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

Stephanie Kinser

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/skins/Documents/EDA/Environmental_Data_Analytics_2022/Assignments"
```

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1

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

```
library(agricolae)
library(lubridate)
```

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

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

```
library(ggridges)


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

#are there other columns that need to be formatted as date?

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

```
## [1] "Date"
```

```
#2
A6_theme <- theme_light(base_size = 12)+
   theme(axis.text = element_text(color = "black"),
         legend.position = "right", panel.grid.minor = element_blank())

theme_set(A6_theme)
```

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 does not change with depth in July. (Mean temperature across depths in July = 0)

   H1: The mean lake temperature changes with depth in July (Mean temperature across depths in July != 0)

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)

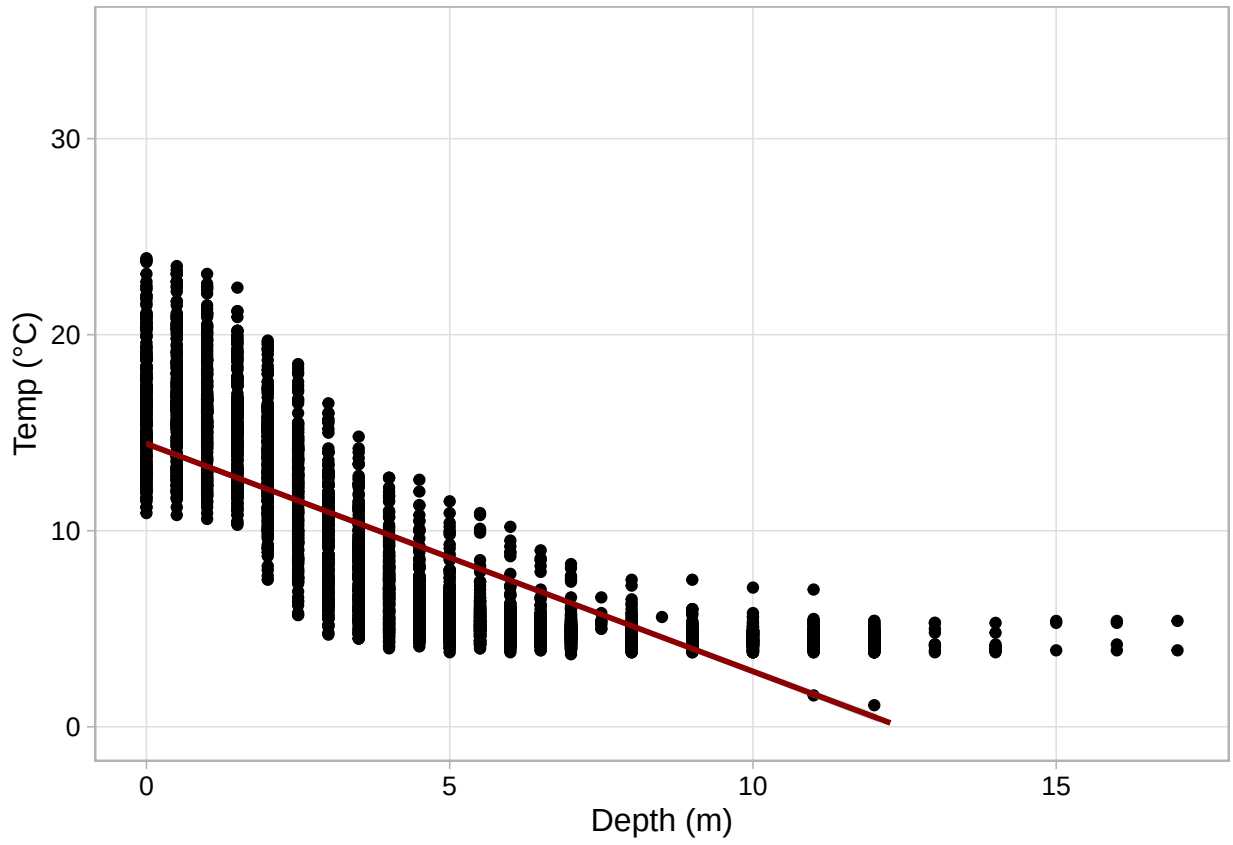
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
#4

NTL_LTER_tidy <- NTL_LTER %>%
  mutate(month = month(sampledate))%>%
  filter(month == 05) %>%
  select(month, lakename, year4, daynum, depth, temperature_C)%>%
  na.omit()

#5
plot_Q5<- ggplot(NTL_LTER_tidy, aes(x = depth, y = temperature_C)) +
  geom_point()+
  ylim(0,35)+
  geom_smooth(method = lm, color = "darkred")+
  labs(x = "Depth (m)", y = "Temp (°C)")
print(plot_Q5)
```

```
## `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: The figure suggests that temperatures generally decrease with a lower subsurface depth, demonstrating a negative relationship between temperature and depth. While there is a general downward sloping trend, the temperatures vary significantly with each depth measurement suggesting that a linear trend may not be the best model to account for all the variability.

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

```
#7
temp_regression <- lm(NTL_LTER_tidy$temperature_C ~ NTL_LTER_tidy$depth)
summary(temp_regression)
```

```
##
## Call:
## lm(formula = NTL_LTER_tidy$temperature_C ~ NTL_LTER_tidy$depth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.2541 -2.2895 -0.4476  2.0652 10.7042
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          14.43808    0.08093  178.41   <2e-16 ***
## NTL_LTER_tidy$depth -1.16131    0.01369  -84.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.958 on 4079 degrees of freedom
## Multiple R-squared:  0.6381, Adjusted R-squared:  0.638
## F-statistic:  7193 on 1 and 4079 DF,  p-value: < 2.2e-16
```

```
cor.test(NTL_LTER_tidy$temperature_C, NTL_LTER_tidy$depth)
```

```
##
##  Pearson's product-moment correlation
##
## data:  NTL_LTER_tidy$temperature_C and NTL_LTER_tidy$depth
## t = -84.809, df = 4079, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.809661 -0.787441
## sample estimates:
##        cor
## -0.7988233
```

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 results shows that we can reject the null hypothesis ($p < 0.05$) with a 95% confidence interval and that the mean temperature of lakes for July does change with depth. There is a negative correlation between temperature and depth; for every 1 additional meter in subsurface

depth, we expect a change of approximately -1.16 degrees Celsius in temperature. The model explains 63.8% of the variabilty between temperature and depth and there is a 2.96 residual standard error on 4079 degrees of freedom.

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
temp_AIC <- lm(data = NTL_LTER_tidy, temperature_C ~ year4 + daynum + depth)

#Choose a model by AIC in a Stepwise Algorithm
step(temp_AIC) #removes a variable one by one and runs a regression each time
```

```
## Start:  AIC=8745.05
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq   RSS     AIC
## <none>                34718  8745.1
## - year4   1        45 34763  8748.3
## - daynum  1       898 35615  8847.2
## - depth   1     63008 97726 12966.5

##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER_tidy)
##
## Coefficients:
## (Intercept)        year4       daynum        depth
##    23.31590     -0.01097      0.09041     -1.16222
```

```
#10
temp_model <- lm(data = NTL_LTER_tidy, temperature_C ~ year4 + daynum + depth)
summary(temp_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER_tidy)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4799 -2.2645 -0.4029  2.0595 10.3542
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.315899   9.709021   2.401   0.0164 *
## year4       -0.010970   0.004782  -2.294   0.0218 *
## daynum       0.090413   0.008806  10.267   <2e-16 ***
## depth       -1.162220   0.013511 -86.019   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.918 on 4077 degrees of freedom
## Multiple R-squared:  0.6479, Adjusted R-squared:  0.6476
## F-statistic:  2501 on 3 and 4077 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 that the AIC model suggests are all the variables we added to the model: year4, daynum, and depth. This model explains approximately 65% of variance between the explanatory variables and the dependent variables, and the model has a residual standard error of 2.918. This is an improvement over our model with only depth because the amount of variance explained is improved and the residual standard error is slightly lower.

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
#linear model
laketemp_lm <- lm(data = NTL_LTER_tidy, temperature_C ~ lakename)
summary(laketemp_lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_LTER_tidy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.725 -3.925 -2.464  3.704 15.836
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 12.9065     0.7179  17.978  < 2e-16 ***
## lakenameCrampton Lake       -3.1428     0.8406  -3.739 0.000187 ***
## lakenameEast Long Lake      -4.9425     0.7551  -6.546 6.65e-11 ***
## lakenameHummingbird Lake    -5.3675     1.0457  -5.133 2.99e-07 ***
```

```
## lakenamePaul Lake         -3.6106     0.7329  -4.926 8.71e-07 ***
## lakenamePeter Lake        -4.0813     0.7312  -5.582 2.54e-08 ***
## lakenameTuesday Lake      -4.9540     0.7451  -6.649 3.34e-11 ***
## lakenameWard Lake         -3.1084     0.8561  -3.631 0.000286 ***
## lakenameWest Long Lake    -4.1856     0.7565  -5.533 3.35e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.869 on 4072 degrees of freedom
## Multiple R-squared:  0.02096,    Adjusted R-squared:  0.01903
## F-statistic:  10.9 on 8 and 4072 DF,  p-value: 2.609e-15
```

