Assignment 7: GLMs (Linear Regressios, ANOVA, & t-tests)

Max Hermanson

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 Tuesday, March 2 at 1:00 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
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.5     v dplyr   1.0.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0
```

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

```
#install.packages("agricolae")
library(agricolae)
library(ggridges)
library(lubridate)
```

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

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

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

#2
maxstheme2 <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(maxstheme2)
```

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 does not change with depth across lakes Ha: mean lake temperature changes with depth across different 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
lakechem.df$sampledate <- as.Date(lakechem.df$sampledate, format = "%m/%d/%Y")

#lakechem.df2 <- mutate(lakechem.df, month=month(sampledate))

lakechem.filtered <- lakechem.df %>%
  mutate(lakechem.df, month=month(sampledate)) %>%
  filter(month == 7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()

unique(lakechem.filtered$month)   #verify only July was filtered for
```

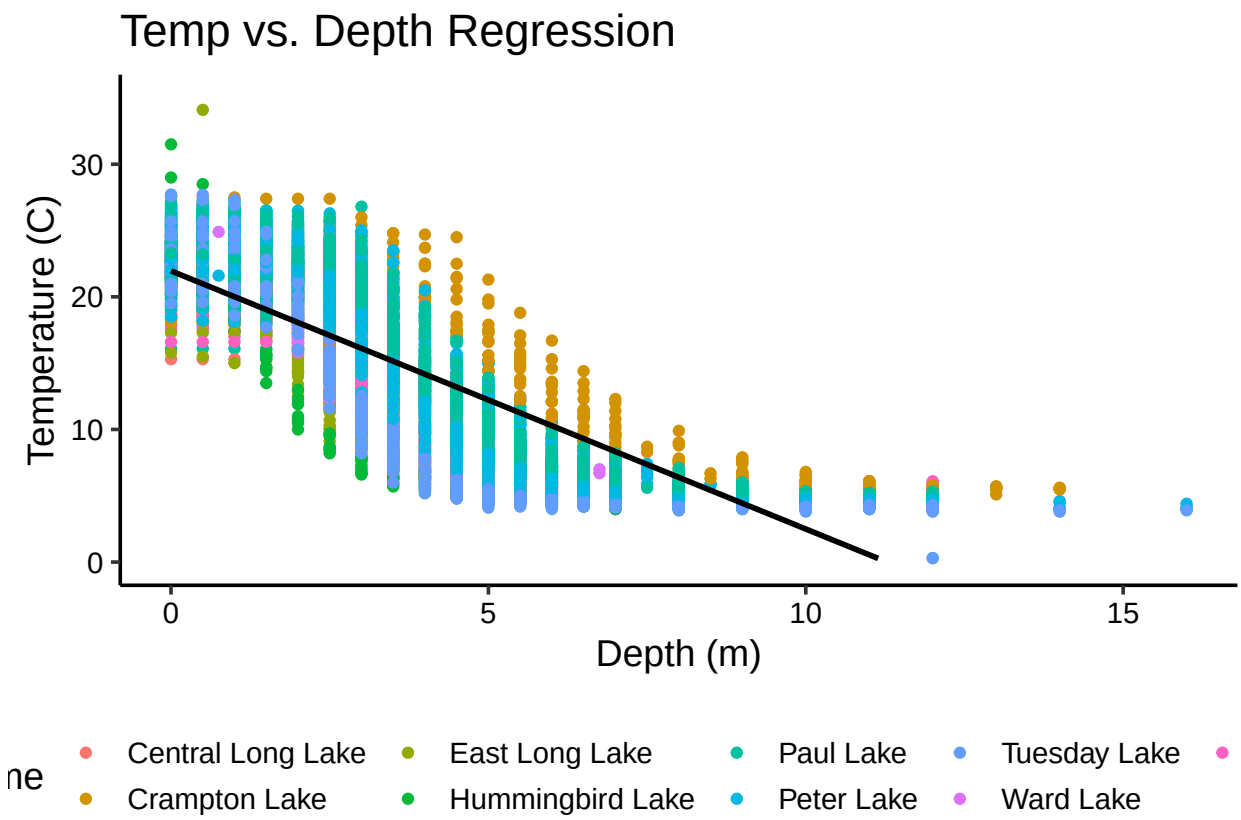
```
## NULL
```

```
regress_lakes <- ggplot(lakechem.filtered, aes(x=depth, y= temperature_C))+
  geom_point(aes(color = lakename))+
  xlab("Depth (m)")+
  ylab("Temperature (C)")+
  ylim(0,35)+
  geom_smooth(method=lm, color = "black")+
  theme(legend.position = "bottom")+
  ggtitle("Temp vs. Depth Regression")

print(regress_lakes)
```

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

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




```
#5
```

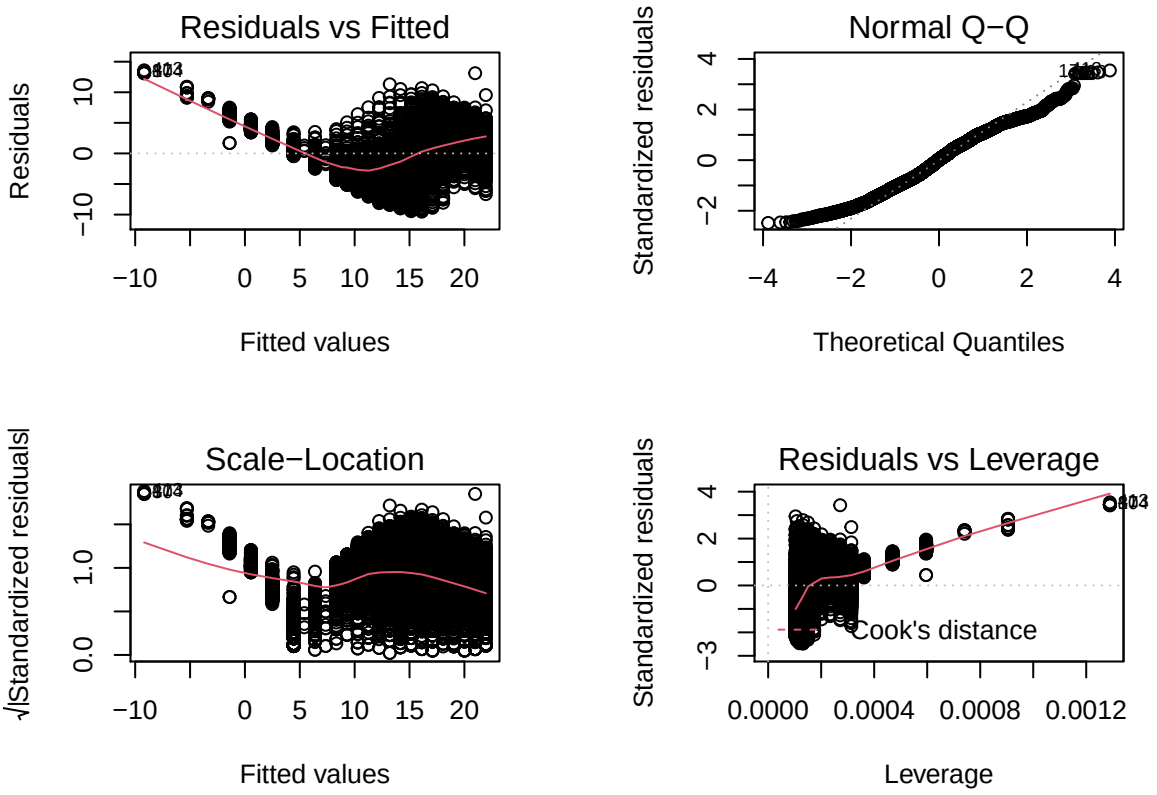
6. Interpret the figure. What does it suggest with regards to the response of temperature to depth? Do the distribution of points suggest about anything about the linearity of this trend?

   Answer: There is a clear moderate negative correlation between temperature and depth. The distribution of points indicate this is a non-linear model, based on their curviture. An inverse logistic function would be a more appropriate estimation of their relationship, and the model's equation should be modified to account for this.

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

