Assignment 6: Generalized Linear Models

Sebastian Bognar

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
setwd("/Users/Seabass/Documents/Duke/spring_2019/env_872L/lesson_2/ENV_872L")
getwd()
```

```
## [1] "/Users/Seabass/Documents/Duke/spring_2019/env_872L/lesson_2/ENV_872L"
```

```
library(viridis)
```

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

```
library(RColorBrewer)
library(colormap)
library(ggplot2)
library(dunn.test)
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------- tidyverse 1.2.
```

```
## v tibble  2.0.1     v purrr   0.2.5
## v tidyr   0.8.2     v dplyr   0.7.8
## v readr   1.3.1     v stringr 1.3.1
## v tibble  2.0.1     v forcats 0.3.0
```

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

```
EPA_ECOTOX<-read.csv("./Data/Raw/ECOTOX_Neonicotinoids_Mortality_raw.csv", header = TRUE) # import ecot

NTL_LTER_Chem.physics <- read.csv("./Data/Processed/NTL-LTER_Lake_ChemistryPhysics_PeterPaul_Processed.

# change the format of date from factor to "DAte"

NTL_LTER_Chem.physics$sampledate<- as.Date(as.character(NTL_LTER_Chem.physics$sampledate), "%m/%d/%y")



#2

theme_A6 <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right",
        axis.ticks = element_line(colour = "black"),
        panel.border = element_rect(fill= NA,color="black", size=0.5,
                                    linetype="solid"),
        panel.grid.major.y =element_line(color = "grey"),
        panel.grid.minor.y = element_blank(),
        panel.grid.major.x  = element_blank(),
        panel.grid.minor.x  = element_blank())
```

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.

   answer: The publication years that are associated with each chemical are not approximated by a normal distribution, which is shown by the shapiro.test and qq plot (shapiro test; $p < 0.001$).

5. Is there equal variance among the publication years for each chemical? Hint: var.test is not the correct function.

   answer: There is not equal variance among the publication years for each chemical (Barlett Test; df =8, p <0.0001).

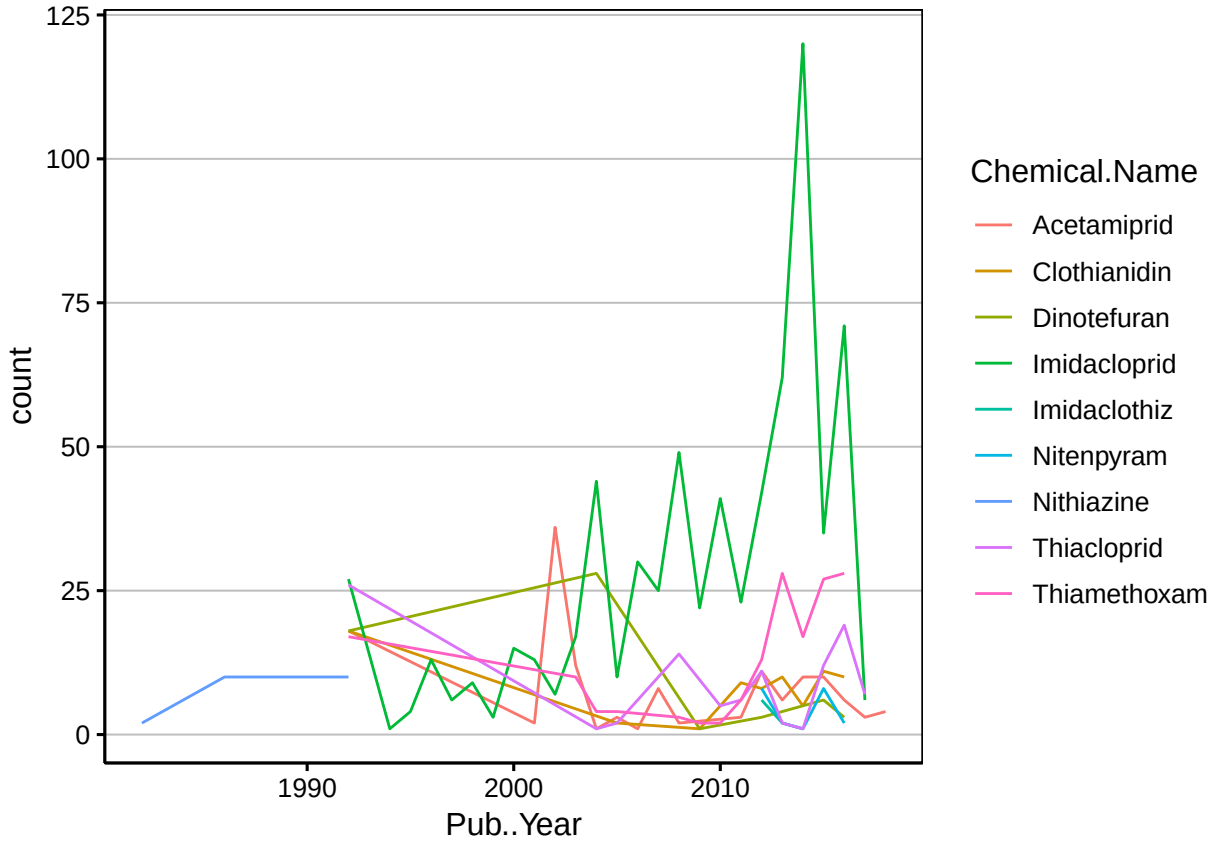
```
#3

