Bayesian Analysis of Data for Admission in the University

Adeepa Gustinna Wadu
Aash Makwana
Department of Mathematics and Statistics,
University Of Calgary

04/04/2024

Outline

- Introduction
- Methodology
- Results and Discussion

Introduction

- In today's education world there are many number of students
- We are focusing on only the students who want to pursue their higher education in universities.
- But we are focusing on only the students who want to do their Masters in America.
- Students who want to do masters in America have to write GRE and TOEFL

Introduction

UNIVERSITY OF CALGARY

- Among these tests CGPA is another important factors consider my universities to grant admission to universities.

- This study focus on figure out underlying probability distributions of these features

Dataset

UNIVERSITY OF CALGARY

- Dataset has been sourced from Kaggle
- Includes various attributes related to university admissions.
- The total number of observations is 400
- We are considering to perform Bayesian analysis for GRE score, TOEFL score and CGPA variables

Objective

- Main objective of this study is to figure out which data generating model is most suitable for these features

Methodology- Preprocessing

UNIVERSITY OF CALGARY

- Checked for missing values
- Divide training(75% of data) and testing datasets

Descriptive Analysis

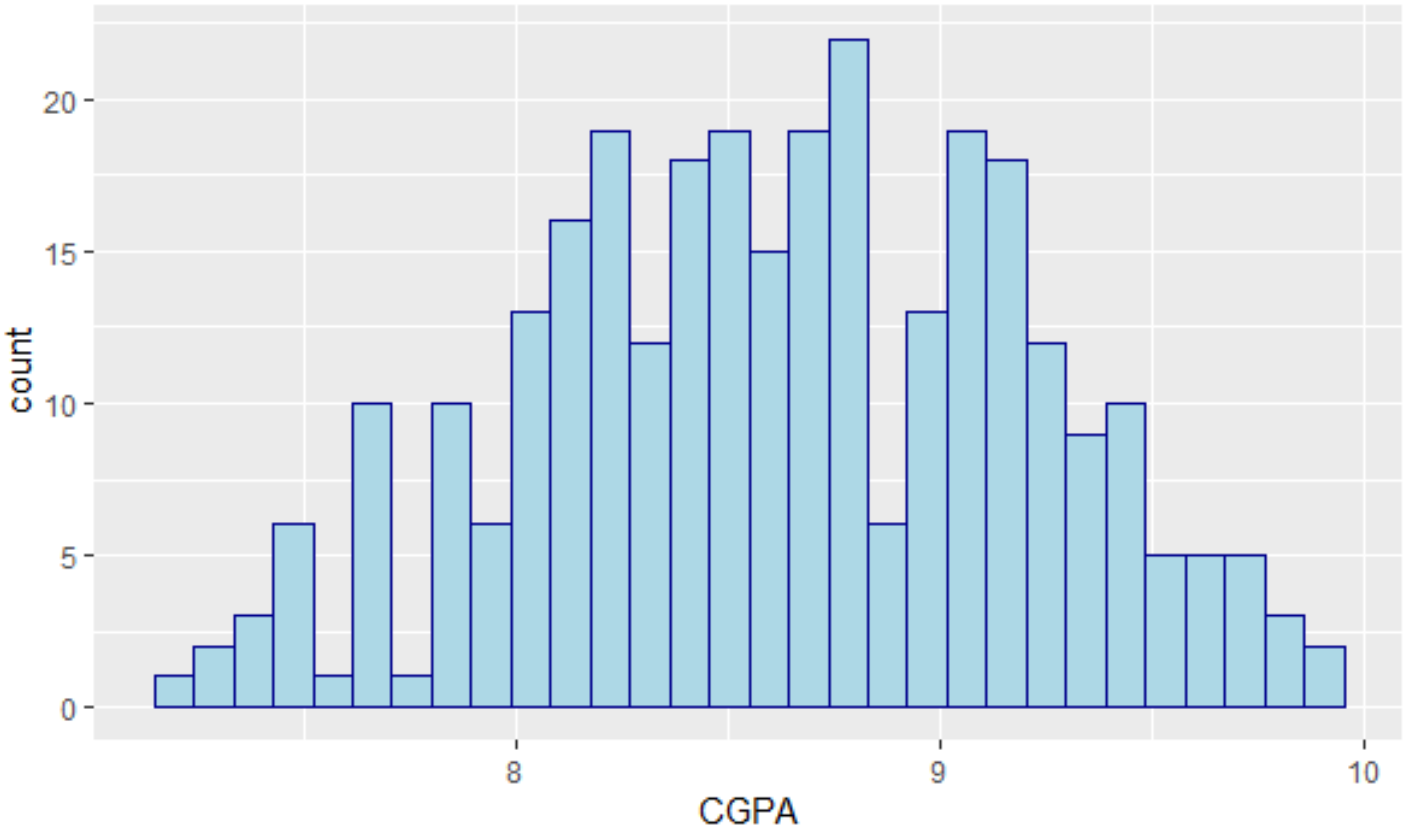
UNIVERSITY OF CALGARY

- Summary of all features were calculated

- Density curves used for explore the distribution of the featues

Descriptive Analysis

- CGPA

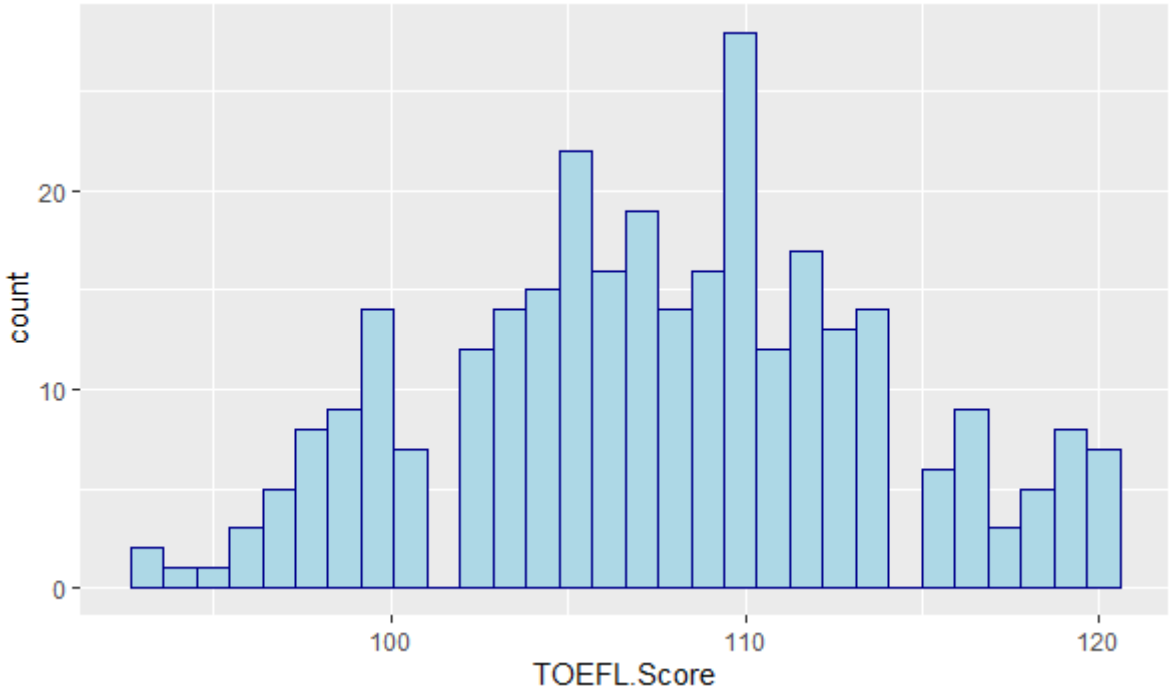


Descriptive Analysis

- GRE Score

Descriptive Analysis

- TOEFL score



Descriptive Analysis

UNIVERSITY OF CALGARY

- Summary

Feature	Median	Mean	Standrad Deviation
CGPA	8.640	8.625	0.584
GRE score	317.5	317.2	11.408
TOEFL score	108.0	107.7	6.014

Analysis – Approach 1

- For a given feature, assume data generated from normal distribution.

$$P(X_{1:n}|\mu) \sim N(\mu, \sigma^2)$$

- Assume standard deviation is fixed

- Priors were defined for μ
  - $P(\mu) \sim N(\mu_1, \sigma_1^2)$
  - $P(\mu) \propto 1$

Analysis – Approach 1

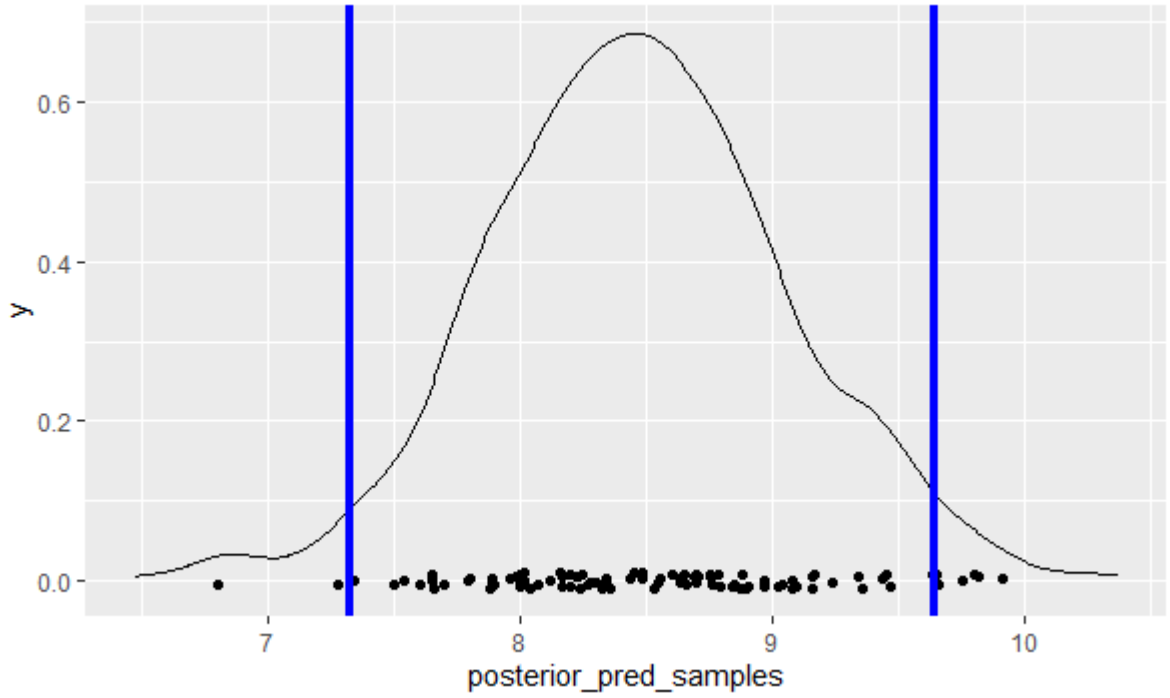
- Posterior distribution is calculated for the training data.
- Posterior predictive distribution was calculated for testing data
- 95% credible interval was calculated for Posterior predictive distribution
- Finally number of testing sample within the credible interval was calculated

Analysis – Approach 1

Prior	Posterior	Posterior Predictive
$P(\mu) \sim N(\mu_1, \sigma_1^2)$	$P(\mu \mid X_{1:n}) \sim N(\mu_2, \sigma_2^2)$ $\sigma_2^2 = \frac{\sigma_1^2 \sigma^2}{n\sigma_1^2 + \sigma^2} \quad \mu_2 = \frac{n\sigma_1^2 \bar{X} + \mu_1 \sigma^2}{n\sigma_1^2 + \sigma^2}$	$P(X_{new} \mid X_{1:n}) \sim N(\mu_3, \sigma_3^2)$ $\mu_3 = \mu_2$ $\sigma_3^2 = \sigma_2^2 + \sigma^2$
$P(\mu) \propto 1$	$\mu_2 = \mu \; P(\mu \mid X_{1:n}) \sim N(\mu_2, \sigma_2^2)$ $\sigma_2^2 = \sigma^2$	$P(X_{new} \mid X_{1:n}) \sim N(\mu_3, \sigma_3^2)$ $\mu_3 = \mu_2$ $\sigma_3^2 = \sigma_2^2 + \sigma^2$

