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

Mara Michel

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
#Check working directory
getwd()
```

```
## [1] "C:/Users/marga/OneDrive/Documents/EDE_Fall2023"
```

```
#Load packages
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 1.9.3     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
```

```r
library(agricolae)
library(here)
```

```
## here() starts at C:/Users/marga/OneDrive/Documents/EDE_Fall2023
```

```r
library(lubridate)
library(dplyr)

#Import data
NTLRaw <- read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
                   stringsAsFactors = TRUE)

#Set date columns to date objects
NTLRaw$sampledate <- as.Date(NTLRaw$sampledate, format = "%m/%d/%y")

#2
#Modify theme and set it as default
mytheme <- theme_classic() +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top",
        panel.background = element_rect(fill = "aliceblue"),
        legend.title = element_text(size = 10),
        legend.text = element_text(size = 8))
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: Mean lake temperature recorded during July does not change with depth across all lakes.
   Ha: Mean lake temperature recorded during July changes with depth across all lakes.

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.

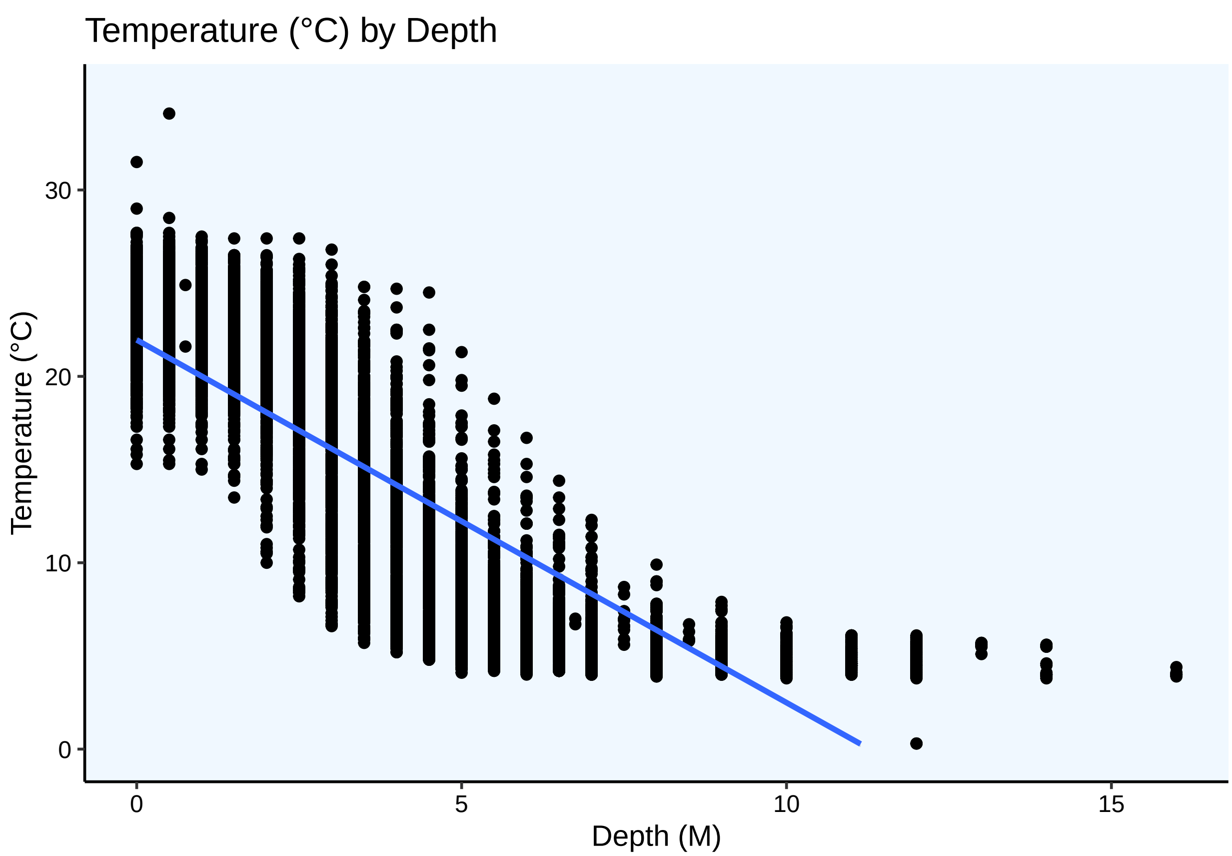
```r
#4 Wrangle Data using pipe function
NTLJuly <- NTLRaw %>%
  filter(month(NTLRaw$sampledate) %in% 7:7) %>%
  select('lakename','year4','daynum', 'depth', 'temperature_C')%>%
  na.omit()
```

```r
#5 Create and display scatterplot of wrangled data
NTLJulyPlot <- ggplot(NTLJuly,aes(x=depth,y=temperature_C))+
  geom_point()+
  geom_smooth(method="lm")+
  ylim(0,35)+
  labs(x="Depth (M)",y="Temperature (\u00B0C)",title="Temperature (\u00B0C) by Depth")

NTLJulyPlot
```

```
## `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: There is a negative correlation between temperature and depth: the deeper the water, the lower the temperature.The distribution of the points is downward sloping from left to right which indicates this linear trend. The points also appear to be stacked in columns, signaling that the same depths may have different temperatures.This phenonomnom could be due to differences in temperatures across lakes or across years since the only consistency across measurements is the month.

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

