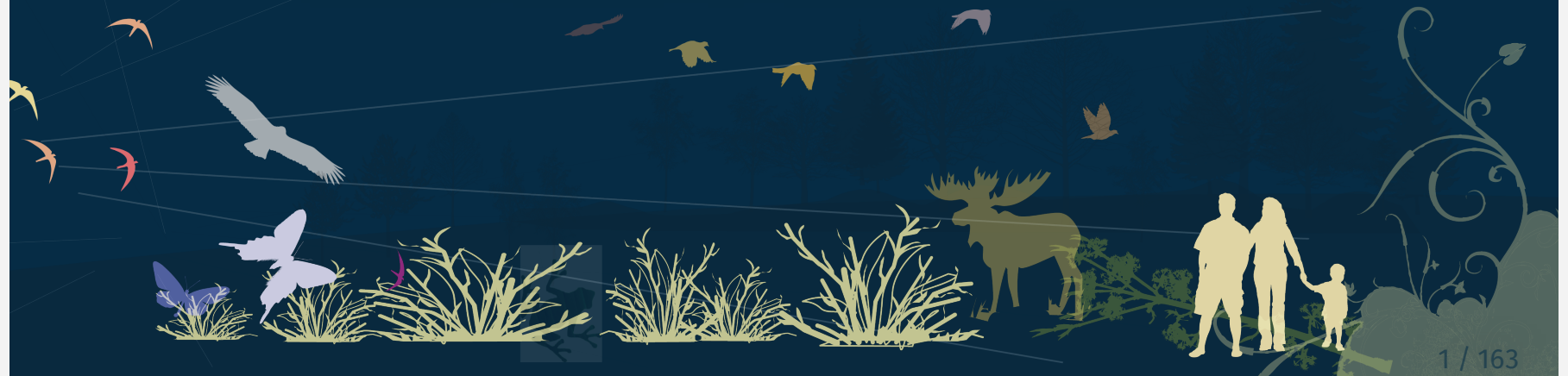
Workshop 4: Linear models

QCBS R Workshop Series

Québec Centre for Biodiversity Science



About this workshop

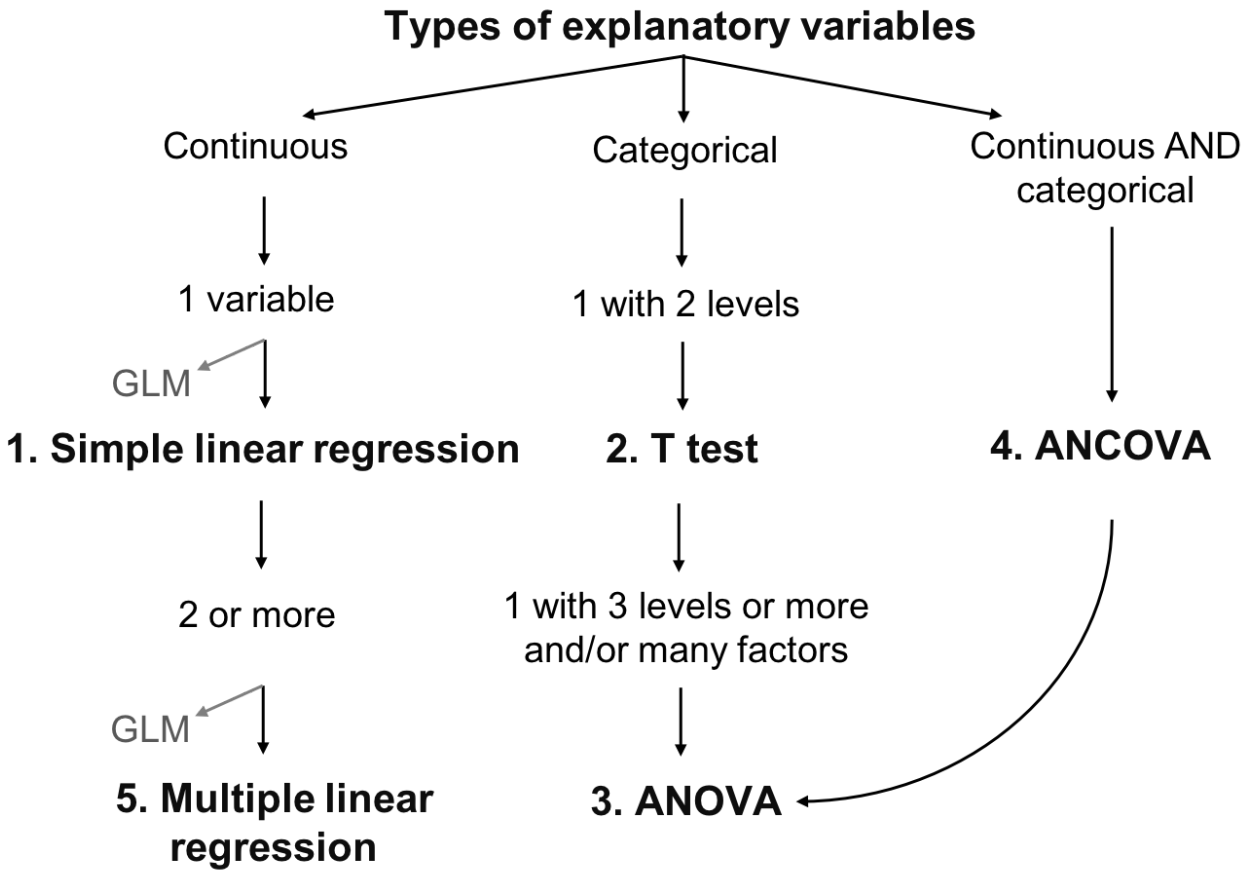
Required packages

- dplyr
- vegan
- e1071
- MASS
- car
- effect

```
install.packages(c('dplyr', 'vegan', 'e1071', 'MASS', 'car', 'effect'))
```

Linear models

-Learn the structure of a linear models and its *different variants*



Learning objectives

- Learn the structure of a linear models and its different variants
- Learn how to perform a linear model with R with `lm()` and `anova()`
- Learn how to identify when assumptions are not met and ways to fix it

What is a linear model?

A **linear model ...**

... describes the relationship between one variable (the **response**) and one or more other variables (the **predictors**).

... is used to investigate a **well-formulated hypothesis**, usually based on a more general research question.

... is used to make inferences about the **direction** and **strength** of a relationship, and our **confidence** in the effect estimates.

Example: Abundance and mass of bird species

Hypothesis

> For bird species, the average mass of an individual has an effect on the maximum abundance of the species, due to ecological constraints (food sources, habitat availability, etc.).

Prediction

> Species with larger individuals have a lower maximum abundance.

Group discussion

Which variable is the response? Which the predictor?

What is the expected direction and strength of the relationship?

Example: Abundance and mass of bird species

Let's have a look at the data ...

Import the `birdsdiet` dataset:

```
bird <- read.csv("birdsdiet.csv", stringsAsFactors = TRUE)
```

Visualize the data using the structure `str()` command:

```
str(bird)
# 'data.frame':    54 obs. of  7 variables:
#  $ Family   : Factor w/ 53 levels "Anhingas","Auks& Puffins",..: 18 25 23 21 2 10
#  $ MaxAbund : num  2.99 37.8 241.4 4.4 4.53 ...
#  $ AvgAbund : num  0.674 4.04 23.105 0.595 2.963 ...
#  $ Mass     : num  716 5.3 35.8 119.4 315.5 ...
#  $ Diet     : Factor w/ 5 levels "Insect","InsectVert",..: 5 1 4 5 2 4 5 1 1 5 ...
#  $ Passerine: int  0 1 1 0 0 0 0 0 0 0 ...
#  $ Aquatic  : int  0 0 0 0 1 1 1 0 1 1 ...
```

Example: Abundance and mass of bird species

Let's have a look at the data ...

Common **measures of location** (central tendency):

- Arithmetic **mean** $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

```
mean(bird$MaxAbund)
# [1] 44.90577
```

- **Median** (value separating higher half from lower half of a sample)

```
median(bird$MaxAbund)
# [1] 24.14682
```

Common **measures of spread** (dispersion):

- **Variance** $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2$

```
var(bird$MaxAbund)
# [1] 5397.675
```

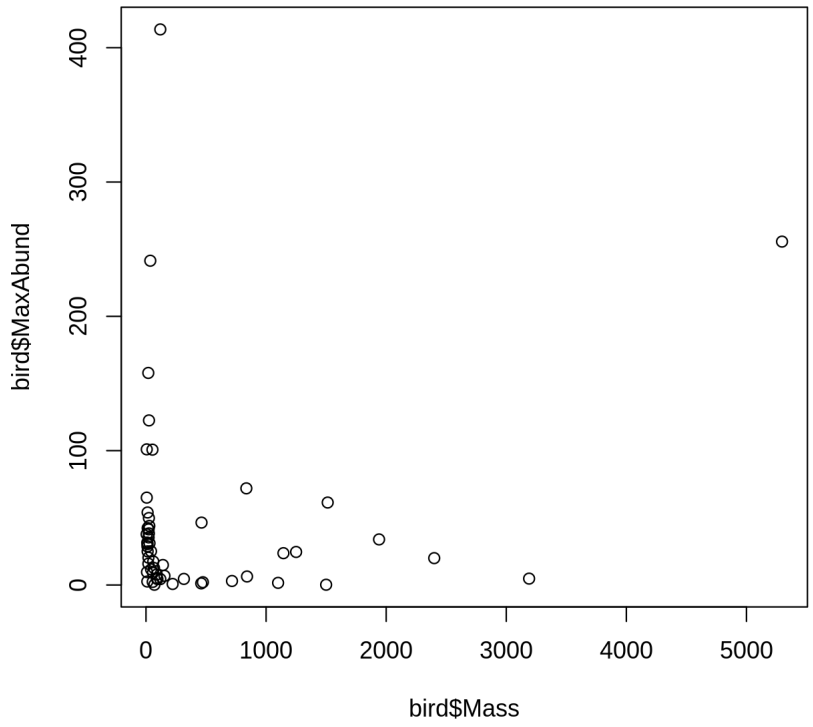
- **Standard deviation** σ

```
sd(bird$MaxAbund)
# [1] 73.46887
```

Example: Abundance and mass of bird species

Plot the response against the predictor:

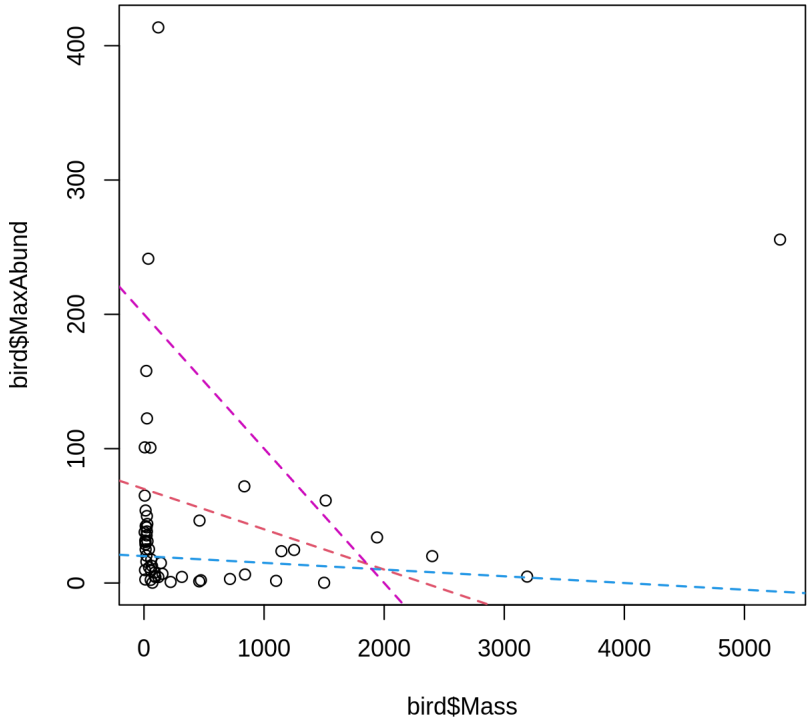
```
plot(bird$Mass, bird$MaxAbund)
```



Example: Abundance and mass of bird species

How do we find the "best" estimate of the relationship?

```
plot(bird$Mass, bird$MaxAbund)
```



Formulation of a linear model

Variables

- y_i is an observation of the **response** y
  (e.g. maximum abundance of species i)
- x_i is a corresponding observation of the **predictor** x
  (e.g. average weight of an individual of species i)

Assumed relationship

$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i$$

- Paramter β_0 is the **intercept**
- Parameter β_1 quantifies the **effect** of x on y
- The residual ϵ_i captures **unexplained** variation
- The **fitted** (or predicted) value of y_i is defined as: $\hat{y}_i = \beta_0 + \beta_1 \times x_i$

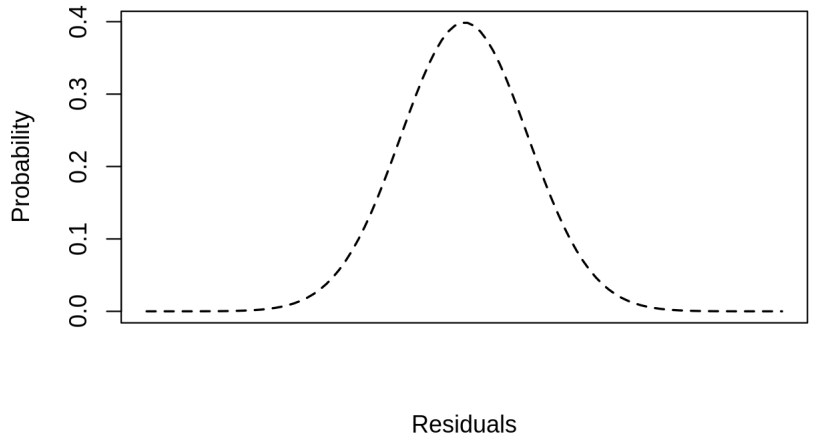
Assumptions of the linear model

$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i$$

Normal distribution

The **residuals** ϵ follow a **normal distribution** with mean 0 and variance σ^2:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$



Assumptions of the linear model

$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i$$

Normal distribution

The **residuals** ϵ follow a **normal distribution** with mean 0 and variance σ^2:

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

This means: Each *single* observation y_i follows a normal distribution, with mean $\hat{y} = \beta_0 + \beta_1 \times x_i$ and variance σ^2:

$$y_i \sim \mathcal{N}(\hat{y}, \sigma^2)$$

This does not mean that the whole set of observed values y ~~must follow a normal distribution~~.

Assumptions of the linear model

$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Homoskedasticity

- All residuals ϵ follow the same distribution, the **variance** σ^2 stays **constant**.

Independence of residuals

- Each residual ϵ_i is **indepedent** from all other residuals.

Assumptions of the linear model

$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Summary of assumptions

- Linear relationship between response and predictor
- Residuals follow a normal distribution with mean 0
- Residuals are identically distributed (*homoscedasticity*)
- Residuals are independent from each other

Notation for linear models

Mathematical notation (for manuscripts)

- Individual observations:
  $y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i \quad \text{with} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$

- All observations (matrix notation, intercept included in $\mathbf{X}$ and $\boldsymbol{\beta}$):
  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon \quad \text{with} \quad \epsilon_i \sim \mathcal{N}(0, I_n\sigma^2)$

R notation

- Model formula:

```
y ~ 1 + x
```

- Or even simpler:

```
y ~ x
```

(also includes intercept)

Never mix different kinds of notation!

