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

Dwiti Bagadia

Spring 2023

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

library(tidyverse)
library(agricolae)
library(here)
library(cowplot)
library(ggplot2)
library(lubridate)

getwd()
```

```
## [1] "C:/Users/dwiti/OneDrive - University of North Carolina at Chapel Hill/EDA/EDA-Spring2023"
```

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

class(NTL_LTER$sampledate)
```

```
## [1] "factor"
```

```r
NTL_LTER$sampledate = ymd(NTL_LTER$sampledate)

#2 creating theme
library(ggthemes)
tweetheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "darkgrey"),
        axis.ticks = element_line(color = "darkgrey"),
        plot.background = element_rect(color = "white"),
        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: The mean lake temperature recorded during July doesn't change with depth across all lakes. Ha: The 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)

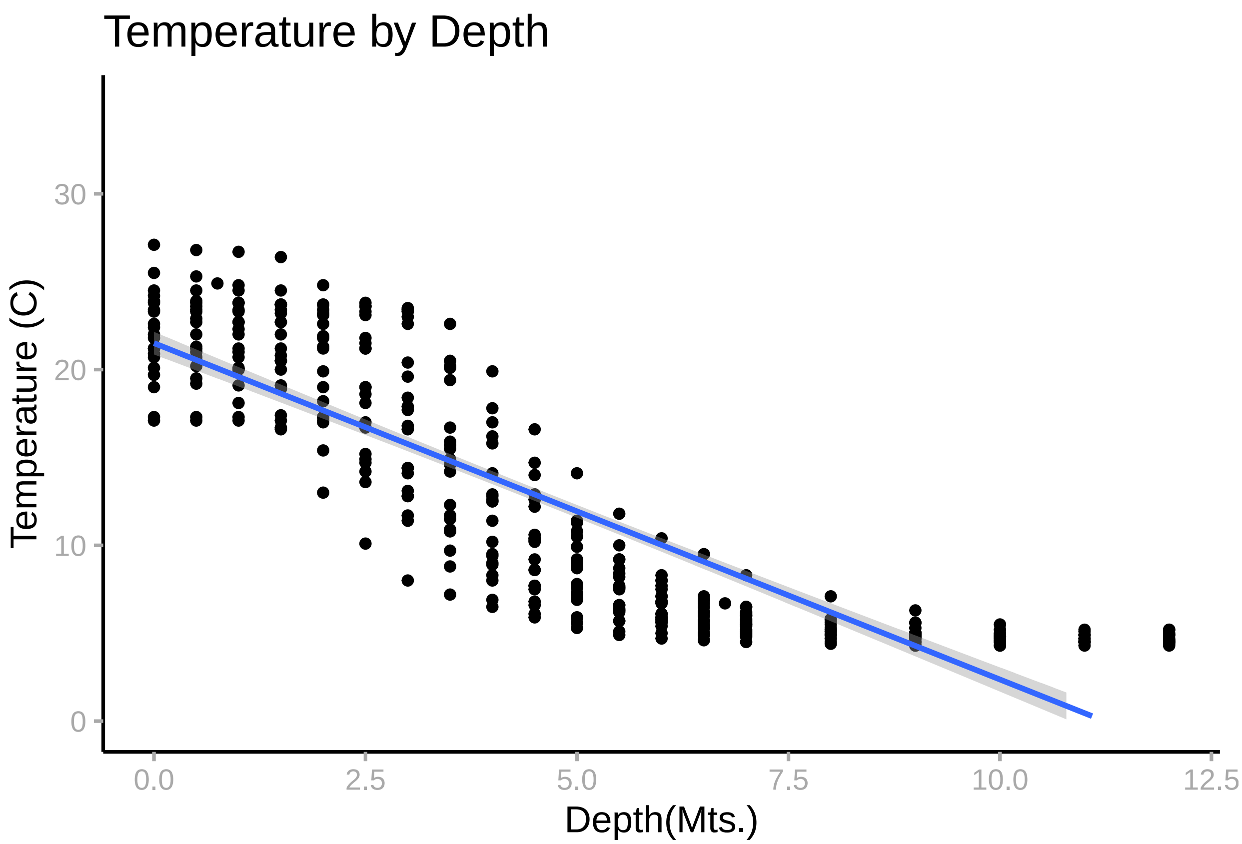
5. Visualize the relationship among the two continuous variables with a scatter plot of temperature by depth. Add a smoothed line showing the linear model, and limit temperature values from 0 to 35 °C. Make this plot look pretty and easy to read.

