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

Rachael Stephan

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
#load libraries
library(tidyverse); library(agricolae); library(here); library(lubridate)

#check working directory
getwd()
```

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

```
here()
```

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

```
#set up chunk options
knitr::opts_chunk$set(message = FALSE, warning = FALSE)

#load in data
NTL_chemphys <- read.csv(here("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"), stringsAsFactors = 
  mutate(sampledate = mdy(sampledate))

#2
#create default theme
mytheme <- theme_bw(base_size = 10)+
  theme(axis.title = element_text(size = 10, hjust = 0.5),
        plot.title.position = "panel",
        legend.box = "vertical",
        legend.location = "plot",
        axis.gridlines = element_line(color = "grey", linewidth = 0.25),
        axis.ticks = element_line(color = "black", linewidth = 0.5))
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:

    > H0: The mean lake temperature does not change with depth during July across all lakes (slope = 0)
    >
    > Ha: The mean lake temperature does change with depth during July across all lakes (slope =/= 0)

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

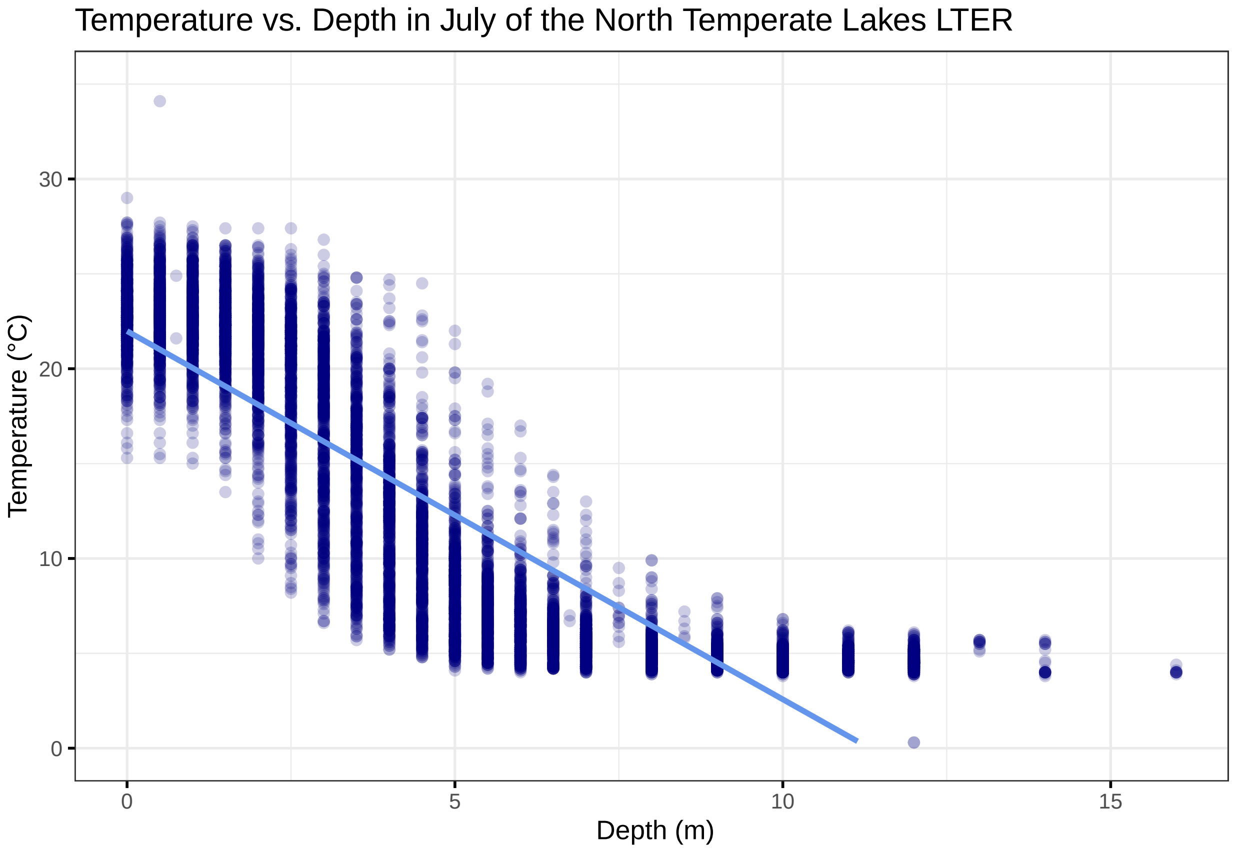
- Only dates in July - day number 183 to 213.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