Fitting a linear model

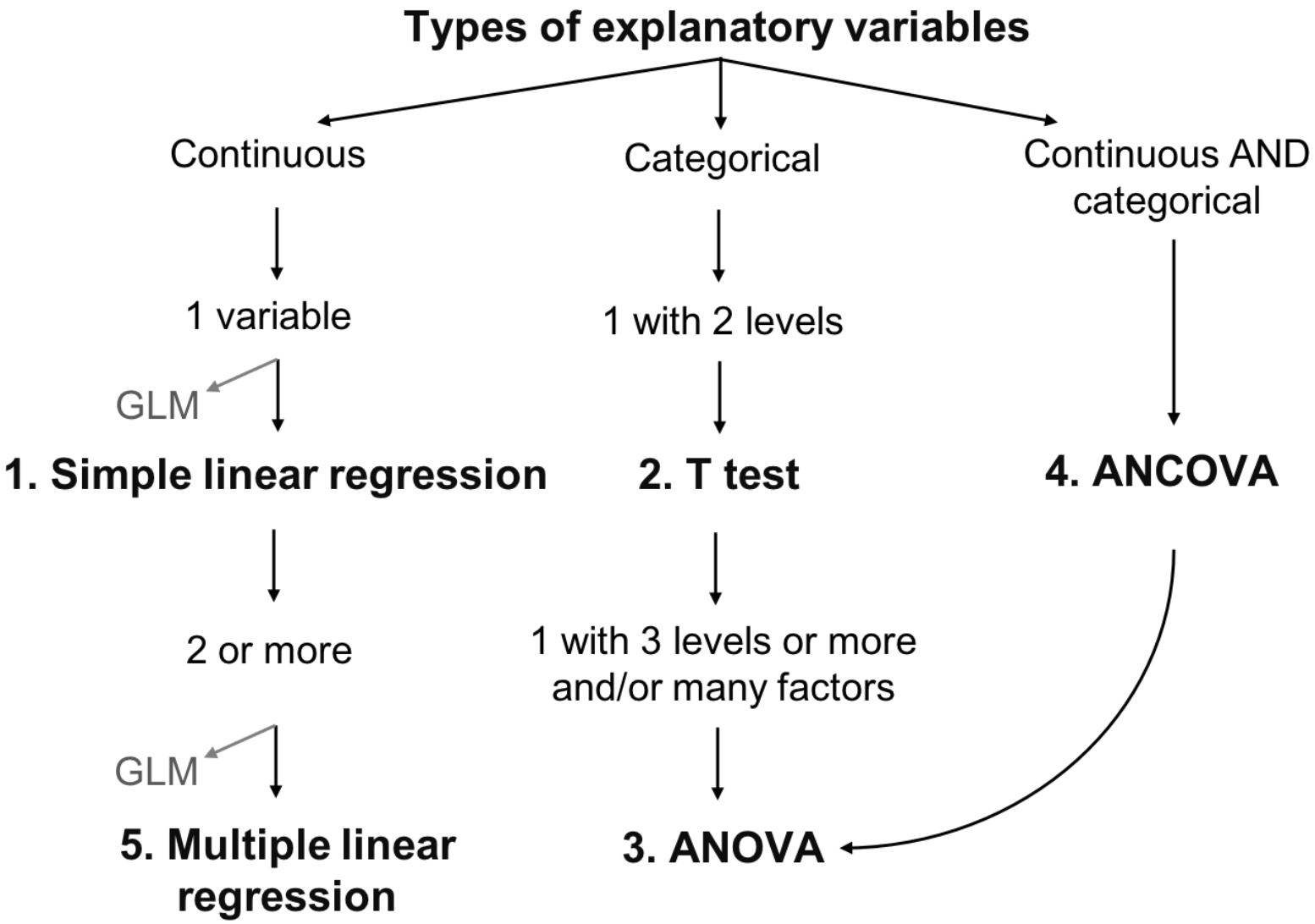
$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Model estimation

- Find the "best" estimates of the parameters β_0, β_1
- The "best" parameters are those, that minimize the sum of the squared residuals $\sum \epsilon_i^2$
- This method is called **ordinary least squares** (OLS)

Learning objectives

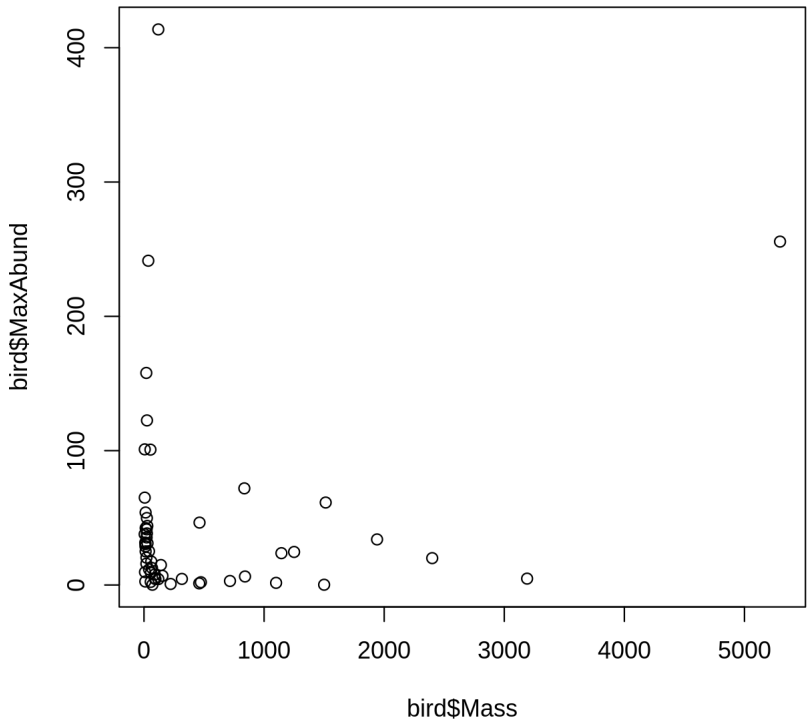


Linear regression in R

Linear regression in R

Back to the birds ...

```
plot(bird$Mass, bird$MaxAbund)
```



Model formulation

> *Hypothesis*: For bird species, the **average size of an individual has an effect on the maximum abundance** of the species, due to ecological constraints (food sources, habitat availability, etc.).

Model equation

$$\text{MaxAbund}_i = \beta_0 + \beta_1 \times \text{Mass}_i + \epsilon_i \ , \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Model formula in R:

```
MaxAbund ~ Mass
```

Linear regression in R

Step 1

Fit a linear model based on a hypothesis

Step 2

Verify assumptions of the linear model

↓

Assumptions are met?

Step 3

- Analyze regression parameters
- Plot your model
- Test for significance of parameter estimates (if necessary)

↓

Assumptions are not met?

Consider using a *Generalized Linear Model* (GLM) or transforming the data

↓

Use a GLM that is better suited for the data

↓

Go back to Step 1 with transformed variables

Linear regression in R

Step 1. Fit a linear model

The function `lm()` is used to fit a linear model, providing an R *model formula* as the first argument:

```
lm1 <- lm(MaxAbund ~ Mass, data = bird)
```

- `lm1` : New object containing the linear model we created
- `MaxAbund ~ Mass` : Model formula
- `bird` : object holding the variables

Linear regression in R

Step 1. Fit a linear model

Let's look at the parameter estimates:

```
lm1
#
# Call:
# lm(formula = MaxAbund ~ Mass, data = bird)
#
# Coefficients:
# (Intercept)         Mass
#    38.16646      0.01439
```

How do the parameters compare to our prediction?

Can we trust these estimates?

Step 2. Verify assumptions using diagnostic plots of the residuals

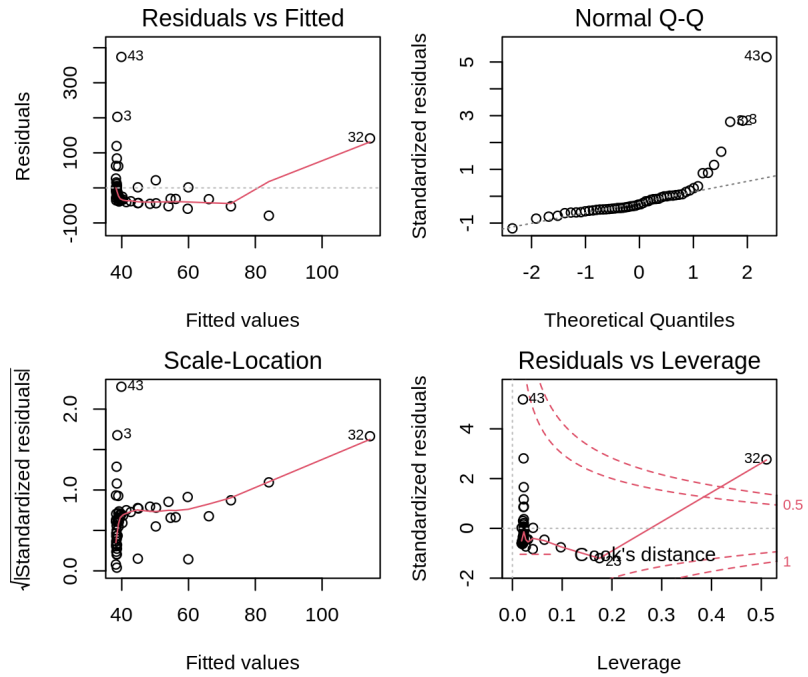
We can produce **four diagnostic plots** of an `lm` object:

```
par(mfrow=c(2,2))
plot(lm1)
```

- `par()` : Function to set graphical parameters
- `mfrow=c(2,2)`: Graphical parameter to display a grid of 2 x 2 at once
- `plot()`: The generic function to plot graphics in R

Step 2. Verify assumptions using diagnostic plots of the residuals

```
par(mfrow=c(2,2)
plot(lm1)
```



How do we interpret these plots?

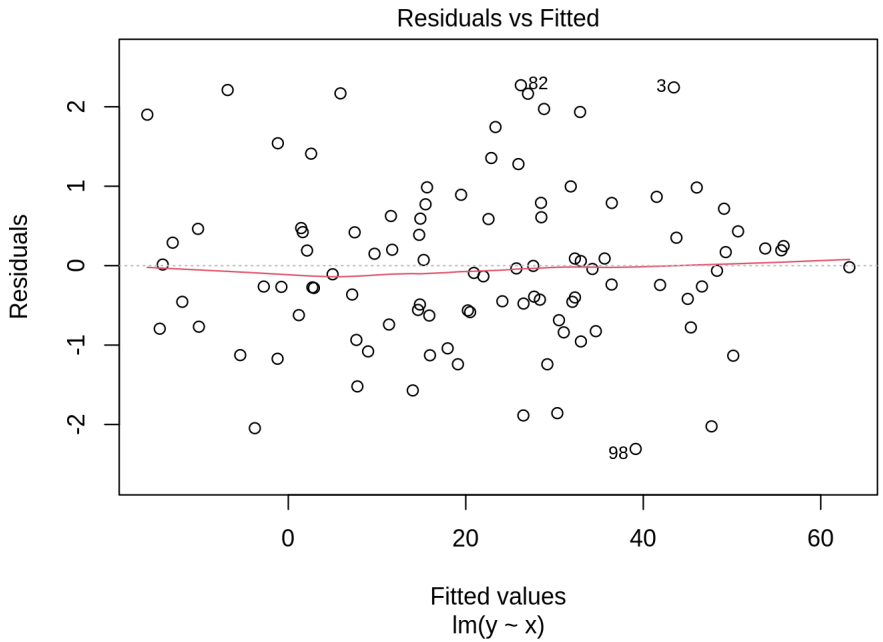
Diagnostic plot # 1 - Residuals vs Fitted

What we see:

- **Y**-axis: Residuals ϵ_i
- **X**-axis: Fitted values $\hat{y}_i = \beta_0 + \beta_1 \times x_i$

What we hope to see: Random scatter, no pattern

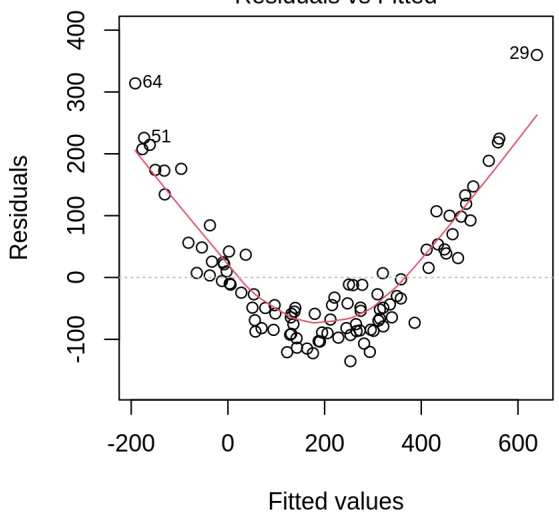
Why: Shows whether residuals are *independent* and *identically distributed*



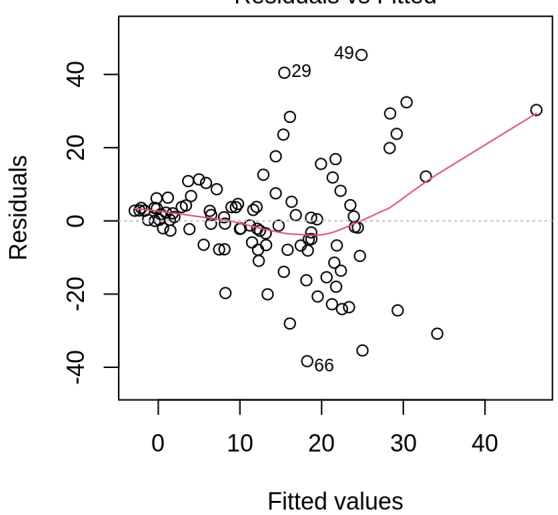
Diagnostic plot # 1 - Residuals vs Fitted

What should make use suspicious:

Nonlinear



Heteroscedastic



What to do:

- Use a **generalized linear model** (GLM) instead that allows for other distributions: Poisson, binomial, negative binomial, etc.)
- Attempt **transformation** of the response and/or predictor variables

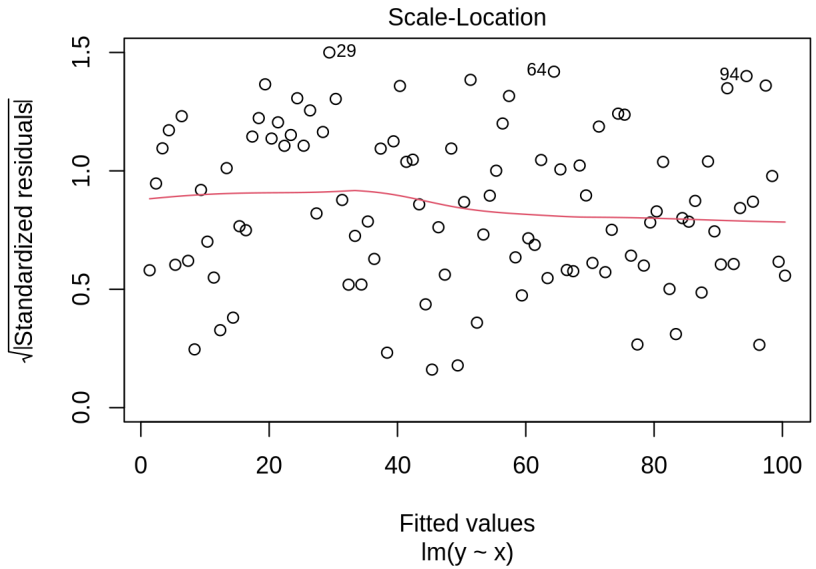
Diagnostic plot # 2 - Scale Location

What we see:

- **Y**-axis: Square root of standardized residuals $\sqrt{\frac{\epsilon_i}{\sigma}}$
- **X**-axis: Fitted values $\hat{y}_i = \beta_0 + \beta_1 \times x_i$

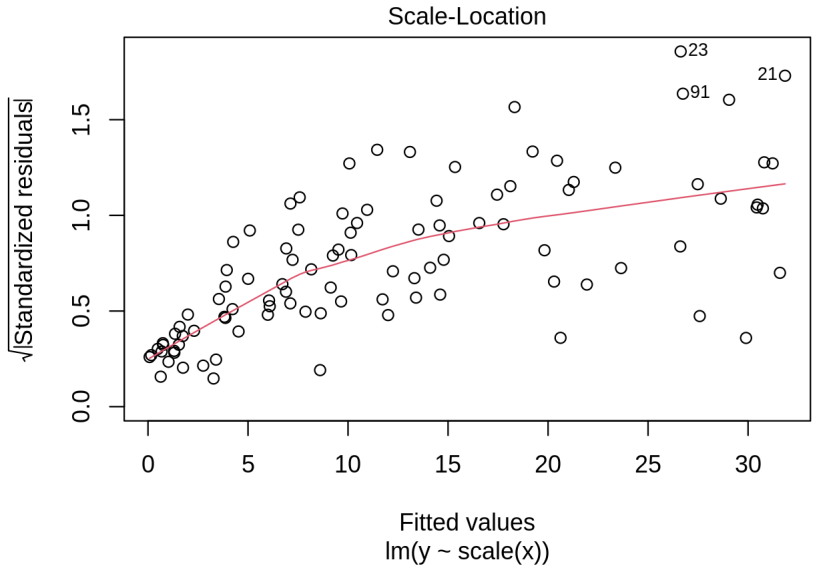
What we hope to see: Random scatter, no pattern

Why: Violations of assumptions are sometimes easier to detect than in the first plot, especially when the predictor is not uniformly distributed.



Diagnostic plot # 2 - Scale Location

What should make use suspicious:



Strong pattern in the residuals

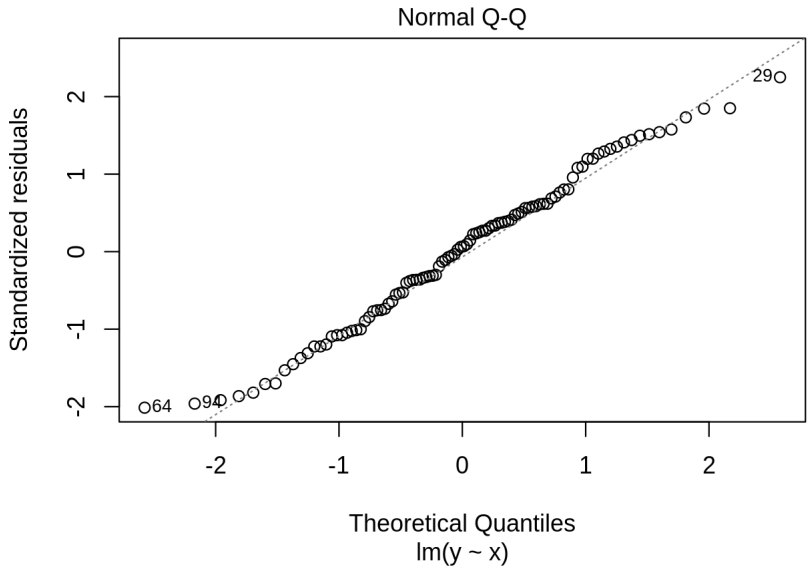
Diagnostic plot # 3 - Normal QQ

What we see:

- **Y**-axis: Standardized residuals $\frac{\epsilon_i}{\sigma}$
- **X**-axis: Standard normal distribution $\mathcal{N}(0, \sigma^2)$

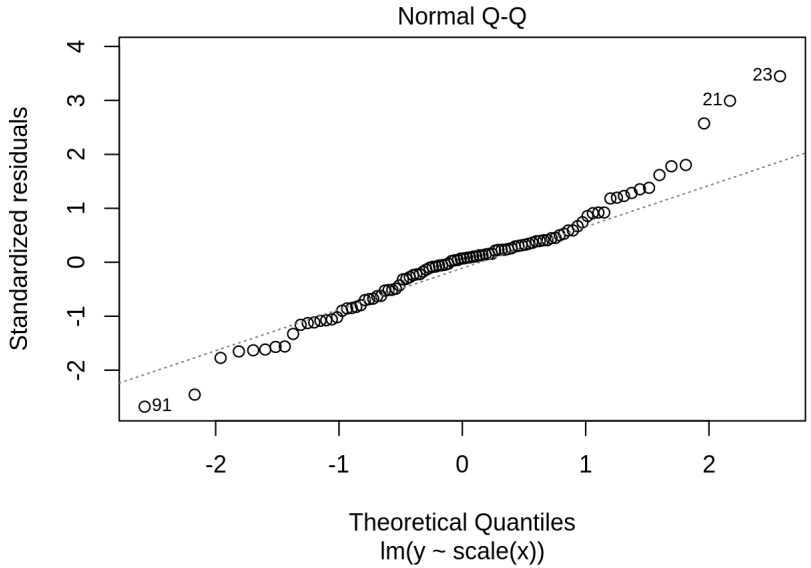
What we hope to see: Points clearly on the 1:1 line

Why: Compares the distribution (quantiles) of the residuals with a standard normal distribution



Diagnostic plot # 3 - Normal QQ

What should make use suspicious:



Residuals do not follow a normal distribution

Diagnostic plot # 4 - Residuals vs Leverage

Why:

- The model should **not depend strongly on single observations**
- **Leverage points** are extreme observations of the predictor.
- The **model passes close to leverage points**, because they lack neighboring observations.
- Leverage points **may or may not have a high influence on the regression**
- Influence can be quantified by **Cook's distance: greater than 0.5 is problematic**.

Examples: Leverage and influence

These plots show the response vs. the predictor.
They are **not** *the diagnostic plots.*

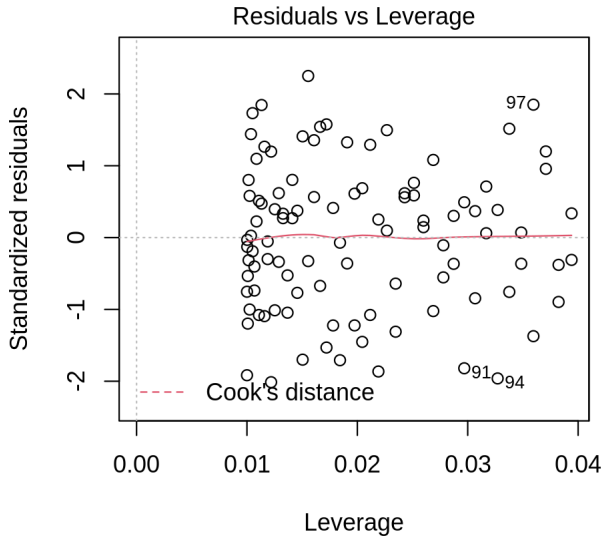
Diagnostic plot # 4 - Residuals vs Leverage

What we see:

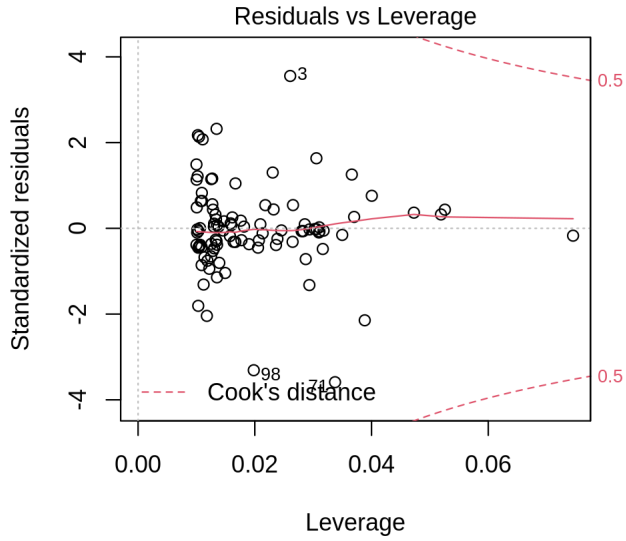
- **Y**-axis: Standardized residuals $\frac{\epsilon_i}{\sigma}$
- **X**-axis: Leverage
- Dashed red line: Cook's distance of 0.5

What we hope to see: No leverage points with high influence

No influential observations

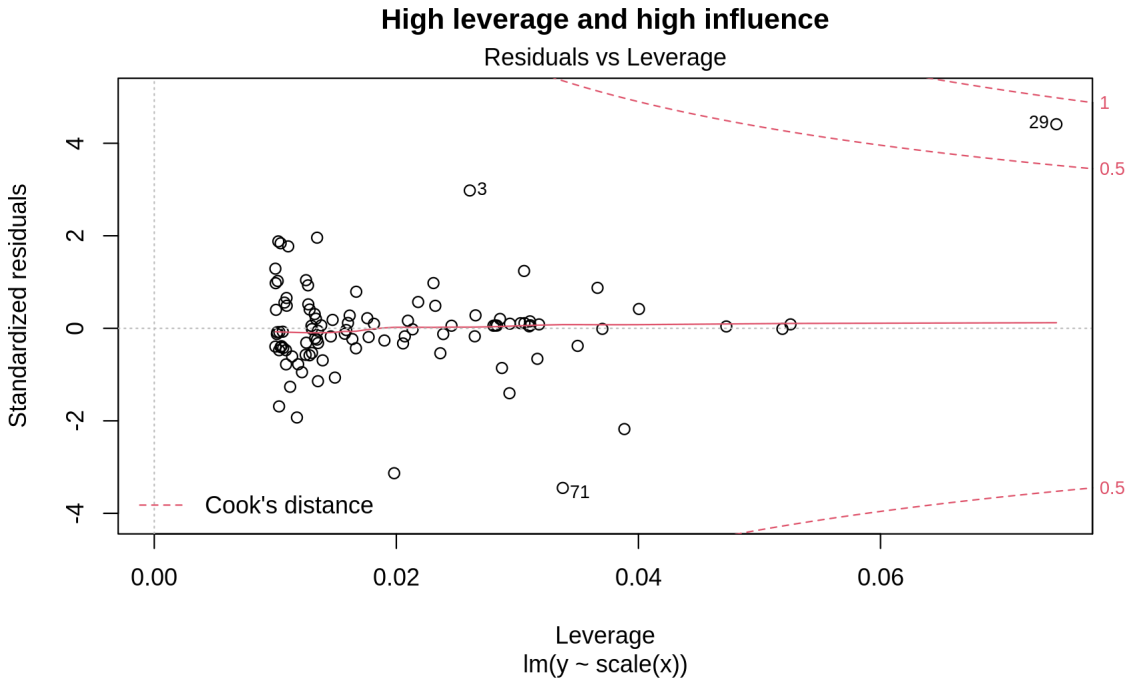


Leverage point with low influence



Diagnostic plot # 4 - Residuals vs Leverage

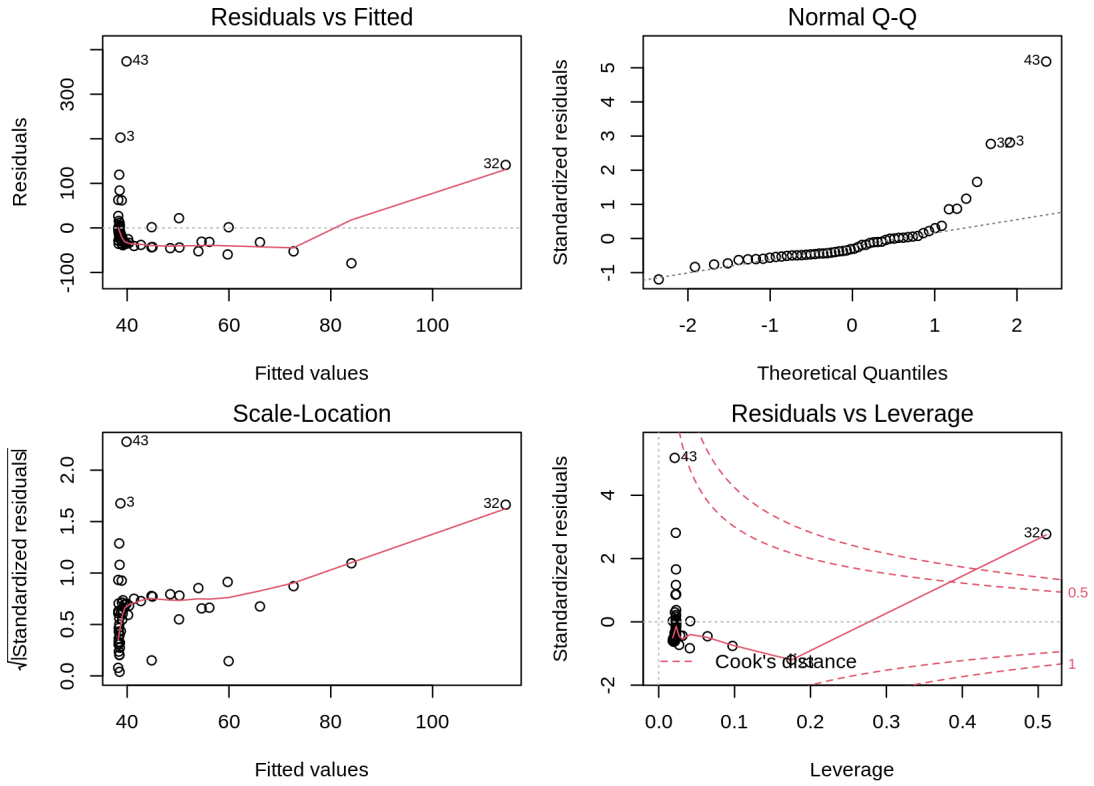
What should make use suspicious:



You should never remove outliers unless you have very good reasons to do so

Step 2. Verify assumptions of `lm1`

```
par(mfrow=c(2,2))
plot(lm1)
```

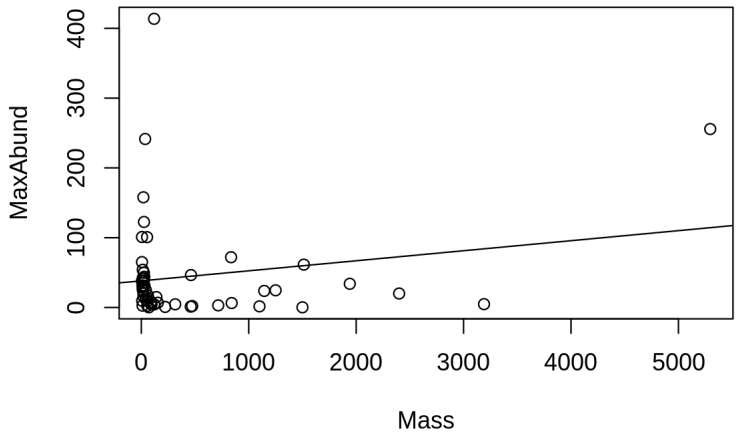


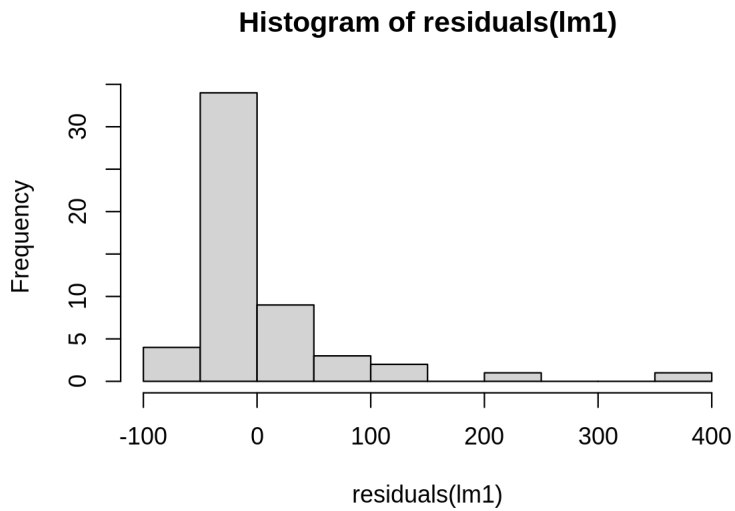
Group discussion: Does `lm1` violate any assumptions of the linear model?

Assumptions not met - what is wrong?

Let's plot our model on top of the observations

```
par(mfrow = c(1,2))
coef(lm1) # intercept and slope
# (Intercept)        Mass
# 38.16645523  0.01438562
plot(MaxAbund ~ Mass, data=bird) # left plot
abline(lm1) # line described by model parameters
hist(residuals(lm1)) # right plot: distribution of residuals
```





Assumptions not met - what is wrong?

To see if the residuals follow a normal distribution, we can also use the *Shapiro-Wilk* and *Skewness* tests:

```
shapiro.test(residuals(lm1))
#
#     Shapiro-Wilk normality test
#
# data:  residuals(lm1)
# W = 0.64158, p-value = 3.172e-10

library(e1071)
skewness(residuals(lm1))
# [1] 3.154719
```

Distribution is significantly different from a normal distribution, and left-skewed (positive skewness)

Assumptions not met - how to proceed?

There are two options when assumptions of the linear model are not satisfied:

1. Use a **different type of model** better suited to the hypothesis and data (QCBS R workshops 6 - 8).
2. Attempt **transformation** of the predictor and / or response variables
    - **Several types of transformations exist**. Their usefulness depends on the distribution of the variable and the type of model.
    - Transformation can **fix some** problems but might **create others**.
    - The **results of statistical tests** on transformed data **do not automatically hold** for the untransformed data.

Challenge 1: A model on transformed variables

Lets try fixing our problems with a log-transformation.

Add the log-transformed variables to our data frame :

```
bird$logMaxAbund <- log10(bird$MaxAbund)
bird$logMass <- log10(bird$Mass)
```

Challenge

Step 1. Re-run the analysis with the log-transformed variables `logMaxAbund` and `logMass`. Save the model as `lm2`

Step 2: Verify assumptions of model `lm2` using diagnostic plots.

```
lm2 <- lm(logMaxAbund ~ logMass, data = bird)
```

Challenge 1: A model on transformed variables

Step 1. Re-run the analysis with the log-transformed variables

```
lm2 <- lm(logMaxAbund ~ logMass, data = bird)

lm2
#
# Call:
# lm(formula = logMaxAbund ~ logMass, data = bird)
#
# Coefficients:
# (Intercept)      logMass
#      1.6724      -0.2361
```

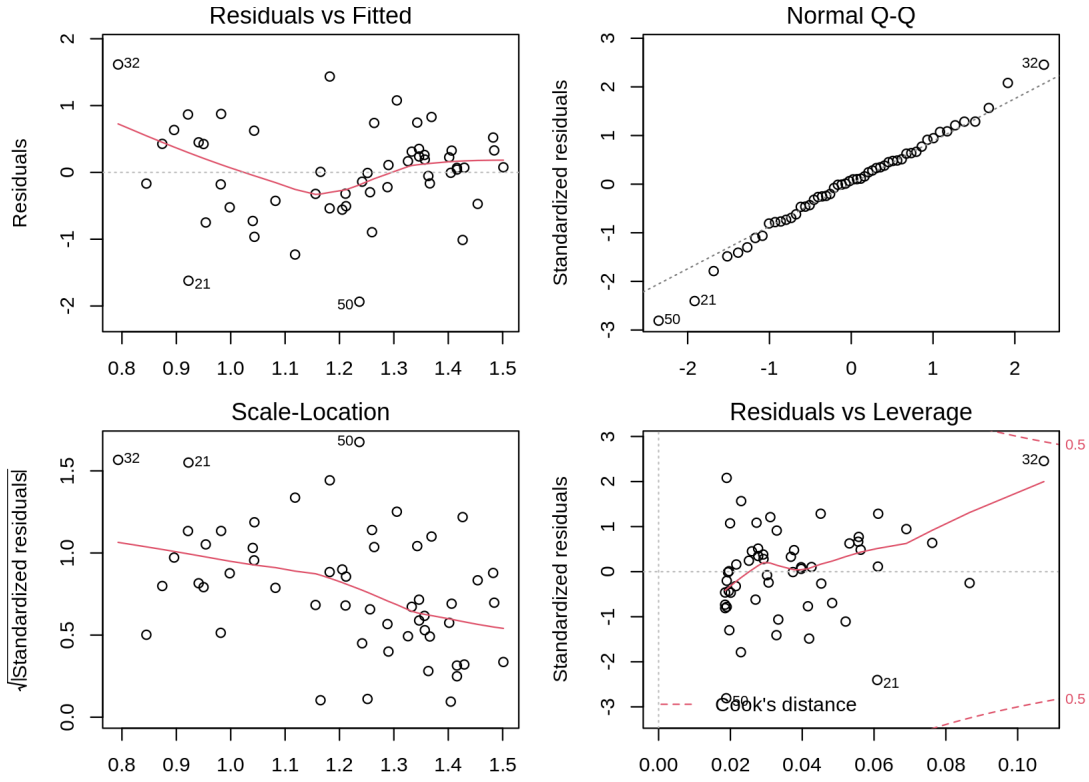
How do the parameters compare to our prediction?

Can we trust these estimates?