```r
#4 wrangling dataset
NTL_LTER.selected <- NTL_LTER %>%
  filter(month(sampledate) == 7) %>%
  select("lakename", "year4", "daynum", "depth", "temperature_C") %>%
  drop_na()

#5 visualizing with scatter plot
temperature.depthplot <- ggplot(NTL_LTER.selected) +
  geom_point(aes(x = depth, y = temperature_C)) +
  geom_smooth(aes(x = depth, y = temperature_C), method = "lm") +
  ylim(0, 35) +
  labs(title = "Temperature by Depth", x = "Depth(Mts.)", y = "Temperature (C)") +
  tweetheme
print(temperature.depthplot)
```

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

```
## Warning: Removed 6 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: The temperature is negatively correlated with depth. The points distribution of points suggests that the linearity of the data is negative.

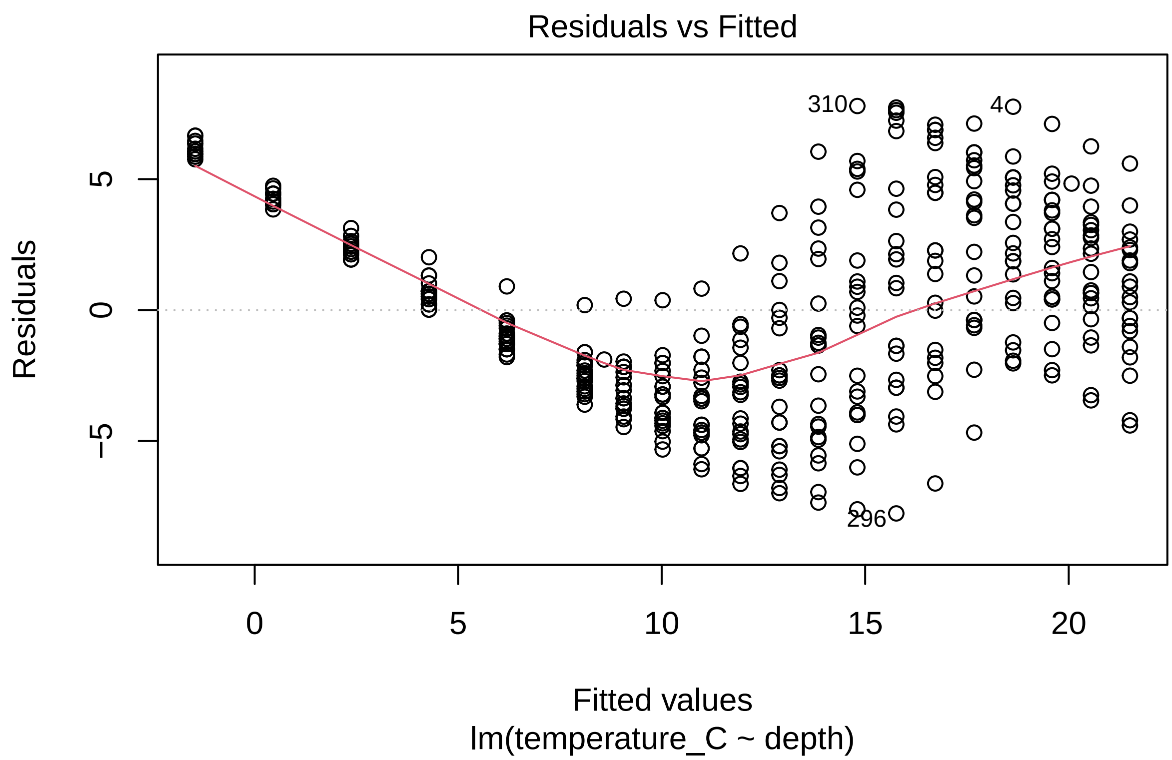
7. Perform a linear regression to test the relationship and display the results

```r
#7 regression models
temperature.depth.regreesion = lm(data = NTL_LTER.selected, temperature_C ~ depth)
summary(temperature.depth.regreesion)
```

