Bicheng (Tony) Zhang

STAT 8670 Final Project Report

Introduction

For my final project, the dataset I chose came from a website called Emcien, and contains data on Automobile Insurance claims including location, policy type and claim amount (in US dollars). The reason why I want to work with this data is that I believe that males, especially younger men, have more expensive/higher car insurance premiums compared with women as men tend to drive more aggressive/reckless compared to women. This dataset has a total of 9,134 observations, and 26 variables. The only variables that I am interested in are the "Claim Amount", "Gender", and "Coverage". Claim Amount is the response variable in the dataset and is a continuous variable. Gender and Coverage are both categorical variables, with Gender taking 2 values (Male & Female) and Coverage taking 3 values ("Basic", "Extended", & "Premium").

In this project we will do a Bayesian data analysis using the Markov Chain Monte Carlo method on this data to hypothesize whether men had higher mean car insurance claims compared to women. Bayesian analysis is a formal method for combining prior beliefs with observed information. It can fit very realistic but complicated models. The questions we are interested in are:

- Is the mean "Claim Amount" for Men under "Basic" car insurance coverage greater than the mean of "Claim Amount" for Women under the same coverage? If so, how much?
- Is the mean "Claim Amount" for Men under "Extended" car insurance coverage greater than the mean of "Claim Amount" for Women under the same coverage? If so, how much?
- Is the mean "Claim Amount" for Men under "Premium" car insurance coverage greater than the mean of "Claim Amount" for Women under the same coverage? If so, how much?

Methods

As mentioned earlier, the data analysis will be done using Markov Chain Monte Carlo. We will be using the Metropolis-Hastings algorithm. We are trying to find an estimate of the

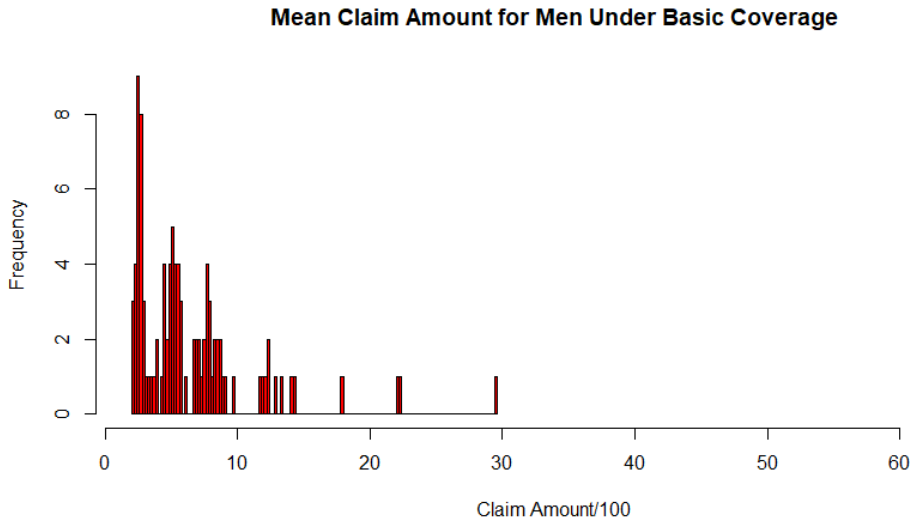
posterior probability that the difference between the mean of men and womens' auto insurance claims is greater than 0, in other words that men have higher claim amounts than women. To accomplish this, we need to construct a likelihood function and prior function, then multiply them to get our posterior. Because the likelihood function involves a lot of successive multiplications with small numbers, we will undoubtedly work with exponentially small numbers. Therefore, instead of using the entire dataset, I will instead take a random sample of 100 male observations and 100 female observations within each coverage category (basic, extended, and premium). This is because when R works with extremely small numbers close to 0, for example 10^{-20}, it will just round the number to 0. So I will just choose a random sample of 100 to prevent this from happening, instead of using all ~9,000 observations.

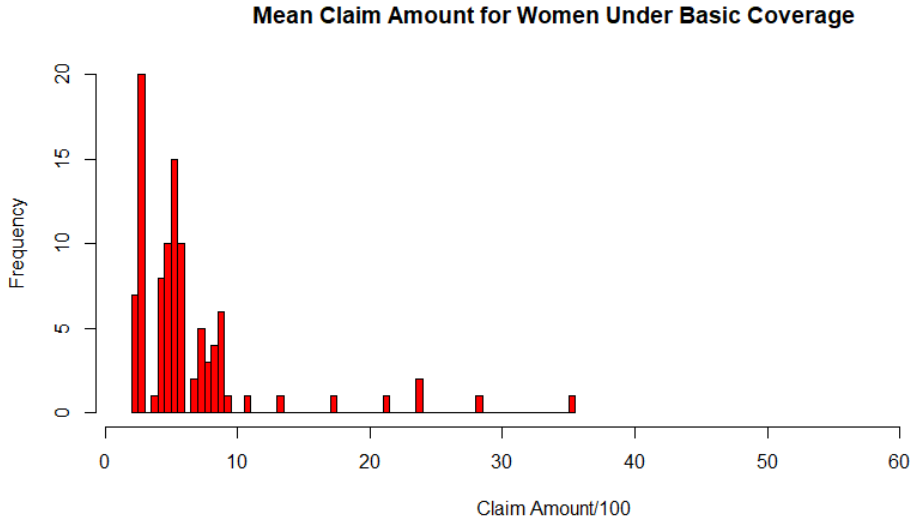
To start, we assume that the claim amounts for men and women under each category are normally distributed with the parameter vector $\theta = (\mu_1, \sigma_1, \mu_2, \sigma_2,)$, where 1 represents men and 2 represents women. We then need to decide on their prior density distributions, so I just chose both μ_1 and μ_2 as normally distributed with the sample mean claim amount for men and women respectively as both the mean and standard deviation and σ_1 and σ_2 as exponential with the same respective sample mean (rate = 1/mean). All 4 variables are independent so we can just multiply them to get the joint density. We chose the priors like this because this way they do not prejudge which is bigger, and they are uniformish over all even remotely plausible values. This way we let the data produce any interesting features in the posterior distribution, rather than introducing them ourselves in the prior.

Data Analyses

Basic Coverage:

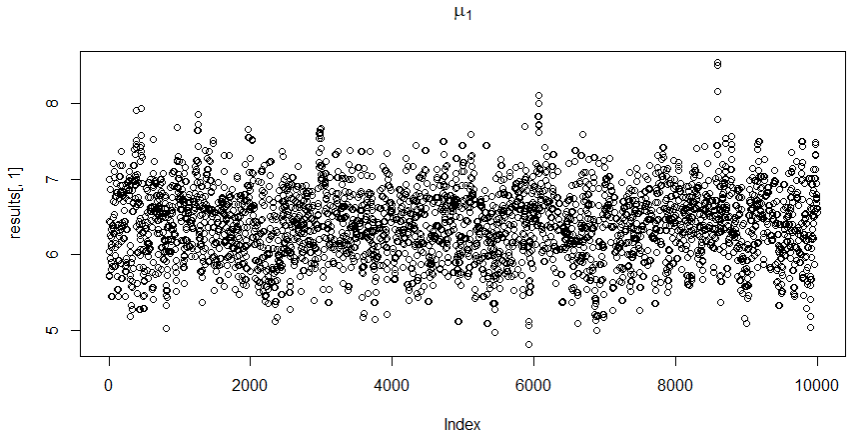
The histograms for mean auto insurance claim amounts for men and women with basic coverage are given below:

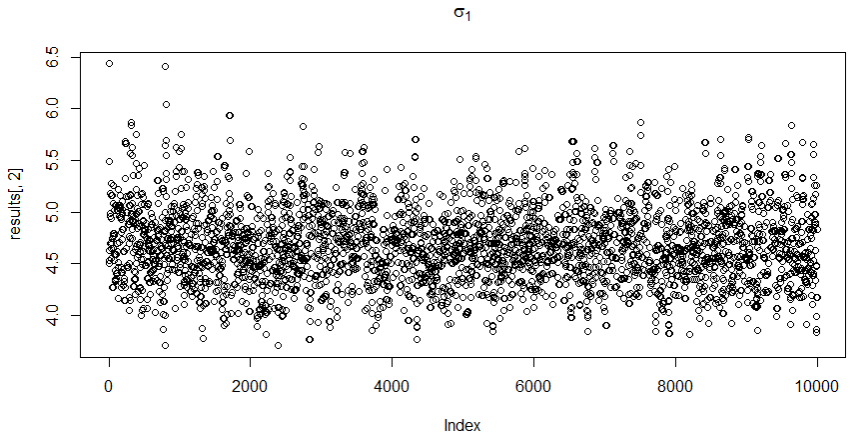


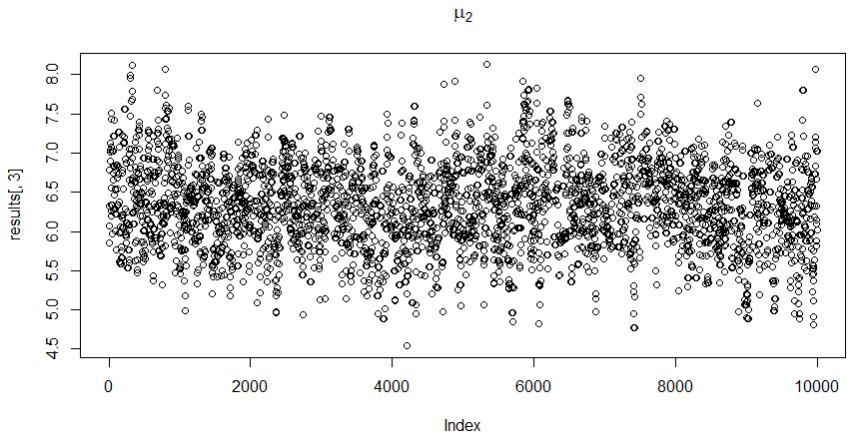


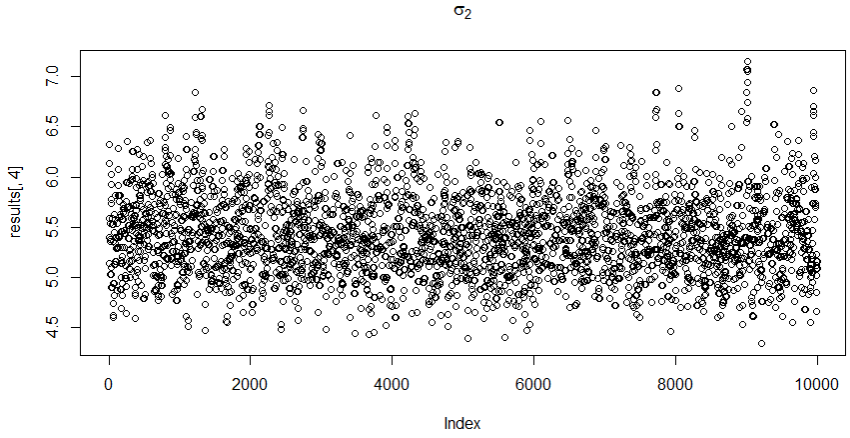
We will run a Markov chain using Metropolis-Hastings algorithm for 10,000 iterations to simulate a sample. From our random sample from the original data, the sample means for men and women respectively were 6.435025 and 6.322932. Therefore we choose our starting values as θ_0 = (6.435025, 6.435025, 6.322932, 6.322932). Given a current state, we decide on a way to propose a "candidate" move, then evaluate the posterior of the candidate move and posterior of the current state and take the ratio of these two. We then record all our results in a big matrix with 10000 rows by 4 columns where the first row will be our original starting values and each successive row will record the next θ vector as we run the Markov chain.

