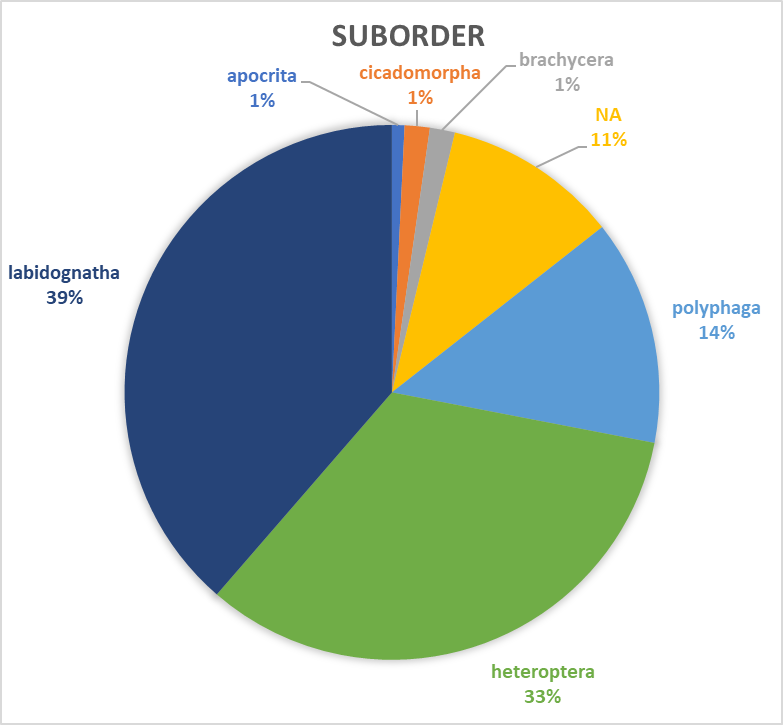
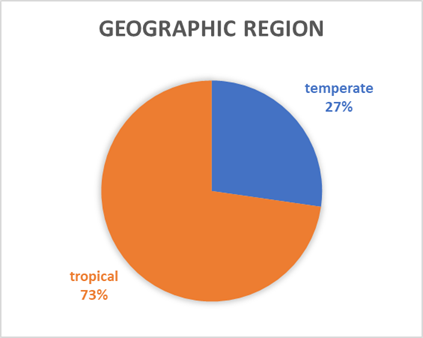
**Projet : application économétrique**

1. **Problématique**

On étudie la base de données qui contient 132 observations concernant des arthropodes. On dispose du poids (en mg) comme la variable endogène y, et des variables explicatives, qui incluent les variables quantitatives comme la longueur (en mm) et la largeur (en mm), et les variables qualitatives comme la région géographique et le sous-ordre. On a transformé les variables de poids, de longueur et de largeur en log afin d’assurer la normalité des données et l'hypothèse d'homoscédasticité.

1. **Analyse descriptive de la base de données**

132 arthropodes dans la base de données sont répartis en 7 catégories de sous-ordre. 39% d'entre eux sont des labidognatha, 33% sont des hétéroptères, 14% sont des polyphages, 11% ne peuvent être identifiés (NA), les 3% restants sont répartis également entre 3 groupes, apocrita, cicadomorpha et brachycera.

En termes de région géographique, 73% des arthropodes ont été collectés dans la région tropicale tandis que le reste a été collecté dans la région tempérée

**Analyse du poids des arthropodes**

Sur un échantillon de 132 arthropodes, le poids moyen des arthropodes est de 22,97 mg.

Le poids minimum observé est de 0,04 mg et le poids maximum est de 152,3 mg.

En regardant les indicateurs de dispersion, tel que l'étendue, la variance et l'écart type, les valeurs respectives sont significatives et révèlent une forte dispersion des poids dans l'échantillon.

Le Coefficient d'asymétrie positif met en évidence une distribution unilatéralement dévalée vers la droite par rapport à la moyenne.

Le Kurtosis ou Coefficient d'aplatissement positif montre que la distribution n'est pas plate mais relativement pointée.

1. **Interprétation des différentes régressions simples**
2. **Ln Longueur et Ln Poids**

* **Interprétation du R2**

À partir de nuage de points on obtient le R2 est égal à 0.9477. Cela signifie que 94,77% des variations de y (Ln Poids) sont expliquée linéairement par les variations de x (Ln Longueur).

