# Introduction to Bayesian Computing

Dipali Vasudev Mestry  Amiya Ranjan Bhowmick

September 23, 2024

# Table of contents

# Preface

Bayesian statistical methods are becoming increasingly popular across all branches of science. With the rapid development of statistical software, Bayesian computations are now accessible to researchers across various domains. These methods are routinely used by practitioners in both industry and academia. However, with the growing availability of software and packages, the fundamental understanding of Bayesian estimation is sometimes compromised.

In this document, we aim to bridge that gap by offering a clear understanding of the Bayesian principles through practical examples and case studies. By simulating data and comparing Bayesian estimates with likelihood-based estimates, we focus on the core concept of bias-variance decomposition of the mean square error (MSE) when evaluating estimators.

This work is the outcome of a series of lectures delivered by the author, Amiya Ranjan Bhowmick, during the summer vacation of 2018 at the Institute of Chemical Technology (ICT), Mumbai. This document would be incomplete without acknowledging the influence of two exceptional books: `Statistical Inference` (Casella and Berger 2002) and `All of Statistics` (Wasserman 2004). These texts provided the foundation for much of the material discussed, and it would have been impossible to develop this work without their insights. We studied many examples and exercises from these books and have put our understanding into words here. And, certainly mistake is a part of our life. We would be grateful if the mistakes are informed to us.

# 1 Introduction

**Bayesian statistics** has emerged as a powerful tool in the realm of statistical analysis, providing a flexible framework for making inferences under uncertainty. The foundation of Bayesian methods dates back to the $18^{\text{th}}$ century when Reverend Thomas Bayes proposed a theorem for updating beliefs with new evidence. This approach offers a probabilistic paradigm that allows the incorporation of prior knowledge, leading to more informed decision-making.

Bayesian analysis, rooted in Bayesian statistical methods, aims to make inferences about **parameters**. Parameters are the quantities we seek to estimate or infer from our data. They represent the underlying characteristics or properties of the system or process under investigation. For instance, in a model predicting the probability of species extinction, the parameter of interest might be the actual likelihood of such an extinction event occurring. Mathematically, the parameter(s) is denoted by $\theta$.

At the heart of Bayesian analysis are three fundamental components: **prior**, **likelihood**, and **posterior**. These elements work together to form a cohesive system that informs probabilistic reasoning. Let us understand the what exactly these terms are in Bayesian statistics;

## Prior

The **prior distribution** or simply **prior** is a probability distribution that represents our initial beliefs or knowledge about the parameters before observing any data. It reflects what we think the values of the parameters could be based on prior experience, expert knowledge, or sometimes, a deliberate choice to remain neutral. Priors can take many forms, from non-informative or vague, which show that we have little or no prior knowledge about the parameter, to informative, which incorporate strong prior knowledge. In this book we denote the prior distribution of parameter $\theta$ by $\pi(\theta)$.

## Likelihood

The **likelihood function** is derived from the data and describes the probability of the observed data given the model parameters. It reflects how well different parameter values explain the observed data. In Bayesian inference, the likelihood is combined with the prior to update our beliefs about the parameters.

Mathematically, if we denote the parameter of interest by $\theta$ and the observed data by $y$, the likelihood function is given by $p(y|\theta)$, which measures the plausibility of $\theta$ given the data.

## Posterior

The **posterior distribution** denoted by $p(\theta|y)$ is the result of updating the prior with the observed data through the likelihood. It is the core output of a Bayesian analysis and represents our updated beliefs about the parameter(s) after accounting for the data.

Using Bayes' Theorem, the posterior is expressed as:

$$p(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{p(y)},$$

where $p(\theta|y)$ is the posterior, $p(y|\theta)$ is likelihood, $\pi(\theta)$ is the prior, and $p(y)$ is the marginal likelihood, which normalize the posterior distribution. The posterior reflects both the information contained in the prior and the evidence from the data, thus offering a balanced view of parameter uncertainty.

## Improper prior

In some cases, a prior distribution $\pi(\theta)$ may not integrate to one, which is referred to as an improper prior:

$$\int_{-\infty}^{\infty} \pi(\theta)\, d\theta = \infty$$

These priors are often used when we want a non-informative or vague prior, such as in situations where no clear prior knowledge exists. Improper priors can sometimes lead to improper posteriors, but when handled correctly, they can still be useful, particularly when the likelihood is highly informative.

For example, a flat prior over an infinite range is an improper prior because it doesn't integrate to a finite value. However, it is often employed when the intention is to let the data dominate the inference. Let us understand it through one example;

A common example of an improper prior is the **uniform prior** over the entire real line:

$$\pi(\theta) = 1$$

for all $\theta \in \mathbb{R}$. This prior is "improper" because:

$$\int_{-\infty}^{\infty} \pi(\theta)\, d\theta = \int_{-\infty}^{\infty} 1\, d\theta = \infty.$$

For the posterior distribution to be proper, the integral of the posterior distribution over all possible values of $\theta$ must be finite and equal to one:

$$\int_{-\infty}^{\infty} \pi(\theta|y)\, d\theta = 1.$$

If the likelihood function $P(y|\theta)$ is such that it ensures this condition, then the improper prior can be used effectively.

## Conjugate prior

Given a likelihood function $p(y \mid \theta)$, if the posterior distribution $p(\theta \mid y)$ belongs to the same family of probability distributions as the prior distribution $p(\theta)$, then the prior and posterior are referred to as conjugate distributions with respect to that likelihood function. In this case, the prior is called a **conjugate prior** for the likelihood function $p(y \mid \theta)$.

In this introductory chapter, we have explored the foundational concepts of Bayesian statistics, including the role of parameters, prior distributions, likelihoods, and the posterior distribution. We have also introduced the idea of improper and conjugate priors. With these core principles in mind, we are now ready to move from theory to practice.

# 2 Illustrative Examples in practice

In this chapter, we will explore five commonly used examples in Bayesian inference. Each example is designed to show, step-by-step, how to calculate the posterior distribution, find the Bayesian estimate of the parameter, and compute the Mean Square Error (MSE) of the estimate.

These examples have been selected to clearly demonstrate how Bayesian methods work in practice for teaching purposes. By following these examples, one can gain a better understanding of how to apply Bayesian techniques and assess the accuracy of the estimates. The detailed calculations and used codes are included to make learning easier and more practical.

## 2.1 Example - I (Poisson rate parameter with gamma prior)

Let $X_1, X_2, \cdots, X_n$ be a random sample of size $n$ from a population following Poisson($\lambda$) distribution. Suppose $\lambda$ have a gamma($\alpha$, $\beta$) distribution, which is the conjugate family for Poisson. So, the statistical model has the following hierarchy:

$$\begin{aligned} X_i|\lambda &\sim \text{Poisson}(\lambda), \quad i = 1, 2, \cdots, n, \\ \lambda &\sim \text{gamma}(\alpha, \beta). \end{aligned}$$