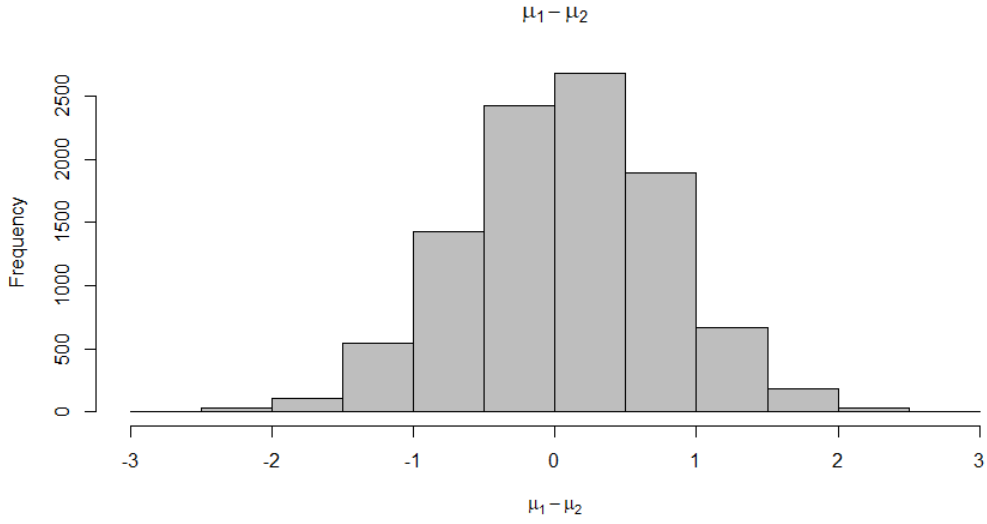
Below are the results of the Markov chain:







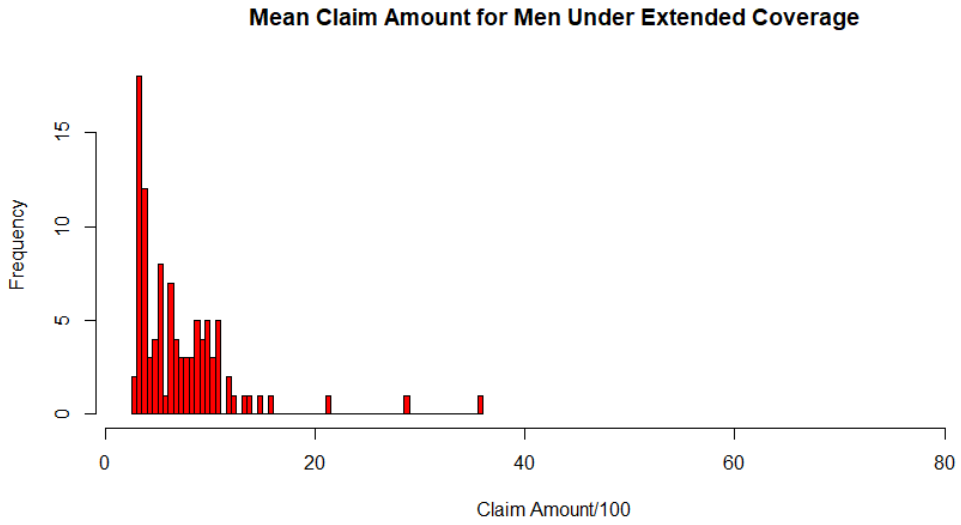


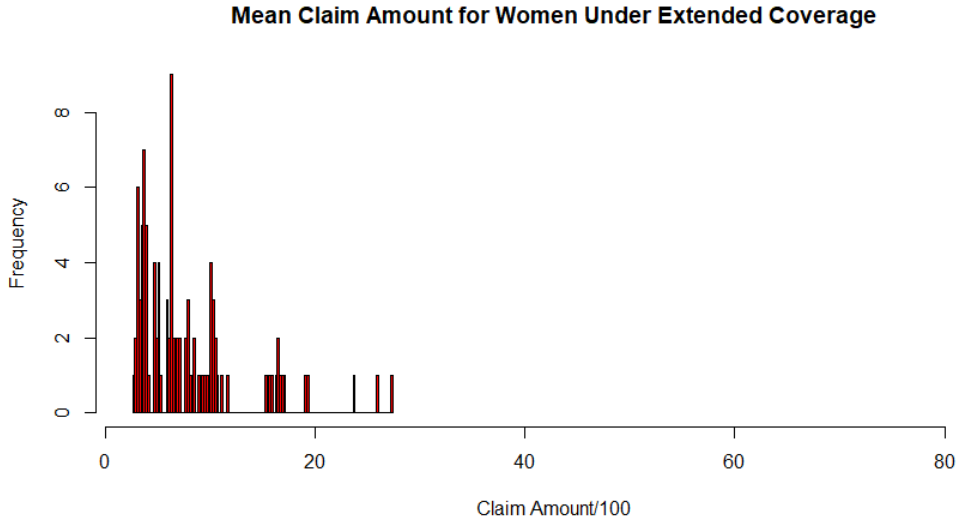


Our estimated posterior probability that the mean claim amount for men under basic coverage is greater than the mean claim amount for women under basic coverage is 0.5456.

