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

Jake Whisler

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 Checking working directory, loading necessary packages, and importing data
getwd()
```

```
## [1] "/home/guest/EDE_Fall2023"
```

```
library(lubridate)
```

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

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
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 purrr   1.0.2     v tidyr   1.3.0

## -- 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(here)
```

```
## here() starts at /home/guest/EDE_Fall2023
```

```
library(dplyr)

NTL <- read.csv("~/EDE_Fall2023/Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")
NTL$sampledate <- as.Date(NTL$sampledate, format = "%m/%d/%Y")
class(NTL$sampledate)
```

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

```
#2
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: Mean lake temperature in July does not change with depth according to our data. Ha: Mean lake temperature in July does change with depth according to our data.
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)

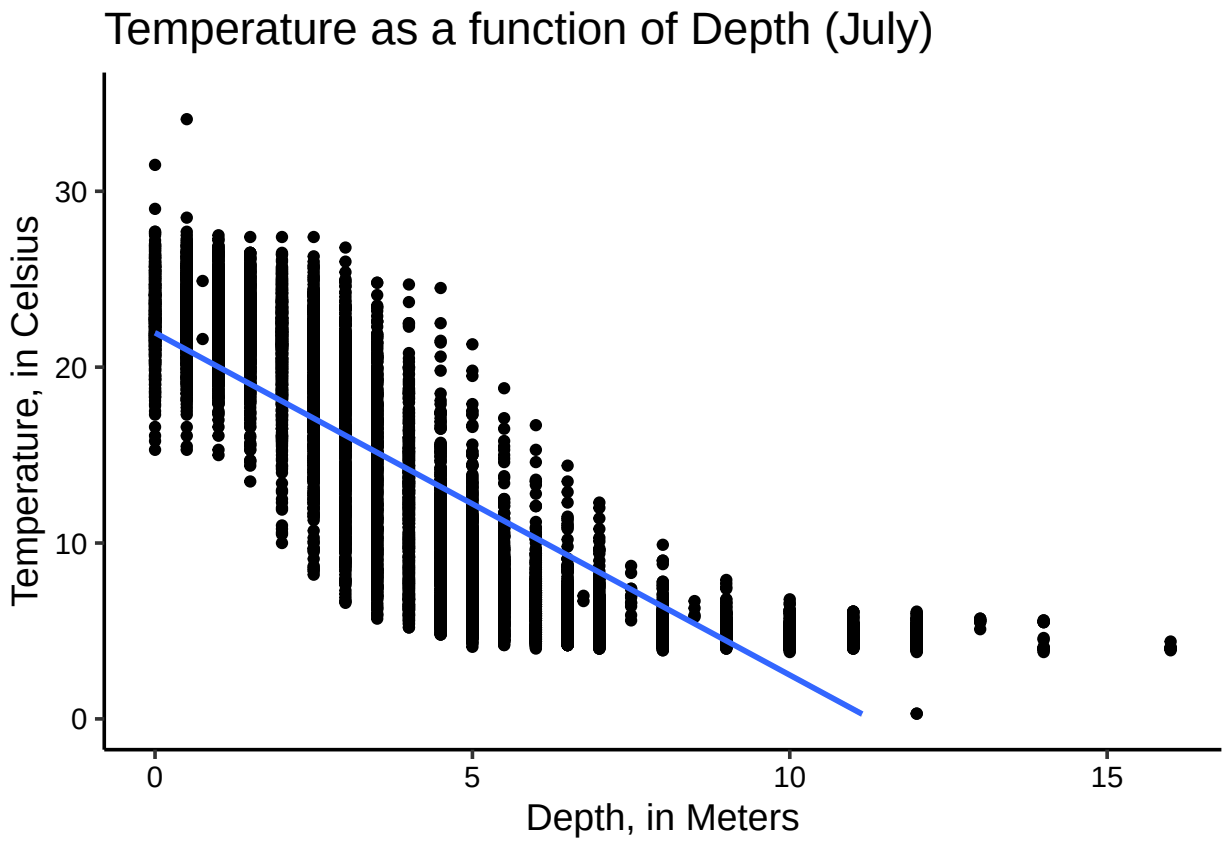
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```r
#4 Wrangling the dataset
NTL <- mutate(NTL, month = month(sampledate))
NTL_wrangled <-
  NTL %>%
  filter(month == "7") %>%
  select("lakename", "year4", "daynum", "depth", "temperature_C") %>%
  na.omit("temperature_C")

#5 Visualizing temperature by depth with linear model
NTL_depth_vs_temp <-
  ggplot(data = NTL_wrangled, aes(x = depth, y = temperature_C)) +
  ylim(0,35) +
  geom_point() +
  geom_smooth(method = "lm") +
  xlab("Depth, in Meters") +
  ylab("Temperature, in Celsius") +
  ggtitle("Temperature as a function of Depth (July)")
print(NTL_depth_vs_temp)
```

```
## `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: Generally speaking, it seems temperature decreases with depth, eventually reaching leveling out around 5 degrees Celsius at the deepest depths recorded. However, there is a considerable amount of variability in temperature for samples taken at the same depth, so the relationship is not entirely linear.

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

