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

Samantha White-Murillo

Spring 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
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.5
## 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(lubridate)
library(here)
```

```
## here() starts at /home/guest/EDA_Spring2024_SamanthaWM
```

```
library(cowplot)
```

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

```
library(agricolae)

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

#2
mytheme <- theme_classic(base_size = 12) +
  theme(plot.title = element_text(face = "bold", size = 12,
color = "black"),
axis.title = element_text(color = "darkblue", )
) +
theme(legend.position = "bottom")
```

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 recorded during July does change with depth across all lakes. Ha: Mean lake temperature recorded during July does change with depth across some lakes. Hb: Mean lake temperature recorded during July does NOT 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
Lake_ChemsPhys_processed <- Lake_ChemsPhys %>%
  filter(month(sampledate) == 7) %>%
  select(lakename, year4, daynum, depth, temperature_C) %>%
  drop_na()
```

```r
#5
Lake_ChemsPhys_plot <- ggplot(Lake_ChemsPhys_processed,
                              aes(x = temperature_C,
                                  y = depth)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "orange") +
  scale_y_log10() +
  xlim(0, 35) +
  labs(x = "Temperature (°C)",
       y = "Depth") +
  mytheme
print(Lake_ChemsPhys_plot)
```

```
## Warning: Transformation introduced infinite values in continuous y-axis
## Transformation introduced infinite values in continuous y-axis

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

## Warning: Removed 533 rows containing non-finite values (`stat_smooth()`).
```

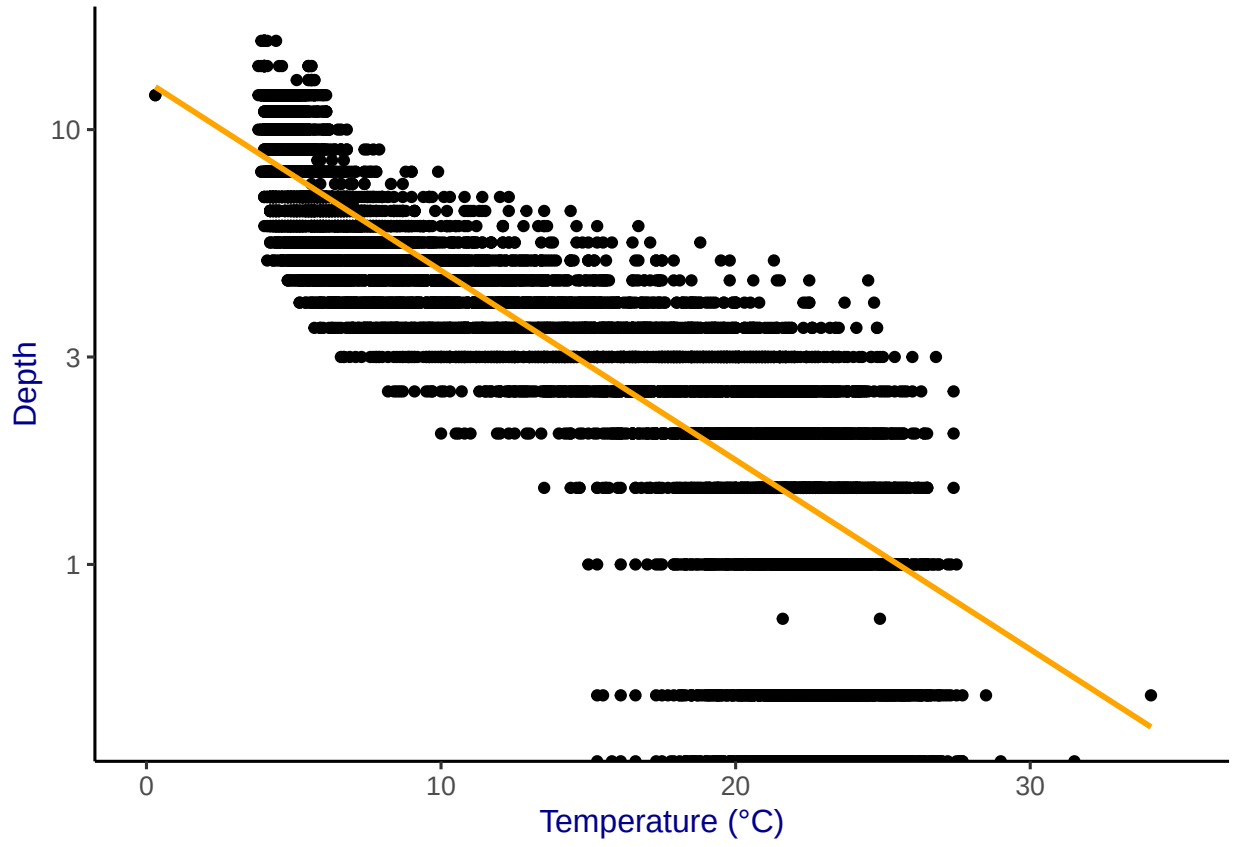


Figure 1: Temperature vs. Depht

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: It shows that the greater the depth, the lower the temperature. Nevertheless, the correlation is not directly proportional, meaning that at a medium depth range the temperature is relatively constant.

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

```
#7
Lake_ChemsPhys_linear <- lm(
  data = Lake_ChemsPhys_processed,
  temperature_C ~ depth
)

summary(Lake_ChemsPhys_linear)
```

```
##
## Call:
## lm(formula = temperature_C ~ depth, data = Lake_ChemsPhys_processed)
##
## 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 estimated mean temperature when the depth is zero is 21.96. Also, the the estimated change in temperature for a one-unit increase in depth is -1.95. On the other hand, approximately 73.87% of the variability in temperature is caused by changes in depth. The F-statistic is 2.75e+04 with a very low p-value, indicating that the regression model is statistically significant.

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.

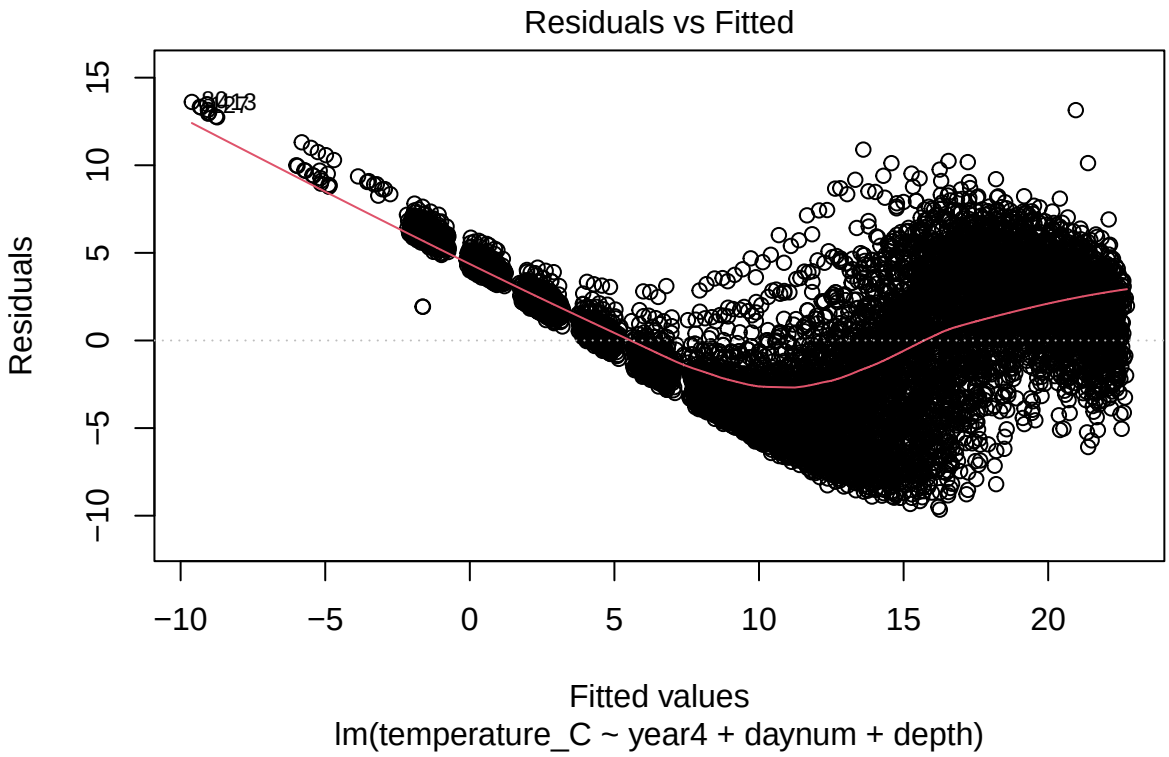
```
#9
library(MASS)
```

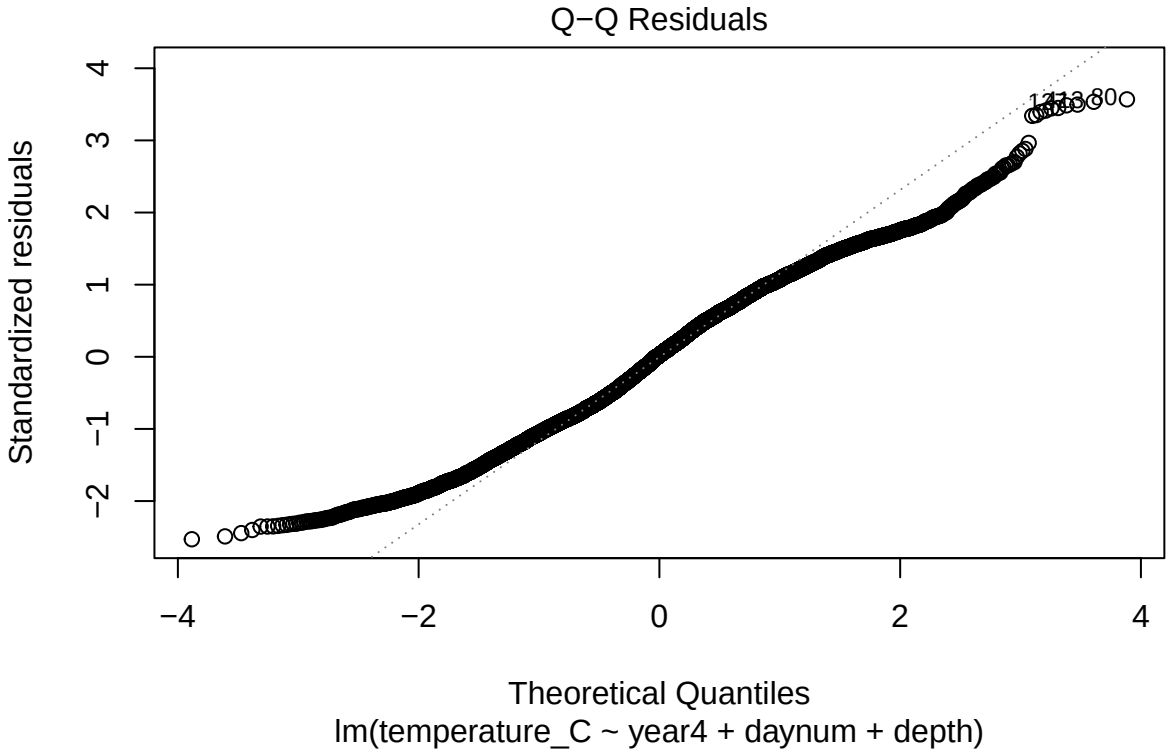
```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

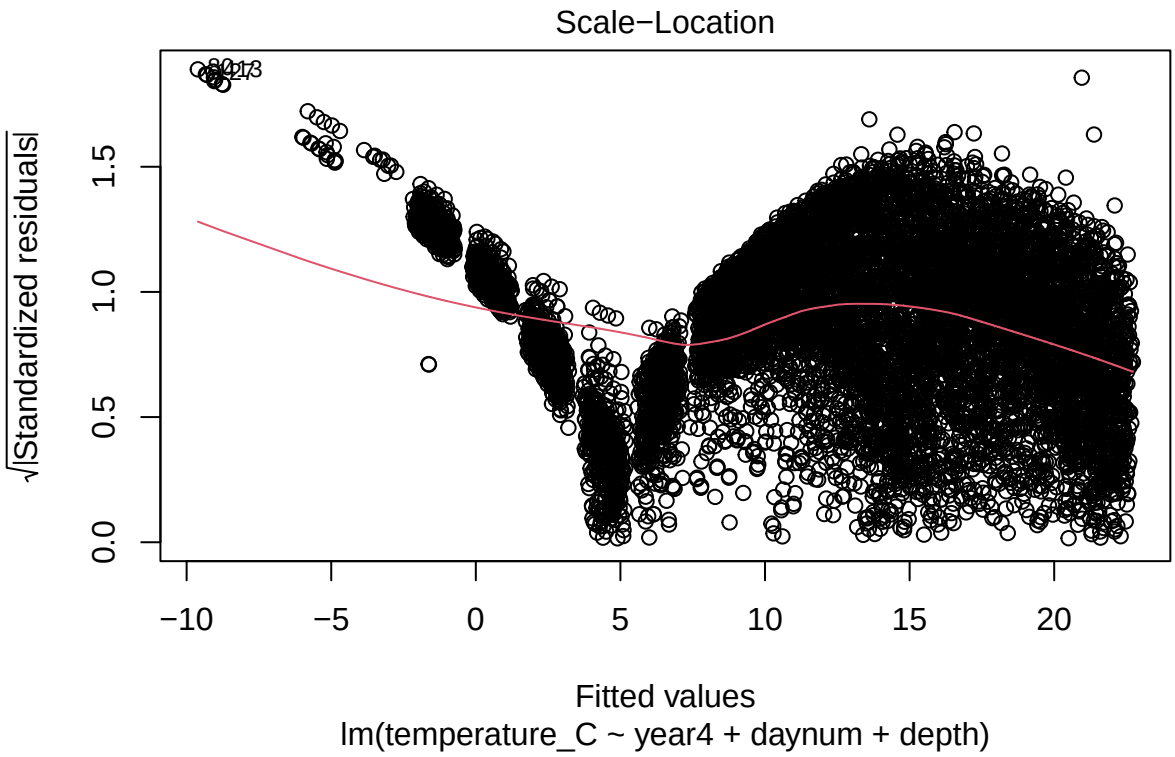
```
Lake_ChemsPhys_AIC <- lm(
  data = Lake_ChemsPhys_processed,
  temperature_C ~ year4 + daynum + depth
)
model <- stepAIC(Lake_ChemsPhys_AIC,
              direction = "both", trace = 0)
summary(model)
```

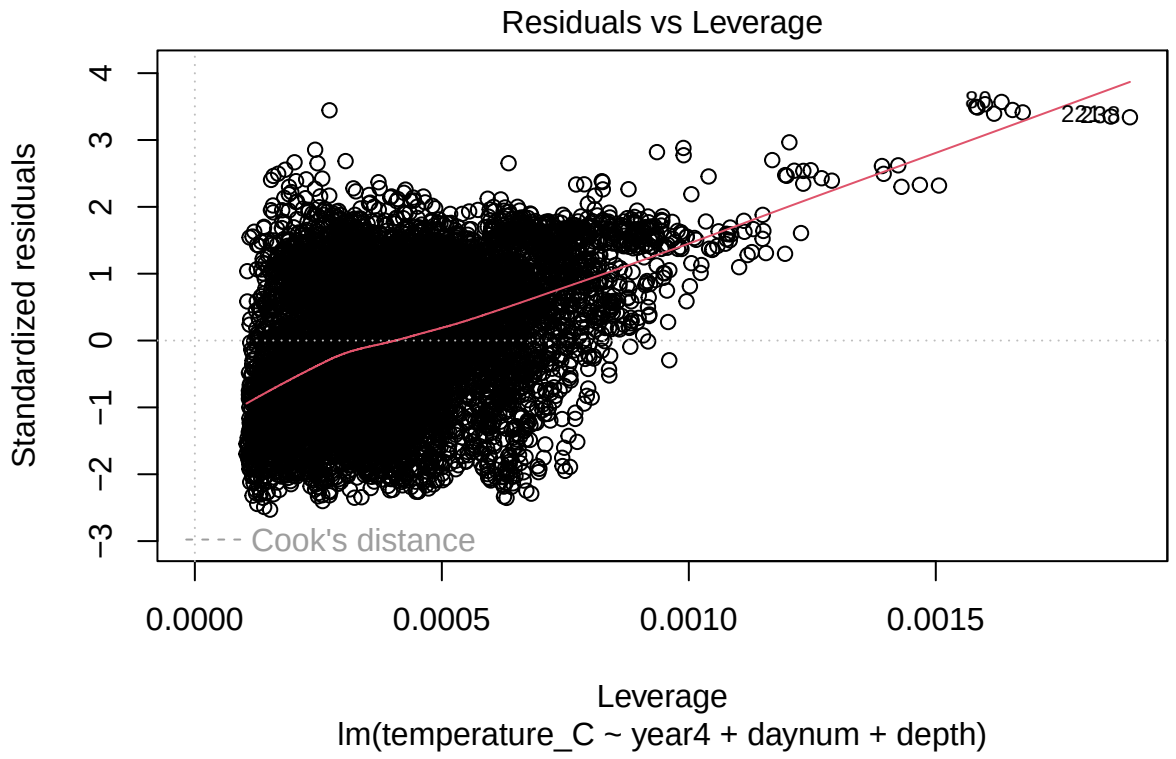
```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = Lake_ChemsPhys_processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6536 -3.0000  0.0902  2.9658 13.6123
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -8.575564   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
```

```
plot(Lake_ChemsPhys_AIC)
```


Residuals vs Fitted
Residuals
15
10
5
0
−5
−10
−10
−5
0
5
10
15
20
Fitted values
lm(temperature_C ~ year4 + daynum + depth)


Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
3
4
−4
−2
0
2
4
Theoretical Quantiles
lm(temperature_C ~ year4 + daynum + depth)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−10
−5
0
5
10
15
20
Fitted values
lm(temperature_C ~ year4 + daynum + depth)

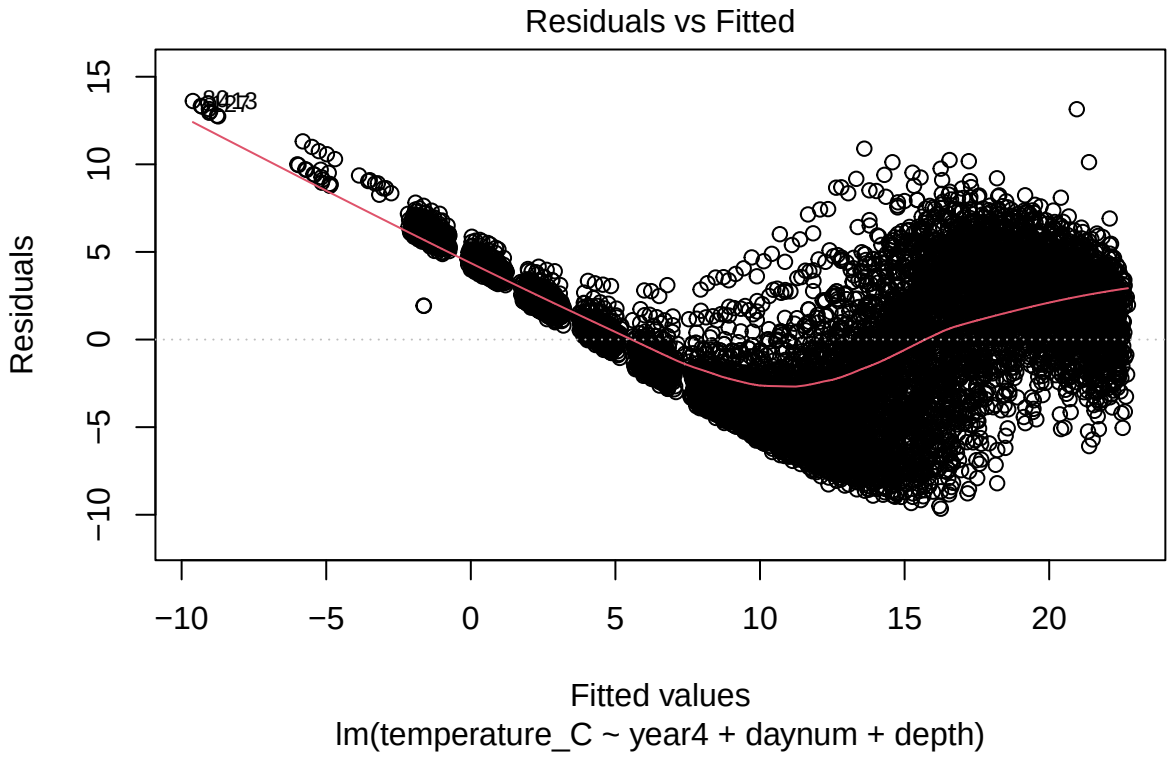


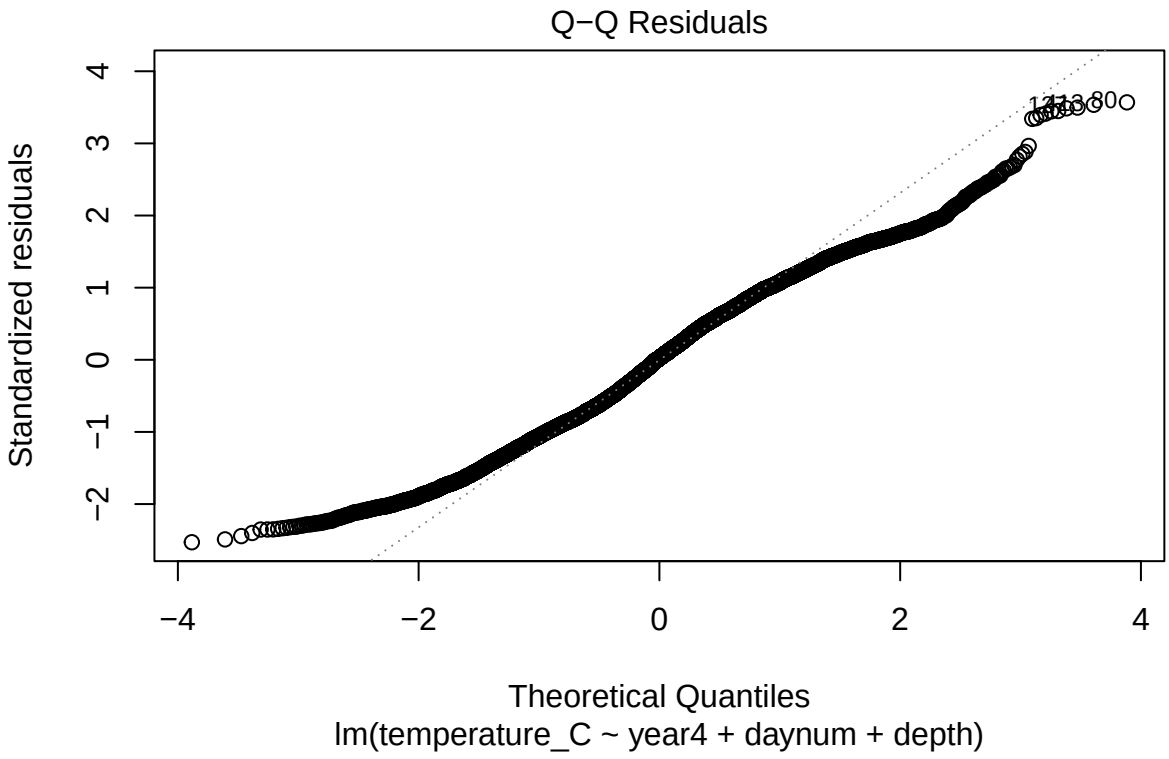

```r
#10
recommended_model <- lm(
  data = Lake_ChemsPhys_processed,
  temperature_C ~ year4 + daynum + depth
  )
summary(recommended_model)
```

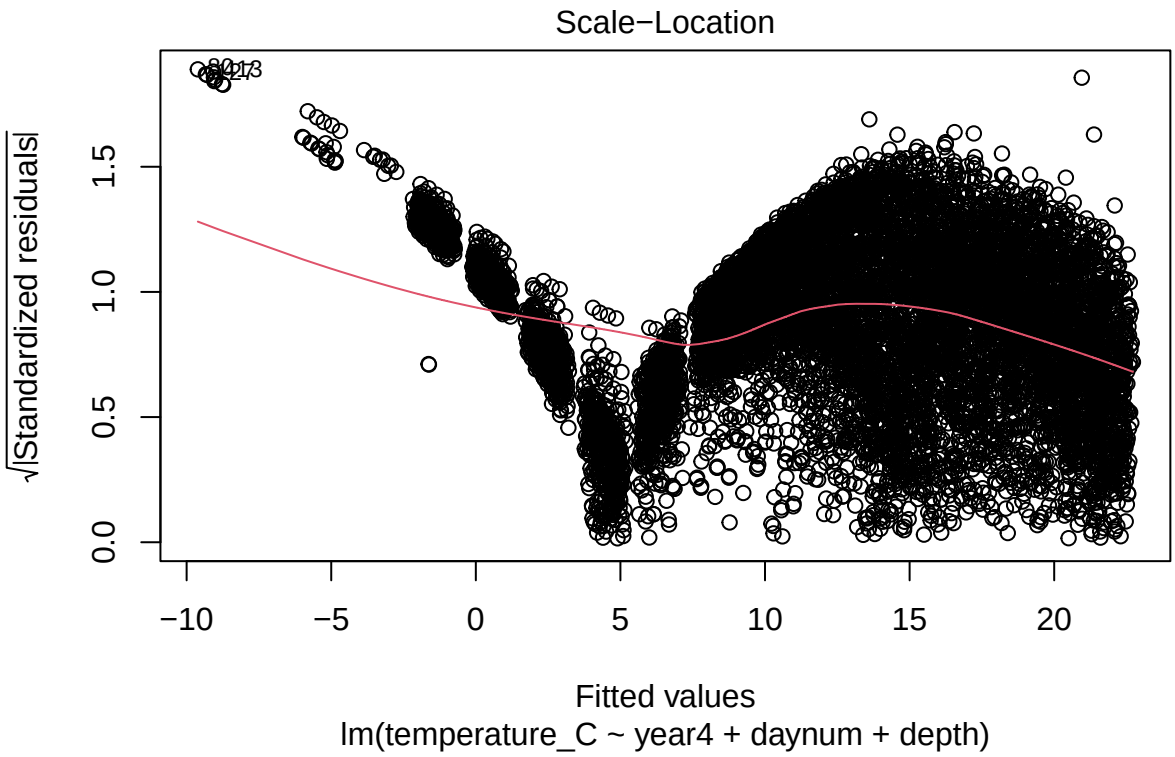
```
##
## Call:
## lm(formula = temperature_C ~ year4 + daynum + depth, data = Lake_ChemsPhys_processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6536 -3.0000  0.0902  2.9658 13.6123
##
## Coefficients:
##              Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -8.575564   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
```

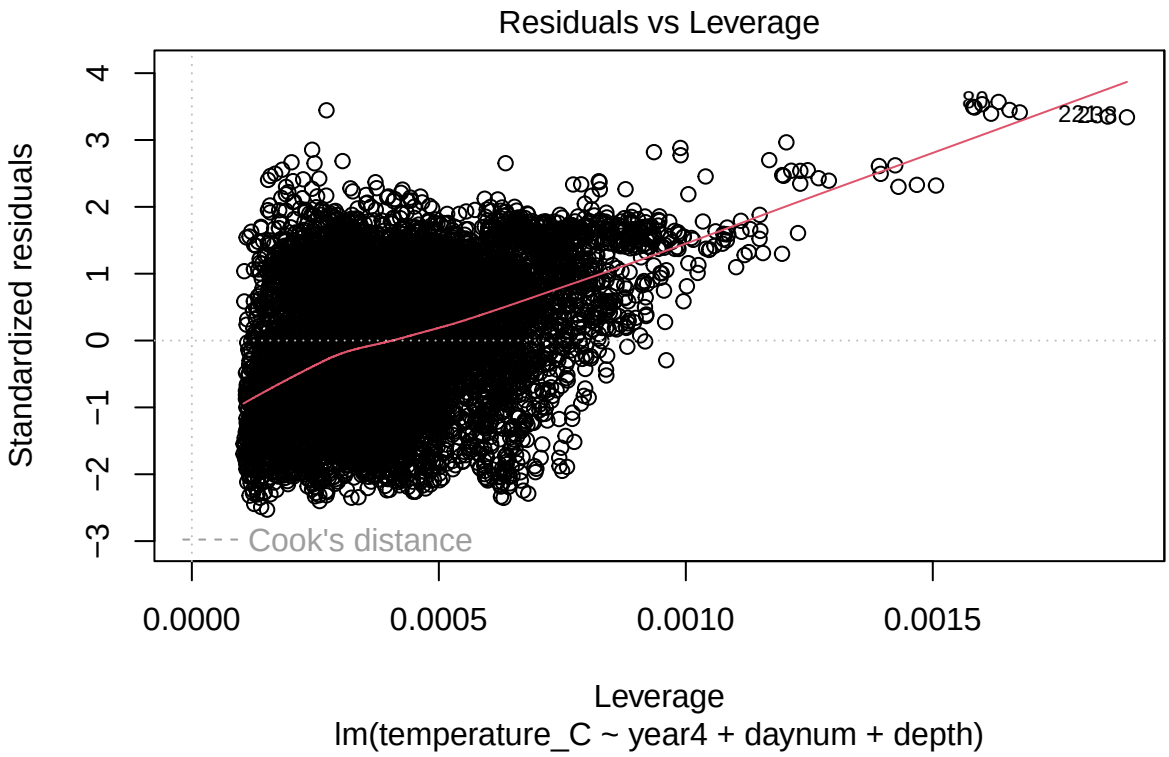
```
## F-statistic:  9283 on 3 and 9724 DF,  p-value: < 2.2e-16
```

```
plot(recommended_model)
```




Q−Q Residuals
Standardized residuals
−2
−1
0
1
2
3
4
−4
−2
0
2
4
Theoretical Quantiles
lm(temperature_C ~ year4 + daynum + depth)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
−10
−5
0
5
10
15
20
Fitted values
lm(temperature_C ~ year4 + daynum + depth)



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:

Analysis of Variance

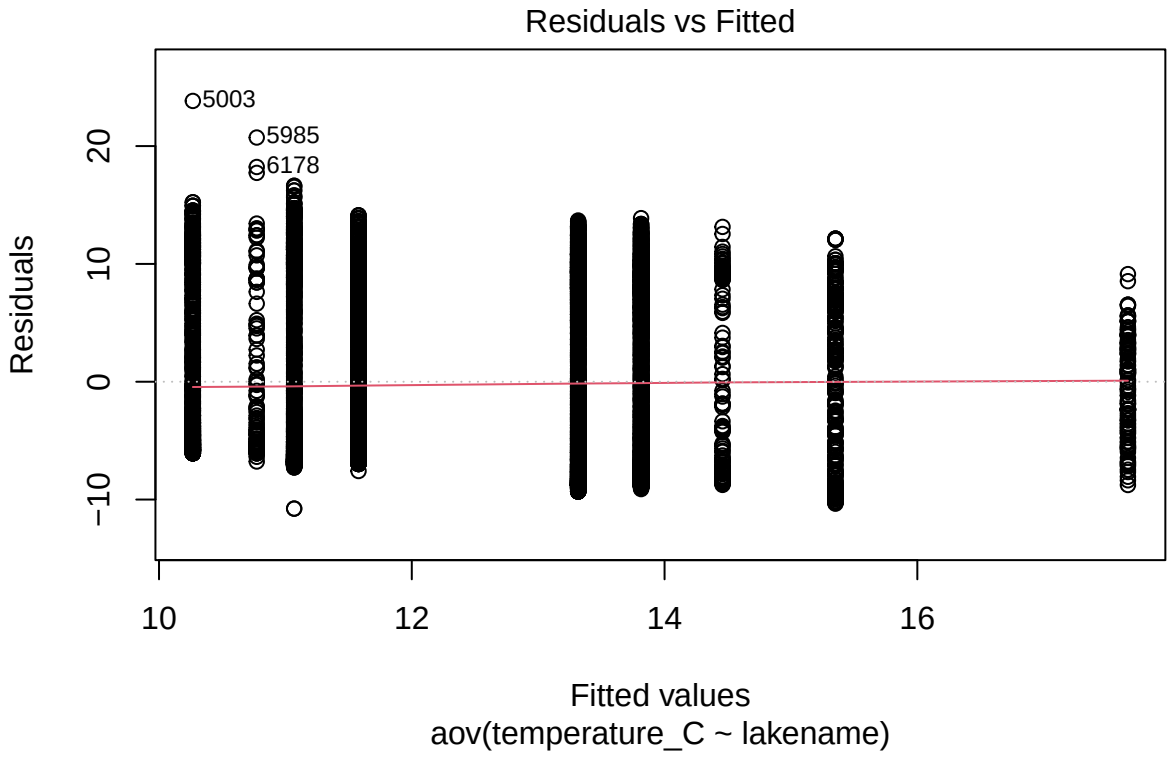
12. Now we want to see whether the different lakes have, on average, different temperatures in the month of July. Run an ANOVA test to complete this analysis. (No need to test assumptions of normality or similar variances.) Create two sets of models: one expressed as an ANOVA models and another expressed as a linear model (as done in our lessons).

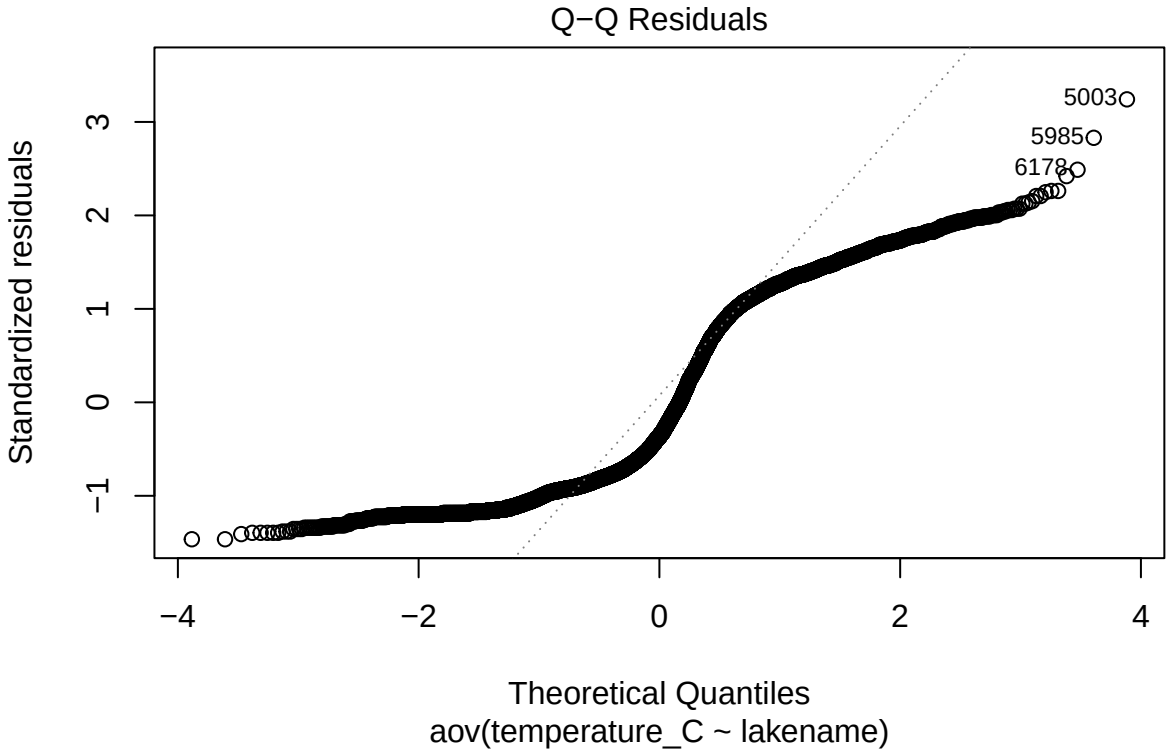
```
#12a
ChemsPhys_anova_model <- aov(data = Lake_ChemsPhys_processed,
                  temperature_C ~ lakename)

summary(ChemsPhys_anova_model)
```

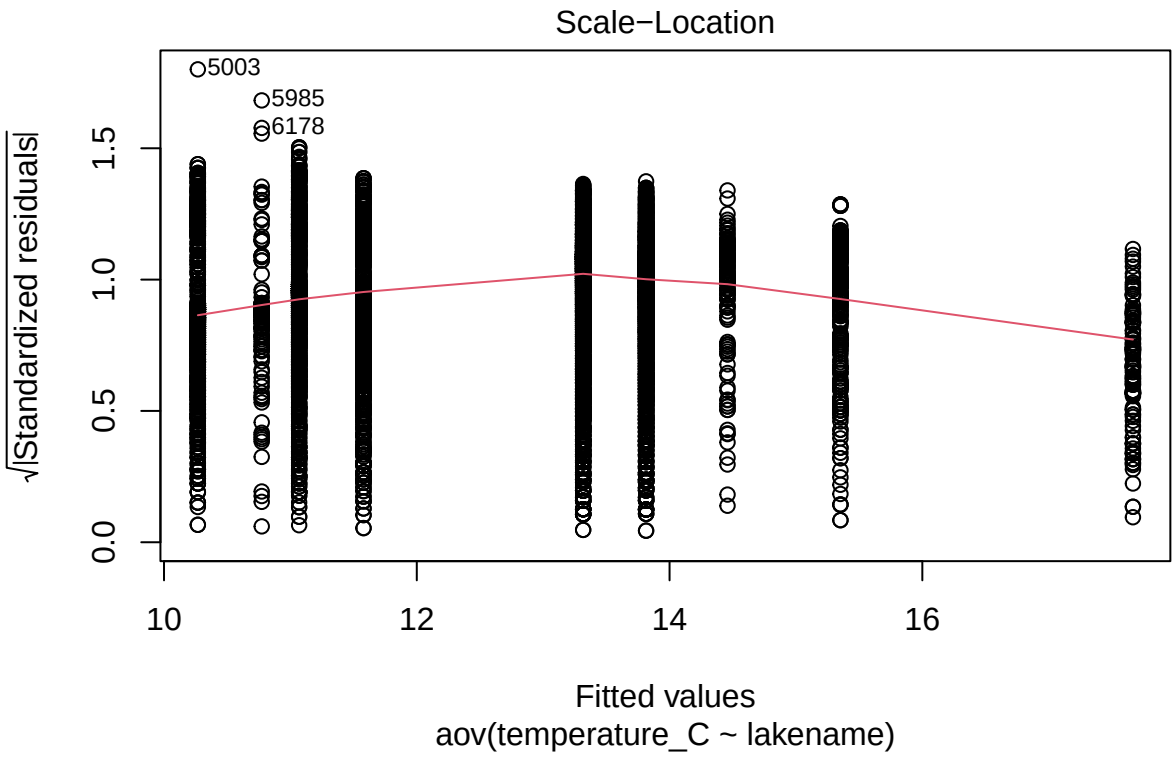
```
##               Df Sum Sq Mean Sq F value Pr(>F)
## lakename       8  21642  2705.2      50 <2e-16 ***
```

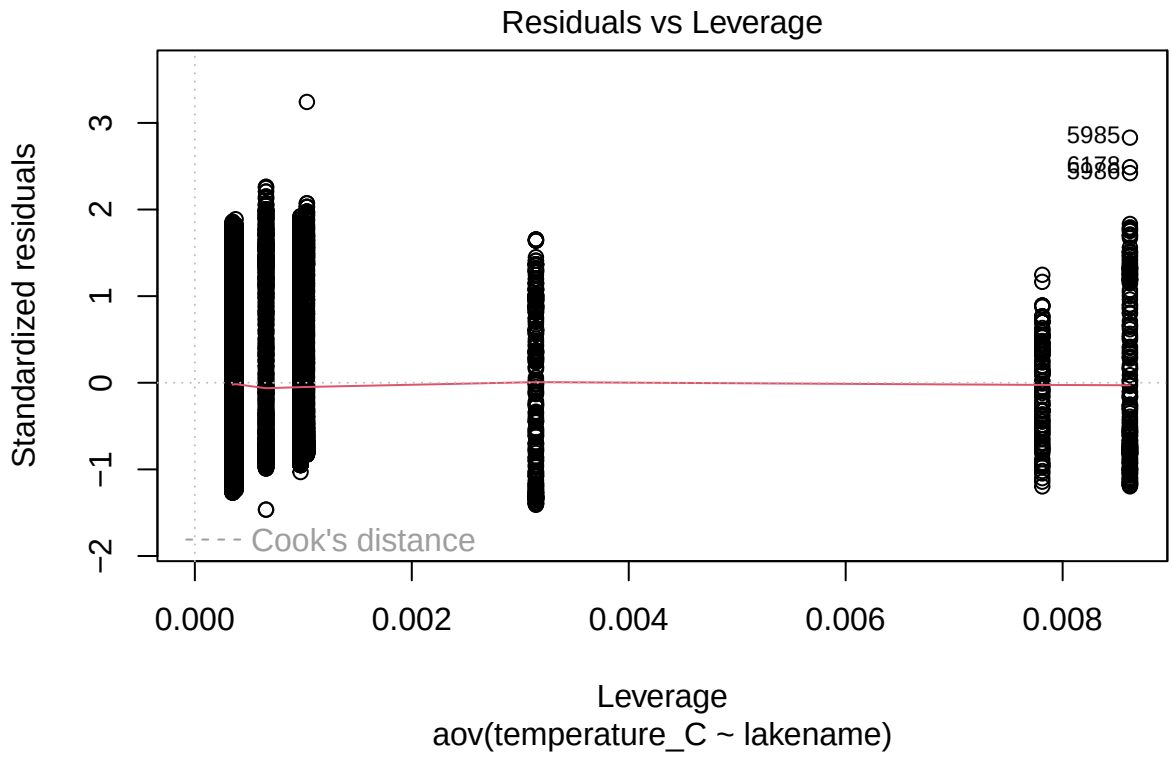
```
## Residuals   9719 525813    54.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(ChemsPhys_anova_model)
```




Q−Q Residuals
Standardized residuals
−1
0
1
2
3
5003
5985
6178
−4
−2
0
2
4
Theoretical Quantiles
aov(temperature_C ~ lakename)


Scale−Location
5003
5985
6178
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
12
14
16
Fitted values
aov(temperature_C ~ lakename)

