

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

GuruBanda Khalsa

OVERVIEW

This exercise accompanies the lessons in Environmental Data Analytics on generalized linear models.

Directions

1. Rename this file <FirstLast>_A06_GLMs.Rmd (replacing <FirstLast> with your first and last name).
2. Change “Student Name” on line 3 (above) with your name.
3. Work through the steps, **creating code and output** that fulfill each instruction.
4. Be sure to **answer the questions** in this assignment document.
5. When you have completed the assignment, **Knit** the text and code into a single PDF file.

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. I checked my working directory using the getwd() function, loaded the
# `tidyverse` and `agricolae` packages, and imported the raw NTL-LTER raw data
# file for chemistry/physics. I also set date columns to date objects.
getwd()
```

```
## [1] "/Users/survivormangb/Desktop/Masters at Duke/Second Year/Fall Semester/Environmental Data Analy
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## -- Attaching packages ----- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble   3.1.8     v dplyr    1.0.10
## v tidyr    1.2.1     v stringr  1.4.1
## v readr    2.1.2     vforcats  0.5.2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

```
## Warning: package 'tibble' was built under R version 4.1.2
```

```

## Warning: package 'tidyverse' was built under R version 4.1.2

## Warning: package 'readr' was built under R version 4.1.2

## Warning: package 'dplyr' was built under R version 4.1.2

## Warning: package 'stringr' was built under R version 4.1.2

## Warning: package 'forcats' was built under R version 4.1.2

## -- Conflicts ----- tidyverse_conflicts() --
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()      masks base::date()
## x dplyr::filter()        masks stats::filter()
## x lubridate::intersect() masks base::intersect()
## x dplyr::lag()          masks stats::lag()
## x lubridate::setdiff()   masks base::setdiff()
## x lubridate::union()    masks base::union()

library(agricolae)
NTL1 <- read.csv("~/Desktop/Masters at Duke/Second Year/Fall Semester/Environmental Data Analytics/EDA-1/View(NTL1)
NTL1$sampleddate <- as.Date(NTL1$sampleddate, format = "%m/%d/%y")

# 2. I assigned a pre-built theme from Ggplot as a common theme across all
# plots in my analysis session.
mytheme <- theme_classic(base_size = 14) + theme(axis.text = element_text(color = "black"),
  legend.position = "top")

```

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 has an effect on depth across all lakes. H_a: Mean lake temperature recorded during July has no effect on 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.

```

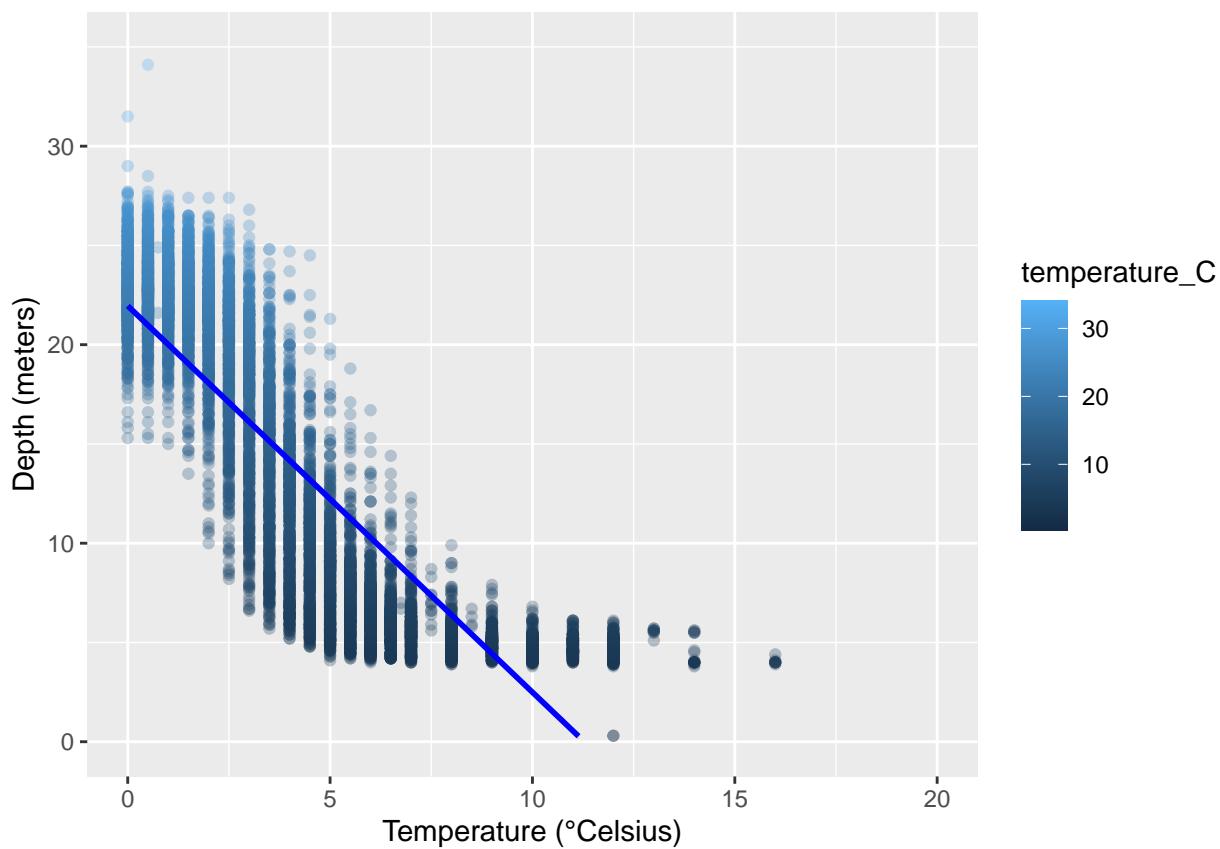
# 4. I wrangled the NTL-LTER dataset with a pipe function so that the records
# follow the criteria listed above.
NTL1.wrangle <- NTL1 %>%
  mutate(month = month(sampledate), year = year(sampledate)) %>%
  filter(month == "7") %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na()

# 5. I visualized the relationship among the two continuous variables with a
# scatter plot of temperature by depth with a smoothed line showing the linear
# model. I also limited values from 0 to 35 °C.
NTL1.wrangle.plot <- ggplot(NTL1.wrangle, aes(x = depth, y = temperature_C)) + geom_point(aes(color = temperature_C,
  alpha = 0.3)) + geom_smooth(method = "lm", col = "blue") + xlim(0, 20) + ylim(0,
  35) + xlab("Temperature (°Celsius)") + ylab("Depth (meters)")
print(NTL1.wrangle.plot)

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

## Warning: Removed 24 rows containing missing values (geom_smooth).

```



- 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: The figure suggests that temperature decreases with depth. The distribution of the points suggests that the relationship is most linear in the temperature range of 4 to 8°C. The trend levels out once 8°C is approached.

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

```
# 7. I performed a linear regression to test the relationship and displayed the
# results. The p-value is less than 2.2e-16, or in other words, it is
# statistically significant.
NTL1.wrangle.regression <- lm(data = NTL1.wrangle, temperature_C ~ depth)
summary(NTL1.wrangle.regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL1.wrangle)
##
## 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: 73.9% of the variability in temperature is explained by changes in depth. This finding is based on the 9,726 degrees of freedom. The result shows a statistical significance of less than 2.2e-16. This means that we reject our null hypothesis that mean lake temperature recorded during July has no effect on depth across all lakes. This means that depth does have an effect on mean temperature of lakes. For every 1 meter change in depth, temperature is predicted to decrease by 1.95 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. I ran an AIC to determine what set of explanatory variables is best suited
# to predict temperature.
NTL1.wrangle.ALLregression <- lm(data = NTL1.wrangle, temperature_C ~ depth + year4 +
  daynum)
summary(NTL1.wrangle.ALLregression)

