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

Hannah Wudke

Fall 2024

OVERVIEW

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

Directions

1. Rename this file `<FirstLast>_A07_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
getwd()
```

```
## [1] "/home/guest/EDE-Class"
```

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(agricolae)
library(here)
```

```
## here() starts at /home/guest/EDE-Class
```

```
library(lubridate)
library(ggplot2)
library(dplyr)
library(cowplot)
```

```
##
## Attaching package: 'cowplot'
##
## The following object is masked from 'package:lubridate':
##
##     stamp
```

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

LakeChemPhys$sampledate <- as.Date(LakeChemPhys$sampledate, format = "%m/%d/%Y")
class(LakeChemPhys$sampledate)
```

```
## [1] "Date"
```

```
#2
mytheme <- theme_classic(base_size = 12) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "right")
theme_set(mytheme)
```

Simple regression

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

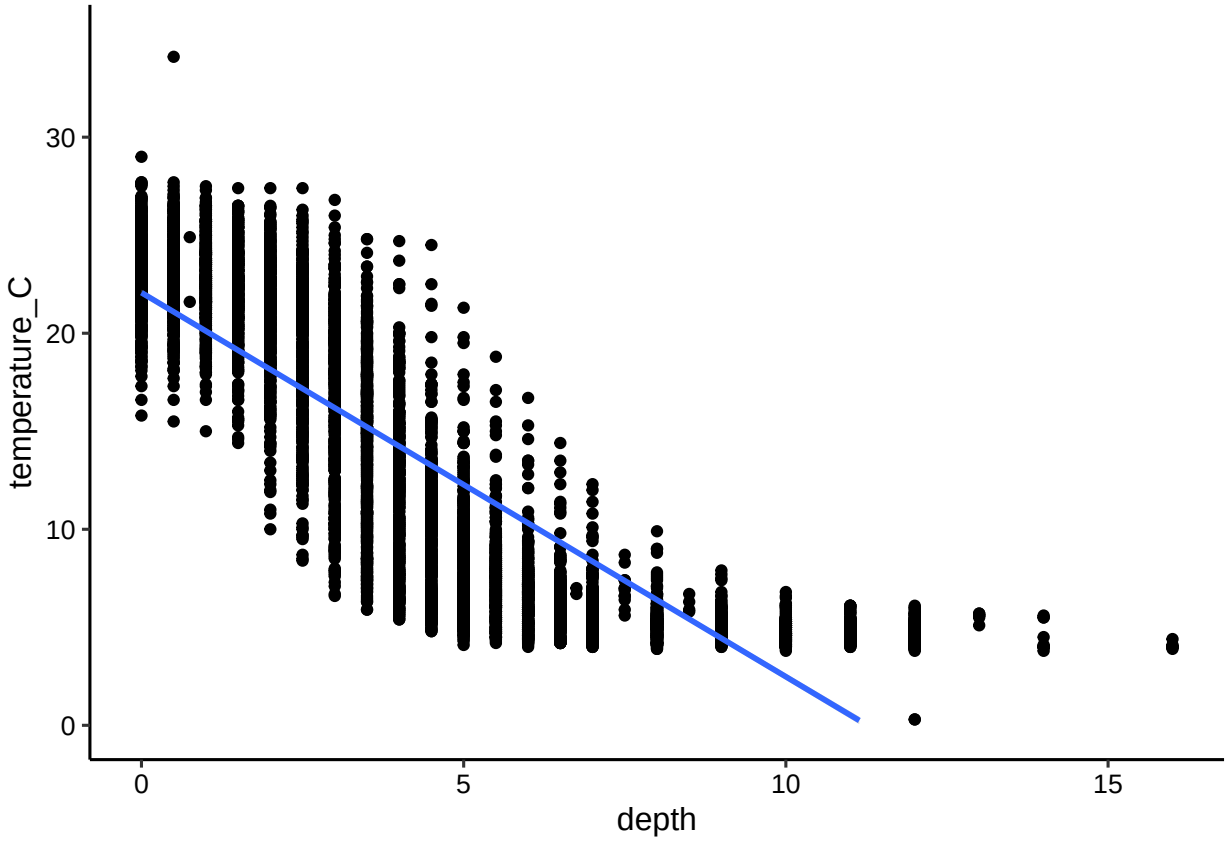
3. State the null and alternative hypotheses for this question: > Answer: H0: Mean lake temperature is consistant at varying depths across all lakes Ha: Mean lake temperature changes with depth across all lakes
4. Wrangle your NTL-LTER dataset with a pipe function so that the records meet the following 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.