Based on the observed sample, we are interested to estimate the mean of the population, that is the value of $\lambda$. The prior distribution of $\lambda$, $\pi(\lambda)$, is given by

$$\pi(\lambda) = \frac{\lambda^{\alpha-1} e^{-\frac{\lambda}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}, \quad \lambda > 0,$$

where $\alpha$ and $\beta$ are positive constants. First, we shall obtain the posterior distribution of $\lambda$. If there were no prior information available about the parameter $\lambda$, then we could use the sample mean $\overline{X}$ to estimate it. However, the exact sampling distribution of $\overline{X}$ is not known in this case. Therefore, we begin with a statistic $Y = \sum_{i=1}^{n} X_i$, whose sampling distribution is known to be Poisson($n\lambda$) (denoted by $f(y|\lambda)$). The posterior distribution, the conditional distribution of $\lambda$ given the sample, $X_1, X_2, \cdots, X_n$, that is, given $Y = y$, is

$$\pi(\lambda|y) = \frac{f(y|\lambda)\pi(\lambda)}{m(y)}, \tag{2.1}$$

where $m(y)$ is the marginal distribution of $Y$ and can be calculated as follows;

$$\begin{aligned}
m(y) &= \int_0^\infty f(y|\lambda)\pi(\lambda)\mathrm{d}\lambda \\
&= \int_0^\infty \frac{e^{-n\lambda}(n\lambda)^y}{y!} \cdot \frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}}{\Gamma(\alpha)\beta^\alpha}\mathrm{d}\lambda \\
&= \frac{n^y}{y!\Gamma(\alpha)\beta^\alpha}\int_0^\infty e^{-\lambda\left(n+\frac{1}{\beta}\right)}\lambda^{y+\alpha-1}\mathrm{d}\lambda \\
&= \frac{n^y\Gamma(y+\alpha)}{y!\Gamma(\alpha)\beta^\alpha\left(y+\frac{1}{\beta}\right)^{y+\alpha}}\int_0^\infty \frac{e^{-\lambda\left(n+\frac{1}{\beta}\right)}\lambda^{y+\alpha-1}\left(n+\frac{1}{\beta}\right)^{y+\alpha}}{\Gamma(y+\alpha)}\mathrm{d}\lambda \\
&= \frac{n^y\Gamma(y+\alpha)}{y!\Gamma(\alpha)\beta^\alpha\left(n+\frac{1}{\beta}\right)^{y+\alpha}}, \quad y = 0, 1, 2, \cdots.
\end{aligned}$$

In the above, the integrand is the kernel of the gamma $\left(y+\alpha, \left(n+\frac{1}{\beta}\right)^{-1}\right)$ density function, hence integrated out to be 1. Now, the posterior density $\pi(\lambda|y)$ is given by

$$\begin{aligned}
\pi(\lambda|y) &= \frac{e^{-n\lambda}(n\lambda)^y}{y!} \cdot \frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}}{\Gamma(\alpha)\beta^\alpha} \cdot \frac{y!\Gamma(\alpha)\beta^\alpha\left(n+\frac{1}{\beta}\right)^{y+\alpha}}{n^y\Gamma(y+\alpha)} \\
&= \frac{\left(n+\frac{1}{\beta}\right)^{y+\alpha}e^{-\lambda\left(n+\frac{1}{\beta}\right)}\lambda^{y+\alpha-1}}{\Gamma(y+\alpha)}, \quad 0 < \lambda < \infty.
\end{aligned}$$

Hence, as expected, the posterior distribution of $\lambda$,

$$\lambda|y \sim \text{gamma}\left(y+\alpha, \left(n+\frac{1}{\beta}\right)^{-1}\right).$$

It also verifies the claim that the gamma($\alpha$,$\beta$) is the conjugate family for Poisson. A closure look in the above calculations reveals that the steps can be heavily reduced, in fact the explicit expression for $m(y)$ is not at all required. Since, it does not depend of $\lambda$, it is a constant, that makes the integral $\int_0^\infty \pi(\lambda|y)\mathrm{d}\lambda$ to be equal to 1, so that it becomes a valid probability density function. Thus, appearance of the posterior density in the integrand of the marginal is not a magic. We shall use this posterior distribution of $\lambda$ to make statements about the parameter $\lambda$. The mean of the posterior can be used as a point estimate of $\lambda$. So, the Bayesian estimator

of $\lambda$ (call it $\hat{\lambda}_B$) is given by $E(\lambda|Y)$. Because of the gamma density, we avoid the computation of the integral and directly write as:

$$\hat{\lambda}_B = \frac{y+\alpha}{n+\frac{1}{\beta}} = \frac{\beta(y+\alpha)}{n\beta+1}.$$

Let us investigate the structure of the Bayesian estimate of $\lambda$. If no data were available, we are forced to use the information from the prior distribution (only). $\pi(\lambda)$ has the mean $\alpha\beta$, which would be our best estimate of $\lambda$. If no prior information were available, we would use $Y/n$ (sample mean $\overline{X}$) to estimate $\lambda$ and draw the conclusion based on the sample values only. Now, the beautiful part is that the Bayesian estimator combines all of these information (if available). We can write $\hat{\lambda}_B$ in the following way.

$$\hat{\lambda}_B = \left(\frac{n\beta}{n\beta+1}\right)\left(\frac{Y}{n}\right) + \left(\frac{1}{n\beta+1}\right)(\alpha\beta).$$

Thus $\hat{\lambda}_B$ is a linear combination of the prior mean and the sample mean. The weights are determined by the values of $\alpha$, $\beta$ and $n$.

Suppose that we increase the sample size such that $n \to \infty$. Then $\frac{n\beta}{n\beta+1} \to 1$ and $\frac{1}{n\beta+1} \to 0$, and $\hat{\lambda}_B \to \overline{X}$. The idea is that as we increase the sample size, the data histogram closely approximates the population distribution itself. As if we have got enough knowledge about the population itself (because of a large number of observations). Hence, the prior information becomes less relevant and the value of the sample mean dominates the Bayesian estimate of $\lambda$. When the size of the sample is small, then it is wise to utilize the prior information about the population parameter. Hence, the term $\left(\frac{\alpha\beta}{n\beta+1}\right)$ contributes significantly in the final estimate. By weak law of large numbers $\overline{X} \to \lambda$ in probability as $n \to \infty$, so $\hat{\lambda}_B \to \theta$ in probability, proving that $\hat{\lambda}_B$ is consistent estimator for $\lambda$. Another interesting point is that, for any finite $n$, $\hat{\lambda}_B$ is a biased estimator of $\lambda$, with $\text{Bias}_\lambda(\hat{\lambda}_B) = \frac{\alpha\beta-\lambda}{n\beta+1} \to 0$ as $n \to \infty$. $\hat{\lambda}_B$ is asymptotically unbiased.

We end up with two estimators for the parameter $\lambda$, viz. Bayesian estimate, $\hat{\lambda}_B$ and Maximum likelihood estimator $\hat{\lambda} = \overline{X}$. It is natural to ask which estimator should we prefer? The efficiency of an estimator is computed by the Mean Square Error (MSE) which measures the average squared difference between the estimator and the parameter. In the present situation, the MSE of $\hat{\lambda}$ is

$$\text{E}_\lambda(\hat{\lambda}-\lambda)^2 = \text{Var}_\lambda(\overline{X}) = \frac{\lambda}{n}.$$

Since, $\overline{X}$ is an unbiased estimator of $\lambda$, the MSE of $\overline{X}$ is equal to its variance. Now, given $Y = \sum_{i=1}^{n} X_i$, the MSE of the Bayesian estimator of $\lambda$, $\hat{\lambda}_B = \frac{\beta(Y+\alpha)}{n\beta+1}$, is

$$\begin{aligned} \mathrm{E}_\lambda(\hat{\lambda}_B - \lambda)^2 &= \mathrm{Var}_\lambda \hat{\lambda}_B + (\mathrm{Bias}_\lambda \hat{\lambda}_B)^2 \\ &= \mathrm{Var}_\lambda \left( \frac{\beta(Y+\alpha)}{n\beta+1} \right) + \left( \mathrm{E}_\lambda \left( \frac{\beta(Y+\alpha)}{n\beta+1} \right) - \lambda \right)^2 \\ &= \frac{\beta^2 n\lambda}{(n\beta+1)^2} + \frac{(\alpha\beta-\lambda)^2}{(n\beta+1)^2}. \end{aligned}$$

For fixed $n$, the above quantity will be minimum if $\alpha\beta = \lambda$, that is the prior mean is equal to the true value. However, in general, MSE is a function of the parameter. So, it is highly unlikely that we would end up with a single estimator which is the best for all parameter values. In general the MSEs of two estimators cross each other. This demonstrates that one estimator is better with respect to another estimator in only a portion of the parameter space. Now let us delve deep further in comparing the above two estimators. If $\lambda = \alpha\beta$, then for large $n$,

$$\mathrm{MSE}_\lambda(\hat{\lambda}_B) = \frac{\beta^2 n\lambda}{(n\beta+1)^2} = \frac{\beta^2 n\lambda}{n^2\beta^2\left(1+\frac{1}{n\beta}\right)^2} \approx \frac{\lambda}{n}.$$

Thus, if the prior is chosen in a such way (choice of $\alpha$ and $\beta$) so that the prior mean is close to the true value, then both the estimators have same variance approximately, for large $n$. This observation is depicted in Figure 2.1. We consider the values of $\alpha$ and $\beta$ to be 4 and $\frac{1}{2}$, respectively, so that the prior mean is $\alpha\beta = 2$. It is clear from the figure that if prior mean is close to the true value, then $\hat{\lambda}_B$ performs better than $\hat{\lambda}$. If the prior mean is much away from the true value, then $\hat{\lambda}$ performs better than $\hat{\lambda}_B$. However, for large sample size ($n \to \infty$) both the estimators have same MSE.

