STAT 578 - Advanced Bayesian Modeling - Fall 2019
Data Analysis Assignment

Xiaoming Ji

Introduction

Marijuana (formally cannabis) is a psychoactive drug from the Cannabis plant used for medical or recreational purposes. The immediate desired effects from consuming arijuana include relaxation and euphoria (the "high" or "stoned" feeling), a general change in perception, heightened mood, and an increase in appetite. Due to this characteristic, it is the most widely used illegal drugs in the world. In 2013, between 128 and 232 million people used Marijuana (2.7% to 4.9% of the global population between the ages of 15 and 65). The short-term side effects by Marijuana may include a decrease in short-term memory, impaired motor skills, red eyes, and feelings of paranoia or anxiety. Long-term side effects may include addiction, decreased mental ability in those who started regular use as teenagers, and behavioral problems in children whose mothers used marijuana during pregnancy. Due to this reason, the possession, use, and cultivation of cannabis is illegal in most countries of the world. [1]

Marijuana use hurts brain functioning in adolescents, this is added concern because adolescence is an important time in development brains by building the connections to improve executive functioning. In addition, research found adolescents who begin using marijuana before age 18 are 4 to 7 times more likely than adults to develop a marijuana use addiction. [2]

A survey conducted in 2016 on adolescents' marijuana usage for at least once in the past month reported the following percentage,

- 5% of students in 8th grade
- 14% of students in 10th grade
- 23% of students in 12th grade
- 22% of college students and young adults

From 2007 to 2017, marijuana use among adolescents has increased in the past 10 years for students in 12th grade. [3]

In this data analysis task, we will use Bayesian model to analyze the marijuana data.

Data

The data file `marijuanause.csv` contains survey data on reported marijuana use during adolescence in a sample of 236 people of approximately the same age. Each row represents a person in the sample, and the columns are as follows: - female: an indicator that the person is female (1 if so, 0 if not) (All people in the survey were either female or male) - use1976: an indicator that the person used marijuana in the year 1976 (1 if so, 0 if not) - use1977: an indicator that the person used marijuana in the year 1977 (1 if so, 0 if not) - use1978: an indicator that the person used marijuana in the year 1978 (1 if so, 0 if not) - use1979: an indicator that the person used marijuana in the year 1979 (1 if so, 0 if not) - use1980: an indicator that the person used marijuana in the year 1980 (1 if so, 0 if not)

The data has 116 records for male and 120 records for female. The frequency of use during this time frame is summarized as,

[1] Based on content from Wikipedia, https://en.wikipedia.org/wiki/Cannabis_(drug)

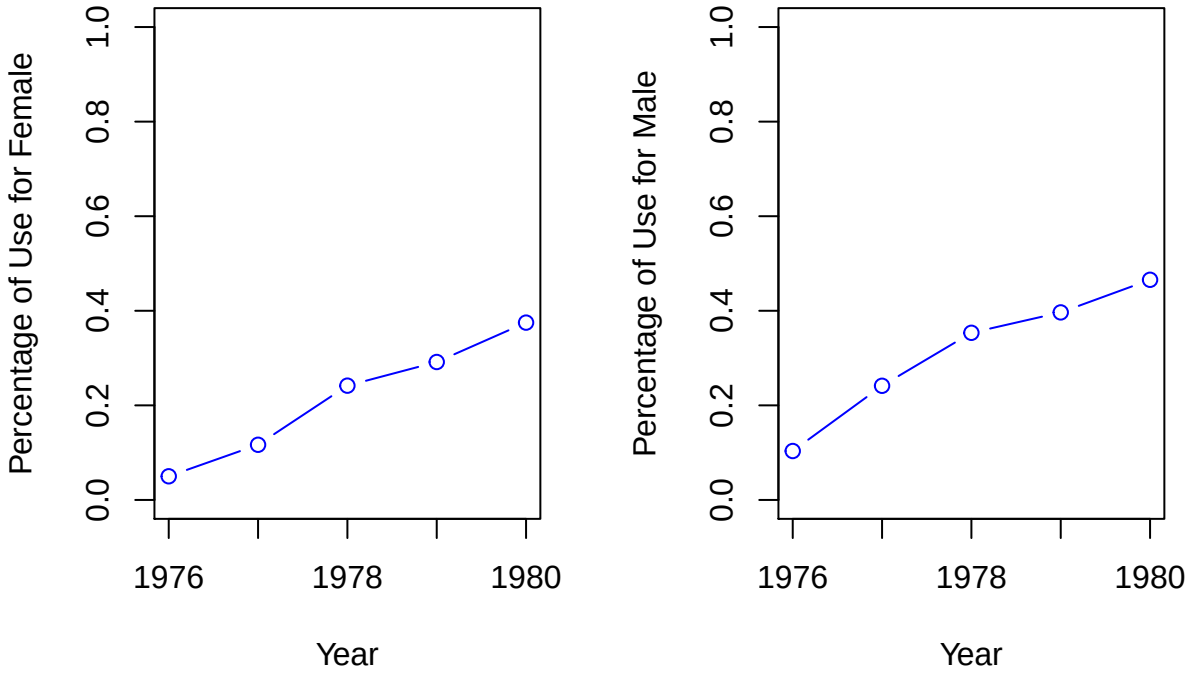
[2] Based on content from U.S. Department of Health & Human Services, https://www.hhs.gov/ash/oah/adolescent-development/substance-use/marijuana/risks/index.html

[3] Based on content from U.S. Department of Health & Human Services, https://www.hhs.gov/ash/oah/adolescent-development/substance-use/marijuana/index.html

	0	1	2	3	4	5
Use Percentage	0.47	0.17	0.11	0.12	0.08	0.04

We see only 47% of adolescents didn't use marijuana during this time frame.

The following two plots show percentage of use versus year: one plot for females and the other for males.



We see the obviously trend of use increase from 1976 to 1980. Male has more percentage of use than female for every year.

First Model

We fit a (univariate) Bayesian logistic regression model to explain marijuana use based on the year and on whether or not the person is female. The model can be described in the following expressions,

$$Use_{ij}|\beta, X_{ij} \sim indep.Bernoulli(p_{ij})$$

$$Logit(p_{ij}) = \beta_0 + \beta_{female} * female_{ij} + \beta_{year} * year_{ij}$$

$$\beta_0 \sim t_1(0, 10^2)$$

$$\beta_{female}, \beta_{year} \sim t_1(0, 2.5^2)$$

- Use_{ij} is the response variable by person (i) and year (j). 1 if marijuana was used, 0 if not.
- $female_{ij}$ is the female explanatory variables by person (i) and year (j). It will be centered to have sample mean of zero, and rescaled to have sample standard deviation of.
- $year_{ij}$ is the year explanatory variables by person (i) and year (j). It will be centered (but not rescaled).
- p_{ij} is the probability by person and by year.

- β_0 is intercept, β_{female} is coefficient for female, β_{year} is coefficient for year.

(a)

We will use JAGS to build and run the model.

```
model {
    for (i in 1:length(female_centered)) {
        for(j in 1:length(year_scaled)) {
            drug_use[i,j] ~ dbern(prob[i,j])
            logit(prob[i,j]) <- beta_0 + beta_female * female_centered[i] +
                                beta_year * year_scaled[j]
        }
    }

    beta_0 ~ dt(0, 0.01, 1)
    beta_female ~ dt(0, 0.16, 1)
    beta_year ~ dt(0, 0.16, 1)

    #Build posterior replication for female and male
    for(j in 1:length(year_scaled)){
        logit(prob_female_rep[j]) <- beta_0 + beta_female * female_const_centered +
                                     beta_year * year_scaled[j]
        drug_use_female_rep[j] ~ dbern(prob_female_rep[j])

        logit(prob_male_rep[j]) <- beta_0 + beta_female * male_const_centered +
                                   beta_year * year_scaled[j]
        drug_use_male_rep[j] ~ dbern(prob_male_rep[j])
    }
}
```

(b)

We run the model with the following configurations:

- Use 4 chains, the overdispersed starting values for $\beta(s)$ are set from -10 to 10.
- Make 1000 iterations for adaptation.
- Take 1000 iterations for burn-in.