```
#7
NTLJuly.Regression <-lm(data=NTLJuly,temperature_C~depth)
summary(NTLJuly.Regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTLJuly)
##
## 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: The linear regression model measures the effect depth (M) has on temperature (C). The model has an R-squared value of 0.7387, indicating that changes in depth explain 73.87% of the change in temperature. This is based on 9726 degrees of freedom. The P-value is 2.2e-16, signaling extreme statistical significance. The model found that temperatures decrease by -1.94621 degrees C for each additional meter of 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
NTLJuly.RegressionAll <-lm(data=NTLJuly,temperature_C~year4+depth+daynum)
step(NTLJuly.RegressionAll)
```

```
## Start:  AIC=26065.53
## temperature_C ~ year4 + depth + daynum
##
##          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 + depth + daynum, data = NTLJuly)
##
## Coefficients:
## (Intercept)        year4        depth       daynum
##    -8.57556      0.01134     -1.94644      0.03978
```

```
#Including all factors is the best option since <none> has the lowest AIC vaue.

#10
summary(NTLJuly.RegressionAll)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + depth + daynum, data = NTLJuly)
##
## 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 **
## depth       -1.946437   0.011683 -166.611  < 2e-16 ***
## daynum       0.039780   0.004317    9.215  < 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 final set of explanatory variables that the AIC method suggests are year4, depth, and daynum. This model explains 74.12% of the variance which is an improvement of 0.25% over the model with just depth.

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
#Format as AOV
NTLJuly.AOV.Lake <-aov(data=NTLJuly,temperature_C~lakename)
summary(NTLJuly.AOV.Lake)
```

```
##               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
```

```
#Format as LM
NTLJuly.Regression.Lake <-lm(data=NTLJuly,temperature_C~lakename)
summary(NTLJuly.Regression.Lake)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTLJuly)
##
## 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 in mean temeprature among the lakes as determined by the P value of 2e-16, which is less than the statistically significant benchmark of <0.05.