```

```

##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL1.wrangle)
##
## 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

```

```
step(NTL1.wrangle.ALLregression)
```

```

## 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 = NTL1.wrangle)
##
## Coefficients:
## (Intercept)      depth      year4      daynum
## -8.57556      -1.94644     0.01134     0.03978

```

10. I ran a multiple regression on the recommended set of variables. I
removed year4, since it would improve the AIC

```

NTL1.wrangle.BESTregression <- lm(data = NTL1.wrangle, temperature_C ~ depth + daynum)
summary(NTL1.wrangle.BESTregression)

```

```
##
```

```

## Call:
## lm(formula = temperature_C ~ depth + daynum, data = NTL1.wrangle)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -9.6174 -2.9809  0.0845  2.9681 13.4406 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 14.088588  0.855505 16.468 <2e-16 ***
## depth       -1.946111  0.011685 -166.541 <2e-16 ***
## daynum      0.039836  0.004318   9.225 <2e-16 ***  
## ---      
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.818 on 9725 degrees of freedom
## Multiple R-squared:  0.741, Adjusted R-squared:  0.741 
## F-statistic: 1.391e+04 on 2 and 9725 DF, p-value: < 2.2e-16
```

```
step(NTL1.wrangle.BESTregression)
```

```

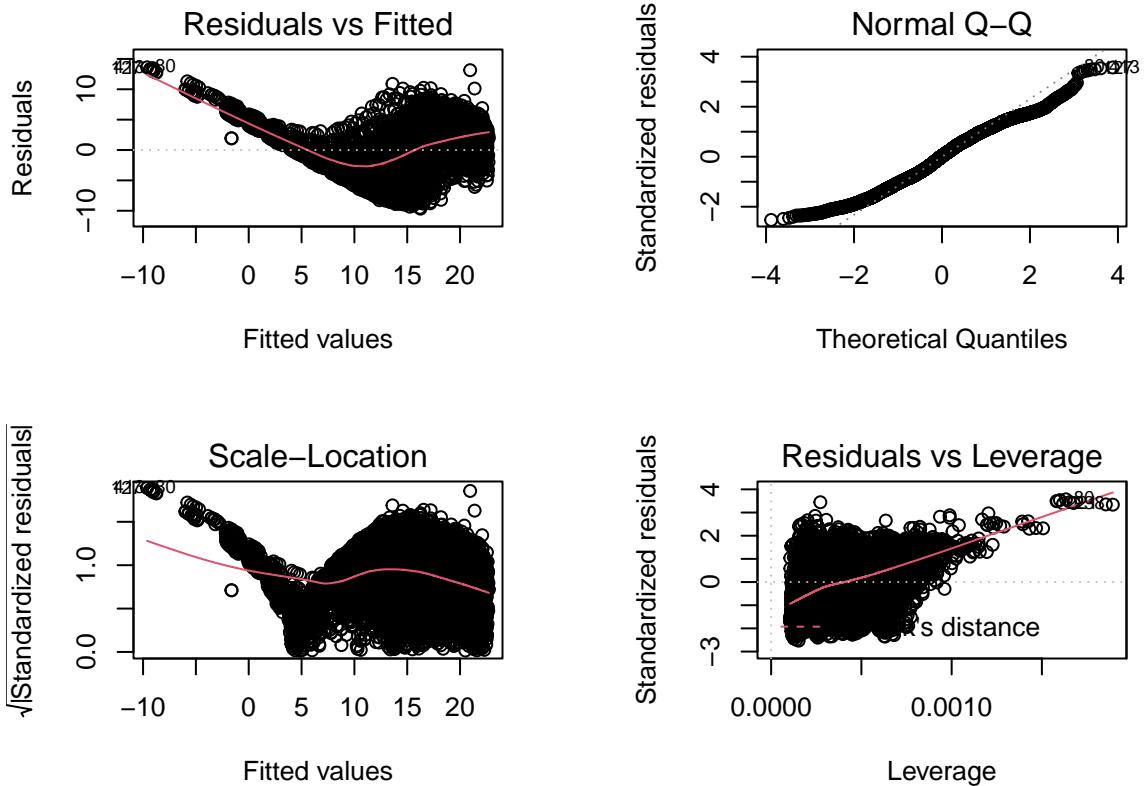
## Start: AIC=26070.49
## temperature_C ~ depth + daynum
##
##          Df Sum of Sq    RSS   AIC
## <none>           141788 26070
## - daynum     1      1241 143029 26153
## - depth      1     404384 546173 39188
```

```

##
## Call:
## lm(formula = temperature_C ~ depth + daynum, data = NTL1.wrangle)
##
## Coefficients:
## (Intercept)      depth      daynum
## 14.08859     -1.94611     0.03984
```

