



# Atelier 4: Modèles linéaires

Série d'ateliers R du CSBQ

Centre de la Science de la Biodiversité du Québec



# À propos de cet atelier

 REPO

 DEV

 WIKI

04



DIAPOS

04



DIAPOS

04

 SCRIPT

04

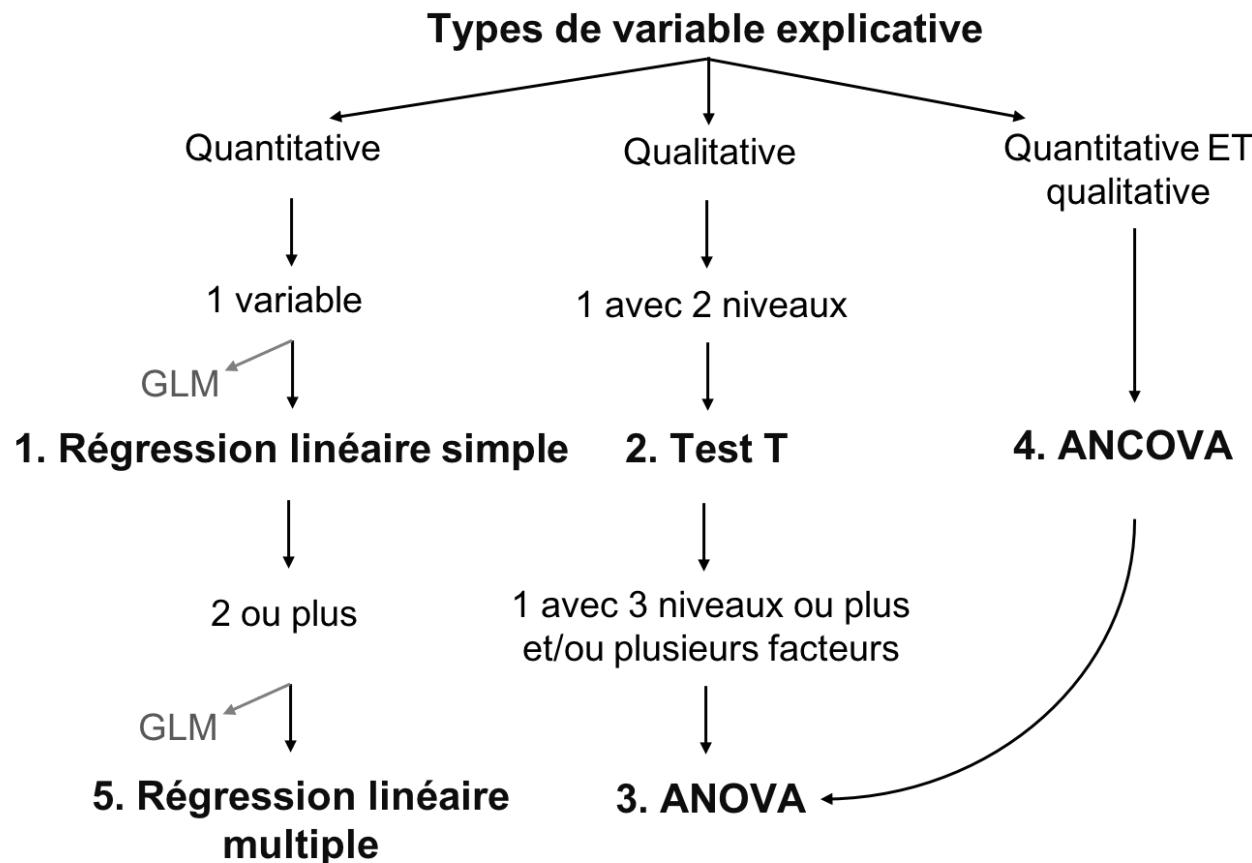
# Packages requis

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

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

# Objectifs d'apprentissage

-Apprendre la structure d'un modèle linéaire et ses *différentes variantes*.



# Objectifs d'apprentissage

- Apprendre la structure d'un modèle linéaire et ses différentes variantes.
- Apprendre comment faire un modèle linéaire dans R avec `lm()` et `anova()`
- Apprendre comment identifier un modèle dont les conditions d'application ne sont pas rencontrées et comment régler le problème.

# Qu'est-ce qu'un modèle linéaire ?

## Un modèle linéaire ...

... décrit la relation entre une variable (la **réponse**) et une ou plusieurs autres variables (les **prédicteurs**).

... est utilisé pour analyser une **hypothèse bien formulée**, souvent associée à une question de recherche plus générale.

... est utilisé pour faire des inférences sur la **direction** et la **force** d'une relation, et notre **confiance** dans les estimations de l'effet.

# Exemple : Abondance et masse des espèces d'oiseaux

## Hypothèse

Pour différentes espèces d'oiseaux, la masse moyenne d'un individu a un effet sur l'abondance maximale de l'espèce, en raison de contraintes écologiques (sources de nourriture, disponibilité de l'habitat, etc.).

## Prédiction

Les espèces caractérisées par des individus plus grands ont une abondance maximale plus faible.

### Discussion en groupe

*Quelle variable est la réponse ? Quelle est le prédicteur ?*

*Quelles sont nos attentes concernant la direction et la force de la relation ?*

# Exemple : Abondance et masse des espèces d'oiseaux

## Regardons les données ...

Importer le jeu de données "birdsdiet" :

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

Visualiser le tableau de la structure des données en utilisant la fonction `str()` :

```
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 ...
```

# Exemple : Abondance et masse des espèces d'oiseaux

Regardons les données ...

Mesures communes de **localisation** (tendance centrale) :

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

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

- **Médiane** (valeur séparant la moitié supérieure de la moitié inférieure d'un échantillon)

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

Mesure communes de **variation** (dispersion) :

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

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

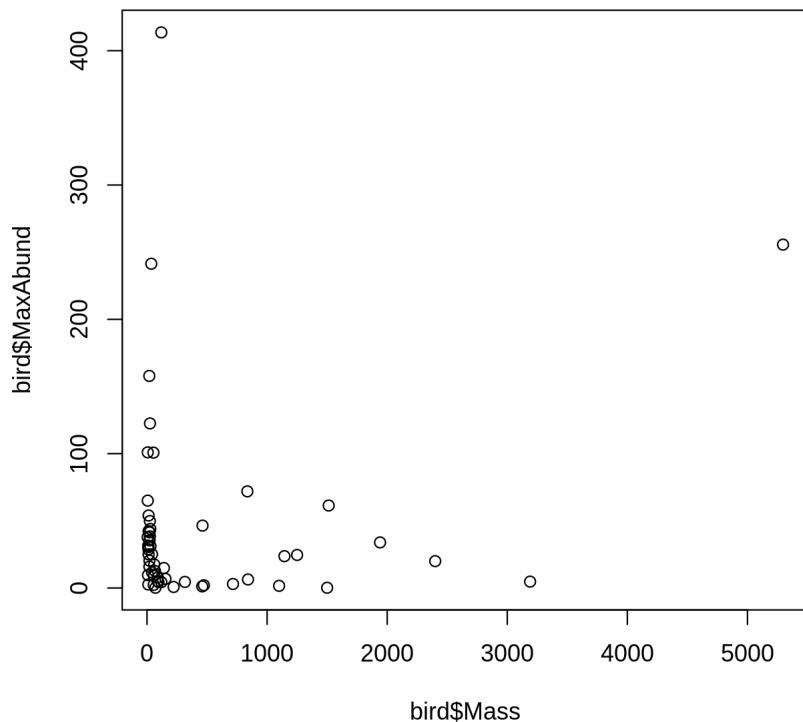
- **Écart type**  $\sigma$

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

# Exemple : Abondance et masse des espèces d'oiseaux

Tracer la réponse en fonction du prédicteur:

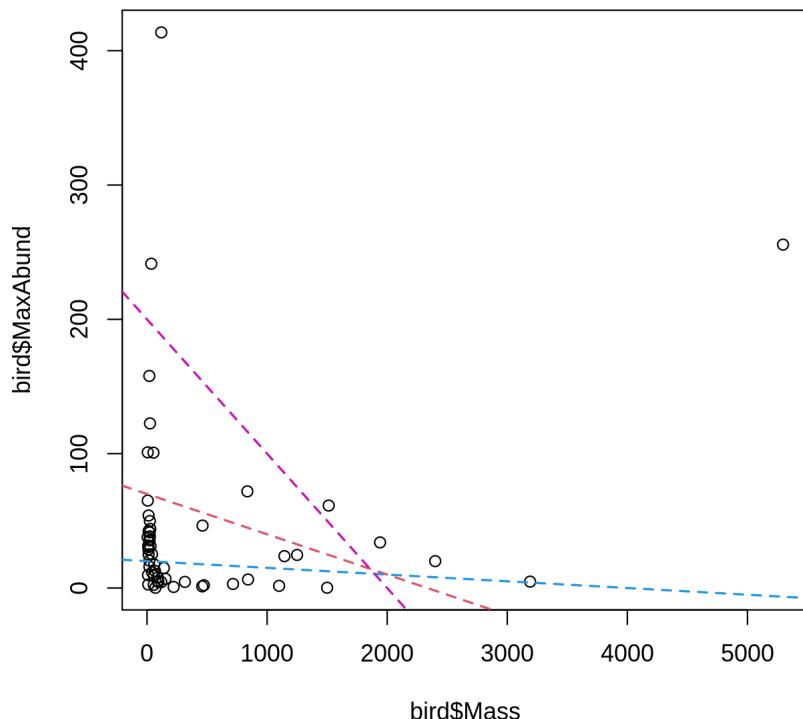
```
plot(bird$Mass, bird$MaxAbund)
```



# Exemple : Abondance et masse des espèces d'oiseaux

Comment trouver la "meilleure" estimation de la relation ?

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



# Formulation d'un modèle linéaire

## Variables

- $y_i$  est une observation de la **réponse  $y$**   
(par exemple, l'abondance maximale des espèces  $i$ )
- $x_i$  est une observation correspondante du **prédicteur  $x$**   
(par exemple, le poids moyen d'un individu d'une espèce  $i$ )

## Relation supposée

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

- Le paramètre  $\beta_0$  est **l'ordonnée à l'origine** (ou constante)
- Le paramètre  $\beta_1$  quantifie **l'effet** de  $x$  sur  $y$ .
- Le résidu  $\epsilon_i$  représent la variation **non expliquée**
- La **valeur prédictive** de  $y_i$  se définit comme :  $\hat{y}_i = \beta_0 + \beta_1 \times x_i$

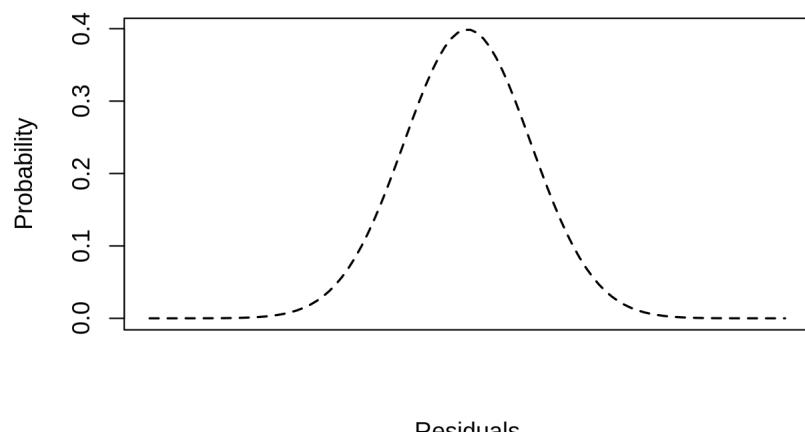
# Conditions d'application du modèle linéaire

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

## Distribution normale

Les **résidus**  $\epsilon$  suivent une **distribution normale** avec une moyenne de 0 et une variance de  $\sigma^2$

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



# Conditions d'application du modèle linéaire

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

## Distribution normale

Les **résidus**  $\epsilon$  suivent une **distribution normale** avec une moyenne de 0 et une variance de  $\sigma^2$

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

**Cela veut dire :** Chaque observation  $y_i$  suit une distribution normale, avec moyenne  $\hat{y} = \beta_0 + \beta_1 \times x_i$  et variance  $\sigma^2$ :

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

**Cela ne veut pas dire** que l'ensemble des valeurs observées  $y$  ~~doit suivre une distribution normale~~.

# Conditions d'application du modèle linéaire

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

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

## Homoscédasticité

- Tous les résidus  $\epsilon$  suivent la même distribution, la **variance  $\sigma^2$  reste constante**.

## Indépendance des résidus

- Chaque résidu  $\epsilon_i$  est **indépendant** de tout autre résidu.

# Conditions d'application du modèle linéaire

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

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

## Résumé des conditions d'application

- Relation linéaire entre la réponse et le prédicteur
- Les résidus suivent une distribution normale avec une moyenne de **0**
- Les résidus sont distribués de manière identique (*homoscédasticité*)
- Les résidus sont indépendants les uns des autres

# Notation des modèles linéaires

## Notation mathématique (pour des manuscrits)

- Observations individuelles :

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

- Toutes les observations (notation matricielle, interception incluse dans  $\mathbf{X}$  et  $\boldsymbol{\beta}$ ) :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad \text{with} \quad \epsilon_i \sim \mathcal{N}(0, I_n \sigma^2)$$

## Notation en R

- Formule du modèle :

```
y ~ 1 + x
```

- Ou encore plus simple :

```
y ~ x
```

(inclut aussi la constante)

**Il ne faut jamais mélanger les différentes types de notation !**

# Effectuer une modèle linéaire

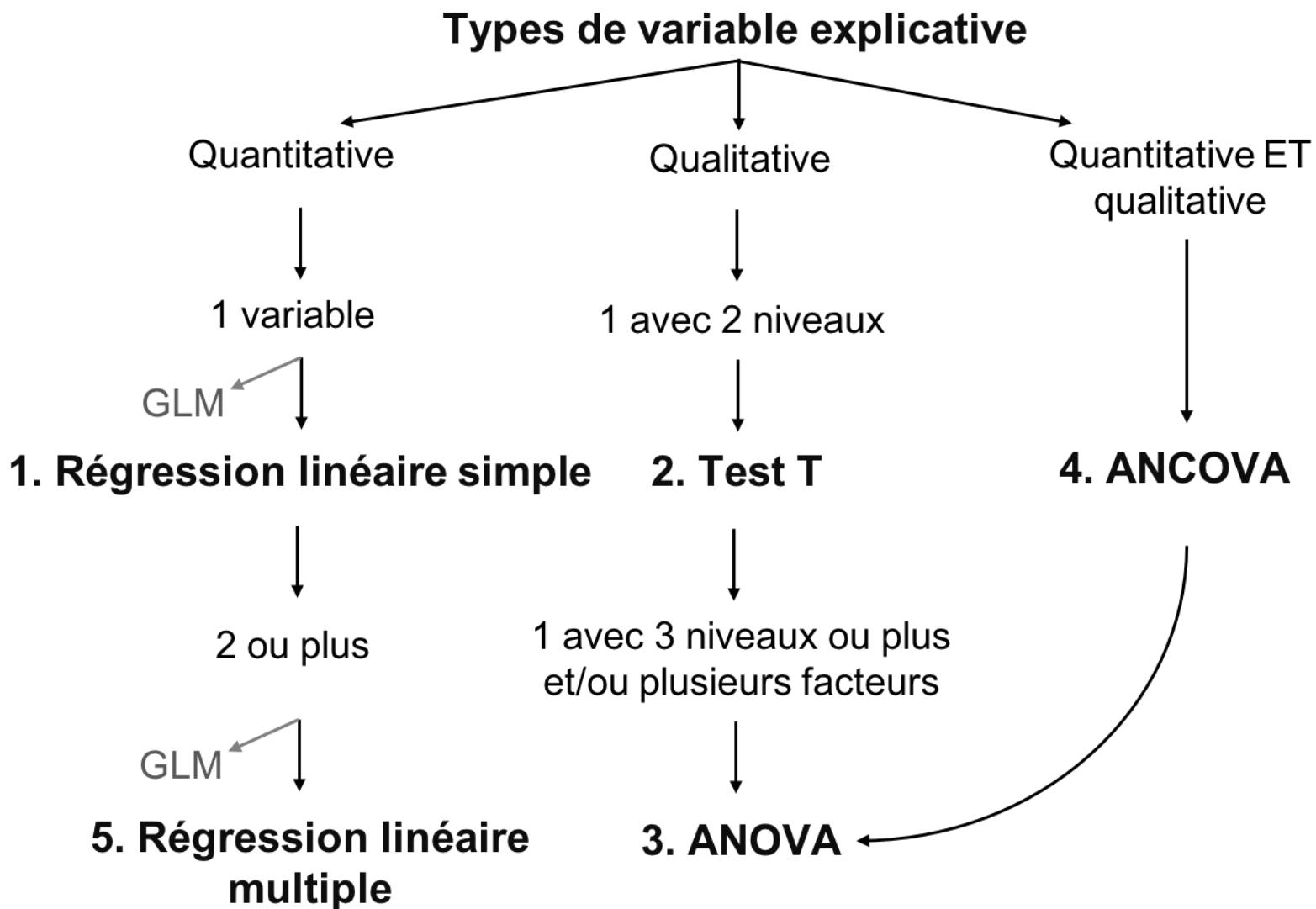
$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i$$