We have seen that the posterior mean is a linear combination of prior mean and the sample mean. Basically, the Bayesian estimate lies between the sample mean and prior mean. That is, the linear combination is in fact a convex combination. This can also be better visualized if we plot the three distributions in a single plot window, viz. the data histogram, prior distribution, posterior distribution (Figure 2.2).

**R Code for Figure 2.1**

```
par(mfrow=c(1,3))
alpha = 4; beta = 1/2;
lambda = seq(0.1, 5, length.out = 50)
n_vals = c(10, 25, 50)
for(n in n_vals){
  mse_lambda_cap = lambda/n
  mse_lambdaB_cap = (beta^2*n*lambda + (alpha*beta-lambda)^2)/(n*beta+1)^2
  plot(lambda, mse_lambda_cap, col = "red", lwd=3, lty=1, type="l",
       xlab = expression(lambda), cex.lab = 1.2, ylab = "MSE", main = paste("n = ", n), ce
```

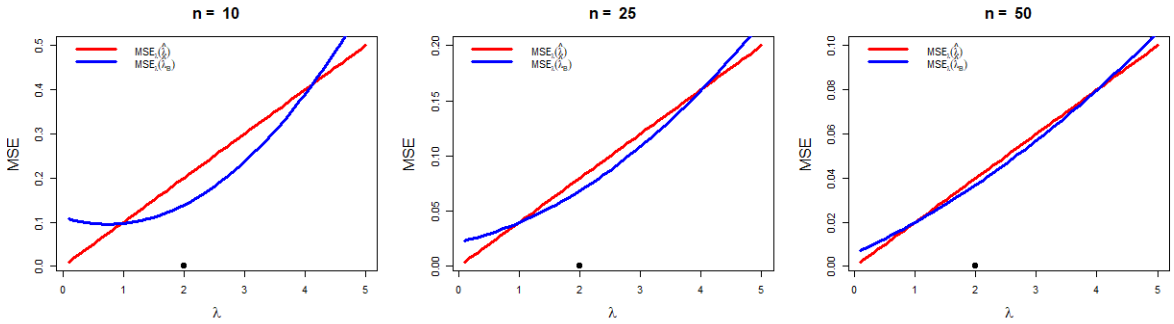



Figure 2.1: MSE of $\hat{\lambda} = \overline{X}$ and $\hat{\lambda}_B$ at different values of $\lambda$ for different choices of the sample size, $n$. The prior distribution of $\lambda$ is chosen as gamma(4, $\frac{1}{2}$). The prior mean is 2 and is indicated by a black dot. For small sample size $n = 10$, the $\text{MSE}_\lambda(\hat{\lambda}_B)$ is lower than the $\text{MSE}_\lambda(\hat{\lambda})$ for all true values of $\lambda$ in a neighborhood of 2. As $\lambda$ increases, after a certain value, $\text{MSE}_\lambda(\hat{\lambda})$ crosses the MSE of $\hat{\lambda}$. The same is observed for very small values of $\lambda$. However, for large $n$ values, the MSE of both the estimators merges, essentially the prior information becomes redundant ($n = 50$).

```
  lines(lambda, mse_lambdaB_cap, col = "blue", lwd=3, lty=1)
  legend = c( expression(paste("MSE"[lambda], (hat(lambda)))),
              expression(paste("MSE"[lambda], (hat(lambda)[B])))
  legend("topleft",legend = legend, lwd = c(3,3),col = c("red","blue"),
         lty = c(1,1), cex = 1, bty = "n")
  points(alpha*beta, 0, lwd=2, pch=20, cex=2)
}
```

**R Code for Figure 2.2**

```
par(mfrow=c(1,1))

# sampling distribution of sample mean
n = 5                                          # sample size
lambda = 2                                     # true mean values
set.seed(123)                                  # reporducibility of simualtion
rep = 100                                      # number of replication
mean_vals = numeric(length = rep)
for(i in 1:rep){
    mean_vals[i] = mean(rpois(n = n, lambda = lambda))
}
```