Then we take 1000 samples from the model and check Gelman-Rubin statistics for these β parameters. All give us value lower than 1.1. The plot of the model for each parameter also shows all 4 chains cover the same value ranges. This indicates a convergence of the simulation.

For sampling, we take another 2000 iterations. The effective sample size for these parameters are all greater than 4000.

(c)

The approximate of posterior mean, posterior standard deviation, and 95% central posterior interval for each parameter is listed in the following table.

	Mean	SD
beta_0	-1.1539593	0.0745854

	Mean	SD
beta_female	-0.5437077	0.1383312
beta_year	1.5070601	0.1676980

	2.5%	97.5%
beta_0	-1.3007705	-1.0057880
beta_female	-0.8108107	-0.2707166
beta_year	1.1865353	1.8478821

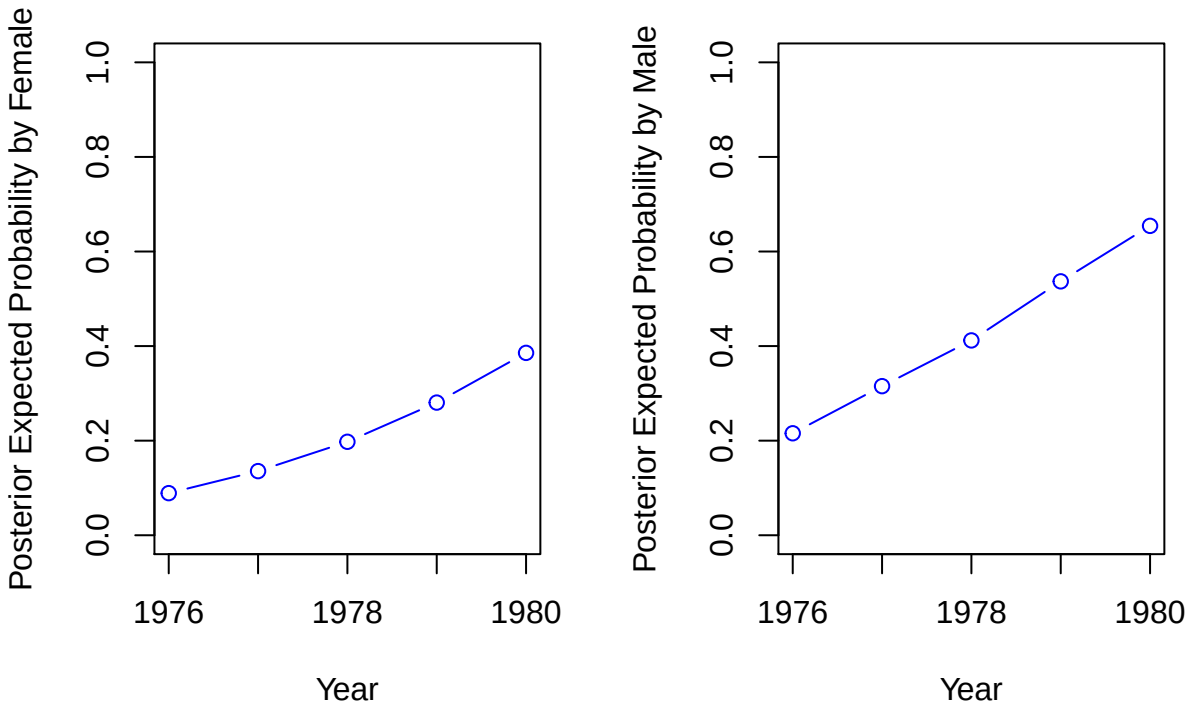
- beta_female is less than 0, means female has less probability to use drug.
- beta_year is greater than 0, means the drug use would increase as year increases.

(d)

To approximate the posterior probability that the coefficient for the (centered) indicator of being female exceeds zero, we take sample average of $\beta_{female} > 0$. This gives us value of **0**. This means, for a given year, females are **less** likely to use marijuana than males.

(e)

For each year (1976 through 1980), we approximate the posterior expected probability (according to the model) that a newly sampled female and male person would use marijuana in that year. We plot these probabilities versus year, separately for females and males in the following figures.



(f)

To evaluate the model, we approximate the value of (Plummer's) DIC using function `dic.samples` with 100,000 iterations. We got DIC score as:1258 and effective number of parameters as: 2.99. The effective number of parameters is very close to the the actual number of parameters, which is 3.

Second Model

Now we extend our first model by allowing each person to have a separate additive random effect. The model can be described as,

$$Use_{ij}|\beta, X_{ij} \sim indep.\ Bernoulli(p_{ij})$$
$$Logit(p_{ij}) = \beta_0 + \beta_{female} * female_{ij} + \beta_{year} * year_{ij} + \epsilon_i$$
$$\beta_0 \sim t_1(0, 10^2)$$
$$\beta_{female}, \beta_{year} \sim t_1(0, 2.5^2)$$
$$\epsilon_i \sim iid\ N(0, \sigma^2_{person})$$
$$\sigma_{person} \sim flat(0, \infty)$$

(a)

We modify the first JAGS model as,

```
model {
    for (i in 1:length(female_centered)) {
        for(j in 1:length(year_scaled)) {
            drug_use[i, j] ~ dbern(prob[i, j])
            logit(prob[i, j]) <- beta_0 + beta_female * female_centered[i] +
                                 beta_year * year_scaled[j] + epsilon[i]
            drug_use_rep[i, j] ~ dbern(prob[i, j])
        }
        epsilon[i] ~ dnorm(0, 1 / sigma_person^2)
    }

    beta_0 ~ dt(0, 0.01, 1)
    beta_female ~ dt(0, 0.16, 1)
    beta_year ~ dt(0, 0.16, 1)
    sigma_person ~ dunif(0,10)
}
```

(b)

This model adds 236 ϵ_i parameters and 1 σ_{person} hyperparameter. Thus is more complex than the first model and take more time to run. We run the model with the following configurations:

- Use 4 chains, the overdispersed starting values for $\beta(s)$ are set from -4 to 4 and σ_{person} from 2 to 4. The starting values are less extreme than the first model to reduce the time for model convergence.
- Take 1000 iterations for adaptation.
- Take 4000 iterations for burn-in.

Then we take 10,000 samples from the model and check Gelman-Rubin statistics for these $\beta(s)$ and σ_{person} parameters. All give us value lower than 1.1. The plot of the model for each parameter also shows all 4 chains cover the same value ranges. This indicates a convergence of the simulation.

For sampling, we take another 30,000 iterations. The effective sample size for these parameters are all greater than 4000.

(c)

The approximate of the posterior mean, posterior standard deviation, and 95% central posterior interval for the **intercept** for the coefficient of the (centered and rescaled) year number, for the coefficient of the (centered) indicator of being female, and for σ_{person} are listed in the following table.

	Mean	SD
beta_0	-2.403081	0.2821893
beta_female	-1.010033	0.4554922
beta_year	2.970627	0.2909500
sigma_person	2.955831	0.3134598

	2.5%	97.5%
beta_0	-2.984454	-1.8817809
beta_female	-1.918084	-0.1242713
beta_year	2.419904	3.5600958
sigma_person	2.390136	3.6252755

(d)

To evaluate the model, we approximate the value of (Plummer's) DIC using function `dic.samples` with 100,000 iterations. We got DIC score as:842 and effective number of parameters as: 167.899 which is smaller than the actual number of parameters (240).

Given the DIC score for second model is much smaller than the first one, we believe the second model is better than the first model.

Conclusions

In this exercise, we uses 2 Bayesian statistical models to analyze the marijuana use data by adolescence from 1976 to 1980. The first model is simple (univariate) Bayesian logistic regression model. And the second model extend the first model by adding random effect for each person.

Both models give us a reasonable fit to the data. According to the inference results, we can draw the following conclusions:

1. The use of marijuana increases every year for both female and male adolescents.
2. Male adolescent usage of marijuana is higher than female adolescent.
3. The second model fit the data better than the first one by introducing the additional parameters. But it takes much longer time to simulate.

For future exploration, we probably can further extend the model by making a hierarchical model and give different coefficient for each person, although the random effect also provide similar effect.

Appendix 1 - R Code for data preparation

```
marijuana_data = read.csv("marijuanause.csv", header=TRUE)

year = seq(1976, 1980)

marijuana_data_male = marijuana_data[marijuana_data$female==0,-1]
marijuana_data_female = marijuana_data[marijuana_data$female==1,-1]

male_count = nrow(marijuana_data_male)
female_count = nrow(marijuana_data_female)
```

Appendix 2 - R Code for model 1

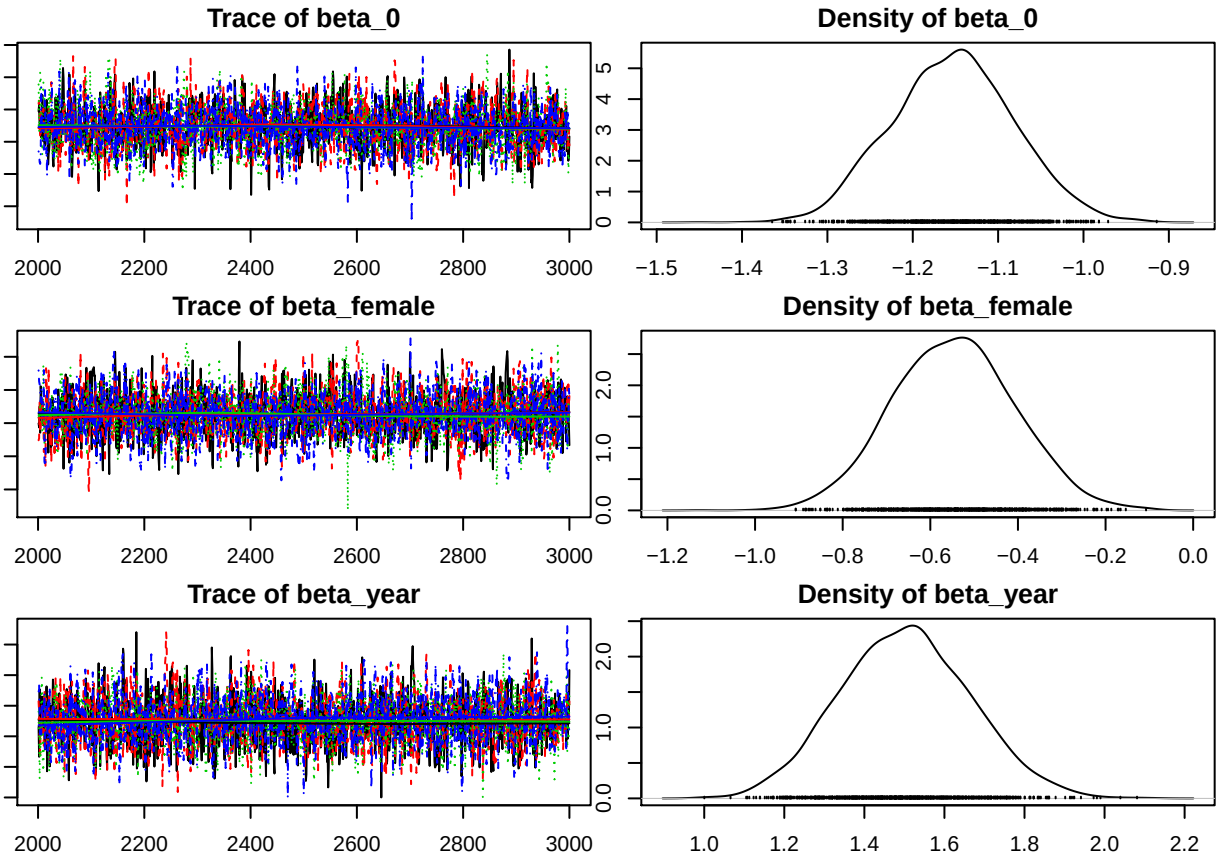
```
library(rjags)

# Setup model data, year need to be centered and scaled (by 2 sd), female only need
# to be centered.
df_jags_1 <- list( female_centered = as.vector(scale(marijuana_data$female,
                                                     center = TRUE, scale = FALSE)),
                   year_scaled = as.vector(scale(year, scale=2*sd(year))),
                   drug_use = marijuana_data[, 2:6],
                   female_const_centered = 1 - mean(marijuana_data$female),
                   male_const_centered = 0 - 1 - mean(marijuana_data$female))

# Set the starting value from -10 to 10
```

```
initial_vals_1 <- list(list(beta_0 = 10,  beta_female = 10, beta_year = 10),
                       list(beta_0 = 10,  beta_female = -10, beta_year = 10),
                       list(beta_0 = 10,  beta_female = 10, beta_year = -10),
                       list(beta_0 = -10,  beta_female = -10, beta_year = -10))

model_1 <- jags.model("drug_1.bug", df_jags_1, initial_vals_1, n.chains = 4,
                      n.adapt = 1000)

#burn-in
update(model_1, 1000)

# Take sample for Gelman-Rubin statistics
x1 <- coda.samples(model_1, c("beta_0","beta_female","beta_year"), n.iter = 1000)
```

```
# Calculate Gelman-Rubin statistics
gelman.diag(x1, autoburnin=FALSE)
```

```
## Potential scale reduction factors:
##
##             Point est. Upper C.I.
## beta_0               1       1.00
## beta_female          1       1.00
## beta_year            1       1.01
##
## Multivariate psrf
##
## 1
```

```
# Plot the convergence diagrams
par(mar=c(2,1,2,1))
plot(x1)
```




```
# Take sample for statistical analysis
coef_sample_1 <- coda.samples(model_1, c("beta_0","beta_female","beta_year",
                                         "drug_use_female_rep",
                                         "drug_use_male_rep"), n.iter = 2000)
```

```
# Calculate effective sample size for parameters
effectiveSize(coef_sample_1[,c("beta_0","beta_female","beta_year")])
```

```
##      beta_0 beta_female   beta_year
##    4575.561    4841.459    4071.410
```

Appendix 3 - R Code for model 2

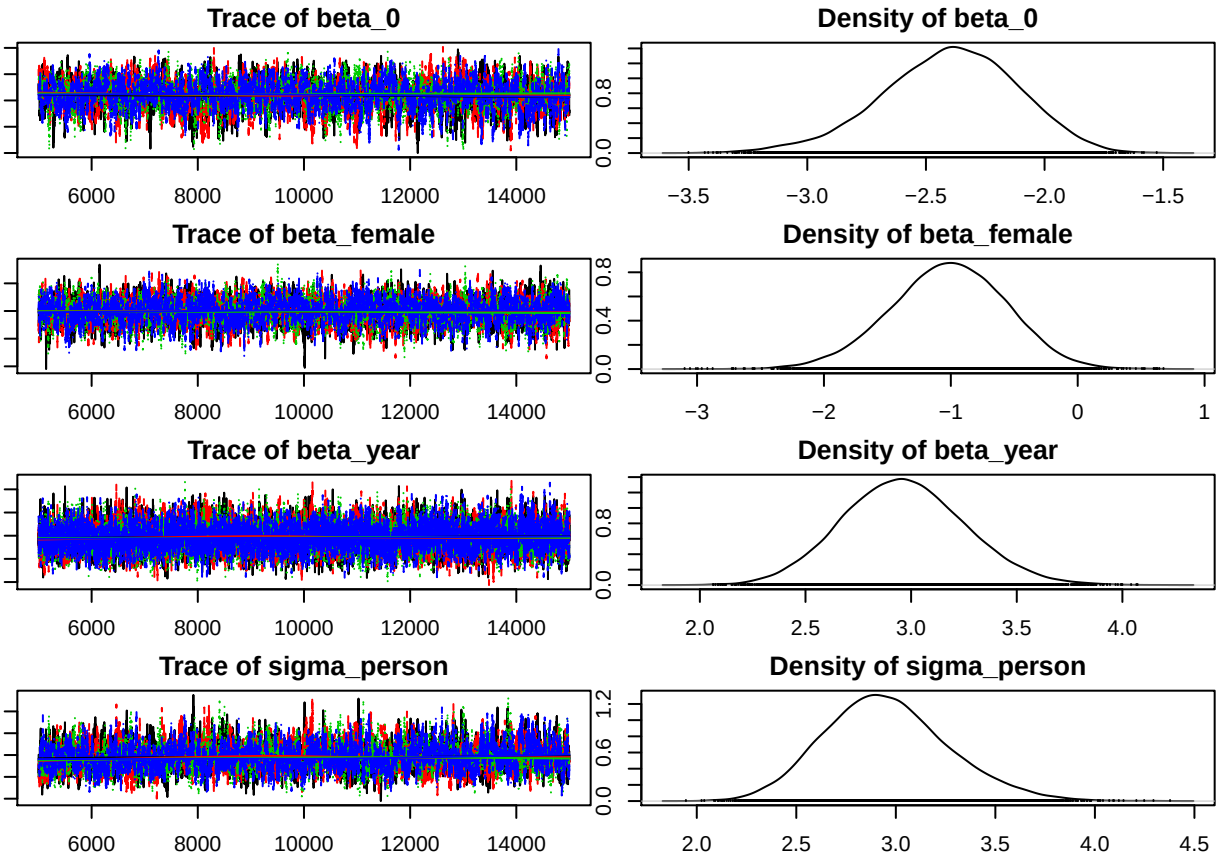
```
# Setup model data, year need to be centered and scaled (by 2 sd), female only need
# to be centered.
df_jags_2 <- list( female_centered = as.vector(scale(marijuana_data$female,
                                                     center = TRUE, scale = FALSE)),
                   year_scaled = as.vector(scale(year, scale=2*sd(year))),
                   drug_use = as.matrix(marijuana_data[, 2:6]))

# Set the starting value for beta(s) from -4 to 4 and sigma_person from 2 to 4
initial_vals_2 <- list(list(beta_0 = 4,  beta_female = 4, beta_year = 4, sigma_person = 4),
                       list(beta_0 = 4,  beta_female = -4, beta_year = 4, sigma_person = 2),
                       list(beta_0 = -4,  beta_female = 4, beta_year = -4, sigma_person = 4),
                       list(beta_0 = -4,  beta_female = -4, beta_year = -4, sigma_person = 2))
```

```
model_2 <- jags.model("drug_2.bug", df_jags_2, initial_vals_2, n.chains = 4,
                      n.adapt = 1000)

#burn-in
update(model_2, 4000)

# Take sample for Gelman-Rubin statistics
x2 <- coda.samples(model_2, c("beta_0","beta_female","beta_year", "sigma_person"),
                   n.iter = 10000)
```

```
# Calculate Gelman-Rubin statistics
gelman.diag(x2, autoburnin=FALSE)
```

```
## Potential scale reduction factors:
##
##              Point est. Upper C.I.
## beta_0                1          1
## beta_female           1          1
## beta_year             1          1
## sigma_person          1          1
##
## Multivariate psrf
##
## 1
```

```
# Plot the convergence diagrams
par(mar=c(2,1,2,1))
plot(x2)
```




```
# Take sample for statistical analysis
coef_sample_2 <- coda.samples(model_2, c("beta_0","beta_female","beta_year", "sigma_person",
                                         "drug_use_rep"), n.iter = 30000)
```

```
# Calculate effective sample size for parameters
effectiveSize(coef_sample_2[,c("beta_0","beta_female","beta_year", "sigma_person")])
```

```
##       beta_0  beta_female    beta_year sigma_person
##     4169.789     7394.954    11811.595     4697.939
```