Extended Coverage:

The histograms for mean auto insurance claim amounts for men and women with extended coverage are given below:

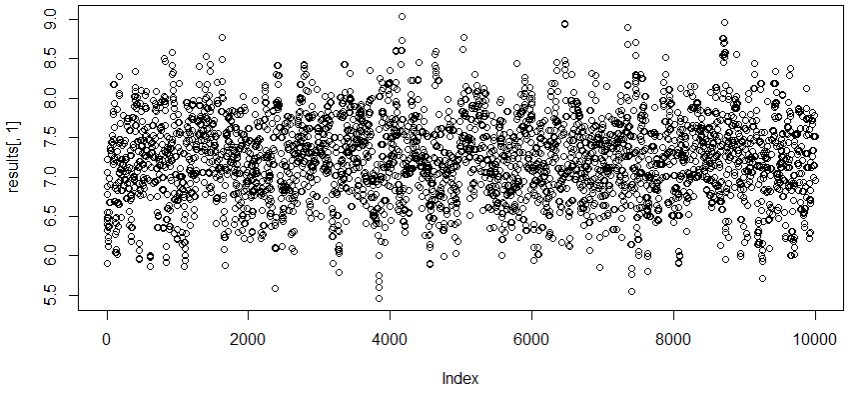




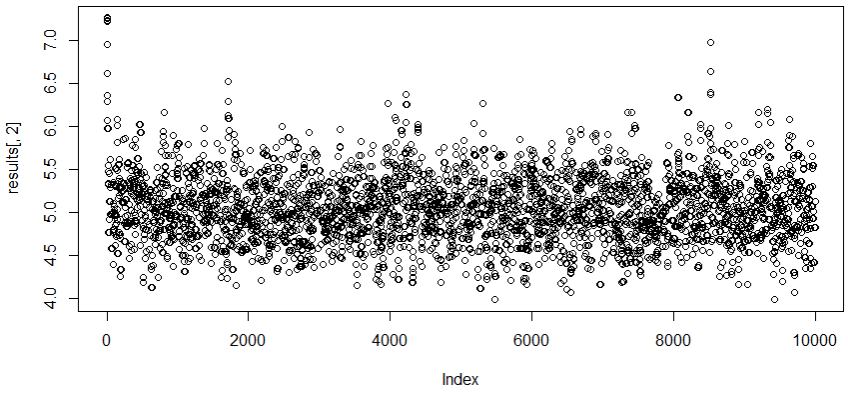
Again, we will run a Markov chain using Metropolis-Hastings algorithm for 10,000 iterations to simulate a sample. From our random sample from the original data, the sample means for men and women respectively were 7.226277 and 7.822649. Therefore we choose our starting values as θ_0 = (7.226277, 7.226277, 7.822649, 7.822649).

Below are the results of the Markov chain:

