# Assignment 6: Generalized Linear Models

*Xin Wang*

## 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("Y:/19spring/872/Environmental_Data_Analytics/Assignments")
ecotox <- read.csv("../Data/Raw/ECOTOX_Neonicotinoids_Mortality_raw.csv")
ntl <- read.csv("../Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.5.2
```

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

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     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
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
## Warning: package 'tibble' was built under R version 3.5.2
```

```
## Warning: package 'tidyr' was built under R version 3.5.2
```

```
## Warning: package 'readr' was built under R version 3.5.2
```

```
## Warning: package 'purrr' was built under R version 3.5.2
```

```
## Warning: package 'dplyr' was built under R version 3.5.2
```

```
## Warning: package 'stringr' was built under R version 3.5.2
```

```
## Warning: package 'forcats' was built under R version 3.5.2
```

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

```
#2
wang <- theme_classic() +
  theme(plot.title=element_text(size = 20,hjust = 0.5),
        panel.background=element_rect(fill="white",color="grey30"),
        axis.title = element_text(size = 15),
        legend.title = element_text(size = 15), legend.text = element_text(size = 12),
        legend.margin=margin(6,6,6,6))
```

## 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.

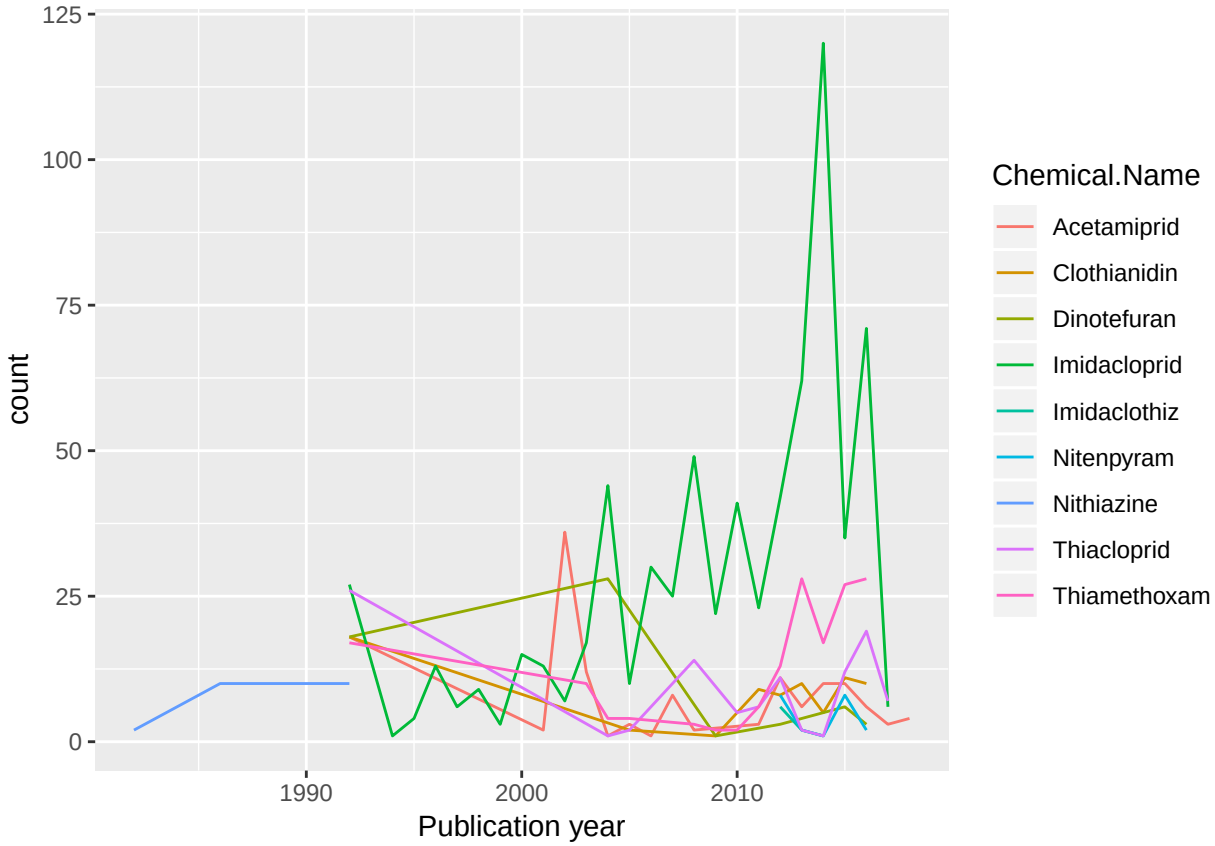
```
#3
nlevels(ecotox$Chemical.Name)
```

