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

Jess Garcia

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. Set up session
getwd()
```

```
## [1] "C:/Users/93jes/Documents/ENV872/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()
```

```
#install.packages("agricolae")
library(agricolae)

Lake_ChemPhys_Raw <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv", stringsAsFactors = TR

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

```
#2.Build theme & set as default
JG_default_theme <- theme_bw(base_size = 12)+
  theme(axis.text = element_text(color = "black"),
        plot.title = element_text(color = "black", size = 16, face = "bold",
        hjust = 0.5))
```

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 mean lake temperature recorded during July does not change with depth across all lakes Ha: The mean lake temperature recorded during July does change 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)

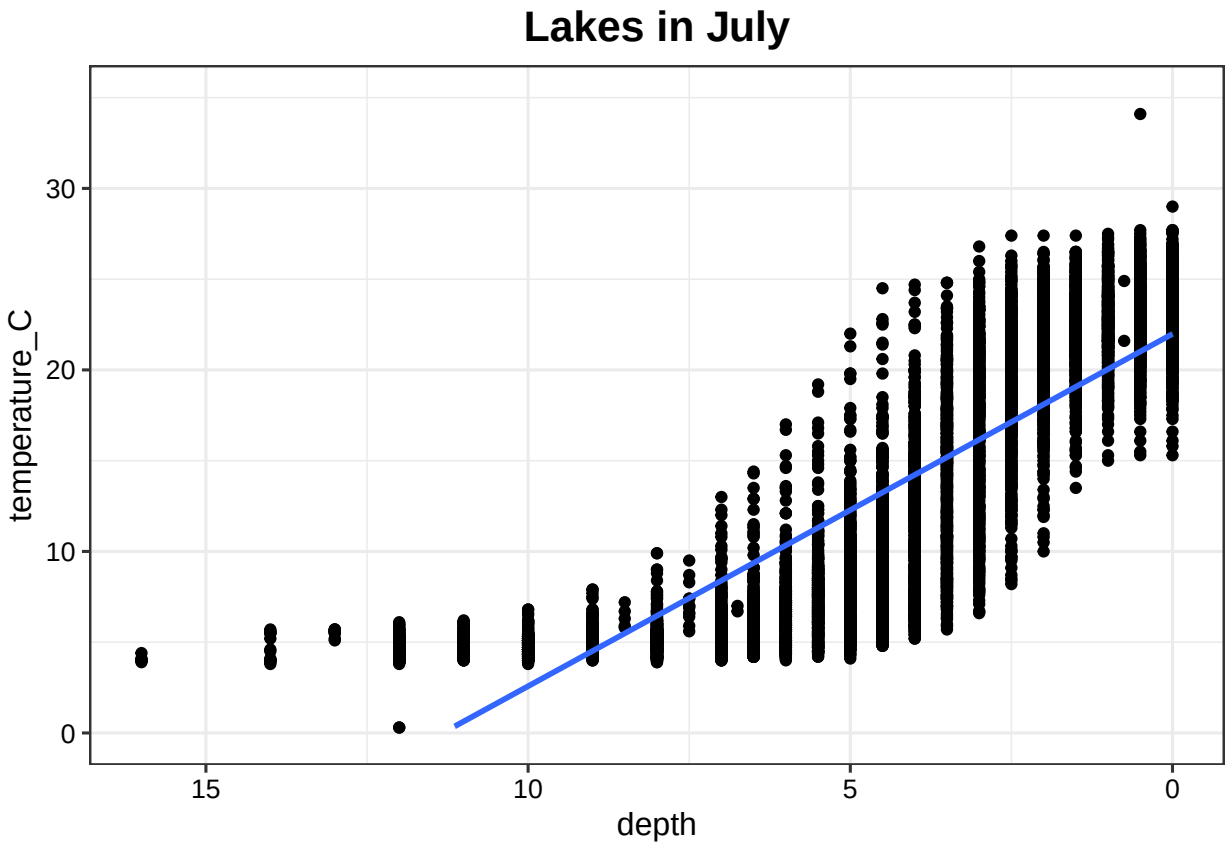
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
#4. Wrangle data
Lake_ChemPhys_Raw_Wrangled <-Lake_ChemPhys_Raw %>%
  filter(daynum>=183 & daynum<=213)%>%
  select(lakename:daynum, depth:temperature_C)%>%
  drop_na()


