Quinn & Keough Exercises

Mick Keough & Gerry Quinn

2023-09-07

Contents

Introduction

These exercises are intended to get you familiar with the concepts underlying the various statistical approaches we introduced in the book *and* build your skills in taking a design, interrogating the data, and drawing some conclusions.

You won't find standard revisionary questions, such as "What are the assumptions underlying x?", or "Why would you use method x over method y?". These exercises will build your understanding of model structure, assessing assumptions, making decisions about how models fit and how to describe effects, but we think it better when you need to answer these questions while looking at data used to answer biological questions.

You will find more of the kinds of worked examples we used throughout the book - real world data analysis situations, taken from published papers. We'll give you a bit of context for the data, and then ask you a series of questions that will result in your translating the question to a statistical model, identifying the best way to fit that model, assessing assumptions, and describing the patterns that you infer from the data analysis.

For these data sets,

- You'll find the full reference, including a link to the paper, at the end of each chapter's exercises.
- We've used papers that are Open Access, as far as we know.
- The papers have been chosen because the authors followed the good practice of making the raw data available publicly. They are setting a great example.
- Getting the data.
  - In the ideal cases, you can just download the raw data and import it to R. For us, the ideal case means a csv file (or equivalent), with one row of headers, then the data. The variables have simple, short names, without spaces or other things that make us use quotes everywhere in our code.
  - In other cases, the files aren't quite as straightforward, and we've given you a bit of R code to import the file as a well-behaved dataframe. That step might include skipping rows or columns, renaming variables, etc.
  - In some cases, we've found it simpler to create our own data file and give you a link to it.
- In the worked examples, we showed you an approach that matched what the original paper's authors did. We did this to allow you to see how the authors put their data analysis into a larger story. In the exercises, we haven't been constrained that way - we've seen opportunities to make the analysis simpler or more complex, we've asked you to use more than one approach to a data set, and so on. As a result, you shouldn't worry about your results exactly matching what's in a paper.

Folder structures

If you're assembling the material into a local directory, you should take note of the folder structure we've used for material.

- The home folder includes:
  - the home page (index.html)
  - **data** folder
  - **media** folder, which includes images, word documents

R-wrangling

We've not provided the code for these exercises, other than to help you with getting hold of the data. The code you'll want will depend on the solutions you identify. However,...

- The worked examples provide a “bank” of code for particular tasks, and you may well have identified ways to improve or streamline that code
- Instructors have access to suggested solutions to the exercises, which include the code. If you’re working through these exercises in a way linked to a class, ask your instructor about code and solutions. We’ve run those solutions in *R* 4.3.1.
- We’ve often kept a bit of code, particularly when opening a file we’ve produced locally or to tidy up a data set, and we’ll show you the first few lines of the file as an illustration.

Chapter exercises

Exercises are mostly grouped by chapters …but 4 chapters (1-3, 14) don’t have any

Chapter 1

Relax and enjoy the scenery. This chapter sets the scene for the rest.

Chapter 2

This chapter is revisionary, and we'd use exams to check students' understanding. We tend to make these questions a gentle start to the inquisition.

Chapter 3

This chapter covers some basic design principles, which we pick up in some other exercises and assessment exercises.

Chapter 4

The aim of these exercises is to help you make the link between verbal descriptions of sample collection and the appropriate linear model for analysis.

Below you will find brief descriptions of several research studies. For each study, you should be able to answer the following questions:

1. What is the biological question? Think about the plausible cause-effect relationship between variables.
2. What is/are the response variable(s)?
   a. What distribution(s) might the response variable(s) follow? If you specified more than one distribution, how would you decide which one is appropriate?
3. What is the predictor variable?
   a. Is it continuous or categorical?
   b. If it is categorical, is it fixed or random?
4. Write a linear model for this data set
5. What link function would you use?
6. What estimation method would you use?

A. Foraging in elephant seals

Le Boeuf *et al.* (2000) examined the foraging behaviour of northern elephant seals (*Mirounga angustirostris*) that breed along the west coast of Mexico and the USA. They attached platform satellite transmitter terminals (PTTs) to 22 male seals and recorded, for each seal, the distance (km) to its main feeding area offshore and the amount of time (duration in days) it spent at the feeding area.

B. Invertebrates in mussel clumps

Relationships between species richness and habitat area has long been of interest to ecologists. Peake and Quinn (1993) investigated the relationship between the number of individuals and number of species of invertebrates living in amongst clumps of mussels on a rocky intertidal shore and the area of those mussel clumps.

C. Honeydew production in aphids

Vosteen *et al.* (2016) examined patterns of production of honeydew by different races of pea aphids (*Acyrthosiphon pisum*) and how that attracts ovipositioning hoverflies (*Episyrphus balteatus*) to create enemy-free space for the aphids. They measured honeydew production (mg) over 24 hours by three races of aphids (representing the native hosts they were collected from: *Triflolium*, *Pisum*, *Medicago* races) on plants of their native hosts and also the universal host plant *Vicia faba* in a climate chamber. There were six combinations of aphid race and host (native vs universal) plant that Vosteen et al (2016) treated as a single factor.

D. Effects of parasites on fish swimming

Binning *et al.* (2013) studied the effect of ectoparasitic isopods (*Anilocra nemipteri*) on the swimming ability of a tropical species of bream (*Scolopsis bilineatus*). They collected 18 unparasitized and 20 parasitized fish from Lizard Island on the Great Barrier Reef and created four treatment groups in the laboratory: eight unparasitized fish, 10 parasitized fish, 10 parasitized fish that had the parasites removed, and ten unparasitized fish that had model parasites made of plastic added. They recorded the swimming speed (body lengths per second) and oxygen consumption (mgO_2 per kg per hour) of each fish, in a respirometer.

E. Plant biomass in different grassland types

Dai *et al.* (2020) examined the relationships between above ground (AGB) and below ground (BGB) plant biomass from 80 sites across four types of grassland (temperate grassland, desert grassland, alpine meadow, meadow steppe) in a region of China.

F. Inbreeding and susceptibility to disease

One of the consequences of inbreeding in animal populations is that it is thought to increase the susceptibility of individuals to disease. To assess this Townsend *et al.* (2018) did genetic analysis of 178 crows found across California, and assessed their level of homozygosity (homozygosity by locus - HL), a measure of how inbred they were (higher levels of HL = more inbred). They also took blood samples from the same crows to check for the presence of avian malaria (*Plasmodium*), a common disease.

G. Vitamins and fish growth

Senadheera *et al.* (2012) tested whether vitamin B_6 affects fatty acid metabolism in cultured Rainbow trout, as part of a research program optimising the feeding of these fish. Fish were fed one of four experimental diets and the diets only differed in the amount of B_6. Treatments were applied to tanks containing 20 small fish, with three tanks per treatment. For each tank, they determined mean fatty acid metabolism, adjusted for fish biomass, as nmol g^{-1} d^{-1}.

Note: not open source, and need to get data if any further exploration is used

H. PFAS and quolls

Quolls are carnivorous marsupials of Australia. In southeastern Australia they live in dense forests and are endangered. A recent environmental concern has been raised about the risks from PFAS (Per- and Polyfluoroalkyl Substances). PFAS is a group of compounds widely used in firefighting, and they can be present around rural airports, firefighting training facilities, etc, which may be near quoll habitat. We wish to understand the risk posed by PFAS and we will visit sites with a wide range of PFAS concentrations. It’s been suggested that PFAS affects reproductive performance, so we will sample quolls during their reproductive season. Our target species, spot-tailed quolls, have litters of up to 6, carried in a simple pouch. Data collection involves collecting individual female quolls, counting their young, and taking a soil sample to estimate PFAS levels.

Chapter 5

The aim of these exercises is to help you get comfortable with running exploratory data analysis - taking a data set and a potential model, and evaluating whether the two are compatible. This activity will be part of most of the exercises in the later chapters. Particular things to look for are:

- Does the data set have extreme values (outliers)?
- How do you identify those values?
    - Are the observations legitimate or mistakes?
    - Are extreme values influential?
- What statistical model is proposed for this data set?
    - What are the important assumptions?
    - How can you evaluate whether those assumptions are satisfied?
    - If they aren't satisfied, what are your options?

Exploratory data analysis can be counter-intuitive. We emphasise its value as something to do before going ahead with fitting the model, assessing parameters, and testing hypotheses, **but** some assumptions can only be assessed by fitting the model. At this stage, we're aiming to fit the model without bothering to look at parameters of interest, confidence intervals, or P-values, but software doesn't always make this easy. We're trying to avoid the temptation of P-hacking or its equivalent - fitting a model that will give us desired results, rather than one that is appropriate for the particular data set. If you wanted to be super-rigorous about this step, you could write some Rmarkdown code where the results of model-fitting are suppressed (using RESULTS= 'Hide', etc.) and you just generate, e.g. residual plots!

A. Elephant seal foraging

Continue with the Le Boeuf *et al.* (2000) elephant seal example from Chapter 4's exercises.

For the linear model you've specified, what are the assumptions?

Open the data file and check those assumptions.

```
df <- read.csv(”data/leboeuf.csv”)
knitr::kable(head(df,10), booktabs=TRUE) %>%
    kableExtra::kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

male	departwt	distance	FFAduration	durationto	durationfrom
Pop	NA	534	31	18	11
Alt	973	755	89	9	8
Pro	977	1210	77	12	18
Hal	1121	NA	NA	NA	NA
Blu	NA	1297	76	19	25
Dua	996	1487	68	18	23
Rov	1100	2073	69	29	25
Ric	1068	2181	46	21	42
Ori	1097	NA	NA	NA	NA
Jer	1199	NA	NA	NA	NA

The data also include information on departure weight. Have a look to see if that variable might also be linked to foraging duration, and whether it might also be linked to distance travelled.

Hint: here's your chance at a scatterplot matrix

B. Invertebrates in mussel clumps

Continue with the Peake and Quinn (1993) example from Chapter 4's exercises, the relationships between the number of individuals and number of species (response variables) against mussel clump area (predictor variable).

For the linear model you've specified, what are the assumptions?

Open the **data file and check those assumptions.**

```
df <- read.csv(”data/peakquinn.csv”)
knitr::kable(head(df,10), booktabs=TRUE) %>%
     kableExtra::kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

area	indiv	species
516.00	18	3
469.06	60	7
462.25	57	6
938.60	100	8
1357.15	48	10
1773.66	118	9
1686.01	148	10
1786.29	214	11
3090.07	225	16
3980.12	283	9

C. Honeydew production in aphids

Continue with the Vosteen *et al.* (2016) study examining patterns of production of honeydew by different races of pea aphids (*Acyrthosiphon pisum*) from Chapter 4's exercises. You should have described a model to fit to these data.

For the linear model you've specified, what are the assumptions?

Read in the data and check the assumptions

```
df <- read.csv(”data/vosteen.csv”)
knitr::kable(head(df,10), booktabs=TRUE) %>%
     kableExtra::kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

clone_plant_combination	honeydew	clone	plant	hostplant
T_Vicia	1.08	T	Vicia	Universal
T_Vicia	2.21	T	Vicia	Universal
T_Vicia	2.63	T	Vicia	Universal
T_Vicia	1.63	T	Vicia	Universal
T_Vicia	3.51	T	Vicia	Universal
T_Vicia	2.53	T	Vicia	Universal
T_Vicia	2.92	T	Vicia	Universal
T_Vicia	0.98	T	Vicia	Universal
T_Vicia	2.39	T	Vicia	Universal
T_Vicia	2.05	T	Vicia	Universal

D. Parasites and fish swimming

Again, you examined in Chapter 4’s exercises the work of Binning *et al.* (2013) who studied the effect of ectoparasitic isopods on the swimming ability of a tropical species of bream. They created four treatment groups in the laboratory: eight unparasitized fish, 10 parasitized fish, 10 parasitized fish that had the parasites removed, and ten unparasitized fish that had model parasites made of plastic added. They recorded the swimming speed (body lengths per second) and oxygen consumption (mgO_2 per kg per hour) of each fish. The data are available from datadryad. Within dryad, the dataset you want is “binning etal 2012 one way anova.txt”; in this dataset, SMR is Standard Metabolic Rate (O_2 consumption), AS = factorial aerobic scope, and Ucrit is swimming speed.

In Chapter 4’s exercises, you should have described a model to fit to these data.

What are the assumptions associated with that model?

Read in the data and check the assumptions

```
df <- read.csv(”data/binning.csv”)
knitr::kable(head(df,10), booktabs=TRUE) %>%
      kableExtra::kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

Fish	Treatment	SMR	MMR	AS	Ucrit
P10	P	110.20	535.80	4.86	3.79
P12	P	140.11	471.45	3.36	3.66
P27	P	135.84	573.08	4.22	2.47
P42	P	140.44	379.69	2.70	3.65
P15	P	120.54	561.93	4.66	3.52
P23	P	108.02	375.57	3.48	3.39
P26	P	119.23	534.75	4.49	3.74
P37	P	152.73	494.95	3.24	3.37
P72	P	100.86	429.73	4.26	3.36
P75	P	134.09	434.60	3.24	3.68

E. Neuroanatomy of insectivours mammals

Kaufman *et al.* (2013) examined the neuroanatomy of a recently described species of sengi, which are small insectivorous mammals also known as elephant or jumping shrews. These animals are interesting, having been originally placed with the mammalian order Insectivora, along with shrews, hedgehogs, moles, etc., but this group is now known to be polyphyletic, and sengis are more appropriately grouped with elephants, dugongs, and hyraxes. They are in the order Macroscelidea, within the Afrotheria. The Afrotheria includes another order of small insectivores, the Tenrecoidea (tenrecs and golden moles). The Laurasiatheria also includes several families of small insectivores

Small insectivores are generally thought to have small brain mass (when adjusted for overall body mass), but there has been some question of whether sengis fit this pattern, and Kaufman and colleagues were curious whether the new species, *Rhynchocyon udzungwensis*, fitted with other sengi. They assembled data from 56 small insectivores, 5 sengi, 14 afrotherian species, and 37 laurasiatherians. For each species, they calculated brain mass (in mg) and total body mass (g).

Data are all presented in Table 1 of the paper. We've extracted it from the paper and it's here kaufman.csv

In the exercises for this chapter, we'll just think about brain size relative to body size, and we'll pick this example up again in Chapter 8.

Load the data file, and look at the relationship between brain mass and body mass.

```
df <- read.csv(”data/kaufman.csv”)
knitr::kable(head(df,10), booktabs=TRUE) %>%
     kableExtra::kable_styling(latex_options = c(”HOLD_position”,”scale_down”))
```

family	genus	species	bodymass	brainmass	relation	relation2
Solenodontidae	Solenodon	paradoxus	672.0	4723	laurasiatherian	other insectivore
Tenrecidae	Tenrec	ecaudatus	852.0	2588	afrotherian	other insectivore
Tenrecidae	Setifer	setosus	237.0	1516	afrotherian	other insectivore
Tenrecidae	Hemicentetes	semispin	116.0	839	afrotherian	other insectivore
Tenrecidae	Echinops	telfairi	87.5	623	afrotherian	other insectivore
Tenrecidae	Oryzorictes	talpoides	44.2	580	afrotherian	other insectivore
Tenrecidae	Microgale	cowani	15.2	420	afrotherian	other insectivore
Tenrecidae	Limnogale	mergulus	92.0	1150	afrotherian	other insectivore
Tenrecidae	Microgale	dobsoni	31.9	557	afrotherian	other insectivore
Tenrecidae	Microgale	talazaci	48.2	766	afrotherian	other insectivore

What kind of model are we intending to fit to these data?

Look at the relationship between the two variables? Are there any steps you'd recommend we take?

Note that the original researchers used a reduced major axis regression, as they considered both variables measured with error. Note the discussion in Chapter 6 about whether to consider X random or fixed. For our purposes here, we'll treat it as fixed

Chapter 6

The aim of these exercises is to make sure you're comfortable with fitting simple (one predictor) linear models to data and to emphasise the similarity between regression and "ANOVA". We will complete some of the analyses we introduced in Chapter 4's exercises.

For each of the 5 examples below, you should follow the sequence we've used previously:

1. What is the biological question and what are the response and predictor variables?
2. What distribution do you expect the response variable to follow?
3. Are the predictors continuous or categorical?
4. Write out the linear model corresponding to the biological question.
5. What are the assumptions behind the statistical model you'll fit?
    1. Are those assumptions satisfied?
6. Fit the model
    1. How will you assess whether the model fits well?
    2. Can you detect an effect of the predictor?
    3. How do you measure the effect?
7. What, if any, next steps would you suggest?
8. What do you conclude (including any cautions)?

Continuous predictor (regression)

A. Foraging elephant seals

Le Boeuf *et al.* (2000) examined the foraging behaviour of northern elephant seals (*Mirounga angustirostris*) that breed along the west coast of Mexico and the USA. They attached platform satellite transmitter terminals (PTTs) to 22 male seals and recorded, for each seal, the distance (km) to its main feeding area offshore and the amount of time (duration in days) it spent at the feeding area.

You can get this data file here.

B. Invertebrates in mussel clumps

Peake and Quinn (1993) investigated the relationship between the number of individuals and number of species (response variables) of invertebrates living in amongst clumps of mussels on a rocky intertidal shore and the area of those mussel clumps (predictor).

You can get this data file here.

C. Plant biomass in different grassland types

Dai *et al.* (2020) examined the relationships between aboveground (AGB) and belowground (BGB) plant biomass from 80 sites across four types of grassland (temperate grassland, desert grassland, alpine meadow, meadow steppe) in a region of China. Their research questions did not distinguish response and predictor variables so they used reduced major axis (RMA) regression to determine the slopes of the relationships between log(ABG) and log(BGB) for each grassland type and compare these to a slope of 1, indicating an (isometric) allometric relationship.

Categorical predictor ("ANOVA")

D. Honeydew production in aphids

Vosteen *et al.* (2016) examined patterns of production of honeydew by different races of pea aphids (*Acyrthosiphon pisum*) and how that attracts ovipositioning hoverflies (*Episyrphus balteatus*) to create enemy-free space for the aphids. They measured honeydew production (mg) over 24 hours by three races of aphids (representing the native hosts they were collected from: *Triflolium*, *Pisum*, *Medicago* races) on plants of their native hosts and also the universal host plant *Vicia faba* in a climate chamber. There were six combinations of aphid race and host (native vs universal) plant that Vosteen et al (2016) treated as a single factor.

The data from this paper are available through *datadryad*. For this question, the data set we'll use is for Figure 2e. The authors' analysis is provided and uses the data set ...2e...collection.txt, but we want to do a different analysis in Chapter 7, so the file with additional factors is here.

Describe the terms in the linear model if this study was treated as a two-factor (aphid race and host plant) crossed design (Chapter 7).

E. Effects of parasites on fish swimming

Binning *et al.* (2013) studied the effect of ectoparasitic isopods (*Anilocra nemipteri*) on the swimming ability of a tropical species of bream (*Scolopsis bilineatus*). They collected 18 unparasitized and 20 parasitized fish from Lizard Island on the Great Barrier Reef and created four treatment groups in the laboratory: eight unparasitized fish, 10 parasitized fish, 10 parasitized fish that had the parasites removed, and ten unparasitized fish that had model parasites made of plastic added. They recorded the swimming speed (body lengths per second) and oxygen consumption (mgO_2 per kg per hour) of each fish, in a respirometer.

The data are available from datadryad. Within dryad, the dataset you want is "binning etal 2012 one way anova.txt"; in this dataset, SMR is Standard Metabolic Rate (O_2 consumption), AS = factorial aerobic scope, and Ucrit is swimming speed.

You should be able to use the answer from the previous section to generate the code you'll need

F. Effects of herbal supplements on rat physiology

Here is an example from Kiss *et al.* (2017). They were interested in the effects of intake of an invasive weed (ragweed) on health of humans as this species is now being used as a herbal supplement. They did an experiment using Wistar laboratory rats. Twenty-four rats were randomly allocated to one of three treatment groups, with eight rats in each group. Group 1 was a control group which was just fed cookie dough, group 2 rats were fed cookie dough with a low dose of ragweed and group 3 rats were fed cookie dough with a high dose of ragweed. The total amount of feed was the same in each group and the experiment ran for 28 days, with rats fed daily. At the end of the experiment, a range of blood parameters were measured and we will focus on aspartate aminotransferase (ast).