```
#7
regression_1 <- lm(lakechem.filtered$temperature_C ~ lakechem.filtered$depth)

par(mfrow = c(2,2), mar=c(4,4,4,4)) #displays all 4 graphs at once
plot(regression_1)
```




```
par(mfrow = c(1,1)) # this cycles you through individual graphs

#depth_temp_reg <- lm(data = lakechem.filtered, temperature_C ~ depth)

summary(regression_1)
```

```
##
## Call:
## lm(formula = lakechem.filtered$temperature_C ~ lakechem.filtered$depth)
##
## 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 ***
## lakechem.filtered$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: R2 = 0.7387 (74% of the variance in temperature can be explained by depth); df = 9726;

p value of fstat < 0.05; for every 1m increase in depth, temperature falls by 1.94621 degrees C (p <0.05).

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
#create new filtered dataframe that includes all 12 explanatory variables.
#lakechem.filtered_2 <- lakechem.df %>%
#  mutate(lakechem.df, month=month(sampledate)) %>%
#  filter(month == 7) %>%
#  filter(lakeid %in% c("L", "R", "T", "E")) %>%
#  drop_na(dissolvedOxygen, irradianceWater, irradianceDeck, lakeid)
#filtered with only year4, daynum, depth
lakechem.filtered3 <- lakechem.df %>%
  mutate(lakechem.df, month=month(sampledate)) %>%
  filter(month == 7) %>%
  filter(lakeid %in% c("L", "R", "T", "E")) %>%    # filter for LTRE North Temperate Lakes
  drop_na( lakeid, depth, lakename, year4, temperature_C) %>%
  select(lakeid, lakename, year4, daynum, depth, temperature_C) %>%
  na.omit()

unique(lakechem.filtered3$lakeid)   #verification
```

```
## [1] L R T E
## Levels: C E H L M R T W Ward
```

```
unique(lakechem.df$lakeid)
```

```
## [1] L    R    T    E    W    C    H    M    Ward
## Levels: C E H L M R T W Ward
```

```
#AIC_lakevars <- lm(data = lakechem.filtered_2, temperature_C ~ year4 + daynum + depth)   #should i use
#step(AIC_lakevars)

#AIC_lakevars2 <- lm(data = lakechem.filtered_2, temperature_C ~ year4 + daynum + depth + dissolvedOxyg
#step(AIC_lakevars2)

AIC_lakevars3 <- lm(data = lakechem.filtered3, temperature_C ~ year4 + daynum + depth)
step(AIC_lakevars3)
```