Approach 1 - Results

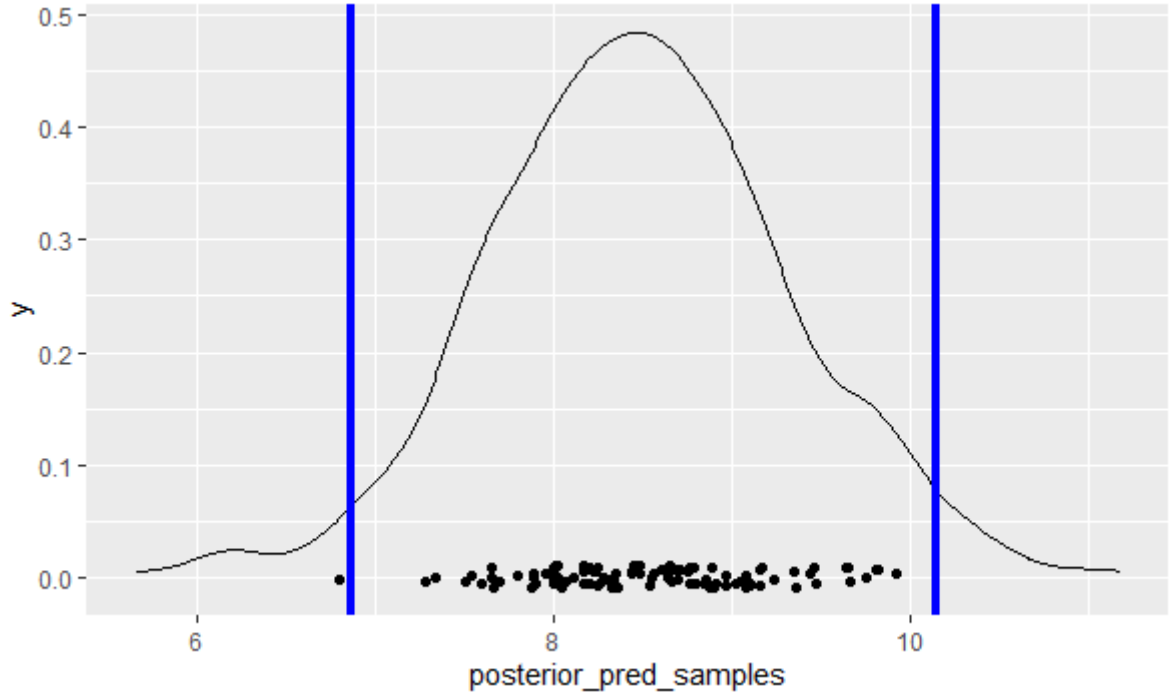
- CGPA - $P(X_{1:n}|\mu) \sim N(8.5, 0.35)$ &$P(\mu) \sim N(2,1)$



- 92 test samples were inside the credible interval

Approach 1 - Results

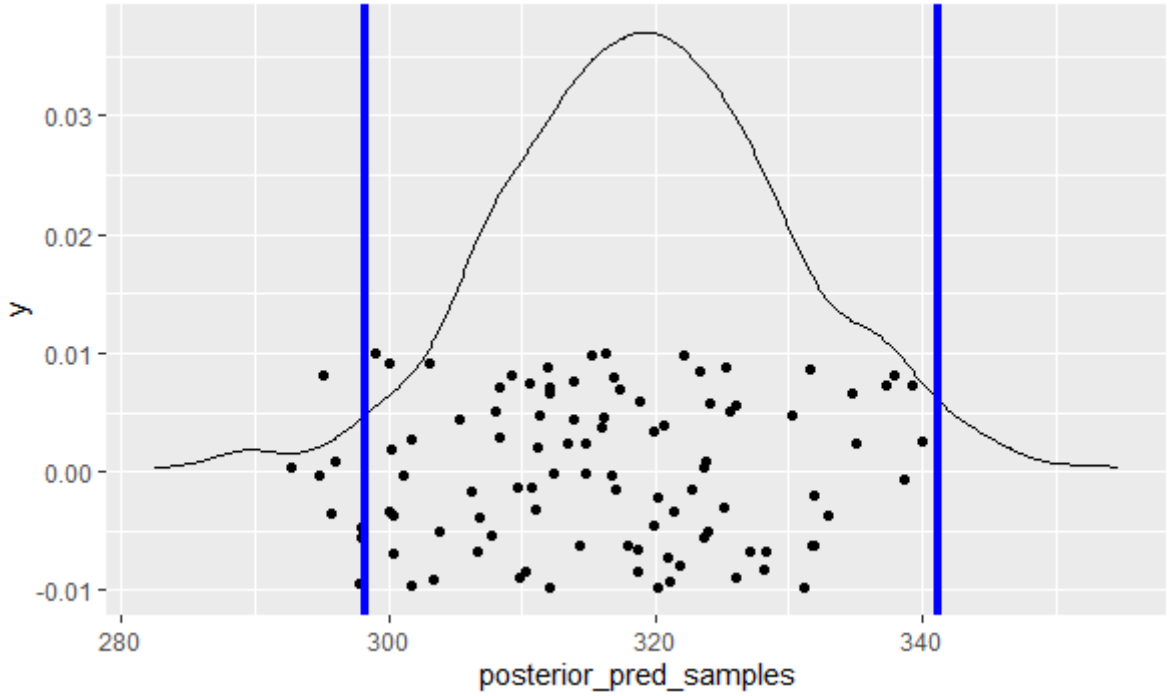
- CGPA - $P(X_{1:n}|\mu) \sim N(8.5, 0.35)$ & $P(\mu) \propto 1$



- 99 test samples were inside the credible interval

Approach 1 - Results

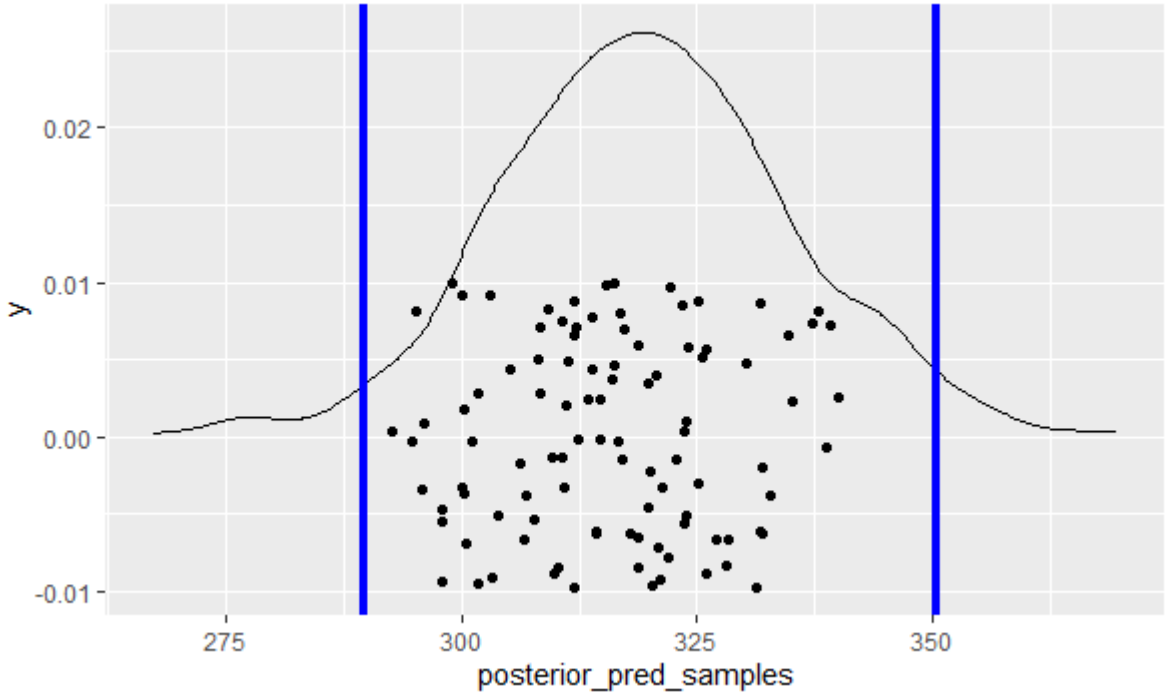
- GRE score - $P(X_{1:n}|\mu) \sim N(320, 120)$ & $P(\mu) \sim N(250, 100)$



- 92 test samples were inside the credible interval

Approach 1 - Results

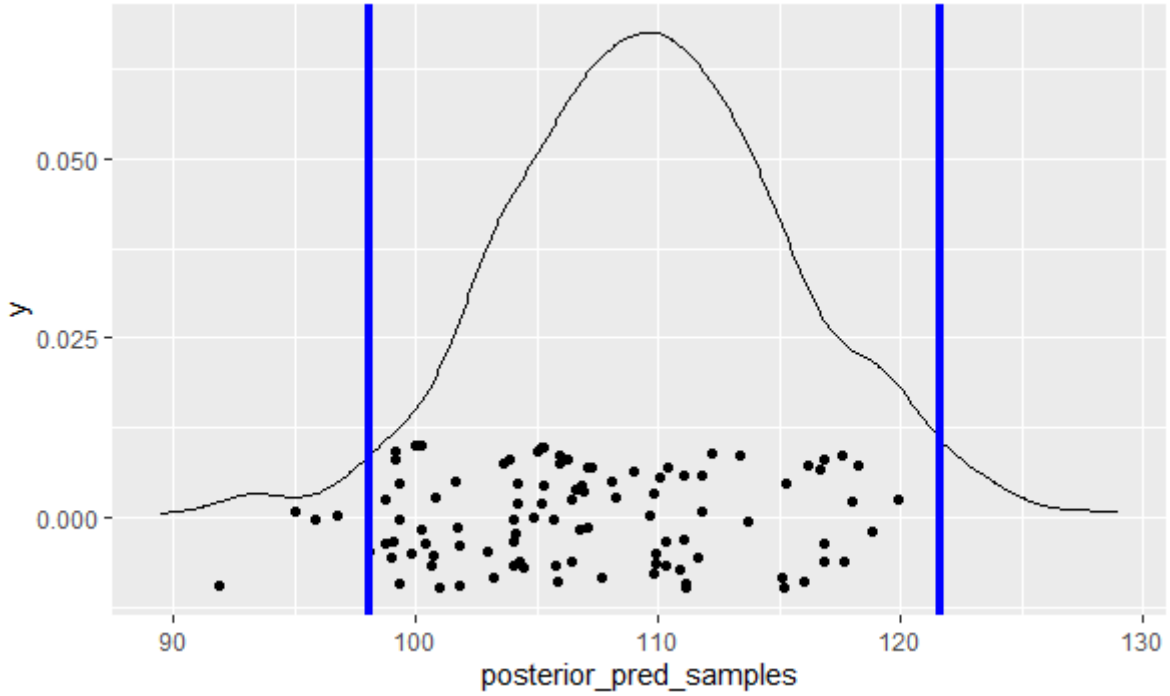
- GRE score - $P(X_{1:n}|\mu) \sim N(320, 120)$ & $P(\mu) \propto 1$



- 100 test samples were inside the credible interval

Approach 1 - Results

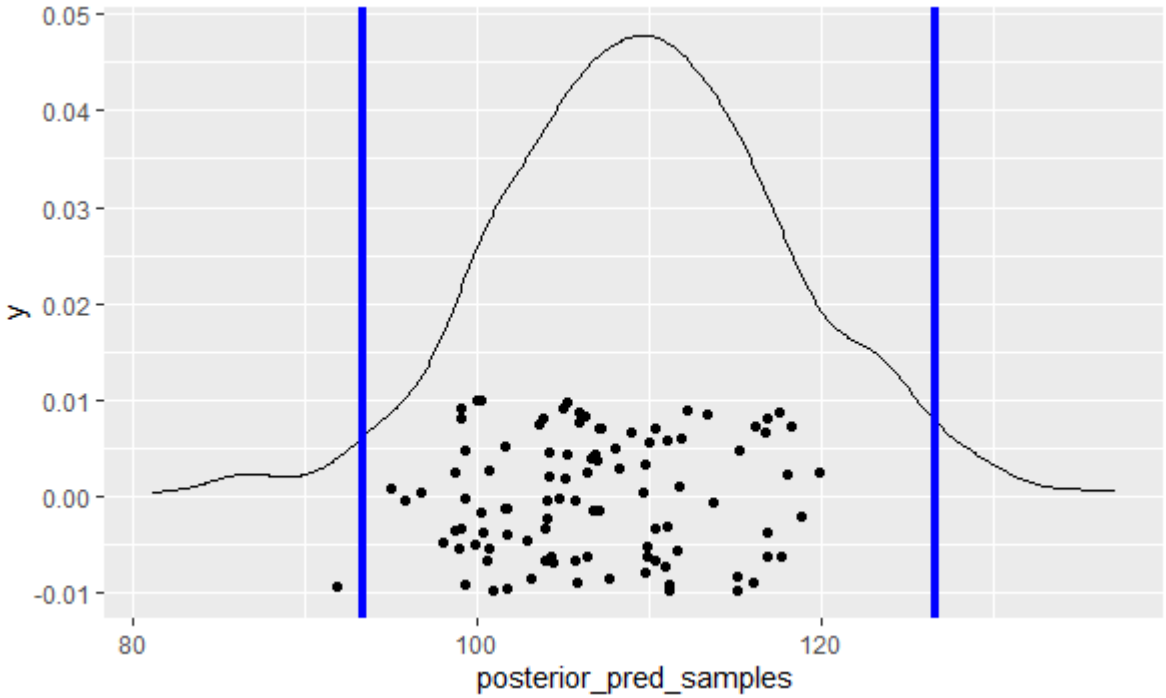
- TOEFL score- $P(X_{1:n}|\mu) \sim N(110, 36)$ &$P(\mu) \sim N(100,25)$



