

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

Matthew Vining

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
getwd()

## [1] "/home/guest/R/EDA-Fall2022"

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   1.2.0      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(agricolae)
library(lubridate)

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

```

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

class(chemistry_physics$sampleddate)

## [1] "factor"

chemistry_physics$sampleddate <- as.Date(chemistry_physics$sampleddate, format = "%m/%d/%y")
class(chemistry_physics$sampleddate)

## [1] "Date"

# 2

# Set Standard Theme
mytheme <- theme_classic(base_size = 10) + theme(axis.text = element_text(color = "black"),
  legend.position = "top")
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: There is no significant change in mean lake temperature recorded during July as the depth across all lakes changes. Ha: There is a significant change in mean lake temperature recorded during July as the depth across all lakes changes.
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.

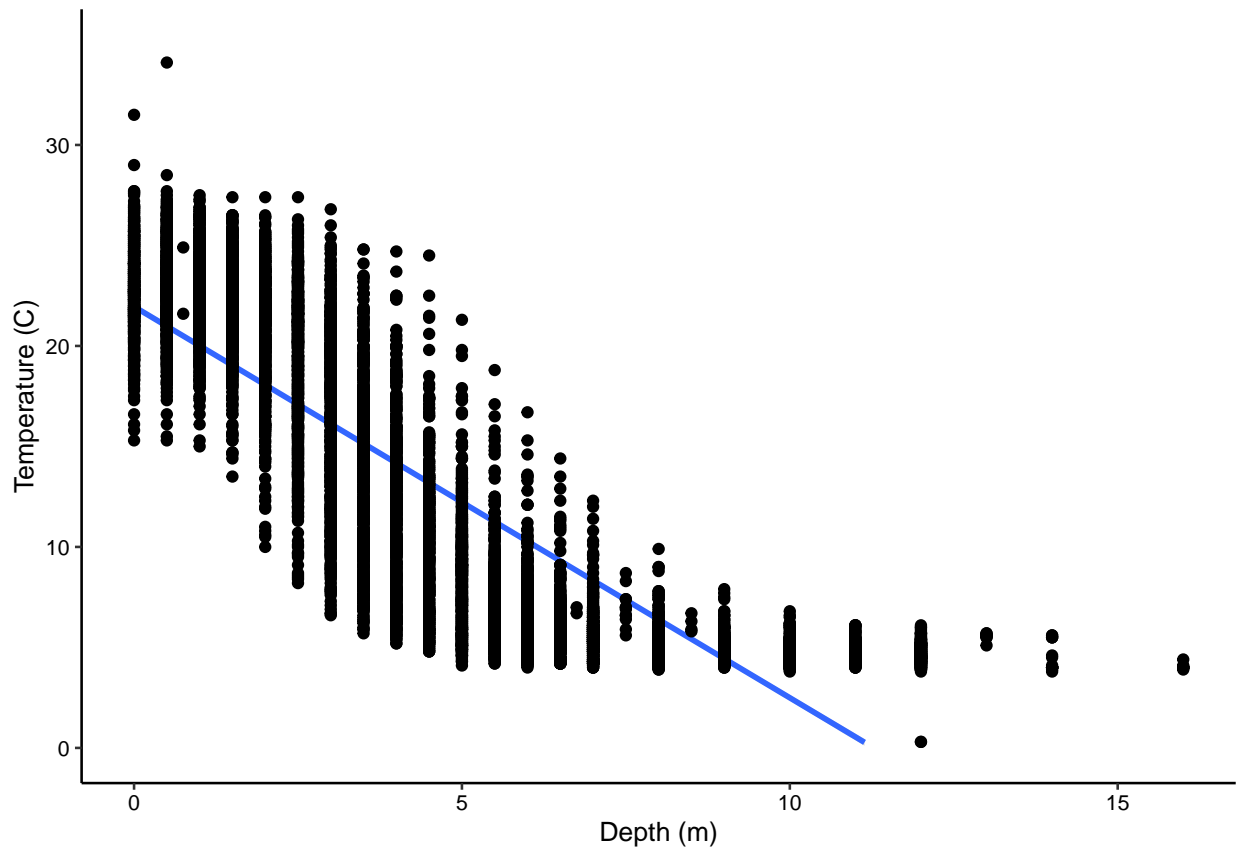
```

# 4
wrangle.P4 <- chemistry_physics %>%
  filter(month(sampleddate) == 7) %>%
  select(lakename:daynum, depth, temperature_C) %>%
  drop_na()

# 5
plot.5 <- ggplot(wrangle.P4, aes(x = depth, y = temperature_C)) + geom_smooth(method = "lm") +
  ylim(0, 35) + geom_point() + ylab("Temperature (C)") + xlab("Depth (m)")
print(plot.5)

## `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 anything about the linearity of this trend?

Answer: The figure suggests that as depth of the lakes increase, the temperature decreases, forming an inverse relationship. Looking at the distribution of points, it appears to start off linear, but possibly turns quadratic or to a higher degree than just a linear relationship after a depth of about 5m.

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

```
# 7
P7.regression <- lm(wrangle.P4$temperature_C ~ wrangle.P4$depth)
summary(P7.regression)

##
## Call:
## lm(formula = wrangle.P4$temperature_C ~ wrangle.P4$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 ***
## wrangle.P4$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: Based on the regression output, the model predicts that for every unit increase in depth, temperature decreases by 1.946 degrees C on average. Our R-squared value tells us that about 74% of the variability in temperature is explained by changes in depth. The degrees of freedom on which this finding is based is 9726 degrees of freedom. Looking at the p-value of 2.2e-16, we can assume that there is statistical significance assuming that the confidence level is 95%.

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
P9AIC <- lm(data = wrangle.P4, temperature_C ~ depth + year4 + daynum)
step(P9AIC) #try for lowest score

## 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 = wrangle.P4)
##
## Coefficients:
## (Intercept)      depth      year4      daynum
##   -8.57556    -1.94644     0.01134     0.03978
# results indicate best variables are daynum and depth
```

```
# 10
P10.regression <- lm(data = wrangle.P4, temperature_C ~ depth + daynum)
summary(P10.regression)

##
## Call:
## lm(formula = temperature_C ~ depth + daynum, data = wrangle.P4)
```

```
##
## 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 ***
## depth       -1.946111   0.011685  -166.541  <2e-16 ***
## 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
```

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 suggested by the AIC method are depth (m) and daynum. The model explains about 74% of the observed variance and both depth and daynum are significant variables based on their p-values and a 95% confidence level. While only slightly, it is an improvement over our simple regression based off the R-squared value and significance values (p-values).

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
Lakes.anova <- aov(data = wrangle.P4, temperature_C ~ lakename)
summary(Lakes.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

Lakes.regression <- lm(data = wrangle.P4, temperature_C ~ lakename)
summary(Lakes.regression)