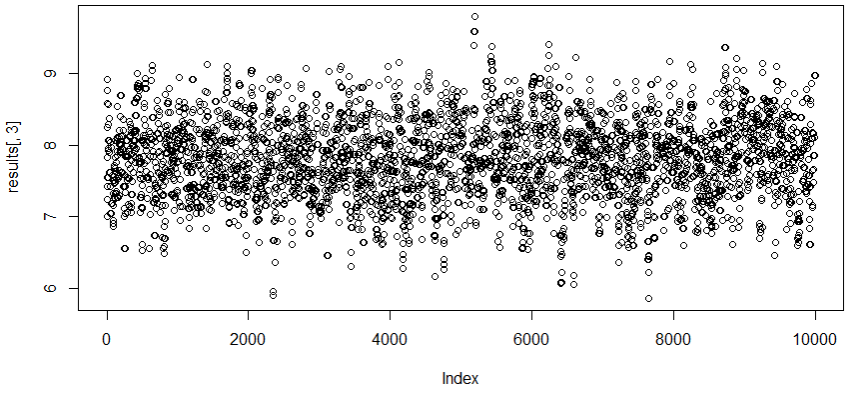
μ_1

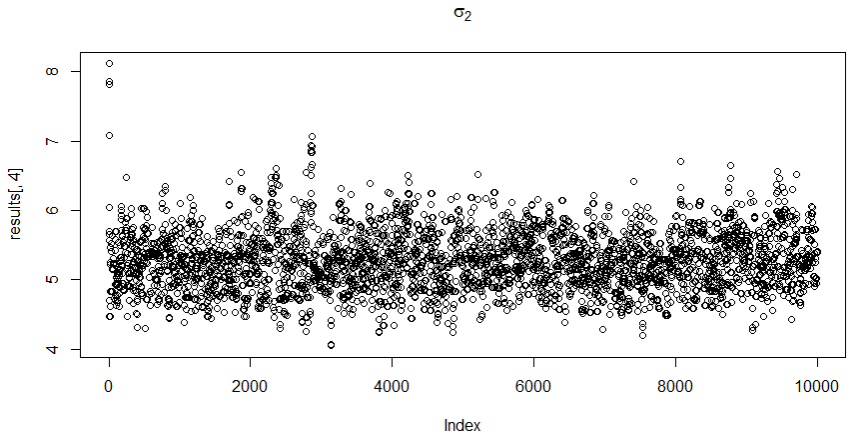


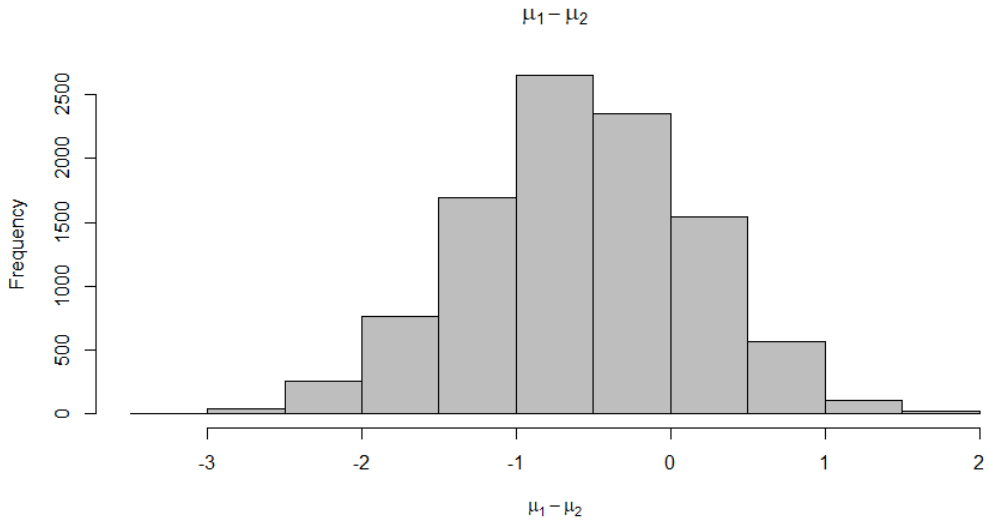
σ_1



μ_2



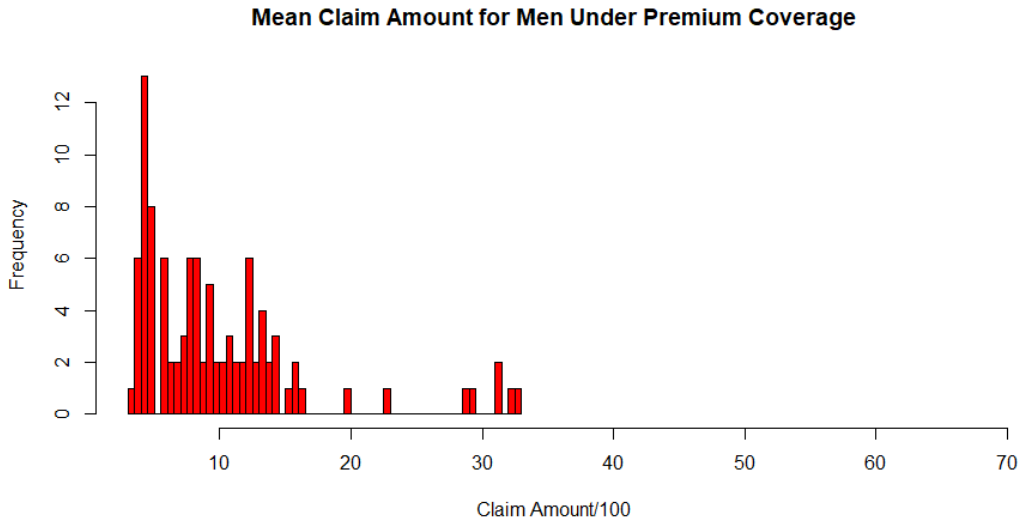


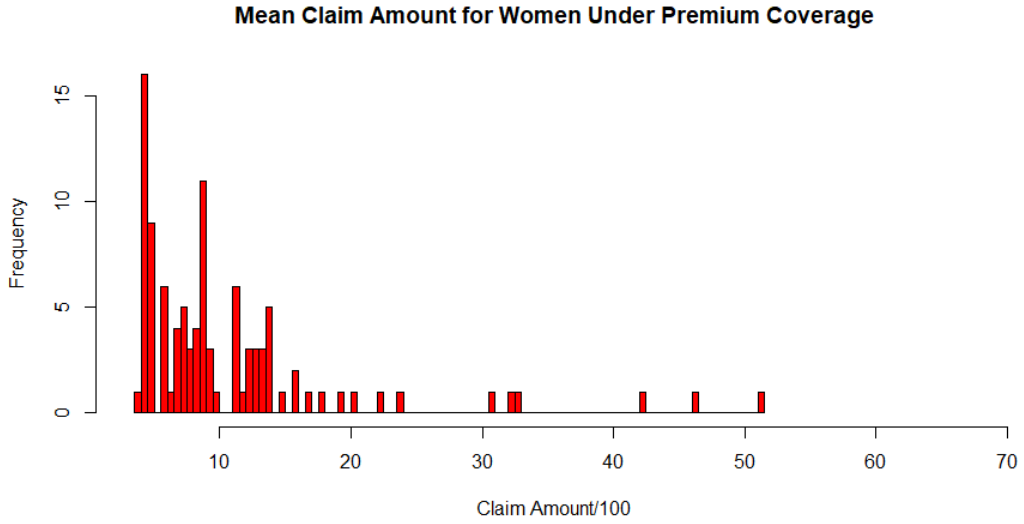


Our estimated posterior probability that the mean claim amount for men under extended coverage is greater than the mean claim amount for women under extended coverage is 0.2242.

