R Demonstration – Model Selection

Objective: In this session we will perform model selection and model averaging using Akaike's Information Criteria in R.

Download the Indigo Snake habitat selection data from Breininger et al. (2011) as a text file (`snake_data.txt`) from the course website and save it in your `PCB6466` folder. You may also wish to download the script `model_average_snakes.R` (written by Eric D. Stolen).

Part I. Model Selection

Start the R software. From the menu bar, select *File→Change dir…* and then browse to the `PCB6466` folder on the Desktop. Enter the following commands to load and attach the Indigo Snake habitat selection data from Breininger et al. (2011):

```
snake <- read.delim("snake_data.txt", row.names="ID")
```

Now we will use the ***`factor`*** function to specify variables *landc* and *SEX* as categorical:

```
snake$landc <- factor(snake$landc)
snake$SEX <- factor(snake$SEX)
```

We will create two new variables. First, we will transform values of the dependent variable *ha* to their natural logarithm and name the new variable *ha.ln*:

```
snake$ha.ln <- log(snake$ha)
```

We will change the original categories in variable *landc* (1, 2, 3) from three to two (1, 2) using the ***`ifelse`*** function to convert "3" to "2", then we will specify this variable as categorical:

```
snake$lc2   <- ifelse(snake$landc==3,2,1)
snake$lc2   <- factor(snake$lc2)
```

Next, we will use the ***`summary`*** function to produce summary statistics for our new data matrix and explore the first five cases of the data. The output for this instruction is not shown in this demo, but you should explore it on your own:

```
summary(snake)
snake[1:5,]
```

We use the function ***`library`*** (AICcmodavg) to call the program ***`AICcmodavg,`*** which helps us evaluate the regression models and their relative information. It is necessary that you have installed this package in order to run the rest of this demo.

```
library(AICcmodavg)
```

We will start by setting-up the candidate model set. This is one the most critical parts of the procedure. For this exercise we will include eight models with three independent variables: a null model, three single independent variable models, all three possible two-way additive models and the three-way additive model. We concatenate the models in the vector `ms`, and define the variable `Cand.models` as an array. We use the function `for` to create a loop to sequentially evaluate each model with the function `glm` assuming Gaussian errors:

```
ms  <- c(
            "ha.ln ~ 1 ",
              "ha.ln ~ SEX ",
            "ha.ln ~ lc2",
            "ha.ln ~ weeks ",
            "ha.ln ~ SEX + lc2",
            "ha.ln ~ SEX + weeks",
            "ha.ln ~ lc2 + weeks",
            "ha.ln ~ SEX + lc2 + weeks"  #no comma at end!
             )
Cand.models<-list()
for (i in 1:length(ms)) {
Cand.models[[i]] <- glm(as.formula(ms[i]), family=gaussian, data=snake)
}
```

We will create a vector of names to trace back models in the set. The function `paste` concatenates the names of the models in the vector `Modnames` after converting the model numbers to characters. It uses `sep=" "` to create space after the name.

```
Modnames<-paste("model", 1:length(Cand.models), sep="")
```

Now, we will generate and display an AIC summary table to present the main results. We use the function *`aictab`* to generate the table. We sort the models based on their relative information. We round the results to 4 digits after the decimal place and provide the log-likelihood. Notice that the estimate of variance is included as a parameter for each model.

```
print(aictab(cand.set = Cand.models, modnames = Modnames, sort =
TRUE),digits = 4, LL = TRUE)
```

```
Model selection based on AICc :

       K     AICc Delta_AICc AICcWt Cum.Wt       LL
model5 4 112.9789     0.0000 0.7089 0.7089 -51.9894
model8 5 114.8876     1.9087 0.2730 0.9819 -51.6746
model2 3 121.5228     8.5439 0.0099 0.9918 -57.4687
model6 4 122.2718     9.2929 0.0068 0.9986 -56.6359
model3 3 125.9252    12.9463 0.0011 0.9997 -59.6699
model7 4 128.3374    15.3585 0.0003 1.0000 -59.6687
model1 2 139.2762    26.2973 0.0000 1.0000 -67.4953
model4 3 141.2594    28.2805 0.0000 1.0000 -67.3370
```

We can also set up a data table for predicting the outcome of four different scenarios. In the array we define the value of each variable for each scenario in order, so for example if our first case is one where sex="1", lc2="1" and week="52"; you will find these numbers in positions 1, 5 and 9:

```
fitdata<-array(c(1,2,1,2,1,1,2,2,52,52,52,52),dim=c(4,3))
fitdata<-data.frame(fitdata)
names(fitdata)[1] <- "SEX"
names(fitdata)[2] <- "lc2"
names(fitdata)[3] <- "weeks"
fitdata$SEX<-factor(fitdata$SEX)
fitdata$lc2<-factor(fitdata$lc2)
fitdata
```

```
> fitdata
  SEX lc2 weeks
1   1   1    52
2   2   1    52
3   1   2    52
4   2   2    52
```

With the following command we can get model-averaged predictions for the *ln* of home range for each candidate scenario:

```
modavgpred(cand.set= Cand.models, modnames= Modnames, newdata=fitdata,
type = "link", c.hat = 1, gamdisp = NULL, second.ord = TRUE, nobs =
NULL, uncond.se = "revised")
```

```
Model-averaged predictions on the link scale based on entire model set:

  mod.avg.pred uncond.se
1         4.78      0.21
2         3.74      0.26
3         3.85      0.32
4         2.80      0.27
```

And we can also get model-averaged estimates for the slopes of the regressions for each variable:

```
modavg(cand.set= Cand.models, parm = "lc22", modnames= Modnames)
```

```
Multimodel inference on " lc22 " based on AICc

 AICc table used to obtain model-averaged estimate:

       K   AICc Delta_AICc AICcWt Estimate   SE
model3 3 125.93      12.95   0.00    -1.33 0.31
model5 4 112.98       0.00   0.72    -0.96 0.28
model7 4 128.34      15.36   0.00    -1.33 0.32
model8 5 114.89       1.91   0.28    -0.92 0.29

Model-averaged estimate: -0.95
Unconditional SE: 0.29
```

```
 95 % Unconditional confidence interval: -1.51 , -0.39

modavg(cand.set= Cand.models, parm = "SEX2", modnames= Modnames)

Multimodel inference on " SEX2 " based on AICc

 AICc table used to obtain model-averaged estimate:

       K   AICc Delta_AICc AICcWt Estimate   SE
model2 3 121.52       8.54   0.01    -1.30 0.26
model5 4 112.98       0.00   0.71    -1.03 0.25
model6 4 122.27       9.29   0.01    -1.34 0.26
model8 5 114.89       1.91   0.27    -1.07 0.25

Model-averaged estimate: -1.05
Unconditional SE: 0.25
 95 % Unconditional confidence interval: -1.55 , -0.55

modavg(cand.set= Cand.models, parm = "weeks", modnames= Modnames)

Multimodel inference on " weeks " based on AICc

 AICc table used to obtain model-averaged estimate:

       K   AICc Delta_AICc AICcWt Estimate   SE
model4 3 141.26      26.37   0.00     0.00 0.01
model6 4 122.27       7.38   0.02     0.01 0.01
model7 4 128.34      13.45   0.00     0.00 0.01
model8 5 114.89       0.00   0.97     0.00 0.00

Model-averaged estimate: 0
Unconditional SE: 0
 95 % Unconditional confidence interval: -0.01 , 0.01
```