You can access the data file here. However, it's an Excel file not really formatted for *R*. You'll need to delete some rows and redo the treatment labels to be able to use it as a data frame. Our reformatting of the file is here. Check your file against it if you have problems.

Again, you should use the same approach as for the two previous questions (although the answers will be different!)

Chapter 7

The aim of these exercises is to make sure you're comfortable with fitting linear models with two or more categorical predictors. We will focus on normally distributed responses.

For each of the 5 examples below, you should follow the sequence we've used previously:

1. What is the biological question and what are the response and predictor variables?

2. What distribution do you expect the response variable to follow?

3. Are the predictors continuous or categorical?

4. Write out the linear model corresponding to this question.

5. What are the assumptions behind the statistical model you'll fit?

    1. Are those assumptions satisfied?

6. Fit the model

    1. How will you assess whether the model fits well?

    2. Can you detect an effect of the predictor?

    3. How do you measure the effect?

7. What next steps would you suggest?

8. What do you conclude (including any cautions)?

A. Psychostimulant effects on vole physiology

Hostetler *et al.* (2016) examined the effects of methamphetamine (MA) on hypothalamic neuropeptides in the brains of laboratory-housed prairie voles. Individual voles were the experimental units, and there were two fixed factors. Treatment, representing access to MA or water, was an experimental factor, crossed with sex (male vs. female), an observational factor, with five or six voles in each of the four combinations.

We'll start with the oxytocin data as the response variable. The data are in the supplementary files of the paper or use our version. Import hostetler_OT data file (hostetler OT.csv), and we'll make sure that sex and treatment are treated as factors

```
host_ot <- read.csv(”data/hostetler␣OT.csv”)
kable(head(host_ot,10), booktabs=TRUE) %>%
      kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

Animal.ID	trtmt	sex	pvn1	pvn2	pvn3	pvn4	pvn5	pvn6	otmean
231.15-5	0	1	35	32	56	51	NA	NA	43.5
220.17-2	1	1	39	44	41	29	34	27	35.7
248.1-4	0	1	39	30	55	34	80	73	51.8
240.7-4	1	1	26	26	70	56	NA	NA	44.5
220.17-4	0	1	48	55	68	70	NA	NA	60.3
231.15-6	1	1	66	51	38	52	50	39	49.3
220.17-5	0	1	57	59	59	47	NA	NA	55.5
240.7-3	1	1	25	29	39	37	35	31	32.7
232.17-2	0	1	42	57	NA	NA	NA	NA	49.5
220.17-3	1	1	43	56	59	83	NA	NA	60.3

```
# make sex & trtmt factors
host_ot$sex<-as.factor(host_ot$sex)
host_ot$trtmt<-as.factor(host_ot$trtmt)
```

Hint. To help you analyse this data set in R, have a look at the online version of Box 7.3, which includes some relevant R code.

Would the same approach work for the ATV response?

B. Metabolic responses of salmon to temperature change

Kraskura *et al.* (2020) examined sex-specific differences in the metabolic responses of coho salmon (*Oncorhynchus kisutch)* to three different temperatures: 9°C (typical), 14°C (current maximum) and 18°C (climate change scenario). Fish were collected from a hatchery in British Columbia, Canada, and returned to a research laboratory and held in outdoor tanks at 9°C. Temperatures were raised slowly to the relevant test temperature and held for 2 days. Fish were then transferred individually to a swim tunnel respirometer where critical swim speed was measured by increasing the water velocity until the fish could no longer maintain its position in the water column. In addition, the maximum metabolic rate was also recorded during each trial. The fish were subsequently euthanized and their sex determined.

Analyze these data to examine the effects of temperature, sex and the interaction between these two fixed factors on absolute critical swim speed (cm s^{-1}).

The data are available on Dryad. They are in an Excel document that needs some tidying before using in R, as sheets with data include columns for summary statistics, etc. A tidied version is here.

```
krask <- read.csv(”data/krask_swim.csv”)
kable(head(krask,10), booktabs=TRUE) %>%
    kable_styling(latex_options = c(”HOLD_position”,”striped”))# make sex &
        temperature factors
```

fish_id	sex	temperature	bm_test_kg	ucrit_bl_s	ucrit_cm_s
Oct06fish1	F	9	2.461	2.836	167.756
Oct07fish2	F	9	1.525	2.488	113.567
Oct08fish1	F	9	1.885	3.592	135.208
Oct09fish1	F	9	1.614	3.184	112.879
Oct10fish2	F	9	1.984	3.647	155.070
Oct14fish1	F	9	2.142	3.132	160.883
Oct14fish2	F	9	2.028	2.767	105.653
Oct16fish1	F	9	2.581	2.936	168.327
Oct16fish2	F	9	2.708	2.638	140.878
Oct16fish3	F	9	3.255	3.086	154.580

```
krask$sex<-as.factor(krask$sex)
krask$temperature<-as.factor(krask$temperature)
```

Would the same approach work for the maximum metabolic rate ($mgO_2\ kg^{-1}\ min^{-1}$) response?

The tidied metabolic rates file is here; the variable we're interested in is total_epoc

```
krask <- read.csv("data/krask_metab.csv")
kable(head(krask,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c("HOLD_position","striped", "scale_down"))#
         make sex & temperature factors
```

fish_id	sex	temperature	bm_test_kg	fl_test_cm	mmr	rmr	AAS	FAS	total_epoc
10_16_b1_1	F	9	1.63	52.7	12.160	1.380	10.780	8.812	512.49
10_16_b1_2	F	9	2.10	41.1	11.980	1.127	10.853	10.630	616.91
10_16_b1_3	F	9	1.62	55.0	9.750	1.099	8.651	8.872	367.30
10_16_b2_3	F	9	2.61	59.6	8.840	1.280	7.560	6.906	387.92
10_22_b1_1	F	9	2.54	60.8	11.916	1.262	10.654	9.442	612.14
10_22_b1_2	F	9	2.54	59.3	11.501	1.080	10.421	10.649	690.88
10_22_b2_1	F	9	2.56	60.1	13.595	0.989	12.606	13.746	485.28
10_22_b2_3	F	9	2.74	61.7	14.832	1.368	13.464	10.842	2782.92
10_30_b1_1	F	9	2.17	57.8	10.100	1.225	8.875	8.245	698.77
10_30_b1_2	F	9	2.31	58.4	7.839	1.308	6.531	5.993	622.76

```
# make sex & temperature factors
krask$sex<-as.factor(krask$sex)
krask$temperature<-as.factor(krask$temperature)
```

C. Dietary supplements and osteoarthritis in dogs

This exercise is a little more complex, designed to start thinking about using additional predictors to give a clearer picture of any signals from predictors of interest, i.e. to reduce background noise in the data. Two things to note:

- We'll assume you've completed the previous two exercises, so you're familiar with multifactor models, the questions they answer, etc. We won't go through the initial checklist again, but recommend that you do it anyway.
- We'll come back to this example in exercises for later chapters (10, 11, and 13).

The example we'll use is from Martello *et al.* (2022), who tested whether dietary supplements, which are often expensive, can help with symptoms of osteoarthritis in dogs. Their study was experimental, with 40 dogs with chronic pain symptoms. Dogs were also screened and excluded if, for example, they had recent surgery, were taking steroidal medication, etc. The dogs were then allocated randomly to one of two treatments, a dietary supplement that included glucosamine sulfate and chondroitin, and a placebo. The study was completely blind, so administration of treatments and recording of data was done with no knowledge of what each dog had received. The supplements were administered for 60 days, and dogs received treatments orally at 2g/10kg body weight. The treatments are the variable GROUP.

Two response variables were used, both involving scoring of dogs' behaviour and activity:

1. Dog owners used a questionnaire, responses to which generated a value on a 40 point Helsinki Chronic Pain Index, and these assessments were done at the start of treatment, and after 40 and 60 days. In the data file, these measures are the variables HCPI, HCP.4 and HCP.6
2. Veterinary assessments were done at the same 3 times. These assessments were on a 5-point scale reflecting increasing degrees of lameness. These variables are SEGNI.OA.VET, SEGNI.OA.VET.4, SEGNI.OA.VET.6

We won't deal with them here, but veterinarians also measured a range of physiological and biochemical variables from a blood sample at 0 and 60 days.

The authors stated that osteoarthritis is affected by a range of other factors, including sex, body weight, breed, etc., and they also considered whether other things like desexing, with its hormonal changes, might be important. They measured several other potential predictors:

- Sex
- Sterilization (Yes/No)
- Body weight (PESO.KG)
- Breed (RAZZA)
- Estimated age (ETA)
- Body condition score, which combines all of these to produce a 4-point scale (BCS).

The data are in the supporting information (and a copy is here).

```
df <- read.csv(”data/martello.csv”)
kable(head(df,10), booktabs=TRUE) %>%
    kable_styling(latex_options = c(”HOLD_position”,”striped”,”scale_down”))
```

group	PAZIENTE	FAR	breed	wt	eta	ster	sex	REGIONE.ANATOMICA	Note	ESAMI.EMATICI	RX	bcs	hcp0	hcp40	hcp60	vet0	vet40	vet60	clin.BASELINE.MV.0	clin.FUP.MV.2	clin.FUP.MV.4	clin.FUP.MV.6	clin.BASELINE.PR.0	clin.FUP.PR.2	clin.FUP.PR.4	clin.FUP.PR.6
CTR	21	B	METICCIO	21	10	N	F	ANCA DX		NRM	SI	3	25	25	25	2	2	2	1	1	1	2	1	1	1	1
CTR	22	B	DOGUE DEB	47	6	N	M	GOMITI		NRM	SI	3	26	26	28	2	2	2	1	1	1	2	1	1	1	1
CTR	23	B	PASTORE T	23	9	N	M	GOMITO DX		NRM	SI	2	32	32	30	3	3	3	2	1	2	2	2	0	1	2
CTR	24	B	METICCIO	32	8	S	F	ANCA DX/SX		NRM	SI	3	34	34	32	3	3	3	2	2	1	1	3	3	3	3
CTR	25	B	METICCIO	29	4	N	M	GOMITO DX/SX		NRM	SI	3	20	20	22	1	1	2	0	0	1	2	1	1	1	1
CTR	26	B	METICCIO	35	12	N	M	ANCA DX/SX		NRM	SI	4	26	26	28	2	2	2	1	1	1	2	1	1	1	1
CTR	27	B	METICCIO	30	8	N	F	GOMITO DX	CARPO DX	NRM	SI	4	28	28	28	2	2	2	1	1	1	1	1	0	0	1
CTR	28	B	LABRADOR R	28	11	N	F	GINOCCHIO DX		NRM	SI	3	22	22	20	1	1	1	0	0	1	1	0	1	0	0
CTR	29	B	GOLDEN R	24	7	S	F	ANCA SX		NRM	SI	3	38	28	28	3	3	2	0	0	1	2	1	1	0	0
CTR	30	B	LAGOTTO	18	9	S	F	GINOCCHIO DX	TARSO DX	NRM	SI	3	40	36	40	3	2	3	3	3	3	3	2	2	2	1

```
# make sex & temperature factors
#Tidy up the names
#df_names<-c(group=”GROUP”, ster=”sterilizzato”, bcs=”BCS”,eta=”ETA”, breed=”RAZZA
    ”,
#                    hcp0=”HCPI”, hcp40=”HCPI.4”, hcp60=”HCP.6”,
#                    vet0=”SEGNI.OA.VET”, vet40=”SEGNI.AO.VET.4”, vet60=”SEGNI.AO.VET
    .6”)
#df<-rename(df, !!!df_names)
# make sex, treatment, and sterilization factors
df$sex<-as.factor(df$sex)
df$group<-as.factor(df$group)
df$ster<-as.factor(df$ster)
```

The question of most interest is whether the dietary supplements result in an improvement after 60 days, and for now we'll include sex in the analysis, as there are good grounds for expecting a range of sex-based differences.

Identify an appropriate model for examining whether the dog's condition, as assessed by their owners, after 60 days, depends on whether a dog received the supplement or a placebo and whether those results depended on sex of the dog.

Before starting analysis, we'd do a little tidying up of the data file, to stay consistent with using lower-case for variable names and doing away with periods, etc. It doesn't really matter, but it keeps things tidy. As we read in the data, we'll also declare a few of the variables as factors.

Start with boxplots, then fit model and look at residuals

Would including desexing in the analysis improve the model fit or the conclusions?

Anything else worth checking?

Have a think about the analysis and the experimental design. We've analyzed the dogs' status at the end of the experiment, and relied on randomization to keep us away from misleading results. Think about how you might use the measurements at time 0 to do some additional checks or analyses. Think of a couple of ways and then investigate them.

For interest, you could try using the body condition score (bcs) instead of sex and sterilization, as they were chosen by the authors to play the same role.

Other things to consider

As you work through later chapters, we'll use this data set to explore a few other things: 1. See whether body weight plays a role in the response. Dosage rates were adjusted for body weight, but there are other considerations around forces acting on joints, etc. 2. Make use of the 3 measurements on each dog and see if there are more subtle aspects of the response 3. Take a look at the vet data; with only 5 values, it might be better not to try and force it into an analysis that presumes a normal distribution.

D. Honeydew production in aphids

We'll return to the example from the Chapter 6 exercises, where Vosteen *et al.* (2016) examined patterns of production of honeydew by different races of pea aphids (*Acyrthosiphon pisum*) and how that attracts ovipositioning hoverflies (*Episyrphus balteatus*) to create enemy-free space for the aphids. They measured honeydew production in response to combinations of aphid races and native/universal hosts, and treated these combinations as a single factor.

We could view aphid race and whether the aphids were on their native host or a universal one as two separate factors. Use the modified data file to see what happens when you separate these factors.

```
#Read in the data and assign it to an object df
df <- read.csv("data/vosteen.csv")
kable(head(df,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c("HOLD_position","striped"))
```

clone_plant_combination	honeydew	clone	plant	hostplant
T_Vicia	1.08	T	Vicia	Universal
T_Vicia	2.21	T	Vicia	Universal
T_Vicia	2.63	T	Vicia	Universal
T_Vicia	1.63	T	Vicia	Universal
T_Vicia	3.51	T	Vicia	Universal
T_Vicia	2.53	T	Vicia	Universal
T_Vicia	2.92	T	Vicia	Universal
T_Vicia	0.98	T	Vicia	Universal
T_Vicia	2.39	T	Vicia	Universal
T_Vicia	2.05	T	Vicia	Universal

```
#Make sure clone and hostplant are factors
df$clone<-as.factor(df$clone)
df$hostplant<-as.factor(df$hostplant)
```

Chapter 8

The aim of these exercises is to improve your ability to deal with multi-predictor linear models where we have a single continuous response and the two or more predictors are all continuous or a mixture of continuous and categorical.

A. Dietary supplements and canine osteoarthritis

Let's continue with the Martello *et al.* (2022) example from Chapter 7 assessing whether dietary supplements improve the perceived health of dogs with osteoarthritis. The model we focused on at the end of that exercise was one modelling the pain index of dogs after 60 days as a function of whether they received dietary supplements or a placebo and the sex of the dog. The dogs unavoidably varied in body weight, ranging from 14-47 kg. To partly account for this, the authors adjusted the doses to a constant amount per kg of body weight. However, you can probably think of a range of ways in which weight might affect osteoarthritis. The model using treatment and sex fitted the data fairly well, with r^2 around 0.6. We detected a strong treatment effect, but it is possible that if we reduced background noise, we might see sex-specific responses and we'd also get a more precise estimate of the effects.

Think about the steps you'd take to see if it would be helpful to include body weight in the model, then go back to the data and run the analysis.

```
df <- read.csv(”data/martello.csv”)
knitr::kable(head(df,10), booktabs=TRUE) %>%
  kable_styling(latex_options = c(”HOLD_position”,”scale_down”))
```

group	PAZIENTE	FAR	breed	wt	eta	ster	sex	REGIONE.ANATOMICA	Note	ESAMI.EMATICI	RX	bcs	hcp0	hcp40	hcp60	vet0	vet40	vet60	clin.BASELINE.MV.0	clin.FUP.MV.2	clin.FUP.MV.4	clin.FUP.MV.6	clin.BASELINE.PR.0	clin.FUP.PR.2	clin.FUP.PR.4	clin.FUP.PR.6
CTR	21	B	METICCIO	21	10	N	F	ANCA DX		NRM	SI	3	25	25	25	2	2	2	1	1	1	2	1	1	1	1
CTR	22	B	DOGUE DEB	47	6	N	M	GOMITI		NRM	SI	3	26	26	28	2	2	2	1	1	1	2	1	1	1	1
CTR	23	B	PASTORE T	23	9	N	M	GOMITO DX		NRM	SI	2	32	32	30	3	3	3	2	1	2	2	2	0	1	2
CTR	24	B	METICCIO	32	8	S	F	ANCA DX/SX		NRM	SI	3	34	34	32	3	3	3	2	2	1	1	3	3	3	3
CTR	25	B	METICCIO	29	4	N	M	GOMITO DX/SX		NRM	SI	3	20	20	22	1	1	2	0	0	1	2	1	1	1	1
CTR	26	B	METICCIO	35	12	N	M	ANCA DX/SX		NRM	SI	4	26	26	28	2	2	2	1	1	1	2	1	1	1	1
CTR	27	B	METICCIO	30	8	N	F	GOMITO DX	CARPO DX	NRM	SI	4	28	28	28	2	2	2	1	1	1	1	1	0	0	1
CTR	28	B	LABRADOR R	28	11	N	F	GINOCCHIO DX		NRM	SI	3	22	22	20	1	1	1	0	0	1	1	0	1	0	0
CTR	29	B	GOLDEN R	24	7	S	F	ANCA SX		NRM	SI	3	38	28	28	3	3	2	0	0	1	2	1	1	0	0
CTR	30	B	LAGOTTO	18	9	S	F	GINOCCHIO DX	TARSO DX	NRM	SI	3	40	36	40	3	2	3	3	3	3	3	2	2	2	1

B. Effects of floral traits on milkweed fitness components

La Rosa and Conner (2017) examined effects of several floral traits on fitness components of milkweeds, *Asclepias* spp. The fitness components were male and female pollination success and female reproductive success.

In the paper, their analysis focused on 6 predictors, They measured six floral traits, although one of them, hood height, was not relevant for **Asclepias viridiflora**, which was the species with the largest sample size:

- gynostegium width,
- hood length,
- hood height,
- horn reach,
- slit length, and
- gap width

Their Figure 2 shows what these measurements represent on flowers.

The data are available from Dryad here.

Fitness component estimates were relativized by dividing by the mean, and the traits were standardized to a mean of zero and standard deviation of one.

Start by looking at *A. syriaca*, then for comparison, look at how these floral traits affect *A. viridiflora*

First look at the removals per flower

- What checks should you do before assessing the predictors' effects?
- If you're happy with your pre-flight checks, fit the model and make some conclusions about the effects of each predictor, including any notes of caution

Run through same sequence for the other two life-history traits

- What would you need to check in doing analyses on three different fitness components as response variables?
- What do you conclude about the floral traits' influence on fitness components of this species?

Now have a look at the data for the more common species *Asclepias viridiflora*

- What do you conclude about the role of floral traits in these two species?
- Is there anything you'd be cautious about in making this comparison?

C. Brain and body mass in insectivorous mammals

Recall the sengi example (Kaufman *et al.* (2013)) from Chapter 5 (or go back and look at it ;-)) where we looked at the relationship between brain mass and body mass in small insectivore species. Now we will look at how this relationship varies between families of insectivores, including whether sengis are different from the rest. There are 3 groups of insectivores, sengis and closely (afrotherian) and more distantly (laurasiatherian) species, and the research question is about where sengis fit. We can frame this as 2 or 3 questions.

1. Does the new species (*udzugwensis*) fit within the pattern of other sengi?
2. Are sengi different from the other small insectivores in their brain size?
    1. sengi vs all others, or
    2. sengi vs closely-related vs distantly related insectivores

Get started by loading the kaufman data.

