

# Final Exam — Bayesian Data Analysis

Kári Hlynsson

November 20, 2023

## Exercise 1

Let  $Y_i$  denote the  $i$ th observed count of pulses from a Geiger counter that was pointed toward a specimen of mineral over a period of  $t_i$  minutes,  $i = 1, \dots, n$ , where  $n$  is the number of measurement periods. The purpose is to measure radiation. The intensity of the radiation is measured in pulses per minute.

Assume that the counts,  $Y_i$ , are independent and follow a Poisson distribution. The probability mass function of  $Y_i$  is

$$\mathcal{L}(Y_i|\theta) = \frac{\exp(-t_i\theta)(t_i\theta)^{Y_i}}{Y_i!}, \quad Y_i \in \{0, 1, 2, \dots\}, \quad i \in \{1, \dots, n\},$$

where  $\theta$  is an unknown parameter such that  $\theta > 0$ , representing the intensity of the radiation per minute. Assume that the prior density of  $\theta$  is a gamma density with parameters  $\alpha$  and  $\beta$ .

- Find the posterior distribution of  $\theta$  (with the normalising constant). Hint: conjugate distributions.
- The following data were observed.

$i$	1	2	3	4	5	6	7	8	9	10
$t_i$	2.9	2.0	2.2	3.4	2.1	2.4	1.9	1.8	1.7	2.7
$Y_i$	16	10	6	8	3	9	8	11	5	6

Based on these data, calculate the posterior mean, standard deviation and 95% interval of  $\theta$ . Set the prior density such that  $\alpha = 1$ , and  $\beta = 0.01$ .

- Draw the posterior density of  $\theta$ .

## Solution

### Part (a)

By Bayes' theorem, one has

$$\begin{aligned} p(\theta|\mathbf{Y}) &\propto \mathcal{L}(\mathbf{Y}|\theta)\pi(\theta) = \prod_{i=1}^n \mathcal{L}(Y_i|\theta)\pi(\theta) \\ &\propto \left[ \prod_{i=1}^n \theta^{Y_i} \exp(-t_i\theta) \right] \theta^{\alpha-1} \exp(-\beta\theta) \\ &= \theta^{\sum_{i=1}^n Y_i + \alpha - 1} \exp \left[ - \left( \sum_{i=1}^n t_i + \beta \right) \theta \right] \end{aligned}$$

where we used that the  $Y_i$ s are i.i.d. Thus we see that  $\theta \sim \text{Gamma}(\sum_{i=1}^n Y_i + \alpha, \sum_{i=1}^n t_i + \beta)$ . This is to be expected, since a gamma prior paired with a Poisson likelihood is conjugate, yielding a gamma posterior. The full form of the distribution is

$$p(\theta|\mathbf{Y}) = \frac{B^A}{\Gamma(A)} \theta^{A-1} \exp(-B\theta) = \frac{(\sum_{i=1}^n t_i + \beta)^{\sum_{i=1}^n Y_i + \alpha}}{\Gamma(\sum_{i=1}^n Y_i + \alpha)} \theta^{\sum_{i=1}^n Y_i + \alpha - 1} \exp \left[ - \left( \sum_{i=1}^n t_i + \beta \right) \theta \right]$$

where  $A := \sum_{i=1}^n Y_i + \alpha$  and  $B := \sum_{i=1}^n t_i + \beta$ .

### Part (b)

This is a matter of computation. First generate draws from the posterior using Monte Carlo sampling:

```

alpha <- 1
beta  <- 0.01
t     <- c(2.9, 2.0, 2.2, 3.4, 2.1, 2.4, 1.9, 1.8, 1.7, 2.7)
Y     <- c(16, 10, 6, 8, 3, 9, 8, 11, 5, 6)

S     <- 1e+5
post  <- rgamma(n = S, shape = sum(Y) + alpha, rate = sum(t) + beta)

```

Summarise the posterior as follows:

```

mean_post <- mean(post)
sd_post  <- sd(post)
ci_post  <- quantile(post, c(0.025, 0.975))

```

The results of the summarisation can be seen in Table 1.

Mean	Standard deviation	95% credible set
3.592168	0.3939559	[2.8602, 4.4060]

Table 1: Summary statistics of the model parameter  $\theta$ .

One can even verify that this is truly a 95% credible set for  $\theta$  by drawing Monte Carlo samples and computing the proportion of draws included in the interval specified in the table:

```

ci_check <- rgamma(n = S, shape = sum(Y) + alpha, rate = sum(t) + beta)
mean(ci_check > ci_post[1] & ci_check < ci_post[2])

```

```
## [1] 0.9508
```

Evidently, this interval is the 95% credible set since it contains approximately 95% of the samples drawn.

## Part (c)

The posterior density of  $\theta$  is shown in Figure 1.

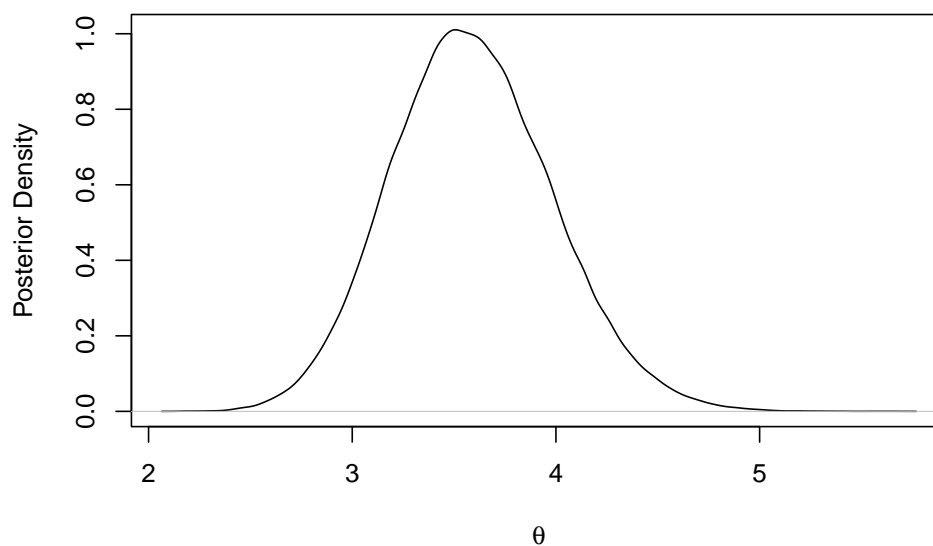


Figure 1: The posterior density of the model parameter  $\theta$ .

## Exercise 2

The data set `problem2.txt` contains measurements on the annual maximum daily precipitation (mm per 24 hours, from 9:00 AM to 9:00 AM the next day) recorded at Reykjavík over the years 1926 to 2022. Reykjavík is located in the southwest part of Iceland (64°07.648' N, 21°54.166' W).

- (a) Assume the data follow a normal distribution, that is

$$Y_i \sim \mathcal{N}(\mu, \sigma^2), \quad i = 1, \dots, n,$$

where  $\mu \in \mathbb{R}$ ,  $\sigma^2 > 0$ . Assign the following prior densities of  $\mu$  and  $\sigma^2$ ,

$$\mu \sim \mathcal{N}(0, 100^2), \quad \sigma^2 \sim \text{InvGamma}(0.1, 0.1).$$

Sample from the posterior density of  $(\mu, \sigma^2)$  by using a Gibbs sampler. Present the formulas for the Gibbs sampler. Write code for a Gibbs sampler that samples from the posterior density of  $(\mu, \sigma^2)$  (do not use JAGS or other modeling libraries). Provide posterior summary for  $\mu$  and  $\sigma^2$ , that is, present the posterior mean, the posterior standard deviation and the 95% equal-tailed posterior interval (based on 4 chains, each of length 13000 with 3000 for burn-in).

- (b) Plot the theoretical cumulative density function (c.d.f.) of the normal distribution using the posterior means of  $\mu$  and  $\sigma^2$ . Plot the empirical cumulative density of the data on the same graph. Does the proposed normal model capture the shape of the empirical cumulative density?
- (c) Draw the trace plots for  $\mu$ , and  $\sigma^2$ , using the `plot` command in `rjags`. Do these plots indicate that the chains of the two parameters are mixing well?
- (d) Use the `autocorr.plot` command in `rjags` to plot the autocorrelation function of each of the sampled parameters. Is the autocorrelation strong or weak? Do the autocorrelation functions indicate that the chains of the parameters will converge quickly or slowly?
- (e) Compute the following diagnostic statistics: the lag 1 autocorrelation, the effective sample size, and the Gelman-Rubin statistic for each of the two parameters. Interpret these diagnostic statistics.
- (f) Assume here that the data follow a lognormal distribution. This is equivalent to assuming that  $U_i = \log(Y_i)$  follows a normal distribution, that is

$$U_i \sim \text{Normal}(\eta, \kappa^2), \quad i = 1, \dots, n,$$

where  $\eta \in \mathbb{R}$ ,  $\kappa^2 > 0$ . Assign the following prior densities to  $\eta$  and  $\kappa^2$ ,

$$\eta \sim \text{Normal}(0, 100^2), \quad \kappa^2 \sim \text{InvGamma}(0.1, 0.1).$$

Sample from the posterior density of  $(\eta, \kappa^2)$  by using a Gibbs sampler (same formulas as in (a) but a different response variable). Provide posterior summary for  $\eta$  and  $\kappa^2$ , that is, present their posterior means, the posterior standard deviations and their equal-tailed posterior intervals (based on 4 chains, each of length 13000 with 3000 for burn-in).

- (g) Plot the theoretical cumulative density function (c.d.f.) of the normal distribution using the posterior means of  $\eta$  and  $\kappa^2$ . Plot the empirical cumulative density of the log-transformed data on the same graph. Does the proposed normal model capture the shape of the empirical cumulative density of the log-transformed data?
- (h) Based on the plots in (b) and (g), which of the two models appears to be better suited for the annual maximum daily precipitation data?