#5. Create scatter plot
Lake_ChemPhys_Raw_Wrangled.Scatter<-
ggplot(Lake_ChemPhys_Raw_Wrangled, aes(x=depth, y=temperature_C))+
  geom_point()+
  geom_smooth(method = lm)+
  ylim(0,35)+
  scale_x_reverse()+
  JG_default_theme+
  ggtitle("Lakes in July")

print(Lake_ChemPhys_Raw_Wrangled.Scatter)
```

```
## `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 anything about the linearity of this trend?

   Answer: This figure suggests that the greater the depth, the lower the temperature of the lake at that point. This trend seems fairly accurately linear because there is about an equal distribution (or mirroring) of points above and below the best fit line at least at shallower depths. However, there seems to be less data points from the deeper depths and it appears less linear at that end. Some of the lakes are also just not as deep as others, at least in the observed measurements.

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

```
#7. Linear regression
Temperature_Depth.Regression <-lm(data=Lake_ChemPhys_Raw_Wrangled, temperature_C ~ depth)

summary(Temperature_Depth.Regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = Lake_ChemPhys_Raw_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: This linear regression suggests a rejection of the null, which means that there is a difference of temperature values that are not because of depth, because the p-value is less than .05. About 73.71% of the temperature at the lakes can be explained by changes in depth. This finding is based on 9,671 degrees of freedom. For every 1m change in depth the temperature is predicted to decrease by 1.94 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. Run AIC on explanatory variables
Lake_ChemPhys.AIC <- lm(data = Lake_ChemPhys_Raw_Wrangled, temperature_C ~ year4 + daynum + depth)

step(Lake_ChemPhys.AIC)
```

```
## Start:  AIC=25998.22
## temperature_C ~ year4 + daynum + depth
##
##          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 ~ year4 + daynum + depth, data = Lake_ChemPhys_Raw_Wrangled)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##   -18.19700      0.01611      0.04024     -1.94133
```

```
#10. Run multiple regression
Lake_ChemPhys.AIC <- lm(data = Lake_ChemPhys_Raw_Wrangled, temperature_C ~ year4 + daynum + depth)
