Assignment 6: GLMs (Linear Regressios, ANOVA, & t-tests)

Garrett Pearce

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 Monday, February 28 at 7:00 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] "C:/Users/dgp20/Documents/ENV 872/Environmental_Data_Analytics_2022/Assignments"
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v stringr 1.4.0
## v tidyr   1.1.4     v forcats 0.5.1
## v readr   2.1.1
```

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

```
library(htmltools)
library(agricolae)
library(cowplot)
#Loading some extra packages, just in case. Note- I had to install the agricolae package.

chemphys <- read.csv('../Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv', stringsAsFactors = TRUE)

chemphys$sampledate <- as.Date(chemphys$sampledate, format = "%m/%d/%Y")

#2
newTheme <- theme(legend.position = "right", axis.text = element_text(color = "green", size = 10),
                  legend.title = element_text(size = 15, color = "orange"))
```

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 NOT a statistically significant difference with depth in lake temperature recurded during July in all lakes. Ha: There is a statistically significant difference with depth in lake temperature recurded during July in all lakes.
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:

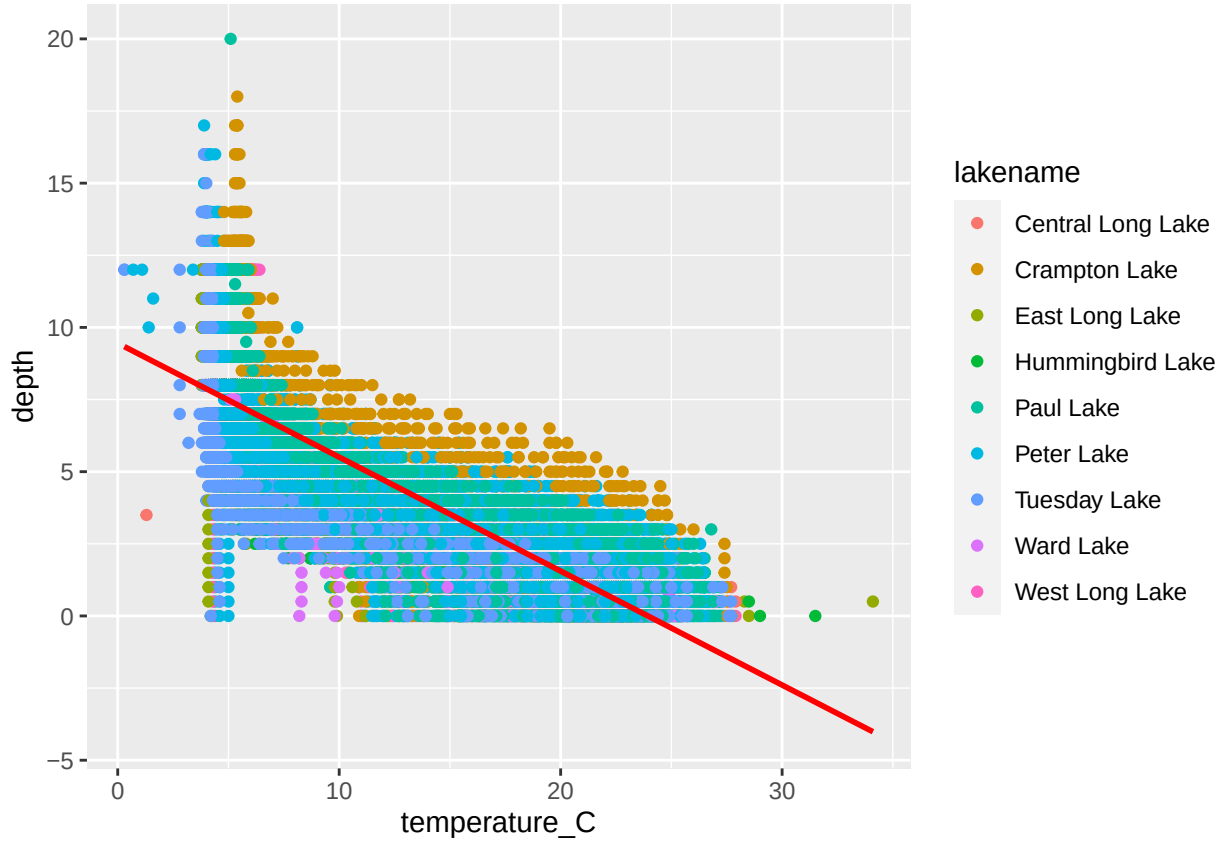
- Only dates in July.
- Only the columns: `lakename`, `year4`, `daynum`, `depth`, `temperature_C`
- Only complete cases (i.e., remove NAs)

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.