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

```
#4
normtest <- ecotox %>%
  group_by(Chemical.Name) %>%
  summarise(W = shapiro.test(Pub..Year)$statistic, p = shapiro.test(Pub..Year)$p.value)
```

```
## Warning: package 'bindrcpp' was built under R version 3.5.2
```

```
ggplot(ecotox,aes(x=Pub..Year,color=Chemical.Name)) +
  geom_freqpoly(stat = "count") + xlab("Publication year")
```




```
#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
```

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

   ANSWER: Variance among the publication years for each chemical is not equal (Bartlett test; df=8, p<2.2e-16). Therefore, the non-parametric method, Kruskal-Wallis test should be adopted.

7. Run this test below.

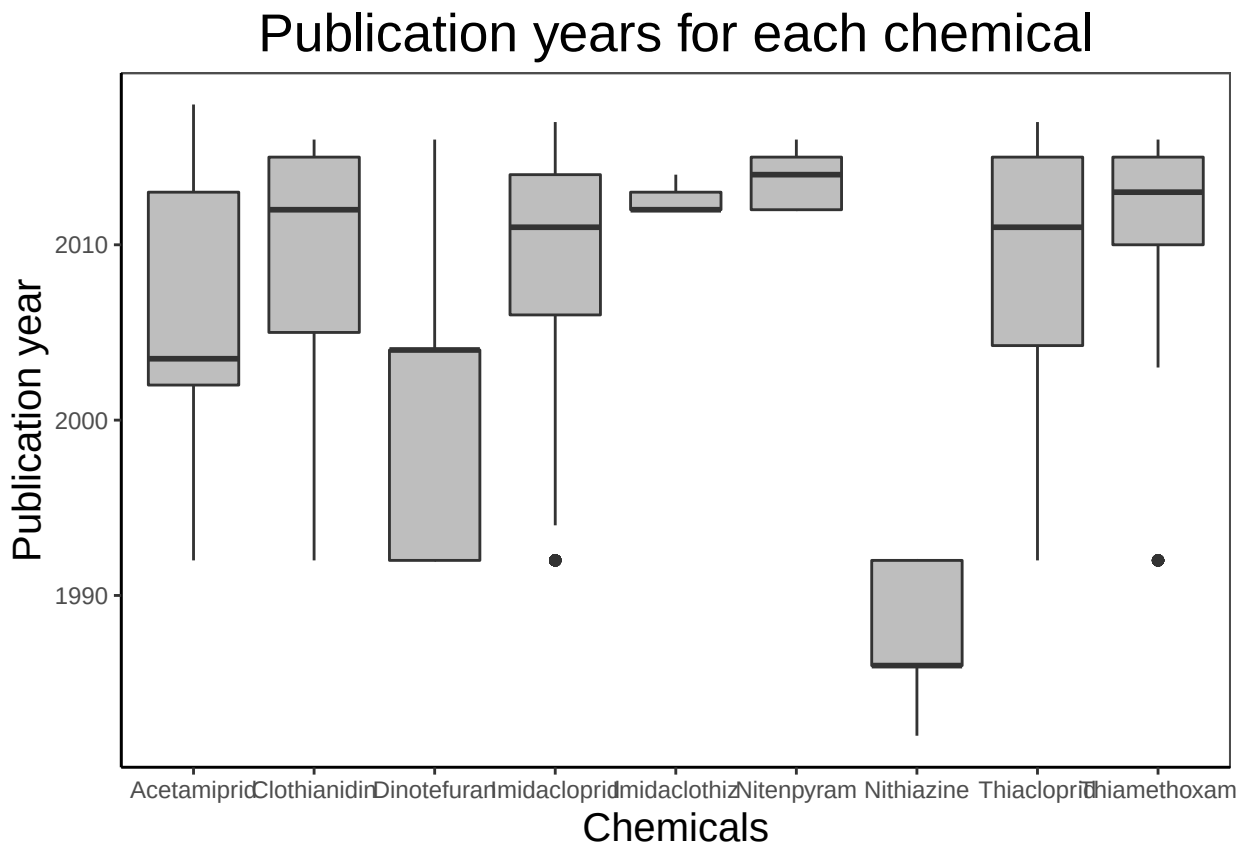
8. Generate a boxplot representing the range of publication years for each chemical. Adjust your graph to make it pretty.

```
#7
kruskal.test(ecotox$Pub..Year ~ ecotox$Chemical.Name)
```

```
##
##  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
```

```
#8
ggplot(ecotox,aes(x=Chemical.Name,y=Pub..Year)) +
```

```
  geom_boxplot(fill="gray") + wang +
  labs(title = "Publication years for each chemical", x="Chemicals", y="Publication year")
```




