Régression linéaire simple

6 octobre 2021

Ce laboratoire doit être remis le **13 octobre à 17h sur Moodle**. Dans votre réponse pour chaque question, veuillez inclure le code R utilisé (s'il y a lieu) et les résultats obtenus.

1. Croissance d'une espèce de pin

Le tableau de données `Loblolly` inclus avec R indique la hauteur en pieds (*height*) mesurée à six valeurs de l'âge (3 à 25 ans) pour 14 individus de l'espèce *Pinus taeda*. Chaque individu est indiqué par une valeur différente dans la colonne *Seed*.

```
head(Loblolly)
```

```
##    height age Seed
## 1    4.51   3  301
## 15  10.89   5  301
## 29  28.72  10  301
## 43  41.74  15  301
## 57  52.70  20  301
## 71  60.92  25  301
```

a) Estimez et interprétez les paramètres d'un modèle linéaire de la hauteur en fonction de l'âge pour ces pins. Est-ce que l'ordonnée à l'origine a un sens biologique pour ce modèle?

b) Ajoutez au jeu de données une colonne correspondant aux résidus du modèle en (a), obtenus avec la fonction `residuals`, puis produisez un nuage de points ce ces résidus en fonction de l'arbre (*Seed*). D'après ce résultat, quelle supposition de la régression linéaire pourrait être invalide ici?

c) À partir des graphiques de diagnostic du modèle en (a), indiquez si les suppositions de linéarité, d'égalité des variances et de normalité semblent être respectées.

d) Expliquez comment chacune des suppositions non-respectées identifiées en (b) et (c) affecte la validité des conclusions du modèle. Autrement dit, de quelle façon les estimés et prédictions du modèle pourraient différer de la réalité?

2. Diversité des plantes sur des îles britanniques

Le tableau de données britain_species.csv provient de l'étude de Johnson et Simberloff (1974), "Environmental determinants of island species numbers in the British Isles". Ces données indiquent le nombre d'espèces de plantes vasculaires (*species*) pour 42 îles britanniques en fonction de différents prédicteurs, incluant la surface de l'île en km^2 (*area*).

```
iles <- read.csv("britain_species.csv")
str(iles)
```

```
## 'data.frame':    42 obs. of  7 variables:
##  $ island      : chr  "Ailsa" "Anglesey" "Arran" "Barra" ...
##  $ area        : num  0.8 712.5 429.4 18.4 31.1 ...
##  $ elevation   : int  340 127 874 384 226 1343 210 103 143 393 ...
```

```
##  $ soil_types  : int  1 3 4 2 1 16 1 3 1 1 ...
##  $ latitude    : num  55.3 53.3 55.6 57 60.1 54.3 57.1 56.6 56.1 56.9 ...
##  $ dist_britain: num  14 0.2 5.2 77.4 201.6 ...
##  $ species     : int  75 855 577 409 177 1666 300 443 482 453 ...
```

a) Supposons qu'une théorie prédit que le nombre d'espèces (S) dépend de la surface d'une île (A) en fonction de l'équation suivante, où c et z sont des paramètres à déterminer:

$$S = cA^z$$

Comment pouvez-vous transformer cette équation en un modèle linéaire?

b) Ajustez le modèle en (a) aux données et vérifiez les graphiques de diagnostic. Quel est l'estimé de z?

c) Supposons que la théorie prédit que $z = 0.25$. Calculez un intervalle de confiance pour z et déterminez si cette hypothèse peut être rejetée ou non à un seuil de signification $\alpha = 0.05$.

d) À partir du modèle en (b), donnez un intervalle de prédiction à 90% du nombre d'espèces pour (i) une île de 0.5 km^2 et (ii) une île de 20 km^2.

Notes:

- Modifiez le % de l'intervalle de prédiction avec l'argument `level` de `predict`.
- Si la réponse du modèle est une transformation de la variable `species`, vous pouvez appliquer la transformation inverse aux bornes inférieure et supérieure des intervalles obtenus avec `predict` pour retrouver le nombre d'espèces correspondant.