LINEAR AND HIERARCHICAL LINEAR MODELS

Mixed Models, and Multilevel Models, GLMMs, etc…

Camila Pacheco Katrín Björnsdóttir

TODAY

- Linear Models
- What are hierarchical linear models
- Identify situations in which the use of mixed effects is appropriate
- Implement basic linear mixed models (LMM) with R
- Break 
- Practice

REQUIRED MATERIAL

https://github.com/lacapary/BIO503/

REQUIRED MATERIAL

You are required to have downloaded and installed

```
install.packages(c("lme4",
                   "ggeffects",
                   "stargazer"
                   ))
```

REQUIRED MATERIAL

DO NOT HESITATE TO ASK QUESTIONS!

LINEAR MODELS

What is a linear model?

- A mathematical tool used to describe the relationship between a *response variable* and an *explanatory variable*.
- In the simplest form is expressed as:
- $Y = \beta_0 + \beta_1 X + \varepsilon$.

LINEAR MODELS

A simple example:

Does soil nitrogen content influence plant productivity?

Plant biomass = $\beta_0 + \beta_1 \cdot$ Soil nitrogen + ε

LINEAR MODELS

A simple example:

Does soil nitrogen content influence plant productivity?

Plant biomass = $\beta_0 + \beta_1 \cdot$ Soil nitrogen + ε

↑ Response variable

LINEAR MODELS

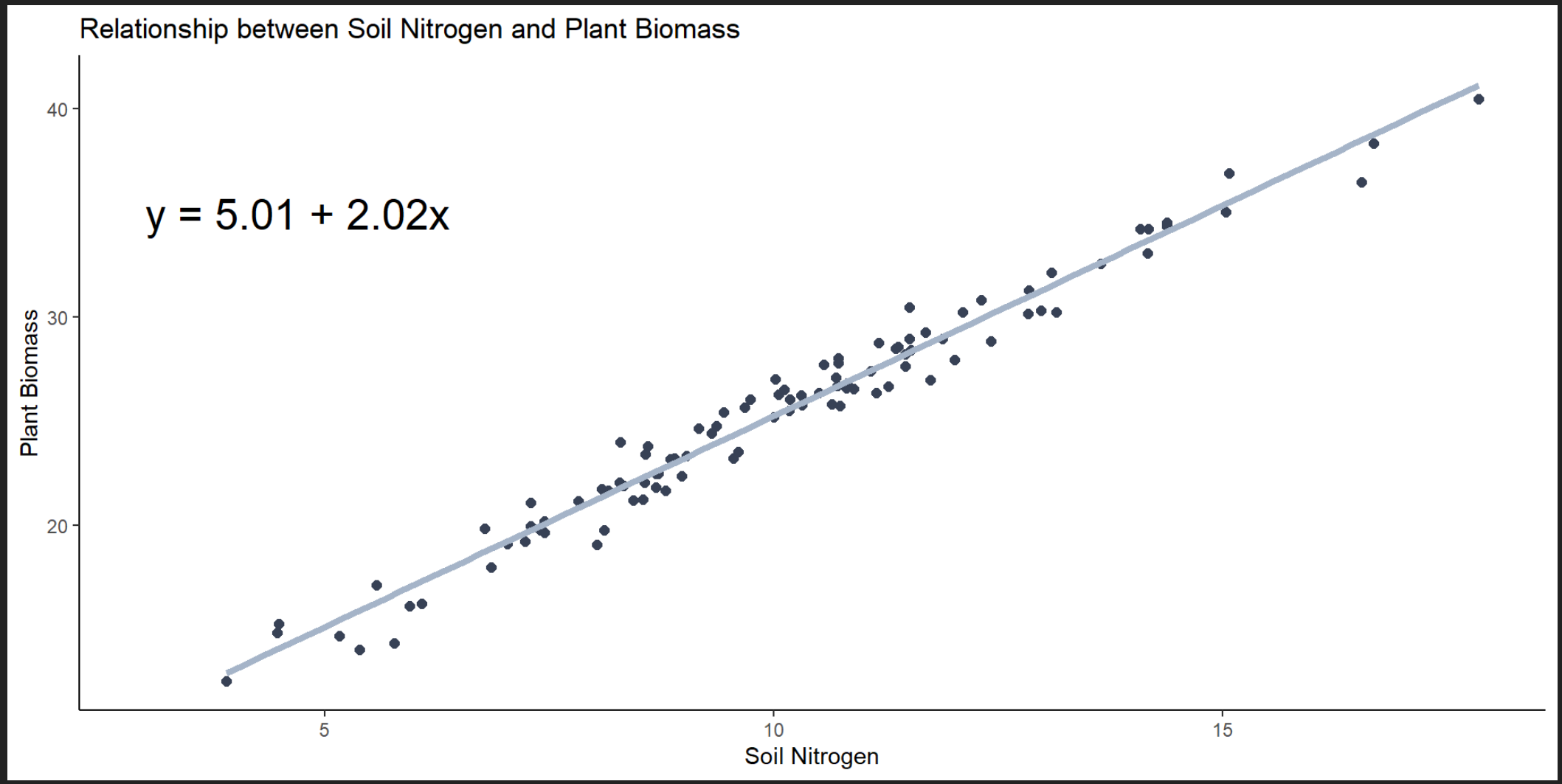
A simple example:

Does soil nitrogen content influence plant productivity?

Plant biomass = $\beta_0 + \beta_1 \cdot$ **Soil nitrogen** $+ \varepsilon$

↑ Explanatory variable

LINEAR MODELS



LINEAR MODELS

Why do we use linear models in ecology?

- Ecological systems are *complex*, but linear models give us a *simplified* framework to ask and answer key ecological questions.
  - ✅ Simple and interpretable: Easy to understand and communicate.
  - ✅ Flexible: Can be extended .
  - ✅ Hypothesis testing: Relationships based on ecological theory.
  - ✅ Help to test ecological hypotheses with real-world data.
  - ✅ Foundation for advanced models

DATA FOR THIS PART OF THE SESSION

Coding Club Mixed models

EXPLORE THE DATA

```
1 library(tidyverse)
2 load("CC-Linear-mixed-models/dragons.RData")
3
4 dragons_Bavarian <- dragons |>
5   filter(mountainRange == "Bavarian")
```

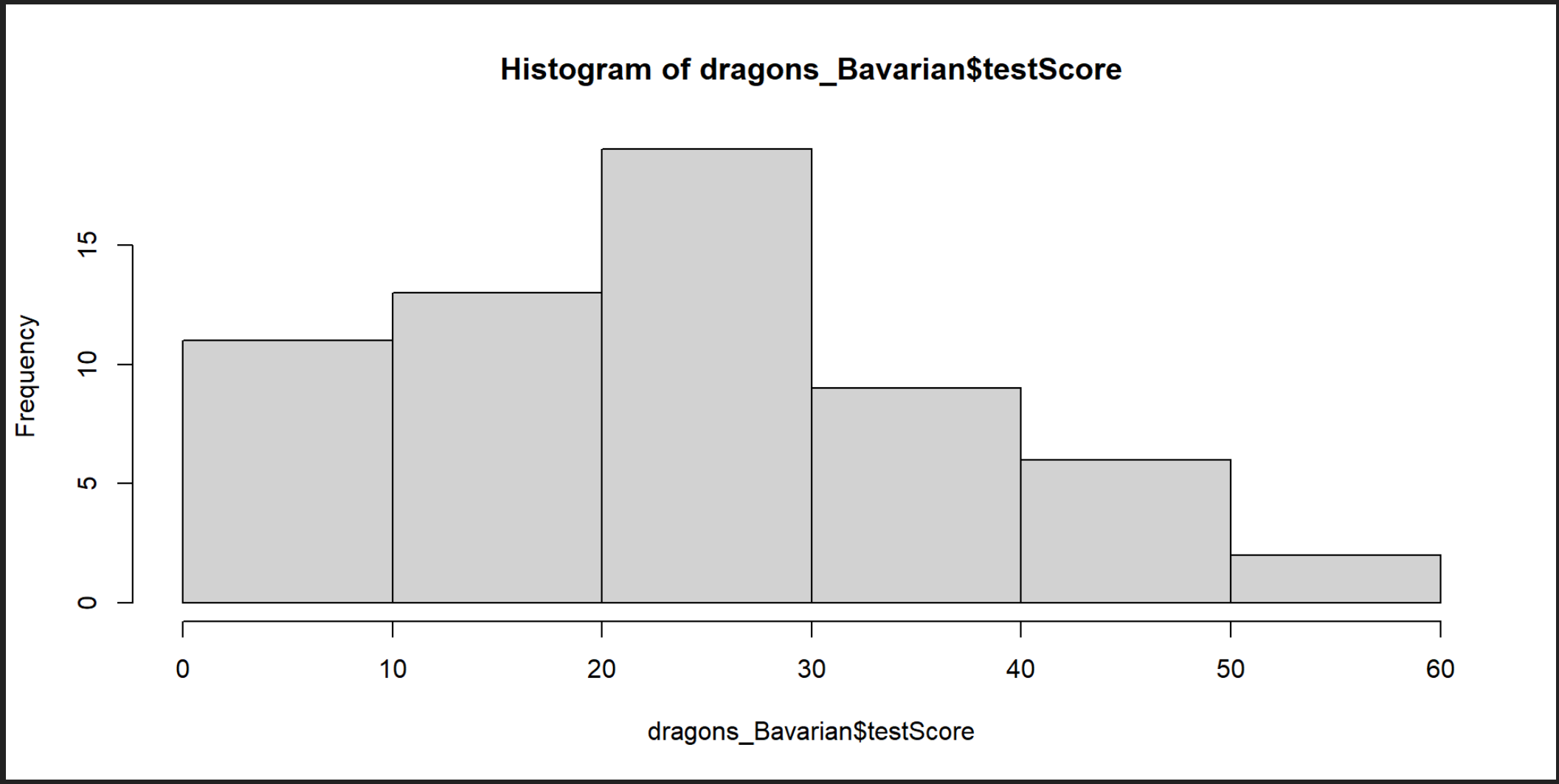
EXPLORE THE DATA

```
1 head(dragons_Bavarian)
```

```
  testScore bodyLength mountainRange site
1 16.147309   165.5485      Bavarian    a
2 33.886183   167.5593      Bavarian    a
3  6.038333   165.8830      Bavarian    a
4 18.838821   167.6855      Bavarian    a
5 33.862328   169.9597      Bavarian    a
6 47.043246   168.6887      Bavarian    a
```

