TP 5
Analyse de Variance
Analyse de Covariance

On va s'intéresser dans ce TP aux données contenues dans le fichier `penguins.csv` (plus de détails ici) listant différentes quantités mesurées sur 344 manchots en Antarctique au cours de 3 études différentes menées. Les 8 variables suivantes ont été relevées :

- `species` : espèce de l'individu
- `island` : île de résidence
- `bill_length_mm` : longueur du bec
- `bill_depth_mm` : hauteur du bec
- `flipper_length_mm` : longueur de la nageoire
- `body_mass_g` : masse corporelle
- `sex` : sexe de l'individu
- `year` : année de l'étude menée

Dans ce TP, le but est de choisir une des variables **quantitatives** accessibles et de l'étudier en fonction d'autres variables. Tout au long de ce sujet, on appellera `Quant1`la variable que vous aurez choisie.

1 Traitement et visualisation des données de base

Le jeu de données contient des `NA`, qui correspondent à des données manquantes (Not Available). Il existe différentes méthodes permettant de gérer ces données manquantes, mais on se contentera ici de retirer les individus concernés. Effectuer ce nettoyage à l'aide de la fonction `is.na` en vous servant de l'exemple ci-dessous.

```
X = c(1, NA, 123) ## Exemple
is.na(X)
```

```
## [1] FALSE  TRUE FALSE
```

Une fois le nettoyage effectué, commencer par représenter graphiquement la variable qui vous intéresse à l'aide (par exemple) des commandes ci-dessous, et identifier des relations potentiellement intéressantes à étudier.

- Si un data-frame `donnees` contient deux variables nommées `VarX` et `VarY`, on peut visualiser leur relation avec

  ```
  plot(VarY ~ VarX, data = donnees)
  ```

  On peut le lien d'une seule variable en fonction de toutes les autres avec

  ```
  plot(VarY ~ . , data = donnees)
  ```

  On peut alors faire défiler les graphiques successifs avec ⏎ (ctrl+C pour stoper le défilement). La totalité des relations peut également être tracée avec `plot(donnees)` (mais ce n'est pas très lisible).
- Il est possible de faire apparaître sur un diagramme de dispersion (cçd représentation de deux variables quantitatives) de faire apparaître une variable qualitative `VarZ` comme ceci :

  ```
  plot(VarY ~ VarX, data = donnees, col = VarZ)  ## VarZ représentée par la couleur des points
  plot(VarY ~ VarX, data = donnees, pch = as.numeric(VarZ)) ## VarZ représentée par la forme des points
  ```

  Remarque : ces commandes n'ont de sens que si `VarX` est également une variable quantitative.

2 Analyse de variance

Rappel : l'utilisation des différentes commandes R pour la mise en place d'une Anova sont données dans les diapos du cours (Chap 1 et 2).

1. Choisir une variable qualitative `Qual1` et
   (a) Représenter graphiquement votre variable `Quant1`en fonction de `Qual1`, et émettre une hypothèse à vérifier en fonction de ce que vous observez sur le graphique ovtenu.

(b) Construire le modèle linéaire lien `Quant1`avec `Qual1` avec `lm`. Commenter les estimations des coefficients du modèle construit.

(c) Extraire la table récapitulative à l'aide de la fonction `anova` (ou `Anova` du package `car`). Commenter la table ainsi obtenue.

(d) Éprouver les hypothèses de modélisation à partir de graphiques et des tests appropriés sur les résidus.

2. Choisir une deuxième variable qualitative `Qual2` et

(a) Répéter les 4 étapes précédentes de la question précédente pour `Qual2` seule.

(b) Tracer les profils d'interaction à l'aide de la fonction `interaction.plot`.

(c) Répéter les 4 étapes précédentes de la question précédente pour l'ajustement de `Quant1`en fonction `Qual1` et `Qual2`.

3. Effectuer la même analyse en fonction de **toutes** les variables qualitatives. **Note** : les deux formules `y~(a+b)^2` et `y~a+b+a:b` sont équivalentes.

3 Analyse de covariance

L'analyse de covariance permet d'étudier une possible interaction entre un prédicteur qualitatif et un prédicteur variable quantitatif dans leur influence sur une sortie quantitative.

1. Commencer par choisir une seconde variable quantitative `Quant2` et

(a) Représenter à nouveau graphiquement votre variable `Quant1`en fonction de `Quant2`, et émettre une hypothèse à vérifier en fonction de ce que vous observez sur le graphique ovtenu.

(b) Construire le modèle linéaire lien `Quant1`avec `Quant2` avec `lm`. Commenter les estimations des coefficients du modèle construit, et ajouter le tracé de la droite de régression sur

(c) Éprouver les hypothèses de modélisation à partir de graphiques et des tests appropriés sur les résidus.

2. Considérons à présent la variable qualitative `Qual1`

(a) Pour chaque modalité de `Qual1`, effectuer la régression linéaire de `Quant1` en fonction de `Quant2`.

(b) Sur le même graphique que précédemment, représenter les différentes droites de régression obtenues.

(c) Comparer ces droites entre elles, ainsi qu'avec la droite de régression "globale". Que pourriez-vous émettre comme hypothèse ?

(d) Extraire la table récapitulative et la commenter.

(e) Éprouver les hypothèses de modélisation à partir de graphiques et des tests appropriés sur les résidus.