Assignment 6: GLMs (Linear Regressios, ANOVA, & t-tests)

Sammy DiLoreto

Spring 2023

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A06_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);library(agricolae);library(here);library(lubridate)
library(viridis);library(RColorBrewer);library(colormap);library(ggthemes)
#Check working directory
getwd()
```

```
## [1] "/Users/sammydiloreto/Library/CloudStorage/Box-Box/ENV872-EDA/EDA-Spring2023"
```

```
#Upload data sets
raw_data = "Data/Raw"
lake.chem.phys <- read.csv(here(raw_data,
       "NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
       stringsAsFactors = T)
#Change date columns to date format
lake.chem.phys$sampledate <- mdy(lake.chem.phys$sampledate)
class(lake.chem.phys$sampledate)
```

```
## [1] "Date"
```

```r
#2
# Set theme
my_theme <- theme_base() +
  theme(
    plot.background = element_rect(
      linewidth = 1
      ),
    plot.title = element_text(
      size = rel(1),
      face = "bold"
      ),
    axis.title = element_text(
      size = rel(0.8),
      face = "bold"
      ),
    axis.text = element_text(
        size = rel(0.6),
      face = "bold"
      ),
    legend.title = element_text(
      size = rel(0.7),
      face = "bold"
      ),
    legend.text = element_text(
      size = rel(0.6)
      ),
    legend.position = 'top',
    complete = TRUE
  )
theme_set(my_theme)
```

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: mean lake temps at all depths are equal (mu1 = mu2 = mu3 etc) Ha: mean lake temps at all depths are not equal (mu1 != mu2 != mu3 etc)

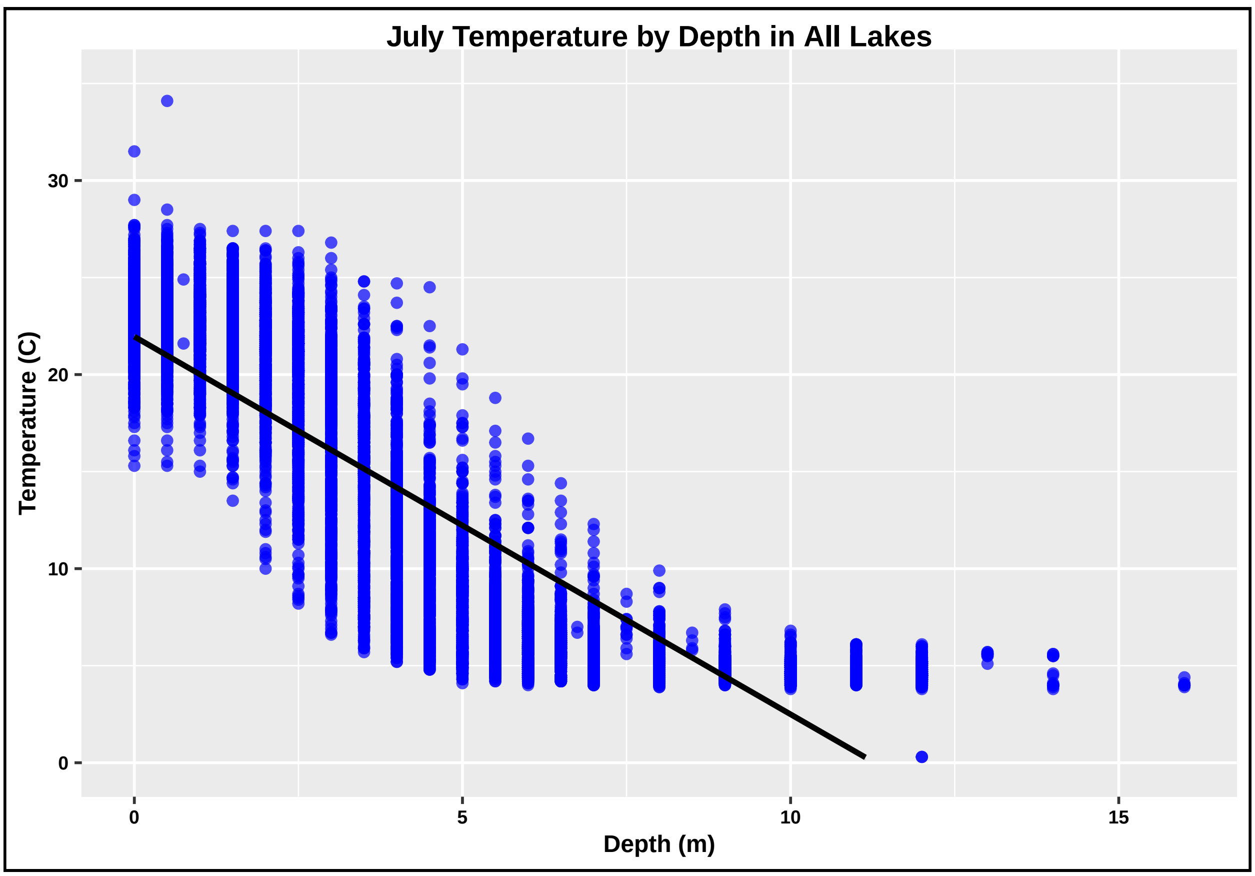
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

- 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.