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

## Estimation du modèle

- Trouver les "meilleures" estimations des paramètres  $\beta_0, \beta_1$ .
- Les "meilleurs" paramètres sont ceux qui minimisent la somme des résidus au carré  $\sum \epsilon_i^2$
- Cette méthode est appelée la méthode de **moindres carrés ordinaires** (MCO)

# Modèles linéaires

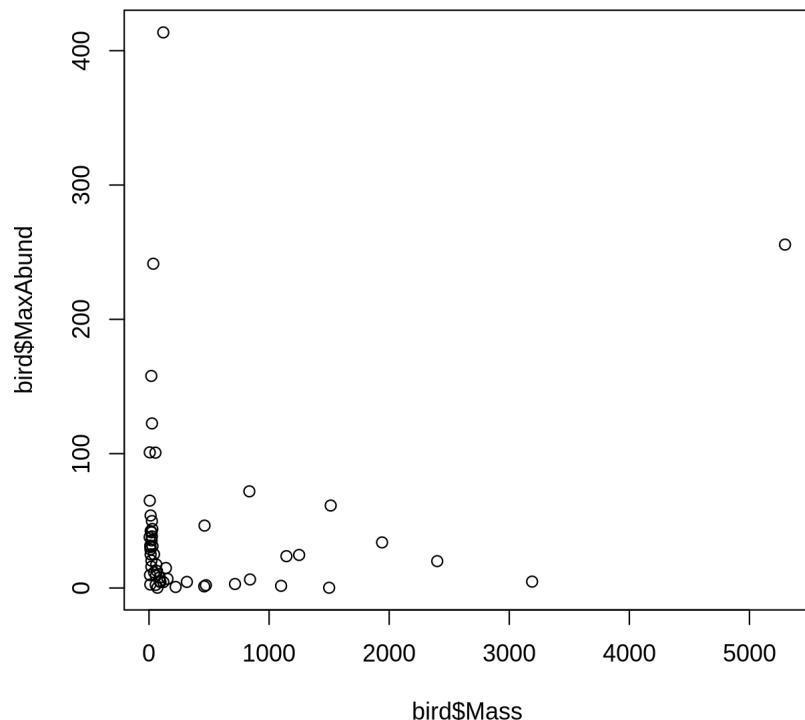


# Régression linéaire avec R

# Régression linéaire avec R

Revenons sur les oiseaux ...

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



# Régression linéaire avec R

## Formulation du modèle

*Hypothèse:* Pour différentes espèces d'oiseaux, la **masse moyenne d'un individu a un effet sur l'abondance maximale** de l'espèce, en raison de contraintes écologiques (sources de nourriture, disponibilité de l'habitat, etc.).

## Équation du modèle

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

## Formule du modèle en R

MaxAbund ~ Mass

# Régression linéaire avec R

## Étape 1

Formuler et exécuter un modèle linéaire basé sur un hypothèse

## Étape 2

Vérifier les conditions d'application du modèle linéaire



*Conditions sont satisfaites ?*

## Étape 3

- Analyser les paramètres de régression
- Tracer le modèle
- Effectuer des tests de signification sur les estimations des paramètres (si nécessaire)



*Conditions non satisfaites ?*

Envisager l'utilisation d'un *Modèle linéaire généralisé (GLM)* ou la transformation des données



Utiliser un GLM mieux adapté aux données



Retourner à l'Étape 1 avec des variables transformées

# Régression linéaire avec R

## Étape 1. Formuler et exécuter un modèle linéaire

La fonction `lm()` est utilisée pour ajuster un modèle linéaire, en fournissant la formule du modèle comme premier argument ::

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

- `lm1` : Nouvel objet contenant le modèle linéaire
- `MaxAbund ~ Mass` : Formule du modèle
- `bird` : objet contenant les variables

# Régression linéaire avec R

## Étape 1. Formuler et exécuter un modèle linéaire

Examinons les estimations des paramètres :

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

*Comment les paramètres se comparent-ils à nos prédictions ?*

**Peut-on se fier aux estimations du modèle ?**

# Régression linéaire avec R

## Étape 2. Vérifier les conditions d'application avec les graphiques diagnostics

Nous pouvons produire **quatre graphiques diagnostics** d'un objet `lm` :

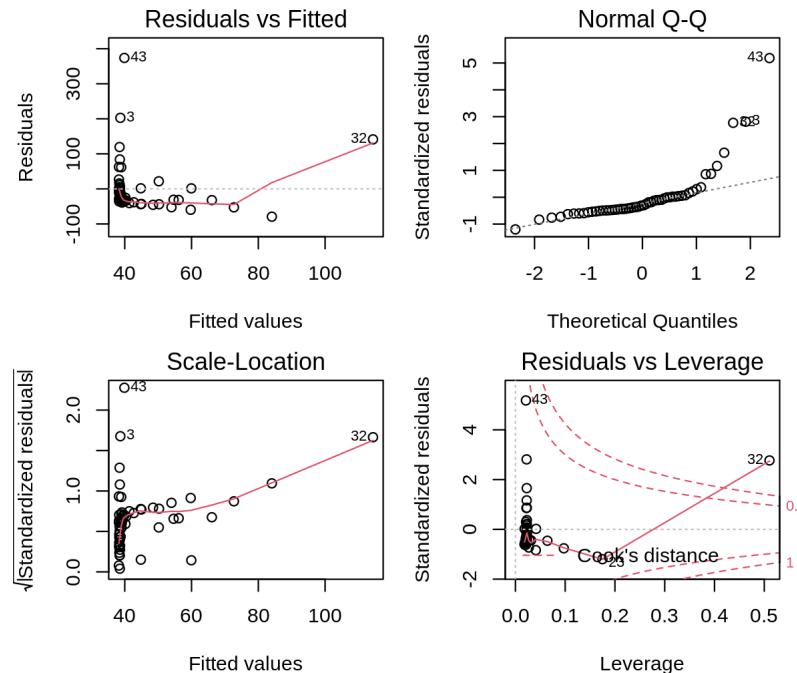
```
par(mfrow=c(2, 2))  
plot(lm1)
```

- `par()` : Fonction pour définir les paramètres du graphique
- `mfrow=c(2, 2)`: Paramètre graphique permettant d'afficher une grille de  $2 \times 2$  graphiques à la fois
- `plot()`: Fonction pour produire les graphiques

# Régression linéaire avec R

## Étape 2. Vérifier les conditions d'application avec les graphiques diagnostics

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



**Comment interpréter ces graphiques ?**

# Graph. #1 - Résidus vs valeurs prédictes

**Ce qu'on voit :**

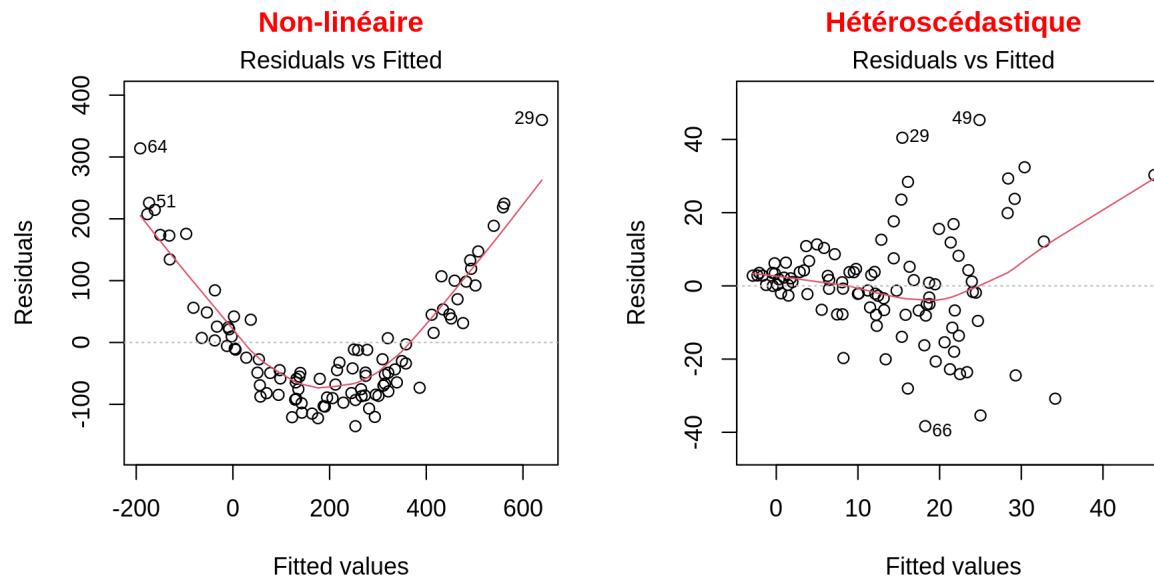
- Axe des **Y** : Résidus  $\epsilon_i$
- Axe des **X** : Valeurs prédictes  $\hat{y}_i = \beta_0 + \beta_1 \times x_i$

**Ce qu'on espère voir :** Dispersion de points sans patron

**Motivation :** Indication si les résidus sont *indépendants et uniformément distribués*.

# Graphique # 1 - Résidus vs valeurs prédictées

Ce qui devrait nous rendre méfiants :



Quoi faire ?

- Utiliser plutôt un **modèle linéaire généralisé** (MLG) qui permet d'autres distributions : Poisson, binomial, binomial négatif, etc.)
- Essayer de **transformer** la réponse et/ou prédicteurs

# Graphique # 2 - Échelle localisé

## Ce qu'on voit :

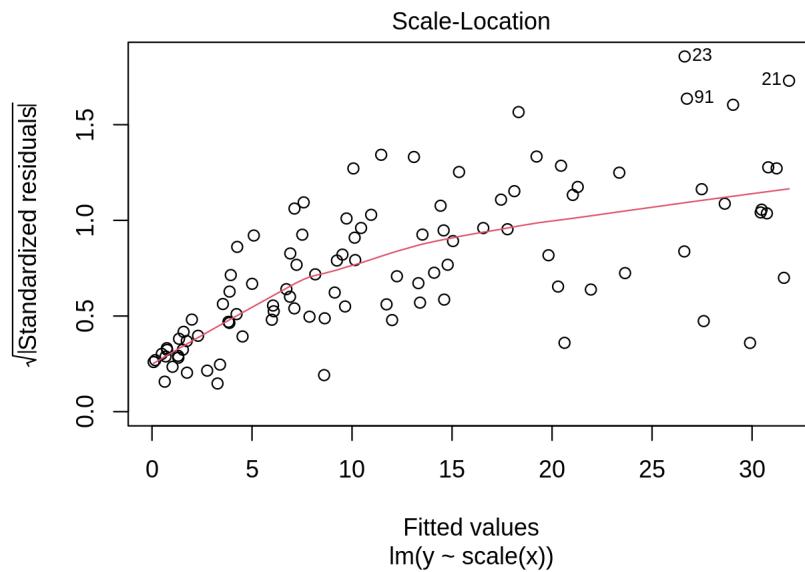
- Axe des **Y** : Racine carrée des résidus standardisés  $\sqrt{\frac{\epsilon_i}{\sigma}}$
- Axe des **X** : Valeurs prédictes  $\hat{y}_i = \beta_0 + \beta_1 \times x_i$

**Ce qu'on espère voir :** Dispersion de points sans patron

**Motivation :** Parfois plus facile de détecter si les conditions d'application ne sont pas respectées, surtout quand le prédicteur est distribué de manière inégale.

# Graphique # 2 - Échelle localisé

Ce qui devrait nous rendre méfiants :



*Forte tendance dans les résidus*

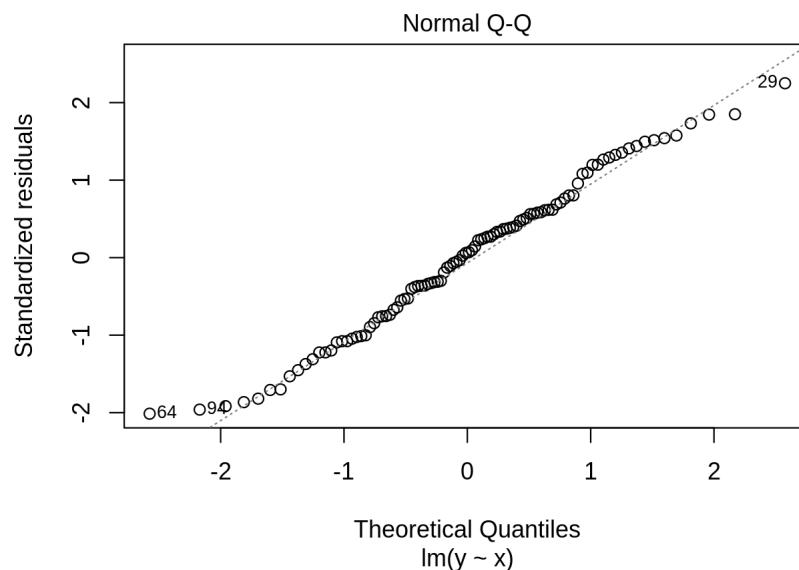
# Graphique # 3 - Normal QQ

Ce qu'on voit :

- Y-axis: Residus standardisés  $\frac{\epsilon_i}{\sigma}$
- X-axis: Distribution normale standard  $\mathcal{N}(0, \sigma^2)$

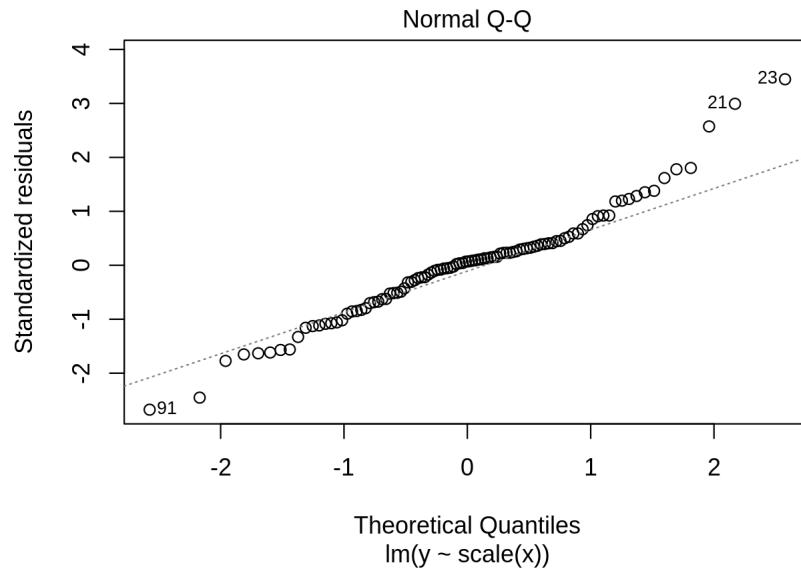
Ce qu'on espère voir : Points sur la ligne 1:1

**Motivation :** Comparer la distribution (quantiles) des résidus à une distribution normale standard



# Graphique # 3 - Normal QQ

Ce qui devrait nous rendre méfiants :



Les résidus ne suivent pas une distribution normale

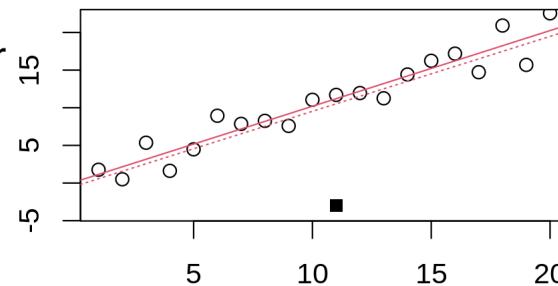
# Graphique # 4 - Résidus vs effet de levier

## Motivation :

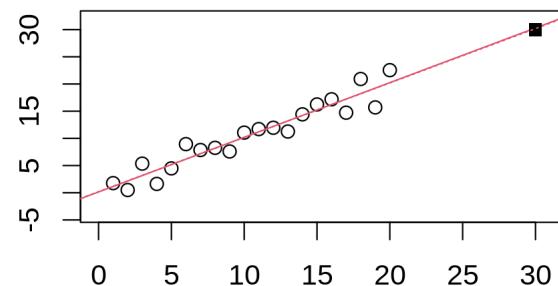
- Le modèle ne devrait **pas dépendre fortement d'observations isolées**.
- Les **points de levier** sont des observations extrêmes du prédicteur.
- Le **modèle passe près des points de levier**, car ils manquent d'observations voisines.
- Les points de levier **peuvent (ou pas) avoir une grande influence sur la régression**
- L'influence peut être quantifiée par **la distance Cook : plus de 0,5 est problématique**.