EXPLORE THE DATA

```
1  hist(dragons_Bavarian$testScore)
```



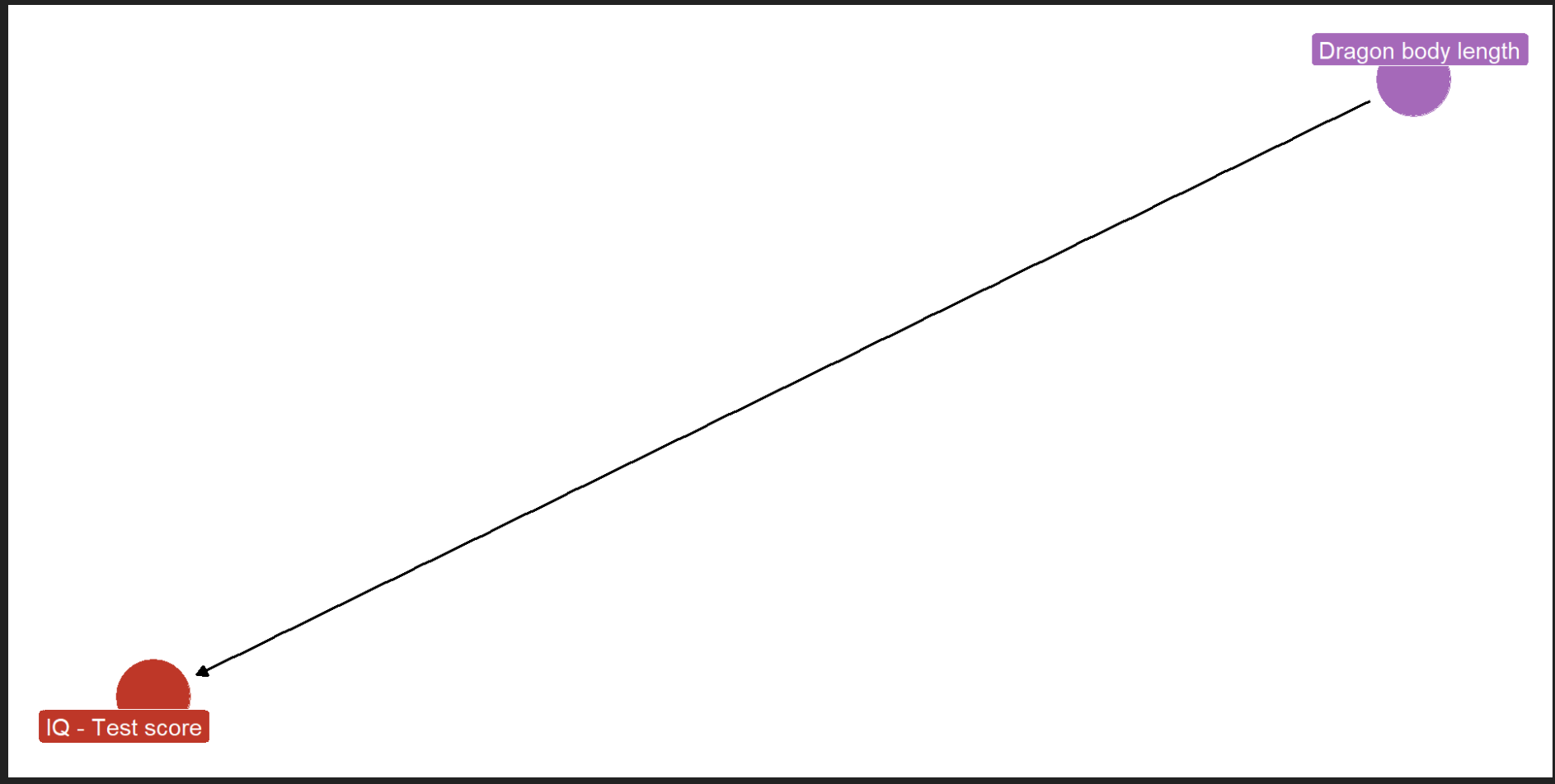
OUR QUESTION

How does body size of dragons influence their intelligence?

OUR QUESTION

How does body size of dragons influence their intelligence?

First we will plot our DAG:



OUR QUESTION

How does body size of dragons influence their intelligence?

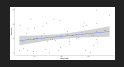
Then our model:

$$\text{Test score} = \beta_0 + \beta_1 \cdot \text{Body length} + \varepsilon$$

OUR QUESTION

Now lets visualize the relationship using ggplot.

```r
(p <- dragons_Bavarian %>%
  ggplot(aes(
    x = bodyLength,
    y = testScore)) + # to assign a color to each group
  geom_point() +
  stat_smooth(method = "lm") +
  theme_classic() +
  labs(
    x = "Body length",
    y = "Test score")
)
```



Note that putting your entire ggplot code in brackets (`()`) creates the graph and then shows it in the plot viewer. If you don't have the brackets, you've only created the object, but haven't visualized it. You would then have to call the object such that it will be displayed by just typing `p` after you've created the "p" object.

LETS MODEL IT

```r
basic.lm <- lm(testScore ~ bodyLength, data = dragons_Ba
summary(basic.lm)
```

```
Call:
lm(formula = testScore ~ bodyLength, data = dragons_Bavarian)

Residuals:
    Min      1Q  Median      3Q     Max
-26.433  -8.228  -0.594   6.893  31.637

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -44.8343    32.3344  -1.387   0.1709
bodyLength    0.3777     0.1795   2.104   0.0397 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.35 on 58 degrees of freedom
Multiple R-squared:  0.07093,   Adjusted R-squared:  0.05491
F-statistic: 4.428 on 1 and 58 DF,  p-value: 0.0397
```

FACTORIAL ANOVA: IT'S JUST A LINEAR MODEL

- *Factorial ANOVA* is a special kind of analysis used when you have two or more categorical predictors.
- But in R, you don't need a special ANOVA function. You can just use the `lm()` function.
- That's because factorial ANOVA is just a *linear model* with categorical predictors.

FACTORIAL ANOVA: EXAMPLE

Using the `iris` data set.

This fits a linear model where Species is treated as a factor.

```r
model <- lm(Sepal.Length ~ Species, data = iris)
summary(model)
```

```
Call:
lm(formula = Sepal.Length ~ Species, data = iris)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6880 -0.3285 -0.0060  0.3120  1.3120

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         5.0060     0.0728  68.762  < 2e-16 ***
Speciesversicolor   0.9300     0.1030   9.033 8.77e-16 ***
Speciesvirginica    1.5820     0.1030  15.366  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5148 on 147 degrees of freedom
Multiple R-squared:  0.6187,    Adjusted R-squared:  0.6135
F-statistic: 119.3 on 2 and 147 DF,  p-value: < 2.2e-16
```

