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

Clara Fast

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()
require(tidyverse)
```

```
## Loading required package: tidyverse
```

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

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

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

```
require(agricolae)
```

```
## Loading required package: agricolae
```

```
require(lubridate)
```

```
## Loading required package: lubridate
```

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

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

```
#Import raw data file
Litter <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv",
                   stringsAsFactors = TRUE)

#Set date column to date object
Litter$sampledate <- as.Date(Litter$sampledate, format = "%m/%d/%y")

#2
#Build ggplot theme
mytheme <- theme_classic(base_size = 14) +
  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: Mean lake temperature recorded during July does not change with depth across all lakes Ha: Mean lake temperature recorded during July changes 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)

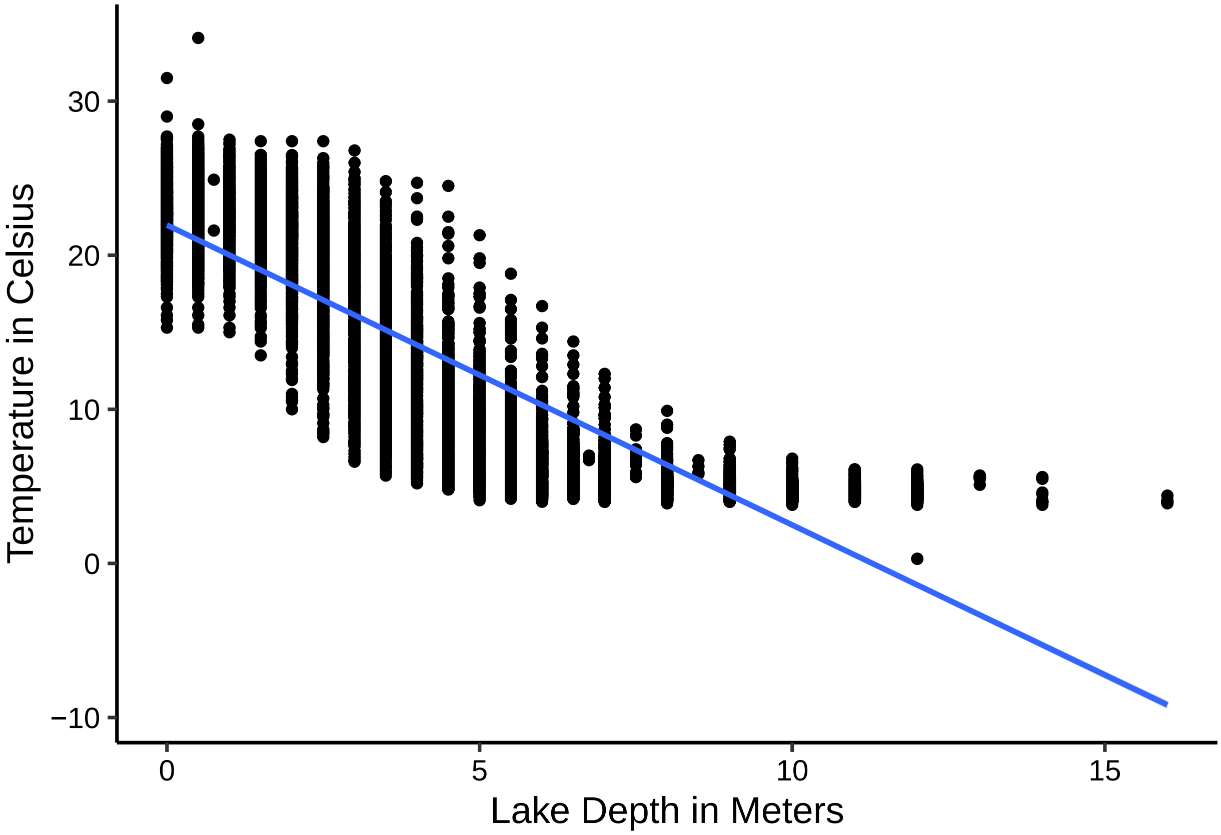
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```
#4
#Wrangle dataset to meet criteria
Litter_qfour <-
  Litter %>%
  filter(month(sampledate)==7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na(lakename, year4, daynum, depth, temperature_C)

#5
#Visualize relationship among temperature and depth with a scatter plot
Litter_tempdepth<-ggplot(Litter_qfour, aes(x = depth, y = temperature_C)) +
  geom_point() +
  geom_smooth(method='lm', formula= y~x) +
  labs(x="Lake Depth in Meters", y = "Temperature in Celsius")
  xlim(0,35)
```