```
#anova
laketemp_anova <- aov(data = NTL_LTER_tidy, temperature_C ~ lakename)
summary(laketemp_anova)
```

```
##               Df Sum Sq Mean Sq F value   Pr(>F)
## lakename       8   2066  258.29   10.89 2.61e-15 ***
## Residuals   4072  96533   23.71
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

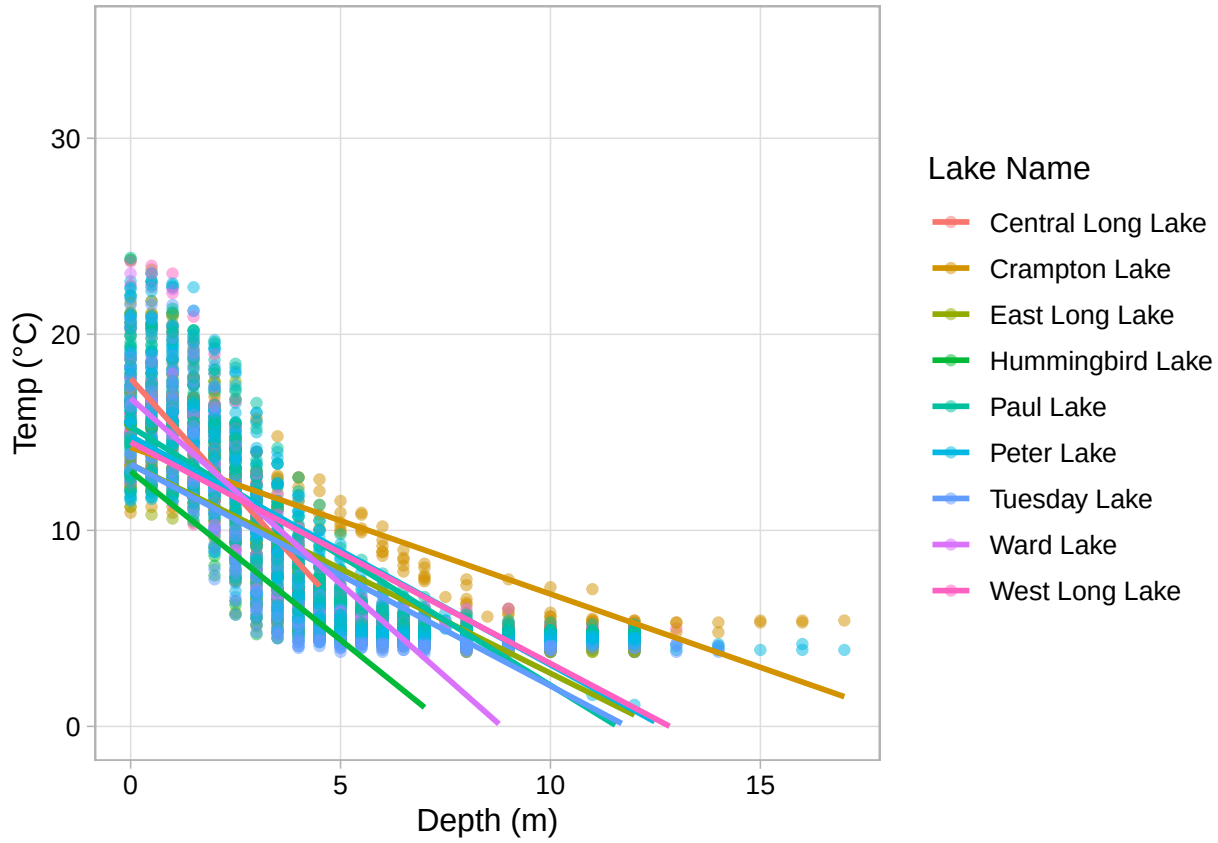
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: The results show that we can reject the null hypothesis and that there is a statistically significant (p-value <= .05) difference in means between temperatures across the lakes. The lm and aov models report a similar p-value of approximately <2.6e-15.

