13: Generalized Linear Models (ANCOVA and mixed effects)

Environmental Data Analytics | Kateri Salk

Spring 2020

Objectives

2. Apply special cases of the GLM (ANCOVA, mixed effects models) to real datasets
3. Interpret and report the results of linear regressions in publication-style formats
4. Apply model selection methods to choose model formulations

Set up

```
getwd()
```

```
## [1] "/Users/ks501/Documents/GitHub_Repos/Environmental_Data_Analytics_2020"
```

```
library(tidyverse)
library(lubridate)
library(viridis)
#install.packages("nlme")
library(nlme)
#install.packages("piecewiseSEM")
library(piecewiseSEM)

PeterPaul.chem.nutrients <- read.csv("./Data/Processed/NTL-LTER_Lake_Chemistry_Nutrients_PeterPaul_Proc
NTL.chem <- read.csv("./Data/Raw/NTL-LTER_Lake_ChemistryPhysics_Raw.csv")

NTL.chem$sampledate <- as.Date(NTL.chem$sampledate, format = "%m/%d/%y")

# Set theme
mytheme <- theme_classic(base_size = 14) +
  theme(axis.text = element_text(color = "black"),
        legend.position = "top")
theme_set(mytheme)
```

ANCOVA

Analysis of Covariance consists of a prediction of a continuous response variable by both continuous and categorical explanatory variables. We set this up in R with the `lm` function, just like prior applications in this lesson.

Let's say we wanted to predict total nitrogen concentrations by depth and by lake. We could represent these explanatory variables as main effects (two intercepts, same slope) or as interaction effects (two intercepts and two slopes).

```
# main effects
TNancova.main <- lm(data = PeterPaul.chem.nutrients, tn_ug ~ lakename + depth)
```

```r
summary(TNancova.main)
```

```
##
## Call:
## lm(formula = tn_ug ~ lakename + depth, data = PeterPaul.chem.nutrients)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -794.23 -116.79  -38.13   77.30 2324.10
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         284.437     12.638  22.507  < 2e-16 *** intercept: Paul lake at depth=o
## lakenamePeter Lake   66.449     16.053   4.139 3.69e-05 *** 284.+66.449437
## depth                68.559      2.269  30.214  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 302 on 1422 degrees of freedom
##   (21583 observations deleted due to missingness)
## Multiple R-squared:  0.4023, Adjusted R-squared:  0.4015
## F-statistic: 478.6 on 2 and 1422 DF,  p-value: < 2.2e-16
```

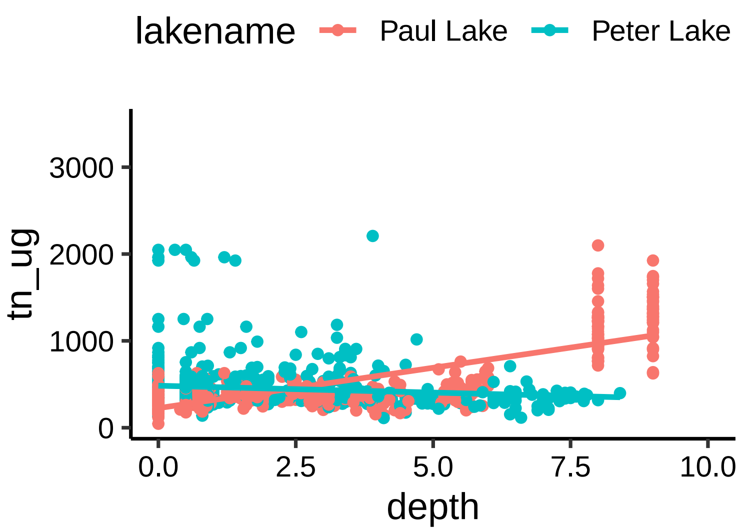
```r
# interaction effects
TNancova.interaction <- lm(data = PeterPaul.chem.nutrients, tn_ug ~ lakename * depth)
summary(TNancova.interaction)
```

```
##
## Call:
## lm(formula = tn_ug ~ lakename * depth, data = PeterPaul.chem.nutrients)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -679.85 -133.72  -26.35   80.15 2438.49
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               228.744     14.395  15.891  < 2e-16 ***
## lakenamePeter Lake        161.607     20.109   8.037 1.92e-15 ***
## depth                      90.894      3.685  24.669  < 2e-16 ***paul  depth increase 1 m depth, 90.894 increa
## lakenamePeter Lake:depth  -35.156      4.623  -7.605 5.15e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 296.1 on 1421 degrees of freedom
##   (21583 observations deleted due to missingness)
## Multiple R-squared:  0.4257, Adjusted R-squared:  0.4245
## F-statistic: 351.1 on 3 and 1421 DF,  p-value: < 2.2e-16
```

```r
TNplot <- ggplot(PeterPaul.chem.nutrients, aes(x = depth, y = tn_ug, color = lakename)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  xlim(0, 10)
print(TNplot)
```

```
## Warning: Removed 21694 rows containing non-finite values (stat_smooth).

## Warning: Removed 21694 rows containing missing values (geom_point).
```




```
# Make the graph attractive
```

