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

Amanda Booth

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

#Check working directory and load packages
getwd()
```

```
## [1] "/Users/amandabooth/Documents/GitHub/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   1.0.8
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
```

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

```
library(agricolae)

#Import data

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

#2

#Build theme
mytheme <- theme_linedraw(base_size = 10) +
  theme(axis.text = element_text(color = "rosybrown"),
        legend.position = "bottom")

#Set as default theme
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 the mean lake temperature recorded in July due to the change in depth == 0 Ha: The change in the mean lake temperature recorded in July due to the change in depth != 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_July <- NTL.data %>%
  select(lakename:daynum, depth:temperature_C) %>%
  filter(daynum == 183:213) %>%
  na.omit()
```

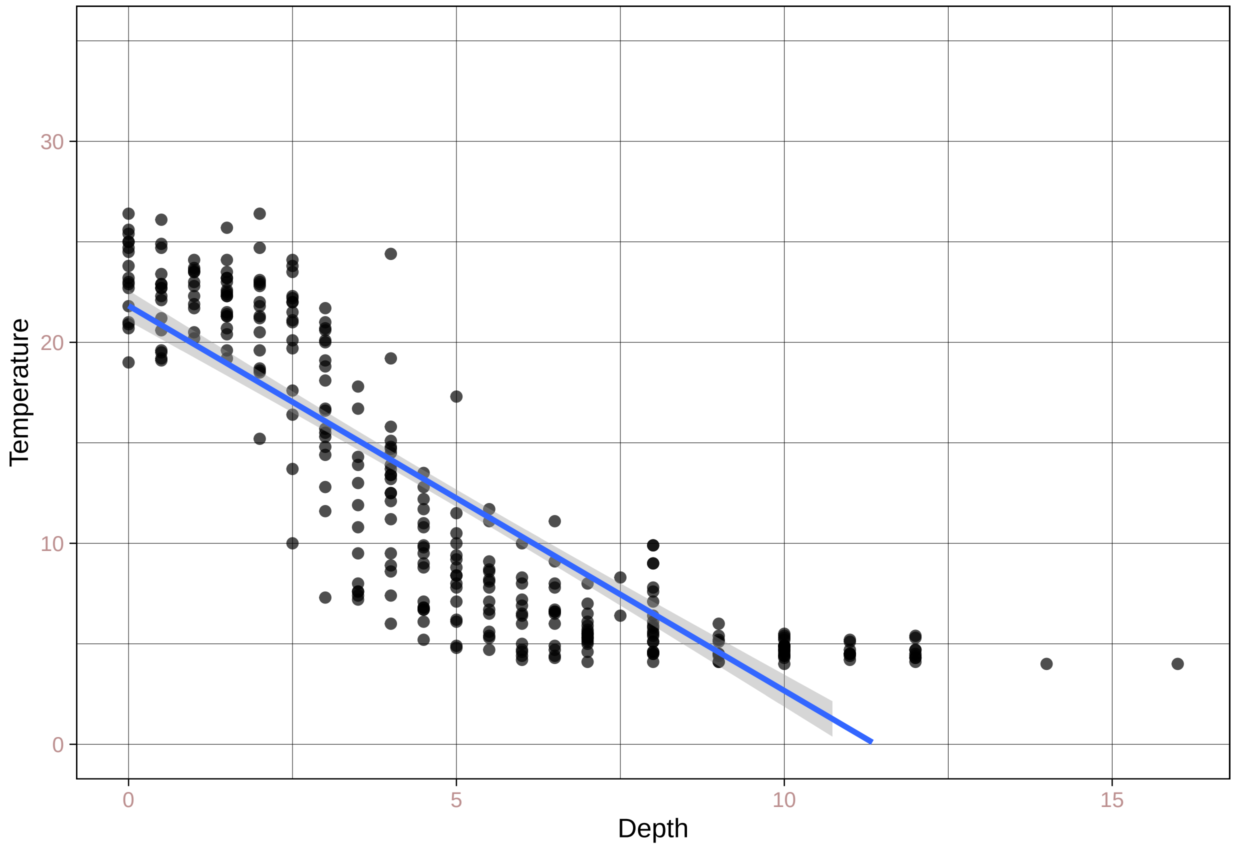
```
## Warning in daynum == 183:213: longer object length is not a multiple of shorter
## object length
```

```r
#5

Plot1 <- ggplot(NTL_July, aes(x = depth, y = temperature_C)) +
  geom_point(alpha = 0.7)+
  geom_smooth(method = "lm") +
  ylim(0,35) +
  labs(x = "Depth", y = "Temperature")
print(Plot1)
```

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