FACTORIAL ANOVA: EXAMPLE

When we run the ANOVA, we're checking if the average sepal length is about the same for all the species, or if at least one species is different from the others.

```
1  anova(model)
```

```
Analysis of Variance Table

Response: Sepal.Length
           Df Sum Sq Mean Sq F value    Pr(>F)
Species     2 63.212  31.606  119.26 < 2.2e-16 ***
Residuals 147 38.956   0.265
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

FACTORIAL ANOVA: POST-HOC ANALYSIS

- Let's say we run an ANOVA and find a significant result.
- The test: "There's a difference somewhere"
- The post-hoc test: "Who's different from who"

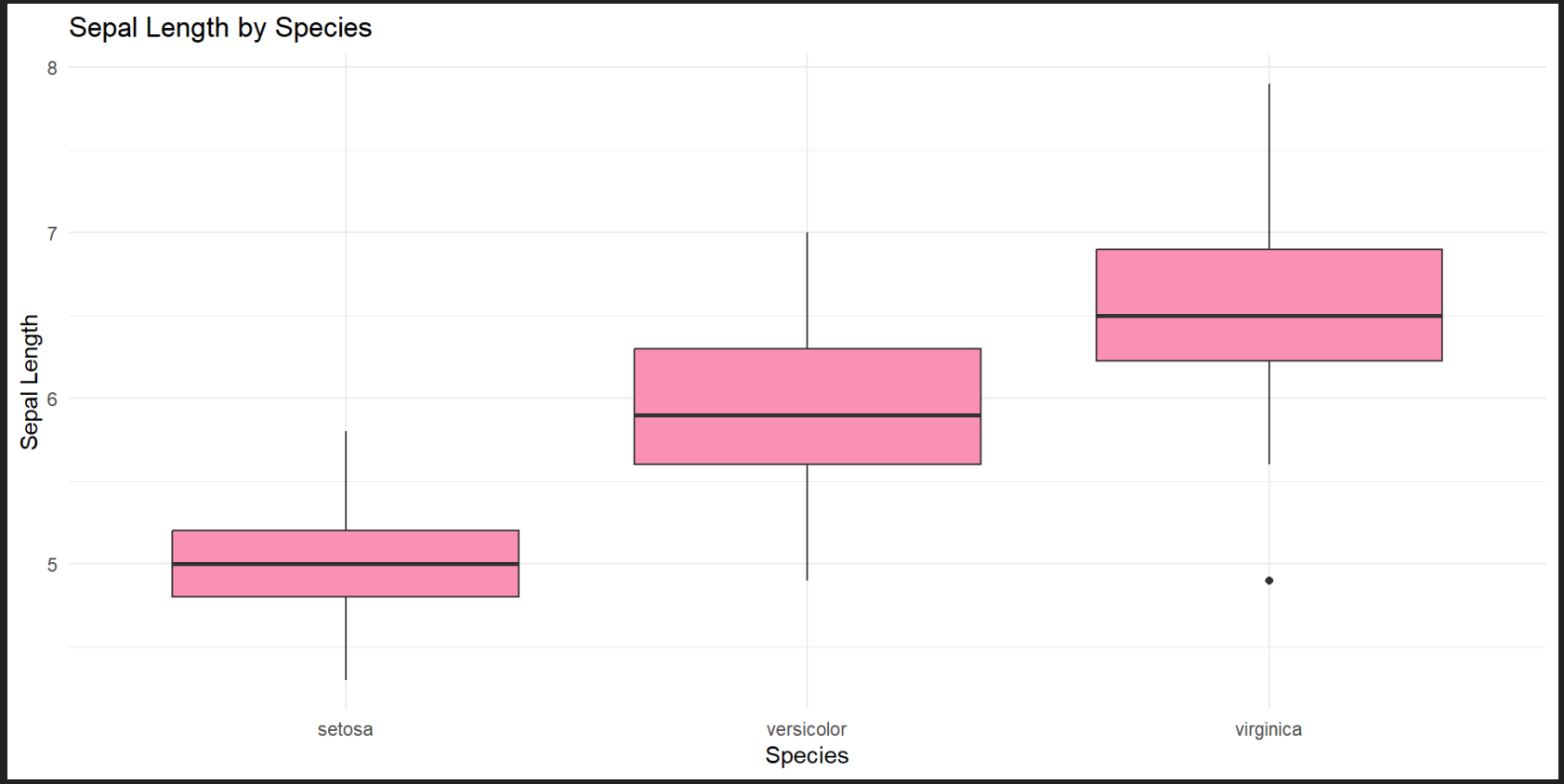
```r
library(emmeans)
# Get estimated marginal means
em <- emmeans(model, ~ Species)

# Pairwise comparisons with Tukey adjustment
pairs(em, adjust = "tukey")
```

```
 contrast              estimate    SE  df t.ratio p.value
 setosa - versicolor     -0.930 0.103 147  -9.033  <.0001
 setosa - virginica      -1.582 0.103 147 -15.366  <.0001
 versicolor - virginica  -0.652 0.103 147  -6.333  <.0001

P value adjustment: tukey method for comparing a family of 3 estimates
```

FACTORIAL ANOVA: POST-HOC ANALYSIS

Important

Use post-hoc tests only if your ANOVA result is significant.



GENERALIZE LINEAR MODELS

Get familiar with different data distributions Here is a brief summary of the data distributions you might encounter most often.

- Gaussian - Continuous data (normal distribution and homoscedasticity assumed)
- Poisson - Count abundance data (integer values, zero-inflated data, left-skewed data)
- Binomial - Binary variables (TRUE/FALSE, 0/1, presence/absence data)

Choosing the right statistical test for your analysis is an important step about which you should think carefully.

GENERALIZE LINEAR MODELS

(a non-exhaustive list of…)

COMMON DATA TYPES IN ENVIRONMENTAL SCIENCES AND THEIR ASSOCIATED DISTRIBUTIONS & TESTS

DATA TYPE	DISTRIBUTION: TESTS	EXAMPLES
Continous data	Gaussian (normal): lm, mixed-effects models	Traits like height, weight, nutrient content, and anything else you can measure across a range centered on a mean value
Count data (whole numbers)	Poisson: glm, glmm	Population counts, number of species in an area, of youngs in a litter, and other discrete measures
Proportion data	If 2 possible outcomes, Binomial: glm, glmm	Survival, seed germination trials, other events that can be described as “successes” and “failures”
	If more outcomes: chi-squared test	Habitat selection (does a species utilise a type of habitat in greater proportion than its availability?), differences in vegetation types between sites or over time

GENERALIZE LINEAR MODELS

Syntax for a model with poisson (count) distribution

```
1  glm(y ~ x, family = poisson, data = df)
```

Syntax for a model with binomial distribution

```
1  glm(y ~ x, family = binomial, data = df)
```

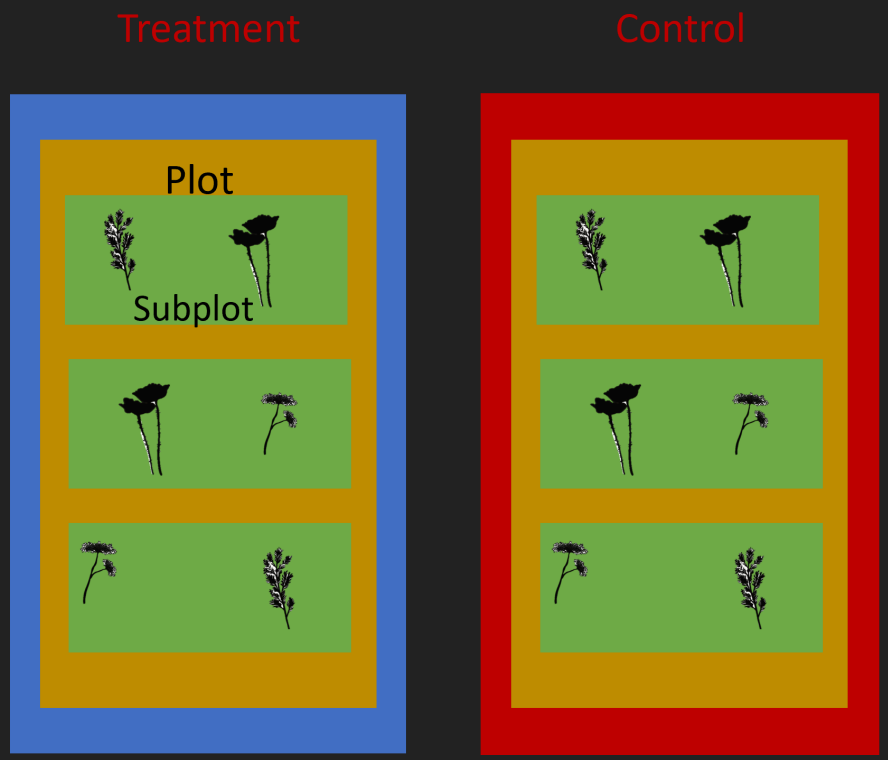
BREAK

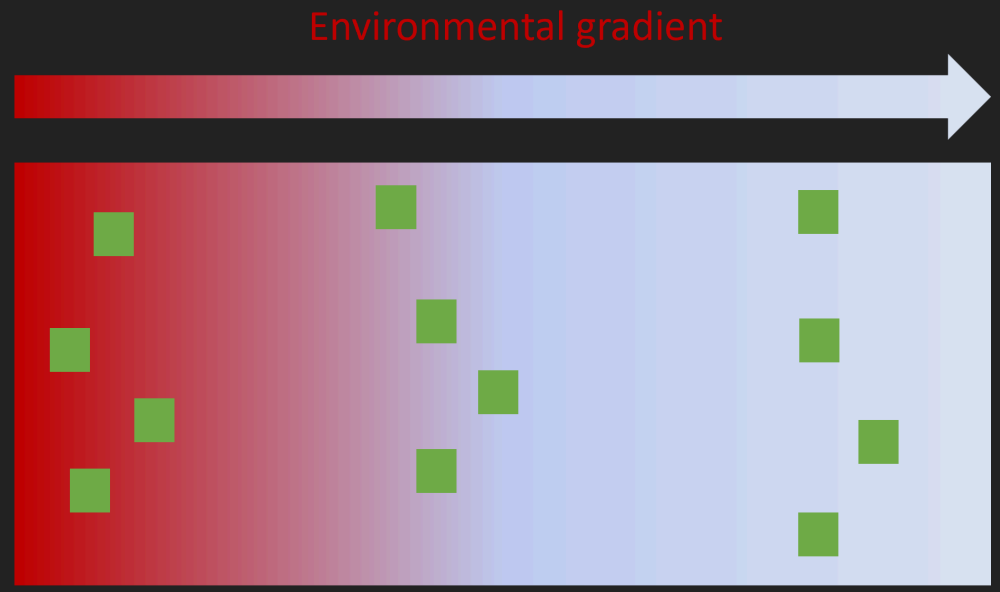
HIERARCHICAL LINEAR MODELS

ECOLOGICAL AND BIOLOGICAL DATA CAN BE COMPLEX!

- Hierarchical structure in the data
- Many covariates and grouping factors
- Unbalanced study/experimental design

HIERARCHICAL STRUCTRES EXAMPLE





WHAT IS INDEPENDENCE ASSUMPTION?

In basic statistical models, we assume each observation gives us **new, independent information**.

What happens in one sample shouldn't depend on what happened in another.

HIERARCHICAL STRUCTRES IN ECOLOGY

- Measurements from the **same plot or site** are often more similar.
- Observations from the **same species** might behave similarly.
- Samples collected **close in time or space** can be correlated.

ANALOGY 💡

> Measuring the same tree's height 3 times isn't the same as measuring 3 different trees.

We call this **dependence** — and we need special models to handle it!

DIAGNOSING DEPENDENCE

Ask yourself:

Are there clusters in my data?

Are data points grouped by site, species, individual, or time?

HOW COULD WE ANALYZE THIS DATA?

We need to account for correlation in the data.

- LINEAR MIXED MODELS

The *random effects* structure will aid correct inference about *fixed effects*, depending on which level of the system's hierarchy is being manipulated.

- What is random effects ?
- What is fixed effects?
- When to used them ?

FIXED EFFECTS : DETERMINISTIC PROCESSES

- They are like drivers, categories or groups that you think might directly influence the outcome you're studying.
- They're called "*fixed*" because you want to estimate the effect of each level

  `"I want to know how these different things (fixed effects)`
- `affect the outcome."`

RANDOM EFFECTS : STOCHASTIC PROCESSES

- Name random doesn’t have much to do with mathematical randomness
- Grouping variables you want to control for
- You’re not interested in each level individually
- but we know that they might be influencing the patterns we see

