Exercice 1 : la régression simple

La régression linéaire est une méthode d'analyse qui permet, à partir de données expérimentales x et y mesurées sur un échantillon de taille n, de répondre (entre autres) à la question suivante :

> *Quelle est la force du lien linéaire entre y et x ?*

Pour répondre à cette question, on estime un modèle linéaire.

Exemple sur données réelles

Les données prises comme exemple ont été recueillies sur 62 espèces de mammifères et représentent les poids moyens de leur corps et de leur cerveau (Weisberg 1980).

Les données sont représentées en Figure 1. On se pose la question de savoir s'il existe un lien linéaire entre le poids du corps (x) et le poids du cerveau y.

Modèle sans "intercept"

Dans un premier temps, on fait l'hypothèse que si le poids du corps est nul, alors le poids du cerveau également sera nul.

Le modèle linéaire peut alors se formuler de la manière suivante : le poids du cerveau est égal au poids du corps multiplié par une constante (β).

Nous allons calculer β par la méthode des moindres carrés : nous allons minimiser l'écart quadratique entre les mesures y_i et leur prédiction $\hat{y}_i = \beta x_i$.

Le coefficient du modèle linéaire est donc estimé est résolvant le problème de minimisation suivant.

$$\min_{\beta \in \mathbb{R}} f(\beta) = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

1. Montrez que f est une fonction convexe en β.
2. Calculez analytiquement la valeur de β qui minimise f.
3. A partir des données, estimez cette valeur de β.
4. Faites une représentation graphique des données et du modèle.

Modèle avec "intercept"

Dans un deuxième temps, nous allons nous intéresser aux mêmes données, mais après transformation logarithmique (cf Fig. 2). Cette représentation semble en effet beaucoup plus correspondre au modèle linéaire !

Après transformation logarithmique, il n'y a aucune raison de penser que le modèle a un *intercept* nul. Le problème précédent se transforme donc en

$$\min_{\beta_0, \beta} g(\beta) = \sum_{i=1}^{n} (y'_i - \beta x'_i - \beta_0)^2,$$

avec $y' = \log(y)$ et $x' = \log(x)$.

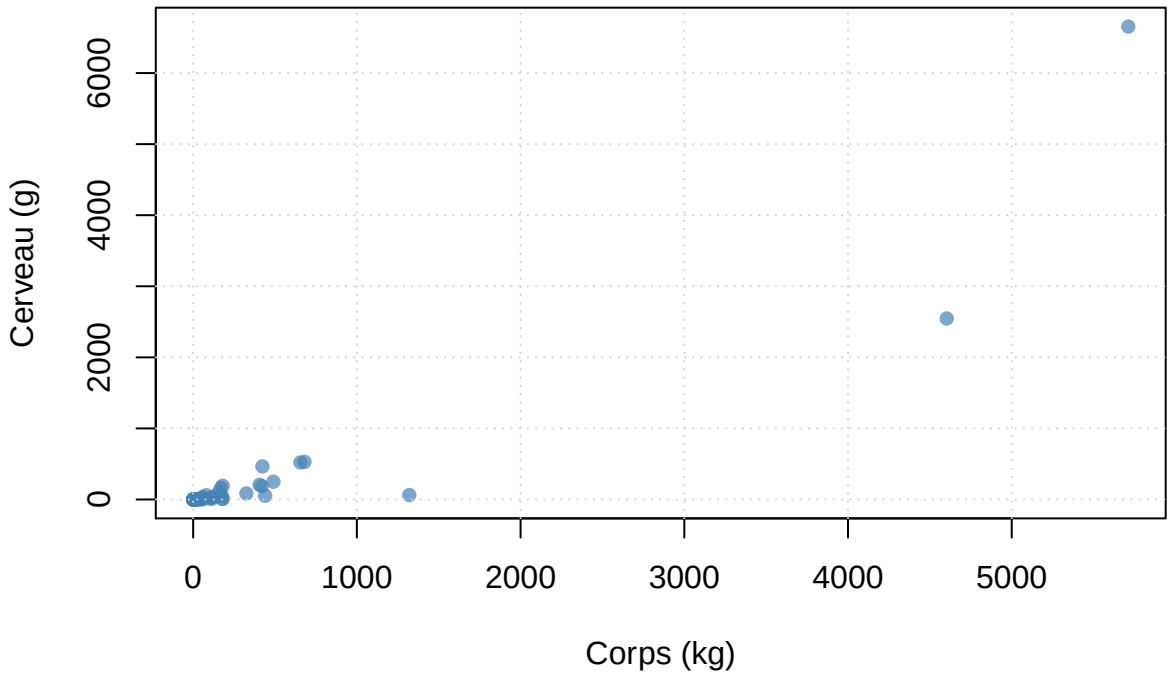

Figure 1: Poids du cerveau en fonction du poids du corps moyens chez 62 espèces de mammifères.

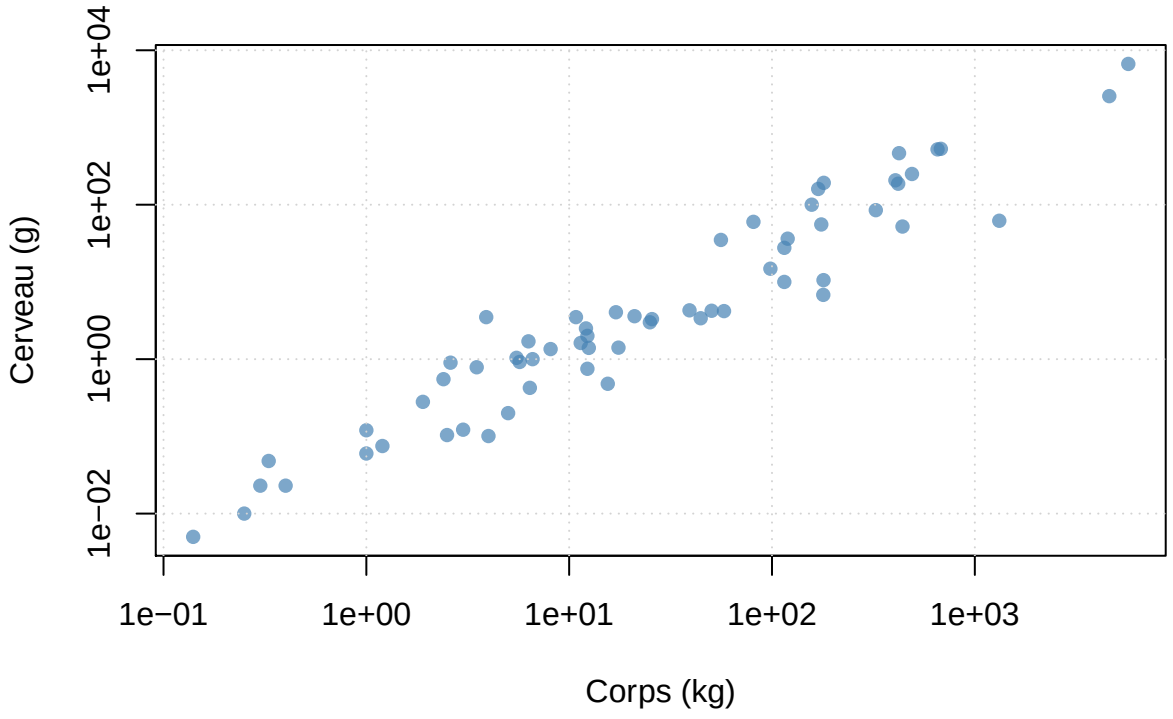

Figure 2: Poids du cerveau en fonction du poids du corps moyens chez 62 espèces de mammifères, en échelle logarithmique.

5. Calculez les paramètres optimaux du nouveau modèle.
6. Faites une représentation graphique du nouveau modèle en échelle naturelle.

Références

Weisberg, Sanford. 1980. *Mathematical Algorithms for Linear Regression. Wiley.* Nature Publishing Group.