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

Hanbin Lyu

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_Fall2024"
```

```
install.packages("agricolae")
library(tidyverse)
library(lubridate)
library(here)
library(cowplot)
library(agricolae)
library(ggplot2)

chemphy = read.csv(
  here("Data","Raw", "NTL-LTER_Lake_ChemistryPhysics_Raw.csv"),
  stringsAsFactors = TRUE) %>%
  mutate(sampledate = mdy(sampledate))
```

```
#2
my_theme = theme(
    line = element_line(color = "darkseagreen4", linewidth = 2,
                        linetype = "solid", lineend = "round"),
    rect = element_rect(fill = "honeydew",color = "darkolivegreen4",
                        linewidth = 1, linetype = "solid"),
    text = element_text(family = "serif", face = "plain",
                        color = "darkolivegreen",size = 12, hjust = 0.5,
                        vjust = 0.5, angle = 0, lineheight = 1.5),

    # Modified inheritance structure of text element
    plot.title = element_text(family = "serif", face = "bold",
                              color = "darkolivegreen", size = 16, hjust = 0.5,
                              vjust = 0.5, angle = 0, lineheight = 1.5),
    axis.title.x = element_text(family = "serif", face = "plain",
                              color = "darkolivegreen", size = 12, hjust = 0.5,
                              vjust = 0.5, angle = 0, lineheight = 1.5),
    axis.title.y = element_text(family = "serif", face = "plain",
                              color = "darkolivegreen", size = 12, hjust = 0.5,
                              vjust = 0.5, angle = 0, lineheight = 1.5),
    axis.text = element_text(family = "serif", face = "plain",
                             color = "darkolivegreen", size = 12, hjust = 0.5,
                             vjust = 0.5, angle = 0, lineheight = 1.5),

    # Modified inheritance structure of line element
    axis.ticks = element_blank(),
    panel.grid.major =  element_line(color = "gray85", linewidth = 0.5,
                                     linetype = "solid", lineend = "square"),
    panel.grid.minor =  element_blank(),

    # Modified inheritance structure of rect element
    plot.background = element_rect(fill = "ivory"),
    panel.background = element_rect(fill = "lightyellow"),
    legend.key = element_rect(fill = "honeydew", linewidth = 0.5),

    # Modifiying legend.position
    legend.position = 'right',

    complete = TRUE
    )
```

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 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.

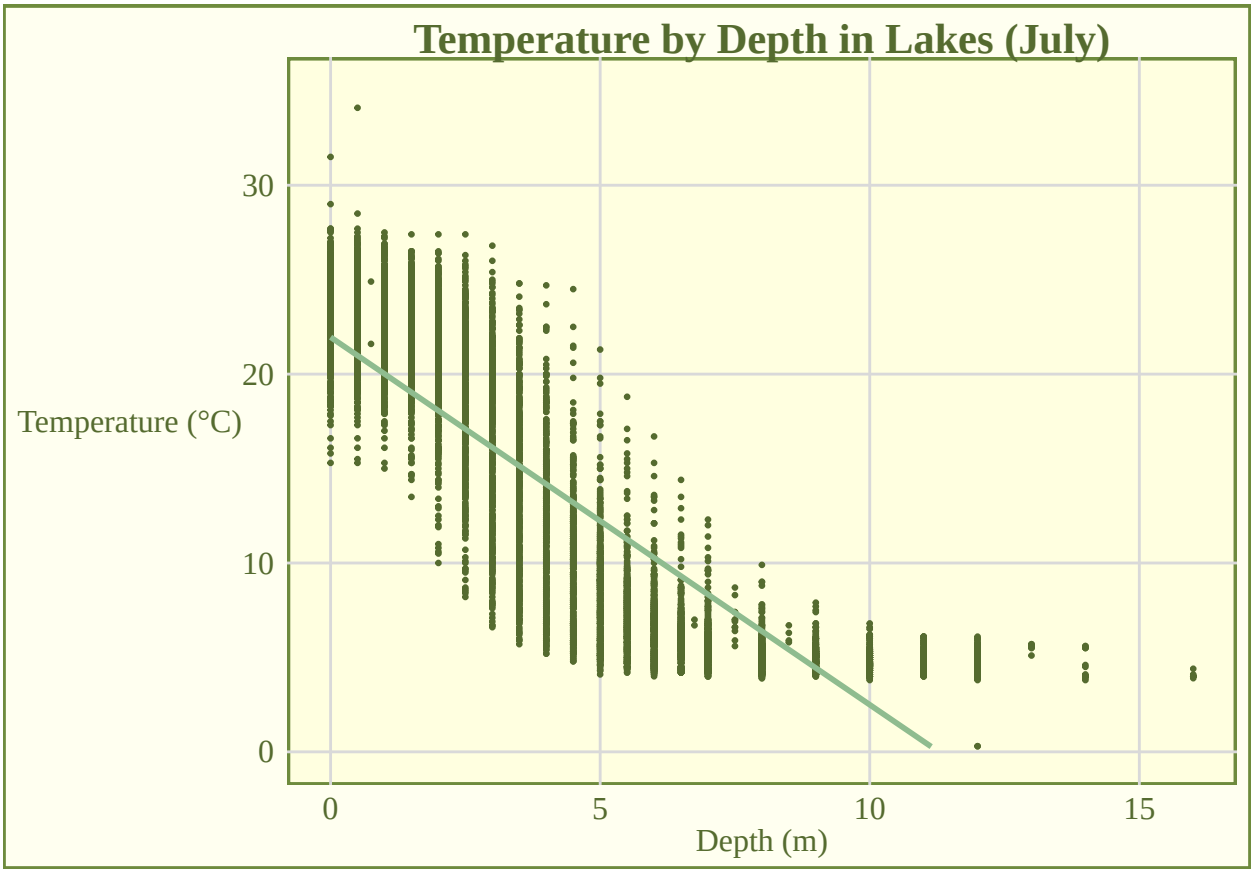
```
#4
summary_chemphy = chemphy %>%
  filter(month(sampledate) == 7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na()

#5
tem_depth_plot = ggplot(summary_chemphy, aes(x = depth, y = temperature_C)) +
  geom_point(size = 0.5, color = "darkolivegreen") +
  geom_smooth(method = "lm", se = FALSE, color = "darkseagreen") +
  ylim(0, 35) +
  labs(
    title = "Temperature by Depth in Lakes (July)",
    x = "Depth (m)",
    y = "Temperature (°C)") +
  my_theme

tem_depth_plot
```

```
## `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: the figure shows a negative relationship between temperature and depth, with temperature decreasing as depth increases. The points suggest a rapid decline in temperature at shallower depths, followed by a slower decline at greater depths. The linear model captures the overall trend of temperature-depth relationship.

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

```
#7
linear_regression = lm(data = summary_chemphy, temperature_C ~ depth)

summary(linear_regression)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = summary_chemphy)
##
## 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: the linear regression model shows a significant relationship between temperature and depth because p-value smaller than 2.2e-16. The coefficient for depth is -1.94621, indicating that for every 1-meter increase in depth, the temperature is predicted to decrease by approximately 1.95°C. The Multiple R-squared value is 0.7387, meaning that about 73.87% of the variability in temperature is explained by changes in depth. The model is based on 9726 degrees of freedom, with a residual standard error of 3.835, indicating some variability in temperature that is not accounted for by depth alone.

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
models = lm(data = summary_chemphy, temperature_C ~ year4 + daynum + depth)
best_model = step(models)
```

```
## Start:  AIC=26065.53
## temperature_C ~ year4 + daynum + depth
##
##          Df Sum of Sq    RSS   AIC
## <none>                141687 26066
## - year4   1       101 141788 26070
## - daynum  1      1237 142924 26148
## - depth   1    404475 546161 39189
```

```
#10
summary(best_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = summary_chemphy)
##
## 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
## year4        0.011345   0.004299    2.639  0.00833 **
## daynum       0.039780   0.004317    9.215  < 2e-16 ***
## depth       -1.946437   0.011683 -166.611  < 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 AIC method suggests using year4, daynum, and depth as the final set of explanatory variables to predict lake temperature. The multiple regression model explains 74.12% of the variance in temperature (R-squared = 0.7412), which is a slight improvement over the model using only depth as a predictor (R-squared = 0.7387). This indicates that including year4 and daynum adds additional explanatory power, although the improvement is small.

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).

```r
#12
anova_model = aov(data = summary_chemphy, temperature_C ~ lakename)
summary(anova_model)
```

```
##               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
```

```r
lm_model = lm(data = summary_chemphy, temperature_C ~ lakename)
summary(lm_model)
```

```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = summary_chemphy)
##
## 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: based on the ANOVA results, there is a significant difference in mean temperature among the lakes. This suggests that at least one of the lakes has a mean temperature that is significantly different from the others. The linear model results also support this conclusion, with several lake coefficients being different from other lakes. The significant p-values for the lake names indicate that many of the lakes differ in their average temperature compared to the intercept.

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.

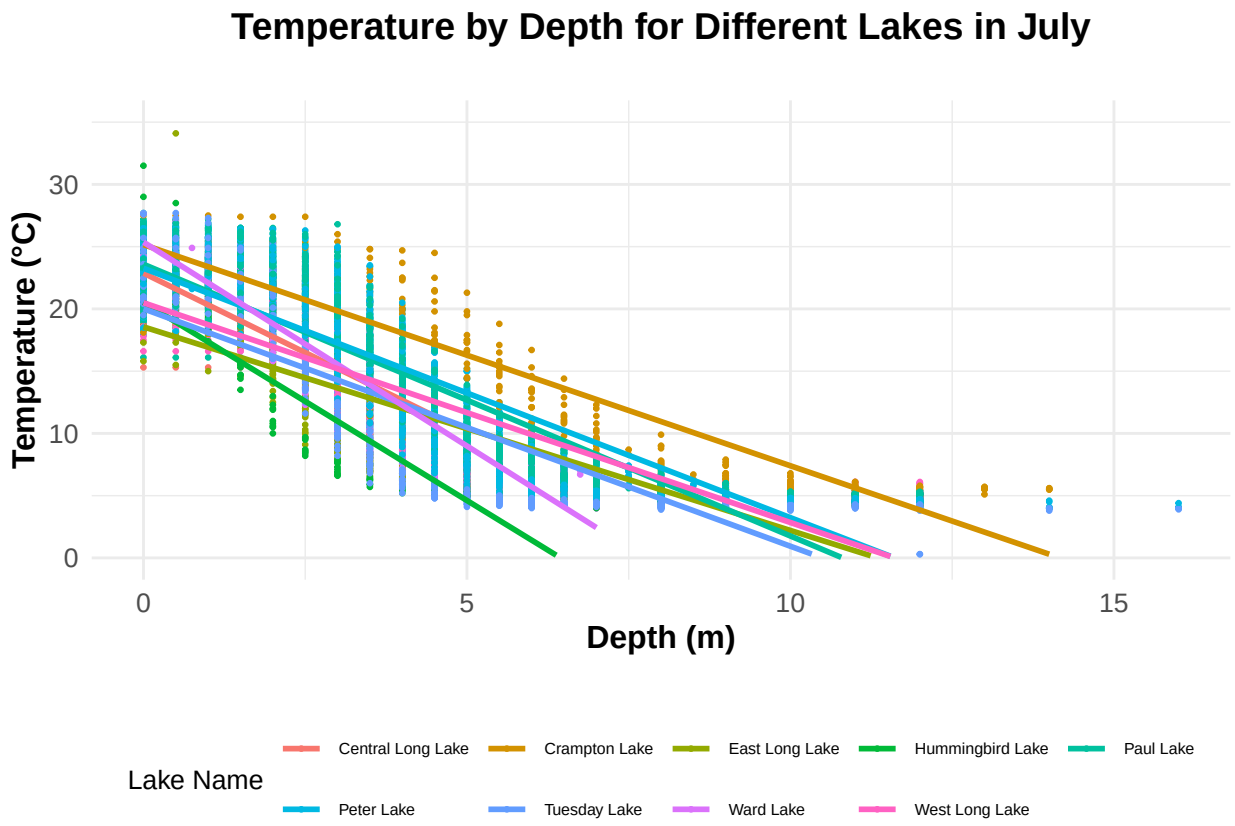
```r
#14.
all_lakes = ggplot(summary_chemphy, aes(x = depth, y = temperature_C,
                                        color = lakename)) +
  geom_point(size = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  ylim(0, 35) +
  labs(
    title = "Temperature by Depth for Different Lakes in July",
    x = "Depth (m)",
    y = "Temperature (°C)",
    color = "Lake Name") +
  theme_minimal() +
  theme(
```

```
    text = element_text(size = 12),
    plot.title = element_text(hjust = 0.5, size = 14,
                              face = "bold", margin = margin(b = 20)),
    axis.title = element_text(face = "bold"),
    axis.text = element_text(size = 10),
    legend.title = element_text(size = 10),
    legend.text = element_text(size = 6),
    legend.position = "bottom",
    legend.box.margin = margin(t = 10)) +
  guides(color = guide_legend(nrow = 2, byrow = TRUE))

all_lakes
```

```
## `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.

```
#15
tukey = HSD.test(anova_model, "lakename", group = TRUE)

tukey
```

```
## $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: the Tukey's HSD test results indicate that Peter Lake has the same mean temperature as Paul Lake and Ward Lake, since they all fall within group c. Central Long Lake is the only

lake that has a mean temperature distinct from all the other lakes. Although the Crampton Lake is group ab, its temperature is still much different from that of Central Long 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: if we are just looking at Peter Lake and Paul Lake, another test to determine if they have distinct mean temperatures is the two-sample t-test. This test can compare the means of the two lakes to see if there is a significant difference.

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?

```
two_lakes = summary_chemphy %>%
  filter(lakename == c("Crampton Lake", "Ward Lake"))
t_test = t.test(data = two_lakes, temperature_C ~ lakename)

t_test
```

```
##
##  Welch Two Sample t-test
##
## data:  temperature_C by lakename
## t = 0.98673, df = 95.77, p-value = 0.3263
## alternative hypothesis: true difference in means between group Crampton Lake and group Ward Lake is n
## 95 percent confidence interval:
##  -1.130614  3.365610
## sample estimates:
## mean in group Crampton Lake     mean in group Ward Lake
##                    15.37107                    14.25357
```

    Answer: the two-sample t-test indicates that there is no significant difference in the mean temperatures between Crampton Lake and Ward Lake (p-value = 0.3263). Since the p-value is greater than 0.05, we conclude that their mean temperatures are statistically equal. This result is consistent with the findings from part 16, where Tukey's HSD test also showed no significant difference between the two lakes, as they both had group b.