# Exemples: Effet levier et influence

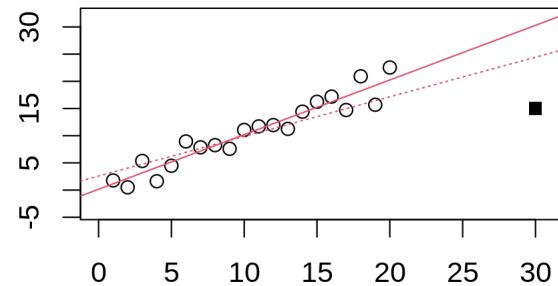
- \* Pas d'effet de levier
- \* Faible influence



- \* Effet de levier
- \* Pas d'influence



- \* Effet de levier
- \* Influence élevée

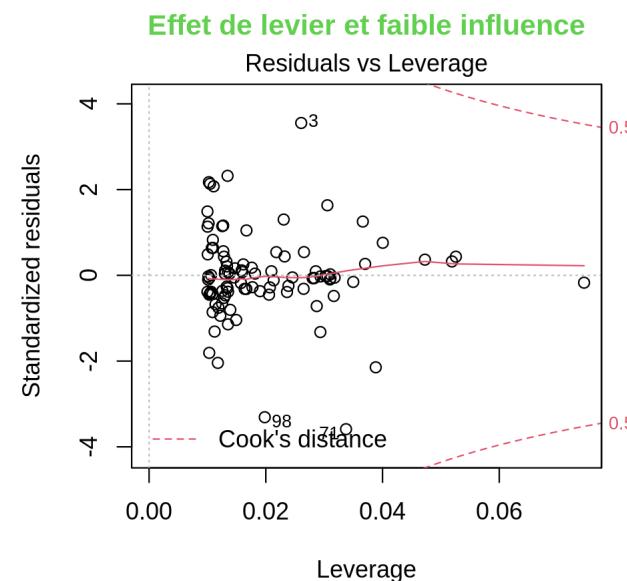
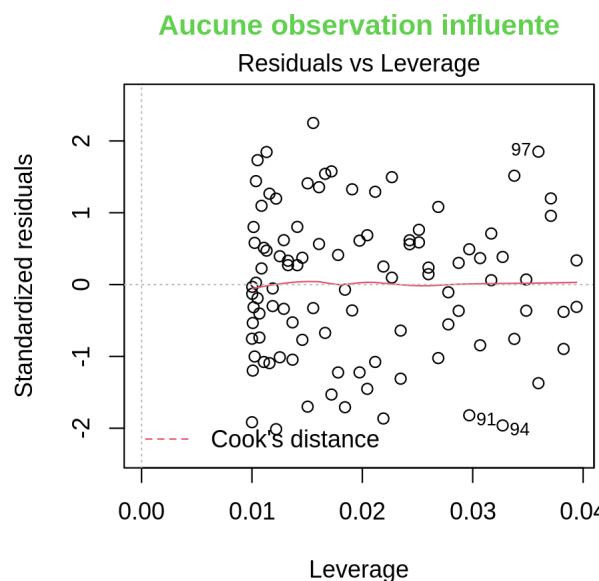


# Graphique # 4 - Résidus vs effet de levier

## Ce qu'on voit :

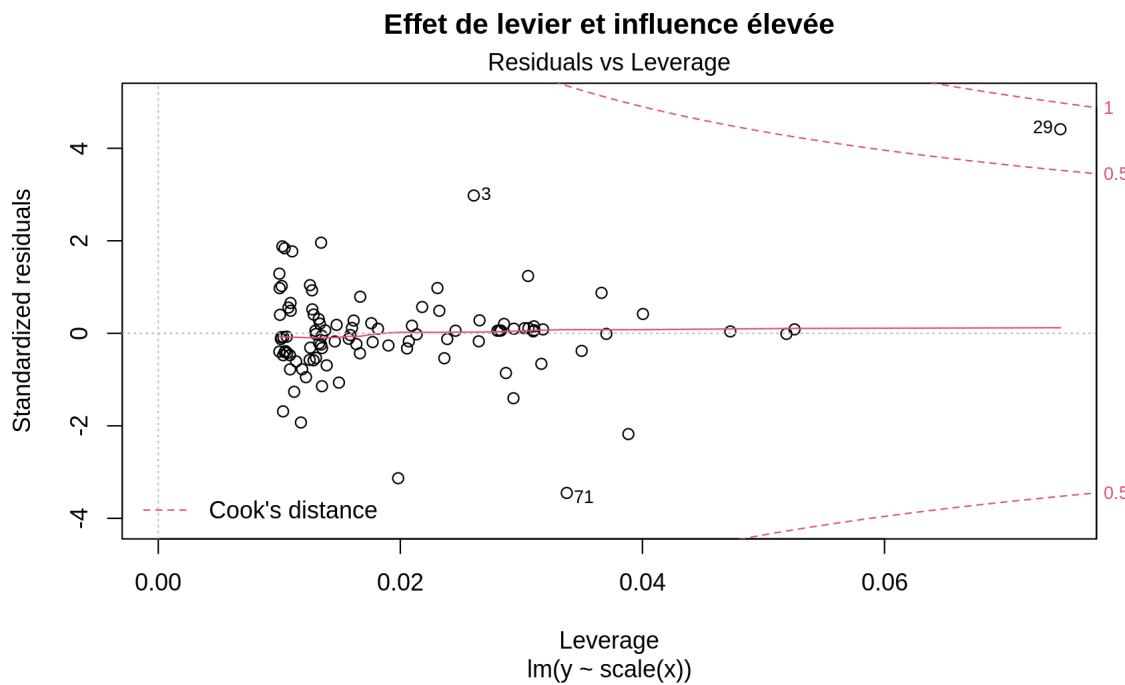
- Axe des **Y** : Residus standardisés  $\frac{\epsilon_i}{\sigma}$
- Axe des **X** : Effet de levier
- Ligne rouge en tirets : distance Cook de 0.5

**Ce qu'on espère voir :** Pas de points de levier avec influence élevée



# Graphique # 4 - Résidus vs effet de levier

Ce qui devrait nous rendre méfiants :

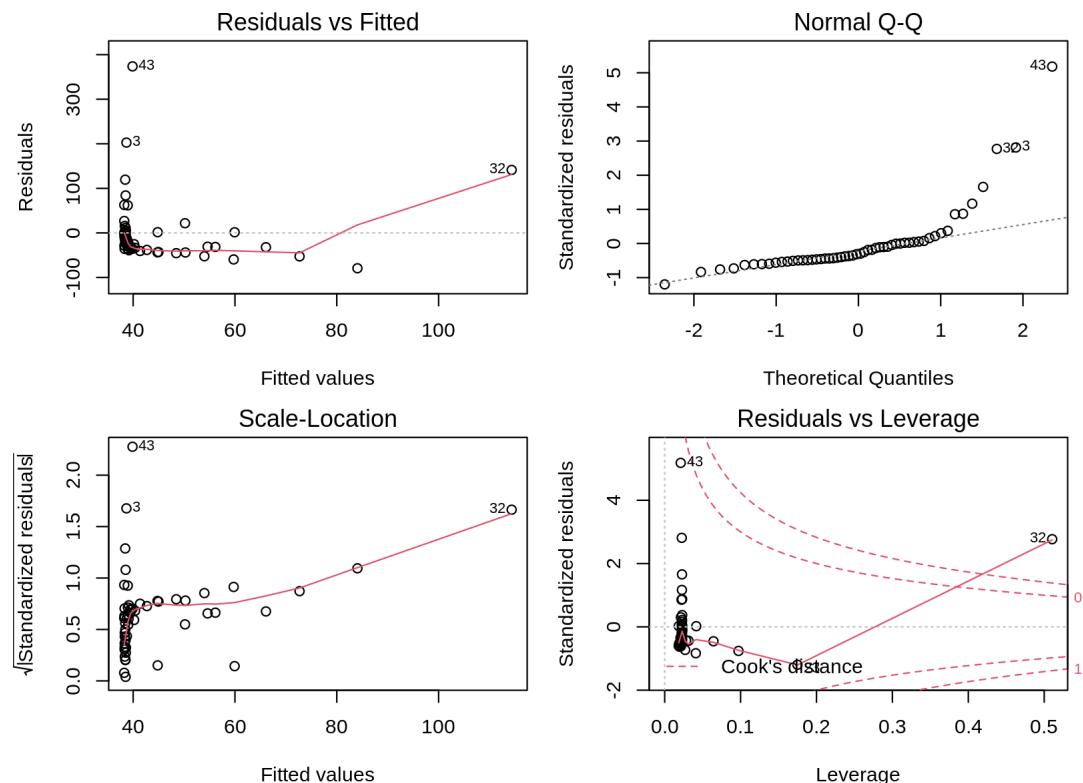


Il ne faut jamais supprimer les valeurs aberrantes sans avoir des bonnes raisons de le faire

# Étape 2. Vérifier les conditions d'application pour lm1

lm1

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

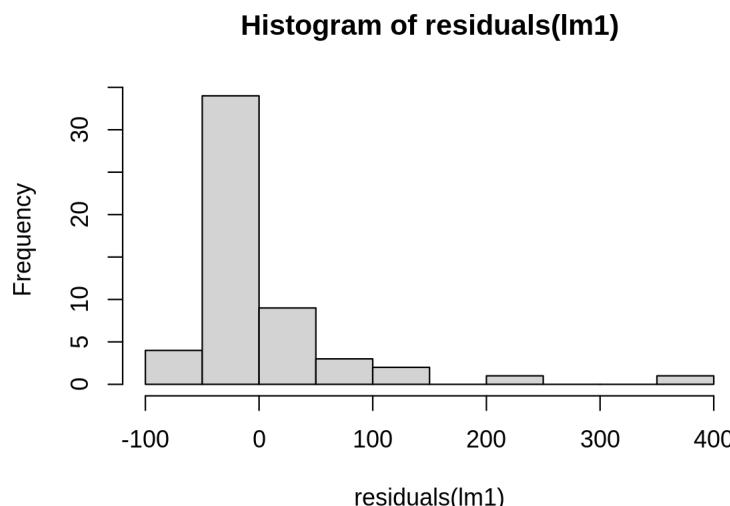
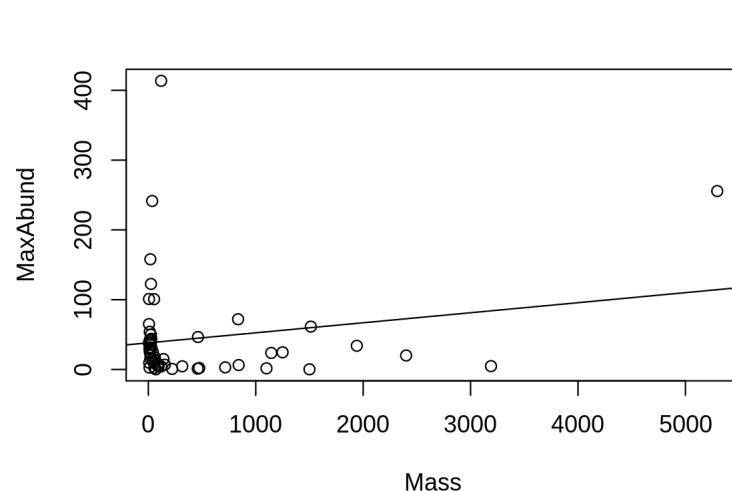


**Discussion :** Le modèle lm1 respecte-t-il les conditions du modèle linéaire ?

# Conditions non-respectées - Quelle est la cause ?

Traçons le modèle avec les observations ...

```
par(mfrow = c(1, 2))
coef(lm1) # constante et pente
# (Intercept)      Mass
# 38.16645523  0.01438562
plot(MaxAbund ~ Mass, data=bird) # graphique à gauche
abline(lm1) # ligne définie par les paramètres du modèle
hist(residuals(lm1)) # graphique à droite : distribution des résidus
```



# Conditions non-respectées - Quelle est la cause ?

On peut vérifier si les résidus suivent une distribution normale à l'aide d'un test de *Shapiro-Wilk* et d'un test d'asymétrie (*skewness*) :

```
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
```

*La distribution est significativement différente d'une distribution normale, décalée vers la gauche (valeur positive d'asymétrie)*

# Conditions non-respectées - Comment proéeder ?

*Il y a deux options quand les conditions d'application du modèle linéaire ne sont pas respectées :*

1. Utiliser un **autre type de modèle** mieux adapté à l'hypothèse et aux données (ateliers 6 - 8 du CSBQ R).
2. Essayer de **transformer** la réponse et / ou le prédicteurs
  - Il existe **plusieurs types de transformations** et leur utilité dépend de la distribution de la variable et du type de modèle.
  - La transformation peut **régler certains** problèmes mais peut en **créer d'autres**.
  - Les **résultats des tests de signification** sur les données transformées ne sont **pas automatiquement valables** pour les données non transformées.

# Défi 1: Un modèle sur variables transformées



Essayons de résoudre nos problèmes avec une transformation logarithmique.

Ajoutons des variables transformées à notre jeu de données :

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

## Défi

**Étape 1.** Exécuter une régression linéaire sur les variables transformées `logMaxAbund` et `logMass`. Sauvegarder l'objet du modèle sous `lm2`

**Étape 2:** Vérifier les conditions pour `lm2` en utilisant les graphiques diagnostiques.

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

# Défi 1: Un modèle sur variables transformées



**Étape 1.** Exécuter une régression linéaire sur les variables transformées

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

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

*Comment les paramètres se comparent-ils à nos prédictions ?*

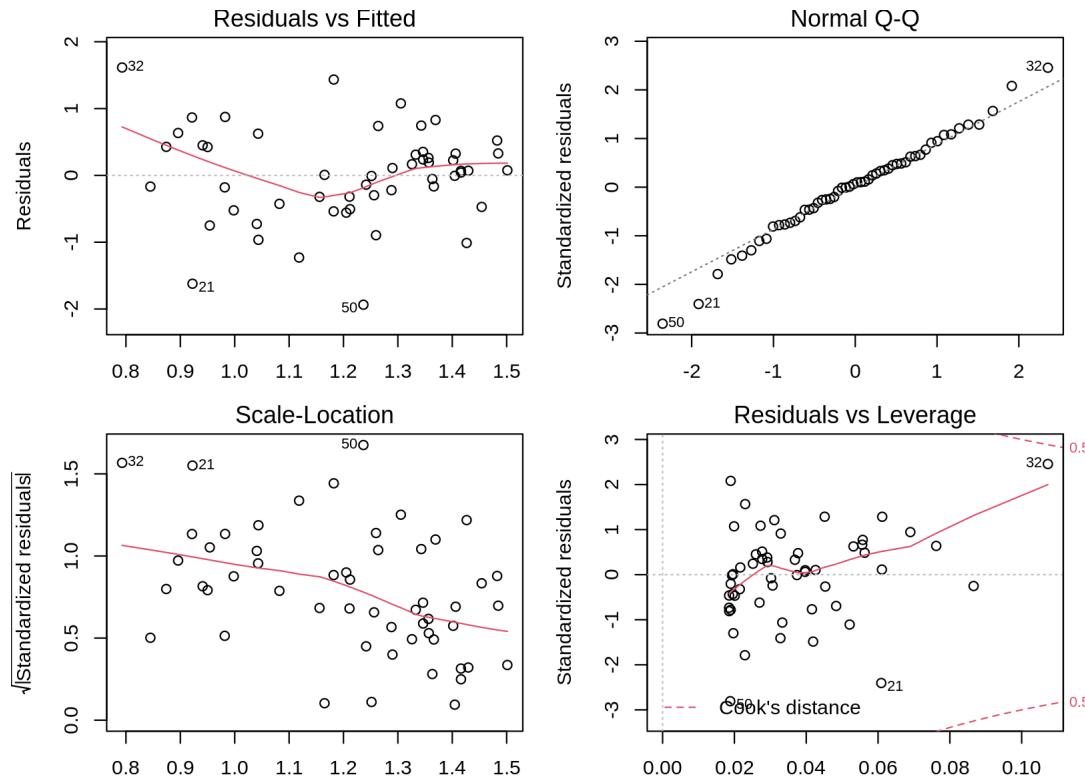
**Peut-on se fier aux estimations du modèle ?**

# Défi 1: Un modèle sur variables transformées



Étape 2: Vérifier les conditions pour `lm2`

```
par(mfrow=c(2, 2), mar=c(3, 4, 1.15, 1.2))  
plot(lm2)
```

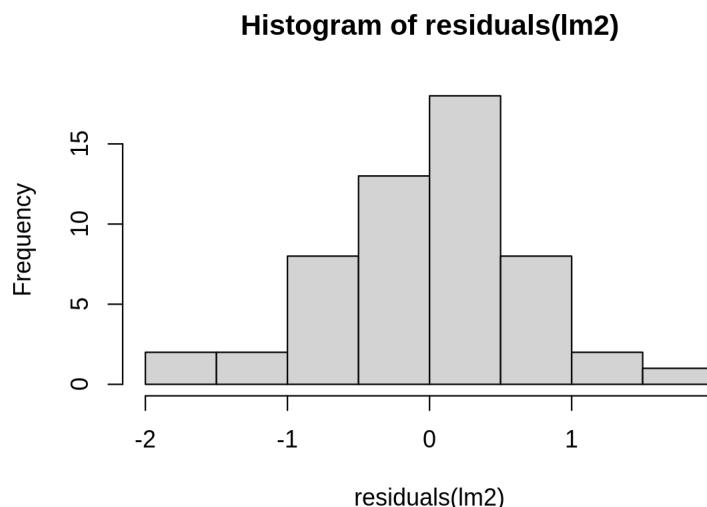
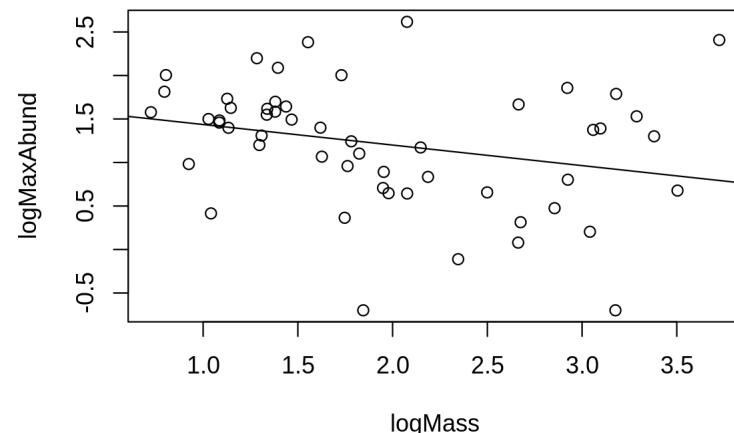


*Beaucoup mieux, mais il reste des problèmes*

# Étape 2. Vérifier les conditions d'application pour lm2

lm2

```
par(mfrow = c(1, 2))
coef(lm2) # constante et pente
# (Intercept)      logMass
# 1.6723673 -0.2361498
plot(logMaxAbund ~ logMass, data=bird) # graphique à gauche
abline(lm2) # ligne définie par les paramètres du modèle
hist(residuals(lm2)) # graphique à droite : distribution des résidus
```



# Étape 3. Analyser les paramètres

La fonction `summary()` est utilisée pour obtenir plus d'informations sur le modèle ajusté.

```
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

# Étape 3. Analyser les paramètres

Nous pouvons aussi extraire les paramètres du modèle et des autres résultats :

```
# Vecteurs de résidus et valeurs prédictives
e <- residuals(lm2)
y <- fitted(lm2)

coefficients(lm2) # coefficients
# (Intercept) logMass
# 1.6723673 -0.2361498

summary(lm2)$coefficients # coefficients avec test de t
#             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 # R au carré ajusté
# [1] 0.05483696
```

# Interprétation du modèle

*Dans quelle mesure le modèle soutient-il notre hypothèse ?*

## Hypothèse

Pour différentes espèces d'oiseaux, la **masse moyenne d'un individu a un effet sur l'abondance maximale** de l'espèce, en raison de contraintes écologiques (sources de nourriture, disponibilité de l'habitat, etc.).

# Interprétation du modèle

*Dans quelle mesure le modèle soutient-il notre hypothèse ?*

```
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

# Interprétation du modèle

*Dans quelle mesure le modèle soutient-il notre hypothèse ?*

Il n'y a que très **peu de preuves à l'appui** de notre hypothèse parce que :

- Le modèle n'explique pas bien la réponse (*faible R au carré ajusté*)
- Le modèle n'est que légèrement meilleur qu'un modèle sans variables prédictives (*F-test à peine significatif*)
- L'estimation du paramètre "logMass" est à peine différente de 0 (*valeur de t à peine significatif*)

# Trouver un meilleur modèle : oiseaux terrestres

*Peut-être devrions-nous formuler une hypothèse plus précise ?*

## Hypothèse

Pour différentes espèces d'oiseaux **terrestres**, la **masse moyenne d'un individu a un effet sur l'abondance maximale** de l'espèce, en raison de contraintes écologiques (sources de nourriture, disponibilité de l'habitat, etc.).

# Trouver un meilleur modèle : oiseaux terrestres

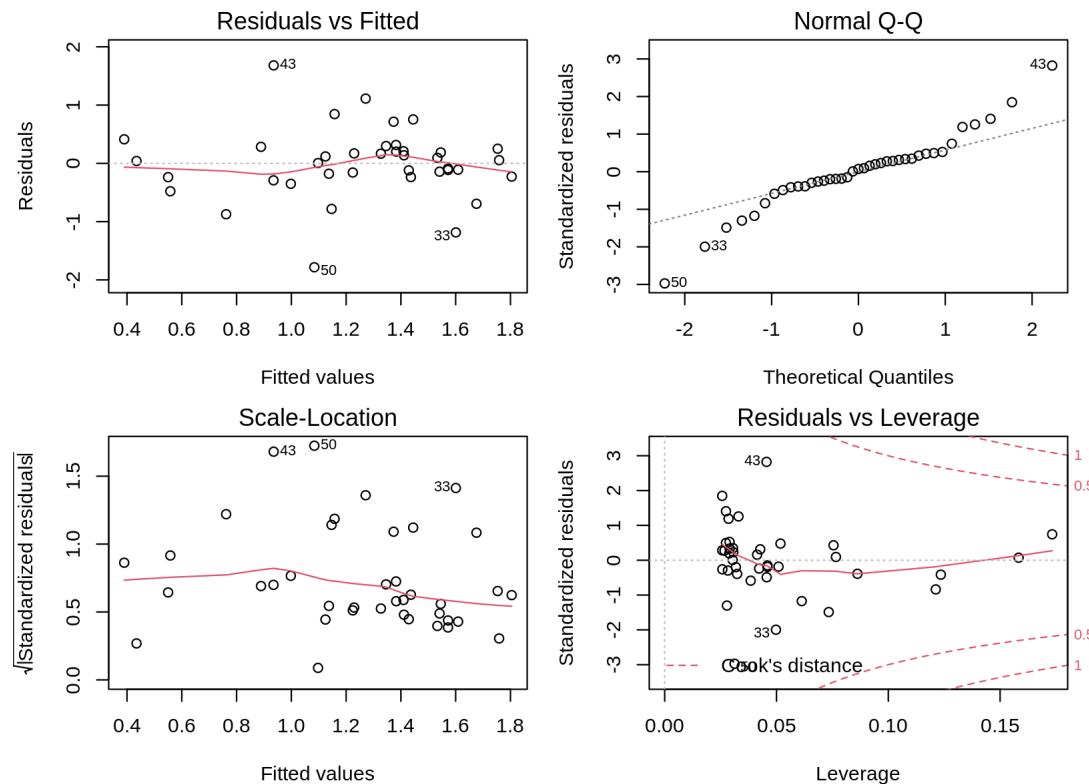
Exclure tous les oiseaux aquatiques (en utilisant `!`) et ajuster un modèle linéaire :

```
lm3 <- lm(logMaxAbund~logMass, data=bird, subset=!bird$Aquatic)
# exclut les oiseaux aquatiques (!birdsAquatic == TRUE)
# ou de façon équivalente :
# lm3 <- lm(logMaxAbund~logMass, data=bird, subset=bird$Aquatic == 0)

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

# Trouver un meilleur modèle : oiseaux terrestres

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



*Conditions d'application respectées*

# Trouver un meilleur modèle : oiseaux terrestres

*Dans quelle mesure le modèle soutient-il notre hypothèse ?*

```
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

# Trouver un meilleur modèle : oiseaux terrestres

*Dans quelle mesure le modèle soutient-il notre hypothèse ?*

Le modèle fournit des **preuves à l'appui** de notre hypothèse, parce que :

- Le modèle est raisonnablement bien ajusté aux données ( $R$  au carré ajusté)
- Le modèle est clairement meilleur qu'un modèle sans variables prédictives ( $F$ -test)
- L'estimation du paramètre "logMass" est clairement différente de 0 ( $t$ -test)



# Défi 2

Rassemblons tout les étapes :

1. Formuler une autre hypothèse similaire sur l'abondance maximale et la masse moyenne d'un individu, cette fois pour les **passereaux** ("passerine birds").
2. Ajuster un **modèle** pour évaluer cette hypothèse, en utilisant les variables transformées (c'est-à-dire `logMaxAbund` et `logMass`). Sauvegarder le modèle sous le nom de `lm4`.
3. **Vérifier les conditions d'application** du modèle linéaire à l'aide des graphiques diagnostics.
4. Interpréter les résultats : Le modèle fournit-il des **preuves à l'appui de l'hypothèse ?**

Indice : Comme les espèces aquatiques, les passereaux (variable `Passerine`) sont codées 0/1 (vérifier avec `str(bird)`)

# Défi 2 - Solution



## Hypothèse

Pour différentes espèces de **passereaux**, la **masse moyenne d'un individu a un effet sur l'abondance maximale** de l'espèce, en raison de contraintes écologiques (sources de nourriture, disponibilité de l'habitat, etc.).



# Défi 2 - Solution

ajuster le modèle :

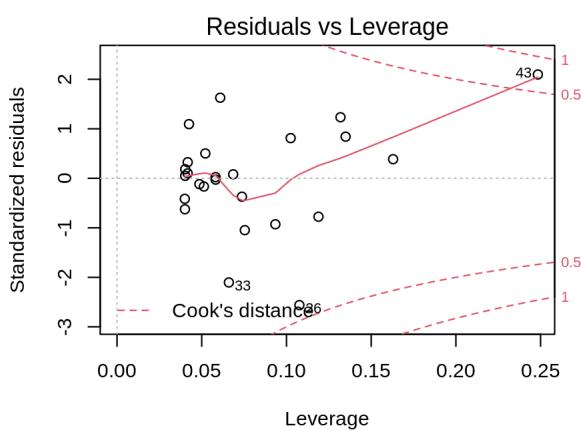
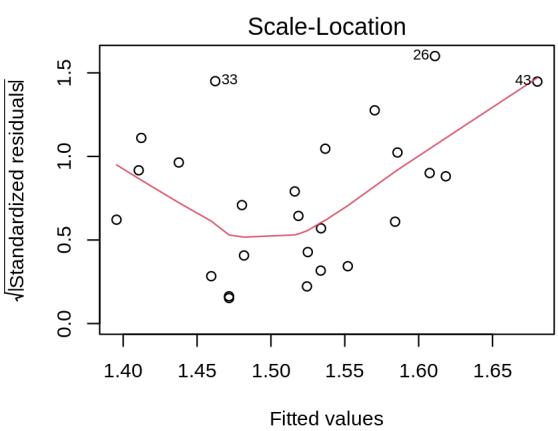
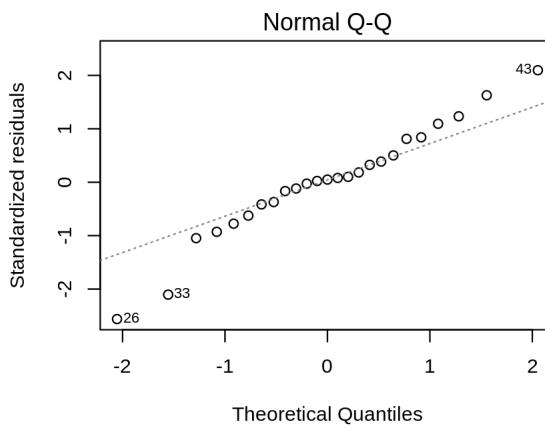
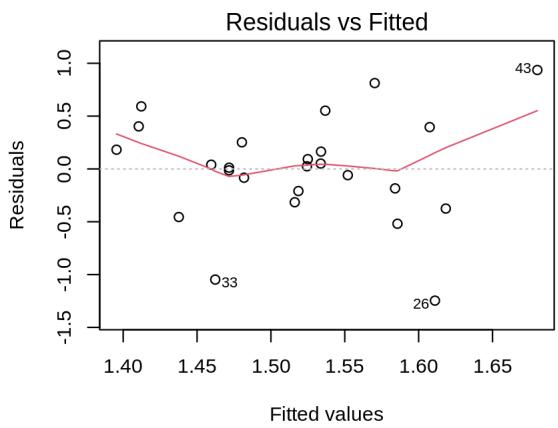
```
lm4 <- lm(logMaxAbund ~ logMass, data=bird, subset=bird$Passerine == 1)
lm4
#
# Call:
# lm(formula = logMaxAbund ~ logMass, data = bird, subset = bird$Passerine ==
#     1)
#
# Coefficients:
# (Intercept)      logMass
#           1.2429        0.2107
```



# Défi 2 - Solution

Vérifier les conditions d'application :