```
#NOTE- I included results='hide' for this chunk because my knitted PDF was displaying
#every value in my filtered dataset for #4,and I wasn;t sure how to turn it off.
#I don't believe it affected any other aspects here.

#4
chemphys.processed <-
  chemphys%>%
  filter(daynum >= 181|daynum <= 212)
  select(chemphys, lakename, year4, daynum, depth, temperature_C) %>%
  filter(lakename != "", year4 != "", daynum != "", depth != "" |  temperature_C != "")


#5

chemphys.processed.scatter <-
  chemphys.processed%>%
  filter(temperature_C <= 35)

#Setting a new piped dataset to further filter the data for the scatterplot-
#there may be more elegant ways to do this
```

```
tempdepth.plot <- ggplot(chemphys.processed.scatter, aes(x = temperature_C, y = depth, color = lakename)
    geom_point() +
    geom_smooth(method= 'lm', color = 'red')

#Added color by lakename to make graph a bit more aesthetically pleasing and readable.

print(tempdepth.plot)
```

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




```
#Hopefully as pretty as the graph can be with so many values.
```

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 suggests that temperature tends to decrease with depth across all lakes. Distribution suggests a much higher temperature range at low-depth areas.The distribution does not seem to be completely linear, but rather more proportional, with an almost exponential-looking decrease in temp at deeper depths.Regardless, this does suggest some relationship between temperature and depth, at a glance, which supports Ha.

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

```
#7
linregtest <- lm(data = chemphys.processed.scatter, temperature_C ~ depth)
#Setting up a variable to run linear regression

summary(linregtest)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = chemphys.processed.scatter)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7864  -3.1363  -0.1219   3.1815  19.2568
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.986395   0.037166   537.8   <2e-16 ***
## depth       -1.707162   0.006366  -268.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.961 on 34754 degrees of freedom
## Multiple R-squared:  0.6742, Adjusted R-squared:  0.6742
## F-statistic: 7.192e+04 on 1 and 34754 DF,  p-value: < 2.2e-16
```

```
#Printing linear regression variables.
```

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 numbers here suggest that, with an average temp of around 19 degrees C, a 1-unit change in depth produces a roughly 1.71-unit decrease in temperature, with 34754 degrees of freedom. Our R-squared value of 0.6742 tells us that 67.42% of temp variation is explained by depth. In terms of statistical significance, given the small p-value and large F-statistic, it is safe to reject the null hypothesis here. We can conclude a statistically significant link between temperature and depth based on this data.

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
#setting this ahead of #10.
multregtest <- lm(data = chemphys.processed, temperature_C ~ year4 + daynum + depth)


step(multregtest)
```

```
## Start:  AIC=92515.66
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq     RSS    AIC
## <none>                  497693  92516
## - year4   1       167  497861  92525
## - daynum  1     47378  545071  95674
```

```
## - depth   1   1130140 1627834 133700
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = chemphys.processed)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##   -2.268081     0.007734     0.034990    -1.708651
```

```
#Since removing depth seems to drive up the AIC by the largest amount, it would seem that depth is best

#10
multregtest <- lm(data = chemphys.processed, temperature_C ~ year4 + daynum + depth)

summary(multregtest)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = chemphys.processed)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -19.7228  -2.8606  -0.1706   2.9267  17.8338
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -2.2680808  4.5365880   -0.500 0.617111
## year4        0.0077342  0.0022622    3.419 0.000629 ***
## daynum       0.0349904  0.0006083   57.517  < 2e-16 ***
## depth       -1.7086514  0.0060824 -280.915  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.784 on 34752 degrees of freedom
##   (3858 observations deleted due to missingness)
## Multiple R-squared:  0.7026, Adjusted R-squared:  0.7026
## F-statistic: 2.737e+04 on 3 and 34752 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: While depth is the best predictor, the AIC model seems to suggest that year, day, and depth are all useful in predicting temp, as removal of any of these three drives the AIC up. Removing nothing gets us the best AIC (92515.66), though just barely. Based on this model's r squared value, this model explains about 70.26% of the observed variance. This is better than our depth-only model, which explains about 67.42% of observed variance.

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).

