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

Shiqi Zheng

Fall 2023

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 load data
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## 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 /Users/shiqizheng/Desktop/ENV872/EDE_Fall2023
```

```r
here()
```

```
## [1] "/Users/shiqizheng/Desktop/ENV872/EDE_Fall2023"
```

```r
NTL.raw <- read.csv(here("Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
                    stringsAsFactors = TRUE)
# Set date to date format
NTL.raw$sampledate <- as.Date(NTL.raw$sampledate, format = "%m/%d/%y")

#2 build theme
mytheme <- theme_gray() +
  theme(axis.text = element_text(color = "black"),
        axis.ticks = element_line(color = "gray", linewidth = 0.5),
        panel.grid = element_line(color = "white", linewidth = 0.2),
        legend.position = "right",
        legend.background = element_rect(fill = "lightyellow"),
        plot.title = element_text(color = "black", size = 16, face = "bold", hjust = 0.5)
        )
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: The mean lake temperature recorded during July is the same across all lake depths Ha: The mean lake temperature recorded during July is different across different lake depths

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.

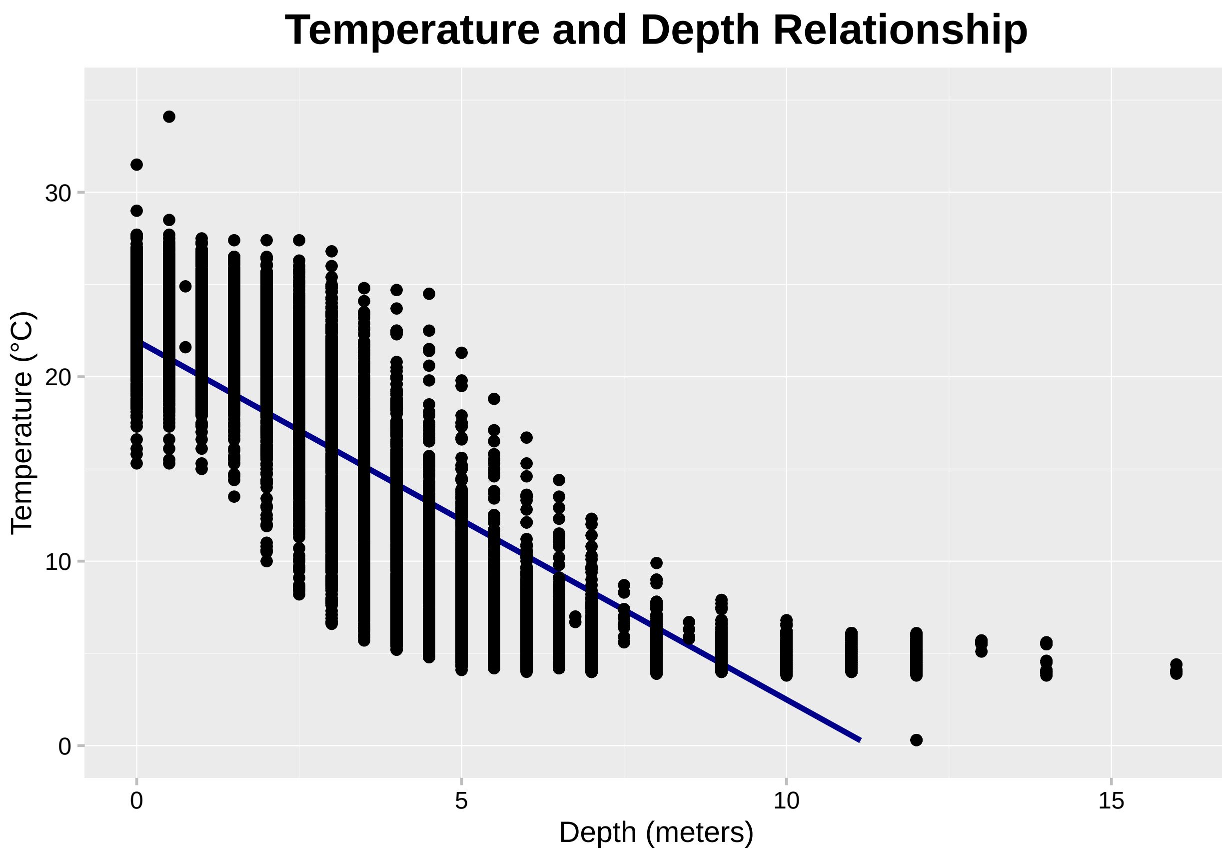
```r
#4 wrangle data
NTL <-
  NTL.raw %>%
  mutate(Month = month(sampledate)) %>%
  filter(Month == 7) %>%
  select(lakename,year4,daynum,depth,temperature_C) %>%
  na.omit()

#5 scatter plot
ggplot(NTL, aes(x = depth, y = temperature_C)) +
  geom_smooth(method = "lm",color = "darkblue", se = TRUE) +
  geom_point()+
```

```r
  ylim(0,35)+
  labs(x="Depth (meters)", y="Temperature (°C)",
       title="Temperature and Depth Relationship")
```

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

```
## Warning: Removed 24 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: There is a general negative relationship between temperature depth. The negative slope of linear model line (the blue line) also comfirms that temperature decreases with depth increases. The linear model line is not fitting perfectly with the general trend of the data points. There may be some variation and noise in the data, as indicated by the spread of data points around the linear model line. Also there may be some outlier data when depth beyond 10 m. This implies that while there is a positive relationship, it's not a perfect linear relationship, and other factors may also influence the temperature at different depths.

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

```r
#7
linear_model <- lm(data = NTL, temperature_C ~ depth)
summary(linear_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = NTL)
##
## 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: According to the R^2 value, 73.87% of the total variance in temperature is explained by changes in depth. The degree of freedon is 9726. The p-value is < 2.2e-16 which is less than the confidence level 0.05, so this model is statistically significant and a meaningful regression. The temperature is predicted to decrease 1.9 degree with every 1m change in depth.

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

```
## 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 = NTL)
##
## Coefficients:
## (Intercept)        depth        year4       daynum
##    -8.57556     -1.94644      0.01134      0.03978
```

```
#10 multiple regression
AICmodel <- lm(data = NTL, temperature_C ~ depth + year4 + daynum)
summary(AICmodel)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth + year4 + daynum, data = NTL)
##
## 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 from AIC is depth, year, and day number. The p-value is less than the confidence level so this model is statistically significant and a meaningful regression. It explains 74.11% of variance which is slightly higher than the single variable regression model that explains 73.87% of the total variance. So there is an small improvement by using multiple variables.

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 ANOVA as aov
Lake.anova <- aov(data = NTL, temperature_C ~ lakename)
summary(Lake.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
```

```
# Format ANOVA as lm
Lake.anova.lm <- lm(data = NTL, temperature_C ~ lakename)
summary(Lake.anova.lm)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = NTL)
##
## 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
```

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

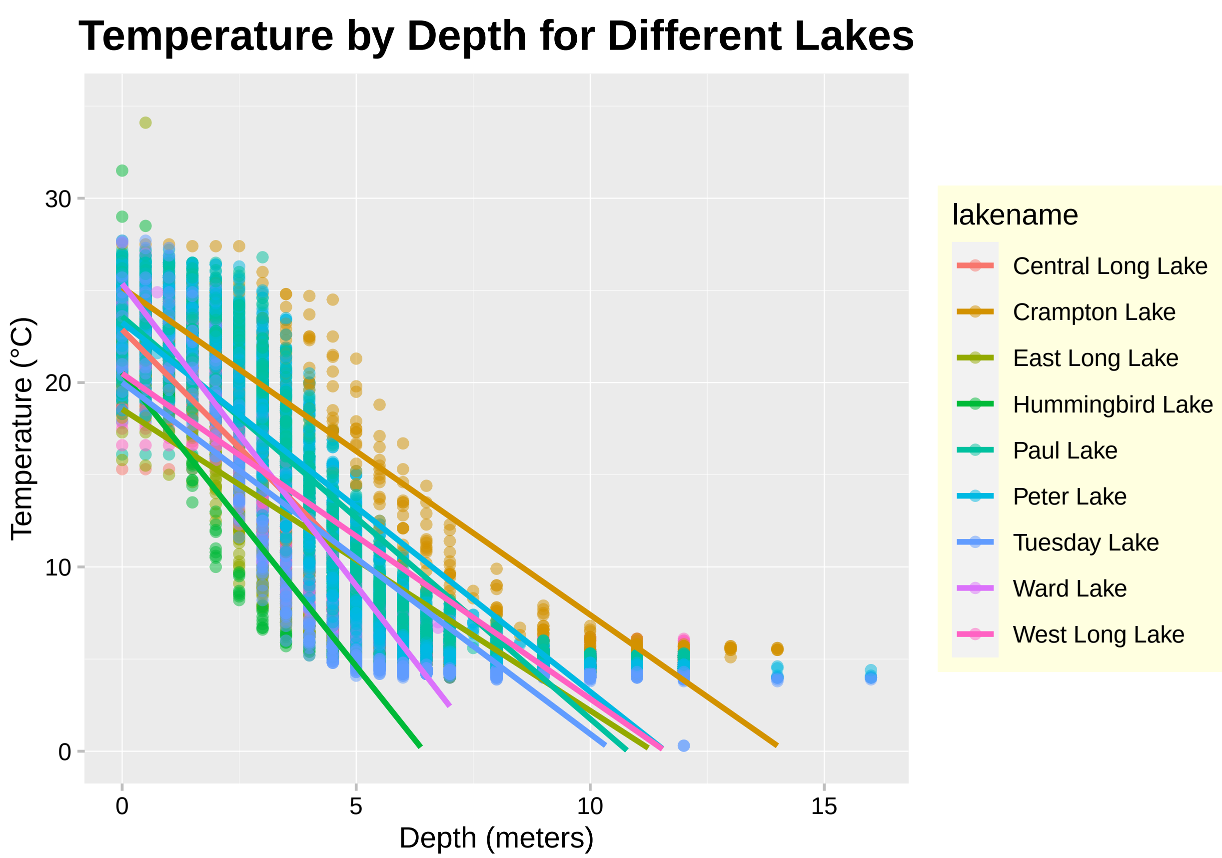
    Answer: Both p value from ANOVA model and linear model are less than 2e-16, indicating high level of statistical significance and strong evidence against the null hypothesis. So there is a significant difference in mean temperature among the lakes.

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.
temp_lake <- ggplot(NTL, aes(x = depth, y=temperature_C, color = lakename)) +
  geom_point(alpha=0.5)+ #transparency
  geom_smooth(method = "lm", se = FALSE) +
   ylim(0,35)+
   labs(x="Depth (meters)", y="Temperature (°C)",
        title = "Temperature by Depth for Different Lakes")
print(temp_lake)
```

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

## Warning: Removed 73 rows containing missing values (`geom_smooth()`).
```



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

```
#15
# Extract groupings for pairwise relationships
Lake.groups <- HSD.test(Lake.anova, "lakename", group = TRUE)
Lake.groups
```

```
## $statistics
##   MSerror   Df     Mean       CV
##   54.1016 9719 12.72087 57.82135
##
## $parameters
##    test   name.t ntr StudentizedRange alpha
##   Tukey lakename   9         4.387504  0.05
##
## $means
##                    temperature_C      std    r        se Min  Max    Q25   Q50
## Central Long Lake      17.66641 4.196292  128 0.6501298 8.9 26.8 14.400 18.40
## Crampton Lake          15.35189 7.244773  318 0.4124692 5.0 27.5  7.525 16.90
## East Long Lake         10.26767 6.766804  968 0.2364108 4.2 34.1  4.975  6.50
## Hummingbird Lake       10.77328 7.017845  116 0.6829298 4.0 31.5  5.200  7.00
## Paul Lake              13.81426 7.296928 2660 0.1426147 4.7 27.7  6.500 12.40
## Peter Lake             13.31626 7.669758 2872 0.1372501 4.0 27.0  5.600 11.40
## Tuesday Lake           11.06923 7.698687 1524 0.1884137 0.3 27.7  4.400  6.80
## Ward Lake              14.45862 7.409079  116 0.6829298 5.7 27.6  7.200 12.55
## West Long Lake         11.57865 6.980789 1026 0.2296314 4.0 25.7  5.400  8.00
##                       Q75
## Central Long Lake 21.000
## Crampton Lake     22.300
## East Long Lake    15.925
## Hummingbird Lake  15.625
## Paul Lake         21.400
## Peter Lake        21.500
## Tuesday Lake      19.400
## Ward Lake         23.200
## West Long Lake    18.800
##
## $comparison
## NULL
##
## $groups
##                    temperature_C groups
## Central Long Lake      17.66641      a
## Crampton Lake          15.35189     ab
## Ward Lake              14.45862     bc
## Paul Lake              13.81426      c
## Peter Lake             13.31626      c
## West Long Lake         11.57865      d
## Tuesday Lake           11.06923     de
## Hummingbird Lake       10.77328     de
## East Long Lake         10.26767      e
##
## 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: Lakes with the same letter are not statistically distinct from each other. So Paul Lake (c) and Ward Lake (bc) have the same mean temperature statistically as Peter Lake (c). There is

no lake is statistically distinct from all the other lakes as they all share at least one same letters with one another lake.

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 can use the Two-sample 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_lake <-
  NTL %>%
  filter(lakename %in% c("Crampton Lake", "Ward Lake"))

t.test <- t.test(NTL_lake$temperature_C ~ NTL_lake$lakename)
print(t.test)
```

```
##
##  Welch Two Sample t-test
##
## data:  NTL_lake$temperature_C by NTL_lake$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 n
## 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 t-test has p-value larger than 0.05 so the null hypothesis is rejected. The test accept the alternative hypothesis that true difference in means between group Crampton Lake and group Ward Lake is not equal to 0. So there is a different between these two lakes' temperature. This is not match with HSD test result as Crampton Lake and Ward Lake share a same letter b indicating they are not statistically distinct.