Assignment 6: Generalized Linear Models

Kat Horan

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 Set up session

# Set working directory:
# setwd("/Users/kathleenhoran/Desktop/Duke/Spring 2019/Env. Data Analytics/Env_Data_Analytics")

# Load packages:
library(ggplot2)
library(viridis)
```

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

```
library(RColorBrewer)
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------- tidyverse 1.2.1 --
## v tibble  2.0.1       v purrr   0.3.0
## v tidyr   0.8.2       v dplyr   0.8.0.1
## v readr   1.3.1       v stringr 1.4.0
## v tibble  2.0.1       v forcats 0.4.0
```

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

```
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(dunn.test)

# Load datasets:
# NTL-LTER for chemistry/physics raw data
NTL.lake.chem.phys.raw <-
  read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

# Ecotox for Neonicotinoids raw data
Ecotox.neo.mort.raw <- read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Mortality_raw.csv")

#2 Set ggplot theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
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 Check different chemicals listed in Chemical.Name column
levels(Ecotox.neo.mort.raw$Chemical.Name)
```

```
## [1] "Acetamiprid"  "Clothianidin" "Dinotefuran"  "Imidacloprid"
## [5] "Imidaclothiz" "Nitenpyram"   "Nithiazine"   "Thiacloprid"
## [9] "Thiamethoxam"
```

```
summary.factor(Ecotox.neo.mort.raw$Chemical.Name)
```

```
##  Acetamiprid Clothianidin  Dinotefuran Imidacloprid Imidaclothiz
##          136           74           59          695            9
##   Nitenpyram   Nithiazine  Thiacloprid Thiamethoxam
##           21           22          106          161
```

```r
#4
# Shapiro test for normality for publication years associated with each of the nine chemicals
shapiro.test(Ecotox.neo.mort.raw$Pub..Year[Ecotox.neo.mort.raw$Chemical.Name == "Acetamiprid"])
```

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

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

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

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

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

```r
shapiro.test(Ecotox.neo.mort.raw$Pub..Year[Ecotox.neo.mort.raw$Chemical.Name == "Imidacloprid"])
```

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

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

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

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

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

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

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

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

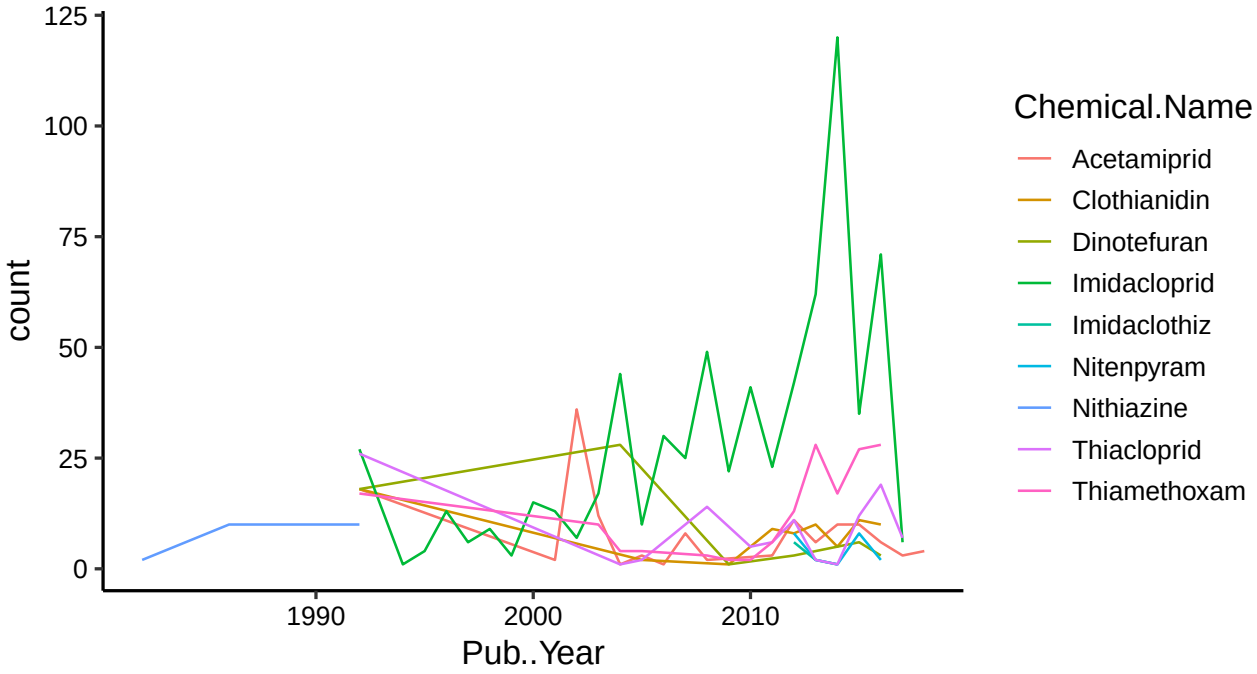
```
##
##  Shapiro-Wilk normality test
##
## data:  Ecotox.neo.mort.raw$Pub..Year[Ecotox.neo.mort.raw$Chemical.Name ==     "Thiacloprid"]
## W = 0.7669, p-value = 1.118e-11
```

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

```
##
##  Shapiro-Wilk normality test
##
## data:  Ecotox.neo.mort.raw$Pub..Year[Ecotox.neo.mort.raw$Chemical.Name ==     "Thiamethoxam"]
## W = 0.7071, p-value < 2.2e-16
```

```r
# The publication years are not well-approximated by a normal distribution.
# Since the p-values generated by the Shapiro-Wilk tests for pub years associated
# with each chemical are all smaller than .05, we reject the null hypothesis of
# normality.

# Frequency polygon
ggplot(Ecotox.neo.mort.raw, aes(x = Pub..Year, color = Chemical.Name)) +
  geom_freqpoly(stat = "count") +
  theme(legend.position = "right")
```




```r
#5 Test for equal variance
bartlett.test(Ecotox.neo.mort.raw$Pub..Year ~ Ecotox.neo.mort.raw$Chemical.Name)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Ecotox.neo.mort.raw$Pub..Year by Ecotox.neo.mort.raw$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, as
# indicated by the low p-value from the Bartlett test. We accept the Bartlett
# test's alternative hypothesis that the variances are not equal
```

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

   ANSWER: I would select the Kruskal Wallace test because it is not normally distributed.

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 Kruskal Wallace test
chem.pub.kw <- kruskal.test(Ecotox.neo.mort.raw$Pub..Year ~ Ecotox.neo.mort.raw$Chemical.Name)
chem.pub.kw
```

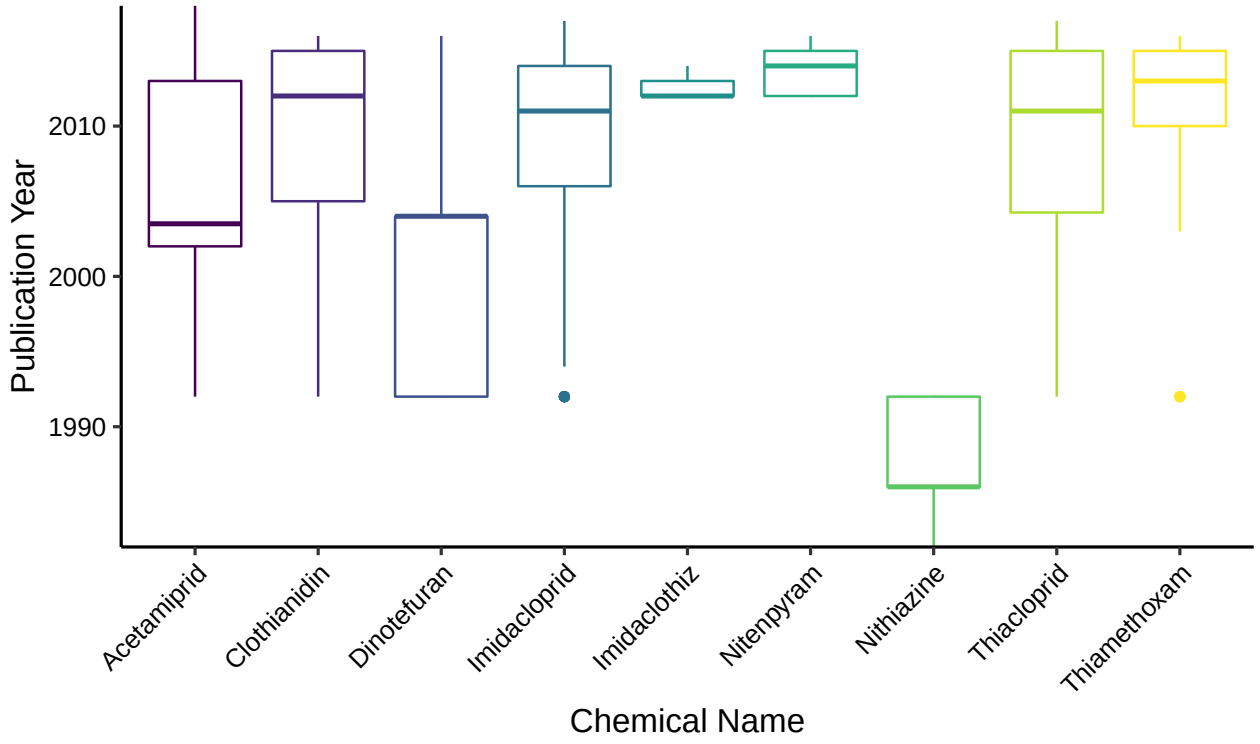
```
##
##  Kruskal-Wallis rank sum test
##
## data:  Ecotox.neo.mort.raw$Pub..Year by Ecotox.neo.mort.raw$Chemical.Name
## Kruskal-Wallis chi-squared = 134.15, df = 8, p-value < 2.2e-16
```

```
dunn.test(Ecotox.neo.mort.raw$Pub..Year, Ecotox.neo.mort.raw$Chemical.Name, kw = T,
          table = F, list = T, method = "holm", altp = T)
```