```r
#4
Selected_LakeChemPhys <- select(LakeChemPhys, lakename, year4, daynum, depth,
                                temperature_C)

WrangledLakeChemPhys <- Selected_LakeChemPhys %>% filter(daynum %in% (186:209)) %>% drop_na()

#5

TempDepthPlot <- ggplot(WrangledLakeChemPhys, aes(x = depth,
    y = temperature_C)) + geom_point() + ylim(0, 35) + geom_smooth(method = "lm")
print(TempDepthPlot)
```

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

## Warning: Removed 24 rows containing missing values or values outside the scale range
## (`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: Temperature decreases with increasing depth. Although the relationship is very linear at low depths, it becomes more exponential at high depths. However, the lm line shows that a linear relationship is evident and can still be derived overall.

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

```
#7

LakeRegression <- lm(WrangledLakeChemPhys$temperature_C ~ WrangledLakeChemPhys$depth)
summary(LakeRegression)
```

```
##
## Call:
## lm(formula = WrangledLakeChemPhys$temperature_C ~ WrangledLakeChemPhys$depth)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5925 -3.0359  0.0616  2.9807 13.6748
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  22.06953    0.07715   286.0   <2e-16 ***
## WrangledLakeChemPhys$depth -1.95902    0.01331  -147.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.842 on 7592 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7405
## F-statistic: 2.167e+04 on 1 and 7592 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 relationship between temperature and depth is negatively correlated, meaning that an increase in depth results in a decrease in temperature. 67.42% of variability in temperature is explained by changes in depth (R-squared). This is based on 7592 degrees of freedom. For every 1m change in depth, the temperature is expected to decrease 1.95 degrees Celsius. The p-vale is 2.2e-16, which is significantly smaller than 0.05. This means the results are statistically signficant, and we reject the null hypothesis. This means that temperature is not the same across varying depths.

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
LakeAIC <- lm(data = WrangledLakeChemPhys, temperature_C ~ depth + daynum + year4)
step(LakeAIC)
```

```
## Start:  AIC=20406.3
## temperature_C ~ depth + daynum + year4
##
##          Df Sum of Sq    RSS   AIC
## <none>                 111438 20406
## - year4   1       132 111569 20413
## - daynum  1       496 111934 20438
## - depth   1    320055 431493 30685

##
## Call:
## lm(formula = temperature_C ~ depth + daynum + year4, data = WrangledLakeChemPhys)
##
## Coefficients:
## (Intercept)        depth       daynum        year4
##   -14.29474     -1.95967      0.03654      0.01458
```

```r
#10
LakeMultRegression <- lm(data = WrangledLakeChemPhys,
                   temperature_C ~ depth + daynum + year4)
summary(LakeMultRegression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + daynum + year4, data = WrangledLakeChemPhys)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7213 -3.0163  0.0931  2.9965 13.7457
##
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -14.294744   9.803364   -1.458  0.14484
## depth        -1.959669   0.013273 -147.645  < 2e-16 ***
## daynum        0.036544   0.006288    5.812 6.42e-09 ***
## year4         0.014579   0.004867    2.995  0.00275 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.832 on 7590 degrees of freedom
## Multiple R-squared:  0.742,  Adjusted R-squared:  0.7419
## F-statistic:  7277 on 3 and 7590 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 variables is all of the variables - depth, daynum, and year4. This is an improvement. This new model has an R-squared value of 0.7419, indicating 74.19% of variability

in temperature is explained by the x variables. The original model has an R-squared value of 0.7405, meaning 74.05% of change was explained before. This model is only a slight improvement over the original model, which suggests that depth is the main variable impacting temperature, and the other variables have a somewhat negligble impact on temperature.

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
LakeRegressionANOVA <- aov(data = WrangledLakeChemPhys, temperature_C ~ lakename)
summary(LakeRegressionANOVA)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  17428  2178.5   39.86 <2e-16 ***
## Residuals   7585 414517    54.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
LakeRegressionLM <- lm(data = WrangledLakeChemPhys, temperature_C ~ lakename)
summary(LakeRegressionLM)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = WrangledLakeChemPhys)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.765  -6.613  -2.711   7.689  23.798
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 17.9848     0.7707  23.335  < 2e-16 ***
## lakenameCrampton Lake       -2.3882     0.9097  -2.625 0.008676 **
## lakenameEast Long Lake      -7.6824     0.8144  -9.433  < 2e-16 ***
## lakenameHummingbird Lake    -7.0984     1.1023  -6.440 1.27e-10 ***
## lakenamePaul Lake           -4.0864     0.7879  -5.187 2.20e-07 ***
## lakenamePeter Lake          -4.6743     0.7865  -5.943 2.92e-09 ***
## lakenameTuesday Lake        -6.9196     0.8001  -8.648  < 2e-16 ***
## lakenameWard Lake           -3.5262     1.0321  -3.417 0.000637 ***
## lakenameWest Long Lake      -6.2715     0.8124  -7.720 1.32e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.393 on 7585 degrees of freedom
## Multiple R-squared:  0.04035,    Adjusted R-squared:  0.03934
## F-statistic: 39.86 on 8 and 7585 DF,  p-value: < 2.2e-16
```

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

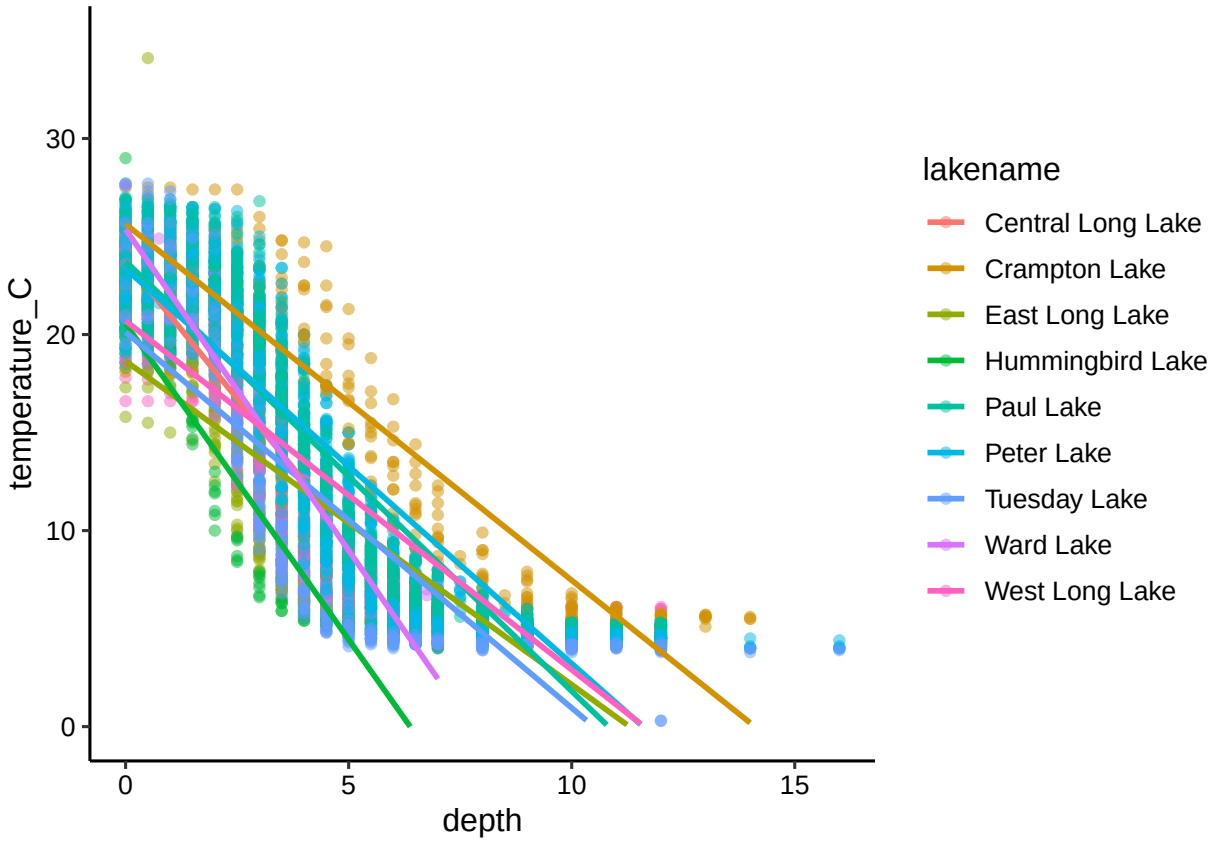
    Answer: Yes, the Pr(>F) value is 2e-16, which is well below 0.05. This means we reject the null hypothesis, which in this case, is that the mean temperature is the same among various lakes.

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.
MultiLakePlot <- ggplot(WrangledLakeChemPhys, aes(x = depth,
y = temperature_C, color = lakename)) + geom_point(alpha = 0.5) + ylim(0,
                                      35) + geom_smooth(method = "lm", se = FALSE)

print(MultiLakePlot)
```

