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

Ellie Harrigan

Fall 2024

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     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## 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(ggplot2)
library(dplyr)
library(here)
```

```
## here() starts at /home/guest/EDE_Fall2024
```

```r
getwd()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
here()
```

```
## [1] "/home/guest/EDE_Fall2024"
```

```r
NTL.LTR <- read.csv(
  file = here('Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv'),
stringsAsFactors = TRUE)

NTL.LTR$sampledate <- as.Date(NTL.LTR$sampledate, format = "%m/%d/%y")

#2
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
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: There is no difference in mean lake temperature recorded during July across different depths for all lakes Ha: The 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.

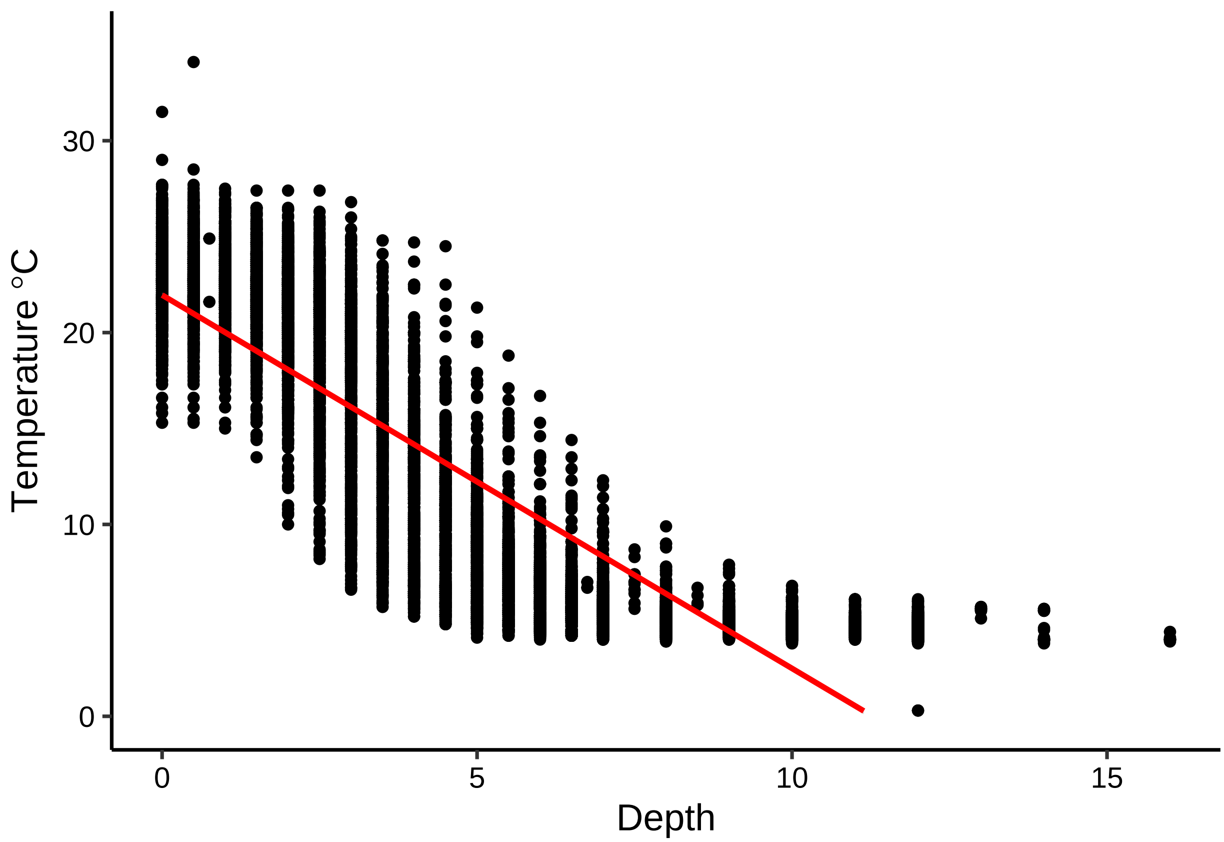
```r
#4
NTL.LTR.wrangle <- NTL.LTR %>%
  filter(format(sampledate, "%m") == "07") %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()

#5
NTL.LTR.temp.plot <-
  ggplot(NTL.LTR.wrangle, aes(x = depth, y = temperature_C)) +
```

```r
  geom_point() +
  geom_smooth(method = "lm", col="red") +
  ylim(0,35) +
  xlab("Depth") +
  ylab(expression(Temperature ~ degree*C))
print(NTL.LTR.temp.plot)
```

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

## Warning: Removed 24 rows containing missing values or values outside the scale range
## (`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: As depth increases temperature decreases.

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

```r
#7
TempbyDepth <- lm(
  data = NTL.LTR.wrangle, temperature_C  ~ depth)
summary(TempbyDepth)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL.LTR.wrangle)
##
## 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
```

```
#step(TempbyDepth)
```

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: 73.87% of variability in water temperature is explained by changes in depth, as highlighted by the Multiple R-Squared Value. This is based on the 9726 degrees of freedom for the residual error, and there is a statistically significant relationship between depth and temperature. For every 1m increase in depth, temperature is predicted to change by about 1.95 degrees Celcius, as shown by the depth coefficient of -1.946. This model shows a very significant, negative relationship between depth and temperature, explaining a big portion of the temperature variability is the dataset.

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 -
temp.aic <- lm(data = NTL.LTR.wrangle, temperature_C ~ year4 + daynum + depth)
step(temp.aic)
```

```
## 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 = NTL.LTR.wrangle)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    -8.57556      0.01134      0.03978     -1.94644
```

```
#10
temp.model <- lm(data = NTL.LTR.wrangle, temperature_C ~ year4 + daynum + depth)
summary(temp.model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL.LTR.wrangle)
##
## 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 final set of explanatory variables the AIC method suggests are year4, daynum, and depth. The observed variance is 74% (from the Multiple R-Squared value). This model is a slight improvement from the model using only depth, which an increase of about .25% of variance from the Multiple R-Squared value and a lower AIC.

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
NTL.LTR.anova <- aov(data = NTL.LTR.wrangle, temperature_C  ~ lakename)
summary(NTL.LTR.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.LTR.model <- lm(data = NTL.LTR.wrangle, temperature_C ~ lakename)
summary(NTL.LTR.model)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL.LTR.wrangle)
##
## 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.

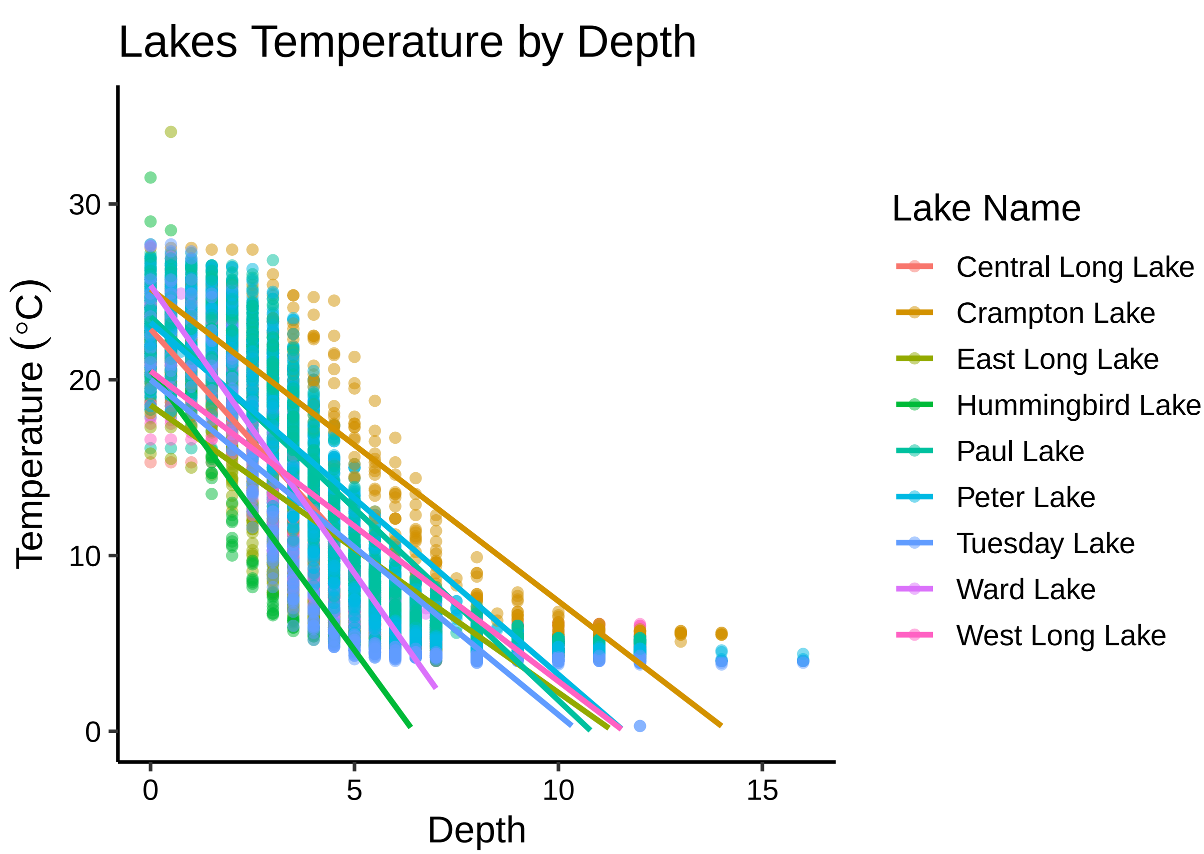
    Answer: For the dates in July, there is statistical significance between mean temperature among the lakes. The anova test has an extremely small p-value of <2e-16 which indicates the mean temperatures among lakes are statistically significant. The linear model shows the coefficients for each lake compared to the intercept. Each coefficient for the lakes has a negative estimate which indicates that the lakes have a lower average temperature compared to the intercept. These findings rejects the null hypothesis that the lakes have the same mean 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.

```r
#14.
NTL.temp.depth.plot <- ggplot(NTL.LTR.wrangle,
                              aes(x = depth, y = temperature_C,
                                  color = lakename)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  ylim(0, 35) +
  labs(
    title = "Lakes Temperature by Depth",
    x = "Depth",
    y = expression("Temperature" ~ (degree * C)),
    color = "Lake Name"
  ) +
  mytheme
print(NTL.temp.depth.plot)
```