- 95 test samples were inside the credible interval

Approach 1 - Results

- TOEFL - $P(X_{1:n}|\mu) \sim N(110, 36)$ & $P(\mu) \propto 1$



- 99 test samples were inside the credible interval

Approach 1 – Summary - CGPA

Data generating model	Prior	Number of samples inside the credible interval
$P(X_{1:n}\|\mu)\sim N(8.5, 0.35)$	$P(\mu)\sim N(2,1)$	92
$P(X_{1:n}\|\mu)\sim N(8.5, 0.35)$	$P(\mu) \propto 1$	99
$P(X_{1:n}\|\mu)\sim N(8.5, 0.35)$	$P(\mu)\sim N(20,10)$	92
$P(X_{1:n}\|\mu)\sim N(8.5, 0.35)$	$P(\mu)\sim N(200,100)$	92
$P(X_{1:n}\|\mu)\sim N(8.5, 0.35)$	$P(\mu) \propto 100$	99

Approach 1 – Summary - GRE

UNIVERSITY OF CALGARY

Data generating model	Prior	Number of samples inside the credible interval
$P(X_{1:n}\|\mu)\sim N(320,120)$	$P(\mu)\sim N(250,100)$	92
$P(X_{1:n}\|\mu)\sim N(320,120)$	$P(\mu)\propto 1$	100
$P(X_{1:n}\|\mu)\sim N(320,120)$	$P(\mu)\sim N(2.5,1)$	95
$P(X_{1:n}\|\mu)\sim N(320,120)$	$P(\mu)\sim N(2500,1000)$	92
$P(X_{1:n}\|\mu)\sim N(320,120)$	$P(\mu)\propto 100$	100

Approach 1 – Summary - TOEFL

UNIVERSITY OF CALGARY

Data generating model	Prior	Number of samples inside the credible interval
$P(X_{1:n}\|\mu)\sim N(110,36)$	$P(\mu)\sim N(100,25)$	95
$P(X_{1:n}\|\mu)\sim N(110,36)$	$P(\mu)\propto 1$	99
$P(X_{1:n}\|\mu)\sim N(110,36)$	$P(\mu)\sim N(1,0.25)$	96
$P(X_{1:n}\|\mu)\sim N(110,36)$	$P(\mu)\sim N(1000,250)$	95
$P(X_{1:n}\|\mu)\sim N(110,36)$	$P(\mu)\propto 100$	99

Analysis – Metropolis Hastings MCMC

- M – H can be treated with two main ingredients.
- A proposal distribution and an acceptance probability.
- Acceptance probability is given as,

$$\alpha(\theta_{new}, \theta_{t-1}) = min(1, \frac{Posteriorprobabilityof\,\theta_{new}}{Posteriorprobabilityof\,\theta_{t-1}})$$

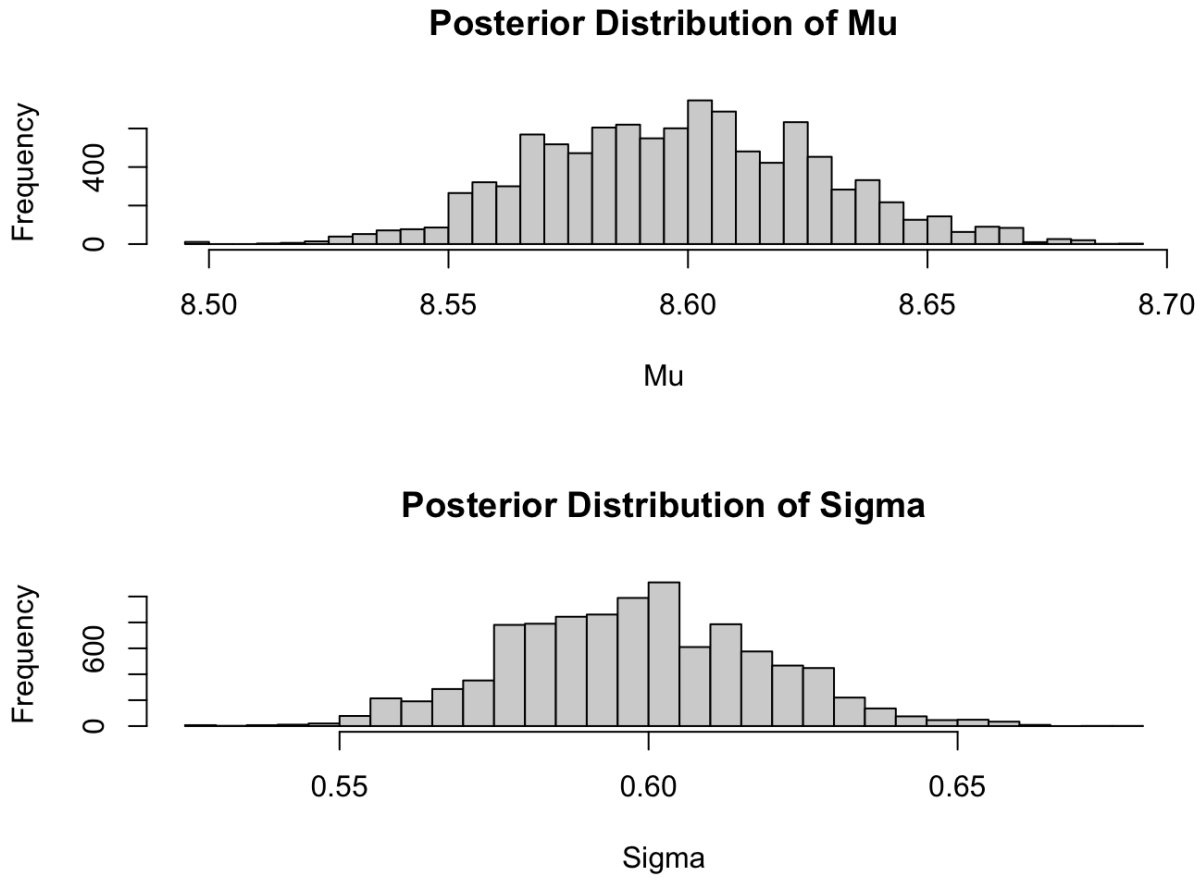
M-H Algorithm

- Initialize θ_0, number of iterations.
- Given the current state θ_t, generate new state θ_{new} from proposal distribution.
- Calculate acceptance probability $\alpha(\theta_{new}, \theta_t)$.
- With probability $\alpha(\theta_{new}, \theta_t)$, set $\theta_{t+1} = \theta_{new}$, else set $\theta_{t+1} = \theta_t$.
- Iterate.
- Result: Realizations of dependent samples $\{\theta_1, \theta_2, \dots\}$ from the target distribution $\pi(\theta)$.

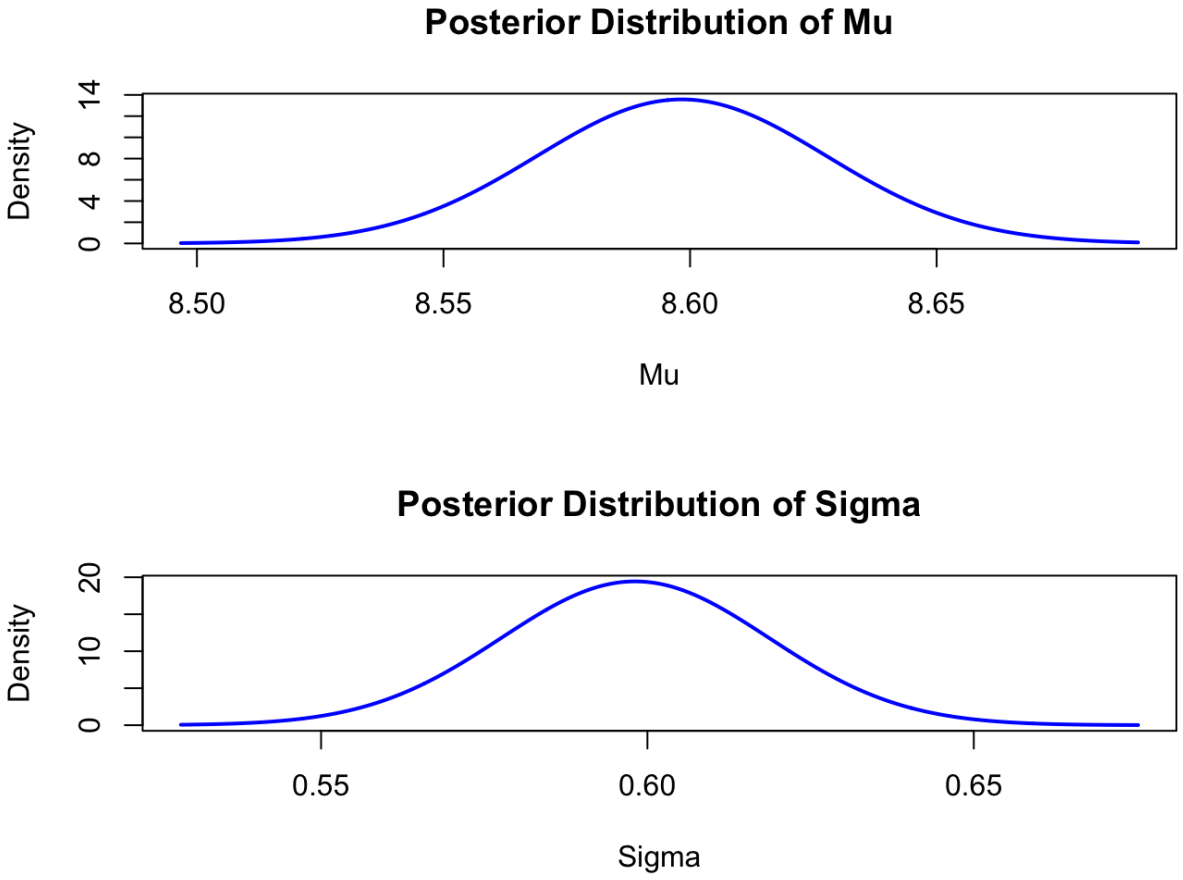
For CGPA parameter

- $Likelihood \sim N(8.59, 0.59)$
- $Prior\ for\ Mean \sim N(8,1)$
- $Prior\ for\ SD \sim N(1,1)$
- Iterations: 10000

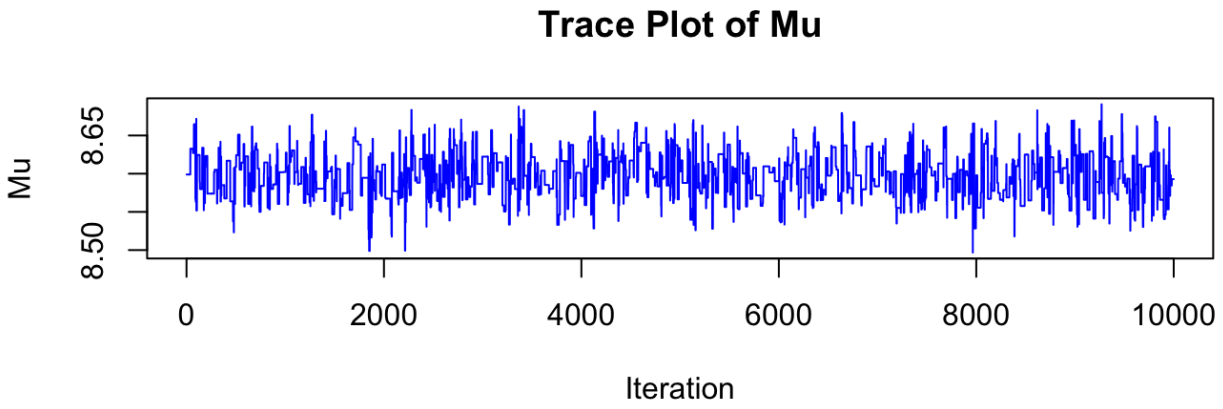
Posterior Distribution (Histogram)

