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

Ruiqing Li

Spring 2023

OVERVIEW

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

Directions

1. Rename this file `<RuiqingLi>_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)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

```
library(here)
```

```
## here() starts at /Users/ruiqingli/Desktop/DataAnalytics/RWORK/EDA-Spring2023
```

```
library(agricolae)
getwd()
```

```
## [1] "/Users/ruiqingli/Desktop/DataAnalytics/RWORK/EDA-Spring2023"
```

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

Raw.NTL.LTER$sampledate <- as.Date(Raw.NTL.LTER$sampledate, format = "%m/%d/%y")
#2
Mytheme.A6 <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "pink"),
        legend.position = "top")

theme_set(Mytheme.A6)
```

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 does not change with depth across all lakes. Ha: The mean lake temperature recorded during July does change 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.

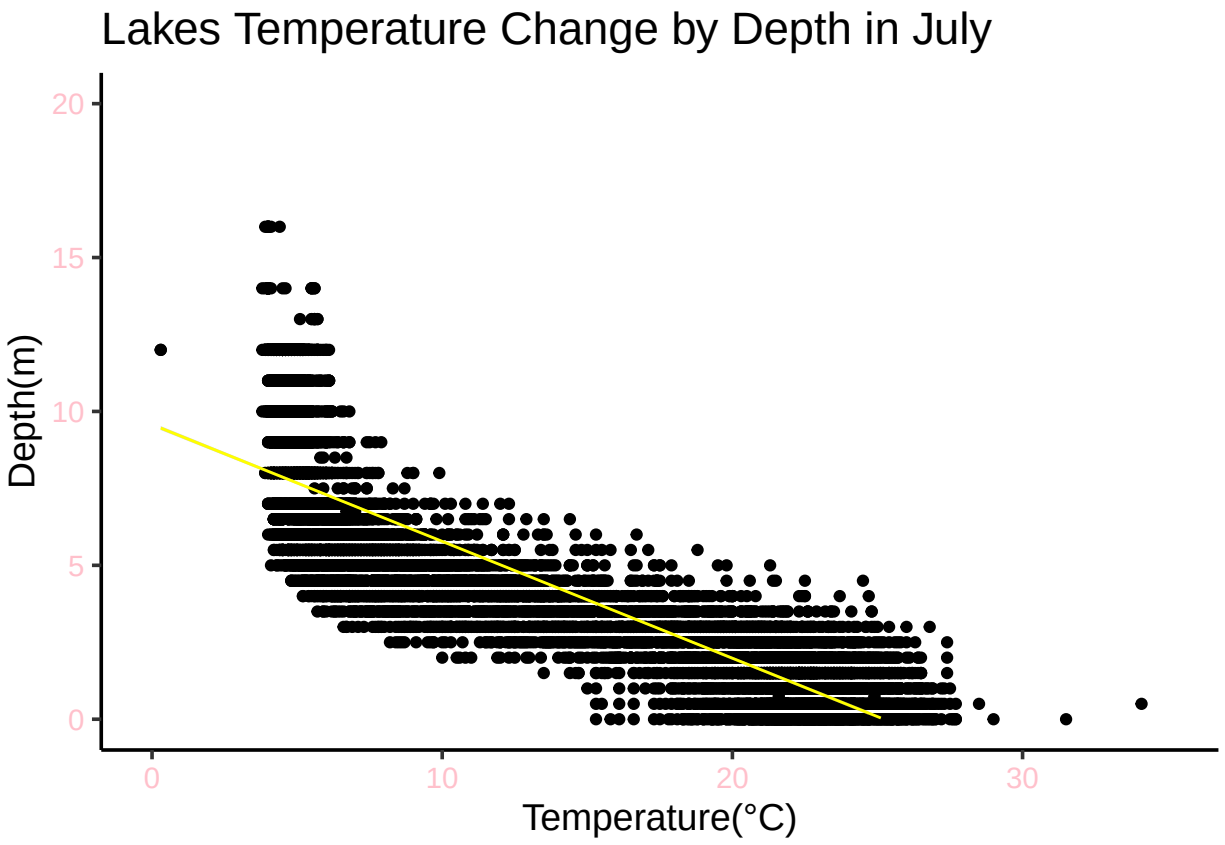
```
#4
July.NTL.LTER <-
 Raw.NTL.LTER %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  filter(month(Raw.NTL.LTER$sampledate) %in% 7)%>%
  na.omit()

#5
TemperatureDepth.NTL.LTER <-
  ggplot(July.NTL.LTER, aes(x = temperature_C, y=depth)) +
  geom_point()+
 xlim(0, 35) +
```

```
  ylim(0, 20)+
  geom_smooth(method=lm,color="yellow",size=0.5)+
                xlab("Temperature(°C)")+
                ylab("Depth(m)")+
  ggtitle("Lakes Temperature Change by Depth in July")
print(TemperatureDepth.NTL.LTER)
```

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

```
## Warning: Removed 21 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 increase as the depth decrease. The distrubution of the trend suggests a negative linear relationship between temperature and depth.

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

```
#7
depth.JulyLakes.Regression <- lm(July.NTL.LTER$temperature_C ~ July.NTL.LTER$depth)
summary(depth.JulyLakes.Regression)
```

```
##
## Call:
## lm(formula = July.NTL.LTER$temperature_C ~ July.NTL.LTER$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 ***
## July.NTL.LTER$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: There is a significant negative correlation between temperature and depth (higher temperature at lower depths), and that this model explains about 73.87 % of the total variance in temperature. The residual standard errror is 3.835 on 9726 degrees of freedom. The F-statistic is 2.75e+04 on 1 and 9726 DF. The temperature is predicted to change 1.95°C 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.

```
year4.daynum.depth <- lm(data = July.NTL.LTER, July.NTL.LTER$temperature_C ~ July.NTL.LTER$depth + July
step(year4.daynum.depth)
```

```
## Start:  AIC=26065.53
## July.NTL.LTER$temperature_C ~ July.NTL.LTER$depth + July.NTL.LTER$year4 +
##     July.NTL.LTER$daynum
##
##                         Df Sum of Sq    RSS   AIC
## <none>                               141687 26066
```

```
## - July.NTL.LTER$year4   1       101 141788 26070
## - July.NTL.LTER$daynum  1      1237 142924 26148
## - July.NTL.LTER$depth   1    404475 546161 39189

##
## Call:
## lm(formula = July.NTL.LTER$temperature_C ~ July.NTL.LTER$depth +
##     July.NTL.LTER$year4 + July.NTL.LTER$daynum, data = July.NTL.LTER)
##
## Coefficients:
##          (Intercept)   July.NTL.LTER$depth   July.NTL.LTER$year4
##              -8.57556              -1.94644               0.01134
## July.NTL.LTER$daynum
##              0.03978
```

```
#9
depth.year4<-
  lm(July.NTL.LTER$temperature_C ~ July.NTL.LTER$depth + July.NTL.LTER$year4 )
AIC(depth.year4)
```

```
## [1] 53756.97
```

```
summary(depth.year4)
```

```
##
## Call:
## lm(formula = July.NTL.LTER$temperature_C ~ July.NTL.LTER$depth +
##     July.NTL.LTER$year4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5543 -3.0227  0.0981  2.9492 13.7469
##
## Coefficients:
##                       Estimate Std. Error  t value Pr(>|t|)
## (Intercept)         -1.104769   8.629545   -0.128  0.89813
## July.NTL.LTER$depth -1.946542   0.011733 -165.906  < 2e-16 ***
## July.NTL.LTER$year4  0.011538   0.004317    2.672  0.00754 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.834 on 9725 degrees of freedom
## Multiple R-squared:  0.7389, Adjusted R-squared:  0.7389
## F-statistic: 1.376e+04 on 2 and 9725 DF,  p-value: < 2.2e-16
```

```
depth.daynum<-
  lm(July.NTL.LTER$temperature_C ~ July.NTL.LTER$depth + July.NTL.LTER$daynum )
AIC(depth.daynum)
```

```
## [1] 53679.36
```

```
summary(depth.daynum)
```

```
##
## Call:
## lm(formula = July.NTL.LTER$temperature_C ~ July.NTL.LTER$depth +
##     July.NTL.LTER$daynum)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6174 -2.9809  0.0845  2.9681 13.4406
##
## Coefficients:
##                        Estimate Std. Error  t value Pr(>|t|)
## (Intercept)          14.088588   0.855505   16.468   <2e-16 ***
## July.NTL.LTER$depth  -1.946111   0.011685 -166.541   <2e-16 ***
## July.NTL.LTER$daynum  0.039836   0.004318    9.225   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.818 on 9725 degrees of freedom
## Multiple R-squared:  0.741,  Adjusted R-squared:  0.741
## F-statistic: 1.391e+04 on 2 and 9725 DF,  p-value: < 2.2e-16
```

```
year4.daynum<-
  lm(July.NTL.LTER$temperature_C ~ July.NTL.LTER$daynum + July.NTL.LTER$year4 )
AIC(year4.daynum)
```

```
## [1] 66798.34
```

```
summary(year4.daynum)
```

```
##
## Call:
## lm(formula = July.NTL.LTER$temperature_C ~ July.NTL.LTER$daynum +
##     July.NTL.LTER$year4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.279  -7.158  -2.591   8.072  21.402
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -2.827705  16.944033  -0.167    0.867
## July.NTL.LTER$daynum  0.040484   0.008475   4.777 1.81e-06 ***
## July.NTL.LTER$year4   0.003779   0.008439   0.448    0.654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.494 on 9725 degrees of freedom
## Multiple R-squared:  0.002363,   Adjusted R-squared:  0.002158
## F-statistic: 11.52 on 2 and 9725 DF,  p-value: 1.007e-05
```

```
#10
```

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: depth and daynum

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
```

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

    Answer:

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.
```

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

```
#15
```

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:

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:

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?

    Answer: