

Bayesian Methods - JAGS - Heteroscedastic normal linear regression

Jan Vávra

Full assignment in PDF

Send your solutions to vavraj@karlin.mff.cuni.cz by 1. 12. 2025, 9:00.

Data description

We will use a small dataset (47 observations and 8 different covariates) containing some pieces of information about fires (**fire**) that occurred in Chicago in 1970 while distinguishing for some information provided in the data. Particularly, we will focus on its relation to the proportion of minority citizens (**minor**) and possibly the locality in Chicago (**side** - north vs. south). The dataset is loaded from the website and some brief insight is provided below.

```
chicago <- read.csv("http://www.karlin.mff.cuni.cz/~vavraj/nmst431/data/chicago.csv",
                    header=T, stringsAsFactors = TRUE)

chicago <- transform(chicago,
                      fside = factor(side, levels = 0:1, labels=c("North", "South")))
```

```
head(chicago)
```

```
##   minor fire theft  old insur income side fside
## 1  10.0  6.2   29 60.4   0.0 11.744   0 North
## 2  22.2  9.5   44 76.5   0.1  9.323   0 North
## 3  19.6 10.5   36 73.5   1.2  9.948   0 North
## 4  17.3  7.7   37 66.9   0.5 10.656   0 North
## 5  24.5  8.6   53 81.4   0.7  9.730   0 North
## 6  54.0 34.1   68 52.6   0.3  8.231   0 North
```

```
dim(chicago)
```

```
## [1] 47  8
```

```
summary(chicago[,c("fire", "minor", "fside")])
```

```
##           fire           minor           fside
## Min.      : 2.00   Min.      : 1.00   North:25
## 1st Qu.: 5.65   1st Qu.: 3.75   South:22
## Median :10.40   Median :24.50
## Mean     :12.28   Mean     :34.99
## 3rd Qu.:16.05   3rd Qu.:57.65
## Max.     :39.70   Max.     :99.70
```

Task 1 - Realizing the problem with variance (as frequentists)

Let us denote by

- Y_i the number of fires in area i ,
- X_i the minority percentage in area i .

Plot the estimated line from linear regression model $Y_i \sim X_i$. Comment on the problem of homoscedasticity assumption. Remind yourself modelling techniques that are able to reflect this issue.

Task 2 - Full-conditional distributions

For the following tasks consider the following heteroscedastic linear model:

$$Y_i|X_i, \beta, \tau, \lambda \sim \mathcal{N}(\beta_0 + \beta_1 \log X_i, \tau^{-1} \exp\{\lambda \log(X_i)\})$$

where $\beta \in \mathbb{R}^2$, $\tau > 0$ and $\lambda \in \mathbb{R}$ are primary model parameters. Assume the following independent prior distributions:

- $\beta_j \sim \mathcal{N}(0, 10^2)$ for $j = 0, 1$,
- $\tau \sim \Gamma(1, 0.005)$,
- $\lambda \sim \mathcal{N}(0, 10^2)$.

Write down all pdfs and derive the pdfs of full-conditional distributions. Do some of them belong to well-known distributional families?

Task 3 - Model implementation in JAGS

Prepare the data for the use in JAGS. Write down the model in JAGS (include its implementation in your report).

Task 4 - Convergence assessment

Run MCMC (two chains) for sufficient number of iterations. Assess the convergence to the stationary distribution. Produce trace and autocorrelation plots to support your decision. If any problems occur, try to solve them appropriately (increase `burnin` or `thin`). The final number of iterations used for inference should be 10 000.

Task 5 - Posterior summary

Summarize the posterior distribution of the model parameters including plots. Provide both ET and HPD credible intervals. Compare with the frequentistic approach from Task 1.

Task 6 - Relationship between number of fires and minority percentage

For several chosen values of variable `minor` estimate the posterior distribution of a parametric functions $E[Y|X = x]$ and $\text{var}[Y|X = x]$, the expected value and variance of the number of fires in an area of minority percentage x . Plot posterior estimates of location and capture also the variability of the posterior.

BONUS Tasks

- Assume different curves for different city sides (north and south).
- Implement the model for log-transformation of the response and compare the results on original scale.