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

Karly Nocera

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on generalized linear models.

Directions

1. Change "Student Name" on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., "Fay_A06_GLMs.Rmd") prior to submission.

The completed exercise is due on Tuesday, March 2 at 1:00 pm.

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
getwd()
```

```
## [1] "Z:/Environmental_Data_Analytics_2021"
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(htmltools)
library(agricolae)
```

```
## Warning: package 'agricolae' was built under R version 4.0.4
```

```
library(lubridate)
```

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

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

```
library(viridis)
```

```
## Loading required package: viridisLite
```

```
ChemPhys <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

class(ChemPhys$sampledate)
```

```
## [1] "character"
```

```
ChemPhys$sampledate <- ymd(ChemPhys$sampledate)
```

```
## Warning: 33138 failed to parse.
```

```
#2

mytheme <- theme_classic(base_size = 10, base_family = "sans") +
  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: The change in mean lake temperatures recorded during July across all lakes is zero. Ha: The change in mean lake temperatures recorded during July across all lakes is different than zero.

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

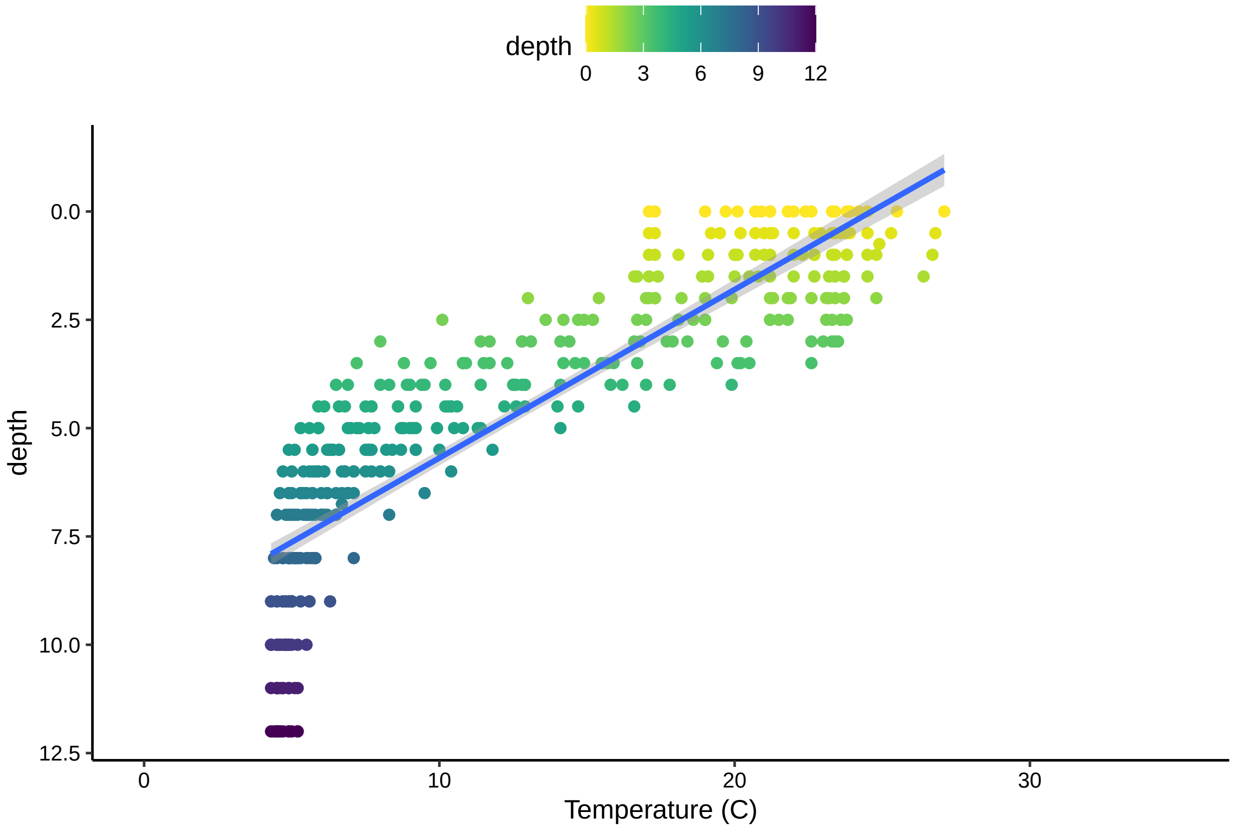
- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

