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

Jazmine Pritchett

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

library(tidyverse)
install.packages("agricolae")
library(agricolae)
library(dplyr)
library(ggplot2)
library(here)
here()
```

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

```
Chemphys <- read.csv(file = here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
stringsAsFactors = TRUE)


# Set date to date format
Chemphys$sampledate <- ymd(Chemphys$sampledate)
```

```r
#2

mytheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "darkgrey"),
        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: There is no change in mean lake temperature with depth recorded during July across all lakes. Ha: Mean lake temperature recorded during July changes with depth across 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)

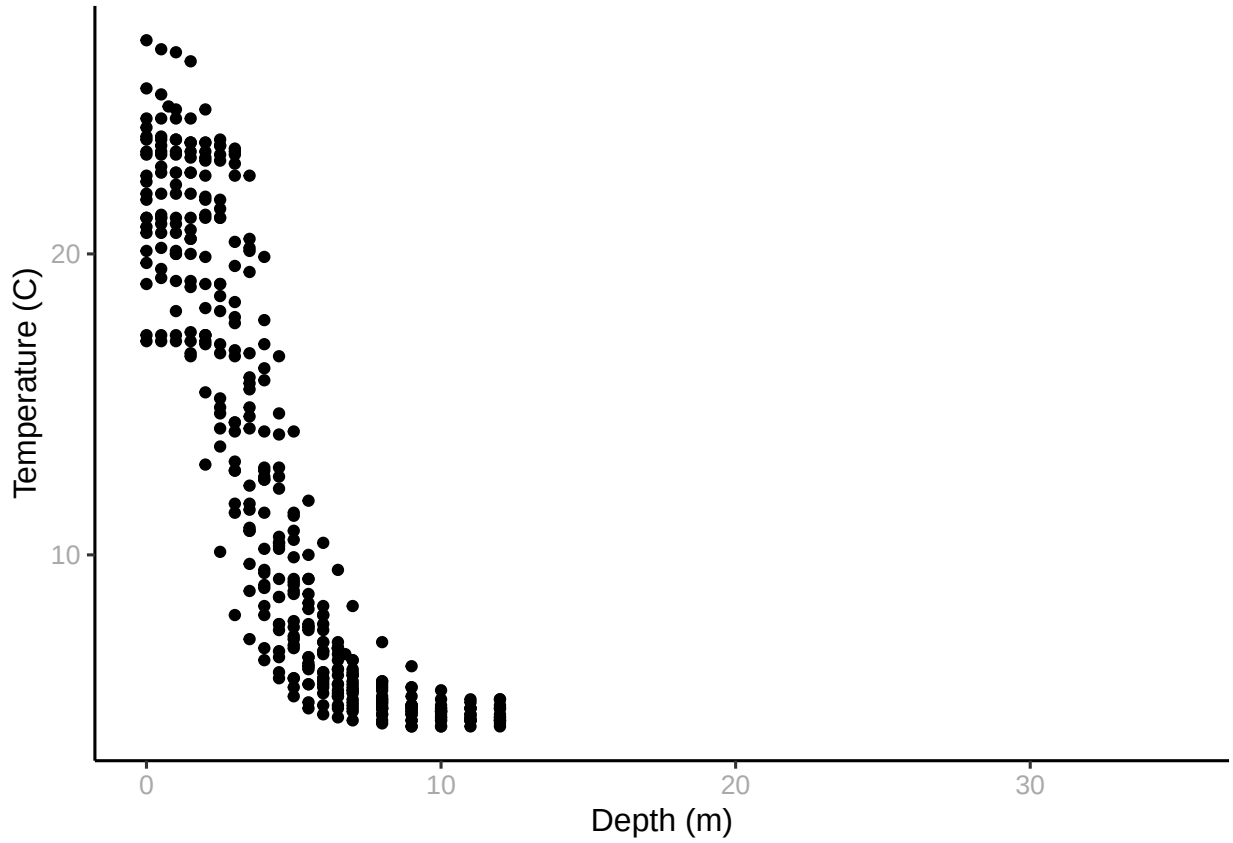
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```r
#4

