

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

Catherine Otero

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/Catherine/OneDrive - Duke University/GitHub_Spot/Environmental_Data_Analytics_2021/Env"
```

```
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)
library(lubridate)

##
## Attaching package: 'lubridate'

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

LakeChemistry <- read.csv("../Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

class(LakeChemistry$sampdate)

## [1] "character"

#Use fwd slashes in format = "%m/%d/%y" instead of dashes. Dashes yields NAs
LakeChemistry$sampdate <- as.Date(LakeChemistry$sampdate, format = "%m/%d/%y")

class(LakeChemistry$sampdate)

## [1] "Date"

#2
theme_set(theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"), legend.position = "bottom"))

```

Simple regression

Our first research question is: Does mean lake temperature recorded during July change with depth across all lakes?

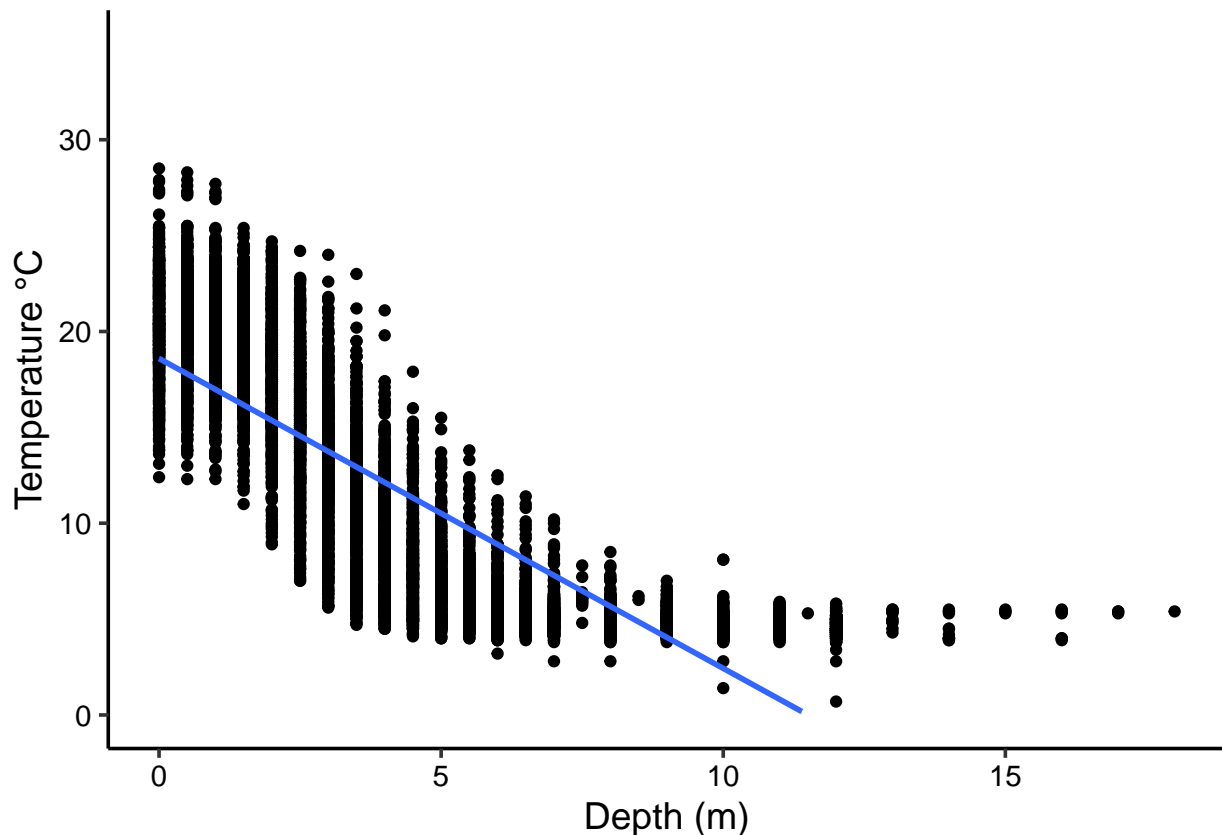
3. State the null and alternative hypotheses for this question: > Answer: H0: Lake temperature does not change with depth across all lakes. Ha: Lake temperature changes with depth across all lakes.
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following criteria:
 - 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
LakeChemistry.July <- LakeChemistry %>%
  mutate(Month = month(sampledate)) %>%
  filter(Month == 6) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na()

#5
TempbyDepth <- ggplot(LakeChemistry.July, aes(x = depth, y = temperature_C)) +
  geom_point() +
  ylim(0,35) +
  ylab("Temperature \u00B0C") +
  xlab("Depth (m)") +
  geom_smooth(method = lm)
print(TempbyDepth)
```

