



UNIVERSITÉ
CAEN
NORMANDIE

Université de Caen
Normandie



IUT Grand Ouest Normandie

Bachelor Universitaire de Technologie
Science des Données
Campus de Lisieux

Science des Données 1 - SAE 2.03
Régression sur données réelles

Analyse et régression sur une population de manchots



Auteurs

Prieur Noam
Diop Mandir
Yon Anthony

Année universitaire 2023-2024

Table des matières

Introduction	3
1 Analyse exploratoire et régression linéaire de l'association entre le poids des manchots et les variables explicatives	4
1.1 Analyse exploratoire	4
1.2 Modélisation	6
2 Analyse selon sexe	8
2.1 Analyse exploratoire	8
2.2 Modélisation	9
3 Analyse selon l'espèce	11
3.1 Analyse exploratoire	11
3.2 Modélisation	13
Conclusion	16
Annexe 1	17
Annexe 2	18
Annexe 3	19
Table des figures	20

Introduction

Un ornithologue présent dans une station située dans les **archipels palmer** en Antarctique souhaite étudier une éventuelle association entre le poids des manchots et 5 variables telles que la longueur de leurs nageoires, la longueur et la profondeur de leur crête supérieure du bec, le sexe ou l'espèce. Dans ce cadre, il dispose d'un fichier contenant des données portant sur un échantillon de 344 manchots.

Ce jeu de données (Fig. 1) contient en particulier les informations suivantes :

- l'espèce¹
- l'île d'origine
- la longueur de la crête supérieur du bec
- la profondeur de la crête supérieur du bec
- la longueur des nageoires
- le poids
- le sexe
- l'année d'étude

#	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
1	Adelie	Torgersen	39.1	18.7	181	3750	male	2007
2	Adelie	Torgersen	39.5	17.4	186	3800	female	2007
3	Adelie	Torgersen	40.3	18.0	195	3250	female	2007
5	Adelie	Torgersen	36.7	19.3	193	3450	female	2007
6	Adelie	Torgersen	39.3	20.6	190	3650	male	2007
7	Adelie	Torgersen	38.9	17.8	181	3625	female	2007
8	Adelie	Torgersen	39.2	19.6	195	4675	male	2007
13	Adelie	Torgersen	41.1	17.6	182	3200	female	2007
14	Adelie	Torgersen	38.6	21.2	191	3800	male	2007
15	Adelie	Torgersen	34.6	21.1	198	4400	male	2007
16	Adelie	Torgersen	36.6	17.8	185	3700	female	2007
17	Adelie	Torgersen	38.7	19.0	195	3450	female	2007
18	Adelie	Torgersen	42.5	20.7	197	4500	male	2007
19	Adelie	Torgersen	34.4	18.4	184	3325	female	2007
20	Adelie	Torgersen	46.0	21.5	194	4200	male	2007
21	Adelie	Biscoe	37.8	18.3	174	3400	female	2007
22	Adelie	Biscoe	37.7	18.7	180	3600	male	2007
23	Adelie	Biscoe	35.9	19.2	189	3800	female	2007

Fig. 1. Extrait des données

On se propose ensuite de réaliser une analyse de ces données en trois parties. Une première partie sera consacrée à l'analyse exploratoire et la modélisation de la régression linéaire de l'association entre le poids des manchots et les variables explicatives. Une deuxième partie mettra en perspective l'étude de la première partie selon le sexe et la troisième, selon l'espèce.

1. informations complémentaires en annexe

1 Analyse exploratoire et régression linéaire de l'association entre le poids des manchots et les variables explicatives

Cette première section est consacrée à l'étude d'une éventuelle association entre le poids et différentes caractéristiques des manchots que sont la longueur de leurs nageoires, la longueur et la profondeur de leur crête supérieure du bec . On va donc : réaliser une analyse exploratoire, créer pour chaque situation un modèle de régression linéaire puis en déduire une interprétation.

1.1 Analyse exploratoire

Premièrement, croisons nos données afin d'observer la distribution des variables et de commencer à visualiser une potentielle association entre le poids des manchots et chaque caractéristiques énoncées précédemment.

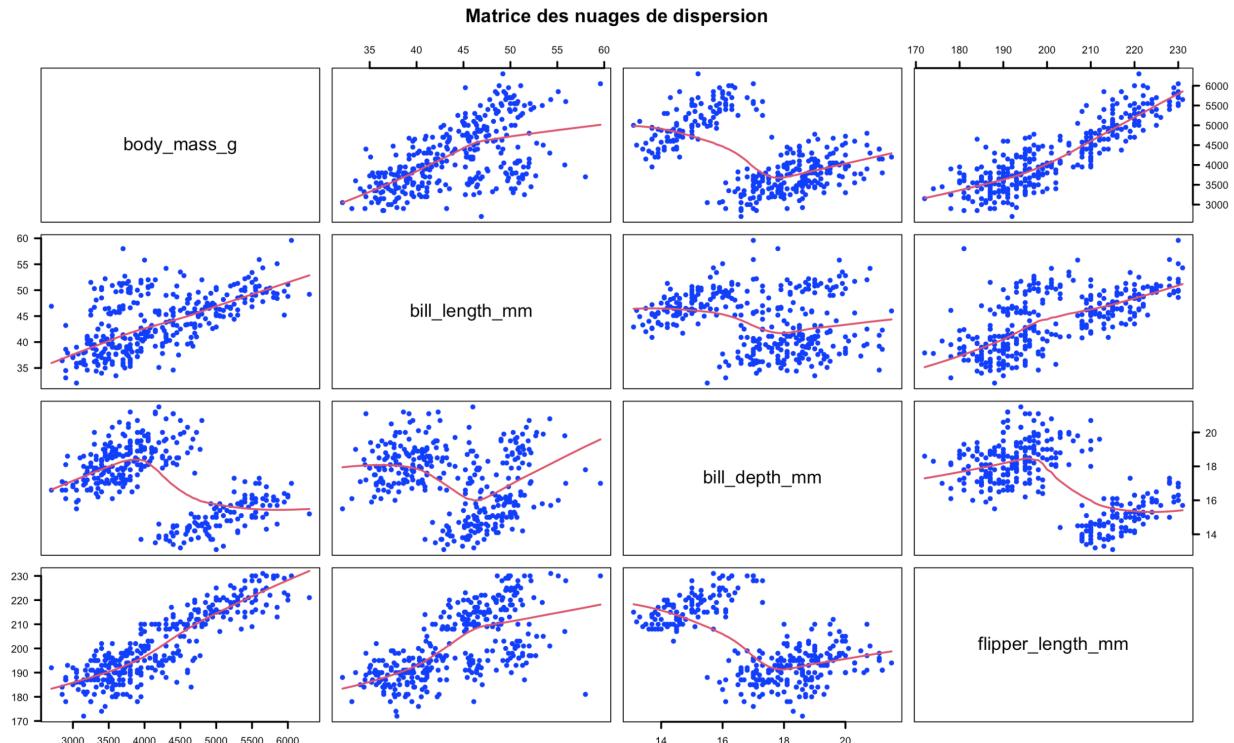


Fig. 2. Matrice des nuages de dispersion

Grâce à cette sortie on peut observer (Fig. 2) ci-dessus qu'il existe bel et bien une association linéaire qui est positive entre le poids et la longueur du bec, mais aussi entre le poids et la longueur des nageoires puis négative entre la variable poids et la variable profondeur de la crête supérieure du bec puisque les points des nuages de dispersion sont sensiblement distribués le long d'une droite constante. Cependant, vu la forme de la courbe de régression lissée on pourra accepter l'hypothèse qu'il existe une variable de confusion qui est sans doute l'espèce (Species) ou le sexe .

