# Assignment 6: Generalized Linear Models

*Siying Chen*

## 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.

```r
#1
getwd()
```

```
## [1] "/Users/Sylvia/Downloads/ENV872/ENV872"
```

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------- tidy
```

```
## 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_
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(RColorBrewer)

Ecotox <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Mortality_raw.csv")
NTL_LTER.ChemPhys <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
```

```
#2
mytheme <- theme_bw(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "bottom",
        panel.grid.major = element_line(size = 0.5, linetype = 'solid'),
        panel.grid.minor = element_line(size = 0.25, linetype = 'dashed'),
        title = element_text(face = "bold"))
theme_set(mytheme)
```

## Neonicotinoids test

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

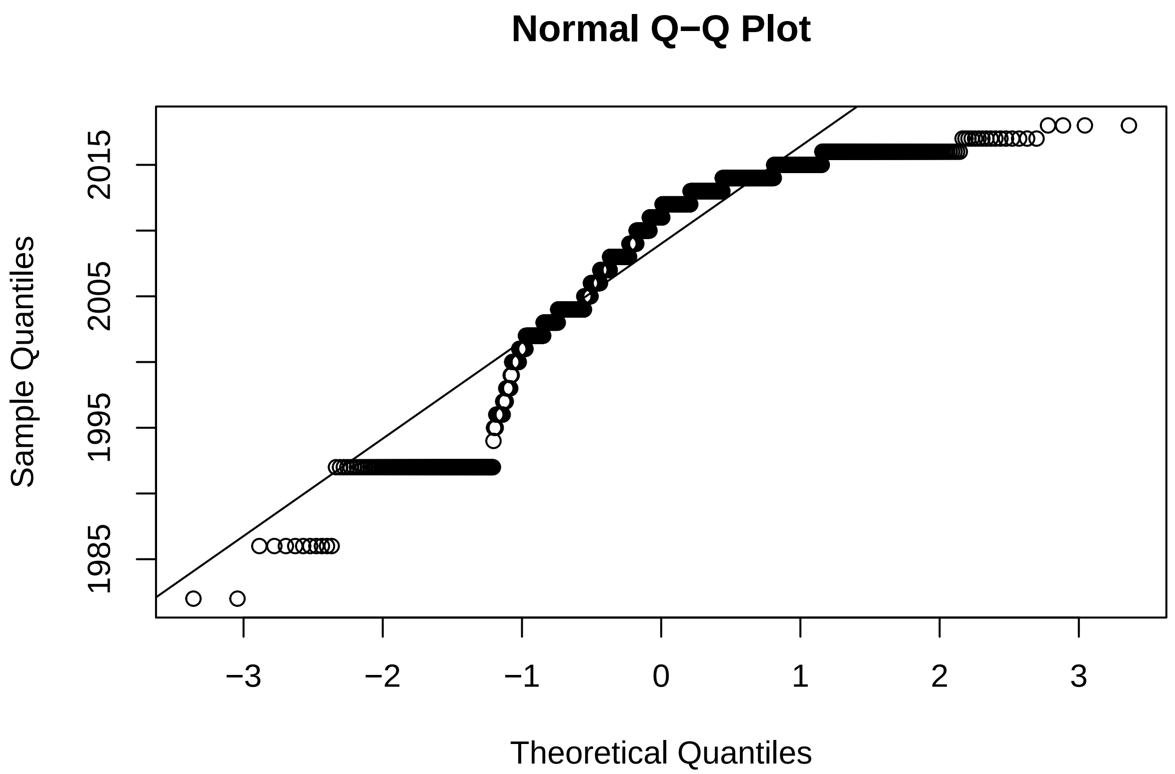
3. Generate a line of code to determine how many different chemicals are listed in the Chemical.Name column.
4. Are the publication years associated with each chemical well-approximated by a normal distribution? Run the appropriate test and also generate a frequency polygon to illustrate the distribution 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(Ecotox$Chemical.Name))
```

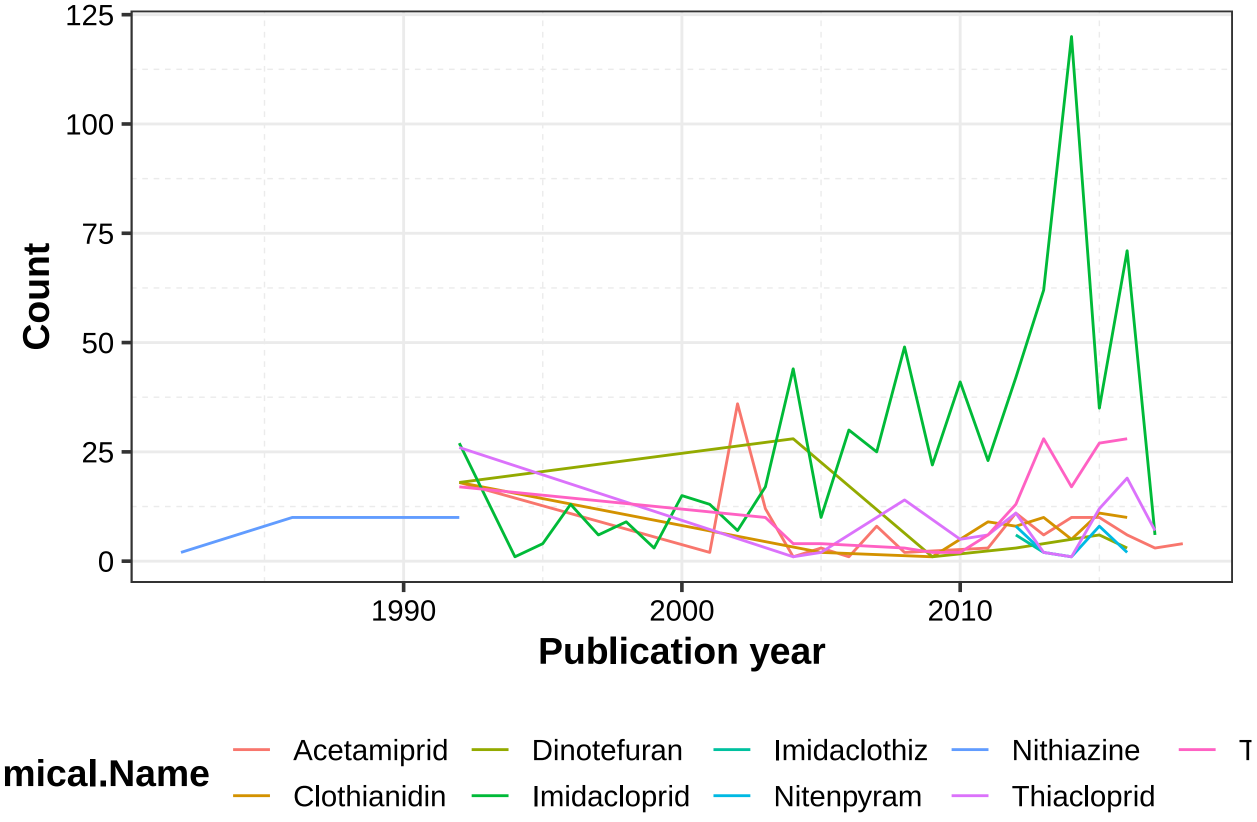
```
## [1] 9
```

```
#4
Ecotox_normality <- Ecotox %>%
  group_by(Chemical.Name) %>%
  summarise(W = shapiro.test(Pub..Year)$statistic,
            p.value = shapiro.test(Pub..Year)$p.value)
# All p < 0.0001
# reject the null hypothesis and conclude that none of the data is normally distributed

qqnorm(Ecotox$Pub..Year); qqline(Ecotox$Pub..Year) # Q-Q plot confirms the previous conclusion
```



```
ggplot(Ecotox, aes(x = Pub..Year, color = Chemical.Name)) +
  geom_freqpoly(stat = "count") +
  labs(x = "Publication year", y = "Count")
```



```
#5
Ecotox_variance <- bartlett.test(Ecotox$Pub..Year ~ Ecotox$Chemical.Name)
# p < 0.0001
# reject the null hypothesis and conclude that not all the variances for different chemical names are t
```

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

   ANSWER: I would choose to run the one-way ANOVA test, because the there are multiple categories in the chemical names, and a one-way ANOVA test is similar to a two-sample t-test but for three or more groups.

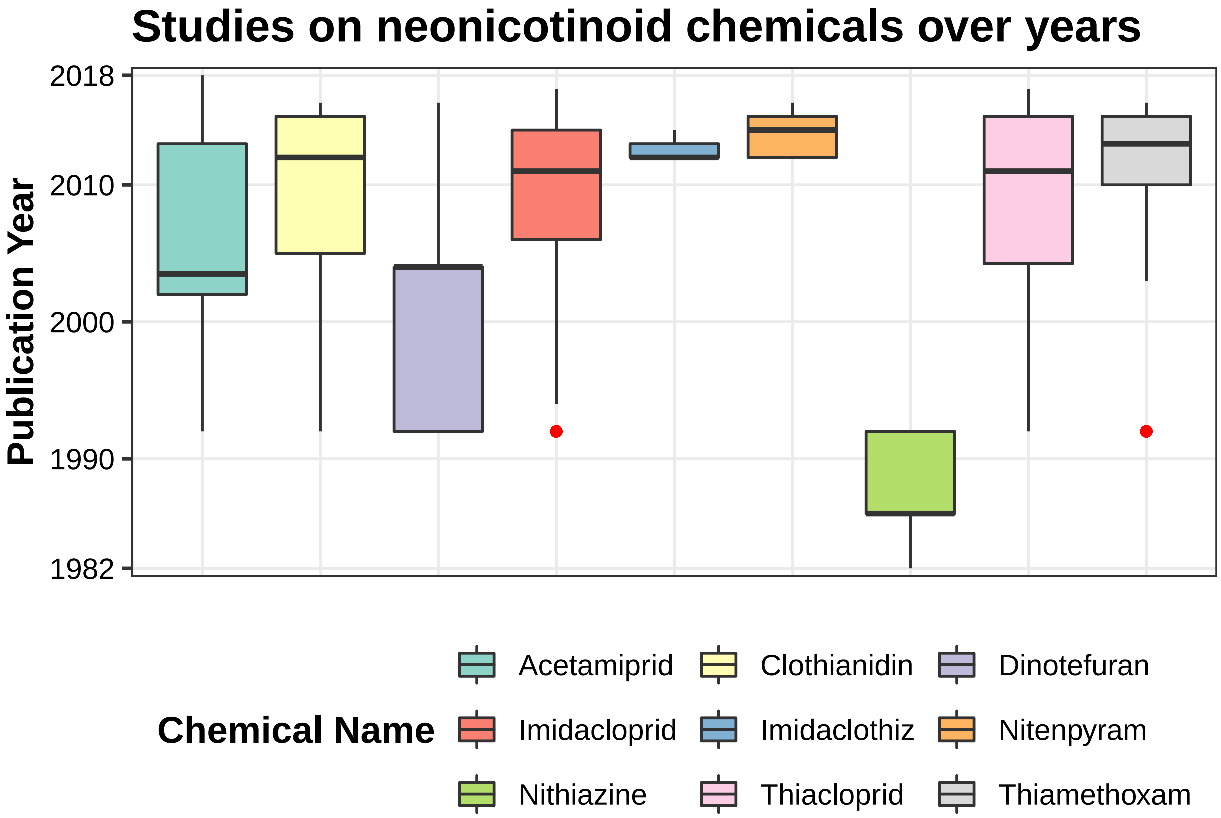
7. Run this test below.

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

```
#7
Ecotox.anova <- lm(Ecotox$Pub..Year ~ Ecotox$Chemical.Name)
summary(Ecotox.anova)
```

```
##
## Call:
## lm(formula = Ecotox$Pub..Year ~ Ecotox$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
## Ecotox$Chemical.NameClothianidin    2.0479     1.0246    1.999  0.04584
## Ecotox$Chemical.NameDinotefuran    -3.4333     1.1057   -3.105  0.00194
## Ecotox$Chemical.NameImidacloprid    3.1181     0.6651    4.689 3.05e-06
## Ecotox$Chemical.NameImidaclothiz    6.4518     2.4412    2.643  0.00832
## Ecotox$Chemical.NameNitenpyram      7.7216     1.6630    4.643 3.78e-06
## Ecotox$Chemical.NameNithiazine    -17.6290     1.6299  -10.816  < 2e-16
## Ecotox$Chemical.NameThiacloprid     1.6394     0.9190    1.784  0.07467
## Ecotox$Chemical.NameThiamethoxam    4.3738     0.8261    5.295 1.40e-07
##
## (Intercept)                      ***
## Ecotox$Chemical.NameClothianidin *
## Ecotox$Chemical.NameDinotefuran  **
## Ecotox$Chemical.NameImidacloprid ***
## Ecotox$Chemical.NameImidaclothiz **
## Ecotox$Chemical.NameNitenpyram   ***
## Ecotox$Chemical.NameNithiazine   ***
## Ecotox$Chemical.NameThiacloprid  .
## Ecotox$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
```

```r
#8
Ecotox.anova.plot <- ggplot(Ecotox, aes(x = Chemical.Name, y = Pub..Year, fill = Chemical.Name)) +
  geom_boxplot(outlier.colour = "red") +
  labs(y = "Publication Year", title = "Studies on neonicotinoid chemicals over years") +
  theme(axis.text.x = element_blank(), axis.title.x = element_blank(), axis.ticks.x = element_blank()) +
  scale_y_discrete(limits = c(1982, 1990, 2000, 2010, 2018)) +
  scale_fill_brewer(palette = "Set3", name = "Chemical Name") +
  guides(fill = guide_legend(nrow = 3,byrow = TRUE))
print(Ecotox.anova.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: The publication years of each neonicotinoid chemicals are not normally distributed, and not all the variances for each neonicotinoid chemicals are the same. Most of the studies on neonicotinoid chemicals are published around 2014. However, studies on Nithiazine are mostly published around 1988, which is also the earlieststudies on neonicotinoid chemicals among these chosen chemicals. (one-way ANOVA; $p < 0.0001$, df = 1274, F = 33.21)

## 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_July <- NTL_LTER.ChemPhys %>%
  filter(daynum >= 182 & daynum <= 212) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()

#12
NTL_July_AIC <- lm(data = NTL_July, temperature_C ~ year4 + daynum + depth)
step(NTL_July_AIC)
```

```
## Start:  AIC=26016.31
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                141118 26016
## - year4   1        80 141198 26020
## - daynum  1      1333 142450 26106
## - depth   1    403925 545042 39151
```

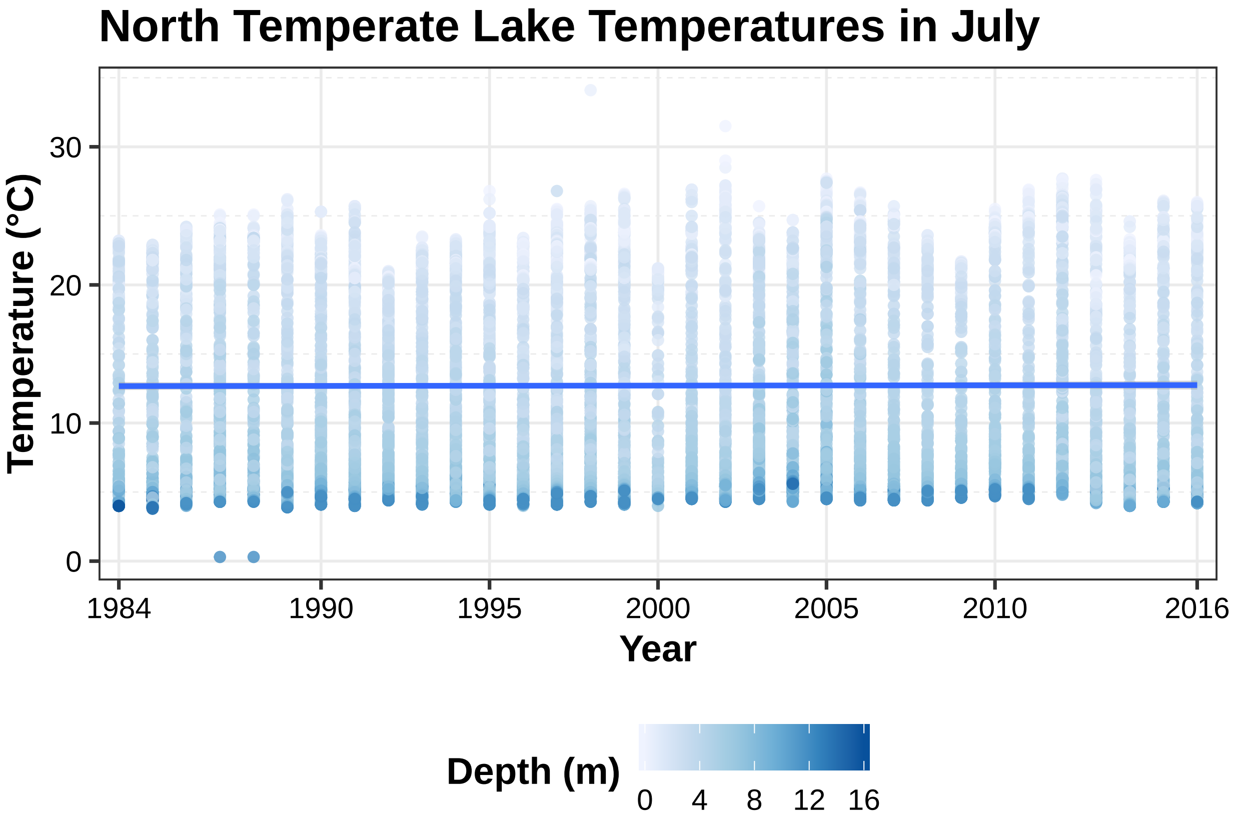
```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_July)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -6.45556      0.01013      0.04134     -1.94726
```

```
# year4 has the lowest AIC, which make it the best candidate
# year4 and daynum daynum have similar AIC, which can mean they are redundant
# also daynum is already included in the initial data filter
# I would choose year and depth as explanatory variables

NTL_July_MR <- lm(data = NTL_July, temperature_C ~ year4 + depth)
summary(NTL_July_MR)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + depth, data = NTL_July)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.541 -3.016  0.098  2.946 13.751
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)  1.268845   8.641177    0.147   0.8833
## year4        0.010346   0.004323    2.393   0.0167 *
## depth       -1.947320   0.011730 -166.013   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.828 on 9719 degrees of freedom
## Multiple R-squared:  0.7393, Adjusted R-squared:  0.7392
## F-statistic: 1.378e+04 on 2 and 9719 DF,  p-value: < 2.2e-16
```

```
NTL_July_MR_plot <- ggplot(NTL_July, aes(x = year4, y = temperature_C, color = depth)) +
  geom_point(alpha = 0.8) +
  labs(x = "Year", y = "Temperature (\u00B0C)", title = "North Temperate Lake Temperatures in July") +
  scale_x_discrete(limits = c(1984, 1990, 1995, 2000, 2005, 2010, 2016)) +
  scale_color_distiller(palette = "Blues", direction = 1, name = "Depth (m)") +
  geom_smooth(method = "lm")
print(NTL_July_MR_plot)
```



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 = 1.27 + 0.01(year) - 1.9(depth) + error. This model explains about 73.92% 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_July, temperature_C ~ lakename * depth)
summary(Temp_ancova.interaction)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename * depth, data = NTL_July)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6455 -2.9133 -0.2879  2.7567 16.3606
```

```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       22.9455     0.5861  39.147  < 2e-16 ***
## lakenameCrampton Lake              2.2173     0.6804   3.259  0.00112 **
## lakenameEast Long Lake            -4.3884     0.6191  -7.089 1.45e-12 ***
## lakenameHummingbird Lake          -2.4126     0.8379  -2.879  0.00399 **
## lakenamePaul Lake                  0.6105     0.5983   1.020  0.30754
## lakenamePeter Lake                 0.2998     0.5970   0.502  0.61552
## lakenameTuesday Lake              -2.8932     0.6060  -4.774 1.83e-06 ***
## lakenameWard Lake                  2.4180     0.8434   2.867  0.00415 **
## lakenameWest Long Lake            -2.4663     0.6168  -3.999 6.42e-05 ***
## depth                             -2.5820     0.2411 -10.711  < 2e-16 ***
## lakenameCrampton Lake:depth        0.8058     0.2465   3.268  0.00109 **
## lakenameEast Long Lake:depth       0.9465     0.2433   3.891  0.00010 ***
## lakenameHummingbird Lake:depth    -0.6026     0.2919  -2.064  0.03903 *
## lakenamePaul Lake:depth            0.4022     0.2421   1.662  0.09664 .
## lakenamePeter Lake:depth           0.5799     0.2418   2.398  0.01649 *
## lakenameTuesday Lake:depth         0.6605     0.2426   2.723  0.00648 **
## lakenameWard Lake:depth           -0.6930     0.2862  -2.421  0.01548 *
## lakenameWest Long Lake:depth       0.8154     0.2431   3.354  0.00080 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.471 on 9704 degrees of freedom
## Multiple R-squared:  0.7861, Adjusted R-squared:  0.7857
## F-statistic:  2097 on 17 and 9704 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: Yes, lake temperature is associated with depth and lake name. This model can explain about 78.57% of the variance.

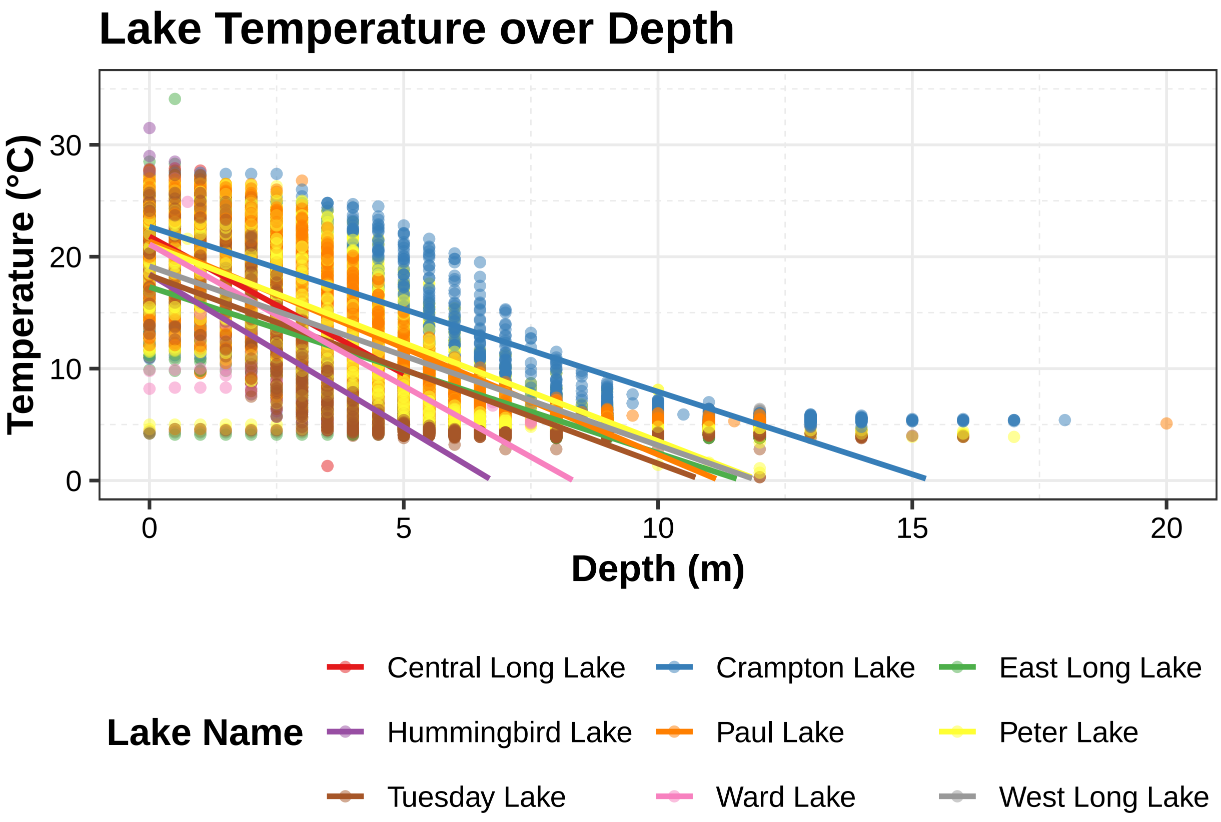
16. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#16
Temp_ancova_plot <- ggplot(NTL_LTER.ChemPhys, 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 (m)", y = "Temperature (\u00B0C)", title = "Lake Temperature over Depth") +
  scale_color_brewer(palette = "Set1", name = "Lake Name") +
  guides(color = guide_legend(nrow = 3,byrow = TRUE))
print(Temp_ancova_plot)
```

```
## Warning: Removed 3858 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 3858 rows containing missing values (geom_point).
```

```
## Warning: Removed 126 rows containing missing values (geom_smooth).
```


Lake Temperature over Depth
Temperature (°C)
30
20
10
0
0
5
10
15
20
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