Chemphys.Totals <- Chemphys %>% filter(month(sampledate) == 7) %>% select(lakename, year4, daynum,
depth, temperature_C) %>% na.omit()


#5

ggplot(Chemphys.Totals, aes(x= depth, y= temperature_C)) + geom_point() + geom_smooth(method = "1m",
se = FALSE) + xlim(0, 35) + labs(x= "Depth (m)", y = "Temperature (C)")
```



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 that as depth increases, meaning the lake becomes deeper, ttemperature overtime decreases. It seems that there is a linear relationship between the two variables as an inverse relationship.

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

```
#R-squared and correlation

Chemphys.regression <- lm(
  data = Chemphys.Totals,
  temperature_C ~ depth)
summary(Chemphys.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = Chemphys.Totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7641 -2.8586 -0.3779  2.6155  7.7928
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.5054     0.3261   65.95   <2e-16 ***
## depth        -1.9138     0.0567  -33.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.65 on 392 degrees of freedom
## Multiple R-squared:  0.744,  Adjusted R-squared:  0.7434
## F-statistic:  1139 on 1 and 392 DF,  p-value: < 2.2e-16
```

```
cor.test(Chemphys.Totals$temperature_C, Chemphys.Totals$depth)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Chemphys.Totals$temperature_C and Chemphys.Totals$depth
## t = -33.755, df = 392, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8858738 -0.8349256
## sample estimates:
##       cor
## -0.8625706
```

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 statistically signifigant relationship between variable of depth with temperature. The R-squared value is 0.7387 (73.87% variability) with 9726 degrees of freedom. The statistical signifigance of the result is p-value < 2.2e-16 (less than signifigance of 0.05). Looking at the slope of the regression line, -1.94621, which means that for every 1m increase in depth, the temperature is predicted to decrease by 1.94621 degrees Celsius.

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
TPAIC <- lm(data = Chemphys.Totals, temperature_C ~ depth + year4 + daynum)

#Choose a model by AIC in a Stepwise Algorithm
step(TPAIC)
```

```
## Start:  AIC=966.18
## temperature_C ~ depth + year4 + daynum
##
##          Df Sum of Sq     RSS     AIC
## - year4   1      21.7  4505.8  966.08
## <none>                 4484.0  966.18
## - daynum  1     638.3  5122.3 1016.62
## - depth   1   15263.4 19747.5 1548.29
##
## Step:  AIC=966.08
## temperature_C ~ depth + daynum
##
##          Df Sum of Sq     RSS     AIC
## <none>                 4505.8  966.08
## - daynum  1       717  5222.8 1022.27
## - depth   1     15242 19747.5 1546.29

##
## Call:
## lm(formula = temperature_C ~ depth + daynum, data = Chemphys.Totals)
##
## Coefficients:
## (Intercept)        depth       daynum
##    11.19860     -1.91767      0.05466
```

```
TPmodel <-lm(data = Chemphys.Totals, temperature_C ~ depth + year4 + daynum)
summary(TPmodel)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = Chemphys.Totals)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3086 -2.7609 -0.3194  2.5294  8.1964
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 123.85217   81.96700   1.511    0.132
## depth        -1.92045    0.05271 -36.435  < 2e-16 ***
## year4        -0.05591    0.04067  -1.375    0.170
## daynum        0.05268    0.00707   7.451 6.02e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.391 on 390 degrees of freedom
## Multiple R-squared:  0.7802, Adjusted R-squared:  0.7785
## F-statistic: 461.5 on 3 and 390 DF,  p-value: < 2.2e-16
```

```
#10

#lowest AIC value best fit, year4= 26066, daynum= 26148, depth = 39189
Chemphysregression <- lm(data = Chemphys.Totals, temperature_C ~ + year4 + daynum)
summary(Chemphysregression)
```

```
##
## Call:
## lm(formula = temperature_C ~ +year4 + daynum, data = Chemphys.Totals)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.694 -6.035 -2.522  7.606 13.399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.845e-01  1.716e+02   0.004 0.996820
## year4       9.197e-04  8.518e-02   0.011 0.991391
## daynum      5.232e-02  1.482e-02   3.531 0.000463 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.107 on 391 degrees of freedom
## Multiple R-squared:  0.03216,    Adjusted R-squared:  0.02721
## F-statistic: 6.497 on 2 and 391 DF,  p-value: 0.001677
```

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 smaller the AIC value, the better. The final values suggested is that of 'year4' and 'daynum' variables. To determine how much observed variance the model explains, we look at the R-squared value. The muktiple R-squared is 0.002363. Comparing to linear regression of just depth, the R-squared is 0.7387. This is not an improvement as for depth it is closer to 1, meaning a stronger relationship.

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
Chemphys.Totals.anova <- aov(data = Chemphys.Totals, temperature_C ~ lakename)
summary(Chemphys.Totals.anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## lakename      4    228   56.96   1.098  0.357
## Residuals   389  20176   51.87
```

```
Chemphys.Totals.anova2 <- lm(data = Chemphys.Totals, temperature_C ~ lakename)
summary(Chemphys.Totals.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = Chemphys.Totals)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.331 -6.756 -2.550  7.338 14.536
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.5556     1.6975   6.218  1.3e-09 ***
## lakenamePaul Lake       1.9008     1.7856   1.065    0.288
## lakenamePeter Lake      2.0088     1.7904   1.122    0.263
## lakenameTuesday Lake   -0.4389     2.4006  -0.183    0.855
## lakenameWard Lake       3.3755     2.1610   1.562    0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.202 on 389 degrees of freedom
## Multiple R-squared:  0.01117,    Adjusted R-squared:  0.0009981
## F-statistic: 1.098 on 4 and 389 DF,  p-value: 0.3571
```

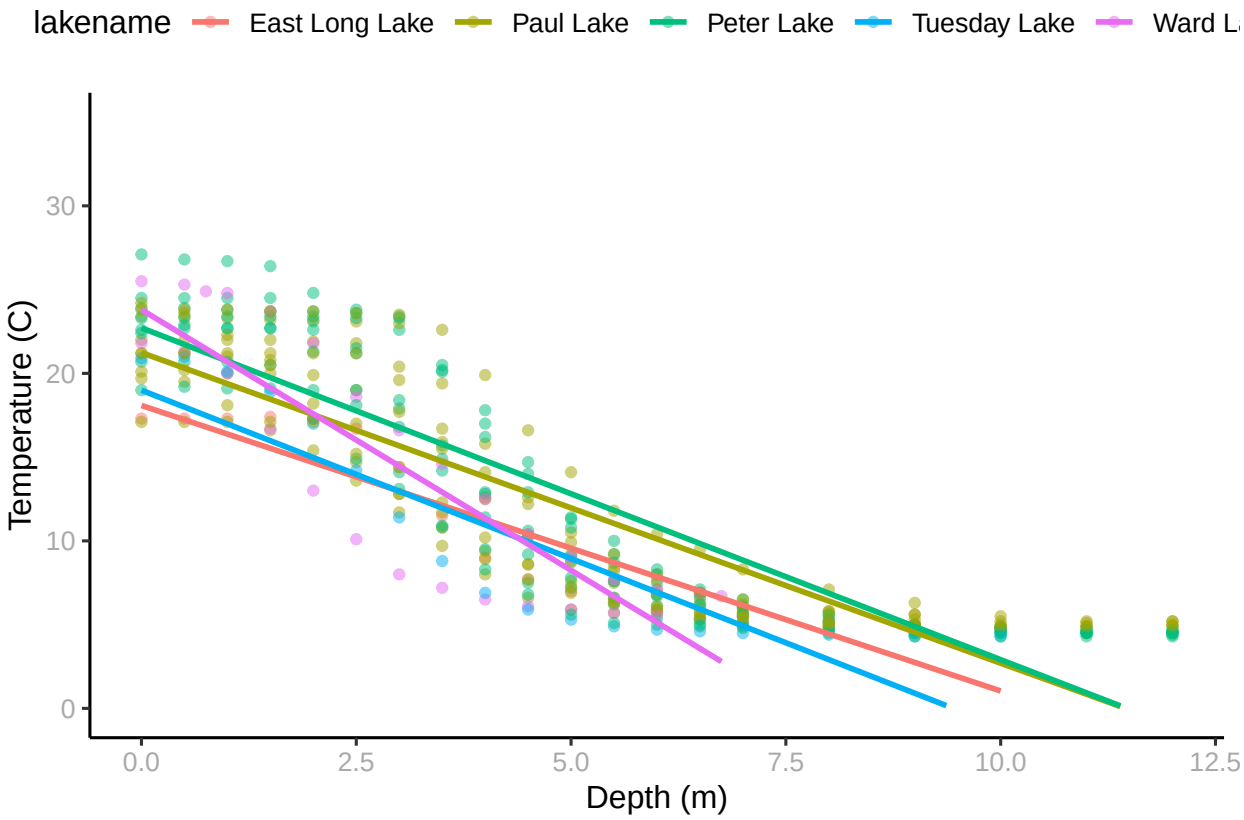
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: The models both show that there is a signifignt difference between mean temperature and lakes as p- value is < 0.05 being $<$ 2.2e-16.

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.

ggplot(Chemphys.Totals, aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha = 0.5) +  ylim(0, 35) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Depth (m)", y = "Temperature (C)", color = "lakename") +
  scale_color_discrete(name = "lakename") + mytheme
```



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

```
#15
TukeyHSD(Chemphys.Totals.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = Chemphys.Totals)
##
## $lakename
##                                    diff       lwr      upr     p adj
## Paul Lake-East Long Lake      1.9007758 -2.992848 6.794400 0.8245987
## Peter Lake-East Long Lake     2.0088194 -2.898035 6.915674 0.7948864
## Tuesday Lake-East Long Lake -0.4388889 -7.018014 6.140236 0.9997496
## Ward Lake-East Long Lake      3.3754789 -2.546994 9.297952 0.5227817
## Peter Lake-Paul Lake          0.1080436 -2.069085 2.285172 0.9999228
## Tuesday Lake-Paul Lake       -2.3396647 -7.233289 2.553959 0.6849800
## Ward Lake-Paul Lake           1.4747031 -2.492456 5.441862 0.8466692
## Tuesday Lake-Peter Lake      -2.4477083 -7.354562 2.459146 0.6491273
## Ward Lake-Peter Lake          1.3666595 -2.616808 5.350127 0.8810112
## Ward Lake-Tuesday Lake        3.8143678 -2.108105 9.736840 0.3955788
```

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: Central long lake, Crampton lake, East Long lake, Hummingbird, Tuesday, West Long statistically are similar to Peter Lake. Ward lake is distinct.

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 t-test would be efficient at comparing mean temperatures of Peter and Paul lakes since it compares the means of two groups for statistical signifigance.

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?

```
CW_July <- subset(Chemphys.Totals, lakename == "Crampton Lake" | lakename == "Ward Lake")
#Execution becomes halted when knitting because of this code I would use to perform the t-test:
#t.test(CW_July$temperature_C ~ CW_July$lakename)
#generates

#Welch Two Sample t-test

#data:  CW_July$temperature_C by CW_July$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
#not equal to 0
#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 test gives a p-value of 0.2649 and a 95% confidence interval of -0.6821129 to 2.4686451. As p-value is greater than 0.05, null hypothesis is rejected. There is not enough evidende to show that mean temperatures are equal. THe answer I have for question 16 also found that there was not enough evidence to suggest a difference.