Désormais, afin de compléter cette première étape nous représentons l'histogramme et la densité lissée pour chaque variable

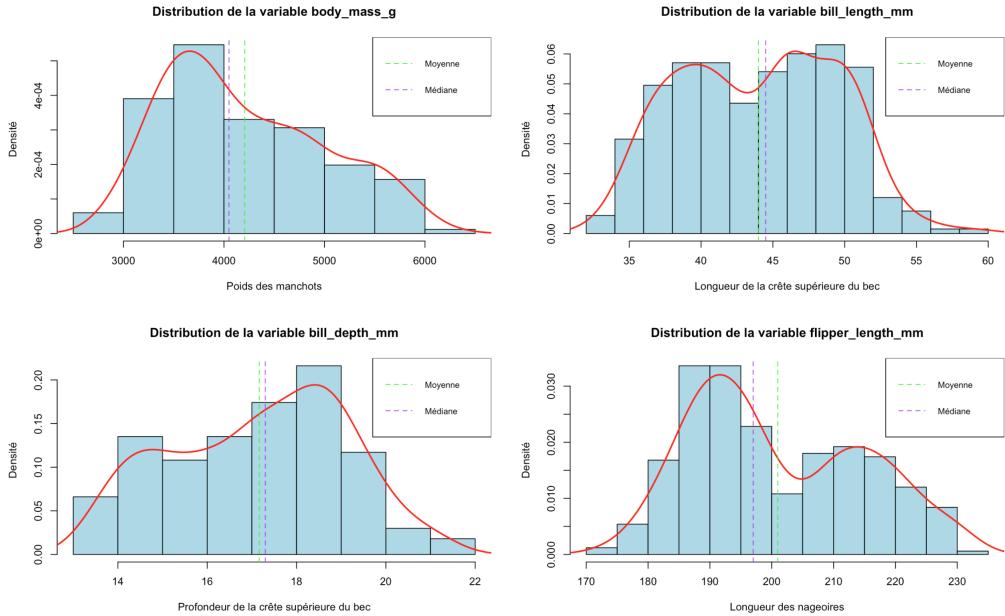


Fig. 3. Histogrammes et densités lissées

On observe ici (Fig. 3) des distributions légèrement dissymétriques mais pas suffisamment pour déranger notre phase de modélisation par la suite. On souhaite tout de même vérifier cela alors on se propose de réaliser le même exercice en représentant les nuages de dispersion avec une régression lissée.

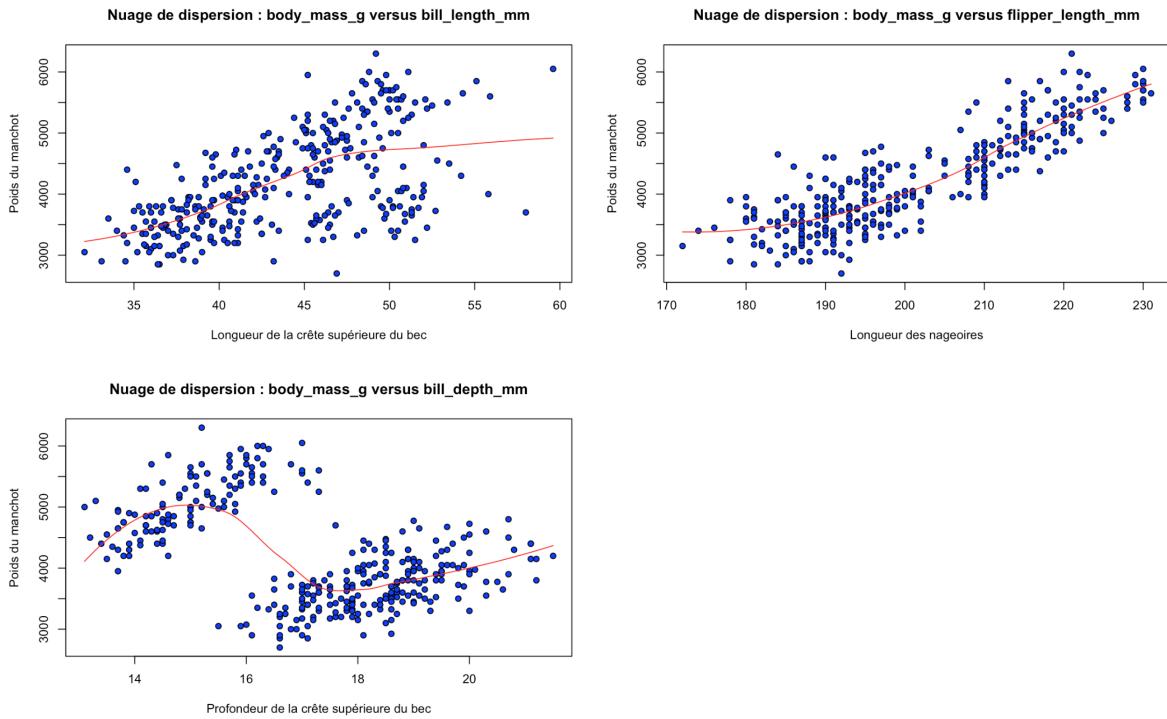


Fig. 4. Nuages de dispersions et régression lissée

Ces graphiques viennent confirmer la première analyse réalisée sur la matrice des nuages de dispersion, à savoir une association positive entre les variables "longueur du bec" et "longueur de la nageoire" et la variable "poids", association plutôt linéaire comme en témoigne la régression lissée.

Cependant, lorsqu'on observe le nuage de dispersion entre la poids et la variable profondeur de la crête supérieur, on constate une absence d'association linéaire car les points semble distribués selon une structure différente de celle d'une droite. Cette absence d'association pourrait s'expliquer par la présence d'une variable de confusion ou variable cachée.

Maintenant que l'étape de l'analyse exploratoire est terminée et que nous avons prouvé que la modélisation est possible, passons à l'étape suivante.

1.2 Modélisation

Dans un second temps, on se propose de construire un modèle de régression linéaire simple entre la variable à expliquer "poids" et chacune des variables explicatives puis de déterminer le meilleur modèle en s'appuyant sur le coefficient de détermination.

Pour commencer, représentons les nuages de dispersion précédents en y ajoutant la droite des moindres carrés.

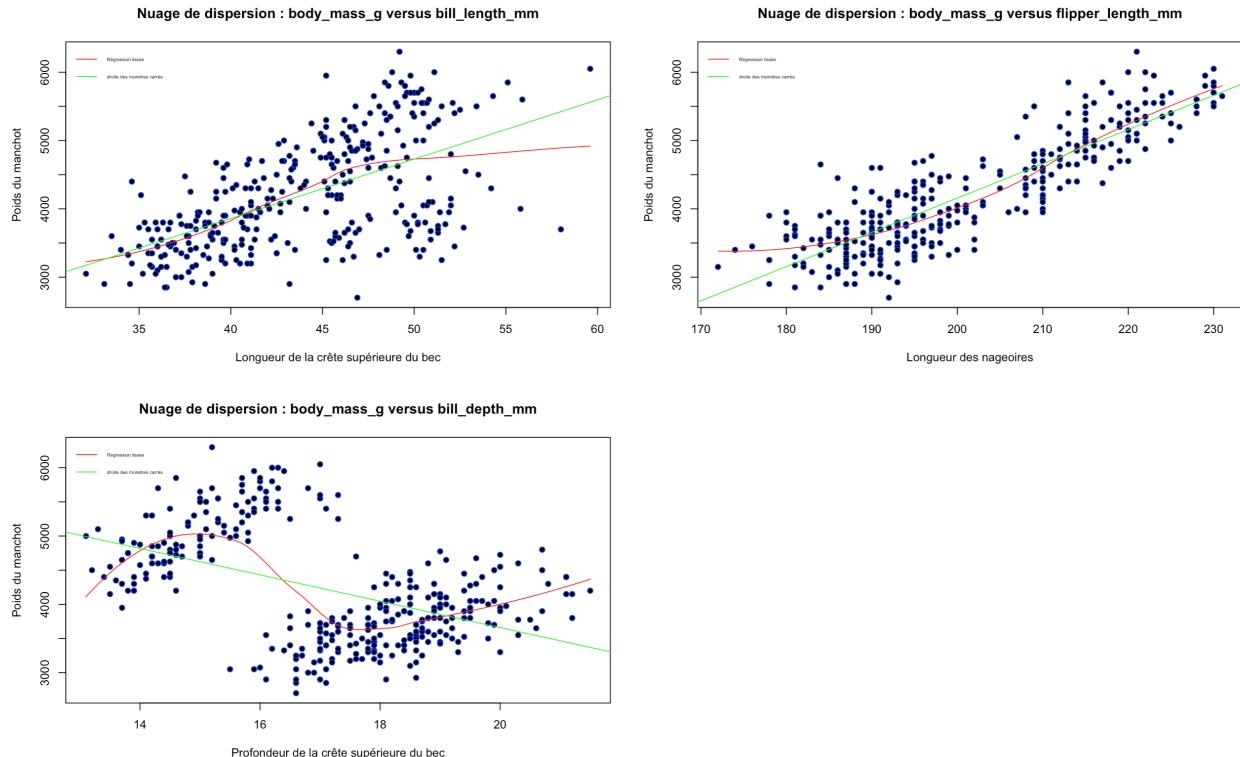


Fig. 5. Nuages de dispersions avec régression lissée et droites des moindres carrés

On se propose de calculer le coefficient de détermination pour les trois associations. Celle qui affiche le coefficient le plus élevé est l'association entre le "poids" et la "longueur de la crête", avec un coefficient de détermination de 78,29% soit une association linéaire forte. Cela suggère que 78,29% de la variation observée dans le poids peut être expliquée par la variation de la longueur de la crête dans notre modèle de régression.

