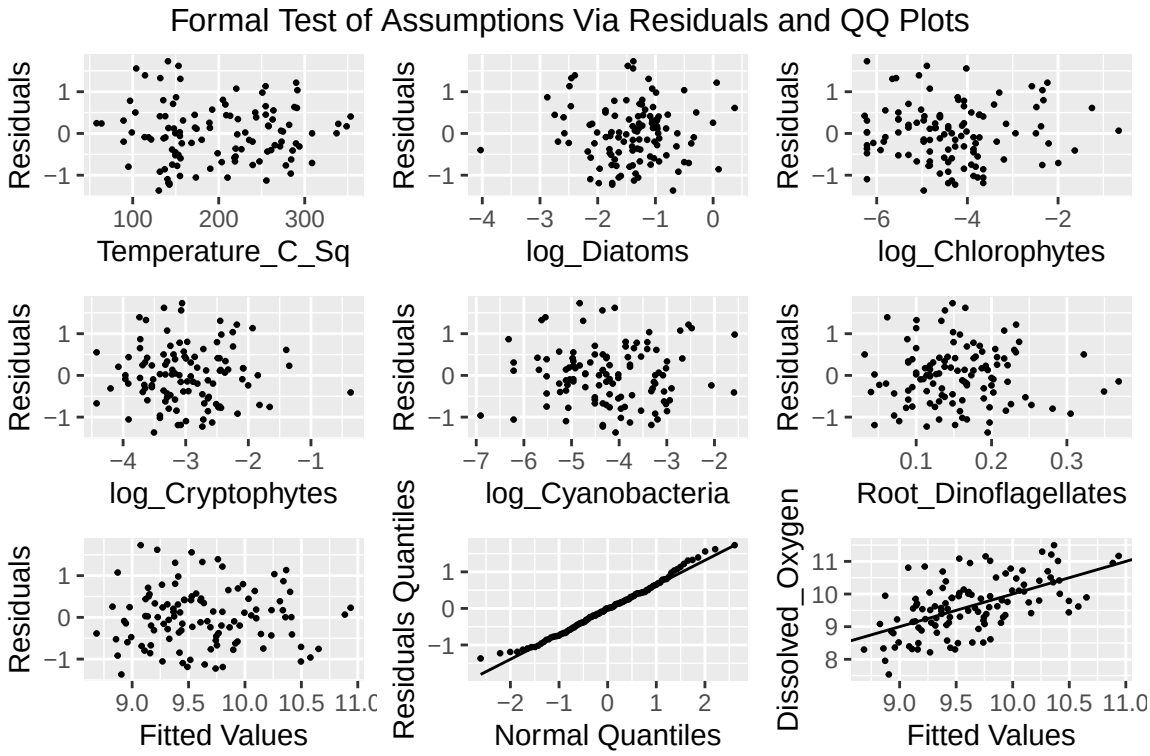

Figure 4: Formal Test Assumptions for the Chosen Model on the Training Data Set. The QQ Plot represents the normality assumption and the Residual Vs Predictor Graphs can show violations of the Constant Variance Assumption. The Response vs Fitted plot is used to verify a condition of using the residual plots to interpret the assumption validity

6.3 C

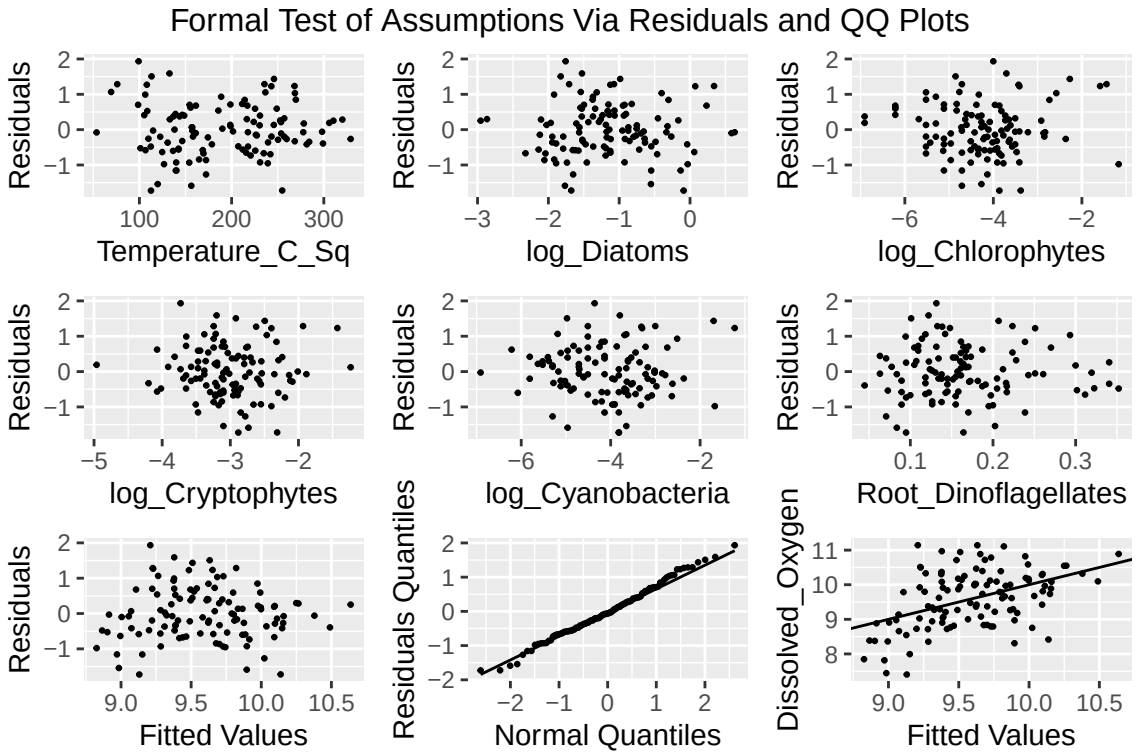

Figure 5: Formal Test Assumptions for the Chosen Model on the Testing Data Set. The QQ Plot represents the normality assumption and the Residual Vs Predictor Graphs can show violations of the Constant Variance Assumption. The Response vs Fitted plot is used to verify a condition of using the residual plots to interpret the assumption validity