

Monte Carlo Assignment 6 and Final

Kylie Taylor

5/9/2019

Introduction

We have two models

Model 1 (M1):

$$y_i = \log(1 + \exp(\beta_0 + \beta_1 x_{1i})) + \sigma \epsilon_i$$

Model 2 (M2):

$$y_i = \log(1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})) + \sigma \epsilon_i$$

The data shows up as (y_i, x_{1i}, x_{2i}) , each having $n=50$ observations. The error term, ϵ_i is distributed i.i.d. standard normal, $N(0,1)$.

We are given the priors for each model as $P(M1) = P(M2) = \frac{1}{2}$. Each β term $(\beta_0, \beta_1, \beta_2)$ have priors

$$f(\beta_0) = f(\beta_1) = f(\beta_2) = N(0, 5^2)$$

The goal of this assignment is to find $P(M1|y_i, x_{1i}, x_{2i})$, thus $P(M2|y_i, x_{1i}, x_{2i}) = 1 - P(M1|y_i, x_{1i}, x_{2i})$ using two Monte Carlo methods

- 1) Bayes Factor to compare models
- 2) Metropolis Hastings algorithm to jump between models

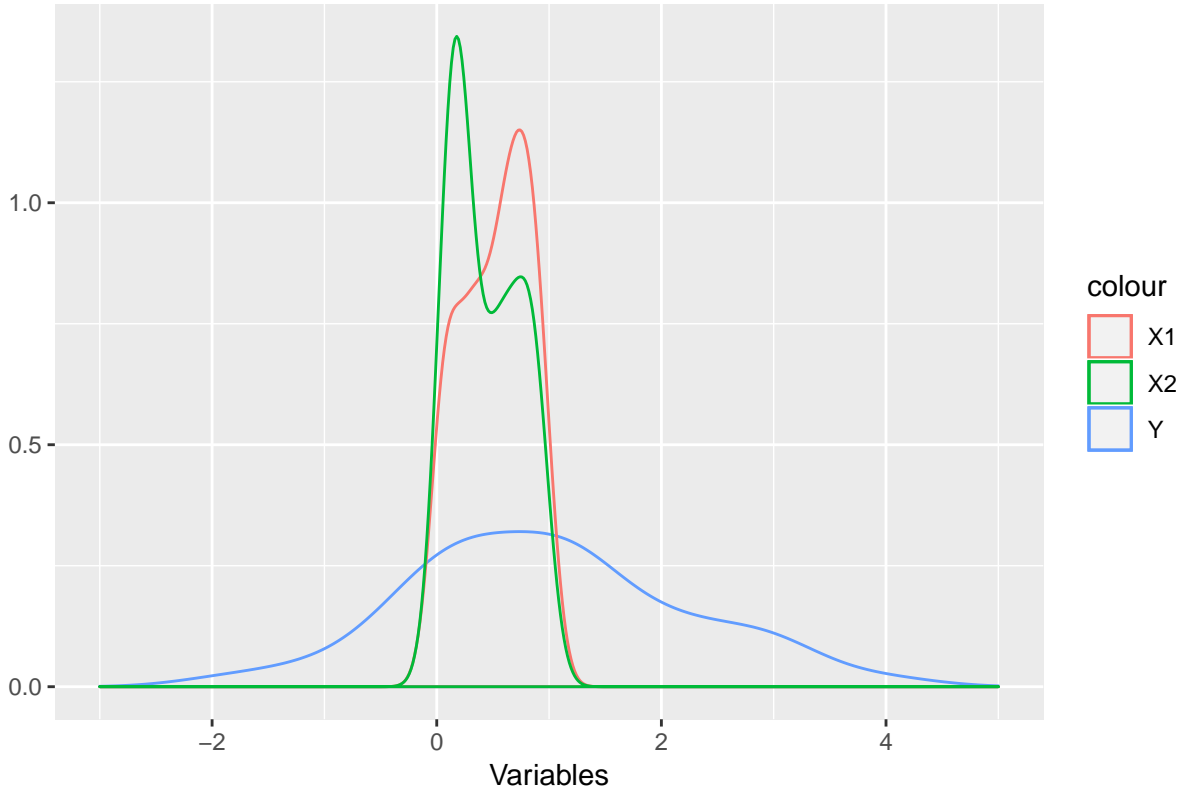
After accomplishing that, our goal is to find the predictive density for somebody with $x_{1,51} = 0.23$ and $x_{2,51} = -1.2$.

Data

The first step I will take is to inspect the data. “V1” corresponds to y_i values, “V2” corresponds to x_{1i} values, and “V3” corresponds to x_{2i} values. y_i has mean 0.986, minimum value -1.8619, and max value 4.0021. x_{1i} has mean 0.52240, minimum value 0.03637, and max value 0.98170. x_{2i} has mean 0.435406, minimum value 0.004233, and max value 0.998834.

I plotted the density of the three variables in one plot below. We see that x_{1i} and x_{2i} have much smaller variance than y_i .

Histograms of Y, X1 and X2



Conditional Posterior Densities

My next step will be to find the conditional posterior densities of the parameters. The parameters in these models are $(\beta_0, \beta_1, \lambda)$ for model 1, and $(\beta_0, \beta_1, \beta_2, \lambda)$ for model 2.

For model 1, the joint density is

$$f(\beta_0, \beta_1, \lambda, y_i | x_{1i}) = f(\lambda) f(\beta_0) f(\beta_1) \prod_i k(y_i | \beta_0, \beta_1, \lambda)$$

The prior for λ will be given as

$$f(\lambda) = Ga\left(\frac{1}{2}, \frac{1}{2}\right)$$

.

The likelihood is

$$\prod_i N(y_i | \log(1 + \exp(\beta_0 + \beta_1 x_{1i})), \sqrt{\frac{1}{\lambda}})$$

The posterior for β_0 is

$$f(\beta_0 | \lambda, \beta_1, y_i) = f(\beta_0) \prod_i k(y_i | \beta_0, \beta_1, \lambda)$$

$$\propto e^{-\frac{\beta_0^2}{50}} e^{-\frac{\lambda n}{2} \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i})))^2}$$

$$\propto e^{-\frac{1}{2} \left(\frac{\beta_0^2}{25} + \lambda n \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i})))^2 \right)}$$

This follows for β_1

$$f(\beta_1|\beta_1, \lambda, y_i) \propto e^{-\frac{1}{2}(\frac{\beta_1^2}{25} + \lambda n \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i})))^2)}$$

The posterior for λ is

$$\begin{aligned} f(\lambda|\beta_0, \beta_1, y_i) &\propto \lambda^{\frac{1}{2}-1} e^{-\frac{1}{2}\lambda \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i})))^2} \\ &\propto \lambda^{\frac{1}{2}-\frac{n}{2}} e^{-\lambda(\frac{1}{2} + \frac{n}{2} \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i})))^2)} \\ f(\lambda|\beta_0, \beta_1, y_i) &= Ga(\frac{1}{2} + \frac{n}{2}, \frac{1}{2} + \frac{1}{2} \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i})))^2) \end{aligned}$$

For model 2, the joint density is

$$f(\beta_0, \beta_1, \beta_2, \lambda, y_i|x_{1i}) = f(\lambda)f(\beta_0)f(\beta_1)f(\beta_2) \prod_i k(y_i|\beta_0, \beta_1, \beta_2, \lambda)$$

The posterior for β_0 is

$$f(\beta_0|\beta_1, \beta_2, \lambda, y_i) \propto e^{-\frac{1}{2}(\frac{\beta_0^2}{25} + \lambda \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})))^2)}$$

For β_1

$$f(\beta_1|\beta_0, \beta_2, \lambda, y_i) \propto e^{-\frac{1}{2}(\frac{\beta_1^2}{25} + \lambda \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})))^2)}$$

For β_2

$$f(\beta_2|\beta_0, \beta_1, \lambda, y_i) \propto e^{-\frac{1}{2}(\frac{\beta_2^2}{25} + \lambda \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})))^2)}$$

For λ

$$f(\lambda|\beta_0, \beta_1, \beta_2, y_i) = Ga(\frac{1}{2} + \frac{n}{2}, \frac{1}{2} + \frac{1}{2} \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})))^2)$$

Posterior Estimations

In order to estimate every β term in both model 1 and 2, I must introduce a Metropolis Hastings step. To sample some β_k for $k = (0,1,2)$, I will implement the following MH step

```
given current state b_n
take proposal, b' ~ q
take u ~ U(0,1)
if u < a(b,b')
    b_{n+1} = b'
else
    b_{n+1} = b_n
```

I will use a normal proposal density, and the

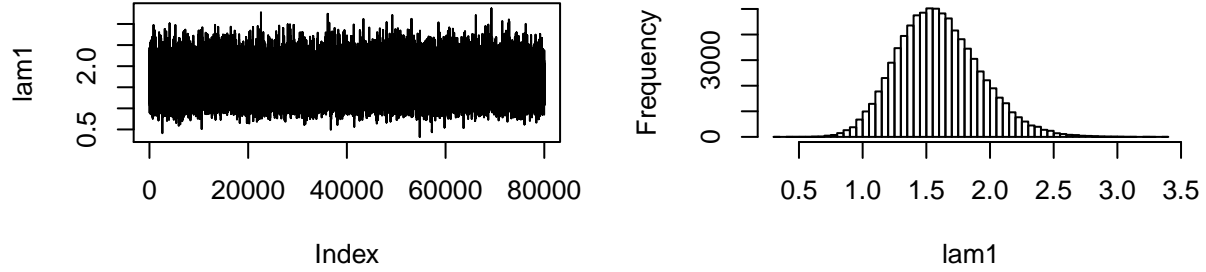
$$\alpha(\beta_k, \beta'_k) = \min(1, \frac{f(\beta'_k | \beta_{-k}, \lambda, y_i) * q(\beta_k | \beta'_k, e)}{f(\beta_k | \beta_{-k}, \lambda, y_i) * q(\beta'_k | \beta_k, e)})$$

$$\propto \min(1, \frac{e^{-\frac{1}{2}(\frac{\beta'^2_k}{25} + \lambda \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})))^2)}}{e^{-\frac{1}{2}(\frac{\beta^2_k}{25} + \lambda \sum_i (y_i - \log(1 + \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i})))^2)}})$$