HIERARCHICAL MODELS

Hierarchical models, or **mixed-effects models,** are a type of linear model in which explanatory variables are given a model whose parameters are also estimated by the data. The coefficients associated with explanatory variables thus may not be a single value but instead be sampled from a distribution, called the hyper-distribution, which is defined by the modeler. The advantage of the hierarchical model is that it builds capacity to describe multiple layers of stochasticity, which enables accounting of all aspects of uncertainty in a system. Specifically, we can separately model the process of interest and the sampling process.

The coefficients of a hierarchical model are divided into two categories: **fixed effects** and **random effects.** A **fixed effect** is a factor whose levels are experimentally determined or whose interest lies in the effects of each level (e.g., covariates, treatments, interactions). A **random effect** is a factor whose levels are sampled from a larger population, or whose interest lies in the variation among them rather than the specific effect of each level. In choosing whether you are dealing with a fixed or a random effect, consider the following questions:

- Do you have a particular interest in the studied factor level?
- Have you included all possible levels in the study?
- Do you have interest in the variance among levels?
- Do you have interest in generalizing to factor levels that you did not study?

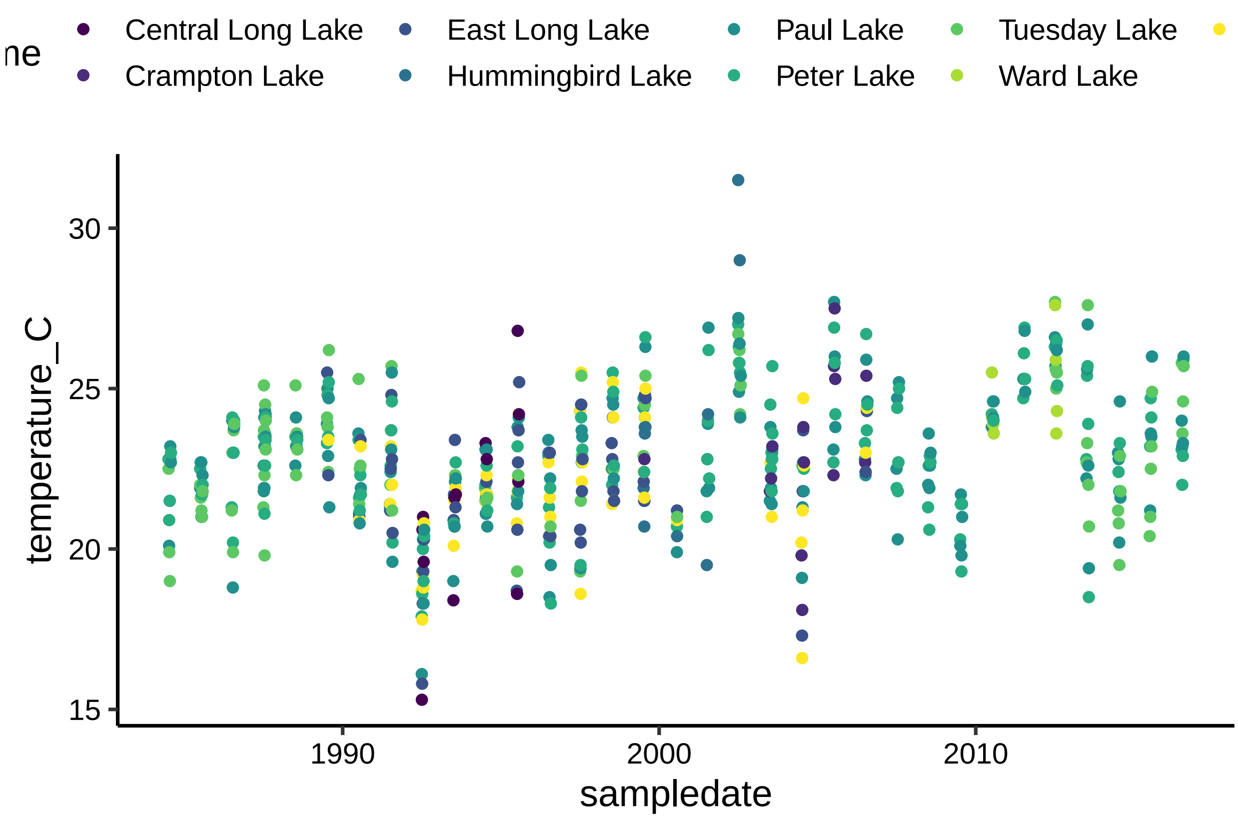
A common variable in hierarchical models is **space.** In many situations, we may want to infer conditions beyond the sites that we have sampled. By treating space as a random variable, we may be able to extrapolate conditions of the response variable across a spatial gradient.