```
## Start:  AIC=21196.5
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## - year4   1        26 112540 21196
## <none>                 112514 21196
## - daynum  1      1105 113619 21273
## - depth   1    345513 458027 32459
```

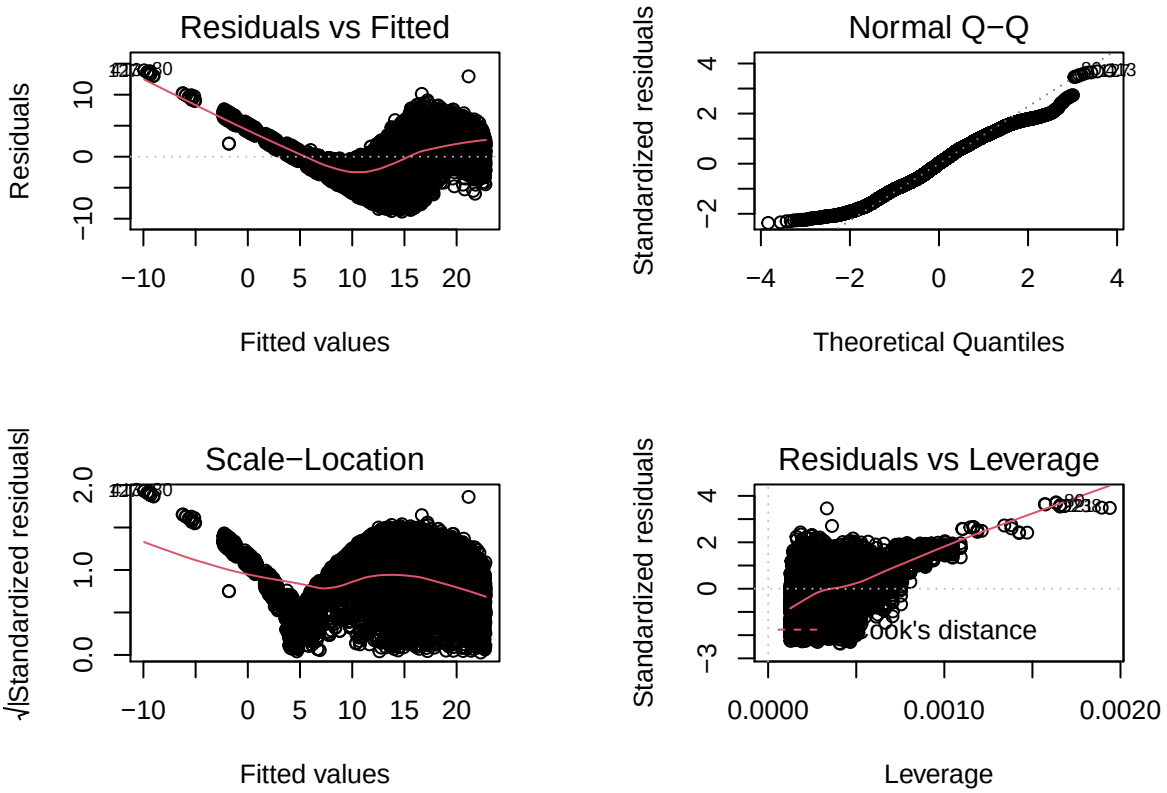
```
##
## Step:  AIC=21196.34
## temperature_C ~ daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                 112540 21196
## - daynum  1      1107 113647 21273
## - depth   1    345493 458032 32457
```

```
##
## Call:
## lm(formula = temperature_C ~ daynum + depth, data = lakechem.filtered3)
##
## Coefficients:
## (Intercept)       daynum        depth
##    14.01906      0.04121     -1.98869
```

```
#10
multiReg_lake <- lm(data = lakechem.filtered3, temperature_C ~ daynum + depth)
summary(multiReg_lake)
```

```
##
## Call:
## lm(formula = temperature_C ~ daynum + depth, data = lakechem.filtered3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8604 -2.8586  0.0484  2.8902 13.9281
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) 14.019059   0.919351   15.249   <2e-16 ***
## daynum       0.041214   0.004639    8.883   <2e-16 ***
## depth       -1.988689   0.012673 -156.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.746 on 8021 degrees of freedom
## Multiple R-squared:  0.7549, Adjusted R-squared:  0.7549
## F-statistic: 1.235e+04 on 2 and 8021 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2), mar=c(4,4,4,4)) #displays all 4 graphs at once
plot(multiReg_lake)
```