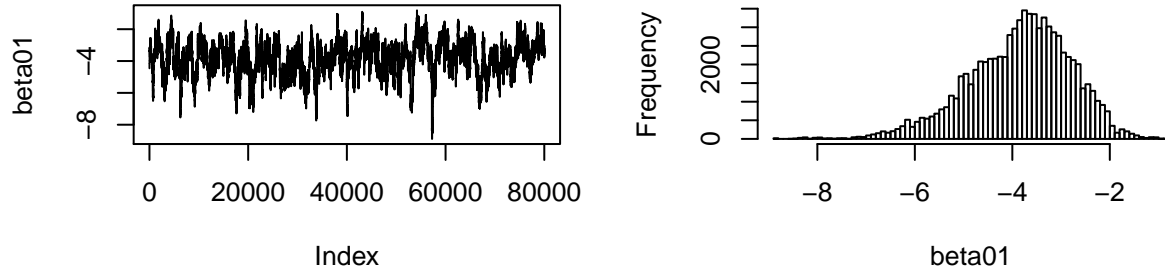
In order to sample the λ I can use a Gibbs Framework. I iterate through 100,000 samples of the parameters with a 20,000 burnin, resulting in 80,000 final samples of all the parameters.

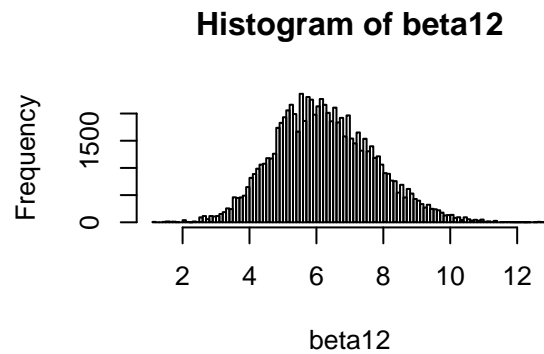
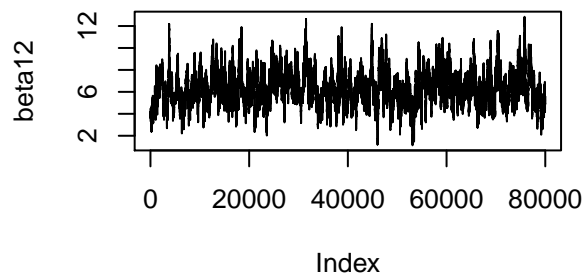
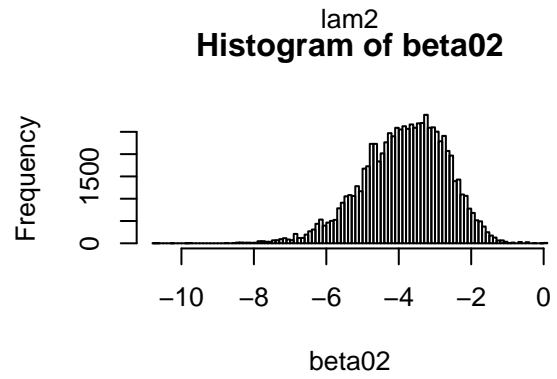
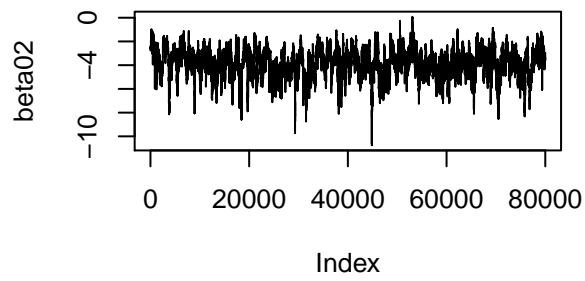
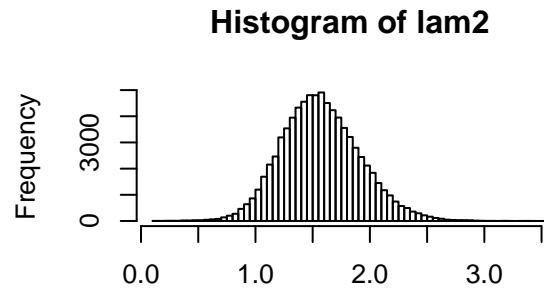
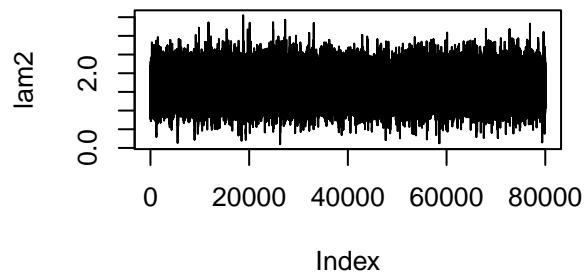
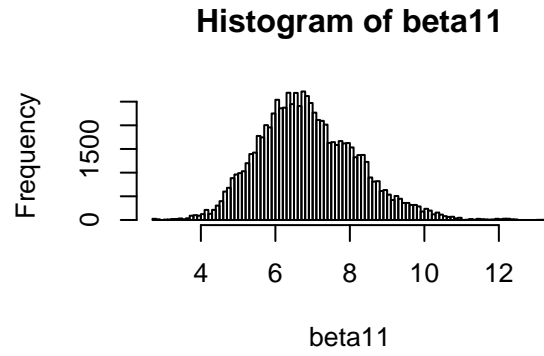
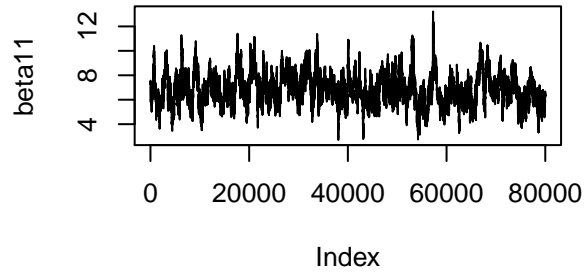
Trace plots and histograms of all the parameters are included below. The trace plots below reveal that the Metropolis Hastings step for sampling the β 's and Gibbs step for sampling the λ appear to move well and have an appropriate proposal variance. The histograms reveal there are no out-of-ordinary distributions for the parameters, and are centered around some high density point, likely the mean.

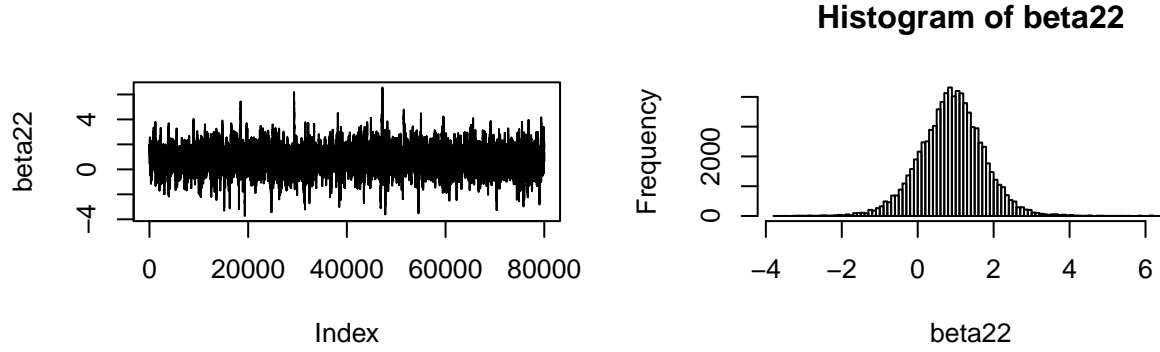
Histogram of lam1



Histogram of beta01







Bayes Factor

Remember that the posterior distribution for some θ is $f(\theta|data) \propto f(\theta)k(data|\theta)$. The goal in Bayes Factor is to estimate $f(\hat{\theta}|data)$ given the posterior samples $(\theta^{(1)}, \dots, \theta^{(m)})$, which I obtained above using a Gibbs sampler for λ and Metropolis Hastings framework for $(\beta_0, \beta_1, \beta_2)$.

Using the estimates from the posteriors for Model 1 and Model 2, I will calculate the Bayes Factor to find $P(M1|data)$. Recall that the equation for the posterior odds is

$$\frac{P(M1|y_i, x_{1i}, x_{2i})}{P(M2|y_i, x_{1i}, x_{2i})} = \frac{K(y_i, x_{1i}, x_{2i}|M1)}{K(y_i, x_{1i}, x_{2i}|M2)} * \frac{P(M1)}{P(M2)}$$

Again, the data shows up as (y_i, x_{1i}, x_{2i}) . The posterior odds is

$$\frac{P(M1|y_i, x_{1i}, x_{2i})}{P(M2|y_i, x_{1i}, x_{2i})}$$

The Bayes Factor is

$$\frac{K(y_i, x_{1i}, x_{2i}|M1)}{K(y_i, x_{1i}, x_{2i}|M2)}$$

And the Prior Odds are

$$\frac{P(M1)}{P(M2)} = \frac{\frac{1}{2}}{\frac{1}{2}} = 1$$

This term will drop away from the equation, also $P(M2|data) = 1 - P(M1|data)$, giving us the Posterior Odds equal to the Bayes Factor.

$$\frac{P(M1|y_i, x_{1i}, x_{2i})}{1 - P(M1|y_i, x_{1i}, x_{2i})} = \frac{K(y_i, x_{1i}, x_{2i}|M1)}{K(y_i, x_{1i}, x_{2i}|M2)}$$

The Bayes Factor is calculated using the following equations. The integral $I = K(y_i, x_{1i}, x_{2i}|M_i)$ is known as the marginal likelihood for a given model.

For model 1:

$$I = K(y_i, x_{1i}, x_{2i}|M1) = \int f(\lambda)f(\beta_0)f(\beta_1) \prod_i k(y_i|\beta_0, \beta_1, \lambda)d\lambda d\beta_0 d\beta_1$$

Since the marginal likelihood is the normalizing constant of the posterior density, we can write the following equation, which is referred to as the basic marginal likelihood identity.

$$K(y_i, x_{1i}, x_{2i}|M1) = \frac{f(\tilde{\lambda})f(\tilde{\beta}_0)f(\tilde{\beta}_1)k(y_i|\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\lambda})}{f(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\lambda}, y_i|x_{1i}, x_{2i})}$$

Taking the log of the marginal likelihood identity and evaluating it at some $\tilde{\theta}$, returns

$$\log(K(y_i, x_{1i}, x_{2i}|M1)) = \log(f(\tilde{\lambda})) + \log(f(\tilde{\beta}_0)) + \log(f(\tilde{\beta}_1)) + \log(k(y_i|\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\lambda})) - \log(f(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\lambda}, y_i|x_{1i}, x_{2i}))$$

The value $\tilde{\theta}$ is taken as some high density point in the support of the posterior, which I use as the mean of the posterior estimates

The same steps follow natural for model 2:

$$I = K(y_i, x_{1i}, x_{2i}|M2) = \int f(\lambda)f(\beta_0)f(\beta_1)f(\beta_2) \prod_i k(y_i|\beta_0, \beta_1, \beta_2, \lambda) d\lambda d\beta_0 d\beta_1 d\beta_2$$

$$K(y_i, x_{1i}, x_{2i}|M2) = \frac{f(\tilde{\lambda})f(\tilde{\beta}_0)f(\tilde{\beta}_1)f(\tilde{\beta}_2)k(y_i|\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\lambda})}{f(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\lambda}, y_i|x_{1i}, x_{2i})}$$

$$\log(K(y_i, x_{1i}, x_{2i}|M2)) = \log(f(\tilde{\lambda})) + \log(f(\tilde{\beta}_0)) + \log(f(\tilde{\beta}_1)) + \log(f(\tilde{\beta}_2)) + \log(k(y_i|\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\lambda})) - \log(f(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\lambda}, y_i|x_{1i}, x_{2i}))$$

The major issue with finding a value of I is that $f(\tilde{\theta}|data)$ is hard to compute. Therefore I will implement the following equation to get an estimate for $f(\tilde{\theta}|data)$

$$f(\tilde{\theta}|data) = \frac{\int \alpha(\theta, \tilde{\theta}) * q(\tilde{\theta}|\theta) * f(\theta|data) d\theta}{\int \alpha(\tilde{\theta}, \theta) * q(\theta|\tilde{\theta}) d\theta}$$

Where $\tilde{\theta} = (\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\lambda}, \tilde{y}_i)$, $\theta' = (\beta'_0, \beta'_1, \beta'_2, \lambda', y'_i)$ and $data = (y_i, x_{1i}, x_{2i})$.

This is a clever way of approximating $f(\tilde{\theta}|data)$, because the numerator satisfies the reversibility condition. The reversibility condition entails that $p(\theta^{(m)}, \tilde{\theta})f(\theta^{(m)}|data) = f(\tilde{\theta}|data)p(\tilde{\theta}, \theta^{(m)})$, where $p(\theta^{(m)}, \tilde{\theta}) = \alpha(\theta, \tilde{\theta}) * q(\tilde{\theta}|\theta)$. This allows us to use the following estimations for the above integrals.

$$f(\tilde{\theta}|data) = \frac{\frac{1}{M} \sum_m \alpha(\theta^{(m)}, \tilde{\theta}) * q(\tilde{\theta}|\theta^{(m)})}{\frac{1}{M} \sum_m \alpha(\tilde{\theta}, \theta^{(m)})}$$

Here, the $\alpha(\theta^{(m)}, \tilde{\theta})$ is equal to

$$\alpha(\theta^{(m)}, \tilde{\theta}) = \min(1, \frac{f(\tilde{\theta}|data) * q(\tilde{\theta}|\theta^{(m)}, \epsilon)}{f(\theta^{(m)}|data) * q(\theta^{(m)}|\tilde{\theta}, \epsilon)})$$

Plugging in for $f(\theta^{(m)}|data)$ and $f(\tilde{\theta}|data)$ we have the following fraction:

$$\frac{f(\tilde{\theta}|data)}{f(\theta^{(m)}|data)} = \frac{\frac{k(data|\tilde{\theta}) * f(\tilde{\theta})}{f(data)}}{\frac{k(data|\theta^{(m)}) * f(\theta^{(m)})}{f(data)}} = \frac{k(data|\tilde{\theta}) * f(\tilde{\theta})}{k(data|\theta^{(m)}) * f(\theta^{(m)})}$$

In terms of model 1 parameters,

$$\frac{f(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\lambda}, y_i | x_{1i}, x_{2i})}{f(\beta_0^{(m)}, \beta_1^{(m)}, \lambda^{(m)}, y_i | x_{1i}, x_{2i})} = \frac{\prod_i N(y_i | \log(1 + \exp(\tilde{\beta}_0 + \tilde{\beta}_1 x_{1i})), \sqrt{\frac{1}{\tilde{\lambda}}}) f(\tilde{\beta}_0) f(\tilde{\beta}_1) f(\tilde{\lambda})}{\prod_i N(y_i | \log(1 + \exp(\beta_0^{(m)} + \beta_1^{(m)} x_{1i})), \sqrt{\frac{1}{\lambda^{(m)}}}) f(\beta_0^{(m)}) f(\beta_1^{(m)}) f(\lambda^{(m)})}$$

In terms of model 2 parameters,

$$\frac{f(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, \tilde{\lambda}, y_i | x_{1i}, x_{2i})}{f(\beta_0^{(m)}, \beta_1^{(m)}, \beta_2^{(m)}, \lambda^{(m)}, y_i | x_{1i})} = \frac{\prod_i N(y_i | \log(1 + \exp(\tilde{\beta}_0 + \tilde{\beta}_1 x_{1i} + \tilde{\beta}_2 x_{2i})), \sqrt{\frac{1}{\tilde{\lambda}}}) f(\tilde{\beta}_0) f(\tilde{\beta}_1) f(\tilde{\beta}_2) f(\tilde{\lambda})}{\prod_i N(y_i | \log(1 + \exp(\beta_0^{(m)} + \beta_1^{(m)} x_{1i} + \beta_2^{(m)} x_{2i})), \sqrt{\frac{1}{\lambda^{(m)}}}) f(\beta_0^{(m)}) f(\beta_1^{(m)}) f(\beta_2^{(m)}) f(\lambda^{(m)})}$$