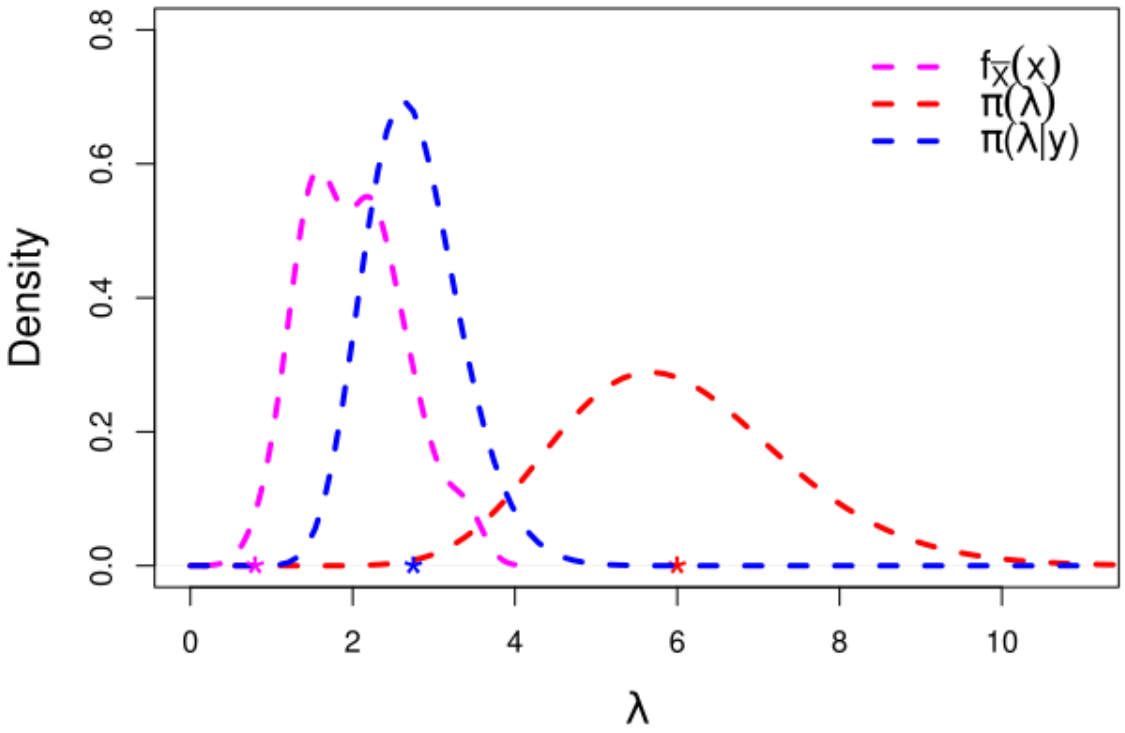



Figure 2.2: A sample of size $n = 5$ were simulated from the Poisson distribution with parameter $\lambda = 2$ and the sample mean has been computed. The process has been replicated 100 times to obtain the sampling distribution of the sample meanand approximated by a kernel density estimator (magenta colour). The maximum likelihood estimate of $\lambda$ is denoted by magenta coloured '*'. Similarly exact prior density and exact posterior density functions are plotted using blue and red colour, respectively. Similarly the prior mean and posterior mean values are also marked. The depicted picture clearly verifies the theoretical calculations performed in the text.

```r
plot(density(mean_vals), lwd=3, col = "magenta",  xlim = c(0, 11), ylim =c(0,.8), main = "
    xlab = expression(lambda), cex.lab = 1.4, lty = 2)

# Prior density of lambda
alpha = 18; beta = 1/3                      # hyperparameters
prior_density = function(x){                # prior density
    exp(-x/beta)*x^(alpha-1)/(beta^alpha * gamma(alpha))
}
curve(prior_density(x), 0, 13, col = "red", lwd=3, add = TRUE, lty = 2)
points(alpha*beta, 0, lwd=2, col ="red", pch = "*", cex=2)

y = sum(rpois(n = n, lambda = lambda))      # sufficient statistic
posterior_density = function(x){            # posterior density
    (n+1/beta)^(y+alpha)*exp(-x*(n+1/beta))*x^(y+alpha-1)/gamma(y+alpha)
}
curve(posterior_density(x), col = "blue", lwd=3, add = TRUE, lty = 2)
points(beta*(y+alpha)/(n*beta + 1),0, lwd=2, col ="blue", pch = "*", cex=2)     # posterio
points(y/n, 0, lwd=2, col ="magenta", pch = "*", cex=2)
legend(8,0.8, c(expression(f[bar(X)](x)), expression(pi(lambda)),
                expression(paste(pi,"(",lambda, "|", y,")" ))),
lwd=rep(3,3), col = c("magenta", "red", "blue"),  bty="n", cex = 1.3, lty = rep(2, 3))
```

## 2.2 Example - II (Normal prior for normal mean)

Suppose we observe a sample of size 1, $X \sim \mathcal{N}(\theta, \sigma^2)$ and suppose that the prior distribution of $\theta$ is $\mathcal{N}(\mu, \tau^2)$. We assume that the quantities, $\sigma^2$, $\mu$ and $\tau^2$ are all known. We are interested to obtain the posterior distribution of $\theta$. The prior distribution is given as;

$$\pi(\theta) = \frac{1}{\tau\sqrt{2\pi}} e^{-\frac{(\theta-\mu)^2}{2\tau^2}}, \quad -\infty < \mu, \theta < \infty, \; 0 < \tau < \infty.$$

The posterior density function of $\theta$ is given as follows;

$$\begin{aligned} \pi(\theta|x) &= \frac{f(x|\theta)\pi(\theta)}{\int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)\mathrm{d}\theta} \\ &= \frac{\frac{1}{\sigma\tau(\sqrt{2\pi})^2} \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\theta}{\sigma}\right)^2 + \left(\frac{\theta-\mu}{\tau}\right)^2\right]\right\}}{\int_{-\infty}^{\infty} f(x|\theta)\pi(\theta)\mathrm{d}\theta}. \end{aligned}$$

The exponent can be expressed as

$$\frac{\sigma^2+\tau^2}{\sigma^2\tau^2}\cdot\left[\left(\theta-\frac{x\tau^2+\mu\sigma^2}{\sigma^2+\tau^2}\right)^2+\frac{\tau^2x^2+\mu^2\sigma^2}{\sigma^2+\tau^2}-\left(\frac{x\tau^2+\mu^2\sigma^2}{\sigma^2+\tau^2}\right)^2\right].$$

Note that, all the terms except containing the expression of $\theta$ will be cancelled with the denominator, resulting the posterior density of $\theta$ given as follows;

$$\pi(\theta|x)=\frac{\sqrt{\sigma^2+\tau^2}}{\sigma\tau\sqrt{2\pi}}\exp\left[-\frac{\sigma^2+\tau^2}{2\sigma^2\tau^2}\left(\theta-\frac{x\tau^2+\mu\sigma^2}{\sigma^2+\tau^2}\right)^2\right].$$

The posterior distribution of $\theta$ is normal, showing that the normal family is its own conjugate when indexed by the mean ($\theta$). The posterior mean and variance of $\theta$ are as follows;

$$\mathrm{E}(\theta|x)=\frac{x\tau^2+\mu\sigma^2}{\sigma^2+\tau^2},\qquad\text{and}\qquad\mathrm{Var}(\theta|x)=\frac{\sigma^2\tau^2}{\sigma^2+\tau^2}.$$

As discussed earlier, $\mathrm{E}(\theta|x)$ is a point of estimate of $\theta$, thus the Bayesian estimator of $\theta$ based on a single sample is given by

$$\hat{\theta}_B=\frac{X\tau^2+\mu\sigma^2}{\sigma^2+\tau^2}=\left(\frac{\tau^2}{\sigma^2+\tau^2}\right)X+\left(\frac{\sigma^2}{\sigma^2+\tau^2}\right)\mu.$$

Again, the Bayesian estimate of $\theta$ is a linear combination of the prior mean and the sample value (which is in fact the estimate of $\theta$, as the size of the sample is 1). We shall treat this problem based on a sample of size $n$ and obtain the Bayesian estimator of $\theta$ using it. A tedious calculation is required to obtain the posterior distribution. However, that will help us to obtain the the distribution of $\theta|\overline{X}$ easily. Suppose that we draw a sample $X_1, X_2, \cdots, X_n$ of size $n$ from $\mathcal{N}(\theta,\sigma^2)$, then the sample mean $\overline{X}\sim\mathcal{N}(\theta,\sigma^2/n)$. The same calculation will follow with $X$ would be replaced by $\overline{X}$ and $\sigma^2$ will be replaced by $\frac{\sigma^2}{n}$, respectively. Then the posterior distribution of $\theta$ is normal, with mean and variance given by

$$\mathrm{E}(\theta|\overline{x})=\frac{\overline{x}n\tau^2+\mu\sigma^2}{\sigma^2+n\tau^2},\qquad\mathrm{Var}(\theta|\overline{x})=\frac{\sigma^2\tau^2}{\sigma^2+n\tau^2},$$

and the Bayesian estimator is given as follows;

$$\hat{\theta}_B=\left(\frac{n\tau^2}{\sigma^2+n\tau^2}\right)\overline{X}+\left(\frac{\sigma^2}{\sigma^2+n\tau^2}\right)\mu.$$

As $n \to \infty$, $\frac{n\tau^2}{\sigma^2+n\tau^2} \to 1$ and $\frac{\sigma^2}{\sigma^2+n\tau^2} \to 0$, so that $\hat{\theta}_B \approx \overline{X}$. But, when $n$ is small, use of prior information improves the estimate. Since, $\overline{X} \to \theta$ in probability as $n \to \infty$, which follows that $\hat{\theta}_B \to \theta$ in probability as $n \to \infty$ (By Slutsky's theorem). MSE of $\hat{\theta}_B$ is given by

$$\begin{aligned}
\mathrm{E}_\theta(\hat{\theta}_B - \theta)^2 &= \mathrm{Var}_\theta(\hat{\theta}_B) + \left(\mathrm{Bias}_\theta(\hat{\theta}_B)\right)^2 \\
&= \mathrm{Var}_\theta\left(\frac{\overline{X}n\tau^2 + \mu\sigma^2}{\sigma^2 + n\tau^2}\right) + \left(\mathrm{E}_\theta\left(\frac{\overline{X}n\tau^2 + \mu\sigma^2}{\sigma^2 + n\tau^2}\right) - \theta\right)^2 \\
&= \frac{\tau^4 n\sigma^2}{(\sigma^2 + n\tau^2)^2} + \left(\frac{(\mu-\theta)\sigma^2}{\sigma^2 + n\tau^2}\right)^2 \\
&= \frac{\sigma^2}{(\sigma^2 + n\tau^2)^2} \cdot (n\tau^4 + (\mu-\theta)^2\sigma^2)\,.
\end{aligned}$$

The estimators $\hat{\theta}_B$ and $\overline{X}$ are compared for different sample sizes and also for different choices of the prior variance $\tau^2$ (Figure 4.2a and Figure 4.2b). The estimator can also be compared by using the relative efficiency, which is defined as the MSE of the two estimators. So the efficiency of $\hat{\theta}_B$ relative to $\overline{X}$, $\mathrm{eff}_\theta(\hat{\theta}_B|\overline{X})$, is

$$\frac{\mathrm{MSE}_\theta(\hat{\theta}_B)}{\mathrm{MSE}_\theta(\overline{X})} = \frac{n\tau^4 + (\mu-\theta)^2\sigma^2}{(\sigma^2 + n\tau^2)^2}.$$

If $\mu = \theta$, for a given sample size $n$, relative efficiency is minimum and less than 1. Hence, $\hat{\theta}_B$ is more efficient than $\overline{X}$ for any given $n$. If we move $\theta$ values away from $\mu$, then $\mathrm{eff}_\theta(\hat{\theta}_B|\overline{X})$ crosses the line $y = 1$. For theta values beyond that point, $\overline{X}$ is better than $\hat{\theta}_B$. We encourage the reader to draw this picture and compare two estimators based on the relative efficiency. It is clearly understood that, essentially we are describing the same thing, only with a different graphical representation. Of course, the function `curve` from `R` is a very useful tool, which has been utilized throughout this material.

**R Code for Figure 4.2a and Figure 4.2b**

```
sigma = sqrt(1)      # population sd, known
mu = 3               # prior mean value
tau = sqrt(0.5)      # prior sd
n_vals = c(4, 10, 20) # sample size

par(mfrow=c(2,3))    # space for six plots in a single window
for(n in n_vals){
  curve(sigma^2/(sigma^2 + n*tau^2)^2*(n*tau^4 + (mu-x)^2*sigma^2),
        0, 6, col = "blue", lwd=3, ylab = "MSE", ylim=c(0,0.3),
```

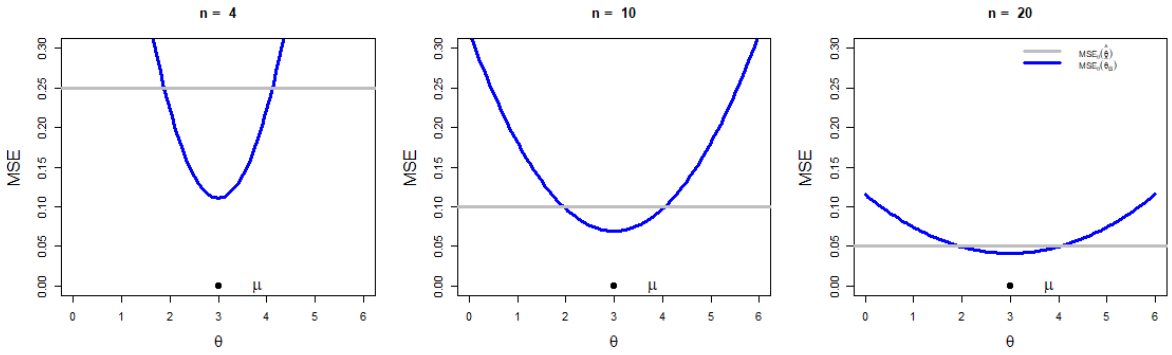



Figure 2.3: $\text{MSE}_\theta(\hat{\theta})$ and $\text{MSE}_\theta(\hat{\theta}_B)$ are plotted as a function of $\theta$ for different sample size $n$. The prior distribution of $\theta$ is considered to be $\mathcal{N}(\mu = 3, \tau^2 = 0.5)$. The prior mean $\mu$ is indicated by a black dot. For sample size, $n = 4$, $\text{MSE}_\theta(\hat{\theta}_B)$ is smaller than $\text{MSE}_\theta(\hat{\theta})$ at all values of $\theta$ in a neighborhood of $\mu$. As we move away from $\mu$, $\hat{\theta}$ is more preferable outsize the neighborhood of $\mu$. However, as we increase $n$, MSE of both the estimators become closer in every neighborhood of $\mu$ ($n = 20$), reducing the impact of prior information.

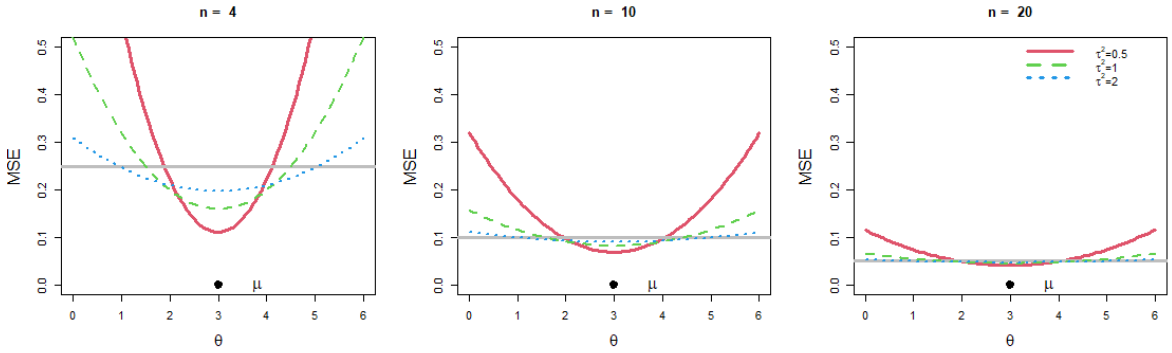



Figure 2.4: The MSE of the two estimators are compared at different values of $\tau^2$. For small sample size ($n = 4$) and small $\tau^2 < \sigma^2$, $\hat{\theta}_B$ performs better that $\hat{\theta}$. For large values of $\tau^2 > \sigma^2$, $\text{MSE}_\theta(\hat{\theta}_B)$ increases. Basically, prior information becomes vague for large $\tau^2$ values. As expected, for large sample size ($n = 20$) both the estimators perform at par irrespective of the prior variance.

```
        main = paste("n = ", n), xlab = expression(theta), cex.lab = 1.5)
  abline(h=sigma^2/n,col = "grey", lwd=3)
  points(mu, 0, lwd=2.5, pch=20, cex=2)
  text(mu+0.8, 0, expression(mu), cex = 1.5 )
}
legend = c( expression(paste("MSE"[theta], (hat(theta)))),
    expression(paste("MSE"[theta], (hat(theta)[B]))))
legend("topright",legend = legend, lwd = c(3,3),col = c("grey","blue"),
       lty = c(1,1), cex = 0.8, bty = "n")

par(mfrow=c(2,3))
n_vals = c(4, 10, 20)
sigma = sqrt(1)
tau_vals = sqrt(c(0.5, 1, 2))  # varying prior sd
for(n in n_vals){
  for(i in 1:length(tau_vals)){
    tau = tau_vals[i]
    if(i == 1){
      curve(sigma^2/(sigma^2 + n*tau^2)^2*(n*tau^4 + (mu-x)^2*sigma^2),
            0, 6, col = i+1, lwd=3, ylab = "MSE", ylim=c(0,0.5), lty=i,
            xlab = expression(theta), main = paste("n = ", n), cex.lab = 1.5)
      abline(h=sigma^2/n, lwd=3, col = "grey")
      points(mu, 0, lwd=3, pch=20, cex=2)
      text(mu+0.8, 0, expression(mu), cex=1.5 )
    }
    else
      curve(sigma^2/(sigma^2 + n*tau^2)^2*(n*tau^4 + (mu-x)^2*sigma^2),
            0, 6, col = i+1, lwd=2.5, add = TRUE, lty=i)
  }
}
legend = c(expression(paste(tau^2,"=", 0.5)), expression(paste(tau^2,"=", 1)),
        expression(paste(tau^2,"=", 2)))
legend("topright", legend, col = c(2,3,4), lwd=c(3,3,3), lty=1:3, bty = "n")
```

## 2.3 Example - III (Beta prior for Bernoulli($p$))

Let $X_1, X_2, \cdots, X_n$ be iid Bernoulli($p$). Then $Y = \sum_{i=1}^{n} X_i$ is binomial($n, p$). We assume the prior distribution of $p$ is Beta($\alpha, \beta$). Then the joint distribution of $Y$ and $p$ is given by

$$\begin{aligned} f(y,p) &= \left[\binom{n}{y}p^y(1-p)^{n-y}\right]\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1-p)^{\beta-1}\right] \\ &= \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{y+\alpha-1}(1-p)^{n-y+\beta-1}. \end{aligned}$$

The marginal pdf of $Y$ is

$$f(y) = \int_0^1 f(y,p)\mathrm{d}p = \binom{n}{y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)},$$

which is the beta-binomial distribution. The posterior distribution of $p$ is given as follows;

$$\pi(p|y) = \frac{f(y,p)}{f(y)} = \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)}p^{y+\alpha-1}(1-p)^{n-y+\beta-1},$$

which is beta$(y+\alpha, n-y+\beta)$. So, the Bayes estimator of $p$ is taken as a mean of the posterior distribution under a squared error loss, and it is given as follows;

$$\hat{p}_B = \frac{y+\alpha}{\alpha+\beta+n}.$$

We can write $\hat{p}_B$ in the following way;

$$\hat{p}_B = \left(\frac{n}{\alpha+\beta+n}\right)\left(\frac{Y}{n}\right) + \left(\frac{\alpha+\beta}{\alpha+\beta+n}\right)\left(\frac{\alpha}{\alpha+\beta}\right). \tag{2.2}$$

Thus $\hat{p}_B$ is a linear combination of the prior mean and the sample mean. The weights are determined by the values of $\alpha$, $\beta$ and $n$. In the present situation, the MSE of $\hat{p}$ is

$$\mathrm{E}_p(\hat{p}-p)^2 = \mathrm{Var}((\overline{X})) = \frac{p(1-p)}{n}. \tag{2.3}$$

Since, $\overline{X}$ is an unbiased estimator of $p$, hence the MSE is equal to the variance. Now, given $Y = \sum_{i=1}^n X_i$, the MSE of the Bayesian estimator of $p$, $\hat{p}_B = \frac{Y+\alpha}{\alpha+\beta+n}$ is

$$\begin{aligned} \mathrm{E}_p(\hat{p}_B-p)^2 &= \mathrm{Var}_p(\hat{p}_B) + (\mathrm{Bias}_p\hat{p}_B)^2 \\ &= \mathrm{Var}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) + \left(\mathrm{E}_p\left(\frac{Y+\alpha}{\alpha+\beta+n}\right) - p\right)^2 \\ &= \frac{np(1-p)}{(\alpha+\beta+n)^2} + \frac{(\alpha-p(\alpha+\beta))^2}{(\alpha+\beta+n)^2}. \end{aligned}$$

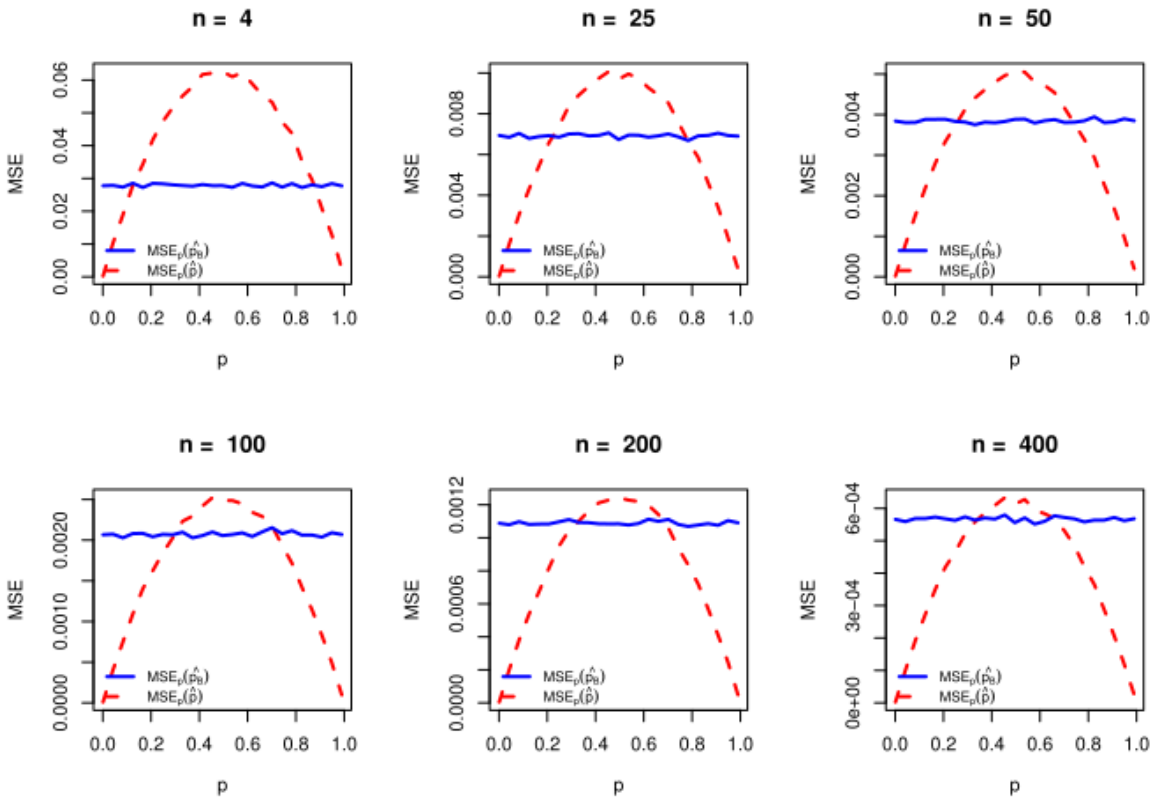



Figure 2.5: $\text{MSE}_p(\hat{p}) = \text{MSE}_p(\overline{X}) = \frac{p(1-p)}{n}$ and $\text{MSE}_p(\hat{p}_B) = \frac{np(1-p)+[\alpha-p(\alpha+\beta)]^2}{(\alpha+\beta+n)^2}$ are computed based on simulated data from a $\text{Bin}(n,p)$ distribution. The prior distribution $\pi(p)$ is assumed to be $\text{Beta}(\alpha,\beta)$. Here we consider $\alpha = \beta = \sqrt{\frac{n}{4}}$ so that $\text{MSE}(\hat{p}_B) = \frac{n}{4(n+\sqrt{n})^2}$, which is constant for all values of $p(0 < p < 1)$, for fixed $n$. The computation was done for different sample sizes as depicted in the figure. At each value of $p$, average MSEs of both the estimators were computed using 1000 replications. The MSEs are plotted against different values of $p$. For small sample size $n = 4$, $\hat{p_B}$ has more precision in estimating the true proportion than $\hat{p}$ for almost all $p \in (0,1)$ except the values closer to 0 and 1, where $\hat{p}$ is better. However, as sample size increases, the interval in which $\hat{p_B}$ is better than $\hat{p}$ decreases substantially. For $n = 400$, $\hat{p_B}$ works better in a very small interval about 0.5.

**R Code for Figure 2.5**

```
par(mfrow = c(2,3))
n_vals = c(4, 25, 50, 100, 200, 400)  # size of the sample
rep = 10^4   # number of replication
P = seq(0.001, 0.99, length.out = 25) # values of the true probability
for(n in n_vals){
  alpha = sqrt(n/4); beta = sqrt(n/4) # Beta(alpha, beta) to make the MSE constant
  mse_p_cap = numeric(length = length(P))
  mse_pB_cap = numeric(length = length(P))

  for(i in 1:length(P)){
    p_cap = numeric(rep)
    pB_cap = numeric(rep)

    for(j in 1:rep){
      d = rbinom(n, 1, P[i])
      p_cap[j] = sum(d)/n
      pB_cap[j] = (sum(d) + alpha)/(alpha + beta + n )
    }
    mse_p_cap[i] = mean((p_cap - P[i])^2)
    mse_pB_cap[i] = mean((pB_cap - P[i])^2)
  }
  ylim = c(min(mse_p_cap, mse_pB_cap), max(mse_p_cap, mse_pB_cap))
  plot(P, mse_p_cap, col = "red", lwd=2, main = paste("n = ", n),
       ylim = ylim, type = "l", ylab = "MSE", lty =2, xlab = expression(p))
  lines(P, mse_pB_cap, col = "blue", lwd=2, lty=1)
  legend = c(expression(paste("MSE"[p], (hat(p[B])))),
    expression(paste("MSE"[p], (hat(p)))))
  legend("bottomleft", legend = legend, lwd=c(2,2), col = c("blue", "red"),
    bty = "n", lty = c(1,2), cex = 0.8)
```

