

# Assignment 7: GLMs week 2 (Linear Regression and beyond)

Kristine Swann

## OVERVIEW

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

## Directions

1. Change “Student Name” on line 3 (above) with your name.
2. Work through the steps, **creating code and output** that fulfill each instruction.
3. Be sure to **answer the questions** in this assignment document.
4. When you have completed the assignment, **Knit** the text and code into a single PDF file.
5. After Knitting, submit the completed exercise (PDF file) to the dropbox in Sakai. Add your last name into the file name (e.g., “Salk\_A06\_GLMs\_Week1.Rmd”) prior to submission.

The completed exercise is due on Tuesday, February 25 at 1:00 pm.

## Set up your session

1. Set up your session. Check your working directory, load the tidyverse, nlme, and piecewiseSEM packages, import the *raw* NTL-LTER raw data file for chemistry/physics, and import the processed litter dataset. You will not work with dates, so no need to format your date columns this time.
2. Build a ggplot theme and set it as your default theme.

```
## [1] "C:/Users/krist/Box Sync/Spring 2020/R/Environmental_Data_Analytics_2020"
```

## NTL-LTER test

Research question: What is the best set of predictors for lake temperatures in July across the monitoring period at the North Temperate Lakes LTER?

3. Wrangle your NTL-LTER dataset with a pipe function so that it contains only the following criteria:
  - Only dates in July (hint: use the daynum column). No need to consider leap years.
  - Only the columns: lakename, year4, daynum, depth, temperature\_C
  - Only complete cases (i.e., remove NAs)
4. Run an AIC to determine what set of explanatory variables (year4, daynum, depth) is best suited to predict temperature. Run a multiple regression on the recommended set of variables.

```
## Start:  AIC=24461.34
## temperature_C ~ lakename + year4 + daynum + depth
##
##           Df Sum of Sq    RSS   AIC
## <none>                 120062 24461
## - year4      1         184 120245 24474
## - daynum     1         1346 121407 24568
## - lakename   8         21056 141118 26016
## - depth     1        403139 523201 38770
```

```
##
## Call:
## lm(formula = temperature_C ~ lakenamename + year4 + daynum + depth,
##     data = cp)
##
## Coefficients:
##             (Intercept)      lakenamenameCrampton Lake      lakenamenameEast Long Lake
##                45.17306                4.71362                -1.46041
## lakenamenameHummingbird Lake      lakenamenamePaul Lake      lakenamenamePeter Lake
##                -4.73042                0.99422                1.44048
##      lakenamenameTuesday Lake      lakenamenameWard Lake      lakenamenameWest Long Lake
##                -1.38445                -0.46590                -0.16847
##                year4                daynum                depth
##                -0.01588                0.04157                -1.96540
##
## Call:
## lm(formula = temperature_C ~ lakenamename + year4 + daynum + depth,
##     data = cp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.8938 -3.0274 -0.2114  2.7781 15.2926
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   45.17306    8.248578   5.476 4.45e-08 ***
## lakenamenameCrampton Lake   4.713617    0.382185  12.333 < 2e-16 ***
## lakenamenameEast Long Lake  -1.460406    0.343271  -4.254 2.12e-05 ***
## lakenamenameHummingbird Lake -4.730421    0.459795 -10.288 < 2e-16 ***
## lakenamenamePaul Lake       0.994222    0.331643   2.998 0.002726 **
## lakenamenamePeter Lake      1.440479    0.331406   4.347 1.40e-05 ***
## lakenamenameTuesday Lake    -1.384450    0.336476  -4.115 3.91e-05 ***
## lakenamenameWard Lake       -0.465900    0.464619  -1.003 0.316003
## lakenamenameWest Long Lake  -0.168474    0.341961  -0.493 0.622257
## year4          -0.015885    0.004118  -3.857 0.000115 ***
## daynum         0.041574    0.003985  10.432 < 2e-16 ***
## depth         -1.965403    0.010885 -180.566 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.516 on 9710 degrees of freedom
## Multiple R-squared:  0.7803, Adjusted R-squared:  0.78
## F-statistic: 3135 on 11 and 9710 DF, p-value: < 2.2e-16
```

5. What is the final set of explanatory variables that predict temperature from your multiple regression? How much of the observed variance does this model explain?

Answer: Wow. The model explains 78% of the variance. That's crazy high! Final set of explanatory variables that predict temp: lakes (crampton lake, east long lake, hummingbird lake, paul lake, peter lake, tuesday lake, ward lake, and long lake), year, day within July, and depth. The only lake that didn't have a specific coefficient listed was Central Long Lake.