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





# Défi 2 - Solution

Vaut-il la peine d'interpréter les résultats ?

```
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

# Régression linéaire avec R

## Étape 1

Formuler et exécuter un modèle linéaire basé sur un hypothèse

## Étape 2

Vérifier les conditions d'application du modèle linéaire



*Conditions sont satisfaites ?*

## Étape 3

- Analyser les paramètres de régression
- Tracer le modèle
- Effectuer des tests de signification sur les estimations des paramètres (si nécessaire)



*Conditions non satisfaites ?*

Envisager l'utilisation d'un *Modèle linéaire généralisé (GLM)* ou la transformation des données



Utiliser un GLM mieux adapté aux données



Retourner à l'Étape 1 avec des variables transformées

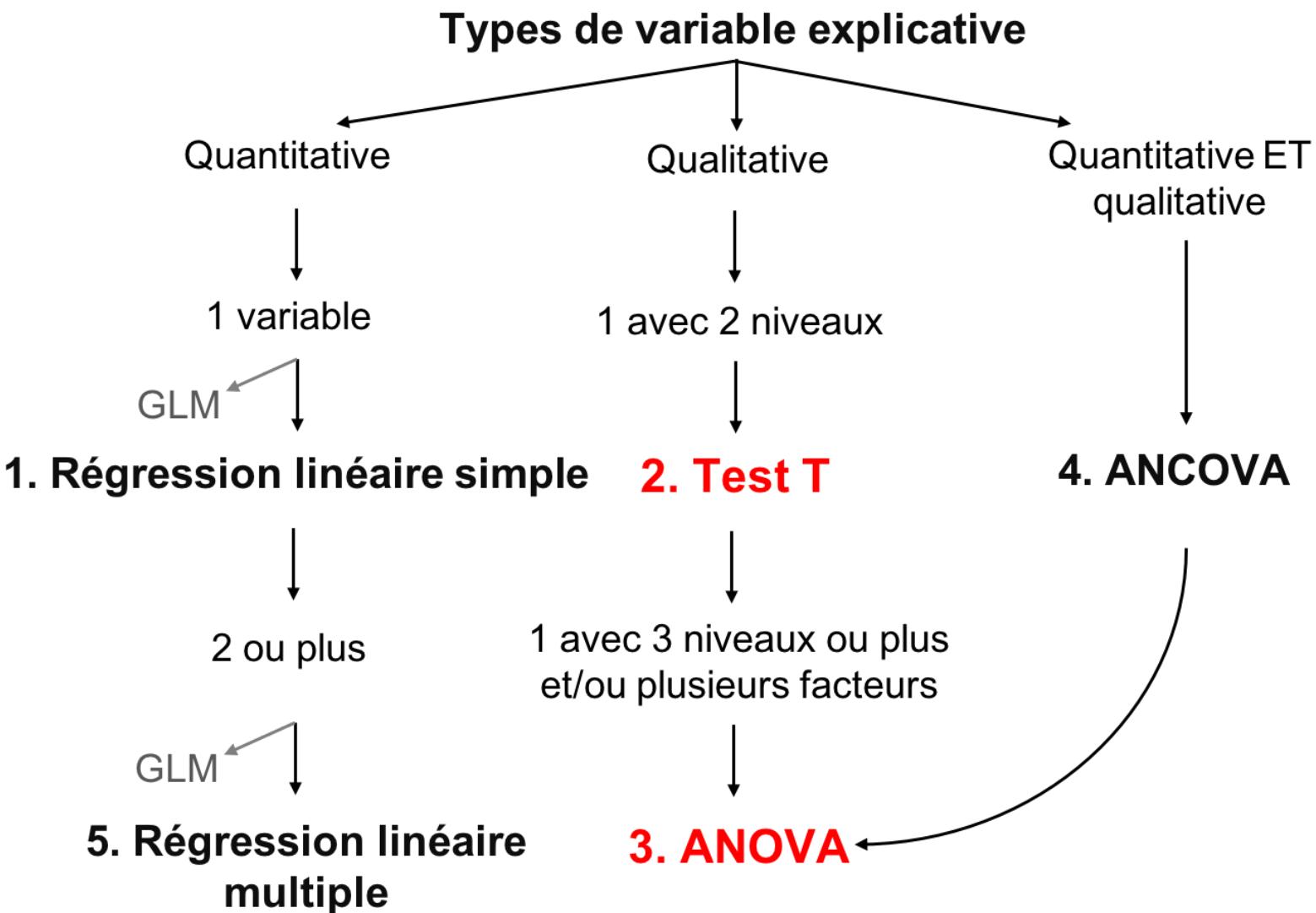
# Noms de variables

Des termes différents sont utilisés pour la *réponse* et le *prédicteur*, , en fonction du contexte et du domaine scientifique (les termes ne sont pas toujours synonymes).

<b>réponse</b>	<b>prédicteur</b>
var. expliqué	var. explicatif
	covariable
var. endogène	var. exogène
var. dépendante	var. indépendante

<b>response</b>	<b>predictor</b>
	explanatory var.
	covariate
outcome	
output var.	input var.
dependent var.	independent var.

# Modèles linéaires



# ANOVA

## Test-t

ANOVA à un critère de classification

ANOVA à deux critères de classification

# ANOVA

Variable réponse continue

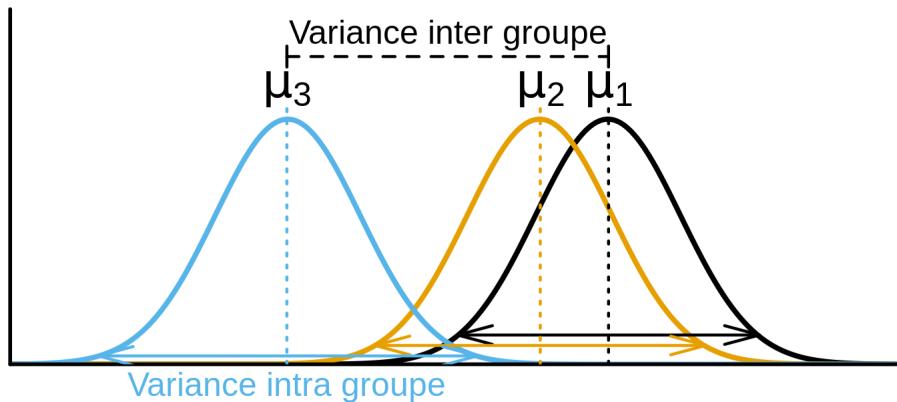
## Variables explicatives catégoriques

- Deux niveaux ou plus (groupes)

Compare la variation intra-groupe et inter-groupe afin de déterminer si les moyennes des groupes diffèrent

# ANOVA

Compare la variation intra-groupe et inter-groupe afin de déterminer si les moyennes des groupes diffèrent



Somme des carrés : variance intra-traitement vs variance inter-traitement

Si variance inter traitements > variance intra traitements:

- la variable explicative a un effet plus important que l'erreur aléatoire
- variable explicative est donc susceptible d'influencer significativement la variable réponse

# Types d'ANOVA

1. ANOVA à un critère de classification
  - Une variable explicative catégorique avec au moins 2 niveaux
  - S'il y a 2 niveaux, un **test de t** peut être utilisé alternativement
2. ANOVA à deux critères de classification
  - Deux variables explicatives catégoriques ou plus
  - Chaque facteur peut avoir plusieurs niveaux
  - Les interactions entre chaque facteur doivent être testées

Mesures répétées ?

- L'ANOVA peut être utilisée pour des mesures répétées, mais ce sujet n'est pas abordé dans cet atelier
- Modèle linéaire mixte peut également être utilisé pour ce type de données (voir l'atelier 6)

# Test de t

# Test de t

- **Variable réponse →** quantitative
- **Variable explicative →** qualitative avec **2 niveaux**

## conditions d'application

- Les résidus suivent une distribution normale
- Les variances des groupes sont homogènes

*Le test est plus robuste lorsque la taille de l'échantillon est plus élevée et lorsque les groupes ont des tailles égales*

# Exécuter un test de t dans R

Vous pouvez utiliser la fonction `t.test()`

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

- `Y`: variable réponse
- `X2`: facteur (2 niveaux)
- `data`: nom du jeu de données
- hypothèse `alternative` : `"two.sided"` (par défaut), `"less"`, ou `"greater"`

Le test de t est un modèle linéaire et un cas spécifique de l'ANOVA avec un facteur à 2 niveaux

Vous pouvez donc aussi utiliser la fonction `lm()`

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

# Exécuter un test de t dans R

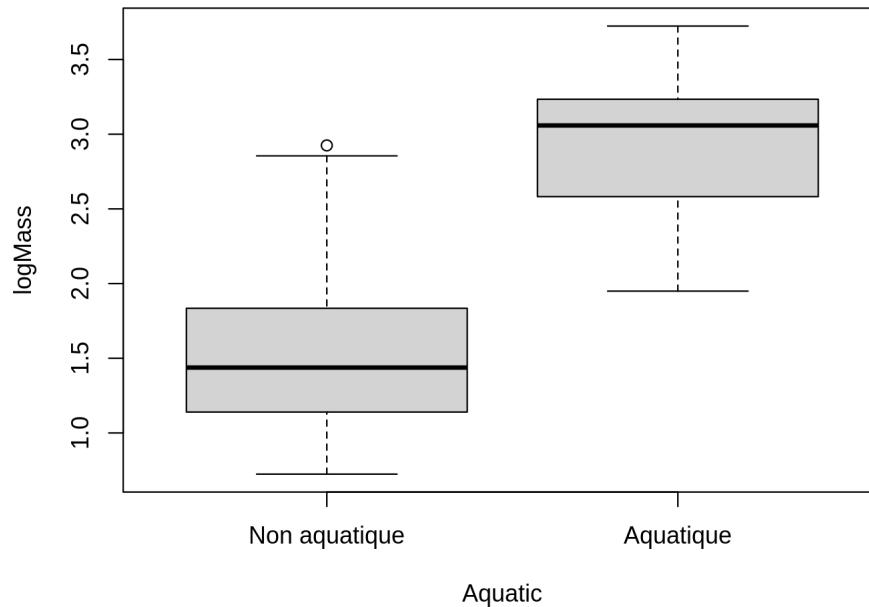
Les oiseaux aquatiques sont-ils plus lourds que les oiseaux terrestres ?

- Variable réponse : **Bird mass** → num: continue
- Variable explicative : **Aquatic** → 2 niveaux : 1 ou 0 (oui ou non)

# Exécuter un test de t dans R

Premièrement, visualiser les données à l'aide de la fonction `boxplot()`

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



# Exécuter un test de t dans R

Testons l'homogénéité des variances avec la fonction `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
```

*Le rapport des variances n'est pas statistiquement différent de 1, celles-ci peuvent donc être considérées comme égales*

*Nous pouvons maintenant procéder au test de t !*

# Exécuter un test de t dans R

```
ttest1 <- t.test(logMass ~ Aquatic, var.equal = TRUE, data = bird)  
# Or use lm()  
ttest.lm1 <- lm(logMass ~ Aquatic, data=bird)
```

*Spécifie que l'homogénéité des variances est respectée*

Vérifiez que **t.test()** et **lm()** donnent le même modèle :

```
ttest1$statistic^2  
#      t  
# 60.3845  
anova(ttest.lm1)$`F value`  
# [1] 60.3845      NA  
# réponse : F=60.3845 dans les deux cas
```

*Lorsque la condition d'égalité de variance est confirmée,  $t^2 = F$*

# Exécuter un test de t dans R

Si  $p < 0,01$  (ou  $0,05$ ), l'hypothèse de l'absence de différence entre les moyenne des 2 groupes ( $H_0$ ) peut être rejetée, avec un risque de  $0,01$  (ou  $0,05$ ) de se tromper

```
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
```

Il existe une différence entre la masse des oiseaux aquatiques et terrestres -  $p\text{-value}$

Regardez les moyennes des 2 groupes

# Non respect des conditions d'application

- **Correction de Welch** : lorsque les écarts entre les groupes ne sont pas égaux (par défaut dans R !)
- **Test de Mann-Whitney** : l'équivalent **non paramétrique** du test de t lorsque les conditions d'application ne sont pas respectées
- **Test de t apparié** : lorsque les deux groupes ne sont **pas indépendants** (par exemple, des mesures sur la même personne récoltées lors de 2 années différentes)

# Sondage

Avec un test de t, il est possible d'être plus précis et de donner une direction à notre hypothèse avec `alternative`.

Nous voulons tester si **les oiseaux aquatiques sont plus lourds que les oiseaux terrestres**.

Lequel devrait être utilisé? `alternative = "???"`

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

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

# Réponse

```
# 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
```

Pourquoi **"greater"** n'aurait pas marché?

*Indice: retournez à vos données. Quelle est la nature de la variable Aquatic?*

# ANOVA

# Analyse de Variance (ANOVA)

Généralisation du test de t à  $> 2$  groupes, et/ou  $\geq 2$  facteurs explicatifs

Décomposition de la variance observée de la variable réponse en effets additifs d'un ou de plusieurs facteurs et de leurs interactions

$$Y = \underbrace{\mu}_{\text{moyenne globale de la variable réponse sur tous les individus}} + \underbrace{\tau_i}_{\text{Le résultat moyen sur tous les individus du groupe } i} + \underbrace{\epsilon}_{\text{Résidus}}$$

# Rappel : ANOVA

conditions d'application

- Normalité des résidus
- L'égalité de la variance inter-groupes

Test complémentaire

- Lorsque l'ANOVA détecte une différence significative entre les groupes, l'analyse n'indique pas quel(s) groupe(s) diffère(nt) de(s) l'autre(s)
- Un test couramment utilisé *a posteriori* pour répondre à cette question est le **Test de Tukey**

# Exécuter une ANOVA dans R

## Est-ce que l'abondance maximale dépend du régime alimentaire ?

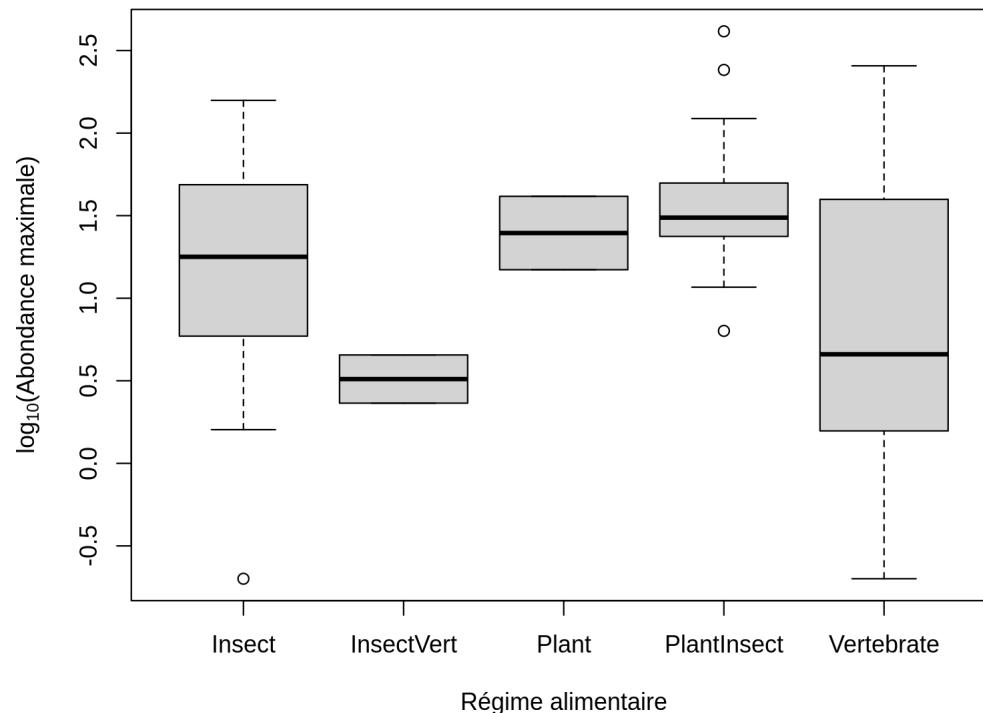
- Variable réponse : **MaxAbund** → num: quantitative
- Variable explicative : **Diet** → facteur avec 5 niveaux

```
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 ...
# $ 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 ...
```