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
#wrangle data
NTL_chemphys_wrangled <- NTL_chemphys %>%
  filter(daynum %in% 183:213) %>% #get days in july
  select(lakename, year4, daynum, depth, temperature_C) %>% # select wanted
  na.omit() #remove empty cases
```

```
#5
#create scatterplot
NTL_vis <- ggplot(data = NTL_chemphys_wrangled, aes(x=depth, y=temperature_C))+
  geom_point(alpha = 0.2, colour = "navy")+
  geom_smooth(method = lm, colour = "cornflowerblue", se = FALSE)+
  ylim(0, 35)+
  labs(x = " Depth (m)",
       y = "Temperature (°C)",
       title = "Temperature vs. Depth in July of the North Temperate Lakes LTER")+
  mytheme

NTL_vis
```



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 figure suggests there is an inverse relationship between temperature and depth (i.e., at larger depths, there is a lower temperature). However, the points don't seem to follow a linear trend. There seems to be nonlinearity to the relationship (points create a bit of a backwards s-shaped curve).

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

```
#7
#create regression
NTL.regression <- lm(data = NTL_chemphys_wrangled, temperature_C ~ depth)

#view regression
summary(NTL.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL_chemphys_wrangled)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5606 -3.0380  0.0872  2.9872 13.4706
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.98318    0.06840   321.4   <2e-16 ***
## depth       -1.94086    0.01179  -164.7   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.852 on 9671 degrees of freedom
## Multiple R-squared:  0.7371, Adjusted R-squared:  0.7371
## F-statistic: 2.712e+04 on 1 and 9671 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 p-value for depth is <0.05. Therefore, depth is a significant factor for water temperature. The R-squared value is 0.7371, meaning 73.71% of the variation is explained by depth in this model. This is based on 9671 degrees of freedom, which is calculated with the number of obsevations and the number of factors in the model. The intercept is 21.98318, indicating the surface temperature of the lakes is about 22°C. The depth coefficient is -1.94086, indicating that for every increase in depth of 1m, the temperature drops by ~1.94°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
#create regression with all variables
NTL.regression.aic <- lm(data = NTL_chemphys_wrangled, temperature_C ~ depth + year4 + daynum)

#use AIC to determine best formula
step(NTL.regression.aic)
```

```
## Start:  AIC=25998.22
## temperature_C ~ depth + year4 + daynum
##
##          Df Sum of Sq    RSS   AIC
## <none>                142056 25998
## - year4   1       201 142257 26010
## - daynum  1      1237 143293 26080
## - depth   1    402549 544605 38995
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL_chemphys_wrangled)
##
## Coefficients:
## (Intercept)        depth        year4       daynum
##   -18.19700     -1.94133      0.01611      0.04024
```

```
#10
#create best regression
NTL.regression.best <- lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL_chemphys_wrangle

summary(NTL.regression.best)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL_chemphys_wrangled)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6857 -3.0267  0.1055  2.9937 13.6038
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -18.196998   8.741236   -2.082 0.037392 *
## depth        -1.941328   0.011728 -165.528  < 2e-16 ***
## year4         0.016113   0.004353    3.701 0.000216 ***
## daynum        0.040237   0.004385    9.176  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.833 on 9669 degrees of freedom
## Multiple R-squared:  0.7398, Adjusted R-squared:  0.7397
## F-statistic:  9162 on 3 and 9669 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 methods suggests that all variables (depth, daynum, and year) should be included in the model. This new model explains 73.97% of the data variance. This is a very slight improvement over the last model.

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
#run ANOVA
NTL.anova.lakename <- aov(data = NTL_chemphys_wrangled, temperature_C ~ lakename)
summary(NTL.anova.lakename)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## lakename        8  22188  2773.5   51.18 <2e-16 ***
## Residuals    9664 523706    54.2
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
#run regression
NTL.regression.lakename <- lm(data = NTL_chemphys_wrangled, temperature_C ~ lakename)
summary(NTL.regression.lakename)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_chemphys_wrangled)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.773  -6.612  -2.673   7.657  23.813
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               17.6664     0.6507  27.151  < 2e-16 ***
## lakenameCrampton Lake     -2.1851     0.7565  -2.889 0.003879 **
## lakenameEast Long Lake    -7.3795     0.6915 -10.671  < 2e-16 ***
## lakenameHummingbird Lake  -6.6828     0.9571  -6.982 3.09e-12 ***
## lakenamePaul Lake         -3.8234     0.6666  -5.735 1.00e-08 ***
## lakenamePeter Lake        -4.3162     0.6652  -6.489 9.08e-11 ***
## lakenameTuesday Lake      -6.5937     0.6777  -9.730  < 2e-16 ***
## lakenameWard Lake         -3.2078     0.9437  -3.399 0.000679 ***
## lakenameWest Long Lake    -6.0542     0.6893  -8.783  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.361 on 9664 degrees of freedom
## Multiple R-squared:  0.04064,    Adjusted R-squared:  0.03985
## F-statistic: 51.18 on 8 and 9664 DF,  p-value: < 2.2e-16
```

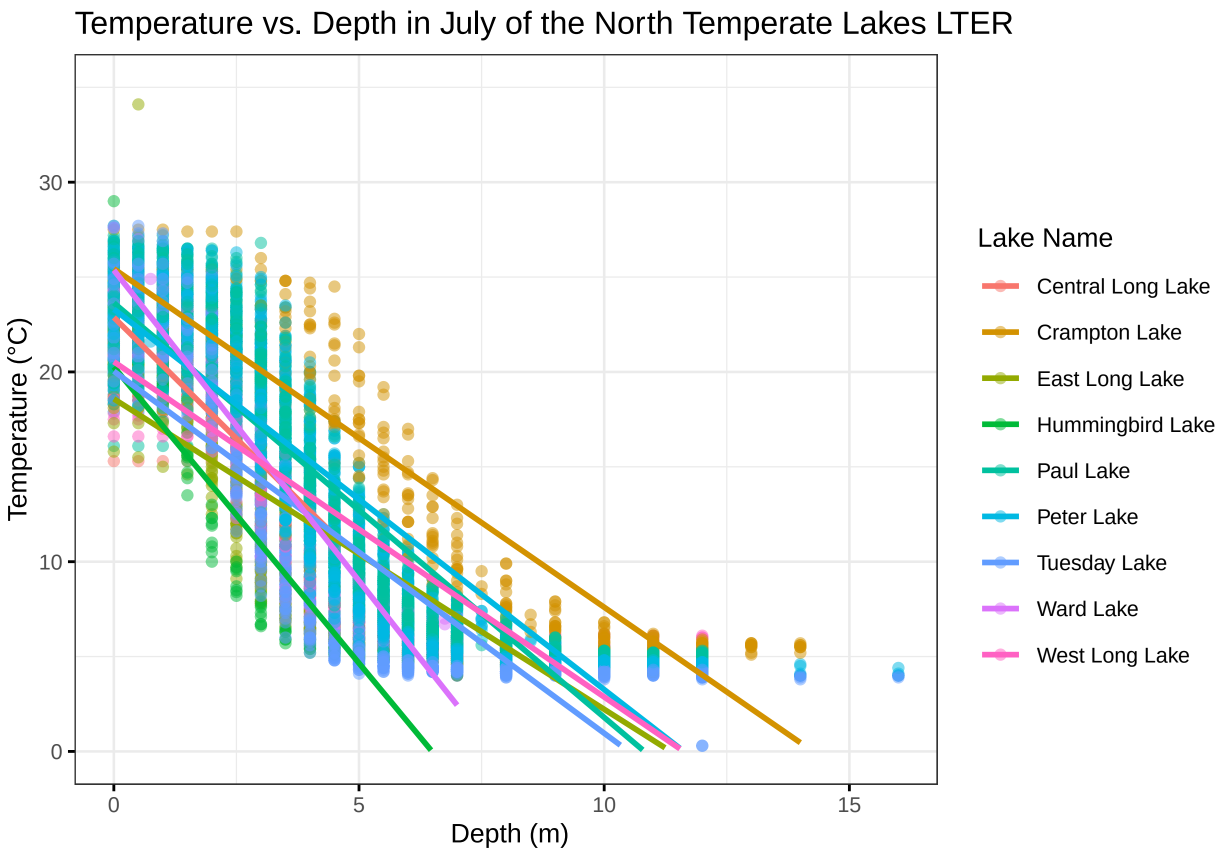
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: There is a significant difference among the lakes. The ANOVA indicates that lake is a significant factor for temperature variation. However, it does not indicate which means are different. The linear model shows that each lake name has a significant coefficient compared to the reference level, but this also does not indicate which lakes have the same means.

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.
NTL_vis2 <- ggplot(data = NTL_chemphys_wrangled, aes(x=depth, y=temperature_C, colour = lakename))+
  geom_point(alpha = 0.5)+
  geom_smooth(method = lm, se = FALSE)+
  ylim(0, 35)+
  labs(x = " Depth (m)",
       y = "Temperature (°C)",
       title = "Temperature vs. Depth in July of the North Temperate Lakes LTER",
       colour = "Lake Name")+
  mytheme

NTL_vis2
```