```
par(mfrow = c(1,1))
```

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: R2 = 0.755 (76% of the variance in temperature can be explained by the explanatory variables); df = 4564; p value of fstat <0.05; for every 1m increase in depth, temperature falls by 1.99 degrees celcius. The higher R2 value is an improvement from the model that just used Depth (original r2 = 74%); the fstat holds and all of the individual explanatory variable coefficients are significant (p<0.05).

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
#test normality of each different lake population's temperature
#use shaprio test,
#histograms
#bartlett test for equal variance

#summary stats
library(agricolae)
library(htmltools)
unique(lakechem.filtered3$lakename) # test
```

```
## [1] Paul Lake      Peter Lake     Tuesday Lake   East Long Lake
## 9 Levels: Central Long Lake Crampton Lake East Long Lake ... West Long Lake
```

```
laketemp.df <- lakechem.filtered3 %>%
  group_by(lakename) %>%
  summarise(avgtemp = mean(temperature_C))

summary(laketemp.df)
```

```
##               lakename    avgtemp
##  East Long Lake   :1   Min.   :10.27
##  Paul Lake        :1   1st Qu.:10.87
##  Peter Lake       :1   Median :12.19
##  Tuesday Lake     :1   Mean   :12.12
##  Central Long Lake:0   3rd Qu.:13.44
##  Crampton Lake    :0   Max.   :13.81
##  (Other)          :0
```

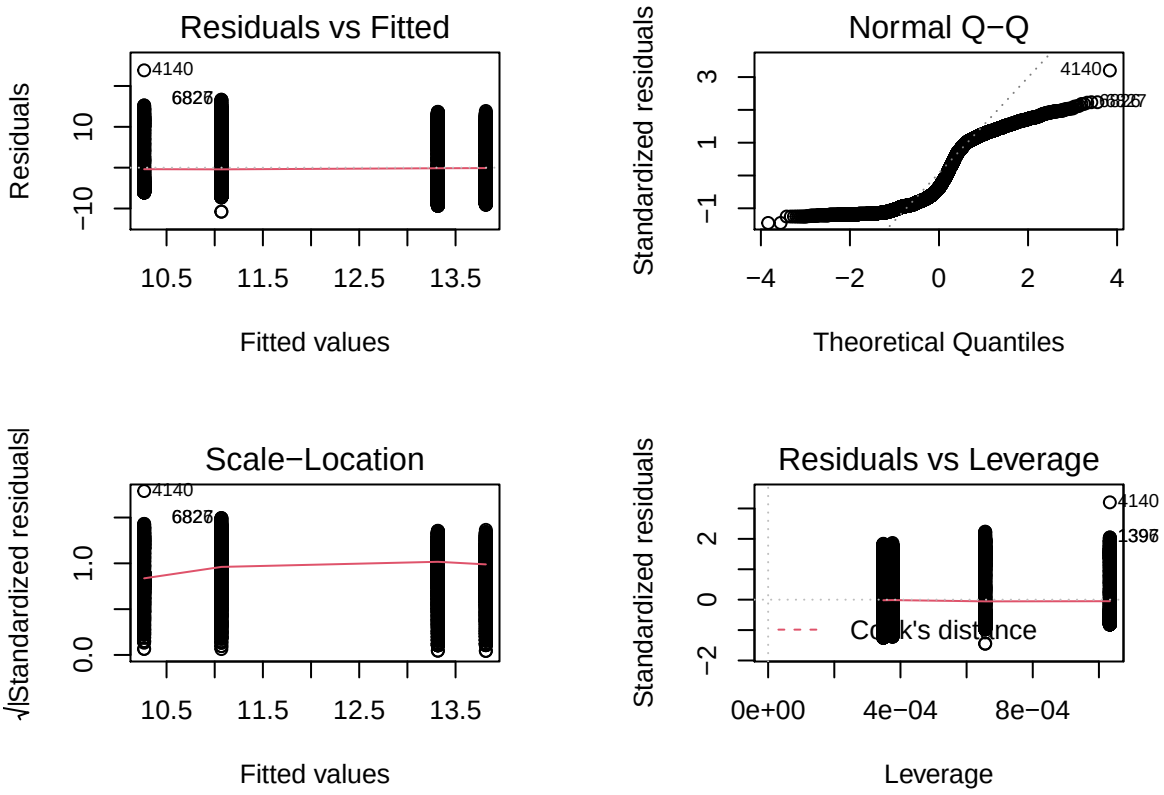
```
#ANOVA with aov
lake.anova <- aov(data = lakechem.filtered3, temperature_C ~ lakename)
summary(lake.anova)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       3  14172    4724   85.13 <2e-16 ***
## Residuals   8020 445012      55
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#lm
lake.anova.lm <- lm(formula = temperature_C ~ lakename, data = lakechem.filtered3)
summary(lake.anova.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = lakechem.filtered3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.769  -6.769  -2.768   7.884  23.832
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.2677     0.2394  42.885  < 2e-16 ***
## lakenamePaul Lake       3.5466     0.2796  12.684  < 2e-16 ***
## lakenamePeter Lake      3.0486     0.2768  11.012  < 2e-16 ***
## lakenameTuesday Lake    0.8016     0.3062   2.618  0.00886 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.449 on 8020 degrees of freedom
## Multiple R-squared:  0.03086,    Adjusted R-squared:  0.0305
## F-statistic: 85.13 on 3 and 8020 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2), mar=c(4,4,4,4)) #displays all 4 graphs at once
plot(lake.anova.lm)
```




```
par(mfrow = c(1,1))
```

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

    Answer: based on the p-value of the ANOVA test being less than 0.05, there is strong evidence to support there is a significant difference in means among the lakes. This also suggests variance within each subpopulation (each lake) may not be great.

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.

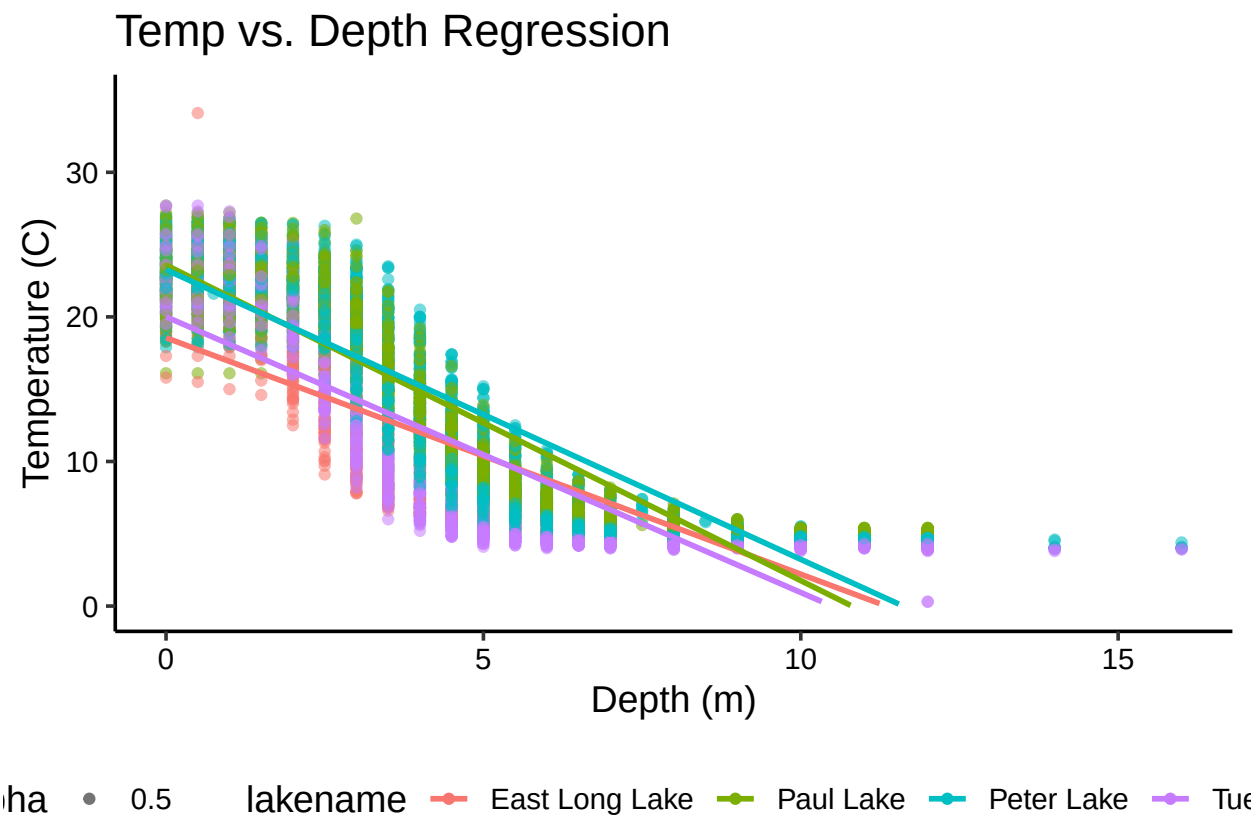
```
#14.
depth_graph <- ggplot(lakechem.filtered3, aes(x= depth, y=temperature_C , color = lakename))+
  geom_point(aes(alpha= 0.5))+
  geom_smooth(method = "lm", se = FALSE)+
  ylim(0,35)+
  xlab("Depth (m)")+
  ylab("Temperature (C)")+
  theme(legend.position = "bottom")+
  ggtitle("Temp vs. Depth Regression")



print(depth_graph)
```