```r
#12
#First, wrangling data for ANOVA model:
laketemps <- chemphys.processed %>%
  group_by(lakename, sampledate) %>%
  summarise(temperature_C = sum(temperature_C))
```

```
## `summarise()` has grouped output by 'lakename'. You can override using the `.groups` argument.
```

```r
summary(laketemps)
```

```
##            lakename     sampledate         temperature_C
##  Peter Lake    :530   Min.   :0000-05-24   Min.   : 71.5
##  Paul Lake     :524   1st Qu.:0011-07-13   1st Qu.:118.3
##  Tuesday Lake  :311   Median :0087-08-04   Median :154.8
##  West Long Lake:193   Mean   :0059-11-24   Mean   :167.2
##  East Long Lake:180   3rd Qu.:0093-09-05   3rd Qu.:216.0
##  Crampton Lake : 51   Max.   :0099-08-31   Max.   :313.1
##  (Other)       :113                        NA's   :1828
```

```r
laketemps.anova <- aov(data = laketemps, temperature_C ~ lakename)

summary(laketemps.anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## lakename     6  55205    9201   3.075 0.0102 *
## Residuals   67 200504    2993
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 1828 observations deleted due to missingness
```

```r
#Linear regression model:

laketemps.lm <- lm(data = laketemps, temperature_C ~ lakename)

summary(laketemps.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = laketemps)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -104.883  -39.121    4.256   41.939  101.950
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  211.15      38.68   5.459 7.54e-07 ***
## lakenameEast Long Lake       -96.92      49.94  -1.941   0.0565 .
## lakenameHummingbird Lake     -62.82      49.94  -1.258   0.2128
## lakenamePaul Lake            -50.15      41.03  -1.222   0.2259
## lakenamePeter Lake           -16.27      39.95  -0.407   0.6852
## lakenameTuesday Lake         -73.06      40.77  -1.792   0.0777 .
## lakenameWest Long Lake       -82.70      54.70  -1.512   0.1353
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 54.7 on 67 degrees of freedom
##   (1828 observations deleted due to missingness)
## Multiple R-squared:  0.2159, Adjusted R-squared:  0.1457
## F-statistic: 3.075 on 6 and 67 DF,  p-value: 0.01018
```

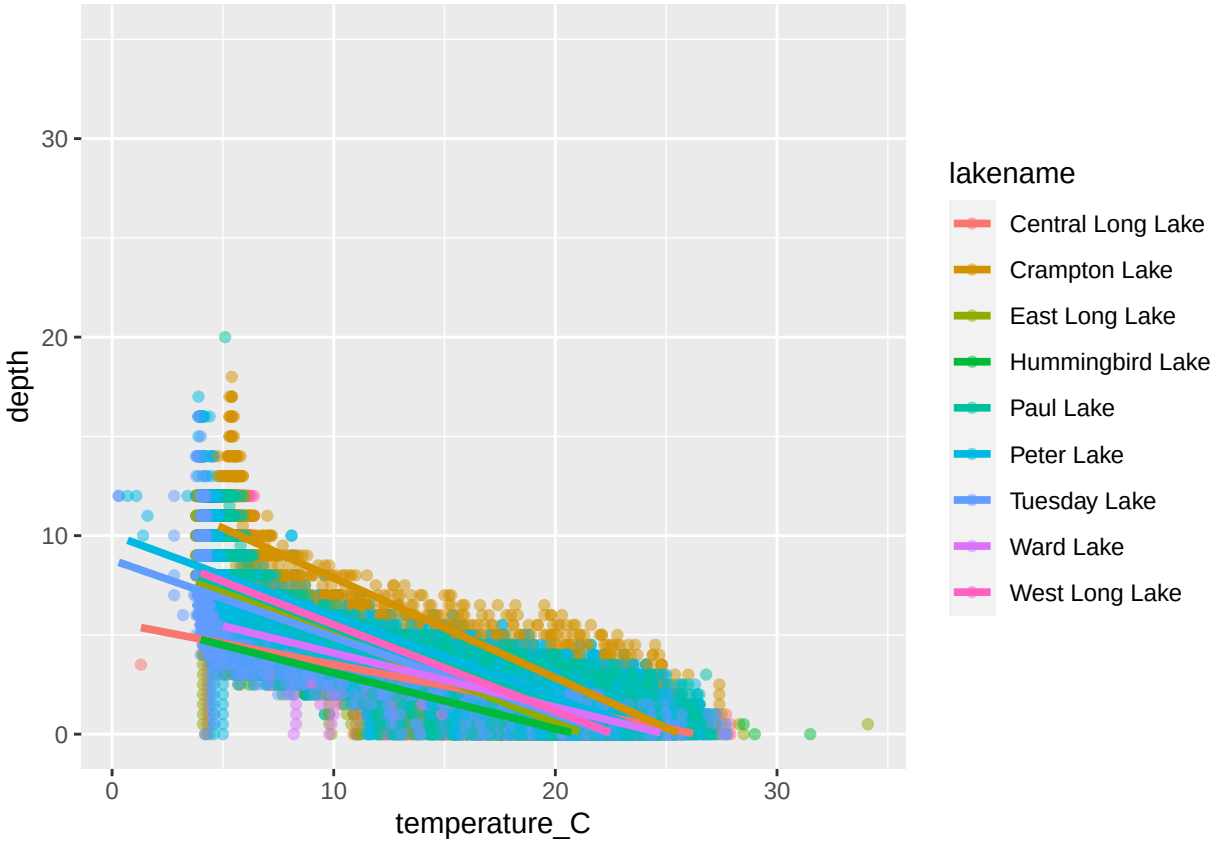
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: The ANOVA model has a small p-value (<0.05), suggesting that there is significant difference in mean temperature between lakes. The linear regression model shows a large F-statistic and a small p-statistic, suggesting that the null hypothesis can be rejected, and that there is in fact a significant variation in mean temperature across lakes. Both tests have very similar p-values, both implying a rejection of the null hypothesis. We can therefore say that there is a significant difference.

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.
#reusing some code from #6- looks like I was thinking ahead in using separate colors for each lake!

laketemps.plot <- ggplot(chemphys.processed, aes(x = temperature_C, y = depth, color = lakename), inher
    geom_point(alpha = 0.5) +
    geom_smooth(method= 'lm', se = FALSE, size = 1.35) +
    ylim(0, 35)
#Changed line sizes, because even with 50% opacity, the sheer number of points was making it difficult

print(laketemps.plot)
```