## Solution

### Part (a)

Derivations from the course book show that given the model

$$Y_i \sim \mathcal{N}(\mu, \sigma^2), \quad \mu \sim \mathcal{N}(\gamma, \tau^2) \quad \text{and} \quad \sigma^2 \sim \text{InvGamma}(\alpha, \beta)$$

it can be shown that

$$\begin{aligned} \mu | \sigma^2, \mathbf{Y} &\sim \mathcal{N} \left( \frac{\sum_{i=1}^n Y_i / \sigma^2 + \gamma / \tau^2}{n / \sigma^2 + 1 / \tau^2}, \frac{1}{n / \sigma^2 + 1 / \tau^2} \right), \\ \sigma^2 | \mu, \mathbf{Y} &\sim \text{InvGamma} \left( n/2 + \alpha, \sum_{i=1}^n (Y_i - \mu)^2 / 2 + \beta \right). \end{aligned}$$

Thus, in our case where  $\gamma = 0$ ,  $\tau^2 = 100^2$ ,  $\alpha = \beta = 0.1$ , the full conditionals for model parameters  $\mu$  and  $\sigma^2$  are given by

$$\begin{aligned} \mu | \sigma^2, \mathbf{Y} &\sim \mathcal{N} \left( \frac{\sum_{i=1}^n Y_i / \sigma^2}{n / \sigma^2 + 1 / 100^2}, \frac{1}{n / \sigma^2 + 1 / 100^2} \right) \\ \sigma^2 | \mu, \mathbf{Y} &\sim \text{InvGamma} \left( n/2 + 0.01, \sum_{i=1}^n (Y_i - \mu)^2 / 2 + 0.01 \right) \end{aligned}$$

We now have the formulae needed to set up a Gibbs sampler to summarise the posterior. All that is left is to implement the algorithm in code.

```
# set seed for reproducibility
set.seed(42)

data2 <- read.table("data/problem2.txt", header=T)
Y      <- data2[,2]
year   <- data2[,1]

n <- length(Y)

# no. samples + burn-in
burn <- 3000
S    <- 1e+4 + burn

compute_chain <- function() {

  # Create data frame for model params
  theta      <- matrix(NA, nrow = S, ncol = 2)
  theta[1,]  <- c(mean(Y), var(Y))

  colnames(theta) <- c("mu", "sigma2")

  # Gibbs sampling
  for (s in 2:S) {
    sigma2 <- theta[s - 1, "sigma2"]
```

```

N <- sum(Y) / sigma2
D <- n/sigma2 + 1/100^2
mu <- rnorm(1, mean = N/D, sqrt(D))

A <- n/2 + 0.01
B <- sum((Y - mu)^2)/2 + 0.01
sig2 <- rinvgamma(1, A, B)

  theta[s,] <- c(mu, sig2)
}
return(theta[(burn + 1):S,])
}

```

Using the `compute_chain` function defined above, we can create four chains, convert them into `mcmc` objects from `rjags`, and combine them into a `mcmc.list` which can be summarised using functions from `rjags`.

```

chains <- lapply(1:4, function(i) mcmc(compute_chain(i)))
chains_jags <- mcmc.list(chains)
sum_jags <- summary(chains_jags)

```

```
sum_jags
```

```

##
## Iterations = 1:10000
## Thinning interval = 1
## Number of chains = 4
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean      SD Naive SE Time-series SE
## mu         31.29   1.136  0.00568      0.005591
## sigma2     78.13  11.670  0.05835      0.057471
##
## 2. Quantiles for each variable:
##
##           2.5%   25%   50%   75%   97.5%
## mu          29.07 30.53 31.29 32.06 33.54
## sigma2      58.61 69.81 77.02 85.19 104.00

```

From the summary output above one sees that the posterior mean for  $\mu$  is 30.64, with posterior standard deviation 0.8256 and 95% credible set [29.02, 32.28]. The full summary of the model parameters are shown below in Table 2.

	Mean	Standard deviation	95% credible set
$\mu$	31.2923739	1.1359336	[29.0727376, 33.543175]
$\sigma^2$	78.1299623	11.6695249	[58.6128057, 104.0041172]

Table 2: Summary statistics for parameters  $\mu$  and  $\sigma^2$  in the model.

### Part (b)

Figure 2 shows a comparisonal plot of the empirical cumulative density function of the precipitation data compared with the theoretical c.d.f. which was constructed using the posterior mean estimates of model parameters  $\mu$  and  $\sigma^2$ . Judging from the graph, it seems like the normal model is able to capture patterns in the data with considerable accuracy.

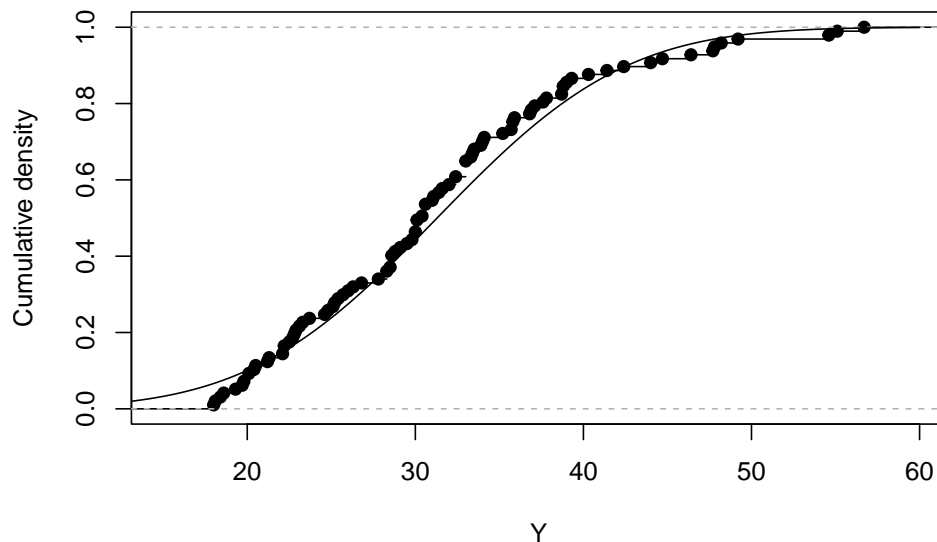


Figure 2: Empirical and theoretical cumulative density functions.

### Part (c)

Trace and density plots of model parameters  $\mu$  and  $\sigma^2$  are shown in Figure 3. Linear structure in the trace plots is indicative of the fact that the parameter estimates converge to a probabilistic centre (some prefer to describe the structure as caterpillar or barcode-like). Thus the parameters seems to be mixing well.

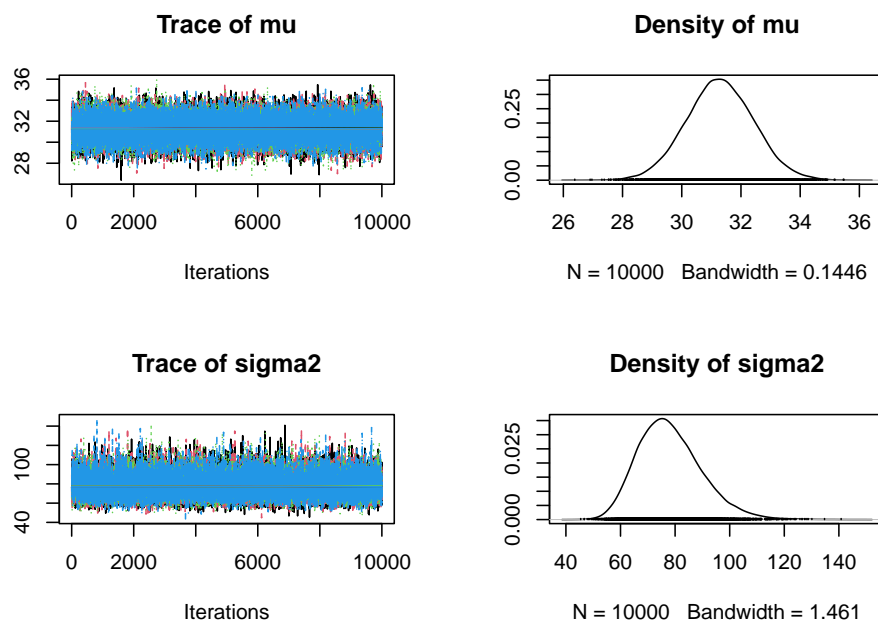


Figure 3: Trace plots of model parameters  $\mu$  and  $\sigma^2$  in the normal model.

**Part (d)**

Figure 4 shows autocorrelation plots for each of the chains generated in part (a). The first line of the plots shows the autocorrelation plots for the first chain, and so forth. The graphs show very little correlation besides at lag  $\ell = 0$  which is obviously to be expected. This implies a rapid rate of convergence, since for lag  $\ell > 0$  it seems that the dependence or correlation between samples is very little at best, indicating that the samples being drawn from the posterior almost instantly become independent. One could raise point out that this autocorrelation analysis was performed on burned-in samples and investigating the raw samples would be more appropriate, but they yield the exact same results since we picked good initial values.

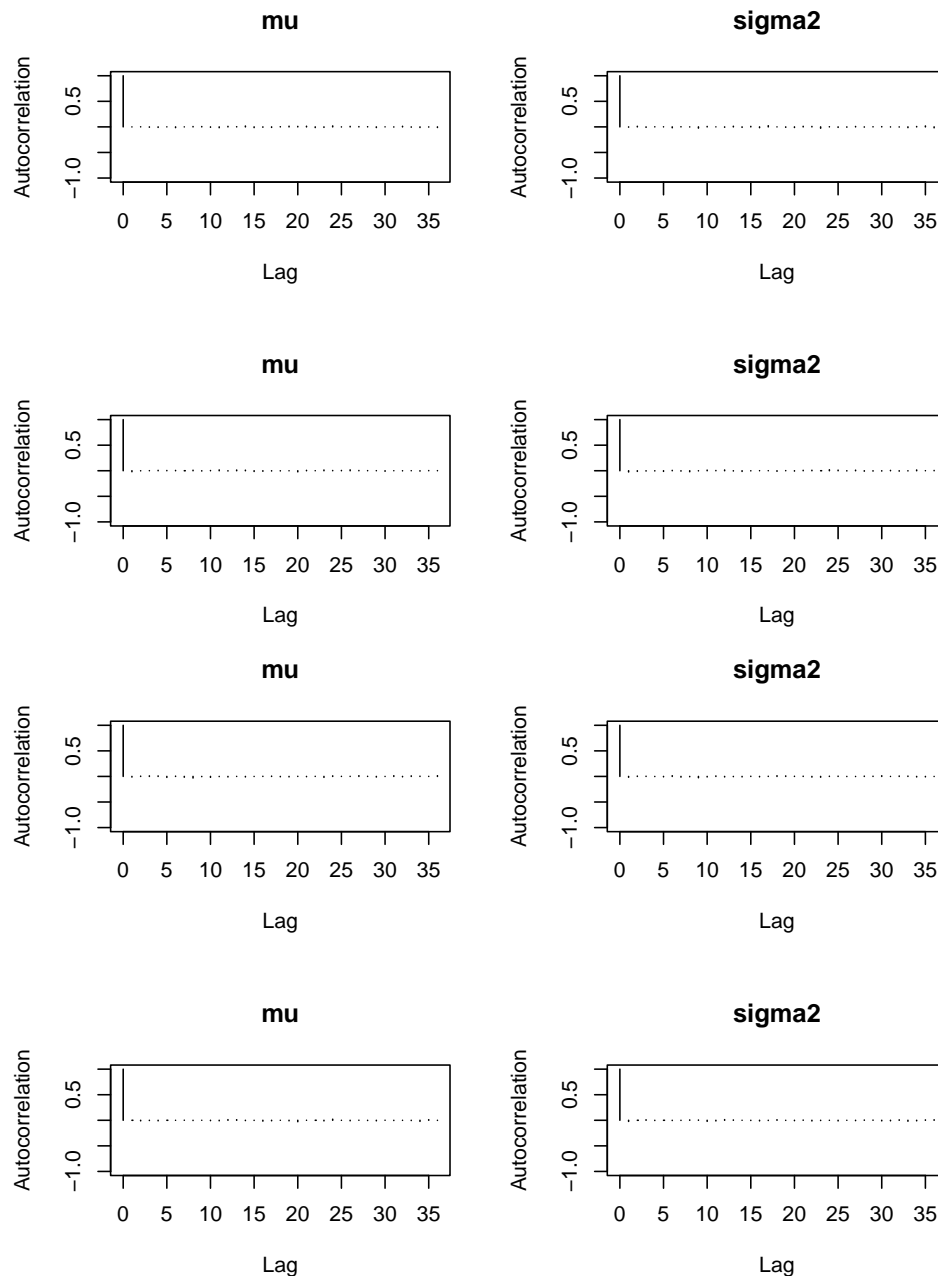


Figure 4: Autocorrelation plots for the chains generated in part (a).



## Part (e)

Start with the autocorrelation with  $\ell = 1$ :

```
autocorr.diag(chains_jags, lag = 1)
```

```
##              mu      sigma2
## Lag 1 -0.007896202 -0.01662788
```

There is very little correlation between samples for both  $\mu$  and  $\sigma^2$ , virtually none. This indicates that for a given sample, its predecessor has gives no bearing on how its successor evolves, which implies that the assumption of independent sampling holds. Let us next look at the effective sample size diagnostic:

```
effectiveSize(chains_jags)
```

```
##      mu      sigma2
## 41304.98 41245.44
```

In both cases it is very high, exceeding the “couple thousand” rule-of-thumb threshold for all parameters. In fact, it is centred around the Monte Carlo sample size  $S$  which was used in the code for the Gibbs sampler, which implies that the information per sample contained in the samples in the four chains is one-to-one. Thus there is no redundancy in the samples which is equivalent to the samples being uncorrelated (this also follows from the definition of the *ESS* diagnostic). Finish with the Gelman-Rubin diagnostic:

```
gelman.diag(chains_jags)
```

```
## Potential scale reduction factors:
##
##      Point est. Upper C.I.
## mu           1           1
## sigma2       1           1
##
## Multivariate psrf
##
## 1
```

In similar fashion, Gelman-Rubin diagnostic indicates excellent convergence. The Gelman-Rubin diagnostic is essentially the scaled result of an ANOVA test which compares the means between chains, or in other words whether they agree. In this case it is evident.

## Part (f)

The Gibbs sampler is set up below, this time using JAGS for convenience:

```
init_log <- textConnection("model{
  for (i in 1:n) {
    Y[i] ~ dlnorm(eta, tau)
  }

  eta ~ dnorm(0, 0.0001)
  tau ~ dgamma(0.1, 0.1)
  kappa <- 1/tau
}")

model_log <- jags.model(init_log, data = list(Y = Y, n = n), n.chains = 4, quiet = TRUE)
update(model_log, n.iter = 3000, progress.bar = "none")
samples_log <- coda.samples(model_log, variable.names = c("eta", "kappa"), n.iter = 10000,
  progress.bar = "none")
```

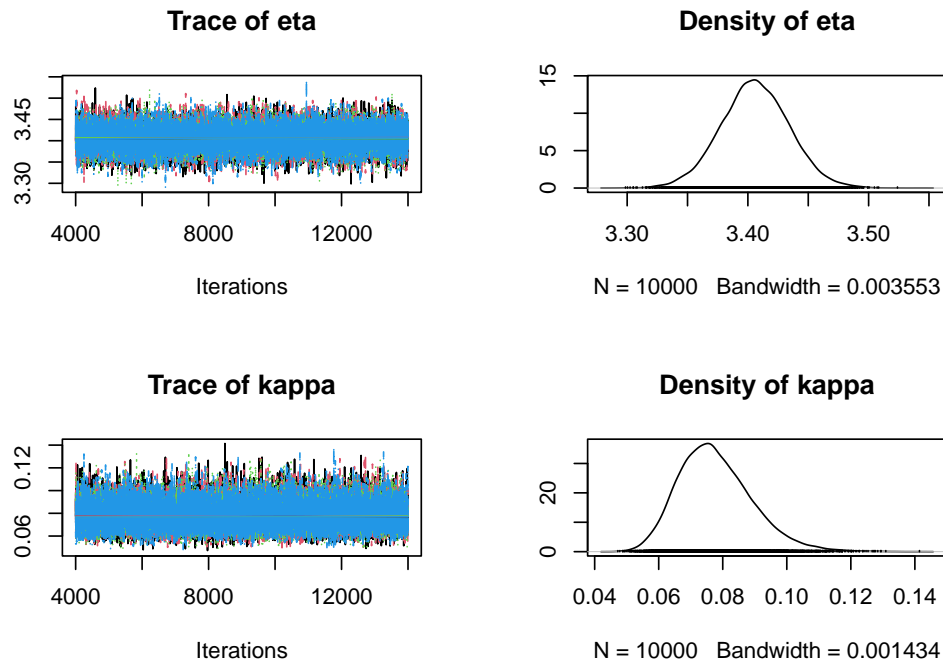


Figure 5: Trace and density plots of model parameters  $\eta$  and  $\kappa^2$  in the log-normal model.

Figure 5 shows trace and density plots of model parameters  $\eta$  and  $\kappa^2$ . The chains are obviously convergent. Let us look at some summary statistics for the log-normal model, also shown in Table 3.

```
sum_precip <- summary(samples_log)
sum_precip

##
## Iterations = 4001:14000
## Thinning interval = 1
## Number of chains = 4
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##      Mean      SD Naive SE Time-series SE
## eta   3.40598 0.02835 0.0001417      0.0001838
## kappa 0.07774 0.01146 0.0000573      0.0000572
##
## 2. Quantiles for each variable:
##
##      2.5%    25%    50%    75%   97.5%
## eta   3.34963 3.38738 3.40584 3.42478 3.4621
## kappa 0.05852 0.06957 0.07657 0.08466 0.1034
```

	Mean	Standard deviation	95% credible set
$\eta$	3.4059848	0.0283456	[3.3496337, 3.4621452]
$\kappa^2$	0.0777426	0.0114592	[0.0585218, 0.103382]

Table 3: Summary statistics for parameters  $\eta$  and  $\kappa^2$  in the log-normal model, rounded to two decimals.

**Part (g)**

Figure 6 shows the empirical and theoretical cumulative density functions.

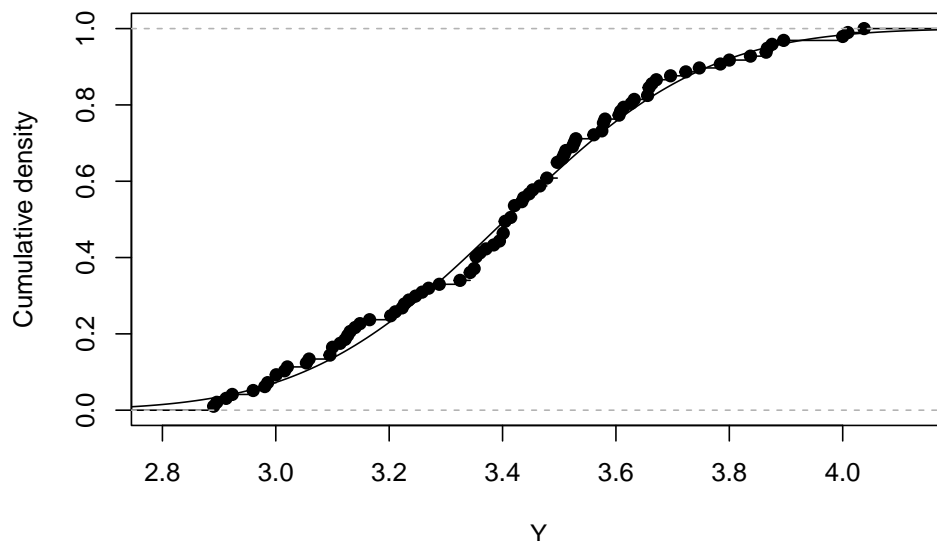


Figure 6: Empirical and theoretical cumulative density functions in the log-normal model.

**Part (h)**

A comparison of Figures 2 and 6 shows that the discrepancy is lessened when working with the log-transformed data, i.e. the difference between the empirical cumulative density and the theoretical one is more similar in the log-normal model. Therefore, this model is preferable to the one using the untransformed data.

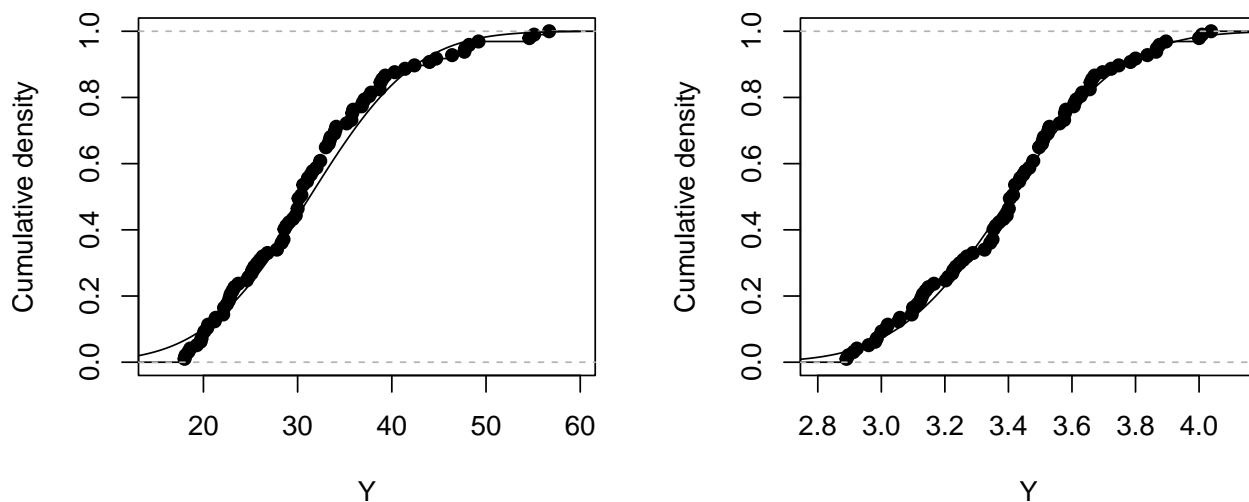


Figure 7: Visual comparison of the untransformed and log-normal models.

## Exercise 3

Here we explore two methods to compare proportions of individuals with a particular trait in two populations, namely DIC and Bayes factor. Assume that we have independent samples from each of the populations, and the sample sizes are  $n_1 = 1053$  and  $n_2 = 976$ , respectively. Let  $Y_j$  denote the number of individuals with the trait in the sample from population  $j \in \{1, 2\}$ . Assume that the samples are such that  $Y_1 = 26$  and  $Y_2 = 45$ .

To compare the two proportions, we compare two models. Model 1,  $\mathcal{M}_1$ , is such that

$$Y_1 \sim \text{Bin}(n_1, \theta), \quad Y_2 \sim \text{Bin}(n_2, \theta), \quad \theta \sim \text{Beta}(1, 1).$$

Model 2,  $\mathcal{M}_2$ , is such that

$$Y_1 \sim \text{Bin}(n_1, \theta_1), \quad Y_2 \sim \text{Bin}(n_2, \theta_2), \quad \theta_1 \sim \text{Beta}(1, 1), \quad \theta_2 \sim \text{Beta}(1, 1).$$

- Use DIC to compare the two models. Use simulation to estimate DIC for each of the two models. In the case of the above data, does the difference in DIC indicate that Model 2 is a more suitable model?
- Use Bayes factor to compare the models. Assume *a priori* that the two models are equally likely. Here we use the result for Bayes factor directly. The Bayes factor of Model 2 relative to Model 1 is given by

$$\text{BF} = \frac{Y_1!(n_1 - Y_1)!Y_2!(n_2 - Y_2)!(n_1 + n_2 + 1)!}{(n_1 + 1)!(n_2 + 1)!(Y_1 + Y_2)!(n_1 + n_2 - Y_1 - Y_2)!}.$$

In the case of the above data, does the value of BF indicate that Model 2 is a more suitable model?

- Derive the formula for the Bayes factor given in (b).

## Solution

### Part (a)

The code for comparing the models via DIC is shown below.

```
n <- c(1053, 976)
Y <- c(26, 45)

init_M1 <- textConnection("model{
  for (i in 1:2) {
    Y[i] ~ dbin(theta, n[i])
  }
  theta ~ dbeta(1, 1)
}")

init_M2 <- textConnection("model{
  for (i in 1:2) {
    Y[i] ~ dbin(theta[i], n[i])
    theta[i] ~ dbeta(1, 1)
  }
}")

data <- list(n = n, Y = Y)

model_M1 <- jags.model(init_M1, data = data, n.chains = 2, quiet = TRUE)
model_M2 <- jags.model(init_M2, data = data, n.chains = 2, quiet = TRUE)

update(model_M1, n.iter = 3000, progress.bar = "none")
```

```

update(model_M2, n.iter = 3000, progress.bar = "none")

DIC_M1 <- dic.samples(model_M1, n.iter = 10000, progress.bar = "none")
DIC_M2 <- dic.samples(model_M2, n.iter = 10000, progress.bar = "none")

DIC_M1; DIC_M2

## Mean deviance: 18.6
## penalty 0.975
## Penalized deviance: 19.58

## Mean deviance: 12.67
## penalty 1.984
## Penalized deviance: 14.66

```

From the output we see that  $\text{DIC}_{\mathcal{M}_1} = 19.58$  and  $\text{DIC}_{\mathcal{M}_2} = 14.66$ . This indicates that model  $\mathcal{M}_2$  is favoured. We can verify this by comparing the result rendered by frequentist proportion test:

```

prop.test(Y, n)

##
## 2-sample test for equality of proportions with continuity correction
##
## data: Y out of n
## X-squared = 6.2596, df = 1, p-value = 0.01235
## alternative hypothesis: two.sided
## 95 percent confidence interval:
## -0.038556503 -0.004273895
## sample estimates:
## prop 1 prop 2
## 0.02469136 0.04610656

```

Again the test concludes that  $\mathcal{M}_2$  is favourable (or from a frequentist perspective, that the proportions are unequal between groups), but the result is not decisive as can be seen from the  $p$ -value of the test.

## Part (b)

The Bayes factor for comparing the two models is given by

$$\text{BF} = \frac{\Pr(\mathcal{M}_2|\mathbf{Y})/\Pr(\mathcal{M}_1|\mathbf{Y})}{\Pr(\mathcal{M}_2)/\Pr(\mathcal{M}_1)} = \frac{\Pr(\mathcal{M}_2|\mathbf{Y})}{\Pr(\mathcal{M}_1|\mathbf{Y})}$$

where the expression was reduced to the form to the right in light of the fact that  $\Pr(\mathcal{M}_1) = \Pr(\mathcal{M}_2)$ . Also, it is known to be

$$\begin{aligned} \text{BF} &= \frac{Y_1!(n_1 - Y_1)!Y_2!(n_2 - Y_2)!(n_1 + n_2 + 1)!}{(n_1 + 1)!(n_2 + 1)!(Y_1 + Y_2)!(n_1 + n_2 - Y_1 - Y_2)!} \\ &= \frac{\Gamma(Y_1 + 1)\Gamma(n_1 - Y_1 + 1)\Gamma(Y_2 + 1)\Gamma(n_2 - Y_2 + 1)\Gamma(n_1 + n_2 + 2)}{\Gamma(n_1 + 2)\Gamma(n_2 + 2)\Gamma(Y_1 + Y_2 + 1)\Gamma(n_1 + n_2 - Y_1 - Y_2 + 1)} \end{aligned}$$

since  $\Gamma(x + 1) = x!$ . So computing the Bayes factor for  $\mathcal{M}_1$  and  $\mathcal{M}_2$  can be carried out as follows:

```

numer <- lgamma(Y[1] + 1) + lgamma(n[1] - Y[1] + 1) + lgamma(Y[2] + 1) +
  lgamma(n[2] - Y[2] + 1) + lgamma(n[1] + n[2] + 2)
denom <- lgamma(n[1] + 2) + lgamma(n[2] + 2) + lgamma(Y[1] + Y[2] + 1) +
  lgamma(n[1] + n[2] - Y[1] - Y[2] + 1)

```

```
res <- exp(number - denom)
res
```

```
## [1] 0.6331682
```

Interestingly, the Bayes factor of  $\mathcal{M}_2$  relative to  $\mathcal{M}_1$  seems to slightly favour the simpler model,  $\mathcal{M}_1$ .

### Part (c)

Start by deriving the quantity in the denominator:

$$\begin{aligned}
\Pr(\mathcal{M}_1|\mathbf{Y}) &= \int_{\Omega} \Pr(\mathcal{M}_1, \mathbf{Y}) d\theta \\
&= \int_{\Omega} \mathcal{L}(\mathbf{Y}|\mathcal{M}_1) \pi(\mathcal{M}_1) d\theta \\
&= \int_{\Omega} \mathcal{L}(Y_1|\mathcal{M}_1) \mathcal{L}(Y_2|\mathcal{M}_1) \pi(\mathcal{M}_1) d\theta \\
&= \int_{\Omega} \binom{n_1}{Y_1} \theta^{Y_1} (1-\theta)^{n_1-Y_1} \binom{n_2}{Y_2} \theta^{Y_2} (1-\theta)^{n_2-Y_2} d\theta \\
&= \binom{n_1}{Y_1} \binom{n_2}{Y_2} \int_{\Omega} \theta^{(Y_1+Y_2+1)-1} (1-\theta)^{(n_1+n_2-Y_1-Y_2+1)-1} d\theta \\
&= \binom{n_1}{Y_1} \binom{n_2}{Y_2} \frac{\Gamma(Y_1+Y_2+1) \Gamma(n_1+n_2-Y_1-Y_2+1)}{\Gamma(n_1+n_2+2)} \int_{\Omega} \mathcal{L}(\vartheta|\alpha, \beta) d\vartheta \\
&= \binom{n_1}{Y_1} \binom{n_2}{Y_2} \frac{\Gamma(Y_1+Y_2+1) \Gamma(n_1+n_2-Y_1-Y_2+1)}{\Gamma(n_1+n_2+2)}
\end{aligned}$$

where  $\vartheta \sim \text{Beta}(\alpha, \beta)$  such that  $\alpha = Y_1 + Y_2 + 1$  and  $\beta = n_1 + n_2 - Y_1 - Y_2 + 1$ . The constant induced by the prior  $\pi(\mathcal{M}_i)$  is equal to one since  $\Gamma(a+b) = \Gamma(2) = 1$  and  $\Gamma(a)\Gamma(b) = \Gamma(1)^2 = 1$ . Move on to the numerator:

$$\begin{aligned}
\Pr(\mathcal{M}_2|\mathbf{Y}) &= \iint_{\Omega} \Pr(\mathcal{M}_2, \mathbf{Y}) d\theta \\
&= \iint_{\Omega} \mathcal{L}(\mathbf{Y}|\mathcal{M}_2) \pi(\mathcal{M}_2) d\theta \\
&= \int_{\Omega_1} \mathcal{L}(Y_1|\mathcal{M}_2) \pi(\mathcal{M}_2) d\theta_1 \int_{\Omega_2} \mathcal{L}(Y_2|\mathcal{M}_2) \pi(\mathcal{M}_2) d\theta_2 \\
&= \int_{\Omega_1} \binom{n_1}{Y_1} \theta_1^{Y_1} (1-\theta_1)^{n_1-Y_1} d\theta_1 \int_{\Omega_2} \binom{n_2}{Y_2} \theta_2^{Y_2} (1-\theta_2)^{n_2-Y_2} d\theta_2 \\
&= \binom{n_1}{Y_1} \binom{n_2}{Y_2} \int_{\Omega_1} \theta_1^{(Y_1+1)-1} (1-\theta_1)^{(n_1-Y_1+1)-1} d\theta_1 \int_{\Omega_2} \theta_2^{(Y_2+1)-1} (1-\theta_2)^{(n_2-Y_2+1)-1} d\theta_2 \\
&= \binom{n_1}{Y_1} \binom{n_2}{Y_2} \frac{\Gamma(Y_1+1) \Gamma(n_1-Y_1+1)}{\Gamma(n_1+2)} \frac{\Gamma(Y_2+1) \Gamma(n_2-Y_2+1)}{\Gamma(n_2+2)} \int_{\Omega_1} \mathcal{L}(\vartheta_1|\alpha_1) d\vartheta_1 \int_{\Omega_2} \mathcal{L}(\vartheta_2|\alpha_2) d\vartheta_2 \\
&= \binom{n_1}{Y_1} \binom{n_2}{Y_2} \frac{\Gamma(Y_1+1) \Gamma(n_1-Y_1+1) \Gamma(Y_2+1) \Gamma(n_2-Y_2+1)}{\Gamma(n_1+2) \Gamma(n_2+2)}
\end{aligned}$$