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

```
## Warning: Removed 63 rows containing missing values (geom_smooth).
```



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

```
#15
TukeyHSD(lake.anova)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = lakechem.filtered3)
##
## $lakename
##                                   diff         lwr         upr     p adj
## Paul Lake-East Long Lake     3.5465903  2.82811054  4.26507016 0.0000000
## Peter Lake-East Long Lake    3.0485952  2.33722582  3.75996449 0.0000000
## Tuesday Lake-East Long Lake  0.8015604  0.01487106  1.58824981 0.0438965
## Peter Lake-Paul Lake        -0.4979952 -1.01306664  0.01707625 0.0624217
## Tuesday Lake-Paul Lake      -2.7450299 -3.35995465 -2.13010518 0.0000000
## Tuesday Lake-Peter Lake     -2.2470347 -2.85363635 -1.64043310 0.0000000
```

```
comparison_ABC <- HSD.test(lake.anova, "lakename", group=TRUE)
comparison_ABC
```

```
## $statistics
##    MSerror   Df     Mean       CV
##   55.48784 8020 12.68679 58.71473
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   4         3.633921  0.05
```

```
##
## $means
##                 temperature_C      std    r Min  Max   Q25  Q50    Q75
## East Long Lake      10.26767 6.766804  968 4.2 34.1 4.975  6.5 15.925
## Paul Lake           13.81426 7.296928 2660 4.7 27.7 6.500 12.4 21.400
## Peter Lake          13.31626 7.669758 2872 4.0 27.0 5.600 11.4 21.500
## Tuesday Lake        11.06923 7.698687 1524 0.3 27.7 4.400  6.8 19.400
##
## $comparison
## NULL
##
## $groups
##                 temperature_C groups
## Paul Lake           13.81426      a
## Peter Lake          13.31626      a
## Tuesday Lake        11.06923      b
## East Long Lake      10.26767      c
##
## 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 and Peter lakes have the same mean temperatures (diff = 0.49, p value > 0.05). Both Tuesday and East Long lakes have mean temperatures statistically distinct from all of the other lakes based on examination of the p-values of the Tukey test. Both of these lakes rejected the null in comparisons with the other three 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: an indepedent samples t-test or Welch's two-sample test could be used to compare the mean temperatures of these lakes.