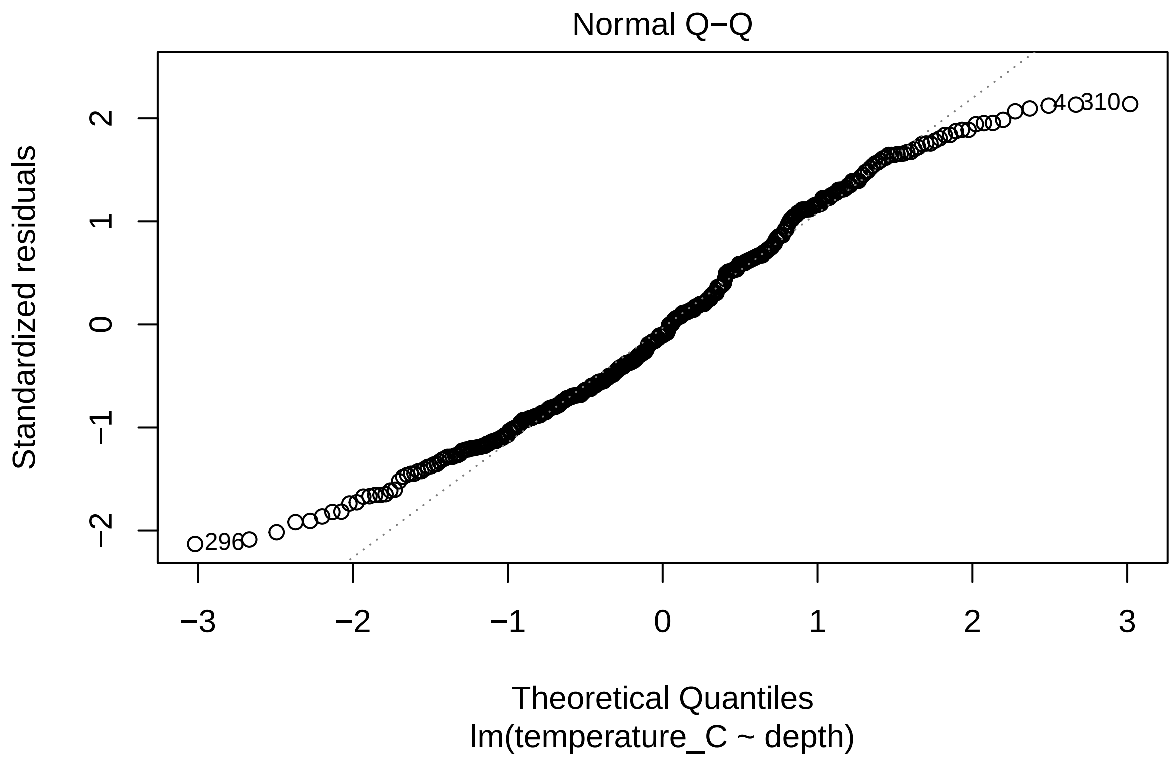
```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL_LTER.selected)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7641 -2.8586 -0.3779  2.6155  7.7928
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.5054     0.3261   65.95   <2e-16 ***
## depth        -1.9138     0.0567  -33.76   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.65 on 392 degrees of freedom
## Multiple R-squared:  0.744,  Adjusted R-squared:  0.7434
## F-statistic:  1139 on 1 and 392 DF,  p-value: < 2.2e-16
```