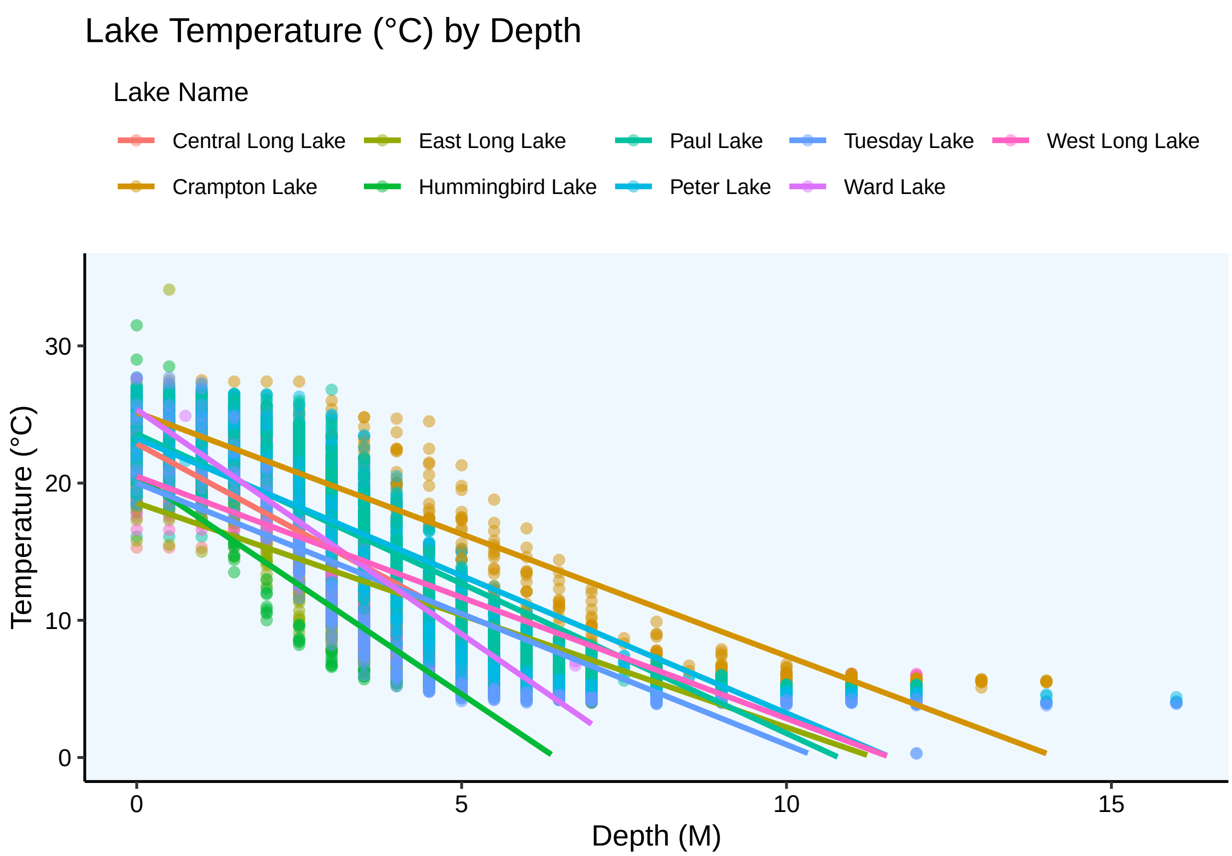
14. Create a graph that depicts temperature by depth, with a separate color for each lake. Add a geom_smooth (method = "lm", se = FALSE) for each lake. Make your points 50 % transparent. Adjust your y axis limits to go from 0 to 35 degrees. Clean up your graph to make it pretty.

```
#14.
NTLJulyPlot.Lake <-ggplot(NTLJuly,aes(x=depth,y=temperature_C,color=lakename))+
  geom_point(alpha=0.5)+
  geom_smooth(method="lm",
              se=FALSE)+
  ylim(0,35)+
  labs(x="Depth (M)",
       y="Temperature (\u00B0C)",
       title="Lake Temperature (\u00B0C) by Depth",colour="Lake Name")+
  guides(colour = guide_legend(title.position = "top"))

NTLJulyPlot.Lake
```

```
## `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
NTL.HSD <-HSD.test(NTLJuly.AOV.Lake,"lakename",group=TRUE)
NTL.HSD
```

```
## $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        se Min  Max    Q25   Q50
## Central Long Lake       17.66641 4.196292  128 0.6501298 8.9 26.8 14.400 18.40
## Crampton Lake           15.35189 7.244773  318 0.4124692 5.0 27.5  7.525 16.90
## East Long Lake          10.26767 6.766804  968 0.2364108 4.2 34.1  4.975  6.50
## Hummingbird Lake        10.77328 7.017845  116 0.6829298 4.0 31.5  5.200  7.00
## Paul Lake               13.81426 7.296928 2660 0.1426147 4.7 27.7  6.500 12.40
## Peter Lake              13.31626 7.669758 2872 0.1372501 4.0 27.0  5.600 11.40
## Tuesday Lake            11.06923 7.698687 1524 0.1884137 0.3 27.7  4.400  6.80
## Ward Lake               14.45862 7.409079  116 0.6829298 5.7 27.6  7.200 12.55
## West Long Lake          11.57865 6.980789 1026 0.2296314 4.0 25.7  5.400  8.00
##                       Q75
## Central Long Lake 21.000
## 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       17.66641      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: According to the results of the HSD.test groups, Paul Lake has the same mean temperature, statistically speaking, as Peter Lake as both are assigned to group "c". Ward Lake, assigned group "bc" also has an overlap with Peter Lake. While Central Long Lake, Crampton Lake, Ward Lake, West Long Lake, and East Long Lake do not have the statistically same mean temperatures as another lake, they all have an overlap with at least one other lake and so no lakes have a statistically distinct mean.

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-way T-test can be run to determine if Peter and Paul Lakes have distinct mean temperatures.

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 your answer for part 16?

```r
NTL.Crampton.Ward <-NTLJuly %>%
  filter(lakename == c("Crampton Lake", "Ward Lake"))

NTL.Crampton.Ward.Ttest <-t.test(NTL.Crampton.Ward$temperature_C~NTL.Crampton.Ward$lakename)
NTL.Crampton.Ward.Ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL.Crampton.Ward$temperature_C by NTL.Crampton.Ward$lakename
## t = 0.98673, df = 95.77, p-value = 0.3263
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -1.130614  3.365610
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.37107                    14.25357
```

> Answer: The p-value of the test (0.3263) is greater than 0.05 and so we fail to reject the null hypothesis. While we cannot formally conclude that the means are the same (because one cannot technically accept a null hypothesis), the two-way t-test estimates that the mean temperature of Crampton Lake is 15.37107 and the mean of Ward Lake is 14.25357 and these are different values. These values are slightly different from the results for question 16. In question 16, Crampton and Ward had calculated means of 15.35189 and 14.45862, respectively, and were assigned different but overlapping group values.