```
## Warning: Removed 23 rows containing missing values (geom_smooth).
```



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: It generally looks as though the deeper the water, the lower the temperature. The distribution of points gives the impression that the the temperature decreases at different rates depending on the depth, and therefore the trend is likely not linear.

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

```r
#7

temp.depth.regression <- lm(data = NTL_July, temperature_C ~ depth)
summary(temp.depth.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL_July)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7745 -2.8711 -0.0203  2.8949 12.8240
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.82033    0.38745   56.32   <2e-16 ***
## depth       -1.91527    0.06628  -28.90   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.839 on 307 degrees of freedom
## Multiple R-squared:  0.7312, Adjusted R-squared:  0.7303
## F-statistic:   835 on 1 and 307 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: The depth is a variable with a statisticlly significant impact on the temperature and can explain approximately 73.03% of the result of the temperature. For every 1 unit that depth increases, temperature decreases by 1.91 degrees C. This finding is based on 307 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.

```r
#9

#Create object
temp.AIC <- lm(data = NTL_July, temperature_C ~ depth + year4 + daynum)
```

```r
#Perform test

step(temp.AIC)
```

```
## Start:  AIC=833.38
## temperature_C ~ depth + year4 + daynum
##
##          Df Sum of Sq     RSS     AIC
## - year4   1       6.5  4473.6  831.83
## <none>                 4467.1  833.38
## - daynum  1      50.9  4518.0  834.88
## - depth   1   12277.1 16744.2 1239.67
##
## Step:  AIC=831.83
## temperature_C ~ depth + daynum
##
##          Df Sum of Sq     RSS     AIC
## <none>                 4473.6  831.83
## - daynum  1      51.1  4524.7  833.35
## - depth   1   12274.2 16747.8 1237.74

##
## Call:
## lm(formula = temperature_C ~ depth + daynum, data = NTL_July)
##
## Coefficients:
## (Intercept)        depth       daynum
##    12.68623     -1.91305      0.04596
```

```r
#10

temp.multi.regression <- lm(data = NTL_July, temperature_C ~ depth + daynum)
summary(temp.multi.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + daynum, data = NTL_July)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.5627 -2.9372  0.1678  2.7994 12.4555
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.68623    4.89920   2.589   0.0101 *
## depth       -1.91305    0.06602 -28.975   <2e-16 ***
## daynum       0.04596    0.02457   1.870   0.0624 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.824 on 306 degrees of freedom
## Multiple R-squared:  0.7342, Adjusted R-squared:  0.7325
## F-statistic: 422.6 on 2 and 306 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 suggests using depth and daynum to predict temperature. This model explains 73.25% of the observed variance, which is a slight improvement from only using depth as an explanatory variable. However, daynum is not significantly different from zero when using a 95% confidence interval.

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

#Linear model

lake.temp.linear <- lm(data =NTL_July, temperature_C ~ lakename)
summary(lake.temp.linear)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_July)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.391 -6.302 -2.993  7.309 14.898
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               14.3000     5.1844   2.758  0.00617 **
## lakenameCrampton Lake     -1.1000     5.5998  -0.196  0.84440
## lakenameEast Long Lake    -4.6067     5.3544  -0.860  0.39028
## lakenameHummingbird Lake  -2.2600     6.1342  -0.368  0.71282
## lakenamePaul Lake         -0.2226     5.2457  -0.042  0.96618
## lakenamePeter Lake        -0.9090     5.2423  -0.173  0.86246
## lakenameTuesday Lake      -3.5979     5.2913  -0.680  0.49705
## lakenameWard Lake         -2.1000     6.3495  -0.331  0.74108
## lakenameWest Long Lake    -2.5829     5.3304  -0.485  0.62835
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.332 on 300 degrees of freedom
## Multiple R-squared:  0.04187,    Adjusted R-squared:  0.01632
## F-statistic: 1.639 on 8 and 300 DF,  p-value: 0.1133
```

```r
#ANOVA model

lake.temp.anova <- aov(data =NTL_July, temperature_C ~ lakename)
summary(lake.temp.anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## lakename      8    705   88.09   1.639  0.113
## Residuals   300  16127   53.76
```

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

    Answer: There is not a statistically significant difference in mean temperature among the lakes. Therefore, we do not reject the null hypothesis. The variable lakename can only explain 1.63% of the result of the temperature. This finding is based on 300 degrees of freedom.

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.