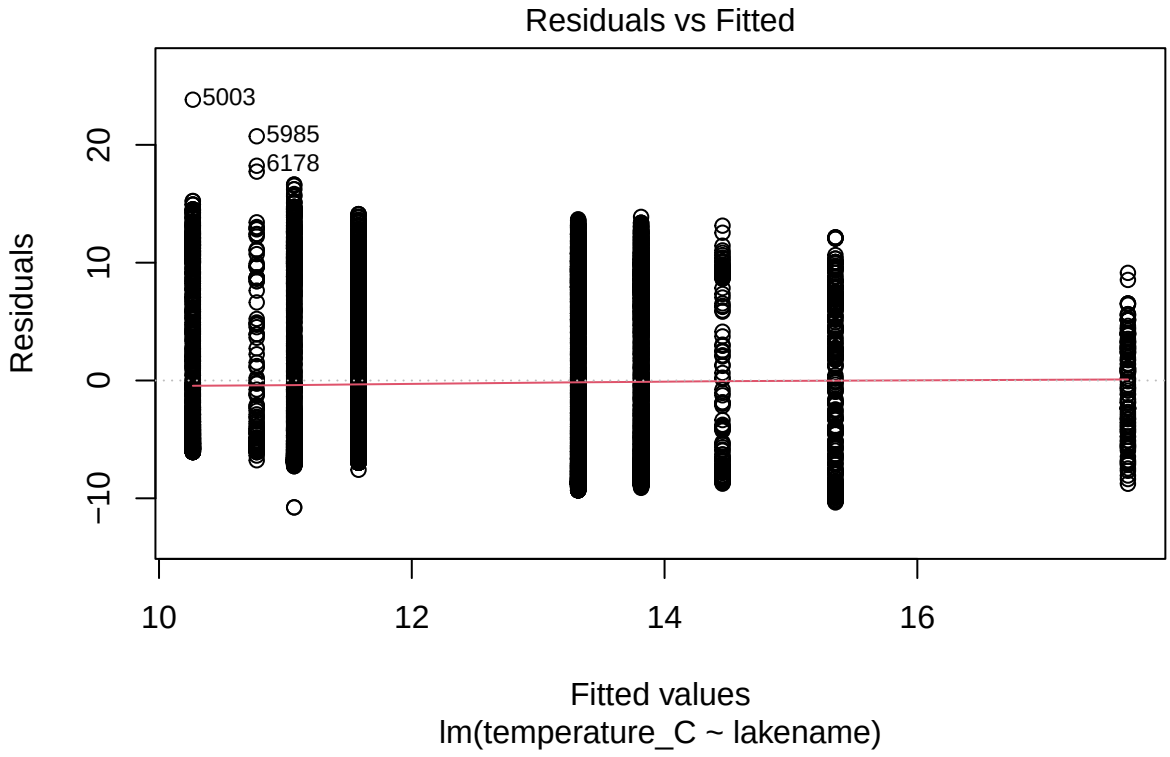


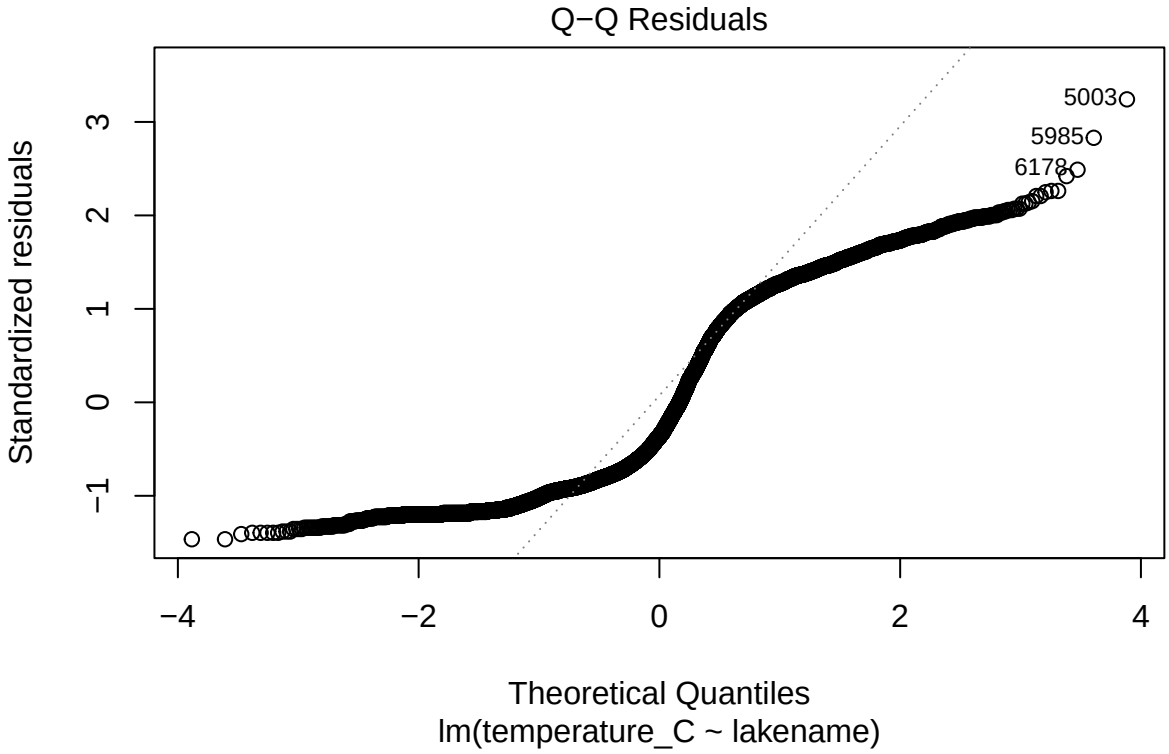

```
#12b
ChemsPhys_lm_model <- lm(data = Lake_ChemsPhys_processed,
               temperature_C ~ lakename)

summary(ChemsPhys_lm_model)
```

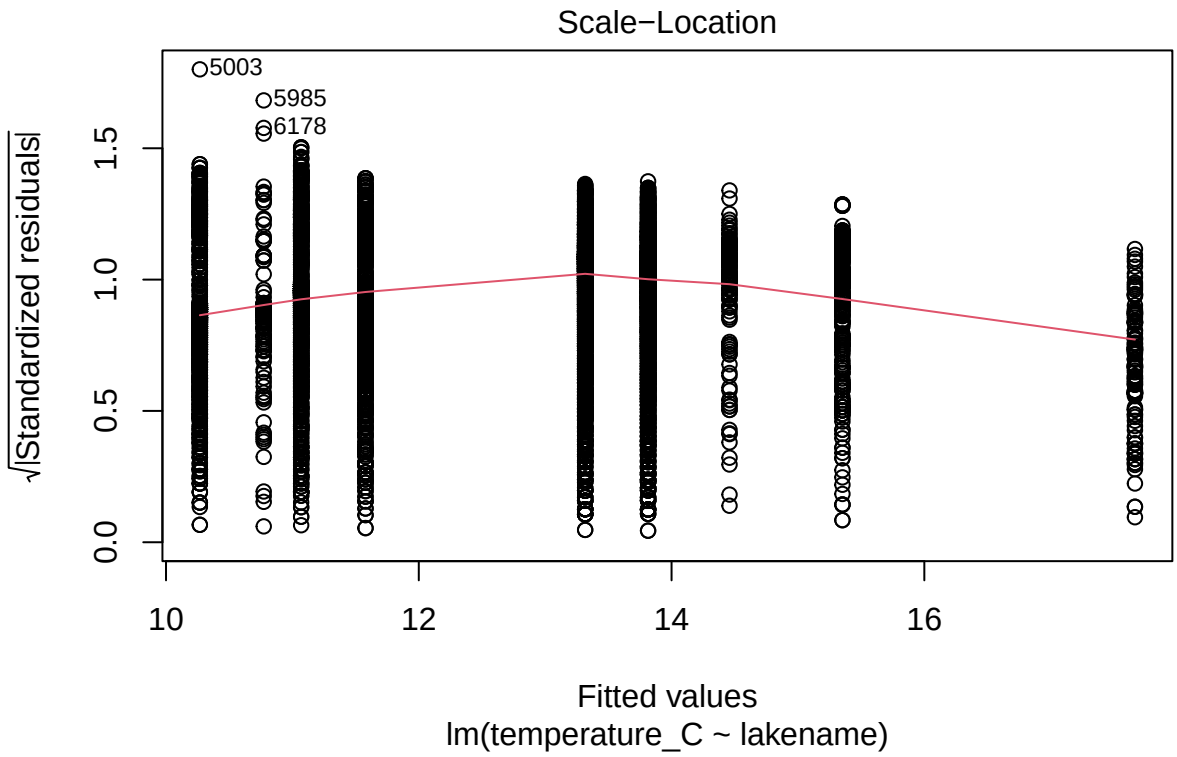
```
##
## Call:
## lm(formula = temperature_C ~ lakename, data = Lake_ChemsPhys_processed)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.769  -6.614  -2.679   7.684  23.832
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 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 ***
## ---
```

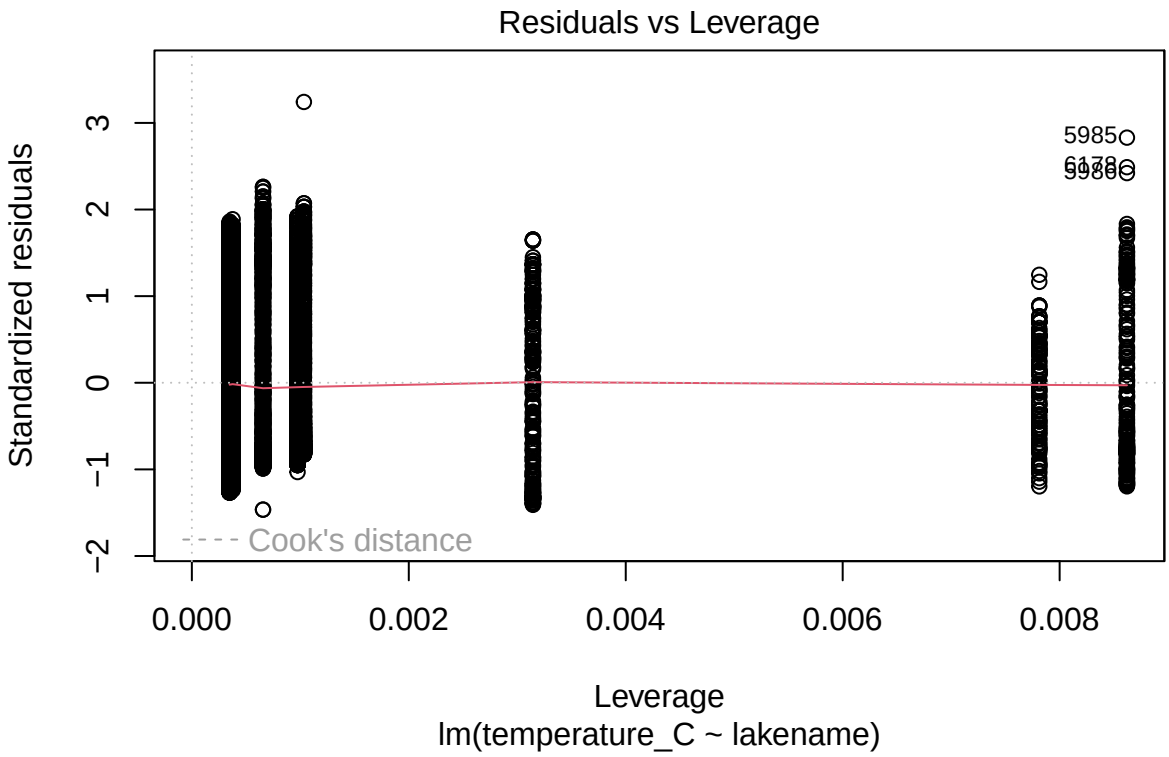
```
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 7.355 on 9719 degrees of freedom
## Multiple R-squared:  0.03953,    Adjusted R-squared:  0.03874
## F-statistic:    50 on 8 and 9719 DF,  p-value: < 2.2e-16
```

```
plot(ChemsPhys_lm_model)
```




Q−Q Residuals
Standardized residuals
−1
0
1
2
3
5003
5985
6178
−4
−2
0
2
4
Theoretical Quantiles
lm(temperature_C ~ lakename)


Scale−Location
5003
5985
6178
√|Standardized residuals|
0.0
0.5
1.0
1.5
10
12
14
16
Fitted values
lm(temperature_C ~ lakename)



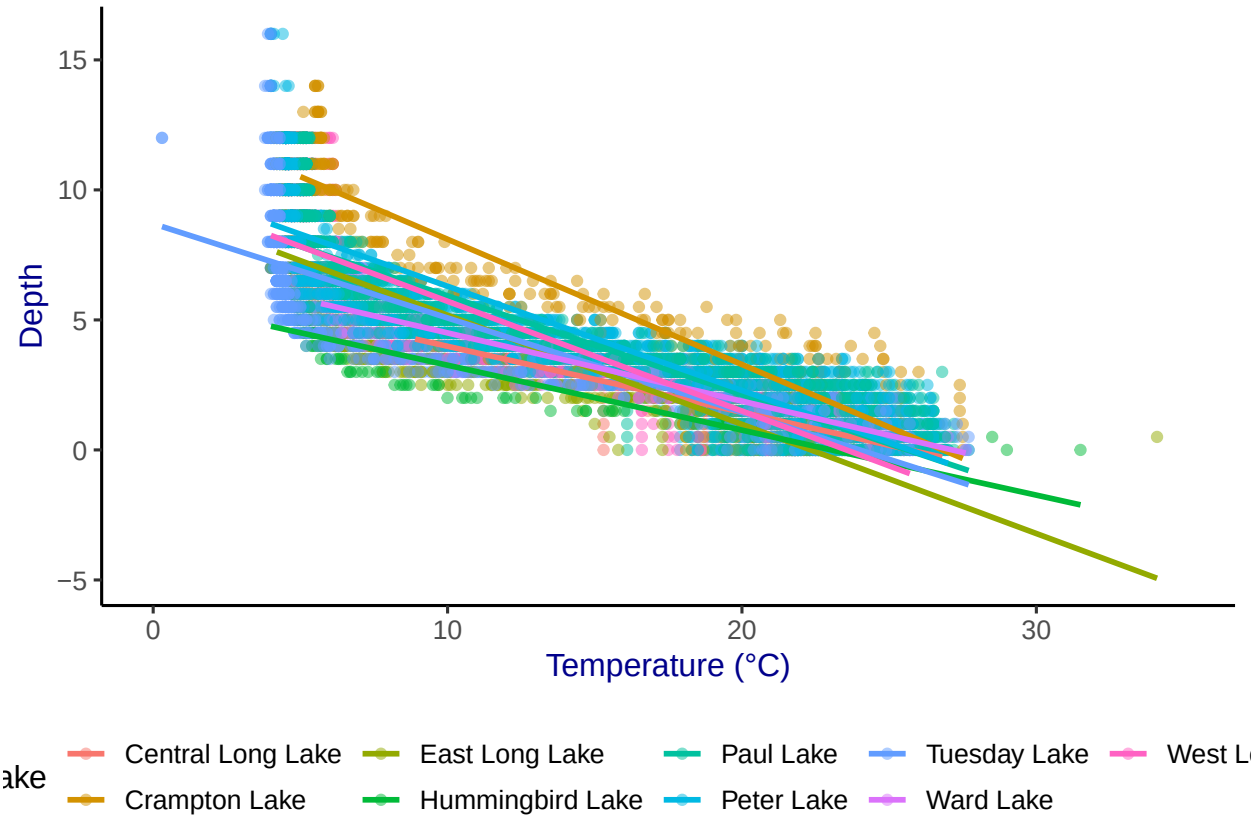
13. Is there a significant difference in mean temperature among the lakes? Report your findings.

    Answer: Yes (there is a significant difference). The coefficient for the different lakes are all significantly different from zero, suggesting that the mean temperature varies significantly across lakes. Also, the F-statistic is significant with a very small p-value ($p < 2.2e\text{-}16$), indicating that 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.
library(ggplot2)

ggplot(Lake_ChemsPhys_processed,
       aes(x = temperature_C,
           y = depth,
           color = lakename)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  xlim(0, 35) +
  labs(x = "Temperature (°C)",
       y = "Depth",
       color = "Lake") +
  mytheme +
  theme(legend.position = "bottom")
```

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



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

```
#15
lakes_means <- TukeyHSD(ChemsPhys_anova_model, "lakename")
summary(lakes_means)
```

```
##          Length Class  Mode
## lakename 144    -none- numeric
```

```
anova_model <- aov(data = Lake_ChemsPhys_processed,
                   temperature_C ~ lakename)
tukey_lakes_means <- TukeyHSD(anova_model, "lakename")
summary(tukey_lakes_means)
```

```
##          Length Class  Mode
## lakename 144    -none- numeric
```

```
view(tukey_lakes_means)
```

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 no statistically significant difference in mean temperature compared to Peter Lake are - Ward Lake - Crampton Lake - Tuesday Lake - West Long Lake Several pairs of lakes have p-values less than 0.05, indicating statistically significant differences in mean temperatures between them. However, there isn't a single lake that is statistically distinct from all the 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: We can consider conducting an independent samples t-test. This test would allow us to assess whether there is a statistically significant difference in mean temperatures between these 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 your answer for part 16?

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

CramWarLake_Ttest <- t.test(data = CramWarLake,
                            temperature_C ~ lakename)
view(CramWarLake_Ttest)
```

    Answer: It shows that Crampton Lake and Ward Lake have the same temperature values. These result is interesting because from section 16, these lakes have the same the same mean temperature as Peter Lake.