Model selection - Solutions

1. Root biomass as a function of the environment

The `environment.csv` dataset (from Beckerman and Petchey's textbook, *Getting started with R: An introduction for biologists*) includes measures of root biomass (in g/m^2) for 10 sites as a function of altitude (in m), temperature (in degrees C) and rainfall (in 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
```

a) Estimate the parameters of the model including the three predictors: `biomass ~ altitude + temperature + rainfall`. Does the inclusion of the three predictors in the same model cause problems? Justify your answer.

Solution

Altitude and temperature are strongly correlated (VIF of 11 for temperature and correlation of -0.92 between temperature and altitude), so it is preferrable to not include them in the same model.

```
mod_comp <- lm(biomass ~ altitude + temperature + rainfall, enviro)

library(car)
vif(mod_comp)
```

```
##    altitude temperature    rainfall
##    7.258410   11.178113    2.878935
```

```
cor(enviro$temperature, enviro$altitude)
```

```
## [1] -0.9174924
```

Note that `rainfall` is somewhat correlated with the other predictors, but the VIF is not too large for a model with `temperature + rainfall` or `altitude + rainfall`.

b) Propose several alternative models for this dataset, including the null model (0 predictor) and models with 1 or 2 predictors (without interactions). Avoid using highly correlated predictors in the same model. Create a table comparing these models according to their AICc.

Solution

Out of 8 possible models with 3 predictors, we exclude those that contain both temperature and altitude, leaving 6 models.

```
liste_mod <- list(
    nul = lm(biomass ~ 1, enviro),
    alt = lm(biomass ~ altitude, enviro),
```

```
    temp = lm(biomass ~ temperature, enviro),
    rain = lm(biomass ~ rainfall, enviro),
    altrain = lm(biomass ~ altitude + rainfall, enviro),
    temprain = lm(biomass ~ temperature + rainfall, enviro)
)

library(AICcmodavg)
aictab(liste_mod)
```

```
##
## Model selection based on AICc:
##
##          K   AICc Delta_AICc AICcWt Cum.Wt     LL
## temp     3 105.64       0.00   0.51   0.51 -47.82
## alt      3 106.05       0.41   0.42   0.93 -48.03
## altrain  4 110.97       5.32   0.04   0.96 -47.48
## temprain 4 111.03       5.39   0.03   1.00 -47.51
## rain     3 120.37      14.73   0.00   1.00 -55.19
## nul      2 122.05      16.41   0.00   1.00 -58.17
```

c) What is the best model for predicting root biomass at a new site similar to those sampled? Would it be useful to make average predictions from several models here? Justify your answer.

Solution

The model with only temperature has the best AICc, closely followed by the model with only altitude. Generally, when two models have almost the same AICc, it is useful to average their predictions. However, since temperature and altitude are strongly correlated, both models contain almost the same information here.

2. Predictions of the migration of bird species

The file `migration.csv` contains data from Rubolini et al. (2005) on 28 bird species that migrate between Europe and Africa.

```
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 ...
```

We are looking to predict the date of arrival in Europe (*migDate*, measured in days from April 1st) based on the following predictors:

- Latitude of the breeding site in Europe (*latBreed*)
- Latitude of the wintering site in Africa (*latWntr*). *Note*: Latitude is positive if north of the equator, negative if south.
- Whether the species nests in existing cavities (*nestSite*, 0 = no, 1 = yes)
- Whether the species moults at the wintering site (*moult*, 0 = no, 1 = yes)

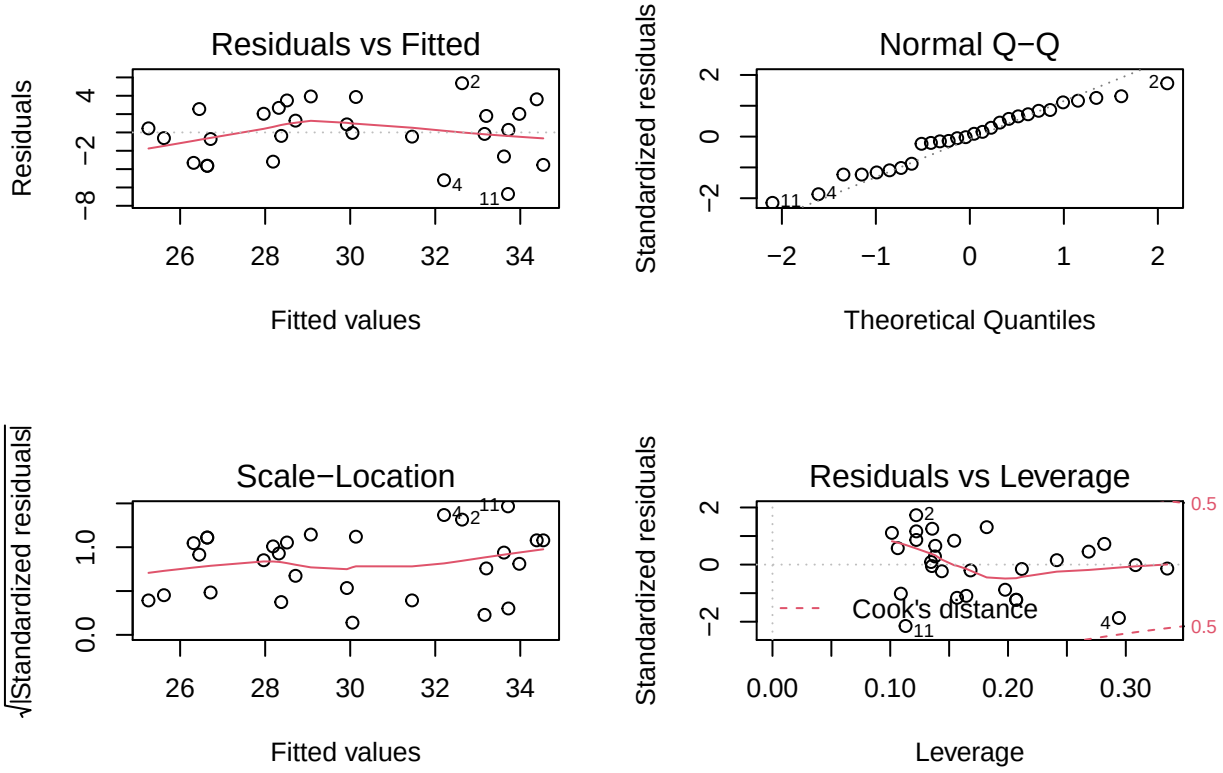
In theory, birds are expected to arrive later if their breeding site is further north (due to climate and distance) and if they moult at the wintering site. Birds are expected to arrive earlier if their wintering grounds are at a higher latitude in Africa (less distance to travel) and if they nest in existing cavities.

a) Check the fit of the complete linear model including the 4 predictors. Interpret the values obtained for each of the coefficients of these predictors (but not the intercept). Are these results consistent with those expected in theory?

Solution