```
#7 Creating a linear regression of temperature as a function of depth
NTL_depth_vs_temp_regression <-
  lm(data = NTL_wrangled, temperature_C ~ depth)
summary(NTL_depth_vs_temp_regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL_wrangled)
##
## 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: First of all, the intercept is estimated to be 21.956 degrees Celsius, meaning that at zero meters of depth (i.e. the surface of the lake), the water temperature is predicted to be 21.956 degrees Celsius. Approximately 73.87% of the variability in temperature is explained by changes in depth. This finding is based on 9727 degrees of freedom. The p-value returned by this linear model is less than 2.2e-16, meaning that it is statistically significant and we can assume our alternative hypothesis to be true. For each additional 1m change in depth, there is an estimated decrease of 1.946 degrees Celsius in temperature.

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 Choosing variables based on AIC Stepwise algorithm
TPAIC <- lm(data = NTL_wrangled, temperature_C ~ depth + daynum + year4)
step(TPAIC)
```

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

```
#10 Running the multiple regression
NTL_day_year_vs_temp <-
  lm(data = NTL_wrangled, temperature_C ~ daynum + year4)
summary (NTL_day_year_vs_temp)
```

```
##
## Call:
## lm(formula = temperature_C ~ daynum + year4, data = NTL_wrangled)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.279  -7.158  -2.591   8.072  21.402
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.827705  16.944033  -0.167    0.867
## daynum       0.040484   0.008475   4.777 1.81e-06 ***
## year4        0.003779   0.008439   0.448    0.654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.494 on 9725 degrees of freedom
## Multiple R-squared:  0.002363,   Adjusted R-squared:  0.002158
## F-statistic: 11.52 on 2 and 9725 DF,  p-value: 1.007e-05
```

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 gave us lower AIC values for "daynum" and "year4", meaning that we should use those variables to predict temperature in our multiple regression. However, this model explains approximately 0.2% of the variability in temperature. That being said, this is not an improvement upon our original model 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).

```
#12 Creating the ANOVA and linear models
NTL.anova <- aov(data = NTL_wrangled, temperature_C ~ lakename)
summary(NTL.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
```

```
NTL.lm <-
  lm(data = NTL_wrangled, temperature_C ~ lakename)
summary(NTL.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_wrangled)
##
## 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.