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 different years (Kruskal-Wallis test; df=8, p < 2.2e-16).

## 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$sampledate <- as.Date(ntl$sampledate,"%m/%d/%y")
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.5.2
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
ntl <- mutate(ntl,month=month(sampledate))
ntl.skinny <- ntl %>%
  filter(month==7) %>%
  select(lakename,year4,daynum,depth,temperature_C) %>%
  na.omit()
#12
TAIC <- lm(temperature_C ~ year4+daynum+depth, data = ntl.skinny)
step(TAIC)
```

```
## Start:  AIC=26065.53
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                141687 26066
## - year4   1       101 141788 26070
## - daynum  1      1237 142924 26148
## - depth   1    404475 546161 39189
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = ntl.skinny)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
#year4, daynum and depth are all best suited explanatory variables
summary(TAIC)
```

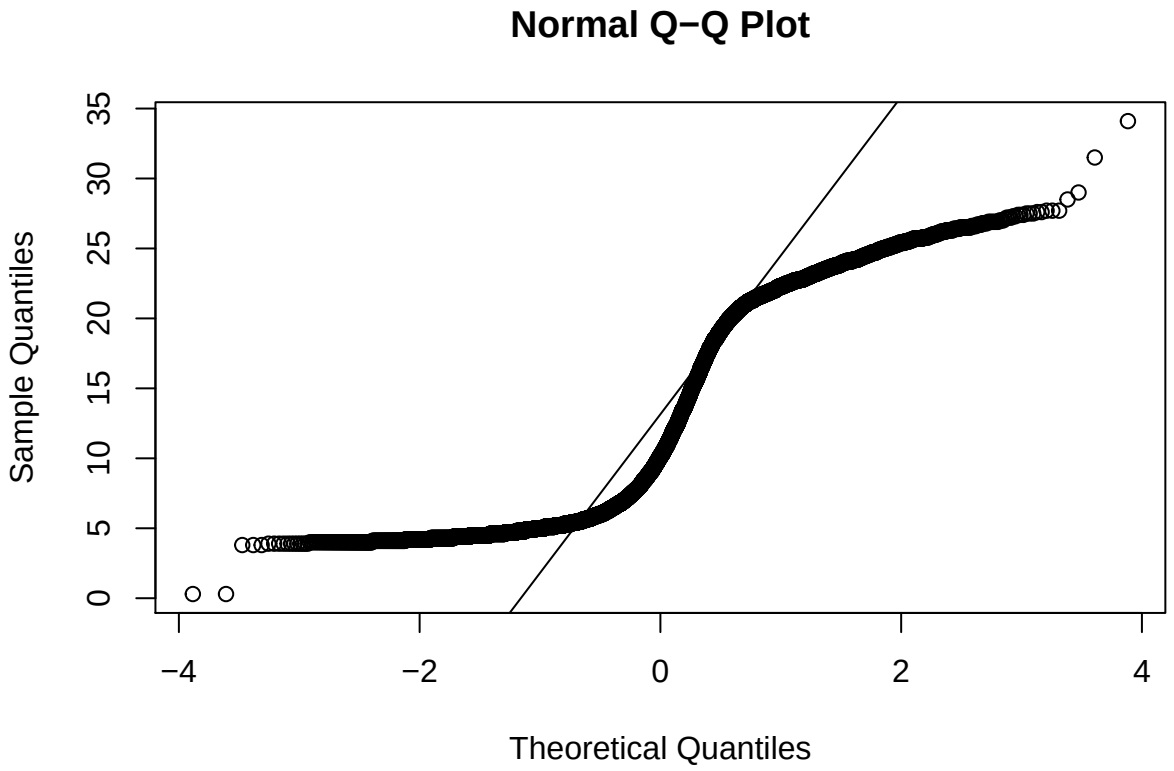
```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = ntl.skinny)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6536 -3.0000  0.0902  2.9658 13.6123
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -8.575564   8.630715   -0.994  0.32044
## year4        0.011345   0.004299    2.639  0.00833 **
## daynum       0.039780   0.004317    9.215  < 2e-16 ***
## depth       -1.946437   0.011683 -166.611  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.817 on 9724 degrees of freedom
## Multiple R-squared:  0.7412, Adjusted R-squared:  0.7411
## F-statistic:  9283 on 3 and 9724 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?

