Assignment 6: Generalized Linear Models

Yifei Zhang

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/yifeizhang/R/Environmental Data Analytics"
```

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## 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 -------------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(colormap)
EPAEcotox <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Mortality_raw.csv")
NTL_LTER_Lake <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
#2
Yifeitheme <- theme_light(base_size = 14) +
```

```
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
theme_set(Yifeitheme)
```

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
length(unique(EPAEcotox$Chemical.Name))
```

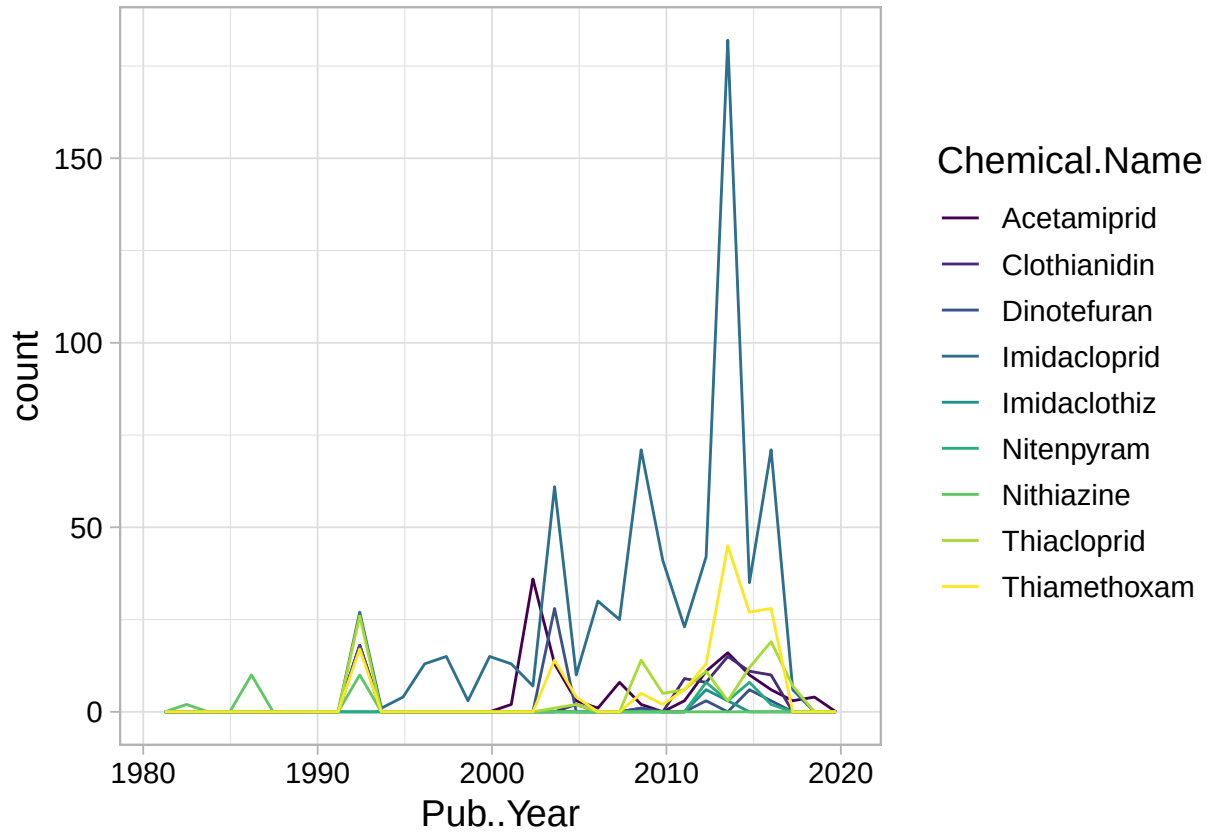
```
## [1] 9
```

```
#4
EPAEcotox %>%
group_by(Chemical.Name)%>%
summarise(statistic=shapiro.test(Pub..Year)$statistic, p.value = shapiro.test(Pub..Year)$p.value)
```