```r
#4 wrangle your NTL-LTER dataset
lake.chem.phys.processed <-lake.chem.phys %>%
  filter(month(sampledate) == 07) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na()

#5 scatter plot of temperature by depth
TempByDepth.plot <- lake.chem.phys.processed %>%
  ggplot(
    mapping = aes(x=depth,
                  y=temperature_C)) +
  geom_point(color="blue",
             alpha=0.7,
             size=1.5)+
  ylim(0,35)+
  geom_smooth(method = "lm",
              color = "black")+
  labs(x = "Depth (m)" ,
       y = "Temperature (C)",
       title = "July Temperature by Depth in All Lakes")
print(TempByDepth.plot)
```



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 of temperature by depth suggests that as depth increases, temperature decreases. In other words, the deeper the lake, the colder it is. The distribution of these points suggests that this trend is a exponential, with a plateau at ~5 degrees C starting at 10 m. This is also shown by the lack of fit of the line.

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

```
#7 Perform a linear regression to test the relationship
TempByDepth.regression <- lm(data = lake.chem.phys.processed, temperature_C ~ depth)
summary(TempByDepth.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = lake.chem.phys.processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5173 -3.0192  0.0633  2.9365 13.5834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.95597    0.06792   323.3   <2e-16 ***
## depth       -1.94621    0.01174  -165.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.835 on 9726 degrees of freedom
## Multiple R-squared:  0.7387, Adjusted R-squared:  0.7387
## F-statistic: 2.75e+04 on 1 and 9726 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: There is a significant negative correlation between temperature and depth (lower temperature levels at greater depths) (p-value $< 2.2e-16$, df = 9726). 73.87% of the total variance in temperature is explained by changes in depth. For every 1 m change in depth there is a decrease in 1.95 degrees C.

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.

```r
#9 Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to p
Temp.ALL <- lm(data = lake.chem.phys.processed, temperature_C ~ year4 + daynum +
              depth)
summary(Temp.ALL)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = lake.chem.phys.processed)
##
## 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
```

```r
step(Temp.ALL)
```

```
## 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 = lake.chem.phys.processed)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```r
#10 Run a multiple regression on the recommended set of variables.
Temp.Best <- lm(data = lake.chem.phys.processed, temperature_C ~ year4 + daynum +
              depth)
summary(Temp.Best)
```

```
##
```

```
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = lake.chem.phys.processed)
##
## 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
```