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.
plot_Q14 <- ggplot(NTL_LTER_tidy, aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha = 0.5)+
  ylim(0,35)+
  geom_smooth(aes(group = lakename), method = lm, se = FALSE)+
  labs(x = "Depth (m)", y = "Temp (°C)", color = "Lake Name")
print(plot_Q14)
```

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



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

```
#15
TukeyHSD(laketemp_anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = NTL_LTER_tidy)
##
## $lakename
##                                          diff        lwr        upr     p adj
## Crampton Lake-Central Long Lake    -3.14281206 -5.7514206 -0.5342035 0.0058482
## East Long Lake-Central Long Lake   -4.94254945 -7.2858020 -2.5992969 0.0000000
## Hummingbird Lake-Central Long Lake -5.36749735 -8.6128582 -2.1221365 0.0000106
## Paul Lake-Central Long Lake        -3.61056214 -5.8850273 -1.3360970 0.0000307
## Peter Lake-Central Long Lake       -4.08133837 -6.3506138 -1.8120629 0.0000009
## Tuesday Lake-Central Long Lake     -4.95400496 -7.2662810 -2.6417289 0.0000000
## Ward Lake-Central Long Lake        -3.10835660 -5.7650895 -0.4516237 0.0087221
## West Long Lake-Central Long Lake   -4.18560828 -6.5334543 -1.8377622 0.0000012
## East Long Lake-Crampton Lake       -1.79973739 -3.3387671 -0.2607076 0.0087808
## Hummingbird Lake-Crampton Lake     -2.22468529 -4.9468438  0.4974732 0.2148424
## Paul Lake-Crampton Lake            -0.46775008 -1.8998714  0.9643713 0.9847007
## Peter Lake-Crampton Lake           -0.93852630 -2.3623911  0.4853385 0.5113037
## Tuesday Lake-Crampton Lake         -1.81119290 -3.3026353 -0.3197505 0.0052244
```

```
## Ward Lake-Crampton Lake             0.03445546 -1.9494824  2.0183934 1.0000000
## West Long Lake-Crampton Lake       -1.04279622 -2.5888108  0.5032183 0.4780238
## Hummingbird Lake-East Long Lake    -0.42494790 -2.8939843  2.0440885 0.9998375
## Paul Lake-East Long Lake            1.33198731  0.4735202  2.1904544 0.0000534
## Peter Lake-East Long Lake           0.86121109  0.0165898  1.7058324 0.0416184
## Tuesday Lake-East Long Lake        -0.01145551 -0.9656015  0.9426905 1.0000000
## Ward Lake-East Long Lake            1.83419285  0.2149326  3.4534531 0.0131761
## West Long Lake-East Long Lake       0.75694118 -0.2804378  1.7943201 0.3642027
## Paul Lake-Hummingbird Lake          1.75693521 -0.6469159  4.1607863 0.3618321
## Peter Lake-Hummingbird Lake         1.28615898 -1.1127823  3.6851003 0.7686653
## Tuesday Lake-Hummingbird Lake       0.41349239 -2.0261651  2.8531499 0.9998553
## Ward Lake-Hummingbird Lake          2.25914075 -0.5091689  5.0274503 0.2165387
## West Long Lake-Hummingbird Lake     1.18188907 -1.2915073  3.6552854 0.8637646
## Peter Lake-Paul Lake               -0.47077622 -1.0998582  0.1583058 0.3286138
## Tuesday Lake-Paul Lake             -1.34344282 -2.1133477 -0.5735379 0.0000023
## Ward Lake-Paul Lake                 0.50220554 -1.0158072  2.0202183 0.9833881
## West Long Lake-Paul Lake           -0.57504613 -1.4459733  0.2958811 0.5088458
## Tuesday Lake-Peter Lake            -0.87266660 -1.6271021 -0.1182311 0.0101222
## Ward Lake-Peter Lake                0.97298176 -0.5372441  2.4832076 0.5438174
## West Long Lake-Peter Lake          -0.10426991 -0.9615526  0.7530128 0.9999887
## Ward Lake-Tuesday Lake              1.84564836  0.2715481  3.4197486 0.0084773
## West Long Lake-Tuesday Lake         0.76839668 -0.1969752  1.7337686 0.2468682
## West Long Lake-Ward Lake           -1.07725168 -2.7031521  0.5486487 0.5038692
```

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 a mean temperature that is statistically similar to Tuesday Lake. The mean temperatures only differ by -0.87 degrees Celsius and the p-value is 0.01 less than 0.05. Central Long Lake has a mean temperature that is statistically distinct from every other lake. Each of the p-values is less than 0.05 and the means between lake temperatures are several degrees different.

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 is another test we can use to explore whether Peter and Paul Lake have distinct mean temperatures. The HSD test extracts groupings to understand pairwise relationships.