```
## # A tibble: 9 x 3
##   Chemical.Name statistic  p.value
##   <fct>             <dbl>    <dbl>
## 1 Acetamiprid       0.902 5.71e- 8
## 2 Clothianidin      0.696 4.29e-11
## 3 Dinotefuran       0.828 8.83e- 7
## 4 Imidacloprid      0.882 1.38e-22
## 5 Imidaclothiz      0.684 9.30e- 4
## 6 Nitenpyram        0.796 5.69e- 4
## 7 Nithiazine        0.759 1.24e- 4
## 8 Thiacloprid       0.767 1.12e-11
## 9 Thiamethoxam      0.707 1.57e-16
```

```
  ggplot(EPAEcotox) +
  geom_freqpoly(aes(x = Pub..Year, color = Chemical.Name)) +
  scale_color_colormap(discrete = T)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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

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

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

   ANSWER: I would choose one-way ANOVA.

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
PubYear.anova <- lm(EPAEcotox$Pub..Year ~ EPAEcotox$Chemical.Name)
summary(PubYear.anova)
```

```
##
## Call:
## lm(formula = EPAEcotox$Pub..Year ~ EPAEcotox$Chemical.Name)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.366  -3.993   1.889   4.889  13.441
```

```
##
## Coefficients:
##                                    Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                       2005.9926     0.6082 3298.222  < 2e-16
## EPAEcotox$Chemical.NameClothianidin    2.0479     1.0246    1.999  0.04584
## EPAEcotox$Chemical.NameDinotefuran    -3.4333     1.1057   -3.105  0.00194
## EPAEcotox$Chemical.NameImidacloprid    3.1181     0.6651    4.689 3.05e-06
## EPAEcotox$Chemical.NameImidaclothiz    6.4518     2.4412    2.643  0.00832
## EPAEcotox$Chemical.NameNitenpyram      7.7216     1.6630    4.643 3.78e-06
## EPAEcotox$Chemical.NameNithiazine    -17.6290     1.6299  -10.816  < 2e-16
## EPAEcotox$Chemical.NameThiacloprid     1.6394     0.9190    1.784  0.07467
## EPAEcotox$Chemical.NameThiamethoxam    4.3738     0.8261    5.295 1.40e-07
##
## (Intercept)                       ***
## EPAEcotox$Chemical.NameClothianidin *
## EPAEcotox$Chemical.NameDinotefuran  **
## EPAEcotox$Chemical.NameImidacloprid ***
## EPAEcotox$Chemical.NameImidaclothiz **
## EPAEcotox$Chemical.NameNitenpyram   ***
## EPAEcotox$Chemical.NameNithiazine   ***
## EPAEcotox$Chemical.NameThiacloprid  .
## EPAEcotox$Chemical.NameThiamethoxam ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.093 on 1274 degrees of freedom
## Multiple R-squared:  0.1726, Adjusted R-squared:  0.1674
## F-statistic: 33.21 on 8 and 1274 DF,  p-value: < 2.2e-16
```