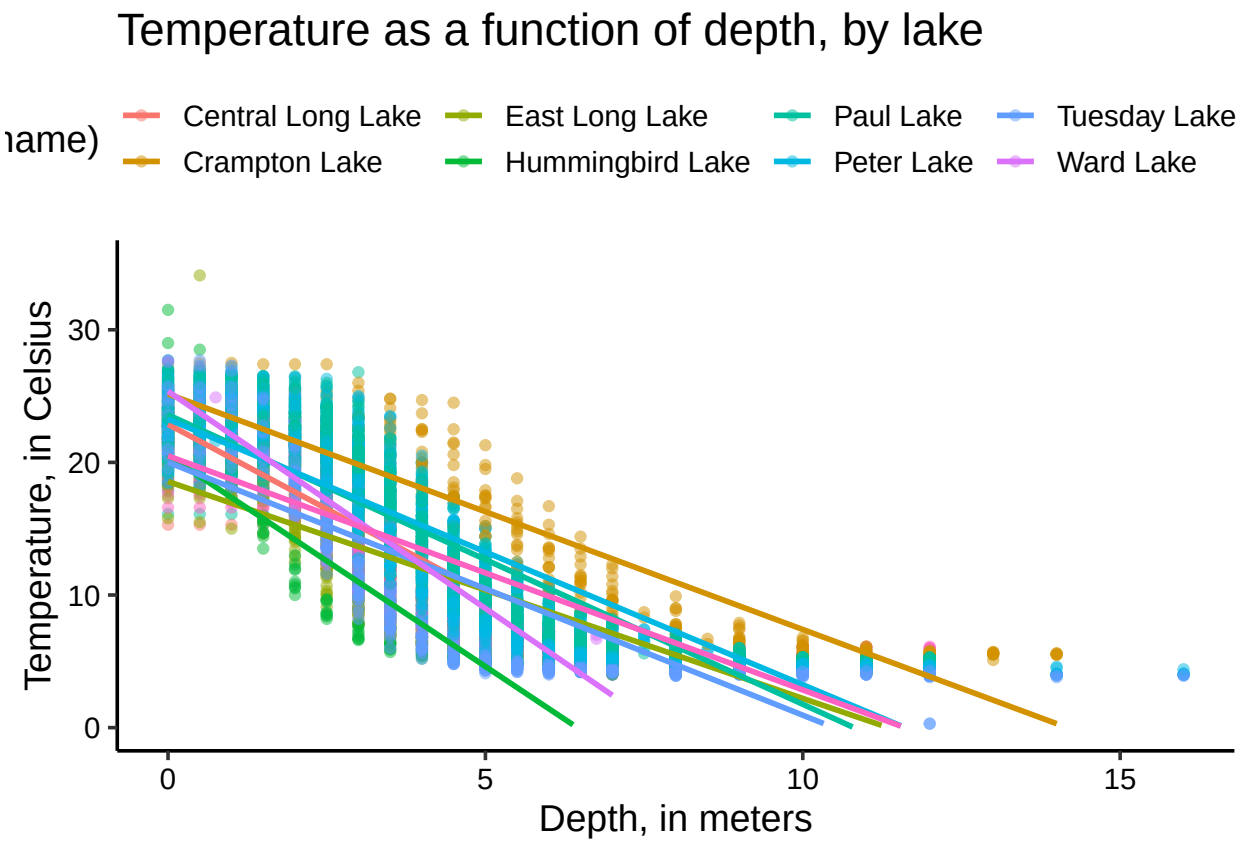
    Answer: The p-value for both the ANOVA and the linear model indicate that there is a significant difference in mean temperature among the lakes, since it is well below the average benchmark for statistical significance (0.05).

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. Creating graph of temperature by depth with colors separating lakes
NTL_temp_by_lake_graph <-
  ggplot(data = NTL_wrangled, aes(x = depth, y = temperature_C, color = factor(lakename))) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  ylim(0,35) +
  xlab("Depth, in meters") +
  ylab("Temperature, in Celsius") +
  ggtitle("Temperature as a function of depth, by lake")
print(NTL_temp_by_lake_graph)
```

```
## `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 Performing the Tukey HSD test
NTLgroups <- HSD.test(NTL.anova, "lakename", group = TRUE)
print(NTLgroups)
```

```
## $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: There are two groups of two lakes that have statistically the same mean as one another: Peter Lake and Paul Lake have a statistically similar mean temperature, and Tuesday Lake and Hummingbird Lake share a statistically similar mean as well. All the rest of the lakes have statistically distinct mean temperatures from one another.

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: We would run a two-sample t-test to see if their mean temperatures in July match.

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?

```
NTL_wrangled_CrWa <-
  NTL_wrangled %>%
  filter(lakename %in% c("Crampton Lake","Ward Lake"))

NTL.twosample <- t.test(NTL_wrangled_CrWa$temperature_C ~ NTL_wrangled_CrWa$lakename)
NTL.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL_wrangled_CrWa$temperature_C by NTL_wrangled_CrWa$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 mean temperatures for Crampton Lake and Ward Lake are not equal, as the t-test returned a significance value of 0.26, which is much higher than the normal 0.05 for statistical significance, meaning that there is insufficient evidence to say that they are the same. This indicates that they are not the same, which matches the results from the Tukey HSD groupings.