4. Data exploration & model selection

- Understand the importance of data exploration and how to do it
- Understand the difference between exploration and fishing expeditions (and when each is appropriate)
- Learn and apply model selection procedures (usually based on good biological knowledge)
- Understand which factors can be included as random effects

Okay we have data!

Now what?

How to do statistics

1 – Clearly state/write out your hypothesis

- Hint: your hypothesis *is* your statistical model
- Draw the graph you eventually want to publish

2 – Explore your data for familiarity and problems

3 – Proper model selection

- determine random effects
- determine fixed effects
- validate model assumptions

6 – Graph the results!

7 – Publish!

Experimental versus exploratory research

- **Do you want to test a hypothesis or do you want to find the best fit model to your data?**
  - Experimental – has clear a priori hypothesis. Data is collected based on this hypothesis. Statistics will test this specific hypothesis
  - Observational/Explorative – possibly no clear hypothesis. Data may be explored for promising patterns that will then be used to guide future work. Results should be interpreted with caution and not strongly generalizable

For experimental research

Your hypothesis *is* your statistical model

Your statistical model *is* your hypothesis

Your life will be dramatically improved if you start thinking about this ***before*** you collect any data

Data exploration

- So many problems can be avoided by doing proper data exploration BEFORE any analysis.

Data exploration

- **TO BE VERY CLEAR:**
- For *experimental* research - Proper data exploration does NOT mean that you are doing hypothesis testing or searching (fishing) for any ol' significant effect.
  - It is simply investigating whether your model can be trusted (is valid).
  - DO NOT USE data exploration to generate or change your hypothesis – that is WRONG

1 Formulate biological hypothesis
Carry out experiment & collect data

Data exploration

2

1. Outliers Y & X	*boxplot & Cleveland dotplot*
2. Homogeneity Y	*conditional boxplot*
3. Normality Y	*histogram or QQ-plot*
4. Zero trouble Y	*frequency plot or corrgram*
5. Collinearity X	*VIF & scatterplots* *correlations & PCA*
6. Relationships Y & X	*(multi-panel) scatterplots* *conditional boxplots*
7. Interactions	*coplots*
8. Independence Y	*ACF & variogram* *plot Y versus time/space*

3 Apply statistical model

Zuur et al. 2010 *MEE*

Exploration examples

- Loyn data set
- RIKZ data set

Model selection

- K.I.S.S. – Keep it simple, stupid
  - Don’t use ‘fancy’ stats to cover up flawed designs or ‘boring’ results
- Mixed models are powerful tools but data hungry and (sometimes) difficult to explain
  - Only use them when appropriate & make sure you understand what they are doing

General procedure for model selection

1. Determine optimal random structure using "beyond optimal" (when possible) fixed structure model
   - Test nested models using LLR tests with REML estimation
   - Selection criteria (AIC; BIC)

2. Then, determine optimal fixed effects
   - Lots of different philosophies on how to reduce fixed effects
   - Test nested models using LLR test with ML estimation

3. Run final model
   - Use REML to get parameter estimates on random effects
   - Use ML to ~~get parameter estimates on fixed effects~~
   - Use LLR to get p-values for overall effects

Zuur pg 90 & 121

ML vs. REML

- Maximum likelihood (ML)
  - Only use when testing nested models that differ in fixed effects
  - Can underestimate the error in random effects so shouldn't use for random effects
- Restricted maximum likelihood (REML)
  - should be used for mixed modeling because properly determines degrees of freedom
  - Use when testing nested models that differ in random effects
  - (REML is the default for most R mixed modeling packages)

Proper model selection - determine random structure

- Random effects are normally introduced by the experimental design
  - Biological units
    - Individuals
    - Lakes
    - Incubators

  Design units
  - blocks
  - split-plots
- If they are introduced by design, then you should really really (*really*) include them in the model
  - Many statisticians would argue it is actually inappropriate to EVER remove these terms from a model (even if they are “non-significant”)

Proper model selection: determine random structure

- Start with “beyond optimal” model that contains all/most potential predictor variables and interactions
  - This ensures that the model first pulls out any and all variation attributable to any potential fixed effects first (because inherently interested in fixed effects usually, and not so much random effects)
  - Of course, if you have LOTS of predictors a full model may not be possible

What would our initial model for the RIKZ dataset probably look like?

Proper model selection: determine (initial) fixed effects – HOW?

- A number of different philosophies on how to do this:
  1. Start with a model with no interactions. Apply model and validate. Check residuals and include interactions if needed to explain patterns in residuals.
  2. Decide using biological knowledge of system which effects and interactions to include
  3. Use good data exploration to see which interactions are important
  4. Include only main terms and all two-way interactions
  5. Include all interactions by default and reduce

- This applies for non-mixed (regular linear) models too!

Determining initial fixed effects is tricky

- If you have lots of potential response variables and lots of predictors, you can almost ALWAYS find a significant effect *somewhere*.
  - ‘seeing what sticks’
  - ‘fishing expeditions’
  - ‘p-hacking’

https://fivethirtyeight.com/features/science-isnt-broken/#part1

  - compare GDP ~ all Repubs vs. Employment ~ Dems

Pitfalls of (blind) backward selection

- Once predictors chosen, you remove the non-significant ones right?
- This is a tricky and potentially dangerous thing to do (especially with small data sets):
    - Cryptic hypothesis testing (increases overall error rate)
        - This is why including 'TIME' as a variable can be especially problematic
    - Biases effect estimates upwards (away from H0 = 0)
    - Overfitting of your model (N/k >> 3)
    - Might end up with a result that makes no biological sense
  - NEVER DO THIS IS YOU HAVE MISSING DATA IN AN UNBALANCED WAY

Pitfalls of backward selection

d13C ~ TL + d15N + Spawner + ~~Home.range~~

	A	B	C	D	E	F	G	H	I	J	K	L
1	ID	TL.x	d13C	d15N	Spawner	Num.offspr	Origin	Sex	Weight	CondFact	Growth.rate	Home.range
2	41500	504	-23.479	10.63	0	0	a	f	807	0.63	10.216948	NA
3	41600	437	-23.698	11.037	NA	NA	NA	NA	NA	NA	10.164064	NA
4	41700	478	-24.054	11.402	0	0	a	f	630	0.577	11.341065	NA
5	41800	446	-24.518	10.85	NA	NA	NA	NA	NA	NA	NA	NA
6	41900	528	-22.519	10.955	NA	NA	NA	NA	NA	NA	NA	6483.9346
7	42100	470	-24.578	9.777	0	0	b	f	576	0.555	10.536503	11463.864
8	42200	438	-24.731	9.959	1	1	a	m	583	0.694	9.5074825	6951.8474
9	42300	409	-23.896	9.854	0	0	a	m	419	0.612	10.404927	5548.109
10	42400	651	-22.953	10.76	0	0	a	f	1750	0.634	11.191544	13803.428
11	42500	455	-24.265	9.828	1	2	b	f	516	0.548	12.996628	NA
12	42600	451	-24.133	10.678	0	0	a	f	509	0.555	NA	2807.4768
13	42700	436	-24.85	11.096	1	1	a	f	453	0.547	11.230324	10026.703
14	42800	NA	-23.82	11.309	NA	NA	NA	NA	NA	NA	9.8668315	NA
15	42900	415	-24.016	10.236	0	0	a	m	416	0.582	10.248096	935.82561
16	43000	464	-24.107	10.441	1	1	a	f	546	0.547	12.002849	NA
17	43100	465	-23.945	10.86	0	0	a	f	572	0.569	9.9208767	NA
18	43300	432	-23.627	10.081	1	1	a	f	442	0.548	11.0801	NA
19	43400	498	-23.551	10.823	1	1	a	f	673	0.545	NA	5180.4632
20	43500	535	-23.944	10.865	0	0	a	f	872	0.569	NA	NA
21	43600	469	-23.464	10.321	1	2	a	f	578	0.56	10.195964	22860.883
22	43700	572	-22.408	11.706	NA	NA	NA	NA	NA	NA	NA	3475.9237
23	43800	445	-24.052	10.614	0	0	a	f	440	0.499	13.485118	13770.005

Pitfalls of backward selection

- Philosophically, backward selection can be problematic:
  - For a **planned experiment**, you should know a priori what your fixed predictors of interest are
  - For an observational study, you should have good reasons why some predictor should be included
    - UNLESS you are truly doing an exploratory analysis to develop hypotheses (and then later plan an actual experiment to test those hypotheses)

Proper backward selection

- Experimental research
  - Start with a strong a priori hypothesis (this is why stats helps you become a better scientist!)
  - Never remove variables that you are inherently interested in (e.g. treatment effects)
- More exploratory work
  - Can present your initial full model and your reduced model so as to be transparent

Don't remove your tent if you need to go camping!

How to decide what to remove from your model

```
> 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
```

Is 'Month' a significant predictor of clam size?

Significance tests for model effects

- We want a single test statistic for the overall effect of MONTH
- Remove MONTH and see if it has a major effect on the model
  - If so, then it is important so you should NOT remove it

```
# log likelihood ratio test
mod2      <- lm(AFD ~ LENGTH + fMONTH, data = Clams)
mod2.test <- lm(AFD ~ LENGTH          , data = Clams)
```

Using log likelihood ratio tests

- Remove a term to see if it significantly changes the model

Log likelihood ratio tests

- Compare the likelihood of two models:
  - One with the effect of interest
  - One without the effect of interest

- $-2ln\left(\frac{L_{reduced}}{L_{full}}\right), \sim\chi^2$
- If the LLR is sufficiently large then the full model is better supported, if the LLR is small then the reduced model is supported

Model validation

- Plot residuals versus fitted
- Plot qqnorm graphs
- Plot residuals versus all fixed effects

- Go through model selection & validation with RIKZ dataset

Model interpretation

```
> summary(mod.final)
Linear mixed-effects model fit by REML
 Data: RIKZ
       AIC      BIC    logLik
  240.5538 249.2422 -115.2769

Random effects:
 Formula: ~1 | Beach
        (Intercept) Residual
StdDev:    1.907175 3.059089

Fixed effects: Richness ~ NAP + fExposure
                Value Std.Error DF   t-value p-value
(Intercept)  8.601088 1.0594875 35  8.118158  0.0000
NAP         -2.581708 0.4883901 35 -5.286160  0.0000
fExposure11 -4.532777 1.5755610  7 -2.876929  0.0238
 Correlation:
            (Intr) NAP
NAP         -0.136
fExposure11 -0.655 -0.037

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-1.5163203 -0.4815106 -0.1218700  0.2922854  3.8777562

Number of Observations: 45
Number of Groups: 9
```

What do you write in your paper?

- Estimates, errors and df from this REML model (remember these are unstandardized estimates – do you want to standardize them?)
- Also a good idea to include the variance estimates for random effects

Model interpretation

```
> mod.1 <- lme(Richness ~ NAP + fExposure, random = ~1|Beach,
+              method = "ML", data = RIKZ)
> mod.1a <- lme(Richness ~ NAP, random = ~1|Beach,
+               method = "ML", data = RIKZ)
>
> mod.1b <- lme(Richness ~ fExposure, random = ~1|Beach,
+               method = "ML", data = RIKZ)
>
> anova(mod.1, mod.1a)
       Model df      AIC      BIC    logLik   Test  L.Ratio p-value
mod.1      1  5 244.7589 253.7922 -117.3795
mod.1a     2  4 249.8291 257.0557 -120.9145 1 vs 2 7.070141  0.0078
> anova(mod.1, mod.1b)
       Model df      AIC      BIC    logLik   Test L.Ratio p-value
mod.1      1  5 244.7589 253.7922 -117.3795
mod.1b     2  4 265.4332 272.6599 -128.7166 1 vs 2 22.6743  <.0001
>
```

What do you write in your paper?

- For overall significance of an effect it is most powerful to use the LRT here – report the likelihood ratio (probably the df too) and its p-value
- Generally better to report this than the individual t-tests from summary()

Effect	Estimate (s.e.)	Df	T-value	LLR	P-value
Fixed effects (marginal R^2 = XX, conditional R^2 = xx)					
Intercept†	8.60 (1.06)	35	8.11		
NAP	-2.58 (0.49)	35	-5.28	22.67	<0.001
Exposure: 11	-4.53 (1.57)	7	-2.87	7.07	<0.001
Random effects (Proportion of variance explained by Beach = 0.28)					
Beach	3.61				
Residual	9.35				

Table X. Results of linear mixed model testing the effects of NAP and exposure on species richness.
† intercept taken at NAP = 0 and exposure level 10

Table 1. Linear mixed effect model predicting mean velocity and maximum sustained swimming speed in the **Atlantic mollies**. Responses and length were first centered and scaled to unit variance, and test temperature was centered prior to analysis. Significance of effects were estimated using a log-likelihood ratio test on nested models; in models where a two-way interaction was significant, we did not test the significance of an involved main effect (see methods for more details). Estimates significant at the $p < 0.05$ level are bolded.