Premium Coverage:

The histograms for mean auto insurance claim amounts for men and women with extended coverage are given below:

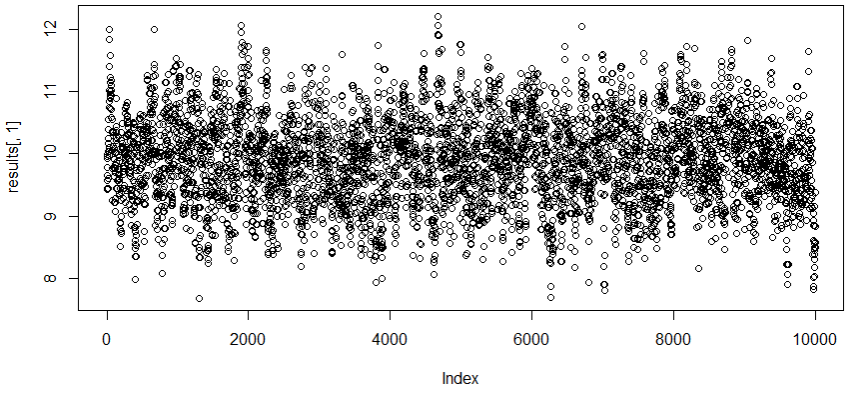




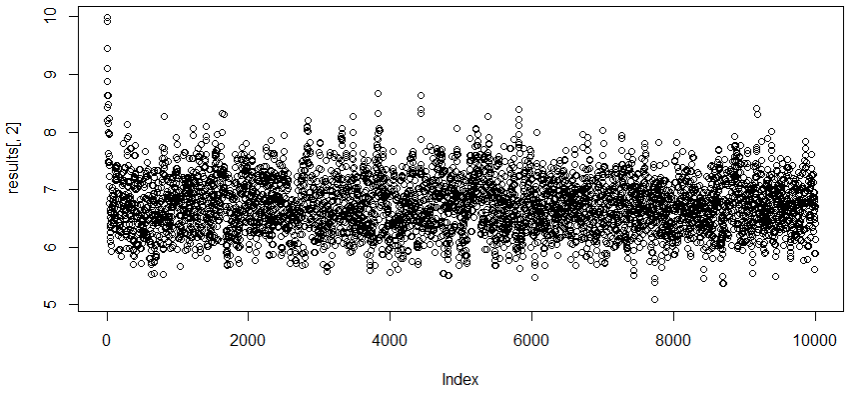
Again, we will run a Markov chain using Metropolis-Hastings algorithm for 10,000 iterations to simulate a sample. From our random sample from the original data, the sample means for men and women (premium coverage) respectively were 9.922662 and 10.70696. Therefore we choose our starting values as θ_0 = (9.922662, 9.922662, 10.70696, 10.70696).

Below are the results of the Markov chain:

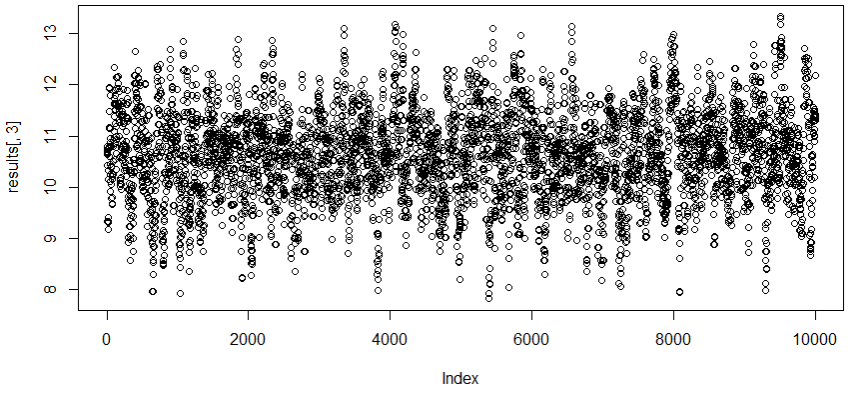
μ_1

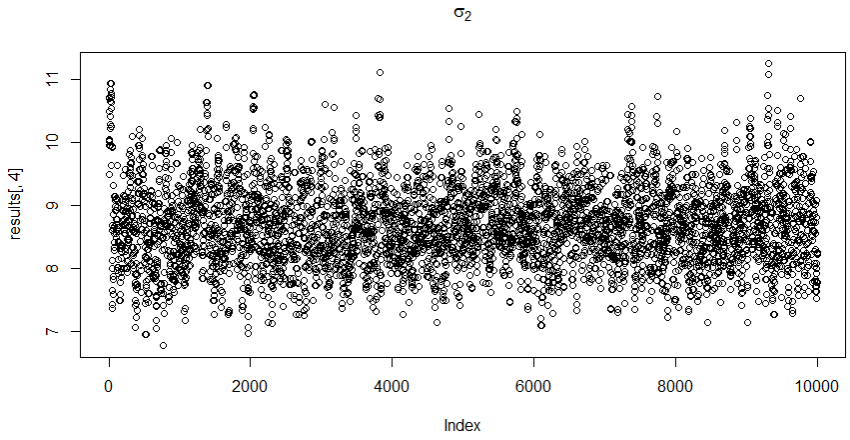


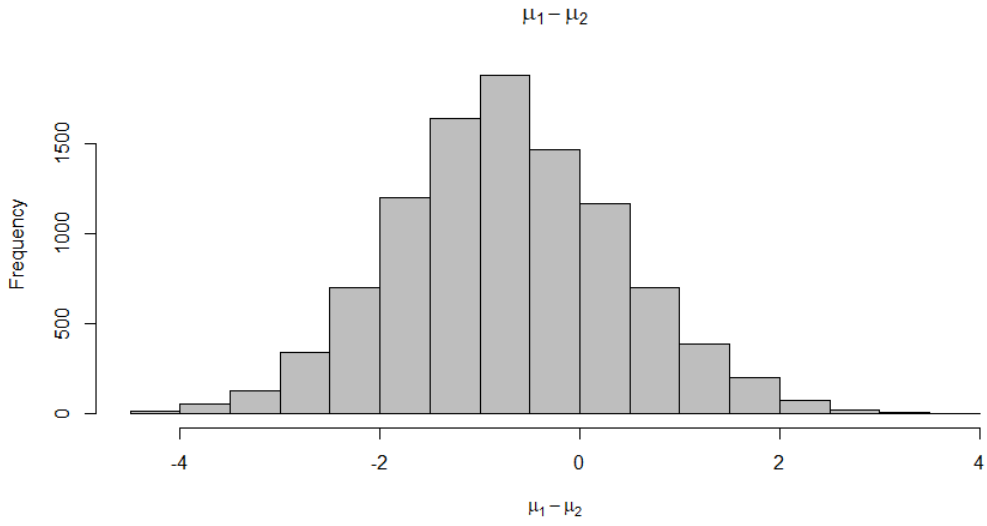
σ_1



μ_2







Our estimated posterior probability that the mean claim amount for men under premium coverage is greater than the mean claim amount for women under premium coverage is 0.2572.

Conclusion:

It was interesting to note that none of the posterior probabilities in each of the 3 coverage categories were relatively high. In fact, we could say with high probability that the mean of male claim amounts is less than the mean of female claim amounts in the extended and premium coverage categories (their posterior probabilities for $\mu_1 - \mu_2 > 0$ were 0.2242 and 0.2572 respectively).

Appendix (R Code):

```
# First, read in and explore the data:

raw_data <- read.csv("C:\\Users\\tzhan\\Google Drive\\GSU Graduate School\\STAT 8670 Computational Methods in Statistics\\Project\\Auto_Insurance_Claims_Sample.csv",header=T)

data <- data.frame(raw_data$Claim.Amount,raw_data$Gender,raw_data$Coverage)

names(data) <- c("Claim Amount","Gender","Coverage")

plot(data$Coverage,data$`Claim Amount`)

Basic_Coverage <- data[data$Coverage=='Basic',]

Extended_Coverage <- data[data$Coverage=='Extended',]

Premium_Coverage <- data[data$Coverage=='Premium',]

table(data$Gender)

table(Basic_Coverage$Gender)

table(Extended_Coverage$Gender)

table(Premium_Coverage$Gender)


Basic_Coverage_M <- subset(Basic_Coverage, Gender=='M')

Basic_Coverage_F <- subset(Basic_Coverage, Gender=='F')

Extended_Coverage_M <- subset(Extended_Coverage, Gender=='M')

Extended_Coverage_F <- subset(Extended_Coverage, Gender=='F')

Premium_Coverage_M <- subset(Premium_Coverage, Gender=='M')

Premium_Coverage_F <- subset(Premium_Coverage, Gender=='F')


# Basic Coverage:

M_sample_Basic <- sample(nrow(Basic_Coverage_M), 100)

F_sample_Basic <- sample(nrow(Basic_Coverage_F), 100)

Amt_Male_Basic <- Basic_Coverage_M[c(M_sample_Basic),1]/100

Amt_Female_Basic <- Basic_Coverage_F[c(F_sample_Basic),1]/100


mean(Amt_Male_Basic) # 6.435025

mean(Amt_Female_Basic) # 6.322932
```

```
# draw a picture

xlim = c(min(Basic_Coverage[,1]/100),max(Basic_Coverage[,1]/100))

hist(Amt_Male_Basic,100,col="red",xlim=xlim,main='Mean Claim Amount for Men Under Basic Coverage',xlab='Claim Amount/100')

hist(Amt_Female_Basic,100,col="red",xlim=xlim,main='Mean Claim Amount for Women Under Basic Coverage',xlab='Claim Amount/100')


# likelihood function

lik = function(th){

  mu1 <- th[1]; sig1 <- th[2]; mu2 <- th[3]; sig2 <- th[4]


prod(dnorm(Amt_Male_Basic,mean=mu1,sd=sig1))*prod(dnorm(Amt_Female_Basic,mean=mu2,sd=sig2))

}


# prior function

prior <-  function(th){

  mu1 <- th[1]; sig1 <- th[2]; mu2 <- th[3]; sig2 <- th[4]

  if(sig1 <= 0 | sig2 <= 0) return(0)


dnorm(mu1,6.435025,6.435025)*dnorm(mu2,6.322932,6.322932)*dexp(sig1,rate=1/6.435025)*dexp(sig2,rate=1/6.322932)

}


# posterior function

post <-  function(th){prior(th) * lik(th)}


# Starting values

mu1 <-  6.435025; sig1 <-  6.435025; mu2 <-  6.322932; sig2 <-  6.322932

th0 <- c(mu1,sig1,mu2,sig2)

# Here is what does the MCMC (Metropolis method):
```

```
nit <- 10000
results <-  matrix(0, nrow=nit, ncol=4)
th <-  th0
results[1,] <-  th0
for(it in 2:nit){
  cand <-  th + rnorm(4,sd=.5)
  ratio <-  post(cand)/post(th)
  if(runif(1)<ratio) th <- cand
  results[it,] <-  th
}
# Take a peek at what we got
edit(results)

plot(results[,1], main=expression(mu[1]))
plot(results[,2], main=expression(sigma[1]))
plot(results[,3], main=expression(mu[2]))
plot(results[,4], main=expression(sigma[2]))

mu1s <-  results[,1]
sig1s <-  results[,2]
mu2s <- results[,3]
sig2s <- results[,4]
hist(mu1s-mu2s, xlab=expression(mu[1]-mu[2]), main=expression(mu[1]-mu[2]), col = 'gray')
mean(mu1s-mu2s > 0) # 0.5456

# Extended Coverage:
M_sample_Ext <- sample(nrow(Extended_Coverage_M), 100)
F_sample_Ext <- sample(nrow(Extended_Coverage_F), 100)
Amt_Male_Ext <- Extended_Coverage_M[c(M_sample_Ext),1]/100
Amt_Female_Ext <- Extended_Coverage_F[c(F_sample_Ext),1]/100
```

```
mean(Amt_Male_Ext) # 7.226277

mean(Amt_Female_Ext) # 7.822649


# draw a picture

xlim = c(min(Extended_Coverage[,1]/100),max(Extended_Coverage[,1]/100))

hist(Amt_Male_Ext,100,col="red",xlim=xlim,main='Mean Claim Amount for Men Under
Extended Coverage',xlab='Claim Amount/100')

hist(Amt_Female_Ext,100,col="red",xlim=xlim,main='Mean Claim Amount for Women Under
Extended Coverage',xlab='Claim Amount/100')


# likelihood function

lik = function(th){

  mu1 <- th[1]; sig1 <- th[2]; mu2 <- th[3]; sig2 <- th[4]


prod(dnorm(Amt_Male_Ext,mean=mu1,sd=sig1))*prod(dnorm(Amt_Female_Ext,mean=mu2,sd
=sig2))

}


# prior function

prior <-  function(th){

  mu1 <- th[1]; sig1 <- th[2]; mu2 <- th[3]; sig2 <- th[4]

  if(sig1 <= 0 | sig2 <= 0) return(0)


dnorm(mu1,7.226277,7.226277)*dnorm(mu2,7.822649,7.822649)*dexp(sig1,rate=1/7.226277)*
dexp(sig2,rate=1/7.822649)

}


# posterior function

post <-  function(th){prior(th) * lik(th)}


# Starting values
```

```
mu1 <-  7.226277; sig1 <-  7.226277; mu2 <-  7.822649; sig2 <-  7.822649
th0 <- c(mu1,sig1,mu2,sig2)
# Here is what does the MCMC (Metropolis method):
nit <- 10000
results <-  matrix(0, nrow=nit, ncol=4)
th <-  th0
results[1,] <-  th0
for(it in 2:nit){
  cand <-  th + rnorm(4,sd=.5)
  ratio <-  post(cand)/post(th)
  if(runif(1)<ratio) th <- cand
  results[it,] <-  th
}
# Take a peek at what we got
edit(results)

plot(results[,1], main=expression(mu[1]))
plot(results[,2], main=expression(sigma[1]))
plot(results[,3], main=expression(mu[2]))
plot(results[,4], main=expression(sigma[2]))

mu1s <-  results[,1]
sig1s <-  results[,2]
mu2s <- results[,3]
sig2s <- results[,4]
hist(mu1s-mu2s, xlab=expression(mu[1]-mu[2]), main=expression(mu[1]-mu[2]), col = 'gray')
mean(mu1s-mu2s > 0) # 0.2242

# Premium Coverage:
M_sample_Prem <- sample(nrow(Premium_Coverage_M), 100)
```

```
F_sample_Prem <- sample(nrow(Premium_Coverage_F), 100)

Amt_Male_Prem <- Premium_Coverage_M[c(M_sample_Prem),1]/100

Amt_Female_Prem <- Premium_Coverage_F[c(F_sample_Prem),1]/100


mean(Amt_Male_Prem) # 9.922662

mean(Amt_Female_Prem) # 10.70696


# draw a picture

xlim = c(min(Premium_Coverage[,1]/100),max(Premium_Coverage[,1]/100))

hist(Amt_Male_Prem,100,col="red",xlim=xlim,main='Mean Claim Amount for Men Under Premium Coverage',xlab='Claim Amount/100')

hist(Amt_Female_Prem,100,col="red",xlim=xlim,main='Mean Claim Amount for Women Under Premium Coverage',xlab='Claim Amount/100')


# likelihood function

lik = function(th){

  mu1 <- th[1]; sig1 <- th[2]; mu2 <- th[3]; sig2 <- th[4]


prod(dnorm(Amt_Male_Prem,mean=mu1,sd=sig1))*prod(dnorm(Amt_Female_Prem,mean=mu2,sd=sig2))

}


# prior function

prior <-  function(th){

  mu1 <- th[1]; sig1 <- th[2]; mu2 <- th[3]; sig2 <- th[4]

  if(sig1 <= 0 | sig2 <= 0) return(0)


dnorm(mu1,9.922662,9.922662)*dnorm(mu2,10.70696,10.70696)*dexp(sig1,rate=1/9.922662)*dexp(sig2,rate=1/10.70696)

}


# posterior function
```

```
post <-  function(th){prior(th) * lik(th)}

# Starting values
mu1 <-  9.922662; sig1 <-  9.922662; mu2 <-  10.70696; sig2 <-  10.70696
th0 <- c(mu1,sig1,mu2,sig2)
# Here is what does the MCMC (Metropolis method):
nit <- 10000
results <-  matrix(0, nrow=nit, ncol=4)
th <-  th0
results[1,] <-  th0
for(it in 2:nit){
  cand <-  th + rnorm(4,sd=.5)
  ratio <-  post(cand)/post(th)
  if(runif(1)<ratio) th <- cand
  results[it,] <-  th
}
# Take a peek at what we got
edit(results)

plot(results[,1], main=expression(mu[1]))
plot(results[,2], main=expression(sigma[1]))
plot(results[,3], main=expression(mu[2]))
plot(results[,4], main=expression(sigma[2]))

mu1s <-  results[,1]
sig1s <-  results[,2]
mu2s <- results[,3]
sig2s <- results[,4]
hist(mu1s-mu2s, xlab=expression(mu[1]-mu[2]), main=expression(mu[1]-mu[2]), col = 'gray')
mean(mu1s-mu2s > 0) # 0.2572
```