```
"I want to know how these factors (random effects)
contribute to the overall variation in the outcome."
```

- They are always categorical
- Examples: Site, Observer, Year

WHEN DO WE USE THEM?

- Whenever your data has hierarchical structure.
- The choice depends on the research question
- Is the variable a source of variation or the target of inference?

ACTIVE BREAK 💡

Scenario: You measured plant height across 6 sites, with different treatments at each site.

Questions:

- What could be a fixed effect?
- What could be a random effect?
- What would happen if you ignored site?

DRAGON DATA

We went to the field and sampled more mountains with dragons.

```
# A tibble: 24 × 3
# Groups:   mountainRange [8]
   mountainRange site  samples
   <fct>         <fct>   <int>
 1 Bavarian      a          20
 2 Bavarian      b          20
 3 Bavarian      c          20
 4 Central       a          20
 5 Central       b          20
 6 Central       c          20
 7 Emmental      a          20
 8 Emmental      b          20
 9 Emmental      c          20
10 Julian        a          20
# ℹ 14 more rows
```

RESEARCH QUESTION

IS THE TEST SCORE AFFECTED BY BODY LENGTH?

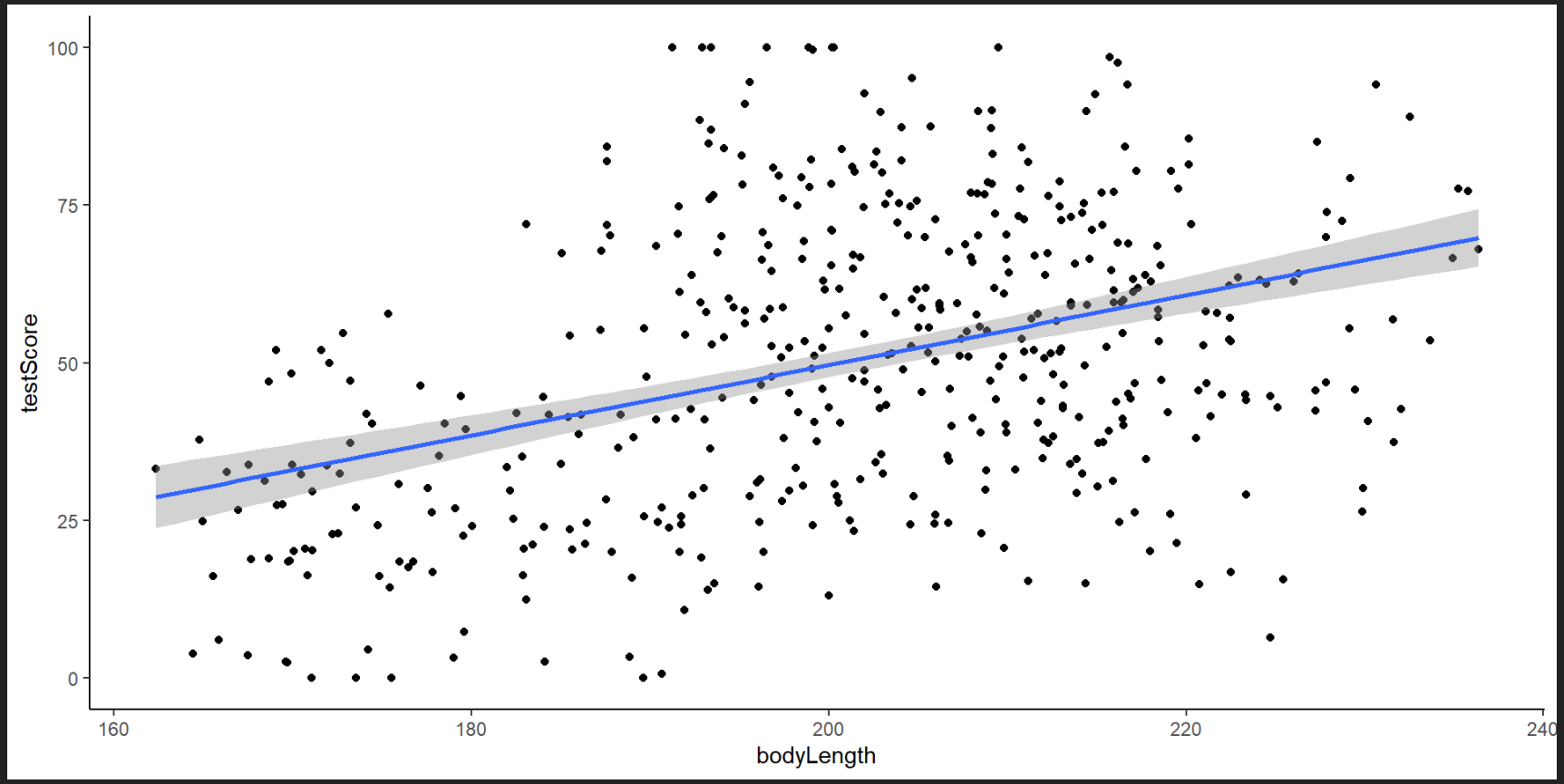
DRAGONS LINEAR MODEL

```r
basic.lm <- lm(testScore ~ bodyLength2, data = dragons)
summary(basic.lm)
```

```
Call:
lm(formula = testScore ~ bodyLength2, data = dragons)

Residuals:
    Min      1Q  Median      3Q     Max
-56.962 -16.411  -0.783  15.193  55.200

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.3860     0.9676  52.072   <2e-16 ***
bodyLength2   8.9956     0.9686   9.287   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.2 on 478 degrees of freedom
Multiple R-squared:  0.1529,    Adjusted R-squared:  0.1511
F-statistic: 86.25 on 1 and 478 DF,  p-value: < 2.2e-16
```

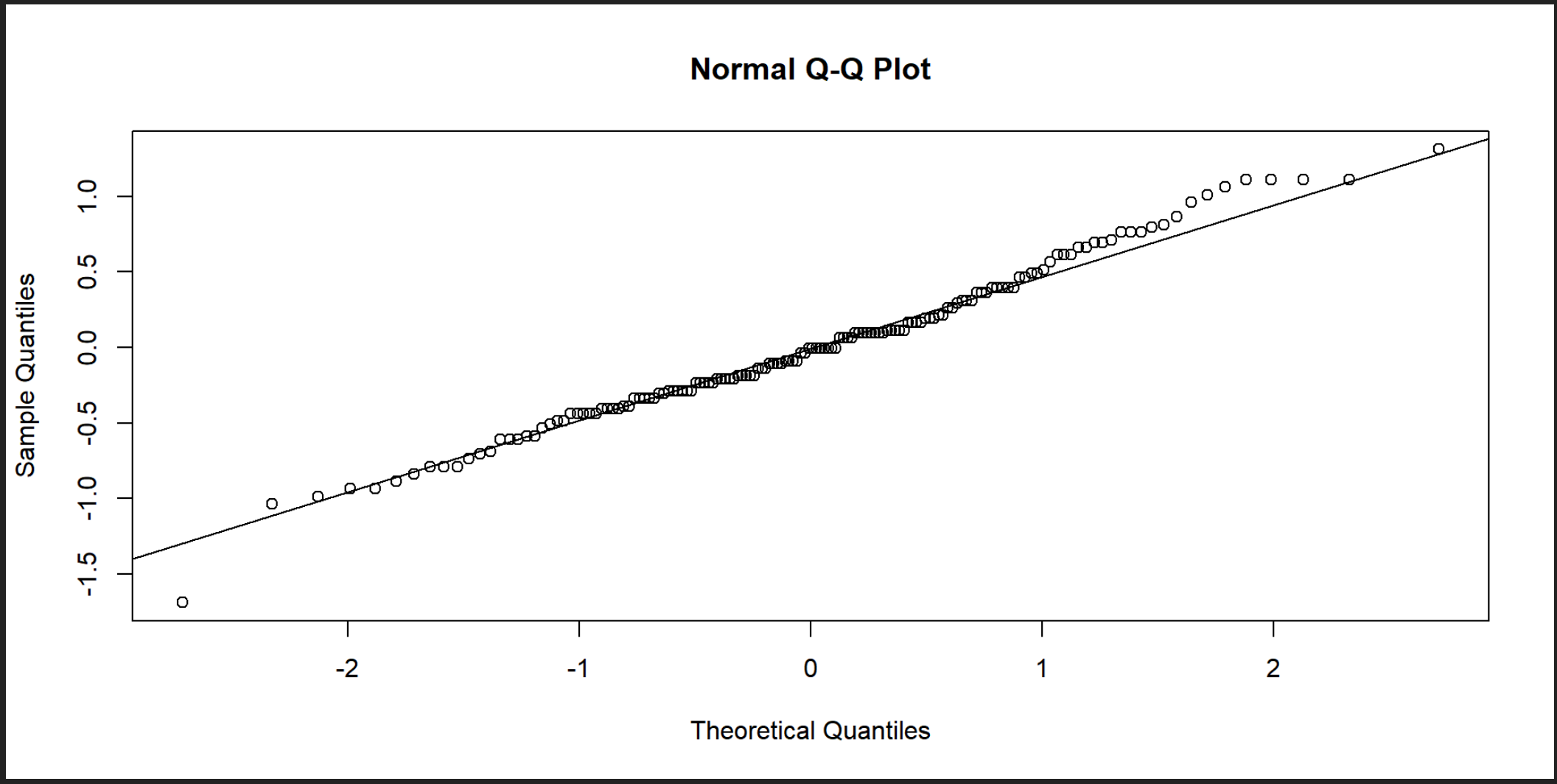
DRAGONS LINEAR MODEL

```
(prelim_plot <- ggplot(dragons, aes(x = bodyLength, y =
  geom_point() +
  geom_smooth(method = "lm"))+
  theme_classic()
```



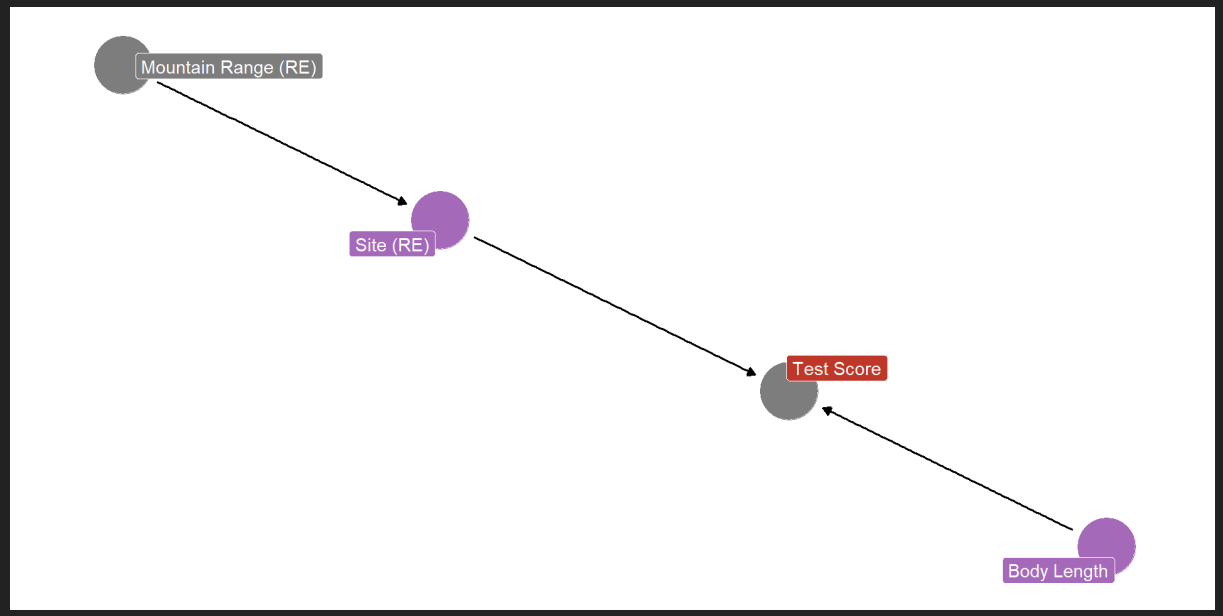
ASSUMPTIONS CHECK

```
1  qqnorm(resid(model)); qqline(resid(model))
```



ASSUMPTIONS CHECK

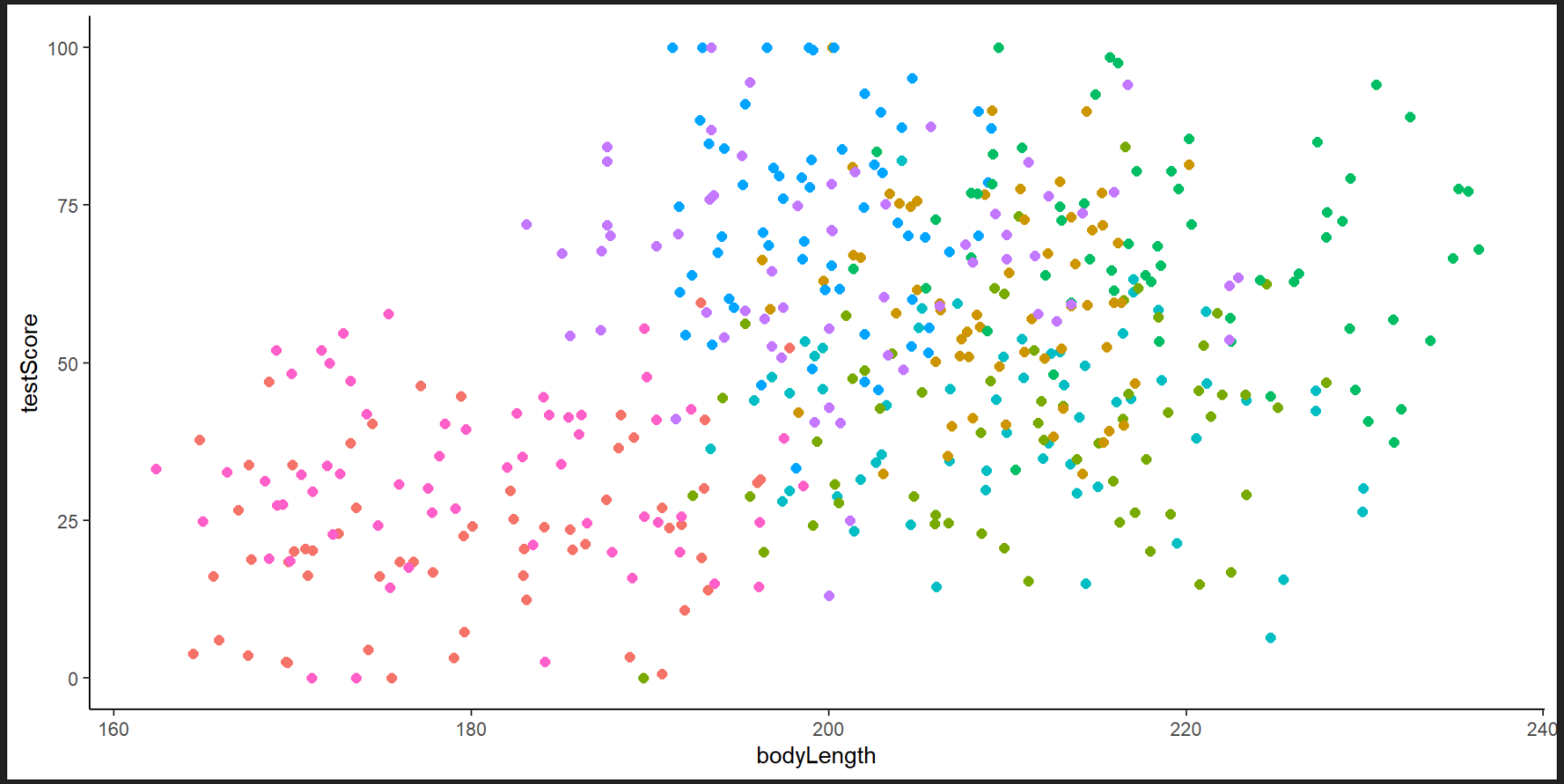
ARE OUR DATA INDEPENDENT?



- They are Hierarchical!
- Our research question:
  - Does a dragon's body length predict its test score, after accounting for variation across mountain ranges?

ASSUMPTIONS CHECK

```
(colour_plot <- ggplot(dragons, aes(x = bodyLength, y =
  geom_point(size = 2) +
  theme_classic() +
  theme(legend.position = "none"))
```



HOW TO IMPLEMENT MIXED MODELS IN R?

- Step 1: Model building
- Step 2: Model validation
- Step 3: Model interpretation
- Step 4: Model visualization

STEP 1: MODEL BUILDING

Hierarchical linear models do is they essentially fit a separate regression line for each and every cluster. And then estimates what we call the *Fixed slope*. Average slope between x and y across my clusters.

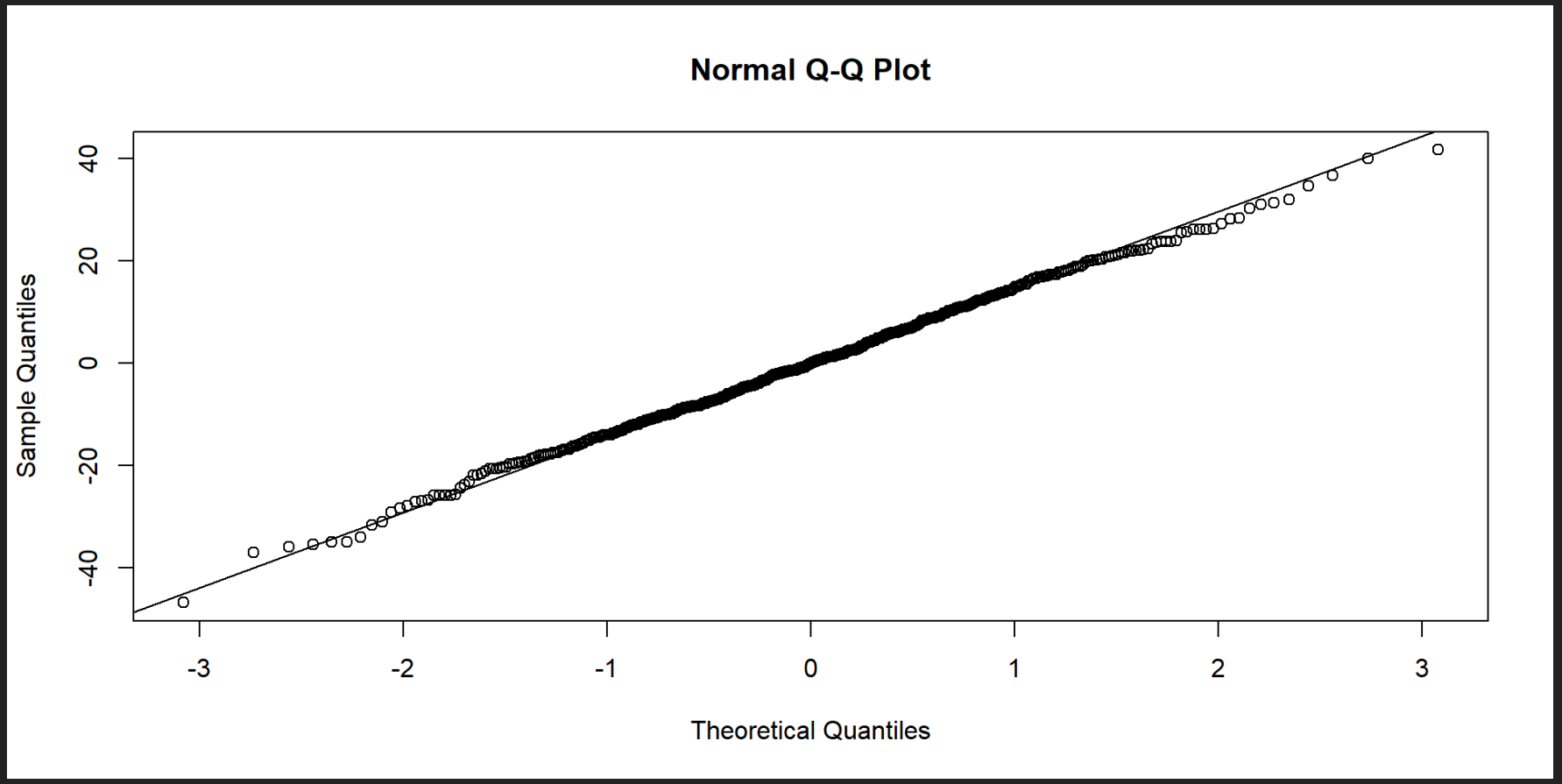
STEP 1: MODEL BUILDING DRAGONS

```
1 library(lme4) #"linear mixed model" function from lme4 p
2 mixed.lmer <- lmer(testScore ~ bodyLength2 +
3                      (1|mountainRange/site), #random eff
4                    data = dragons,
5                    REML = TRUE #estimation method other
6                    )
```

STEP 2: MODEL VALIDATION DRAGONS

Assumptions check

```
1  qqnorm(resid(mixed.lmer))
2  qqline(resid(mixed.lmer))
```



STEP 3: MODEL INTERPREATION DRAGONS

```
Linear mixed model fit by REML ['lmerMod']
Formula: testScore ~ bodyLength2 + (1 | mountainRange/site)
   Data: dragons

REML criterion at convergence: 3970.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2425 -0.6752 -0.0117  0.6974  2.8812

Random effects:
 Groups              Name        Variance Std.Dev.
 site:mountainRange (Intercept)  23.09    4.805
 mountainRange      (Intercept) 327.56   18.099
 Residual                       208.58   14.442
Number of obs: 480, groups:  site:mountainRange, 24; mountainRange, 8

Fixed effects:
            Estimate Std. Error t value
(Intercept)   50.386      6.507   7.743
```

```r
library(MuMIn)
r.squaredGLMM(mixed.lmer)
```

```
             R2m       R2c
[1,] 0.001233396 0.6274854
```

STEP 4. VISUALIZATION

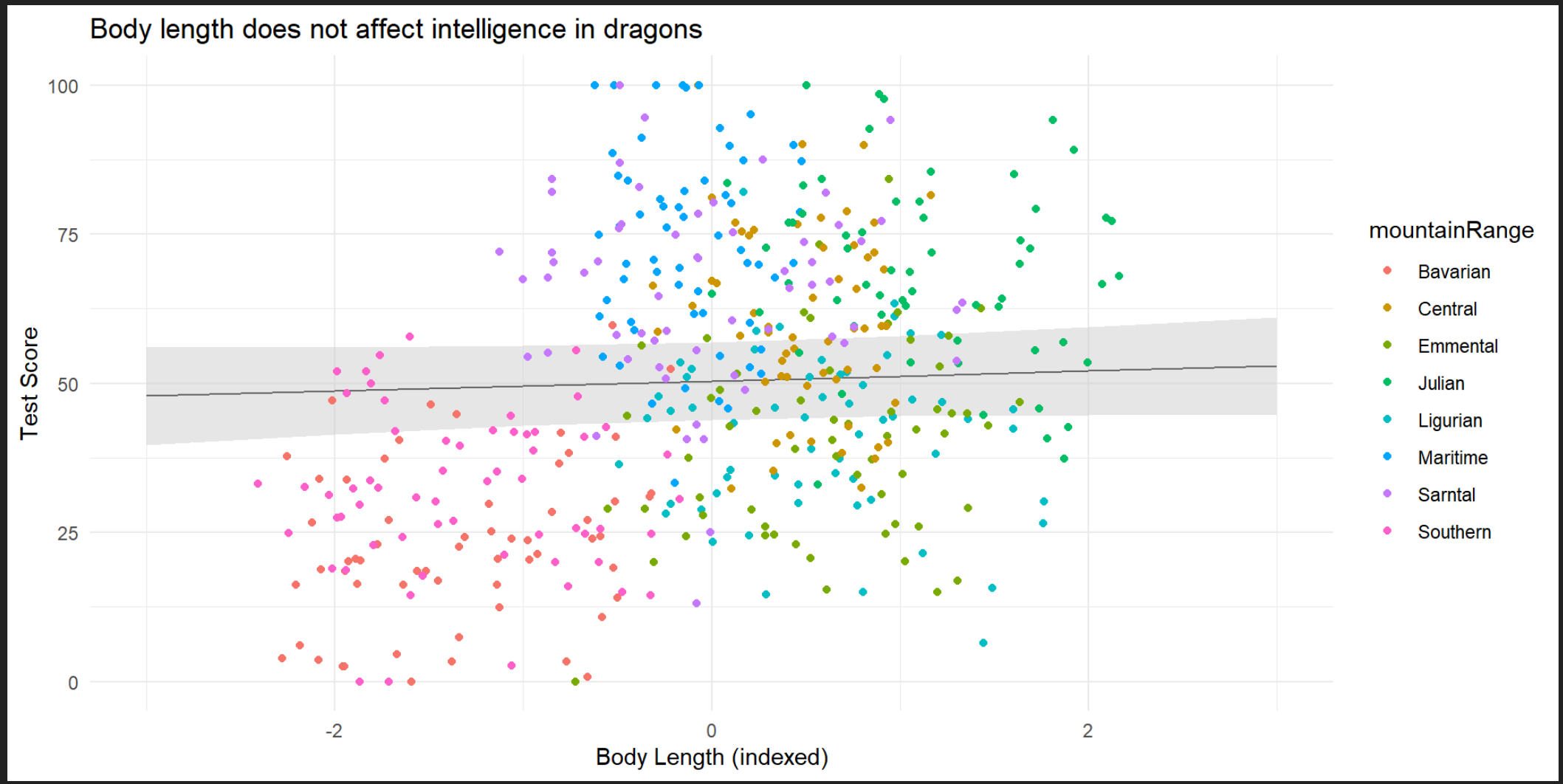
```r
library(ggeffects)

