Advanced statistics: General linear mixed models

9 – 16 January 2019
Kate Laskowski (laskowski@igb-berlin.de)
Leibniz Institute of Freshwater Ecology & Inland Fisheries
Berlin, Germany
Course Topics

1- Refresher on linear modeling (ANOVA + regression)
2- Model assumptions
3- Intro to mixed models
4- Data exploration and model selection
5- Crossed vs nested effects
6- Estimating R2 from mixed models
7- Centering predictors to aid interpretation
8- Dealing with variance heterogeneity & autocorrelation
9- Random regression models
<table>
<thead>
<tr>
<th>Wednes</th>
<th>Thurs</th>
<th>Fri</th>
<th>Mon</th>
<th>Tues</th>
<th>Wednes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1- Linear modeling refresher</td>
<td>4 – Data exploration and model selection</td>
<td>7 – Centering &amp; interpretation</td>
<td>Wrap-up Group work</td>
<td>Group work</td>
<td>Final presentations!</td>
</tr>
<tr>
<td>2- Model assumptions</td>
<td>5 - Repeatability and other random effects</td>
<td>8 – variance heterogeneity &amp; temporal autocorr</td>
<td>Group work</td>
<td>Consults with Kate</td>
<td></td>
</tr>
<tr>
<td>3 – Intro to mixed models</td>
<td>6 – Estimating R2</td>
<td>9 – random slopes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Book


- Available as e-book through the library catalog

- I also try to list useful further reading at the end of each section
Other useful books
Useful online resources

- [https://ourcodingclub.github.io/](https://ourcodingclub.github.io/)
  - Seriously helpful tutorials on lots of stats/R stuff (written by ecologists!)

- [http://m-clark.github.io/documents.html](http://m-clark.github.io/documents.html)
  - More in depth tutorials, also more advanced topics (Bayesian, SEM, Generalized)

  - Wasserman’s ‘All of Statistics’ book available as a pdf online
(Refresher on)
Linear regression/modeling

- YOUR HYPOTHESIS = YOUR MODEL = YOUR GRAPH
- Envision your analysis in terms of the graph you want to draw
- Understand the parameters in a basic linear model and what they mean
- Understand the difference between predictors (fixed) variables and response variables
- Understand how to interpret an R summary of a linear model
Statistics are **fundamental** to biological research

Understanding statistics helps you do better science
Clouds of variation
What is the effect of x on y?

• Is x categorical or continuous?
  – Categorical variables: logical grouping, no linear relationship expected among levels
    • E.g. treatments, sex, habitat types
    • More expensive in terms of df – estimates intercept for each level
  – Continuous: logical relationship along some gradient, expect increase/decrease with gradient
    • E.g. body size, growth rate
    • Only costs 1 df – estimates the slope across the gradient
Time as a categorical fixed effect with 6 levels. Sex and Treatment as categorical fixed effects with 2 levels each. Two-way interaction.
Behavior

Predators absent
Predators present

Temperature as a continuous fixed effect (or covariate)
that interacts with Predation as a categorical fixed effect
How the data is graphed should match up with your statistical model.
Linear modeling

• Catch-all term for ANOVA, ANCOVA, regression
• How do your predictor(s) relate to your response variable?

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]
Linear modeling

• Catch-all term for ANOVA, ANCOVA, regression
• How do your predictor(s) relate to your response variable?

\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

\[ y_i = \beta_0 + \left\{ \beta_1 x_i \right\} + \epsilon_i \]

\[ \beta_{1-n} = \text{intercepts/means} \]
What is the effect of x on y?

• Is ‘time’ a continuous or categorical variable?
  – E.g. years, seasons, months, observations

・ More time points increases odds of false positives because you are increasing the number of comparisons you make.

・ Are you actually interested in the effects of time? Then you should probably increase the number of time points.

・ Or are you more interested in treatment effects? Then don’t bother with more time points, but increase your replicates.

・ How many df can you burn?
  – Think about this before you plan your experiment!
Linear models – error terms

• In order to estimate $\beta_0$ or $\beta_1$ you need multiple observations to make an error term
  – Intercepts: many observations at that factor level
  – Slopes: many observations around line (2 points make a line)

• If you do not have multiple observations, then you do not have an error term

YOU CANNOT DO STATISTICS
Where is the error coming from?

- Predators absent
- Predators present
Writing models (in R)

• You want to measure whether daphnia abundance is influenced by proximity to an inflow pipe from a power plant. So you think you will collect samples from 47 sites - 23 that are near the pipe, 24 that are far away. Additionally you know that calcium concentration varies in the lake across site and this could also influence abundance.
Draw graph here
Writing models (in R)

• You will measure the abundance of birds in 56 different forest patches in Australia. You think these abundances may be related to a number of forest characteristics, so you also want to measure the size of the forest patch, the distance to the nearest patch, the distance to the nearest larger patch, the altitude of the patch and the intensity of grazing.
What affects clam size?

- We collected a whole boatload of clams over 6 different months in a year. We measured each clam’s length and its AFD (Ash Free Dry weight). We want to see what factors affect AFD.
How to interpret (a simple) R model

```r
> summary(mod1)

Call:
  lm(formula = AFD ~ LENGTH, data = Clams)

Residuals:
   Min     1Q   Median     3Q    Max
-0.079253 -0.014092 -0.004987  0.008578  0.286588

Coefficients:  
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            -0.1313364  0.0049097  -26.75  <2e-16 ***
LENGTH                  0.0119844  0.0002598   46.13  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03004 on 396 degrees of freedom 
Multiple R-squared:  0.8431,  Adjusted R-squared:  0.8427
F-statistic: 2128 on 1 and 396 DF,  p-value: < 2.2e-16
```
• (Intercept) – what does this mean?