* **Interprétation du nuage centré**

La fonction affine y = 2.744 x - 1.687 se simplifie en y = 2.744 x lorsqu'il s'agit des variables centrées. Dans ce dernier cas, la pente passe l'origine. Cette origine correspond au point moyen sur le graphique d'origine : une moyenne de Ln Longueur de 1.2999 pour une moyenne de Ln Poids de 1.8804.

* **Interprétation du test de nullité :**

On cherche à tester qu’il n'y a pas de relation linéaire entre Ln Poids et Ln Longueur.

H0 : a = 0

H1 : a ≠ 0

La statistique de test, sous H0 est distribuée comme suit

T ~ St (130)

La réalisation de la statistique de test

tc = (2.744 – 0)/ 0.0565 ≈ 48.5

La p-valeur vaut

2×P(T > 48.5) ≈ 0+

On voit que la p-valeur est pratiquement nulle. On rejette donc H0. Il y a bien une relation linéaire entre Ln Poids et Ln Longueur.

La région de rejet de 5% sera

RR5% =] - ∞; -1,9784 [∪] 1,9784; + ∞ [

Comme tc = 48.5 ∈ RR5%, on confirme le rejet à 5%.

* **Tests pertinents sur la pente**
* On cherche à tester que la pente est égale à 1.

H0 : a = 1

H1 : a ≠ 1

La statistique de test, sous H0 est distribuée comme suit

T ~ St (130)

La réalisation de la statistique de test

tc = (2.744 – 1)/ 0.0565 ≈ 30.85

La p-valeur vaut

2×P(T > 30.85) ≈ 0+

* On cherche à tester que la pente est égale à 2.8.

H0 : a = 2.8

H1 : a ≠ 2.8

La statistique de test, sous H0 est distribuée comme suit

T ~ St (130)

La réalisation de la statistique de test

tc = (2.744 – 2.8)/ 0.0565 ≈ -0.9901

La p-valeur vaut

2×P(T < -0.9901) ≈ 0.324

1. **Ln Largeur et Ln Poids**

* **Interprétation du R2**

À partir de nuage de points on obtient le R2 = 0.8925. Cela signifie que 89,25% des variations de y (Ln Poids) sont expliquée linéairement par les variations de x (Ln Largeur).

* **Interprétation du nuage centré**

La fonction affine y = 2.6395 x - 0.6437 se simplifie en y = 2.6395 x lorsqu'il s'agit des variables centrées. Dans ce dernier cas, la pente passe l'origine. Cette origine correspond au point moyen sur le graphique d'origine : une moyenne de Ln Largeur de 0.956 pour une moyenne de Ln Poids de 1.8804.

* **Interprétation du test de nullité :**

On cherche à tester qu’il n'y a pas de relation linéaire entre Ln Poids et Ln Largeur.

H0 : a = 0

H1 : a ≠ 0

La statistique de test, sous H0 est distribuée comme suit

T ~ St (130)

La réalisation de la statistique de test

tc = (2.6395 – 0)/ 0.0803 ≈ 32.86

La p-valeur vaut

2×P(T > 32.86) ≈ 0+

On voit que la p-valeur est pratiquement nulle. On rejette donc H0. Il y a bien une relation linéaire entre Ln Poids et Ln Largeur.

La région de rejet de 5% sera

RR5% =] - ∞; -1,9784 [∪] 1,9784; + ∞ [

Comme tc = 32.86 ∈ RR5%, on confirme le rejet à 5%.

* **Tests pertinents sur la pente**
* On cherche à tester que la pente est égale à 1.

H0 : a = 1

H1 : a ≠ 1

La statistique de test, sous H0 est distribuée comme suit

T ~ St (130)

La réalisation de la statistique de test

tc = (2.6395 – 1)/ 0.0803 ≈ 20.4

La p-valeur vaut

2×P(T > 20.4) ≈ 0+

On voit que la p-valeur est pratiquement nulle. On rejette donc H0.

* On cherche à tester que la pente est égale à 2.5.

