6: Part 1 - Generalized Linear Models

Environmental Data Analytics | John Fay and Luana Lima | Developed by Kateri Salk

Spring 2022

Objectives

1. Describe the components of the generalized linear model (GLM)
2. Apply special cases of the GLM (linear regression) to real datasets
3. Interpret and report the results of linear regressions in publication-style formats
4. Apply model selection methods to choose model formulations

Generalized Linear Models (GLMs)

The analysis of variance (ANOVA), and linear regression are special cases of the **generalized linear model** (GLM). The GLM also includes analyses not covered in this class, including logistic regression, multinomial regression, chi square, and log-linear models. The common characteristic of general linear models is the expression of a continuous response variable as a linear combination of the effects of categorical or continuous explanatory variables, plus an error term that expresses the random error associated with the coefficients of all explanatory variables. The explanatory variables comprise the deterministic component of the model, and the error term comprises the stochastic component of the model.

> Data (continuous variable y) = model (type of model depends on one or more variables x1, x2, . . . , xn, which can be continuous or categorical) + residual or error term (GLM is based on the assumption that the data residuals approximate a normal distribution)

Historically, artificial distinctions were made between linear models that contained categorical and continuous explanatory variables, but this distinction is no longer made. The inclusion of these models within the umbrella of the GLM allows models to fit the main effects of both categorical and continuous explanatory variables as well as their interactions.

Choosing a model from your data: A "cheat sheet"

One-way ANOVA (Analysis of Variance): Continuous response, one categorical explanatory variable with more than two categories.

Two-way ANOVA (Analysis of Variance) Continuous response, two categorical explanatory variables.

Single (Simple) Linear Regression Continuous response, one continuous explanatory variable.

Multiple Linear Regression Continuous response, two or more continuous explanatory variables.

ANCOVA (Analysis of Covariance) Continuous response, categorical explanatory variable(s) and continuous explanatory variable(s).

If multiple explanatory variables are chosen, they may be analyzed with respect to their **main effects** on the model (i.e., their separate impacts on the variance explained) or with respect to their **interaction effects,** the effect of interacting explanatory variables on the model.

Review of Hypothesis Testing

We use hypothesis testing to (1) analyze evidence provided by data and (2) make decisions based on data. A **statistical hypothesis** is an assumption about a population parameter that may or may not be true (i.e. is the mean and variance of our sample representative of the mean and variance of the population?). We usually have **H0** the null hypothesis and **H1** the alternative hypothesis (opposite of the null hypothesis). The results of our test will tell us whether we can accept or reject the null hypothesis.

This is the basic procedure of hypothesis testing: 1. State the hypothesis and identify the claim 2. Find the critical value(s) from the appropriate table 3. Compute the test value 4. Make the decision to reject or not reject the null hypothesis. If P-value =< alpha, there is enough evidence to **reject** the null hypothesis. If P-value > alpha, there is not enough evidence to reject the null hypothesis, so we **do not reject** the null hypothesis.

Assumptions of the GLM

The GLM is based on the assumption that the data residuals approximate a normal distribution (or a linearly transformed normal distribution). We will discuss the non-parametric analogues to several of these tests if the assumptions of normality are violated. For tests that analyze categorical explanatory variables, the assumption is that the variance in the response variable is equal among groups. Note: environmental data often violate the assumptions of normality and equal variance, and we will often proceed with a GLM even if these assumptions are violated. In this situation, justifying the decision to proceed with a linear model must be made.

Set up

```
getwd()
```

```
## [1] "/Users/ataliefischer/Desktop/EDA/Environmental_Data_Analytics_2022/Lessons"
```

```
library(tidyverse)
options(scipen = 4)

PeterPaul.chem.nutrients <- read.csv("../Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Pro

# Set theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

Linear Regression

A linear regression is comprised of a continuous response variable, plus a combination of 1+ continuous response variables (plus the error term). The deterministic portion of the equation describes the response variable as lying on a straight line, with an intercept and a slope term. The equation is thus a typical algebraic expression:

$$y = \alpha + \beta * x + \epsilon$$

Regression is a technique for fitting a line to a set of data points. The goal for the linear regression is to find a **line of best fit**, which is the line drawn through the bivariate space that minimizes the total distance of points from the line. This is also called a "least squares" regression. The remainder of the variance not explained by the model is called the **residual error.**

The linear regression will test the null hypotheses that

1. The intercept (alpha) is equal to zero.

2. The slope (beta) is equal to zero

Whether or not we care about the result of each of these tested hypotheses will depend on our research question. Sometimes, the test for the intercept will be of interest, and sometimes it will not.

Important components of the linear regression are the correlation and the R-squared value. The **correlation** is a number between -1 and 1, describing the relationship between the variables. Correlations close to -1 represent strong negative correlations, correlations close to zero represent weak correlations, and correlations close to 1 represent strong positive correlations. The **R-squared value** is the correlation squared, becoming a number between 0 and 1. The R-squared value describes the percent of variance accounted for by the explanatory variables.

Simple Linear Regression

A **simple linear regression** is the simplest form of regression that involves a linear relationship between two variables. The object of a simple linear regression is to obtain an equation of a straight line that minimizes the sum of squared vertical deviations from the line (i.e., the least squares criterion).

The results of a simple linear regression will always show the standard error of the estimate. The **standard error** is a measure of the scatter of points around a regression line. If the standard error is relatively small, the predictions using the linear equation will tend to be more accurate than if the standard error is larger.

$$S_e = \sqrt{(\frac{\sum y - y_c^2}{n-2})}$$

where S_{e} = standard error of estimate y = y value of each data point n = number of data points

The **correlation coefficient, r** is a measure of the strength and direction of relationship between two variables, and ranges between -1.00 and +1.00.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{([n(\sum x)^2] * [n(\sum y^2) - (\sum y)^2])}}$$

There is a positive correlation when the slope is positive, no correlation when the slope is zero, and a negative correlation when the slope is negative.

r^2, square of the correlation coefficient is a measure of the percentage of variability in the values of y that is "explained" by the independent variable, and ranges between 0 and 1.00. An r^2 value close to 1 means that almost all the variation in the y-variable is explained by the x-variable.

After fitting a regression model, check the **residual** plots first to be sure that you have unbiased estimates. This will help us answer questions such as: * Do the residuals follow a normal distribution? * Do we have unbiased estimates? * Is there a correlation in our residual series? which will help us determine if we need more explanatory variables in our model.

The **p-values** for the coefficients indicate whether these relationships are statistically significant (i.e., determine whether the relationships that you observe in your sample also exist in the larger population). The p-value for each independent variable tests the null hypothesis (H0: independent variable as no correlation with the dependent variable). If the p-value for a variable is less than your significance level, your sample data provide enough evidence to reject the null hypothesis for the entire population (if the p-value < alpha, there is significant correlation with the dependent variable, and the addition of x to the model is worthwhile).

For the NTL-LTER dataset, can we predict irradiance (light level) from depth?

```
irradiance.regression <- lm(PeterPaul.chem.nutrients$irradianceWater ~ PeterPaul.chem.nutrients$depth)

# another way to format the lm() function
irradiance.regression <- lm(data = PeterPaul.chem.nutrients, irradianceWater ~ depth)
summary(irradiance.regression)
```

```
##
## Call:
## lm(formula = irradianceWater ~ depth, data = PeterPaul.chem.nutrients)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -458.9  -144.1   -41.2    90.3 23813.0
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  486.818      4.063  119.82   <2e-16 ***
## depth        -95.890      1.153  -83.14   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 303.4 on 15449 degrees of freedom
##   (7557 observations deleted due to missingness)
## Multiple R-squared:  0.3091, Adjusted R-squared:  0.3091
## F-statistic:  6912 on 1 and 15449 DF,  p-value: < 2.2e-16
```

```
# Correlation
cor.test(PeterPaul.chem.nutrients$irradianceWater, PeterPaul.chem.nutrients$depth)
```

```
##
##  Pearson's product-moment correlation
##
## data:  PeterPaul.chem.nutrients$irradianceWater and PeterPaul.chem.nutrients$depth
## t = -83.137, df = 15449, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.5667674 -0.5449776
## sample estimates:
##       cor
## -0.555968
```

Question: How would you report the results of this test (overall findings and report of statistical output)?

> H0: Irradiance is not related to depth. HA: Irradiance is related to depth. The line of best fit through our sample data is Irradiance = -95.890*Depth + 486.818, meaning that irradiance decreases with depth. With a p-value < 2.2e-16, our results are significant and there is sufficient evidence to reject the null hypothesis. The residual standard error is 303.4 and the r^2 value is 0.3091, meaning that approximately 30% of the variance in irradiance can be explained by depth. The correlation coefficient is -0.556, meaning that there is significant negative correlation between irradiance and depth.

So, we see there is a significant negative correlation between irradiance and depth (lower light levels at greater depths), and that this model explains about 31% of the total variance in irradiance. Let's visualize this relationship and the model itself.

An exploratory option to visualize the model fit is to use the function `plot`. This function will return four graphs, which are intended only for checking the fit of the model and not for communicating results. The plots that are returned are:

1. **Residuals vs. Fitted.** The value predicted by the line of best fit is the fitted value, and the residual is the distance of that actual value from the predicted value. By definition, there will be a balance of positive and negative residuals. Watch for drastic asymmetry from side to side or a marked departure from zero for the red line - these are signs of a poor model fit. > We want the red line to be as flat as

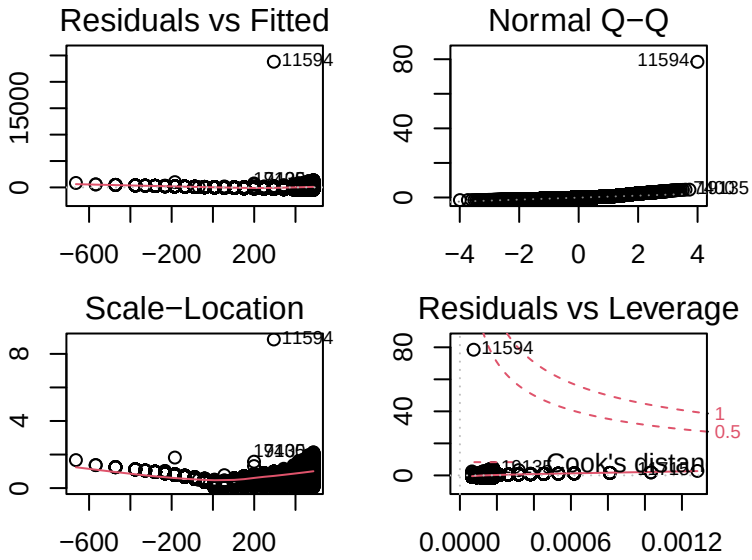
possible and symmetrical. Look for outliers!

2. **Normal Q-Q.** The points should fall close to the 1:1 line if the data follow a normal distribution. We often see departures from 1:1 at the high and low ends of the dataset, which could be outliers. > The dashed line shows the normal quantile-quantile for a normal distribution. We want our data to be as close to this line as possible. Look for outliers!

3. **Scale-Location.** Similar to the residuals vs. fitted graph, this will graph the squared standardized residuals by the fitted values. > Fitted values compared to the squared residuals. Look for symmetry in the red line.

4. **Residuals vs. Leverage.** This graph will display potential outliers. The values that fall outside the dashed red lines (Cook's distance) are outliers for the model. Watch for drastic departures of the solid red line from horizontal - this is a sign of a poor model fit. > This plot show outliers outside the two dashed red lines. Points that drastically deviate from the solid red line will indicate poor model fit.

```
par(mfrow = c(2,2), mar=c(2,2,2,2)) #mfrow --> divide plot screen into four cell in a 2x2 grid. mar -->
plot(irradiance.regression)
```

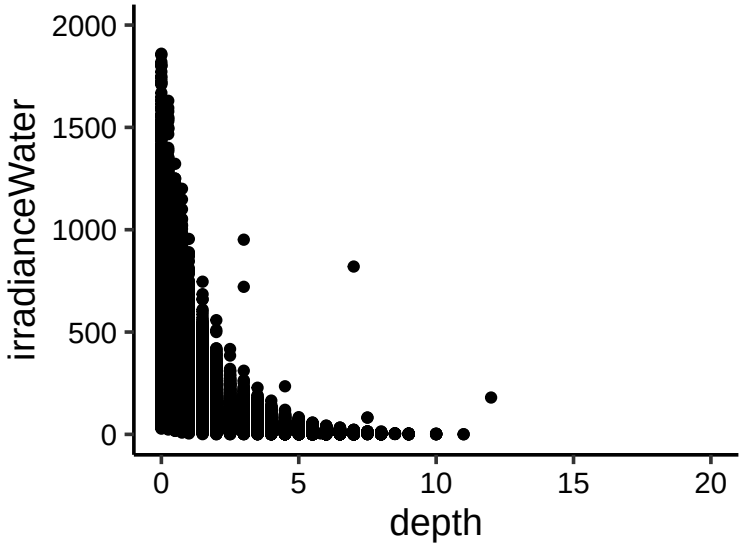



```
par(mfrow = c(1,1))
```

The option best suited for communicating findings is to plot the explanatory and response variables as a scatterplot.

```
# Plot the regression
irradiancebydepth <-
  ggplot(PeterPaul.chem.nutrients, aes(x = depth, y = irradianceWater)) +
  ylim(0, 2000) + #does not remove the outlier from the data, but easier to view the data without the o
  geom_point()
print(irradiancebydepth)
```

```
## Warning: Removed 7558 rows containing missing values (geom_point).
```

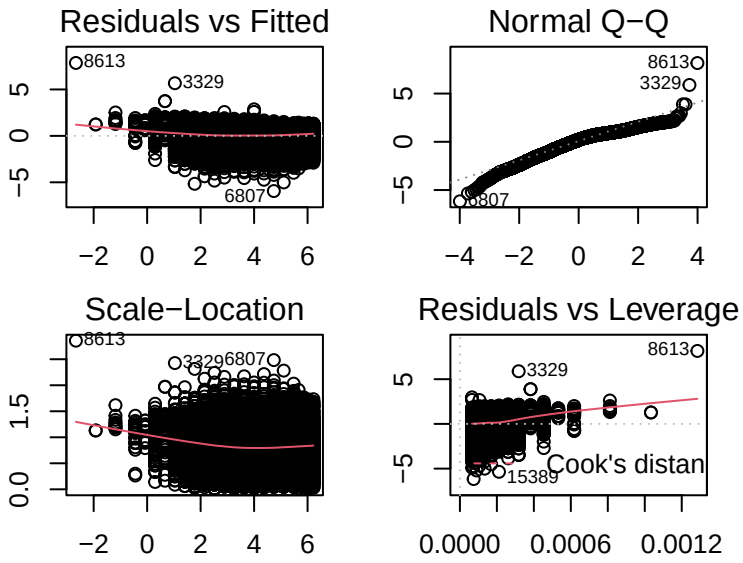


Given the distribution of irradiance values, we don't have a linear relationship between x and y in this case. The data is highly skewed right. Let's try log-transforming the irradiance values. Note we also remove the observations that seems to ba an outlier.

```
PeterPaul.chem.nutrients <- filter(PeterPaul.chem.nutrients,
                                   irradianceWater != 0 & irradianceWater < 5000) #remove outliers and
irradiance.regression2 <- lm(data = PeterPaul.chem.nutrients, log(irradianceWater) ~ depth)
summary(irradiance.regression2)
```

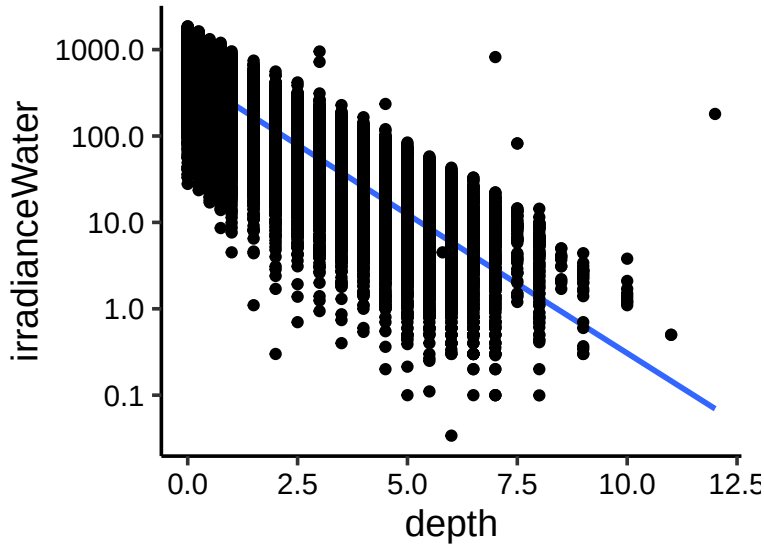
```
##
## Call:
## lm(formula = log(irradianceWater) ~ depth, data = PeterPaul.chem.nutrients)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9421 -0.5745  0.1930  0.7215  7.8568
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.218503   0.012903   481.9   <2e-16 ***
## depth       -0.740198   0.003664  -202.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9634 on 15445 degrees of freedom
## Multiple R-squared:  0.7255, Adjusted R-squared:  0.7254
## F-statistic: 4.081e+04 on 1 and 15445 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2), mar=c(2,2,2,2))
plot(irradiance.regression2)
```



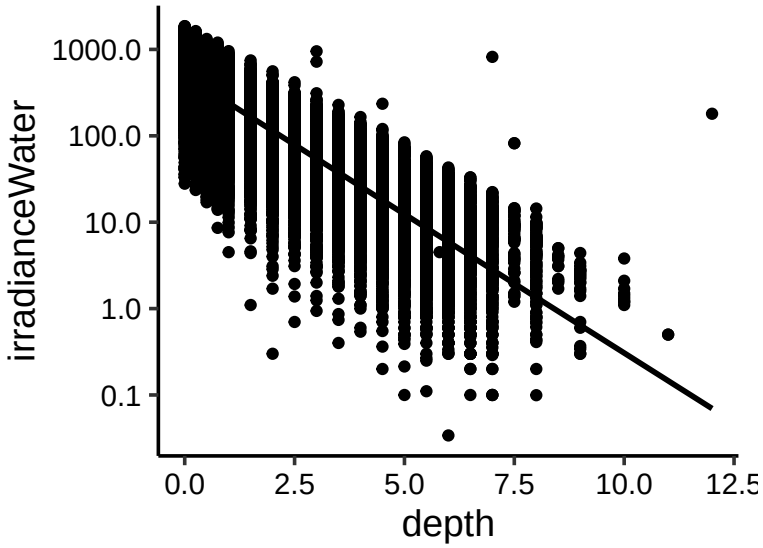

```
par(mfrow = c(1,1))

# Add a line and standard error for the linear regression
irradiancebydepth2 <-
  ggplot(PeterPaul.chem.nutrients, aes(x = depth, y = irradianceWater)) +
  geom_smooth(method = "lm") +
  scale_y_log10() +
  geom_point()
print(irradiancebydepth2)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# SE - confidence interval around smooth can also be removed
irradiancebydepth2 <-
    ggplot(PeterPaul.chem.nutrients, aes(x = depth, y = irradianceWater)) +
    geom_point() +
    scale_y_log10() +  #log transformed values for irradianceWater
    geom_smooth(method = 'lm', se = FALSE, color = "black") #remove confidence interval on plot
print(irradiancebydepth2)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# Make the graph attractive
```

Multiple Linear Regression

It is possible, and often useful, to consider multiple continuous explanatory variables at a time in a linear regression. For example, total phosphorus concentration in Paul Lake (the unfertilized lake) could be dependent on depth and dissolved oxygen concentration:

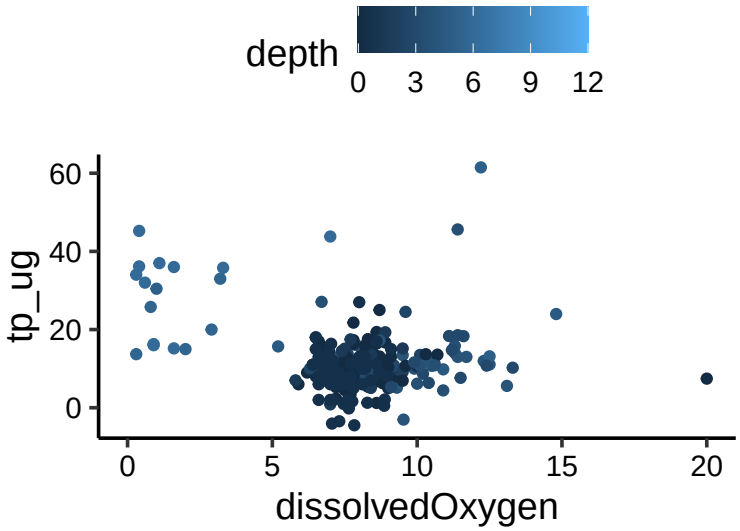
```
TPregression <- lm(data = subset(PeterPaul.chem.nutrients, lakename == "Paul Lake"),
                   tp_ug ~ depth + dissolvedOxygen)
summary(TPregression)
```

```
##
## Call:
## lm(formula = tp_ug ~ depth + dissolvedOxygen, data = subset(PeterPaul.chem.nutrients,
##     lakename == "Paul Lake"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.266  -3.436  -0.534   2.425  44.559
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      12.7869     1.4985   8.533 8.27e-16 ***
## depth             2.1691     0.2222   9.762  < 2e-16 ***
## dissolvedOxygen  -0.5494     0.1726  -3.184  0.00161 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.405 on 288 degrees of freedom
##   (7283 observations deleted due to missingness)
## Multiple R-squared:  0.2985, Adjusted R-squared:  0.2936
## F-statistic: 61.28 on 2 and 288 DF,  p-value: < 2.2e-16
```

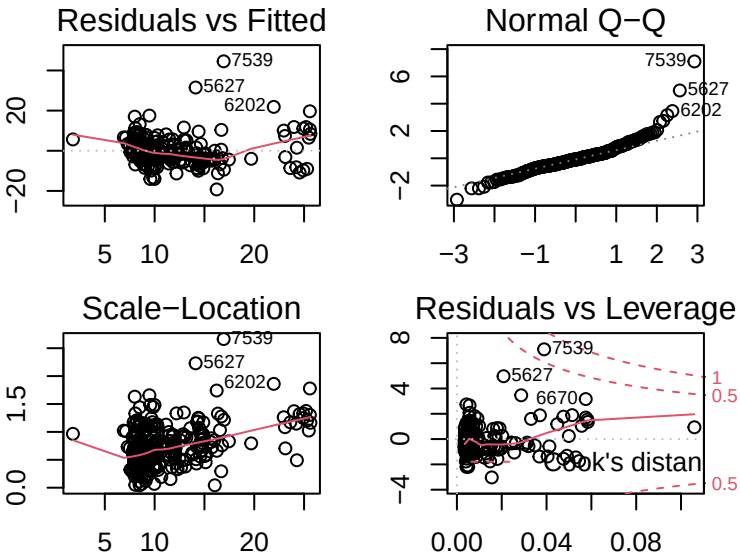
```
TPplot <- ggplot(subset(PeterPaul.chem.nutrients, lakename == "Paul Lake"),
                 aes(x = dissolvedOxygen, y = tp_ug, color = depth)) +
```

```
  geom_point() +
  xlim(0, 20)
print(TPplot)
```

```
## Warning: Removed 7283 rows containing missing values (geom_point).
```




```
par(mfrow = c(2,2), mar=c(2,2,2,2))
plot(TPregression)
```




```
par(mfrow = c(1,1))
```

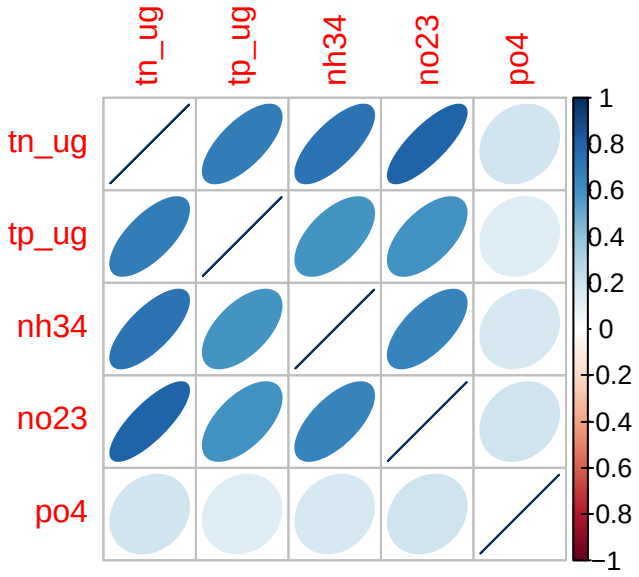
Correlation Plots

We can also make exploratory plots of several continuous data points to determine possible relationships, as well as covariance among explanatory variables.

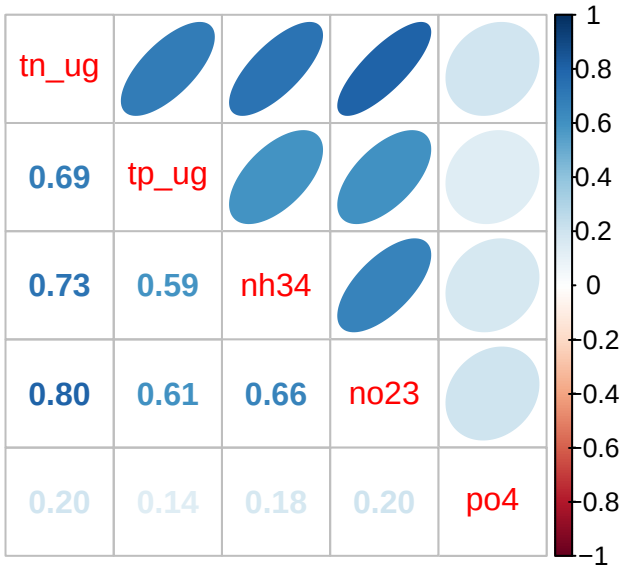
```
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
PeterPaulnutrients <-
  PeterPaul.chem.nutrients %>%
  select(tn_ug:po4) %>%
  na.omit()

# look at correlation coefficients to find additional variables to add to the model
PeterPaulCorr <- cor(PeterPaulnutrients)
corrplot(PeterPaulCorr, method = "ellipse") # correlation plot matrix. Diagonal line represents perfect
```




```
corrplot.mixed(PeterPaulCorr, upper = "ellipse") # mixed correlation plot of both ellipses and numbers
```



AIC to select variables

However, it is possible to over-parameterize a linear model. Adding additional explanatory variables takes away degrees of freedom, and if explanatory variables co-vary the interpretation can become overly complicated. Remember, an ideal statistical model balances simplicity and explanatory power! To help with this trade-off, we can use the **Akaike's Information Criterion (AIC)** to compute a stepwise regression that either

adds explanatory variables from the bottom up or removes explanatory variables from a full set of suggested options. The smaller the AIC value, the better.

Let's say we want to know which explanatory variables will allow us to best predict total phosphorus concentrations. Potential explanatory variables from the dataset could include depth, dissolved oxygen, temperature, PAR, total N concentration, and phosphate concentration.

```
Paul.naomit <- PeterPaul.chem.nutrients %>%
  filter(lakename == "Paul Lake") %>%
  na.omit()

TPAIC <- lm(data = Paul.naomit, tp_ug ~ depth + dissolvedOxygen +
              temperature_C + tn_ug + po4) #consider all variables to be explanatory variables

#Choose a model by AIC in a Stepwise Algorithm
step(TPAIC)
```

```
## Start:  AIC=353.95
## tp_ug ~ depth + dissolvedOxygen + temperature_C + tn_ug + po4
##
##                   Df Sum of Sq    RSS    AIC
## - po4              1     1.111 2330.8 352.00
## - depth            1    16.739 2346.4 352.76
## - tn_ug            1    30.674 2360.3 353.43
## <none>                         2329.7 353.95
## - temperature_C    1   178.398 2508.1 360.29
## - dissolvedOxygen  1   232.194 2561.9 362.68
##
## Step:  AIC=352
## tp_ug ~ depth + dissolvedOxygen + temperature_C + tn_ug
##
##                   Df Sum of Sq    RSS    AIC
## - depth            1    15.745 2346.5 350.76
## - tn_ug            1    38.270 2369.1 351.84
## <none>                         2330.8 352.00
## - temperature_C    1   194.143 2524.9 359.04
## - dissolvedOxygen  1   250.981 2581.8 361.56
##
## Step:  AIC=350.76
## tp_ug ~ dissolvedOxygen + temperature_C + tn_ug
##
##                   Df Sum of Sq    RSS    AIC
## <none>                         2346.5 350.76
## - tn_ug            1     46.09 2392.6 350.96
## - dissolvedOxygen  1    305.67 2652.2 362.60
## - temperature_C    1    562.60 2909.1 373.05
```

```
##
## Call:
## lm(formula = tp_ug ~ dissolvedOxygen + temperature_C + tn_ug,
##     data = Paul.naomit)
##
## Coefficients:
##     (Intercept)  dissolvedOxygen    temperature_C            tn_ug
##       23.905781        -0.834405        -0.487603         0.007002
```

```r
TPmodel <- lm(data = Paul.naomit, tp_ug ~ dissolvedOxygen + temperature_C + tn_ug)
summary(TPmodel) # removing variables to see how much the AIC reduces. oxygen, temperature, and nitroge
```

```
##
## Call:
## lm(formula = tp_ug ~ dissolvedOxygen + temperature_C + tn_ug,
##     data = Paul.naomit)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.9443  -2.8674  -0.4834   2.1666  14.4713
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     23.905781   3.586389   6.666 1.12e-09 ***
## dissolvedOxygen -0.834405   0.221438  -3.768 0.000267 ***
## temperature_C   -0.487603   0.095382  -5.112 1.37e-06 ***
## tn_ug            0.007002   0.004786   1.463 0.146304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.64 on 109 degrees of freedom
## Multiple R-squared:  0.2774, Adjusted R-squared:  0.2575
## F-statistic: 13.95 on 3 and 109 DF,  p-value: 0.00000009208
```