Homework #02

Statistical Methods in Data Science II & Lab 2021/2022

Michelangelo Saveriano

06/15/2022

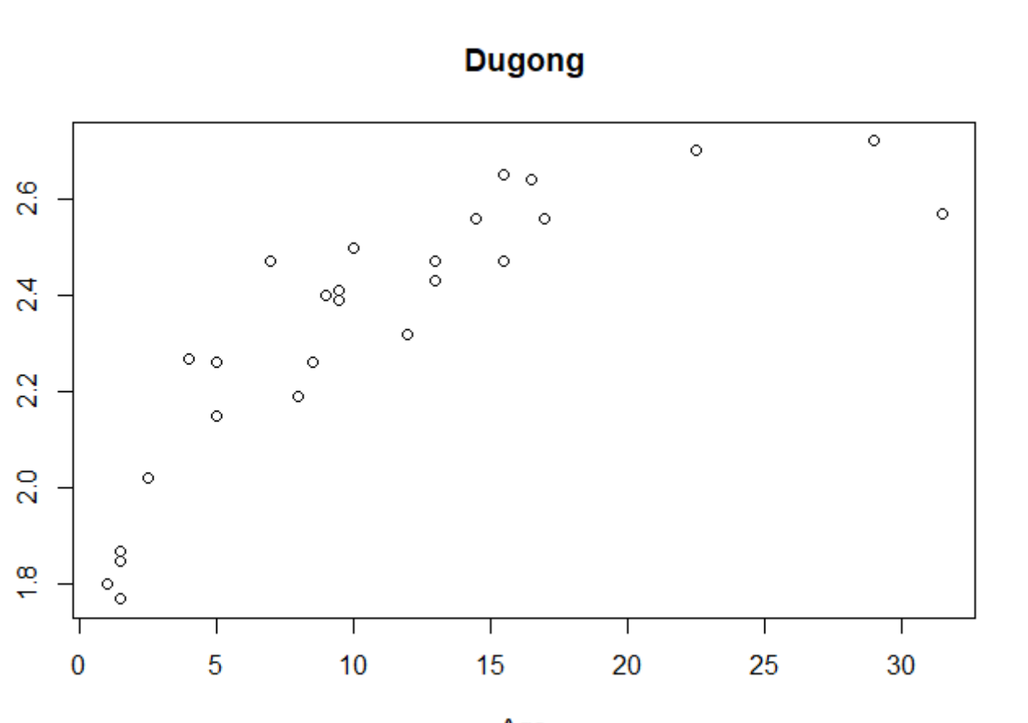
Michelangelo Saveriano

Excercise 1

1.a

Illustrate the characteristics of the statistical model for dealing with the *Dugong*'s data

The data we are considering describe the lengths (Y_i) and ages (x_i) of 27 dugongs captured off the coast of Queensland.



As we can see from the plot above age and length are **not linearly related**, that's why the following regression model is considered:

$$Y_i \sim N(\mu_i, \tau^2)$$
$$\mu_i = f(x_i) = \alpha - \beta\gamma^{x_i}$$

Model parameters are $\alpha \in (1,\infty), \beta \in (1,\infty), \gamma \in (0,1), \tau^2 \in (0,\infty)$.

Let us consider the following prior distributions:

$$\alpha \sim N(0, \sigma_\alpha^2)$$
$$\beta \sim N(0, \sigma_\beta^2)$$
$$\gamma \sim Unif(0,1)$$
$$\tau^2 \sim IG(a,b)(InverseGamma)$$

1.b

Derive the corresponding likelihood function

The Likelihood function, up to a proportionality constant, is:

$$L_y(\alpha,\beta,\gamma,\tau^2) = \prod_{i=1}^{n} f(y_i|\alpha,\beta,\gamma,\tau^2) \propto \prod_{i=1}^{n} \frac{1}{\tau^2} \exp\left(-\frac{(y_i-\mu_i)^2}{2\tau^2}\right) \propto \tau^{-2n} \exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}(y_i - \alpha + \beta\gamma^{x_i})^2\right)$$

1.c

Write down the expression of the joint prior distribution of the parameters at stake and illustrate your suitable choice for the hyperparameters.

We can assume the parameter being independent each other and therefore write the joint prior distribution as the product of the marginal distributions:

$$\pi(\alpha,\beta,\gamma,\tau^2) = \pi(\alpha)\pi(\beta)\pi(\gamma)\pi(\tau^2) \propto \frac{\tau^{2(-a-1)}}{\sigma_\alpha\sigma_\beta}\exp\left(-\frac{\alpha^2}{2\sigma_\alpha} - \frac{\beta^2}{2\sigma_\beta} - \frac{b}{\tau^2}\right) 1_{[1,\infty]}(\alpha)\, 1_{[1,\infty]}(\beta)\, 1_{[0,1]}(\gamma)\, 1_{[0,\infty]}(\tau)$$

The values we choose for the hyperparameters are the following:

- Since we have no prior information on the parameter α we impose an high value for its variance: $\sigma_\alpha = 1$
- Since we have no prior information on the parameter β we impose an high value for its variance: $\sigma_\beta = 1$
- Since τ^2 describes the length variance, assuming $\mu_i = c \,\forall\, i$, we impose $\mathbb{E}[\tau^2] = \mathbb{V}ar[Y_i]$ and $\mathbb{V}ar[\tau^2] = 1$ for the same reason above. Given these constraints we get $a = 2.00569, b = 0.07586$

1.d/e

Derive the functional form (up to proportionality constants) of all *full-conditionals*. Which distribution can you recognize within standard parametric families so that direct simulation from full conditional can be easily implemented ?

- $\pi(\alpha|\beta,\gamma,\tau^2,y^{obs}),\ \alpha \in [1,+\infty]$

$$\begin{aligned}\pi(\alpha|\beta,\gamma,\tau^2,y^{obs}) &= \pi(\alpha)L_{y^{obs}}(\alpha,\beta,\gamma,\tau^2)\\ &\propto \exp\left(-\frac{\alpha^2}{2\sigma_\alpha}\right)\exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}(y_i-\alpha+\beta\gamma^{x_i})^2\right)\\ &= \exp\left(-\frac{\alpha^2}{2\sigma_\alpha}\right)\exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}\left(y_i^2 - 2y_i(\alpha-\beta\gamma^{x_i}) + (\alpha-\beta\gamma^{x_i})^2\right)\right)\\ &\propto \exp\left(-\frac{\alpha^2}{2\sigma_\alpha}\right)\exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}\left(-2y_i\alpha + \alpha^2 - 2\alpha\beta\gamma^{x_i}\right)\right)\\ &= \exp\left(-\frac{\alpha^2}{2\sigma_\alpha}\right)\exp\left(-\frac{1}{\tau^2}\left[\left(-\sum_{i=1}^{n}(y_i+\beta\gamma^{x_i})\right)\alpha + \frac{n}{2}\alpha^2\right]\right)\\ &= \exp\left(\alpha\left(\frac{1}{\tau^2}\sum_{i=1}^{n}(y_i+\beta\gamma^{x_i})\right) - \frac{\alpha^2}{2}\left(\frac{n}{\tau^2}+\frac{1}{\sigma_\alpha^2}\right)\right)\end{aligned}$$

Please notice that if the θ's pdf can be expressed as $p(\theta) \sim \exp(-\frac{1}{2}a\theta^2 - b\theta)$ then $\theta \sim N(\mu = \frac{b}{a}, \sigma^2 = \frac{1}{a})$.

Using the relation above we can derive:

$$\pi(\alpha|\beta,\gamma,\tau^2,y^{obs}) \sim N\left(\mu = \frac{b}{a}, \sigma^2 = \frac{1}{a}\right) \quad where \begin{cases} a = \frac{n}{\tau^2} + \frac{1}{\sigma_\alpha^2} \\ b = \frac{1}{\tau^2}\sum_{i=1}^{n}(y_i + \beta\gamma^{x_i}) \end{cases}$$

- $\pi(\beta|\alpha,\gamma,\tau^2,y^{obs}),\ \beta \in [1,+\infty]$

$$\begin{aligned}\pi(\beta|\alpha,\gamma,\tau^2,y^{obs}) &= \pi(\beta)L_{y^{obs}}(\alpha,\beta,\gamma,\tau^2)\\ &\propto \exp\left(-\frac{\beta^2}{2\sigma_\beta}\right)\exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}(y_i-\alpha+\beta\gamma^{x_i})^2\right)\\ &= \exp\left(-\frac{\beta^2}{2\sigma_\beta}\right)\exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}\left(y_i^2 - 2y_i(\alpha-\beta\gamma^{x_i}) + (\alpha-\beta\gamma^{x_i})^2\right)\right)\\ &\propto \exp\left(-\frac{\beta^2}{2\sigma_\beta}\right)\exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}\left(2y_i\beta\gamma^{x_i} - 2\alpha\beta\gamma^{x_i} + \beta^2\gamma^{2x_i}\right)\right)\\ &= \exp\left(-\frac{\beta^2}{2\sigma_\beta}\right)\exp\left(-\frac{1}{\tau^2}\left[\left(-\sum_{i=1}^{n}(y_i\gamma^{x_i} - \alpha\gamma^{x_i})\right)\beta + \frac{\gamma^{2x_i}}{2}\beta^2\right]\right)\\ &= \exp\left(\beta\left(\frac{1}{\tau^2}\sum_{i=1}^{n}(\alpha\gamma^{x_i} - y_i\gamma^{x_i})\right) - \frac{\beta^2}{2}\left(\frac{1}{\tau^2}\sum_{i=1}^{n}\gamma^{2x_i} + \frac{1}{\sigma_\beta^2}\right)\right)\end{aligned}$$

Using the relation above we can derive:

$$\pi(\beta|\alpha,\gamma,\tau^2,y^{obs}) \sim N\left(\mu = \frac{b}{a}, \sigma^2 = \frac{1}{a}\right) \quad where \begin{cases} a = \frac{1}{\tau^2}\sum_{i=1}^{n}\gamma^{2x_i} + \frac{1}{\sigma_\beta^2} \\ b = \frac{1}{\tau^2}\sum_{i=1}^{n}(\alpha\gamma^{x_i} - y_i\gamma^{x_i}) \end{cases}$$

Please notice that, due to their support constraints, both α, β follow a **truncated normal distribution**.

- $\pi(\gamma|\alpha,\beta,\tau^2,y^{obs})$

$$\begin{aligned}\pi(\gamma|\alpha,\beta,\tau^2,y^{obs}) &\propto \exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}(y_i-\mu_i)^2\right)1_{[0,1]}(\gamma)\\ &= \exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}(y_i-(\alpha-\beta\gamma^{x_i}))^2\right)1_{[0,1]}(\gamma)\\ &\propto \exp\left(-\frac{1}{2\tau^2}\sum_{i=1}^{n}(2y_i\beta\gamma^{x_i} - 2\alpha\beta\gamma^{x_i} + \beta^2\gamma^{2x_i})\right)1_{[0,1]}(\gamma)\end{aligned}$$

Differently from the other parameters the γ full-conditional doesn't follow a standard distribution, so we have to use an algorithm like Metropolis Hasting to draw values from it.

- $\pi(\tau^2|\alpha,\beta,\gamma,y^{obs})$

$$\begin{aligned}\pi(\tau^2|\alpha,\beta,\gamma,y^{obs}) &\propto (\tau^2)^{-a-1}\exp\left(-\frac{b}{\tau^2}\right)\tau^{-2\frac{n}{2}}\exp\left(-\frac{1}{2\tau^2}\left(\sum_{i=1}^{n}(y_i-\alpha+\beta\gamma^{x_i})^2\right)\right)\\ &= (\tau^2)^{-a-\frac{n}{2}-1}\exp\left(-\frac{1}{2\tau^2}\left(2b+\sum_{i=1}^{n}(y_i-\alpha+\beta\gamma^{x_i})^2\right)\right)\end{aligned}$$

From this we can clearly see that the τ^2 full-conditional has an **Inverse Gamma** shape:

$$\pi(\tau^2|\alpha,\beta,\gamma,y^{obs}) \sim IG\left(a+\frac{n}{2}, b+\frac{1}{2}\sum_{i=1}^{n}(y_i-\alpha+\beta\gamma^{x_i})^2\right)$$

1.f

Using a suitable Metropolis-within-Gibbs algorithm simulate a Markov chain ($T = 10000$) to approximate the posterior distribution for the above model

Gibbs Sampling

The Gibbs sampling is an algorithm which allows us to simulate a sequence of dependent random variables starting from all the full conditionals. The algorithm:

1. fix the starting values of the parameter components at time $t = 0$:

$$\boldsymbol{\theta}^0 = (\theta_1, \theta, \ldots, \theta_k)$$

2. for $t = 1, \ldots, T$ iterate the following cycle:

$$\theta_j^{t+1} \sim \pi(\theta_j, \theta_{(j)}) = \pi(\theta_j|\theta_1^{t+1}, \ldots, \theta_{j-1}^{t+1}, \theta_{j+1}^{t}, \ldots, \theta_k^t) \quad j = 1, \ldots, k$$

Metropolis-Hastings

Metropolis-Hastings is an algorithm which allows us to simulate a sequence of dependent random variables drawn from a proposal distribution $p_x(y)$, where x is the current state of the chain. Unlike the Gibbs sampler which relies on conditional distribution, the Metropolis-Hastings algorithm uses the joint distribution to generate a candidate draws. The algorithm works as follow:

1. Draw a candidate $Y_{t+1} = y \sim p_x(y)$ 2 Decide whether or not the candidate is accepted as the next state of the chain at time $t+1$

$$X_{t+1} = \begin{cases} y & \text{with probability } \alpha(x,y) \\ x & \text{with probability } 1-\alpha(x,y) \end{cases}$$

where

$$\alpha(x,y) = \min\left\{\frac{\pi(y)p_y(x)}{\pi(x)p_x(y)}, 1\right\}$$

Metropolis-within-Gibbs

Gibbs sampling requires the knowledge of the full conditionals to work, unfortunately, since γ does not follow a well-known parametric distribution, we're lacking this information. To overcome this issue we can use the Metropolis-within-Gibbs algorithm which allows us to replace the original i-th kernel K_i with one related to the Metropolis algorithm $K_i^{MET_i}$.

To perform such simulations we'll use the Jags library.

```
parameters <- c("alpha", "beta", "gamma", "tau2",
                "Ypred_20", "cond_exp_20", "Ypred_30", "cond_exp_30")

dugongjags <- jags(data=mydata,
                   # let JAGS choose the initial values
                   parameters.to.save=parameters,
                   model.file="dugong_jags_model.txt",
                   DIC = F,
                   n.chains=1,
                   n.burnin=1,
                   n.iter=10000)
```

```
## module glm loaded
```

```
## module dic loaded
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 27
##    Unobserved stochastic nodes: 6
##    Total graph size: 134
##
## Initializing model
```

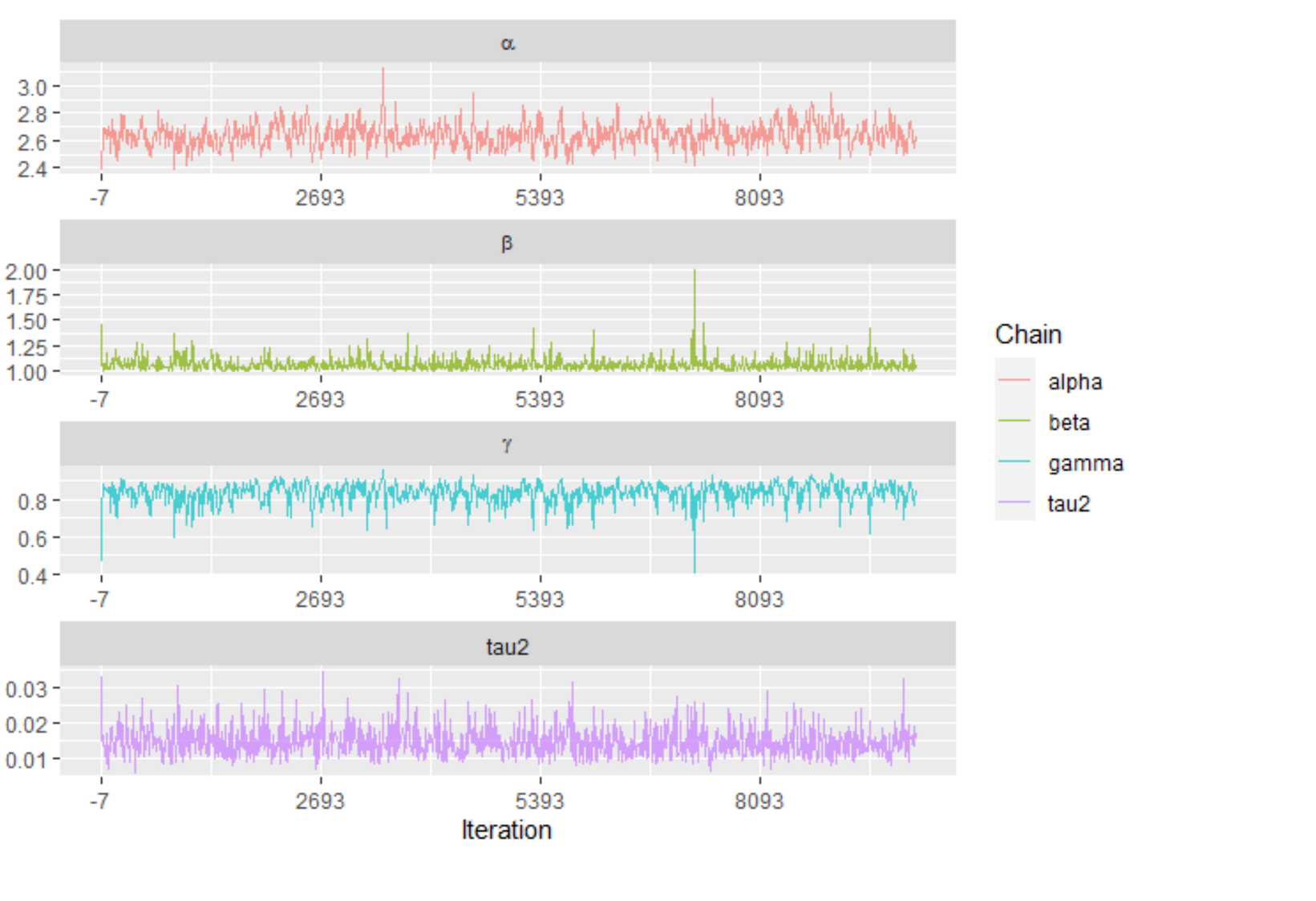
```
print(dugongjags)
```

```
## Inference for Bugs model at "dugong_jags_model.txt", fit using jags,
##  1 chains, each with 10000 iterations (first 1 discarded), n.thin = 9
##  n.sims = 1111 iterations saved
##             mu.vect sd.vect  2.5%   25%   50%   75% 97.5%
## Ypred_20      2.592   0.135 2.301 2.507 2.594 2.678 2.853
## Ypred_30      2.620   0.142 2.345 2.527 2.622 2.714 2.893
## alpha         2.642   0.087 2.479 2.584 2.637 2.697 2.822
## beta          1.065   0.067 1.002 1.023 1.049 1.086 1.230
## cond_exp_20   2.589   0.047 2.478 2.564 2.594 2.622 2.669
## cond_exp_30   2.627   0.070 2.479 2.582 2.629 2.678 2.756
## gamma         0.838   0.055 0.709 0.814 0.848 0.876 0.914
## tau2          0.015   0.004 0.009 0.012 0.014 0.017 0.025
```

1.g

Show the 4 univariate trace-plots of the simulations of each parameter

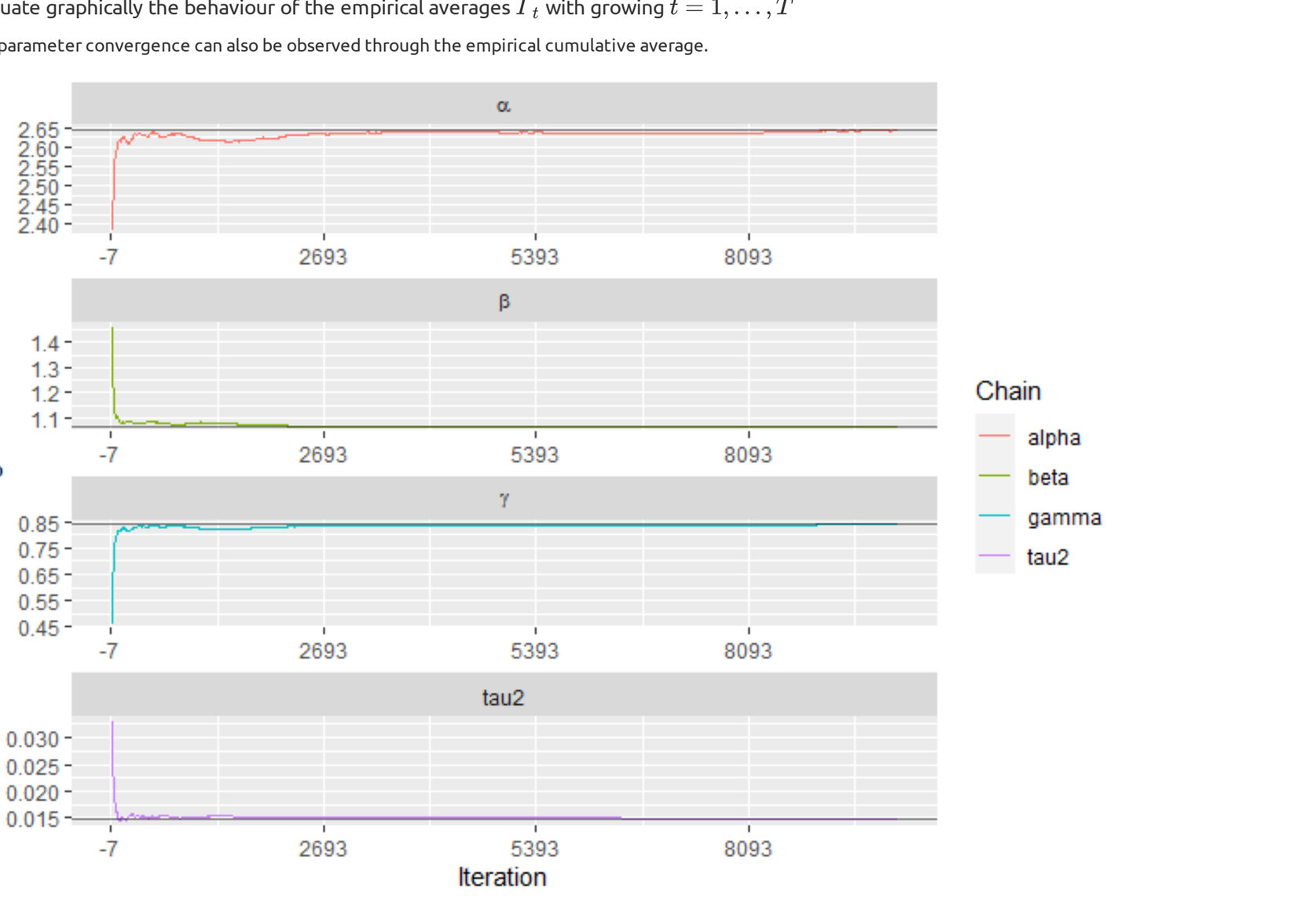
As we can see from the trace-plots below all the parameters converge to some stationary distribution.



1.h

Evaluate graphically the behaviour of the empirical averages $\hat{I}_t$ with growing $t = 1, \ldots, T$

The parameter convergence can also be observed through the empirical cumulative average.



1.i

Provide estimates for each parameter together with the approximation error and explain how you have evaluated such error

The estimates for each parameter can be easily computed using the empirical mean.

To evaluate the approximation error we have to take into account the temporal dependency between the samples: contrary to standard MC, where the samples are iid, in a MCMC setting the current state depends on the previous one. Therefore we have to compute the MCMC variance as the MC variance divided by a penalisation term that depends on the correlation between the samples:

$$\mathbb{V}ar_{MCMC}[\hat{I}] = \frac{\mathbb{V}ar[\hat{I}]}{S_{eff}}$$

where S_{eff} measures the effective sample size. This quantity can be interpreted as the number of independent Monte Carlo samples necessary to carry the same amount of information.

Parameters	Estimates	Approximation_Error	Variance	ESS
alpha	2.6424961	0.0047367	0.0076045	338.9453
beta	1.0645952	0.0027528	0.0044558	587.9958
gamma	0.8387293	0.0026113	0.0030258	443.7499
tau2	0.0147211	0.0001357	0.0000180	976.8022

1.l

Which parameter has the largest posterior uncertainty? How did you measure it?

To measure the posterior uncertainty we can use the equal-tails 95% confidence interval.

	lower_limit	upper_limit	length
alpha	2.4791811	2.8224715	0.3432904
beta	1.0020391	1.2295500	0.2275109
gamma	0.7087873	0.9135787	0.2047914
tau2	0.0087796	0.0249764	0.0161968

As we can see the parameter with the highest uncertainty (the parameter whose 95% confidence interval length is the largest) is α.

1.m

Which couple of parameters has the largest correlation (in absolute value)?

From the correlation matrix below we can see that the most correlated couple is (α, γ).

	alpha	beta	gamma	tau
alpha	1.0000000	-0.1148434	0.8734139	-0.1863524
beta	-0.1148434	1.0000000	-0.4312648	0.2778105
gamma	0.8734139	-0.4312648	1.0000000	-0.2769318
tau	-0.1863524	0.2778105	-0.2769318	1.0000000

1.n/o/p

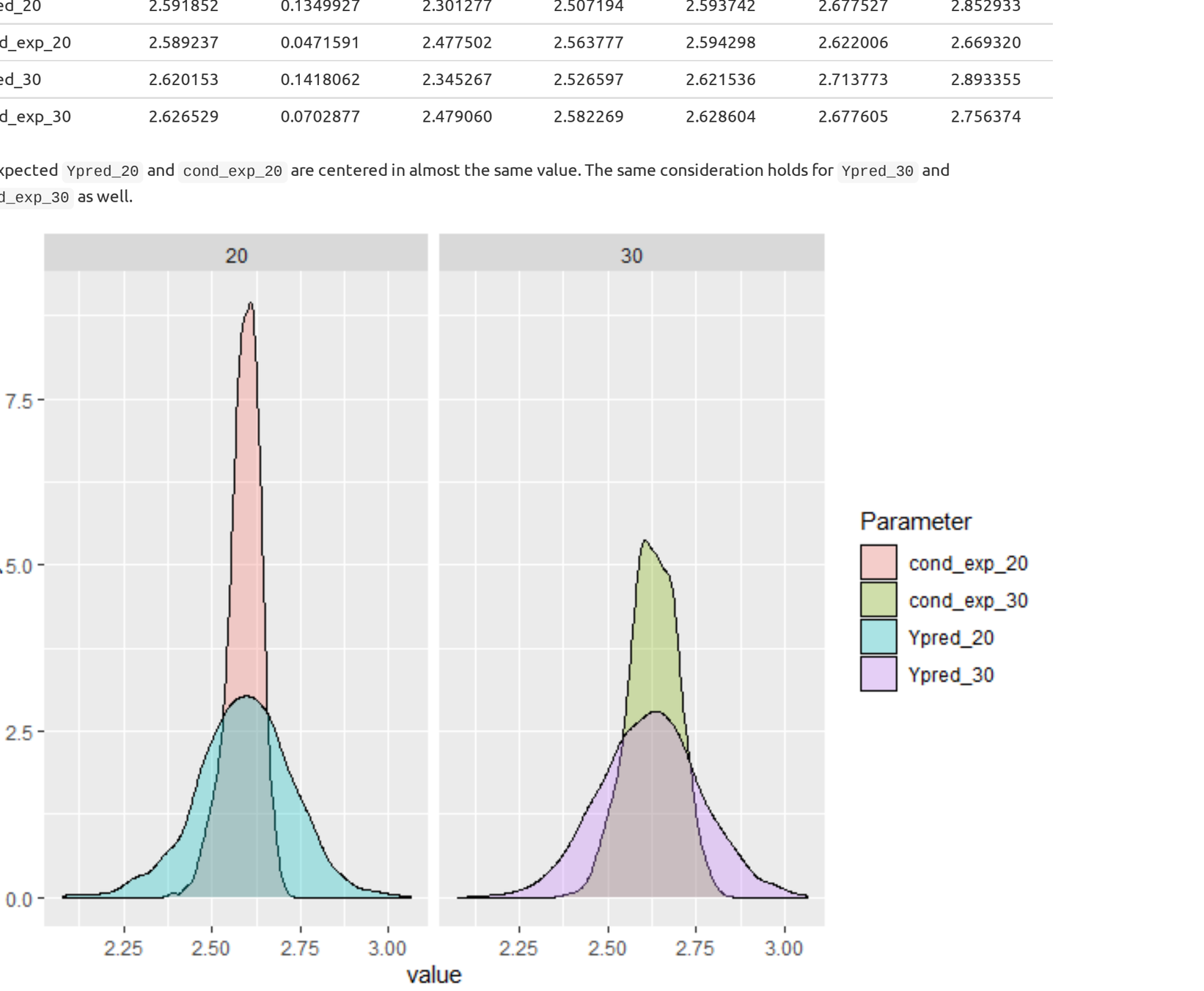
Use the Markov chain to approximate the posterior predictive distribution of the length of a dugong with age of 20 years. Provide the prediction of a different dugong with age 30. Which prediction is less precise?

In order to perform such predictions we ask Jags to provide us samples drawn from the following distributions:

- `Ypred_20`, the predicted dugong length, normally distributed, centered in `cond_exp_20`
- `cond_exp_20`, the conditional expectation regressed for $x = 20$
- `Ypred_30`, the predicted dugong length, normally distributed, centered in `cond_exp_30`
- `cond_exp_30`, the conditional expectation regressed for $x = 30$

	mean	sd	2.5%	25%	50%	75%	97.5%
Ypred_20	2.591852	0.1349927	2.301277	2.507194	2.593742	2.677527	2.852933
cond_exp_20	2.589237	0.0471591	2.477502	2.563777	2.594298	2.622006	2.669320
Ypred_30	2.620153	0.1418062	2.345267	2.526597	2.621536	2.713773	2.893355
cond_exp_30	2.626529	0.0702877	2.479060	2.582269	2.628604	2.677605	2.756374

As expected `Ypred_20` and `cond_exp_20` are centered in almost the same value. The same consideration holds for `Ypred_30` and `cond_exp_30` as well.

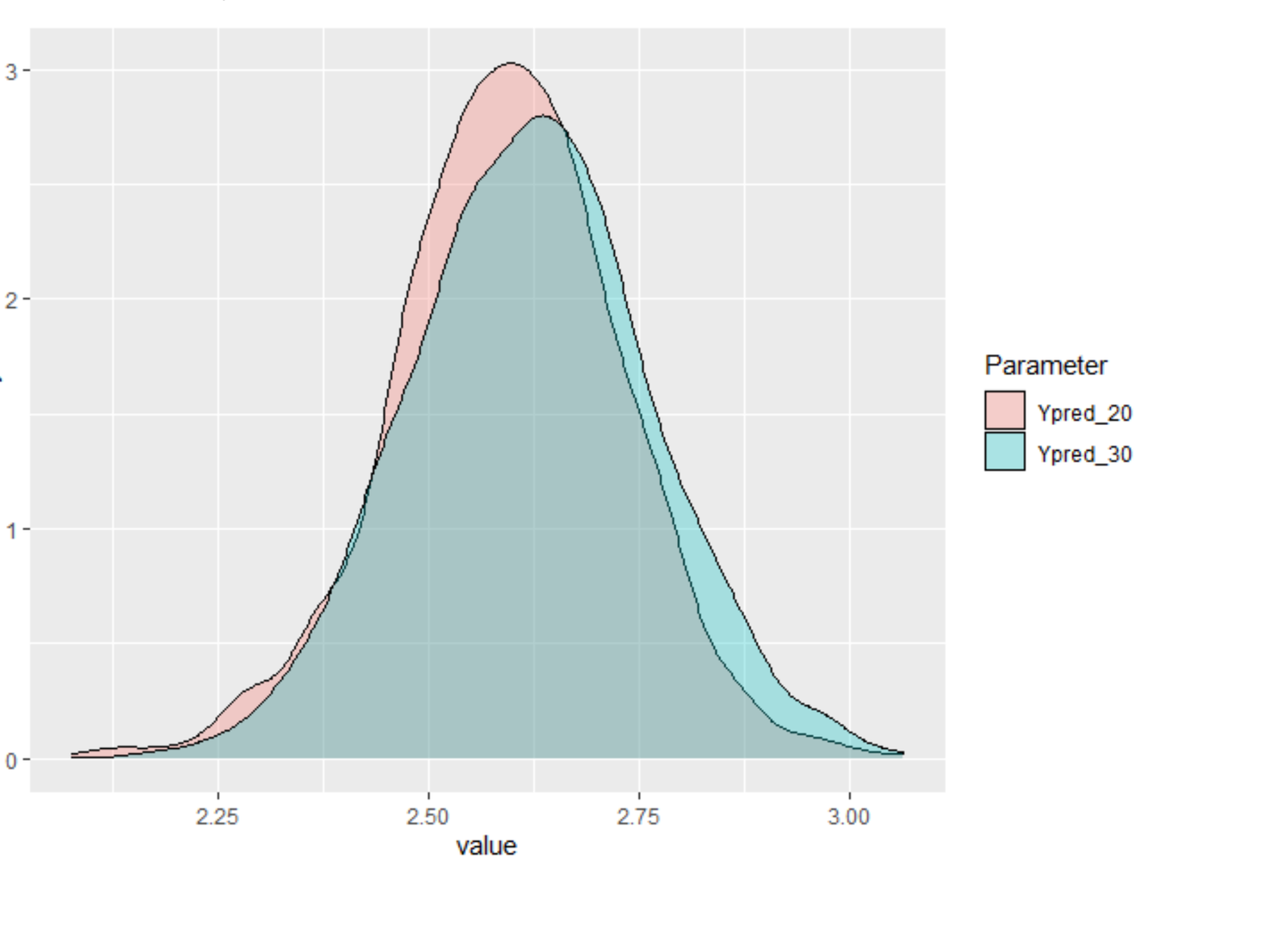


In order to compare the precision of the two predictions we can compare the length of 95% confidence interval as we did before.

	length
Ypred_20	0.5516561
Ypred_30	0.5480885

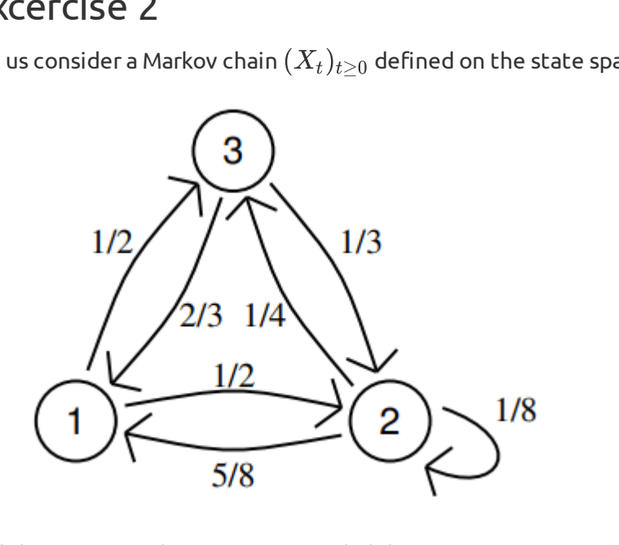
From the table above we see that the prediction `Ypred_20` is slightly less precise than the one for `Ypred_30`.

Below we can see how the two predictions are distributed.



Excercise 2

Let us consider a Markov chain $(X_t)_{t\geq 0}$ defined on the state space $\mathcal{S} = \{1,2,3\}$ with the following transition



and the corresponding transition probability matrix:

$$P = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{5}{8} & \frac{1}{8} & \frac{1}{4} \\ \frac{2}{3} & \frac{2}{3} & 0 \end{pmatrix}$$

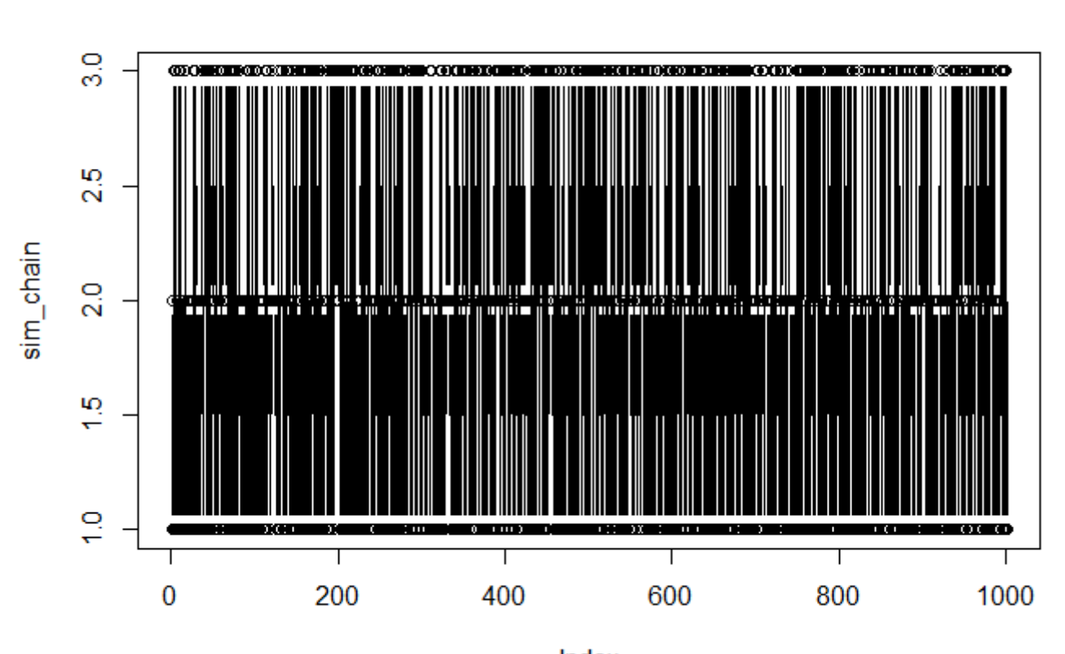
2.a

Starting at time $t = 0$ in the state $X_0 = 1$ simulate the Markov chain with distribution assigned as above for $t = 1000$ consecutive times

```
# State space
S = 1:3

# Transition probability matrix
P <- matrix(c(0, 0.5, 0.5,
              5/8, 1/8, 1/4,
              2/3, 1/3, 0),
            nrow = 3,
            byrow = T)

P
```

```
##           [,1]      [,2] [,3]
## [1,] 0.0000000 0.5000000 0.50
## [2,] 0.6250000 0.1250000 0.25
## [3,] 0.6666667 0.3333333 0.00
```

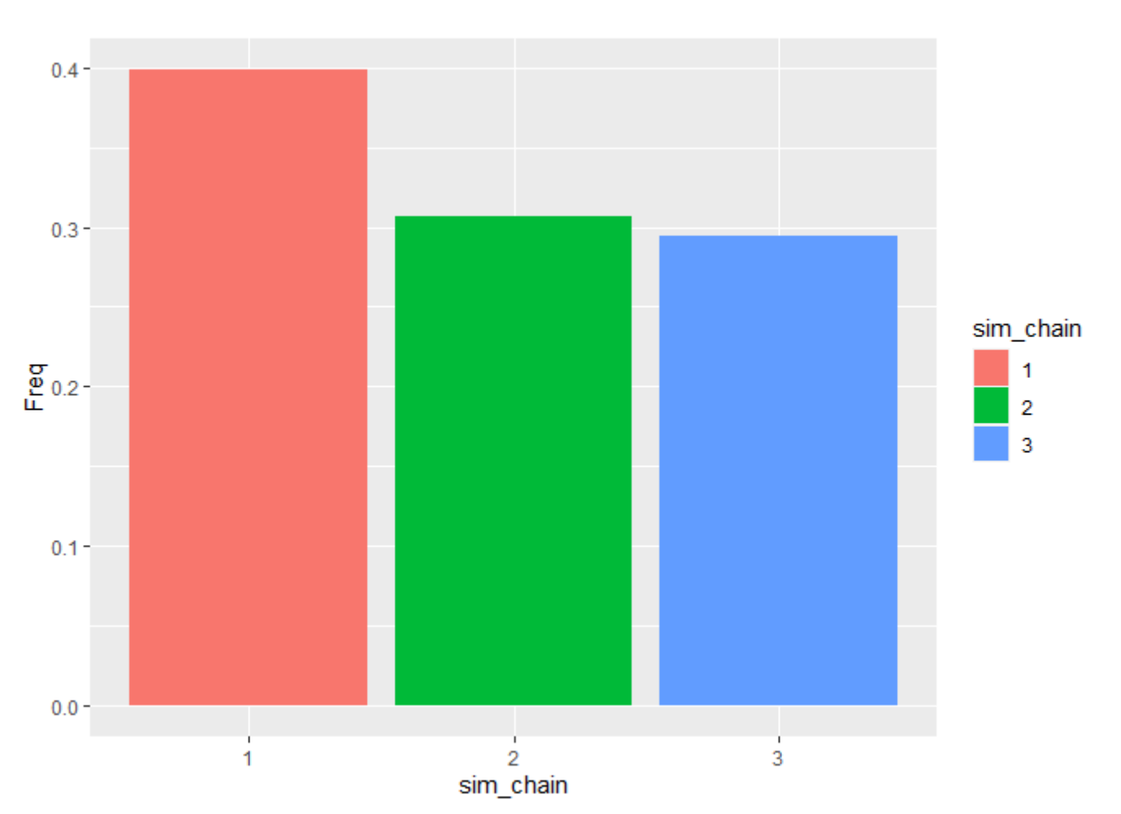
```
set.seed(1234)

sim_Markov_chain <- function(space, P, x0, N_steps){
  # Vector to store the simulated values
  sim_chain <- rep(NA, N_steps+1)
  sim_chain[1] <- x0

  # Simulation
  for(t in 1:N_steps){
    sim_chain[t+1] <- sample(space, size=1, prob=P[sim_chain[t], ])
  }

  return(sim_chain)
}

# Number of steps
N_steps <- 1000

# Initial state
x_0 <- 1

# Simulated chain
sim_chain <- sim_Markov_chain(S, P, x_0, N_steps)

plot(sim_chain, type = 'b')
```



2.b

compute the empirical relative frequency of the two states in your simulation

sim_chain	Freq
1	0.3986014
2	0.3066933
3	0.2947053

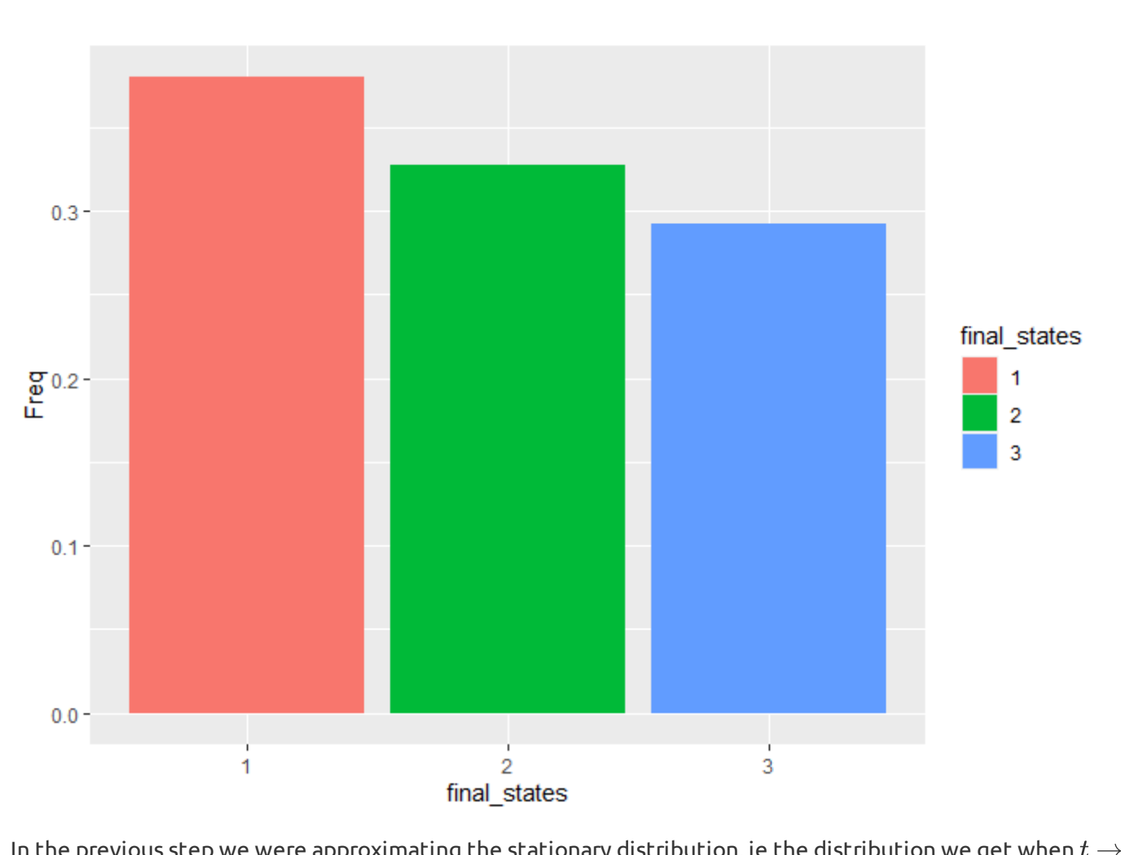


2.c

repeat the simulation for 500 times and record only the final state at time $t = 1000$ for each of the 500 simulated chains. Compute the relative frequency of the 500 final states. What distribution are you approximating in this way? Try to formalize the difference between this point and the previous point.

```
M <- 500
t <- 1000

# Simulate M chains
final_states <- sapply(1:M, function(x) sim_Markov_chain(S, P, x_0, N_steps)[t + 1])
```

final_states	Freq
1	0.380
2	0.328
3	0.292



In the previous step we were approximating the stationary distribution, ie the distribution we get when $t \to +\infty$, while in this step we're approximating the distribution for a fixed $t = 1000$.

2.d/e

compute the theoretical stationary distribution π and explain how you have obtained it. Is it well approximated by the simulated empirical relative frequencies computed in (b) and (c)?