```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 --   35
```

```
  print(Litter_tempdepth)
```



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 there is a negative relationship between temperature and depth. That is, as depth increases, temperature decreases. However, the graph clearly demonstrates that for one depth, there are an array of temperatures possible. This suggests that other variables may play a part in explaining the variability in temperature at decreasing depths. In other words, the wide distribution of points for one depth suggests that the trend is not linear. Moreover, the data points in general do not closely follow the linear model generated.

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

```
#7
#Generate linear regression
Litter_tempdepth_lm <- lm(data = Litter_qfour, temperature_C ~ depth)
summary(Litter_tempdepth_lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = Litter_qfour)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5173 -3.0192  0.0633  2.9365 13.5834
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.95597    0.06792   323.3   <2e-16 ***
## depth       -1.94621    0.01174  -165.8   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.835 on 9726 degrees of freedom
## Multiple R-squared:  0.7387, Adjusted R-squared:  0.7387
## F-statistic: 2.75e+04 on 1 and 9726 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: About 74% of the variability in temperature is explained by changes in depth (Adjusted R-Squared: 0.7387). This is based on 9726 degrees of freedom. The result of this linear model is statistically significant, as demonstrated by the p-value of < 2.2e-16, indicating that the mean is significantly different from 0. Depth is a statistically significant indicator of temperature, with a p-value of <2e-16 ***. For every 1 meter change in depth, temperature changes by 21.95597 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
Litter_lm_test <- lm(data = Litter_qfour, temperature_C ~ depth + year4 +
                       daynum)

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

```
## Start:  AIC=26065.53
## temperature_C ~ depth + year4 + daynum
## 
##          Df Sum of Sq    RSS   AIC
## <none>                 141687 26066
## - year4   1       101 141788 26070
## - daynum  1      1237 142924 26148
## - depth   1    404475 546161 39189
```

```
## 
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = Litter_qfour)
## 
## Coefficients:
## (Intercept)        depth        year4       daynum  
##    -8.57556     -1.94644      0.01134      0.03978
```

```
#10
#Run multiple regression on the recommended set of variables
Litter_lm_new<-lm(formula = temperature_C ~ depth + year4 + daynum,
                  data = Litter_qfour)
summary(Litter_lm_new)
```

```
## 
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = Litter_qfour)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -9.6536 -3.0000  0.0902  2.9658 13.6123 
## 
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)    
## (Intercept) -8.575564   8.630715   -0.994  0.32044    
## depth       -1.946437   0.011683 -166.611  < 2e-16 ***
## year4        0.011345   0.004299    2.639  0.00833 ** 
## daynum       0.039780   0.004317    9.215  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 3.817 on 9724 degrees of freedom
## Multiple R-squared:  0.7412, Adjusted R-squared:  0.7411 
## F-statistic:  9283 on 3 and 9724 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 method suggests we use to predict temperature are depth, year, and daynum. This model explains about 74% of the observed variance (Adjusted R-squared: 0.7411). This is not much of an improvement over the previous model (Adjusted R-squared: Adjusted R-Squared: 0.7387).

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
#Create ANOVA and linear models
aovlitter_lakename<-aov(data=Litter_qfour, temperature_C~lakename)
summary(aovlitter_lakename)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)    
## lakename       8  21642  2705.2      50 <2e-16 ***
## Residuals   9719 525813    54.1                   
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
lmlitter_lakename<-lm(data=Litter_qfour, temperature_C~lakename)
summary(lmlitter_lakename)
```

```
## 
## Call:
## lm(formula = temperature_C ~ lakename, data = Litter_qfour)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -10.769  -6.614  -2.679   7.684  23.832 
## 
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               17.6664     0.6501  27.174  < 2e-16 ***
## lakenameCrampton Lake     -2.3145     0.7699  -3.006 0.002653 ** 
## lakenameEast Long Lake    -7.3987     0.6918 -10.695  < 2e-16 ***
## lakenameHummingbird Lake  -6.8931     0.9429  -7.311 2.87e-13 ***
## lakenamePaul Lake         -3.8522     0.6656  -5.788 7.36e-09 ***
## lakenamePeter Lake        -4.3501     0.6645  -6.547 6.17e-11 ***
## lakenameTuesday Lake      -6.5972     0.6769  -9.746  < 2e-16 ***
## lakenameWard Lake         -3.2078     0.9429  -3.402 0.000672 ***
## lakenameWest Long Lake    -6.0878     0.6895  -8.829  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.355 on 9719 degrees of freedom
## Multiple R-squared:  0.03953,    Adjusted R-squared:  0.03874
## F-statistic:    50 on 8 and 9719 DF,  p-value: < 2.2e-16
```