```

par(mfrow = c(2, 2), mar = c(4, 4, 4, 4))
plot(NTL1.wrangle.ALLregression)
```



```
par(mfrow = c(1, 1))
summary(NTL1.wrangle.ALLregression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL1.wrangle)
##
## 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 in our multiple regression includes only the relationship between temperature and depth as well as “daynum.” The observed variance is 74.1% of the variance in temperature. This is an improvement from 73.9% when only depth was included.

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. I created an ANOVA model and a linear model to see whether the different
# lakes have, on average, different temperatures in the month of July.
```

```
# ANOVA Model
NTL1.wrangle.anova <- aov(data = NTL1.wrangle, temperature_C ~ lakename)
summary(NTL1.wrangle.anova)
```

```
##           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
```

```
par(mfrow = c(2, 2), mar = c(4, 4, 4, 4))
plot(NTL1.wrangle.anova)
par(mfrow = c(1, 1))
```

```
# Linear Model
NTL1.wrangle.anova2 <- lm(data = NTL1.wrangle, temperature_C ~ lakename)
summary(NTL1.wrangle.anova2)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL1.wrangle)
##
## 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 ***
```

```

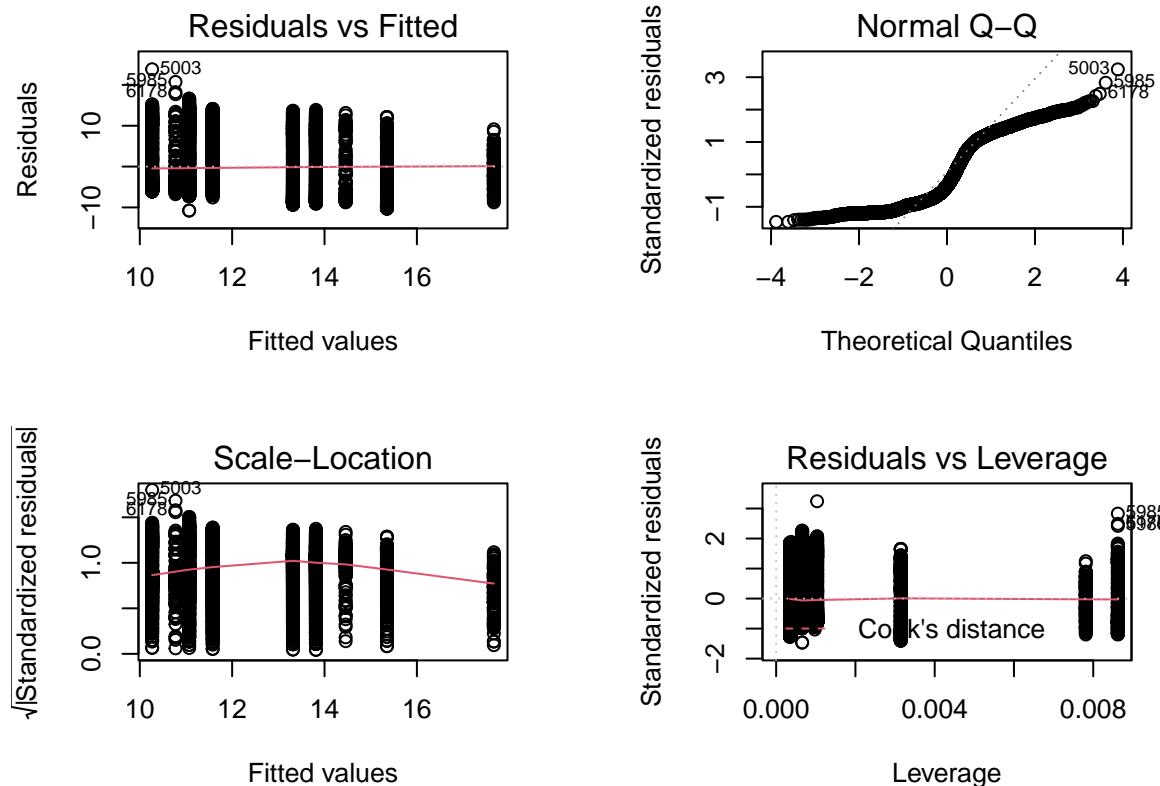
## lakenamePaul Lake      -3.8522    0.6656   -5.788 7.36e-09 ***
## lakenamePeter Lake     -4.3501    0.6645   -6.547 6.17e-11 ***
## lakenameTuesday Lake   -6.5972    0.6769   -9.746 < 2e-16 ***
## lakenameWard Lake      -3.2078    0.9429   -3.402 0.000672 ***
## lakenameWest Long Lake -6.0878    0.6895   -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