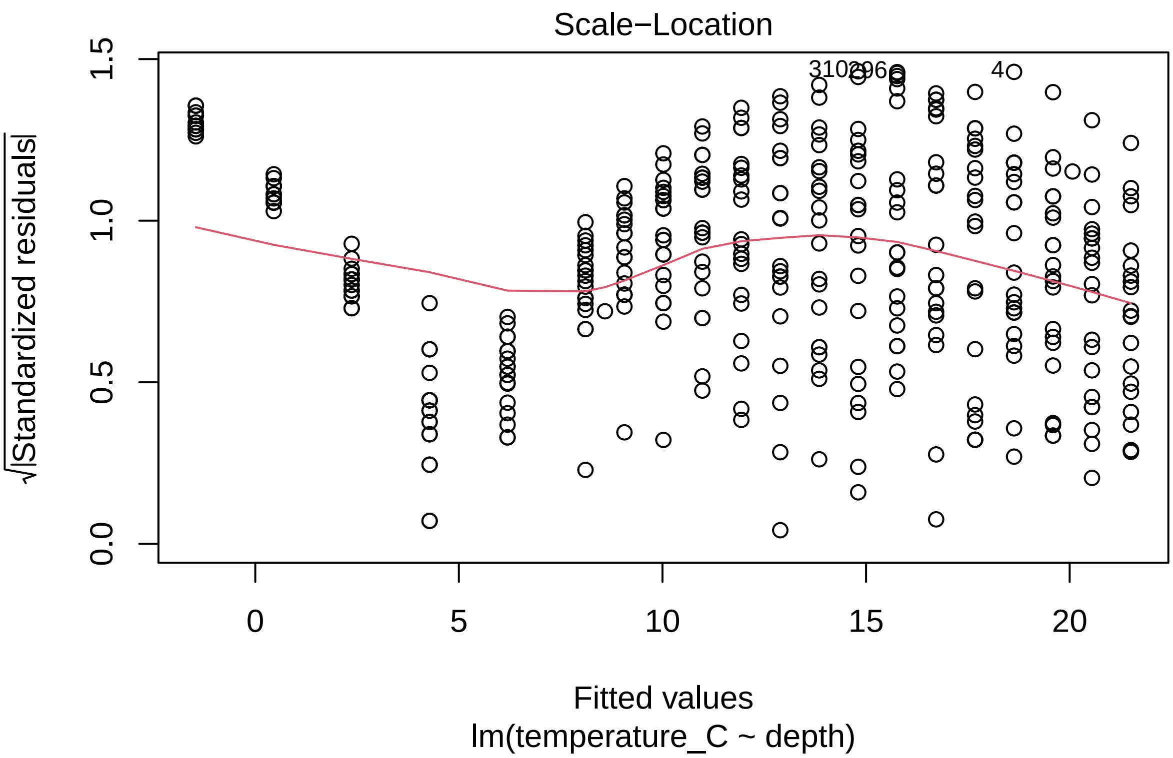
```
plot(temperature.depth.regreesion, 1)
```