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

```
## Warning: Removed 73 rows containing missing values or values outside the scale range
## (`geom_smooth()`).
```



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

```r
#15
TukeyHSD(LakeRegressionANOVA)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = WrangledLakeChemPhys)
##
## $lakename
##                                         diff          lwr        upr     p adj
## Crampton Lake-Central Long Lake    -2.3882014  -5.21068009  0.4342773 0.1761464
## East Long Lake-Central Long Lake   -7.6823745 -10.20922327 -5.1555257 0.0000000
## Hummingbird Lake-Central Long Lake -7.0984190 -10.51840802 -3.6784299 0.0000000
## Paul Lake-Central Long Lake        -4.0863881  -6.53091350 -1.6418627 0.0000078
## Peter Lake-Central Long Lake       -4.6742918  -7.11459249 -2.2339912 0.0000001
## Tuesday Lake-Central Long Lake     -6.9196093  -9.40212796 -4.4370907 0.0000000
## Ward Lake-Central Long Lake        -3.5261619  -6.72824408 -0.3240798 0.0184282
## West Long Lake-Central Long Lake   -6.2714976  -8.79212435 -3.7508708 0.0000000
## East Long Lake-Crampton Lake       -5.2941731  -7.00149376 -3.5868524 0.0000000
## Hummingbird Lake-Crampton Lake     -4.7102176  -7.57837101 -1.8420641 0.0000126
## Paul Lake-Crampton Lake            -1.6981867  -3.28112180 -0.1152516 0.0247011
## Peter Lake-Crampton Lake           -2.2860904  -3.86249343 -0.7096874 0.0002379
## Tuesday Lake-Crampton Lake         -4.5314079  -6.17240698 -2.8904088 0.0000000
## Ward Lake-Crampton Lake            -1.1379605  -3.74243921  1.4665182 0.9141312
## West Long Lake-Crampton Lake       -3.8832962  -5.58139465 -2.1851977 0.0000000
## Hummingbird Lake-East Long Lake     0.5839555  -1.99381155  3.1617226 0.9987642
## Paul Lake-East Long Lake            3.5959864   2.63460072  4.5573721 0.0000000
## Peter Lake-East Long Lake           3.0080827   2.05749061  3.9586747 0.0000000
## Tuesday Lake-East Long Lake         0.7627652  -0.29149134  1.8170217 0.3765590
## Ward Lake-East Long Lake            4.1562126   1.87544776  6.4369774 0.0000006
## West Long Lake-East Long Lake       1.4108769   0.26977337  2.5519805 0.0040081
## Paul Lake-Hummingbird Lake          3.0120309   0.51490798  5.5091538 0.0057402
## Peter Lake-Hummingbird Lake         2.4241271  -0.06886015  4.9171144 0.0641832
## Tuesday Lake-Hummingbird Lake       0.1788097  -2.35551810  2.7131374 0.9999998
## Ward Lake-Hummingbird Lake          3.5722571   0.32984306  6.8146710 0.0183282
## West Long Lake-Hummingbird Lake     0.8269214  -1.74474687  3.3985896 0.9862013
## Peter Lake-Paul Lake               -0.5879037  -1.29101468  0.1152072 0.1889721
## Tuesday Lake-Paul Lake             -2.8332212  -3.67119290 -1.9952496 0.0000000
## Ward Lake-Paul Lake                 0.5602262  -1.62898097  2.7494333 0.9970561
## West Long Lake-Paul Lake           -2.1851095  -3.13002061 -1.2401984 0.0000000
## Tuesday Lake-Peter Lake            -2.2453175  -3.07088355 -1.4197514 0.0000000
## Ward Lake-Peter Lake                1.1481299  -1.03635874  3.3326186 0.7880103
## West Long Lake-Peter Lake          -1.5972057  -2.53113284 -0.6632786 0.0000041
## Ward Lake-Tuesday Lake              3.3934474   1.16189587  5.6249989 0.0000845
## West Long Lake-Tuesday Lake         0.6481117  -0.39114347  1.6873669 0.5892990
## West Long Lake-Ward Lake           -2.7453357  -5.01920522 -0.4714661 0.0056640
```

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: Peter Lake is statistically different than Long Lake, Crampton Lake, East Long Lake, Tuesday Lake, and West Long Lake. No lake has a mean temperature that is statistically distinct from all 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: A two-sample t-test, after filtering the dataset to only include those two lakes.

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?

```
WrangledLakeChemPhys2 <- WrangledLakeChemPhys %>% filter(lakename %in% c("Crampton Lake",
                                                                         "Ward Lake"))

TwoLakeTTest <- t.test(WrangledLakeChemPhys2$temperature_C ~ WrangledLakeChemPhys2$lakename)
TwoLakeTTest
```

```
##
##  Welch Two Sample t-test
##
## data:  WrangledLakeChemPhys2$temperature_C by WrangledLakeChemPhys2$lakename
## t = 1.3547, df = 228.5, p-value = 0.1769
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -0.5172467  2.7931678
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.59658                    14.45862
```

```
TwoLakeTTestLM <- lm(WrangledLakeChemPhys2$temperature_C ~ WrangledLakeChemPhys2$lakename)

summary(TwoLakeTTestLM)
```

```
##
## Call:
## lm(formula = WrangledLakeChemPhys2$temperature_C ~ WrangledLakeChemPhys2$lakename)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4966  -7.4336   0.0224   6.9034  13.1414
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           15.5966     0.4829  32.301   <2e-16
## WrangledLakeChemPhys2$lakenameWard Lake  -1.1380     0.8387  -1.357    0.176
##
## (Intercept)                           ***
## WrangledLakeChemPhys2$lakenameWard Lake
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.386 on 348 degrees of freedom
## Multiple R-squared:  0.005262,   Adjusted R-squared:  0.002403
## F-statistic: 1.841 on 1 and 348 DF,  p-value: 0.1757
```

Answer: The p-value is 0.1769, which means we fail to reject the null hypothesis. The means in each lake are statistically the same. This does match my answer for part 16, which gives a p-value of 0.91. Although the p-values are different, the conclusions are the same.