```
PubYear.anova2 <- aov(EPAEcotox$Pub..Year ~ EPAEcotox$Chemical.Name)
summary(PubYear.anova2)
```

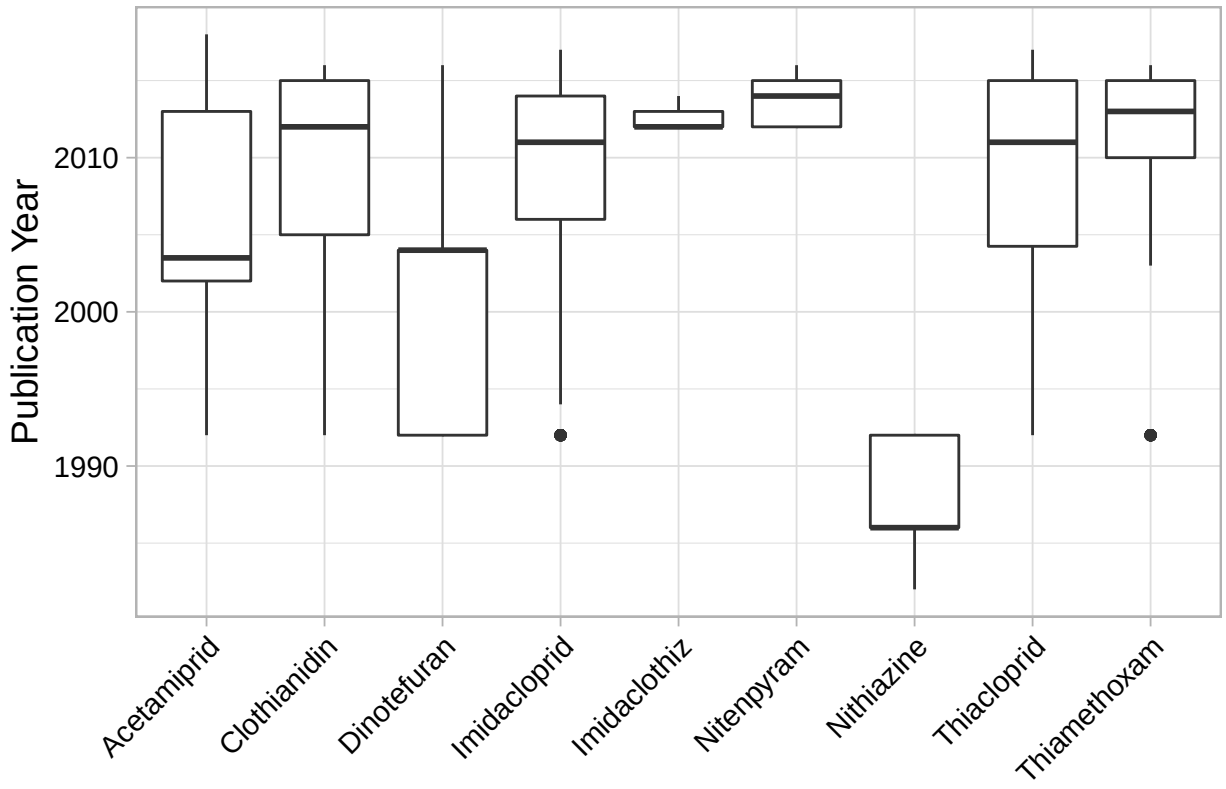
```
##                            Df Sum Sq Mean Sq F value Pr(>F)
## EPAEcotox$Chemical.Name     8  13365  1670.7   33.21 <2e-16 ***
## Residuals                1274  64093    50.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(PubYear.anova2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = EPAEcotox$Pub..Year ~ EPAEcotox$Chemical.Name)
##
## $`EPAEcotox$Chemical.Name`
##                                  diff          lwr           upr     p adj
## Clothianidin-Acetamiprid    2.0478935  -1.13556203   5.231348994 0.5444735
## Dinotefuran-Acetamiprid    -3.4333250  -6.86887521   0.002225165 0.0502982
## Imidacloprid-Acetamiprid    3.1181443   1.05175043   5.184538190 0.0001059
## Imidaclothiz-Acetamiprid    6.4517974  -1.13341497  14.037009746 0.1700689
## Nitenpyram-Acetamiprid      7.7216387   2.55455876  12.888718547 0.0001312
## Nithiazine-Acetamiprid    -17.6290107 -22.69334307 -12.564678323 0.0000000
## Thiacloprid-Acetamiprid     1.6394284  -1.21592420   4.494781028 0.6929485
## Thiamethoxam-Acetamiprid    4.3738126   1.80714109   6.940484045 0.0000050
```