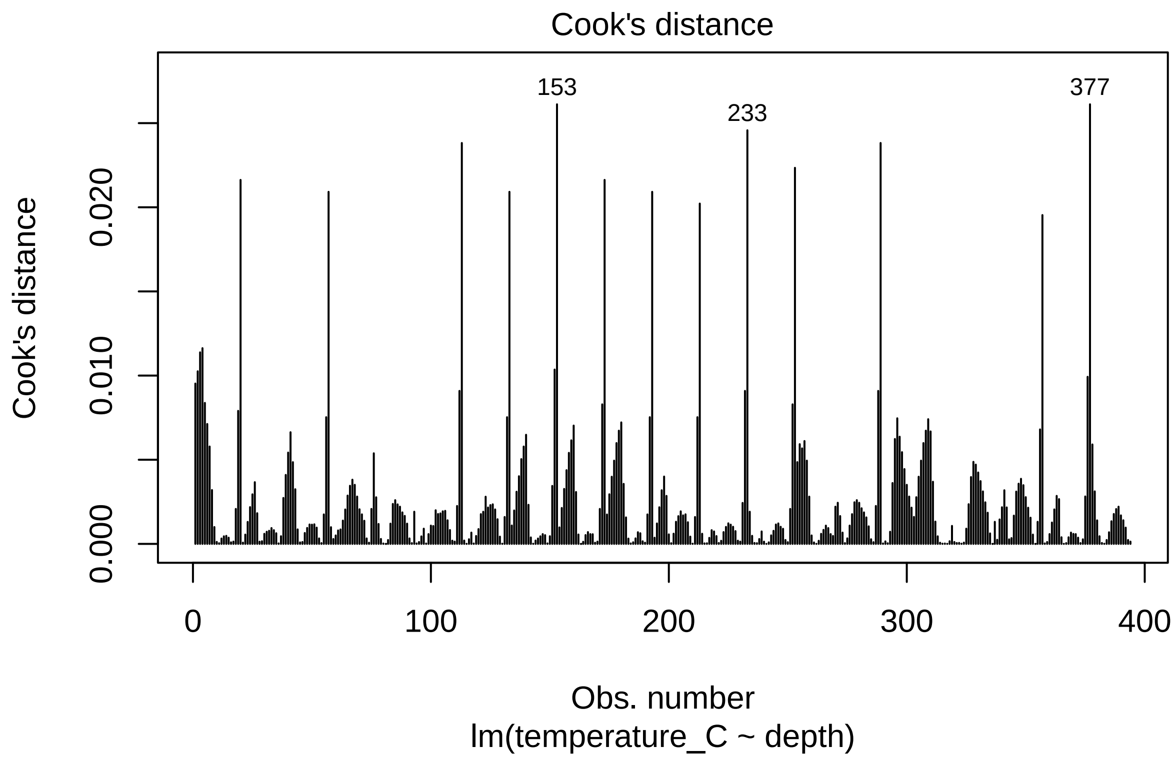

```
plot(temperature.depth.regreesion, 2)
```




```
plot(temperature.depth.regreesion, 3)
```




```
plot(temperature.depth.regreesion, 4)
```



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: There is a significant negative correlation (p value < 2.2e-16) between temperature and depth with around 9726 degrees of freedom(df). This model helps to explain 73.87% of variance in 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.

```r
#9 running AIC to determine best suited set of variables to predict temeprature
NTL_LTER.aic <- lm(data = NTL_LTER.selected, temperature_C ~ year4 + daynum + depth)
step(NTL_LTER.aic)
```

```
## Start:  AIC=966.18
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq     RSS     AIC
## - year4   1      21.7  4505.8  966.08
## <none>                 4484.0  966.18
## - daynum  1     638.3  5122.3 1016.62
## - depth   1   15263.4 19747.5 1548.29
##
## Step:  AIC=966.08
## temperature_C ~ daynum + depth
##
##          Df Sum of Sq     RSS     AIC
## <none>                 4505.8  966.08
## - daynum  1       717  5222.8 1022.27
## - depth   1     15242 19747.5 1546.29

##
## Call:
## lm(formula = temperature_C ~ daynum + depth, data = NTL_LTER.selected)
##
## Coefficients:
## (Intercept)       daynum        depth
##    11.19860      0.05466     -1.91767
```

```r
#10 running multiple regression on the recommended set of variables
temperature.best <- lm(data = NTL_LTER.selected, temperature_C ~ year4 + daynum + depth)
summary(temperature.best)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = NTL_LTER.selected)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3086 -2.7609 -0.3194  2.5294  8.1964
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 123.85217   81.96700   1.511    0.132
## year4        -0.05591    0.04067  -1.375    0.170
## daynum        0.05268    0.00707   7.451 6.02e-13 ***
## depth        -1.92045    0.05271 -36.435  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.391 on 390 degrees of freedom
## Multiple R-squared:  0.7802, Adjusted R-squared:  0.7785
## F-statistic: 461.5 on 3 and 390 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 are year, day number and depth. This model explains 74% of the total observed variance. This is a slight improvement from the previous model of just depth as the singular explanatory variable, increasing the R-squared by .01.

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
library(htmltools)

NTL_LTER.ANOVA <- aov(data = NTL_LTER.selected, temperature_C ~ lakename)
summary(NTL_LTER.ANOVA)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## lakename      4    228   56.96   1.098  0.357
## Residuals   389  20176   51.87
```

```
#rejecting null hypothesis

NTL_LTER.linreg <- lm(data = NTL_LTER.selected, temperature_C ~ lakename)
summary(NTL_LTER.linreg)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL_LTER.selected)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.331 -6.756 -2.550  7.338 14.536
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           10.5556     1.6975   6.218  1.3e-09 ***
## lakenamePaul Lake      1.9008     1.7856   1.065    0.288
## lakenamePeter Lake     2.0088     1.7904   1.122    0.263
## lakenameTuesday Lake  -0.4389     2.4006  -0.183    0.855
## lakenameWard Lake      3.3755     2.1610   1.562    0.119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.202 on 389 degrees of freedom
## Multiple R-squared:  0.01117,    Adjusted R-squared:  0.0009981
## F-statistic: 1.098 on 4 and 389 DF,  p-value: 0.3571
```