# Visualiser les données

Visualisons tout d'abord les données avec la fonction `boxplot()`

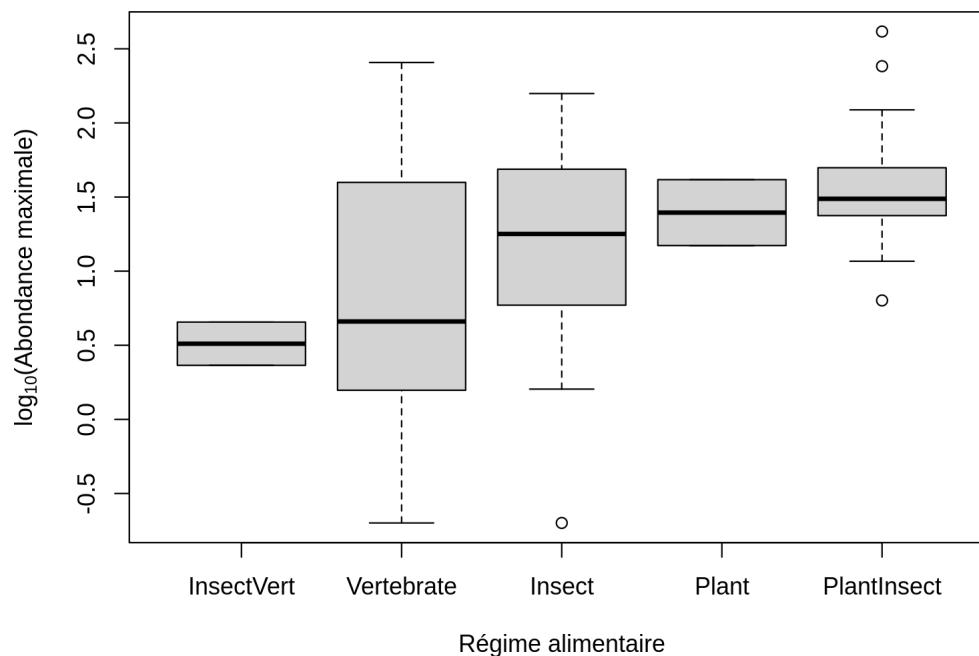
```
boxplot(logMaxAbund ~ Diet, data = bird,  
       ylab = expression("log"[10]*"(Abondance maximale)"), xlab = 'Régime alimentaire')
```



# Visualiser les données

Nous pouvons changer l'ordre des niveaux afin qu'il suivent l'ordre croissant de leurs médianes respectives en utilisant les fonctions `tapply()` et `sort()`

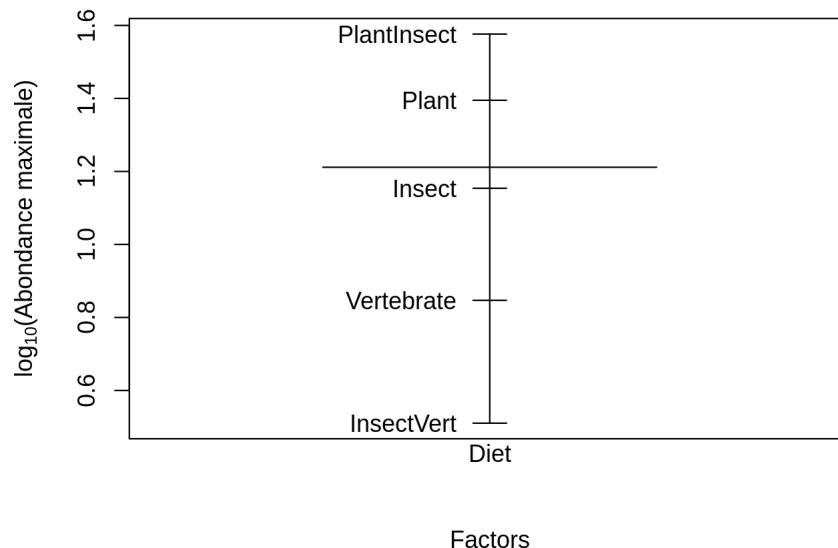
```
med <- sort(tapply(bird$logMaxAbund, bird$Diet, median))
boxplot(logMaxAbund ~ factor(Diet, levels = names(med)), data = bird,
        ylab = expression("log"[10]^*(Abondance maximale)), xlab = 'Régime alimentaire')
```



# Visualiser les données

Une autre façon de visualiser graphiquement les tailles d'effet est d'utiliser la fonction `plot.design()`

```
plot.design(logMaxAbund ~ Diet, data = bird,  
           ylab = expression("log"[10]^"(Abondance maximale)"))
```



Les niveaux d'un facteur le long d'une ligne verticale, et la valeur globale de la réponse dans une ligne horizontale

# ANOVA à un critère de classification dans R

Il est de nouveau possible d'utiliser la fonction `lm()`

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

Il y a aussi une fonction spécifique pour l'analyse de la variance dans R `aov()`

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

*Essayez-les et comparez les sorties !*

# Exécuter une ANOVA

## À un critère de classification dans R

Comparer les sorties

```
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
```

# Vérifier les conditions d'application

**Test de Bartlett:** égalité de la variance entre les groupes

```
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
```

# Vérifier les conditions d'application

**Test de Levene** pour l'homogénéité de la variance:

```
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
```

*Le test de Levene performe mieux, mais a une erreur de Type II un peu plus élevée.*

# Vérifier les conditions d'application

**Test de Shapiro-Wilk:** normalité des résidus

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

Les deux tests sont non-significatifs; les résidus du modèle peuvent être considérés normaux et les variances homogènes

# Et si les conditions d'application ne sont pas respectées...

**Transformer vos données** : pourrait égaliser les variances et normaliser les résidus, et peut convertir un effet multiplicatif en un effet additif

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

- Voir le wiki de l'atelier 1 pour les règles de transformation de données
- ré-exécuter votre modèle avec la variable transformée et vérifier à nouveau les hypothèses

**Test de Kruskal-Wallis**: équivalent non paramétrique de l'ANOVA si vous ne pouvez pas (ou ne voulez pas) transformer les données

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

# Sorties de notre modèle ANOVA

Triage en ordre alphabétique des niveaux et comparaison au niveau de référence (**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) 1.09647   0.14629   7.495 1.14e-09 ***
# Diet1        -0.32172   0.24876  -1.293  0.2020    
# Diet2         0.18757   0.17854   1.051  0.2986    
# Diet3         0.13911   0.06960   1.999  0.0512 .  
# Diet4        -0.06239   0.05238  -1.191  0.2393    
# ---
# 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
```

# Sorties de notre modèle ANOVA

D'autre part, si nous utilisons `lm()`

```
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) 1.09647   0.14629   7.495 1.14e-09 ***
# Diet1       -0.32172   0.24876  -1.293  0.2020
# Diet2        0.18757   0.17854   1.051  0.2986
# Diet3        0.13911   0.06960   1.999  0.0512 .
# Diet4       -0.06239   0.05238  -1.191  0.2393
#
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
#
# Residual standard error: 0.6709 on 49 degrees of freedom
# Multiple R-squared:  0.188,    Adjusted R-squared:  0.
# F-statistic: 2.836 on 4 and 49 DF,  p-value: 0.0341
```

*Différence significative entre les groupes, mais nous ne savons pas lesquels !*

# Test a posteriori

Lorsque l'ANOVA détecte un effet significatif de la variable explicative, un test post-hoc avec la fonction `TukeyHSD()`, doit être effectué pour déterminer quel(s) traitement(s) diffère(nt)

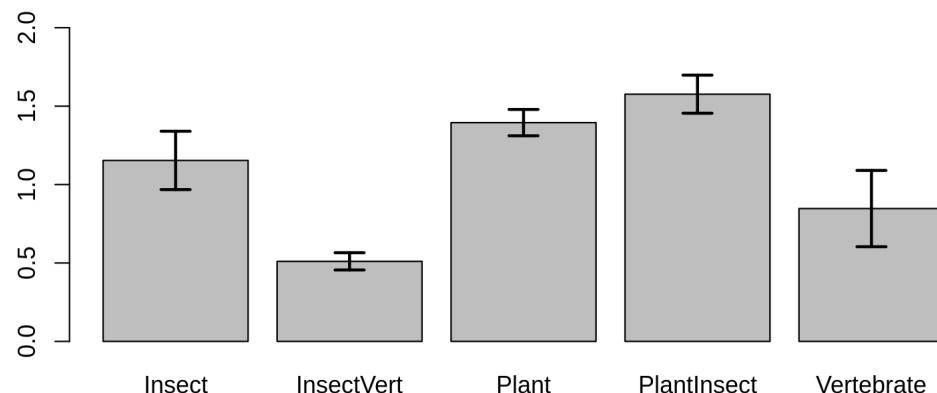
```
TukeyHSD(aov(anov1), ordered = TRUE)
# Tukey multiple comparisons of means
# 95% family-wise confidence level
# factor levels have been ordered
#
# Fit: aov(formula = anov1)
#
# $Diet
#
#           diff      lwr      upr
# Vertebrate-InsectVert 0.3364295 -1.11457613 1.787435
# Insect-InsectVert     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
```

Seuls `Vertebrate` et  
`PlantInsect`  
diffèrent

# Représentation graphique

Représentation graphique de l'ANOVA à l'aide de la fonction `barplot()`

```
sd <- tapply(bird$logMaxAbund, bird$Diet, sd)
means <- tapply(bird$logMaxAbund, bird$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) # barres verticales
segments(bp - epsilon, means - se, bp + epsilon, means - se, lwd = 2) # barres horiz
segments(bp - epsilon, means + se, bp + epsilon, means + se, lwd = 2) # barres horiz
```



# ANOVA à deux critères de classification

# ANOVA à deux critères de classification

Plus d'un facteur

- ANOVA avec un facteur:

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

- ANOVA avec deux ou plus facteurs:

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

Lorsque vous utilisez le symbole "\*" avec `lm()`, le modèle inclut les effets de chaque facteur séparément, ainsi que leur interaction

Lorsque vous utilisez le symbole "+" avec `lm()`, le modèle inclut les effets de chaque facteur séparément (pas d'interaction)

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

# ANOVA à deux critères de classification

Exemple d'interaction non significative

```
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
```

Selon le principe de **parcimonie**, vous voulez que votre modèle explique le plus possible de la variance observée dans les données, avec le moins de termes possible

- Enlever le terme d'interaction s'il n'est pas significatif, et ré-exécuter le modèle

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



# Défi 3

Testez si l'abondance maximale `log(MaxAbund)` varie à la fois en fonction du régime alimentaire (`Diet`) et de l'habitat (`Aquatic`).

INDICE: Assurez-vous d'ajouter une interaction avec `*`

**Salle de réunion!**



```
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)  0.97239  0.17539  5.544 1.48e-06 ***
# Diet1        -0.43494  0.33549 -1.296  0.2014    
# Diet2         0.19846  0.18934  1.048  0.3002    
# Diet3         0.14752  0.07830  1.884  0.0660    
# Diet4        -0.17311  0.07123 -2.430  0.0191 *  
# Aquatic      0.97186  0.37956  2.560  0.0139 *  
# Diet1:Aquatic 0.28017  0.48488  0.578  0.5663    
# Diet2:Aquatic 1.35536  0.73530  1.843  0.0719    
# Diet3:Aquatic -0.39509  0.25582 -1.544  0.1295    
# Diet4:Aquatic      NA       NA       NA       NA      
# ---
# Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 0.6482 on 45 degrees of freedom
```

# Défi 3 - Solution



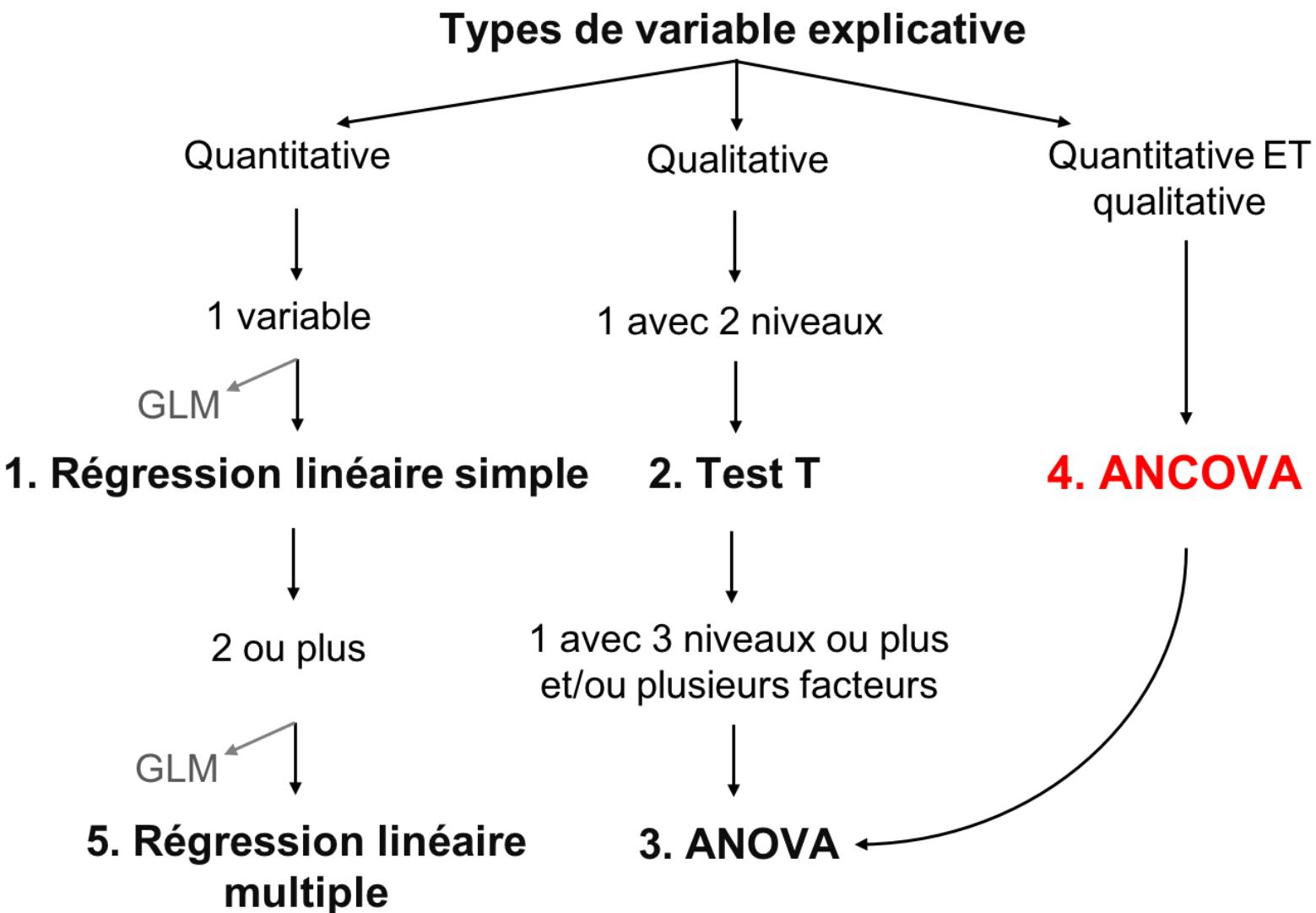
```
anov2 <- lm(logMaxAbund ~ Diet*Aquatic, data = bird)
anova(anov2)
# Analysis of Variance Table
#
# Response: logMaxAbund
#
#             Df  Sum Sq Mean Sq F value    Pr(>F)
# Diet          4  5.1059 1.27647  3.0378 0.02669 *
# Aquatic       1  0.3183 0.31834  0.7576 0.38870
# Diet:Aquatic  3  2.8250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