```

```

par(mfrow = c(2, 2), mar = c(4, 4, 4, 4))
plot(NTL1.wrangle.anova2)

```



```

par(mfrow = c(1, 1))

```

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

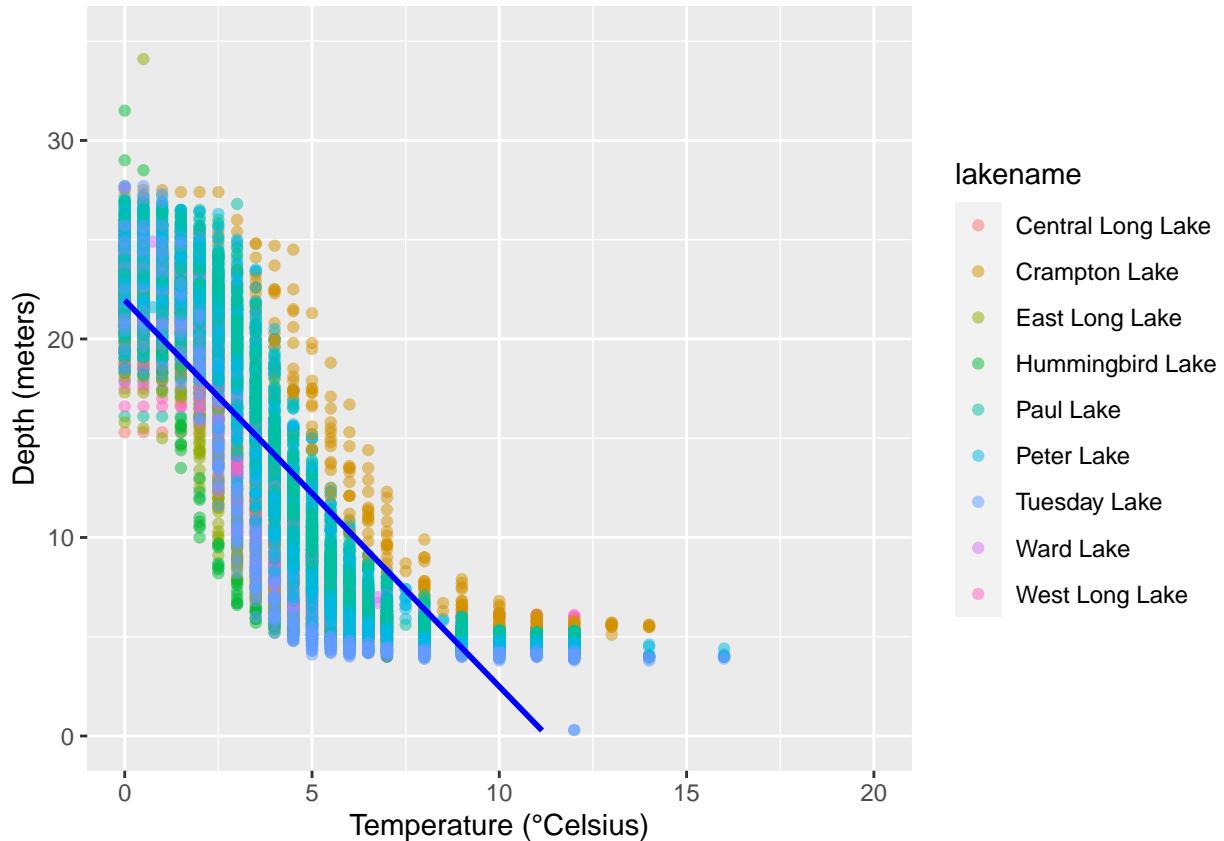
Answer: There is not a significant difference in mean temperature among the lakes. The ANOVA model found that the difference between the pair of group means is statistically significant, so we reject the null hypothesis. The p-value for our AOC model is less than 0.05, so we reject our null hypothesis that both lakes have the same mean temperature.