I will take normal proposal densities of the Metropolis step as

$$q(\theta^{(m)} | \tilde{\theta}, \epsilon) \sim N(\tilde{\theta}, \epsilon)$$

and

$$q(\tilde{\theta} | \theta^{(m)}, \epsilon) \sim N(\theta^{(m)}, \epsilon)$$

Since they have the same proposals, the $q(\theta^{(m)} | \tilde{\theta}, \epsilon)$ and $q(\tilde{\theta} | \theta^{(m)}, \epsilon)$ will cancel each other out, rendering an $\alpha(\theta^{(m)}, \tilde{\theta})$ as

$$\alpha(\theta^{(m)}, \tilde{\theta}) = \min(1, \frac{f(\tilde{\theta} | \text{data})}{f(\theta^{(m)} | \text{data})}) = \min(1, \frac{k(\text{data} | \tilde{\theta}) * f(\tilde{\theta})}{k(\text{data} | \theta^{(m)}) * f(\theta^{(m)})})$$

Jumping Between Models

I will now try to estimate $P(M=1|\text{data})$ using a Metropolis jump between the two models. For a move from model 1 to model 2, the MH will follow the following steps:

```
sample u ~ U(0,1)
if u < a(M1,M2)
  move to M2
  sample posteriors from M2
else
  stay at M1
  sample posteriors from M1
```

This follows naturally for the move from model 2 to model 1, except that the α values are different. The samples of the posteriors are obtained by using a Gibbs framework for λ and a Metropolis Hastings step for $(\beta_0, \beta_1, \beta_2)$.

The α for moving from model 1 to model 2 will be

$$\alpha_{12} = \min(1, \frac{f(M=2, \beta_0, \beta_1, \beta_2, \lambda | \text{data})}{f(M=1, \beta_0, \beta_1, \beta_2, \lambda | \text{data})})$$

$$\alpha_{12} = \min(1, \frac{P(M=2) f(\lambda | M=2) f(\beta_0 | M=2) f(\beta_1 | M=2) f(\beta_2 | M=2) \prod_i k(y_i | \beta_0, \beta_1, \beta_2, \lambda, M=2)}{P(M=1) f(\lambda | M=1) f(\beta_0 | M=1) f(\beta_1 | M=1) f(\beta_2 | M=1) \prod_i k(y_i | \beta_0, \beta_1, \lambda, M=1)})$$

Many of the prior densities will cancel out, because they are the same for both model 1 and model 2. This leaves us with the following equation.

$$\alpha_{12} = \min(1, \frac{f(\beta_2|M=2) \prod_i k(y_i|\beta_0, \beta_1, \beta_2, \lambda, M=2)}{f(\beta_2|M=1) \prod_i k(y_i|\beta_0, \beta_1, \lambda, M=1)})$$

The α for moving from model 2 to model 1 will be

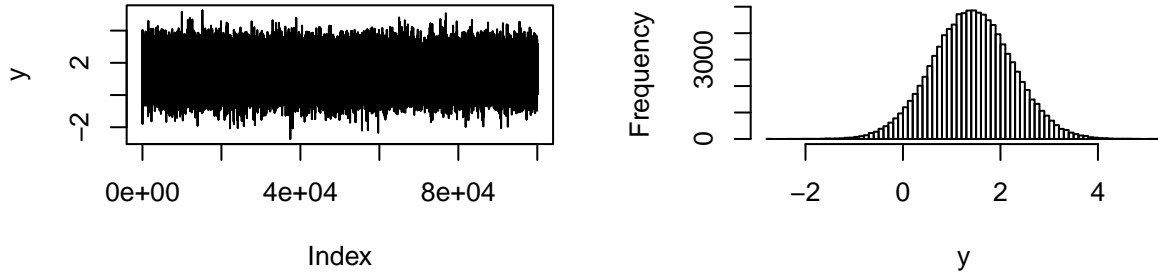
$$\alpha_{21} = \min(1, \frac{f(M=1, \beta_0, \beta_1, \beta_2, \lambda|data)}{f(M=2, \beta_0, \beta_1, \beta_2, \lambda|data)})$$

$$\alpha_{21} = \min(1, \frac{P(M=1)f(\lambda|M=1)f(\beta_0|M=1)f(\beta_1|M=1) \prod_i k(y_i|\beta_0, \beta_1, \lambda, M=1)}{P(M=2)f(\lambda|M=2)f(\beta_0|M=2)f(\beta_1|M=2)f(\beta_2|M=2) \prod_i k(y_i|\beta_0, \beta_1, \beta_2, \lambda, M=2)})$$

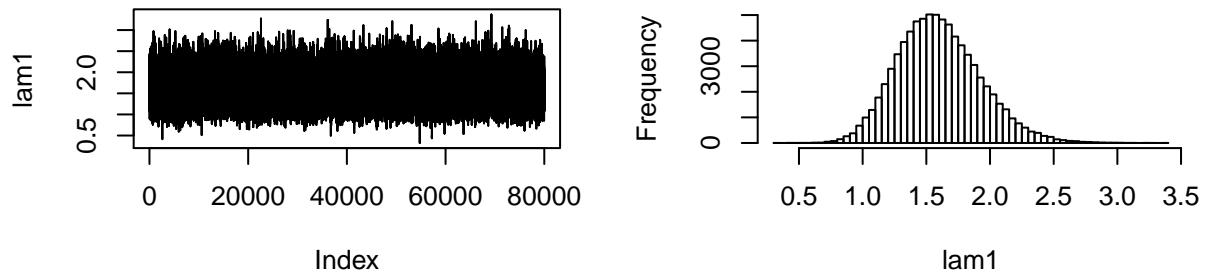
$$\alpha_{21} = \min(1, \frac{f(\beta_2|M=1) \prod_i k(y_i|\beta_0, \beta_1, \lambda, M=1)}{f(\beta_2|M=2) \prod_i k(y_i|\beta_0, \beta_1, \beta_2, \lambda, M=2)})$$

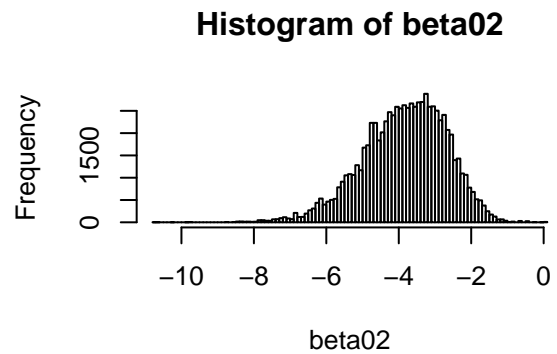
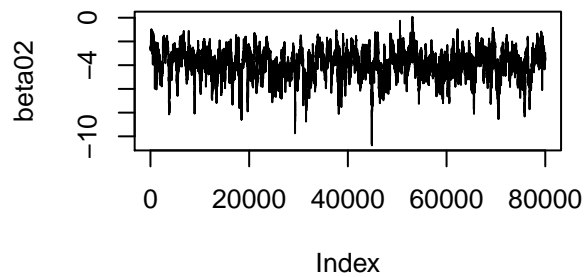
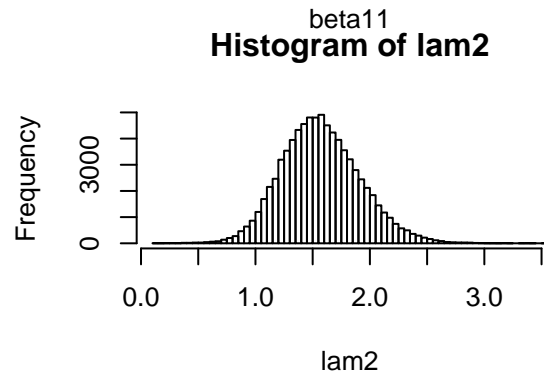
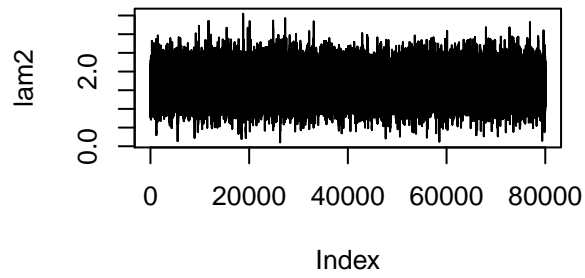
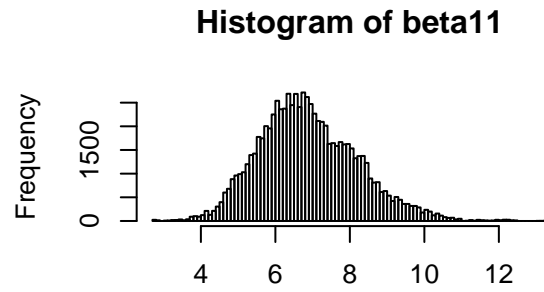
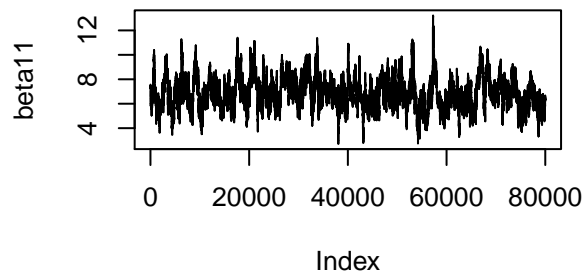
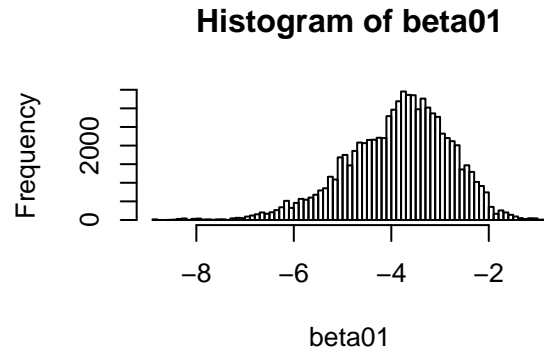
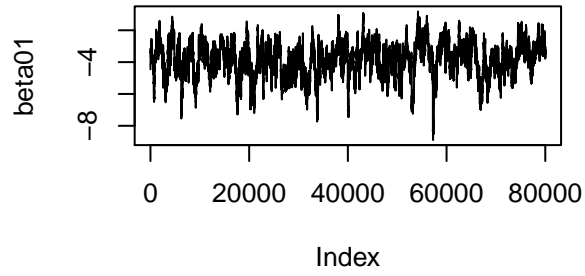
I will take $f(\beta_2|M=1)$ to be a normal centered around 0, with a variance of 1, since we cannot make too many assumptions as to what the value of β_2 is from β_0 and β_1 from model 1.