H0 : a = 2.5

H1 : a ≠ 2.5

La statistique de test, sous H0 est distribuée comme suit

T ~ St (130)

La réalisation de la statistique de test

tc = (2.6395 – 2.5)/ 0.0803 ≈ 1.736

La p-valeur vaut

2×P(T > 1.736) ≈ 0.085

1. **Interprétation d’une régression multiple :**
2. **Estimation MCO au moyen de l’utilitaire d’analyse d’Excel**

On a le modèle estimé :

Ln Poids= β0 + β1(Ln Longueur) + β2(Ln Largeur) + β3(tropical) + β4(sous-ordre 1) + … + β9(sous-ordre 6) + ε

Où

* tropical = 1 si le région graphique est tropical, = 0 sinon
* sous-ordre 1 = 1 si le sous-ordre est cicadomorpha, = 0 sinon
* sous-ordre 2 = 1 si le sous-ordre est brachycera, = 0 sinon
* sous-ordre 3 = 1 si le sous-ordre est NA, = 0 sinon
* sous-ordre 4 = 1 si le sous-ordre est polyphaga, = 0 sinon
* sous-ordre 5 = 1 si le sous-ordre est heteroptera, = 0 sinon
* sous-ordre 6 = 1 si le sous-ordre est labidognatha, = 0 sinon

1. **Interprétation du R2**

À partir de l’estimation MCO, on voit que le R2 est approximativement égal à 97%, ce qui signifie que 97% des variations de Ln Poids sont expliquées par les variations des variables Ln Longueur, Ln Largeur, tropical et sous-ordre. En d'autres termes, 97% des valeurs de Ln Poids correspondent au modèle.

En lançant une régression pour chaque variable explicative, on obtient que 95% des variations de Ln Poids sont expliqué par la variable Ln Longueur, 89% sont expliqué par la variable Ln Largeur, 34% sont expliqué par la variable tropical et enfin 33% sont expliqué par la variable sous-ordre. Si on calcule la somme des R2 des régressions simples, on trouve 95 + 89 + 34 + 33 = 251%. Ce résultat n’est pas similaire que celui qu’on a trouvé en lançant la régression multiple et n’est pas possible car le coefficient de détermination appartient à l'intervalle de 0 à 100%. C'est parce que la variation totale de la variable endogène constitue également une part due à la corrélation interne entre les variables explicatives. Par exemple, les variables Ln Longueur et Ln Largeur sont très linéairement corrélées (R2 = 93%). Ainsi, le pouvoir d’explicatif d’une variable entre deux sera plus faible dans la régression multiple que dans la régression simple. Une partie de l’information d’une variable explicative est présente dans l’autre.

1. **Interprétation du test de nullité de la régression**

Maintenant on cherche à tester que toutes les variables explicatives n’ont pas d’impact linéaire sur Ln Poids. Les hypothèses nulle et alternative sont les suivantes

H0 : β1 = β2 = … = β9 = 0

H1 : au moins un des neuf paramètres est non nul

On a le modèle contraint

Ln Poids = β0 + ε

La réalisation de la statistique de test s’écrit

fc = [(SCRc − SCRnc )/9]/ sigma2hat

= [(SCTnc − SCRnc )/9]/ sigma2hat

= [(436.93 - 12.25)/9]/ 0.1004 ≈ 470.93

Sachant que la statistique de test, sous H0 est distribuée comme suit

F ∼ F (9, 122)

La p-valeur vaut

P(F > 470.93) ≈ 0+

D’où on rejette très largement H0 car la p-valeur est presque nulle. Cela signifie que le modèle non contraint offre un meilleur ajustement aux données qu'un modèle qui ne contient aucune variable explicative. En d’autres termes, le modèle non contraint mieux explique la variance de Ln Poids que le modèle qui contient uniquement la constant.

1. **Interprétation des tests de nullité de chaque pente**

* **Test de nullité d’un paramètre associé à la variable Ln Longueur**

On cherche à tester si la variable Ln Longueur a un impact linéaire sur Ln Poids.

Les hypothèses nulle et alternative sont les suivantes :

H0 : β1 = 0

H1 : β1 ≠ 0

On a le modèle contraint

Ln Poids = β0 + β2(Ln Largeur) + β3(tropical) + β4(sous-ordre 1) + … + β9(sous-ordre 6) + ε

La réalisation de la statistique de test s’écrit

fc = [(SCRc − SCRnc )/1]/ sigma2hat

= [(39.0502 - 12.2461)/1]/ 0.1004 ≈ 267.0329

Sachant que la statistique de test, sous H0 est distribuée comme suit

F ∼ F (1, 122)

La p-valeur vaut

P(F > 267.0329) ≈ 0+

On déduit que l’augmentation de la SCR est trop importante pour être due au hasard et la p-valeur est presque nulle. On rejette donc très largement H0 que la pente de la variable Ln Longueur est nulle. Ln Longueur a bien un impact linéaire sur Ln Poids.