Let's think about the situation of temperature monitoring in the NTL-LTER lakes. We might be interested to know whether surface temperatures in July have increased over time in response to climate change. However, we know that there may be variability across lakes that may obscure the trend we see in temperature. We can set lake as a random effect to account for the across-lake variability and also enable us to extrapolate across lakes in northern Wisconsin.

Let's wrangle our data and visualize a preliminary relationship between our variables of interest.

```
NTL.summertemp <-
  NTL.chem %>%
  select(lakename:temperature_C) %>%
  #filter for Julian days in July and surface measurements
  filter(daynum > 181 & daynum < 213 & depth == 0 ) %>%
  #code won't work if there are NAs
  na.exclude()

NTLtemps <-
ggplot(NTL.summertemp, aes(x = sampledate, y = temperature_C, color = lakename)) +
  geom_point() +
  scale_color_viridis_d()
print(NTLtemps)
```



Next, we will build a hierarchical model. We will use the package `nlme` for our analyses. Another good package for running hierarchical, or mixed-effects, models is `lme4`. For the basic types of hierarchical models, these packages have about the same functionality. We will set year year (continuous) as a fixed effect and lake (categorical) as a random effect. Remember that we are interested in assessing if summer surface temperatures have increased in response to climate change and to account for the inter-lake variability within the model.

```
TempTest.mixed <- lme(data = NTL.summertemp,
                     temperature_C ~ year4,
                     random = ~1|lakename)
summary(TempTest.mixed)
```

```
## Linear mixed-effects model fit by REML
```

```
##  Data: NTL.summertemp
##        AIC      BIC    logLik
##   2277.005 2294.097 -1134.503
##
## Random effects:
##  Formula: ~1 | lakename
##         (Intercept) Residual
## StdDev:    0.448605 2.008743
##
## Fixed effects: temperature_C ~ year4
##                 Value Std.Error  DF   t-value p-value
## (Intercept) -97.72204 19.499332 522 -5.011559       0
## year4         0.06026  0.009755 522  6.177274       0
##  Correlation:
##       (Intr)
## year4 -1
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -3.26300038 -0.56995649 -0.02118468  0.67113577  4.11268759
##
## Number of Observations: 532
## Number of Groups: 9
```

```r
rsquared(TempTest.mixed)
```

```
##        Response   family     link method Marginal Conditional
## 1 temperature_C gaussian identity   none 0.066243   0.1106014
```

```r
# Compare the random effects model with the fixed effects model
TempTest.fixed <- gls(data = NTL.summertemp,
                      temperature_C ~ year4)
summary(TempTest.fixed)
```