Effect	Estimate (± s.e.)	d.f.	t-value	LLR	p-value
Critical sustained swimming speed in a flume (marginal R^2 = 0.41; conditional R^2 = 0.72) [a]					
Intercept	0.80 (0.20)	6.49	4.04		
Length	**-0.22 (0.09)**	**46.1**	**-2.45**	**6.00**	**0.014**
Observation	-0.008 (0.025)	191.74	-0.31	0.15	0.70
Dev.temp (warm)	-0.68 (0.20)	75.16	-3.36		
Test.temp	**0.05 (0.006)**	**191.65**	**7.43**	**101.07**	**<0.001**
Test.temp^2	-0.01 (0.001)	191.60	-9.139		
Dev.temp x Test.temp	0.01 (0.009)	191.73	1.137	1.32	0.25
Dev.temp x Test.temp^2	**0.005 (0.001)**	**191.63**	**2.66**	**7.12**	**0.007**
Individual variance	0.219				
Mother variance	0.103				
Residual variance	0.303				
Adjusted repeatability[b]	0.35				
Mean velocity in an open field (marginal R^2 = 0.12, conditional R^2 = 0.55) [a]					
Intercept	-0.15 (0.20)	13.99	-0.77		
Length	**-0.35 (0.11)**	**45.91**	**-3.146**	**9.59**	**0.002**
Observation	-0.05 (0.11)	195.00	-1.71	3.17	0.07
Dev.temp (warm)	0.35 (0.25)	75.39	1.38	1.22	0.27
Test.temp	**0.01 (0.008)**	**195.00**	**1.34**	**3.76**	**0.05**
Test.temp^2	0.0001 (0.001)	195.00	0.107	0.66	0.42
Dev.temp x Test.temp	0.0008 (0.01)	195.00	0.07	0.004	0.94
Dev.temp x Test.temp^2	-0.002 (0.002)	195.00	-0.99	1.01	0.31
Individual intercepts variance	0.381				
Mother variance	0.055				
Residual	0.464				
Adjusted repeatability[b]	0.27				

[a] Marginal R^2 describes the proportion of the total variance that is explained by the fixed effects in the model whereas conditional R^2 describes the proportion of total variance that is explained by the combined fixed and random effects in the model .

[b] Repeatability was estimated as the proportion of the remaining variance (not explained by the fixed effects) that was attributable to differences in individual intercepts.

Posthoc testing

```
> 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
```

Are clams from Month 3 different from clams from Month 4?

You CAN'T tell that looking at this summary!!

Post-hoc testing

- Log likelihood ratio test tells you
  - Whether slope (continuous covariate) = 0
  - Whether two levels of a single factor are equal
- But if your factor has >2 levels, how to tell which ones are different?
  - Summary() only tells if each level different from the overall **intercept**

Post-hoc testing

- LOTS of options with different +/-
- They basically just all adjust p-values for multiple comparisons
  - Tukey's
  - Fisher's least significant difference
  - Duncan's multiple range
  - Newman-Keuls
  - Dunnett's correction

Example with Clams and RIKZ

Do it on your own – begging owls

- We are interested in investigating what factors influence begging behavior in adorable baby owls. We place cameras in 27 different nest boxes and record begging behavior many times over two nights.

- We are interested in the effects of food availability, so half the nests receive extra food and the other half of the nests we removed food. We also measured things like the sex of the parent (doing the feeding), and the arrival time of the parent.

- We think that the parents might adjust their arrival times differently based on the food treatment, and also that the food treatment might have different effects on the two (sexes) parents

What should our initial model look like?

Begging owls example – on your own!

- Explore the data
  - Histogram and dotplots of all variables
  - Boxplot of response against all predictors
  - Pairplots of responses and predictors
- Determine the proper random structure (using REML)
  - Compare model with and without nest
- Determine the proper fixed structure (using ML)
  - Start with a model including the ~~three~~ 2 2-way interactions
  - Check whether each 2-way can be removed
  - Check whether main effects can be removed
  - Re-run final model with
- Validate model assumptions
  - Plot residuals versus fitted
  - Plot residuals versus all fixed effects
  - Plot qqnorm of residuals
- Interpret your final output
  - Make a small table with the effect name, estimate, stn. error, LLR and p-value

Data exploration

- Understand the importance of data exploration and how to do it
- Understand the difference between exploration and fishing expeditions (and when each is appropriate)

- FURTHER READING
  - Zuur book – Appendix A2
  - Zuur et al. 2010. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology & Evolution* 1, 3-14
  - Zuur & Ieno. 2016. A protocol for conducting and presenting results of regression-type analyses. *Methods Ecol Evol* 7.

Model selection

- Learn and apply model selection procedures (usually based on good biological knowledge)
- Understand which factors can be included as random effects

FURTHER READING on model selection

- Zuur Chapter 5
- Schielzeth & Nakagawa. 2013. Nested by design: model fitting and interpretation in a mixed model era. *Methods Ecol Evol 4.*
- Forstmeier & Schielzeth. 2011. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behav Ecol Sociobiol 65*
- Engqvist. 2005. Mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. *Anim Behav 70.*