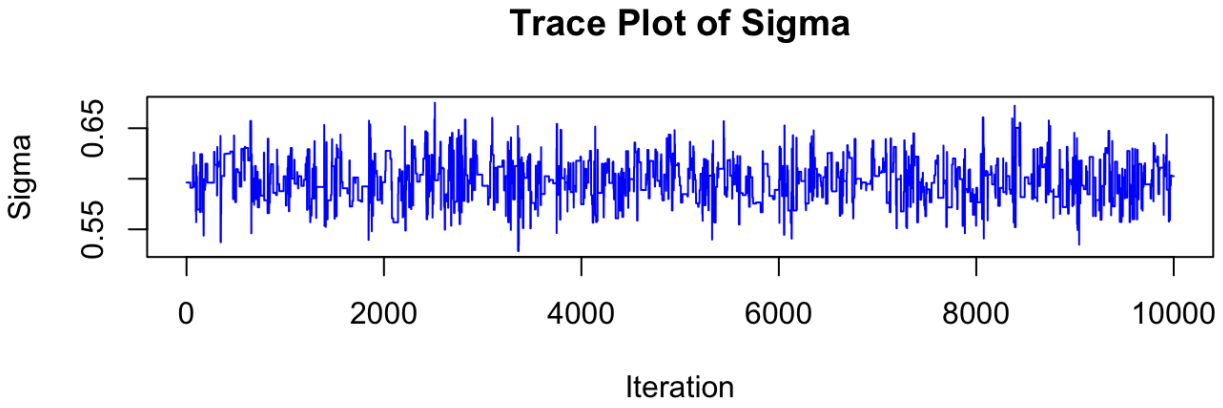


Posterior Distribution (Curve)

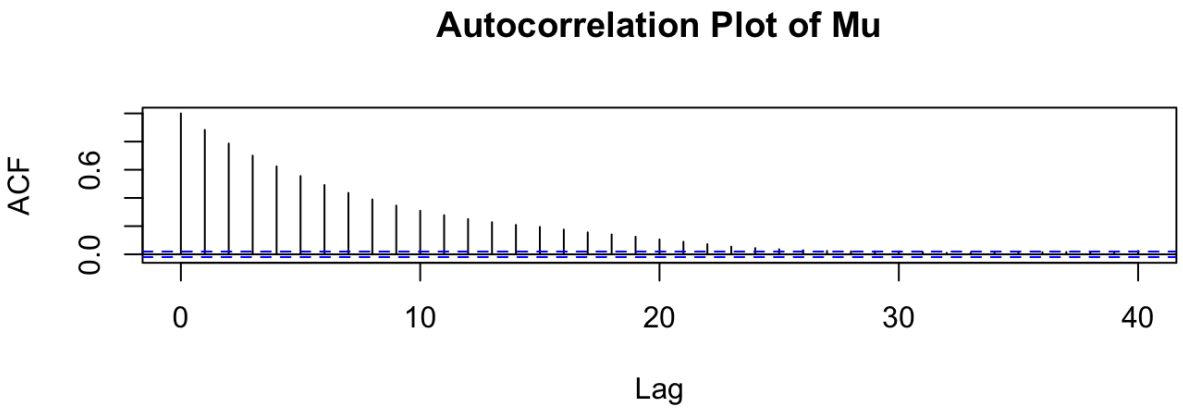


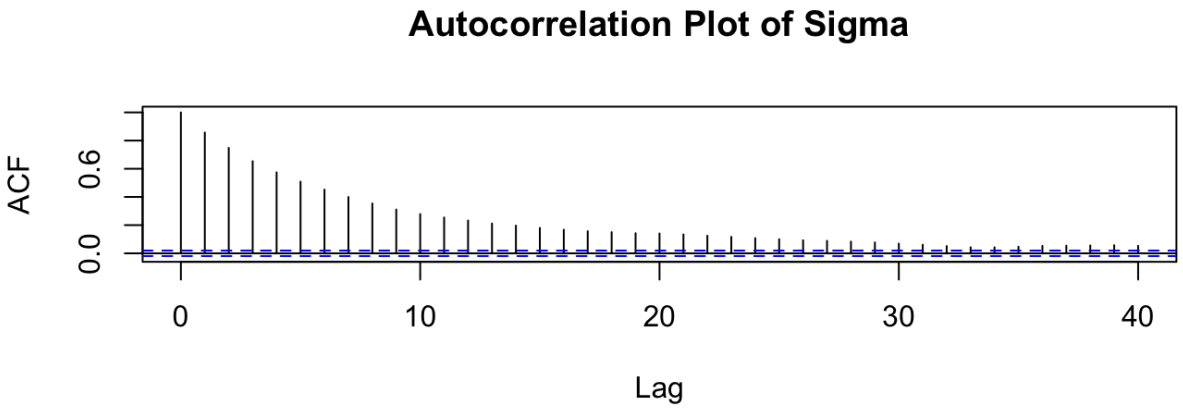
Trace plot


Trace Plot of Mu
Mu
8.50
8.65
0
2000
4000
6000
8000
10000
Iteration


Trace Plot of Sigma
Sigma
0.55
0.65
0
2000
4000
6000
8000
10000
Iteration

Autocorrelation plot

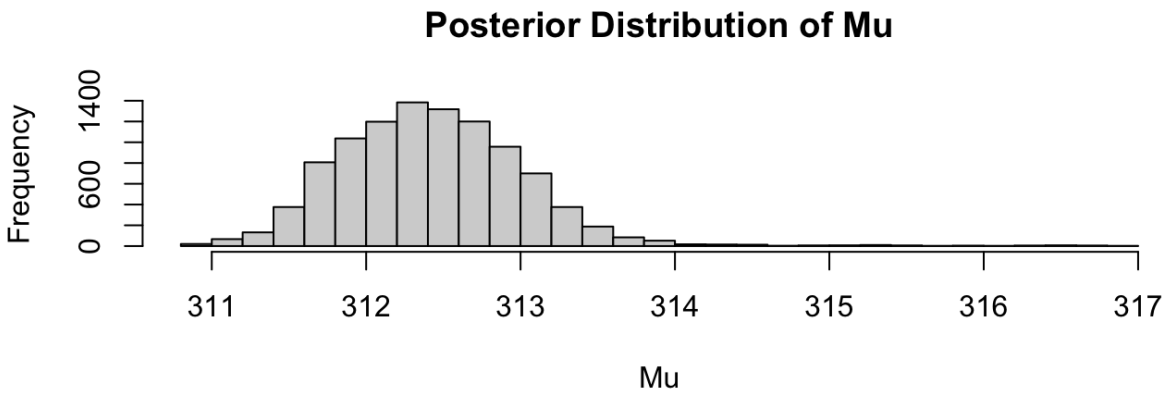


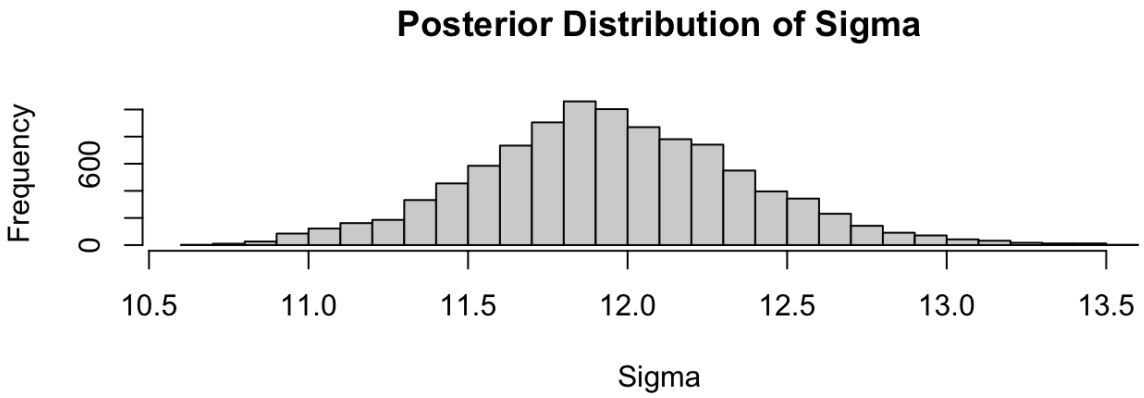


For GRE score parameter

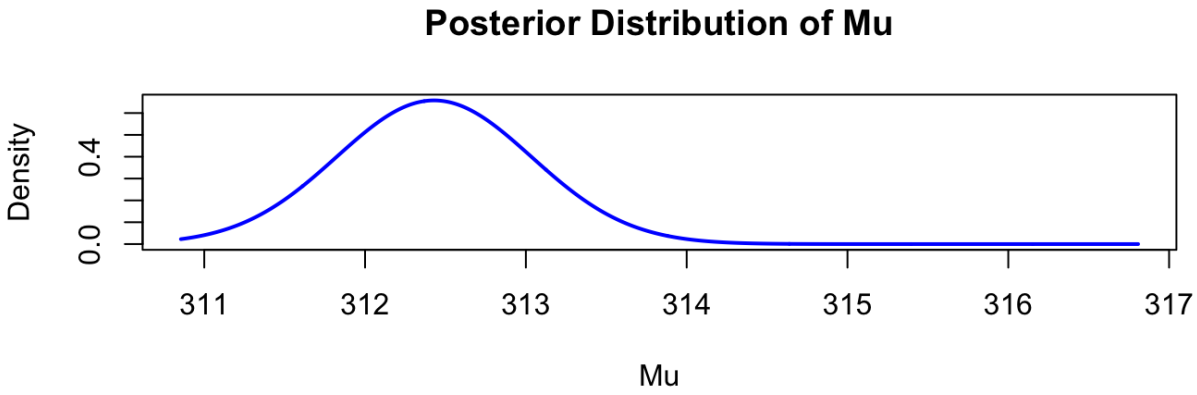
- $Likelihood \sim N(316.81, 11.47)$
- $Prior\ for\ Mean \sim N(300, 1)$
- $Prior\ for\ SD \sim N(10, 1)$
- Iterations: 10000

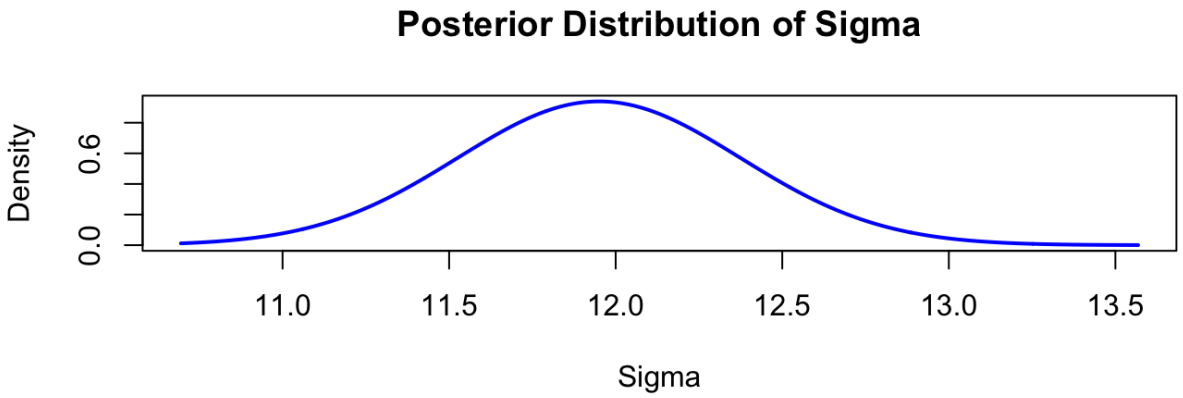
Posterior Distribution (Histogram)