De plus on remarque ici que dans le deuxième nuage de dispersion, la droite des moindres carrés est presque confondue avec la droite de la régression lisée, ce qui témoigne d'une bonne modélisation.

On se propose alors de se concentrer sur ce nuage de dispersion et d'en extraire son équation de la droite de régression lissée.

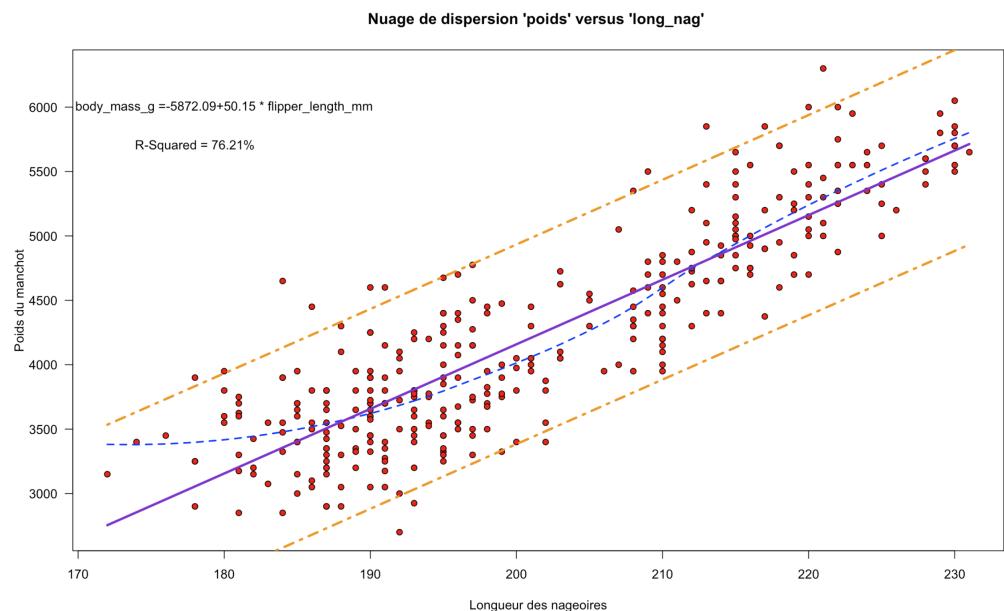


Fig. 6. Nuage de dispersion "poids" versus "longueur des nageoires"

Sur ce graphique en interprétant la droite des moindres carrés on peut constater que lorsque la longueur des nageoires augmente de 1 millimètre alors le poids du manchot augmente en moyenne de 50,15 grammes.

$$\hat{m}(\text{poids}) = -5872.09 + 50,15 * \text{longueur des nageoires}$$

2 Analyse selon sexe

Dans cette deuxième section, on se propose de mettre en perspective l'étude précédente selon le sexe c'est-à-dire analyser la présence d'une éventuelle association entre les caractéristiques physiques des manchots étudiées précédemment et le sexe.

2.1 Analyse exploratoire

Sur ce graphique (Fig. 7), nous pouvons observer qu'il existe une association linéaire positive entre la variable poids et les variables suivantes : la longueur de la crête supérieure du bec pour le sexe masculin et la longueur des nageoires, ainsi qu'une association négative entre la variable poids et la profondeur de la crête supérieure du bec.

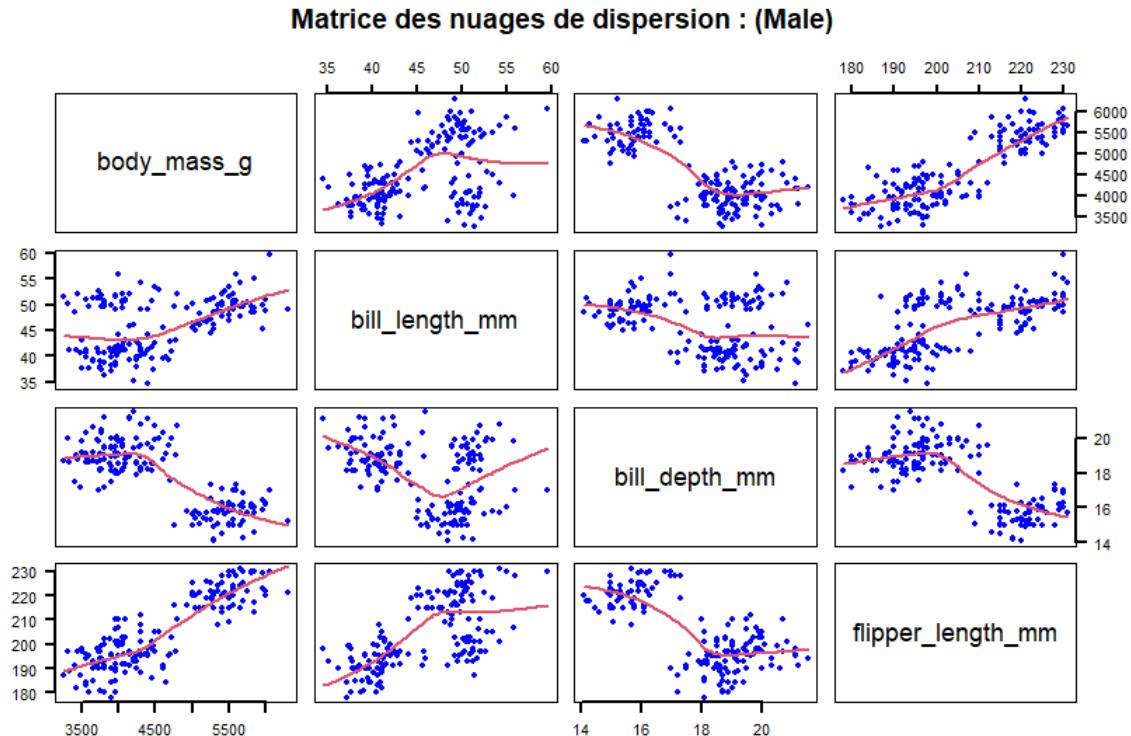


Fig. 7. Matrice des nuages de dispersion - Mâle

Via la Fig. 8, qui représente la matrice des nuages de dispersion du sexe féminin, on constate qu'il existe une association linéaire positive entre la variable poids et les variables suivantes : la longueur de la crête supérieure du bec pour le sexe masculin et la longueur des nageoires, ainsi qu'une association négative entre la variable poids et la profondeur de la crête supérieure du bec.