```
mod_comp <- lm(migDate ~ latBreed + latWntr + nestSite + moult, migr)

par(mfrow = c(2, 2)) # Show 4 graphs in a 2x2 matrix
plot(mod_comp)
```



The diagnostics don't show any major problem.

```
summary(mod_comp)
```

```
##
## Call:
## lm(formula = migDate ~ latBreed + latWntr + nestSite + moult,
##     data = migr)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7117 -2.7590  0.1129  2.1641  5.3685
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.65761    6.04429   3.418  0.00236 **
## latBreed     0.19281    0.12473   1.546  0.13582
## latWntr     -0.08874    0.09007  -0.985  0.33476
## nestSite    -2.98943    1.62521  -1.839  0.07880 .
## moult        2.71921    1.74632   1.557  0.13310
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.315 on 23 degrees of freedom
## Multiple R-squared:  0.4967, Adjusted R-squared:  0.4092
## F-statistic: 5.676 on 4 and 23 DF,  p-value: 0.002497
```

Interpreting the coefficients:

- On average, migration occurs 0.19 days later for each degree of *latBreed* and 0.09 days earlier for each degree of *latWntr*.
- On average, migration occurs 3.0 days earlier for birds nesting in cavities and 2.7 days later for birds that moult at the wintering site.

The direction of these effects corresponds to the theory.

b) Using AICc, compare models including each of the following combinations of the 4 predictors:

- latBreed
- latWntr
- latBreed + latWntr
- latBreed + nestSite
- latWntr + nestSite
- latBreed + latWntr + nestSite
- latBreed + nestSite + moult
- latWntr + nestSite + moult
- latBreed + latWntr + nestSite + moult (complete model)

How many models have a $\Delta AIC \leq 2$? According to the Akaike weights, what is the probability that the best model is among those?

Solution

```
liste_mod <- list(
    breed = lm(migDate ~ latBreed, migr),
    wntr = lm(migDate ~ latWntr, migr),
    breed_wntr = lm(migDate ~ latBreed + latWntr, migr),
    breed_nest = lm(migDate ~ latBreed + nestSite, migr),
    wntr_nest = lm(migDate ~ latWntr + nestSite, migr),
    breed_wntr_nest = lm(migDate ~ latBreed + latWntr + nestSite, migr),
    breed_nest_moult = lm(migDate ~ latBreed + nestSite + moult, migr),
    wntr_nest_moult = lm(migDate ~ latWntr + nestSite + moult, migr),
    comp = lm(migDate ~ latBreed + latWntr + nestSite + moult, migr)
)

aictab(liste_mod)
```

```
##
## Model selection based on AICc:
##
##                   K   AICc Delta_AICc AICcWt Cum.Wt     LL
## breed_nest_moult  5 154.94       0.00   0.31   0.31 -71.11
## wntr_nest         4 156.07       1.12   0.18   0.49 -73.17
## wntr_nest_moult   5 156.55       1.61   0.14   0.63 -71.91
## breed_wntr_nest   5 156.59       1.65   0.14   0.77 -71.93
## comp              6 157.06       2.12   0.11   0.88 -70.53
## breed_nest        4 158.25       3.30   0.06   0.94 -74.25
## wntr              3 159.01       4.07   0.04   0.98 -76.01
## breed_wntr        4 160.67       5.73   0.02   1.00 -75.47
## breed             3 163.90       8.95   0.00   1.00 -78.45
```

Four models have a $\Delta AIC \leq 2$. These models have a combined weight (cumulative) of 77%.

c) Load the dataset `migr_test.csv` which contains the data of 10 other species from the 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
```

Calculate the mean of the square prediction error *(observation - prediction)*^{2} for these 10 new observations according to (i) the best model identified in (b) and (ii) the weighted average prediction of all models.

Tip: To obtain a vector of the average predictions, choose the `mod.avg.pred` component of the object produced by the `modavgPred` function.

Solution

```
pred_best <- predict(liste_mod$breed_nest_moult, newdata = migr_test)
pred_average <- modavgPred(liste_mod, newdata = migr_test)

err_best <- mean((migr_test$migDate - pred_best)^2)
err_average <- mean((migr_test$migDate - pred_average$mod.avg.pred)^2)

err_best
```

```
## [1] 22.8912
```

```
err_average
```

```
## [1] 17.76706
```

The mean square error is smaller for the model-averaged predictions.