*Le seul terme significatif du modèle est le facteur régime alimentaire*

*Selon le principe de parcimonie, nous devrions supprimer le terme d'interaction:*

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

# Modèles linéaires



# ANCOVA

# Analyse de covariance (ANCOVA)

- Combinaison de l'ANOVA et de la régression linéaire
- Les variables explicatives sont un mélange de variables quantitatives (covariable) et qualitatives (facteurs)

$$Y = \mu + \text{Effets principaux des facteurs} + \\ \text{Interactions entre facteurs} + \\ \text{Effets principaux des covariables} + \\ \text{Interactions entre covariables et facteurs} + \epsilon$$

# Rappel : ANCOVA

En plus des conditions d'application des modèles linéaires, les modèles **ANCOVA** doivent respecter :

- Les covariables ont toutes la **même étendue de valeurs**
- Les variables sont **fixes**
- Les variables catégoriques et continues sont **indépendantes**

Un variable **fixe** est une variable d'intérêt pour une étude (e.g. la masse des oiseaux). En comparaison, une variable aléatoire représente surtout une source de bruit qu'on veut contrôler (i.e. le site où les oiseaux ont été échantillonnés)

Voir l'atelier 6 sur les modèles linéaires mixtes

# Types d'ANCOVA

Vous pouvez avoir n'importe quel nombre de facteurs et / ou variables, mais lorsque leur nombre augmente, l'interprétation des résultats devient de plus en plus complexe

ANCOVA fréquemment utilisées

1. **Une covariable et un facteur**
2. Une covariable et deux facteurs
3. Deux covariables et un facteur

*Nous ne considérerons que le premier cas aujourd'hui, mais les deux autres sont similaires*

# ANCOVA avec 1 covariable et 1 facteur

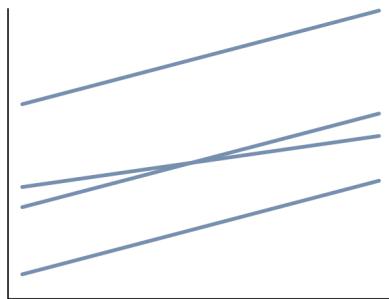
Objectifs de l'analyse :

1. Déterminer l'effet du facteur et de la covariable sur la variable réponse
2. Déterminer l'effet du facteur sur la variable réponse après avoir enlevé l'effet de la covariable
3. Déterminer l'effet de la covariable sur la variable réponse en contrôlant l'effet du facteur

**Si vous avez une interaction significative entre votre facteur et votre covariable, vous ne pouvez pas atteindre ces objectifs !**

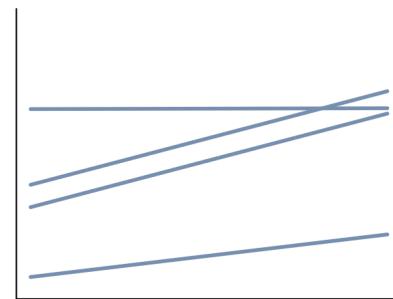
# ANCOVA avec 1 covariable et 1 facteur

Un niveau du facteur  
a une pente différente

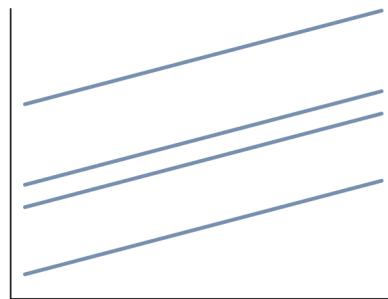


Si l'interaction est significative, vous aurez un scénario qui ressemble à ceci

Des nombreux niveaux ont des pentes différentes



Pas d'interaction



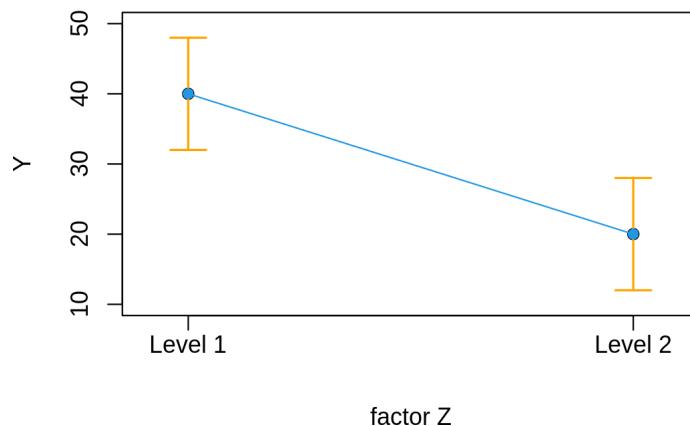
Si votre covariable et votre facteur sont significatifs, vous avez un cas comme celui-ci

# Comparez ANCOVA - moyennes ajustées

Si vous voulez comparer les moyennes des différents facteurs, vous pouvez utiliser les **moyennes ajustées**

La fonction `effect()` utilise les équations données par l'ANCOVA pour estimer les moyennes de chaque niveau, corrigées pour l'effet de la covariable

```
ancova.exemple <- lm(Y ~ X*Z, data=data) # X = quantitative; Z = qualitative  
library(effects)  
adj.means.ex <- effect('Z', ancova.exemple)  
plot(adj.means.ex)
```



# ANCOVA avec 1 covariable et 1 facteur

- Si seulement votre facteur est significatif, éliminer la covariable -> vous avez une **ANOVA**
- Si seulement votre covariable est significative, éliminer le facteur -> vous avez une **régression linéaire simple**
- Si votre interaction covariable \* facteur est significative, vous voudrez peut-être tester quel(s) niveau(x) du facteur a(ont) des pentes différentes

## Vérifier vos conditions d'application !

- Très similaire à ce que vous avez fait précédemment

# Exécuter une ANCOVA dans R

L'abondance maximale varie-t-elle en fonction du régime alimentaire et la masse des oiseaux ?

Variable réponse : **MaxAbund** → num : quantitative continue

Variable explicatives :

- **Diet** → facteur à 5 niveaux
- **Mass** → numérique continue

```
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 ...
```



## Défi 4

1- Exécutez un modèle pour tester les effets du régime alimentaire (`Diet`), de la masse (`logMass`) ainsi que leur interaction sur l'abondance maximale des oiseaux (`logMaxAbund`)

2- Vérifiez si votre interaction est significative

**Salle de réunion!**

# Défi 4 - 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 entre **logMass** et **Diet** n'est pas significative

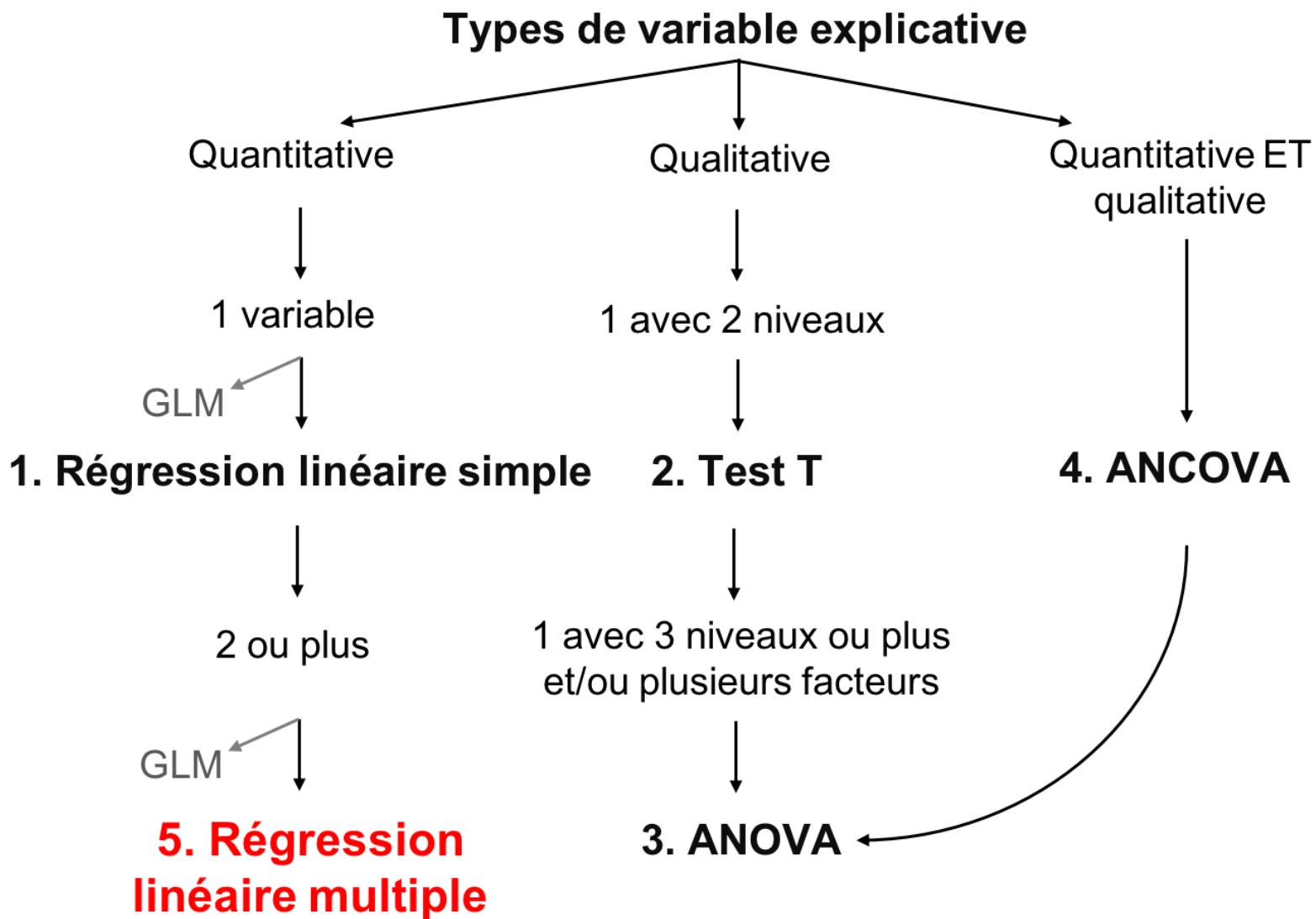
# Défi 4 - Solution



Éliminer le terme d'interaction, puis ré-évaluer le modèle contenant les effets simples de **logMass** et **Diet**

```
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
```

# Modèles linéaires



# Régression linéaire multiple

# Régression linéaire multiple

- **Variables explicatives** → 2 ou plusieurs variables continues
- **Variable réponse** → 1 variable continue

Seule différence avec la régression linéaire simple : **plusieurs variables explicatives** sont incluses dans le modèle.

## Variables

- $y$  : Variable réponse (**continue**)
- $x$  : Plusieurs variables explicatives (**continues** ou **catégoriques**)

## Relation supposée

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

- Le paramètre  $\beta_0$  est **l'ordonnée à l'origine** (ou constante)
- Les paramètre  $\beta_1$  quantifie **l'effet** de  $x$  sur  $y$ .
- Le résidu  $\epsilon_i$  représent la variation **non expliquée**
- La **valeur prédictive** de  $y_i$  se définit comme :

$$\hat{y}_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \dots + \beta_k x_{k,i}$$

# Régression linéaire multiple

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

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

## Conditions d'application

En plus des conditions d'application habituelles des modèles linéaires :

- **Relation linéaire** entre **chaque** variable explicative et la variable réponse.
- Les variables explicatives sont indépendantes les unes des autres (il n'y a pas de **colinéarité**).

# Régression linéaire multiple

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

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

## En cas de colinéarité

- Garder seulement une des variables colinéaires
- Essayer une analyse multidimensionnelle (voir l'atelier 9)
- Essayer une analyse pseudo-orthogonale

# Régression linéaire multiple dans R

En utilisant le jeu de données `Dickcissel` comparez l'importance relative du climat (`clTma`), de la productivité (`NDVI`) et de la couverture du sol (`grass`) comme prédicteurs de l'abondance de dickcissels (`abund`)

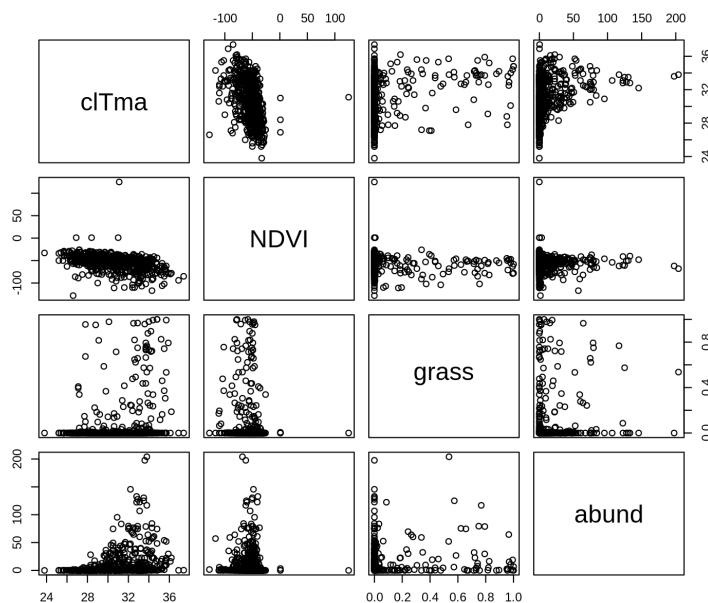
```
Dickcissel = read.csv("data/dickcissel.csv")
str(Dickcissel)
# 'data.frame': 646 obs. of 15 variables:
# $ abund      : num  5 0.2 0.4 0 0 0 0 0 0 0 ...
# $ Present    : chr "Absent" "Absent" "Absent" "Present" ...
# $ clDD       : num  5543 5750 5395 5920 6152 ...
# $ clFD       : num  83.5 67.5 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 ...
```

# Vérifier les conditions d'application

La colinéarité :

- Vérifier la colinéarité de toutes les variables explicatives et d'intérêt

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



*Si vous observez un patron entre vos deux variables explicatives, elles peuvent être colinéaires!*

*Vous devez éviter ceci, sinon leurs effets sur la variable réponse seront confondus*

# Régression linéaire multiple dans R

Exécuter la régression multiple de l'abondance (`abund`) en fonction des variables

`clTma + NDVI + grass`

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

Vérifiez les autres conditions d'application, comme pour la régression linéaire simple

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

# Régression linéaire multiple dans R

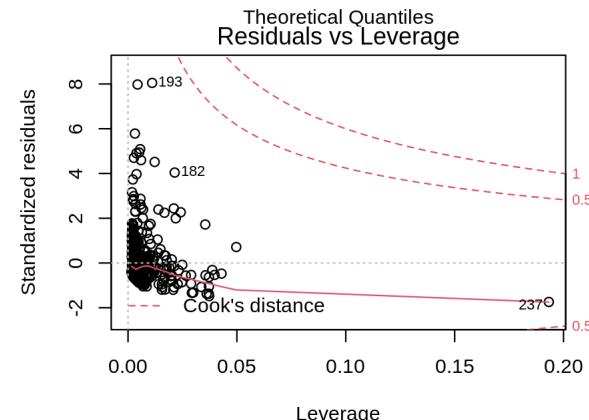
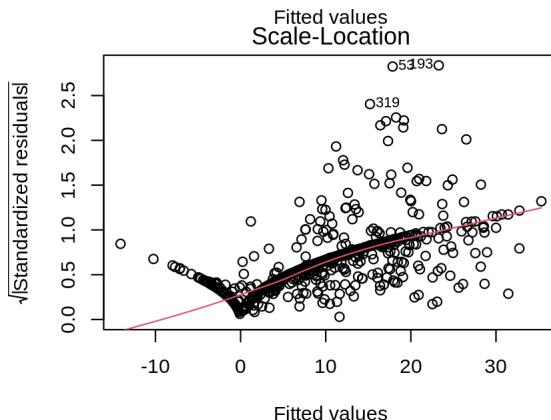
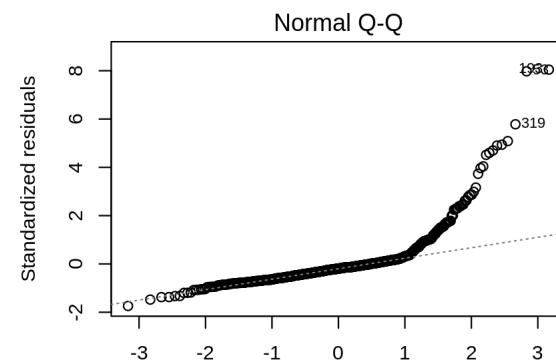
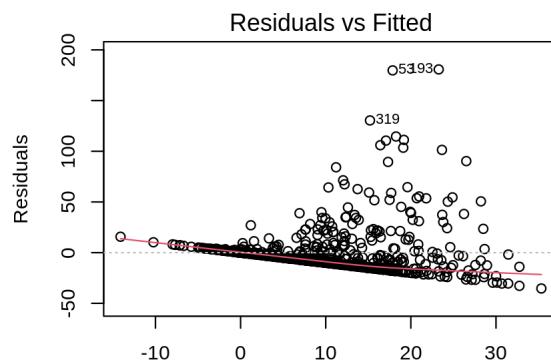
Exécuter la régression multiple de l'abondance (`abund`) en fonction des variables  
`clTma + NDVI + grass`

```
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
```

# Régression linéaire multiple dans R

Vérifiez les autres conditions d'application, comme pour la régression linéaire simple

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



# Quel est le meilleur modèle ?

Souvenez-vous du principe de parcimonie: expliquer le plus de variation avec le plus petit nombre de termes dans votre modèle → enlevez la variable qui est la moins significative

```
summary(lm.mult)$coefficients
#               Estimate Std. Error   t value   Pr(>|t| )
# (Intercept) -83.6081274 11.5774529 -7.221634 1.458749e-12
# c1Tma        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
```

Les 3 variables sont importantes. On garde tout !