where  $\vartheta_1$  and  $\vartheta_2$  are Beta-distributed random variables with parameters  $\alpha_1$  and  $\alpha_2$ , respectively. Now,

$$\text{BF} = \frac{\Pr(\mathcal{M}_2|\mathbf{Y})}{\Pr(\mathcal{M}_1|\mathbf{Y})} = \frac{\Gamma(Y_1+1) \Gamma(n_1-Y_1+1) \Gamma(Y_2+1) \Gamma(n_2-Y_2+1) \Gamma(n_1+n_2+2)}{\Gamma(n_1+2) \Gamma(n_2+2) \Gamma(Y_1+Y_2+1) \Gamma(n_1+n_2-Y_1-Y_2+1)},$$

which is what we wanted to show.

## Exercise 4

For one year, the consumption of petrol,  $Y$  (in millions of gallons) was measured in 48 states. The variables that may affect the consumption of petrol are; the petrol tax,  $X_2$  (in cents per gallon); the per capita income,  $X_3$  (in 1000 dollars per month); the number of miles of paved highway,  $X_4$  (in 1000 miles); and the percentage of the population with driver's licenses,  $X_5$ . The file `problem4.txt` contains data of these variables. The columns contain  $X_2$ ,  $X_3$ ,  $X_4$ ,  $X_5$  and  $Y$ , respectively.

The following linear model is proposed

$$Y_i = \sum_{j=1}^5 X_{ij}\beta_j + \varepsilon_i, \quad i = 1, \dots, 48$$

with  $X_{i1} = 1$  for all  $i$ . Assume that the  $Y$ s are independent, normally distributed and have equal variance, that is

$$Y_i | \beta_1, \dots, \beta_5, \sigma^2 \sim \mathcal{N} \left( \sum_{j=1}^5 X_{ij}\beta_j, \sigma^2 \right), \quad i = 1, \dots, 48$$

*A priori*, the  $\beta_j$ s are independent of each other, and the prior distribution for each  $\beta_j$  is such that  $\beta_j \sim \mathcal{N}(0, 100^2)$ . The prior distribution of  $\sigma^2$  is such that  $\sigma^2 \sim \text{InvGamma}(0.1, 0.1)$ .

- Plot  $Y$  versus  $X_2, X_3, \dots, X_5$ , a total of 4 figures. Which explanatory variables show a clear relationship with  $Y$ , and which do not?
- The per capita income,  $X_3$ , and the percentage of the population with driver's licenses,  $X_5$ , are known to have an effect on the consumption of petrol,  $Y$ . The effect of the petrol tax,  $X_2$ , and the number of miles of paved highway,  $X_4$ , is not as clear. Therefore we test 4 models for the mean, namely

$$\begin{aligned} \mathcal{M}_1 : \mu_i &= \beta_1 + X_{i3}\beta_3 + X_{i5}\beta_5, & \mathcal{M}_2 : \mu_i &= \beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i5}\beta_5, \\ \mathcal{M}_3 : \mu_i &= \beta_1 + X_{i3}\beta_3 + X_{i4}\beta_4 + X_{i5}\beta_5, & \mathcal{M}_4 : \mu_i &= \beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i4}\beta_4 + X_{i5}\beta_5. \end{aligned}$$

Create a table with these 4 models where the columns of the table show the models ( $\mathcal{M}_1$  to  $\mathcal{M}_4$ ), DIC of the models, and the effective number of parameters of the models. Based on the table, select one model for these data. This model will be used below. Sample from the posterior densities of these models using JAGS.

- Draw a normal probability plot of the residuals. Do the residuals appear to follow a normal distribution. Hint: Create a vector of residuals, call it `resid` and use the commands `qqnorm(resid)` and `abline(0, sd(resid))`.
- Plot the residuals versus the predictions of the  $Y_i$ s according to the model selected in (b). Also, plot the residuals versus each of the explanatory variables. Does the variance appear to be fixed when the residuals are plotted against these variables? Is it possible that the expected value of the residuals as a function of these variables is not equal to zero?
- Compute the posterior mean and 95% marginal posterior intervals for the parameters in the final model selected in (b), that is, the  $\beta$ s and  $\sigma^2$ .
- Interpret the parameters in the model, that is, explain the effect of each explanatory variable on the expected value of the mean consumption of petrol by looking at the posterior mean of  $\beta_j$  when increasing the  $j$ th explanatory variable by one unit while holding the other explanatory variables fixed.
- Based on the model found in (b), sample from the posterior predictive distribution (PPD) of the petrol consumption for a state that has the following values;  $x_2 = 7$  cents per gallon,  $x_3 = 3.3 \times 10^3$  dollars per month,  $x_4 = 6.7 \times 10^3$ ,  $x_5 = 51\%$ . Based on these samples, compute the mean of the posterior predictive distribution and its 95% equal-tailed prediction interval.

## Solution

### Part (a)

Read in the data:

```
data4 <- read.table("data/problem4.txt", header=T)
Y <- data4[,5]
X2 <- data4[,1]
X3 <- data4[,2]
X4 <- data4[,3]
X5 <- data4[,4]
```

The plot of the response  $Y$  against the covariates  $X_i$ ,  $i = 2, \dots, 5$  is shown below in Figure 8. Judging by the figures,  $X_3$  and  $X_5$  seem to share a linear relationship with the response whereas  $X_2$  and  $X_4$  are less clear.  $X_2$  contains quite an amount of extreme values, but could possibly be related to  $Y$ , possibly a non-linear relationship. On the other hand,  $X_4$  and  $Y$  seem to be uncorrelated.

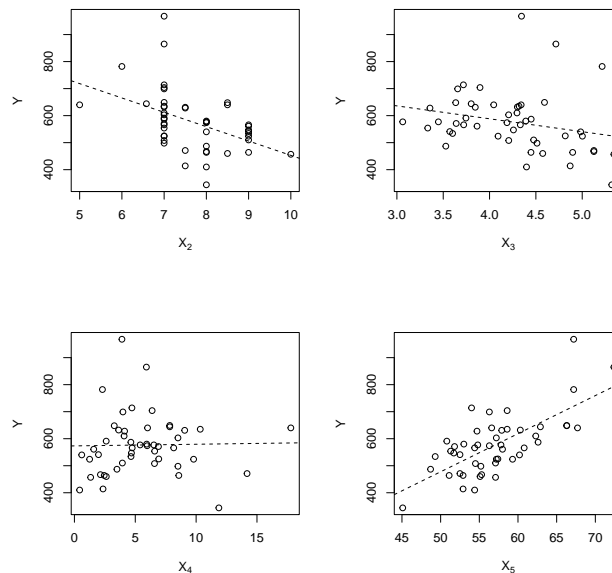


Figure 8: Plots of the response against covariates in the data.

### Part (b)

Start by fitting all the models. We also add a predictive function to the model:

```
# model string
init_model <- "model{
  for (i in 1:n) {
    Y[i] ~ dnorm(mu[i], tau)
    Yp[i] ~ dnorm(mu[i], tau)
    mu[i] <- inprod(X[i,], beta)
  }

  for (j in 1:p) { beta[j] ~ dnorm(0, 0.0001)}

  tau ~ dgamma(0.1, 0.1)
  sigma <- 1/tau
}"
```



```

# construct model matrices for M1, ..., M4 on scaled data
X_all <- cbind(1, scale(X2), scale(X3), scale(X4), scale(X5))
X_M1 <- X_all[,c(1, 3, 5)]
X_M2 <- X_all[,c(1, 2, 3, 5)]
X_M3 <- X_all[,c(1, 3, 4, 5)]
X_M4 <- X_all

# create data vectors for fitting JAGS models
model_matrices <- list(X_M1, X_M2, X_M3, X_M4)
data_vectors <- lapply(1:4, function(i) {
  X <- model_matrices[[i]] # extract model matrix
  data <- list(X = X, Y = Y, n = length(Y), p = ncol(X))
})

jags_models <- lapply(data_vectors, function(data) {
  model <- jags.model(textConnection(init_model), data, n.chains = 4, quiet = TRUE)
  update(model, n.iter = 3000, progress.bar = "none")
  return(model)
})

```

The DICs are computed for the models below, shown in Table 4. Evidently, model  $\mathcal{M}_2$  is preferred.

```

jags_dics <- lapply(jags_models, function(model) {
  dic.samples(model, n.iter = 10000, progress.bar = "none")
})
jags_dics

## [[1]]
## Mean deviance: 546.5
## penalty 4.205
## Penalized deviance: 550.7
##
## [[2]]
## Mean deviance: 539.7
## penalty 5.215
## Penalized deviance: 544.9
##
## [[3]]
## Mean deviance: 546.7
## penalty 5.23
## Penalized deviance: 551.9
##
## [[4]]
## Mean deviance: 540.3
## penalty 6.254
## Penalized deviance: 546.5

```

Model	Model form	DIC	$p_D$
$\mathcal{M}_1$	$\mu_i = \beta_1 + X_{i3}\beta_3 + X_{i5}\beta_5$	550.69	4.21
$\mathcal{M}_2$	$\mu_i = \beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i5}\beta_5$	544.94	5.22
$\mathcal{M}_3$	$\mu_i = \beta_1 + X_{i3}\beta_3 + X_{i4}\beta_4 + X_{i5}\beta_5$	551.89	5.23
$\mathcal{M}_4$	$\mu_i = \beta_1 + X_{i2}\beta_2 + X_{i3}\beta_3 + X_{i4}\beta_4 + X_{i5}\beta_5$	546.53	6.25

Table 4: Model fit criteria for the linear models  $\mathcal{M}_1, \dots, \mathcal{M}_4$ .