summary(Lake_ChemPhys.AIC)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = Lake_ChemPhys_Raw_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 *
## year4         0.016113   0.004353    3.701 0.000216 ***
## daynum        0.040237   0.004385    9.176  < 2e-16 ***
## depth        -1.941328   0.011728 -165.528  < 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 method suggested using all three explanatory variables from the wrangled data set: year4, daynum, and depth.This multiple regression model says that those three explanatory variables account for 73.97% of the variance in lake temperature.This is a tiny improvement (more variation accounted for) over the model that used only depth as an explanatory variable for temperature.

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 & linear model for variance between lakes

#ANOVA model
Lake_ChemPhys.anova <- aov(data = Lake_ChemPhys_Raw_Wrangled, temperature_C ~ lakename)
summary(Lake_ChemPhys.anova)
```

```
##               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
```

```
#Linear model
Lake_ChemPhys.lm <- lm(data = Lake_ChemPhys_Raw_Wrangled, temperature_C ~ lakename)
summary(Lake_ChemPhys.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = Lake_ChemPhys_Raw_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.

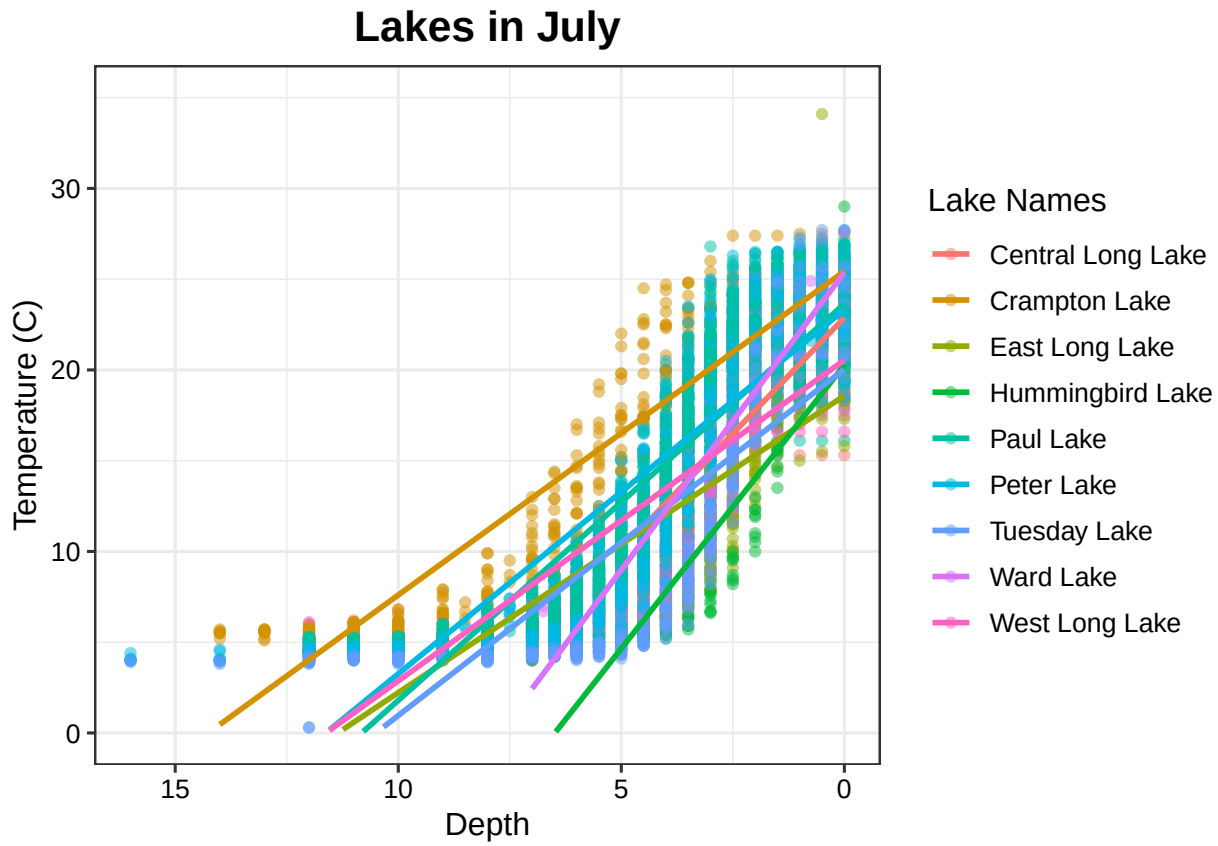
    Answer: There is a significant difference in mean temperatures across lakes. We know this because the p-value when running either the anova or linear model is very small (<2.2e-16), which means that we can reject the null hypothesis. That null hypothesis is that there is not difference in mean temperatures between the lakes. There are differences between the lakes's mean temperatures, and those differences are significant.

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.
Lake_temp_depth.scatter <-
  ggplot(Lake_ChemPhys_Raw_Wrangled, aes(x=depth, y=temperature_C, color =
      lakename))+
  geom_point(alpha = 0.5)+
  geom_smooth(method=lm, se= FALSE)+
  JG_default_theme+
  ggtitle("Lakes in July")+
  scale_x_reverse()+
  ylim(0,35)+
  labs(x= "Depth", y ="Temperature (C)", color = "Lake Names")

print(Lake_temp_depth.scatter)
```

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



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

```
#15. Run Tukey's HSD test
TukeyHSD(Lake_ChemPhys.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = Lake_ChemPhys_Raw_Wrangled)
##
## $lakename
##                                          diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -2.18508757 -4.5319912  0.1618160 0.0915179
## East Long Lake-Central Long Lake   -7.37946293 -9.5249061 -5.2340198 0.0000000
## Hummingbird Lake-Central Long Lake -6.68276989 -9.6520798 -3.7134600 0.0000000
## Paul Lake-Central Long Lake        -3.82336936 -5.8915944 -1.7551443 0.0000004
## Peter Lake-Central Long Lake       -4.31624752 -6.3799766 -2.2525184 0.0000000
## Tuesday Lake-Central Long Lake     -6.59373914 -8.6961647 -4.4913136 0.0000000
## Ward Lake-Central Long Lake        -3.20778556 -6.1355040 -0.2800671 0.0195251
## West Long Lake-Central Long Lake   -6.05416916 -8.1927792 -3.9155591 0.0000000
## East Long Lake-Crampton Lake       -5.19437536 -6.5946962 -3.7940545 0.0000000
## Hummingbird Lake-Crampton Lake     -4.49768232 -6.9825914 -2.0127732 0.0000007
## Paul Lake-Crampton Lake            -1.63828179 -2.9171590 -0.3594045 0.0023129
## Peter Lake-Crampton Lake           -2.13115995 -3.4027534 -0.8595665 0.0000072
## Tuesday Lake-Crampton Lake         -4.40865157 -5.7421303 -3.0751729 0.0000000
## Ward Lake-Crampton Lake            -1.02269799 -3.4577560  1.4123600 0.9307880
## West Long Lake-Crampton Lake       -3.86908159 -5.2589107 -2.4792525 0.0000000
```

```
## Hummingbird Lake-East Long Lake     0.69669304 -1.5988991  2.9922852 0.9905616
## Paul Lake-East Long Lake            3.55609357  2.7014024  4.4107847 0.0000000
## Peter Lake-East Long Lake           3.06321541  2.2194620  3.9069688 0.0000000
## Tuesday Lake-East Long Lake         0.78572379 -0.1486934  1.7201409 0.1828556
## Ward Lake-East Long Lake            4.17167737  1.9301428  6.4132120 0.0000003
## West Long Lake-East Long Lake       1.32529377  0.3120836  2.3385039 0.0016418
## Paul Lake-Hummingbird Lake          2.85940053  0.6358062  5.0829949 0.0021745
## Peter Lake-Hummingbird Lake         2.36652237  0.1471092  4.5859355 0.0263810
## Tuesday Lake-Hummingbird Lake       0.08903074 -2.1664094  2.3444709 1.0000000
## Ward Lake-Hummingbird Lake          3.47498433  0.4355186  6.5144501 0.0117238
## West Long Lake-Hummingbird Lake     0.62860073 -1.6606065  2.9178079 0.9952002
## Peter Lake-Paul Lake               -0.49287816 -1.1146082  0.1288519 0.2521516
## Tuesday Lake-Paul Lake             -2.77036979 -3.5104805 -2.0302590 0.0000000
## Ward Lake-Paul Lake                 0.61558380 -1.5521583  2.7833259 0.9939613
## West Long Lake-Paul Lake           -2.23079980 -3.0681906 -1.3934090 0.0000000
## Tuesday Lake-Peter Lake            -2.27749162 -3.0049438 -1.5500394 0.0000000
## Ward Lake-Peter Lake                1.10846196 -1.0549910  3.2719149 0.8108720
## West Long Lake-Peter Lake          -1.73792164 -2.5641457 -0.9116976 0.0000000
## Ward Lake-Tuesday Lake              3.38595358  1.1855572  5.5863500 0.0000641
## West Long Lake-Tuesday Lake         0.53956999 -0.3790495  1.4581895 0.6673292
## West Long Lake-Ward Lake           -2.84638360 -5.0813788 -0.6113884 0.0025399
```

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: Ward Lake has the same mean temperature as Peter Lake statistically speaking, because even though there is a mean difference of 1.1 degrees celsius that difference is not statistically significant because the p-value is .81 which is greater than the .05 significance threshold. Therefore, we keep the null hypothesis that the mean temperatures are the same. Paul Lake also has the same mean temperature as Peter Lake statistically speaking, because even though there is a mean difference of -.49 degrees celsius that difference is not statistically significant because the p-value is .25 which is greater than the .05 significance threshold. None of these lakes have a mean temperature that is statistically distinct from all the other lakes. You can tell that because for every lake there as at least one pairing with a p-value greater than 0.5, which means keeping the null that the mean temperatures are statistically significantly the same for that pairing.

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: If we are just looking at Peter Lake and Paul Lake we can do a two-sample T-test, which is designed for when you're looking at categorical values (lake names) with only two levels (Peter and Paul). The two sample t-test would also tell us if Peter Lake and Paul Lake have distinct mean temperatures.