Exercices supplémentaires – Régression linéaire

Correction

Exercice 1

Le code utilisé pour obtenir les résultats présentés ici est fourni dans un fichier à part.

1. Le diagramme de dispersion du poids en fonction de la taille est repris à la Figure 1. Il est clair que le poids d'une observation correspondant à une taille de 180cm a été mal encodé. Il s'agit de la $20^{\text{ème}}$ observation.

2. La droite de régression estimée par la technique des moindres carrés est ajoutée au diagramme de dispersion à la Figure 2. Cette droite est clairement attirée par l'observation erronée, et ne traduit pas la tendance linéaire globale présente parmi le reste des observations. Afin de quantifier la qualité de l'ajustement, on peut calculer le coefficient de détermination, qui vaut dans ce cas-ci 0.02. Autrement dit, la régression n'explique que 2% de la variabilité de la variable `Poids`, ce qui est très peu.

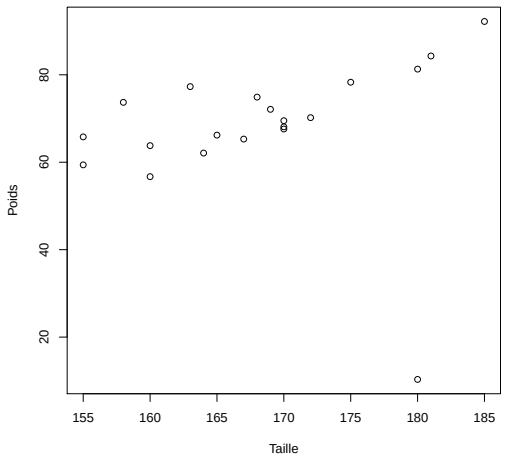


Figure 1: Diagramme de dispersion du poids en fonction de la taille

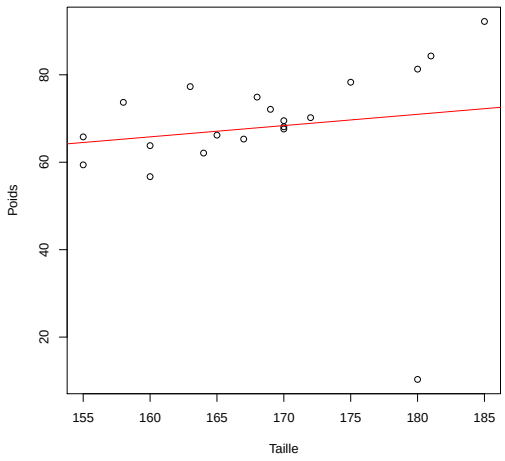
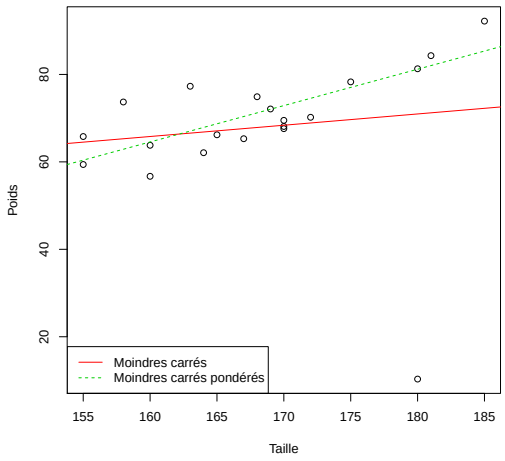


Figure 2: Droite de régression des moindres carrés

3. La droite de régression des moindres carrés estimée sur l'ensemble des observations privé de l'observation erronée est ajoutée au diagramme de dispersion à la Figure 3. Puisque cette dernière n'intervient pas dans l'estimation des paramètres de la droite, elle colle beaucoup mieux à la tendance linéaire globale.



(a)

Figure 3: Droites de régression des moindres carrés et des moindres carrés pondérés

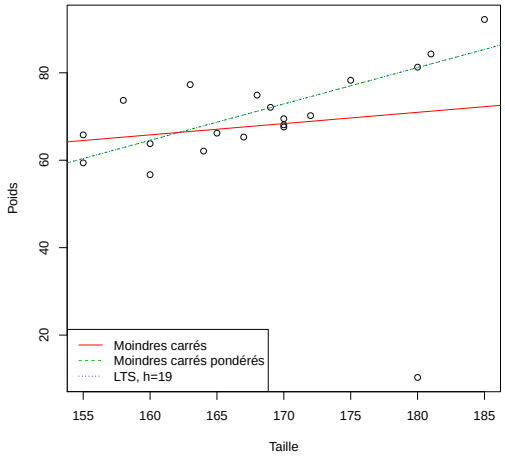


Figure 4: Droites de régression des moindres carrés, des moindres carrés pondérés et LTS

(b) La droite de régression LTS est ajoutée au diagramme de dispersion à la Figure 4. On constate dans ce cas-ci qu'elle est exactement la même que la droite obtenue par la technique des moindres carrés pondérés (mais attention, ce n'est pas nécessairement le cas !) et qu'elle colle donc beaucoup mieux à la tendance linéaire générale.

Exercice 2

1. Le diagramme de dispersion de l'intensité de la lumière en fonction de la température ainsi que la droite des moindres carrés sont représentés à la Figure 5. Quatre étoiles ressortent très clairement du nuage de points principal et se détachant de la tendance linéaire observée dans le reste des observations. Elles ont un impact très important sur l'estimation de la droite de régression par la technique des moindres carrés. En effet, ces quatre étoiles à elles seules renversent la tendance linéaire attendue puisque la droite estimée a une pente négative alors que la tendance dans le nuage principal est croissante.

2. La boîte à moustaches de la variable `LogTemp` est représentée à la Figure 6. On y aperçoit 3 valeurs extérieures, qui sont en fait associées à 5 observations (plusieurs observations ont la même valeur). Il s'agit des observations d'indices $7, 11, 20, 30$ et 34. En attribuant un poids nul à chacune de ces observations, on obtient la droite de régression représentée à la Figure 7. Les 4 étoiles situées dans le coin supérieur gauche ainsi que l'étoile ayant une valeur de `LogTemp` égale à 3.84, qui se détache également du nuage de points principal, ne jouent plus aucun rôle dans l'estimation de la droite et celle-ci caractérise donc beaucoup mieux la tendance linéaire générale.

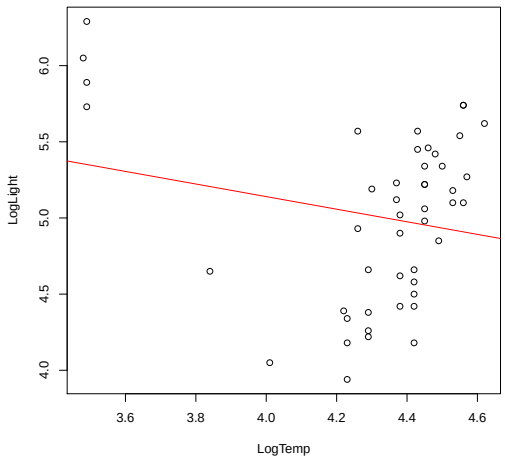


Figure 5: Diagramme de dispersion et droite de régression par moindres carrés

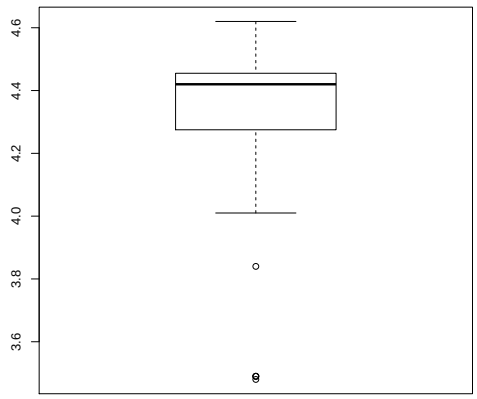


Figure 6: Boîte à moustaches de la variable `LogTemp`

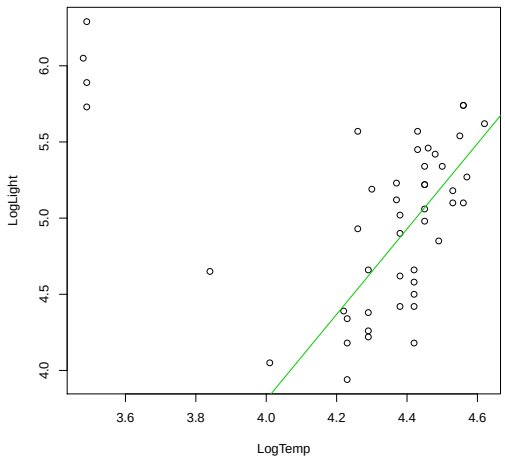


Figure 7: Droite de régression par moindres carrés pondérés

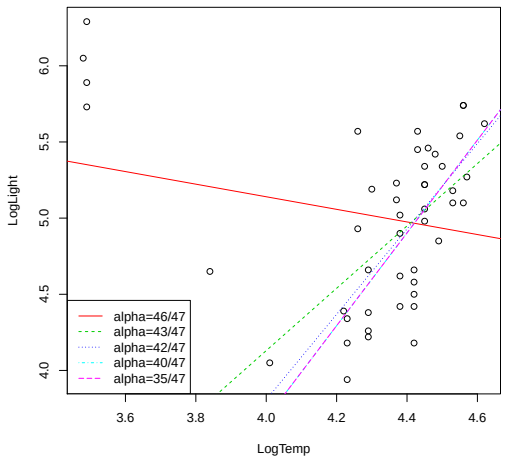


Figure 8: Droite de régression LTS pour différentes valeurs de `alpha`

3. Quatre valeurs différentes de `alpha` ont été choisies et les droites correspondantes sont toutes représentées à la Figure 8. Parmi les valeurs choisies, au plus `alpha` diminue, au plus la pente de la droite de régression augmente et la droite finit par se stabiliser (les deux droites correspondant aux valeurs `alpha`= 40/47 et `alpha`= 35/47 coïncident).