## Part (c)

Based on the data in Table 4, the second model,  $\mathcal{M}_2$  is the one we will use. Therefore register it in a variable, before drawing some samples to summarise the posterior. Also, compute the residual vector  $\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}}$ .

```
model_sel <- jags_models[[2]]
samples_sel <- coda.samples(model_sel, variable.names = c("beta", "sigma", "Yp"),
                           n.iter = 10000, progress.bar = "none")

Yp <- summary(samples_sel)$statistics[1:48, 1]
resid <- Y - Yp
```

Figure 9 below shows some diagnostic plots for the residuals. The structure of the residuals in the index plot (left) is possibly non-linear and thus not distributed around zero (dashed line), as is indicated by the LOWESS line (red). A normal Q-Q plot (middle) shows that the residuals stray quite a bit from the theoretical line, especially at the extremes. A histogram of the residuals (right) shows that the distribution of residuals is quite right-skewed. These observations put together might indicate a problem with the model assumptions, namely that the residuals are zero-centred normal variables with constant variance.

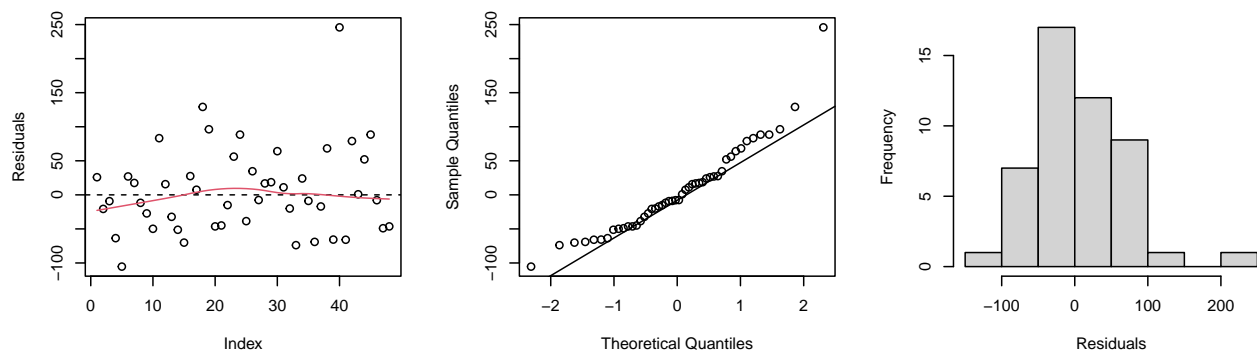


Figure 9: Residual plots for the DIC optimal model,  $\mathcal{M}_2$ .

## Part (d)

The requested plots are shown below in Figures 10 and 11. When plotted against  $X_2$ , there is possible indication of non-constant variance and non-linearity in residuals. Additionally, the structure in the plot to the right ( $X_5$ ) indicates non-linearity. The middle plot ( $X_3$ ) is a bit more vague. We can not say for certain that the expectation of the residuals as a function of  $X_2$  and  $X_5$  is zero, which is problematic. Figure 11 also seems to indicate that the residual structure is normally distributed around zero, and it is not clear whether variance is constant. However it also shows that the model seems to predict fairly well, so we use it.

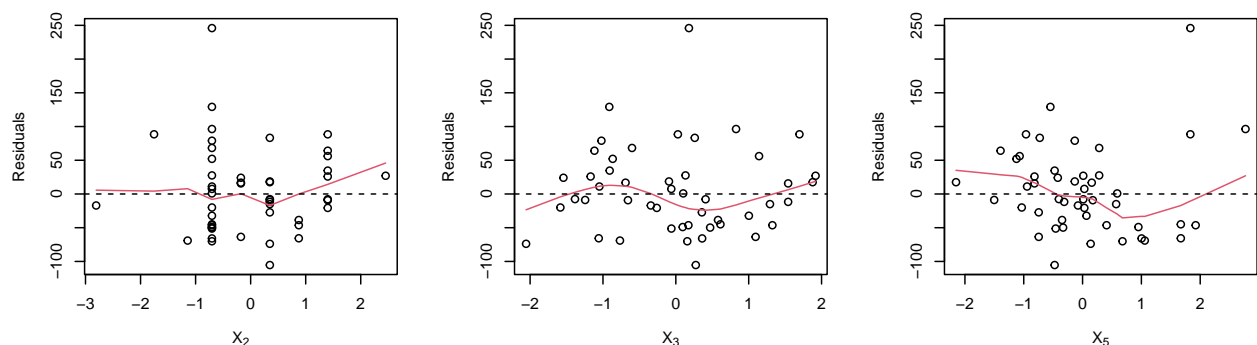


Figure 10: Residuals against each of the covariates in model  $\mathcal{M}_2$ .

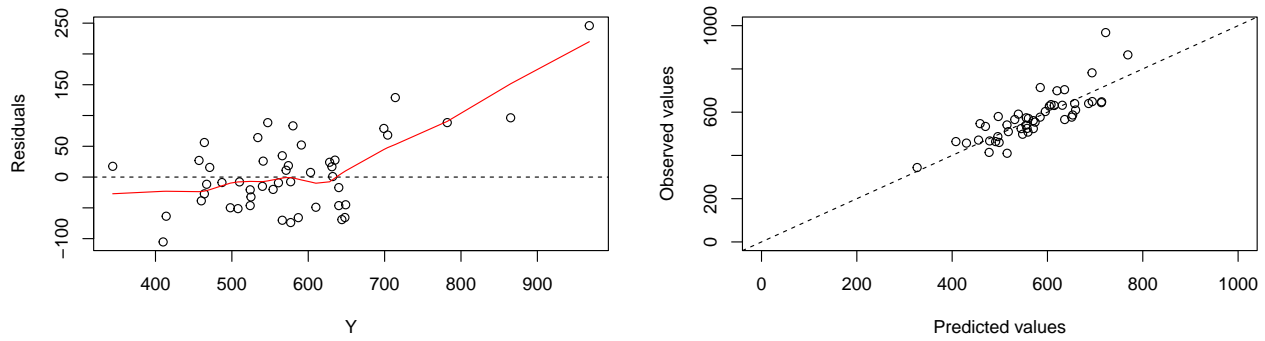


Figure 11: Residuals against fitted values in model  $\mathcal{M}_2$  (left), observed versus fitted values (right).

### Part (e)

Here is a summary of the samples, trimmed to only display the relevant parameters:

```
sum_samples_sel <- summary(samples_sel); sum_samples_sel

##
## Iterations = 13001:23000
## Thinning interval = 1
## Number of chains = 4
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean          SD   Naive SE Time-series SE
## beta[1]  571.33255    9.714891 0.04857446    0.04958464
## beta[2]  -27.90382   10.238787 0.05119393    0.05575517
## beta[3]  -38.51123    9.938218 0.04969109    0.05087810
## beta[4]   75.57987   10.346895 0.05173447    0.05819385
## sigma  4550.36786 1018.409355 5.09204677    5.57017845
##
## 2. Quantiles for each variable:
##
##           2.5%        25%        50%        75%        97.5%
## beta[1]  551.77878  564.90555  571.50234  577.86255  590.18644
## beta[2]  -48.04657  -34.71973  -27.94898  -21.05806  -7.75176
## beta[3]  -58.27586  -45.07145  -38.50538  -31.96702  -18.94362
## beta[4]   54.92778   68.79402   75.58126   82.46751   95.95370
## sigma  2980.10463 3829.89212 4409.98354 5108.85005 6961.06665
```

These are also shown in Table 5 below:

	Mean	Standard deviation	95% credible set
$\beta_1$	571.3325501	9.7148912	[551.7787761, 590.1864427]
$\beta_2$	-27.9038219	10.238787	[-48.0465686, -7.7517599]
$\beta_3$	-38.51123	9.9382177	[-58.2758558, -18.9436242]
$\beta_5$	75.5798728	10.3468949	[54.9277794, 95.9537027]
$\sigma^2$	4550.3678643	1018.4093545	[2980.1046342, 6961.0666507]

Table 5: Summary statistics for the coefficients in the linear model  $\mathcal{M}_2$ .

## Part (f)

We will use Table 5 for this purpose. We see that the intercept coefficient is  $\beta_1 = 571.33$ , which tells us that if all the other scaled covariates were set to zero, the consumption of petrol would be 571.33 millions of gallons. We must take some care in interpreting the coefficients in the normal units because we scaled the data when performing Bayesian linear regression. Due to the nature of scaling, a unit increase in standard unit corresponds to a unit standard deviation increment in the original covariate.

Credible sets follow directly from the ones presented in the table. For  $X_2$ , the coefficient is  $\beta_2 = -27.9$ , which means that increasing the petrol tax by one standard deviation would result in a decrease of approximately  $-27.9$  millions of gallons of petrol consumption per year (which may seem contradictory, but the relationship is reinforced in the data). Increasing the covariate  $X_3$  by one standard deviation, or the per capita income by a thousand dollars, results in a decrease of  $-38.51$  millions of gallons of petrol consumption per year. A unit standard deviation increase in  $X_5$  (percentage of population with drivers license) causes fuel consumption to increase by  $\beta_5 = 75.58$  millions of gallons.

