Paths foward: Multiple and logistic regression

Stat 340: Bayesian Statistics

Example

- Barberan and Leff (2019) published data on dust samples taken from the ledges above doorways in the continental US.

- Bioinformatics processing detects the presence or absence of 763 species (technically operational taxonomic units) of fungi.

- The log of the number of fungi species present in the sample, which is a measure of species richness.

- Our objective: determine which factors influence a home's species richness.

Data

For each home, eight explanatory variables (i.e. covariates) are included in this example:

Variable	Description
`lnspecies`	natural log of the number of fungi species present in the sample
`long`	longitude
`lat`	latitude
`temp`	annual mean temperature of the location
`precip`	annual mean precipitation of the location
`NPP`	net primary productivity (rate at which all the autotrophs in an ecosystem produce net useful chemical energy)
`elev`	elevation
`house`	indicator that the house is a single-family home
`bedrooms`	number of bedrooms in the home

Multiple regression model

Sampling model: $Y_i | x_1, \ldots, x_p \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i,\ \sigma^2)$

Link function: $\mu_i = \beta_0 + \beta_1 x_{i1}, + \cdots + \beta_p x_{ip}$

Priors: We need to place a prior on each coefficient, β_j, and σ

JAGS implementation

```
mlr_string <- "model{
## Sampling model
for(i in 1:n) {
  y[i] ~ dnorm(mu[i], invsigma2)
  mu[i] <- beta0 + beta1 * temp[i] + beta2 * precip[i]
}

## Weakly informative priors
beta0 ~ dnorm(0, 0.0025)
beta1 ~ dnorm(0, 0.0025)
beta2 ~ dnorm(0, 0.0025)
invsigma2 ~ dgamma(0.001, 0.001)
sigma <- pow(invsigma2, -1/2)
}"
```

JAGS implementation

Be sure that there are no NAs in the data set

```
mlr_data <- list(
  y = homes$lnspecies,
  temp = homes$temp,
  precip = homes$precip,
  n = nrow(homes)
)
```

JAGS implementation

```r
mlr_posterior <- run.jags(
  mlr_string,
  n.chains = 3,
  data = mlr_data,
  monitor = c("beta0", "beta1", "beta2", "sigma"),
  adapt = 1000,
  burnin = 5000,
  sample = 5000,
  thin = 30,
  silent.jags = TRUE
)
```

Posterior summary

```r
print(mlr_posterior, digits = 3)
```

```
##
## JAGS model summary statistics from 15000 samples (thin = 30; chains = 3; ad
##
##          Lower95  Median Upper95    Mean       SD    Mode    MCerr MC%ofSD SS
## beta0       6.06    6.15    6.25    6.15   0.0484    6.16 0.000441     0.9 12
## beta1    -0.0285 -0.0228 -0.0168 -0.0228  0.00297 -0.0227 2.49e-05     0.8 14
## beta2   0.000725 0.00162 0.00255 0.00162 0.000465 0.00161 3.95e-06     0.9 13
## sigma      0.353   0.368   0.385   0.369  0.00801   0.368 6.67e-05     0.8 14
##
##           AC.300 psrf
## beta0    0.00959    1
## beta1      0.013    1
## beta2   -0.00824    1
## sigma   -0.00702    1
##
## Total time taken: 42.2 seconds
```

Logistic regression

Arthritis clinical trial

A double-blind clinical trial investigated a new treatment for rheumatoid arthritis

We'll focus on a subset of the variables:

Variable	Description
`Better`	whether the drug improved symptoms 1 = yes, 0 = no
`Treatment`	Placebo or Treated
`Sex`	Male or Female
`Age`	Age in years

Logistic regression model

Sampling model: $Y_i | x_1, \ldots, x_p \overset{\text{ind}}{\sim} \text{Bern}(p_i)$

Link function: $\log\left(\dfrac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 x_{i1}, + \cdots + \beta_p x_{ip}$

Priors: We need to place a prior on each coefficient, β_j

Data preparation

Load and manipulate data

```r
arthritis <- read.csv("http://aloy.rbind.io/data/arthritis.csv")
```

JAGS needs numeric variables, so convert factors to indicators

```r
arthritis <- arthritis %>%
  mutate(
  Treatment = as.numeric(Treatment == "Treated"),
  Sex = as.numeric(Sex == "Male")
)

data_list <- list(
  Y = arthritis$Better,
  Treatment = arthritis$Treatment,
  Sex = arthritis$Sex,
  Age = arthritis$Age,
  n = nrow(arthritis)
)
```

Model fitting

```
# Logistic regression specification
model_string <- "model{
## Sampling model
for(i in 1:n){
  Y[i] ~ dbern(p[i])
  logit(p[i]) <- beta0 + beta_sex * Sex[i] + beta_age * Age[i] +
      beta_trt * Treatment[i]
}

## Priors
beta0 ~ dnorm(0,0.001)
beta_sex ~ dnorm(0,0.001)
beta_age ~ dnorm(0,0.001)
beta_trt ~ dnorm(0,0.001)

}"
```

Model fitting

```r
# Compile model
logistic_model <- run.jags(
  model_string,
  data = data_list,
  monitor = c("beta0", "beta_sex", "beta_age", "beta_trt"),
  n.chains = 3,
  sample = 5000,
  thin = 30,
  silent.jags  = TRUE)
```

Results

term	Mean	SD	2.5%	97.5%
beta0	-3.22	1.21	-5.75	-0.99
beta_sex	-1.57	0.62	-2.86	-0.43
beta_age	0.05	0.02	0.01	0.10
beta_trt	1.86	0.55	0.81	3.00

$$e^{\widehat{\beta}_{\text{sex}}} = 0.2078664$$

- For subjects in the same treatment group of the same age, the odds of improved symptoms are 0.21 times lower (i.e. about 79% lower) for males than females.

Results

term	Mean	SD	2.5%	97.5%
beta0	-3.22	1.21	-5.75	-0.99
beta_sex	-1.57	0.62	-2.86	-0.43
beta_age	0.05	0.02	0.01	0.10
beta_trt	1.86	0.55	0.81	3.00

$$e^{\widehat{\beta}_{\text{age}}} = 1.0533767$$

- For subjects in the same treatment group of the same sex, a one-year increase in age is associated with an increase in the odds of improved symptoms by a factor of 1.05 (i.e. about a 5% increase).

Results

term	Mean	SD	2.5%	97.5%
beta0	-3.22	1.21	-5.75	-0.99
beta_sex	-1.57	0.62	-2.86	-0.43
beta_age	0.05	0.02	0.01	0.10
beta_trt	1.86	0.55	0.81	3.00

$$e^{\widehat{\beta}_{\text{treat}}} = 6.4467075$$

- For subjects of the same sex and age, the odds of improved symptoms are 6.45 times higher (i.e. about 645% higher) for the treatment group than the placebo group.