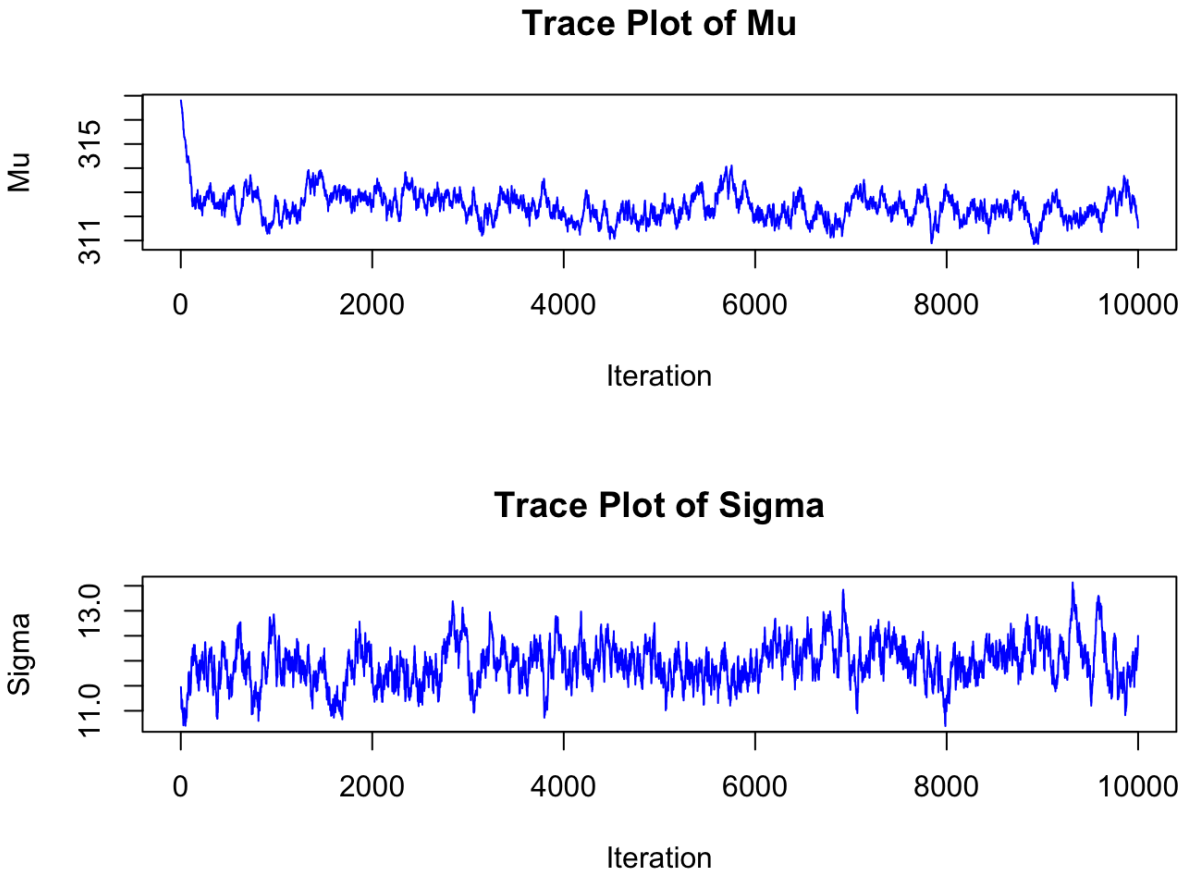


Posterior Distribution (Curve)

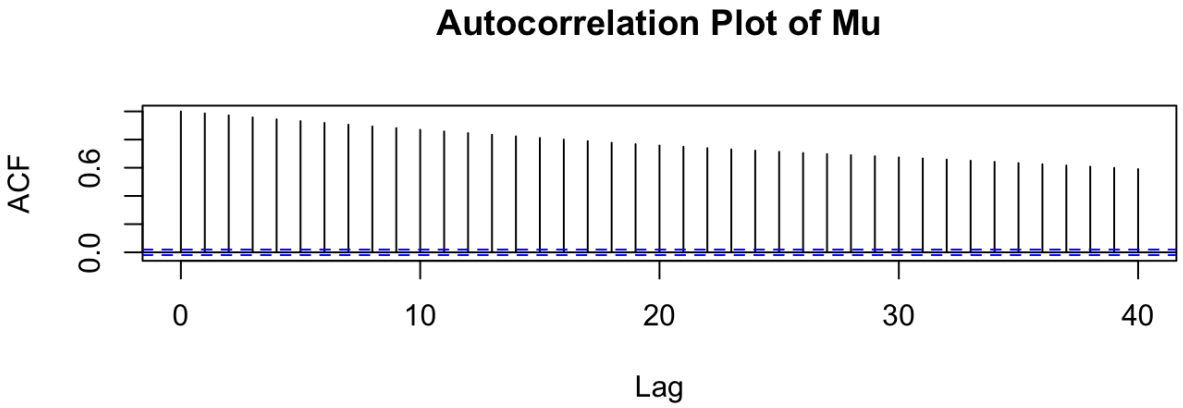


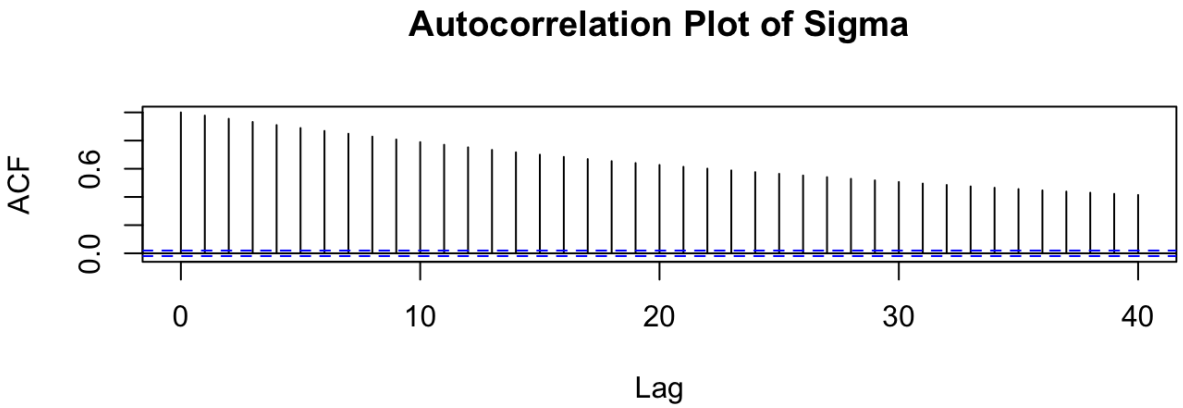


Trace plot


Trace Plot of Mu
Mu
311
315
0
2000
4000
6000
8000
10000
Iteration
Trace Plot of Sigma
Sigma
11.0
13.0
0
2000
4000
6000
8000
10000
Iteration

Autocorrelation plot

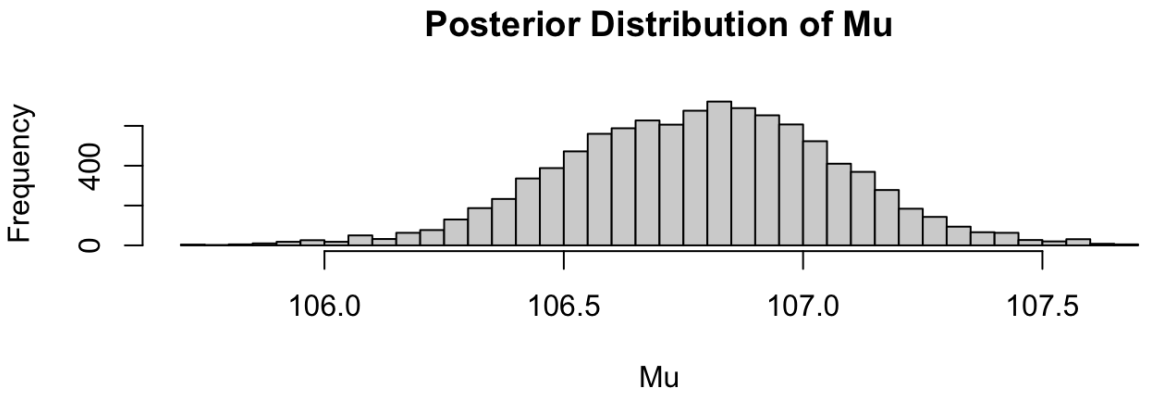


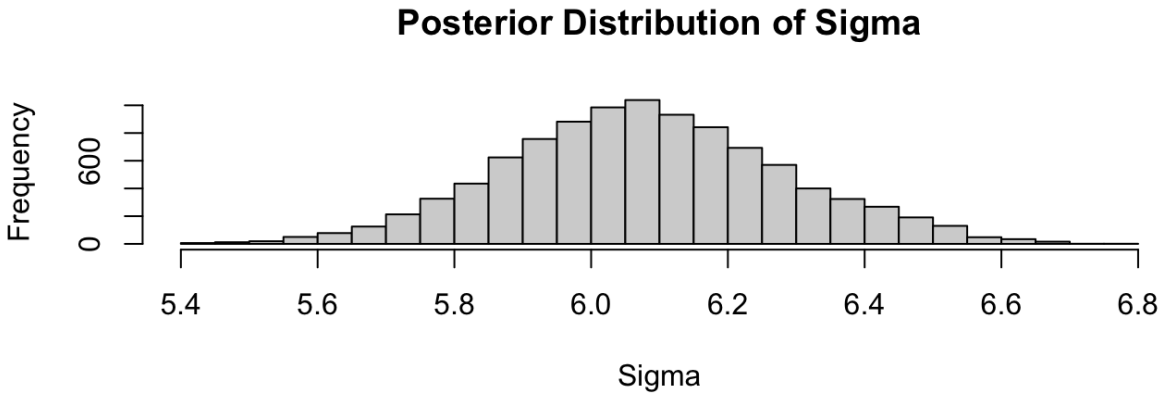


For TOEFL score parameter

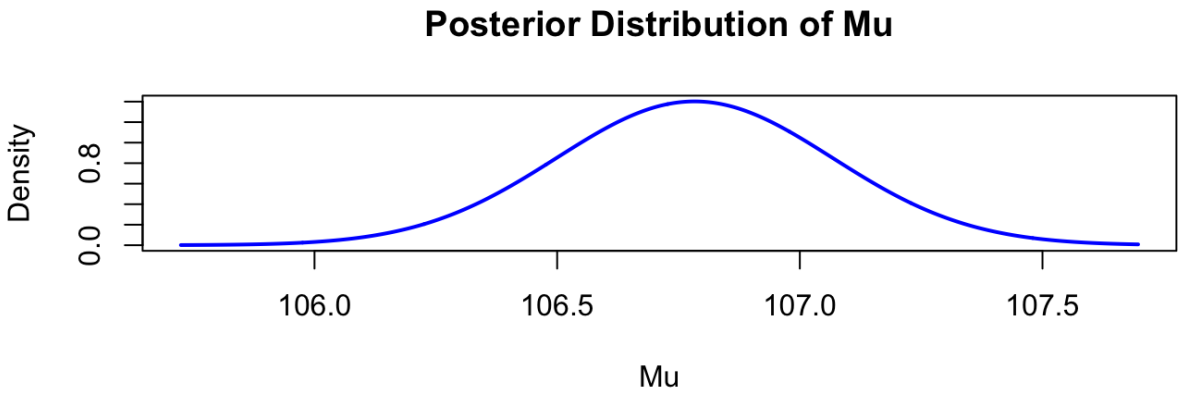
- $Likelihood \sim N(107.41, 6.07)$
- $Prior\ for\ Mean \sim N(100, 1)$
- $Prior\ for\ SD \sim N(5, 1)$
- Iterations: 10000

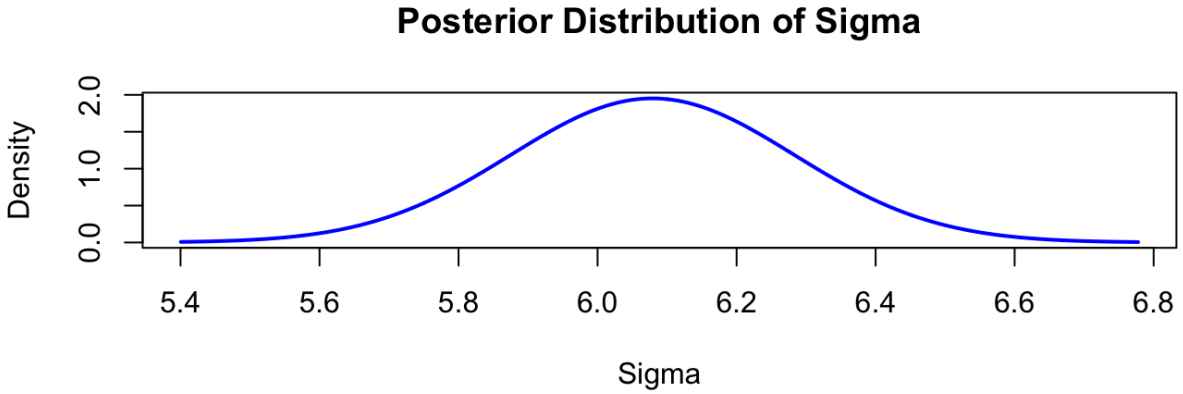
Posterior Distribution (Histogram)



