> *“Data dredging (also called data snooping, data mining, post hoc data analysis) should generally be avoided, except in (1) the early stages of exploratory work or (2) after a more confirmatory analysis has been done. In this latter case, the investigator should fully admit to the process that led to the post hoc results and should treat them much more cautiously than those found under the initial, a priori, approach.”*

Burnham and Anderson, 2002, Model Selection and Multimodel Inference: A Practical Information-theoretic Approach

Model selection in R, part 2

Information Criteria and statistical inference

Timothée Bonnet

June 13, 2019

BDSI / RSB

If you get bored:

Scroll down the slides and find ”Challenge exercises”.

Model selection and automation

Model selection and causal inference

Reminder: Why model selection

- Adding predictors increases fit to the response, in the current data
- But too many predictors:
  - DECREASE fit in new data (from the same population)
  - Hinder biological interpretation
  - Increases esimtation uncertainty (larger SE and p-values)
- Model selection aims to balance fit and generalisation

Information criteria

Akaike information criterion (AIC)

- AIC $=2\times$Number of parameters - $2\times$ log(model likelihood)

Information criteria

Akaike information criterion (AIC)

- AIC $=2\times$Number of parameters - $2\times$ log(model likelihood)
- Smaller is better

Information criteria

Akaike information criterion (AIC)

- AIC $= 2\times$Number of parameters - $2\times$ log(model likelihood)
- Smaller is better
- Only relative measure, no absolute meaning

Information criteria

Akaike information criterion (AIC)

- AIC $=$ 2×Number of parameters - 2× log(model likelihood)
- Smaller is better
- Only relative measure, no absolute meaning
- In R: `AIC(model)`

Practice: reminder AIC

Load `VoleWeight.csv`.

We want to understand what factors explain variation in individual body weight. Compare a few (plausible) models with AIC.

Different information criteria?

2 most important:

1. AICc (`MuMIn::AICc()`)
    - Small sample size correction for AIC
    - Can always been used instead of AIC
    - **Maximizes prediction** (of new data)

Different information criteria?

2 most important:

1. AICc (`MuMIn::AICc()`)
    - Small sample size correction for AIC
    - Can always been used instead of AIC
    - **Maximizes prediction** (of new data)
2. Bayesian Information Criterion (`stat::BIC()`)
    - More penalty per parameter
    - Simpler models than AIC / AICc
    - **Maximizes consistency** (=effects you find in model selection data likely to be present in new data)

Package `MuMIn`

```r
install.packages("MuMIn") library(MuMIn)
```

Package `MuMIn`

```r
install.packages("MuMIn") library(MuMIn)
```

1. AICc

- AIC is biased for small sample size
- AICc ("second-order AIC") when sample size / number of parameters is less than 40
- `MuMIn::AICc()`

Package `MuMIn`

```r
install.packages("MuMIn") library(MuMIn)
```

1. AICc

- AIC is biased for small sample size
- AICc ("second-order AIC") when sample size / number of parameters is less than 40
- `MuMIn::AICc()`

2. dredge

- Automate model selection
- Many competing models, some may not make sense
- `MuMIn::dredge()`

Try automated model selection

Try to use dredge(), with selection based on AICc, to automate model selection on the vole data. Start from a model including all predictors (plus some interactions).

For some reason you first need to run:

```
options(na.action="na.fail")
```

Do you find the same result as on slide 6?

dredge() best practices

```
dredge(global.model= ,fixed = ,varying= ,subset= ,rank=)
```

- `global.model` makes sense; not too complicated

dredge() best practices

```
dredge(global.model= ,fixed = ,varying= ,subset= ,rank=)
```

- `global.model` makes sense; not too complicated
- `fixed` coefficients across models

dredge() best practices

```
dredge(global.model= ,fixed = ,varying= ,subset= ,rank=)
```

- `global.model` makes sense; not too complicated
- `fixed` coefficients across models
- `subset` and `varying` for more complex subset of models to test

dredge() best practices

```
dredge(global.model= ,fixed = ,varying= ,subset= ,rank=)
```

- `global.model` makes sense; not too complicated
- `fixed` coefficients across models
- `subset` and `varying` for more complex subset of models to test
- `rank=AICc` or `BIC`; decide before you look!

Practice: constrain dredge

Still using the vole data, use dredge to select among models of the effects of humidity and temperature while controling for sex and age. Use the arguments `fixed` or `varying`.

Also check the difference between AICc and BIC based model selection.

Your starting model is:

```r
m0 <- lm(Weight ~ Sex*Age + humidity*temperature +
           as.factor(Year) , data = voles)
```

Model selection and automation

Model selection and causal inference

Principle of model selection based causal inference

1. List biological hypotheses

Principle of model selection based causal inference

1. List biological hypotheses
2. Define statistical models corresponding to hypotheses

Principle of model selection based causal inference

1. List biological hypotheses
2. Define statistical models corresponding to hypotheses
3. Importantly, the models do not have to be nested or look the same

Principle of model selection based causal inference

1. List biological hypotheses
2. Define statistical models corresponding to hypotheses
3. Importantly, the models do not have to be nested or look the same
4. (AIC/)BIC model selection

Principle of model selection based causal inference

1. List biological hypotheses
2. Define statistical models corresponding to hypotheses
3. Importantly, the models do not have to be nested or look the same
4. (AIC/)BIC model selection
5. Which models/hypotheses do better? Is one clearly best?

