

Sélection de modèles

28 octobre 2020

1. Biomasse racinaire en fonction de l'environnement

Le jeu de données `environment.csv` (tiré du manuel de Beckerman et Petchey, *Getting started with R: An introduction for biologists*) inclut des mesures de biomasse racinaire (*biomass*, en g/m²) pour 10 sites en fonction de l'altitude (en m), de la température (en degrés C) et de la précipitation annuelle (*rainfall*, en m).

```
enviro <- read.csv("environment.csv")
str(enviro)
```

```
## 'data.frame':    10 obs. of  5 variables:
## $ site          : int   1 2 3 4 5 6 7 8 9 10
## $ altitude      : int  13 160 100 205 45 84 349 509 399 30
## $ temperature   : int   24 18 17 15 20 21 14 11 13 19
## $ rainfall      : num   0.01 0.5 0.6 1.1 0.09 0.2 1.2 0.6 0.8 0.5
## $ biomass       : int   20 120 110 200 45 70 150 275 220 38
```

- Estimez les paramètres du modèle incluant les trois prédicteurs: `biomass ~ altitude + temperature + rainfall`. L'inclusion des trois prédicteurs dans le même modèle cause-t-elle des problèmes? Justifiez votre réponse.
- Proposez plusieurs modèles alternatifs pour ce jeu de données, incluant le modèle nul (0 prédicteur) et des modèles à 1 ou 2 prédicteurs (sans interactions). Évitez d'inclure des prédicteurs fortement corrélés dans le même modèle. Créez un tableau comparant ces modèles selon leur AICc.
- Quel est le meilleur modèle pour prédire la biomasse racinaire à un nouveau site semblable à ceux échantillonnés? Serait-ce utile de faire des prédictions moyennes à partir de plusieurs modèles ici? Justifiez votre réponse.

2. Prédictions de la migration d'espèces d'oiseaux

Le fichier `migration.csv` contient les données de Rubolini et al. (2005) sur 28 espèces d'oiseaux qui migrent entre l'Europe et l'Afrique.

```
migr <- read.csv("migration.csv")
str(migr)
```

```
## 'data.frame':    28 obs. of  14 variables:
## $ speciesID : int   1 3 4 5 7 8 9 11 12 13 ...
## $ species1  : chr   "Acrocephalus" "Acrocephalus" "Anthus" "Anthus" ...
## $ species2  : chr   "arundinaceus" "scirpaceus" "campestris" "trivialis" ...
## $ migDate   : num   33 38 32 27 35 30 31 30.8 30 28 ...
## $ latBreed  : num   46 48 43.5 55.3 47.5 50.3 51 51.5 48.8 59 ...
## $ latWntr   : num  -10.3 0 6 -10 -7.5 18.5 -15 7.5 -10 7.5 ...
## $ sexDchrmt : num   0 0 0 0 4.3 2 2.3 7 17.3 16 ...
## $ nestSite  : int   0 0 0 0 0 0 0 0 1 1 ...
## $ moult     : int   1 1 0 0 1 0 1 0 0 0 ...
## $ mWngLn    : num   96.8 66.8 91.6 88.7 192.1 ...
```

```
## $ fWngLn      : num  92.3 66 86.9 84.7 194.3 ...
## $ numSpecies: int   641 546 140 3531 269 104 166 101 737 12837 ...
## $ X           : num  -10.3 0 6 -10 -7.5 18.5 -15 7.5 -10 7.5 ...
## $ Y           : num   33 38 32 27 35 30 31 30.8 30 28 ...
```

Nous cherchons à prédire la date d'arrivée en Europe (*migDate*, mesurée en jours à partir du 1er avril) en fonction des prédictors suivants:

- Latitude du site de reproduction en Europe (*latBreed*)
- Latitude du site d'hivernage en Afrique (*latWntr*). *Note*: La latitude est positive au nord de l'équateur, négative au sud.
- Si l'espèce niche dans des cavités existantes (*nestSite*, 0=non, 1=oui)
- Si l'espèce mue au site d'hivernage (*moult*, 0=non, 1=oui)

En théorie, on s'attend à ce que les oiseaux arrivent plus tard si leur site de reproduction est plus nordique (dû au climat et à la distance à parcourir) et s'ils muent au site d'hivernage. On s'attend à ce que les oiseaux arrivent plus tôt si leur site d'hivernage est plus au nord en Afrique (moins de distance à parcourir) et s'ils nichent dans des cavités existantes.

- Vérifiez l'ajustement du modèle linéaire complet incluant les 4 prédictors. Interprétez les valeurs obtenues pour chacun des coefficients de ces prédictors (mais pas l'ordonnée à l'origine). Ces résultats sont-ils conformes à ceux attendus en théorie?
- Comparez avec l'AICc les modèles incluant les combinaisons suivantes des 4 prédictors:

- latBreed
- latWntr
- latBreed + latWntr
- latBreed + nestSite
- latWntr + nestSite
- latBreed + latWntr + nestSite
- latBreed + nestSite + moult
- latWntr + nestSite + moult
- latBreed + latWntr + nestSite + moult (modèle complet)

Combien de modèles ont un $\Delta AIC \leq 2$? Selon les poids d'Akaike, quelle est la probabilité que le meilleur modèle se trouve parmi ceux-ci?

- Chargez le jeu de données `migr_test.csv` qui contient les données de 10 autres espèces de l'étude de Rubolini et al.

```
migr_test <- read.csv("migr_test.csv")
str(migr_test)
```

```
## 'data.frame':   10 obs. of  14 variables:
## $ speciesID : int   2 6 10 14 18 22 26 30 34 38
## $ species1  : chr   "Acrocephalus" "Calandrella" "Delichon" "Hippolais" ...
## $ species2  : chr   "schoenobaenus" "brachydactyla" "urbica" "icterina" ...
## $ migDate   : num   35 27.5 29 39 31.2 28 35 27 22 22
## $ latBreed  : num   57.5 39.5 48.5 56 54.5 49 45.5 56.5 48 44
## $ latWntr   : num   -7.5 15.5 -15 -19 13 -7.5 -12 -9 11 16
## $ sexDchrmt : num    0 0 0 0 0 9 19.3 0 5.7 2.3
## $ nestSite  : int    0 0 0 0 0 0 0 0 0 1
## $ moult     : int    1 0 1 1 1 0 1 1 0 1
## $ mWngLn    : num   67.2 93.4 111.1 78.9 64.6 ...
## $ fWngLn    : num   64.7 89.8 110 78 63.6 ...
## $ numSpecies: int  2524 138 1624 10297 63 1163 1525 24767 2658 410
## $ X         : num   -7.5 15.5 -15 -19 13 -7.5 -12 -9 11 16
## $ Y         : num   35 27.5 29 39 31.2 28 35 27 22 22
```

Calculez la moyenne de l'erreur carrée de prédiction $(\textit{observation} - \textit{prédiction})^2$ pour ces 10 nouvelles observations selon (i) le meilleur modèle identifié en (b) et (ii) la prédiction moyenne pondérée de tous les modèles.

Indice: Pour obtenir un vecteur des prédictions moyennes, choisissez la composante `mod.avg.pred` de l'objet produit par la fonction `modavgPred`.