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. I created a graph that depicts temperature by depth, with a separate
# color for each lake with a geom_smooth for each lake.
NTL1.wrangle.plot <- ggplot(NTL1.wrangle, aes(x = depth, y = temperature_C)) + geom_point(aes(color = lakename),
alpha = 0.5) + geom_smooth(method = "lm", se = FALSE, col = "blue") + xlim(0,
20) + ylim(0, 35) + xlab("Temperature (°Celsius)") + ylab("Depth (meters)")
print(NTL1.wrangle.plot)

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

## Warning: Removed 24 rows containing missing values (geom_smooth).
```



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

```
# 15. I used the Tukey's HSD test to determine which lakes have different
# means.
TukeyHSD(NTL1.wrangle.anova)
```

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
```

```

## 
## Fit: aov(formula = temperature_C ~ lakename, data = NTL1.wrangle)
##
## $lakename
##          diff      lwr      upr    p adj
## Crampton Lake-Central Long Lake -2.3145195 -4.7031913  0.0741524 0.0661566
## East Long Lake-Central Long Lake -7.3987410 -9.5449411 -5.2525408 0.0000000
## Hummingbird Lake-Central Long Lake -6.8931304 -9.8184178 -3.9678430 0.0000000
## Paul Lake-Central Long Lake -3.8521506 -5.9170942 -1.7872070 0.0000003
## Peter Lake-Central Long Lake -4.3501458 -6.4115874 -2.2887042 0.0000000
## Tuesday Lake-Central Long Lake -6.5971805 -8.6971605 -4.4972005 0.0000000
## Ward Lake-Central Long Lake -3.2077856 -6.1330730 -0.2824982 0.0193405
## West Long Lake-Central Long Lake -6.0877513 -8.2268550 -3.9486475 0.0000000
## East Long Lake-Crampton Lake -5.0842215 -6.5591700 -3.6092730 0.0000000
## Hummingbird Lake-Crampton Lake -4.5786109 -7.0538088 -2.1034131 0.0000004
## Paul Lake-Crampton Lake -1.5376312 -2.8916215 -0.1836408 0.0127491
## Peter Lake-Crampton Lake -2.0356263 -3.3842699 -0.6869828 0.0000999
## Tuesday Lake-Crampton Lake -4.2826611 -5.6895065 -2.8758157 0.0000000
## Ward Lake-Crampton Lake -0.8932661 -3.3684639  1.5819317 0.9714459
## West Long Lake-Crampton Lake -3.7732318 -5.2378351 -2.3086285 0.0000000
## Hummingbird Lake-East Long Lake 0.5056106 -1.7364925  2.7477137 0.9988050
## Paul Lake-East Long Lake 3.5465903  2.6900206  4.4031601 0.0000000
## Peter Lake-East Long Lake 3.0485952  2.2005025  3.8966879 0.0000000
## Tuesday Lake-East Long Lake 0.8015604 -0.1363286  1.7394495 0.1657485
## Ward Lake-East Long Lake 4.1909554  1.9488523  6.4330585 0.0000002
## West Long Lake-East Long Lake 1.3109897  0.2885003  2.3334791 0.0022805
## Paul Lake-Hummingbird Lake 3.0409798  0.8765299  5.2054296 0.0004495
## Peter Lake-Hummingbird Lake 2.5429846  0.3818755  4.7040937 0.0080666
## Tuesday Lake-Hummingbird Lake 0.2959499 -1.9019508  2.4938505 0.9999752
## Ward Lake-Hummingbird Lake 3.6853448  