Matrice des nuages de dispersion : (Female)

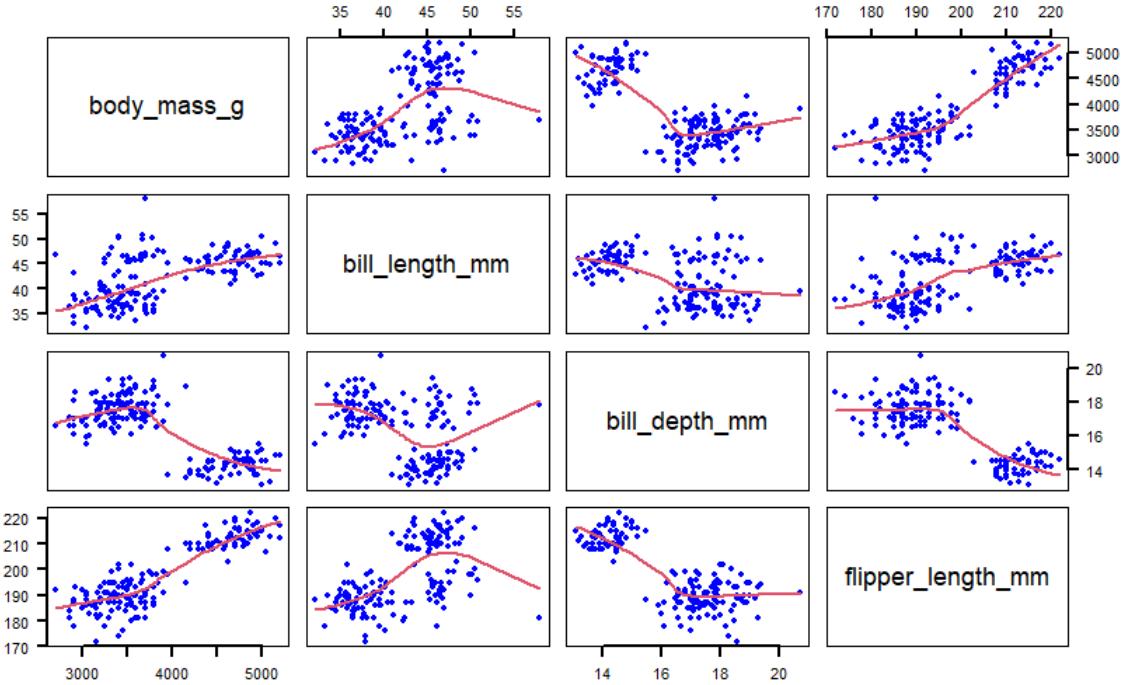


Fig. 8. Matrice des nuages de dispersion - Femelle

2.2 Modélisation

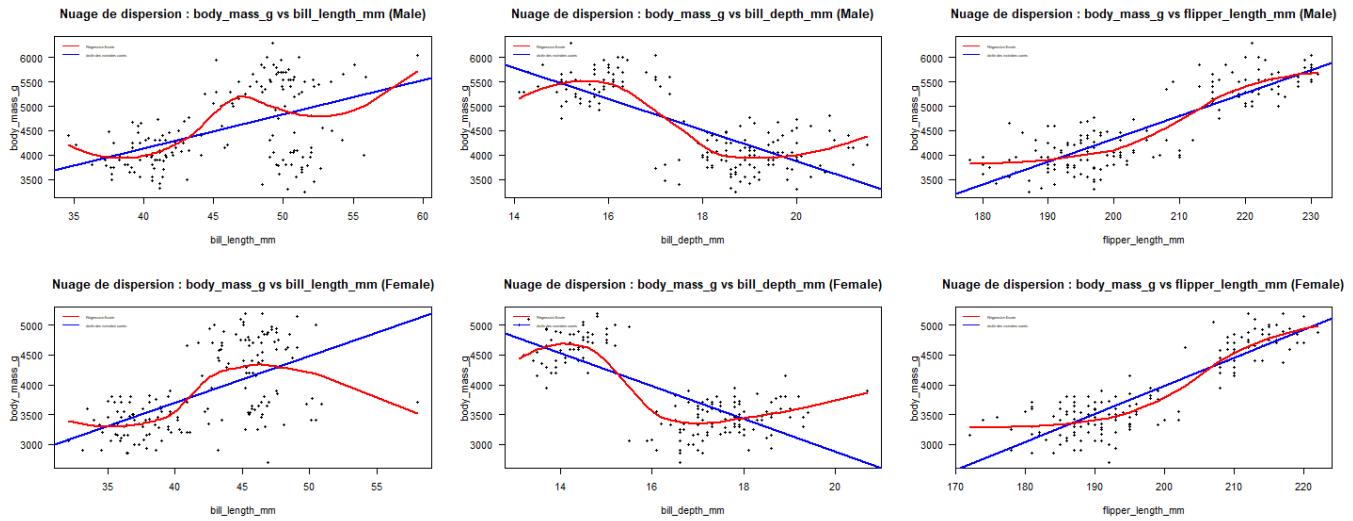


Fig. 9. Nuages de dispersion avec régression lissée et droites des moindres carrés - Sexe

En observant les nuages de dispersion des deux sexes (mâles et femelles) croisant la variable à expliquer, le poids, avec les variables explicatives telles que la longueur de la crête supérieure du bec et la longueur des nageoires, nous pouvons affirmer qu'il existe une association linéaire positive entre la variable poids et ces variables. Cela est soutenu par la répartition des points le long d'une droite non horizontale, ainsi que par la proximité de la courbe de régression lissée avec une droite.

Cependant, lorsqu'on observe le nuage de dispersion entre la variable poids et la variable profondeur de la crête supérieure, on constate une absence d'association linéaire car les points semblent distribués selon une structure différente de celle de droite. Cette absence d'association pourrait s'expliquer par la présence d'une variable de confusion ou variable cachée.

On se propose alors de se concentrer sur les modèles les plus pertinents et d'en extraire les équations de la droite des moindres carrés.

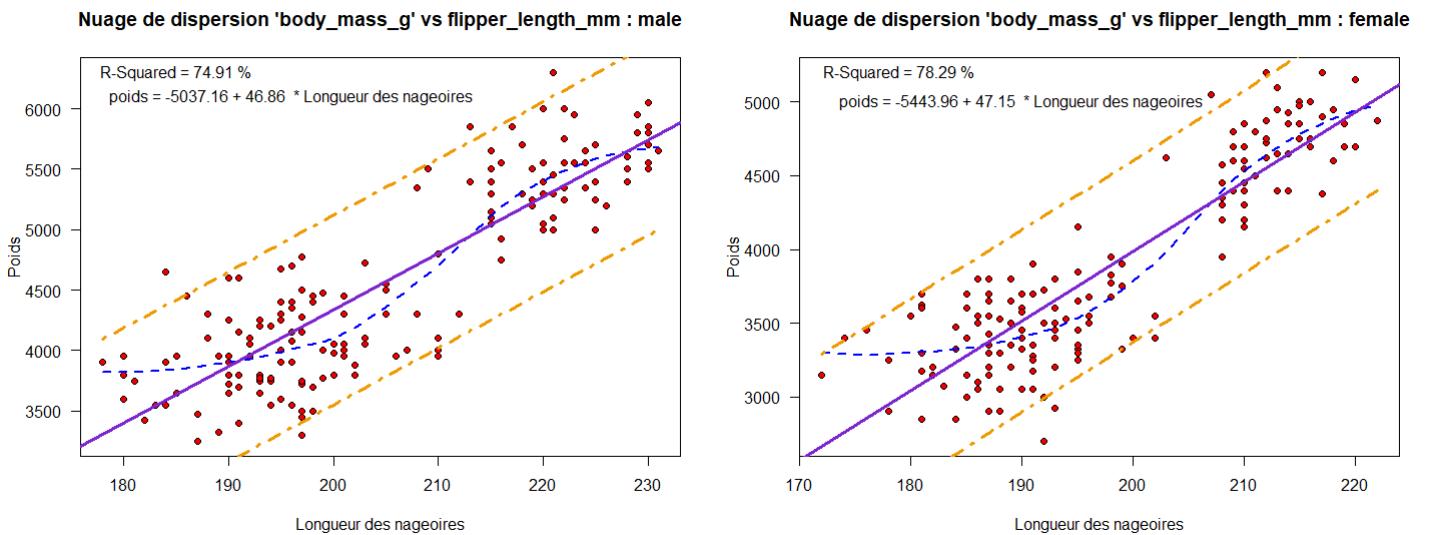


Fig. 10. Nuages de dispersion : poids versus longueur des nageoires - Mâle et Femelle

Ensuite, nous proposons de calculer le coefficient de détermination des trois modèles pour chaque sexe. L'association qui affiche le coefficient le plus élevé est celle entre le "poids" et la "longueur des nageoires", avec un coefficient de détermination de 74.91% pour les mâles et de 78.29% pour les femelles, soit des associations linéaires fortes. Cela suggère que 74.91% de la variation observée dans le poids peut être expliquée par la variation de la longueur des nageoires dans notre modèle de régression pour les mâles et 78.29% respectivement pour les femelles.

Sur ce graphique en interprétant la droite des moindres carrés on peut constater que lorsque la longueur des nageoires augmente de 1 millimètre alors le poids du manchot augmente en moyenne de 46.86 grammes pour les mâles et de 47.15 pour les femelles.

$$\hat{m}(\text{poids_male}) = -5037.16 + 46.86 * \text{longueur des nageoires}$$

$$\hat{m}(\text{poids_female}) = -5439.96 + 47.15 * \text{longueur des nageoires}$$

Quel que soit le sexe, le modèle croisant la variable poids et la variable longueur des nageoires reste le plus pertinent pour expliquer les variations du poids des manchots. Reste à supposer qu'il existe une association entre le poids et l'espèce des manchots.

3 Analyse selon l'espèce

Dans cette troisième section, tout comme la deuxième, on se propose de mettre en perspective notre première étude mais cette fois ci, selon l'espèce (Adélie, Chinstrap, Gentoo).

3.1 Analyse exploratoire

Via ces trois graphiques représentant les matrices de dispersion des espèces (Fig. 11), (Fig. 12), et (Fig. 13), nous pouvons observer qu'il existe une association linéaire positive entre la variable poids et les variables explicatives suivantes : la longueur de la crête supérieure du bec, la profondeur de la crête supérieure du bec et la longueur des nageoires.

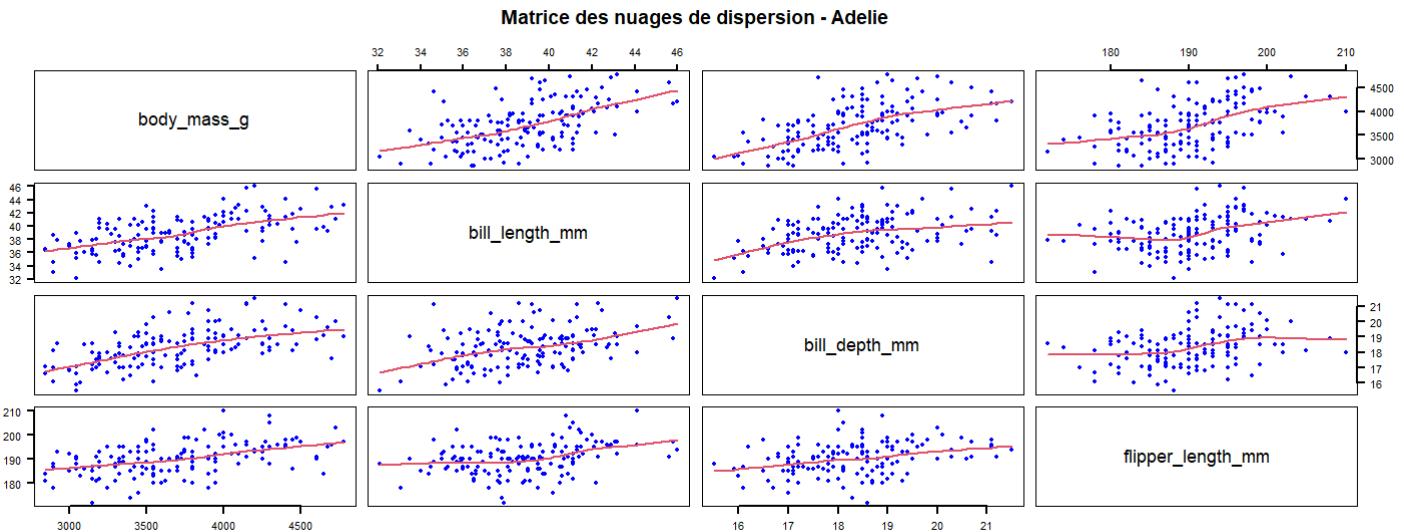


Fig. 11. Matrice des nuages de dispersion - espèce Adélie

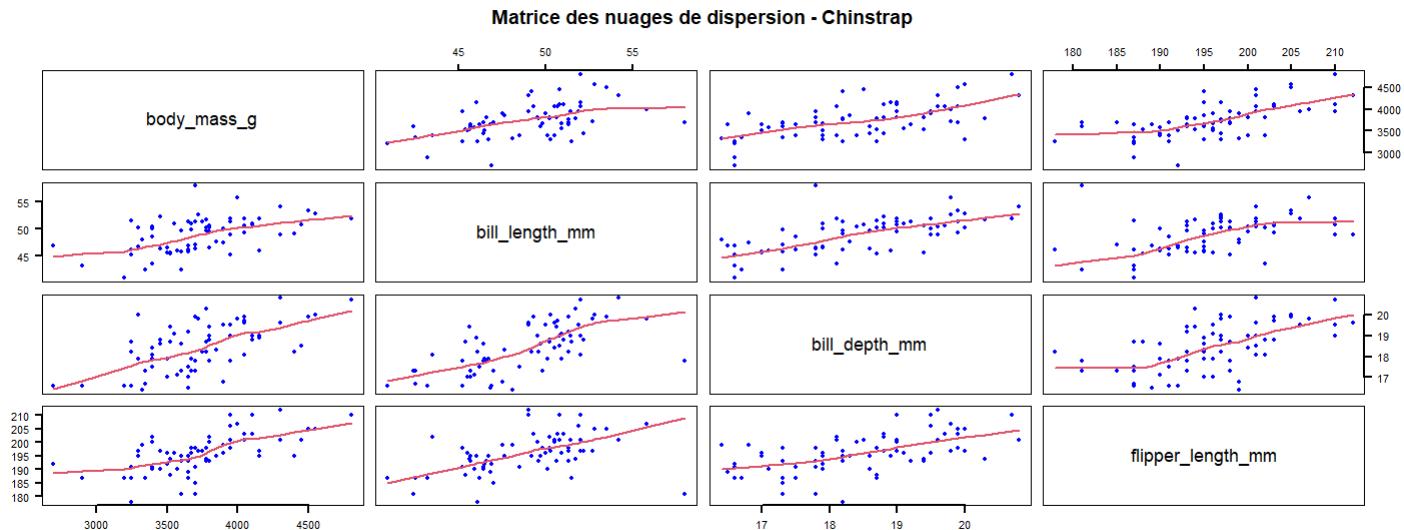


Fig. 12. Matrice des nuages de dispersion - espèce Chinstrap

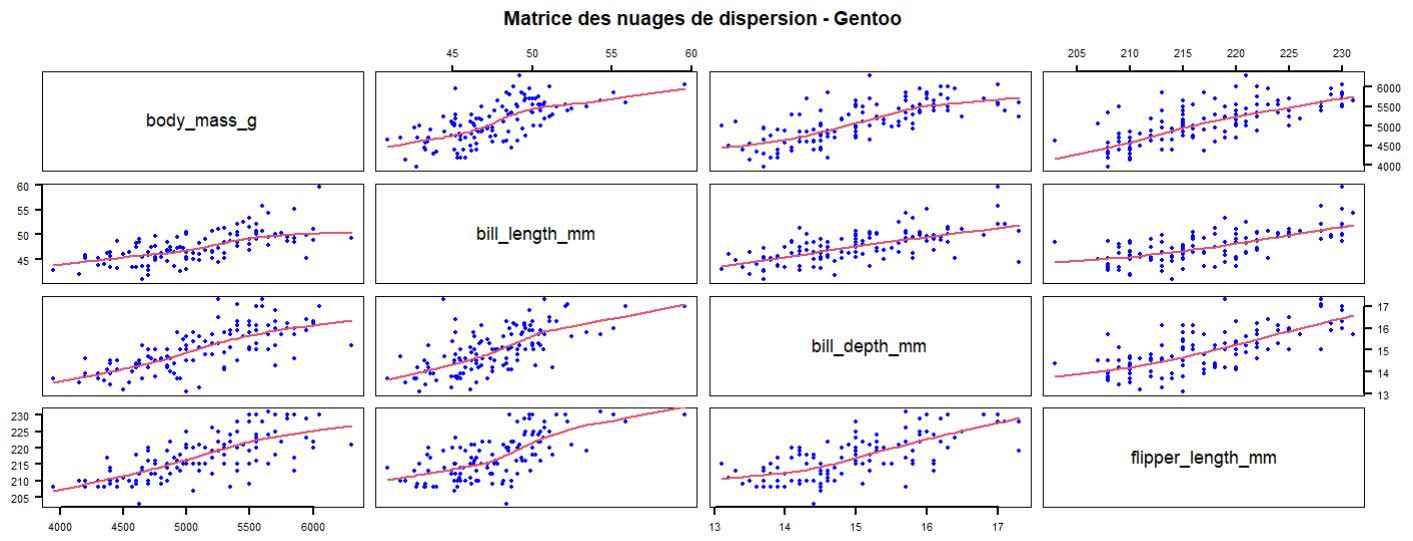
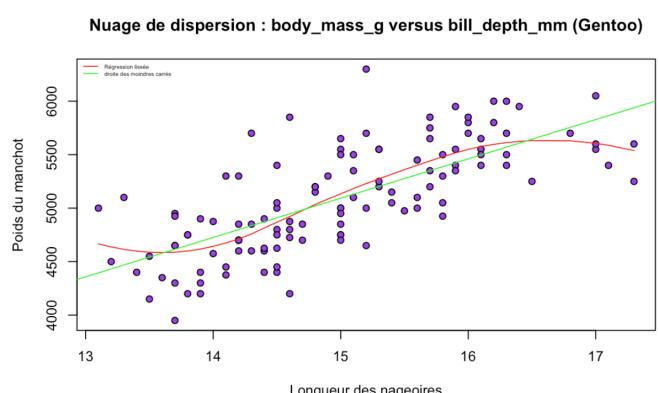
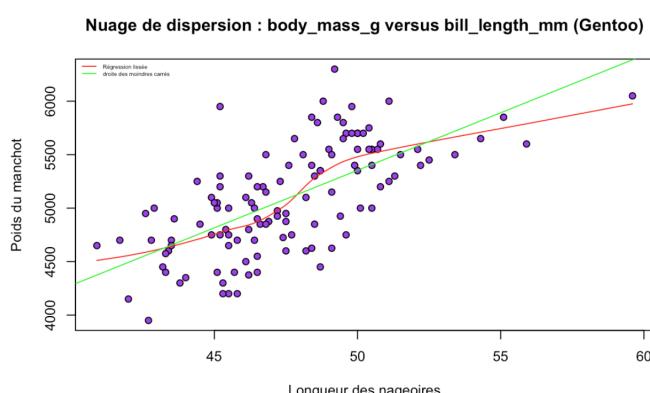
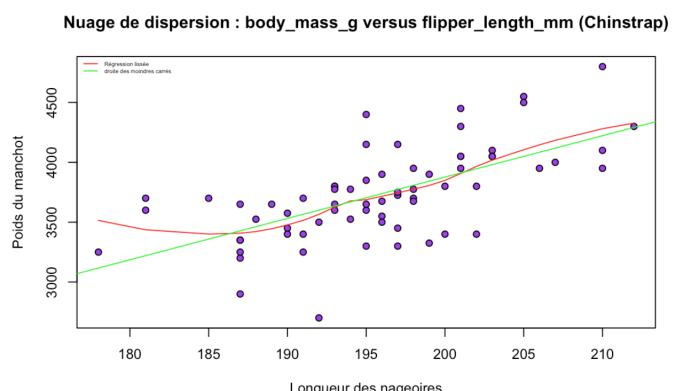
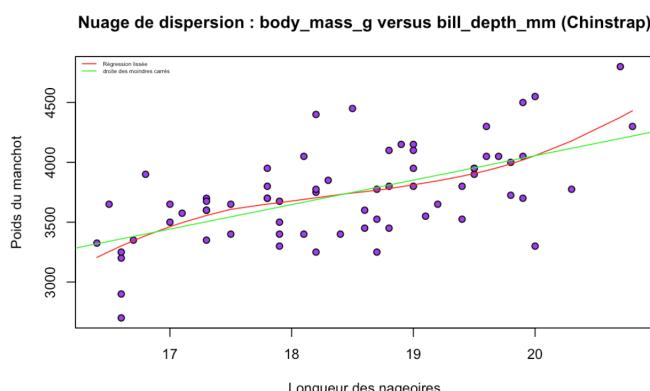
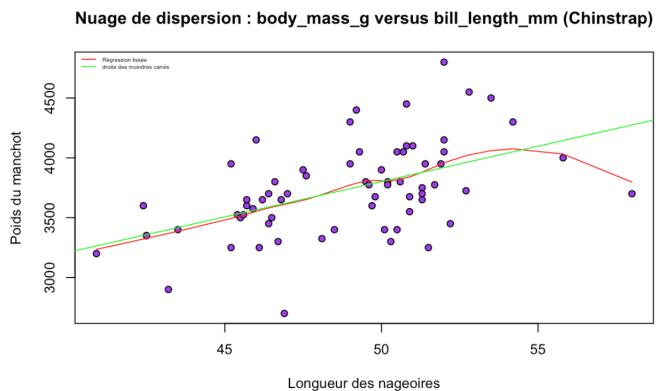
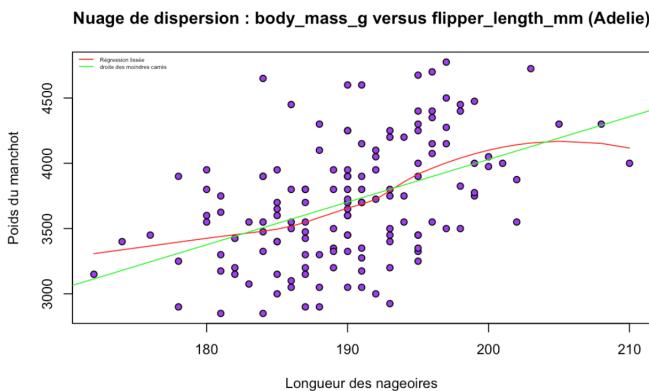
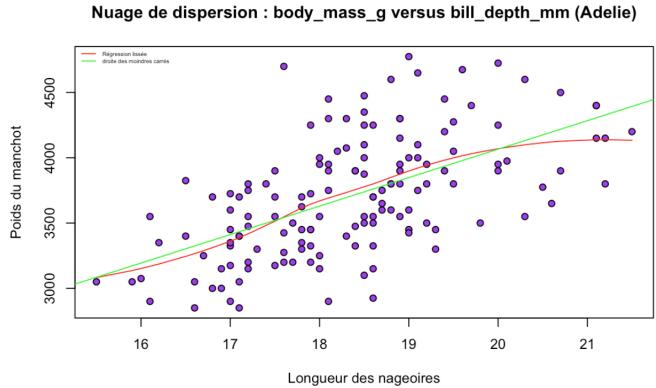
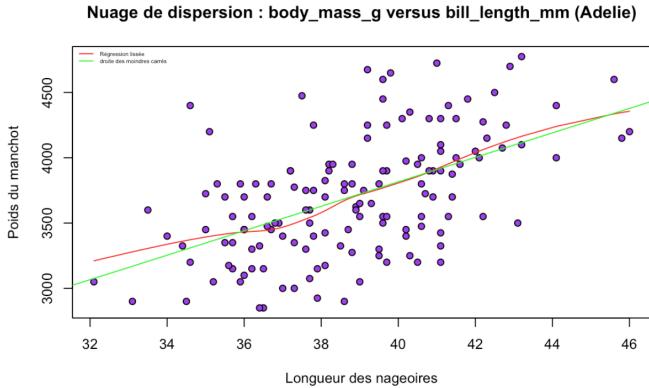


Fig. 13. Matrice des nuages de dispersion - espèce Gentoo

3.2 Modélisation

En examinant les nuages de dispersions des trois espèces croisant la variable à expliquer "poids" avec les variables explicatives, on peut remarquer qu'il y a une association linéaire positive entre ces variables. (Fig. 14)



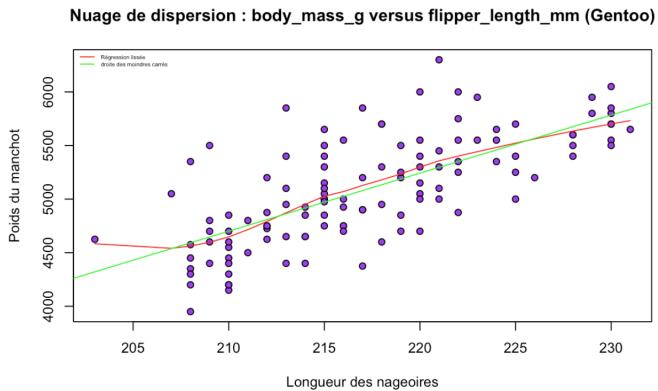


Fig. 14. Nuages de dispersion avec régression lissée et droites des moindres carrés - Espèces

On se propose de calculer le coefficient de détermination pour les neuf modèles. L'association qui affiche le coefficient le plus élevé est celle entre le poids et la profondeur du bec de l'espèce Gentoo, avec un coefficient de détermination de 52.27% soit une association linéaire modérée. Cependant, nous procédons ensuite à la représentation des modèles les plus pertinents pour chaque espèce.

En observant le graphique ci-après et l'équation de la droite des moindres carrés, nous constatons qu'une augmentation de 1 millimètre de la profondeur de la crête supérieure du bec entraîne en moyenne une augmentation de 218.21 grammes dans le poids du manchot Adelie. On observe également que dans notre modèle de régression, 33.66% de la variation observée dans le poids peut être expliquée par la profondeur de la crête supérieure du bec.

$$\hat{m}(\text{poids}) = -297.38 + 218.21 * \text{profondeur du bec}$$

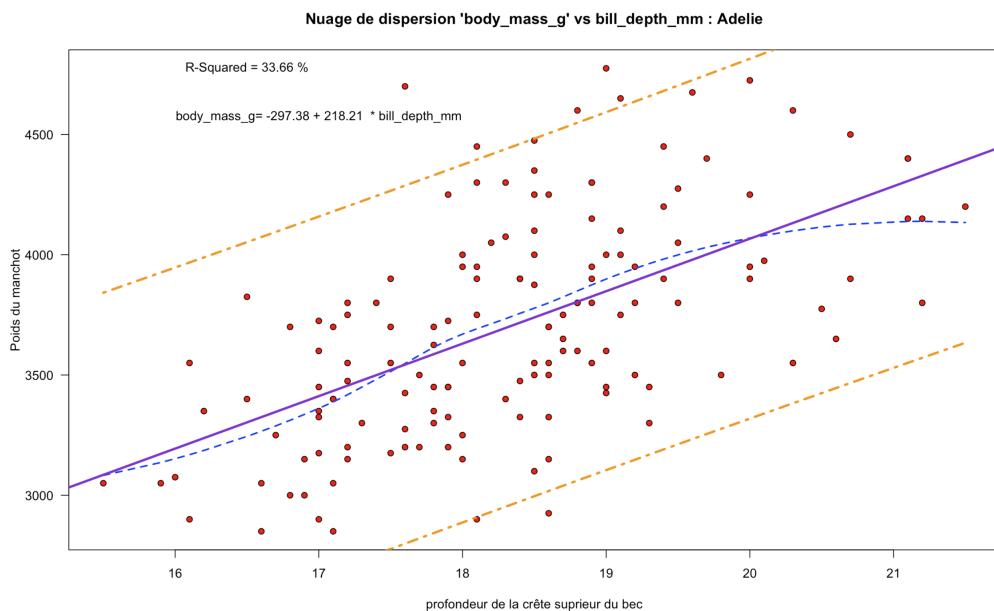


Fig. 15. Nuage de dispersion : poids versus profondeur de la crête supérieure du bec : Adelie

Après examen du graphique ci-dessous et de l'équation de la droite des moindres carrés, il apparaît qu'une augmentation de 1 millimètre de la longueur des nageoires entraîne en moyenne une augmentation de 218.21 grammes dans le poids du manchot Chinstrap. On constate également que 41.16% de la variation observée pour le poids peut être expliquer par la variation de la longueur des nageoires dans notre modèle de régression.

$$\hat{m}(\text{poids}) = -3037.2 + 34.57 * \text{longueur des nageoires}$$

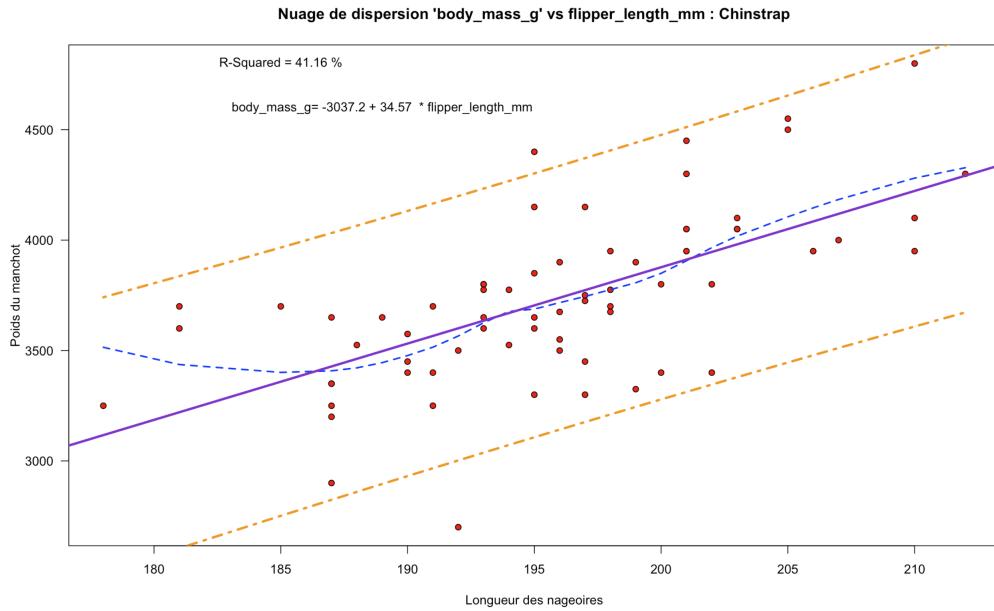


Fig. 16. Nuage de dispersion : poids versus longueur des nageoires : Chinstrap

Grâce à ce graphique ci-dessus et à l'équation de la droite des moindres carrés, on peut constater que lorsque la profondeur du bec augmente de 1 millimètre alors le poids du manchot Gentoo augmente en moyenne de 367.7 grammes.

$$\hat{m}(\text{poids}) = -421.82 + 367.7 * \text{profondeur du bec}$$

Ainsi 52.27% de la variation observée pour le poids peut être expliquer par la variation de la profondeur du bec dans notre modèle de régression.

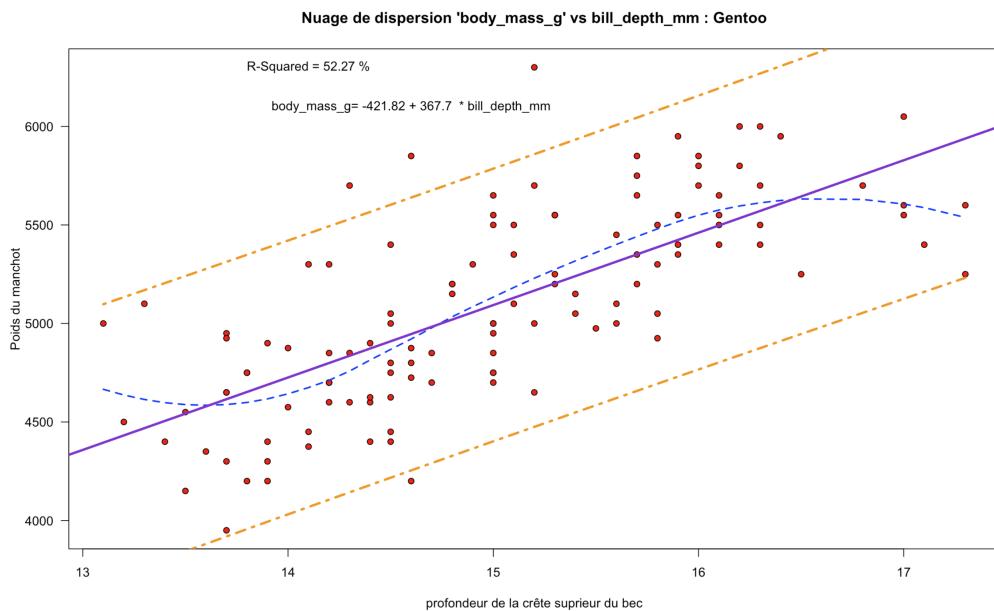


Fig. 17. Nuage de dispersion : poids versus profondeur de la crête supérieur du bec : Gentoo

Cette analyse confirme l'hypothèse selon laquelle le modèle établi sur le poids en fonction des variables explicatives, en tenant compte de l'espèce, est plus pertinent pour déterminer la variation du poids des manchots. Sur l'ensemble de ces modèles, la droite des moindres carrés est presque confondue avec la régression lissée.

Conclusion

On remarque tout d'abord des associations linéaires entre le poids et la longueur de la crête, la profondeur de la crête et la longueur de la nageoire. Grâce au modèle de régression linéaire on sait que c'est la longueur de la nageoire qui a l'incidence la plus importante sur le poids.

Nous avons ensuite observer ces associations en selon du sexe, les résultats de la modélisation suggèrent que la longueur de la nageoire est la variable ayant l'impact le plus significatif sur le poids, tant pour les manchots mâle que pour les manchots femelles.

Enfin, nous avons examiné les données par espèce et avons observé des associations linéaires entre le poids et la longueur de la crête, la profondeur de la crête et la longueur de la nageoire pour les trois espèces. Les résultats de la modélisation montrent que la profondeur du bec est la variable ayant l'impact le plus important sur le poids pour l'espèce Gentoo, tout comme pour l'espèce Adélie, tandis que pour l'espèce Chinstrap, c'est la longueur de la nageoire qui est la variable la plus importante. En conclusion, l'espèce est la variable qui détermine le mieux la variation du poids des manchots.

Annexe 1

Informations complémentaires sur le manchot Adélie

Description

Le manchot Adélie est l'image d'Epinal du manchot: il est entièrement noir sur le dos, blanc sur le ventre, la tête est noire avec un anneau blanc autour de l'oeil. Son nom provient du prénom Adèle de la femme de l'explorateur français Dumont d'Urville. Il fait partie des manchots les mieux connus et étudiés (177 points de présence ont été répertoriés).

C'est l'un des rares manchots (avec le manchot empereur) dont la présence se limite à l'Antarctique et les eaux environnantes (limitée par le 'pack ice'). Du fait de sa petite taille et pour éviter une trop forte déperdition de sa chaleur, ses plumes sont plus longues que celles des autres espèces et elles recouvrent une grande partie du bec.

En dehors de la période de reproduction, ils sont en mer. Ils se reposent sur les icebergs et les glaces flottants dérivant au grès des courants marins. Ils viennent à terre pour préparer le nid, donner naissance et élever ses poussins.

Ils ont une espérance de vie estimée entre 10 et 20 ans.



© www.manchots.com

Manchot Adélie – île du Roi-George (îles Shetland du Sud)

Classification

Nom scientifique: *Pygoscelis adeliae*

Genre: *Pygoscelis*

Sous-espèces: Aucune

Nom anglais: *Adelie Penguin*

Autre nom français: Manchot d'Adélie

Autres dénominations: *Pingüino de Adelia* (espagnol)

Fig. 18. Extrait de la page "manchot Adélie" sur le site "manchots.com"

Annexe 2

Informations complémentaires sur le manchot Chinstrap (à jugulaire)

Manchot à jugulaire

📁 Les espèces de manchots

Description

Le manchot à jugulaire est le deuxième manchot le plus nombreux (après le gorfou Macaroni) principalement présent sur les îles Sandwich du Sud (très peu visitées). Il est facilement reconnaissable à la fine bande de plumes noires autour du menton et de la gorge (d'où son nom). Globalement, il a le ventre blanc et le dos noir. Parmi les Pygoscelis, cette espèce est réputée pour son intrépidité, sa combativité et son agilité.



© www.manchots.com

Manchot à jugulaire –
Île de l'Eléphant (îles
Shetland du Sud)

Classification

Nom scientifique: *Pygoscelis antarctica / Pygoscelis antarcticus*

Genre: *Pygoscelis*

Sous-espèces: Aucune

Nom anglais: *Chinstrap Penguin*

Autre nom français: Manchot Antarctique, Manchot barbu

Autres dénominations: *Pinguino de barbijo* (espagnol)

Fig. 19. Extrait de la page "manchot Chinstrap" sur le site "manchots.com"

Annexe 3

Informations complémentaires sur le manchot Gentoo (Papou)

Description

Le manchot papou fait parti du genre *Pygoscelis* dont elle est l'espèce la plus craintive et la plus colorée avec son bec orange et ses pattes entre l'orange et le rose.

Sa première description remonte à 1776 par le naturaliste français Pierre Sonnerat. Il le décrivit dans son ouvrage *Voyage à la Nouvelle-Guinée*. Ses noms français et scientifiques ont pour origine la Papouasie Nouvelle-Guinée bien qu'il n'y ait aucun manchot dans ce pays.

Cette espèce est présente sur les îles sub-antarctiques et sur la Péninsule Antarctique.

Les manchots papous restent dans leur colonie tout au long de l'année. Les colonies restent rarement au même endroit d'une année à l'autre car à la fin de la saison de reproduction, l'activité des oiseaux a détruit la végétation, et par conséquent, les manchots doivent se déplacer l'année suivante.



Manchot papou – île de la Déception (îles Shetland du Sud)

Classification

Nom scientifique: *Pygoscelis papua*

Genre: *Pygoscelis*

Sous-espèces:

- *Pygoscelis papua papua* (la plus au nord)

Présente sur les îles sub-antarctiques.

Elle représente environ 75% des manchots papous.

- *Pygoscelis papua ellsworthii* (la plus au sud)

Plus petite que l'autre sous-espèce et avec des plumes plus courtes

Présente sur la Péninsule Antarctique et les îles environnantes.

Elle représente environ 25% des manchots papous.

Nom anglais: *Gentoo Penguin*

Le mot *gentoo* fait référence aux habitants d'Inde qui portent des bonnets blancs leur faisant ressembler aux bandes blanches de la tête du manchot papou.

Fig. 20. Extrait de la page "manchot Gentoo" sur le site "manchots.com"

Table des figures

1	Extrait des données	3
2	Matrice des nuages de dispersion	4
3	Histogrammes et densités lissées	5
4	Nuages de dispersions et régression lissée	5
5	Nuages de dispersions avec régression lissée et droites des moindres carrés	6
6	Nuage de dispersion "poids" versus "longueur des nageoires"	7
7	Matrice des nuages de dispersion - Mâle	8
8	Matrice des nuages de dispersion - Femelle	9
9	Nuages de dispersion avec régression lissée et droites des moindres carrés - Sexe	9
10	Nuages de dispersion : poids versus longueur des nageoires - Mâle et Femelle	10
11	Matrice des nuages de dispersion - espèce Adélie	11
12	Matrice des nuages de dispersion - espèce Chinstrap	11
13	Matrice des nuages de dispersion - espèce Gentoo	12
14	Nuages de dispersion avec régression lissée et droites des moindres carrés - Espèces	14
15	Nuage de dispersion : poids versus profondeur de la crête supérieur du bec : Adelie	14
16	Nuage de dispersion : poids versus longueur des nageoires : Chinstrap	15
17	Nuage de dispersion : poids versus profondeur de la crête supérieur du bec : Gentoo	15
18	Extrait de la page "manchot Adélie" sur le site "manchots.com"	17
19	Extrait de la page "manchot Chinstrap" sur le site "manchots.com"	18
20	Extrait de la page "manchot Gentoo" sur le site "manchots.com"	19