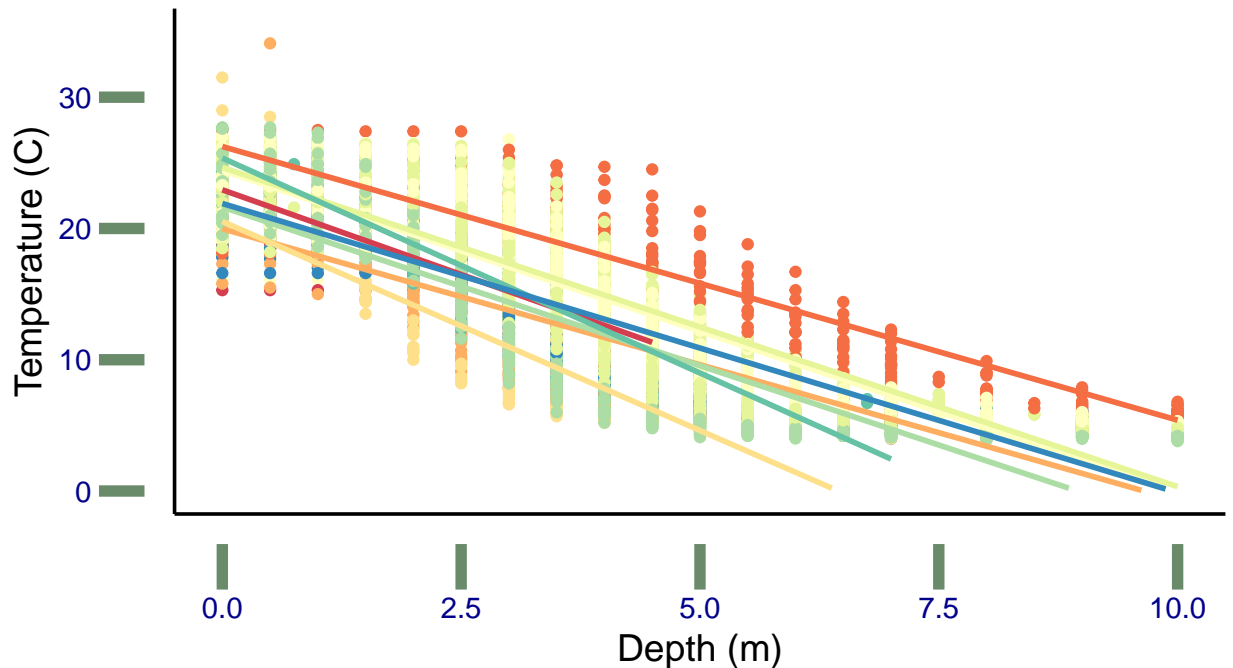
6. Run an interaction effects ANCOVA to predict temperature based on depth and lakenamename from the same wrangled dataset.

```
##
## Call:
## lm(formula = temperature_C ~ lakename * depth, data = cp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -7.6455 -2.9133 -0.2879  2.7567 16.3606
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      22.9455     0.5861  39.147 < 2e-16 ***
## lakenameCrampton Lake      2.2173     0.6804   3.259  0.00112 **
## lakenameEast Long Lake    -4.3884     0.6191  -7.089 1.45e-12 ***
## lakenameHummingbird Lake  -2.4126     0.8379  -2.879  0.00399 **
## lakenamePaul Lake         0.6105     0.5983   1.020  0.30754
## lakenamePeter Lake        0.2998     0.5970   0.502  0.61552
## lakenameTuesday Lake    -2.8932     0.6060  -4.774 1.83e-06 ***
## lakenameWard Lake        2.4180     0.8434   2.867  0.00415 **
## lakenameWest Long Lake   -2.4663     0.6168  -3.999 6.42e-05 ***
## depth              -2.5820     0.2411 -10.711 < 2e-16 ***
## lakenameCrampton Lake:depth  0.8058     0.2465   3.268  0.00109 **
## lakenameEast Long Lake:depth  0.9465     0.2433   3.891  0.00010 ***
## lakenameHummingbird Lake:depth -0.6026     0.2919  -2.064  0.03903 *
## lakenamePaul Lake:depth    0.4022     0.2421   1.662  0.09664 .
## lakenamePeter Lake:depth    0.5799     0.2418   2.398  0.01649 *
## lakenameTuesday Lake:depth  0.6605     0.2426   2.723  0.00648 **
## lakenameWard Lake:depth    -0.6930     0.2862  -2.421  0.01548 *
## lakenameWest Long Lake:depth  0.8154     0.2431   3.354  0.00080 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.471 on 9704 degrees of freedom
## Multiple R-squared:  0.7861, Adjusted R-squared:  0.7857
## F-statistic: 2097 on 17 and 9704 DF, p-value: < 2.2e-16
```

7. Is there a significant interaction between depth and lakename? How much variance in the temperature observations does this explain?

Answer: There is a significant interaction between depth and lakename, with an overall model p-value < 0.0001. The adjusted R<sup>2</sup> value is 0.79, which is even higher than the other lm! The interactions between depth and individual lakes have varied p-values, with Hummingbird being the only one without significant interaction with depth (p-values > 0.05). I suppose all this suggests that site specific conditions include total depth, so there's an interaction.

8. 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.



● Central Long Lake    ● East Long Lake    ● Paul Lake    ● Tuesday Lake    ● We  
● Crampton Lake    ● Hummingbird Lake    ● Peter Lake    ● Ward Lake

9. Run a mixed effects model to predict dry mass of litter. We already know that `nlcdClass` and `functionalGroup` have a significant interaction, so we will specify those two variables as fixed effects with an interaction. We also know that litter mass varies across plot ID, but we are less interested in the actual effect of the plot itself but rather in accounting for the variance among plots. Plot ID will be our random effect.

- a. Build and run a mixed effects model.
- b. Check the difference between the marginal and conditional R<sup>2</sup> of the model.

```
## Response family link method Marginal Conditional
## 1 dryMass gaussian identity none 0.2465822 0.2679023
```

- b. continued... How much more variance is explained by adding the random effect to the model?

Answer: approximately 2 %, with marginal  $r^2 = 0.24$  and conditional = 0.26.

- c. Run the same model without the random effect.
- d. Run an anova on the two tests.

```
## Response family link method R.squared
## 1 dryMass gaussian identity none 0.2515836

## Model df AIC BIC logLik Test L.Ratio p-value
## litter.mixed 1 26 9038.575 9179.479 -4493.287
## litter.fixed 2 25 9058.088 9193.573 -4504.044 1 vs 2 21.51338 <.0001
```

- d. continued... Is the mixed effects model a better model than the fixed effects model? How do you know?

Answer: The fixed effect is slightly better than the mixed effect because the AIC is slightly lower (mixed = 9038.6; fixed = 9038.1).