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 AIC method suggests to use all given explanatory variables, year, day number, and depth, to predict temperature. This model explains 74.11% of the total variance in temperature. This is a very slight improvement in just using depth as the explanatory 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).

```r
#12
#Run ANOVA
Lake.Temps.anova <- aov(data = lake.chem.phys.processed, temperature_C ~ lakename)
summary(Lake.Temps.anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21642  2705.2      50 <2e-16 ***
## Residuals   9719 525813    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Format ANOVA as lm
Lake.Temps.anova2 <- lm(data = lake.chem.phys.processed, temperature_C ~ lakename)
summary(Lake.Temps.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = lake.chem.phys.processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.769  -6.614  -2.679   7.684  23.832
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                17.6664     0.6501  27.174  < 2e-16 ***
## lakenameCrampton Lake      -2.3145     0.7699  -3.006 0.002653 **
## lakenameEast Long Lake     -7.3987     0.6918 -10.695  < 2e-16 ***
## lakenameHummingbird Lake   -6.8931     0.9429  -7.311 2.87e-13 ***
## lakenamePaul Lake          -3.8522     0.6656  -5.788 7.36e-09 ***
## lakenamePeter Lake         -4.3501     0.6645  -6.547 6.17e-11 ***
## lakenameTuesday Lake       -6.5972     0.6769  -9.746  < 2e-16 ***
## lakenameWard Lake          -3.2078     0.9429  -3.402 0.000672 ***
## lakenameWest Long Lake     -6.0878     0.6895  -8.829  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.355 on 9719 degrees of freedom
## Multiple R-squared:  0.03953,    Adjusted R-squared:  0.03874
## F-statistic:    50 on 8 and 9719 DF,  p-value: < 2.2e-16
```

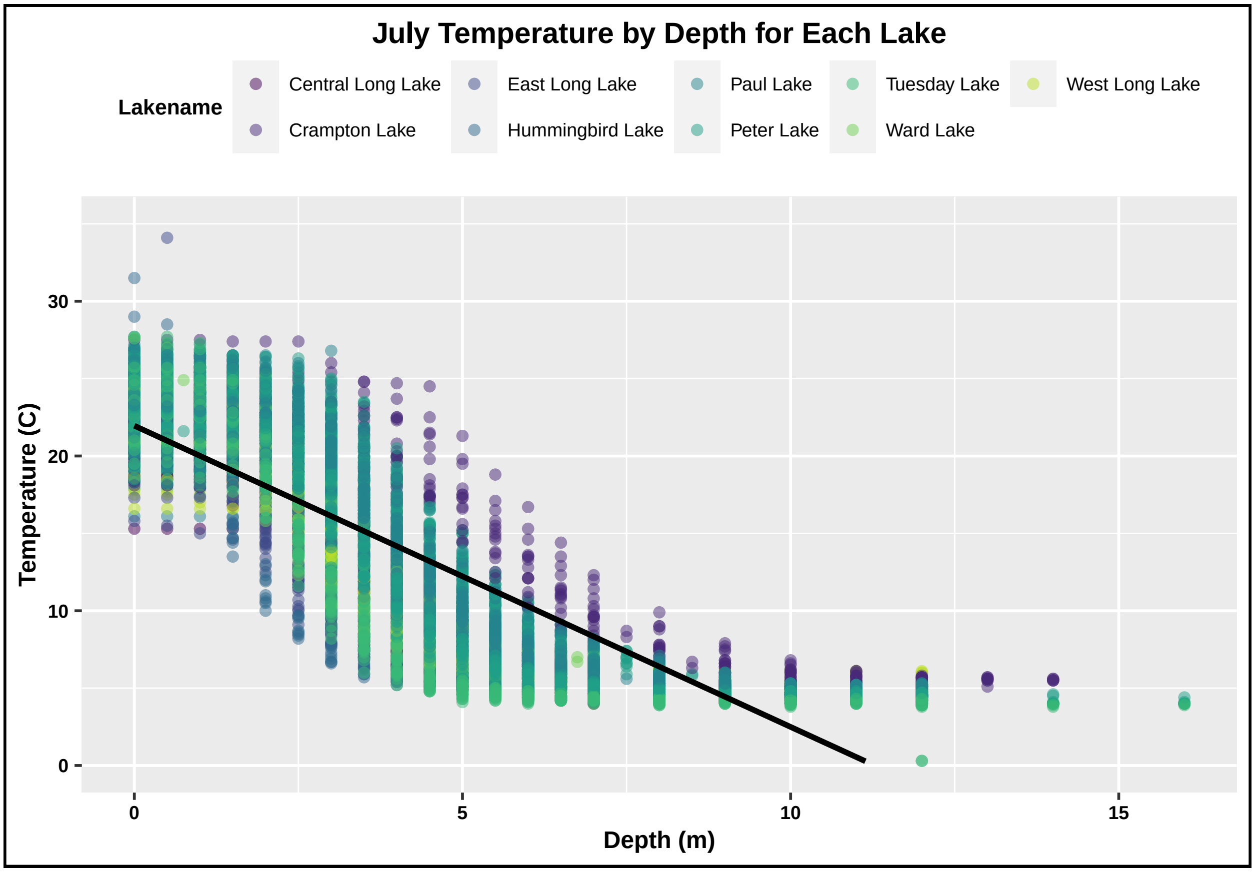
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: There is a significant difference between temperature among lakes (p-value $<$ 2.2e-16, df = 9719). This model explains about 4% of the total variance in temperature.