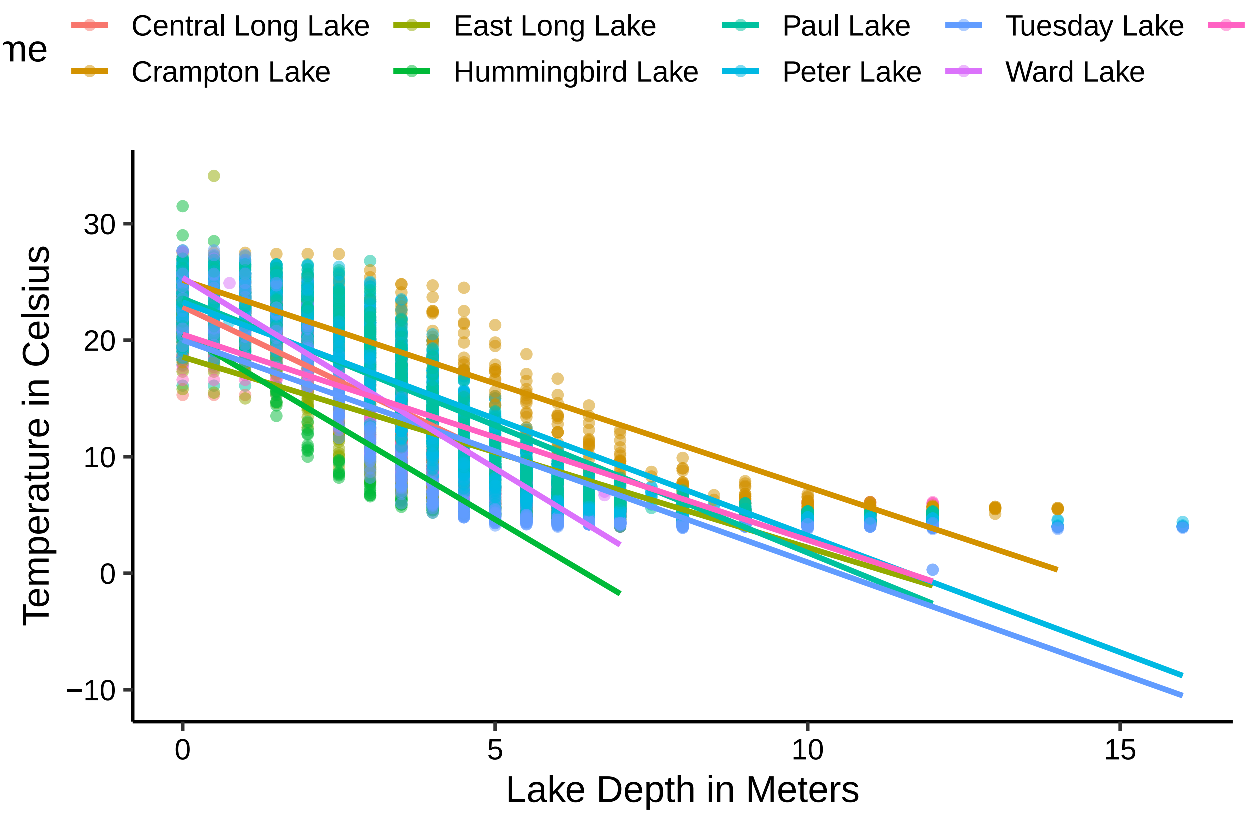
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: With 8 degrees of freedom, the summary of the ANOVA model shows that there is a significant difference in mean temperatures among the lakes, as illustrated by the p-value of <2e-16 *** for the variable of lakename. This does not tell us however which lakes are different, as such post-hoc tests are needed. This will allow us to compare all possible group pairings. As for the linear model, it is visible that the lakes possess different means, and these means are statistically significant. The linear model however only explains about 4% of the variance in temperatures (Adjusted R-squared: 0.03874).