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

```
## Warning: Removed 29 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: At shallower depths there is more temperature variance than at lower depths where there is less variance in temperatures. Temperature and depth have an inverse relationship. As

depth increases temperature decreases until about 8-10 m where the relationship levels out. The distribution suggests that there is a linear relationship but that it is not super tight.

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

```
#7
LakeChemistry.July.regression <- lm(LakeChemistry.July$temperature_C ~ LakeChemistry.July$depth)
summary(LakeChemistry.July.regression)

##
## Call:
## lm(formula = LakeChemistry.July$temperature_C ~ LakeChemistry.July$depth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.2359 -2.8873 -0.2792  2.6694 15.8990
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    18.59256     0.06427   289.3   <2e-16 ***
## LakeChemistry.July$depth -1.61620     0.01100  -146.9   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.582 on 9501 degrees of freedom
## Multiple R-squared:  0.6943, Adjusted R-squared:  0.6942
## F-statistic: 2.158e+04 on 1 and 9501 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: We reject the null based on our results. The results are statistically significant with a p-value of $< 2.2e-16$. About 69% of the observed variance is explained by depth. The degrees of freedom are 9501. Temperature is predicted to change by 1.6 degrees for every 1m of change in depth. The results suggest a negative correlation between depth and temperature.

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

```
JulyAIC <- lm(data = LakeChemistry.July, temperature_C ~ depth + daynum + year4)
```

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

```
## Start:  AIC=23932.45  
## temperature_C ~ depth + daynum + year4  
##  
##           Df Sum of Sq    RSS   AIC  
## <none>                 117822 23932  
## - year4    1           31 117853 23933  
## - daynum   1          4040 121862 24251  
## - depth    1         276513 394335 35410  
  
##  
## Call:  
## lm(formula = temperature_C ~ depth + daynum + year4, data = LakeChemistry.July)  
##  
## Coefficients:  
## (Intercept)      depth      daynum      year4  
##   18.805558   -1.615432    0.074440   -0.006318
```

#10

```
JulyModel <- lm(data = LakeChemistry.July, temperature_C ~ depth + daynum + year4)  
summary(JulyModel)
```

```
##  
## Call:  
## lm(formula = temperature_C ~ depth + daynum + year4, data = LakeChemistry.July)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -8.6228 -2.8170 -0.1937  2.7410 15.7528   
##  
## Coefficients:  
##              Estimate Std. Error  t value Pr(>|t|)      
## (Intercept) 18.805558   7.973540   2.358    0.0184 *     
## depth       -1.615432   0.010819 -149.308 <2e-16 ***   
## daynum       0.074440   0.004125  18.047 <2e-16 ***   
## year4       -0.006318   0.003970  -1.591    0.1116      
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 3.522 on 9499 degrees of freedom  
## Multiple R-squared:  0.7045, Adjusted R-squared:  0.7044   
## F-statistic: 7549 on 3 and 9499 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 is depth, daynum4, and year4. The model explains about 70% of the observed variance. This model with additional explanatory variables is an improvement over the model by a little bit. Depth explained about 69% of the 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).

#12

#Format as AOV

```
July.anova.aov <- aov(data = LakeChemistry.July, temperature_C ~ lakename)
summary(July.anova.aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## lakename      8 12019  1502.4   36.88 <2e-16 ***
## Residuals    9494 386708    40.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Format as lm

```
July.anova.lm <- lm(data = LakeChemistry.July, temperature_C ~ lakename)
summary(July.anova.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = LakeChemistry.July)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.405  -5.358  -2.914   5.886  19.302
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    15.6992     0.5802  27.058 < 2e-16 ***
## lakenameCrampton Lake    -2.8291     0.6918  -4.090 4.36e-05 ***
## lakenameEast Long Lake   -6.5010     0.6169 -10.538 < 2e-16 ***
## lakenameHummingbird Lake -6.7827     0.8428  -8.048 9.44e-16 ***
## lakenamePaul Lake        -4.0856     0.5935  -6.884 6.20e-12 ***
## lakenamePeter Lake       -4.5938     0.5926  -7.752 9.99e-15 ***
## lakenameTuesday Lake    -6.1413     0.6036 -10.174 < 2e-16 ***
## lakenameWard Lake        -2.5904     0.8139  -3.183 0.00146 **
## lakenameWest Long Lake   -5.3559     0.6124  -8.746 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6.382 on 9494 degrees of freedom
## Multiple R-squared:  0.03014,    Adjusted R-squared:  0.02933
## F-statistic: 36.88 on 8 and 9494 DF,  p-value: < 2.2e-16
```