```
df <− read.csv(”data/kaufman.csv”)
kable(head(df,10), booktabs=TRUE) %>%
  kable_styling(latex_options = c(”HOLD_position”,”striped”, ”scale_down”))
```

family	genus	species	bodymass	brainmass	relation	relation2
Solenodontidae	Solenodon	paradoxus	672.0	4723	laurasiatherian	other insectivore
Tenrecidae	Tenrec	ecaudatus	852.0	2588	afrotherian	other insectivore
Tenrecidae	Setifer	setosus	237.0	1516	afrotherian	other insectivore
Tenrecidae	Hemicentetes	semispin	116.0	839	afrotherian	other insectivore
Tenrecidae	Echinops	telfairi	87.5	623	afrotherian	other insectivore
Tenrecidae	Oryzorictes	talpoides	44.2	580	afrotherian	other insectivore
Tenrecidae	Microgale	cowani	15.2	420	afrotherian	other insectivore
Tenrecidae	Limnogale	mergulus	92.0	1150	afrotherian	other insectivore
Tenrecidae	Microgale	dobsoni	31.9	557	afrotherian	other insectivore
Tenrecidae	Microgale	talazaci	48.2	766	afrotherian	other insectivore

In Chapter 5, you should have decided that log-transforming both variables was sensible, so lets also start by defining new variables logbrain and logbody. That will make the coding tidier, without having to log things repeatedly.

```
df$logbrain <- log(df$brainmass)
df$logbody <- log(df$bodymass)
```

For the first question, cast your mind back to Chapter 6. How would you decide whether the new species is unusual?

Now lets compare sengis to the other insectivores. Use three groups for comparison (sengi, Afrotherian and Laurasiatherian). These groups are defined in the variable *relation*

** You could make this comparison in two ways:

- fit a linear model including the groups as a categorical factor and log body mass as a covariate, i.e. an analysis of covariance
- look at the patterns in the residuals for the relationship between log-brain and log-body

**Before you start, are there any things to check in the original data, linked to the assumptions of the linear model you'll fit?

Analysis of covariance **Outline the steps you'll take**

Run the analysis

Use residuals from a regression of all data and compare residuals between groups

D. Elephant seal foraging

We'll return to the elephant seal example in the study by Le Boeuf *et al.* (2000) and see whether body weight plays any role in foraging. In Chapter 5, you should have noticed that while the focus of the initial analysis was on the relationship between time spend on the foraging grounds and distance travelled, the authors recorded weight on departure for each animal. Your exploratory data analysis should have shown a relationship between body weight and the original predictor and response variables. Now try and make some sense of what's going on here.

Think about how body mass might influence distance travelled and how it might contribute to time on foraging areas

How will you assess whether including body weight as a second predictor helps us understand feeding time better?

- Write out the linear model you'd apply

- What checks do you need when fitting the model?

```
#Get the data file back
df <- read.csv("data/leboeuf.csv")
kable(head(df,10), booktabs=TRUE) %>%
  kable_styling(latex_options = c("HOLD_position","striped"))
```

male	departwt	distance	FFAduration	durationto	durationfrom
Pop	NA	534	31	18	11
Alt	973	755	89	9	8
Pro	977	1210	77	12	18
Hal	1121	NA	NA	NA	NA
Blu	NA	1297	76	19	25
Dua	996	1487	68	18	23
Rov	1100	2073	69	29	25
Ric	1068	2181	46	21	42
Ori	1097	NA	NA	NA	NA
Jer	1199	NA	NA	NA	NA

Fit the appropriate model to the data, interpret the results, and explain whether body weight helps us.

Is there anything else you might look at?

Chapter 9

The aim of these exercises is to extend the analyses from Chapter 8, focusing on identifying the relative importance of predictors in linear models that include at least one continuous predictor.

A. Floral traits and fitness components in milkweed

Recall the example of La Rosa and Conner (2017) from the Chapter 8 exercises. They examined effects of up to six floral traits on fitness components of milkweeds, *Asclepias* spp. The fitness components were male and female pollination success and female reproductive success.

The data are available from Dryad here. Fitness component estimates were relativized by dividing by the mean, and the traits were standardized to a mean of zero and standard deviation of one. You can also get the data from larosa.csv.

```
df <- read.csv(”data/larosa.csv”)
knitr::kable(head(df,10), booktabs=TRUE) %>%
     kableExtra::kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

species	plant.id	gyn.width	hood.length	hood.height	horn.reach	slit.length	gap.width	display.flowers.1day	remo
Asyr	AS01	0.2190	0.3557	0.5095	0.2038	0.1918	0.0464	168	
Asyr	AS02	0.2042	0.2776	0.5293	0.1820	0.1843	0.0597	120	
Asyr	AS03	0.2233	0.2771	0.5083	0.1809	0.1687	0.0664	117	
Asyr	AS04	0.2081	0.2403	0.4241	0.1520	0.1796	0.0515	76	
Asyr	AS05	0.2136	0.3544	0.5368	0.1940	0.1842	0.0459	148	
Asyr	AS07	0.2276	0.2984	0.4867	0.2161	0.1953	0.0840	61	
Asyr	AS08	0.2103	0.3155	0.4390	0.1806	0.1967	0.0581	69	
Asyr	AS09	0.2261	0.3385	0.5744	0.2101	0.1898	0.0597	388	
Asyr	AS10	0.2214	0.3508	0.5406	0.2295	0.1997	0.0525	47	
Asyr	AS11	0.2175	0.3166	0.4320	0.1501	0.1746	0.0535	35	

```
df_syr<-subset(df,species==”Asyr”)
df_vir<-subset(df, species=='Avir')
df_tub<-subset(df, species=='Atub')
```

For each fitness component and for each species separately (i.e. 6 models in total):

Refit the linear models from the Chapter 8 exercises but now use the recommended methods from this chapter (hierarchical partitioning or LMD, PMVD) to assess each predictor's relative importance in each of the models.

If you're really keen, there's a third species in the dataframe (*Atub*).

Now use AIC and Aikake weights to find the most parsimonious model (best fit with fewest predictors) for each combination of fitness component and species. If there are multiple models with similar AICs, then use full (zero) model averaging to produce a final model.

B. Diet and land-use drives mercury accumulation in wolverines

Peraza *et al.* (2023) studied what factors drove mercury accumulation in muscle tissues of a high-altitude carnivore, the wolverine (*Gulo gulo*). Wolverine muscle (for Hg) and hair (for N and C stable isotopes) samples were obtained from carcasses submitted by trappers and hair snags across four Canadian provinces. We will focus on total Hg concentration (µg.gdw) in muscle as the response and 14 predictor variables measured at the point of collection:

- delta15N and delta13C from hair samples,
- longitude,
- latitude,
- precipitation,
- mean maximum and mean minimum temperatures,
- elevation,
- soil organic C,
- distance from nearest coast,
- mean soil pH at 10cm and 60cm depths, and
- Hg net and Hg wet deposition rates.

Start by reading in the data.

```
peraza <- read.csv("data/peraza␣clean.csv")
knitr::kable(head(peraza,10), booktabs=TRUE) %>%
     kableExtra::kable_styling(latex_options = c("HOLD_position","scale_down","
        striped"))
```

ageclass	sex	long	lat	thg	delta15n	delta13c	hgdep	hgwet	prec	tempmax	tempmin	elev	dist	soc	sph10	sph60	X
Adult	Male	-139.2	64.4	0.095	5.775	-24.480	8.644	5.731	30.459	0.242	-10.445	1205	496304.4	19.000	56	58	NA
Yearling	Female	-136.5	62.2	1.184	9.570	-25.438	10.364	3.209	22.764	3.282	-8.797	954	721306.5	144.104	56	60	NA
Adult	Female	-137.4	62.8	0.498	6.692	-25.216	12.601	2.918	25.705	2.845	-8.578	687	653101.6	114.319	54	60	NA
Yearling	Female	-133.1	60.7	0.308	5.747	-25.773	9.165	3.253	28.955	4.186	-7.525	926	883503.6	69.691	56	58	NA
Adult	Male	-135.4	63.3	0.058	6.330	-25.684	9.067	2.700	31.001	3.142	-9.092	598	590826.8	106.367	66	70	NA
Yearling	Male	-139.2	64.4	0.084	5.594	-23.896	8.644	5.731	30.459	0.242	-10.445	1205	496304.4	19.000	56	58	NA
Adult	Male	-140.6	64.6	0.120	4.000	-33.459	11.238	2.964	25.016	3.363	-8.605	442	501614.8	93.346	57	61	NA
Yearling	Male	-137.4	62.8	0.595	7.223	-26.486	12.601	2.918	25.705	2.845	-8.578	687	653101.6	114.319	54	60	NA
Yearling	Male	-137.4	62.8	0.309	7.280	-25.037	12.601	2.918	25.705	2.845	-8.578	687	653101.6	114.319	54	60	NA
Adult	Male	-132.0	61.3	0.052	4.728	-25.333	6.673	3.062	31.631	3.376	-9.474	1048	830036.0	81.455	55	58	NA

We will fit a multiple linear regression model relating total Hg to the predictors.

Do the usual pre-analysis checks of assumptions using boxplots, a scatterplot matrix and VIFs.

Note the strong collinearity between min and max temperature, between soil pH at 10 and 60cm and between distance from coast and latitude. The authors removed min temperature, latitude and pH at 60cm from their model.

Fit a multiple regression model relating log total Hg to the remaining 11 predictors and check the residual plot.

Any indication of outliers of concern?

We recommend you proceed with the model with all data, but note that Perazo et al. omitted 17 observations as outliers so your results will differ somewhat from theirs.

What conclusions would you draw from your linear model?

Now use the methods from the first question to evaluate the relative importance of the different predictors and find the most parsimonious model.

C. Predictors of insect richness in freshwater streams

Tonkin *et al.* (2015) surveyed 80 freshwater stream sites in mid-latitude China to determine how different climate and catchment (watershed) variables predicted the richness of three different insect groups (Ephemeroptera, Plecoptera, Trichoptera; collectively abbreviated as EPT). They recorded 32 predictor variables in total but to avoid collinearity (r > 0.7), only 17 variables were included in the analyses (see their Table 1).

The full dataset is available at http://dx.doi.org/10.6084/m9.figshare.1305679 but a tidied-up version including only the non-collinear predictors is available here. We will also not include region (a categorical variable) as a predictor, resulting in 16 predictors. Tonkin et al used Poisson regression models to link richness to these predictors but for the purposes of this chapter, we will treat richness as normally distributed (its distribution wasn't very skewed - you can use boxplots to see if you agree).

```
tonkin <- read.csv("data/tonkin.csv")
knitr::kable(head(tonkin,10), booktabs=TRUE) %>%
     kableExtra::kable_styling(latex_options = c("HOLD_position","scale_down","
         striped"))
```

sitecode	region	ept	ephem	plec	trich	trees_bl	trees_nl	shrub	herbaceous	cultivated	water	bio1	bio4	bio8	bio15	bio18	ai	pet	elevation	slope	catch_size
BS01	East	7	6	0	1	0.00000	100.00000	0.00000	0	0	0	170	8420	211	55	573	14284	1153	99	2.631222	76
BS02	East	14	10	1	3	0.00000	100.00000	0.00000	0	0	0	157	8406	198	53	588	15367	1062	292	2.318342	11
BS03	East	16	13	0	3	0.00000	84.61538	15.38462	0	0	0	164	8476	205	54	578	14841	1103	186	2.871561	13
BS04	East	17	14	0	3	0.00000	100.00000	0.00000	0	0	0	165	8503	206	54	556	14424	1094	145	1.416681	5
BS05	East	11	6	2	3	0.00000	100.00000	0.00000	0	0	0	164	8354	205	54	585	14670	1124	211	3.630091	1
BS06	East	12	10	1	1	0.00000	100.00000	0.00000	0	0	0	157	8264	197	54	615	15977	1084	338	5.192832	8
BS07	East	17	7	6	4	0.00000	100.00000	0.00000	0	0	0	166	8407	206	55	580	14747	1127	157	2.908305	1
BS08	East	18	12	4	2	42.85714	57.14286	0.00000	0	0	0	165	8400	206	54	582	15317	1104	167	6.563747	7
BS09	East	27	13	6	8	66.66667	33.33333	0.00000	0	0	0	159	8327	199	54	603	15790	1081	285	6.270956	6
BS10	East	9	7	1	1	0.00000	100.00000	0.00000	0	0	0	166	8487	207	54	572	14708	1115	126	3.333718	4

Fit a multiple linear regression model relating total EPT richness to the 16 predictors. Note that the ratio of observations to predictors is very marginal!

Which predictors had the strongest influence on EPT richness?

Now use a standard regression tree to relate richness to the predictors.

How do the results compare?

Tonkin et al used boosted regression trees (BRTs) - Fit a BRT

Use the same settings as they did (with the default bag fraction of 0.5):

- a slow learning rate of 0.0005 (to ensure 1000 trees were reached) and
- tree complexity at 5.
- model validation was based on 10-fold cross-validation.

Chapter 10

The aim of these exercises is to start getting familiar with mixed models, and in these exercises we'll be focused on models with multiple categorical predictors, with at least one predictor a random effect. The random effects can be crossed or be nested within other predictors, creating a nested or multilevel structure.

For each of the examples below, you should follow the sequence we've used previously:

1. What is the biological question and what are the response and predictor variables?
2. What distribution do you expect the response variable to follow?
3. Is each predictor
    1. Continuous or categorical?
    2. Fixed or random?
4. Write out the linear model corresponding to the biological question.
5. What are the assumptions behind the statistical model you'll fit?
    1. Are those assumptions satisfied?
6. Fit the model
    1. How will you assess whether the model fits well?
    2. Can you detect an effect of the fixed predictors?
    3. How much of the variance is attributable to each random predictor?
7. What do you conclude (including any cautions)

Things to look out for

In fitting these models, we'll need to make the initial checklist more complex:

- Think about the relationship between predictors - are they crossed (factorial) or nested (hierarchical)? Make sure that your linear model reflects these relationships!
- When there are random effects, assumptions can be more extensive. We're usually interested in the fixed effects, and different fixed effects can have different assumptions. This is particularly the case with nested designs.

Crossed mixed model designs

Remember from Chapter 10 that these designs are typically used for two purposes.

Random effects are often used to reduce background noise in the data (e.g. randomized blocks designs), where the random effect is of little intrinsic interest. In a sense, we sacrifice degrees of freedom from the residual variance, anticipating that the variance attributed to the random effects will be large enough to make this sacrifice worthwhile. When you see these designs, it's a good idea to look and see if the use of a random effect seems a good decision, e.g. by looking at blocking efficiency.

In other cases, the random effects are used to estimate how consistently fixed effects act, by estimating variance in their effects across, for example, an environmental spectrum (fixed predictor).

A. Leaf domatia and mites

Walter and O'Dowd (1992) were interested in testing the hypothesis that leaves of the shrub *Viburnum tinus* with domatia (small shelters at the juncture of veins on leaves) have more mites than leaves without domatia. Fourteen paired leaves on a shrub of *V. tinus* were randomly chosen and one leaf in each pair had its domatia shaved while the other remained as a control; the number of mites was recorded on each leaf (experimental unit) after two weeks.

The data file is here: walter.csv.

```
df <- read.csv("data/walter.csv")
kable(df) %>%
    kable_styling(latex_options = c("HOLD_position", "striped"))
```

leaf	pair	domatia	mites
a1	a	intact	9
a2	a	shaved	1
b1	b	intact	2
b2	b	shaved	1
c1	c	intact	0
c2	c	shaved	2
d1	d	intact	12
d2	d	shaved	4
e1	e	intact	15
e2	e	shaved	2
f1	f	intact	3
f2	f	shaved	1
g1	g	intact	11
g2	g	shaved	0
h1	h	intact	6
h2	h	shaved	0
i1	i	intact	7
i2	i	shaved	1
j1	j	intact	6
j2	j	shaved	0
k1	k	intact	5
k2	k	shaved	1
l1	l	intact	8
l2	l	shaved	1
m1	m	intact	3
m2	m	shaved	1
n1	n	intact	6
n2	n	shaved	0

This file has four variables:

- *leaf* is just a code to identify individual leaves
- *pair* identifies the pair to which that leaf belongs
- *domatia* describes the experimental treatments, whether domatia were shaved off or left intact
- *mites* is the number of mites recorded from that leaf

Use the sequence at the start of these exercises to make some conclusions about the role of domatia on this shrub species and the variability between leaf pairs.

Focus for now on using an OLS approach, so think about the response variable and how you might treat it.

Why do you think the authors chose to run the experiment with pairs of leaves?

Did that pairing decision improve their ability to say something about domatia?

Extra questions:

Are there alternatives to how you treated the response variable in your analysis.

Hint: there are at least a couple

Would your conclusions have changed?

B. Rainfall intensity and frequency and grassland plants

Didiano *et al.* (2016) assessed the effects of rainfall intensity and frequency on performance of grassland plants in Ontario, attempting to understand potential effects of future climates. The area is projected to receive much more summer rainfall than at present. They established a field experiment in which plants were grown from seeds in pots subjected to one of two rainfall levels (70 and 90 mm per month) and three frequencies (3, 5, and 15 days per month). They were interested in whether responses of plants varied according to whether the plants were monocots (5 species, though only 4 grew successfully) or eudicots (10, species).

The experiment was done using shelters (see their figure S2) that prevented natural rainfall, allowing the researchers to control water received by the plants. They used 10 shelters, each of which had 6 groups of 18 pots (see their figure S1). Each group received one of the 6 frequency/amount combinations. The 18 pots in each group had one of the 15 target species, with three pots used to measure soil moisture levels during the experiment.

For each seedling, they recorded leaf number (which didn't change much), plant height, and above- and below-ground biomass at the end of the experiment. They also calculated a ratio of above:below biomass.

The data are available from dryad; the data you'll use are the first sheet in the Excel file, which you can import directly, or use Excel to save that first sheet as a csv file. To make things a little complicated, the data file uses only common names, while the paper uses formal scientific names. We've made our own version of the file with species names and monocot/eudicot added - it's didiano.csv

Let's keep things simple by focusing on one response variable (above.ground.biomass) and one species (Big bluestem)

```
didiano <- read.csv("data/didiano.csv")
#Now create subset using just one species
df<-subset(didiano, Species=="Big bluestem")
kable(head(df,10), booktabs=TRUE) %>%
    kable_styling(latex_options = c("HOLD_position","scale_down"))
```

Plant.ID	Shelter	Plot	Pot	Species	Amount	Frequency	X1st.leaf.measurement	X1st.plant.height.measurement	X2nd.leaf.measurement	X2nd.plant.height.measurement	Above.ground.biomass	Below.ground.biomass	Above..to.below.ground.ratio	Sp_name	Type
43	1	1	10	Big bluestem	70	3	21	18.5	15	51.0	3.2582	5.11740	15.65281444	Andropogon	monocot
56	1	2	26	Big bluestem	70	15	14	14.0	18	44.0	3.5384	5.13360	12.43499887	Andropogon	monocot
42	1	3	43	Big bluestem	70	30	28	21.0	40	122.5	9.9042	10.92450	12.36849014	Andropogon	monocot
18	1	4	61	Big bluestem	90	3	9	18.0	23	59.0	4.5037	7.07450	13.10033972	Andropogon	monocot
81	1	5	87	Big bluestem	90	15	28	18.5	32	103.0	7.5734	12.62010	13.60023239	Andropogon	monocot
80	1	6	102	Big bluestem	90	30	17	18.0	24	101.5	6.6888	18.50870	15.17462026	Andropogon	monocot
167	2	7	117	Big bluestem	70	3	22	20.0	21	74.0	5.4688	13.18560	13.53130486	Andropogon	monocot
155	2	8	133	Big bluestem	70	15	43	20.0	33	61.0	9.5101	23.75450	6.414233289	Andropogon	monocot
181	2	9	157	Big bluestem	70	30	20	19.0	34	58.0	4.9502	1.07738	11.71669832	Andropogon	monocot
106	2	10	160	Big bluestem	90	3	22	17.0	49	81.0	10.0241	26.37970	8.080525933	Andropogon	monocot

Use the sequence at the start of these exercises to make some conclusions about the role of rainfall amount and frequency on this species and the variability between shelters.

Focus for now on using an OLS approach, so think about the response variable how you might treat it.

Now use REML instead

Why do you think the authors chose to lay out the experiment with 10 shelters?

Did that design decision improve their ability to say something about the rainfall effects?

Simplified model

The appropriate model for full RCB has Amount x Shelter and Frequency x Shelter. Didiano et al. apparently did not include these terms.

