Assignment 6: Generalized Linear Models

Lindsay Roth

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics (ENV872L) on generalized linear models.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Use the lesson as a guide. It contains code that can be modified to complete the assignment.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file. You will need to have the correct software installed to do this (see Software Installation Guide) Press the `Knit` button in the RStudio scripting panel. This will save the PDF output in your Assignments folder.
6. After Knitting, please submit the completed exercise (PDF file) to the dropbox in Sakai. Please add your last name into the file name (e.g., "Salk_A06_GLMs.pdf") prior to submission.

The completed exercise is due on Tuesday, 26 February, 2019 before class begins.

Set up your session

1. Set up your session. Upload the EPA Ecotox dataset for Neonicotinoids and the NTL-LTER raw data file for chemistry/physics.
2. Build a ggplot theme and set it as your default theme.

```
#1
getwd()
```

```
## [1] "/Users/lindsayroth/Documents/MEM 1st Year/Spring 2019/Env_Data_Analytics/Env_Data_Analytics"
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------- ti
```

```
## v ggplot2 3.1.0     v purrr   0.3.0
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.3.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------------------------------------- tidyvers
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
library(RColorBrewer)
library(colormap)

Ecotox <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Mortality_raw.csv")
NTL.chem.physics <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

#2
lindsay_theme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
theme_set(lindsay_theme)
```

Neonicotinoids test

Research question: Were studies on various neonicotinoid chemicals conducted in different years?

3. Generate a line of code to determine how many different chemicals are listed in the Chemical.Name column.
4. Are the publication years associated with each chemical well-approximated by a normal distribution? Run the appropriate test and also generate a frequency polygon to illustrate the distribution of counts for each year, divided by chemical name. Bonus points if you can generate the results of your test from a pipe function. No need to make this graph pretty.
5. Is there equal variance among the publication years for each chemical? Hint: var.test is not the correct function.

```
#3
unique(Ecotox$Chemical.Name)
```

```
## [1] Imidacloprid Thiacloprid  Thiamethoxam Acetamiprid  Clothianidin
## [6] Dinotefuran  Nitenpyram   Nithiazine   Imidaclothiz
## 9 Levels: Acetamiprid Clothianidin Dinotefuran ... Thiamethoxam
```

```
#There are 9 different chemicals in the Chemical.Name column.

#4
shapiro.test(Ecotox$Pub..Year[Ecotox$Chemical.Name == "Imidacloprid"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ecotox$Pub..Year[Ecotox$Chemical.Name == "Imidacloprid"]
## W = 0.88178, p-value < 2.2e-16
```

```
shapiro.test(Ecotox$Pub..Year[Ecotox$Chemical.Name == "Thiacloprid"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ecotox$Pub..Year[Ecotox$Chemical.Name == "Thiacloprid"]
## W = 0.7669, p-value = 1.118e-11
```

```
shapiro.test(Ecotox$Pub..Year[Ecotox$Chemical.Name == "Thiamethoxam"])
```

```
##
##  Shapiro-Wilk normality test
```

```
##
## data:  Ecotox$Pub..Year[Ecotox$Chemical.Name == "Thiamethoxam"]
## W = 0.7071, p-value < 2.2e-16
```

```
shapiro.test(Ecotox$Pub..Year[Ecotox$Chemical.Name == "Acetamiprid"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ecotox$Pub..Year[Ecotox$Chemical.Name == "Acetamiprid"]
## W = 0.90191, p-value = 5.706e-08
```

```
shapiro.test(Ecotox$Pub..Year[Ecotox$Chemical.Name == "Clothianidin"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ecotox$Pub..Year[Ecotox$Chemical.Name == "Clothianidin"]
## W = 0.69577, p-value = 4.287e-11
```

```
shapiro.test(Ecotox$Pub..Year[Ecotox$Chemical.Name == "Dinotefuran"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ecotox$Pub..Year[Ecotox$Chemical.Name == "Dinotefuran"]
## W = 0.82848, p-value = 8.83e-07
```

```
shapiro.test(Ecotox$Pub..Year[Ecotox$Chemical.Name == "Nitenpyram"])
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Ecotox$Pub..Year[Ecotox$Chemical.Name == "Nitenpyram"]
## W = 0.79592, p-value = 0.0005686
```

```
shapiro.test(Ecotox$Pub..Year[Ecotox$Chemical.Name == "Nithiazine"])
```

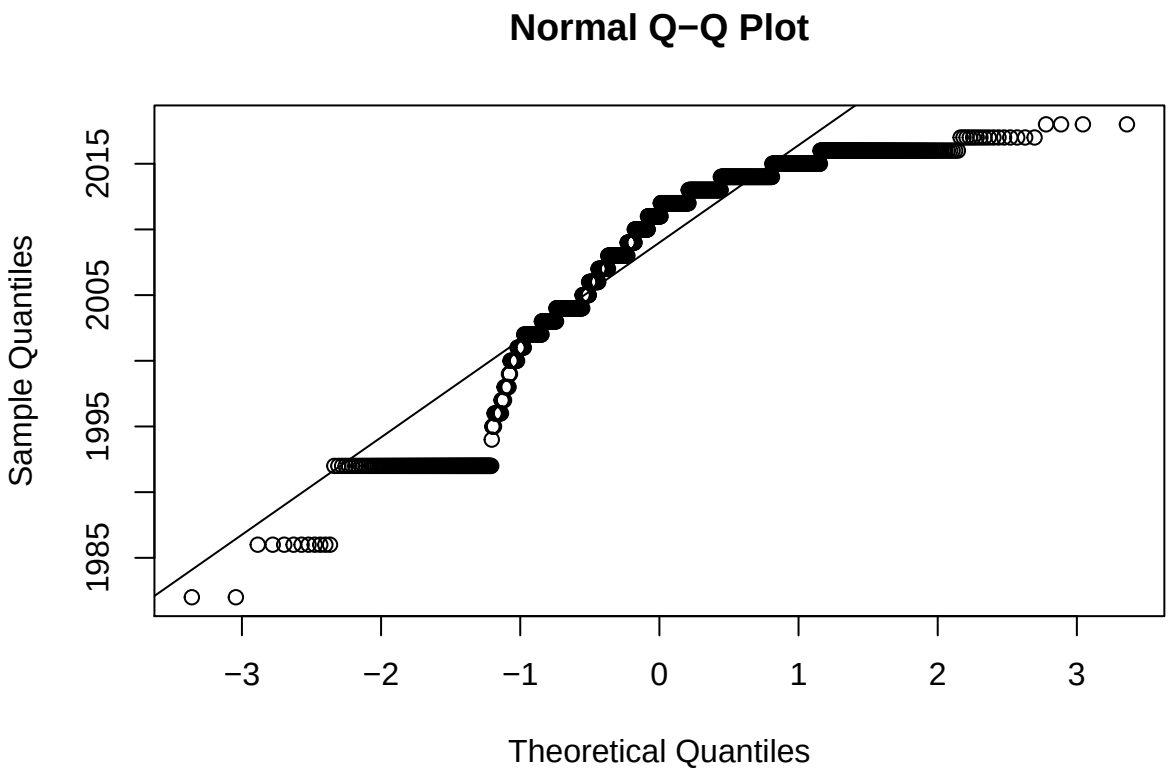
```
##
##  Shapiro-Wilk normality test
##
## data:  Ecotox$Pub..Year[Ecotox$Chemical.Name == "Nithiazine"]
## W = 0.75938, p-value = 0.0001235
```

```
shapiro.test(Ecotox$Pub..Year[Ecotox$Chemical.Name == "Imidaclothiz"])
```

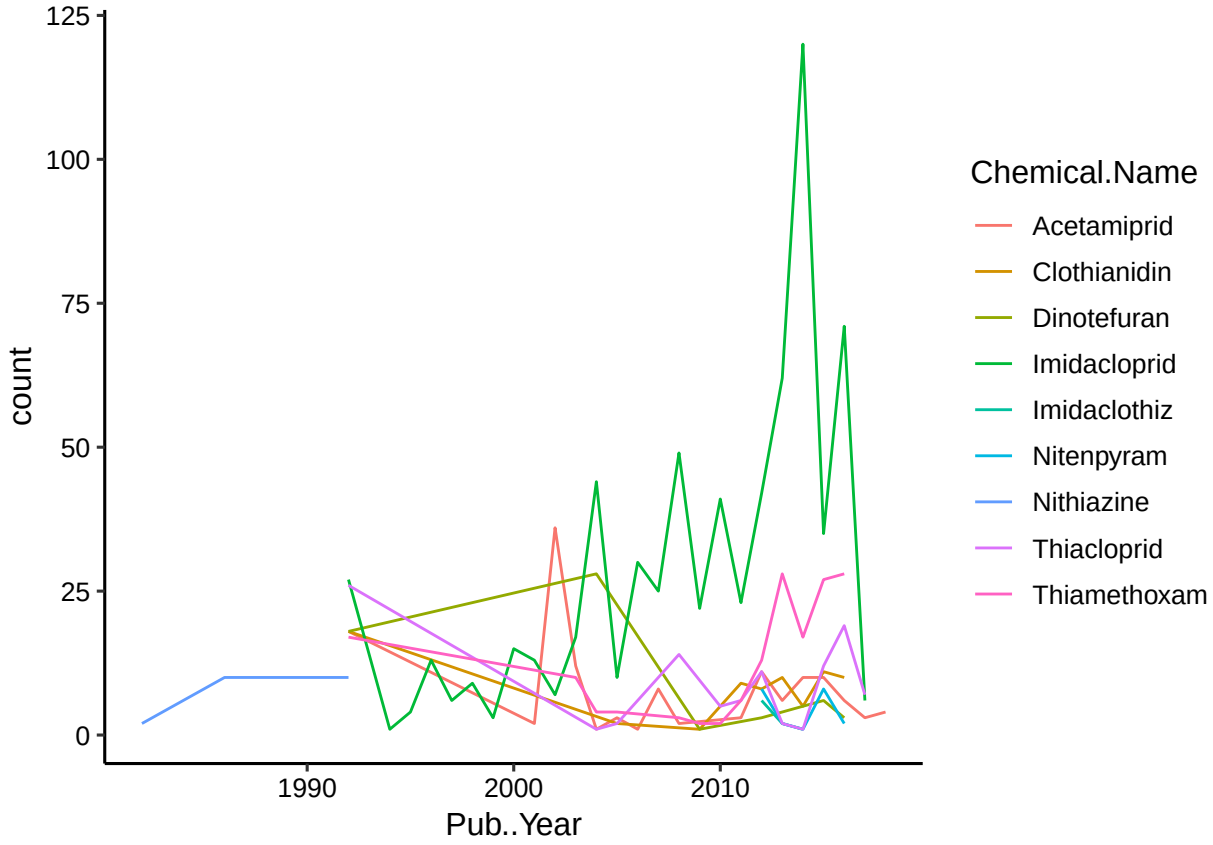
```
##
##  Shapiro-Wilk normality test
##
## data:  Ecotox$Pub..Year[Ecotox$Chemical.Name == "Imidaclothiz"]
## W = 0.68429, p-value = 0.00093
```

```
#The publication years associated with each chemical are not well approximated
#by the normal distribution because all of the p-values resulting from the
#Shapiro-Wilkes test are less than 0.001

qqnorm(Ecotox$Pub..Year); qqline(Ecotox$Pub..Year)
```




```
Ecotox.Freq <- ggplot(Ecotox, aes(x = Pub..Year, color = Chemical.Name)) +
  geom_freqpoly(stat = "count")
print(Ecotox.Freq)
```




```
#5
bartlett.test(Ecotox$Pub..Year ~ Ecotox$Chemical.Name)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Ecotox$Pub..Year by Ecotox$Chemical.Name
## Bartlett's K-squared = 139.59, df = 8, p-value < 2.2e-16
```

```
#There is not equal variance among the publication years for each chemical (p<0.001).
```

6. Based on your results, which test would you choose to run to answer your research question?

   ANSWER: Because the distribution of the data is not normal, I would run a non-parametric test to see if the studies on the various chemicals were conducted in different years. Since the variables are both categorical and there are more than 2, I will use the non-paramentric ANOVA test, Kruskal-Wallis.

7. Run this test below.

8. Generate a boxplot representing the range of publication years for each chemical. Adjust your graph to make it pretty.

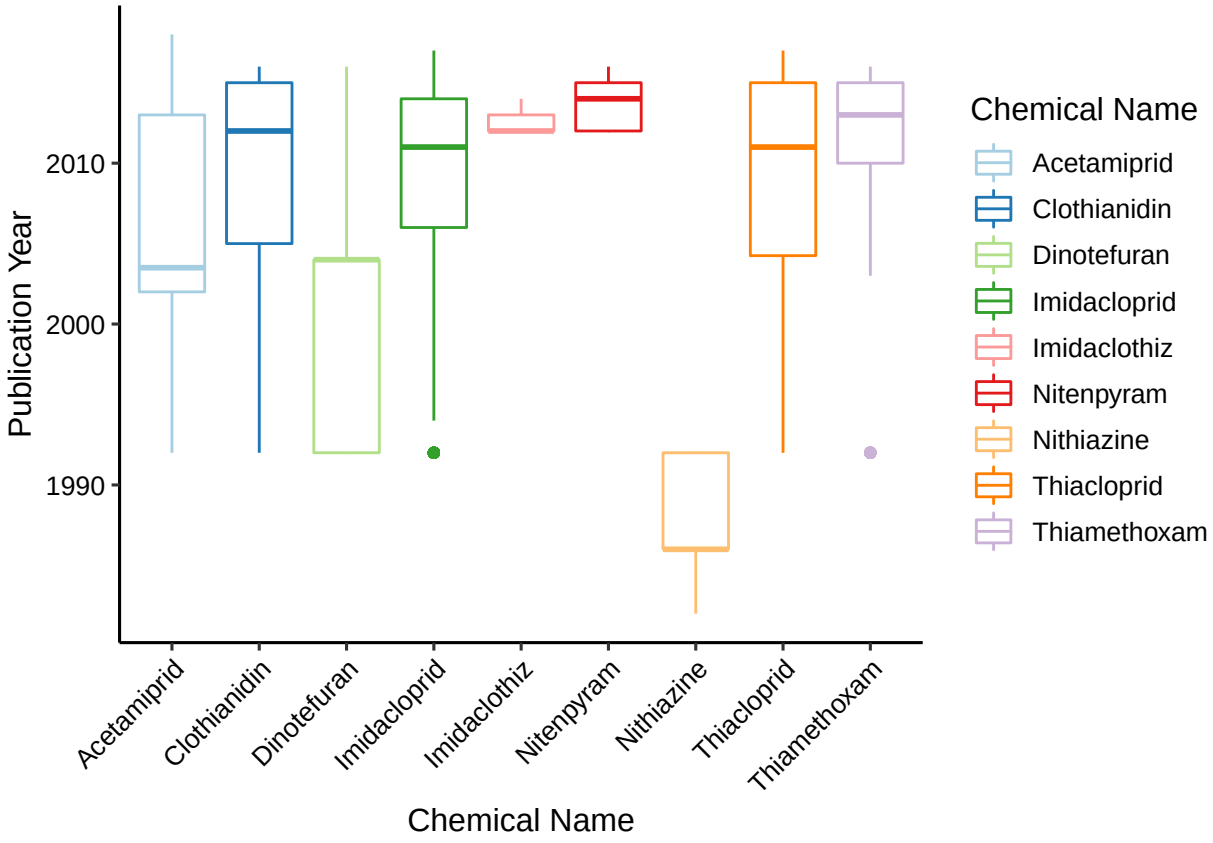
```
#7
Ecotox.test <- kruskal.test(Ecotox$Pub..Year ~ Ecotox$Chemical.Name)
Ecotox.test
```

```
##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  Ecotox$Pub..Year by Ecotox$Chemical.Name
## Kruskal-Wallis chi-squared = 134.15, df = 8, p-value < 2.2e-16
```

```
#There is a significant difference between the publication years of each chemical
#(ANOVA, Kruskal-Wallis; df = 8, p<0.001)

#8
Ecotox.plot <-
  ggplot(Ecotox, aes(x = Chemical.Name, y = Pub..Year, color = Chemical.Name)) +
  geom_boxplot() +
  scale_color_brewer(palette = "Paired") +
  labs(x = "Chemical Name", y = "Publication Year", color = "Chemical Name") +
  theme(axis.text.x = element_text(angle = 45,  hjust = 1))
print(Ecotox.plot)
```



9. How would you summarize the conclusion of your analysis? Include a sentence summarizing your findings and include the results of your test in parentheses at the end of the sentence.

   ANSWER: Studies on various neonicotinoid chemicals were conducted in significantly different years (ANOVA, Kruskal-Wallis;X^2 = 134.15, df = 8, p<0.001).

NTL-LTER test

Research question: What is the best set of predictors for lake temperatures in July across the monitoring period at the North Temperate Lakes LTER?

11. Wrangle your NTL-LTER dataset with a pipe function so that it contains only the following criteria:

- Only dates in July (hint: use the daynum column). No need to consider leap years.
- Only the columns: lakename, year4, daynum, depth, temperature_C
- Only complete cases (i.e., remove NAs)

12. Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to predict temperature. Run a multiple regression on the recommended set of variables.

```
#11
NTL.wrangled <-
  NTL.chem.physics %>%
  filter(daynum > 182 & daynum < 214) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()

#12
NTL.test <- lm(data = NTL.wrangled, temperature_C ~ year4 + daynum + depth)
step(NTL.test)
```

```
## Start:  AIC=25998.22
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                142056 25998
## - year4   1       201 142257 26010
## - daynum  1      1237 143293 26080
## - depth   1    402549 544605 38995
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.wrangled)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##   -18.19700      0.01611      0.04024     -1.94133
```

```
summary(NTL.test)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.wrangled)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6857 -3.0267  0.1055  2.9937 13.6038
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -18.196998   8.741236   -2.082 0.037392 *
## year4         0.016113   0.004353    3.701 0.000216 ***
## daynum        0.040237   0.004385    9.176  < 2e-16 ***
## depth        -1.941328   0.011728 -165.528  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.833 on 9669 degrees of freedom
## Multiple R-squared:  0.7398, Adjusted R-squared:  0.7397
```

```
## F-statistic:  9162 on 3 and 9669 DF,  p-value: < 2.2e-16
```

13. What is the final linear equation to predict temperature from your multiple regression? How much of the observed variance does this model explain?

    ANSWER: Temperature = -5.3 + 0.009(year) + 0.04(daynum) - 1.9(depth) This model explains 73.97% of the variation in temperature for the lakes in July.

14. Run an interaction effects ANCOVA to predict temperature based on depth and lakename from the same wrangled dataset.

```
#14
NTL.interact <- lm(data = NTL.wrangled, temperature_C ~ depth*lakename)
step(NTL.interact)
```

```
## Start:  AIC=24110.68
## temperature_C ~ depth * lakename
##
##                   Df Sum of Sq    RSS   AIC
## <none>                          116535 24111
## - depth:lakename  8      4674 121209 24475
```

```
##
## Call:
## lm(formula = temperature_C ~ depth * lakename, data = NTL.wrangled)
##
## Coefficients:
##                       (Intercept)                              depth
##                           22.8748                            -2.5543
##          lakenameCrampton Lake          lakenameEast Long Lake
##                            2.5625                            -4.2925
##       lakenameHummingbird Lake               lakenamePaul Lake
##                           -2.6059                             0.7623
##             lakenamePeter Lake            lakenameTuesday Lake
##                            0.4321                            -2.8349
##              lakenameWard Lake          lakenameWest Long Lake
##                            2.4887                            -2.3347
##    depth:lakenameCrampton Lake    depth:lakenameEast Long Lake
##                            0.7704                             0.9181
## depth:lakenameHummingbird Lake         depth:lakenamePaul Lake
##                           -0.5692                             0.3698
##       depth:lakenamePeter Lake      depth:lakenameTuesday Lake
##                            0.5495                             0.6462
##        depth:lakenameWard Lake    depth:lakenameWest Long Lake
##                           -0.7207                             0.7870
```

```
summary(NTL.interact)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth * lakename, data = NTL.wrangled)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.683 -2.907 -0.290  2.795 16.336
##
## Coefficients:
```

```
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      22.8748     0.5657  40.435  < 2e-16 ***
## depth                            -2.5543     0.2330 -10.962  < 2e-16 ***
## lakenameCrampton Lake             2.5625     0.6516   3.932 8.47e-05 ***
## lakenameEast Long Lake           -4.2925     0.5992  -7.164 8.40e-13 ***
## lakenameHummingbird Lake         -2.6059     0.8262  -3.154  0.00161 **
## lakenamePaul Lake                 0.7623     0.5787   1.317  0.18779
## lakenamePeter Lake                0.4321     0.5773   0.749  0.45412
## lakenameTuesday Lake             -2.8349     0.5862  -4.836 1.35e-06 ***
## lakenameWard Lake                 2.4887     0.8298   2.999  0.00271 **
## lakenameWest Long Lake           -2.3347     0.5974  -3.908 9.36e-05 ***
## depth:lakenameCrampton Lake       0.7704     0.2379   3.239  0.00121 **
## depth:lakenameEast Long Lake      0.9181     0.2352   3.903 9.57e-05 ***
## depth:lakenameHummingbird Lake   -0.5692     0.2879  -1.977  0.04809 *
## depth:lakenamePaul Lake           0.3698     0.2341   1.580  0.11417
## depth:lakenamePeter Lake          0.5495     0.2338   2.350  0.01878 *
## depth:lakenameTuesday Lake        0.6462     0.2345   2.755  0.00587 **
## depth:lakenameWard Lake          -0.7207     0.2795  -2.578  0.00995 **
## depth:lakenameWest Long Lake      0.7870     0.2351   3.347  0.00082 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.474 on 9655 degrees of freedom
## Multiple R-squared:  0.7865, Adjusted R-squared:  0.7861
## F-statistic:  2093 on 17 and 9655 DF,  p-value: < 2.2e-16
```

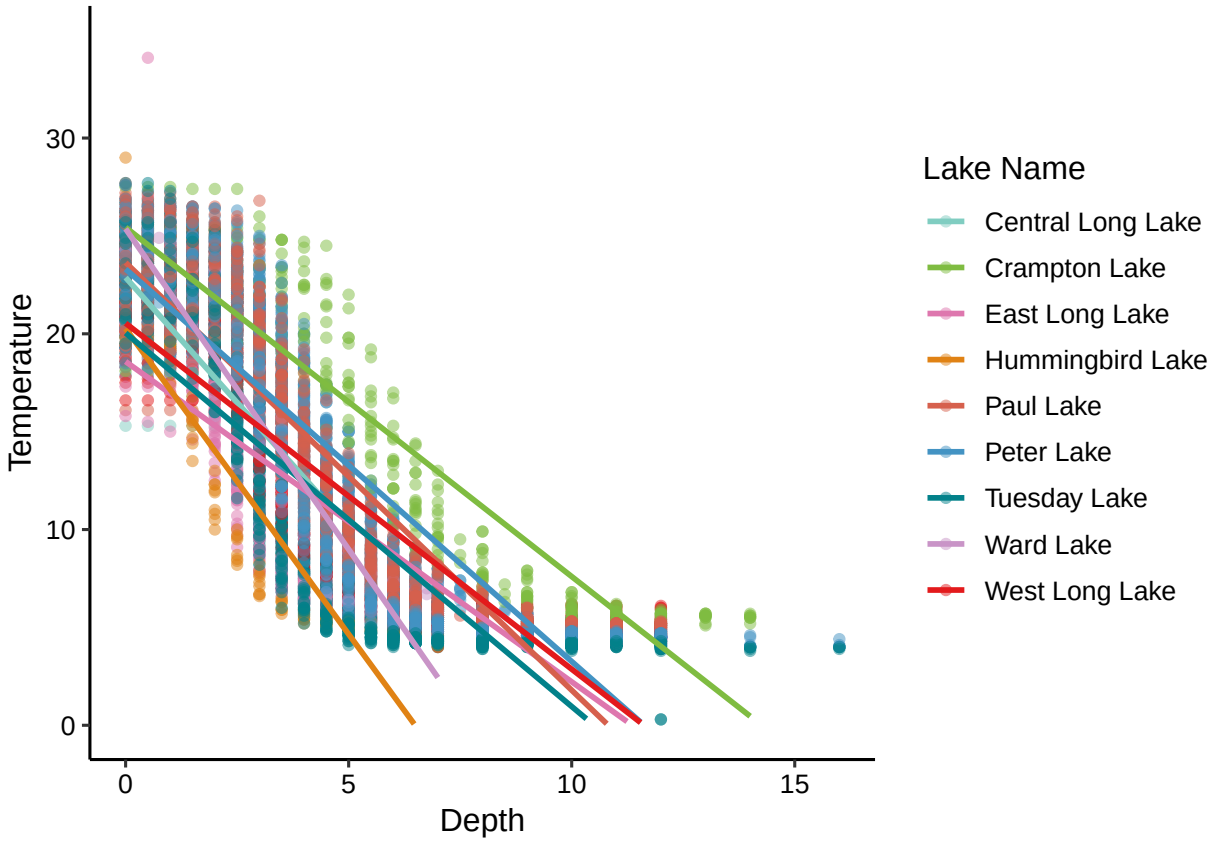
15. Is there an interaction between depth and lakename? How much variance in the temperature observations does this explain?

    ANSWER: There is a significant interaction between depth and lakename. This model explains 78.61% of the variation in tempearture observations.

16. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#16
NTL.plot <-
  ggplot(NTL.wrangled, aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  ylim(0,35) +
  labs(x = "Depth", y = "Temperature", color = "Lake Name")+
  scale_color_manual(values =
    c("#80cdc1", "#7fbc41", "#de77ae", "#e08214", "#d6604d", "#4393c3", "#02818a", "#c994c7", "#e31a1c"
print(NTL.plot)
```

```
## Warning: Removed 72 rows containing missing values (geom_smooth).
```


Temperature
30
20
10
0
0
5
10
15
Depth
Lake Name
Central Long Lake
Crampton Lake
East Long Lake
Hummingbird Lake
Paul Lake
Peter Lake
Tuesday Lake
Ward Lake
West Long Lake