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

```
#15
#conduct test
NTL.tukey.lakename <- TukeyHSD(NTL.anova.lakename)
NTL.HSD <- HSD.test(NTL.anova.lakename, "lakename", group = TRUE)

NTL.HSD$groups
```

```
##                     temperature_C groups
## Central Long Lake        17.66641      a
## Crampton Lake            15.48132     ab
## Ward Lake                14.45862     bc
## Paul Lake                13.84304      c
## Peter Lake               13.35016      c
## West Long Lake           11.61224      d
## Tuesday Lake             11.07267     de
## Hummingbird Lake         10.98364     de
## East Long Lake           10.28694      e
```

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: Peter Lake has the same mean temperature as Ward Lake and Paul Lake. All of the lakes share means 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: You could perform a t-test on Peter Lake and Paul lake. This test is designed to test differences in means between two categories.

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
#wrangle data
NTL_chemphys_wrangled_clwl <- NTL_chemphys_wrangled %>%
  filter(lakename == "Crampton Lake" | lakename == "Ward Lake" )

#conduct ttest
NTL.ttest <- t.test(data = NTL_chemphys_wrangled_clwl, temperature_C ~ lakename)
NTL.ttest
```

```
##
##  Welch Two Sample t-test
##
## data:  temperature_C by lakename
## t = 1.2972, df = 192.4, p-value = 0.1961
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is
## 95 percent confidence interval:
##  -0.5323014  2.5776973
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.48132                    14.45862
```

> Answer: The p-value of the t-test is not significant (>0.05). Therefore, the null hypothesis cannot be rejected. This means that the mean temperature of both lakes is not statistically different from each other. This is the same as in the Tukey test.