Redo the earlier analysis using their simpler model as starting point

Extension activities

1. For the species we've chosen, did rainfall have the same effect on plant height (2nd.plant.height.measurement) and below-ground biomass?
2. Was there a difference in how monocot and eudicot species responded? They used 14 species in their experiment, with the aim of comparing responses of monocot and eudicot species?
   - The monocot species (all Poaceae) were *Andropogon gerardii* (Big bluestem) *, Elymus canadensisa* (Canada wild rye)*, Elymus riparius* (Riverbank rye)*, Panicum virgatum,* and *Sorghastrum nutans* (Indian grass), but *S. nutans* failed to grow.
   - The eudicots were *Asclepias tuberosa* (butterflyweed)*, Desmodium canadense* (Showy tick-trefoil)*, Eupatorium perfoliatum* (Common boneset)*, Euphorbia corollata* (flowering spurge)*, Geum triflorum* (Prairie smoke)*, Oenothera biennis* (Common evening primrose)*, Penstemon hirsutus* (Hairy beardtongue)*, Solidago nemoralis* (Gray goldenrod)*, Verbena stricta* (Hoary vervain), and *Zizia aurea* (Golden alexanders)*,* which were spread over 9 families.
   - Hint: you might want to simplify this question by focusing just on main effects, particularly of rainfall amount
   - If you start working through more analyses, you'll get some experience working with missing values as well for some species! **Note that the simplified model isn't affected to same degree by occasional missed plants

As an example, run the analysis and interpret it using the data for *Zizia*

C. Psychostimulant effects on fruit flies

Highfill *et al.* (2019) examined effects of two psychostimulants of public health concern (cocaine and methamphetamine) using *Drosophila melanogaster* as a “translational” model species. Translational reflected similarities in dopamine transport and the overall effects of exposure in fruitflies and humans. Their interest was in variation in susceptibility to these compounds, so they examined responses of lines of flies chosen from a large set of inbred lines (the Genetic Reference Panel). Groups of 5 flies from a line were placed into test vials and their consumption of sucrose media and sucrose + drug media recorded after 18 h. There’s a nice diagram of the test vial here. Please look at it before proceeding.

It’s not uncommon to find sex-based differences in consumption, so they used groups of males and groups of female flies for each line. Combinations were replicated 10 times.

The final complexity was that they were interested in whether drug sensitivity changed with time, so they allowed 6 hours for flies to feed, then ran another trial, followed by more food, and a third.

What is this design?

As you work your way through the analysis sequence described at the start of these exercises, there are two important questions for you to answer. This is a complex example, so start with a factor diagram (see, e.g. Figure 10.2). It will help to clarify the relationships between the different factors. Look at the data file to see how they’re named. You can download it here

```
library(readxl)
highfill <- read_excel(”data/pgen.1007834.s001.xlsx”)
highfill <- highfill %>%
  dplyr::rename(Line = ”DGRP␣Line”) %>%
  dplyr::rename(vial = Replicate)
kable(head(highfill,10), booktabs=TRUE) %>%
      kable_styling(latex_options = c(”HOLD_position”))
```

Solution	Sex	Line	vial	Consumption	Exposure
Cocaine	F	41	1	34.6	1
Cocaine	F	41	2	23.6	1
Cocaine	F	41	3	22.6	1
Cocaine	F	41	4	38.6	1
Cocaine	F	41	5	11.6	1
Cocaine	F	41	6	37.6	1
Cocaine	F	41	7	54.6	1
Cocaine	F	41	8	13.6	1
Cocaine	F	41	9	59.6	1
Cocaine	F	41	10	81.6	1

How many factors are there in the design?

Hint: it’s more than 3!

What model corresponds to the design?

1. Factorial mixed model, with 3 fixed effects (Solution, Sex, Exposure) and one random effect (Line)
2. Partly-nested mixed model, with vials as plots or subjects. Sex, Exposure and Line as between-plot effects and Drug Effect (the pairs of capillary tubes) as the within-plots effect

3. Partly-nested mixed model, with vials as plots or subjects. Sex and Line as between-plot effects and Drug Effect and Exposure as within-plots effect.

Time for some analysis

We'll approach the analysis of this data set in a couple of ways - simple and complex. The complex way is an extension exercise for Chapter 12. The keenest of you might like to take on the challenge of specifying and running this model. Do it after you've finished the Chapter 12 exercises, and with reference to Box 12.5.

Ways to simplify things

We can also make the main question (variation in sensitivity to a psychostimulant) more tractable. The stimulant effect is the difference between sucrose and the drug solution, so we could calculate that for each vial.

We still have the question of what to do with the three trials (Exposure).

– You could jump ahead to Chapter 12 and see whether sensitivity varies with time , and whether that variation varies with line

– Use the three trials as just a way of estimating sensitivity , i.e. average them

– Choose one of the trials as ”the” endpoint , e.g. Trial 3

For this chapter's exercises, lets stick with the two simplifications - calculate sensitivity and average it across trials, and calculate sensitivity and look at the last trial (Exposure = 3).

Start by rearranging the data so the Sucrose and Drug consumption values are columns, rather than appearing on different rows

```
df <-  highfill %>%
  tidyr::pivot_wider(names_from = Solution, values_from = Consumption)
```

```
## Warning: Values from `Consumption` are not uniquely identified; output will
    contain
## list-cols.
## * Use `values_fn = list` to suppress this warning.
## * Use `values_fn = {summary_fun}` to summarise duplicates.
## * Use the following dplyr code to identify duplicates.
##   {data} %>%
##   dplyr::group_by(Sex, Line, vial, Exposure, Solution) %>%
##   dplyr::summarise(n = dplyr::n(), .groups = ”drop”) %>%
##   dplyr::filter(n > 1L)
```

```
df$Sucrose <-unlist(df$Sucrose)
df$Cocaine <-unlist(df$Cocaine)
df$Pref <-df$Cocaine - df$Sucrose
df$Line <-as.factor(df$Line)
df$Sex <- as.factor(df$Sex)
df$Exposure <- as.factor(df$Exposure)
kable(head(df,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c(”HOLD_position”))
```

Sex	Line	vial	Exposure	Cocaine	Sucrose	Pref
F	41	1	1	34.6	50.0	-15.4
F	41	2	1	23.6	59.0	-35.4
F	41	3	1	22.6	55.0	-32.4
F	41	4	1	38.6	44.0	-5.4
F	41	5	1	11.6	27.0	-15.4
F	41	6	1	37.6	79.8	-42.2
F	41	7	1	54.6	84.8	-30.2
F	41	8	1	13.6	85.8	-72.2
F	41	9	1	59.6	65.8	-6.2
F	41	10	1	81.6	81.8	-0.2

Analyze just one survey

Analyze mean of 3 exposure times

What do you conclude?

Do you see variation among *Drosophila* genetic lines?

Does that variation depend on the sex of flies?

Does your conclusion match that of the original authors in their paper?

Designs with nesting

D. Piscicide effects on trout

Birceanu and Wilkie (2018) examined potential concerns around the use of 3-trifluoromethyl-4-nitrophenol (TFM). This chemical is used to control an invasive species (sea lamprey) in the Great Lakes of North America. Control is directed at larval lampreys, and to be effective, it must be applied at relatively high concentrations. There is a question of risks this may pose to other fish species. Bircenau and Wilkie focused on physiological effects on rainbow trout (*Oncorhynchus mykiss*).

In the data we'll look at, they focused on sub-lethal, environmentally relevant concentrations of TFM, and they considered effects on two pathways involved in detoxification of TFM. The hypothalamic-pituitary-interrenal (HPI) axis mediates the stress response of fishes, and on liver metabolic capacity. Many organisms detoxify pollutants via the liver.

Physiological measurements were done on individual trout. Fish were exposed to TFM (or control) for 9h, then stressed for 12h, after which physiological measurements were taken.

The animal husbandry involved fish being placed in large (180 L) tanks, with 24 fish per tank. Six tanks were used, and after fish had acclimatized for 24 h, TFM was added to three of the tanks. After 9h, all tanks were flushed and fish left overnight. After this time, 6 fish were removed from each tank and assayed, then fish were stressed (by chasing them vigorously for 3 min), then 6 more assayed after 1 and 4 h, then a further 6 sampled after 24h. Their Figure 1 has a diagram of the experimental setup.

```
df <- pone_0200782_s003 <- read_excel(”data/pone.0200782.s003.xlsx”,
    sheet = ”Liver␣glycogen␣on␣SPSS”, col_types = c(”text”,
        ”text”, ”text”, ”numeric”))
```

```
# Relabel glyggen
df <- df %>%
  dplyr::rename(Glycogen = ”Glycogen␣(umol/mg␣protein)”) %>%
  dplyr::rename(Tank = ”Tank#”)
kable(head(df,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

Treatment	Tank	Time	Glycogen
TFM	1	0h	406.0258
TFM	1	0h	1160.8861
TFM	1	0h	250.4860
TFM	2	0h	449.8723
TFM	2	0h	2417.7273
TFM	2	0h	372.0748
TFM	2	0h	1080.8588
TFM	3	0h	3119.3762
TFM	3	0h	3651.7944
TFM	3	0h	3682.6876

Start by going through the steps at the start to be clear about the question, the kinds of predictors, etc. Pay particular attention to experimental units, and what the treatments are.

Let's start with a simple question: what was the situation 4h after stress? Did exposure to TFM affect the stress response at 4h?

Describe two models that you could fit to address this question

Fit both of these models, using OLS (or you can go mixed model/REML if you feel like it)

What do you conclude about TFM?

What is the variation between tanks in Glycogen at 4h?

Do you have any suggestions for alternative ways to do this experiment?

What if we include all 4 stages?

How would your analysis change if we wanted to include very short and longer term responses to stress? Have a quick think about this and then take it up again when you get to Chapter 11.

E. Effects of altered climates on freshwater fish

Cramp *et al.* (2014) examined two aspects of changing climates, altered levels of UVB radiation and increased temperatures. Their interest was in the capacity for these changes to alter physiology (measured as oxygen consumption, VO2) and susceptibility to parasites (measured as prevalence, i.e. whether parasites were present, and intensity, the number of parasites present). They were especially interested in whether UV and temperature acted independently or synergistically, and they investigated this question with a small, widely-distributed freshwater fish, *Gambusia holbrooki*.

The experiment involved wild-caught fish (from the University of Queensland campus) that were brought into the laboratory and housed in 2L tanks. There were 40 tanks, each with 15 fish. After initial acclimitization, tanks were assigned randomly to one of four combinations of temperature (18 or 25 °C) and UVB (high and low, corresponding to a more than 10-fold difference). After 10 days, fish were challenged with a common pathogen, whitespot, which is an ectoparasitic ciliate. Fish were left with the parasites for 8 d, after which the presence or absence of parasites was recorded (prevalence), the number of parasites counted (intensity), and a measure of each fishes metabolic rate (as oxygen consumption) taken.

Follow the steps at the top of these exercises to be sure that you are clear on the design and model

Let's get the data and have a quick look. The data are available from Dryad, and there are two worksheets we're interested in here, VO2 and Parasite intensity.

```
library(readxl)
cramp_vo2 <- read_excel(”data/Cramp␣et␣al␣raw␣data.xlsx”,
      sheet = ”VO2”)
cramp_intens <- read_excel(”data/Cramp␣et␣al␣raw␣data.xlsx”,
      sheet = ”Infection␣Intensity”)

cramp_vo2 <- cramp_vo2 %>%
   dplyr::rename(UV = ”UV␣Level”) %>%
   dplyr::rename(VO2 = ”VO2␣(ml␣h-1)”) %>%
   dplyr::rename(Temperature = 'Temperature␣(oC)')
cramp_intens <- cramp_intens %>%
   dplyr::rename(UV = ”UV␣Level”) %>%
   dplyr::rename(Intensity = ”Whitespots/fish”) %>%
   dplyr::rename(Temperature = 'Temperature␣(oC)')

cramp_intens$Temperature<-as.factor(cramp_intens$Temperature)
cramp_intens$Tank <-as.factor(cramp_intens$Tank)
cramp_vo2$Temperature<-as.factor(cramp_vo2$Temperature)
cramp_vo2$Tank <-as.factor(cramp_vo2$Tank)

kable(head(cramp_vo2, 10)) %>%
      kable_styling(latex_options = c(”HOLD_position”, ”striped”))
```

UV	Temperature	Tank	VO2	Body mass (g)
Low	18	1	0.0405296	0.1245
Low	18	1	0.0386433	0.1079
Low	18	1	0.0263919	0.0746
Low	18	2	0.0596725	0.1291
Low	18	2	0.0437979	0.0990
Low	18	2	0.0073663	0.1132
Low	18	3	0.0198926	0.0468
Low	18	3	0.0196705	0.0505
Low	18	4	0.0131924	0.0502
Low	18	4	0.0255141	0.0682

```
kable(head(cramp_intens, 10)) %>%
      kable_styling(latex_options = c(”HOLD_position”, ”striped”))
```

UV	Temperature	Tank	Intensity
Low	18	1	15
Low	18	1	0
Low	18	1	0
Low	18	1	12
Low	18	1	11
Low	18	1	15
Low	18	1	15
Low	18	1	14
Low	18	1	11
Low	18	2	11

Start by looking at metabolic performance of fish

What are the two main models you could fit to assess the effects of UV and Temperature (Hint: the difference is how tanks are incorporated into the model)?

What assumptions are made in fitting each model?

Use the two approaches to analyse VO2 and outline your conclusions

What are your thoughts about the effects of UV and Temperature on VO2?

And for a bit more, use the approach you've developed to look at parasite intensity

If you're still feeling enthusiastic, keep these data in mind when you get to Chapter 13, and take a look at prevalence (it's a separate sheet in the Excel file from dryad)

Chapter 11

The aim of these exercises is to ...

For each of the examples below, you should follow the sequence we've used previously:

1. What is the biological question?
2. Is the predictor continuous or categorical?
3. Write out the linear model corresponding to this question.
4. What distribution do you expect the response variable to follow?
5. What are the assumptions behind the statistical model you'll fit?
    1. Are those assumptions satisfied?
6. Fit the model
    1. How will you assess whether the model fits well?
    2. Can you detect an effect of the predictor?
    3. How do you measure the effect?
7. What do you conclude (including any cautions)

A. Effects of propagule pressure and water depth on revegetation

Li *et al.* (2015) investigated ways of revegetating damaged freshwater enviroments with vegetation. They used a suite of 4 plant species, and focused on the combined effects of propagule pressure (i.e., the initial seeding event) and water depth on the aquatic vegetation that developed. They reported data for each plant species separately and considered overall plant "community", putting all species together.

The four plant species can propagate clonally, by shoot regeneration, and they used shoot fragments to "seed" experimental pots. Each pot had 1, 2, or 4 fragments of each of the four species (i.e. there were three propagule pressures). The pots, filled with natural lake sediment, were placed in larger plastic tanks, which were filled to a depth of 30 or 70 cm. There were 5 tanks for each depth. The tanks were outside in full sun, and plants were allowed to grow for 7 weeks, after which time plants were harvested and measured.

For each species in a pot, five shoots were selected and their length measured, the number of nodes (indicating clonal growth) counted, and dry biomass calculated. The total biomass for that species was also measured and used to estimate total nodes and total shoot length. Values for each species were then used to calulate total biomass, total shoot length, and total number of nodes for that pot.

Their data can be found in Table S1 of the paper, but we've extracted the data and renamed some of the columns to make things clearer. Our file is here.

```
df <−read.csv(”data/li.csv”)
set_sum_contrasts()

## setting contr.sum globally: options(contrasts=c('contr.sum', 'contr.poly'))
```

```
kable(head(df,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

depth	tank	pressure	biomass	nodes	slength	hvbiomass
30	1	low	3.204	1161.9653	14.095498	2.606
30	1	medium	3.523	1456.9005	16.637120	2.496
30	1	high	7.077	1679.0204	19.736082	5.690
30	2	low	3.969	1257.9132	15.239968	2.789
30	2	medium	4.470	1402.1528	20.200615	3.609
30	2	high	8.041	2172.3806	28.823653	5.935
30	3	low	0.530	371.1823	4.193746	0.339
30	3	medium	1.473	756.2558	8.357996	1.019
30	3	high	1.719	1191.0750	15.192225	1.532
30	4	low	1.879	814.2920	10.042774	1.521

```
df$depth<-as.factor(df$depth)
df$tank<-as.factor(df$tank)
```

How do propagule pressure and water depth address these three measures of the plant community?

Remember to go through the steps at the start of these exercises to make sure you have identified the design and assumptions clearly

B. Effects of predation and sediment type on clam growth

Seitz *et al.* (2016) studied the estuarine clam *Macoma balthica* and set up a field experiment to examine the effects of two different sediment types (shallow-mud with high food availability and muddy-sand with low food availability) and predation by fish and crabs on the growth rate of clams in upper Chesapeake Bay. They used a split-plot design where there were eight shallow-mud and four muddy-sand sites, with more shallow-mud sites because of greater variability; sediment type was the between-plots fixed factor with random sites (“plots”) nested in each type. Within each site, there was one sub-plot (0.25m^2) that excluded predators with a cage and a second sub-plot that had no cage; predation was the within-plots fixed factor. Ten individually marked and measured clams were transplanted into each sub-plot and retrieved 20-22 days later to record growth.

The data are available from dryad. The data for this example are in *GrowSplitPlot.xlsx*, but it needs quite a bit of work for us to work with it in R. Our tidied version is here

Load the data and have a quick look. The variable names should be clear

```
df <- read.csv(”data/seitz.csv”)
df$Plot <-as.factor(df$Plot)
kable(head(df,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

Habitat	Plot	Growth	Predation
Mud	1	0.7	Exclusion
Mud	1	0.7	Exclusion
Mud	1	0.2	Exclusion
Mud	1	1.1	Exclusion
Mud	1	0.9	Exclusion
Mud	1	0.3	Exclusion
Mud	1	0.0	Exclusion
Mud	2	0.1	Exclusion
Mud	2	0.6	Exclusion
Mud	2	0.1	Exclusion

Focus on the relationship between habitat type and predation

Examine this effect using a linear mixed model approach and REML

If we follow the example of Box 11.3, there is some variation in opinion as to whether we'd try a slightly simpler model, omitting the Predation x Plot interaction, which doesn't account for much variation.

What do you think of this option, and would it make a difference?

Can you think of ways to simplify the analysis of this data set?

Simpler is better when explaining your analysis to an audience!

C. Phenotypic plasticity of lake whitefish ecotypes

Dalziel *et al.* (2015) were interested in disentangling the processes of acclimation and adaptation to changing environments, and examined levels of phenotypic plasticity in a small fish, the lake whitefish. This fish has developed two ecotypes - phenotypes that feed in different ways. Dwarf ecotypes are actively swimming and feeding, but slow-growing, while the ancestral normal phenotype is more sedentary. It's thought that these ecotypes evolved separately under past climates, but then came together and can co-occur. There are anatomical and physiological differences between the ecotypes that are linked to swimming ability, and Dalziel and her colleagues were interested in how these ecotypes responed to a changed environment with faster-flowing water, particularly any role of phenotypic plasticity.

Phenotypic plasticity is best assessed by placing different phenotypes into the same set of varied environments and measuring their performance. They did this for whitefish by using roughly size-matched, laboratory-fertilized and reared fish from each ecotype, which were then placed in two environments, still water and fast-flowing water. Fish were kept in these environments for approximately 3 months, then sampled.

Flow environments were produced in 1 m high circular tanks, .6 m in diameter, with a .16 m diameter cylinder in the centre of the tank, creating a circular swimming "race track". Within this track, water flowed at 7.6 cm/s for 6 h/day or 0.5 cm/s. Each tank contained 8 fish of each ecotype.

At the end of the experiment, fish were sacrficed, and the researchers took hematocrit samples, measured heart morphology and took samples of red and white muscle. Red mussel samples were used for enzyme assays and histology, and white muscle to obtain mitochondria, whose performance was assessed.

They provided data and R code through dryad. We'll focus on their data focusing on mitochondrial function in white muscle. These data are in their file *Fig5_MitoRespiration.csv.* If you want to explore some of their other analyses, you can use their R script, which allows easy selection of data files.

```
plasticity<-read.csv(”data/Fig5_MitoRespiration.csv”,header=TRUE, stringsAsFactors =
    FALSE, na.strings = c(”NA”,””))
kable(head(plasticity,10), booktabs=TRUE) %>%
      kable_styling(latex_options = c(”HOLD_position”,”striped”, ”scale_down”))
```

indv	ecotype	treatment	group	tank	mass	length	mitoprotein.ml	mitoprotein.gWM	state2.ml	state2.mg	state3.ml	state3.mg	state4.ml	state4.mg	RCR	ACR	flux1.4.ml	flux1.4.mg	flux2.4.ml	flux2.4.mg	flux.cox.ml	flux.cox.mg	notes1
WP1	dwarf	swim	DS	1	3.2197	7.580	6.25	0.622	74.0	11.8	404.2	64.7	137.8	22.0	2.9	5.5	350.5	56.1	89.8	14.4	955.9	152.9	mitoprotein.ml=mitochondrial protein per mL mito resuspension (from BSA assay)
WP2	dwarf	control	DC	2	4.1880	7.959	5.26	0.526	42.3	8.1	344.9	65.6	128.7	24.5	2.7	8.1	404.8	77.0	121.2	23.0	388.4	73.8	mitoprotein.gWM1= mitochondrial protein per gram tissue based upon calculations from uL buffer added
WP3	normal	swim	NS	3	3.2915	6.501	4.49	0.458	24.3	5.4	261.3	58.2	122.5	27.3	2.1	10.8	286.4	63.8	98.0	21.8	264.0	58.8	NA
WP4	normal	control	NC	4	4.6945	7.840	3.40	0.260	33.5	9.9	331.9	97.6	90.0	26.5	3.7	9.9	405.9	119.4	96.0	28.2	750.7	220.8	NA
WP5	normal	control	NC	5	4.2127	7.582	3.49	0.347	28.3	8.1	266.1	76.2	63.7	18.3	4.2	9.4	310.8	89.1	97.3	27.9	608.3	174.3	NA
WP6	normal	swim	NS	6	3.7213	7.133	4.16	0.435	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
WP7	normal	control	NC	7	3.9910	7.766	5.28	0.541	27.7	5.2	249.2	47.2	77.4	14.7	3.2	9.0	388.8	73.6	47.9	9.1	671.8	127.2	NA
WP8	dwarf	swim	DS	8	5.8855	9.091	4.84	0.484	53.4	11.0	548.8	113.4	69.0	14.2	8.0	10.3	675.8	139.6	171.1	35.4	1143.5	236.3	NA
WP9	normal	swim	NS	8	4.5422	7.742	2.92	0.299	34.9	12.0	248.3	85.0	61.8	21.2	4.0	7.1	281.3	96.3	59.7	20.5	687.1	235.3	NA
WP10	normal	control	NC	7	4.8839	NA	4.22	0.427	39.3	9.3	306.7	72.7	59.1	14.0	5.2	7.8	374.8	88.8	94.9	22.5	748.6	177.4	NA

Look at data measuring fluxes of four oxidative complexes, I-IV. These are the variables flux1.4.mg, flux2.4mg, and flux.cox.mg

flux1.4.mg is total flux across the four complexes, flux2.4.mg is flux across complexes II-IV, and flux.cox.mg is flux across complex IV

```
#Make sure categorical predictors are treated the right way
plasticity$treatment <-as.factor(plasticity$treatment)
plasticity$ecotype <-as.factor(plasticity$ecotype)
plasticity$tank <-as.factor(plasticity$tank)
```

What do you conclude about the ecotypes and their plasticity for these mitochondrial fluxes?

D. Nutrients, warming and browsing as influences on shrub range expansion

Morrissette-Boileau *et al.* (2018) examined factors affecting a shrub (dwarf birch) that is potentially expanding its range northwards with warming. They were interested in modifying factors - enhanced nutrients, warming, and browsing by caribou.

Caribou were not manipulated directly, but were excluded by fencing them out, and their foraging simulated by manual removal of 0, 25, or 75% of available shoots once each spring.

Nutrients were manipulated in 4 x 24 m plots, half of which received a nutrient supplement at bud-burst each year.

Each 4 x 24 area was divided into 4 x 4 m sub-plots. Each of these sub-plots had a 1 m x 1m area that was the subject of a combination of warming and caribou treatments.

Global warming was simulated using open-top hexagonal chambers that trap solar energy. They were used to establish two levels, ambient and warmed. Warming was applied to the target 1 x 1 m area thoughout the growing season.

The experiment ran for 5 years, after which time the authors recorded the total leafy biomass from each birch stem in the 1 x 1 target area. This time was the start of the 2014 growing season.

Their data are available from dryad. The relevant files are inside_productivity.csv and a very clear README_for_inside_productivity.txt file. We won't explain the variables, and just point you to the readme. Figure 1 of the paper also has a nice diagram of the experiment.

Start by taking a quick look at the data file. The variables you may be working with for now are block, fertilization, temp, browsing, and biomass. **Note that this file has a semicolon rather than comma as the delimiter, so make sure you take account of that when importing the data**

```
df <- read_delim("data/inside_productivity.csv",
                 delim = ";", escape_double = FALSE, trim_ws = TRUE)

## Rows: 60 Columns: 8
## -- Column specification ------------------------------------------------------------
## Delimiter: ";"
## chr (4): Block, Fertilization, Temp, Parcelle_id
## dbl (4): Biomass, Browsing, Biomass_beg2009, Biomass_end2009
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message
  .

set_sum_contrasts()
```

```
## setting contr.sum globally: options(contrasts=c('contr.sum', 'contr.poly'))

kable(head(df,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c("HOLD_position","striped"))
```

Biomass	Block	Fertilization	Temp	Browsing	Parcelle_id	Biomass_beg2009	Biomass_end2009
535.30	A	N	H	0	ANH0	41	53
470.52	A	N	H	1	ANH1	34	71
243.51	A	N	H	3	ANH3	20	45
1171.89	A	N	L	0	ANL0	25	61
775.68	A	N	L	1	ANL1	27	56
928.05	A	N	L	3	ANL3	32	64
1273.95	A	S	H	0	ASH0	30	78
851.81	A	S	H	1	ASH1	28	56
1219.47	A	S	H	3	ASH3	22	34
1139.32	A	S	L	0	ASL0	17	65

```
df$Browsing <- as.factor(df$Browsing)
df$Block <- as.factor(df$Block)
df$Temp <- as.factor(df$Temp)
df$Fertilization <-as.factor(df$Fertilization)
```

Outline how you'd assess the combined effects of fertilization, temperature, and browsing on biomass

Make sure you think carefully about scales of experimental units, which factors are between- and which are within-plots, and which are fixed vs random.

Make sure you include initial checking of assumptions

This means telling us what these effects are, and producing a graph (or several)!

In the paper, the authors used initial biomass (Biomass_end2009) as a covariate, to account for initial variation in the number of stems occurring in sub-plots. Does inclusion of initial biomass improve the model fit?

Examine the residuals, and see if you think the response variable needs transformation, and if you do a transform, does the model fit any better?

A simpler model

In the discussion of the paper, there is discussion over the success of the method used to produce Warming. The technique has been used before, but data loggers in place during the experiment suggested no consistent warming was produced.

Would a simpler model omitting Warming lead you to any different conclusions?

Extension activities

Examine other responses

The paper includes other response variables. One you could get practice with is *gr_experiment.csv*. This file has data for 3-6 shoots from each sub-plot, which underwent dendrochronological analysis. The shoots were sectioned and growth increments for each year between 1966-2013 recorded. These increments for each shoot are repeated measures, with the individual shoots as the "subjects".

Think about how you might analyse this data set. If you want, you could try analyzing the full data set. An alternative could be to look at some slightly simpler ways:

- Take just the years of the experiment (2010-2013) and look for trends across this period
- Examine only the final ring (2013)
- Take the average of the ring increments during the experiment.

If you analyze a single ring or an average, you'll then have multiple measurements for each sub-plots. Are these values independent experimental units, or should they be averaged?

More complexity

We've just considered the 4 x 24 m plots in which nutrients were in which nutrients were manipulated, but in the actual experiment, these plots were themselves grouped within 5 caribou fences, each with 12 plots.

You could challenge yourself by thinking about how you'd incorporate this aspect into your model & analysis, including the complications that might arise from adding a second random factor.

Chapter 12

The aim of these exercises is to start getting familiar with mixed models, and in these exercises we'll be focused on models with correlated data.

For each of the examples below, you should follow the sequence we've used previously:

1. What is the biological question?
2. Is the predictor continuous or categorical?
3. Write out the linear model corresponding to this question.
4. What distribution do you expect the response variable to follow?
5. What are the assumptions behind the statistical model you'll fit?
    1. Are those assumptions satisfied?
6. Fit the model
    1. How will you assess whether the model fits well?
    2. Can you detect an effect of the predictor?
    3. How do you measure the effect?
7. What do you conclude (including any cautions)

Things to look out for

In fitting these models, we'll need to make the initial checklist more complex:

- as well as classifying a predictor as continuous or categorical, think clearly about whether it is random or fixed
- think about the structure of the data; are all observations of the response variable independent of each other, or are they linked in some way? If they are linked, how do de account for that linkage in specifying and fitting a statistical model?
- When there are random effects, assumptions can be more extensive. We're usually interested in the fixed effects, and different fixed effects can have different assumptions. This is particularly the case with nested designs.
- When there are correlated effects, we might need to be aware of additional assumptions.

A. Straightforward, with a chance to be complex

Let's start with an example we mentioned briefly in Chapter 12, and work through the data analysis. Allen and Marshall (2014) studied the effects of egg size on larval characteristics of a marine tubeworm. They allocated eggs from a group of large or small female tubeworms to ten jars for each egg size. These eggs were fertilized, and after the embryos hatched, the proportion of larvae that had settled and metamorphosed was recorded for each jar every three days for 15 days. The question here is about maternal effects (using size as a proxy for energy reserves available) on offspring performance (time to settle and initiate metamorphosis from larval form to adult form).

Start by following the steps we've used for other chapters, so you're clear on the questions the details of the design, and the applicable statistical model.

We'll start by reading in the data and taking a quick look at it.

```
# Import data as allen; treat jar and day as character variables
allen <- read_csv(”data/allenmarshalltable1.csv”,
      col_types = cols(jar = col_character(),
      day = col_character()))
allen$jar <- factor(allen$jar)
allen$day <- factor(allen$day, ordered=TRUE)
afex::set_sum_contrasts()

## setting contr.sum globally: options(contrasts=c('contr.sum', 'contr.poly'))

kable(head(allen,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

sizeclass	jar	day	day1	perc
b	1	3	3	0.052
b	2	3	3	0.096
b	3	3	3	0.052
b	4	3	3	0.078
b	5	3	3	0.081
b	6	3	3	0.058
b	7	3	3	0.091
b	8	3	3	0.064
b	9	3	3	0.070
b	10	3	3	0.071

In this data file:

- *sizeclass* describes the mothers, b(ig) or s(mall)
- *jar* is numbered 1-20, but could as easily be a, b, c, etc. or any arbitrary label. We treated it as a character variable just so we don't be fooled into thinking the numbers mean anything quantitative.
- *day* represents sampling times. We've made it categorical, ready for some analyses.
- *day1* represents sampling times, but as a continuous variable
- *perc* is the proportion of larvae settled and undergoing or completed metamorphosis
- *lperc* is log-transformed *perc*. Allen & Marshall used this transformation.

Start by taking a close look at the data and thinking about assumptions

If you have concerns about the raw data, would those concerns be eased if you used log-transformed responses?

Let's start with a straightforward, conventional model

First analyse with day as categorical factor (using the predictor day)

Think of a couple of ways you could look at this predictor

What do you conclude about maternal effects? Illustrate your results graphically

You may (hopefully you did!) have noticed some quite different variances between groups. One way we could deal with this is to allow group-specific variances. For interest, let's use this approach and see if this model fits better

Do your conclusions change?

Extend yourself: treat time as a continuous predictor (use day1)

Think about how you'd specify the model. Describe the random slopes and random intercepts models, and then fit them.

Is the random intercepts approach needed?

Last, if you want to extend yourself a bit, Let's look at accounting for some of the heterogeneity in variances by allowing covariances to differ (we'll use AR1)

Now run as mixed effects random intercept with nlme - use day1 for continuous (not factor) and allow covariances to differ.

Does this approach give us a model that fits better?

B. Organic matter breakdown in streams - a messy example

Kiffer *et al.* (2018) were interested in the nutritional value of leaves from native and introduced trees to "shredders" – invertebrate larvae in streams that help to break down organic material. They focused on one shredder species, the caddisfly *Triplectides gracilis* and its relationship between four native species and the blue gum *Eucalyptus globulus*, an introduced species grown in plantations. They exposed individual caddisflies to one of these species, and recorded their growth; their Figure 1 shows the experimental setup. They measured the tibial length of each larva at weekly intervals, and used a conversion equation to estimate the animal's biomass at this time.

What was their biological question?

What factors are in the design?

Which of these variables would you include:

- ☐ Tree species (S)
- ☐ Plantation type (P)
- ☐ Caddis flies (C)
- ☐ Time (T)

How would you combined them in the linear model?

Look at this list and decide which terms should be included

- ☐ S
- ☐ P(S)
- ☐ C(P(S)
- ☐ C(S)
- ☐ T
- ☐ TS
- ☐ TP(S)
- ☐ TC(P(S))
- ☐ TC(S)

Which terms answer the biological question?

- ☐ S
- ☐ P(S)
- ☐ C(P(S)
- ☐ C(S)
- ☐ T
- ☐ TS
- ☐ TP(S)
- ☐ TC(P(S))

□ TC(S)

What checks would you do before fitting a model to these data?

Now let's get hands-on.

First, the housekeeping - get the data file. The original data are in supplementary file S3, which you can get by following the link above to the original paper. S3 is an xls file, and the experimental data are in the sheet *Larvae growth*. If you want to import it to R, you'll need to skip the first two rows; the variable names are on row 3 and data start on row 4. Name the frame *kiffer*. Alternatively, we have it for you:

```
kiffer <- read.csv("data/kiffer_growth.csv")
kable(head(kiffer,10), booktabs=TRUE) %>%
      kable_styling(latex_options = c("HOLD_position","striped"))
```

subject	species	week1	week2	week3	week4	week5
1	Styrax pohlii	0.04	-0.02	0.02	-0.62	-0.95
2	Styrax pohlii	0.03	0.01	-0.01	-0.05	0.02
3	Styrax pohlii	-0.19	0.00	-0.17	0.18	NA
4	Styrax pohlii	0.30	0.19	0.31	0.19	NA
5	Styrax pohlii	0.17	-0.15	0.94	0.93	NA
6	Styrax pohlii	0.35	0.00	0.12	-0.08	0.18
7	Styrax pohlii	-0.11	0.00	-0.01	0.05	-0.26
8	Styrax pohlii	0.00	0.00	0.20	0.22	0.21
9	Styrax pohlii	-0.70	-0.65	NA	NA	NA
10	Styrax pohlii	0.27	0.10	-0.51	-0.21	0.15

```
# Start by treating week as categorical predictor
kiffer$subject <- factor(kiffer$subject)
kiffer_na <-na.omit(kiffer) #This version omits any case with a missing value in
    one of the columns
```

Is there anything notable about this data set?

If you're not sure, run a "traditional" OLS Repeated Measures analysis.

To make this analysis easier in R, let's start by rearranging the original file with the weekly size measurements as variables to a longer file with one row for each weekly measurement

```
kiffer1 <-  pivot_longer(kiffer,
     cols = starts_with("week"),
     names_to = "weekno",
     names_prefix = "week",
     names_transform = list(weekno = as.integer),
     values_to = "growth",
     values_drop_na = TRUE,
  )
```

Still not sure? Use ezDesign as suggested or ezPrecis.

```
ezPrecis(kiffer)
```

```
## Data frame dimensions: 93 rows, 7 columns

##                type missing values                   min           max
## subject      factor       0     93                     1            97
## species   character       0      5 Eucalyptus globulus Styrax pohlii
## week1    numeric     0    61          −1.22       2.31
## week2    numeric     1    76          −1.65       3.52
## week3    numeric    10    73          −2.03       5.44
## week4    numeric    20    66          −2.23       7.94
## week5    numeric    31    59          −2.32       8.11
```

Suggest two ways of dealing with this problem (or three, if you're really keen!.

Outline any issues with each approach

Assuming that you identified a solution involving dropping subjects and one involving a different way of treating weeks, we'll need to create a couple of other arrangements of the data.

1. A version for running a "traditional" repeated measures OLS. This is a "wide" data file, with measurements at each week as columns. There are missing values, which OLS RM models don't like, so we'll drop any animal without a measurement in every week - *kiffer_na*. We'll then convert it to a long version for running with time as a categorical predictor - *kiffer1_na*

"Traditional" OLS RM

Run the OLS version of the model, with time as a categorical predictor. What other checks would you do at this stage?

Do some exploratory data analysis

What do you conclude from your initial checks?

Here's the split-plot version of this analysis

You might find it helpful to generate linear polynomials

Run the model as a linear mixed effects model

(R script provided). For easy comparison, use the kiffer1_na file, which has the same data.

```
kiffer1_na$weekno<-as.integer(kiffer1_na$weekno)    #weekno was previously a factor
kiffer1_na.mix <- mixed(growth~species*weekno+(weekno|subject), test_intercept=TRUE
    , kiffer1_na)
kiffer1_na.mix
kiffer1.lmer <- lmer(growth~species*weekno+(weekno|subject), REML=TRUE, kiffer1)
```

Get anova results

The advantage of the mixed effects approach is that we can keep all insects in

An important decision is whether to fit a random intercepts model (as we did) or a random intercepts model, which we can do using the mixed model approach.

What do you conclude from your analysis? Would you reach different conclusions from the "traditional" and mixed model approaches?

C (Challenging) - back to psychostimulants and flies

Let's return to the Highfill *et al.* (2019) example from the Chapter 10 exercises. Recall that this was a substantial experiment designed to assess variation in susceptibility to psychosomatic drugs between genetic lines of flies, with the sexes incorporated into the design.

In the earlier discussion of this example, you should have thought about the physical structure of the experiment, in which food with and without the drug in question was offered to small groups of flies, where a group of flies was provided with capillary tubes with one of two solutions. The same group of flies was assessed three times, with feeding/recovery periods in between, to see if drug responses developed through time.

Let’s bring back the data file and look at it again. While we’re at it, we’ll make sure that Lines are treated as factors, rather than continuous predictors:

```
library(readxl)
highfill <− read_excel(”data/pgen.1007834.s001.xlsx”)
highfill <− highfill %>%
  dplyr::rename(Line = ”DGRP␣Line”) %>%
  dplyr::rename(vial = Replicate)
highfill$Line <−as.factor(highfill$Line)
highfill$vial <−as.factor(highfill$vial)
kable(head(highfill,10), booktabs=TRUE) %>%
      kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

Solution	Sex	Line	vial	Consumption	Exposure
Cocaine	F	41	1	34.6	1
Cocaine	F	41	2	23.6	1
Cocaine	F	41	3	22.6	1
Cocaine	F	41	4	38.6	1
Cocaine	F	41	5	11.6	1
Cocaine	F	41	6	37.6	1
Cocaine	F	41	7	54.6	1
Cocaine	F	41	8	13.6	1
Cocaine	F	41	9	59.6	1
Cocaine	F	41	10	81.6	1

The design has vials as an experimental unit, which we might consider to be plots or subjects.

What other possible correlations in the data are there?

The model structure

What are the between-subjects factors?

What are within-subject(s) factors?

Have a crack at specifying the model and fitting it to the data. The worked examples of 12.4 and 12.5 will give you some help, but don’t expect it to be easy :-(

Chapter 13

These exercises will push you a bit to fit GLMs to various kinds of data. You'll need to pay attention to the nature of the predictors and think about the response variables and their likely distributions.

A. Improving biological control of diamondback moths

Uefune *et al.* (2020) studied the use of synthetic herbivory-induced plant volatiles (HIPVs) to attract larval parasitoid wasps (*Plutella xylostella*) to control diamondback moths (DBM: *Cotesia vestalis*), a global pest of cruciferous vegetables. in greenhouses growing mizuna (Japanese mustard) in Japan. They used two groups of greenhouses, the treated group having dispensers for the HIPVs as well as honeyfeeders to attract *C. vestalis*) and a second untreated group. In each greenhouse, a single sticky trap, replaced weekly over 6 months, was used to catch both DBMs and *C. vestalis* and the numbers of both counted. While greenhouse ID could have been included as a random effect in a mixed model analysis, the available data did not record individual greenhouses. We will model numbers of *C. vestalis* against numbers of DBM and treatment using each trap as the units of analysis. The study was done in 2006 and 2008 but we only analyze the 2008 data.

The data are available on Dryad, and you'll want the Excel file Fig3.xls, and within it, the sheet labelled 2008. Before using it in analysis, you'll need to tidy up the first few rows. Alternatively, we provide a simplified version, with three columns, treatment, moth, and parasitoid (uefune.csv).

What distribution would you expect the response variable *parasitoid* to follow?

Import the data and check the properties of the response variable.

```
#Import uefune data file (uefune.csv)
uefune <- read.csv(”data/uefunex.csv”)
kable(head(uefune,10), booktabs=TRUE) %>%
      kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

treatment	moth	parasitoid
treated	0	0
treated	0	0
treated	0	0
treated	0	0
treated	0	0
treated	0	0
treated	0	1
treated	0	0
treated	0	0
treated	0	0

Hint: Try graphical presentation (possibly including a transformation) and look at summary statistics

What do you conclude about the response variable?

Fit a poisson glm and interpret the results

You should have decided that the data are too skewed to fit a linear model assuming a normal distribution, and a poisson is a strong candidate.

Extension question

A look at the means and standard deviations suggests that the variance was much larger than the mean, suggesting some overdispersion. Let's check for it, and see if we should be concerned.

Hint: here's an R code chunk to do that:

presid <- resid(uefune.glm, type="pearson")

ssize <- nrow(uefune1)

params <- length(coef(uefune.glm))

disp <- sum(presid^2)/(ssize-params)

disp

Conclusions?

How does your solution work?

For interest(!). you could also compare the results to the "old" way of transforming response variables closer to normality and fitting an OLS linear model

Try linear with sqrt transform

B. Better coffee-growing for small mammals

Caudill and Rice (2016) examined the effectiveness of methods for making agriculture and habitat preservation more compatible, specifically assessing whether protocols for "Biodiversity-Friendly" coffee growing resulted in positive outcomes for mammals. They compared four habitats, forest, Bird Friendly® shade, conventional shade, and sun coffee, and used a combination of Sherman traps and camera traps to count mammals and assess species richness. Each habitat was represented by multiple sites, with 23 total sites monitored. At each site, they also recorded a series of plant habitat variables, such as cover at canopy, mid- and lower-strata, and ground level, tree basal area. They were interested primarily in differences between habitats, but also the role of vegetation characteristics in influencing mammal diversity. We focus on this latter relationship.

The data were provided as an Excel file through Dryad, which needs a little tidying to use in R; two additional header rows, and some variable names quite long. Use the tidied caudill.csv file for convenience.

Potential explanatory habitat variables are

- % canopy cover,
- % mid-strata vegetation,
- % lower strata vegetation,
- % ground cover,
- tree basal area (m2/ha),
- tree density (number of trees/m2),
- tree richness,
- tree height (m), and
- coffee height (m)

Caudill & Rice examined three response variables, Small Mammal "species density", which was the number of small mammal species recorded, medium-large density, and total species density, which combined small and med-large.

Import the data file and have a look at it.

```
df <- read.csv(”data/caudill.csv”)
kable(head(df,10), booktabs=TRUE) %>%
      kable_styling(latex_options = c(”HOLD_position”,”striped”, ”scale_down”))
```

site	habitat	trapnights	individuals	specdens_small	rrapnights_camera	num_images	specdens_ml	totspecdens	canopy	basalarea	midstrata	lowstrata	groundcov	treecount	elevation	treerichness	treeheight	coffeeheight	X
1	BF	500	3	1	10	8	2	3	90.9	37.9	35.7	30.3	62.2	84	723	10	20.4	2.5	NA
2	BF	500	14	4	10	10	5	9	81.4	20.7	70.3	86.5	2.6	59	711	12	15.1	2.7	NA
3	BF	480	6	3	10	9	3	6	92.7	34.0	58.6	21.4	53.4	80	713	13	11.7	2.8	NA
4	BF	480	13	4	10	24	3	7	77.3	22.9	54.5	67.7	3.2	41	660	9	18.4	2.7	NA
5	BF	480	16	3	10	3	1	4	91.7	22.9	64.5	54.3	65.9	48	641	5	12.9	2.7	NA
6	BF	500	21	5	10	3	2	7	82.3	22.9	35.8	70.6	7.1	35	755	9	19.4	2.4	NA
7	Forest	500	6	2	10	5	2	4	89.0	12.0	76.8	53.8	28.4	171	487	NA	NA	NA	NA
8	Forest	500	16	3	10	5	2	5	96.6	18.4	18.3	65.1	9.4	112	496	NA	NA	NA	NA
9	Forest	340	2	2	10	12	1	3	89.6	38.6	66.7	51.7	24.3	35	650	NA	NA	NA	NA
10	Forest	480	4	1	10	5	3	4	96.0	26.8	44.6	39.2	17.8	203	667	NA	NA	NA	NA

```
#Forest habitat has no coffee trees, and low tree richness, so habitat variables
    not all present. Create subset of this data file
df2<-dplyr::filter(df, habitat != 'Forest')
```

Look at the three response variables. What kind of distribution is most likely for them?

Would you analyse all three response variables? Why?

What checks would you do for the habitat predictor variables, and what do you conclude from those checks?

The authors showed some differences in mammal diversity between habitat types, but for their regression analysis they ignored habitat.

Are there any additional checks you'd do before proceeding?

What is the linear model for species density of small mammals?

Is there any concern about the number of predictors vs number of data points?

Fit this model to the data and assess its fit.

What do you conclude about habitat influences on small mammals?

Now run the analysis for large mammals

C. Why do gnus die?

This exercise expands on the example introduced in Chapter 13, where Sinclair and Arcese (1995) were interested in the causes (predation or other) of death of wildebeeste, more specifically whether predation varied with sex and health of animals. They addressed the question by examining carcasses, and they cross-classified 226 wildebeest carcasses from the Serengeti by three variables: sex (male, female), cause of death (predation, non-predation) and bone marrow type (SWF: solid white fatty; OG: opaque gelatinous; TG: translucent gelatinous; with the first indicating a healthy animal which is not undernourished).

We could look at these data in two ways; as we discussed in the chapter, it is sometimes difficult to identify a clear response variable in situations like this. At other times, one variable is a clear candidate.

The data were extracted from the original paper and used in our first edition. The data file is here.

```
#Import sinclair data file (sinclair.csv)
sinclair <- read.csv(”data/sinclair.csv”)
kable(head(sinclair,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

death	sex	marrow	gnus
Predation	F	SWF	26
Predation	F	OG	32
Predation	F	TG	8
Predation	M	SWF	14
Predation	M	OG	43
Predation	M	TG	10
Other	F	SWF	6
Other	F	OG	26
Other	F	TG	16
Other	M	SWF	7

```
factor(sinclair$sex)   #use these later when running glm
```

```
##  [1] F F F M M M F F F M M M
## Levels: F M
```

```
factor(sinclair$marrow)
```

```
##  [1] SWF OG  TG  SWF OG  TG  SWF OG  TG  SWF OG  TG
## Levels: OG SWF TG
```

```
factor(sinclair$death)
```

```
##  [1] Predation Predation Predation Predation Predation Predation Other
##  [8] Other     Other     Other     Other     Other
## Levels: Other Predation
```

Hint for running analyses

If you’re working with R we’d recommend that you normally just run our small script libraries.R:

In this case, you’ll need to load two other packages, *epitools* and *vcd*

a. The 226 carcasses each could be placed into one of 12 cells of a 2 x 2 x 3 table, and we could model the cell count.

b. We could make the cause of death our focus, recording its death as predation or other, and record values for the other two values.

Write out the linear model you might use in each case

What distribution would you expect to response variable to follow in each case?

Start with situation *a*

What is your first step in fitting the model to these data?

What are your next steps?

What do you conclude about the lives of wildebeeste?

Now do situation *b* - fit logistic model with death as response

There are two kinds of data files that can be used in this case. We could work with a long file in which every gnu carcass is a record, or we could use a summary file, where one column of the file contains a cell count. This second format is what we used for fitting the log-linear model of part *a*, so we'll stay with it.

To run these data with a binary response, we need to do two things:

- Use a *weights* option, so the variable *gnus* in this case indicates that a record like this occurs several times, equal to the value of *gnus* for this record.
- running a glm may require the death variable to be numerical, so we'll define a new variable, *pred* with values 1 if Predation, 0 if Other. We can do that using *ifelse*.

Don't do a full analysis here; just start by looking at the fit of the initial complex model, and see how your results compare to those you obtained in part a

D. Inbreeding effects on parasitism in birds

One of the consequences of inbreeding in animal populations is that it is thought to increase the susceptibility of individuals to disease. To assess this, Townsend *et al.* (2018) did genetic analysis of 178 crows found across California, and assessed their level of homozygosity (homozygosity by locus - HL), a measure of how inbred they were (higher levels of HL = more inbred). They also took blood samples from the same crows to check for the presence of avian malaria (*Plasmodium*), a common disease. They wanted to test whether higher amounts of inbreeding (higher HL) were associated with increased probability of having a *Plasmodium* infection.

The data file is here

```
#Import data file
df <- read.csv(”data/townsend.csv”)
kable(head(df,10), booktabs=TRUE) %>%
      kable_styling(latex_options = c(”HOLD_position”,”striped”))
```

Plasmodium	HL
0	27.2
0	27.0
1	27.2
0	22.6
0	34.6
0	32.4
0	26.0
0	20.5
0	26.2
1	31.8

What kind of response variable do we have here?

Because we are doing a binomial response a simple scatterplot to check assumptions here is a little pointless, but just in case you do want to see it, have a crack

Fit the generalised linear model, Plasmodium = intercept + slope*HL, with a binomial response variable.

Think about how you might check how well the model fits

Examine the results from fitting the GLM and testing the null hypothesis that the slope of the response to predictor equals zero.

Determine and interpret the following:

a. y-intercept
b. slope of the regression
c. z value for main H_0
d. P-value for main H_0
e. Obtain the 95% confidence intervals on the model estimates

From the results above what conclusions would you draw about the relationship between incidence of avian malaria (*Plasmodium*) and the homozygosity level?

We can actually use this model to determine what the probability of a bird having avian malaria is if its homozygosity level is 50%

What is that probability?

Odds

We might want to something a bit more meaningful about the relationship of homozygosity level to incidence of *Plasmodium* and express this in terms of the odds of a bird having avian malaria. Remember that the binomial model uses a logit link function so that it models the response in terms of log odds (remember the odds are the probability of having *Plasmodium* divided by the probability of not having it). To compute how the odds of a bird having *Plasmodium* varies with homozygosity level in R, us *exp(coef(townsend.glm))*. You can also generate 95% confidence intervals on these odds ratios using *exp(confint.default(townsend.glm))*

What do you interpret from these results about the odds of a bird having *Plasmodium* as you increase the homozygosity level?

You might find it helpful to plot the data as well as examining the output

Chapter 14

This chapter introduces multivariate analyses, which are expanded in the following two chapters.

Chapter 15

The aim of these exercises is to get experience with multivariate data sets. In this chapter's, you'll focus on analyses based on dissimilarities/distances.

You'll want the packages *vegan*, *mvabund*, and *DAAG*.

In earlier chapters, we've used the sequence we've used previously as a guide to working through an analysis, but for this chapter and the next, this sequence doesn't work as well, particularly step 3. Keep using the sequence, but skip over steps that aren't appropriate for a particular question.

1. What is the biological question?
2. Is the predictor continuous or categorical?
3. Write out the linear model corresponding to this question.
4. What distribution do you expect the response variable to follow?
5. What are the assumptions behind the statistical model you'll fit?
    1. Are those assumptions satisfied?
6. Fit the model
    1. How will you assess whether the model fits well?
    2. Can you detect an effect of the predictor?
    3. How do you measure the effect?
7. What do you conclude (including any cautions)

A Unconstrained multivariate analysis: PCA and Principal Components Regression

Posthumus *et al.* (2015) were interested in the ecological role of red squirrels in eastern North America. These territorial animals hoard pine cones for food, and create large middens - piles of plant debris - when they return to eat these pine cones, etc (see their Figure 1). These middens are potentially important habitat changes for other vertebrates of these forests. Postumus and colleagues focused on two aspects of this relationship:

How do middens influence birds and other mammals? What habitat conditions influence the use of a site by squirrels?

We'll look at this second aspect of their study - the relationship between habitat variables at each midden location and squirrel residency. They measured six habitat variables: 1. logvol: volume of downed logs > 20 cm diameter 2. cancov: canopy cover 3. basal area of large trees 4. slope (°) 5. aspect (°) 6. lsnags: number of large snags > 40 cm dbh 7. shannon: tree diversity 8. ltrees: number of large trees > 40 cm dbh

The data are available as Supplementary file S1. It's an Excel file, from which you will need the first sheet (Habitat and Squirrel Features), and when you read it in, skip the first 3 rows. The package *readxl* makes this import easy. You might also want to rename the variables to something a bit more tractable. As usual, we've a tidied version of the file that can be used.

```
posthumus <- read_csv("data/posthumus.csv")
```

```
## Rows: 100 Columns: 11
## —— Column specification ————————————————————————————————————————
## Delimiter: ","
## dbl (11): cancov, shannon, logvol, lsnags, ltrees, basarea, slope, aspect, r...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message
    .
```

```
kable(head(posthumus,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c("HOLD_position","striped", "scale_down"))
```

cancov	shannon	logvol	lsnags	ltrees	basarea	slope	aspect	rspres	coneind	resprop
78.468	0.623	101.76	21.93	131.58	455417.7	-10	285	0	0	0.28571
56.716	0.708	117.46	21.93	87.72	305798.3	-8	240	0	0	0.00000
83.876	0.381	31.10	0.00	87.72	528041.9	-22	5	0	0	0.66667
78.208	1.100	44.34	21.93	175.44	636344.8	-5	180	1	4	0.90476
58.500	0.892	153.97	0.00	21.93	319287.6	-14	250	0	0	0.09524
71.708	0.861	101.45	0.00	43.86	520032.0	-12	50	0	0	0.04762
65.104	0.699	124.52	21.93	109.65	584454.9	-10	220	0	0	0.76190
62.348	0.814	173.77	0.00	65.79	440843.2	-4	282	1	1	0.57143
65.832	1.017	74.45	21.93	43.86	628072.8	-8	300	0	0	0.00000
73.164	0.248	92.94	65.79	131.58	928825.2	-14	274	0	0	0.76190

Start by looking at a scatterplot matrix to see if there any issues.

Let's transform the habitat variables, as was done in the original paper, and look at the scatterplot matrix again. Canopy cover was squared, and the other 5 were log-transformed. Note that the slope values in the original data file are negative, so we reversed the sign before transforming. Large snags has zero values, so we'll add 1 to these values (1 being the smallest possible value)

```
# transform as in original paper

posthumus <−posthumus %>%
  dplyr::mutate(
    llogvol = log10(logvol),
    llsnags = log10(lsnags+1),
    lltrees = log10(ltrees),
    lbasarea = log10(basarea),
    lslope = log10(−1∗slope),
    scancov = (cancov^2)
    )
```

Relook at correlations

Use a PCA to simplify the habitat variables

Would you use a correlation or a covariance matrix

Does the PCA help with simplifying the habitat variables?

Summarize the first three PCs:

How much variation do they explain?

How do the original variables contribute to these components?

Does this matter for squirrels?

While recording habitat variables, each location was also surveyed 21 times, and the research team recorded the number of times Red Squirrels were present. With standardised surveying effort, we could look at the number of surveys with squirrels, but they also calculated a proportion of surveys with squirrels present (resprop).

Now that you have some nice independent predictors (we hope ;-)), you can explore whether squirrel presence can be predicted using them. This should be quick, as you know the predictors are independent!

Hint: If you saved your PCA analysis into a file, you can easily access the scores. If you coerce them into a dataframe, using as.data.frame(posthumus.pca$scores), you're good to go.

B. Constrained multivariate analysis (RDA) - amphibians in urban ponds

Hutto and Barrett (2021) were concerned with effects of urbanization and the potential for urban open spaces to mediate those effects. They focused on amphibians as a group of concern, and surveyed 51 ponds in South Carolina, where they recorded a suite of habitat variables relevant to amphibians. The sites were classified into three groups, highly urbanized, low urbanization, and urban open spaces.

At each pond, they also characterized the frog and toad assemblage, using a mixture of dip-net surveys, calls, and deployment of retreats for tree frogs.

The data are in three supplementary files, linked through the paper. You'll be using huttoenv.csv and huttoamph.

```
huttoenv <- read_csv(”data/huttoenv.csv”)

## Rows: 51 Columns: 11
## -- Column specification --------------------------------------------------------
## Delimiter: ”,”
## chr  (1): type
## dbl (10): site, group, ph, cond, wdepth, omdepth, cancov, area, rivernear, w...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message
   .

kable(head(huttoenv,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c(”HOLD_position”,”striped”, ”scale_down”))
```

site	type	group	ph	cond	wdepth	omdepth	cancov	area	rivernear	wetlandnear
1	Low	1	6.54	35.17	1.20	9.49	78	0.29	0.00	357.55
2	High	2	6.32	26.53	1.20	2.50	1	0.22	307.11	55.72
3	Low	1	7.04	157.43	0.33	4.55	1	0.24	21.12	43.91
4	High	2	7.80	36.23	1.20	2.82	1	0.14	449.35	293.66
5	Open Space	3	7.18	47.73	1.20	3.28	0	0.06	21.23	7.97
6	Open Space	3	6.42	35.93	1.20	4.17	0	0.04	26.04	7.97
7	Open Space	3	6.84	101.07	1.20	2.42	6	1.23	0.00	101.83
8	Open Space	3	6.63	49.27	1.20	8.27	19	0.18	0.00	8.90
9	Open Space	3	6.89	43.33	1.20	0.92	9	0.27	0.00	8.90
10	Open Space	3	6.24	44.43	1.20	5.62	33	0.10	1.58	295.02

The variables of interest for our analysis are:

- *type*: Type of area in which the pond occurred (3 categories)
- *area*: wetland area (ha)
- *ph*: pH

- *cond*: Conductivity (µS/m)
- *wdepth*: wetland depth (m)
- *omdepth*: mean depth of the organic layer (cm)
- *rivernear*: linear distance to nearest river (m)
- *wetlandnear*: linear distance to wetland

Now have a quick look at the anuran data.

```
huttoamph <- read_csv(”data/huttoamph.csv”)
```

```
## Rows: 50 Columns: 14
## -- Column specification ------------------------------------------------
## Delimiter: ”,”
## chr  (1): type
## dbl (13): Site, ACRE, BAME, BFOW, GCAR, HCIN, HVER, LCAT, LCLA, LPAL, LSPH, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message
    .
```

```
kable(head(huttoamph,10), booktabs=TRUE) %>%
    kable_styling(latex_options = c(”HOLD_position”,”striped”, ”scale_down”))
```

Site	type	ACRE	BAME	BFOW	GCAR	HCIN	HVER	LCAT	LCLA	LPAL	LSPH	PCRU	PFER
1	Low	9	0	5	0	0	0	1	4	0	1	10	0
2	High	0	0	0	0	4	2	0	4	0	2	5	0
3	Low	0	0	0	3	9	6	2	2	0	1	10	0
4	High	0	0	0	0	0	0	8	0	0	2	1	0
5	Open Space	0	0	0	0	0	0	0	1	0	0	0	0
6	Open Space	1	0	2	0	0	0	4	0	0	1	0	0
7	Open Space	3	0	5	0	3	0	1	2	0	1	1	0
8	Open Space	3	0	0	0	1	0	1	2	0	1	1	0
9	Open Space	0	0	0	0	0	0	0	0	0	0	0	0
10	Open Space	0	0	0	0	0	0	3	6	0	2	1	0

The frog and toad data were essentially standardized by the authors when they combined the different survey methods, with each species receiving an abundance score between 0 and 10.

The anurans were:

- *Acris crepitans* (ACRE)
- *Anaxyrus fowleri* (BFOW)
- *Anaxyrus americana* (BAME)
- *Gastrophryne carolinensis* (GCAR)
- *Dryophytes cinerea* (HCIN)
- *D. versicolor* (HVER)
- *Lithobates clamitans* (LCLA)
- *L. palustris* (LPAL)
- *L. sphenocephalus* (LSPH)
- *L. catesbeianus* (LCAT)
- *Pseudacris feriarum* (PFER)
- *P. crucifer* (PCRU)

They ranged from species present around nearly every pond (*L. clamitans*) to those only present in a few ponds (*P. feriarum*).

Is the frog assemblage related to urbanization?

Start by looking for a relationship between the set of frogs at a pond and the urbanization. This will involve two aspects - taking the data for the 10 anuran species and looking to see if it can be simplified into groupings, and asking whether these groupings differ among pond types.

While we won't do it here, when we group species, we'd be interested in whether the species that contribute to one of the derived variables are a biologically or ecologically sensible set.

You'll be looking at the frog-urbanization link using Redundancy Analysis (see Section 15.4.1).

Use pond type as the predictor, and the 10 frog species as the variables

Look at the frog data.

Do you need to do any standardization before analysis?

Run the analysis, and include a visualization (e.g. a triplot)

Does the kind of urbanization have a strong effect on the frog assemblage?

Are there easily distinguishable frog groupings?

Hint: The visualizations might be a bit hard to see; if that's the case, try a simple PCA on the frogs (i.e. omitting urbanization type)

Try using the environmental variables to constrain the ordination, rather than the habitat categories. Does a clearer pattern emerge?

Think about the distributions of the habitat variables:

Are there correlations between them?

Do any variables need transformation before PCA

Should you standardise the data?

Now do the RDA, once you're happy with the data

Important hint: *One pond has no anuran data, but has environmental data. You'll need to exclude pond 17 from the huttoenv data at some stage*

C. LDA: Using biometric data to distinguish bird species

Militão *et al.* (2014) were interested in how easy it was to distinguish between two closely-related shearwaters, the Yelkouan and the Balearctic. Genetic analyses suggested that these birds diverged relatively recently, and can be hard to identify visually, and they overlap in parts of their ranges in the western Mediterranean. Both are caught as by-catch in commercial longline fisheries, but they differ in conservation status, with the Balearctic being critically endangered and the Yelkouan vulnerable. This difference in conservation status makes it important to identify by-catch, particularly as they may differ in vulnerability to fishing.

Militao and colleagues used several approaches, from biometric measurements to Stable Isotope Analysis to plumage colour to bill morphology, to identify a simple method for identifying bycatch, without having to resort to costly and time-consuming genetic analysis. They applied these approaches to a sample of birds of known species identity.

We'll focus on the biometric data, because this information can be recorded easily while on a fishing vessel, and requires little training. The question of interest is:

Can we use biometric measurements to reliably classify a bird to species?

Their data are in supplement S1 of the paper. They're an Excel file, and you'll need to skip the first 6 rows to get to the variable names in row 7. If you've copied the file into your downloads folder, the next line of code will read in this data frame:

militao <- read_excel("~/Downloads/pone.0115650.s003.xlsx", skip = 6)

We've used a tidied up version (militao.csv)

The variables of relevance are:

- *sex*
- *spp*: species
- *sppsex*: a composite variable combining sex and species to a single variable with four values
- *year*: when bird was collected (not used)
- *bdb*: bill depth at base
- *bdn*: bill depth at nostril
- *bl*: bill length
- *mhl*: maximum head length
- *tl*: tarsus length
- *wl*: wing length

```
militao <- read_csv("data/militao.csv")

## Rows: 194 Columns: 12
## -- Column specification ------------------------------------------------
## Delimiter: ","
## chr (4): dataset, sex, spp, sppsex
## dbl (8): birdid, year, bdb, bdn, bl, mhl, tl, wl
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message
   .

kable(head(militao,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c("HOLD_position","striped", "scale_down"))
```

dataset	birdid	sex	spp	sppsex	year	bdb	bdn	bl	mhl	tl	wl
a	1	m	ys	ysm	2002	10.48	7.55	36.37	85.16	45.98	237.0
a	20	f	bs	bsf	2002	11.54	8.71	38.21	89.42	51.04	251.0
a	21	f	bs	bsf	2002	11.69	8.74	38.02	88.60	48.20	253.0
a	23	f	bs	bsf	2002	11.45	8.40	38.90	90.02	50.88	263.0
a	138	f	bs	bsf	2004	11.29	8.02	38.22	87.82	49.24	250.0
a	140	m	bs	bsm	2004	12.70	9.22	38.99	91.75	51.66	261.0
a	141	f	bs	bsf	2004	10.78	7.95	37.70	88.43	49.89	253.0
a	142	m	u	um	2004	11.42	8.59	39.03	91.46	49.09	241.5
a	143	f	bs	bsf	2004	11.96	9.10	35.68	86.50	47.77	257.0
a	214	f	ys	ysf	2005	10.29	6.77	35.16	82.05	46.59	225.0

Analysis in the paper uses data from this file, but excludes birds that could not be assigned to species (labelled unknown), and for the LDA, uses a subset of the data ("a").

Militao et al. also standardised their biometric variables - "translation" - to remove sexual differences - subtracted sex-species mean for each variable.

```
militao1 <- filter(militao, dataset=='a' & spp %in% c('ys','bs'))
#Centre the morphological variables
militao1 <- militao1 %>%
  dplyr::group_by(sppsex) %>%
  mutate(
    bdbm = mean(bdb),
    bdnm = mean(bdn),
    blm = mean(bl),
    mhlm = mean(mhl),
    tlm = mean(tl),
    wlm = mean(wl)
    ) %>%
  ungroup() %>%
  mutate(
    bdbc = bdb - bdbm,
    bdnc = bdn - bdnm,
    blc = bl - blm,
    mhlc = mhl - mhlm,
    tlc = tl - tlm,
    wlc = wl - wlm
  )
kable(head(militao1,10), booktabs=TRUE) %>%
      kable_styling(latex_options = c(”HOLD_position”,”striped”, ”scale_down”))
```

dataset	birdid	sex	spp	sppsex	year	bdb	bdn	bl	mhl	tl	wl	bdbm	bdnm	blm	mhlm	tlm	wlm	bdbc	bdnc	blc	mhlc	tlc	wlc
a	1	m	ys	ysm	2002	10.48	7.55	36.37	85.16	45.98	237	11.48372	8.335133	37.86885	87.71929	48.86858	246.1106	-1.0037168	-0.7851327	-1.4988496	-2.559292	-2.8885841	-9.11062
a	20	f	bs	bsf	2002	11.54	8.71	38.21	89.42	51.04	251	11.48372	8.335133	37.86885	87.71929	48.86858	246.1106	0.0562832	0.3748673	0.3411504	1.700708	2.1714159	4.88938
a	21	f	bs	bsf	2002	11.69	8.74	38.02	88.60	48.20	253	11.48372	8.335133	37.86885	87.71929	48.86858	246.1106	0.2062832	0.4048673	0.1511504	0.880708	-0.6685841	6.88938
a	23	f	bs	bsf	2002	11.45	8.40	38.90	90.02	50.88	263	11.48372	8.335133	37.86885	87.71929	48.86858	246.1106	-0.0337168	0.0648673	1.0311504	2.300708	2.0114159	16.88938
a	138	f	bs	bsf	2004	11.29	8.02	38.22	87.82	49.24	250	11.48372	8.335133	37.86885	87.71929	48.86858	246.1106	-0.1937168	-0.3151327	0.3511504	0.100708	0.3714159	3.88938
a	140	m	bs	bsm	2004	12.70	9.22	38.99	91.75	51.66	261	11.48372	8.335133	37.86885	87.71929	48.86858	246.1106	1.2162832	0.8848673	1.1211504	4.030708	2.7914159	14.88938
a	141	f	bs	bsf	2004	10.78	7.95	37.70	88.43	49.89	253	11.48372	8.335133	37.86885	87.71929	48.86858	246.1106	-0.7037168	-0.3851327	-0.1688496	0.710708	1.0214159	6.88938
a	143	f	bs	bsf	2004	11.96	9.10	35.68	86.50	47.77	257	11.48372	8.335133	37.86885	87.71929	48.86858	246.1106	0.4762832	0.7648673	-2.1888496	-1.219292	-1.0985841	10.88938
a	214	f	ys	ysf	2005	10.29	6.77	35.16	82.05	46.59	225	11.48372	8.335133	37.86885	87.71929	48.86858	246.1106	-1.1937168	-1.5651327	-2.7088496	-5.669292	-2.2785841	-21.11062
a	215	f	ys	ysf	2005	10.83	7.26	35.43	83.96	49.53	235	11.48372	8.335133	37.86885	87.71929	48.86858	246.1106	-0.6537168	-1.0751327	-2.4388496	-3.759292	0.6614159	-11.11062

Use discriminant analysis (LDA) to decide whether these morphological measurements are enough to distinguish these shearwater species

Sensitivity

The paper’s authors chose to adjust for bird sex by subtracting the mean value for that bird species/sex from each measurement, before running their LDA.

Did that decision improve the discriminatory capacity, i.e. what would have happened if raw data were used?

Can you think of another way of adjusting for sex of bird?

Chapter 16

The aim of these exercises is to continue making sure you're comfortable with handling multivariate data. In this chapter's, you'll focus on analyses based on dissimilarities/distances, including fitting linear models to these kinds of response variables.

You'll need the packages *vegan* and *mvabund*.

For each of the examples below, you should follow the sequence we've used previously, as far as it's sensible:

1. What is the biological question?
2. Is the predictor continuous or categorical?
3. Write out the linear model corresponding to this question.
4. What distribution do you expect the response variable to follow?
5. What are the assumptions behind the statistical model you'll fit?
    1. Are those assumptions satisfied?
6. Fit the model
    1. How will you assess whether the model fits well?
    2. Can you detect an effect of the predictors?
    3. How do you measure the effect?
7. What do you conclude (including any cautions)

A. Does fire history affect reptile assemblages?

Dixon *et al.* (2018) focused on assemblages of reptiles in forests and woodlands of southeastern Australia, with a particular interest in relationships between these assemblages and the fire history of the landscape. They identified 81 sites that varied in time since fire, from 6 months to >96 y. Rather than a continuum, three categories of time since fire were used, 0.5-2y, 6-12y, and >96y. Sites were also classified according to habitat.

Reptile assemblages were sampled with a range of methods, including visual surveys and camera traps, to give counts of 20 reptiles, whose abundance ranged from 0-1 to 0-126, depending on species.

Data are available from dryad, as Rep_abund.csv. You'll want to focus on the columns with reptile numbers, and the one with the fire history (tsf). For convenience, we've extracted those data as dixonbiota.csv

Start by having a quick look at the data.

```
df <- read_csv(”data/dixonbiota.csv”)
```

```
## Rows: 79 Columns: 22
## -- Column specification --------------------------------------------------------
## Delimiter: ”,”
## chr  (2): site, tsf
## dbl (20): adup, aplat, amur, amac, aram, dcor, ecun, esax, eul, htal, ldel, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message
    .
```

```
kable(head(df,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c("HOLD_position","striped", "scale_down"))
```

site	tsf	adup	aplat	amur	amac	aram	dcor	ecun	esax	eul	htal	ldel	lgui	lwhi	ppor	pent	pspe	ptex	rdie	tnig	vros
A01	long	1	0	4	2	0	1	0	0	0	1	1	13	2	0	13	1	0	0	0	0
A02	long	0	0	4	2	0	0	0	0	1	1	0	9	0	0	14	0	0	0	1	0
A04	long	0	3	0	1	0	1	0	0	0	0	0	34	8	0	10	1	0	0	0	0
A05	long	2	0	0	1	0	0	2	0	1	0	2	32	3	0	20	1	1	0	0	0
A06	long	0	1	1	4	1	0	0	1	0	0	5	18	1	0	102	38	0	0	0	0
A07	long	0	0	0	1	0	1	0	0	0	1	2	15	0	0	76	60	0	0	0	0
A08	long	3	1	0	1	0	0	2	0	1	0	2	34	1	0	38	2	1	0	1	0
A09	long	1	1	1	2	0	0	0	0	1	0	3	8	1	0	126	64	0	0	1	1
A10	long	0	2	0	0	0	0	1	2	2	0	0	8	0	0	113	33	0	0	0	0
B01	short	0	2	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	1	0

The file is pretty straightforward, with tsf being the fire category, and columns 3-22 each representing one reptile taxon.

Outline how you'd assess whether the reptile assemblage (the combination of species present and their abundances) is related to fire history, using a dissimilarity-based approach.

You should think about how you'll deal with the data:

- Will you need a transformation, standardization, etc.?
- What are the consequences of any decisions you make for the interpretation of your analysis?
- What measure(s) of dissimilarity are appropriate?

Run the analysis and provide your interpretation

How would you fit a linear model to assess fire effects?

Suppose you think of a reptile assemblage as simply the species that are present, regardless of how common they are.

What would you conclude about the infuence of time since fire now?

Extension activity

Dixon and their colleagues also recorded a range of habitat variables, which we've collected for you in dixonenv.csv

```
dixonenv <- read_csv("data/dixonenv.csv")
```

```
## Rows: 79 Columns: 14
## -- Column specification ------------------------------------------------
## Delimiter: ","
## chr  (4): site, tsf, veg, aspect
## dbl (10): lcwd, shrcov, grcov, litcov, litdep, rocks, elev, warm, cold, twi
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message
    .
```

```
kable(head(dixonenv,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c("HOLD_position","striped", "scale_down"))
```

site	tsf	lcwd	veg	shrcov	grcov	litcov	litdep	rocks	aspect	elev	warm	cold	twi
A01	long	9.720226	SAW	0.077	0.650	0.350	9.9	4.68	N	1050.73	24.97	-2.39	-3.031
A02	long	11.236421	SAW	0.125	0.630	0.465	13.4	0.00	S	969.96	25.65	-2.18	-2.597
A04	long	11.123639	SAW	0.012	0.770	0.320	10.1	1.14	N	1072.86	24.78	-2.49	-2.746
A05	long	10.786490	SAW	0.059	0.345	0.636	9.2	18.90	N	1299.38	22.65	-3.05	-2.489
A06	long	11.388167	DS	0.084	0.685	0.585	23.9	24.34	W	1504.49	20.86	-3.43	-2.414
A07	long	11.786024	DS	0.144	0.900	0.665	22.1	7.66	E	1531.65	20.63	-3.51	-2.859
A08	long	11.950522	DS	0.040	0.725	0.410	10.8	0.00	W	1201.14	23.59	-2.84	-2.103
A09	long	12.131139	DS	0.079	0.740	0.465	17.1	2.86	E	1467.57	21.16	-3.39	-2.376
A10	long	11.502127	DS	0.057	0.575	0.635	21.2	26.14	S	1449.01	21.31	-3.36	-2.252
B01	short	9.227198	DS	0.310	0.175	0.375	8.2	1.37	W	738.71	27.32	0.26	-3.268

They recorded Coarse Woody Debris, which was long-transformed, litter cover (*litcov*), % cover of groundcover (*grcov*), and % cover of shrubs (*shrcov*), and rock cover. In the original paper, Dixon et al. were comfortable that these predictors weren't correlated.

How is reptile assemblage related to time since fire and these five habitat variables

MVABUND

The previous analyses showed differences in dispersion between tsf groups which may affect our permanova interpretations. An alternative is to use mvabund based on a poisson or negative binomial distribution.

B. Simple MDS & Permanova: Frogs and urbanization

Let's return to the Hutto and Barrett (2021) example from the Chapter 15 exercises. There, we used analyses based on associations between variables to explore the relationship between frog assemblages (RDA) in ponds and the degree of urbanization surrounding. This seems a question that could just as easily be examined using distances or dissimilarities.

Return to the data and assess whether frog assemblages (using the standardized abundance scale or presence-absence) differ with urbanization.

Did your conclusions differ from those obtained using RDA?

To answer this question, you'll need to exclude a couple of ponds that had no frogs.

```
df <- read_csv("data/huttoamph.csv")
```

```
## Rows: 50 Columns: 14
## -- Column specification ------------------------------------------------
## Delimiter: ","
## chr  (1): type
## dbl (13): Site, ACRE, BAME, BFOW, GCAR, HCIN, HVER, LCAT, LCLA, LPAL, LSPH, ...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message
   .
```

```
kable(head(df,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c("HOLD_position","striped", "scale_down"))
```

Site	type	ACRE	BAME	BFOW	GCAR	HCIN	HVER	LCAT	LCLA	LPAL	LSPH	PCRU	PFER
1	Low	9	0	5	0	0	0	1	4	0	1	10	0
2	High	0	0	0	0	4	2	0	4	0	2	5	0
3	Low	0	0	0	3	9	6	2	2	0	1	10	0
4	High	0	0	0	0	0	0	8	0	0	2	1	0
5	Open Space	0	0	0	0	0	0	0	1	0	0	0	0
6	Open Space	1	0	2	0	0	0	4	0	0	1	0	0
7	Open Space	3	0	5	0	3	0	1	2	0	1	1	0
8	Open Space	3	0	0	0	1	0	1	2	0	1	1	0
9	Open Space	0	0	0	0	0	0	0	0	0	0	0	0
10	Open Space	0	0	0	0	0	0	3	6	0	2	1	0

```
#ponds 9 and 40 had no frogs. Remove for distance analysis
df <- df %>%
  filter(Site!= 9 & Site !=40)
kable(head(df,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c(”HOLD_position”,”striped”, ”scale_down”))
```

Site	type	ACRE	BAME	BFOW	GCAR	HCIN	HVER	LCAT	LCLA	LPAL	LSPH	PCRU	PFER
1	Low	9	0	5	0	0	0	1	4	0	1	10	0
2	High	0	0	0	0	4	2	0	4	0	2	5	0
3	Low	0	0	0	3	9	6	2	2	0	1	10	0
4	High	0	0	0	0	0	0	8	0	0	2	1	0
5	Open Space	0	0	0	0	0	0	0	1	0	0	0	0
6	Open Space	1	0	2	0	0	0	4	0	0	1	0	0
7	Open Space	3	0	5	0	3	0	1	2	0	1	1	0
8	Open Space	3	0	0	0	1	0	1	2	0	1	1	0
10	Open Space	0	0	0	0	0	0	3	6	0	2	1	0
11	Open Space	0	1	8	0	0	0	5	1	0	2	0	0

The explanation of the variables is in the previous chapter’s exercises.

C. Human oral microbiomes and HIV

Griffen *et al.* (2019) examined the oral microbiome of humans, with a focus on the effects of HIV and AntiRetroviral therapy (ART) on this microbiome. They described the microbiome of 341 patients, who fell into one of two categories: HIV^{-}, and HIV+ with ART. They also matched their samples as far as possible for sex, along with other conditions that can influence oral microbiomes (*Candida* infection, current smoking). We’ll use their records for these other categories as well. There were more HIV+ than - subjects, but within these groups, approximately equal numbers of two sexes, two *Candida* infection status, and two smoking categories.

The bacteriome was recorded separately using 16S RNA sequencing, which overed just over 600 taxa.

The data are available from their paper, as two Excel files in the supplementary information. The metadata gives HIV status, and a raft of demographic information. We’ll just work with HIV, sex, *Candida*, and current smoking. A second sheet has the bacteriome. The code chunk below reads this file (from a Downloads folder - you’ll need to modify the file location). It also screens the file for any bacterial taxa that were not present in this particular study, which reduces the bacterial taxa to 599.

```
library(readxl)
#Read in metadata. Lots of information we don't need, so a couple of iterations to
   get down to a few columns we want - HIV status, candida, current smoking, gender
df <- read_excel(”data/41598_2019_55703_MOESM3_ESM.xlsx”,
      range = ”A1:S342”, na = ”NA”)
df<-dplyr::select(df, HIV, candida, gender, smoking_current)
```

```
kable(head(df,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c("HOLD_position","striped"))
```

HIV	candida	gender	smoking_current
Yes	No	Female	No
Yes	Yes	Male	No
Yes	No	Female	No
Yes	No	Male	No
Yes	No	Male	No
Yes	No	Male	Yes
Yes	Yes	Female	Yes
Yes	No	Female	No
Yes	Yes	Male	Yes
Yes	Yes	Male	Yes

```
#df1 is the bacterial data
df1 <- read_excel("data/41598_2019_55703_MOESM2_ESM.xlsx")

## New names:
## * `` -> `...1`

kable(head(df1,10), booktabs=TRUE) %>%
     kable_styling(latex_options = c("HOLD_position","striped", "scale_down"))
```

[illegible]

```
df1 <- df1 %>%
  dplyr::select(where( ~ is.numeric(.x) && sum(.x) != 0))  #Drops any bacterial
     taxon not recorded in this study (i.e. where it's all zeroes)
df1 <- df1[,-1] #remove col1, which is sample ID
```

Use a dissimilarity based approach to assess the combined effects of HIV-ART, sex, *Candida* infection, and current smoking on the bacteriome

Do these factors act independently on the bacteriome?

Which effects are largest (use some graphical methods)?

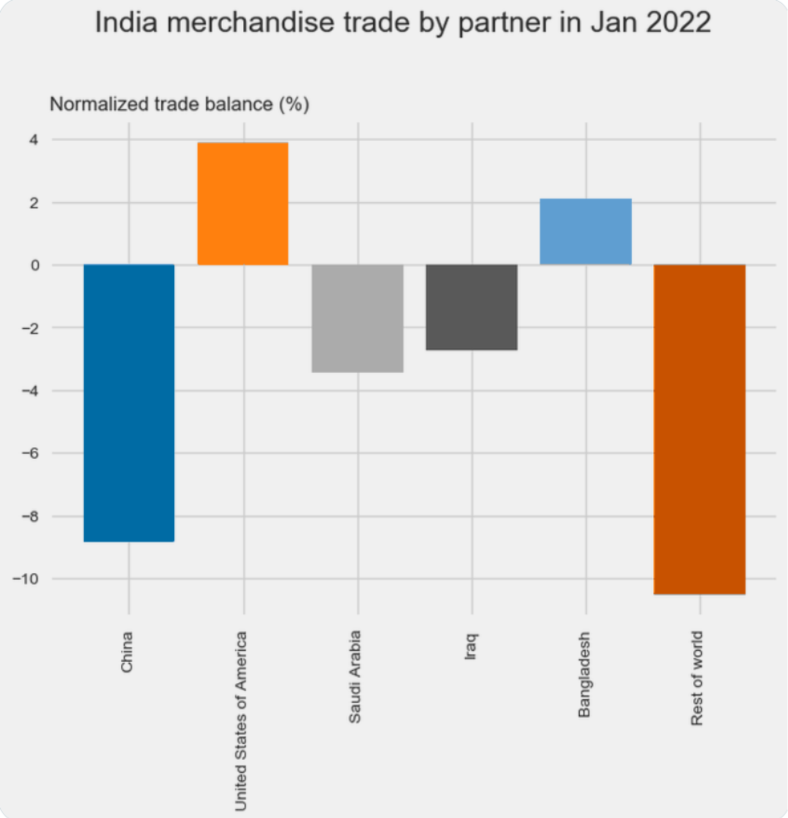
First steps: Think about standardization and which distance measure you'll use. Bray-Curtis seems common for these kind of data, and counts of different taxa vary widely.

Could you fit a simpler model to the data?

Chapter 17

Exercise 1. Improving the reader's experience

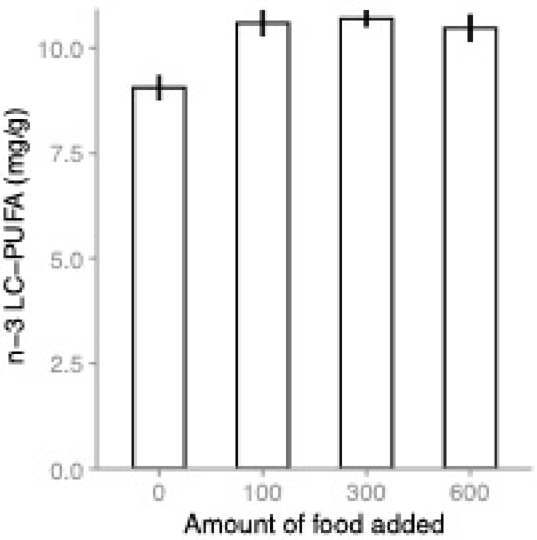
Consider this graph that was produced to show Indian trade balances with prominent partners (link to tweet)



Identify at least two things that could be changed to make this graph more effective.

Exercise 2. Graphics for talks and theses/papers

This exercise is to think about effective figures for different purposes. Let's return to Figure 6.5, which shows means (and SEs) for content of salmon with different feeding regimes (see details in Box 6.4):



... Your task is to produce two versions of this figure. Think about information that would be on and around the figure, and the need for titles and captions.

1. What would you need to do for that figure to appear in a thesis or a manuscript?
2. How would you design the figure for a conference talk or seminar?

Exercise 2. Reporting interactions

This exercise focuses on a marine ecological experiment reported by Brothers and Blakeslee (2021) looking at the interplay between parasitism and habitat structure in affecting survival of flatback mud crabs, *Eurypanopeus depressus*, in the Gulf of Mexico.

The parasite was an introduced barnacle, *Loxothylacus panopaei*, which alters crab behaviour and alter risks of predation. The predators are other crabs, another mud crab, *Panopeus herbstii* and the stone crab *Menippe mercenaria.*

The research question of interest was whether the parasite effect depends on habitat complexity.

The question was addressed with a laboratory experiment. The important details for you

- Mud crabs were infected or uninfected with the parasite
- They lived in simple (gravel) or complex (oyster shell) habitats
- Experimental units were large tanks (“mesocosms”), each with 10 crabs
- 3 replicate tanks for each combination of parasitism and complexity
- Number of surviving crabs recorded daily for 5 days

The task:

Produce a graph or panel of graphs showing what happened, focusing on the research question. You can use the data file provided (crab example for graphing.csv); the file shows mean and s.e. of the proportion of crabs surviving at a given time. Feel free to use all the data or one particular time, e.g. 24 or 48h.

```
crab <- read.csv(”data/crab␣example␣for␣graphing.csv”)
crab

##   Complexity  Parsitism Time Proportion.alive std.error
## 1     Simple Uninfected    0             1.00      0.00
## 2     Simple Uninfected   24             0.50      0.06
## 3     Simple Uninfected   48             0.27      0.13
## 4    Simple Uninfected   72          0.20      0.10
## 5    Simple Uninfected   96          0.07      0.07
## 6    Simple Uninfected  120          0.07      0.07
## 7    Simple   Infected    0          1.00      0.00
## 8    Simple   Infected   24          0.04      0.04
## 9    Simple   Infected   48          0.00      0.00
## 10   Simple   Infected   72          0.00      0.00
## 11   Simple   Infected   96          0.00      0.00
## 12   Simple   Infected  120          0.00      0.00
## 13  Complex Uninfected    0          1.00      0.00
## 14  Complex Uninfected   24          0.44      0.17
## 15  Complex Uninfected   48          0.17      0.03
## 16  Complex Uninfected   72          0.17      0.07
## 17  Complex Uninfected   96          0.08      0.04
## 18  Complex Uninfected  120          0.04      0.03
## 19  Complex   Infected    0          1.00      0.00
## 20  Complex   Infected   24          0.38      0.08
## 21  Complex   Infected   48          0.21      0.05
## 22  Complex   Infected   72          0.07      0.06
## 23  Complex   Infected   96          0.04      0.03
## 24  Complex   Infected  120          0.03      0.03
```

- **Hand drawn or using software – you'll need a single page to share with the class**

Extension question

If the research question was on the value of habitat complexity and whether the parasite changes this value, how would you draw the graph differently?

Brothers, C. A. & Blakeslee, A. M. H. (2021). Alien vs predator play hide and seek: How habitat complexity alters parasite mediated host survival. *Journal of Experimental Marine Biology and Ecology*, 535.

Group activities on specifying linear models

Question 1 (Early on)

Giri et al. (2016) were interested in the capacity of Atlantic salmon (*Salmo salar*), grown in an aquaculture situation, to deliver high levels of omega-3 long-chain polyunsaturated fatty acids (PUFA) in their tissues. Specifically, they tested whether several micronutrients (iron, zinc, magnesium) and coenzymes (riboflavin, biotin and niacin) could increase the conversion from short- to long-chain PUFA. They used four treatments, a diet lacking in these micronutrients and coenzymes (T-0), a diet with normal levels (T-100), and two levels of fortification, where the enzymes and micronutrients were 300% and 600% greater than normal (T-300 and T-600).

The experimental units were 1000 L tanks, each containing 24 fish, with 3 tanks for each of the four diets. After 84 days, fish were euthanized and tissues analyzed. Fish were analyzed individually (the *measurement* unit), and the data were then averaged to produce a single value for each of the large tanks. The tank values were used in statistical analyses, although these data could also have been analyzed by fitting a nested model with tank nested in diet and individual fish values as the observations.

1. What is the biological question?
2. What is the statistical query (hypothesis or prediction)?
3. What are the factors in this experiment?
4. Are these factors fixed or random? Why?
5. Write out the linear model that corresponds to this experiment.
6. What are the experiment units?

Question 2 (Multifactor models)

Long and Porturas (2014) examined the effect of multiple stressors on the performance of a saltmarsh plant that is important for ecological restoration. The plant, *Spartina foliosa*, can be affected by herbivory (from scale insects). Saltmarshes are also environments of varying salinity, and insects do not always cope well with raised salinity. Long and Porturas asked whether *Spartina* was affected by herbivory in the same way in places where salinity was higher. To answer the question, they experimentally removed scale insects or left them intact, on plots with salinity at ambient levels or elevated. The experiment was repeated at two sites chosen to be very different in overall elevation within a marsh in southern California. On each plot, they recorded the senescence time of *Spartina* shoots (they senesce sooner when damaged by insects).

1. What is the biological question?
2. What is the statistical query (hypothesis or prediction)?
3. What are the factors in this experiment?
4. Are these factors fixed or random? Why?
5. Write out the linear model that corresponds to this experiment.
6. What are the experiment units?

Group activity: Scrutinizing data analyses

Part of working as a scientist involves reading published material and using the results. You may be synthesising the results from other papers as part of planning your own study or to make overall recommendations to another group. You may look at the published material to interpret your own data, and to decide what to do next. If you continue in science, you may also find yourself reviewing manuscripts for publication and assessing proposals for funding. In these cases, you need to look at the claims made by authors, generally in abstracts, media releases, etc., and decide whether the data support those claims.

How closely you scrutinize those claims will depend on why you are reading the material. If it's very close to your own research, the results may influence the direction your research takes, and you'll want to be very sure that you should trust the results. If, on the other hand, you're making a broad synthesis, your conclusion may be based on many published papers, and you'll pay less attention to individual ones.

This astivity asks you to look closely at some data analyses, decide whether you accept them, and whether your examination leads you to a different conclusion about the results. It is designed to get you thinking about how to review Results and Methods sections of papers. We have provided a guide to the kinds of questions you should ask about the treatment of the data.

We have asked a series of questions as a guide to your dissection of the paper, and they are provided at the back of this document. Things to pay particular attention to:

- The **biological question.** Is clear, and stated in a way that links to a statistical model?
- The **statistical model.** The authors should give you enough information for you to be satisfied, but it's often useful for you to write out your own model based on their description. Is everything that *should* be in the model there? If not, why not?
- The **experimental (or observational) units.** What are they, and have they been used appropriately?

If it's a complex model, be on the lookout for mixed models, i.e. the presence of random and fixed effects. Remember that the correct hypothesis tests in mixed models are not the same as when all factors are fixed. You can often see mismatches by looking closely at the degrees of freedom for particular tests.

- The **analysis.** Does it support the authors' conclusions? Yes? No? Can't tell?

As an example of how to do this, we'll work through a single paper, and **you'll be asked to answer the following questions**:

- What (biological) question(s) were the data designed to answer?
- What kind of statistical model(s) were fitted to the data?
- What are your preliminary conclusion, based purely on what the authors said about their results?
- What assumptions are associated with the statistical model(s) used?
- Did the authors provide you with enough information to determine whether the data analysis is appropriate?
- If not, what additional information would you like to see presented?
- What changes would you make to the data analysis?
- What would you conclude after your assessment of the data analysis?

References

Allen, RM, and Marshall, D (2014) Egg size effects across multiple life-history stages in the marine annelid Hydroides diramphus. *PLoS One*, **9**(7), e102253. doi:gf8mjs.

Binning, SA, Roche, DG, and Layton, C (2013) Ectoparasites increase swimming costs in a coral reef fish. *Biology Letters*, **9**(1), 20120927. doi:gsgbsx.

Birceanu, O, and Wilkie, MP (2018) Post-exposure effects of the piscicide 3-trifluoromethyl-4-nitrophenol (TFM) on the stress response and liver metabolic capacity in rainbow trout (Oncorhynchus mykiss). *Plos One*, **13**(7). doi:gdw2tk.

Caudill, SA, and Rice, RA (2016) Do Bird Friendly® Coffee Criteria Benefit Mammals? Assessment of Mammal Diversity in Chiapas, Mexico. *PLOS ONE*, **11**(11), e0165662. doi:gsgchq.

Cramp, RL, Reid, S, Seebacher, F, and Franklin, CE (2014) Synergistic interaction between UVB radiation and temperature increases susceptibility to parasitic infection in a fish. *Biology Letters*, **10**(9), 20140449. doi:gr9ckp.

Dai, L, Guo, X, Ke, X, … Du, Y (2020) Biomass allocation and productivity–richness relationship across four grassland types at the Qinghai Plateau. *Ecology and Evolution*, **10**(1), 506–516. doi:gj37pb.

Dalziel, AC, Martin, N, Laporte, M, Guderley, H, and Bernatchez, L (2015) Adaptation and acclimation of aerobic exercise physiology in Lake Whitefish ecotypes (Coregonus clupeaformis). *Evolution*, **69**(8), 2167–2186. doi:f7ptnz.

Didiano, TJ, Johnson, MTJ, and Duval, TP (2016) Disentangling the Effects of Precipitation Amount and Frequency on the Performance of 14 Grassland Species. *Plos One*, **11**(9). doi:f9rkgk.

Dixon, KM, Cary, GJ, Worboys, GL, and Gibbons, P (2018) The disproportionate importance of long-unburned forests and woodlands for reptiles. *Ecology and Evolution*, **8**(22), 10952–10963. doi:gfs587.

Griffen, AL, Thompson, ZA, Beall, CJ, … Fidel, PL (2019) Significant effect of HIV/HAART on oral microbiota using multivariate analysis. *Scientific Reports*, **9**(1), 19946. doi:gsfpz3.

Highfill, CA, Baker, BM, Stevens, SD, Anholt, RRH, and Mackay, TFC (2019) Genetics of cocaine and methamphetamine consumption and preference in Drosophila melanogaster. *PLOS Genetics*, **15**(5). doi:gr2pgk.

Hostetler, CM, Phillips, TJ, and Ryabinin, AE (2016) Methamphetamine Consumption Inhibits Pair Bonding and Hypothalamic Oxytocin in Prairie Voles. *PLoS One*, **11**(7), e0158178. doi:10.1371/journal.pone.0158178.

Hutto, D, and Barrett, K (2021) Do urban open spaces provide refugia for frogs in urban environments? *Plos One*, **16**(1). doi:gr2r9n.

Kaufman, JA, Turner, GH, Holroyd, PA, Rovero, F, and Grossman, A (2013) Brain Volume of the Newly-Discovered Species Rhynchocyon udzungwensis (Mammalia: Afrotheria: Macroscelidea): Implications for Encephalization in Sengis. *PLoS ONE*, **8**(3), e58667. doi:f4qwz3.

Kiffer, WP, Mendes, F, Casotti, CG, Costa, LC, and Moretti, MS (2018) Exotic Eucalyptus leaves are preferred over tougher native species but affect the growth and survival of shredders in an Atlantic Forest stream (Brazil). *PLOS ONE*, **13**(1), e0190743. doi:gr9z2g.

Kiss, T, Szabó, A, Oszlánczi, G, … Csupor, D (2017) Repeated-dose toxicity of common ragweed on rats. *PLOS ONE*, **12**(5), e0176818. doi:f96gk8.

Kraskura, K, Hardison, EA, Little, AG, … Eliason, EJ (2020) Sex-specific differences in swimming, aerobic metabolism and recovery from exercise in adult coho salmon (*Oncorhynchus Kisutch*) across ecologically relevant temperatures. *Conservation Physiology*, **9**(1), coab016. doi:gr9cmm.

La Rosa, RJ, and Conner, JK (2017) Floral function: Effects of traits on pollinators, male and female pollination success, and female fitness across three species of milkweeds (Asclepias). *American Journal of Botany*, **104**(1), 150–160. doi:gr2r9p.

Le Boeuf, BJ, Crocker, DE, Costa, DP, Blackwell, SB, Webb, PM, and Houser, DS (2000) Foraging ecology of northern elephant seals. *Ecological Monographs*, **70**(3), 353–382. doi:fj9rqc.

Li, H-L, Wang, Y-Y, Zhang, Q, Wang, P, Zhang, M-X, and Yu, F-H (2015) Vegetative Propagule Pressure and Water Depth Affect Biomass and Evenness of Submerged Macrophyte Communities. *Plos One*, **10**(11). doi:f8bw5m.

Martello, E, Bigliati, M, Adami, R, … Bruni, N (2022) Efficacy of a dietary supplement in dogs with osteoarthritis: A randomized placebo-controlled, double-blind clinical trial. *PLOS ONE*, **17**(2), e0263971. doi:gr9cmt.

Militão, T, Gómez-Díaz, E, Kaliontzopoulou, A, and González-Solís, J (2014) Comparing Multiple Criteria for Species Identification in Two Recently Diverged Seabirds. *PLoS ONE*, **9**(12), e115650. doi:f6zccr.

Morrissette-Boileau, C, Boudreau, S, Tremblay, J-P, Côté, SD, and Gilliam, F (2018) Simulated caribou browsing limits the effect of nutrient addition on the growth of Betula glandulosa, an expanding shrub species in Eastern Canada. *Journal of Ecology*, **106**(3), 1256–1265. doi:gc8qqn.

Peake, AJ, and Quinn, GP (1993) Temporal variation in species-area curves for invertebrates in clumps of an intertidal mussel. *Ecography*, **16**, 269–277. doi:cwzvgc.

Peraza, I, Chételat, J, Richardson, M, ... Ryjkov, A (2023) Diet and landscape characteristics drive spatial patterns of mercury accumulation in a high-latitude terrestrial carnivore. *PLOS ONE*, **18**(5), e0285826. doi:gsn8rr.

Posthumus, EE, Koprowski, JL, and Steidl, RJ (2015) Red Squirrel Middens Influence Abundance but Not Diversity of Other Vertebrates. *Plos One*, **10**(4). doi:f7jwtd.

Seitz, RD, Lipcius, RN, and Hines, AH (2016) Consumer versus resource control and the importance of habitat heterogeneity for estuarine bivalves. *Oikos*, **126**(1), 121–135. doi:f9mh38.

Senadheera, SD, Turchini, GM, Thanuthong, T, and Francis, DS (2012) Effects of Dietary Vitamin B $_6$ Supplementation on Fillet Fatty Acid Composition and Fatty Acid Metabolism of Rainbow Trout Fed Vegetable Oil Based Diets. *Journal of Agricultural and Food Chemistry*, **60**(9), 2343–2353. doi:f3wnkq.

Sinclair, ARE, and Arcese, P (1995) Population Consequences of Predation-Sensitive Foraging: The Serengeti Wildebeest. *Ecology*, **76**(3), 882–891. doi:fkr5bq.

Tonkin, JD, Shah, DN, Kuemmerlen, M, ... Jähnig, SC (2015) Climatic and Catchment-Scale Predictors of Chinese Stream Insect Richness Differ between Taxonomic Groups. *PLOS ONE*, **10**(4), e0123250. doi:gsn58k.

Townsend, AK, Taff, CC, Wheeler, SS, ... Boyce, WM (2018) Low heterozygosity is associated with vector-borne disease in crows. *Ecosphere*, **9**(10). doi:gfnq8b.

Uefune, M, Abe, J, Shiojiri, K, Urano, S, Nagasaka, K, and Takabayashi, J (2020) Targeting diamondback moths in greenhouses by attracting specific native parasitoids with herbivory-induced plant volatiles. *Royal Society Open Science*, **7**(11), 201592. doi:gsgcg9.

Vosteen, I, Gershenzon, J, and Kunert, G (2016) Hoverfly preference for high honeydew amounts creates enemy-free space for aphids colonizing novel host plants. *Journal of Animal Ecology*, **85**(5), 1286–1297. doi:f9csdq.

Walter, DE, and O'Dowd, DJ (1992) Leaves with domatia have more mites. *Ecology*, **73**(4), 1514–1518. doi:10.2307/1940694.