• LENGTH – is this continuous or categorical? How do you know?

• T-test and Pr(>|t|) – what is the hypothesis this is testing?

• R-squared – what does this mean?

• Overall F-test and p-value – what does this mean?
Draw the graph for this model
```r
> mod2 <- lm(AFD ~ LENGTH + fMONTH, data = Clams)
> summary(mod2)

Call:
  lm(formula = AFD ~ LENGTH + fMONTH, data = Clams)

Residuals:
     Min       1Q   Median       3Q      Max
-0.084185 -0.014142 -0.003910  0.009508  0.274515

Coefficients:  Estimate  Std. Error t value Pr(>|t|)
(Intercept)   -0.142832    0.008056  -17.729   < 2e-16 ***
LENGTH         0.012674    0.000348   36.420   < 2e-16 ***
fMONTH3        0.024168    0.008703    2.777   0.00575  **
fMONTH4        0.001953    0.004339    0.450   0.65291
fMONTH9        0.006264    0.008548   0.733   0.46408
fMONTH11       0.002317    0.006810   0.340   0.73383
fMONTH12      -0.020546    0.004564  -4.501 8.92e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02901 on 391 degrees of freedom
Multiple R-squared:  0.8556,  Adjusted R-squared:  0.8534
F-statistic: 386 on 6 and 391 DF,  p-value: < 2.2e-16
```
• There were 6 MONTHS measured (2, 3, 4, 9, 11, 12) – how many are listed here? WTF?

• MONTHS – what are these estimates?

• What happened to the d.f.? Why?
Draw the graph for this model
```r
> mod3 <- lm(AFD ~ LENGTH*fMONTH, data = Clams)
> summary(mod3)

Call:
  lm(formula = AFD ~ LENGTH * fMONTH, data = Clams)

Residuals:
     Min      1Q  Median      3Q     Max
-0.097255 -0.010519 -0.001542  0.009455  0.187687

Coefficients:  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)    -0.2994176   0.0108672   -27.552  < 2e-16 ***
LENGTH         0.0197932   0.0004869    40.653  < 2e-16 ***
fMONTH3        0.2809834   0.1188598     2.364  0.01857 *
fMONTH4        0.2048829   0.0120065    17.064  < 2e-16 ***
fMONTH9        0.2642918   0.0233638    11.312  < 2e-16 ***
fMONTH11       0.1334337   0.0188923     7.063  7.65e-12 ***
fMONTH12       0.1390291   0.0525140     2.647  0.00844 **
LENGTH:fMONTH3 -0.0169399   0.0115955    -1.461  0.14485
LENGTH:fMONTH4 -0.0103439   0.0005907   -17.512  < 2e-16 ***
LENGTH:fMONTH9 -0.0151597   0.0016534    -9.169  < 2e-16 ***
LENGTH:fMONTH11 -0.0057147   0.0009490   -6.022  4.02e-09 ***
LENGTH:fMONTH12 -0.0072561   0.0023960    -3.028  0.00262 **

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02134 on 386 degrees of freedom
Multiple R-squared:  0.9228,    Adjusted R-squared:  0.9206
F-statistic: 419.6 on 11 and 386 DF,  p-value: < 2.2e-16
```
• What do the interaction terms mean?

• Where is MONTH 2 again??

• What happened to the degrees of freedom?
Draw the graph for this model
Overall significance of effects

```r
> mod2 <- lm(AFD ~ LENGTH + fMONTH, data = Clams)
> summary(mod2)
```

Call:
`lm(formula = AFD ~ LENGTH + fMONTH, data = Clams)`

Residuals:
```
     Min      1Q  Median      3Q     Max
-0.084185 -0.014142 -0.003910  0.009508  0.274515
```

Coefficients:
```
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) -0.142832   0.008056  -17.729  < 2e-16 ***
LENGTH       0.012674   0.000348   36.420  < 2e-16 ***
fMONTH3      0.024168   0.008703    2.777   0.00575 **
fMONTH4      0.001953   0.004339    0.450   0.65291
fMONTH9      0.006264   0.008548    0.733   0.46408
fMONTH11     0.002317   0.006810    0.340   0.73383
fMONTH12    -0.020546   0.004564   -4.501  8.92e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 1
```

Residual standard error: 0.02901 on 391 degrees of freedom
Multiple R-squared:  0.8556, Adjusted R-squared:  0.8534
F-statistic: 386 on 6 and 391 DF,  p-value: < 2.2e-16

Is ‘Month’ a significant predictor of clam size?
Summary() versus anova()

• Use the anova() command with care!
  – Defaults to Type I sum of squares – which are determined sequentially!
  – Type III SS are preferred since these are not sequential

• Can use the Anova() command from the car package to set which Type of SS. But in general, log-likelihood ratio tests are preferred because they better account for degrees of freedom (more on this later on in class)
Linear regression/modeling

- YOUR HYPOTHESIS = YOUR MODEL = YOUR GRAPH
- Envision your analysis in terms of the graph you want to draw
- Understand the parameters in a basic linear model and what they mean
- Understand the difference between predictors (fixed) variables and response variables
- Understand how to interpret an R summary of a linear model

• FURTHER READING on model interpretation:
  – Zuur Appendix
    • We will generally use the “summary()” command in R but there are other ways to get model summaries in R such as “drop1()” and “anova()”. Each does something slightly different and interpreting the results from each is not always valid! Read this appendix to learn more!
  – Gelman & Hill book, Chapter 3