Principle of model selection based causal inference

1. List biological hypotheses
2. Define statistical models corresponding to hypotheses
3. Importantly, the models do not have to be nested or look the same
4. (AIC/)BIC model selection
5. Which models/hypotheses do better? Is one clearly best?
6. Do NOT trust parameter estimates/p-values (need confirmatory model)

Practice: Competing hypotheses

How does respiration rate scale with body mass in mammals? For a while researchers fought over different hypotheses: respiration could increase as a function of $mass^{2/3}$, as a function of $mass^{3/4}$ or as a function of $log(mass)$; while maybe the shape of the animals played a role. Let's find out!

Load `metabo.csv` and compare models through AIC or BIC selection.

NB: you can fit exponents of a predictors using the function `I()`. For instance, for the exponent 0.5 of x:

```
lm(y ~ I(x^(1/2)))
```

Do not choose statistical framework after applying them; do not trust model estimate after model selection

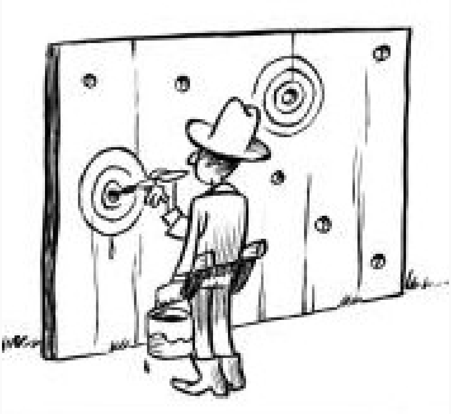

Figure 1: Null-hypothesis testing after model selection ©Dirk Jan-Hoek

Practice

Tell what drives the increase in number of babies in `babies.csv`. Compare AICc vs. BIC model selection.

Practice (challenge): P-values and model selection

When a predictor is independent of the response there is a 5% probability to find a p-value below 0.05 (that's a false positive). But it does not work if we do model selection first! Create a for-loop based on the code below to look at the distribution of p-values after model selection.

```r
nobs <- 60
mainpredictor <- rnorm(nobs)
control1 <- rnorm(nobs) ; control2 <- rnorm(nobs)
control3 <- rnorm(nobs) ; control4 <- rnorm(nobs)
control5 <- rnorm(nobs) ; response <- rnorm(nobs)
mfull <- lm(response ~ mainpredictor +
      control1*control2*control3 + control4*control5)
modall <- dredge(mfull, fixed = "mainpredictor")
summary(get.models(modall, 1)[[1]])$coefficients[2,4]#the pva
```

Summary: Model selection and inference

- AIC best for exploratory / predictive models
- BIC best for robust / consistent models
- AIC/BIC alone can be used for causal inference if models are all meaningful competing hypotheses
- After AIC/AICc selection p-values are wrong
- Parameter estimates, p-values and standard errors MUST be computed on new data (Confirmatory model)

Challenge!

What drives bird abundance (ABUND) in `loyn.csv`?

Challenge!

How do differences in AIC between nested models scale with F-statistic p-values for the extra predictor? Use some simulations to explore the issue.

Take-home messages

1. **Model selection biases estimates/p-values; confirm with new data**

Take-home messages

1. **Model selection biases estimates/p-values; confirm with new data**
2. **Model selection alone can be used for inference; needs careful choice of models**

Take-home messages

1. **Model selection biases estimates/p-values; confirm with new data**
2. **Model selection alone can be used for inference; needs careful choice of models**
3. **Choose method before analysis (AIC or BIC or confirmatory model?)**

Want to know more?

AIC vs. BIC vs. P-values:

- **“AIC does everything”:** Burnham and Anderson, 2002, Model Selection and Multimodel Inference: A Practical Information-theoretic Approach
- **“Sometimes BIC works better”:** Brewer & al. The relative performance of AIC, AICc and BIC in the presence of unobserved heterogeneity. Methods in Ecology and Evolution. 2016, 7, 679–692.
- **“Your goal matters in the choice between AIC, BIC, p-values...:”** Aho & al. A graphical framework for model selection criteria and significance tests: Refutation, confirmation and ecology. Methods in Ecology and Evolution. 2017;8:47–56.

Before you leave:

1. **Write one thing you liked and one you disliked on a sticky note**
2. **Presence sheet! (HDR career development framework)**
3. **Email address to join the Slack channel**
4. **Past workshops at** `https://timotheenivalis.github.io/RSB-R-Stats-Biology/`