## Part (g)

Set up a new JAGS model below, one which predicts the fuel consumption for the unobserved data. We will have to redefine the model matrix to include the unobserved data and scale it, before separating the two. Summarising the data following the fit will allow us to verify that this does not affect the model fit.

```
X_M2_unobs <- rbind(cbind(1, X2, X3, X5), c(1, 7, 3.300, 51))
X_M2_unobs[,2] <- scale(X_M2_unobs[,2])
X_M2_unobs[,3] <- scale(X_M2_unobs[,3])
X_M2_unobs[,4] <- scale(X_M2_unobs[,4])

X_pr      <- X_M2_unobs[49,]
X_M2_unobs <- X_M2_unobs[-49,]

init_model <- textConnection("model{
  for (i in 1:n) {
    Y[i] ~ dnorm(mu[i], tau)
    mu[i] <- inprod(X[i,], beta)
  }

  Y_pr ~ dnorm(inprod(X_pr, beta), tau)

  for (j in 1:p) { beta[j] ~ dnorm(0, 0.0001)}

  tau ~ dgamma(0.1, 0.1)
  sigma <- 1/tau
}")

data <- list(X = X_M2, Y = Y, X_pr = X_pr, n = nrow(X_M2), p = ncol(X_M2))

model_ppd <- jags.model(init_model, data = data, n.chains = 2, quiet = TRUE)
update(model_ppd, n.iter = 3000, progress.bar = "none")

samples_ppd <- coda.samples(model_ppd, variable.names = c("beta", "sigma", "Y_pr"),
                             n.iter = 10000, progress.bar = "none")
```

Summarise the posterior:

```
sum_ppd <- summary(samples_ppd)
sum_ppd
```

```
##
## Iterations = 3001:13000
## Thinning interval = 1
## Number of chains = 2
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##      Mean      SD Naive SE Time-series SE
## Y_pr      571.44  71.009  0.50211      0.50210
## beta[1]    571.25   9.752  0.06896      0.06896
## beta[2]   -27.89  10.267  0.07260      0.07947
## beta[3]   -38.58   9.895  0.06996      0.07131
## beta[4]    75.51  10.399  0.07354      0.08136
## sigma   4569.19 1027.699  7.26693      7.99668
##
## 2. Quantiles for each variable:
##
##      2.5%    25%    50%    75%   97.5%
## Y_pr      429.37 525.00 572.15 618.41 709.81
## beta[1]    551.69 564.84 571.45 577.79 590.00
## beta[2]   -47.88 -34.66 -28.05 -21.08  -7.46
## beta[3]   -57.82 -45.17 -38.69 -32.02 -18.95
## beta[4]    55.07  68.61  75.57  82.38  95.74
## sigma   2984.43 3836.40 4429.08 5144.49 6938.48
```

The estimates of the linear regression seem to be unchanged by our wrangling of the data, which is good. The mean of the posterior predictive distribution is 571.44 with 95% credible interval [429.37, 709.81]. A density plot of the samples drawn from the PPD of the unobserved data is shown in 12 below.

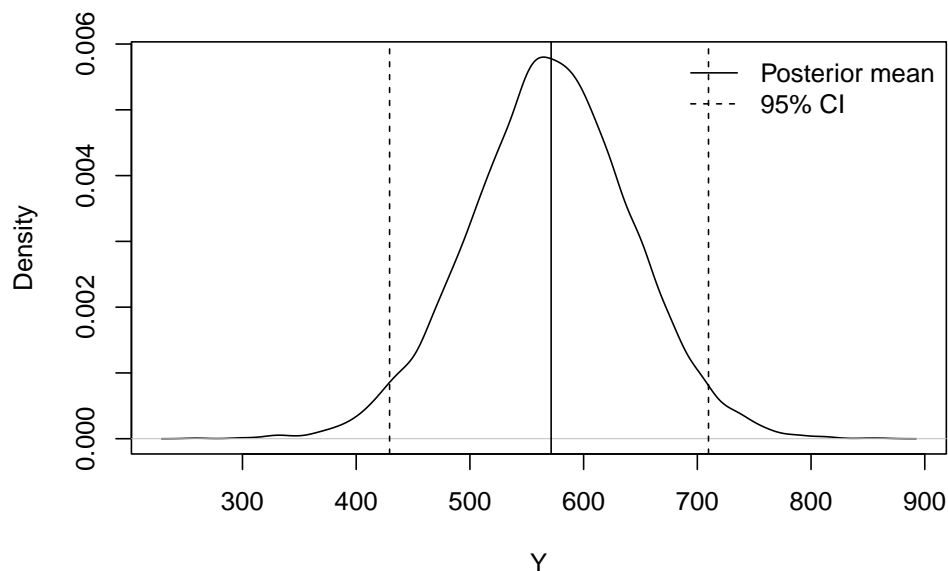


Figure 12: Density plot of samples for the posterior predictive distribution of the unobserved data.

## Exercise 5

The data in the file `problem5.txt` contain the height of 26 boys living in Oxford in England over a two year period. Each boy is measured on nine different occasions over a period of two years. The first column contains the identity number of each boy, the second column contains the time variable (in years, starts at  $-1.0$ , ends at a number close to  $1.0$ ) and the third column contains the height of the boys at the specified time point. The table below illustrates the form of the data.

id	time	height
Boy 1	-1.0000	140.52
Boy 2	-0.7479	143.42
$\vdots$	$\vdots$	$\vdots$
Boy 1	0.9945	155.89
Boy 2	-1.0000	136.91
$\vdots$	$\vdots$	$\vdots$
Boy 2	0.9945	148.39
$\vdots$	$\vdots$	$\vdots$

Let  $Y_{ij}$  be the height of the  $i$ th boy at time  $X_{ij}$  (the  $j$ th time point for the  $i$ th boy). It is assumed that the height increase linearly with time for each boy. However, the growth may vary between boys. Thus, the following model is proposed

$$Y_{ij} = \alpha_{i1} + \alpha_{i2}X_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),$$

$j = 1, \dots, T$ ,  $i = 1, \dots, n$  where  $T$  is the number of measurements taken of each boy  $i$  ( $T = 9$ ), and  $n$  is the number of boys ( $n = 26$ ). Here  $\varepsilon_{ij}$  is the mean-zero deviation of the  $j$ th measurement from the linear model for the  $i$ th boy and  $\sigma^2$  is its variance. It is assumed that the  $\varepsilon_{ij}$ s are independent of each other.

Each  $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2})^\top$  is assigned a bivariate Gaussian distribution with mean  $\boldsymbol{\beta}$  and covariance matrix  $\boldsymbol{\Omega}$ , that is,

$$\boldsymbol{\alpha}_i \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Omega}), \quad i \in \{1, \dots, n\}.$$

The following prior distributions are assigned to  $\sigma^2$ ,  $\boldsymbol{\beta}$  and  $\boldsymbol{\Omega}$ ,

$$\sigma^2 \sim \text{InvGamma}(0.1, 0.1),$$

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 100^2 \mathbf{I}_2),$$

$$\boldsymbol{\Omega} \sim \text{InvWishart}(2.1, \mathbf{I}/2.1).$$

- Find the posterior distribution of the unknown parameters in the hierarchical model. Present it in terms of the densities of the  $Y_{ij}$ s, the  $\boldsymbol{\alpha}_i$ s,  $\sigma^2$ ,  $\boldsymbol{\beta}$  and  $\boldsymbol{\Omega}$ .
- Specify the data layer, the process layer and the prior layer of the hierarchical model, that is, specify which variables/parameters are modeled at each layer, and what probability models are assumed at each level.
- Write code in JAGS within R to sample from the posterior density of the unknown parameters, and present the code. If another programming language than R (with JAGS) is used, please present the code that you use to sample from the posterior density of the unknown parameters.
- Compute the posterior means and 95% posterior intervals for  $\alpha_{11}, \dots, \alpha_{n1}, \alpha_{12}, \dots, \alpha_{n2}, \sigma^2, \boldsymbol{\beta}$  and  $\boldsymbol{\Omega}$ . Use four chains where each chain consists of 13000 iterations and the first 3000 are used for burn-in. Present the  $\alpha_{i1}$ s and the  $\alpha_{i2}$ s in two separate figures, that is, plot the id-number of the boys on the  $x$ -axis and the 95% posterior intervals along with a point for the posterior mean on the  $y$ -axis. Present  $\sigma^2$ ,  $\boldsymbol{\beta}$  and  $\boldsymbol{\Omega}$  in a table.

- (e) For each of the 26 boys plot as dots the observed heights on the  $y$ -axis and corresponding times on the  $x$ -axis. This gives 26 figures. These figures can, for example, be arranged in a seven by four matrix (with two spots empty). In the case of the plot for the  $i$ th boy, draw a line based on the formula  $\alpha_{i1} + \alpha_{i2}x$  where  $x$  goes from  $-1.1$  to  $1.1$ . Use the posterior means of  $\alpha_{i1}$  and  $\alpha_{i2}$ . Does the model appear to fit the data adequately well?

## Solution

### Part (a)

The joint posterior distribution of the unobserved variables  $\alpha_1, \dots, \alpha_n, \beta, \Omega, \sigma^2$  can be decomposed as follows:

$$p(\alpha_1, \dots, \alpha_n, \beta, \Omega, \sigma^2 | \mathbf{Y}) \propto \mathcal{L}(\mathbf{Y} | \alpha_1, \dots, \alpha_n, \sigma^2) \pi(\alpha_1, \dots, \alpha_n | \beta, \Omega) \pi(\beta) \pi(\Omega) \pi(\sigma^2)$$

To see why this is true, it helps to look at Figure 13 and trace through the dependencies for the  $\mathbf{Y}_i$ s. Now,

$$\begin{aligned} p(\alpha_1, \dots, \alpha_n, \beta, \Omega, \sigma^2 | \mathbf{Y}) &\propto \prod_{i=1}^n \left[ \prod_{j=1}^T \mathcal{L}(\mathbf{Y}_i | \alpha_i, \sigma^2) \pi(\alpha_i | \beta, \Omega) \right] \pi(\beta) \pi(\Omega) \pi(\sigma^2) \\ &\propto \prod_{i=1}^n \left[ \prod_{j=1}^T (\sigma^2)^{-1/2} \exp \left( -\frac{(Y_{ij} - \alpha_{i1} - \alpha_{i2}X_{ij})^2}{2\sigma^2} \right) \right] |\Omega|^{-1/2} \dots \\ &\quad \exp \left( -\frac{1}{2} (\alpha_i - \beta)^\top \Omega^{-1} (\alpha_i - \beta) \right) (\sigma^2)^{-a-1} \exp[-b/\sigma^2] \dots \\ &\quad \exp \left( -\frac{1}{2} \beta^\top (m^2 \mathbf{I}_2)^{-1} \beta \right) |\Omega|^{-(p-\nu-1)/2} \exp \left[ -\frac{1}{2} \text{tr} \left( \frac{\mathbf{I}_2}{\nu} \Omega^{-1} \right) \right] \\ &= (\sigma^2)^{-nT/2-a-1} |\Omega|^{-(n+p-\nu-1)/2} \exp \left[ -\frac{1}{2} \left( \frac{\sum_{i=1}^n \sum_{j=1}^T (Y_{ij} - \alpha_{i1} - \alpha_{i2}X_{ij})^2}{\sigma^2} \dots \right. \right. \\ &\quad \left. \left. \dots + \frac{2b}{\sigma^2} + (\alpha_i - \beta)^\top \Omega (\alpha_i - \beta) + \frac{1}{m^2} \beta^\top \beta + \text{tr} \left( \frac{1}{\nu} \Omega^{-1} \right) \right) \right], \end{aligned}$$

where  $a = b = 0.1$ ,  $\nu = 2.1$ ,  $p = 2$  and  $m = 100$ .