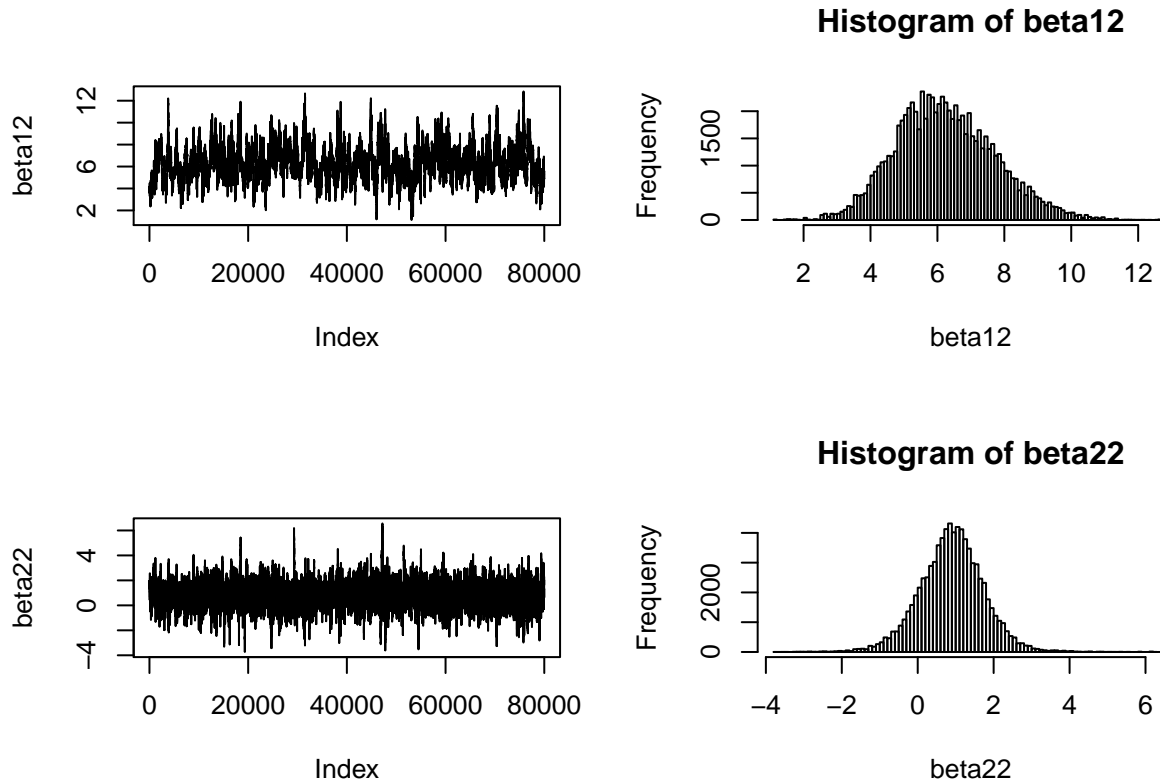
Histogram of y



Histogram of lam1







Results

The Bayes Factor calculation renders a ratio of 3.025. Recall that this is equal to $\frac{P(M1|y_i, x_{1i}, x_{2i})}{P(M2|y_i, x_{1i}, x_{2i})} = \frac{P(M1|y_i, x_{1i}, x_{2i})}{1 - P(M1|y_i, x_{1i}, x_{2i})}$. This means that $P(M1|y_i, x_{1i}, x_{2i}) = 0.752$ and $P(M2|y_i, x_{1i}, x_{2i}) = 0.248$. In lay-mans terms, the chain is estimated to spend approximately 75% of its time taking samples from the first model, and about 25% of its time taking samples from the larger, second model.

The results of the jumping metropolis reveals that the estimated $P(M1|y_i, x_{1i}, x_{2i}) = 0.749$ and $P(M2|y_i, x_{1i}, x_{2i}) = 0.251$. These values were found by calculating the percentage of samples that were taken from model 1, over all iterations of the chain in the jumping metropolis framework. These are very similar results to the Bayes Factor, which we should expect.

These findings confirm that the Monte Carlo Markov chain spends approximately 75% of its time sampling from model 1 and 25% of its time sampling from model 2. This may reveal that majority of the observations in our data are coming from a population that is best represented by model 1.

Estimating y_{51}

The final portion of this assignment is to determine the predictive density for y_{51} , for an observation with $x_{1,51} = 0.23$ and $x_{2,51} = -1.2$ using a Monte Carlo Markov chain method. I looped through the chain using a jumping metropolis step between model 1 and model 1, and for each sampled value, I also sampled y_{51} from $N(y_{i,51} | \log(1 + \exp(\beta_0 + \beta_1 * 0.23)), \sqrt{\frac{1}{\lambda}})$ if the samples came from model 1, and $N(y_{i,51} | \log(1 + \exp(\beta_0 + \beta_1 * 0.23 - \beta_2 * 1.2)), \sqrt{\frac{1}{\lambda}})$ if the samples came from model 2. I collected 100,000 observations with 20,000 burnin of y_{51} , and found the mean to be 0.1000257.

Below is a trace plot and histogram of the predictive density of y_{51} . We see that the chain moves well, and the predictive density is almost symmetrically distributed around the mean of 0.1000257.