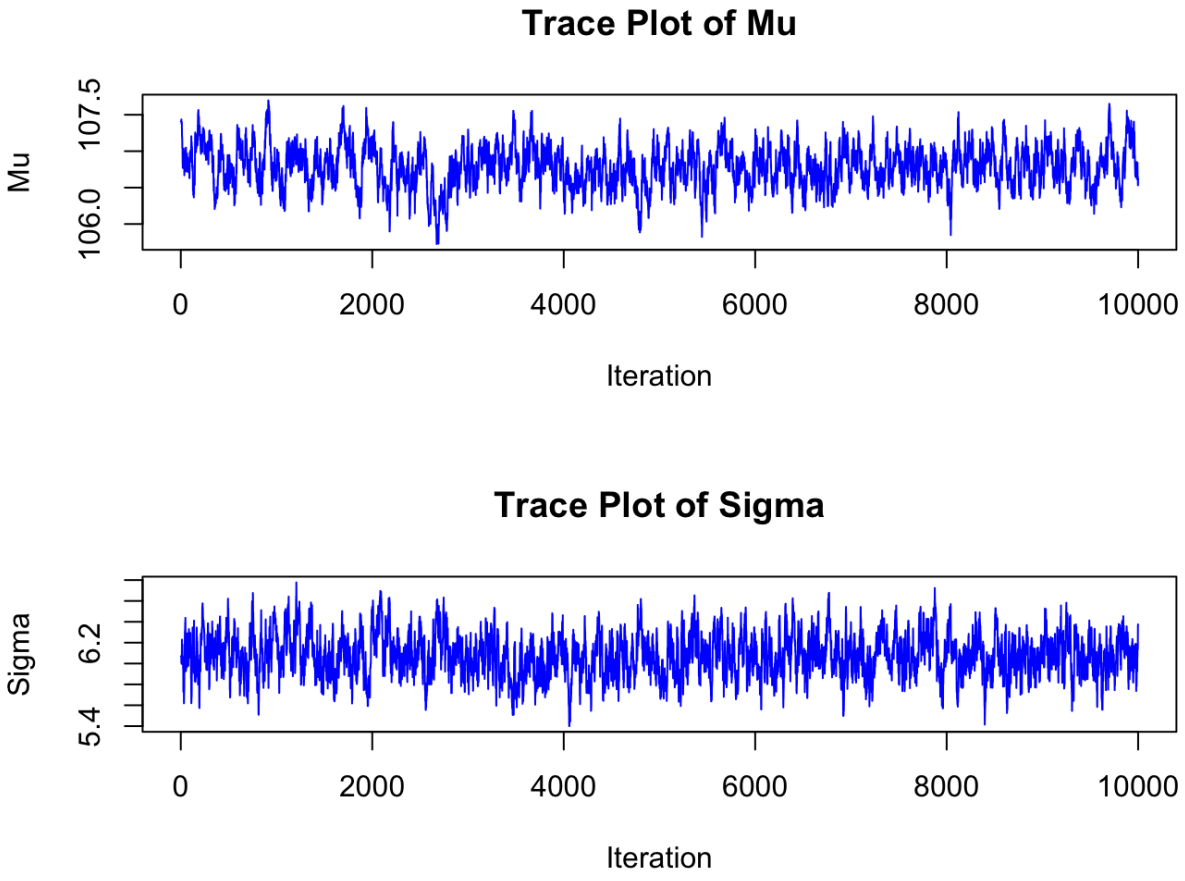


Posterior Distribution (Curve)

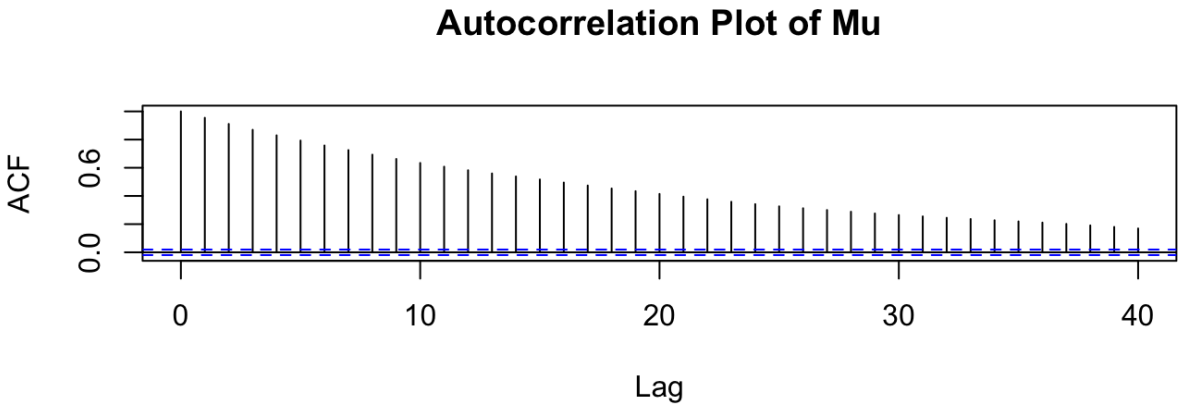


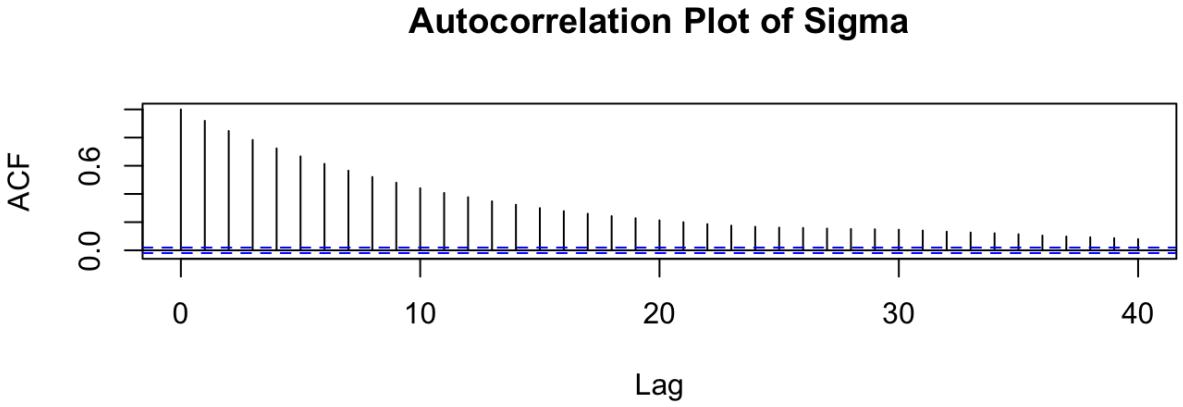


Trace plot


Trace Plot of Mu
Mu
106.0
107.5
0
2000
4000
6000
8000
10000
Iteration
Trace Plot of Sigma
Sigma
5.4
6.2
0
2000
4000
6000
8000
10000
Iteration

Autocorrelation plot





Analysis – Gibbs Sampling

UNIVERSITY OF CALGARY

- Gibbs sampler is an example of the Markov Chain - Monte Carlo (MCMC) technique used to estimate Bayesian models when analytical solution is not feasible.
- Prior distributions reflect your knowledge of the phenomenon prior to the experiment. They are part of the model.
- Initial values are part of the solution (MCMC algorithm) and tell the algorithm where to start looking for the posterior distribution. Initial values can be based on the data. The prior distributions cannot.

Gibbs Algorithm

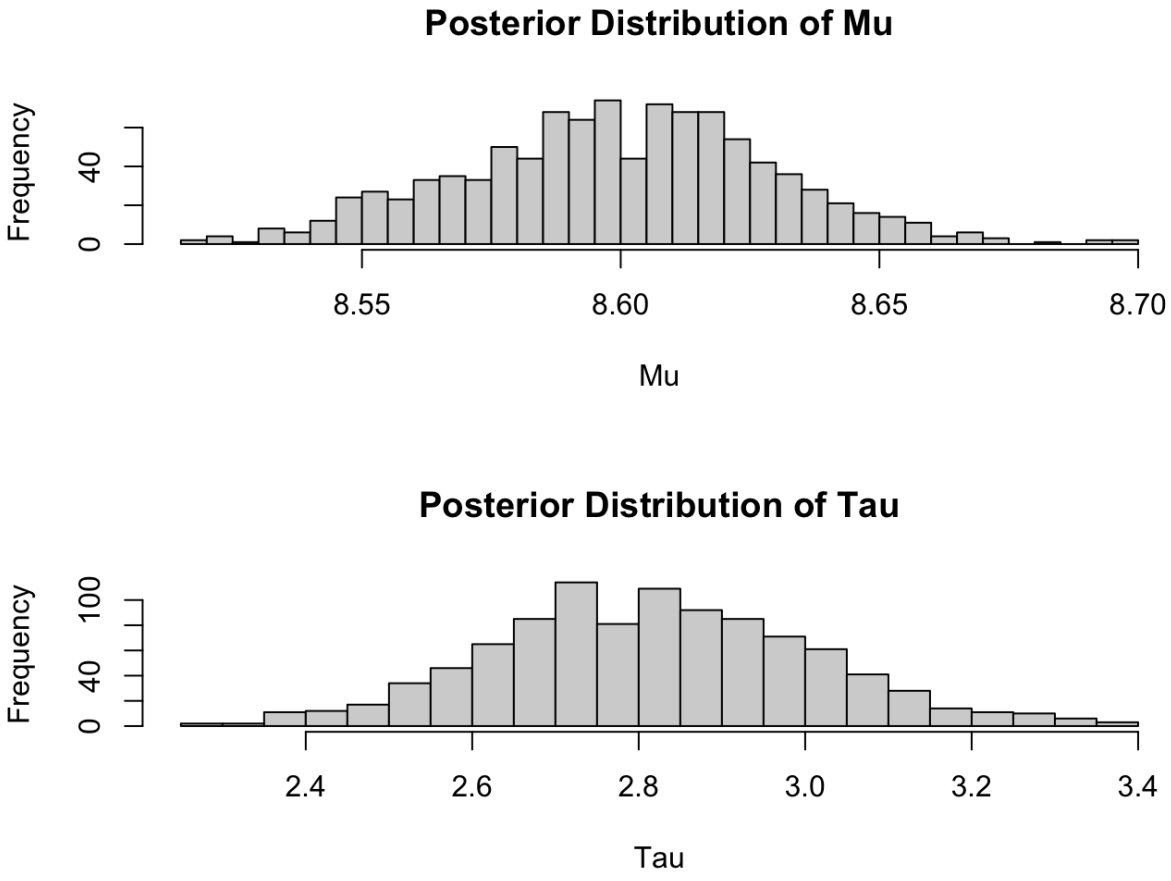
UNIVERSITY OF CALGARY

- Assign some starting values μ^* and τ^* to the parameters of interest.
- Given $\tau = \tau^*$, sample the new value of μ^* from a normal distribution with mean $\frac{n\bar{y}\tau^* + \mu_0\tau_0}{n\tau^* + \tau_0}$ and precision (inverse variance) $n\tau^* + \tau_0$.
- Given $\mu = \mu^*$, sample the new value of τ^* from a gamma distribution with parameters $\alpha + \frac{n}{2}$ and $\frac{n\bar{y}\tau^* + \mu_0\tau_0}{n\tau^* + \tau_0}$.
- Iterate step 2 and 3 many times.

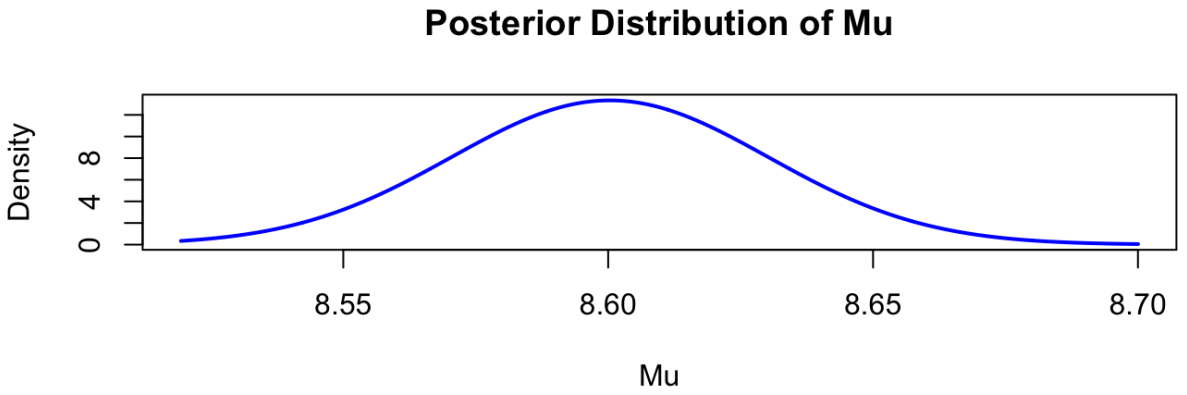
For CGPA score parameter

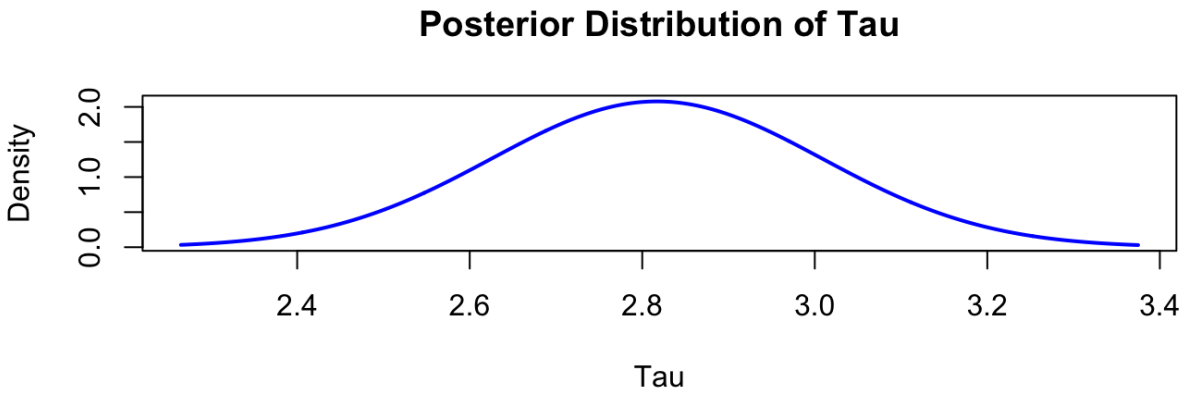
- $Likelihood \sim N(\mu, \tau)$
- $Prior\ for\ Mean \sim N(8, 2.87)$
- $Prior\ for \tau \sim Gamma(0.01, 0.01)$
- Iterations: 1000

Posterior Distribution (Histogram)

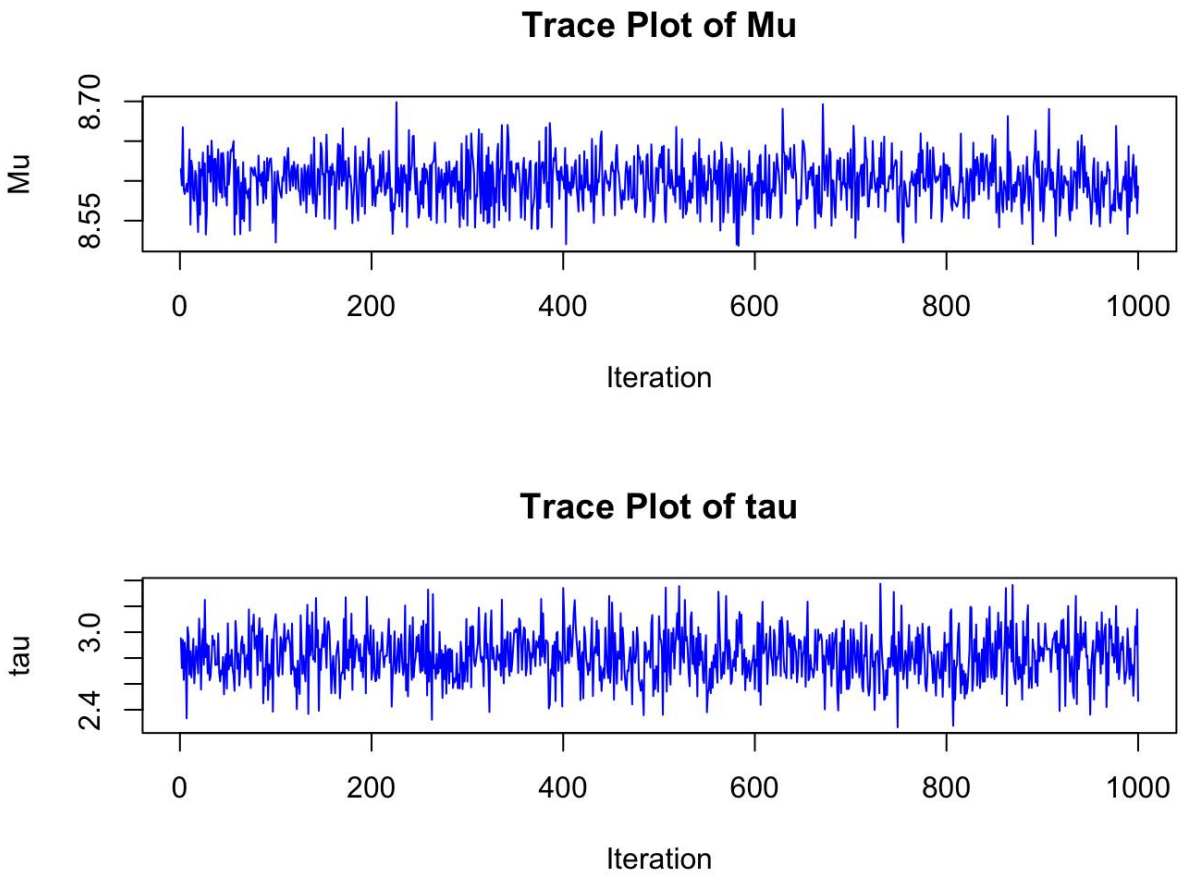


Posterior Distribution (Curve)





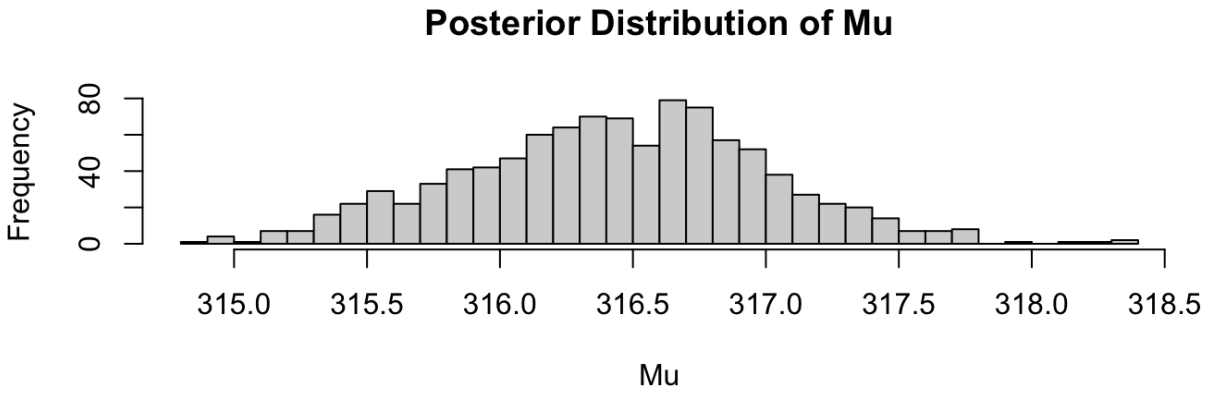
Trace plot

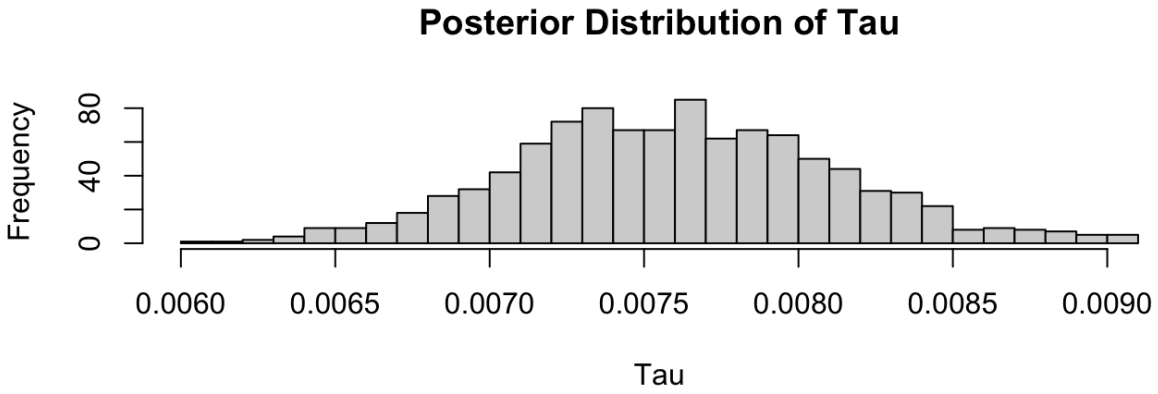

Trace Plot of Mu
Mu
8.55
8.70
0
200
400
600
800
1000
Iteration
Trace Plot of tau
tau
2.4
3.0
0
200
400
600
800
1000
Iteration

For GRE score parameter

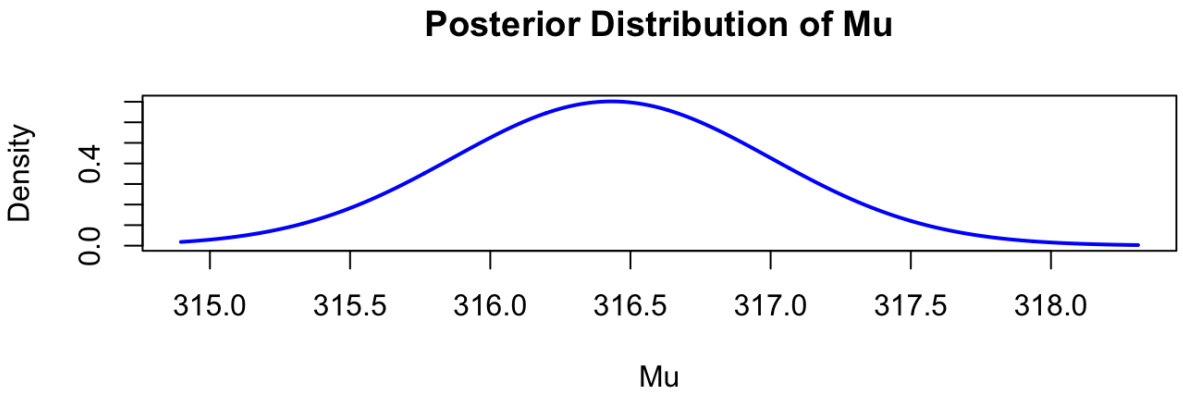
- $Likelihood \sim N(\mu, \tau)$
- $Prior\ for\ Mean \sim N(300, 0.076)$
- $Prior\ for \tau \sim Gamma(0.01, 0.01)$
- Iterations: 1000

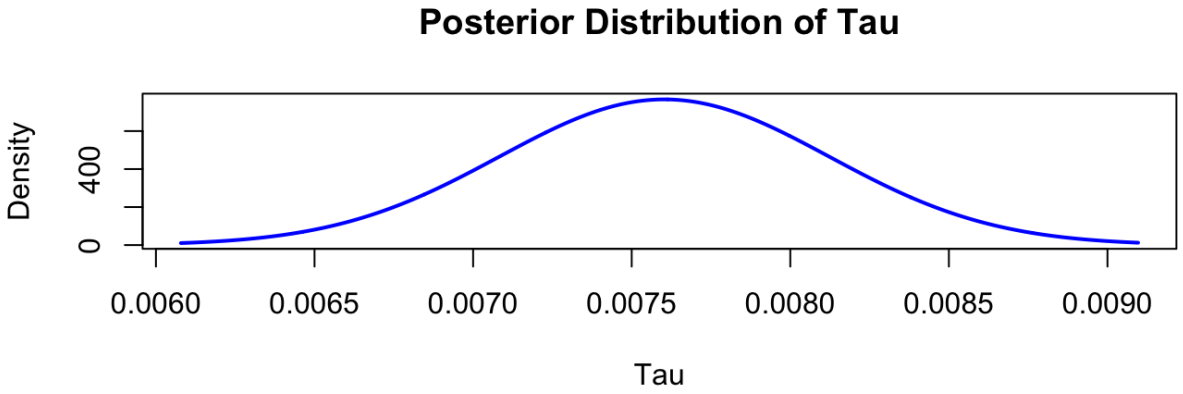
Posterior Distribution (Histogram)



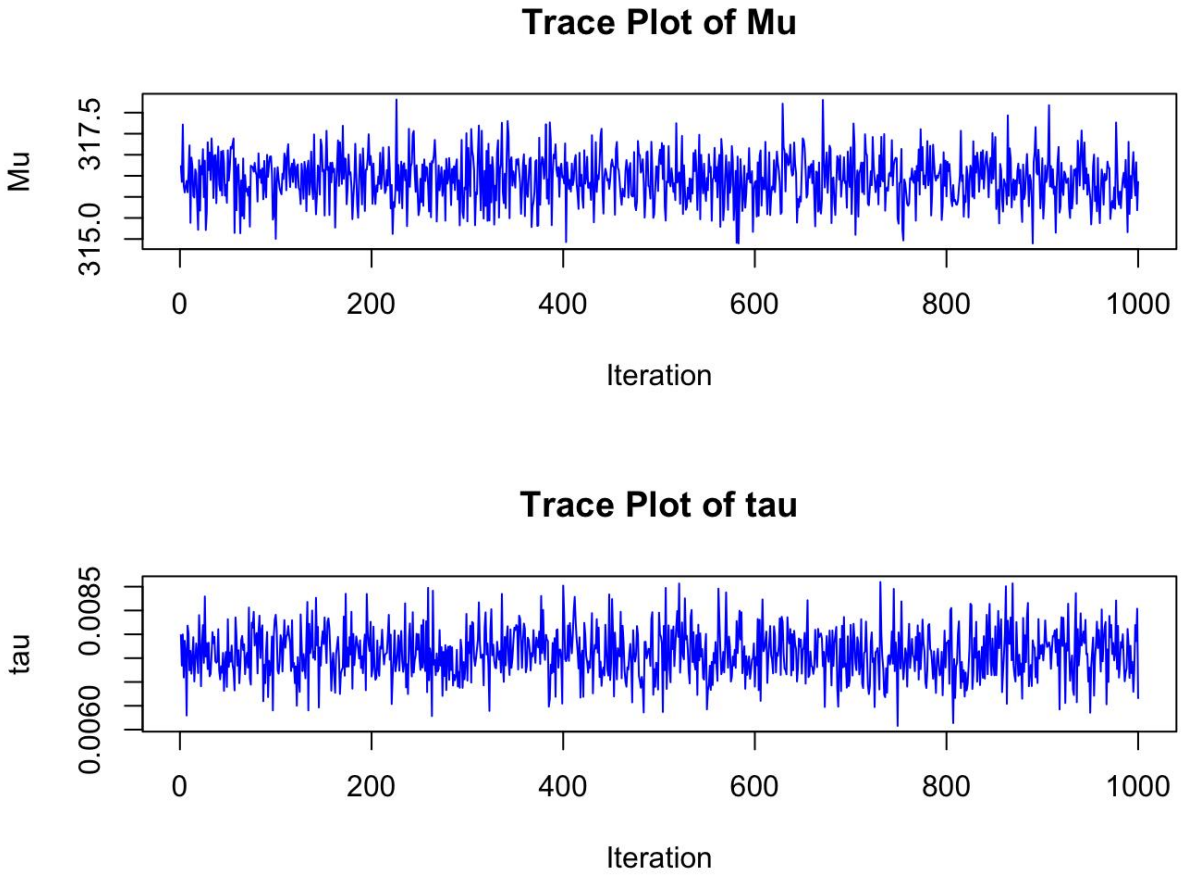


Posterior Distribution (Curve)





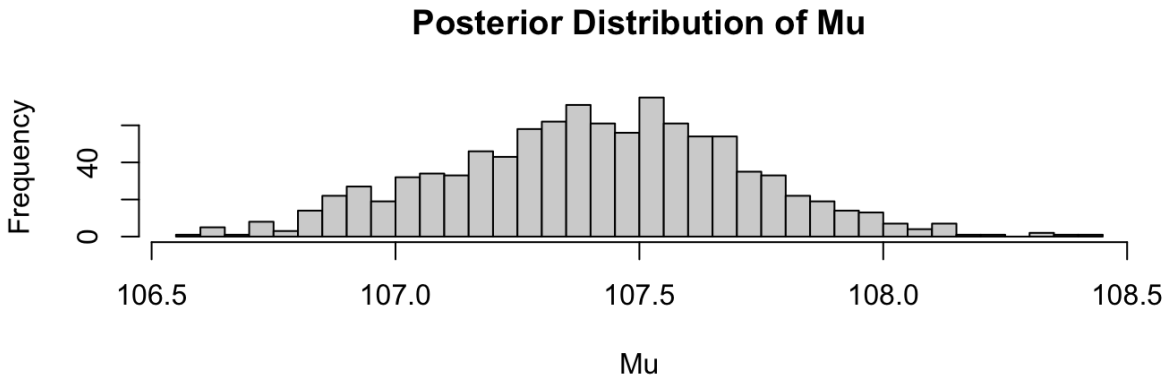
Trace plot

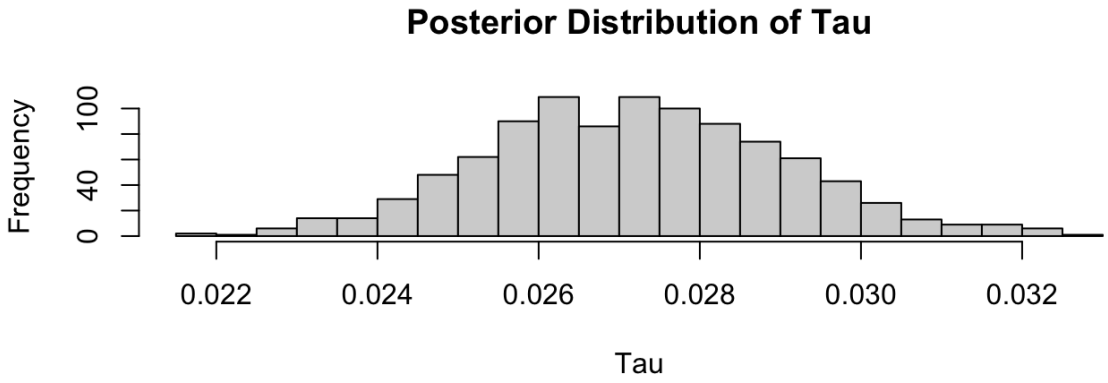

Trace Plot of Mu
Mu
315.0
317.5
0
200
400
600
800
1000
Iteration
Trace Plot of tau
tau
0.0060
0.0085
0
200
400
600
800
1000
Iteration

For TOEFL score parameter

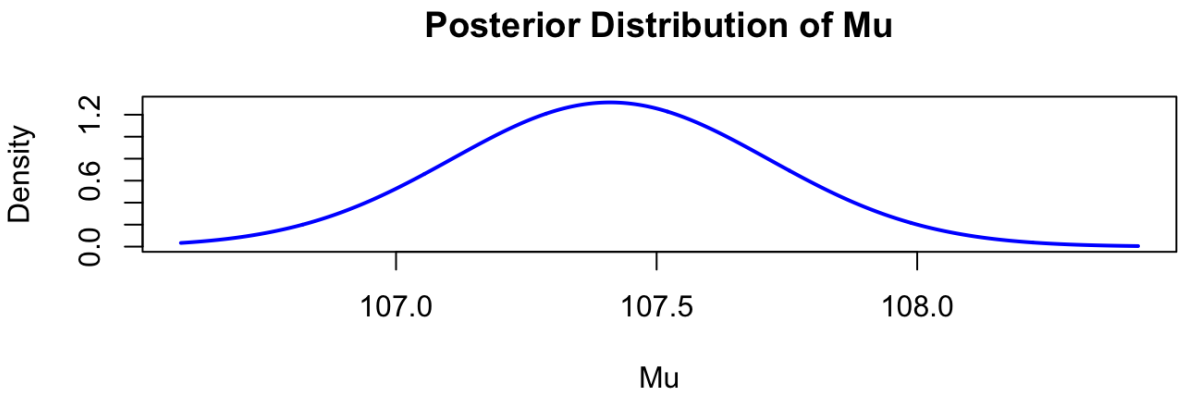
- $Likelihood \sim N(\mu, \tau)$
- $Prior\ for\ Mean \sim N(100, 0.028)$
- $Prior\ for \tau \sim Gamma(0.01, 0.01)$
- Iterations: 1000

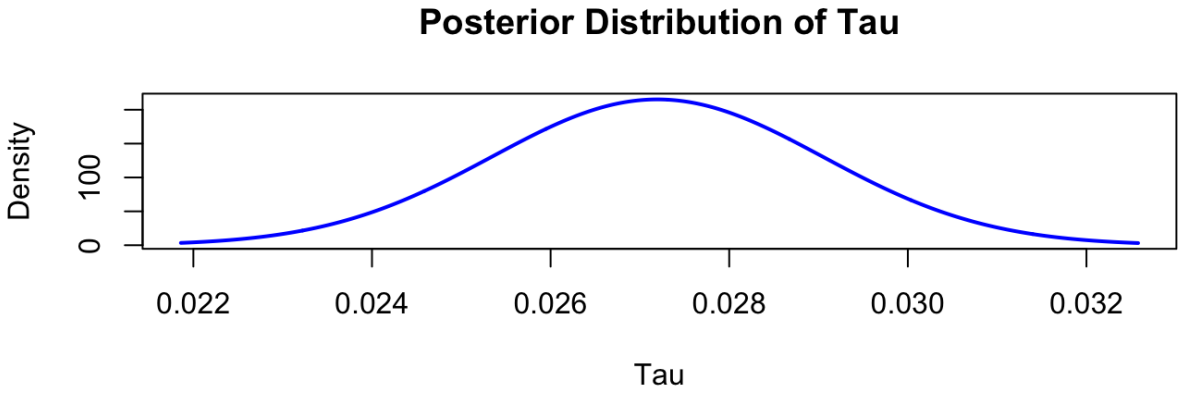
Posterior Distribution (Histogram)



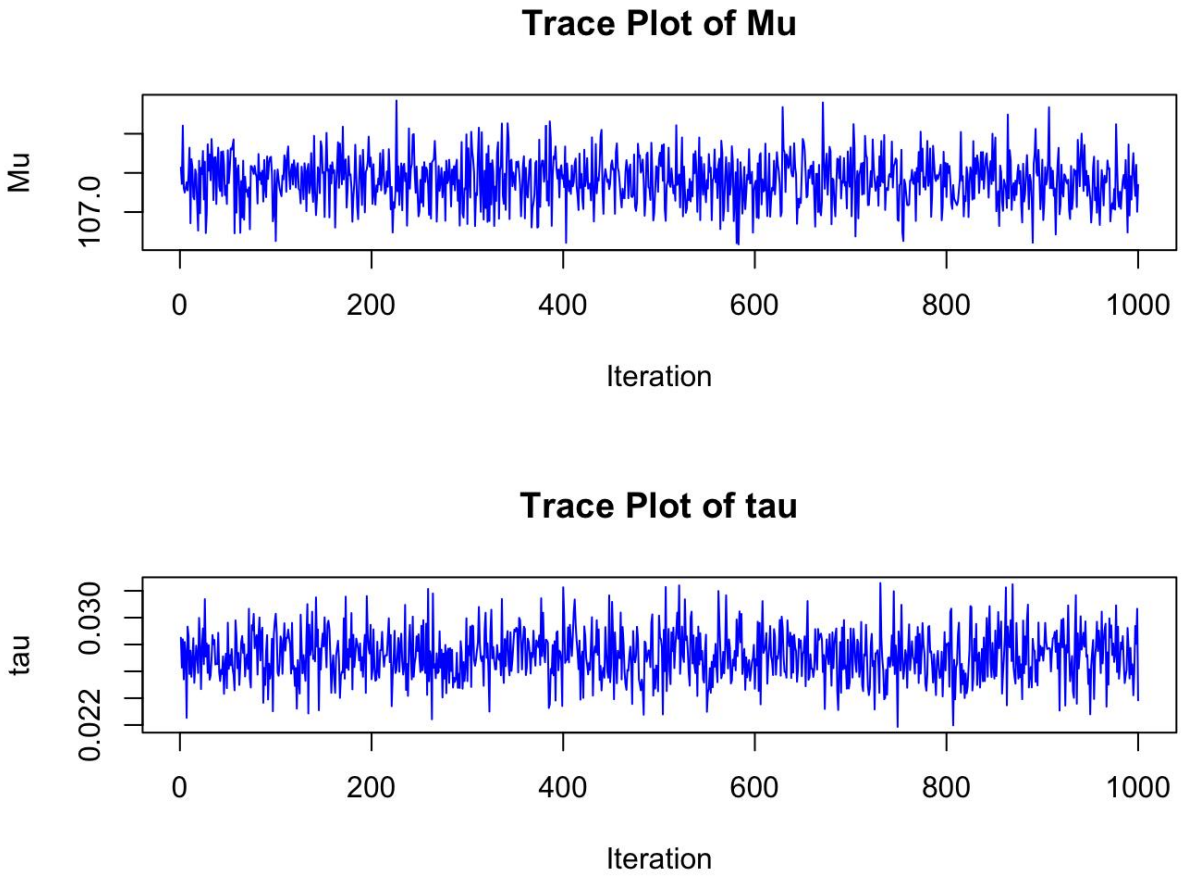


Posterior Distribution (Curve)





Trace plot


Trace Plot of Mu
Mu
107.0
0
200
400
600
800
1000
Iteration
Trace Plot of tau
tau
0.022
0.030
0
200
400
600
800
1000
Iteration

Reference

- Bayesian data analysis, 3rd edition. Gelman et al.

- Admission Prediction in Undergraduate Applications: an Interpretable Deep Learning Approach . Amisha P. & Barbara M.

UNIVERSITY OF CALGARY

Thank You

Adeepa Gustinna Wadu
Aash Makwana