Assignment 7: GLMs week 2 (Linear Regression and beyond)

Xiangtian Wang

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on generalized linear models.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Salk_A06_GLMs_Week1.Rmd") prior to submission.

The completed exercise is due on Tuesday, February 25 at 1:00 pm.

Set up your session

1. Set up your session. Check your working directory, load the tidyverse, nlme, and piecewiseSEM packages, import the *raw* NTL-LTER raw data file for chemistry/physics, and import the processed litter dataset. You will not work with dates, so no need to format your date columns this time.

2. Build a ggplot theme and set it as your default theme.

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

```
## [1] "C:/Timwork/ENV872/Environmental_Data_Analytics_2020/Assignments"
```

```r
library(tidyverse)
library(nlme)
library(piecewiseSEM)
NTL.LTER.CHEM.PHY <- read.csv("../Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
Litter_mass_trap <- read.csv("../Data/Processed/NEON_NIWO_Litter_mass_trap_Processed.csv")
#2
# Set theme
mytheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

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?

3. 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)

4. 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.

```
#3
my.NTL.LTER.data <- NTL.LTER.CHEM.PHY %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  filter( daynum > 152 & daynum < 183) %>%
  na.exclude()
#4
Temp.lm <- lm(temperature_C ~ year4 + daynum + depth,my.NTL.LTER.data)
step(Temp.lm)
```

```
## Start:  AIC=23998.55
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                119008 23999
## - year4   1        45 119053 24000
## - daynum  1      3945 122952 24306
## - depth   1    280487 399494 35482

##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = my.NTL.LTER.data)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##   21.387058    -0.007535     0.073898    -1.629962
```

```
Temp.lm1 <- lm(temperature_C ~ daynum + depth,my.NTL.LTER.data)
AIC(Temp.lm, Temp.lm1)
```

```
##          df      AIC
## Temp.lm   5 50914.98
## Temp.lm1  4 50916.56
```

```
summary(Temp.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = my.NTL.LTER.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6279 -2.8459 -0.1952  2.7873 15.9645
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 21.387058   8.001273    2.673  0.00753 **
## year4       -0.007535   0.003984   -1.891  0.05861 .
## daynum       0.073898   0.004169   17.726  < 2e-16 ***
## depth       -1.629962   0.010904 -149.476  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.543 on 9480 degrees of freedom
## Multiple R-squared:  0.7054, Adjusted R-squared:  0.7053
## F-statistic:  7568 on 3 and 9480 DF,  p-value: < 2.2e-16
```

```
# Temp.lm is the best.
```

5. What is the final set of explanatory variables that predict temperature from your multiple regression? How much of the observed variance does this model explain?

   Answer: Year(year4), day(daynum), and depth. Adjusted R-square is 0.7053, means the model explains 70% observed variance.

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

```
#6
Temp.lm2 <- lm(temperature_C ~ lakename * depth,my.NTL.LTER.data)
summary.aov(Temp.lm2)
```

```
##                   Df Sum Sq Mean Sq  F value Pr(>F)    
## lakename           8  11977    1497   131.54 <2e-16 ***
## depth              1 281002  281002 24690.16 <2e-16 ***
## lakename:depth     8   3298     412    36.22 <2e-16 ***
## Residuals       9466 107734      11                   
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(Temp.lm2)
```

```
## 
## Call:
## lm(formula = temperature_C ~ lakename * depth, data = my.NTL.LTER.data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -7.8106 -2.8094 -0.3609  2.5532 13.8700 
## 
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)                    21.81064    0.56344  38.710  < 2e-16 ***
## lakenameCrampton Lake          -1.02312    0.65966  -1.551 0.120939    
## lakenameEast Long Lake         -5.45328    0.59712  -9.133  < 2e-16 ***
## lakenameHummingbird Lake       -4.18064    0.82723  -5.054 4.41e-07 ***
## lakenamePaul Lake              -1.92030    0.57559  -3.336 0.000852 ***
## lakenamePeter Lake             -2.22934    0.57477  -3.879 0.000106 ***
## lakenameTuesday Lake           -4.65028    0.58319  -7.974 1.72e-15 ***
## lakenameWard Lake              -0.03052    0.79529  -0.038 0.969387    
## lakenameWest Long Lake         -3.57345    0.59349  -6.021 1.80e-09 ***
## depth                          -2.92866    0.22650 -12.930  < 2e-16 ***
## lakenameCrampton Lake:depth     1.59553    0.23133   6.897 5.65e-12 ***
## lakenameEast Long Lake:depth    1.51693    0.22879   6.630 3.54e-11 ***
## lakenameHummingbird Lake:depth  0.33279    0.27798   1.197 0.231265    
## lakenamePaul Lake:depth         1.11248    0.22752   4.890 1.03e-06 ***
## lakenamePeter Lake:depth        1.25097    0.22727   5.504 3.80e-08 ***
## lakenameTuesday Lake:depth      1.33037    0.22799   5.835 5.55e-09 ***
## lakenameWard Lake:depth         0.28497    0.26855   1.061 0.288662    
## lakenameWest Long Lake:depth    1.37717    0.22856   6.025 1.75e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.374 on 9466 degrees of freedom
## Multiple R-squared:  0.7333, Adjusted R-squared:  0.7329
## F-statistic:  1531 on 17 and 9466 DF,  p-value: < 2.2e-16
```

7. Is there a significant interaction between depth and lakename? How much variance in the temperature observations does this explain?

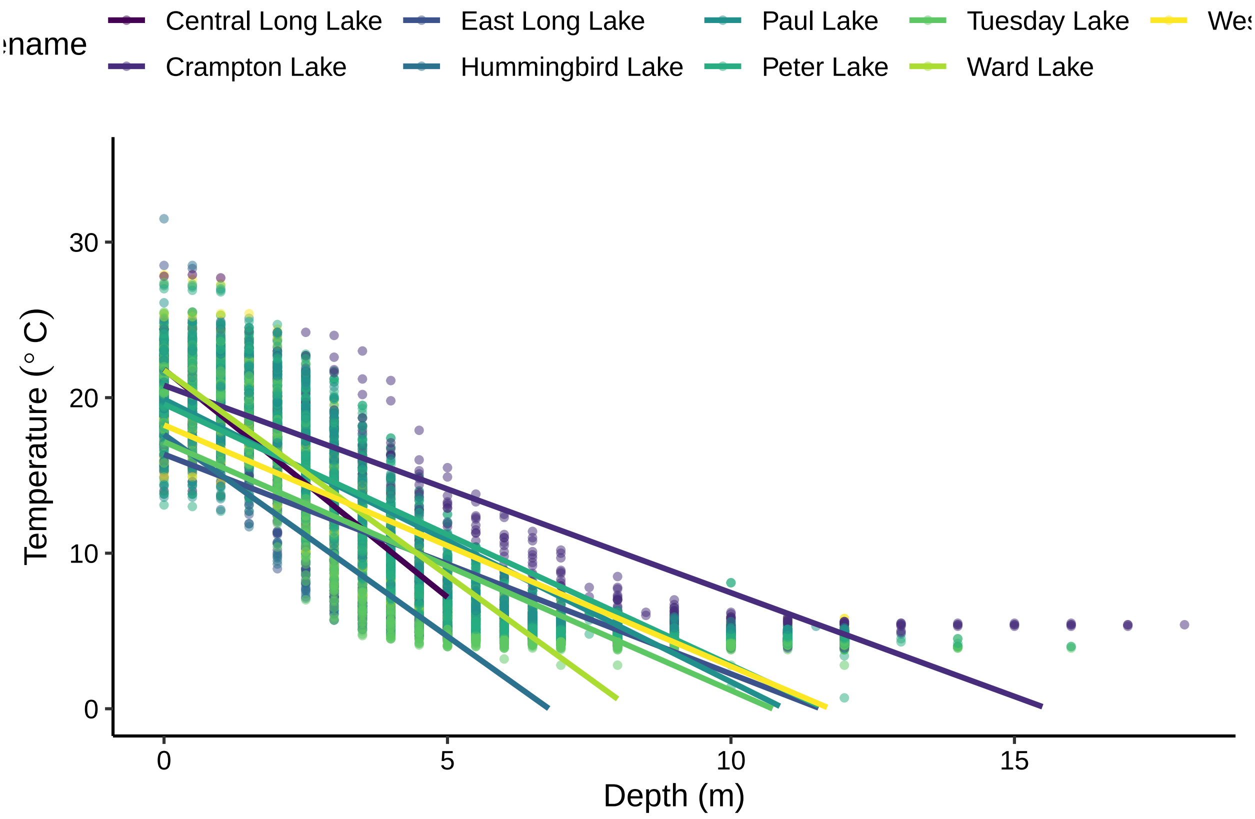
   Answer: The interaction between depth and lakename is significant because the p-value is less than 0.0001(F-value=36.22, df=8). This model(Temp.lm2) explains 73.3% of variance.

8. 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.

```
#8
p.NTL.lm <- ggplot(my.NTL.LTER.data, aes(x = depth, y = temperature_C, color = lakename)) +
  scale_color_viridis_d() +
  geom_point(alpha=0.5,size =1)+
  geom_smooth(method = "lm", se = FALSE)+
  ylim(0,35)+
  labs(x = "Depth (m)", y = expression("Temperature " ( degree~C)))

print(p.NTL.lm)
```

```
## Warning: Removed 89 rows containing missing values (geom_smooth).
```



9. Run a mixed effects model to predict dry mass of litter. We already know that nlcdClass and functionalGroup have a significant interaction, so we will specify those two variables as fixed effects

with an interaction. We also know that litter mass varies across plot ID, but we are less interested in the actual effect of the plot itself but rather in accounting for the variance among plots. Plot ID will be our random effect.

a. Build and run a mixed effects model.
b. Check the difference between the marginal and conditional R2 of the model.

```
Litter.hlm.Random <- lme(dryMass~ nlcdClass+functionalGroup+nlcdClass:functionalGroup, random= ~1|plotI
rsquared(Litter.hlm.Random)
```

```
##   Response   family     link method  Marginal Conditional
## 1  dryMass gaussian identity   none 0.2465822   0.2679023
```

```
 Litter.hlm <- lm(dryMass~ nlcdClass+functionalGroup+nlcdClass:functionalGroup, Litter_mass_trap)
rsquared(Litter.hlm)
```

```
##   Response   family     link method R.squared
## 1  dryMass gaussian identity   none 0.2515836
```

b. continued… How much more variance is explained by adding the random effect to the model?

   Answer: 2 %. Conditional Rsquare - Marginal Rsquare = 2.1%

c. Run the same model without the random effect.
d. Run an anova on the two tests.

```
 Litter.hlm <- lm(dryMass~ nlcdClass+functionalGroup+nlcdClass:functionalGroup, Litter_mass_trap)
rsquared(Litter.hlm)
```

```
##   Response   family     link method R.squared
## 1  dryMass gaussian identity   none 0.2515836
```

```
anova(Litter.hlm.Random,Litter.hlm)
```

```
##                   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## Litter.hlm.Random     1 26 9038.575 9179.479 -4493.287
## Litter.hlm            2 25 9058.088 9193.573 -4504.044 1 vs 2 21.51338  <.0001
```

d. continued… Is the mixed effects model a better model than the fixed effects model? How do you know?

   Answer: Yes. The mixed effects model has a lower AIC value than the fixed one.