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.
TempByDepth.plot2 <- lake.chem.phys.processed %>%
  ggplot(
    mapping = aes(x=depth,
                  y=temperature_C,
                  color=lakename)) +
  geom_point(alpha=0.5,
             size=1.5)+
  ylim(0,35)+
  geom_smooth(method = "lm",
              se = FALSE,
              color = "black")+
  scale_color_viridis(
    discrete = TRUE,
    end = 0.9)+
  labs(x = "Depth (m)" ,
       y = "Temperature (C)",
       title = "July Temperature by Depth for Each Lake",
```

```r
       color = "Lakename")
print(TempByDepth.plot2)
```



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

```r
#15
#TukeyHSD(Lake.Temps.anova)
Lake.Temps.groups <- HSD.test(Lake.Temps.anova, "lakename", group = TRUE)
Lake.Temps.groups
```

```
## $statistics
##   MSerror   Df     Mean       CV
##   54.1016 9719 12.72087 57.82135
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.387504  0.05
##
## $means
##                    temperature_C      std   r Min  Max    Q25   Q50    Q75
## Central Long Lake       17.66641 4.196292 128 8.9 26.8 14.400 18.40 21.000
## Crampton Lake           15.35189 7.244773 318 5.0 27.5  7.525 16.90 22.300
## East Long Lake          10.26767 6.766804 968 4.2 34.1  4.975  6.50 15.925
## Hummingbird Lake        10.77328 7.017845 116 4.0 31.5  5.200  7.00 15.625
```

```
## Paul Lake             13.81426 7.296928 2660 4.7 27.7  6.500 12.40 21.400
## Peter Lake            13.31626 7.669758 2872 4.0 27.0  5.600 11.40 21.500
## Tuesday Lake          11.06923 7.698687 1524 0.3 27.7  4.400  6.80 19.400
## Ward Lake             14.45862 7.409079  116 5.7 27.6  7.200 12.55 23.200
## West Long Lake        11.57865 6.980789 1026 4.0 25.7  5.400  8.00 18.800
##
## $comparison
## NULL
##
## $groups
##                    temperature_C groups
## Central Long Lake      17.66641      a
## Crampton Lake          15.35189     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.81426      c
## Peter Lake             13.31626      c
## West Long Lake         11.57865      d
## Tuesday Lake           11.06923     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 Lake and Paul Lake have the same mean temperature, statistically speaking, as Peter Lake. No, every lake has at least one other lake that has the same mean temperature in statistical terms.

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: A two sample t-test would be the best option to compare temperature between Peter Lake and Paul Lake because it is used to compare a continuous variable between two samples. Temperature is our continuous variable and the two lakes is our categorical variable with two levels.

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?

```
#Wrangle the July data to include only records for Crampton Lake and Ward Lake
lake.chem.phys.processed2 <-lake.chem.phys.processed %>%
  filter(lakename %in% c("Crampton Lake", "Ward Lake"))

#Two-sample T-test to determine whether their July temperature are same or #different
Temp.twosample <- t.test(lake.chem.phys.processed2$temperature_C ~
                           lake.chem.phys.processed2$lakename)
Temp.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  lake.chem.phys.processed2$temperature_C by lake.chem.phys.processed2$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: We cannot reject the null hypothesis that states that the mean temperatures for Crampton Lake and Ward Lake are equal because the p-value is 0.2649. Yes this does match my answer for 16 because the Tukey HSD grouping shows that Crampton Lake and Ward Lake have the same mean temperature, statistically speaking.