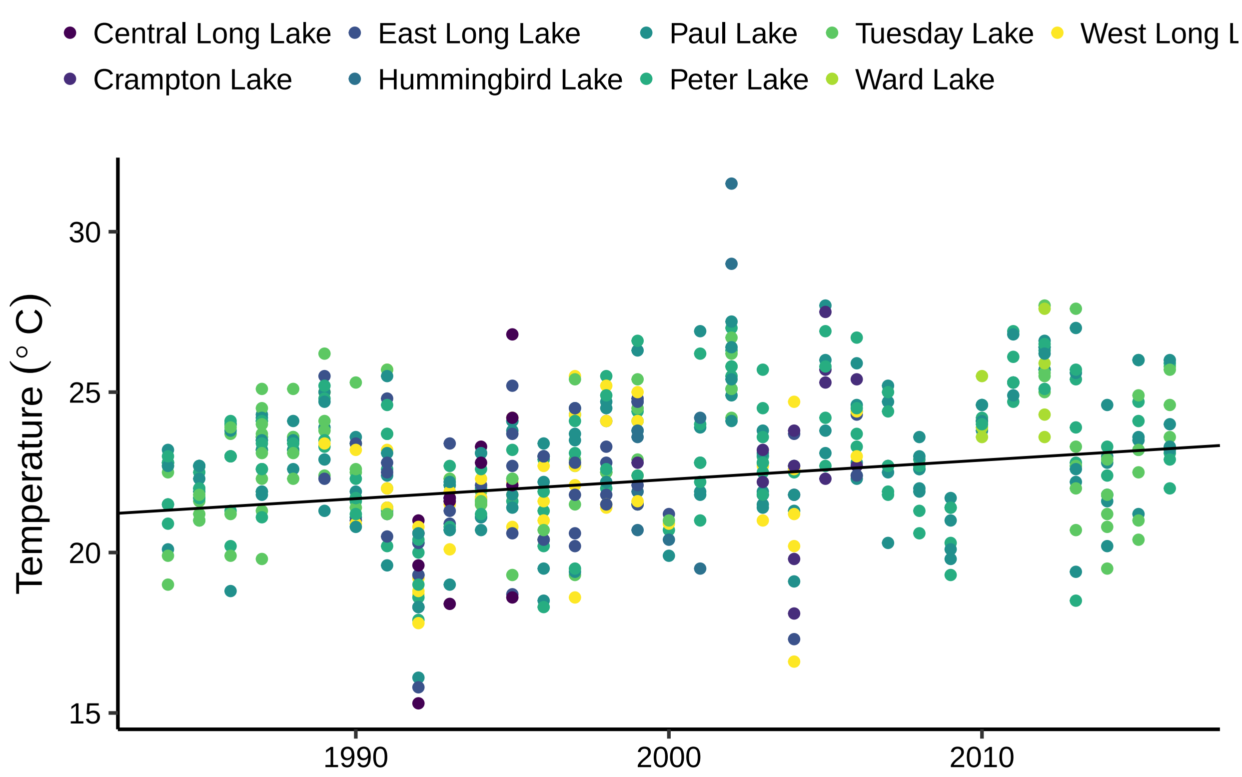
```
## Generalized least squares fit by REML
##   Model: temperature_C ~ year4
##   Data: NTL.summertemp
##        AIC      BIC    logLik
##   2279.845 2292.664 -1136.923
##
## Coefficients:
##                  Value Std.Error   t-value p-value
## (Intercept) -107.22765 19.395640 -5.528441       0
## year4          0.06505  0.009704  6.702821       0    slope = 0.06
##
##  Correlation:
##       (Intr)
## year4 -1
##
## Standardized residuals:
##         Min          Q1         Med          Q3         Max
## -3.46327997 -0.57923635 -0.03005572  0.68423156  4.18009392
##
## Residual standard error: 2.034381
## Degrees of freedom: 532 total; 530 residual
```

```r
anova(TempTest.mixed, TempTest.fixed)
```

```
##                Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## TempTest.mixed     1  4 2277.005 2294.097 -1134.503
## TempTest.fixed     2  3 2279.845 2292.664 -1136.923 1 vs 2 4.839998  0.0278
```

```r
# The lower the AIC, the better.
# The p-value tells us whether those models have a significantly different fit

NTL.tempmodel <-
ggplot(NTL.summertemp, aes(x = year4, y = temperature_C, color = lakename)) +
  geom_point() +
  scale_color_viridis_d() +
  geom_abline(intercept = -97.72, slope = 0.06) +  0.06* 30: over 30 years, the temp will increase 1.8 degree
  # make it look better
  labs(x = "", y = expression("Temperature " ( degree~C)), color = "") +
  theme(legend.spacing.x = unit(0, "cm"))
print(NTL.tempmodel)
```



Question: How would you interpret the collective results of your mixed effects model in the context of the study question?

ANSWER: