All of the Linear Models

Data Science
Week 7

Linear Models

Today we will go through

- Quick recap of simple linear models
- The 3 assumptions
- Generalised linear models (just a way of dealing with non-normal data)
- Linear mixed models (just a way of dealing with non-independent data)

What even is a linear model

Where we want to see if there is a correlation between something and something else…

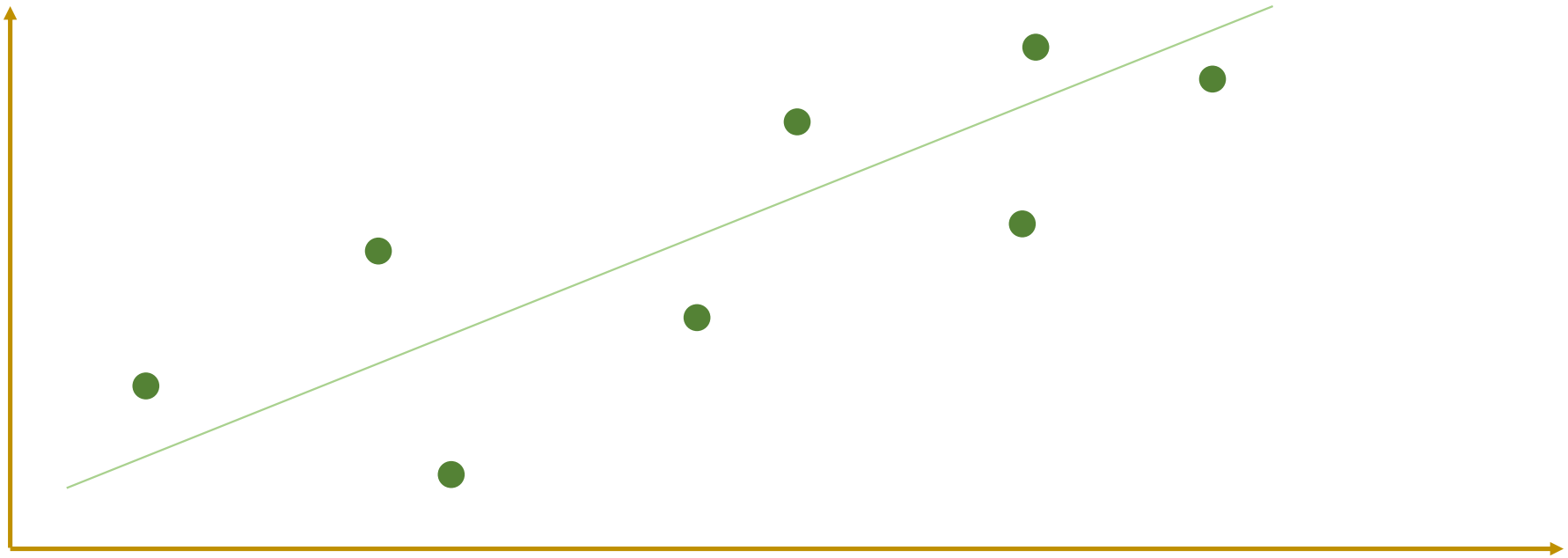
Is species richness higher where there is higher rainfall?

Is a population declining with time?

Linear models involving identifying your response, and one or more predictors, and seeing if they correlate

Beware: correlation does not equal causation! https://www.tylervigen.com/spurious-correlations

A simple linear model



Coding a linear model

```
lm(Count ~ Year, data = MyData)
```

Assumptions

Linear models have 3 main assumptions (aside from the obvious: that there is a linear relationship between predictor and response)

1. The residuals are normally distributed

   (and data is continuous below and above zero)

2. The data are independent

3. Variance is similar ('homogenous') between groups

Assumptions

What do we do if these are violated?! We might need to consider adjusting the model!

Assumptions

What do we do if these are violated?! We might need to consider adjusting the model!

1. The residuals are normally distributed ← GLM (Generalised linear model)

   (and data is continuous below and above zero)

2. The data are independent ← LMM (Linear mixed Model)

3. Variance is the similar ('homogenous') between groups

↑ Harder to fix… affects interpretation

Assumptions

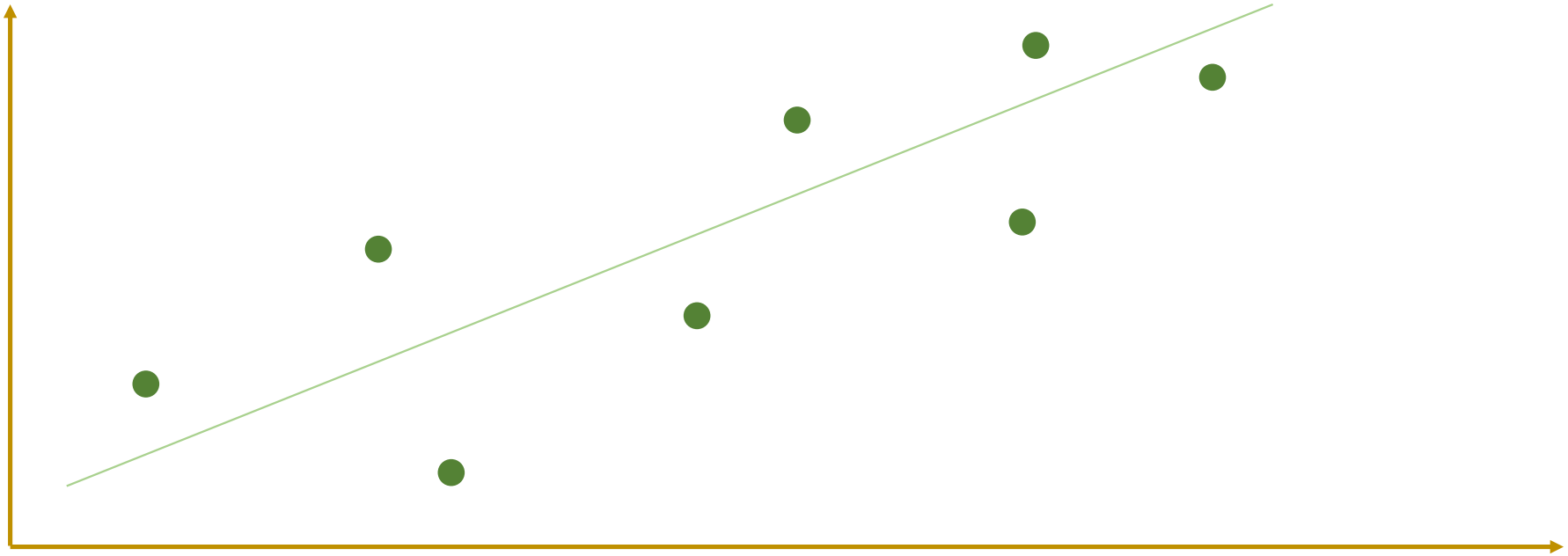
What do we do if these are violated?! We might need to consider adjusting the model!

1. The residuals are normally distributed ← GLM (Generalised linear model)
2. The data are independent ← LMM (Linear mixed Model)
3. Variance is the similar ('homogenous') between groups

↑ Harder to fix… affects interpretation

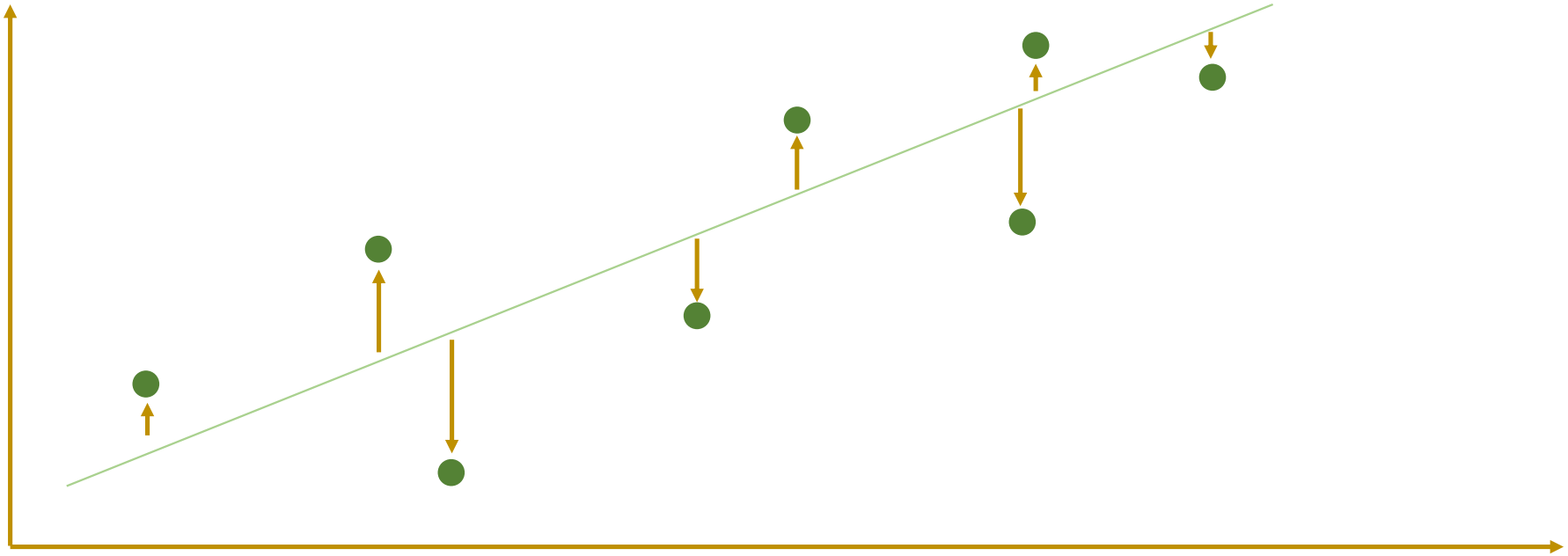
Assumption 1: Normality

1. The residuals are normally distributed



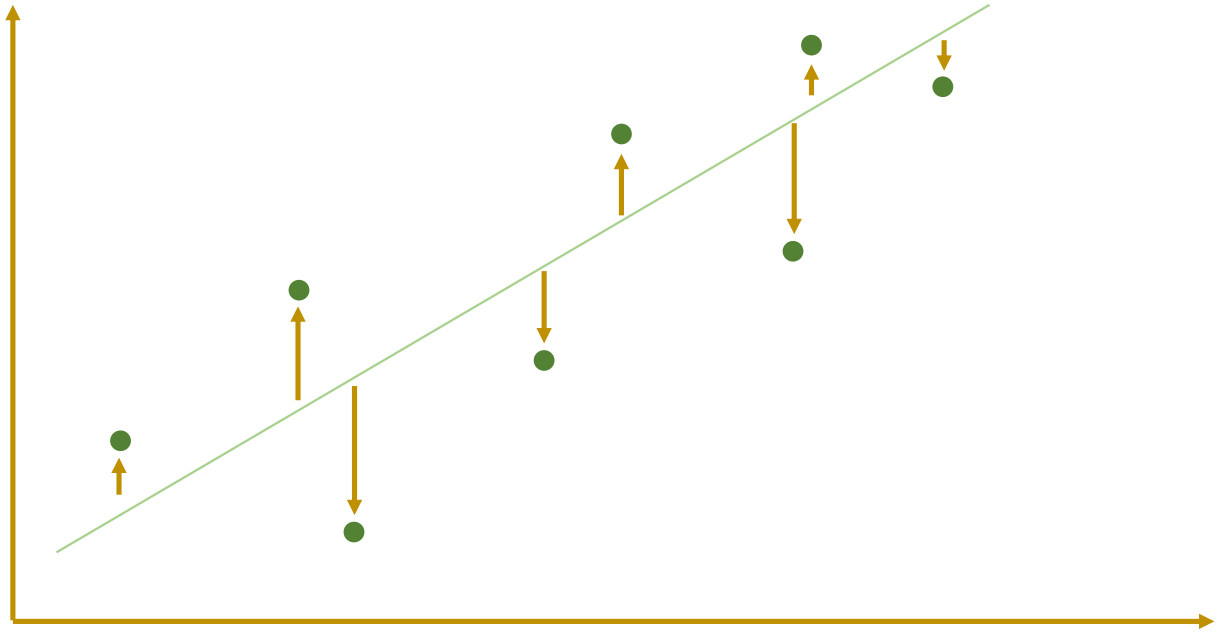
Assumption 1: Normality

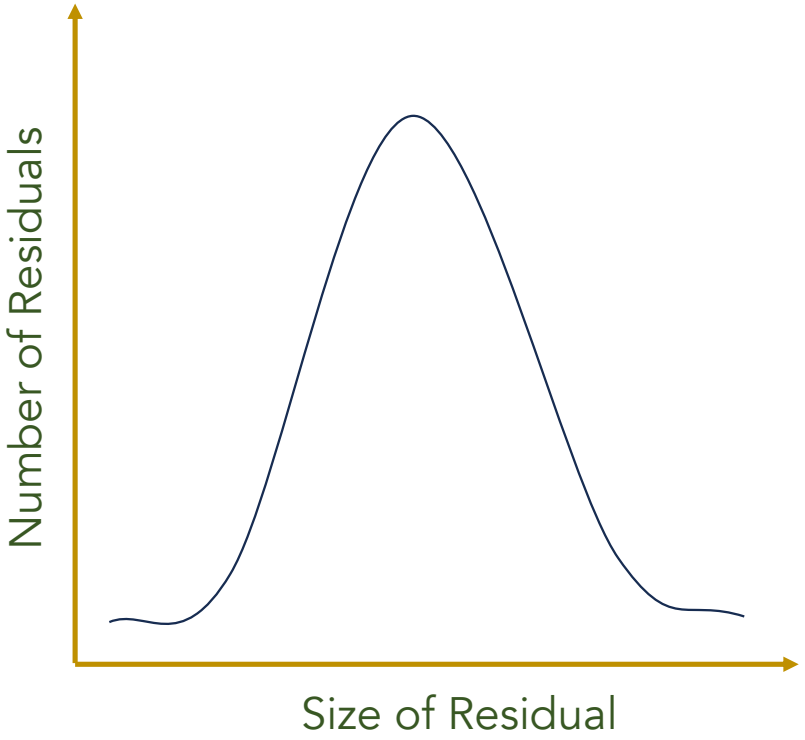
1. The residuals are normally distributed



Assumption 1: Normality

1. The residuals are normally distributed (i.e. a few are very close to the line, a few are very far, most are somewhere in between)





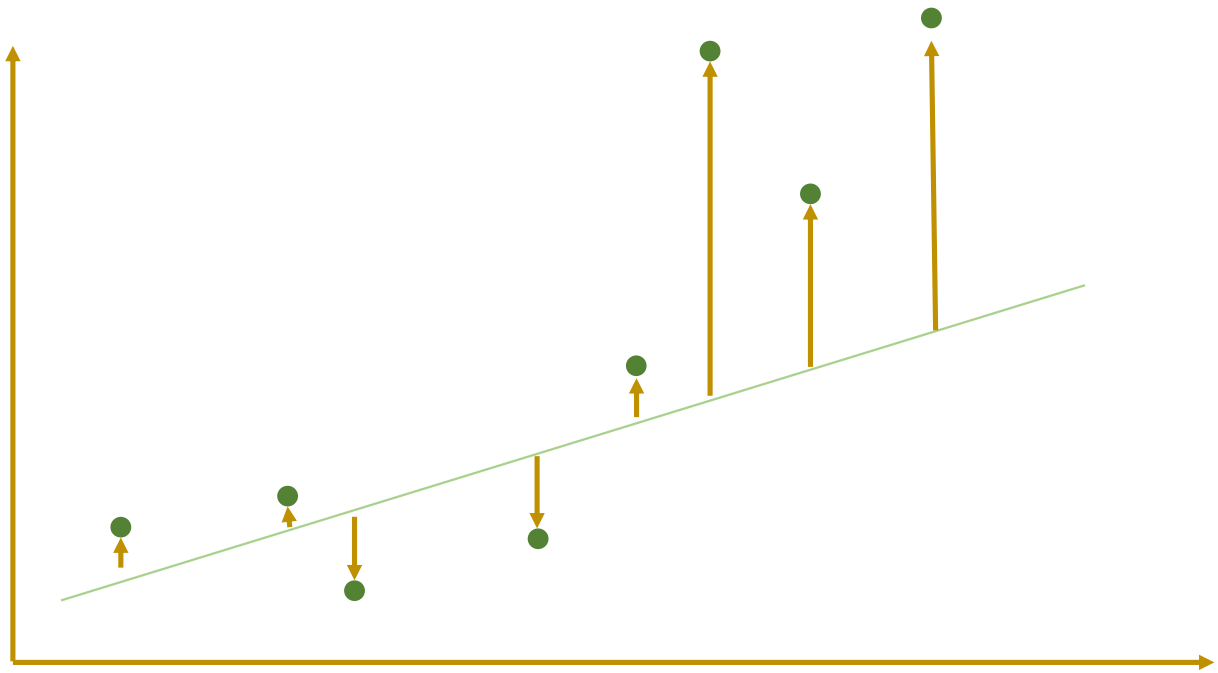
Assumption 1: Normality

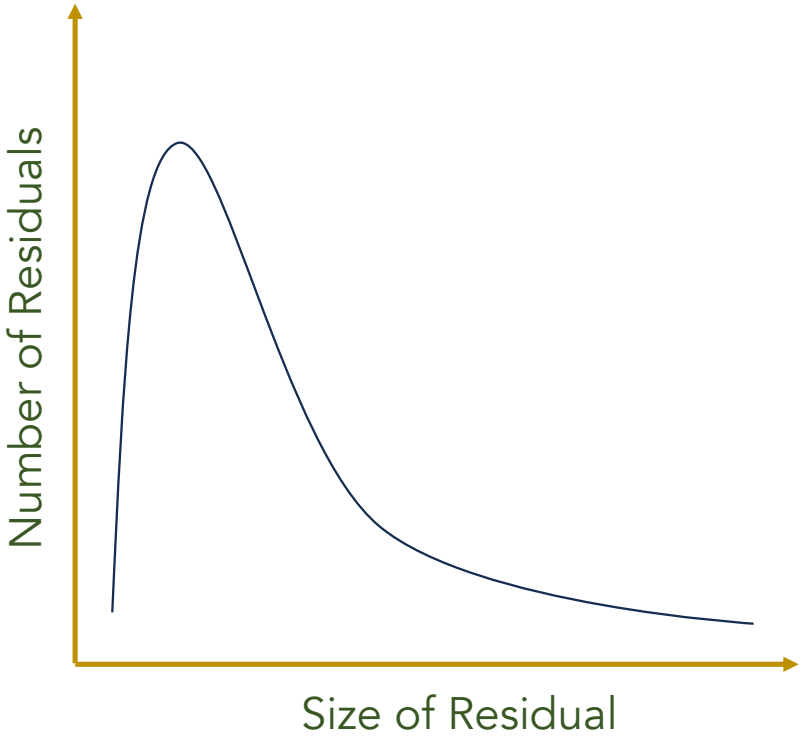
1. The residuals are normally distributed

When might this not be the case?

Assumption 1: Normality

1. An example of non normal residuals (Most are very close to the line, a few are very far away)





Assumption 1: Normality

Non-normal residuals often come with other data issues – like non-continuous data, or data that is only positive.

What's an example of this kind of data?

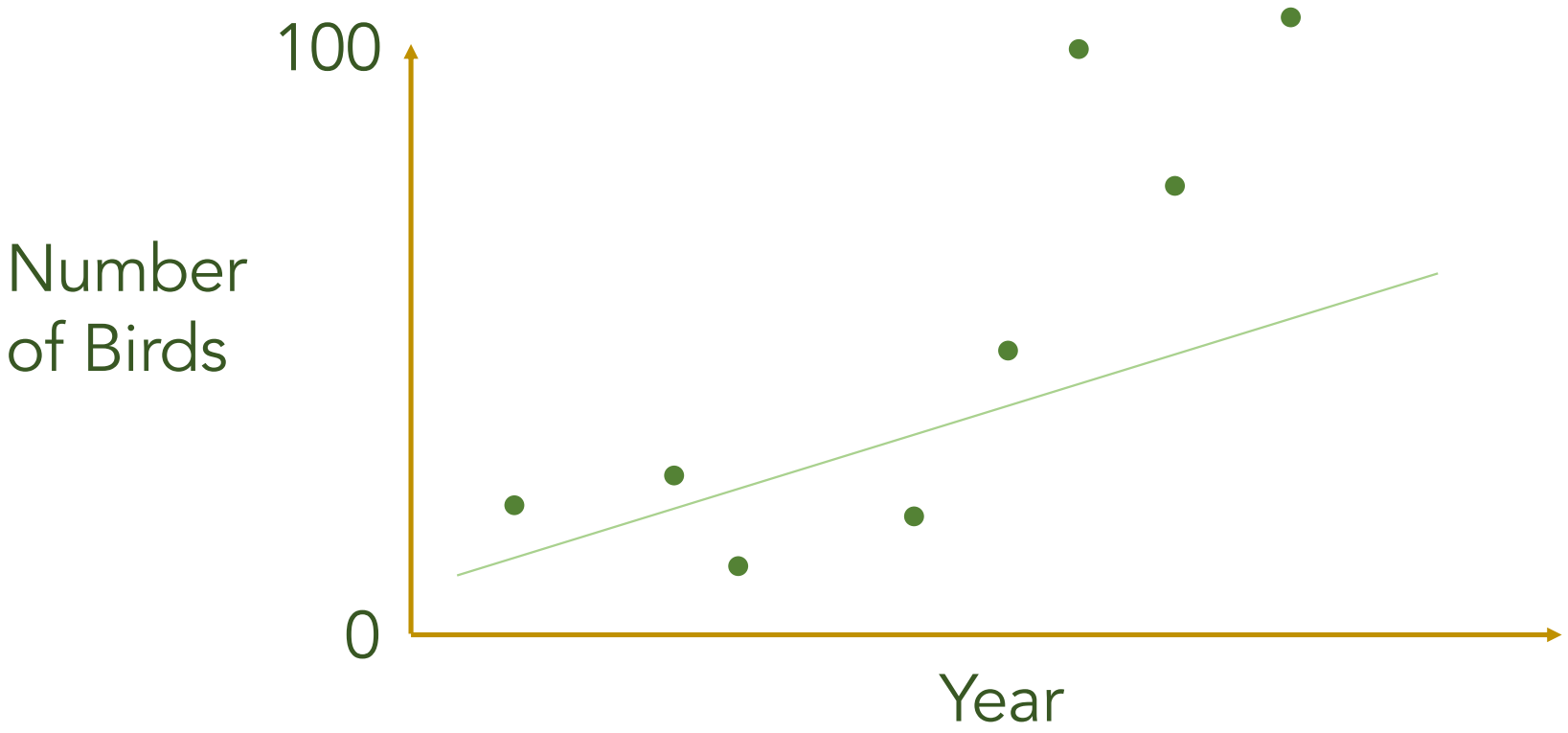
Assumption 1: Normality

Non-normal residuals often come with other data issues – like non-continuous data, or data that is only positive.

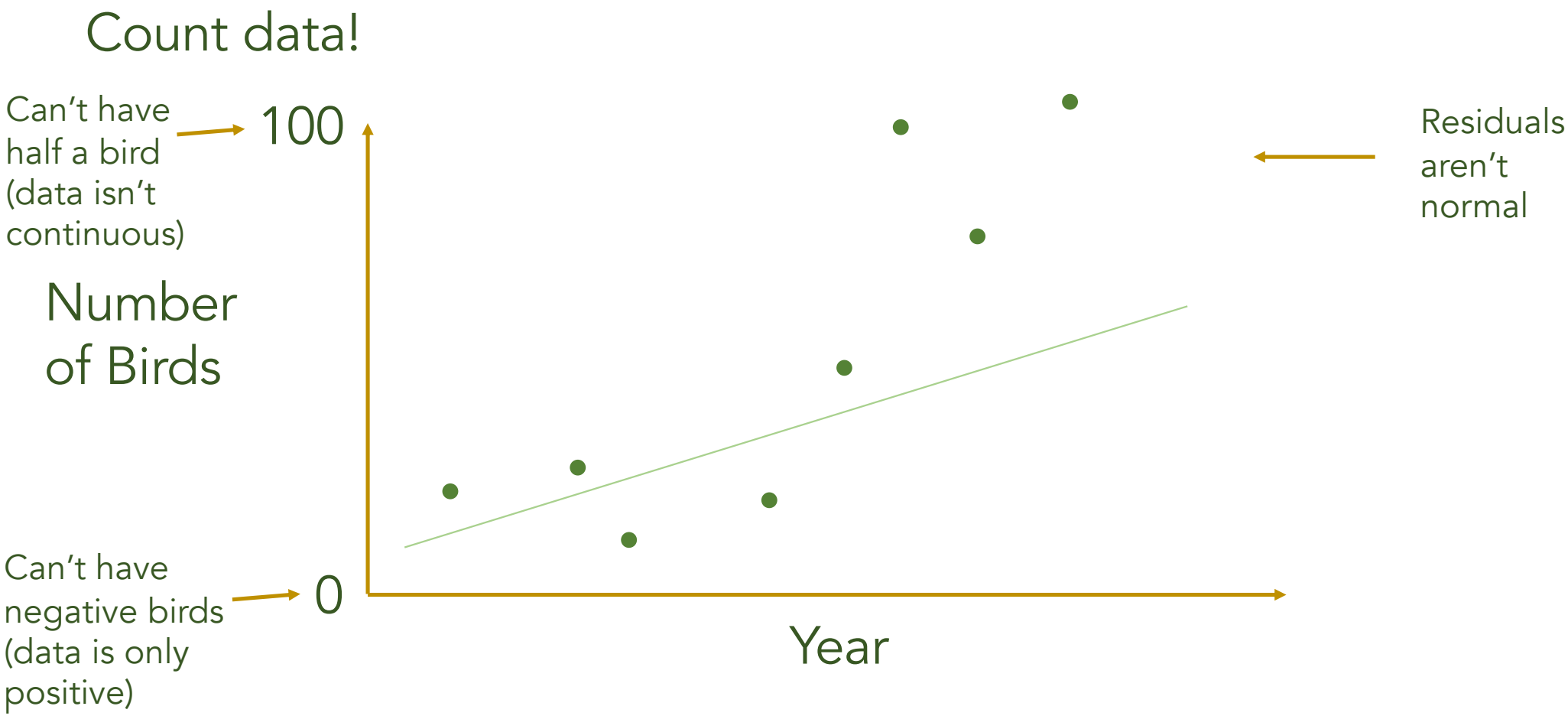
What's an example of this kind of data?

Assumption 1: Normality

Count data!



Assumption 1: Normality



Assumption 1: Normality

Count data is basically a disaster for simple linear models

What's the solution?

Generalised Linear Models!

Generalised Linear Models!

These do maths under the hood to turn our response variable (that is violating assumptions) into one that violates less assumptions.

But the kind of maths depends on the *way* in which the response variable is violating assumptions… and that depends on its distribution

Generalised linear models

Uses a ‘link function’ to tell the model what the distribution of the response variable is

```
lm(Count ~ Year, data = MyData)
```

```
glm(Count ~ Year, data = MyData, family="poisson")
```

Generalised linear models

Uses a ‘link function’ to tell the model what the distribution of the response variable is

```
lm(Count ~ Year, data = MyData)
```

```
glm(Count ~ Year, data = MyData, family="poisson")
```

A generalised linear model

Which family should I specify? (these are the most commonly encountered in ecology)

Response Variable	Distribution aka Family
Count data (e.g. number of lions)	Poisson
Count data where numbers vary widely (e.g. waterbirds at a lake, where the count can be 3 one day, and 300 the next) (technical term for this is 'over-dispersed')	Quasi-poisson or negative binomial
Yes or No (e.g. a site is protected or not protected)	Binomial
Proportions (between 0 and 1)	Beta
Categories (e.g. good, better, best)	Logit (… or use ordinal regression – a class for another day, but look into package 'ordinal' and function clm)

Understanding check

What's the difference between these two models? (Remember, gaussian is another word for normal)

lm(response ~ predictor, data = MyData)

glm(response ~ predictor, data = MyData, family="gaussian")

Understanding check

What's the difference between these two models? (Remember, gaussian is another word for normal)

lm(response ~ predictor, data = MyData)

glm(response ~ predictor, data = MyData, family="gaussian")

They are the same!

Assumptions

What do we do if these are violated?! We might need to consider adjusting the model!

1. The residuals are normally distributed ← GLM (Generalised linear model)
2. The data are independent ← LMM (Linear mixed Model)
3. Variance is the similar ('homogenous') between groups

↑ Harder to fix… affects interpretation

2 min break

Assumptions

What do we do if these are violated?! We might need to consider adjusting the model!

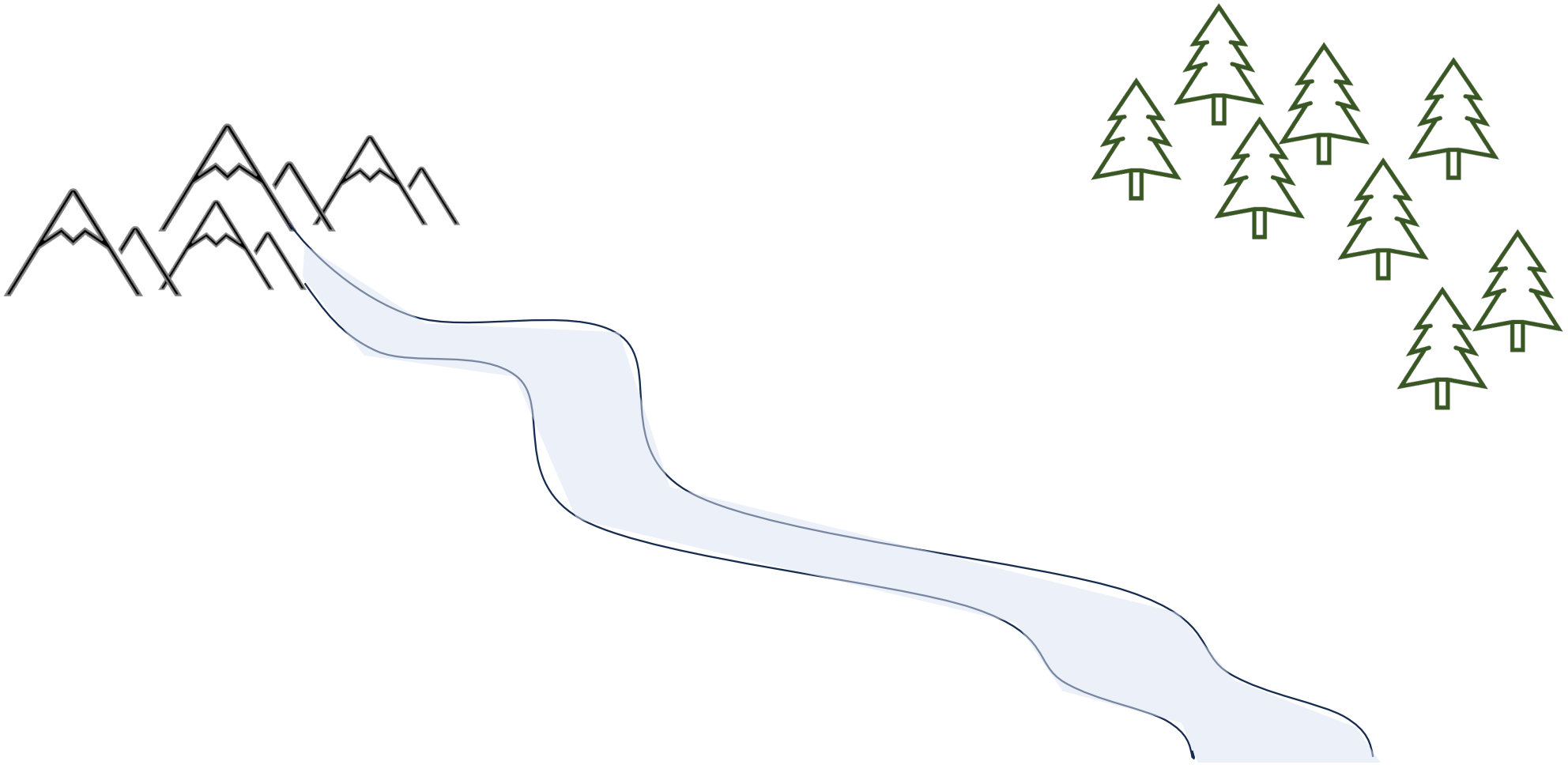
1. The residuals are normally distributed ← GLM (Generalised linear model)
2. The data are independent ← LMM (Linear mixed Model)
3. Variance is the similar ('homogenous') between groups

↑ Harder to fix… affects interpretation

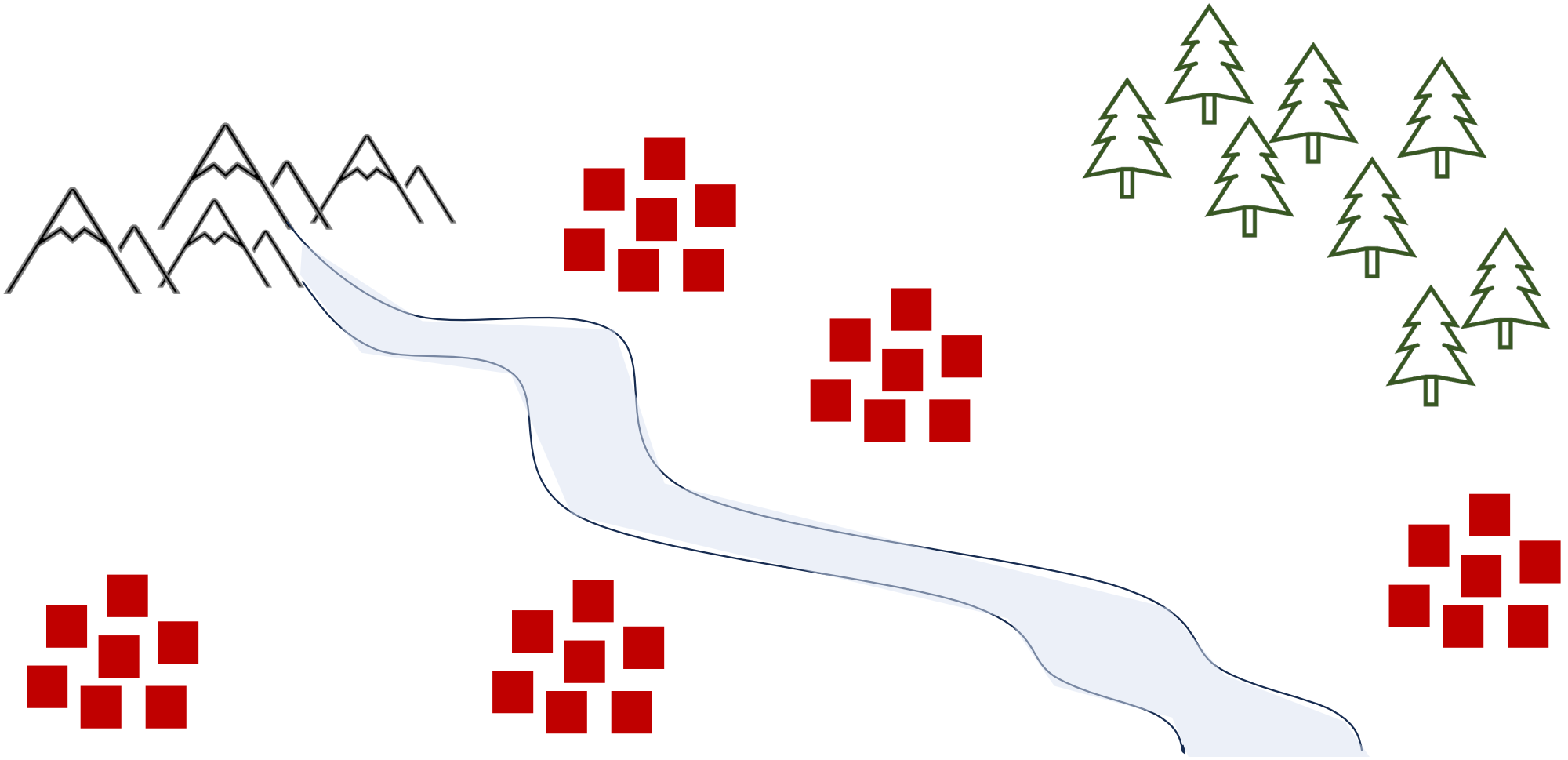
Assumption 2: Independence

When might data not be independent?

Assumption 2: Independence



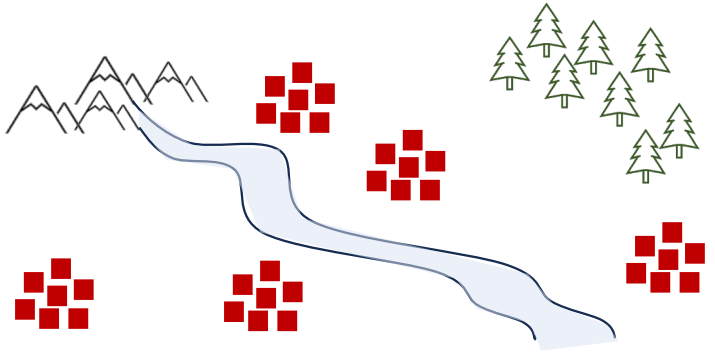
Assumption 2: Independence



Assumption 2: Independence

Our plots are NOT independent.
They are grouped by site.

What do we do?



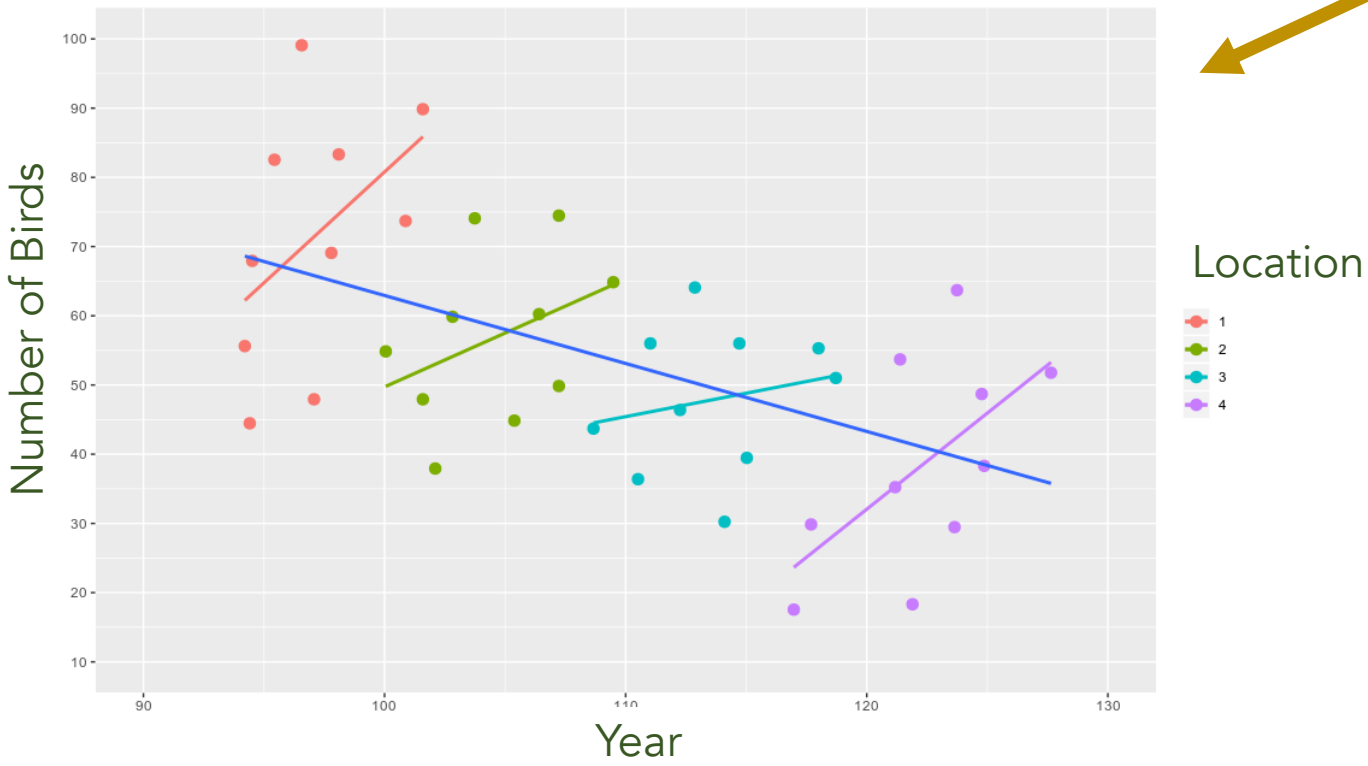
Linear *Mixed* Models!

Linear *Mixed* Models!

These allow you to say: “hey my data has groups”,
and they then calculate relationships by group

Why does this matter?

Why does grouping matter?



If we didn't tell the model there were groups, it would think birds were going down through time

Linear *Mixed* Models!

These allow you to say: “hey my data has groups”,
and they then calculate relationships by group

Why does this matter?

Linear *Mixed* Models

Random effect = how you tell the model there are groups (that you don't care about)

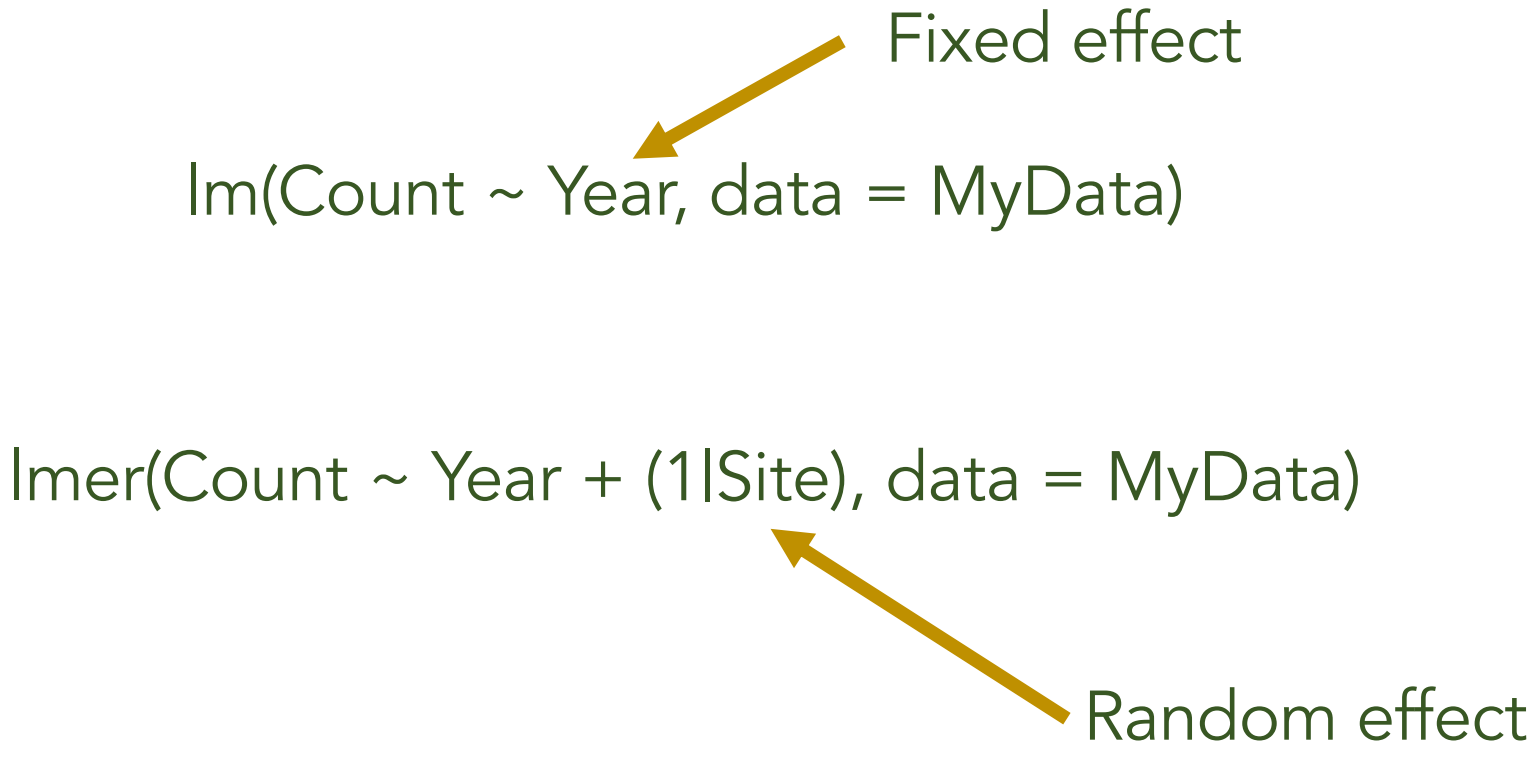
Fixed effect = We now re-term a 'normal' predictor as a 'fixed effect' to distinguish between the two. (Fixed effects can also be groups!! They're just groups you care about)

Linear *Mixed* Models

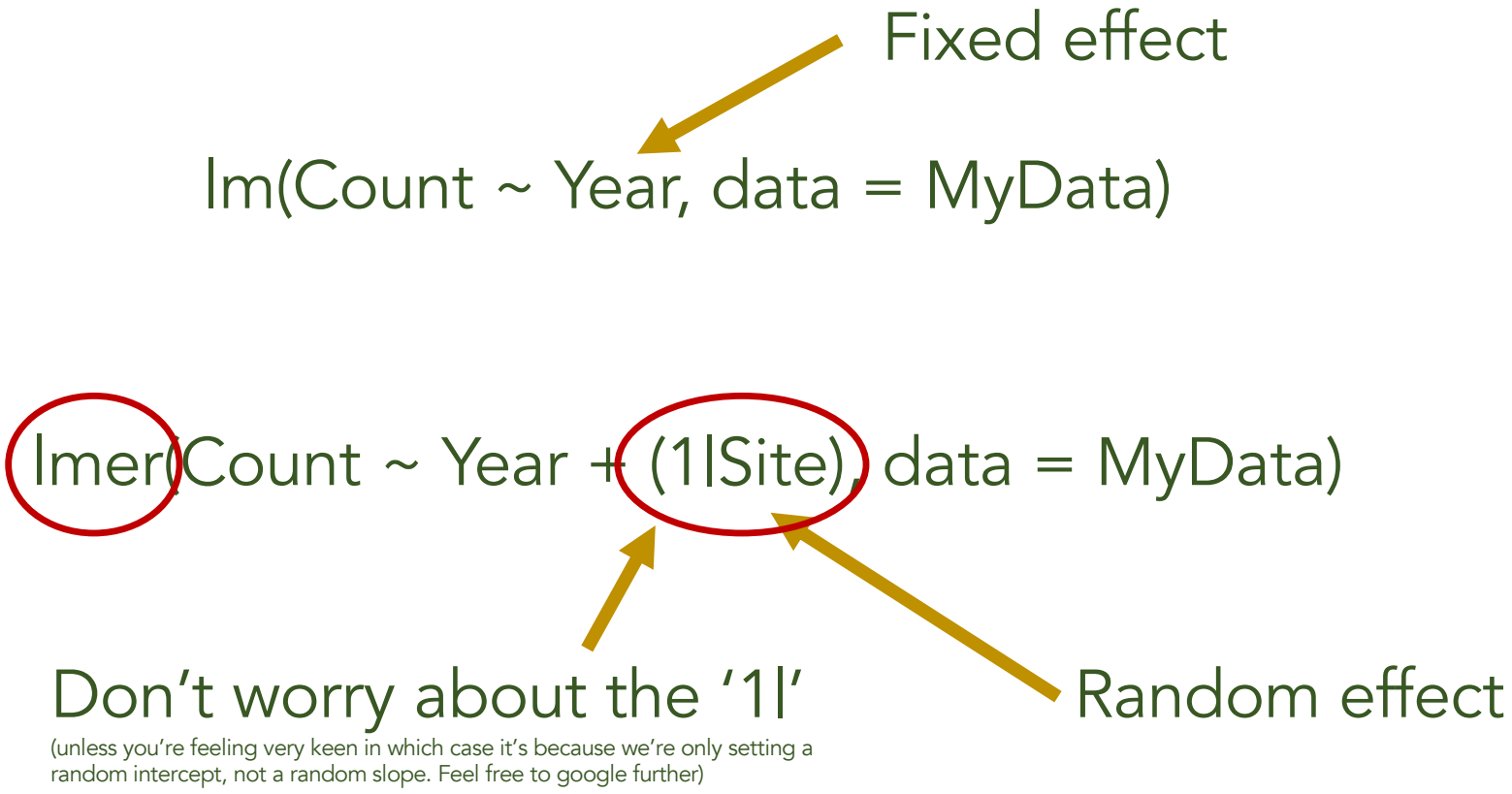
lm(Count ~ Year, data = MyData)

lmer(Count ~ Year + (1|Site), data = MyData)

Linear *Mixed* Models



Linear *Mixed* Models



Linear *Mixed* Models

A rule of random effects: you must have at least 5 groups!!

e.g. if your random effect is “plot” you must have at least 5 plots

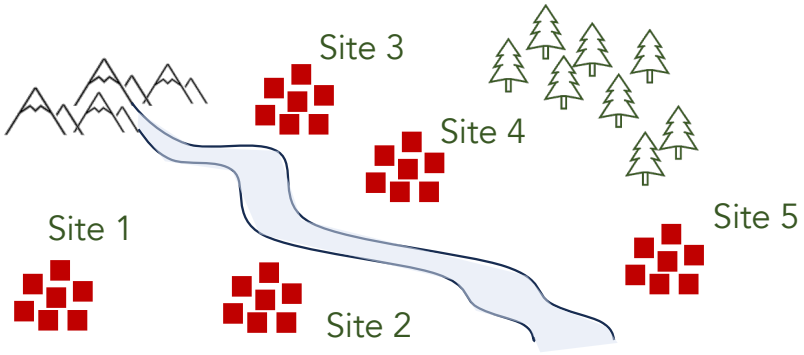
If you have less, just make ‘plot’ a fixed effect

What if I have more than one group?

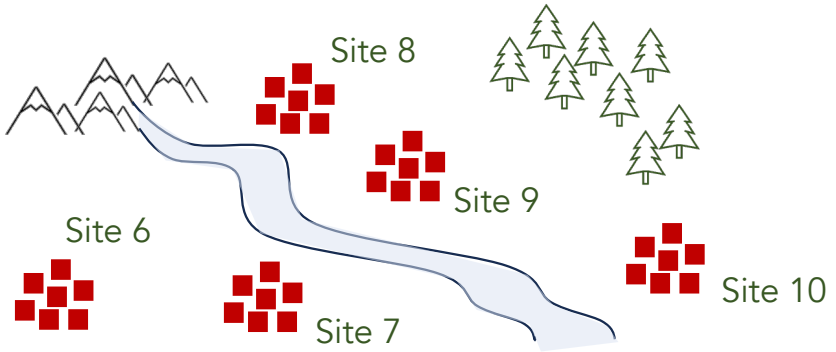
If you have more than one group, they can either be nested or crossed…

Nested Random Effects

Country 1



Country 2 … etc for more countries



Site is *nested* within Country (Site 1 can't occur in Country 2)

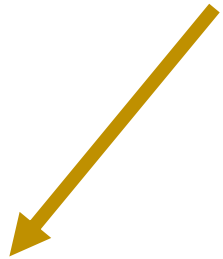
Nested Random Effects

Site is *nested* within Country (Site 1 can't occur in Country 2)

Ask "Can my sites *only* occur in one country each?" If yes – it's nested!

Nested Random Effects

The bigger "nest" goes first



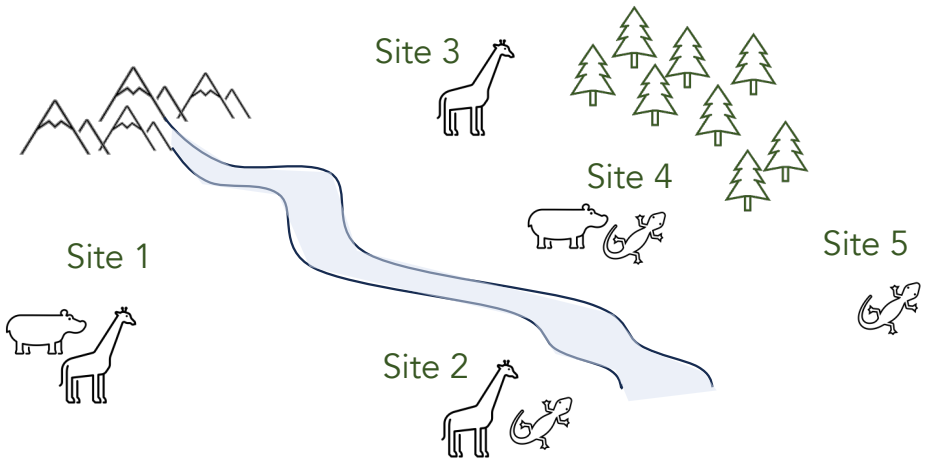
lmer(Count ~ Year + (1|Country/Site), data = MyData)

When you google, you may sometimes see people nest random factors using a colon (":"), imo best practice to use the slash ("/")

Crossed Random Effects

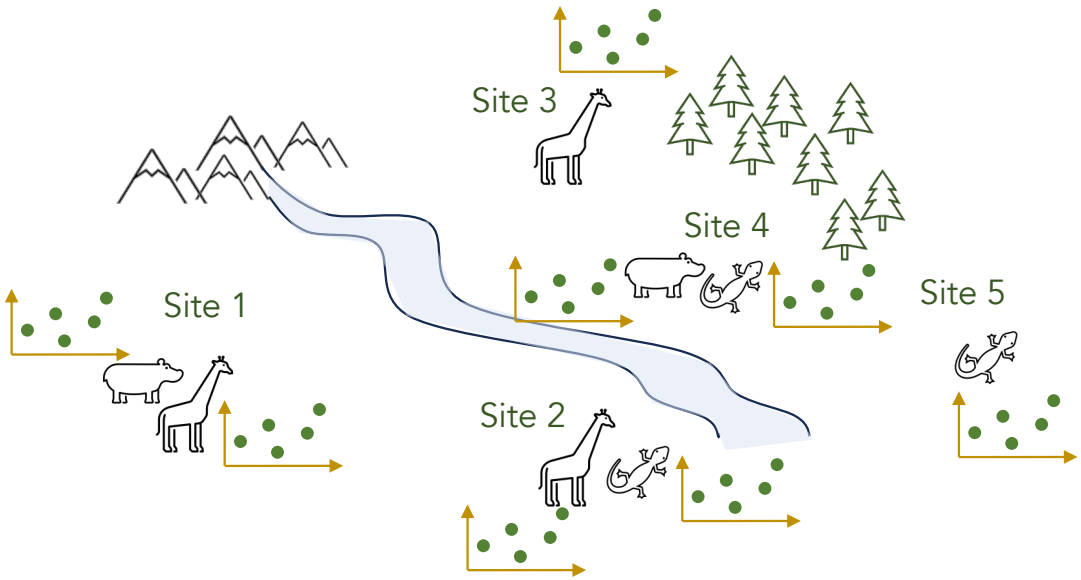
Crossed Random Effects

We have counts through time for each species, at each site



Crossed Random Effects

We have counts through time for each species, at each site



Ask "Can my species *only* occur in one site each?" If no – it's crossed!

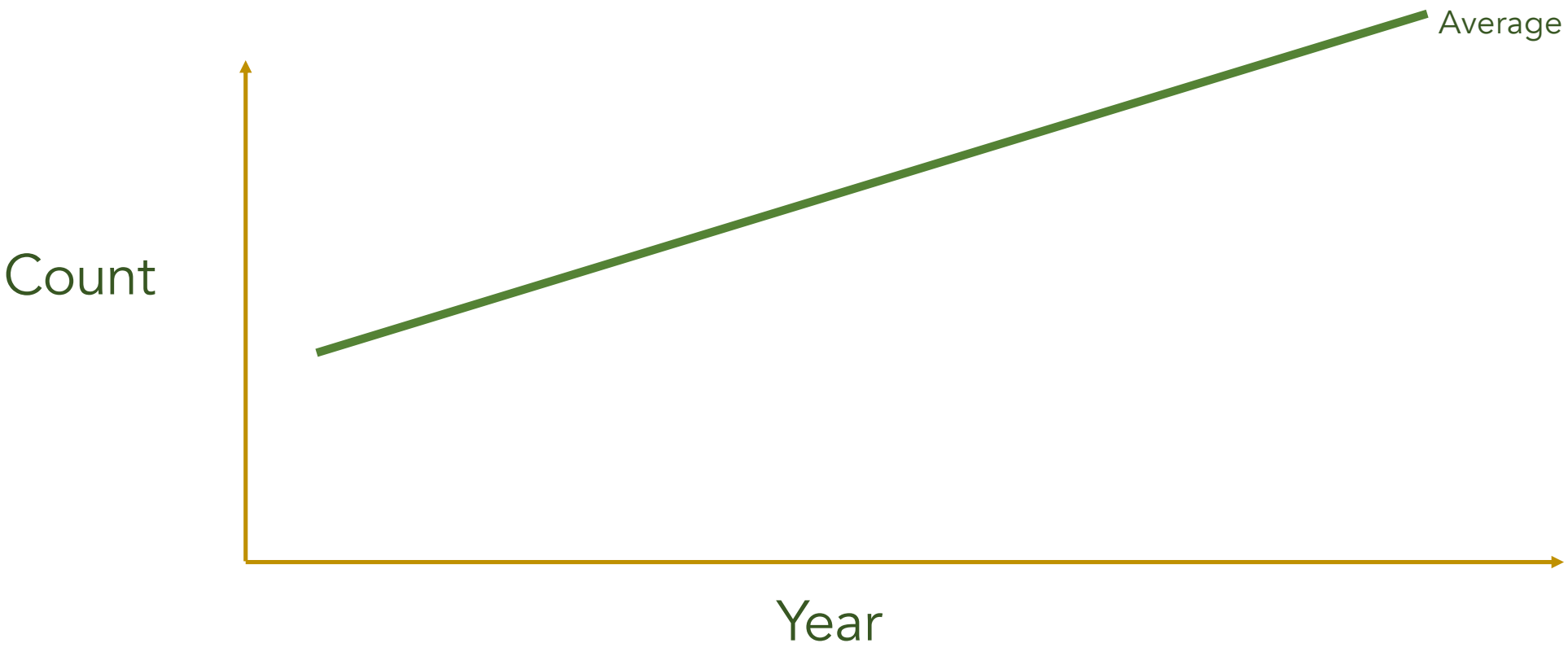
Types of Random Effects

Crossed random effect:

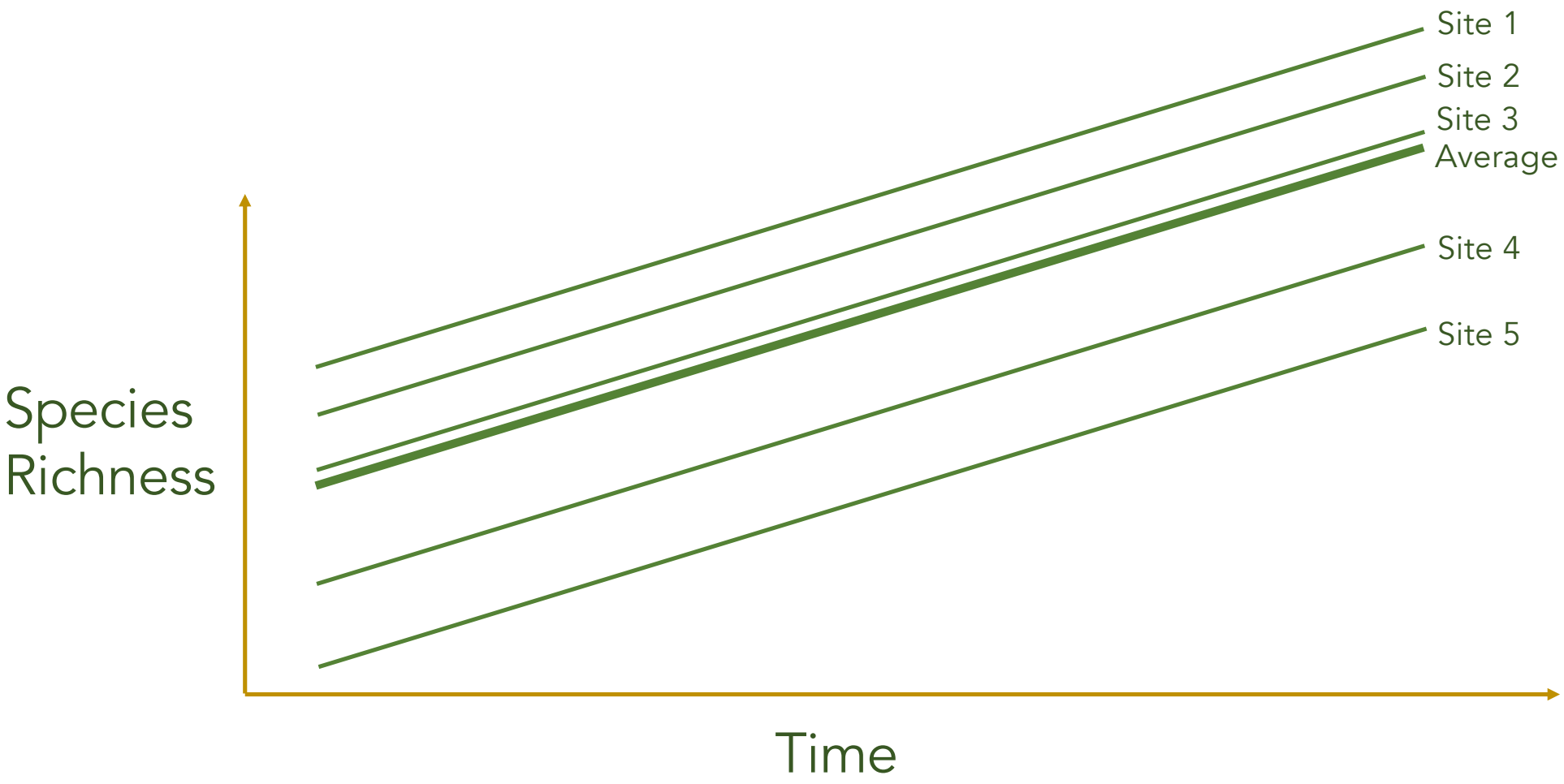
To add a crossed random effect,
we just.. add it on (simple!)

```
lmer(Count ~ Year + (1|Site) + (1|Species), data = MyData)
```

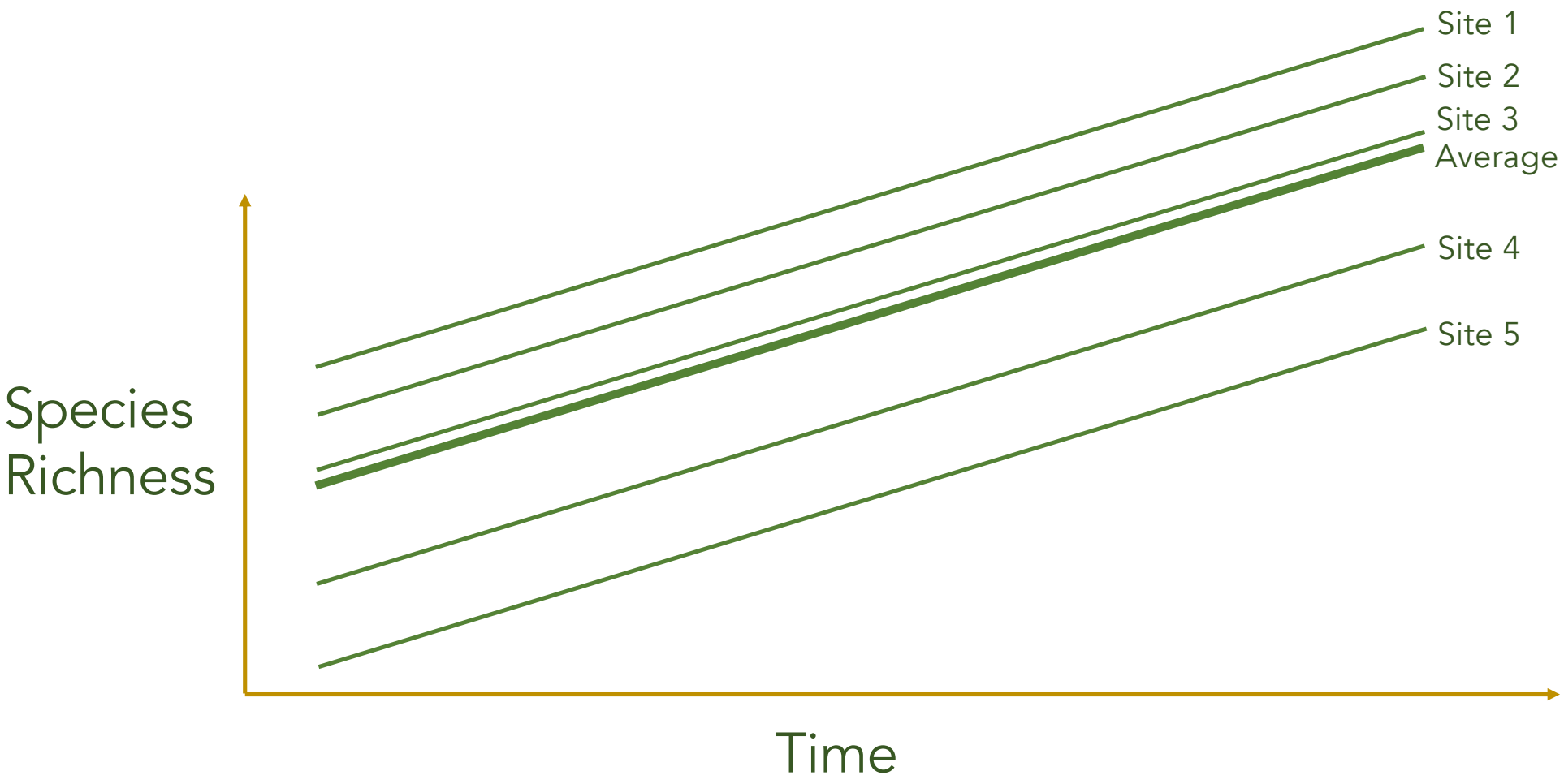
INTERCEPT NOT SLOPE



INTERCEPT NOT SLOPE



INTERCEPT NOT SLOPE



INTERCEPT NOT SLOPE

So far, random effects just change the intercept, *not* the relationship between your predictor and response

You CAN make them vary in slope as well, but that's beyond what we have time for today. But it's important to think about if you expect e.g. the relationship between time and species abundance to be different at each site, and if you do, you also need random slopes!

5 min break

Assumptions

What do we do if these are violated?! We might need to consider adjusting the model!

1. The residuals are normally distributed ← GLM (Generalised linear model)
2. The data are independent ← LMM (Linear mixed Model)
3. Variance is the similar ('homogenous') between groups

↑ Harder to fix… affects interpretation

Assumptions

What do we do if these are violated?! We might need to consider adjusting the model!

1. The residuals are normally distributed ← GLM (Generalised linear model)
2. The data are independent ← LMM (Linear mixed Model)
3. Variance is the similar ('homogenous') between groups

↑ Harder to fix… affects interpretation

Summary

Are your residuals normal?	Does your data have groups? **No**	Does your data have groups? **Yes (you must specify random effects, i.e. your grouping variables)**
Yes	Linear Model (LM)	Linear Mixed Model (LMM)
No (you must specify the distribution, e.g. poisson)	Generalised Linear Model (GLM)	Generalised Linear Mixed Model (GLMM)

Summary – Code calls

Does your data have groups?

Are your residuals normal?	**No**	**Yes (you must specify random effects, i.e. your grouping variables)**
Yes	Linear Model (LM) stats::lm	Linear Mixed Model (LMM) lme4::lmer
No (you must specify the distribution, e.g. poisson)	Generalised Linear Model (GLM) stats::glm	Generalised Linear Mixed Model (GLMM) lme4::glmer

*note! You'll find that other packages offer mixed models. lme4 is a common one, but feel free to shop around. E.g. nlme or glmmTMB

Summary

If your model contains a generalized element

- Identify your link function, i.e. the distribution of the residuals/response (see table earlier in powerpoint or in CC!)
- Code like this: family = “poisson”

If your model contains a mixed element…

- Identify your grouping variables (i.e. your random effects)
- Make sure all have at least 5 levels
- If more than one, figure out if they’re nested or crossed
- Code like this: (1|BigNest/SmallNest) + (1|Crossed)

Terminology confusion

This	Is also referred to as this…
Predictor	Independent Variable, Explanatory variable, Covariate, X
Response	Dependent Variable, Y
A linear model where the predictors are categorial	ANOVA
A 'normal', basic linear model	Linear regression, ordinary linear regression, ordinary least squares (OLS) linear regression (also VERY confusingly, sometimes a 'general' linear model, if you have multiple predictors. Different from a 'generalized' linear model)
A 'normal' predictor in a linear model	A fixed effect
Normal distribution	Gaussian distribution
A linear mixed model	A hierarchical model, mixed effects model, multi-level model
Random effect	Random factor*
Estimate	Effect size*, coefficient*

*one could argue there are technically differences between these terms but in practice they're used interchangeably

Quiz time!

Question 1

I went into the forest and measured the height of 100 plants and the pH of the soil.

I want to know if soil Ph predicts leaf size.

What would my model look like?

I went into the forest and measured the height of 100 plants and the pH of the soil.

I want to know if soil Ph predicts leaf size.

What would my model look like?

$$\text{Height} \sim \text{pH}$$

I went into the forest and measured the height of 100 plants and the pH of the soil.

I want to know if soil Ph predicts leaf size.

What would my model look like?

```
Height ~ pH

lm
```

I forgot to say, I measured my 100 plants at 20 plots, 5 plants each.

I still want to know if soil Ph predicts leaf size.

What would my model look like?

I forgot to say, I measured my 100 plants at 20 plots, 5 plants each.

I still want to know if soil Ph predicts leaf size.

What would my model look like?

$$Height \sim pH + (1|Plot)$$

I forgot to say, I measured my 100 plants at 20 plots, 5 plants each.

I still want to know if soil Ph predicts leaf size.

What would my model look like?

$$Height \sim pH + (1|Plot)$$

$$lmm$$

Question 2

I collected data at four different lakes. At each lake, I took 30 samples of dissolved organic matter (DOM), and water temperature. I want to know if there is a correlation between DOM and temperature, and if different lakes have different levels of DOM.

I collected data at four different lakes. At each lake, I took 30 samples of dissolved organic matter (DOM), and water temperature. I want to know if there is a correlation between DOM and temperature, and if different lakes have different levels of DOM.

DOM ~ Temperature + Lake

lm

I forgot to say, at each lake I went out once a day for 15 days, and each day I took 30 samples. Should I adjust my model to account for this?

DOM ~ Temperature + Lake

I forgot to say, at each lake I went out once a day for 15 days, and each day I took 30 samples. Should I adjust my model to account for this?

$$DOM \sim Temperature + Lake + (1|Day)$$

I forgot to say, at each lake I went out once a day for 15 days, and each day I took 30 samples. Should I adjust my model to account for this?

DOM ~ Temperature + Lake + (1|Day)

lmm

Question 3

I have data on whether or not an area gets protected, and the species richness of that area. They can either get protected (1) or not get protected (0). I think that sites with higher species richness might be more likely to get protected. How would I test this? (hint – yes no data has a binomial distribution)

I have data on whether or not an area gets protected, and the species richness of that area. They can either get protected (1) or not get protected (0). I think that sites with higher species richness might be more likely to get protected. How would I test this? (hint – yes no data has a binomial distribution)

Protected(YesOrNo) ~ SpeciesRichness, family = “binomial”

I have data on whether or not an area gets protected, and the species richness of that area. They can either get protected (1) or not get protected (0). I think that sites with higher species richness might be more likely to get protected. How would I test this? (hint – yes no data has a binomial distribution)

Protected(YesOrNo) ~ SpeciesRichness, family = "binomial"

glm

Question 4

I have a large dataset of waterbird count each year for many years, for many species at many sites. I want to know if waterbirds overall are increasing through time or not. I don't care about responses of individual species or locations.

A) Count ~ Year + (1|Species/Site), family = 'beta'

B) Count ~ Year + (1|Species) + (1|Site), family = 'poisson'

C) Year ~ Count + Species + (1|Site), family = 'gaussian'

D) Count ~ Year + Site + (1|Species), family = 'poisson'

I have a large dataset of waterbird count each year for many years, for many species at many sites. I want to know if waterbirds overall are increasing through time or not. I don't care about responses of individual species or locations.

A) Count ~ Year + (1|Species/Site), family = 'beta'

B) Count ~ Year + (1|Species) + (1|Site), family = 'poisson'

C) Year ~ Count + Species + (1|Site), family = 'gaussian'

D) Count ~ Year + Site + (1|Species), family = 'poisson'

I have a large dataset of waterbird count each year for many years, for many species at many sites. I want to know if waterbirds overall are increasing through time or not. I don't care about responses of individual species or locations.

glmm

A) Count ~ Year + (1|Species/Site), family = 'beta'

B) Count ~ Year + (1|Species) + (1|Site), family = 'poisson'

C) Year ~ Count + Species + (1|Site), family = 'gaussian'

D) Count ~ Year + Site + (1|Species), family = 'poisson'

I've just realised my sites come from 8 countries, and that bird counts might be different on average in different countries. What should I do?

Count ~ Year + (1|Species) + (1|Site), family = 'poisson'

I've just realised my sites come from 8 countries, and that bird counts might be different on average in different countries. What should I do?

Count ~ Year + (1|Species) + (1|Country/Site), family = 'poisson'

Question 5

I have data on how well protected areas are managed, expressed as a proportion. (1 = perfectly managed, 0 = terribly managed). I have this data across many countries. I want to know if the amount of funding a protected area receives predicts how well it is managed. I don't care about country, and I think this relationship will be the same regardless of country. (Hint: proportion data has a 'beta' distribution)

I have data on how well protected areas are managed, expressed as a proportion. (1 = perfectly managed, 0 = terribly managed). I have this data across many countries. I want to know if the amount of funding a protected area receives predicts how well it is managed. I don't care about country, and I think this relationship will be the same regardless of country. (Hint: proportion data has a 'beta' distribution)

ManagementScore ~ Funding + (1|Country), family = "beta"

I have data on how well protected areas are managed, expressed as a proportion. (1 = perfectly managed, 0 = terribly managed). I have this data across many countries. I want to know if the amount of funding a protected area receives predicts how well it is managed. I don't care about country, and I think this relationship will be the same regardless of country. (Hint: proportion data has a 'beta' distribution)

glmm

ManagementScore ~ Funding + (1|Country), family = "beta"

Actually, I think the relationship between funding and management effectiveness WILL be different in different countries. Now what?!

Actually, I think the relationship between funding and management effectiveness WILL be different in different countries. Now what?!

We haven't covered this yet – but the answer is random slopes!! (Or split your data up by country and run separate models per country)

Activity

First – PULL from github, as I've updated some files!!

Then use the paper worksheet, or the Week 7 Worksheet in github

Use the Week_7_In_Class_Activity script to get started, before going through the exercises

Challenge 3

You are a statistical consultant, and we (WWF) have hired you to put together a report from the Living Planet Database

We're interested in how the population trend of a species changes through time.

Take data from the Living Planet Database (link in repo)

4 components…

Challenge 3

1. *Choose a Species* (we want everyone to pick a different one! Bags your species in the Issue on Data Sci hub, if someone else has already used your species pick another)
2. Design your research Q, hypothesis and *fill out a preregistration*
    - There is a template in the "preregistration" folder in the repo. Fill this in, just one or two sentences per question.

Challenge 3

3. *Build and interpret a model (lm, glm, lmm or glmm).* We want you to hear WHY you've chosen your model, why you've chosen your model structure, your understanding of:
    - Meeting model assumptions
    - Model explanatory power
    - Confidence in relationships
    - Strength of relationships
    - Error around your relationships

    Make sure you have ***well commented code*** detailing your work through

Challenge 3

4. *Provide a brief summary report* including (500 words MAX – be succinct!):
   - 1 Question
   - 1-2 Hypotheses (MAX)
   - What model you used and why, incl justification of terms
   - How well your model meets assumptions
   - Statistical summaries (in well formatted tables – we'll cover next week!)
   - Figures of the data and model fit
   - 1-2 sentences interpreting your findings (what does wwf need to know?)

All if this in one markdown file, e.g. the readme.

Challenge 3

Marking Criteria:

1. A well built and justified model, model output is clearly interpreted and communicated to a non-technical audience (34%)
2. Creativity and presentation: clear, professional graph(s), well formatted table, formatting your repo to clearly communicate to WWF (33%)
3. Reproducibility. Clear, logical, critical explanations of workflow. Well commented code and a clearly filled out preregistration document. Data sources properly cited (33%)

Challenge 3

DUE DATE:

Thursday 14th November 12pm