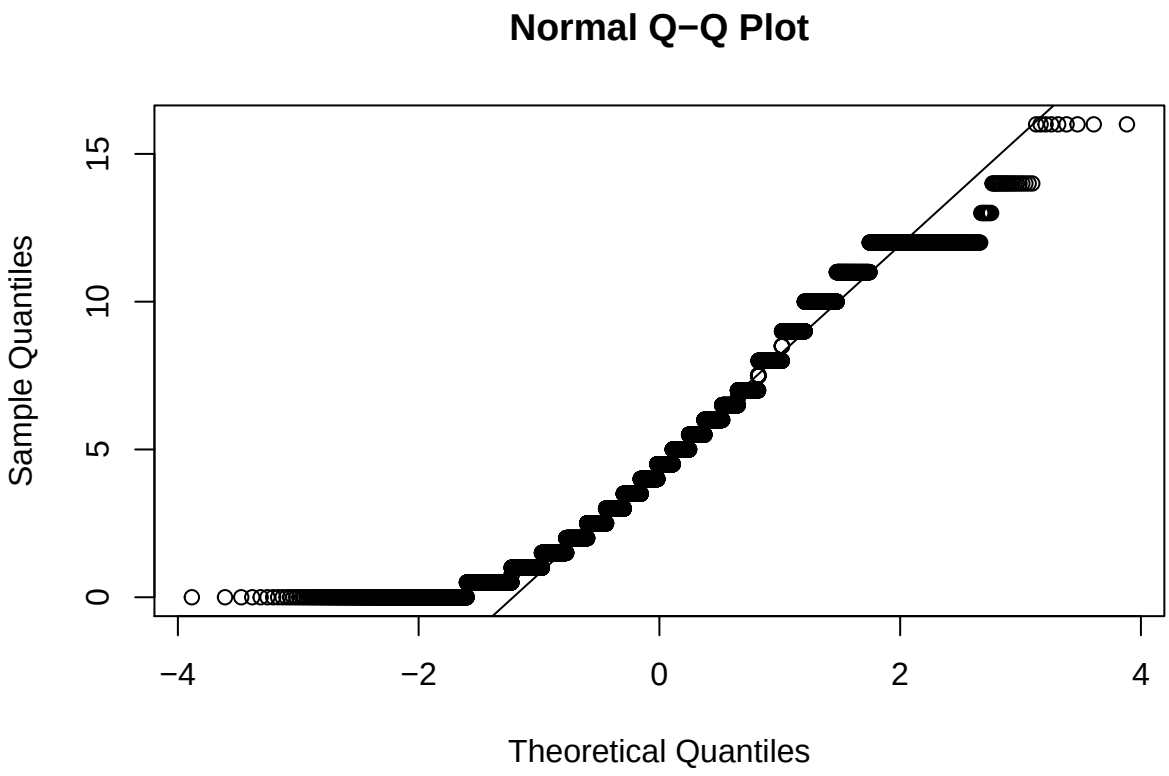
    ANSWER: temperature = -8.58 + 0.01 year4 + 0.04 daynum - 1.95 depth; It explains 74% of the observed variance.