0.6889874  6.6817022 0.0043297
## West Long Lake-Hummingbird Lake 0.8053791 -1.4299320  3.0406903 0.9717297
## Peter Lake-Paul Lake -0.4979952 -1.1120620  0.1160717 0.2241586
## Tuesday Lake-Paul Lake -2.7450299 -3.4781416 -2.0119182 0.0000000
## Ward Lake-Paul Lake 0.6443651 -1.5200848  2.8088149 0.9916978
## West Long Lake-Paul Lake -2.2356007 -3.0742314 -1.3969699 0.0000000
## Tuesday Lake-Peter Lake -2.2470347 -2.9702236 -1.5238458 0.0000000
## Ward Lake-Peter Lake 1.1423602 -1.0187489  3.3034693 0.7827037
## West Long Lake-Peter Lake -1.7376055 -2.5675759 -0.9076350 0.0000000
## Ward Lake-Tuesday Lake 3.3893950  1.1914943  5.5872956 0.0000609
## West Long Lake-Tuesday Lake 0.5094292 -0.4121051  1.4309636 0.7374387
## West Long Lake-Ward Lake -2.8799657 -5.1152769 -0.6446546 0.0021080

```

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: From the findings above, Peter Lake and Ward Lake have the same mean temperature, statistically speaking, as Peter Lake. The p-values given from the Tukey HSD test are all above 0.05 for their relationship with Peter Lake. Central Long Lake 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: If we were just looking at Peter Lake and Paul Lake, we might explore the two-way Bartlett to see whether they have distinct mean temperatures.

18. Wrangle the July data to include only records for Crampton Lake and Ward Lake. Run the two-sample T-test on these data to determine whether their July temperature are same or different. What does the test say? Are the mean temperatures for the lakes equal? Does that match you answer for part 16?

```
# 18. I wrangled the July data to include only records for Crampton Lake and
# Ward Lake.
NTL1.wrangle2 <- NTL1.wrangle %>%
  filter(lakename == "Crampton Lake" | lakename == "Ward Lake")

View(NTL1.wrangle2)

# I then ran the two-sample T-test.
t.test(data = NTL1.wrangle2, temperature_C ~ lakename)

##
##  Welch Two Sample t-test
##
## data: temperature_C by lakename
## t = 1.1181, df = 200.37, p-value = 0.2649
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is
## 95 percent confidence interval:
## -0.6821129  2.4686451
## sample estimates:
## mean in group Crampton Lake      mean in group Ward Lake
##           15.35189                  14.45862
```

Answer: The test says that the p-value is not significant (>0.05), therefore the July temperatures of Crampton Lake and Ward Lake are different. We reject our null hypothesis that they are the same. When looking back at the Tukey HSD test, the relationship between Crampton Lake and Ward Lake has a p-value of 0.97, therefore we rejected the null hypothesis that the two mean temperatures are the same. We determine that the mean temperatures in this t-test are the same means reported in the Tukey HSD test from part 16.