* **Test de nullité d’un paramètre associé à la variable Ln Largeur**

On cherche à tester si la variable Ln Largeur a un impact linéaire sur Ln Poids.

Les hypothèses nulle et alternative sont les suivantes :

H0 : β2 = 0

H1 : β2 ≠ 0

On a le modèle contraint

Ln Poids = β0 + β1(Ln Longueur) + β3(tropical) + β4(sous-ordre 1) + … + β9(sous-ordre 6) + ε

La réalisation de la statistique de test s’écrit

fc = [(SCRc − SCRnc )/1]/ sigma2hat

= [(12.3481 - 12.2461)/1]/ 0.1004 ≈ 1.0162

Sachant que la statistique de test, sous H0 est distribuée comme suit

F ∼ F (1, 122)

La p-valeur vaut

P(F > 1.0162) ≈ 0.315

On déduit que l’augmentation de la SCR est très faible et la p-valeur est importante. On ne rejette pas donc H0 que la pente de la variable Ln Largeur est nulle. Ln Largeur n’a pas d’impact linéaire sur Ln Poids, toutes les choses égales par ailleurs. Autrement dit, le pouvoir explicatif de cette variable n’est pas statiquement significatif.

* **Test de nullité d’un paramètre associé à la variable tropical**

On cherche à tester si la variable tropical a un impact linéaire sur Ln Poids.

Les hypothèses nulle et alternative sont les suivantes :

H0 : β3 = 0

H1 : β3 ≠ 0

On a le modèle contraint

Ln Poids = β0 + β1(Ln Longueur) + β2(Ln Largeur) + β4(sous-ordre 1) + … + β9(sous-ordre 6) + ε

La réalisation de la statistique de test s’écrit

fc = [(SCRc − SCRnc )/1]/ sigma2hat

= [(12.2481 - 12.2461)/1]/ 0.1004 ≈ 0.0202

Sachant que la statistique de test, sous H0 est distribuée comme suit

F ∼ F (1, 122)

La p-valeur vaut

P(F > 0.0202) ≈ 0.8872

On déduit que l’augmentation de la SCR est très faible et la p-valeur est très élevée. On ne rejette pas donc H0 que la pente de la variable tropical est nulle. Cette variable muette n’a pas d’impact linéaire sur Ln Poids, toutes les choses égales par ailleurs. Autrement dit, elle n’est pas statiquement significative.

* **Test de nullité des paramètres associé à la variable sous-ordre**

On cherche à tester si la variable sous-ordre a un impact linéaire sur Ln Poids.

Les hypothèses nulle et alternative sont les suivantes

H0 : β4 = β5 = … = β9 = 0

H1 : au moins un des six paramètres est non nul

On a le modèle contraint

Ln Poids = β0 + β1(Ln Longueur) + β2(Ln Largeur) + β3(tropical) + ε

La réalisation de la statistique de test s’écrit

fc = [(SCRc − SCRnc )/6]/ sigma2hat

= [(22.7347 - 12.2461)/6]/ 0.1004 ≈ 17.415

Sachant que la statistique de test, sous H0 est distribuée comme suit

F ∼ F (6, 122)

La p-valeur vaut

P(F > 17.415) ≈ 0+

On déduit que l’augmentation de la SCR est élevée et la p-valeur est presque nulle. On rejette donc H0 que la pente de la variable sous-ordre est nulle. Cette variable a un impact linéaire sur Ln Poids, toutes les choses égales par ailleurs.

1. **Proposition d’un modèle plus parcimonieux**

Après de faire plusieurs estimations conduisant chacune à la suppression de variables, on constate l’apport des variables Ln Largeur et tropical sur le pouvoir explicatif est très faible. Elles ne sont pas statistiquement significatives par rapport aux variables Ln Longueur et sous-ordre. Par conséquence, on propose un nouveau modèle qui ne contient que les variables explicatives Ln Longueur et sous-ordre. Il explique mieux les variations de Ln Poids et est également statistique significatif.