Statistical Modelling: Understanding Mean Structure

Chapter 3

Terry Neeman and Timothée Bonnet

November 21, 2018

Key components of a statistical model of an experiment

- Outcome measure
  - Response variable
  - Measure of interest
- Experimental factors
  - Conditions that can be manipulated
  - Conditions of interest (e.g. genotype, gender)
  - Main questions: do the conditions impact upon the outcome measure?
- Blocking factors
  - Conditions (not of interest) that may impact upon the outcome measure
  - Sources of variation in the experiment that need to be controlled for
  - Clustering of experimental units

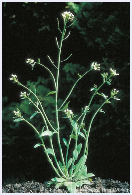
ALWAYS BEGIN WITH A RESEARCH QUESTION

Example 1: Can drought tolerance in Arabidopsis be improved through genetic modification?

Context

Outcome measure: Leaf water retention LWR (%)
Experimental factors:

- Gene A, genotypes (AA/aa)
- Gene B, genotypes (BB/bb)



© K. Lauber - Flora Helvetica - Haupt Verlag

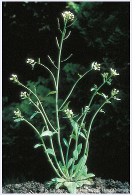
How many parameters to describe the different genotypes combinations?

Example 1: Can drought tolerance in Arabidopsis be improved through genetic modification?

Context

Outcome measure: Leaf water retention LWR (%)
Experimental factors:

- Gene A, genotypes (AA/aa)
- Gene B, genotypes (BB/bb)



How many parameters to describe the different genotypes combinations?

4 treatments		Gene A	
		AA	aa
Gene B	BB	C	$C+A$
	bb	$C+B$	$C+A+B+D$

Two different models

Additive model - 3 parameters

4 treatments		Gene A	
		AA	aa
Gene B	BB	C	$C + A$
	bb	$C + B$	$C + A + B$

Full factorial model / Interactive model - 4 parameters

4 treatments		Gene A	
		AA	aa
Gene B	BB	C	$C + A$
	bb	$C + B$	$C + A + B + D$

What is different? What does the additive model assume?

Which model to use?

Additive model - 3 parameters

4 treatments		Gene A	
		AA	aa
Gene B	BB	C	$C + A$
	bb	$C + B$	$C + A + B$

Full factorial model / Interactive model - 4 parameters

4 treatments		Gene A	
		AA	aa
Gene B	BB	C	$C + A$
	bb	$C + B$	$C + A + B + D$

Analysis in R

1. Import data "Prac3mockLWR.csv"
2. Visualize data
3. Model data
4. Assess model assumptions

Analysis in R

1. Import data ”Prac3mockLWR.csv”

```
LWR <- read.csv(\Prac3mockLWR.csv")
```

Analysis in R

2. Visualise the data

```
ggplot(LWR, aes(GeneB,LWR,colour=GeneA)) +
  geom_boxplot() + geom_point()
```

Full factorial or additive?

Analysis in R

3. Model data

```
lmadditive <- lm(LWR ~ GeneA + GeneB, data = LWR)
summary(lmadditive)
anova(lmadditive)
```

```
lminteraction <- lm(LWR ~ GeneA * GeneB, data = LWR)
summary(lminteraction)
anova(lminteraction)
emmeans(lminteraction, pairwise ~ GeneA|GeneB)
emmeans(lminteraction, pairwise ~ GeneB|GeneA)
```

What are the estimates for A, B, C, D under each models?

Analysis in R

4. Model assumptions

```
plot(lminteraction)
```

Which cabbage cultivar has the higher Vitamin C content on average?

Research context

- 60 cabbage heads
- 2 cultivars: c39 and c52
- 3 planting dates: Days 16, 20, 21

How many parameters to describe our scientific question?

Which cabbage cultivar has the higher Vitamin C content on average?

Research context

- 60 cabbage heads
- 2 cultivars: c39 and c52
- 3 planting dates: Days 16, 20, 21

		Cultivar	
		c39	c52
Planting date	Day 16	A	$A + B$
	Day 20	$A + C$	$A + B + C$
	Day 21	$A + D$	$A + B + D$
Marginal means			

Which cabbage cultivar has the higher Vitamin C content on average?

Fit additive and interactive models in R

Dataset “Prac3cabbagedata.csv”

Are temperature mechanisms modified in a genetically modified tomato plant?

Research context

- 2 tomato plants
- 2 Genotypes: WT/mutant
- Watering condition: Normal/Drought
- Leaf temperature measured

How many parameters to describe our scientific question?

Are temperature mechanisms modified in a genetically modified tomato plant?

Research context

- 2 tomato plants
- 2 Genotypes: WT/mutant
- Watering condition: Normal/Drought
- Leaf temperature measured

		Water condition		
		Normal	Drought	Marginal means
Genotype	WT			
	mutant			
Marginal means				

Are temperature mechanisms modified in a genetically modified tomato plant?

Dataset “Prac3droughtdata.csv”
Fit the appropriate model in R.
Compare genotypes and water conditions with emmeans

Relationship diameter/density differ between tree species?

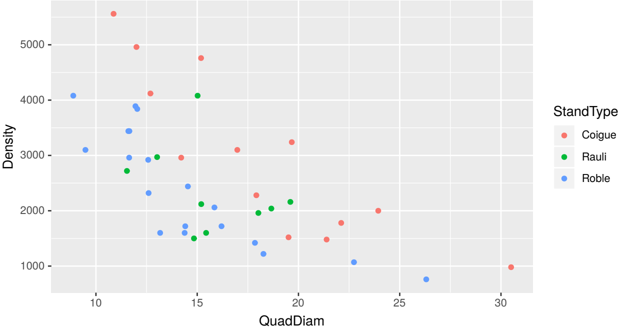
Research context

- *Nothofagus* in the Andes
- 41 plots with 3 species (StandTypes)
- Outcome: Plot density
- Factors: StandType, QuadDiam

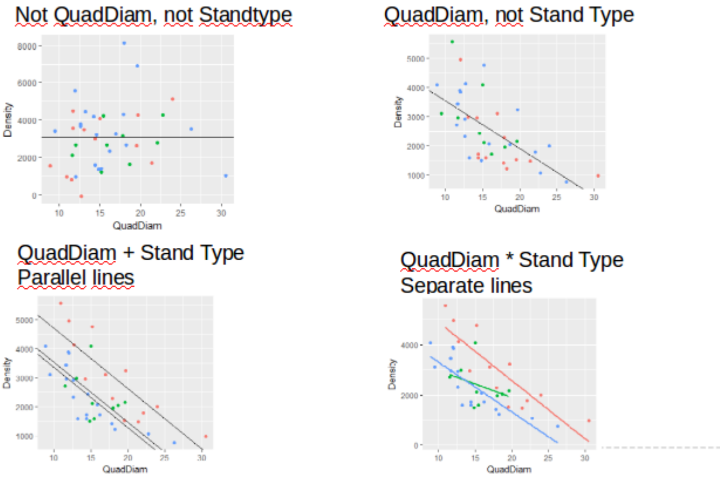
Relationship diameter/density differ between tree species?

Research context

- *Nothofagus* in the Andes
- 41 plots with 3 species (StandTypes)
- Outcome: Plot density
- Factors: StandType, QuadDiam



Which model to use



Not QuadDiam, not Standtype

QuadDiam, not Stand Type

QuadDiam + Stand Type
Parallel lines

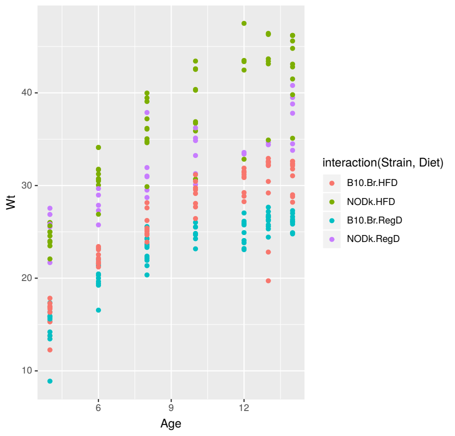
QuadDiam * Stand Type
Separate lines

Fit the models with "Prac3forest.csv" and answer the scientific question

Are NODk mice more susceptible to obesity when exposed to a high fat diet?

Research context

- 37 mice: 16 NODk /21 WT
- Randomised to either regular or high fat diet
- Monitored for 14 weeks
- Outcome measure: Body weight (g)
- Experimental factors: Diet (2), Strain (2), Age (7)



Data "Prac3diabeticmice.csv"