Challenge 1: A model on transformed variables

Step 2. Verify assumptions of model `lm2`

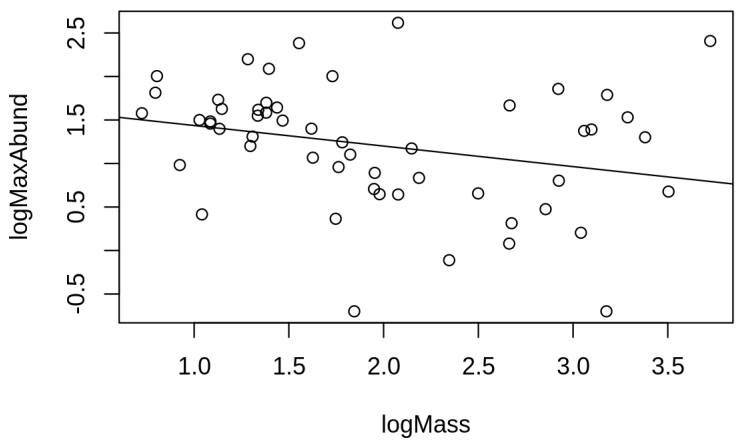
```
par(mfrow=c(2,2))
plot(lm2)
```

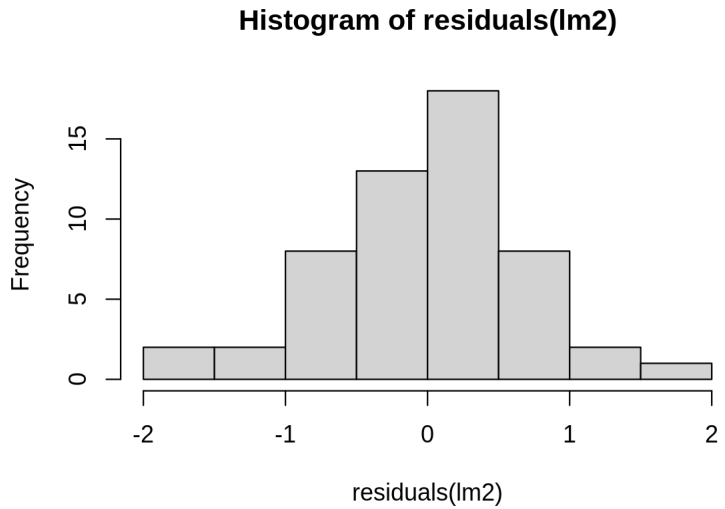


Much improved, but still some problems

Step 2. Verify assumptions of model `lm2`

```
par(mfrow = c(1,2))
coef(lm2) # intercept and slope
# (Intercept)     logMass
#   1.6723673  -0.2361498
plot(logMaxAbund ~ logMass, data=bird) # left plot
abline(lm2) # line described by model parameters
hist(residuals(lm2)) # right plot: distribution of residuals
```





Step 3. Analyze parameter estimates

The `summary()` function provides more information on the fitted model.

```
summary(lm2)

Call:
lm(formula = logMaxAbund ~ logMass, data = bird)

Residuals:
     Min       1Q   Median       3Q      Max
-1.93562 -0.39982  0.05487  0.40625  1.61469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.6724     0.2472   6.767 1.17e-08 ***
logMass      -0.2361     0.1170  -2.019   0.0487 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6959 on 52 degrees of freedom
Multiple R-squared:  0.07267,    Adjusted R-squared:  0.05484
F-statistic: 4.075 on 1 and 52 DF,  p-value: 0.04869
```

Step 3

We can also extract specific parameters and results from the model and summary:

```
# Vectors of residuals and fitted values:
e <- residuals(lm2)
y <- fitted(lm2)

coefficients(lm2) # coefficients
# (Intercept)     logMass
#   1.6723673  -0.2361498
summary(lm2)$coefficients # coefficients with t-tests
#               Estimate Std. Error   t value     Pr(>|t|)
# (Intercept)  1.6723673  0.2471519  6.766557 1.166186e-08
# logMass     -0.2361498  0.1169836 -2.018658 4.869342e-02

summary(lm2)$adj.r.squared # Adjusted R squared
# [1] 0.05483696
```

Model interpretation

How well does the model support our hypothesis?

Hypothesis

> For bird species, the **average size of an individual has an effect on the maximum abundance** of the species, due to ecological constraints (food sources, habitat availability, etc.).

Model interpretation

How well does the model support our hypothesis?

```
summary(lm2)

Call:
lm(formula = logMaxAbund ~ logMass, data = bird)

Residuals:
     Min       1Q   Median       3Q      Max
-1.93562 -0.39982  0.05487  0.40625  1.61469

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.6724     0.2472   6.767 1.17e-08 ***
logMass      -0.2361     0.1170  -2.019   0.0487 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6959 on 52 degrees of freedom
Multiple R-squared:  0.07267,    Adjusted R-squared:  0.05484
F-statistic: 4.075 on 1 and 52 DF,  p-value: 0.04869
```

Model interpretation

How well does the model support our hypothesis?

There is only **very little evidence** in support of the hypothesis because:

- The model does not explain the response well (*low adjusted R-squared*)
- The model is only slightly better than a model without any predictor variables (*F-test barely significant*)
- The parameter estimate for `logMass` is barely dfferent from 0 (*t-test barely significant*)

Finding a better model: terrestrial birds

Maybe we should formulate a more specific hypothesis?

Hypothesis

> For **terrestrial** bird species, the **average size of an individual has an effect on the maximum abundance** of the species, due to ecological constraints (food sources, habitat availability, etc.).

Finding a better model: terrestrial birds

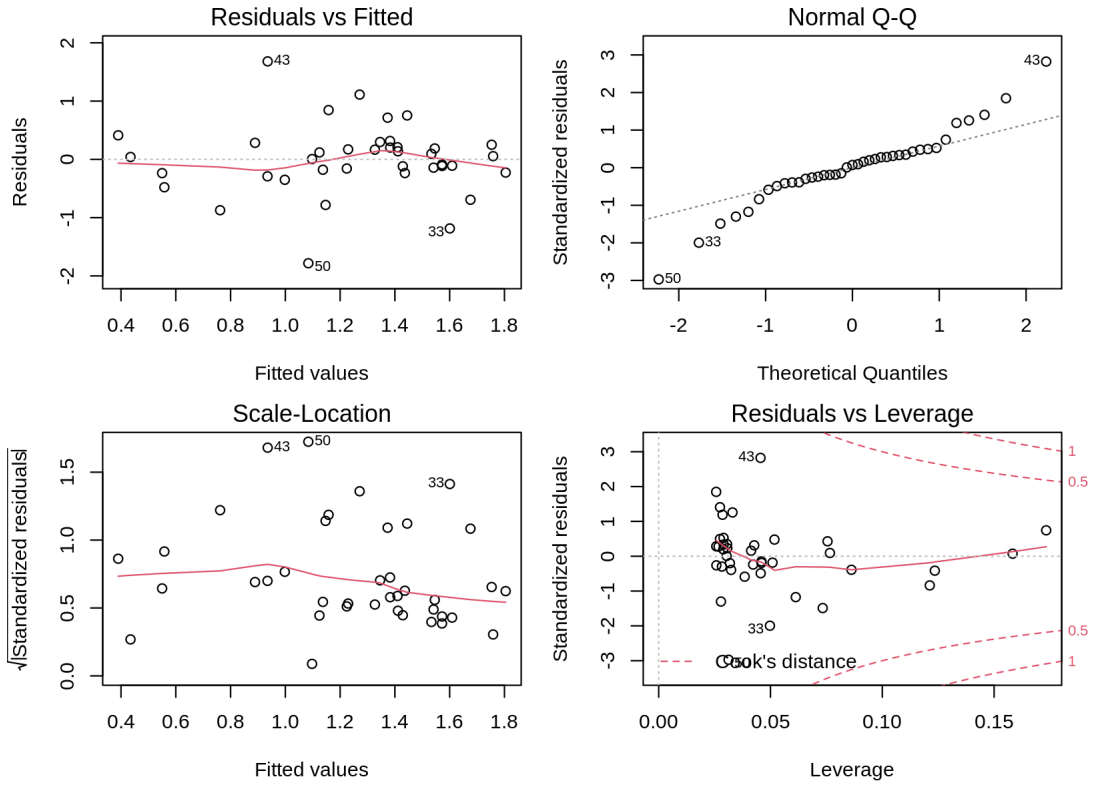
Exclude all aquatic birds (using `!`) and fit a linear model:

```
lm3 <- lm(logMaxAbund~logMass, data=bird, subset=!bird$Aquatic)
# removes aquatic birds (i.e. !birdsAquatic == TRUE)
# or equivalently
# lm3 <- lm(logMaxAbund~logMass, data=bird, subset=bird$Aquatic == 0)

lm3
#
# Call:
# lm(formula = logMaxAbund ~ logMass, data = bird, subset = !bird$Aquatic)
#
# Coefficients:
# (Intercept)      logMass
#      2.2701      -0.6429
```

Finding a better model: terrestrial birds

```
par(mfrow=c(2,2))
plot(lm3)
```



No violation of assumptions

Finding a better model: terrestrial birds

How well does the model support our hypothesis?

```
summary(lm3)

Call:
lm(formula = logMaxAbund ~ logMass, data = bird, subset = !bird$Aquatic)

Residuals:
     Min       1Q   Median       3Q      Max
-1.78289 -0.23135  0.04031  0.22932  1.68109

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2701     0.2931   7.744 2.96e-09 ***
logMass      -0.6429     0.1746  -3.683 0.000733 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6094 on 37 degrees of freedom
Multiple R-squared:  0.2682,    Adjusted R-squared:  0.2485
F-statistic: 13.56 on 1 and 37 DF,  p-value: 0.000733
```

Finding a better model: terrestrial birds

How well does the model support our hypothesis?

The model provides evidence in support of the hypothesis, because:

- The model fits the data reasonably well (*adjusted R-squared*)
- The model is clearly better than a model without any predictor variables (*F-test*)
- The parameter estimate for `logMass` is clearly dfferent from 0 (*t-test*)

Challenge 2

Let's put everything together!

1. Formulate a similar **hypothesis** about maximum abundance and average mass of an indidvidual for **passerine birds**.
2. Fit a **model** to assess this hypothesis, using log-transformed variables (i.e. `logMaxAbund` and `logMass`). Save the model as `lm4`.
3. **Verify assumptions** of the linear model using residual plots.
4. Interpret the results: Does the model provide **evidence for the hypothesis?**

HINT: `Passerine` is also coded 0 and 1 (look at `str(bird)`)

Challenge 2 - Solution

Hypothesis

> For **passerine** bird species, the **average size of an individual has an effect on the maximum abundance** of the species, due to ecological constraints (food sources, habitat availability, etc.).

Challenge 2 - Solution

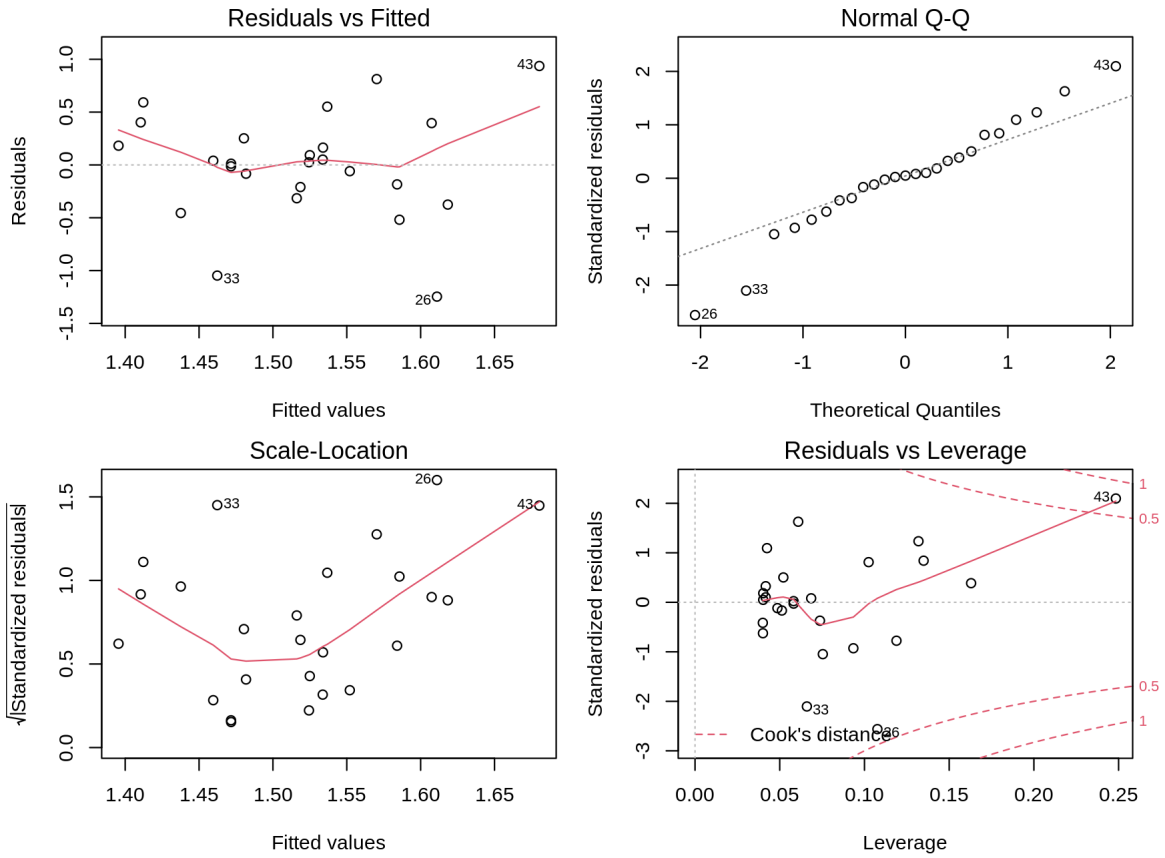
Fit the model:

```
lm4 <- lm(logMaxAbund ~ logMass, data=bird, subset=bird$Passerine == 1)
lm4
#
# Call:
# lm(formula = logMaxAbund ~ logMass, data = bird, subset = bird$Passerine ==
#     1)
#
# Coefficients:
# (Intercept)      logMass
#      1.2429       0.2107
```

Challenge 2 - Solution

Verify assumptions:

```
par(mfrow=c(2,2))
plot(lm4)
```



Challenge 2 - Solution

Should we even interpret the results?

```
summary(lm4)

Call:
lm(formula = logMaxAbund ~ logMass, data = bird, subset = bird$Passerine ==
    1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.24644 -0.20937  0.02494  0.25192  0.93624

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2429     0.4163   2.985  0.00661 **
logMass       0.2107     0.3076   0.685  0.50010
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5151 on 23 degrees of freedom
Multiple R-squared:   0.02,    Adjusted R-squared:  -0.02261
F-statistic: 0.4694 on 1 and 23 DF,  p-value: 0.5001
```

Linear regression in R

Step 1

Fit a linear model based on a hypothesis

Step 2

Verify assumptions of the linear model

↓

Assumptions are met?

Step 3

- Analyze regression parameters
- Plot your model
- Test for significance of parameter estimates (if necessary)

↓

Assumptions are not met?

Consider using a *Generalized Linear Model* (GLM) or transforming the data

↓

Use a GLM that is better suited for the data

↓

Go back to Step 1 with transformed variables

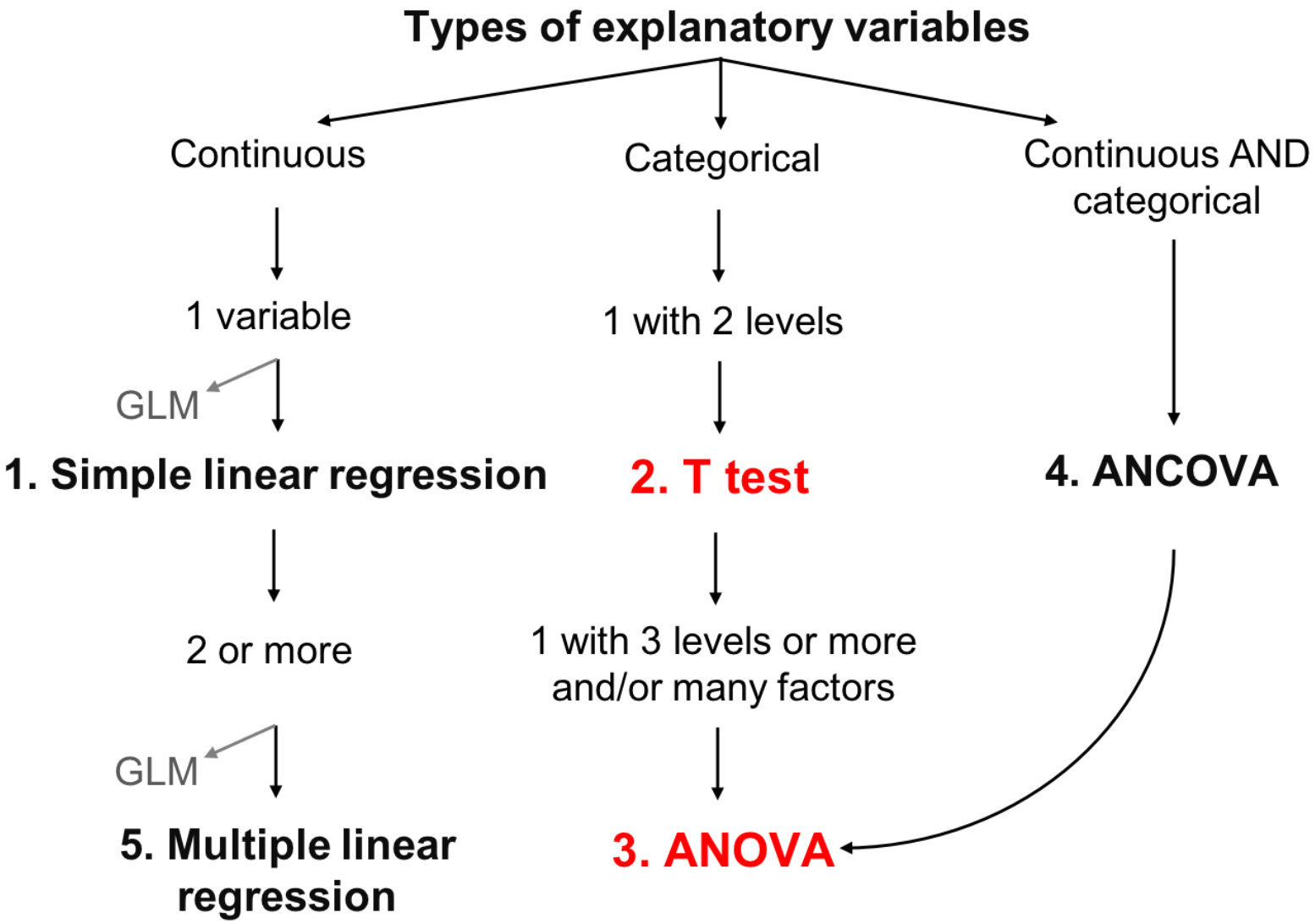
Variable names

Different terms are used for *response* and *predictor* variables, depending on context and scientific field (they are not always synonyms).

response	**predictor**
	explanatory var.
	covariate
outcome	
output var.	input var.
dependent var.	independent var.

réponse	**prédicteur**
var. expliqué	var. explicatif
	covariable
var. endogène	var. exogène
var. dépendante	var. indépendante

Linear models



t-test and ANOVA (Analysis of Variance)

t-test, One-way ANOVA, Two-way ANOVA

Analysis of Variance (ANOVA)

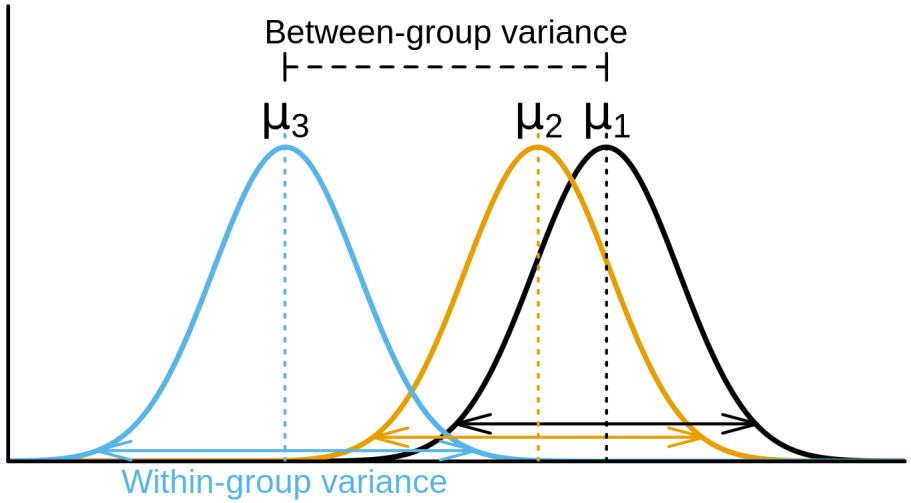
Y: Response variable is **continuous**

X: Explanatory variable(s) is **categorical**, and usually have **two or more levels** (or groups)

ANOVA tests whether the means of the response variable differs between the levels

ANOVA

ANOVA tests whether the means of the response variable differs between the levels



Sum of squares: within-treatment variance *vs* between-treatment variance

If between-treatment variance > within-treatment variance:

- The treatments affect the explanatory variable more than the random error
- The explanatory variable is likely to be significantly influenced by the treatments

Types of ANOVA

1. One-way ANOVA
   - One factor with 2 or more levels
   - If there are 2 levels a **t-test** can be used alternatively
2. Two-way ANOVA
   - 2 factors or more
   - Each factor can have multiple levels
   - The interactions between each factor must be tested

Repeated measures?

- ANOVA can be used for repeated measures, but we won't cover this today
- Linear Mixed-effect Models can also be used for this kind of data (you'll see it on Workshop 6)

T-test

T-test

- **Response variable** → Continuous
- **Explanatory variable** → Categorical with **2 levels**

Assumptions

- Data follow a normal distribution
- Equality of variance between groups (Homoscedasticity)

robustness of this test increases with sample size and is higher when groups have equal sizes

Running a T-test in R

Use the function `t.test()`

```
t.test(Y~X2, data= data, alternative = "two.sided")
```

- `Y`: response variable
- `X2`: factor (2 levels)
- `data`: name of dataframe
- `alternative` hypothesis: `"two.sided"` (default), `"less"`, or `"greater"`

The t-test is still a linear model and a specific case of ANOVA with one factor with 2 levels

Thus the function `lm()` can also be used

```
lm.t <-lm(Y~X2, data = data)
anova(lm.t)
```

Running a T-test in R

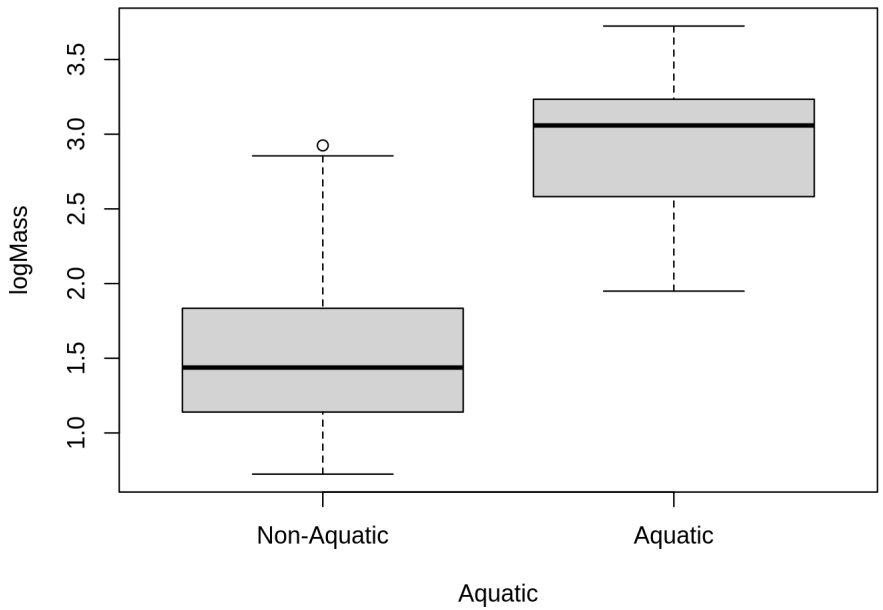
Are aquatic birds heavier than non-aquatic birds?

- Response variable: `Bird mass` ➔ num: Continuous
- Explanatory variable: `Aquatic` ➔ 2 levels: 1 or 0 (yes or no)

Running a T-test in R

First, lets visualize the data using the function `boxplot()`

```
boxplot(logMass ~ Aquatic,
        data = bird, names = c("Non-Aquatic", "Aquatic"))
```



Running a T-test in R

Next, test the assumption of equality of variance using `var.test()`

```
var.test(logMass ~ Aquatic, data = bird)
#
#     F test to compare two variances
#
# data:  logMass by Aquatic
# F = 1.0725, num df = 38, denom df = 14, p-value = 0.9305
# alternative hypothesis: true ratio of variances is not equal to 1
# 95 percent confidence interval:
#  0.3996428 2.3941032
# sample estimates:
# ratio of variances
#           1.072452
```

The ratio of variances is not statistically different from 1, therefore variances are equal

We may now proceed with the t-test!

Running a T-test in R

```
ttest1 <- t.test(logMass ~ Aquatic, var.equal = TRUE, data = bird)

# Or use lm()
ttest.lm1 <- lm(logMass ~ Aquatic, data=bird)
```

Indicates that homogeneity of variance was respected (as we just tested)

You can verify that `t.test()` and `lm()` provide the same model:

```
ttest1$statistic^2
#       t
# 60.3845
anova(ttest.lm1)$`F value`
# [1] 60.3845      NA
# answer: F=60.3845 in both cases
```

When the assumption of equal variance is met `(t^2)` follows an F distribution

Running a T-test in R

If $p < 0.01$ (or 0.05), the null hypothesis of no difference between the two groups (*H0*) can be rejected, with a risk of 0.01 (or 0.05) that we made a mistake in this conclusion

```
ttest1
#
#     Two Sample t-test
#
# data:  logMass by Aquatic
# t = -7.7707, df = 52, p-value = 2.936e-10
# alternative hypothesis: true difference in means is not equal to 0
# 95 percent confidence interval:
#  -1.6669697 -0.9827343
# sample estimates:
# mean in group 0 mean in group 1
#       1.583437        2.908289
```

There exists a difference in mass between the aquatic and terrestrial birds - `p-value`

Look at the mean of the two groups

Violation of Assumptions

- If variances between groups are not equal, can use corrections like the Welch correction (DEFAULT in R!)
- If assumptions cannot be respected, the **non-parametric** equivalent of t-test is the Mann-Whitney test
- If two groups **are not independent** (e.g. measurements on the same individual at 2 different years), you should use a paired t-test

Poll

With T-test we can be more specific and give a direction to our hypothesis with `alternative`.

We want to test whether **aquatic birds are heavier than terrestrial birds**.

Which one should we use? `alternative = "???"`

1. `"two.sided"`
2. `"less"`
3. `"greater"`

```
# Unilateral t-test
uni.ttest1 <- t.test(logMass ~ Aquatic,
                     var.equal = TRUE,
                     data = bird,
                     alternative = "???")
```

Answer

```
# Unilateral t-test
uni.ttest1 <- t.test(logMass ~ Aquatic,
                     var.equal = TRUE,
                     data = bird,
                     alternative = "less")
uni.ttest1
#
#     Two Sample t-test
#
# data:  logMass by Aquatic
# t = -7.7707, df = 52, p-value = 1.468e-10
# alternative hypothesis: true difference in means is less than 0
# 95 percent confidence interval:
#       -Inf -1.039331
# sample estimates:
# mean in group 0 mean in group 1
#        1.583437        2.908289
```

Why would `"greater"` not have work?

Hint: check the dataset. What is the nature of the variable Aquatic?

ANOVA

Analysis of Variance (ANOVA)

Generalization of the t-test for categorical variables with **two or more levels**.

Subsets variation in the response variable into additive effects of one or several factors and the interactions between them:

$$Y = \underbrace{\mu}_{\text{The average outcome over all individuals}} + \overbrace{\tau_i}^{\text{The average outcome over all individuals in group i}} + \underbrace{\epsilon}_{\text{Residuals}}$$

Review: ANOVA

Assumptions

- Normality of residuals
- Equality (*i.e.* homogeneity) of within-group variances, called *homoscedasticity*.
- Plus, independency, random samples.

Complementary test

- When the ANOVA detects a significant difference between groups, it does not tell you which group (or groups) differs from the others.
- A commonly used *post-hoc test* to answer this question is the **Tukey's test**.
- You may also compare between groups using **planned comparisons**. This is more elegant, because it expects that you have an *a priori* expectation for the differences between groups.

Running an ANOVA in R

Does abundance vary across different diets?

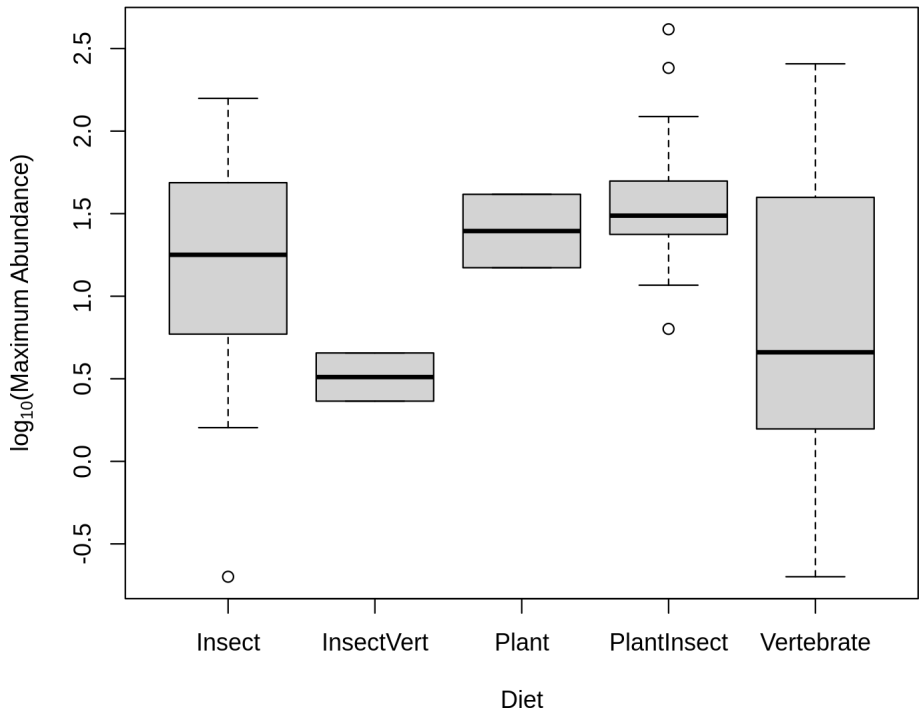
- Response variable: `MaxAbund` ➔ `num`: continuous
- Explanatory variable: `Diet` ➔ `factor` with 5 levels

```
str(bird)
# 'data.frame':    54 obs. of  9 variables:
#  $ Family     : Factor w/ 53 levels "Anhingas","Auks& Puffins",..: 18 25 23 21 2 1
#  $ MaxAbund   : num  2.99 37.8 241.4 4.4 4.53 ...
#  $ AvgAbund   : num  0.674 4.04 23.105 0.595 2.963 ...
#  $ Mass       : num  716 5.3 35.8 119.4 315.5 ...
#  $ Diet       : Factor w/ 5 levels "Insect","InsectVert",..: 5 1 4 5 2 4 5 1 1 5 .
#  $ Passerine  : int  0 1 1 0 0 0 0 0 0 0 ...
#  $ Aquatic    : int  0 0 0 0 1 1 1 0 1 1 ...
#  $ logMaxAbund: num  0.475 1.577 2.383 0.643 0.656 ...
#  $ logMass    : num  2.855 0.724 1.554 2.077 2.499 ...
```

Visualize data

First, visualize the data using `boxplot()` (alphabetical order, by default).

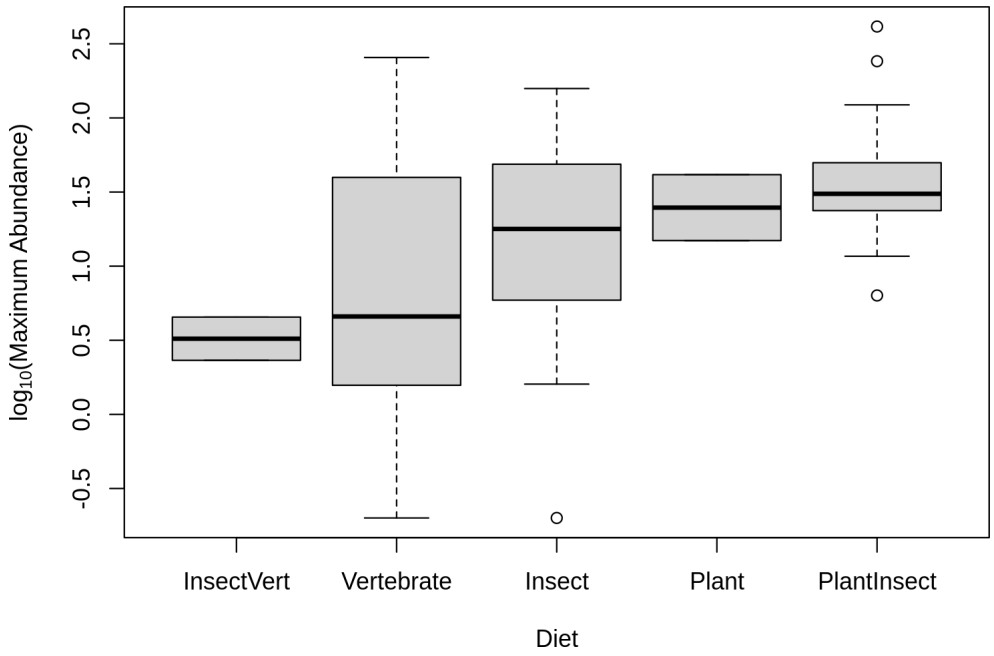
```
boxplot(logMaxAbund ~ Diet,
        data = bird,
        ylab = expression("log"[10]*"(Maximum Abundance)"),
        xlab = 'Diet')
```



Visualize data

We can reorder factor levels according to group medians using the `tapply()` and `sort()` functions.

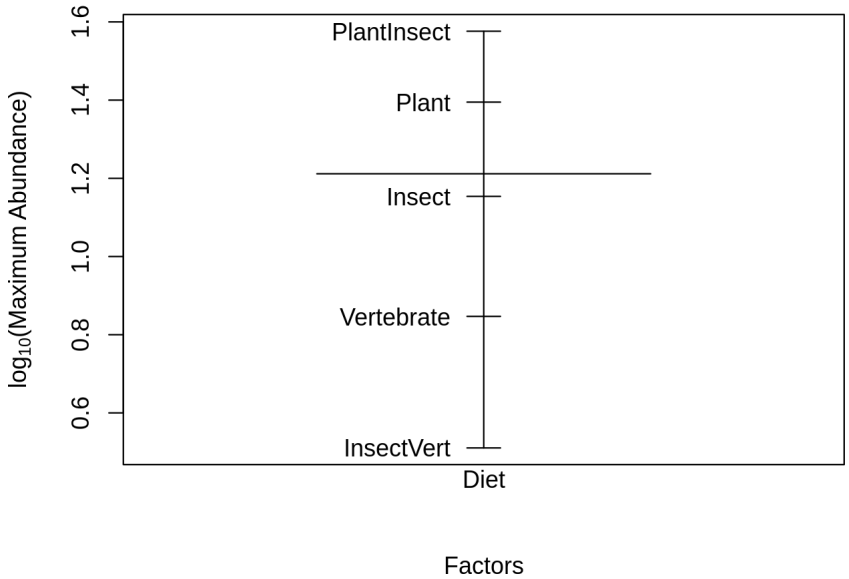
```
med <- sort(tapply(bird$logMaxAbund, bird$Diet, median))
boxplot(logMaxAbund ~ factor(Diet, levels = names(med)), data = bird,
        ylab = expression("log"[10]*"(Maximum Abundance)"), xlab = 'Diet')
```



Visualize data

Another way to represent the effect sizes is to use `plot.design()`.

```
plot.design(logMaxAbund ~ Diet, data=bird,
            ylab = expression("log"[10]*"(Maximum Abundance)"))
```



Levels of a particular factor are displayed along a vertical line, and the overall value of the response variable, in a horizontal line.

Running an One-Way ANOVA in R

We can use the function `stats::lm()` to run an ANOVA:

```
anov1 <- lm(logMaxAbund ~ Diet,
            data = bird)
```

We can also use an specific function called `stats::aov()`:

```
aov1 <- aov(logMaxAbund ~ Diet,
            data = bird)
```

Let's work together: Try both of them and compare the outputs!

Running an One-Way ANOVA in R

And, here are the outputs!

```
anova(anov1)
# Analysis of Variance Table
#
# Response: logMaxAbund
#           Df  Sum Sq Mean Sq F value Pr(>F)
# Diet       4  5.1059 1.27647  2.8363 0.0341 *
# Residuals 49 22.0521 0.45004
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov1)
#             Df Sum Sq Mean Sq F value Pr(>F)
# Diet         4  5.106   1.276   2.836 0.0341 *
# Residuals   49 22.052   0.450
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Did we violate the model assumptions?

Are the variances in each of the groups (samples) the same?

Bartlett's test of homogeneity of variances:

```
bartlett.test(logMaxAbund ~ Diet, data = bird)
#
#     Bartlett test of homogeneity of variances
#
# data:  logMaxAbund by Diet
# Bartlett's K-squared = 7.4728, df = 4, p-value = 0.1129
```

Did we violate the model assumptions?

Levene's test of homogeneity of variances:

```
library(car)
leveneTest(logMaxAbund ~ Diet, data = bird)
# Levene's Test for Homogeneity of Variance (center = median)
#       Df F value  Pr(>F)
# group  4  2.3493 0.06717 .
#       49
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene's test performs better, but has a slightly higher Type II error.

Did we violate the model assumptions?

Are the model residuals normally distributed?

Shapiro-Wilk's test for normality of residuals

```
shapiro.test(resid(anov1))
#
#     Shapiro-Wilk normality test
#
# data:  resid(anov1)
# W = 0.97995, p-value = 0.4982
```

Assumptions of homocedasticity and normality of residuals not violated!

What if the assumptions were **not** met?

Alternatives

1. **Transform your variables** to **normalize residuals** and-or **homogenize variances** and-or **convert a multiplicative effects into additive**. For example:

```
data$logY <- log10(data$Y)
```

- See Workshop 1 for rules on data transformation;
- You must verify the assumptions once again with the transformed data adjusted in your model (*i.e.*, `lm(Y ~ X, data)` changes to `lm(logY ~ X, data)`).

1. **Use a non-parametric test: Kruskal-Wallis' Test** is one non-parametric equivalent to ANOVA.

```
stats::kruskal.test(Y~X, data)
```

Output of our ANOVA model

Factor levels in alphabetical order and all levels are compared to the reference level (`Insect`).

```
summary(anov1)
#
# Call:
# lm(formula = logMaxAbund ~ Diet, data = bird)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1.85286 -0.32972 -0.08808  0.47375  1.56075
#
# Coefficients:
#                Estimate Std. Error t value Pr(>|t|)
# (Intercept)       1.1539     0.1500   7.692 5.66e-10 ***
# DietInsectVert   -0.6434     0.4975  -1.293   0.2020
# DietPlant         0.2410     0.4975   0.484   0.6303
# DietPlantInsect   0.4223     0.2180   1.938   0.0585 .
# DietVertebrate   -0.3070     0.2450  -1.253   0.2161
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.6709 on 49 degrees of freedom
# Multiple R-squared:  0.188,    Adjusted R-squared:  0.1217
```

Output of our ANOVA model

On the other hand, if `lm()` is used:

```
summary.lm(aov1)
#
# Call:
# aov(formula = logMaxAbund ~ Diet, data = bird)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1.85286 -0.32972 -0.08808  0.47375  1.56075
#
# Coefficients:
#                 Estimate Std. Error t value Pr(>|t|)
# (Intercept)       1.1539     0.1500   7.692 5.66e-10 *
# DietInsectVert   -0.6434     0.4975  -1.293   0.2020
# DietPlant         0.2410     0.4975   0.484   0.6303
# DietPlantInsect   0.4223     0.2180   1.938   0.0585 .
# DietVertebrate   -0.3070     0.2450  -1.253   0.2161
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
#
# Residual standard error: 0.6709 on 49 degrees of freed
# Multiple R-squared:  0.188,    Adjusted R-squared:  0.
# F-statistic: 2.836 on 4 and 49 DF,  p-value: 0.0341
```

Significant difference between `Diet` groups, but we do not know which ones!

A posteriori test

If the ANOVA detects significant differences between means, a post-hoc test is required to determine which treatment(s) differ from each other. This can be done with the `TukeyHSD()` function:

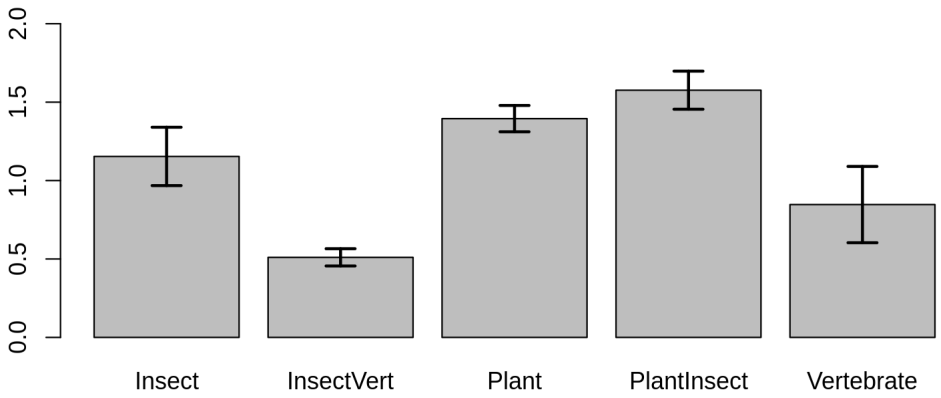
```
TukeyHSD(aov(anov1), ordered = TRUE)
#   Tukey multiple comparisons of means
#     95% family-wise confidence level
#     factor levels have been ordered
#
# Fit: aov(formula = anov1)
#
# $Diet
#                             diff         lwr      upr
# Vertebrate-InsectVert  0.3364295 -1.11457613 1.787435
# Insect-InsectVert      0.6434334 -0.76550517 2.052372
# Plant-InsectVert       0.8844338 -1.01537856 2.784246
# PlantInsect-InsectVert 1.0657336 -0.35030287 2.481770
# Insect-Vertebrate      0.3070039 -0.38670951 1.000717
# Plant-Vertebrate       0.5480043 -0.90300137 1.999010
# PlantInsect-Vertebrate 0.7293041  0.02128588 1.437322
# Plant-Insect           0.2410004 -1.16793813 1.649939
# PlantInsect-Insect     0.4223003 -0.19493574 1.039536
# PlantInsect-Plant      0.1812999 -1.23473664 1.597336
```

Only `Vertebrate` and `PlantInsect` differ

Representation

ANOVA results can be graphically illustrated using the `barplot()` function:

```
sd <- tapply(bird$logMaxAbund, list(bird$Diet), sd)
means <- tapply(bird$logMaxAbund, list(bird$Diet), mean)
n <- length(bird$logMaxAbund)
se <- 1.96*sd/sqrt(n)
bp <- barplot(means, ylim = c(0, max(bird$logMaxAbund) - 0.5))
epsilon = 0.1
segments(bp, means - se, bp, means + se, lwd=2) # vertical bars
segments(bp - epsilon, means - se,
         bp + epsilon, means - se, lwd = 2) # horizontal bars
segments(bp - epsilon, means + se,
         bp + epsilon, means + se, lwd = 2) # horizontal bars
```



Two-way ANOVA

Two-way ANOVA

One-Way ANOVA:

```
aov <- lm(Y ~ X, data)
```

Two-Way ANOVA:

```
aov <- lm(Y ~ X * Z, data)
```

where,

Y: Response variable is **continuous**

X: Explanatory variable(s) is **categorical**, and usually have **two or more levels** (or groups)

Z: Explanatory variable(s) is **categorical**, and usually have **two or more levels** (or groups)

The "`*`" symbol indicates that the main effects, as well, as their interaction will be included in the model.

If use the "`+`" symbol, the main effects, but not their interaction are included.

We can also represent our model with: `aov <- lm(Y ~ X + Z + ..., data)`

Two-Way ANOVA

Always start reading the output from the interaction term, then proceed to the main effects.

```
aov <- lm(Y ~ X * Z, data)

summary(aov)
# Analysis of Variance Table
#
# Response: Y
# Df Sum Sq Mean Sq F value Pr(>F)
# X 4 5.1059 1.27647 3.0378 0.02669 *
# Z 1 0.3183 0.31834 0.7576 0.38870
# X:Z 3 2.8250 0.94167 2.2410 0.10689
# Residuals 45 18.9087 0.42019
# ---
# Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

According to law of **parsimony**, select the model that explain the most variance with the least model parameters as possible: If the multiplicative effect is non-signficant, you may consider a model with only the additive effects:

```
aov <- lm(Y ~ X + Z, data)
```

Challenge 3

Evaluate how `log10(MaxAbund)` varies with `Diet` and `Aquatic`

Hint: add an interaction with `*`,

Break out rooms!

Challenge 2 - Solution

```
anov2 <- lm(logMaxAbund ~ Diet*Aquatic, data = bird)
summary(anov2)
#
# Call:
# lm(formula = logMaxAbund ~ Diet * Aquatic, data = bird)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -1.9508 -0.2447  0.0000  0.3584  1.1558
#
# Coefficients: (1 not defined because of singularities)
#                         Estimate Std. Error t value Pr(>|t|)
# (Intercept)              1.23447    0.17324   7.126 6.64e-09 ***
# DietInsectVert          -0.86989    0.67097  -1.296   0.2014
# DietPlant                0.16043    0.49001   0.327   0.7449
# DietPlantInsect          0.35358    0.23395   1.511   0.1377
# DietVertebrate          -0.95449    0.33772  -2.826   0.0070 **
# Aquatic                 -0.26858    0.31630  -0.849   0.4003
# DietInsectVert:Aquatic   0.56034    0.96976   0.578   0.5663
# DietPlant:Aquatic             NA         NA      NA       NA
# DietPlantInsect:Aquatic  0.05516    0.73821   0.075   0.9408
# DietVertebrate:Aquatic   1.24044    0.49408   2.511   0.0157 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Challenge 3 - Solution

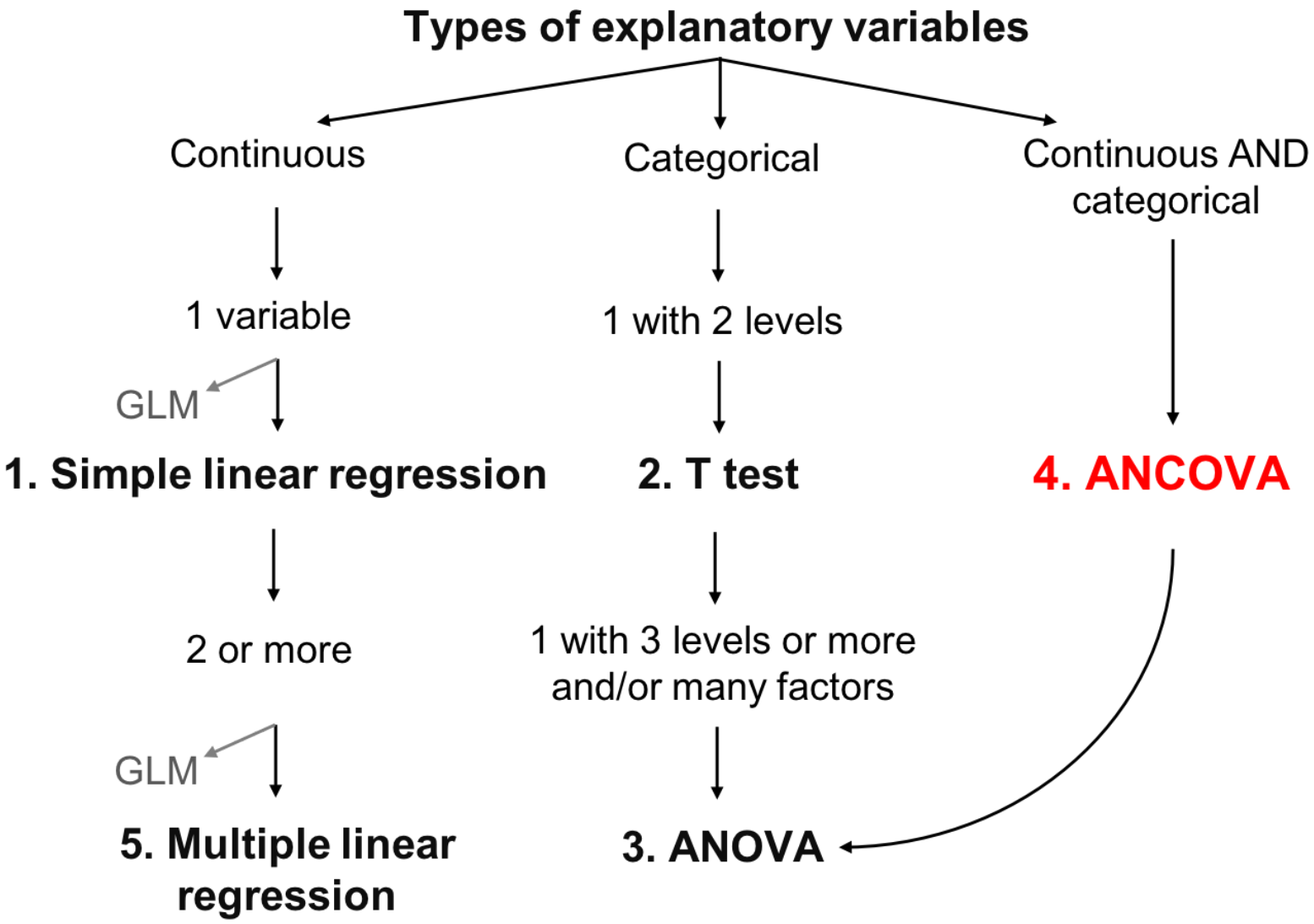
```
anov2 <- lm(logMaxAbund ~ Diet*Aquatic, data = bird)
anova(anov2)
# Analysis of Variance Table
#
# Response: logMaxAbund
#              Df  Sum Sq Mean Sq F value  Pr(>F)
# Diet          4  5.1059 1.27647  3.0378 0.02669 *
# Aquatic       1  0.3183 0.31834  0.7576 0.38870
# Diet:Aquatic  3  2.8250 0.94167  2.2410 0.09644 .
# Residuals    45 18.9087 0.42019
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case, the only significant term of the model is the `Diet` factor.

To adopt the most parsimonious model, we are going to remove the interaction term:

```
anov2 <- lm(logMaxAbund ~ Diet, data = bird)
```

Linear models



ANCOVA

Analysis of Covariance

Analysis of Covariance (ANCOVA)

Here, consider the following:

$$Y = X * Z$$

where,

Y: Response variable is **continuous**

X: Explanatory variable(s) is **categorical**

Z: Explanatory variable(s) is **continuous**

$$\begin{gathered} Y = \mu + \text{Main Effects of Factors} + \\ \text{Interactions between factors} + \\ \text{Main effects of covariates} + \\ \text{Interactions between covariates and factors} + \epsilon \end{gathered}$$

Review: ANCOVA

Assumptions

In addition to the other assumptions of linear models, **ANCOVA** must have:

- The same value range for all covariates;
- Variables that are fixed;
- Categorical and continuous variables that are not "colinear".

*A **fixed** variable is one that you are specifically interested in (e.g., bird mass) whereas a **random** variable is noise that you want to control for (e.g. sites where the birds were sampled at).*

Workshop 6 will cover linear-mixed effects models.

Types of ANCOVA

You can have any number of factors and/or covariates, but as their number increases, the interpretation of results gets more complex.

Frequently used ANCOVA models:

1. **One covariate and one categorical**
2. One covariate and two factors
3. Two covariates and one factor

We will only see the first case today, but this will help you understand the other two kinds!

ANCOVA with 1 covariate and 1 factor

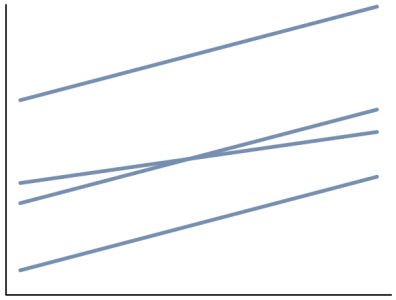
To imagine possible goals of this analysis, you may be interested in the:

1. Effect of factor and covariate on the response variable;
2. Effect of factor on the response variable after removing effect of covariate;
3. Effect of covariate on response variable while controlling for the effect of factor.

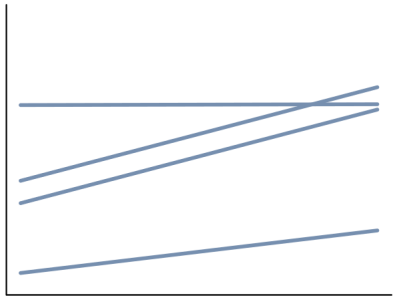
You can only meet these goals if your factor and your covariate are not related!

ANCOVA with 1 covariate and 1 factor

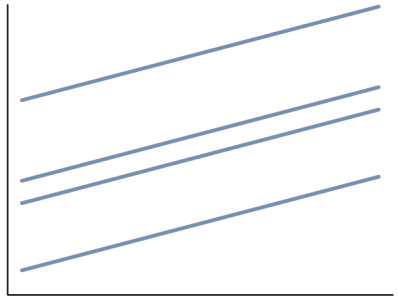
One level of the factor has a different slope



Many levels have different slopes



No interaction



If the interaction is significant, you will have a scenario that looks like these

If your covariate and factor are significant, outputs will look like this

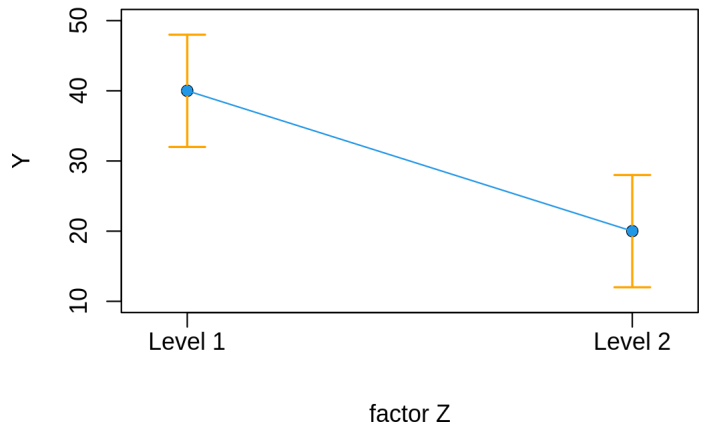
ANCOVA: adjusted mean comparisons

To compare the mean values of each factor, conditional on the effect of the other

The `effects::effect()` function uses the output of the ANCOVA model to estimate the means of each factor level, corrected by the effect of the covariate

```
ancova.example <- lm(Y ~ X*Z, data=data)
# X = quantitative; Z = categorical

library(effects)
adj.means.ex <- effect('Z', ancova.exemple)
plot(adj.means.ex)
```



ANCOVA with 1 covariate and 1 factor

When parsimoniality is the way to go:

- If only your factor is significant, remove the covariate -> you will have a simple **ANOVA**
- If only your covariate is significant, remove your factor -> you will have a **simple linear regression**
- If the interaction between your covariate and factor (`*`) is significant, you should explore which levels of the factor have different slopes from the others.

Verify assumptions!

- This is very similar to what we have done so far!

Running an ANCOVA in R

Is `MaxAbund` a function of `Diet` and `Mass`?

Response variable: **MaxAbund** ➔ num : quantitative continuous

Explanatory variables:

- `Diet` ➔ factor with 5 levels
- `Mass` ➔ numeric, continuous

```
str(bird)
# 'data.frame':    54 obs. of  9 variables:
#  $ Family     : Factor w/ 53 levels "Anhingas","Auks& Puffins",..: 18 25 23 21 2 1
#  $ MaxAbund   : num  2.99 37.8 241.4 4.4 4.53 ...
#  $ AvgAbund   : num  0.674 4.04 23.105 0.595 2.963 ...
#  $ Mass       : num  716 5.3 35.8 119.4 315.5 ...
#  $ Diet       : Factor w/ 5 levels "Insect","InsectVert",..: 5 1 4 5 2 4 5 1 1 5 .
#  $ Passerine  : int  0 1 1 0 0 0 0 0 0 0 ...
#  $ Aquatic    : int  0 0 0 0 1 1 1 0 1 1 ...
#  $ logMaxAbund: num  0.475 1.577 2.383 0.643 0.656 ...
#  $ logMass    : num  2.855 0.724 1.554 2.077 2.499 ...
```

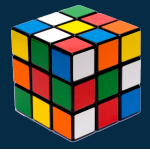
Challenge 4

1- Use an ANCOVA to test the effect of `Diet` and `Mass` (as well as their interaction) on `MaxAbund`

2- Test whether the interaction is significant

Break out rooms!

Challenge 3 - Solution



```
ancov1 <- lm(logMaxAbund ~ logMass*Diet,
             data = bird)
anova(ancov1)
# Analysis of Variance Table
#
# Response: logMaxAbund
#              Df  Sum Sq Mean Sq F value  Pr(>F)
# logMass       1  1.9736 1.97357  4.6054 0.03743 *
# Diet          4  3.3477 0.83691  1.9530 0.11850
# logMass:Diet  4  2.9811 0.74527  1.7391 0.15849
# Residuals    44 18.8556 0.42854
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

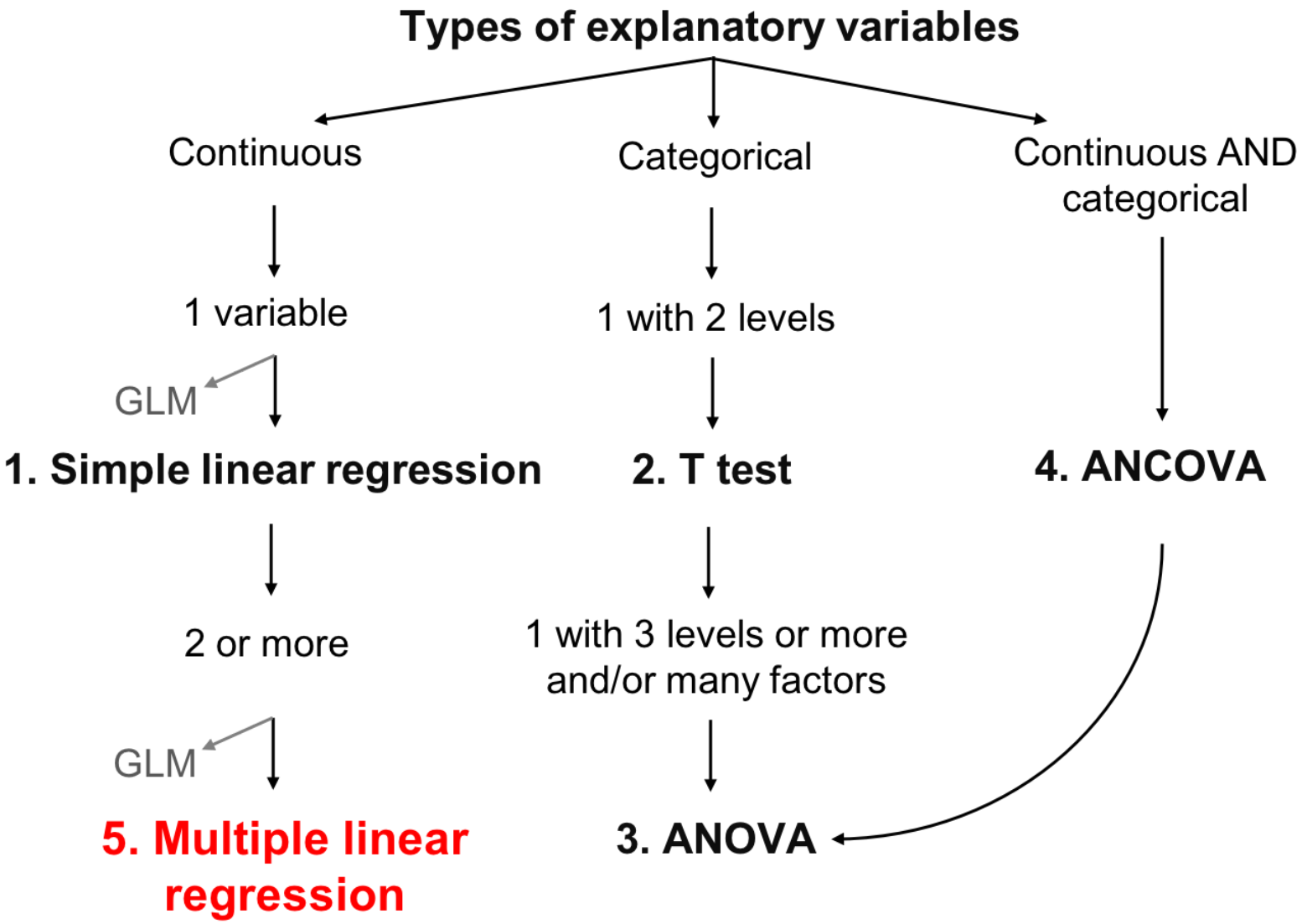
Interaction between `logMass` and `Diet` is not significant

Challenge 4 - Solution

Remove the interaction term and re-evaluate the model (with only the main effects of `Diet` and `logMass`).

```
ancov2 <- lm(logMaxAbund ~ logMass + Diet,
             data = bird)
anova(ancov2)
# Analysis of Variance Table
#
# Response: logMaxAbund
#           Df  Sum Sq Mean Sq F value  Pr(>F)
# logMass    1  1.9736 1.97357  4.3382 0.04262 *
# Diet       4  3.3477 0.83691  1.8396 0.13664
# Residuals 48 21.8367 0.45493
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear models



Multiple linear regression

Multiple linear regression

Only difference to simple linear regression: **several predictor variables** are included in the model.

Variables

- y: Response variable (**continuous**)
- $x_1, x_2, \ldots, x_k$: Several predictor variables (**continuous** or **categorical**)

Assumed relationsip

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + \beta_k x_{k,i} + \epsilon_i$$

- Paramter β_0 is the **intercept**
- Parameters $\beta_1, \beta_2, \ldots, \beta_k$ quantify the **effect** of $x_1, x_2, \ldots, x_k$ on y
- The residual ϵ_i captures **unexplained** variation
- The **fitted** (or predicted) value of y_i is defined as:
  $\hat{y}_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + \beta_k x_{k,i}$

Multiple linear regression

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + \beta_k x_{k,i} + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Assumptions

In addition to the usual assumptions of linear models:

- **Linear relationship** between **each** predictor variable and the response variable.
- Predictor variables are independent of each other (i.e., there is **no collinearity**).

Multiple linear regression

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \ldots + \beta_k x_{k,i} + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

If variables are collinear:

- Reduce the amount of collinear variables.
- Migrate to multidimensional analyses (see Workshop 9).
- Try a pseudo-orthogonal analysis.

Multiple linear regression in R

Using the `Dickcissel` datast to test effect of climate (`clTma`), productivity (`NDVI`) and grassland cover (`grass`) on bird abundance (`abund`):

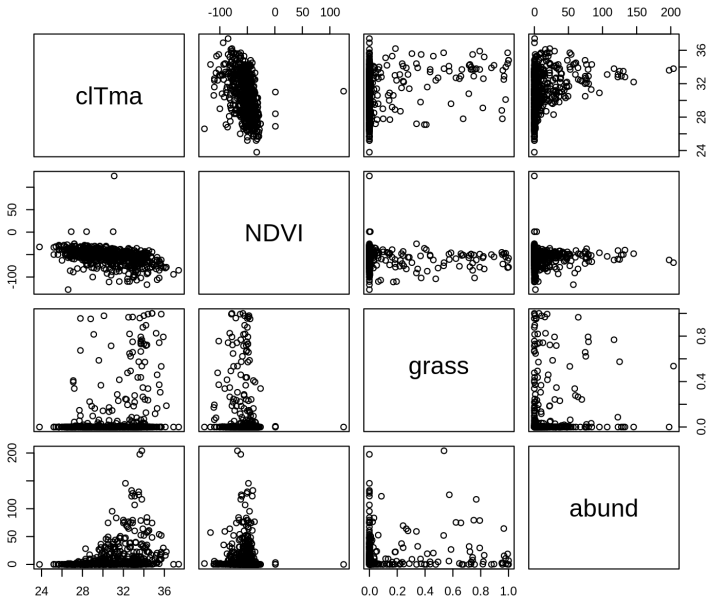
```
Dickcissel = read.csv("data/dickcissel.csv")
str(Dickcissel)
# 'data.frame':    646 obs. of  15 variables:
#  $ abund    : num  5 0.2 0.4 0 0 0 0 0 0 0 ...
#  $ Present  : chr  "Absent" "Absent" "Absent" "Present" ...
#  $ clDD     : num  5543 5750 5395 5920 6152 ...
#  $ clFD     : num  83.5 67.5 79.5 66.7 57.6 59.2 59.5 51.5 47.4 46.3 ...
#  $ clTmi    : num  9 9.6 8.6 11.9 11.6 10.8 10.8 11.6 13.6 13.5 ...
#  $ clTma    : num  32.1 31.4 30.9 31.9 32.4 32.1 32.3 33 33.5 33.4 ...
#  $ clTmn    : num  15.2 15.7 14.8 16.2 16.8 ...
#  $ clP      : num  140 147 148 143 141 ...
#  $ NDVI     : int  -56 -44 -36 -49 -42 -49 -48 -50 -64 -58 ...
#  $ broadleaf: num  0.3866 0.9516 0.9905 0.0506 0.2296 ...
#  $ conif    : num  0.0128 0.0484 0 0.9146 0.7013 ...
#  $ grass    : num  0 0 0 0 0 0 0 0 0 0 ...
#  $ crop     : num  0.2716 0 0 0.0285 0.044 ...
#  $ urban    : num  0.2396 0 0 0 0.0157 ...
#  $ wetland  : num  0 0 0 0 0 0 0 0 0 0 ...
```

Verify assumptions!

Collinearity:

- Examine the degree of collinearity of all explanatory variables and variables of interest using the `plot()` function.

```
# select variables
var <- c('clTma', 'NDVI',
         'grass', 'abund')
plot(Dickcissel[, var])
```



If you see a pattern between any two variables, they may be collinear!

They are likely to explain the same variability of the response variable and the effect of one variable will be masked by the other

Multiple linear regression in R

Run a multiple linear regression with `abund` as a function of `clTma + NDVI + grass`.

```
lm.mult <- lm(abund ~ clTma + NDVI + grass, data = Dickcissel)
summary(lm.mult)
```

Verify diagnostic plots, as you did for the simple linear regression:

```
par(mfrow = c(2, 2))
plot(lm.mult)
```

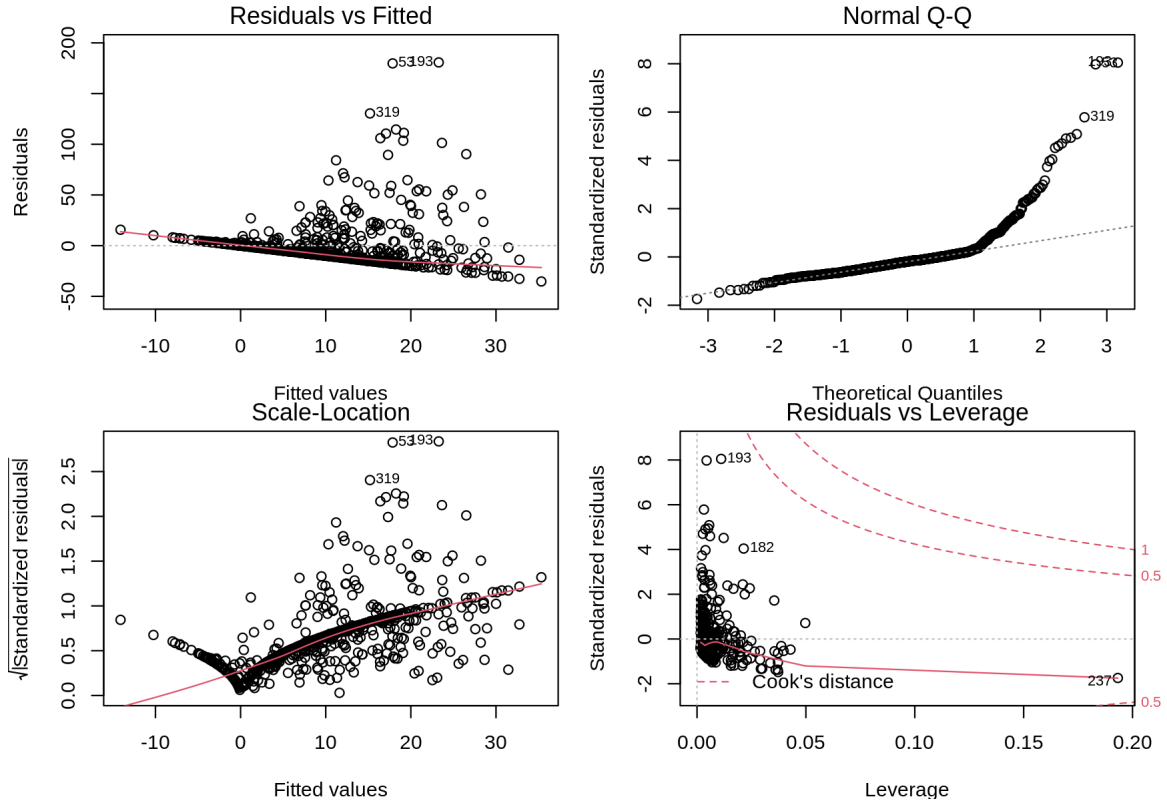
```r
lm.mult <- lm(abund ~ clTma + NDVI + grass, data = Dickcissel)
summary(lm.mult)
#
# Call:
# lm(formula = abund ~ clTma + NDVI + grass, data = Dickcissel)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -35.327 -11.029  -4.337   2.150 180.725
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept) -83.60813   11.57745  -7.222 1.46e-12 ***
# clTma         3.27299    0.40677   8.046 4.14e-15 ***
# NDVI          0.13716    0.05486   2.500   0.0127 *
# grass        10.41435    4.68962   2.221   0.0267 *
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 22.58 on 642 degrees of freedom
# Multiple R-squared:  0.117,    Adjusted R-squared:  0.1128
# F-statistic: 28.35 on 3 and 642 DF,  p-value: < 2.2e-16
```

]

Multiple linear regression in R

Verify diagnostic plots, as you have done for the simple linear regression.

```
par(mfrow = c(2, 2))
plot(lm.mult)
```



Find the best-fit model

Recall the principle of parsimony: we want to explain the most of the variance using the least number of terms as possible.

```
summary(lm.mult)$coefficients
#               Estimate Std. Error   t value     Pr(>|t|)
# (Intercept) -83.6081274 11.5774529 -7.221634 1.458749e-12
# clTma         3.2729872  0.4067706  8.046272 4.135118e-15
# NDVI          0.1371634  0.0548603  2.500231 1.265953e-02
# grass        10.4143451  4.6896157  2.220725 2.671787e-02
```

Parameters for all 3 predictor variables are significantly different from 0.

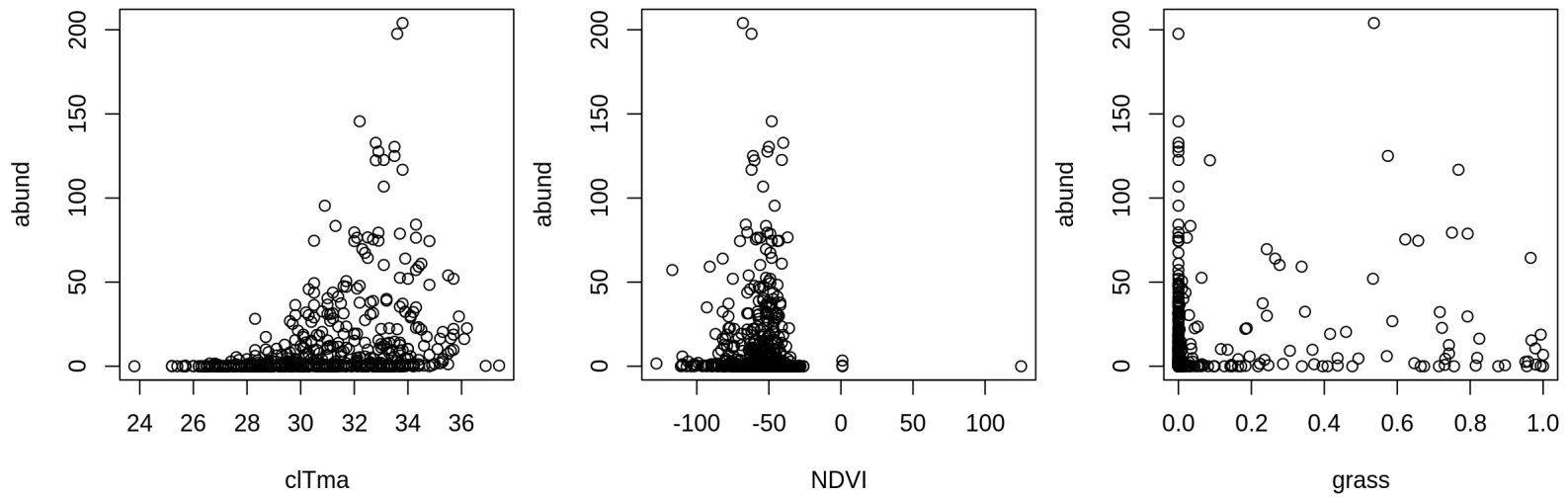
Model explains ~11% of variance in dickcissel abundance $R^2_{adj} = 0.11$.

However: this information is irrelevant because the assumptions of the linear model are not met.

Find the best-fit model

The response variable `abund` is not linearly related to the explanatory variables.

```
plot(abund ~ clTma, data = Dickcissel)
plot(abund ~ NDVI,  data = Dickcissel)
plot(abund ~ grass, data = Dickcissel)
```



*See **Advanced section** on **polynomial regression** for solution!*

Optional section

if time and disposition allows

Optional section

1. Interpreting contrasts
2. Unbalanced ANOVA
3. Polynomial regression
4. Variance partitioning

Interpreting contrasts

Interpreting contrasts

Contrasts compare each level of a factor to a baseline level. We can determine if each level of a factor are significantly different from each other.

The *intercept* is the baseline group and corresponds to the mean of the first (alphabetically) Diet level (`Insect`). **Add Intercept + coefficient estimates of each Diet level**

```
tapply(bird$logMaxAbund, bird$Diet, mean)
#      Insect  InsectVert       Plant PlantInsect  Vertebrate
#   1.1538937   0.5104603   1.3948941   1.5761940   0.8468898
coef(anov1)
#     (Intercept)  DietInsectVert       DietPlant DietPlantInsect  DietVertebrate
#       1.1538937      -0.6434334       0.2410004       0.4223003      -0.3070039
coef(anov1)[1] + coef(anov1)[2] # InsectVert
# (Intercept)
#   0.5104603
coef(anov1)[1] + coef(anov1)[3] # Plant
# (Intercept)
#    1.394894
```

What did you notice?

Interpreting contrasts

We may want to relevel the baseline:

1. Compare the Plant diet to all other diet levels

```
bird$Diet2 <- relevel(bird$Diet, ref="Plant")
anov2 <- lm(logMaxAbund ~ Diet2, data = bird)
summary(anov2)
anova(anov2)
```

1. Reorder multiple levels according to median

```
bird$Diet2 <- factor(bird$Diet, levels=names(med))
anov2 <- lm(logMaxAbund ~ Diet2,
            data = bird)
summary(anov2)
anova(anov2)
```

Did you see change in the significance of each Diet level?

Interpreting contrasts

NOTE: the DEFAULT contrast `contr.treatment` *is NOT orthogonal*

To be orthogonal:

- Coefficients must sum to 0
- Any two contrast columns must sum to 0

```
sum(contrasts(bird$Diet)[,1])
# [1] 1
sum(contrasts(bird$Diet)[,1]*contrasts(bird$Diet)[,2])
# [1] 0
```

Interpreting contrasts

Change the contrast to make levels orthogonal (e.g. Helmert contrast will contrast the second level with the first, the third with the average of the first two, and so on)

```
options(contrasts=c("contr.helmert", "contr.poly"))
sum(contrasts(bird$Diet)[,1])
# [1] 0
sum(contrasts(bird$Diet)[,1]*contrasts(bird$Diet)[,2])
# [1] 0
```

```
anov3 <- lm(logMaxAbund ~ Diet, data = bird)
summary(anov3)
#
# Call:
# lm(formula = logMaxAbund ~ Diet, data = bird)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -1.85286 -0.32972 -0.08808  0.47375  1.56075
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  1.09647    0.14629   7.495 1.14e-09 ***
```

Unbalanced ANOVA

Unbalanced ANOVA

A dataset is considered unbalanced when the sample sizes of two factor levels are not equal.

The `birdsdiet` data is actually unbalanced (number of `Aquatic` and `non-Aquatic` is not equal)

```
table(bird$Aquatic)
#
#  0  1
# 39 15
```

Which means the order of the covariates changes the values of Sums of Squares

```
unb.anov1 <- lm(logMaxAbund ~ Aquatic + Diet, data = bird)
unb.anov2 <- lm(logMaxAbund ~ Diet + Aquatic, data = bird)
```

Unbalanced ANOVA

```
anova(unb.anov1)
# Analysis of Variance Table
#
# Response: logMaxAbund
#           Df  Sum Sq Mean Sq F value  Pr(>F)
# Aquatic    1  0.2316 0.23157  0.5114 0.47798
# Diet       4  5.1926 1.29816  2.8671 0.03291 *
# Residuals 48 21.7337 0.45279
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(unb.anov2)
# Analysis of Variance Table
#
# Response: logMaxAbund
#           Df  Sum Sq Mean Sq F value  Pr(>F)
# Diet       4  5.1059 1.27647  2.8191 0.03517 *
# Aquatic    1  0.3183 0.31834  0.7031 0.40591
# Residuals 48 21.7337 0.45279
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Unbalanced ANOVA

To fix this problem, we can use a different approach to test the effects of each predictor.

Type I : Tests the effects in sequence, starting with the first predictor.

Type II: Tests for the presence of a main effect after the other main effect.

Type III: Tests for the presence of a main effect after the other main effect and interaction.

Type I is the default type used in R which creates our problem with unbalanced data.

If you are considering using Type II or III for your own dataset, you should read more about the subject. You can start with this link

Unbalanced ANOVA

Now try type III ANOVA using the `Anova()` function

```
car::Anova(unb.anov1, type = "III")
# Anova Table (Type III tests)
#
# Response: logMaxAbund
#              Sum Sq Df F value    Pr
# (Intercept) 18.9349  1 41.8186 4.837
# Aquatic      0.3183  1  0.7031   0.4
# Diet         5.1926  4  2.8671   0.0
# Residuals   21.7337 48
# ---
# Signif. codes:  0 '***' 0.001 '**' 0
```

```
car::Anova(unb.anov2, type = "III")
# Anova Table (Type III tests)
#
# Response: logMaxAbund
#              Sum Sq Df F value    Pr
# (Intercept) 18.9349  1 41.8186 4.837
# Diet         5.1926  4  2.8671   0.0
# Aquatic      0.3183  1  0.7031   0.4
# Residuals   21.7337 48
# ---
# Signif. codes:  0 '***' 0.001 '**' 0
```

*What have you noticed when using `Anova()`?*ff

Polynomial regression

Polynomial regression

As we noticed in the section on **multiple linear regression**, `MaxAbund` was non-linearly related to some variables

To test for non-linear relationships, polynomial models of different degrees are compared.

- A polynomial model looks like this:

$$\underbrace{2x^4}_{} + \underbrace{3x}_{} - \underbrace{2}_{}$$

this polynomial has 3 terms

Polynomial regression

For a polynomial with one variable (x), the *degree* is the largest exponent of that variable

this makes the degree 4

$$2x^{\overbrace{4}} + 3x - 2$$

Polynomial regression

When you know a degree, you can also give it a name

Degree	Name	Example
0	Constant	3
1	Linear	$x + 9$
2	Quadratic	$x^2 - x + 4$
3	Cubic	$x^3 - x^2 + 5$
4	Quartic	$6x^4 - x^3 + x - 2$
5	Quintic	$x^5 - 3x^3 + x^2 + 8$

Polynomial regression

Using the `Dickcissel` dataset, test the non-linear relationship between max abundance and temperature by comparing three sets of nested polynomial models (of degrees 0, 1, and 3):

```
lm.linear <- lm(abund ~ clDD, data = Dickcissel)
lm.quad   <- lm(abund ~ clDD + I(clDD^2), data = Dickcissel)
lm.cubic  <- lm(abund ~ clDD + I(clDD^2) + I(clDD^3), data = Dickcissel)
```

Polynomial regression

- Compare the polynomial models and determine which nested model we should keep
- Run a summary of this model, report the regression formula, p-values and R^2-adj

Polynomial regression

Compare the polynomial models; *which nested model we should keep?*

Run a summary of this model, report the regression formula, p-values and R^2-adj

```
print(summ_lm.linear)
# [1] "Coefficients:"
# [2] "            Estimate Std. Error t value Pr(>|t|)   "
# [3] "(Intercept) 1.864566   2.757554   0.676  0.49918   "
# [4] "clDD        0.001870   0.000588   3.180  0.00154 **"
# [5] "Multiple R-squared:  0.01546,\tAdjusted R-squared:  0.01393 "
# [6] "F-statistic: 10.11 on 1 and 644 DF,  p-value: 0.001545"
```

```
print(summ_lm.quad)
# [1] "Coefficients:"
# [2] "              Estimate Std. Error t value Pr(>|t|)    "
# [3] "(Intercept) -1.968e+01  5.954e+00  -3.306    0.001 ** "
# [4] "clDD         1.297e-02  2.788e-03   4.651 4.00e-06 ***"
# [5] "I(clDD^2)   -1.246e-06  3.061e-07  -4.070 5.28e-05 ***"
# [6] "Multiple R-squared:  0.04018,\tAdjusted R-squared:  0.0372 "
# [7] "F-statistic: 13.46 on 2 and 643 DF,  p-value: 1.876e-06"
```

```
print(summ_lm.cubic)
```

Variation Partitioning

Variation Partitioning

Some of the selected explanatory variables in the **multiple linear regression** section were highly correlated

Collinearity between explanatory variables can be assessed using the variance inflation factor `vif()` function of package `car`

- Variable with `VIF > 5` are considered collinearity

```
mod <- lm(clDD ~ clFD + clTmi + clTma + clP + grass, data = Dickcissel)
car::vif(mod)
#      clFD     clTmi     clTma       clP     grass
# 13.605855  9.566169  4.811837  3.196599  1.165775
```

Variation Partitioning

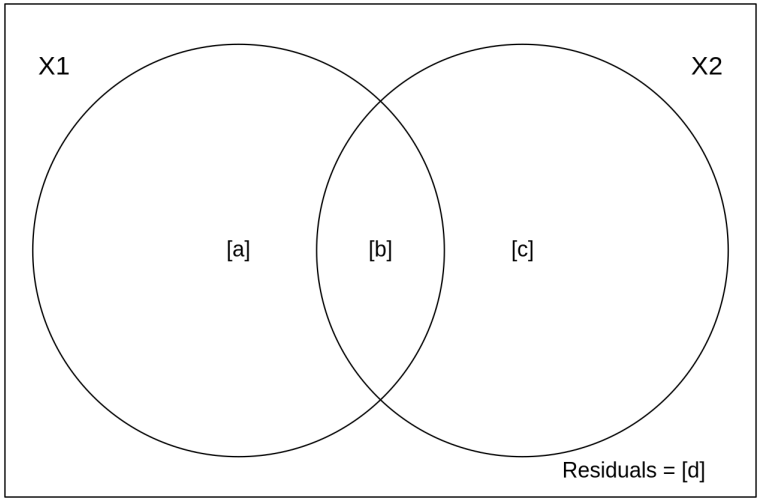
Use `varpart()` to partition the variation in max abundance with all land cover variables in one set and all climate variables in the other set (leaving out NDVI)

```
library(vegan)
part.lm = varpart(Dickcissel$abund, Dickcissel[ ,c("clDD
                  Dickcissel[ ,c("broadleaf","conif","gr
part.lm
#
# Partition of variance in RDA
#
# Call: varpart(Y = Dickcissel$abund, X = Dickcissel[, c
# "clTmi", "clTma", "clP")], Dickcissel[, c("broadleaf",
# "grass", "crop", "urban", "wetland")])
#
# Explanatory tables:
# X1:  Dickcissel[, c("clDD", "clFD", "clTmi", "clTma",
# X2:  Dickcissel[, c("broadleaf", "conif", "grass", "cr
#
# No. of explanatory tables: 2
# Total variation (SS): 370770
#              Variance: 574.84
# No. of observations: 646
#
```

Note: *Collinear variables do not have to be removed prior to partitioning*

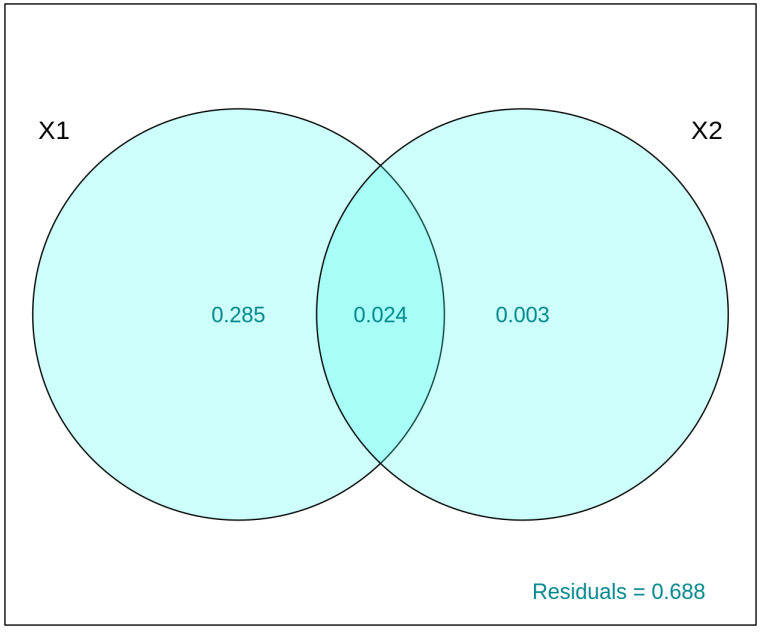
Variation Partitioning

```
showvarparts(2)
```




```
?showvarparts
# With two explanatory tables, the frac
# explained uniquely by each of the two
# are '[a]' and '[c]', and their joint
# is '[b]' following Borcard et al. (19
```

```
plot(part.lm,
     digits = 2,
     bg = rgb(48,225,210,80,
              maxColorValue=225),
     col = "turquoise4")
```



Proportion of variance explained by climate alone is 28.5% (given by X1|X2), by land cover alone is ~0% (X2|X1), and by both combined is 2.4%

Variation Partitioning

Test significance of each fraction

- Climate set

```
out.1 = rda(Dickcissel$abund,
            Dickcissel[ ,c("clDD", "clFD","clTmi","clTma","clP")],
            Dickcissel[ ,c("broadleaf","conif","grass","crop", "urban","wetland")])
```

- Land cover set

```
out.2 = rda(Dickcissel$abund,
            Dickcissel[ ,c("broadleaf","conif","grass","crop", "urban", "wetland")],
            Dickcissel[ ,c("clDD","clFD","clTmi", "clTma","clP")])
```

Variation Partitioning

```
# Climate set
anova(out.1, step = 1000, perm.max = 1(
# Permutation test for rda under reduce
# Permutation: free
# Number of permutations: 999
#
# Model: rda(X = Dickcissel$abund, Y =
#          Df Variance      F Pr(>F)
# Model     5   165.12 53.862  0.001 *
# Residual 634   388.72
# ---
# Signif. codes:  0 '***' 0.001 '**' 0
```

```
# Land cover set
anova(out.2, step = 1000, perm.max = 1(
# Permutation test for rda under reduce
# Permutation: free
# Number of permutations: 999
#
# Model: rda(X = Dickcissel$abund, Y =
#          Df Variance      F Pr(>F)
# Model     6     5.54 1.5063  0.167
# Residual 634   388.72
```

Conclusion: the land cover fraction is non-significant once climate data is accounted for

Thank you for attending!