```
## Dinotefuran-Clothianidin   -5.4812185  -9.32765358  -1.634783428 0.0003529
## Imidacloprid-Clothianidin   1.0702508  -1.62456668   3.765068330 0.9489438
## Imidaclothiz-Clothianidin   4.4039039  -3.37603853  12.183846336 0.7094335
## Nitenpyram-Clothianidin     5.6737452   0.22482121  11.122669133 0.0338611
## Nithiazine-Clothianidin   -19.6769042 -25.02849460 -14.325313751 0.0000000
## Thiacloprid-Clothianidin   -0.4084651  -3.74689228   2.929962144 0.9999879
## Thiamethoxam-Clothianidin   2.3259191  -0.76921583   5.421054007 0.3218154
## Imidacloprid-Dinotefuran    6.5514693   3.56304877   9.539889900 0.0000000
## Imidaclothiz-Dinotefuran    9.8851224   1.99867071  17.771574107 0.0033119
## Nitenpyram-Dinotefuran     11.1549637   5.55501829  16.754909074 0.0000000
## Nithiazine-Dinotefuran    -14.1956857 -19.70096824  -8.690403099 0.0000000
## Thiacloprid-Dinotefuran     5.0727534   1.49312883   8.652378050 0.0003937
## Thiamethoxam-Dinotefuran    7.8071376   4.45326278  11.161012409 0.0000000
## Imidaclothiz-Imidacloprid   3.3336531  -4.05979665  10.727102808 0.8976481
## Nitenpyram-Imidacloprid     4.6034943  -0.27773162   9.484720319 0.0825706
## Nithiazine-Imidacloprid   -20.7471550 -25.51948303 -15.974826981 0.0000000
## Thiacloprid-Imidacloprid   -1.4787159  -3.77669129   0.819259503 0.5440493
## Thiamethoxam-Imidacloprid   1.2556683  -0.67188326   3.183219776 0.5266734
## Nitenpyram-Imidaclothiz     1.2698413  -7.51035405  10.050036590 0.9999561
## Nithiazine-Imidaclothiz   -24.0808081 -32.80093294 -15.360683218 0.0000000
## Thiacloprid-Imidaclothiz   -4.8123690 -12.46391482   2.839176871 0.5758935
## Thiamethoxam-Imidaclothiz  -2.0779848  -9.62655539   5.470585757 0.9950400
## Nithiazine-Nitenpyram     -25.3506494 -32.07402988 -18.627268825 0.0000000
## Thiacloprid-Nitenpyram     -6.0822102 -11.34618420  -0.818236282 0.0103350
## Thiamethoxam-Nitenpyram    -3.3478261  -8.46096463   1.765312458 0.5194316
## Thiacloprid-Nithiazine     19.2684391  14.10528410  24.431594116 0.0000000
## Thiamethoxam-Nithiazine    22.0028233  16.99353853  27.012107998 0.0000000
## Thiamethoxam-Thiacloprid    2.7343842  -0.02215529   5.490923603 0.0538087
```

```
#8
ggplot(EPAEcotox, aes(x = Chemical.Name, y = Pub..Year))+
  geom_boxplot()+
  labs(x = "", y = "Publication Year")+
  theme(axis.text.x = element_text(angle = 45,  hjust = 1))
```



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: The publication years associated with these 9 different chemicals are relatively different(ANOVA; df=1274, F=33.21, p<0.0001)

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_LTER_Tidy <-NTL_LTER_Lake %>%
  filter(daynum>=183 & daynum<=213) %>%
  select(lakename, year4, daynum, depth, temperature_C)%>%
  na.omit
#12
TempAIC <- lm(data = NTL_LTER_Tidy, temperature_C ~ year4 + daynum +
```

```
              depth)
step(TempAIC)
```

```
## 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_LTER_Tidy)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##   -18.19700      0.01611      0.04024     -1.94133
```

```
TempModel <- lm(data = NTL_LTER_Tidy, temperature_C ~ year4 + daynum + depth)
summary(TempModel)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER_Tidy)
##
## 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: The final linear equation is temperature_C = 0.016year4 + 0.04daynum - 1.94depth -18.2, this model explains 74% of the observed variance.

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

```
#14
Temp_ancova.interaction <- lm(data = NTL_LTER_Tidy, temperature_C ~ lakename * depth)
summary(Temp_ancova.interaction)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename * depth, data = NTL_LTER_Tidy)
##
## 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 ***
## 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                            -2.5543     0.2330 -10.962  < 2e-16 ***
## lakenameCrampton Lake:depth       0.7704     0.2379   3.239  0.00121 **
## lakenameEast Long Lake:depth      0.9181     0.2352   3.903 9.57e-05 ***
## lakenameHummingbird Lake:depth   -0.5692     0.2879  -1.977  0.04809 *
## lakenamePaul Lake:depth           0.3698     0.2341   1.580  0.11417
## lakenamePeter Lake:depth          0.5495     0.2338   2.350  0.01878 *
## lakenameTuesday Lake:depth        0.6462     0.2345   2.755  0.00587 **
## lakenameWard Lake:depth          -0.7207     0.2795  -2.578  0.00995 **
## lakenameWest Long Lake:depth      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
```

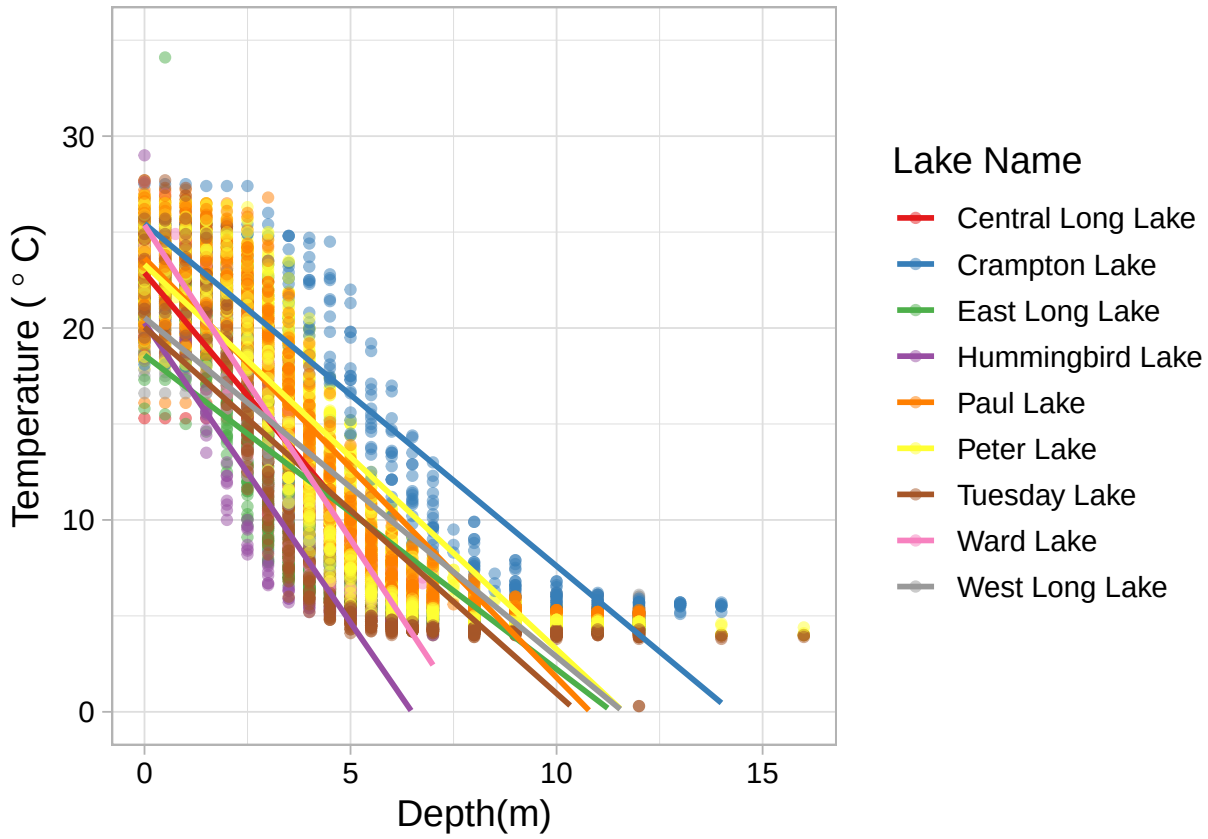
15. Is there an interaction between depth and lakename? How much variance in the temperature observations does this explain?

    ANSWER: Except for Paul lake, there is an interaction between depth and lakename. This explains 78.6% of the variance in the temperature 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
Plot16 <- ggplot(NTL_LTER_Tidy, aes(y = temperature_C, x = depth, color = lakename))+
  geom_point(alpha = 0.5)+
  geom_smooth(method = "lm", se = FALSE)+
  ylim(0,35)+
  labs(x="Depth(m)", y= "Temperature ("~degree~"C)", color = "Lake Name")+
  scale_color_manual(values = c('#e41a1c','#377eb8','#4daf4a','#984ea3','#ff7f00','#ffff33','#a65628','
print(Plot16)
```

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


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