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

```
#14
#Shapiro.test doesn't work for samples larger than 5000.
qqnorm(ntl.skinny$temperature_C)
qqline(ntl.skinny$temperature_C)
```




```
qqnorm(ntl.skinny$depth)
qqline(ntl.skinny$depth)
```




```
TANCOVA <- lm(temperature_C ~ lakename + depth, data = ntl.skinny)
summary(TANCOVA)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename + depth, data = ntl.skinny)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1062 -3.0182 -0.2145  2.8397 15.1605
##
## Coefficients:
##                            Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                21.67335    0.31408   69.006  < 2e-16 ***
## lakenameCrampton Lake       4.53288    0.37298   12.153  < 2e-16 ***
## lakenameEast Long Lake     -1.44524    0.33500   -4.314 1.62e-05 ***
## lakenameHummingbird Lake -4.87775    0.45450  -10.732  < 2e-16 ***
## lakenamePaul Lake           0.93875    0.32184    2.917  0.00354 **
## lakenamePeter Lake          1.40045    0.32179    4.352 1.36e-05 ***
## lakenameTuesday Lake       -1.39244    0.32746   -4.252 2.14e-05 ***
## lakenameWard Lake          -0.67149    0.45458   -1.477  0.13967
## lakenameWest Long Lake     -0.17061    0.33389   -0.511  0.60938
## depth                      -1.96509    0.01096 -179.268  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.544 on 9718 degrees of freedom
## Multiple R-squared:  0.777,  Adjusted R-squared:  0.7768
## F-statistic:  3762 on 9 and 9718 DF,  p-value: < 2.2e-16
```

```
TANCOVA.interaction <- lm(temperature_C ~ lakename * depth, data = ntl.skinny)
summary(TANCOVA.interaction)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename * depth, data = ntl.skinny)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6470 -2.9129 -0.2949  2.7469 16.3606
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      22.8748     0.5660  40.412  < 2e-16 ***
## lakenameCrampton Lake             2.2881     0.6634   3.449 0.000565 ***
## lakenameEast Long Lake           -4.3176     0.6002  -7.194 6.76e-13 ***
## lakenameHummingbird Lake         -2.3418     0.8246  -2.840 0.004523 **
## lakenamePaul Lake                 0.7115     0.5786   1.230 0.218863
## lakenamePeter Lake                0.3884     0.5774   0.673 0.501146
## lakenameTuesday Lake             -2.8656     0.5864  -4.887 1.04e-06 ***
## lakenameWard Lake                 2.4887     0.8302   2.998 0.002728 **
## lakenameWest Long Lake           -2.3819     0.5983  -3.981 6.91e-05 ***
## depth                            -2.5543     0.2331 -10.956  < 2e-16 ***
## lakenameCrampton Lake:depth       0.7781     0.2388   3.258 0.001125 **
## lakenameEast Long Lake:depth      0.9189     0.2354   3.903 9.56e-05 ***
## lakenameHummingbird Lake:depth   -0.6303     0.2856  -2.207 0.027334 *
## lakenamePaul Lake:depth           0.3716     0.2342   1.587 0.112592
## lakenamePeter Lake:depth          0.5511     0.2339   2.356 0.018500 *
## lakenameTuesday Lake:depth        0.6472     0.2347   2.758 0.005826 **
## lakenameWard Lake:depth          -0.7207     0.2797  -2.577 0.009991 **
## lakenameWest Long Lake:depth      0.7892     0.2353   3.354 0.000800 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.476 on 9710 degrees of freedom
## Multiple R-squared:  0.7857, Adjusted R-squared:  0.7853
## F-statistic:  2094 on 17 and 9710 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. It explains about 79% of the variance in 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
ggplot(ntl.skinny,aes(x=depth,y=temperature_C,color=lakename)) + wang +
  geom_point(size=0.8,alpha=0.5) +
  geom_smooth(method = "lm", se=F) +
  scale_color_brewer("Lake name",palette = "Set1") +
```

```
  ylim(c(0,35)) + labs(x="Depth",y="Temperature (°C)")
```

```
## Warning: Removed 73 rows containing missing values (geom_smooth).
```