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

```
## Warning: Removed 73 rows containing missing values or values outside the scale range
## (`geom_smooth()`).
```



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

```
#15 # is the function right?
NTL.LTR.lakemean <- aov(data = NTL.LTR, temperature_C  ~ lakename)

NTL.LTR.means <- HSD.test(NTL.LTR.lakemean, "lakename", group = TRUE)
NTL.LTR.means
```

```
## $statistics
##    MSerror    Df     Mean       CV
##   46.49528 34747 11.80871 57.74336
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.386509  0.05
##
## $means
##                    temperature_C      std     r          se Min  Max  Q25  Q50
## Central Long Lake      16.736343 4.540842   443 0.32396833 1.3 27.9 13.3 16.7
## Crampton Lake          14.192058 6.801706  1108 0.20484933 4.8 27.5  7.1 13.8
## East Long Lake          9.779296 6.304109  3550 0.11444327 3.8 34.1  4.9  6.4
## Hummingbird Lake       10.037831 6.117160   378 0.35071838 4.0 31.5  5.1  6.9
## Paul Lake              12.792275 6.783047  9253 0.07088644 3.9 27.7  6.0 11.3
## Peter Lake             12.252557 7.119817 10189 0.06755207 0.7 27.2  5.2 10.2
## Tuesday Lake           10.346702 7.027998  5503 0.09191887 0.3 27.7  4.4  6.4
## Ward Lake              12.428083 6.575945   527 0.29702918 5.0 27.6  6.6  9.9
## West Long Lake         11.058581 6.555168  3805 0.11054194 4.0 27.9  5.4  7.7
##                      Q75
## Central Long Lake 20.35
## Crampton Lake     20.80
## East Long Lake    14.70
## Hummingbird Lake  14.70
## Paul Lake         19.50
## Peter Lake        19.40
## Tuesday Lake      17.00
## Ward Lake         18.20
## West Long Lake    17.40
##
## $comparison
## NULL
##
## $groups
##                    temperature_C groups
## Central Long Lake      16.736343      a
## Crampton Lake          14.192058      b
## Paul Lake              12.792275      c
## Ward Lake              12.428083     cd
## Peter Lake             12.252557      d
## West Long Lake         11.058581      e
## Tuesday Lake           10.346702      f
## Hummingbird Lake       10.037831     fg
## East Long Lake          9.779296      g
##
## 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: The lake with the statistically similar mean temperature to Peter Lake is Ward Lake. Yes, the lake with a mean temperature that are statistically distinct from all the others is Central Long 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: Another test we might explore would be a two-samle t-test which can test the hypothesis that the mean of two samples (Peter and Paul Lake) have an equivalent mean.

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?

```r
NTL.LTR.crampton.ward <- NTL.LTR.wrangle %>%
  filter(lakename %in% c("Crampton Lake", "Ward Lake"))

NTL.LTR.twosample <- t.test(NTL.LTR.crampton.ward$temperature_C ~
                              NTL.LTR.crampton.ward$lakename)
NTL.LTR.twosample
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL.LTR.crampton.ward$temperature_C by NTL.LTR.crampton.ward$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: Based on the two-sample t-test it appears that the mean temperatures for July from Crampton Lake and Ward Lake are not statistically significantly different, with a p-value of 0.2649 gives evidence to this. The anova test in part 16 suggests that these two lakes are in distinct temperature groups, indicating that trends from this are not show in this specific t-test.