##
## Call:
## lm(formula = temperature_C ~ lakename, data = wrangle.P4)
##
## 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 ***
## lakenamCrampton Lake   -2.3145    0.7699  -3.006 0.002653 **
## lakenamEast Long Lake  -7.3987    0.6918 -10.695 < 2e-16 ***
## lakenamHummingbird Lake -6.8931    0.9429  -7.311 2.87e-13 ***
## lakenamPaul Lake      -3.8522    0.6656  -5.788 7.36e-09 ***
## lakenamPeter Lake     -4.3501    0.6645  -6.547 6.17e-11 ***
## lakenamTuesday Lake   -6.5972    0.6769  -9.746 < 2e-16 ***
## lakenamWard Lake      -3.2078    0.9429  -3.402 0.000672 ***
## lakenamWest 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
```

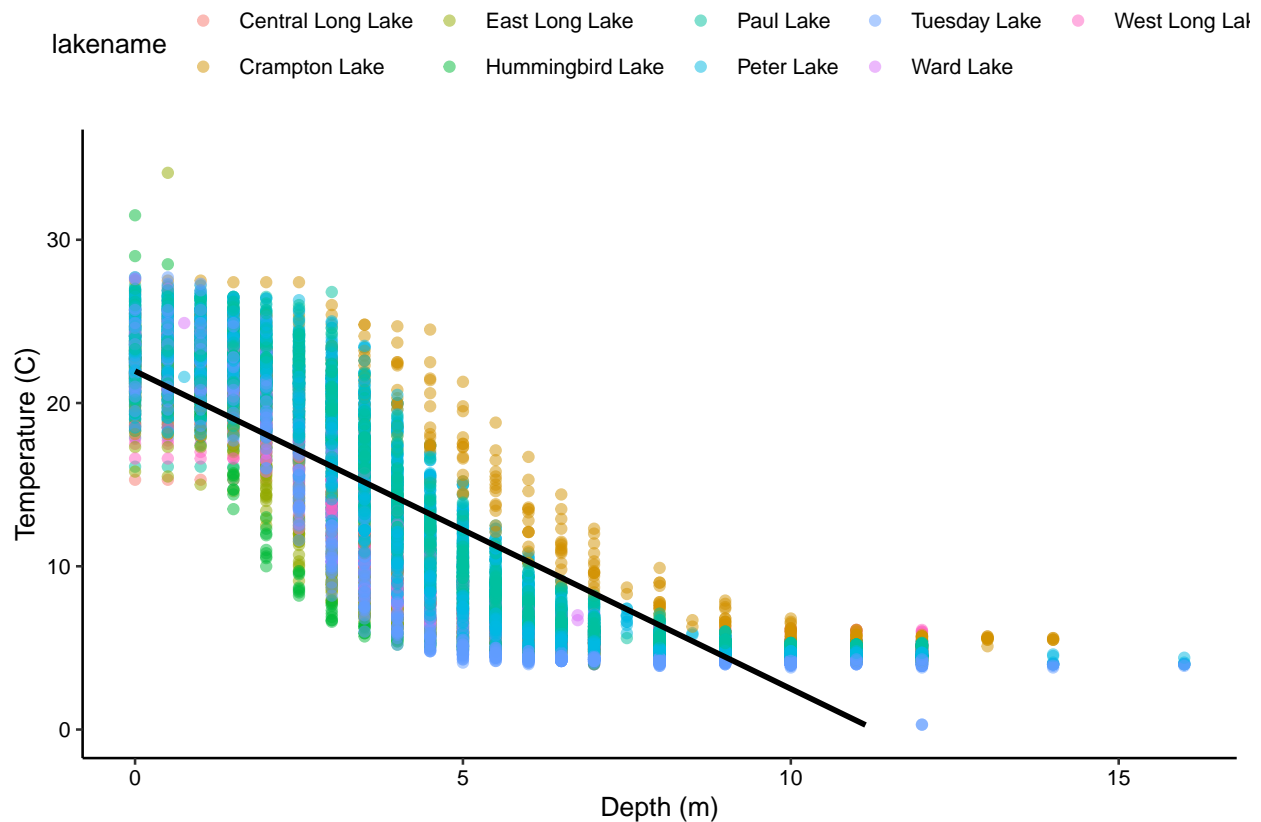
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

Answer: Yes, there appears to be evidence indicating significant differences by lake. Each lake has a significant p-value associated with it, indicating the model is a good predictor with it included as an explanatory variable. Based on the estimates, lakes such as West Long Lake, Tuesday Lake, Hummingbird Lake, and East Long Lake have the coldest temperatures in the group. Lakes such as Crampton Lake, Paul Lake, and Ward Lake have the warmest temperatures of the group.

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.
Plot_P14 <- ggplot(wrangle.P4, aes(x = depth, y = temperature_C, color = lakenam)) +
  geom_point(alpha = 0.5) + geom_smooth(method = lm, col = "black", se = FALSE) +
  ylim(0, 35) + ylab("Temperature (C)") + xlab("Depth (m)")
print(Plot_P14)

## `geom_smooth()` using formula 'y ~ x'
## Warning: Removed 24 rows containing missing values (geom_smooth).
```



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

15

TukeyHSD(Lakes.anova)

```
## Tukey multiple comparisons of means
## 95% family-wise confidence level
##
## Fit: aov(formula = temperature_C ~ lakename, data = wrangle.P4)
##
## $lakename
##
```

	diff	lwr	upr	p adj
Crampton Lake-Central Long Lake	-2.3145195	-4.7031913	0.0741524	0.0661566
East Long Lake-Central Long Lake	-7.3987410	-9.5449411	-5.2525408	0.0000000
Hummingbird Lake-Central Long Lake	-6.8931304	-9.8184178	-3.9678430	0.0000000
Paul Lake-Central Long Lake	-3.8521506	-5.9170942	-1.7872070	0.0000003
Peter Lake-Central Long Lake	-4.3501458	-6.4115874	-2.2887042	0.0000000
Tuesday Lake-Central Long Lake	-6.5971805	-8.6971605	-4.4972005	0.0000000
Ward Lake-Central Long Lake	-3.2077856	-6.1330730	-0.2824982	0.0193405
West Long Lake-Central Long Lake	-6.0877513	-8.2268550	-3.9486475	0.0000000
East Long Lake-Crampton Lake	-5.0842215	-6.5591700	-3.6092730	0.0000000
Hummingbird Lake-Crampton Lake	-4.5786109	-7.0538088	-2.1034131	0.0000004
Paul Lake-Crampton Lake	-1.5376312	-2.8916215	-0.1836408	0.0127491
Peter Lake-Crampton Lake	-2.0356263	-3.3842699	-0.6869828	0.0000999
Tuesday Lake-Crampton Lake	-4.2826611	-5.6895065	-2.8758157	0.0000000
Ward Lake-Crampton Lake	-0.8932661	-3.3684639	1.5819317	0.9714459
West Long Lake-Crampton Lake	-3.7732318	-5.2378351	-2.3086285	0.0000000

## Hummingbird Lake-East Long Lake	0.5056106	-1.7364925	2.7477137	0.9988050
## Paul Lake-East Long Lake	3.5465903	2.6900206	4.4031601	0.0000000
## Peter Lake-East Long Lake	3.0485952	2.2005025	3.8966879	0.0000000
## Tuesday Lake-East Long Lake	0.8015604	-0.1363286	1.7394495	0.1657485
## Ward Lake-East Long Lake	4.1909554	1.9488523	6.4330585	0.0000002
## West Long Lake-East Long Lake	1.3109897	0.2885003	2.3334791	0.0022805
## Paul Lake-Hummingbird Lake	3.0409798	0.8765299	5.2054296	0.0004495
## Peter Lake-Hummingbird Lake	2.5429846	0.3818755	4.7040937	0.0080666
## Tuesday Lake-Hummingbird Lake	0.2959499	-1.9019508	2.4938505	0.9999752
## Ward Lake-Hummingbird Lake	3.6853448	0.6889874	6.6817022	0.0043297
## West Long Lake-Hummingbird Lake	0.8053791	-1.4299320	3.0406903	0.9717297
## Peter Lake-Paul Lake	-0.4979952	-1.1120620	0.1160717	0.2241586
## Tuesday Lake-Paul Lake	-2.7450299	-3.4781416	-2.0119182	0.0000000
## Ward Lake-Paul Lake	0.6443651	-1.5200848	2.8088149	0.9916978
## West Long Lake-Paul Lake	-2.2356007	-3.0742314	-1.3969699	0.0000000
## Tuesday Lake-Peter Lake	-2.2470347	-2.9702236	-1.5238458	0.0000000
## Ward Lake-Peter Lake	1.1423602	-1.0187489	3.3034693	0.7827037
## West Long Lake-Peter Lake	-1.7376055	-2.5675759	-0.9076350	0.0000000
## Ward Lake-Tuesday Lake	3.3893950	1.1914943	5.5872956	0.0000609
## West Long Lake-Tuesday Lake	0.5094292	-0.4121051	1.4309636	0.7374387
## West Long Lake-Ward Lake	-2.8799657	-5.1152769	-0.6446546	0.0021080

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: Both Paul Lake and Ward Lake have the same mean temperatures as Peter Lake, statistically speaking. We know this because of the high p-values that indicate no significance in differences in means at a 95% confidence level. Based on the significance of the p-values, it appears not one lake is statistically distinct from all other lakes in terms of mean temperature. Each lake has at least one other lake in common where their mean temperatures are not statistically distinct from one another.

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 just looking at one categorical variable with two categories, we could perform a t-test to understand whether they have distinct mean temperatures.

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 your answer for part 16?

```
wrangle.P18 <- wrangle.P4 %>%
  filter(lakename == "Crampton Lake" | lakename == "Ward Lake")
```

```
P18.ttest <- t.test(wrangle.P18$temperature_C ~ wrangle.P18$lakename)
P18.ttest
```

```
##
## Welch Two Sample t-test
##
## data: wrangle.P18$temperature_C by wrangle.P18$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 not equal to 0
## 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 results of our two-sample T-test indicate that the mean temperatures for the lake are not statistically different. We know this because the p-value is greater than 0.05 for a 95% confidence interval and the t-value is within the confidence interval range. This outcome matches the conclusion from part 15-17 where Crampton and Ward Lakes had no statistical significance to show any difference in mean temperatures.