summary(EPA_ECOTOX$Chemical.Name)
```

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

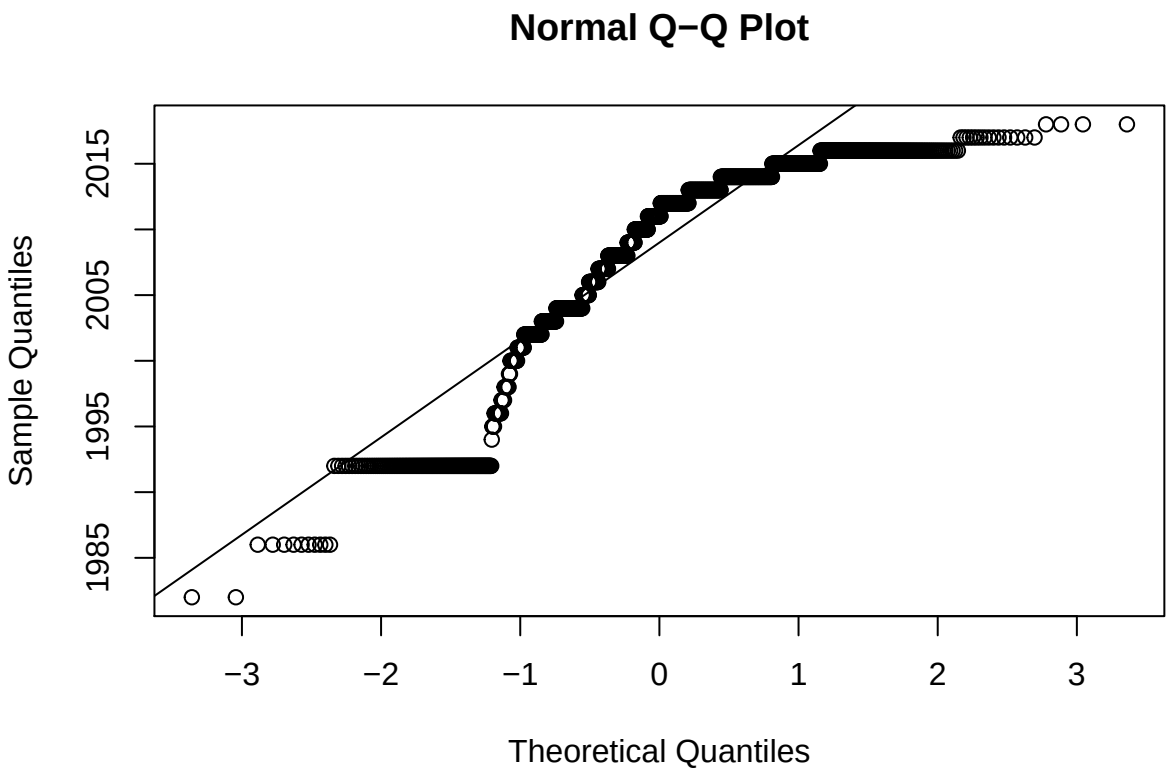
```r
#4

# shapiro test

shapiro.test(EPA_ECOTOX$Pub..Year)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  EPA_ECOTOX$Pub..Year
## W = 0.85472, p-value < 2.2e-16
```

```r
# ggplot of the distribution

PUB_YEAR_PLOT <- ggplot(EPA_ECOTOX, aes( x= Pub..Year, col = Chemical.Name))+
  geom_freqpoly(stat = "count")+
  theme_A6

print(PUB_YEAR_PLOT)
```




```r
# qqplot and qqline

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




```
#5 bartlett test for equal variances

bartlett.test(EPA_ECOTOX$Pub..Year~EPA_ECOTOX$Chemical.Name)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  EPA_ECOTOX$Pub..Year by EPA_ECOTOX$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: The test that should be run is the Kruskal-Wallis Test due to the fact that the data is not normally distributed and you want to determine if studies on various neonicotinoid chemicals were conducted in different years.

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
KT_PUB <- kruskal.test(EPA_ECOTOX$Pub..Year~EPA_ECOTOX$Chemical.Name)
KT_PUB
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  EPA_ECOTOX$Pub..Year by EPA_ECOTOX$Chemical.Name
```

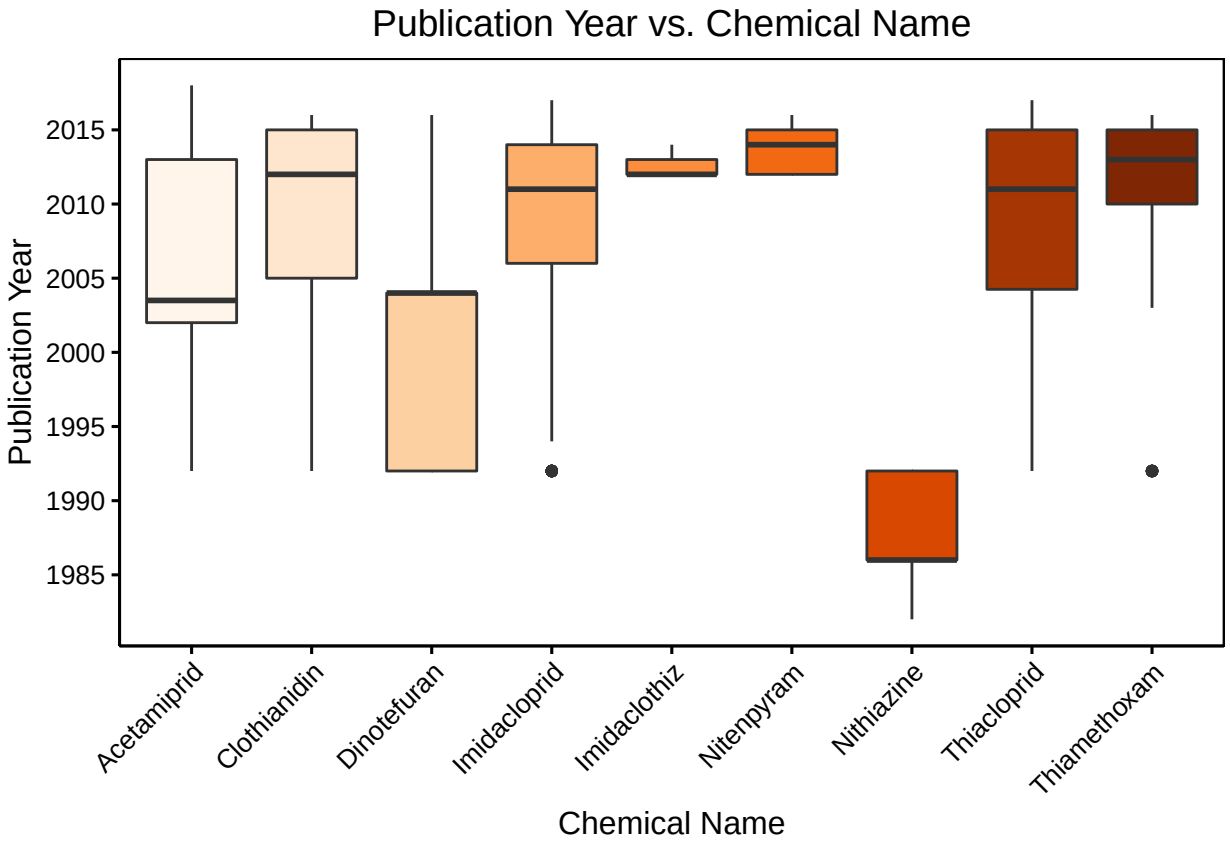
```
## Kruskal-Wallis chi-squared = 134.15, df = 8, p-value < 2.2e-16
```

```
# dunn test


dunn.test(EPA_ECOTOX$Pub..Year, EPA_ECOTOX$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
```

```
#8

PUB_YEAR_PLOT_pretty <- ggplot(EPA_ECOTOX, aes(  x =Chemical.Name, y= Pub..Year, fill =Chemical.Name))+
  geom_boxplot()+
  theme_A6+
  ylab("Publication Year")+
  xlab("Chemical Name")+
  theme(axis.text.x = element_text(angle = 45,  hjust = 1))+
  scale_fill_manual(values = c('#fff5eb','#fee6ce','#fdd0a2','#fdae6b','#fd8d3c','#f16913','#d94801','#a
  theme(legend.position="none")+
  scale_y_continuous( breaks=seq(1980,2018,5))+
  ggtitle("Publication Year vs. Chemical Name")+
  theme(plot.title = element_text(hjust = 0.5))

print(PUB_YEAR_PLOT_pretty)
```