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.
#Create scatterplot
ggplot(Litter_qfour, aes(x = depth, y = temperature_C, color = lakename)) +
  geom_point(alpha=0.5) +
  geom_smooth(method='lm', se=FALSE) +
  labs(x="Lake Depth in Meters", y = "Temperature in Celsius")
```

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

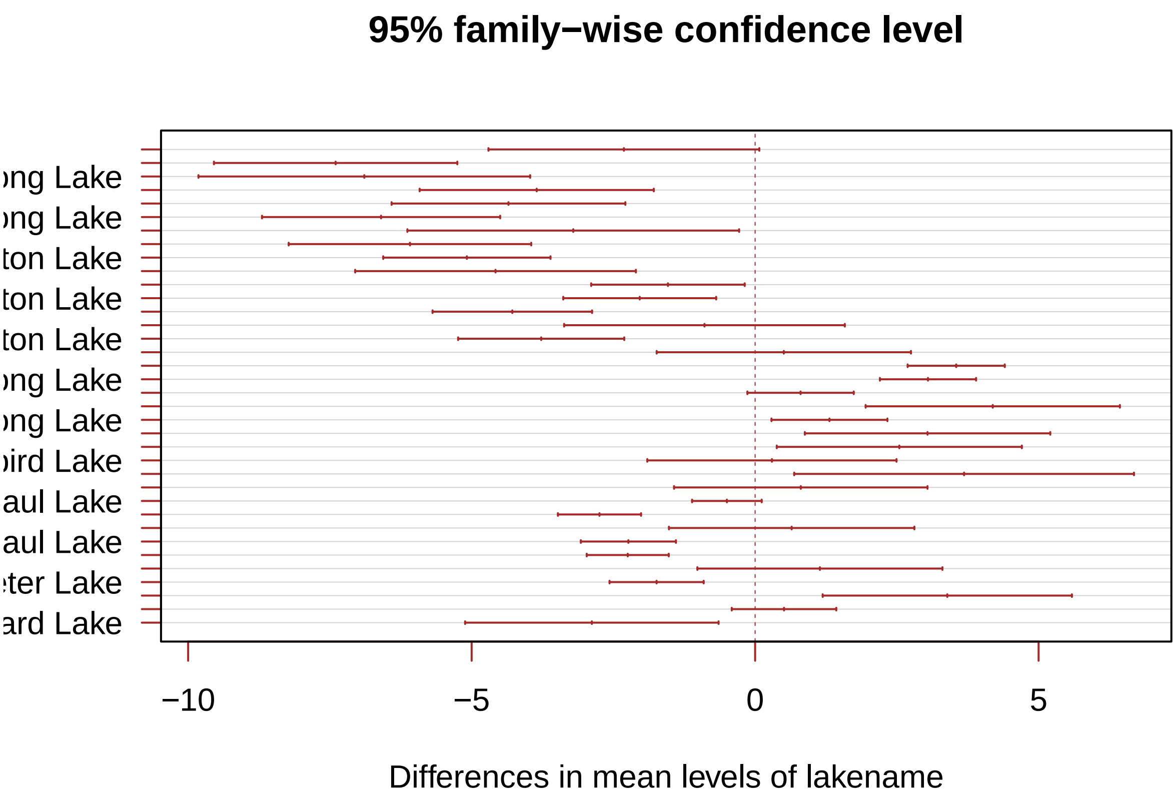


```
  ylim(0,35)
```

```
## <ScaleContinuousPosition>
##  Range:
##  Limits:    0 --   35
```

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

```
#15
#Generate Tukey's HSD test
lakenamestukey<-TukeyHSD(aovlitter_lakename)
plot(lakenamestukey , las=1 , col="brown")
```



```r
#Extract groupings for pairwise relationships
lakenamestukey_groupings <- HSD.test(aovlitter_lakename, "lakename",
                                     group = TRUE)
lakenamestukey_groupings
```

```
## $statistics
##   MSerror   Df     Mean       CV
##   54.1016 9719 12.72087 57.82135
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.387504  0.05
##
## $means
##                     temperature_C      std    r Min  Max    Q25   Q50    Q75
## Central Long Lake        17.66641 4.196292  128 8.9 26.8 14.400 18.40 21.000
## Crampton Lake            15.35189 7.244773  318 5.0 27.5  7.525 16.90 22.300
## East Long Lake           10.26767 6.766804  968 4.2 34.1  4.975  6.50 15.925
## Hummingbird Lake         10.77328 7.017845  116 4.0 31.5  5.200  7.00 15.625
## Paul Lake                13.81426 7.296928 2660 4.7 27.7  6.500 12.40 21.400
## Peter Lake               13.31626 7.669758 2872 4.0 27.0  5.600 11.40 21.500
## Tuesday Lake             11.06923 7.698687 1524 0.3 27.7  4.400  6.80 19.400
## Ward Lake                14.45862 7.409079  116 5.7 27.6  7.200 12.55 23.200
## West Long Lake           11.57865 6.980789 1026 4.0 25.7  5.400  8.00 18.800
##
## $comparison
## NULL
##
```

```
## $groups
##                   temperature_C groups
## Central Long Lake      17.66641      a
## Crampton Lake          15.35189     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.81426      c
## Peter Lake             13.31626      c
## West Long Lake         11.57865      d
## Tuesday Lake           11.06923     de
## Hummingbird Lake       10.77328     de
## East Long Lake         10.26767      e
##
## attr(,"class")
## [1] "group"
```

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: Paul Lake has the same mean temperature, statistically speaking, as Peter Lake, as they have been assigned the same group, "c". No lake has a mean temperature that is statistically distinct from all the other lakes as visible from the groupings - there is no group/letter that is unique.

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 use a two-sample T-test to explore whether the two lakes have distinct means. A two-sample T-test is used to test whether the mean of two samples is equivalent.