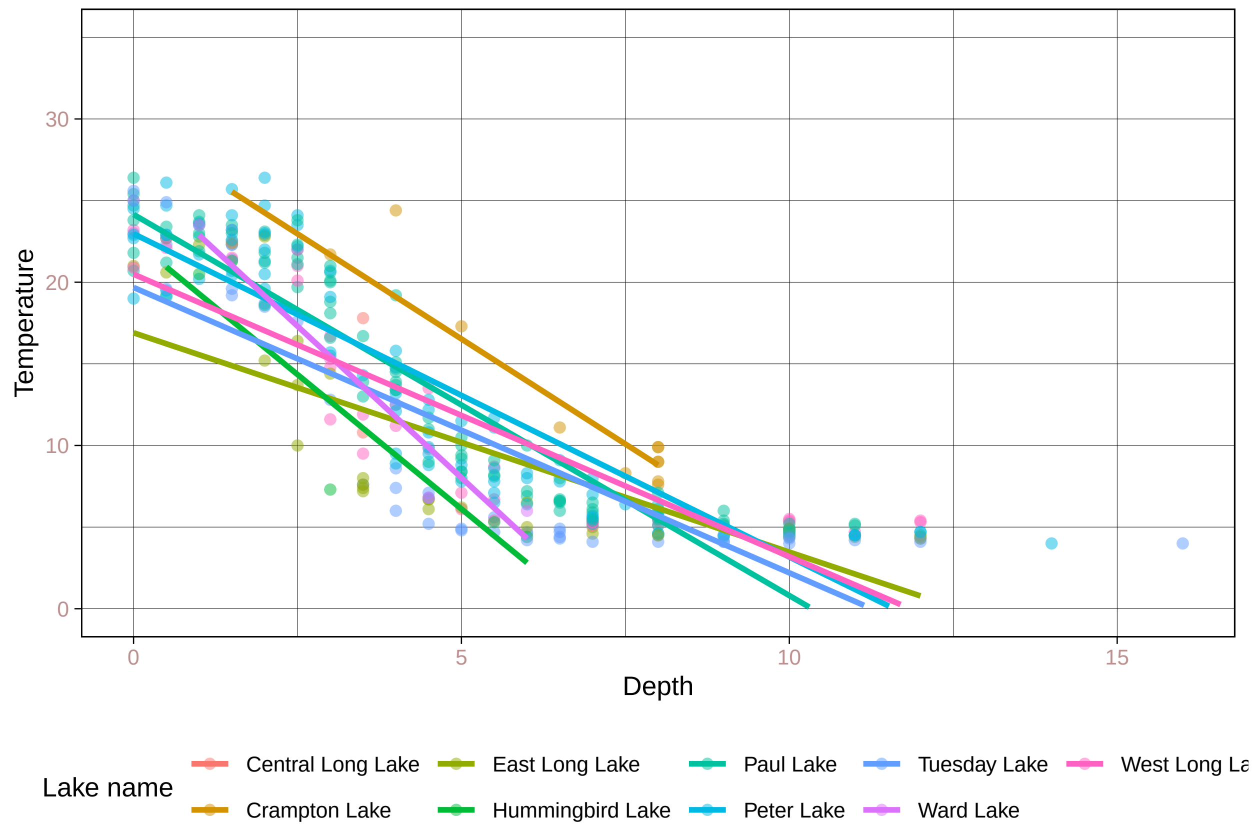
```r
#14.

Plot2 <- ggplot(NTL_July, aes(x = depth, y = temperature_C, color = lakename))+
  geom_point(alpha = 0.5) +
  geom_smooth (method = "lm", se = FALSE) +
  ylim(0, 35) +
   labs(x = "Depth", y = "Temperature", color = "Lake name")


print(Plot2)
```

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

```
## Warning: Removed 45 rows containing missing values (geom_smooth).
```