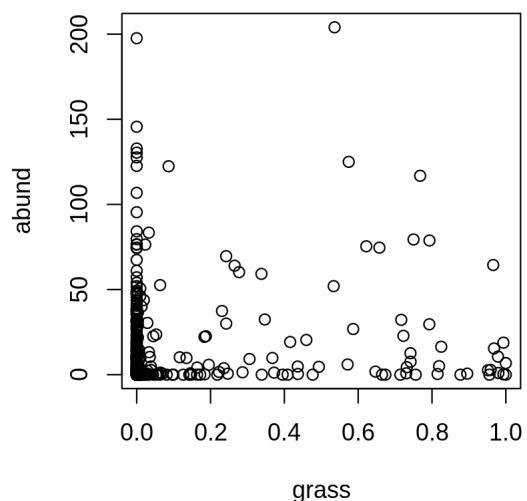
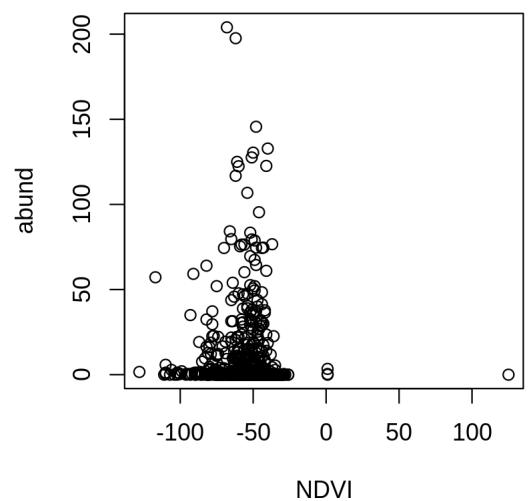
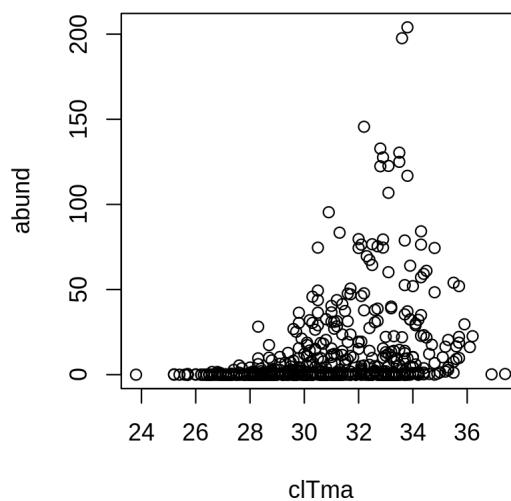
Le modèle explique 11.28% de la variabilité de l'abondance de dickcissels  
 $R^2_{adj} = 0.11$ .

**Toutefois, ces informations ne sont pas valables car les conditions d'application du modèle linéaire ne sont pas respectées.**

# Quel est le meilleur modèle ?

Il est important de noter que la variable réponse ne varie pas de façon linéaire avec les variables explicatives

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



Voir la **section avancée** sur la **régression polynomiale** pour la solution !

# Optionnel

*si le temps le permet*

# Optionnel

1. Interprétation des contrastes
2. ANOVA non équilibrée
3. Régression polynomiale
4. Partitionnement de la variation

# Interprétation des contrastes

# Interprétation des contrastes

Les contrastes servent à comparer chaque niveau du facteur à un niveau de référence, et de détecter des différences significatives entre chaque niveau.

L'estimation de l'ordonnée à l'origine est le niveau de référence et correspond à la moyenne du premier niveau (en ordre alphabétique) du facteur **Diet**

Calculez l'ordonnée à l'origine de référence + l'ordonnée à l'origine de chaque niveau de Diet *Que remarquez-vous ?*

```
tapply(bird$logMaxAbund, bird$Diet, mean)
#      Insect   InsectVert       Plant PlantInsect  Vertebrate
#  1.1538937   0.5104603   1.3948941   1.5761940   0.8468898
coef(anov1)
# (Intercept)     Diet1     Diet2     Diet3     Diet4
# 1.09646639 -0.32171668  0.18757236  0.13911115 -0.06239414
coef(anov1)[1] + coef(anov1)[2] # InsectVert
# (Intercept)
# 0.7747497
coef(anov1)[1] + coef(anov1)[3] # Plant
# (Intercept)
# 1.284039
```

# Interprétation des contrastes

Il se peut que vous vouliez définir un niveau de référence différent

1. Comparez le niveau **Plant** à tous les autres niveaux du facteur **Diet**

```
bird$Diet2 <- relevel(bird$Diet, ref="Plant")
anova2 <- lm(logMaxAbund ~ Diet2, data = bird)
summary(anova2)
anova(anova2)
```

1. Ordonner les niveaux selon leur médiane

```
bird$Diet2 <- factor(bird$Diet, levels=names(med))
anova2 <- lm(logMaxAbund ~ Diet2,
              data = bird)
summary(anova2)
anova(anova2)
```

Observez-vous un changement quant aux niveaux du facteur **Diet** qui sont significatifs ?

# Interprétation des contrastes

*Un point important à remarquer à propos du contraste par défaut dans R (`contr.treatment`) est qu'il n'est PAS orthogonal*

Pour être orthogonal :

- Pour être orthogonal, les propriétés suivantes doivent être respectées:
- La somme du produit de deux colonnes égale 0

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

# Interprétation des contrastes

Changez les contrastes pour mettre les niveaux orthogonaux

```
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 =
summary(anov3)
#
# Call:
# lm(formula = logMaxAbund ~ Diet, data =
#
# Residuals:
#       Min        1Q    Median        3Q
# -1.85286 -0.32972 -0.08808  0.47375
#
# Coefficients:
#             Estimate Std. Error t val
# (Intercept) 1.09647   0.14629  7.4
# Diet1       -0.32172   0.24876 -1.2
# Diet2        0.18757   0.17854  1.0
```

Les contrastes Helmert vont contraster le deuxième niveau avec le premier, le troisième avec la moyenne des deux premiers niveaux, etc.

# ANOVA non équilibrée

# ANOVA non équilibrée

Un jeu de données est considéré non équilibré lorsque le nombre d'échantillons entre deux niveaux n'est pas égal.

Le jeu de données `Birdsdiet` est en réalité non équilibré (le nombre d'espèces aquatiques n'égale pas le nombre d'espèces non-aquatiques)

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

Dans une telle situation, l'ordre des covariable affecte la calculation de la somme des carrés, et donc la valeur de `p`.

Testons-le avec le jeu de données `Birdsdet`.

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

# ANOVA non équilibrée

```
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
```

# ANOVA non équilibrée

Afin de régler ce problème est de prendre une nouvelle approche pour tester les effets de chaque variable.

**Type I** : Teste les effets en séquentiel, en débutant avec la première variable.

**Type II**: Teste les effets de chaque facteur, mais après avoir tester l'autre facteur.

**Type III**: Teste les effets de chaque facteur, mais après avoir tester l'autre facteur et l'interaction.

*Le type I est celui par défaut dans R et qui crée le problème avec des données non équilibré*

**Si vous considérez utiliser le Type II ou III avec vos propres données, vous devriez en lire plus sur le sujet avant de choisir. Vous pouvez commencer avec ce lien**

# ANOVA non équilibrée

Maintenant essayez une `Anova()` de type III

```
car:::Anova(unb.anov1, type = "III")
# Anova Table (Type III tests)
#
# Response: logMaxAbund
#           Sum Sq Df F value    Pr(>F)
# (Intercept) 18.9349  1 41.8186 4.8376
# Aquatic      0.3183  1  0.7031  0.4000
# Diet         5.1926  4  2.8671  0.0350
# 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(>F)
# (Intercept) 18.9349  1 41.8186 4.8376
# Diet         5.1926  4  2.8671  0.0350
# Aquatic      0.3183  1  0.7031  0.4000
# Residuals   21.7337 48
# ---
# Signif. codes:  0 '****' 0.001 '**' 0.
```

Que remarquez-vous en utilisant `Anova()` ?

# Régression polynomiale

# Régression polynomiale

Comme nous l'avons remarqué dans la section sur la **régression linéaire multiple**, certaines variables semblent avoir des relations non-linéaires avec la variable **MaxAbund**

Pour tester des relations non-linéaires, des régressions polynomiales de différents degrés sont comparées

- Un modèle polynômial ressemble à ceci :

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

*Ce polynôme a trois termes*

# Régression polynomiale

Pour un polynôme avec une variable (comme  $x$ ), le *degré* est l'exposant le plus élevé de cette variable

*Nous avons ici un polynôme de degré 4*

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

# Régression polynomiale

Lorsque vous connaissez le degré, vous pouvez lui donner un nom :

degré	Nom	Example
0	Constante	3
1	Linéaire	$x + 9$
2	Quadratique	$x^2 - x + 4$
3	Cubique	$x^3 - x^2 + 5$
4	Quartique	$6x^4 - x^3 + x - 2$
5	Quintique	$x^5 - 3x^3 + x^2 + 8$

# Régression polynomiale

En utilisant le jeu de données `Dickcissel`, testez la relation non-linéaire entre l'abondance et la température en comparant trois modèles polynomiaux groupés (de degrés 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)
```

# Régression polynomiale

- Comparez les modèles polynomiaux et déterminez quel modèle niché nous devrions sélectionner
- Exécutez un résumé de ce modèle, reportez l'équation de la régression, les valeurs de p, et le R carré ajusté

# Régression polynomiale

Comparez les modèles polynomiaux; quel modèle niché nous devrions sélectionner ?

Exécutez un résumé de ce modèle

```
print(summ_lm.linear)
# [1] "Coefficients:
# [2] "
# [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] "
# [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)
```

# Partitionnement de la variation

# Partitionnement de la variation

Certaines variables explicatives de la **régression linéaire multiple** étaient fortement corrélées (c.-à-d. multicolinéarité)

La colinéarité entre variables explicatives peut être détectée à l'aide de critères d'inflation de la variance (fonction `vif()` du packet `car`)

- Les valeurs supérieures à 5 sont considérées colinéaires

```
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
```

# Partitionnement de la variation

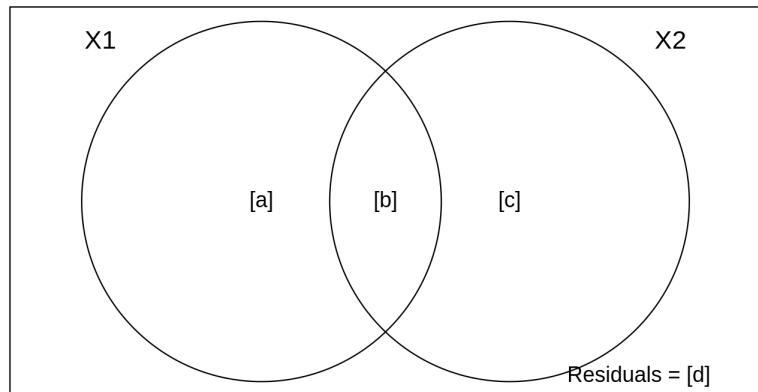
Utilisez `varpart()` afin de partitionner la variation de la variable `abund` avec toutes les variables de la couverture du paysage groupées ensemble et toutes les variables du climat groupées ensemble (laissez NDVI à part)

```
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("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** : les variables colinéaires n'ont pas besoin d'être enlevées avant l'analyse

# Partitionnement de la variation

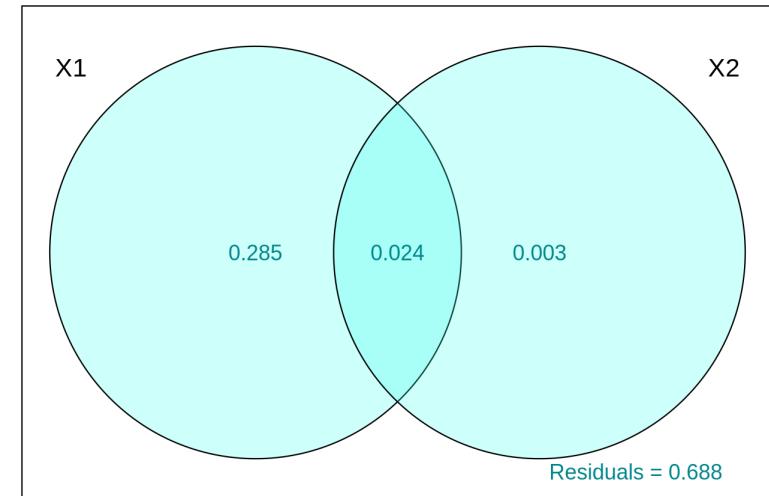
```
showvarparts(2)
```



```
?showvarparts
```

```
# With two explanatory tables, the fractions  
# explained uniquely by each of the two  
# are '[a]' and '[c]', and their joint  
# is '[b]' following Borcard et al. (1992).
```

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



La proportion de la variation de la variable abond expliquée par le climat seulement est 28.5% (obtenu par  $X_1|X_2$ ), par la couverture du paysage seulement est ~0% ( $X_2|X_1$ ), et par les deux combinés est 2.4%

# Partitionnement de la variation

Tester si chaque fraction est significative

- Climat seul

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

- Couverture du paysage seul

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

# Partitionnement de la variation

```
# Climat seul
anova(out.1, step = 1000, perm.max = 1000)
# Permutation test for rda under reduced rank
# Permutation: free
# Number of permutations: 999
#
# Model: rda(X = Dickcissel$abund, Y = clim)
#          Df Variance      F Pr(>F)
# Model      5    165.12 53.862  0.001 ***
# Residual  634    388.72
# ---
# Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

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

Conclusion: la fraction expliquée par la couverture du paysage n'est pas significative une fois que nous avons pris en compte l'effet du climat

Merci d'avoir participé !

