HW5_ModelSelection&linearregression

N. Barrus

2024-02-21

Purpose:

The purpose of this markdown document is to work through Homework 5 in Dr. Babcock's Bayesian Statistics Course at the University of Miami. Homework 5 deals with model selection and linear regression.

General Start to Code

```r
rm(list = ls())

######github#####
#note, only needed after 90 days from 1/16/2024

#  usethis::create_github_token()
#  gitcreds::gitcreds_set()

#####check for r updates#####
#note, updateing may take some time so plan accordingly

#require(installr)

#check.for.updates.R()

#updateR() #only if needed

#######check for package updates#####
#note, updateing may take some time so plan accordingly

#old.packages()

# update.packages() #make the decision to the update the packages
```

Load packages

```r
library(tidyverse)
library(R2jags)
library(rstan)
library(ggmcmc)
library(purrr)
library(magrittr)
library(here)
theme_set(theme_bw(base_size=15))
```

Data

For model selection, the data consist of counts of the number of trees in 30 equal sized quadrats.

And for the linear regression, the data consist of DEET repellent levels and the number of mosquito bits suffered by volunteers.

```r
Y <- c(11, 3, 7, 6, 2, 36, 14, 9, 2, 10, 2, 7, 3, 1, 0, 0, 0, 1,
 5, 0, 2, 11, 5, 3, 0, 3, 3, 27, 0, 11)

tree.counts <- list(Y = Y,
                   N = length(Y))

tree.counts
```

```
## $Y
##  [1] 11  3  7  6  2 36 14  9  2 10  2  7  3  1  0  0  0  1  5  0  2 11  5  3  0  3  3 27  0 11
##
## $N
## [1] 30
```

```r
deet.data <- read_csv(here("data","deet.csv"))
```

```
## Rows: 52 Columns: 2
## -- Column specification ----------------------------------------------------------------------------
## Delimiter: ","
## dbl (2): dose, bites
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```r
knitr::kable(head(deet.data), caption = "Data Preview")
```

Table 1: Data Preview

dose	bites
1.5	4.06
1.4	4.18
2.0	3.54
2.3	3.54
2.3	3.24
2.5	3.24

Problem 1: Model Selection

In this problem we will rerun the negative binomial (either parameterization) and Poisson models from homework 4, now adding calculations of the DIC, WAIC and LOOIC for model selection

A) Calculate the DIC for both the Poisson and negative binomial in JAGS, and make a table with the DIC, pD, deltaDIC and DIC weights. Which model is preferred? *run the JAGS models*

```
treemod.HW4.Q1.poiss <- jags(data = tree.counts,
  parameters.to.save = c("lamda"),
  n.chains = 2,
  n.burnin = 1000,
  n.iter = 20000,
  model.file = here("JAGS_mods","HW3-Q1-Poisson.txt")
)

treemod.HW4.Q3.negbin <- jags(data = tree.counts,
  parameters.to.save = c("p","r","m","v"),
  n.chains = 2,
  n.burnin = 1000,
  n.iter = 20000,
  model.file = here("JAGS_mods","HW3-Q3-NegBinom.txt"))
```

create the model selection table

```
DIC.table <- tibble(model = c("Poisson", "Neg-Binomial"),
                    model.ls = c(list(treemod.HW4.Q1.poiss), list(treemod.HW4.Q3.negbin))) |>
  mutate(DIC = map_dbl(model.ls, c(2,24)),
         pD = map_dbl(model.ls, c(2,23)),
         deltaDIC = DIC - min(DIC),
         weight = round(exp(-2*deltaDIC)/sum(exp(-2*deltaDIC)),digits = 5))

DIC.table |> select(-model.ls) |> knitr::kable(caption = "DIC Table")
```

Table 2: DIC Table

model	DIC	pD	deltaDIC	weight
Poisson	324	1.08	148	0
Neg-Binomial	177	2.12	0	1

The Negative binomial model is preferred.

B) Do the same with the WAIC in STAN. Which model is preferred?

C) Do the same with the LOOIC in STAN. Which model is preferred?

D) Compare the three information criteria in part a, b and c. Do they all give the same results? Is this result consistent with what you learn by looking at the mean and variance that you estimated when you ran the negative binomial model?

Problem 2: Linear Regression

The data in the file called deet.csv is from a study of the effect of DEET insect repellent on the number of mosquito bites suffered by volunteers. The x variable is dose of DEET, and the y variable is the number of bites, square root transformed for normality. This example is from https://whitlockschluter3e.zoology.ubc.ca/chapter17.html

A) Use STAN to run a linear regression to predict bites from dose, using a normal prior for the intercept (a) and slope (b), and an exponential prior for the residual variance. Standardize the X variable by subtracting the mean and dividing by the standard deviation. Give the summary statistics for a, b and the residual variance, along with the probability that the slope is positive.

B) Plot the data and the model fit, with the credible interval of the line. In STAN, you can make a new generated quantity for the mean prediction, say ymean, and then, in R, extract its summary statistics with summary(stanobjectname,par="ymean")

C) Plot the residuals against the predicted values, and the qq normal plot of the residuals. Are the assumptions of the model met?

D) Plot the chi-squared discrepancy plot and calculate the Bayesian P value. Is the model adequate by this metric?

E) Say we want to use the model to predict a plausible range of values for the number of bites a particular volunteer might get with a DEET dose of exactly 2.8. Add this calculation to your STAN code and give the mean and 95% interval of this person's square root transformed number of bites.