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

## Warning: Removed 3858 rows containing non-finite values (stat_smooth).

## Warning: Removed 3858 rows containing missing values (geom_point).

## Warning: Removed 141 rows containing missing values (geom_smooth).
```



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

```
#15
TukeyHSD(laketemps.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = laketemps)
##
## $lakename
##                                       diff        lwr        upr     p adj
## East Long Lake-Crampton Lake     -96.916667 -248.71410  54.880767 0.4614693
## Hummingbird Lake-Crampton Lake   -62.816667 -214.61410  88.980767 0.8682843
## Paul Lake-Crampton Lake          -50.150000 -174.86432  74.564318 0.8830144
## Peter Lake-Crampton Lake         -16.266667 -137.70461 105.171280 0.9996218
## Tuesday Lake-Crampton Lake       -73.055556 -196.99764  50.886530 0.5582194
## West Long Lake-Crampton Lake     -82.700000 -248.98576  83.585757 0.7368665
## Hummingbird Lake-East Long Lake   34.100000 -101.67175 169.871752 0.9876065
## Paul Lake-East Long Lake          46.766667  -57.85249 151.385828 0.8212222
## Peter Lake-East Long Lake         80.650000  -20.04103 181.341026 0.2007073
## Tuesday Lake-East Long Lake       23.861111  -79.83628 127.558500 0.9921992
## West Long Lake-East Long Lake     14.216667 -137.58077 166.014101 0.9999532
## Paul Lake-Hummingbird Lake        12.666667  -91.95249 117.285828 0.9997894
## Peter Lake-Hummingbird Lake       46.550000  -54.14103 147.241026 0.7971086
## Tuesday Lake-Hummingbird Lake    -10.238889 -113.93628  93.458500 0.9999361
## West Long Lake-Hummingbird Lake  -19.883333 -171.68077 131.914101 0.9996677
```

```
## Peter Lake-Paul Lake             33.883333  -17.59368  85.360348 0.4238795
## Tuesday Lake-Paul Lake          -22.905556  -80.04003  34.228923 0.8844843
## West Long Lake-Paul Lake        -32.550000 -157.26432  92.164318 0.9848660
## Tuesday Lake-Peter Lake         -56.788889 -106.36572  -7.212055 0.0146837
## West Long Lake-Peter Lake       -66.433333 -187.87128  55.004614 0.6425652
## West Long Lake-Tuesday Lake      -9.644444 -133.58653 114.297641 0.9999843
```

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: Crampton Actually, it seems that none of the lakes have the same mean temp as Peter Lake- the closest seems to be Crampton Lake with a -16.266667 difference- still not quite the same. Most all of the lakes seem to have mean temps that are statistically distinct from each other, based on the diff estimates and their associated p values.

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 could also use a t-test to explore the difference between these two means.