Bayesian statistics with R
7. Contrast scientific hypotheses with model selection

Olivier Gimenez

March 2021

Model selection

How to select a best model?

- Is there any effect of rain or temperature or both on breeding success?

How to select a best model?

- Is there any effect of rain or temperature or both on breeding success?
- The proportion of explained variance R^2 is problematic, because the more variables you have, the bigger R^2 is.

How to select a best model?

- Is there any effect of rain or temperature or both on breeding success?
- The proportion of explained variance R^2 is problematic, because the more variables you have, the bigger R^2 is.
- Idea: **penalize models with too many parameters**.

Akaike information criterion (AIC)

$$AIC = -2\log(L(\hat{\theta}_1, \ldots, \hat{\theta}_K)) + 2K$$

with L the likelihood and K the number of parameters θ_i.

Akaike information criterion (AIC)

$$\text{AIC} = -2\log(L(\hat{\theta}_1, \ldots, \hat{\theta}_K)) + 2K$$

A measure of goodness-of-fit of the model to the data: the more parameters you have, the smaller the deviance is (or the bigger the likelihood is).

Akaike information criterion (AIC)

$$\mathrm{AIC} = -2\log(L(\hat{\theta}_1, \ldots, \hat{\theta}_K)) + 2K$$

A penalty: twice the number of parameters K

Akaike information criterion (AIC)

- AIC makes the balance between *quality of fit* and *complexity* of a model.

Akaike information criterion (AIC)

- AIC makes the balance between *quality of fit* and *complexity* of a model.
- Best model is the one with lowest AIC value.

Akaike information criterion (AIC)

- AIC makes the balance between *quality of fit* and *complexity* of a model.
- Best model is the one with lowest AIC value.
- Two models are difficult to distinguish if $\Delta\text{AIC} < 2$.

Bayesian version

- Watanabe-Akaike Information Criteria or WAIC:

$$\mathrm{WAIC} = -2\sum_{i=1}^{n} \log E[p(y_i \mid \theta)] + 2p_{\mathrm{WAIC}}$$

- where $E[p(y_i \mid \theta)]$ is the posterior mean of the likelihood of the ith observation and
- p_{WAIC} is the effective number of parameters computed using the posterior variance of the likelihood.
- Relatively new and not yet available in Jags in routine.

WAIC in Jags

```r
# calculate wAIC with JAGS
# https://sourceforge.net/p/mcmc-jags/discussion/610036/thread/8211df61/#e
samples <- jags.samples(storks$model,c("WAIC","deviance"), type = "mean",
                        n.iter = 2000,
                        n.burnin = 1000,
                        n.thin = 1)
```

WAIC in Jags

```r
samples$p_waic <- samples$WAIC
samples$waic <- samples$deviance + samples$p_waic
tmp <- sapply(samples, sum)
waic <- round(c(waic = tmp[["waic"]], p_waic = tmp[["p_waic"]]),1)
waic
#>   waic p_waic
#>  217.3   12.7
```

Your turn

Model selection with WAIC

- Fit models with rainfall effect, temperature effect and without any covariate.
- Rank them with WAIC.

Solution

Model with temperature only

```
# model specification
model <-
paste("
model
{
    for( i in 1 : N)
        {
        nbchicks[i] ~ dbin(p[i],nbpairs[i])
        logit(p[i]) <- a + b * cov[i]
        }

# priors for regression parameters
a ~ dnorm(0,0.001)
b ~ dnorm(0,0.001)
```

```r
# list of lists of initial values (one for each MCMC chain)
init1 <- list(a = -0.5, b = -0.5)
init2 <- list(a = 0.5, b = 0.5)
inits <- list(init1,init2)
# specify parameters that need to be estimated
parameters <- c("a","b")
# specify nb iterations for burn-in and final inference
nb.burnin <- 1000
nb.iterations <- 2000
# read in data
datax <- list(N = 23, nbchicks = nbchicks, nbpairs = nbpairs,
              cov = (temp - mean(temp))/sd(temp))
```

```r
# load R2jags to run Jags through R
storks_temp <- jags(data  = datax,
                    inits = inits,
                    parameters.to.save = parameters,
                    model.file = "code/logtemp.txt",
                    n.chains = 2,
                    n.iter = nb.iterations,
                    n.burnin = nb.burnin)
#> Compiling model graph
#>    Resolving undeclared variables
#>    Allocating nodes
#> Graph information:
#>    Observed stochastic nodes: 23
#>    Unobserved stochastic nodes: 2
#>    Total graph size: 125
#>
```

```
# compute WAIC
samples <- jags.samples(storks_temp$model,c("WAIC","deviance"), type = "mea
                        n.iter = 2000,
                        n.burnin = 1000,
                        n.thin = 1)
samples$p_waic <- samples$WAIC
samples$waic <- samples$deviance + samples$p_waic
tmp <- sapply(samples, sum)
waic_temp <- round(c(waic = tmp[["waic"]], p_waic = tmp[["p_waic"]]),1)
```

Model with rainfall only

```r
# read in data
datax <- list(N = 23, nbchicks = nbchicks, nbpairs = nbpairs,
              cov = (rain - mean(rain))/sd(rain))
```

```r
# load R2jags to run Jags through R
storks_temp <- jags(data  = datax,
                    inits = inits,
                    parameters.to.save = parameters,
                    model.file = "code/logtemp.txt",
                    n.chains = 2,
                    n.iter = nb.iterations,
                    n.burnin = nb.burnin)
#> Compiling model graph
#>    Resolving undeclared variables
#>    Allocating nodes
#> Graph information:
#>    Observed stochastic nodes: 23
#>    Unobserved stochastic nodes: 2
#>    Total graph size: 134
#>
```

```r
# compute WAIC
samples <- jags.samples(storks_temp$model,c("WAIC","deviance"), type = "mea
                        n.iter = 2000,
                        n.burnin = 1000,
                        n.thin = 1)
samples$p_waic <- samples$WAIC
samples$waic <- samples$deviance + samples$p_waic
tmp <- sapply(samples, sum)
waic_rain <- round(c(waic = tmp[["waic"]], p_waic = tmp[["p_waic"]]),1)
```

Model with no effect of covariates

```
# model specification
model <-
paste("
model
{
    for( i in 1 : N)
        {
        nbchicks[i] ~ dbin(p[i],nbpairs[i])
        logit(p[i]) <- a
        }

# priors for regression parameters
a ~ dnorm(0,0.001)
    }
```

```r
# list of lists of initial values (one for each MCMC chain)
init1 <- list(a = -0.5)
init2 <- list(a = 0.5)
inits <- list(init1,init2)
# specify parameters that need to be estimated
parameters <- c("a")
# specify nb iterations for burn-in and final inference
nb.burnin <- 1000
nb.iterations <- 2000
# read in data
datax <- list(N = 23, nbchicks = nbchicks, nbpairs = nbpairs)
```

```r
# load R2jags to run Jags through R
storks_temp <- jags(data  = datax,
               inits = inits,
               parameters.to.save = parameters,
               model.file = "code/lognull.txt",
               n.chains = 2,
               n.iter = nb.iterations,
               n.burnin = nb.burnin)
#> Compiling model graph
#>    Resolving undeclared variables
#>    Allocating nodes
#> Graph information:
#>    Observed stochastic nodes: 23
#>    Unobserved stochastic nodes: 1
#>    Total graph size: 51
#>
```

```
# compute WAIC
samples <- jags.samples(storks_temp$model,c("WAIC","deviance"), type = "mea
                        n.iter = 2000,
                        n.burnin = 1000,
                        n.thin = 1)
samples$p_waic <- samples$WAIC
samples$waic <- samples$deviance + samples$p_waic
tmp <- sapply(samples, sum)
waic_null <- round(c(waic = tmp[["waic"]], p_waic = tmp[["p_waic"]]),1)
```

Compare WAIC

```r
data.frame(model = c('both_covariates', 'temp', 'rain', 'none'),
           waic = c(waic[1],waic_temp[1],waic_rain[1],waic_null[1]),
           p_waic = c(waic[2],waic_temp[2],waic_rain[2],waic_null[2])) %>%
  arrange(waic)
#>             model  waic p_waic
#> 1            rain 212.9    9.2
#> 2            none 215.4    6.4
#> 3 both_covariates 217.3   12.7
#> 4            temp 219.9   10.0
```

Model with rainfall only seems to be better supported by the data. In case models have similar WAIC values, model-averaging might be useful.