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: There was a significant difference between the publication years and the various neonicotinoid chemical studies (Kruskal Wallis Test; chisq = 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_specific <- NTL_LTER_Chem.physics %>%
  filter( daynum > 181 & daynum < 213) %>%
  select(lakename, year4, daynum, depth, temperature_C )%>%
  filter(!is.na(lakename) & !is.na(year4) & !is.na(daynum) & !is.na(depth) & !is.na(temperature_C))

#12

# run the AIC
Temp_AIC <- lm(data = NTL_specific, temperature_C~ year4 + daynum + depth)

step(Temp_AIC)
```

```
## Start:  AIC=12969.4
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                 57559 12969
## - year4   1        46  57604 12972
## - daynum  1       882  58441 13052
## - depth   1    251661 309220 22273
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_specific)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##   33.635999    -0.009502     0.044054    -2.069056
```

```
# best model

Temp_model <- lm(data = NTL_specific, temperature_C~ daynum + depth)
summary(Temp_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ daynum + depth, data = NTL_specific)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9958 -2.8017 -0.1472  2.4700 14.0667
```

```
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 14.640903   0.948310   15.439   <2e-16 ***
## daynum       0.044042   0.004787    9.201   <2e-16 ***
## depth       -2.069980   0.013302 -155.616   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.227 on 5532 degrees of freedom
## Multiple R-squared:  0.8146, Adjusted R-squared:  0.8146
## F-statistic: 1.216e+04 on 2 and 5532 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:

Equation: Temperature (Celsius) = 14.64 + 0.044*(daynum)* - *2.07*(depth)

The model explains 81.46 % of the observed variance of water temperature (Multiple Linear Regression; df = 5532, $p < 0.001$).

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

```
#14

ANCOVA_temp <- lm(data = NTL_specific, temperature_C ~ depth*lakename)
summary(ANCOVA_temp)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth * lakename, data = NTL_specific)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6351 -2.6371 -0.2865  2.4587 13.1875
##
## Coefficients:
##                            Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                23.55608    0.11172  210.845  < 2e-16 ***
## depth                      -2.17974    0.02063 -105.677  < 2e-16 ***
## lakenamePeter Lake         -0.31070    0.15377   -2.021   0.0434 *
## depth:lakenamePeter Lake    0.17768    0.02705    6.568 5.58e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.229 on 5531 degrees of freedom
## Multiple R-squared:  0.8145, Adjusted R-squared:  0.8144
## F-statistic:  8093 on 3 and 5531 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, there is an interaction between depth and lakename. The addition of the interaction between lake name and depth explains 81.44% of the variance in water temperature (ANCOVA; df = 5531, $p < 0.001$).

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_DEPTH_PLOT <- ggplot(NTL_specific, aes( x= depth, y =temperature_C, col = lakename))+
  geom_point(alpha=0.5)+
  theme_A6+
  ylab("Temperature (\u00B0C)")+
  xlab("Depth (m)")+
  ggtitle("Temperature vs. Depth")+
  theme(plot.title = element_text(hjust = 0.5))+
  ylim(0,35)+
  labs(color='Lake Name')+
  geom_smooth(method = "lm", se = FALSE,aes(col = lakename))+
  scale_color_manual(values = c('#bdbdbd','#636363'))




            print(TEMP_DEPTH_PLOT)
```

```
## Warning: Removed 30 rows containing missing values (geom_smooth).
```