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

    Answer: There is a significant difference in mean temperatures among the lakes. This model explains about 4% of the total variance in 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.

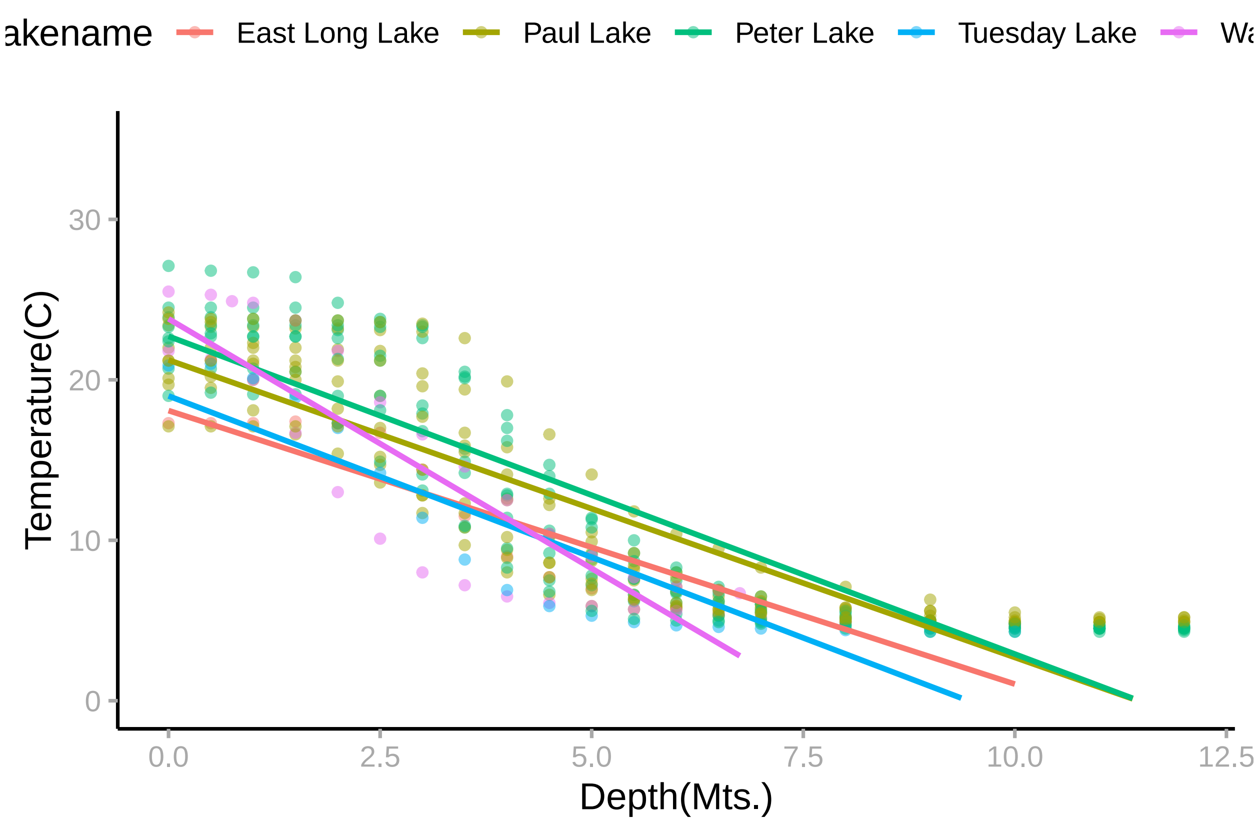
```
#14. scatter plots
unique(NTL_LTER.selected$lakename)
```

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

```
temperature.depth.2 <-
  ggplot(NTL_LTER.selected) +
  geom_point(aes(x = depth, y = temperature_C, color = lakename), alpha = 0.5) +
  geom_smooth(aes(x = depth, y = temperature_C, color = lakename), method = "lm", se = FALSE) +
  ylim(0, 35) +
  labs(x = "Depth(Mts.)", y = "Temperature(C)") +
  tweetheme
print(temperature.depth.2)
```

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

```
## Warning: Removed 13 rows containing missing values (`geom_smooth()`).
```



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

```
#15
TukeyHSD(NTL_LTER.ANOVA)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = NTL_LTER.selected)
##
## $lakename
##                                 diff       lwr      upr     p adj
## Paul Lake-East Long Lake     1.9007758 -2.992848 6.794400 0.8245987
## Peter Lake-East Long Lake    2.0088194 -2.898035 6.915674 0.7948864
## Tuesday Lake-East Long Lake -0.4388889 -7.018014 6.140236 0.9997496
## Ward Lake-East Long Lake     3.3754789 -2.546994 9.297952 0.5227817
## Peter Lake-Paul Lake         0.1080436 -2.069085 2.285172 0.9999228
## Tuesday Lake-Paul Lake      -2.3396647 -7.233289 2.553959 0.6849800
## Ward Lake-Paul Lake          1.4747031 -2.492456 5.441862 0.8466692
## Tuesday Lake-Peter Lake     -2.4477083 -7.354562 2.459146 0.6491273
## Ward Lake-Peter Lake         1.3666595 -2.616808 5.350127 0.8810112
## Ward Lake-Tuesday Lake       3.8143678 -2.108105 9.736840 0.3955788
```

```r
NTL_LTER.group <- HSD.test(NTL_LTER.ANOVA, "lakename", group = TRUE)
NTL_LTER.group
```

```
## $statistics
##    MSerror  Df     Mean       CV
##   51.86594 389 12.41503 58.00875
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   5         3.875817  0.05
##
## $means
##                 temperature_C      std   r Min  Max   Q25   Q50   Q75
## East Long Lake      10.55556 5.428923  18 4.7 17.4 5.525  8.35 17.15
## Paul Lake           12.45633 6.832892 169 4.8 24.2 5.800 10.20 19.70
## Peter Lake          12.56438 7.767335 160 4.3 27.1 5.000  9.35 20.70
## Tuesday Lake        10.11667 6.638502  18 4.3 21.0 4.625  6.40 16.30
## Ward Lake           13.93103 7.293006  29 5.6 25.5 7.100 12.50 21.30
##
## $comparison
## NULL
##
## $groups
##                 temperature_C groups
## Ward Lake           13.93103      a
## Peter Lake          12.56438      a
## Paul Lake           12.45633      a
## East Long Lake      10.55556      a
## Tuesday Lake        10.11667      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: Statistically speaking, Paul lake and Ward Lake have the same mean temperature as Peter Lake. Central Long Lake has a distinct mean temperature from most of the other lakes except from Crampton, hence, no 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 perform a two-way t test.

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?

```
NTL_LTER.ward.crampton <- NTL_LTER.selected %>%
  filter(lakename%in% c("Crampton Lake", "Ward Lake"))
```

Answer: