Assignment 7: GLMs (Linear Regressios, ANOVA, & t-tests)

Jonathan Joyner

Fall 2023

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A07_GLMs.Rmd` (replacing `<FirstLast>` with your first and last name).
2. Change "Student Name" on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

Set up your session

1. Set up your session. Check your working directory. Load the tidyverse, agricolae and other needed packages. Import the *raw* NTL-LTER raw data file for chemistry/physics (`NTL-LTER_Lake_ChemistryPhysics_Raw.csv`). Set date columns to date objects.

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

```
#1
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(agricolae)
library(lubridate)
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(dplyr)
setwd("C:/Users/jbjoy/OneDrive/Documents/Grad School/Fall 2023/Environ 872/EDE_Fall2023/Data/Raw")
NTLRAW<-read.csv("NTL-LTER_Lake_ChemistryPhysics_Raw.csv",stringsAsFactors=TRUE)
NTLRAW$sampledate<-mdy(NTLRAW$sampledate)
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

Simple regression

Our first research question is: Does mean lake temperature recorded during July change with depth across all lakes?

3. State the null and alternative hypotheses for this question: > Answer: H0: July mean water temperatures do not change across all lake depth levels. Ha: July mean water temperatues vary depending on lake depth.

4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

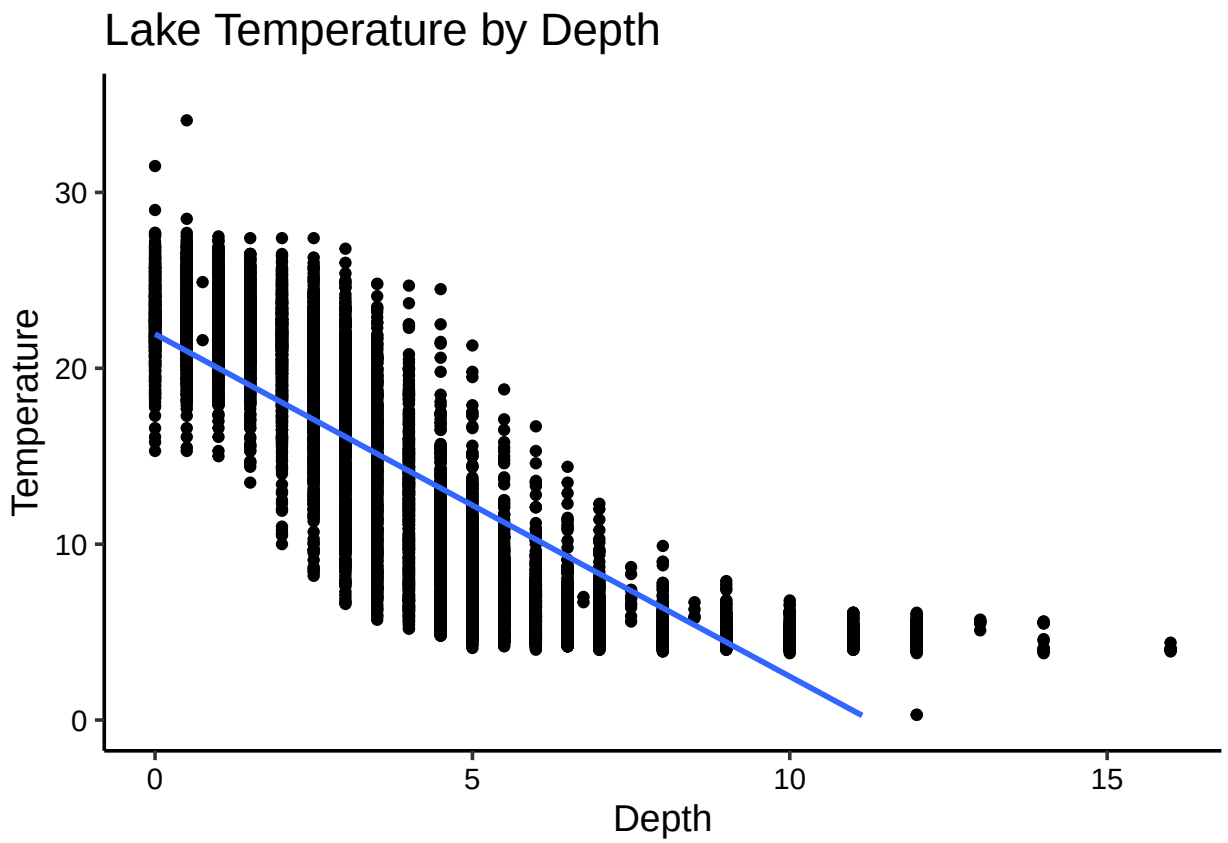
- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
#4
Q4Data<-
  NTLRAW %>%
  select(lakename,year4,daynum,depth,temperature_C) %>%
  filter(daynum%in%c(182:212)) %>%
  drop_na()
#5
Q4Scatter<-
  ggplot(Q4Data,aes(
    x=depth,
    y=temperature_C))+
  labs(title="Lake Temperature by Depth", x="Depth",y="Temperature")+
  geom_point()+
  ylim(0,35)+
  geom_smooth(method = "lm", se = FALSE)
plot(Q4Scatter)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: Removed 24 rows containing missing values (`geom_smooth()`).
```



6. Interpret the figure. What does it suggest with regards to the response of temperature to depth? Do the distribution of points suggest about anything about the linearity of this trend?

   Answer: The scatterplot suggests that across all lakes, the mean temperatures decreases in a similar manner.

7. Perform a linear regression to test the relationship and display the results

```
#7
Q4LR<- lm(
  data = Q4Data, temperature_C ~ depth)
summary(Q4LR)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = Q4Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5077 -3.0182  0.0743  2.9248 13.6033
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.94872    0.06790   323.3   <2e-16 ***
```

```
## depth       -1.94700    0.01173  -166.0   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.829 on 9720 degrees of freedom
## Multiple R-squared:  0.7391, Adjusted R-squared:  0.7391
## F-statistic: 2.754e+04 on 1 and 9720 DF,  p-value: < 2.2e-16
```

8. Interpret your model results in words. Include how much of the variability in temperature is explained by changes in depth, the degrees of freedom on which this finding is based, and the statistical significance of the result. Also mention how much temperature is predicted to change for every 1m change in depth.

   Answer: According to the linear regression results, there is a negative correlation between depth and temperature based on a slope of -1.95 which means that the temperature is decreasing around 1.95 degrees per 1 meter. The p value associated with these results is also less than 0.05 which means the correlation positied in our alternative hypothesis is demonstrably present. Around 74% of the variability can be explained by depth.

Multiple regression

Let's tackle a similar question from a different approach. Here, we want to explore what might the best set of predictors for lake temperature in July across the monitoring period at the North Temperate Lakes LTER.

9. Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to predict temperature.
10. Run a multiple regression on the recommended set of variables.

```
#9
Q4AIC<-lm(data=Q4Data,temperature_C~year4+daynum+depth)
summary(Q4AIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = Q4Data)
##
## 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
```

```
step(Q4AIC)
```

```
## 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 = Q4Data)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -6.45556      0.01013      0.04134     -1.94726
```

```
#10
Q4AIC<-lm(data=Q4Data,temperature_C~year4+daynum+depth)
summary(Q4AIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = Q4Data)
##
## 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
```

11. What is the final set of explanatory variables that the AIC method suggests we use to predict temperature in our multiple regression? How much of the observed variance does this model explain? Is this an improvement over the model using only depth as the explanatory variable?

Answer: The final set is year4,daynum, and depth as all have P values less than 0.05. Adding the year4 and daynum variables increases the R squared value from 73.91% to 74.17% which is a slight improvement over the single regression model with only depth as a variable.

Analysis of Variance

12. Now we want to see whether the different lakes have, on average, different temperatures in the month of July. Run an ANOVA test to complete this analysis. (No need to test assumptions of normality or similar variances.) Create two sets of models: one expressed as an ANOVA models and another expressed as a linear model (as done in our lessons).

```
#12
Q4ANOVA<-aov(data=Q4Data,temperature_C~lakename)
summary(Q4ANOVA)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21214  2651.8   49.04 <2e-16 ***
## Residuals   9713 525188    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

13. Is there a significant difference in mean temperature among the lakes? Report your findings.

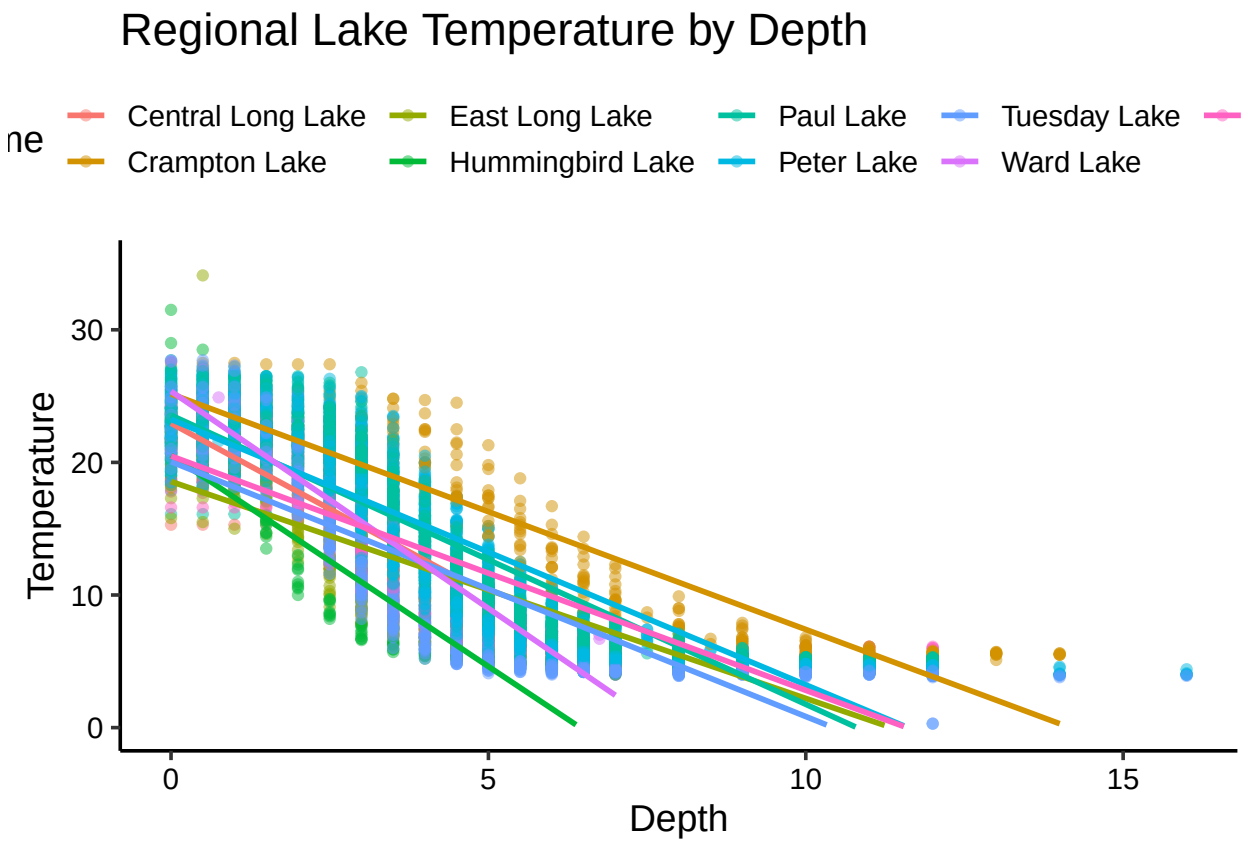
Answer: The anova test results in a P value below 0.05 which shows that the lakes do not have on average, different temperatues in July.

14. 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.

```
#14.
Q4MultiScatter<-
  ggplot(Q4Data,aes(
    x=depth,
    y=temperature_C,
    color=lakename))+
  labs(title="Regional Lake Temperature by Depth", x="Depth",y="Temperature")+
  geom_point(alpha=0.5)+
  ylim(0,35)+
  geom_smooth(method = "lm", se = FALSE)
plot(Q4MultiScatter)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

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



15. Use the Tukey's HSD test to determine which lakes have different means.

```
#15
print(HSD.test(Q4ANOVA,"lakename",group=T))
```

```
## $statistics
##    MSerror   Df     Mean       CV
##   54.07064 9713 12.70646 57.87035
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.387505  0.05
##
## $means
##                   temperature_C      std    r        se Min  Max    Q25   Q50
## Central Long Lake      17.67311 4.273404  119 0.6740735 8.9 26.8 14.400 18.40
## Crampton Lake          15.35189 7.244773  318 0.4123511 5.0 27.5  7.525 16.90
## East Long Lake         10.26767 6.766804  968 0.2363432 4.2 34.1  4.975  6.50
## Hummingbird Lake       10.77328 7.017845  116 0.6827343 4.0 31.5  5.200  7.00
## Paul Lake              13.79180 7.291951 2643 0.1430317 4.7 27.7  6.500 12.40
## Peter Lake             13.30207 7.667550 2892 0.1367356 4.0 27.0  5.600 11.40
## Tuesday Lake           11.06583 7.694274 1507 0.1894192 0.3 27.7  4.400  6.80
## Ward Lake              14.45862 7.409079  116 0.6827343 5.7 27.6  7.200 12.55
## West Long Lake         11.58552 6.963995 1043 0.2276872 4.0 25.7  5.400  8.00
##                       Q75
```

```
## Central Long Lake 21.350
## Crampton Lake     22.300
## East Long Lake    15.925
## Hummingbird Lake  15.625
## Paul Lake         21.400
## Peter Lake        21.500
## Tuesday Lake      19.400
## Ward Lake         23.200
## West Long Lake    18.800
##
## $comparison
## NULL
##
## $groups
##                   temperature_C groups
## Central Long Lake      17.67311      a
## Crampton Lake          15.35189     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.79180      c
## Peter Lake             13.30207      c
## West Long Lake         11.58552      d
## Tuesday Lake           11.06583     de
## Hummingbird Lake       10.77328     de
## East Long Lake         10.26767      e
##
## attr(,"class")
## [1] "group"
```

16.From the findings above, which lakes have the same mean temperature, statistically speaking, as Peter Lake? Does any lake have a mean temperature that is statistically distinct from all the other lakes?

Answer: Ward and Paul Lakes also belong to group c so they are statistically similar to Peter Lake. None of the lakes are statistically unique. They all have a group overlap with at least one other lake.

17. If we were just looking at Peter Lake and Paul Lake. What's another test we might explore to see whether they have distinct mean temperatures?

Answer: The two sample T-test would be a good way to compare the mean temperatures of both Peter and Paul lakes.

18. Wrangle the July data to include only records for Crampton Lake and Ward Lake. Run the two-sample T-test on these data to determine whether their July temperature are same or different. What does the test say? Are the mean temperatures for the lakes equal? Does that match you answer for part 16?

```
Q18Data<-
  NTLRAW %>%
  select(lakename,year4,daynum,depth,temperature_C) %>%
  filter(daynum%in%c(182:212),
           lakename=="Crampton Lake" |
           lakename=="Ward Lake")
t.test(Q18Data$temperature_C~Q18Data$lakename)
```

```
##
##  Welch Two Sample t-test
##
## data:  Q18Data$temperature_C by Q18Data$lakename
## t = 1.1181, df = 200.37, p-value = 0.2649
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -0.6821129  2.4686451
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.35189                    14.45862
```

Answer: The test shows that mean differences in temperature between the lakes are present. The p value is significantly higher than 0.05 and the sample estimate of Crampton Lake is nearly one whole degree higher on average than Ward Lake.