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

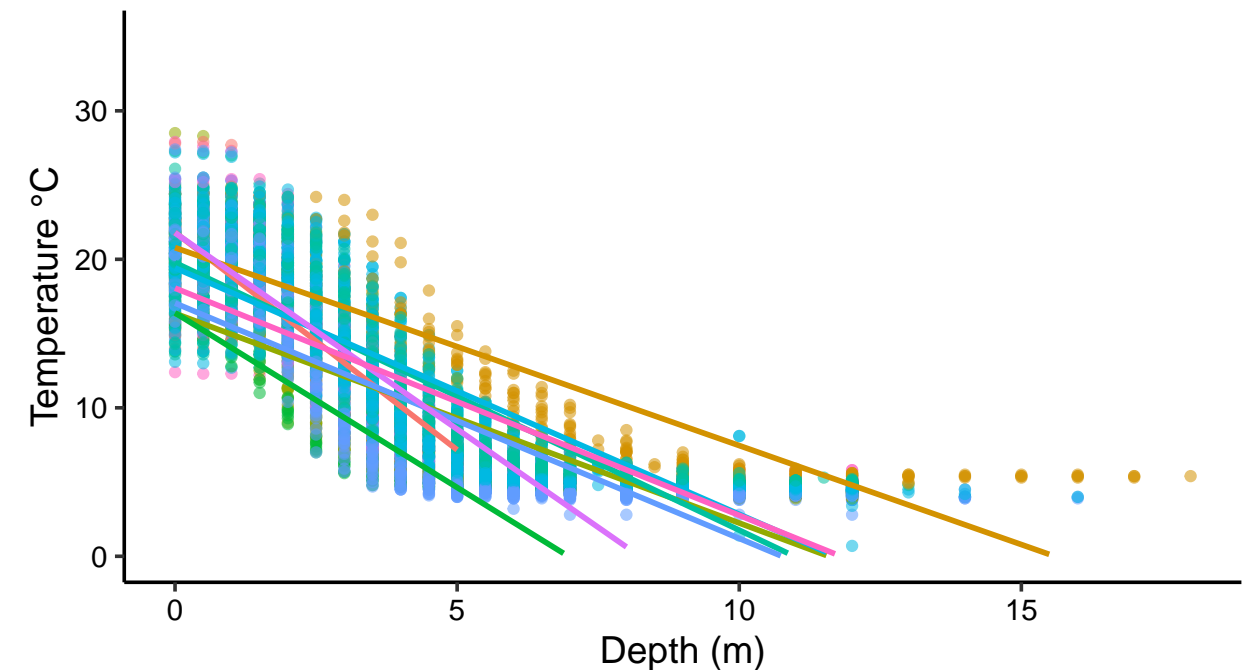
Answer: Yes there is a significant difference in mean temperature among the lakes with a p-value of $< 2.2e-16$.

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.
TempbyDepthColors <- ggplot(LakeChemistry.July, aes(x = depth, y = temperature_C, color = lakename, alpha = 0.5)) +
  geom_point() +
  ylim(0,35) +
  ylab("Temperature \u00B0C") +
  xlab("Depth (m)") +
  geom_smooth(method = lm, se = FALSE)
print(TempbyDepthColors)
```

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