# Extract the prediction data frame
pred.mm <- ggpredict(mixed.lmer, terms = c("bodyLength2"
head(pred.mm)
```

```
# Predicted values of testScore

bodyLength2 | Predicted |       95% CI
-------------------------------------
          -3 |     47.89 | 31.72, 64.07
          -2 |     48.72 | 34.33, 63.12
          -1 |     49.56 | 36.35, 62.76
           0 |     50.39 | 37.60, 63.17
           1 |     51.22 | 38.01, 64.42
           2 |     52.05 | 37.66, 66.44

Adjusted for:
*          site = 0 (population-level)
* mountainRange = 0 (population-level)
```

STEP 4. VISUALIZATION

```r
# Plot the predictions
p<-ggplot(pred.mm) +
    # slope
    geom_line(aes(x = x, y = predicted)) +
    # error band
    geom_ribbon(
      aes(
        x = x,
        ymin = predicted - std.error,
        ymax = predicted + std.error
      ),
      fill = "lightgrey",
      alpha = 0.5
    ) +
    # adding the raw data (scaled values)
    geom_point(data = dragons,
```

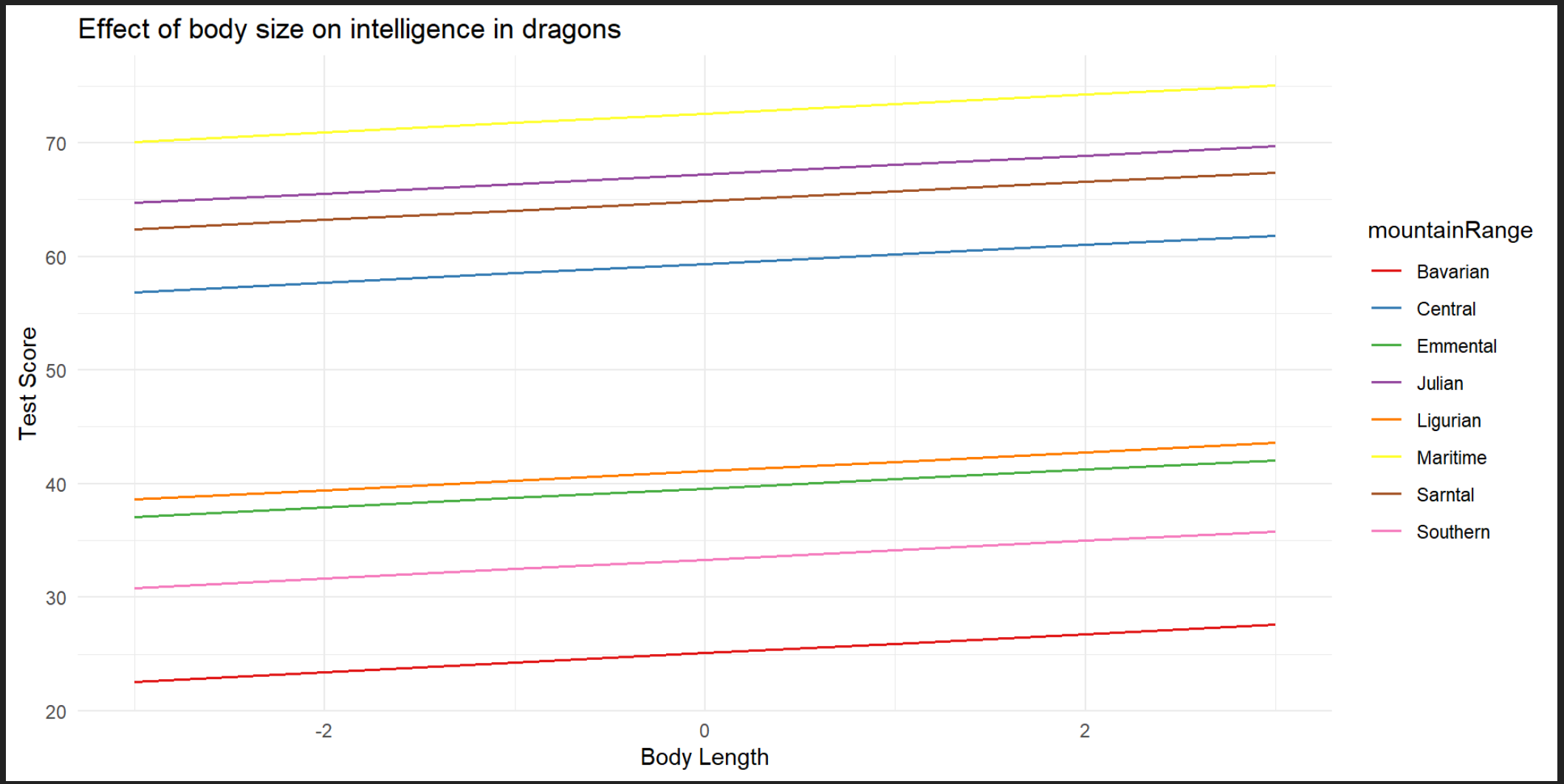
STEP 4. VISUALIZATION



STEP 4. VISUALIZATION

```
p<-
  ggpredict(mixed.lmer, terms = c("bodyLength2", "mounta
            type = "random") %>%
   plot(show_ci = FALSE) +
   labs(x = "Body Length", y = "Test Score",
        title = "Effect of body size on intelligence in
   theme_minimal()
```

STEP 4. VISUALIZATION



RANDOM SLOPES AND RANDOM INTERCEPT

A random-intercept model allows each cluster (e.g. population) to have its own starting point — some may be smarter or dumber overall.

A random-slope model goes further: it lets the effect of a predictor (like body length) vary across clusters. So, one population might show a strong link between body length and intelligence, while another shows little or none.

RANDOM SLOPES AND RANDOM INTERCEPT

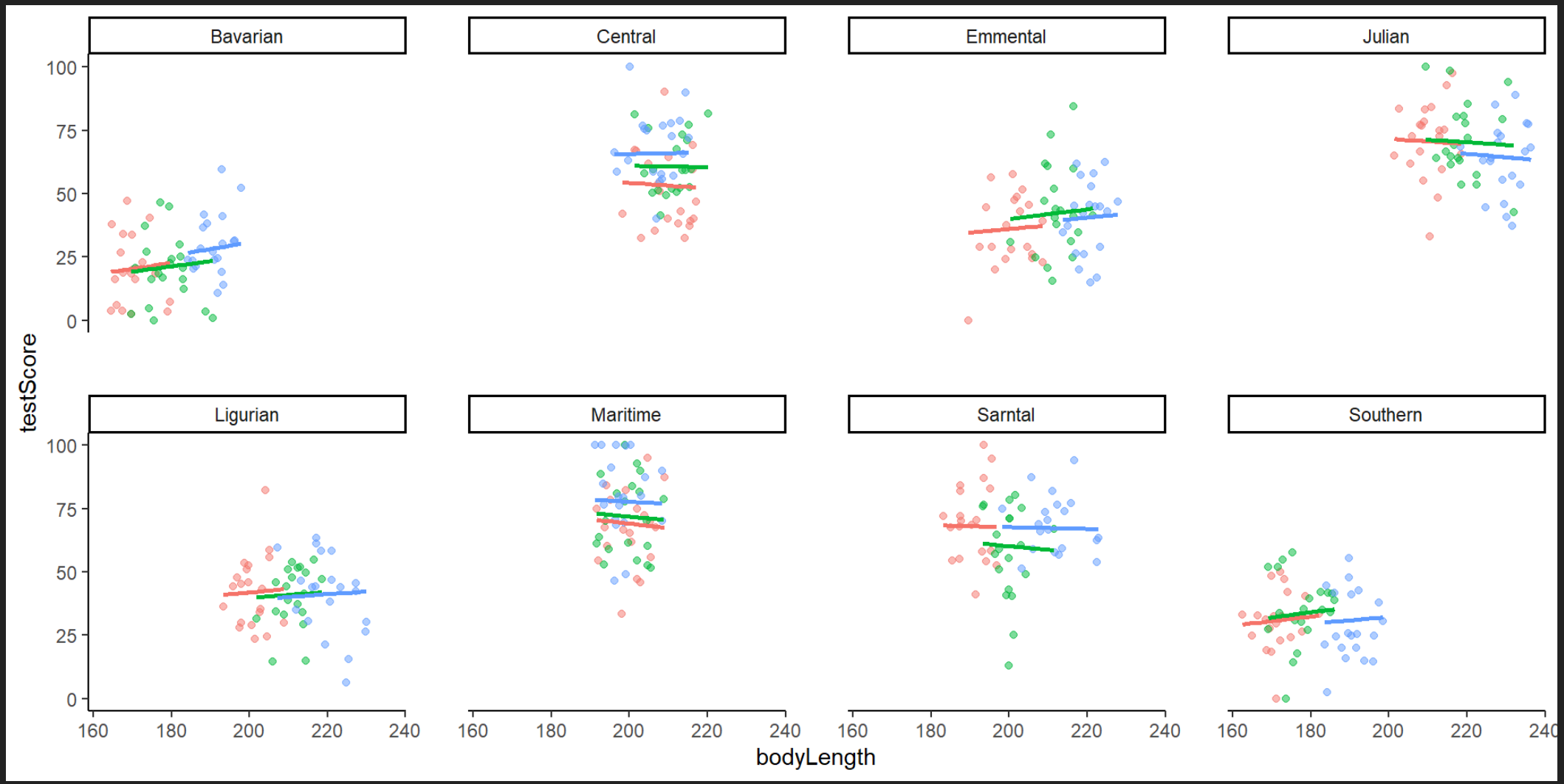
We only need to make one change to our model to allow for random slopes as well as intercept, and that's adding the fixed variable into the random effect brackets:

```
mixed.ranslope <- lmer(testScore ~ bodyLength2 +
                                   (1 + bodyLength2|mountainRange/
                                 data = dragons)

summary(mixed.ranslope)
```

```
Linear mixed model fit by REML ['lmerMod']
Formula: testScore ~ bodyLength2 + (1 + bodyLength2 | mountainRange/site)
   Data: dragons

REML criterion at convergence: 3968.4

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2654 -0.6737 -0.0200  0.6931  2.8432

Random effects:
 Groups              Name        Variance Std.Dev. Corr
 site:mountainRange (Intercept)  19.8156  4.4515
                    bodyLength2   0.7178  0.8472  1.00
```

```
 mountainRange      (Intercept) 310.9691 17.6343
                   bodyLength2   6.1119  2.4722  -1.00
 Residual                      208.5025 14.4396
Number of obs: 480, groups:  site:mountainRange, 24; mountainRange, 8

Fixed effects:
```

RANDOM SLOPES



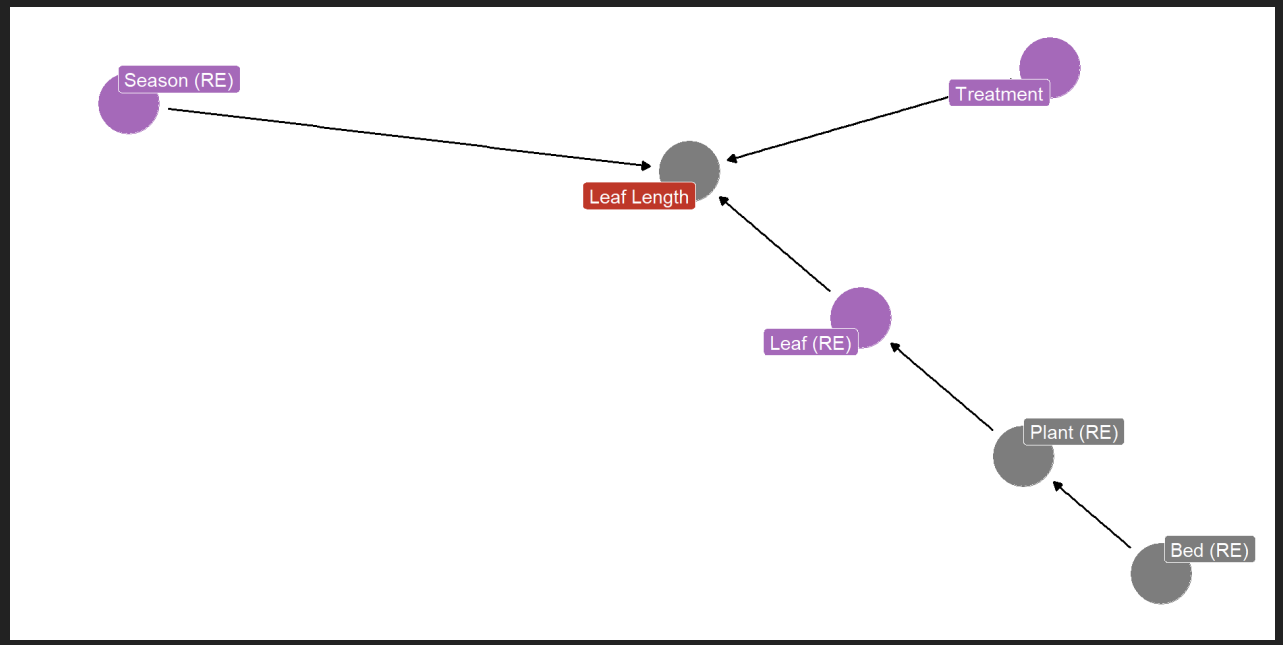
CROSSED EFFECTS

Example:

10 control | 10 experimental

- 3 years
- Each season
- 20 beds
- 50 seedlings
- 5 leaves
- 5 leaves x 50 seedlings x 20 beds x 4 seasons x 3 years = 60 000 measurements per treatment
- *Very hot summer in the second year*

CROSSED EFFECTS




```
1  leafLength ~ treatment + (1|Bed/Plant/Leaf)
```

What about the crossed effects ?

Crossed (or partially crossed) random factors that do not represent levels in a hierarchy.

```
1  leafLength ~ treatment + (1|Bed/Plant/Leaf) + (1|Season)
```

ADDITIONAL RESSOURCES

Popular libraries for (G)LMMs:

- Frequentist : `nlme`, `lme4`, `glmmTMB`
- Bayesian : `brms`, `rstan`, `rstanarm`, `MCMCglmm`
- Nice visualization : link
- Most of the material comes from : Coding Club

BREAK

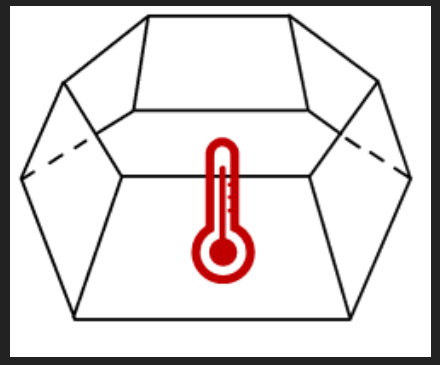
EXERCISE

Now let’s practice …

ITEX

We will be using a dataset from the ITEX network.

- ITEX is a long-term warming experiment that uses standardized protocols to examine impacts of warming on Arctic ecosystems.
- Established in the 1990s - vegetation monitoring over three decades.
- Uses a simple method that is easy to establish in the field - open top chambers



ITEX

LET'S LOOK AT THE DATA

```
1 itex <- read_csv("Data/ITEX_diversity_data.csv")
```

```
1 head(itex)
```

```
# A tibble: 6 × 11
   ...1 SITE      SUBSITE     PLOT   YEAR TRTMT Latitude WarmQuarterTemp SppRich
  <dbl> <chr>     <chr>       <chr> <dbl> <chr>    <dbl>           <dbl>   <dbl>
1     1 ALEXFIORD ALEXFIORD:… Cas.…  2007 CTL       78.9            25.6       9
2     2 ALEXFIORD ALEXFIORD:… Cas.…  2007 CTL       78.9            25.6       9
3     3 ALEXFIORD ALEXFIORD:… Cas.…  2007 CTL       78.9            25.6       7
4     4 ALEXFIORD ALEXFIORD:… Cas.…  2007 CTL       78.9            25.6       8
5     5 ALEXFIORD ALEXFIORD:… Cas.…  2007 CTL       78.9            25.6       6
6     6 ALEXFIORD ALEXFIORD:… Cas.…  2007 CTL       78.9            25.6      10
# ℹ 2 more variables: `row_number()` <dbl>, PlotTemp <dbl>
```

```
1 length(unique(itex$SITE)) #this code tells you how many
```

```
[1] 24
```

```
1 unique(itex$SITE) #you can also do this and then it give
```

```
 [1] "ALEXFIORD"    "ANWR"         "ATQASUK"      "AUDKULUHEIDI" "BARROW"
 [6] "BROOKS"       "BYLOT"        "DOVRE"        "ENDALEN"      "FAROE"
[11] "GAVIA"        "KLUANE"       "KYTALYK"      "LATNJA"       "NIWOT"
[16] "QHI"          "SADVENT"      "STEPSTONES"   "TAISETSU"     "THINGVELLIR"
[21] "TIBET"        "TOOLIK"       "TORNGATS"     "VALBERCLA"
```

EXERCISE 1

Modeling Diversity and Temperature Using the **ITEX diversity dataset** , explore the relationship between **plant diversity** and **temperature** across different sites.

👉 YOUR TASKS:

1. **Model Planning**:
    - What is your **response variable**?
    - Which variables make sense as **fixed effects**?
    - Which variable(s) should be treated as **random effects** (e.g., site, year)?
2. **Fit the Model**:
    - Use `lmer()` to fit a model to the data.
    - Look at the output with `summary()`.
3. **Reflect**:
    - Does the model structure reflect **how the data was collected**?
    - Does the summary output **make sense** given your sampling design?

EXERCISE 2

What is the relationship between diversity and temperature **within sites**?

Now that you've explored overall patterns, focus your model on how **temperature affects diversity within each site**.

👉 **YOUR TASKS:**

1.
   - What changes when we ask about relationships *within sites* instead of *across* them?
   - Should you adjust the way `site` is treated in your model?
2. **Model Check**:
3. **Fit a New Model (if needed)**:
   - Update your `lmer()` call to reflect this new focus.
   - Reinterpret the output