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 134.1455, df = 8, p-value = 0
##
##
##                            Comparison of x by group
##                                    (Holm)
##
## List of pairwise comparisons: Z statistic (adjusted p-value)
## ----------------------------------------------
## Acetamiprid - Clothianidin  : -3.038807 (0.0404)*
## Acetamiprid - Dinotefuran   :  2.117208 (0.4109)
## Clothianidin - Dinotefuran  :  4.406076 (0.0002)*
## Acetamiprid - Imidacloprid  : -4.020498 (0.0013)*
## Clothianidin - Imidacloprid :  0.506889 (1.0000)
## Dinotefuran - Imidacloprid  : -5.214028 (0.0000)*
## Acetamiprid - Imidaclothiz  : -1.805293 (0.7813)
## Clothianidin - Imidaclothiz : -0.516664 (1.0000)
## Dinotefuran - Imidaclothiz  : -2.658649 (0.1177)
## Imidacloprid - Imidaclothiz : -0.728428 (1.0000)
## Acetamiprid - Nitenpyram    : -4.501863 (0.0002)*
## Clothianidin - Nitenpyram   : -2.493626 (0.1770)
## Dinotefuran - Nitenpyram    : -5.452779 (0.0000)*
## Imidacloprid - Nitenpyram   : -3.063483 (0.0394)*
## Imidaclothiz - Nitenpyram   : -1.089720 (1.0000)
## Acetamiprid - Nithiazine    :  5.642529 (0.0000)*
## Clothianidin - Nithiazine   :  7.147325 (0.0000)*
## Dinotefuran - Nithiazine    :  3.869350 (0.0023)*
## Imidacloprid - Nithiazine   :  7.728634 (0.0000)*
```

```
## Imidaclothiz - Nithiazine   :  4.847313 (0.0000)*
## Nitenpyram - Nithiazine     :  7.709981 (0.0000)*
## Acetamiprid - Thiacloprid   : -3.222561 (0.0241)*
## Clothianidin - Thiacloprid  :  0.141491 (0.8875)
## Dinotefuran - Thiacloprid   : -4.602529 (0.0001)*
## Imidacloprid - Thiacloprid  : -0.388871 (1.0000)
## Imidaclothiz - Thiacloprid  :  0.587068 (1.0000)
## Nitenpyram - Thiacloprid    :  2.670974 (0.1210)
## Nithiazine - Thiacloprid    : -7.316688 (0.0000)*
## Acetamiprid - Thiamethoxam  : -5.889886 (0.0000)*
## Clothianidin - Thiamethoxam : -1.758725 (0.7862)
## Dinotefuran - Thiamethoxam  : -6.676212 (0.0000)*
## Imidacloprid - Thiamethoxam : -3.532703 (0.0082)*
## Imidaclothiz - Thiamethoxam : -0.188627 (1.0000)
## Nitenpyram - Thiamethoxam   :  1.592776 (1.0000)
## Nithiazine - Thiamethoxam   : -8.722412 (0.0000)*
## Thiacloprid - Thiamethoxam  : -2.146115 (0.4142)
##
## alpha = 0.05
## Reject Ho if p <= alpha
```

```r
#8 Boxplot with range of publication years for each chemical
chem.pub.boxplot <-
  ggplot(Ecotox.neo.mort.raw, aes(x = Chemical.Name, y = Pub..Year, color = Chemical.Name)) +
  geom_boxplot() +
  labs(x = "Chemical Name", y = "Publication Year") +
  scale_y_continuous(expand = c(0, 0)) +
  scale_color_viridis(discrete = TRUE) +
  theme(legend.position = "none", axis.text.x = element_text(angle = 45,  hjust = 1))
print(chem.pub.boxplot)
```



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 p-value smaller than 0.05 from the Kruskal-Wallis test indicates that there is a significant difference between groups. (Kruskal-Wallis chi-squared = 134.1455, df = 8, p-value = 0).

   Since it does not indicate which groups, I ran a Dunn test to see which pairs had the significant difference. We see that the following pairs are significantly different:

   Acetamiprid - Clothianidin : -3.038807 (0.0404) *Clothianidin - Dinotefuran : 4.406076 (0.0002)* Acetamiprid - Imidacloprid : -4.020498 (0.0013) *Dinotefuran - Imidacloprid : -5.214028 (0.0000)* Acetamiprid - Nitenpyram : -4.501863 (0.0002) *Dinotefuran - Nitenpyram : -5.452779 (0.0000)* Imidacloprid - Nitenpyram : -3.063483 (0.0394) *Acetamiprid - Nithiazine : 5.642529 (0.0000)* Clothianidin - Nithiazine : 7.147325 (0.0000) *Dinotefuran - Nithiazine : 3.869350 (0.0023)* Imidacloprid - Nithiazine : 7.728634 (0.0000) *Imidaclothiz - Nithiazine : 4.847313 (0.0000)* Nitenpyram - Nithiazine : 7.709981 (0.0000) *Acetamiprid - Thiacloprid : -3.222561 (0.0241)* Dinotefuran - Thiacloprid : -4.602529 (0.0001) *Nithiazine - Thiacloprid : -7.316688 (0.0000)* Acetamiprid - Thiamethoxam : -5.889886 (0.0000) *Dinotefuran - Thiamethoxam : -6.676212 (0.0000)* Imidacloprid - Thiamethoxam : -3.532703 (0.0082) *Nithiazine - Thiamethoxam : -8.722412 (0.0000)*

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 Wrangling data set
NTL.lake.chem.phys.processed <-
  NTL.lake.chem.phys.raw %>%
  filter(daynum >= 182 & daynum <= 212) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  filter(!is.na(temperature_C))

#12
# AIC tests with all explanatory variables
NTL.lake.chem.phys.AIC <-
  lm(data = NTL.lake.chem.phys.processed, temperature_C ~ year4 + daynum + depth)
step(NTL.lake.chem.phys.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.lake.chem.phys.processed)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -6.45556      0.01013      0.04134     -1.94726
```

```
# Final multiple regression
NTL.lake.chem.phys.model <-
  lm(data = NTL.lake.chem.phys.processed, temperature_C ~ year4 + daynum + depth)
summary(NTL.lake.chem.phys.model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.lake.chem.phys.processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6517 -2.9937  0.0855  2.9692 13.6171
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -6.455560   8.638808   -0.747   0.4549
## year4        0.010131   0.004303    2.354   0.0186 *
## daynum       0.041336   0.004315    9.580   <2e-16 ***
## depth       -1.947264   0.011676 -166.782   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.811 on 9718 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7417
## F-statistic:  9303 on 3 and 9718 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 also the original linear model. The AIC test found that all three explanatory variables were significant and did not find that it needed to be reduced. 74.17% of the variance is explained in this mode.

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

```
#14 Interaction effects ANCOVA
NTL.lake.chem.phys.ancova.int <-
  lm(data = NTL.lake.chem.phys.processed, temperature_C ~ lakename * depth)
summary(NTL.lake.chem.phys.ancova.int)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename * depth, data = NTL.lake.chem.phys.processed)
##
## 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: The low p-value tells us that there is a significant interaction between depth and lakename, and the we can see the specific significant interactions in the summary. The only interaction that is marginally significant is the interaction of depth and Paul Lake at .0966. 78.57% of the variance in temperature observations can be explained by the interaction between depth and lakename.

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 Plot with temperature by depth for each lake
tempbydepth.plot <- ggplot(NTL.lake.chem.phys.processed,
  aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha = .5) +
  geom_smooth(method = "lm", se = FALSE) +
  xlim(0, 10) +
  ylim(0, 35) +
  scale_color_brewer(palette = "Paired") +
  labs(x = "Depth", y = "Temperature (Celsius)",
              fill = "lakename", color = "Lake Name") +
  theme(legend.position = "right")
print(tempbydepth.plot)
```

```
## Warning: Removed 683 rows containing non-finite values (stat_smooth).
```

```
## Warning: Removed 683 rows containing missing values (geom_point).
```

```
## Warning: Removed 21 rows containing missing values (geom_smooth).
```