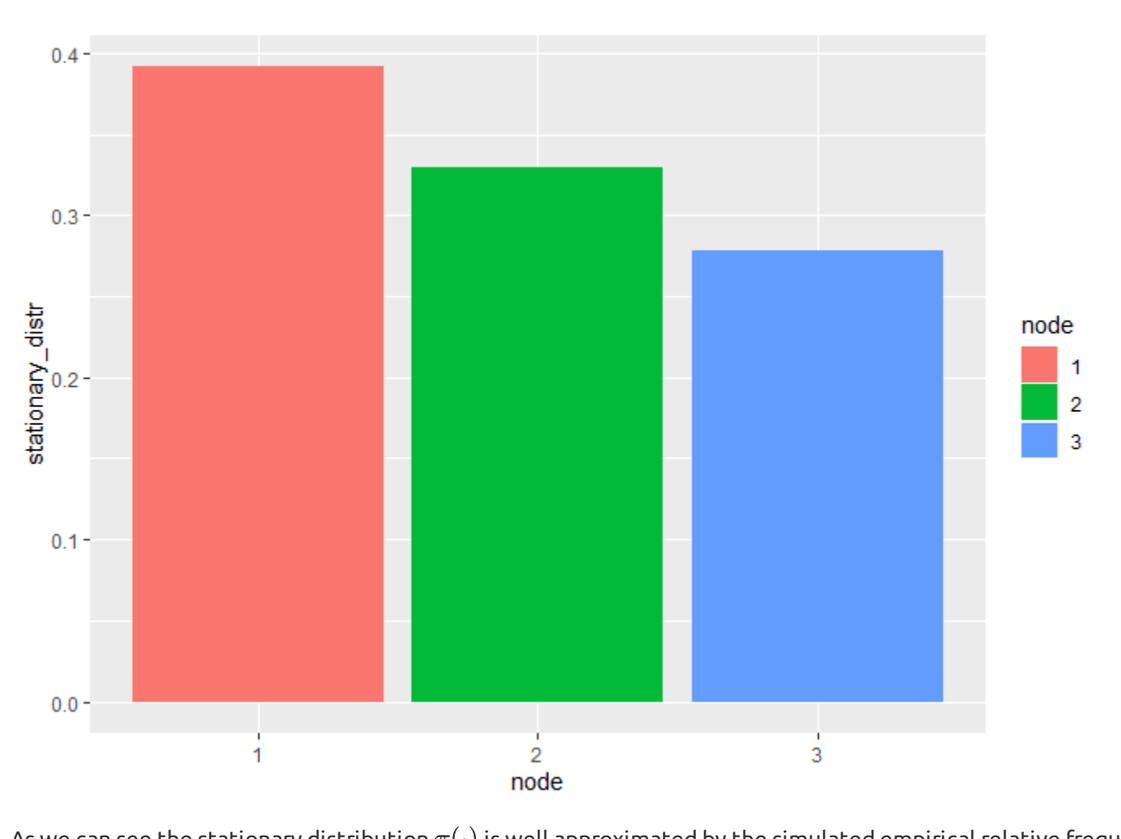
Remembering that the stationary distribution $\pi(\cdot)$ satisfy the following conditions:

- $\pi_j \geq 0 \; \forall \; j \in S$
- $\sum_{j \in S} \pi_j = 1$
- $\pi P = \pi$

we can compute it as the normalized eigenvector relative to the eigenvalue 1.

```
stationary_vect <- eigen(t(P))$vectors[, 1]
stationary_distr <- stationary_vect / sum(stationary_vect)
```

node	stationary_distr
1	0.3917526
2	0.3298969
3	0.2783505



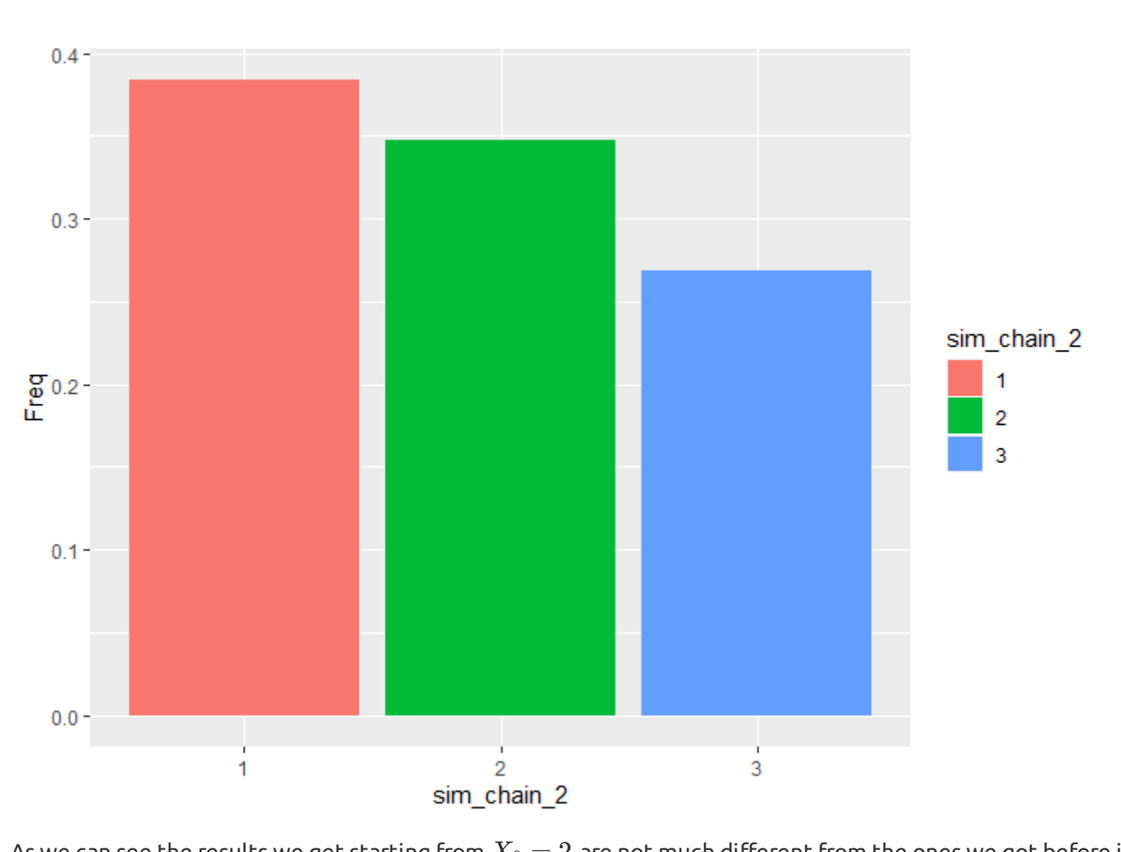
As we can see the stationary distribution $\pi(\cdot)$ is well approximated by the simulated empirical relative frequencies computed in (b) and (c).

2.f

what happens if we start at $t = 0$ from state $X_0 = 2$ instead of $X_0 = 1$?

```
# Initial state
x_0 <- 2

# Simulated chain
sim_chain_2 <- sim_Markov_chain(S, P, x_0, N_steps)
```

sim_chain_2	Freq
1	0.3836164
2	0.3476523
3	0.2687313



As we can see the results we get starting from $X_0 = 2$ are not much different from the ones we got before in the case of $X_0 = 1$. The reason for this is because as $t \to +\infty$ the empirical relative frequencies distribution converges to the stationary distribution.