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.

```
#4

ChemPhys.tidy <-
  ChemPhys %>%
  mutate(month = month(sampledate)) %>%
  filter(month == 7) %>%
  select(lakename, year4, daynum, depth, temperature_C, month) %>%
  drop_na()

### note: when checking data, see that daynum suggests a different day of the year than the sampledate

#5
```

```
tempvdepth <-
  ggplot(ChemPhys.tidy, aes(x = temperature_C, y = depth, color = depth)) +
  geom_point() +
  scale_y_reverse() +
  xlim(0, 35) +
  xlab(expression("Temperature (C)")) +
  geom_smooth(method = lm) +
  scale_color_viridis(direction = -1)
print(tempvdepth)
```

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



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: This figure suggests that shallower depths are associated with warmer temperatures and depper depths with colder temperatures. The distribution of points suggest this may be an exponential rather than linear function, as there are diminishing returns as temperature increases and there is a threshold at which deeper depths are not associated with colder temperatures.

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

```
#7

ChemPhys.regression <- lm(data = ChemPhys.tidy, temperature_C ~ depth)
summary(ChemPhys.regression)
```

```
##
```

```
## Call:
## lm(formula = temperature_C ~ depth, data = ChemPhys.tidy)
##
## 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
```

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: This regression indicates a statistically significant negative relationship between temperature and depth, as seen by p-value below 0.05 for the coefficient with degrees of freedom (392). That is to say, as depth increases every 1m, temperature is predicted to drop -1.9 degrees. The variability in temperature explained by the change in depth is given by the R2 value, at 74.4%.

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

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

step(TPAIC) # identifies daynum and depth as optimal variables
```

```
## Start:  AIC=966.18
## temperature_C ~ year4 + daynum + depth
##
##          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 ~ daynum + depth
##
```

```
##          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 ~ daynum + depth, data = ChemPhys.tidy)
##
## Coefficients:
## (Intercept)       daynum        depth
##    11.19860      0.05466     -1.91767
```

```
#10

TPmodel <- lm(data = ChemPhys.tidy, temperature_C ~ daynum + depth)
summary(TPmodel)
```

```
##
## Call:
## lm(formula = temperature_C ~ daynum + depth, data = ChemPhys.tidy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.6397 -2.6632 -0.3101  2.5174  8.0771
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.198599   1.341363   8.349 1.21e-15 ***
## daynum       0.054660   0.006929   7.888 3.10e-14 ***
## depth       -1.917670   0.052729 -36.368  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.395 on 391 degrees of freedom
## Multiple R-squared:  0.7792, Adjusted R-squared:  0.778
## F-statistic: 689.8 on 2 and 391 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 are daynum and depth to predict temperature in the multiple regression. This model explains 77.9% of the observed variance, which is an improvement from 74.4% in the single regression 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
```

```
## ANOVA model (aov)

ChemPhys.anova <- aov(data=ChemPhys.tidy, temperature_C ~ lakename)
summary(ChemPhys.anova)
```

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

```
## linear model (lm)

chemPhys.lm.aov <- lm(data = ChemPhys.tidy, temperature_C ~ lakename)
summary(chemPhys.lm.aov)
```

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

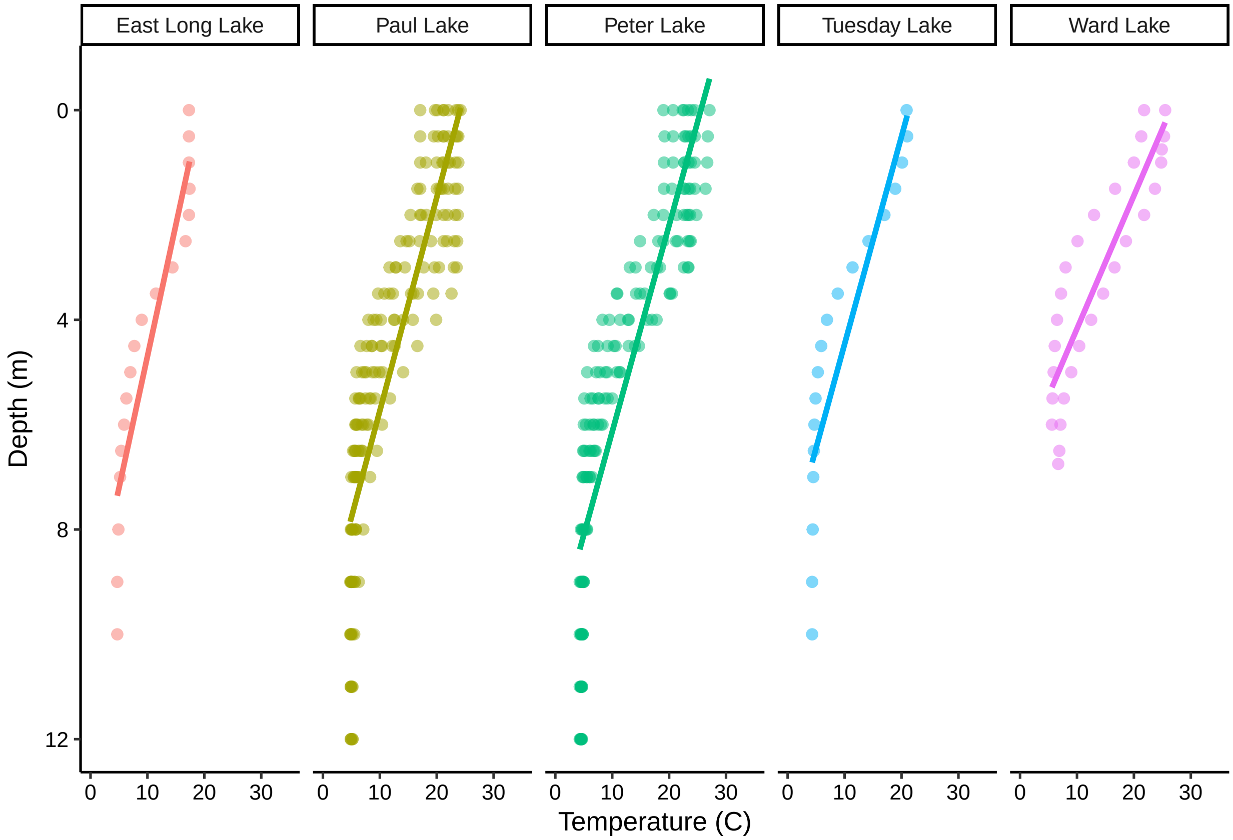
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: Cannot reject the null hypothesis that there is a statistically significant difference in mean temperature among the lakes because the p-value for the ANOVA is above 0.05 (at 0.357).

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.

tempvdepth.lake <-
  ggplot(ChemPhys.tidy, aes(x = temperature_C, y = depth, color = lakename)) +
  geom_point(alpha = 0.5) + #transparency
  geom_smooth(method = lm, se = FALSE) +
  xlim(0, 35) + #x is degree, not y
  scale_y_reverse() +
  facet_wrap(vars(lakename), nrow = 1) +
  theme(legend.position="none") +
  xlab(expression("Temperature (C)")) +
  ylab(expression("Depth (m)"))
print(tempvdepth.lake)
```

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



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

```
#15

TukeyHSD(ChemPhys.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = ChemPhys.tidy)
##
## $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: Two pairs of lakes show similar mean temperatures: 1) Tuesday Lake and East Long Lake; 2) Peter Lake and Paul Lake. However, neither have a p-value less than 0.05 so these results are not statistically significant. In fact, none of the lakes have a mean temperature that is statistically distinct from all the other lakes, because all p-values are above 0.05.

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: The HSD Test can help see if there are any pair-wise relationships and further explore if Peter Lake and Paul lake have distinct mean temperatures.

> author's note: This assignment was complete before deadline, as seen by the github knit date, but due to continued issues with github communicating with my shared drive, was unable to upload PDF until lab.