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

```
#15

#Tukey test

tukey.test <- TukeyHSD(lake.temp.anova)

print(tukey.test)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = NTL_July)
##
## $lakename
##                                          diff         lwr        upr     p adj
## Crampton Lake-Central Long Lake    -1.1000000 -18.5970303 16.3970303 0.9999999
## East Long Lake-Central Long Lake   -4.6066667 -21.3370277 12.1236943 0.9947352
## Hummingbird Lake-Central Long Lake -2.2600000 -21.4270364 16.9070364 0.9999903
## Paul Lake-Central Long Lake        -0.2226190 -16.6134335 16.1681954 1.0000000
## Peter Lake-Central Long Lake       -0.9089888 -17.2890923 15.4711148 1.0000000
## Tuesday Lake-Central Long Lake     -3.5979167 -20.1310563 12.9352230 0.9989988
## Ward Lake-Central Long Lake        -2.1000000 -21.9397676 17.7397676 0.9999958
## West Long Lake-Central Long Lake   -2.5828571 -19.2383616 14.0726473 0.9999199
```

```
## East Long Lake-Crampton Lake       -3.5066667 -11.3315765  4.3182432 0.8971560
## Hummingbird Lake-Crampton Lake     -1.1600000 -13.3542413 11.0342413 0.9999982
## Paul Lake-Crampton Lake             0.8773810  -6.1924869  7.9472488 0.9999855
## Peter Lake-Crampton Lake            0.1910112  -6.8539889  7.2360114 1.0000000
## Tuesday Lake-Crampton Lake         -2.4979167  -9.8917615  4.8959282 0.9798155
## Ward Lake-Crampton Lake            -1.0000000 -14.2265117 12.2265117 0.9999997
## West Long Lake-Crampton Lake       -1.4828571  -9.1464113  6.1806970 0.9995769
## Hummingbird Lake-East Long Lake     2.3466667  -8.7194270 13.4127603 0.9991706
## Paul Lake-East Long Lake            4.3840476  -0.4885226  9.2566178 0.1163541
## Peter Lake-East Long Lake           3.6976779  -1.1387398  8.5340956 0.2938732
## Tuesday Lake-East Long Lake         1.0087500  -4.3230273  6.3405273 0.9996419
## Ward Lake-East Long Lake            2.5066667  -9.6875746 14.7009079 0.9993393
## West Long Lake-East Long Lake       2.0238095  -3.6761002  7.7237192 0.9725248
## Paul Lake-Hummingbird Lake          2.0373810  -8.5083407 12.5831027 0.9995818
## Peter Lake-Hummingbird Lake         1.3510112  -9.1780553 11.8800778 0.9999813
## Tuesday Lake-Hummingbird Lake      -1.3379167 -12.1035172  9.4276839 0.9999854
## Ward Lake-Hummingbird Lake          0.1600000 -15.2078179 15.5278179 1.0000000
## West Long Lake-Hummingbird Lake    -0.3228571 -11.2754494 10.6297351 1.0000000
## Peter Lake-Paul Lake               -0.6863697  -4.1713021  2.7985626 0.9995177
## Tuesday Lake-Paul Lake             -3.3752976  -7.5203751  0.7697799 0.2153195
## Ward Lake-Paul Lake                -1.8773810 -13.6014305  9.8466686 0.9998977
## West Long Lake-Paul Lake           -2.3602381  -6.9692281  2.2487519 0.8042784
## Tuesday Lake-Peter Lake            -2.6889279  -6.7914470  1.4135912 0.5111686
## Ward Lake-Peter Lake               -1.1910112 -12.9000818 10.5180593 0.9999969
## West Long Lake-Peter Lake          -1.6738684  -6.2446215  2.8968847 0.9668554
## Ward Lake-Tuesday Lake              1.4979167 -10.4242999 13.4201332 0.9999841
## West Long Lake-Tuesday Lake         1.0150595  -4.0769634  6.1070825 0.9994728
## West Long Lake-Ward Lake           -0.4828571 -12.5741916 11.6084773 1.0000000
```

```r
#Pairwise test

pairwise.test <- HSD.test(lake.temp.anova, "lakename", group = TRUE)

print(pairwise.test)
```

```
## $statistics
##   MSerror  Df     Mean       CV
##   53.7555 300 12.57249 58.31632
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.418852  0.05
##
## $means
##                   temperature_C      std  r  Min  Max    Q25   Q50    Q75
## Central Long Lake     14.300000 4.949747  2 10.8 17.8 12.550 14.30 16.050
## Crampton Lake         13.200000 6.361389 12  7.6 24.4  8.825  9.90 18.400
## East Long Lake         9.693333 6.386350 30  4.3 22.8  4.750  6.70 14.225
## Hummingbird Lake      12.040000 8.824285  5  4.4 22.7  5.300  7.30 20.500
## Paul Lake             14.077381 7.102237 84  5.1 26.4  6.850 13.40 21.200
## Peter Lake            13.391011 7.587634 89  4.0 26.4  6.400 11.50 21.300
## Tuesday Lake          10.702083 8.056410 48  4.0 25.6  4.375  5.40 19.300
## Ward Lake             12.200000 8.070109  4  6.0 23.5  6.600  9.65 15.250
## West Long Lake        11.717143 6.995355 35  4.7 25.0  5.400  9.50 18.400
```

```
##
## $comparison
## NULL
##
## $groups
##                    temperature_C groups
## Central Long Lake      14.300000      a
## Paul Lake              14.077381      a
## Peter Lake             13.391011      a
## Crampton Lake          13.200000      a
## Ward Lake              12.200000      a
## Hummingbird Lake       12.040000      a
## West Long Lake         11.717143      a
## Tuesday Lake           10.702083      a
## East Long Lake          9.693333      a
##
## 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: All of the lakes have the same mean temperature, statistically speaking, as Peter Lake. None of the lakes have a mean temperature that is statistically distinct from all the other lakes.

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: Two-sample t-test