### Part (b)

Figure 13 shows the DAG representation of the model. The lowest nodes represent the data layer, where we have  $\beta \sim \mathcal{N}(\mathbf{0}, 100^2 \mathbf{I})$  which is can be thought of as the vector which fixes the centre of the distribution of growth curves for the boys. The other matrix parameter  $\Omega \sim \text{InvWishart}(2.1, \mathbf{I}/2.1)$  is used to control the spread of the growth curves, and together they feed into the growth curve parameters for the  $i$ th boy,  $\alpha_i$ . Here the assumption is made that these two parameters can give a reasonable range for each of the growth curves which will then be used to model the data. The parameter  $\sigma^2$  controls the variation of the observed height around the linear predictor  $\alpha_{i1} + \alpha_{i2}X_{ij}$ . This parameter feeds into the nodes  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ , where  $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT})$ .

The process layer is in the middle and consists of the  $\alpha_i$ s. This is where the latent data generating mechanism is constructed, i.e. the growth curves for each of the boys,  $\alpha_i = (\alpha_{i1}, \alpha_{i2})^\top$ . These are multivariate normal random variables with centre  $\beta$  and spread  $\Omega$ , that is  $\alpha_i \sim \mathcal{N}(\beta, \Omega)$ . The assumption made here is that each of the growth curves stems from the same distributions, i.e. that they are independent and identically distributed.

Lastly, the data layer is shown at the top. Here we have nodes  $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ , one for each boy, consisting of their height measurements. Here,  $Y_{ij} \sim \mathcal{N}(\alpha_{i1} + \alpha_{i2}X_{ij}, \sigma^2)$  for each  $j \in \{1, \dots, T\}$ , assuming that a linear regression model is appropriate for these data.

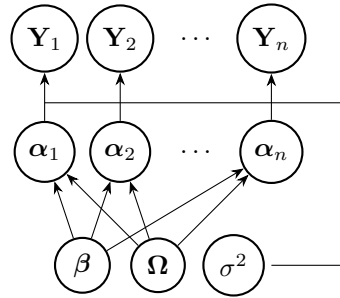


Figure 13: A DAG representation of the model.

**Part (c)**

Start by importing the data:

```

data5 <- read.table("data/problem5.txt", header=TRUE)
id <- data5[,1]
time <- data5[,2]
height <- data5[,3]

T <- 9; n <- 26

X <- matrix(nrow = n, ncol = T)
Y <- matrix(nrow = n, ncol = T)

for(i in 1:n){
  for(j in 1:T){
    k <- j + T*(i - 1)
    X[i,j] <- time[k]
    Y[i,j] <- height[k]
  } }

data <- list(X = X, Y = Y, n = n, T = T)

```

Now fit the model in JAGS:

```

init_height <- textConnection("model{
  for (i in 1:n) {
    for(j in 1:T) {
      Y[i, j] ~ dnorm(alpha[i, 1] + alpha[i, 2] * X[i, j], tau)
    } }

  for (i in 1:n) { alpha[i, 1:2] ~ dnmnorm(beta[1:2], Omega[1:2, 1:2]) }

  for (i in 1:2) { beta[i] ~ dnorm(0, 0.0001) }

  tau ~ dgamma(0.1, 0.1)

  Omega[1:2,1:2] ~ dwish(R[,],2.1)

  R[1,1]<-1/2.1
  R[1,2]<-0
  R[2,1]<-0
  R[2,2]<-1/2.1
}

```



```

sigma <- 1/tau
})")

params <- c("beta","alpha","sigma","Omega")
model <- jags.model(init_height, data = data, n.chains = 4, quiet=TRUE)
update(model, 3000, progress.bar="none")
samples_height <- coda.samples(model, variable.names = params, n.iter = 10000,
                              progress.bar="none")

```

### Part (d)

We will use the samples generated in (c) to summarise the posterior. Firstly, compute the summary:

```
sum_samples_height <- summary(samples_height)
```

Figure 14 shows a plot of the posterior means of model parameters  $\alpha_{i1}, \dots, \alpha_{n1}$  and  $\alpha_{i2}, \dots, \alpha_{n2}$ .

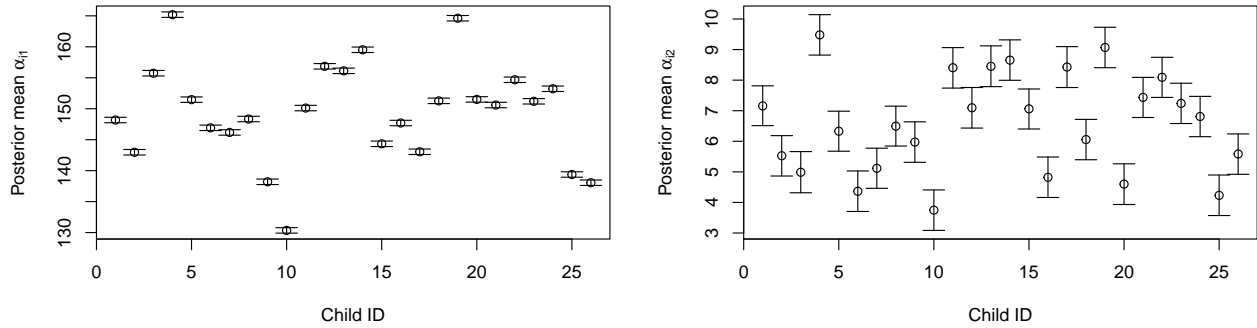


Figure 14: Posterior means of  $\alpha_{i1}$  and  $\alpha_{i2}$  for the children along with 95% credible sets.

	Mean	95% credible set
$\sigma^2$	0.455153	[0.369989, 0.5582852]
$\beta$	$\begin{pmatrix} 149.4005434 \\ 6.578375 \end{pmatrix}$	$\left[ \begin{pmatrix} 146.2616937 \\ 5.9049372 \end{pmatrix}, \begin{pmatrix} 152.528671 \\ 7.2403505 \end{pmatrix} \right]$
$\Omega$	$\begin{pmatrix} 0.0288314 & -0.0907949 \\ -0.0907949 & 0.6785469 \end{pmatrix}$	$\left[ \begin{pmatrix} 0.0152466 & -0.1681471 \\ -0.1681471 & 0.3490058 \end{pmatrix}, \begin{pmatrix} 0.0469607 & -0.0338223 \\ -0.0338223 & 1.1430574 \end{pmatrix} \right]$

Table 6: Summary statistics for the coefficients in the linear model  $\mathcal{M}_2$ .

### Part (e)

Figure 15 shows observed height measurements for each of the boys as a function of time as dots connected by dashed lines. The red line is the predicted change in height over time. As can be seen, the model appears to fit the data quite well.

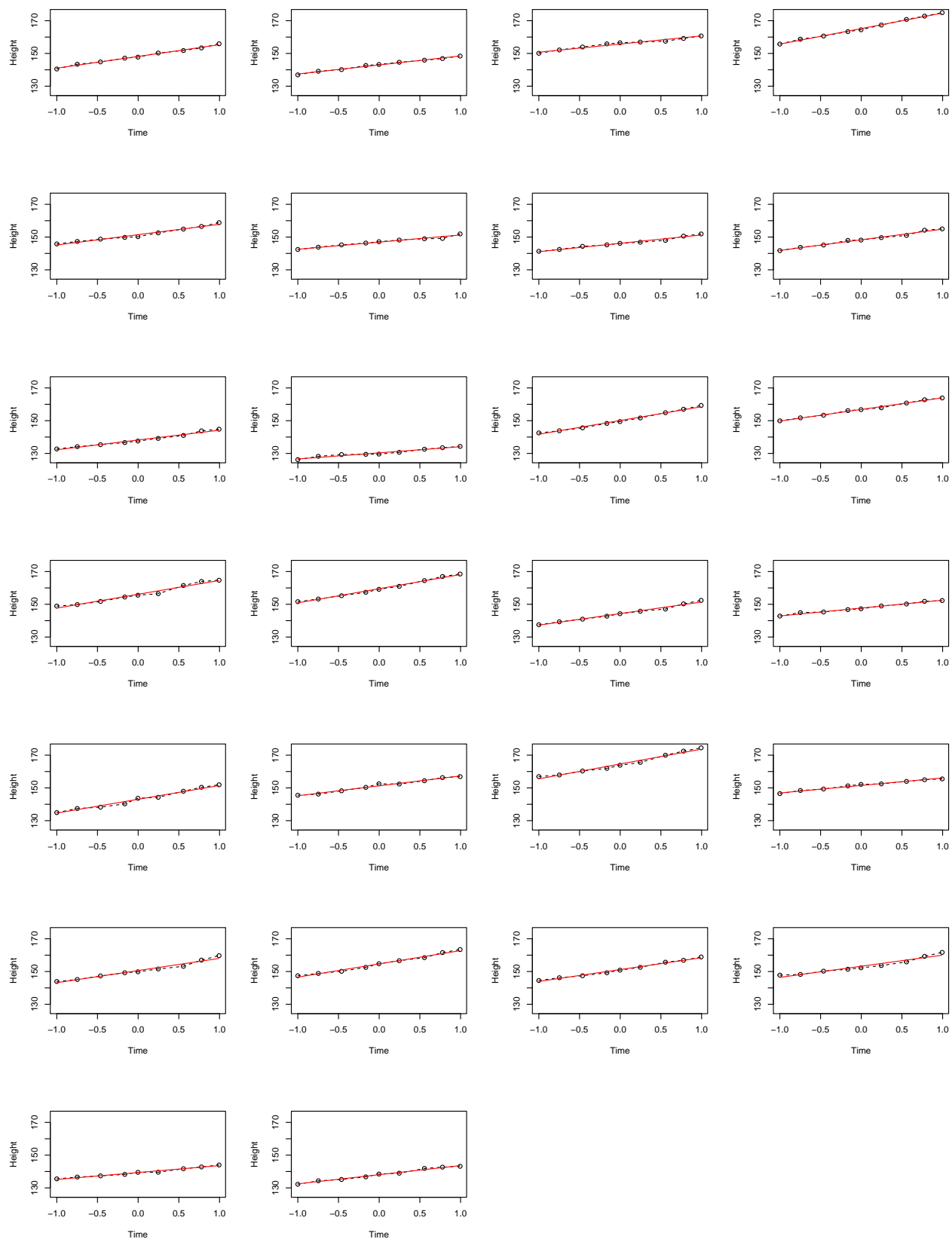


Figure 15: Observed heights in data (dots, dashed line) and the model projection (red line).