```
## Warning: Removed 88 rows containing missing values (geom_smooth).
```



Central Long Lake East Long Lake Paul Lake Tuesday Lake West Long Lake
Hummingbird Lake Peter Lake Ward Lake

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

#15

TukeyHSD(July.anova.aov)

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = LakeChemistry.July)
##
## $lakename
##
```

	diff	lwr	upr	p adj
Crampton Lake-Central Long Lake	-2.8291387	-4.97533050	-0.68294692	0.0014395
East Long Lake-Central Long Lake	-6.5010074	-8.41491151	-4.58710334	0.0000000
Hummingbird Lake-Central Long Lake	-6.7826598	-9.39740915	-4.16791044	0.0000000
Paul Lake-Central Long Lake	-4.0856227	-5.92698164	-2.24426374	0.0000000
Peter Lake-Central Long Lake	-4.5938464	-6.43239794	-2.75529489	0.0000000
Tuesday Lake-Central Long Lake	-6.1412824	-8.01392896	-4.26863584	0.0000000
Ward Lake-Central Long Lake	-2.5903736	-5.11555124	-0.06519587	0.0392367
West Long Lake-Central Long Lake	-5.3558591	-7.25566616	-3.45605194	0.0000000
East Long Lake-Crampton Lake	-3.6718687	-5.00938971	-2.33434772	0.0000000
Hummingbird Lake-Crampton Lake	-3.9535211	-6.18126607	-1.72577609	0.0000013
Paul Lake-Crampton Lake	-1.2564840	-2.48796128	-0.02500667	0.0413505
Peter Lake-Crampton Lake	-1.7647077	-2.99198326	-0.53743215	0.0002831
Tuesday Lake-Crampton Lake	-3.3121437	-4.58993033	-2.03435704	0.0000000
Ward Lake-Crampton Lake	0.2387652	-1.88313397	2.36066428	0.9999938
West Long Lake-Crampton Lake	-2.5267203	-3.84399049	-1.20945020	0.0000001
Hummingbird Lake-East Long Lake	-0.2816524	-2.28658065	1.72327592	0.9999656
Paul Lake-East Long Lake	2.4153847	1.65813576	3.17263372	0.0000000
Peter Lake-East Long Lake	1.9071610	1.15676448	2.65755754	0.0000000
Tuesday Lake-East Long Lake	0.3597250	-0.47071364	1.19016370	0.9180513
Ward Lake-East Long Lake	3.9106339	2.02401108	5.79725666	0.0000000
West Long Lake-East Long Lake	1.1451484	0.25515383	2.03514292	0.0021540
Paul Lake-Hummingbird Lake	2.6970371	0.76123976	4.63283444	0.0005285
Peter Lake-Hummingbird Lake	2.1888134	0.25568630	4.12194045	0.0132248
Tuesday Lake-Hummingbird Lake	0.6413774	-1.32420489	2.60695968	0.9848376
Ward Lake-Hummingbird Lake	4.1922862	1.59743925	6.78713323	0.0000193
West Long Lake-Hummingbird Lake	1.4268007	-0.56467500	3.41827648	0.3907140
Peter Lake-Paul Lake	-0.5082237	-1.04736090	0.03091344	0.0830294
Tuesday Lake-Paul Lake	-2.0556597	-2.70157171	-1.40974770	0.0000000
Ward Lake-Paul Lake	1.4952491	-0.31773721	3.30823548	0.2044231
West Long Lake-Paul Lake	-1.2702364	-1.99111377	-0.54935896	0.0000017
Tuesday Lake-Peter Lake	-1.5474360	-2.18530058	-0.90957138	0.0000000
Ward Lake-Peter Lake	2.0034729	0.19333794	3.81360778	0.0173703
West Long Lake-Peter Lake	-0.7620126	-1.47568845	-0.04833682	0.0259997
Ward Lake-Tuesday Lake	3.5509088	1.70615361	5.39566407	0.0000001
West Long Lake-Tuesday Lake	0.7854233	-0.01198909	1.58283578	0.0573522
West Long Lake-Ward Lake	-2.7654855	-4.63780592	-0.89316507	0.0001611

#check how any unique lake names

unique(LakeChemistry.July[c("lakename")])

lakename


```
## 1          Paul Lake
## 15         Peter Lake
## 31         Tuesday Lake
## 1360       East Long Lake
## 1377       West Long Lake
## 2376       Central Long Lake
## 4982       Hummingbird Lake
## 5165       Crampton Lake
## 7797       Ward Lake
```

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 as Peter Lake statistically speaking. Central Lake has 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: We could use a t-test if we were just looking at Peter and Paul Lake.