## 2.4 Example - IV (Generalization of Hierarchical Bayes)

We examine a generalization of the Hierarchical Bayes model. Let $X_1, X_2, \cdots, X_n$ be an observed random sample of size $n$ such that

$$X_i|\lambda_i \sim \text{Poisson}(\lambda_i), \qquad i = 1, 2, \cdots, n, \quad \text{independent.}$$

Suppose $\lambda_i,\ i = 1, 2, \cdots, n$ have a gamma$(\alpha, \beta)$ distribution, which is the conjugate family for Poisson. So, the hierarchical model is given by,

$$\begin{aligned} X_i|\lambda_i &\sim \text{Poisson}(\lambda_i), \quad i = 1, 2, \cdots, n, \quad \text{independent}, \\ \lambda_i &\sim \text{gamma}(\alpha, \beta), \quad i = 1, 2, \cdots, n, \quad \text{independent}, \end{aligned}$$

where $\alpha$ and $\beta$ are known positive constants, usually provided by the experimenter. Our interest is to estimate the $\lambda_i,\ i = 1, 2, \cdots, n$, based on observed sample. The prior distribution of $\lambda_i$ is given by,

$$\pi(\lambda_i) = \frac{e^{-\frac{\lambda_i}{\beta}}\lambda_i^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha},\ i = 1, 2, \cdots, n.$$

Suppose, we have a situation where the parameter $\beta$ is not provided by the experimenter. However, the value of $\alpha$ is known. To obtain an estimate of $\lambda_i$, we first have to estimate the parameter $\beta$. A critical point is that if we would like to estimate $\beta$, then we require independent observations from the gamma$(\alpha, \beta)$ distribution. Although $\lambda_i$'s are there from the said distribution, but these are not observable quantities. The empirical Bayes analysis makes use of the observed sample $X_1, X_2, \cdots, X_n$ to estimate the parameters of the prior distribution. We First obtain the marginal distribution of $X_i$'s. The use of moment generating function ease the process of computing the marginal distribution greatly. Let $M_{X_i}(t) = \text{E}(e^{tX_i})$, be the mgf of $X_i,\ i = 1, 2, \cdots, n$, which is

$$\begin{aligned} M_{X_i}(t) &= \text{E}\left(e^{tX_i}\right) \\ &= \text{E}\left[\text{E}\left(e^{tX_i}|\lambda_i\right)\right] \\ &= \text{E}\left(e^{\lambda_i(e^t-1)}\right) \\ &= \frac{1}{\left[1-\beta(e^t-1)\right]^\alpha}, \qquad \left[M_{\lambda_i}(t) = (1-\beta t)^{-\alpha}\right] \\ &= \left(\frac{\frac{1}{\beta+1}}{1-\frac{\beta}{\beta+1}e^t}\right)^\alpha. \end{aligned}$$

which is the mgf of Negative Binomial distribution with parameter $r = \alpha$ and $p = \frac{1}{\beta+1}$. Recall that if $X$ follows negative binomial distribution, $X \sim \text{NB}(r, p)$ then mgf of $X$ is given by $\left(\frac{p}{1-(1-p)e^t}\right)^\alpha$ and $\text{E}X = \frac{r(1-p)}{p^2}$ and $\text{Var}(X) = \frac{r(1-p)}{p^2}$. So, $X_i \sim \text{NB}\left(\alpha, \frac{1}{\beta+1}\right),\ i = 1, 2, \cdots, n$, whose pmf of is given by,

$$P(X_i = x_i) = \binom{x_i+\alpha-1}{x_i}\left(\frac{1}{\beta+1}\right)^\alpha\left(\frac{\beta}{\beta+1}\right)^{x_i},\ x_i \in \{0, 1, 2, \cdots\}.$$

A pleasant surprise is that we observe $X_1, X_2, \cdots, X_n$ which are marginally iid and follows $\text{NB}\left(\alpha, \frac{1}{\beta+1}\right)$. Hence, $\beta$ can be estimated by using $X_1, X_2, \cdots, X_n$. We can use the maximum likelihood estimation method to estimate $\beta$. The likelihood function $\mathcal{L}(\beta)$ is given as follows;

$$\mathcal{L}(\beta) = \left[\prod_{i=1}^{n}\binom{x_i+\alpha-1}{x_i}\right]\left(\frac{1}{\beta+1}\right)^{\alpha}\left(\frac{\beta}{\beta+1}\right)^{x_i}, \tag{2.4}$$

and the corresponding log-likelihood function, $l(\beta) = \ln(\mathcal{L}(\beta))$ is

$$l(\beta) = \sum_{i=1}^{n}\ln\binom{x_i+\alpha-1}{x_i} - n\alpha\ln(1+\beta) + \sum_{i=1}^{n}x_i\ln\left(\frac{\beta}{1+\beta}\right). \tag{2.5}$$

Taking first order derivative with respect to $\beta$ and making $\frac{\partial l(\beta)}{\partial\beta} = 0$, we obtain the likelihood equation as

$$\frac{-n\alpha}{1+\beta} + \left(\sum_{i=1}^{n}x_i\right)\left(\frac{1}{\beta} - \frac{1}{1+\beta}\right) = 0.$$

On simplifying we get estimator for prior parameter $\beta$ as, $\widehat{\beta} = \frac{\bar{\mathrm{X}}}{\alpha}$. It can be easily verified that $\frac{\partial^2 l(\beta)}{\partial\beta^2} = \frac{n\alpha}{(\beta+1)^2} + \sum_{i=1}^{n}x_i\left(-\frac{1}{\beta^2} + \frac{1}{(1+\beta)^2}\right)$ at $\beta = \widehat{\beta}$, $\left.\frac{\partial^2(\beta)}{\partial\beta^2}\right|_{\beta=\frac{\bar{\mathrm{X}}}{\alpha}} = \frac{-n\alpha^3\bar{\mathrm{x}}-n\alpha^4}{\bar{\mathrm{x}}(\bar{x}+\alpha)^2} < 0$, showing that $\widehat{\beta}$ is the MLE of $\beta$. Once the estimate of the unknown prior $\widehat{\beta}$ is obtained, the model becomes

$$\begin{aligned} X_i|\lambda_i &\sim \text{Poisson}(\lambda_i), \qquad i = 1,2,\cdots,n, \quad \text{independent,} \\ \lambda_i &\sim \text{gamma}(\alpha,\widehat{\beta}), \qquad i = 1,2,\cdots,n, \quad \text{independent.} \end{aligned}$$

The posterior distribution that is the conditional distribution of $\lambda_i$ given the sample $x_i, i = 1,2,\cdots,n$ is

$$\begin{aligned} \pi(\lambda_i|x_i) &= \frac{f(x_i|\lambda_i)\pi(\lambda_i)}{f(x_i)} \\ &= \frac{\frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}\frac{e^{-\frac{\lambda_i}{\widehat{\beta}}}\lambda_i^{\alpha-1}}{\Gamma(\alpha)\widehat{\beta}^{\alpha}}}{\binom{x_i+\alpha-1}{x_i}\left(\frac{1}{\widehat{\beta}+1}\right)^{\alpha}\left(\frac{\widehat{\beta}}{\widehat{\beta}+1}\right)^{x_i}} \\ &= \frac{\lambda_i^{x_i+\alpha-1}e^{-\lambda_i\left(1+\frac{1}{\widehat{\beta}}\right)}\left(1+\frac{1}{\widehat{\beta}}\right)^{x_i+\alpha}}{\Gamma(x_i+\alpha)}. \end{aligned}$$

So, $\lambda_i|x_i \sim \text{gamma}\left(x_i+\alpha, \left(1+\frac{1}{\widehat{\beta}}\right)^{-1}\right)$. The Bayes estimator of $\lambda_i$ under a squared error loss, that is the posterior mean $\mathrm{E}(\lambda_i|x_i)$ is given by,

$$\hat{\lambda}_{i\,B} = (x_i + \alpha)\left(1 + \frac{1}{\hat{\beta}}\right)^{-1}.$$

Note that the Bayes estimate of $\lambda_i$ contains the term $\hat{\beta}$, which was estimated from data. Hence, there is uncertainty associated with the estimate of the prior parameter $\beta$. We can estimate $\text{Var}(\hat{\beta})$ as follows:

$$\text{Var}(\hat{\beta}) = \frac{1}{\alpha^2 n^2}\text{Var}_\beta(\sum_{i=1}^{n} X_i) = \frac{1}{\alpha^2 n}\text{Var}_\beta(X_1) = \frac{1}{\alpha^2 n}\frac{\alpha\left(1 - \frac{1}{\beta+1}\right)}{\left(\frac{1}{\beta+1}\right)^2} = \frac{\beta(\beta+1)}{n\alpha}.$$

So the estimated variance is $\frac{\hat{\beta}(\hat{\beta}+1)}{n\alpha}$. However, it does not play any role in empirical Bayes estimation. This is a drawback of the empirical Bayes estimation. However, in hierarchical Bayes, the unknown prior parameter is replaced by a distribution. Hence the uncertainty in the hyperparameters gets included in the final estimation of the posterior mean. This is an advantage of hierarchical Bayes over empirical Bayes. The hierarchical formulation of the same problem may be stated as follows:

$$\begin{aligned}
X_i|\lambda_i &\sim \text{Poisson}(\lambda_i), \quad i = 1, 2, \cdots, n, \quad \text{independent} \\
\lambda_i &\sim \text{gamma}(\alpha, \beta), \quad i = 1, 2, \cdots, n, \quad \text{independent}, \\
\beta &\sim \text{uniform}(0, \infty) \quad \text{(noninformative prior)}.
\end{aligned}$$

Coming back to the problem again, the empirical Bayes estimator is given by

$$\hat{\lambda}_{i\,B} = \left(\frac{\hat{\beta}}{1+\hat{\beta}}\right) x_i + \left(\frac{1}{1+\hat{\beta}}\right)(\alpha\hat{\beta}).$$

Again with no surprise, $\hat{\lambda}_{i\,B}$ is a linear combination of the prior mean and sample mean. In above, we estimate $\lambda_i$ using only single observation form Poission$(\lambda_i)$, $i = 1, 2, \cdots, n$. In the discussed example, we have assumed $\alpha$ to be known and $\beta$ unknown. However, it might happen that both $\alpha$ and $\beta$ are known. In such situation, both $\alpha$ and $\beta$ may be estimated from the observed samples $X_1, X_2, \cdots, X_n$, whose marginal distributions are iid NB $\left(\alpha, \frac{1}{\beta+1}\right)$. The likelihood equations are given by

$$\frac{\partial l(\alpha, \beta)}{\partial \alpha} = 0, \quad \frac{\partial l(\alpha, \beta)}{\partial \beta} = 0,$$

where the log-likelihood function $l(\alpha, \beta)$ is same as given in the equation Equation 2.5. These maximization must be carried out using some numerical procedures, for example Newton

Raphson method. The estimates and associated standard errors can be obtained using R. For example, the likelihood function can be passed in the `optim` function or the function `fitdistr`, available in the `MASS` package (Venables and Ripley 2002), may be utilized. Sample code is given below:

```
> x = rgamma(n = 50, shape = 2, rate = 3)
> library(MASS)
> fit = fitdistr(x = x, "gamma", lower = 0.01)
> fit$estimate
   shape     rate
1.615070 2.922565
> fit$sd
    shape      rate
0.2956650 0.6261151
> hist(x, prob = T, col = "lightgrey")
> curve(dgamma(x, shape = fit$estimate[1], rate = fit$estimate[2]),
    col = "red", lwd=2, add = TRUE)
```

In the same example, we may have a one way classification with $n$ groups and $m$ observations per group. Then the example extends to

$$\begin{aligned} X_{ij}|\lambda_i &\sim \text{Poisson}(\lambda_i), \quad i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, m; \quad \text{independent}, \\ \lambda_i &\sim \text{gamma}(\alpha, \beta), \quad i = 1, 2, \cdots, n; \quad \text{independent}. \end{aligned}$$

To estimate $\lambda_i$, we use the statistic $Y_i = \sum_{j=1}^{m} X_{ij}$, then $Y_i|\lambda_i \sim \text{Poisson}(m\lambda_i),\ i = 1, 2, \cdots, n$. The marginal pdf of $Y_i$ is

$$\begin{aligned} f(y_i) &= \int_0^\infty f(y_i|\lambda_i)\pi(\lambda_i)\mathrm{d}\lambda_i \\ &= \int_0^\infty \frac{e^{-m\lambda_i}(m\lambda_i)^{y_i}}{y_i!} \frac{e^{-\frac{\lambda_i}{\beta}}\lambda_i^{\alpha-1}}{\Gamma(\alpha)\hat{\beta}^\alpha}\mathrm{d}\lambda_i \\ &= \int_0^\infty \frac{e^{-\lambda_i\left(m+\frac{1}{\beta}\right)} m^{y_i}\lambda_i^{y_i+\alpha-1}}{y_i!\beta^\alpha\Gamma(\alpha)} \\ &= \frac{m^{y_i}\left\{\left(m+\frac{1}{\beta}\right)^{-1}\right\}^{y_i+\alpha}\Gamma(y_i+\alpha)}{y_i!\beta^\alpha\Gamma(\alpha)} \\ &= \frac{\Gamma(y_i+\alpha+1-1)}{\Gamma(y_i+1)\Gamma(\alpha)}\left(\frac{m\beta}{m\beta+1}\right)^{y_i}\left(\frac{1}{m\beta+1}\right)^{\alpha} \\ &= \binom{y_i+\alpha-1}{y_i}\left(\frac{1}{m\beta+1}\right)^{\alpha}\left(1-\frac{1}{m\beta+1})\right)^{y_i}. \end{aligned}$$

So, $Y_i \sim \text{NB}\left(\alpha, \frac{1}{m\beta+1}\right)$, $i = 1, 2, \cdots, n$. The posterior distribution of $\lambda_i$ is given by,

$$\begin{aligned}
\pi(\lambda_i|y_i) &= \frac{f(y_i|\lambda_i)\pi(\lambda_i)}{f(y_i)} \\
&= \frac{\frac{e^{m\lambda_i}(m\lambda_i)^{y_i}}{y_i!}\frac{e^{-\frac{\lambda_i}{\beta}}\lambda_i^{\alpha-1}}{\beta^\alpha\Gamma(\alpha)}}{f(y_i)} \\
&= \frac{\lambda_i^{y_i+\alpha-1}e^{-\lambda_i\left(m+\frac{1}{\beta}\right)}\left(m+\frac{1}{\beta}\right)^{y_i+\alpha}}{\Gamma(y_i+\alpha)},
\end{aligned}$$

which shows that $\lambda_i|y_i \sim \text{gamma}\left(y_i+\alpha, \left(m+\frac{1}{\beta}\right)^{-1}\right)$. The posterior mean, that is the Bayes estimator of $\lambda_i$ under a squared error loss is

$$\begin{aligned}
\hat{\lambda}_{i_B} &= \text{E}(\lambda_i|y_i) \\
&= (y_i+\alpha)\left(m+\frac{1}{\beta}\right)^{-1} \\
&= \left(\sum_{j=1}^{m} x_{ij}+\alpha\right)\left(m+\frac{1}{\beta}\right)^{-1}, \\
&= \left(\frac{m\beta}{1+m\beta}\right)\bar{X}_i + \left(\frac{1}{1+m\beta}\right)(\alpha\beta).
\end{aligned}$$

Thus, $\hat{\lambda}_{i_B}$ is a linear combination of the prior mean and sample mean. The variance of $\hat{\lambda}_{i_B}$ is given by

$$\begin{aligned}
\text{Var}_{\lambda_i}(\hat{\lambda}_{i_B}) &= \left(\frac{m\beta}{1+m\beta}\right)^2 \text{Var}_{\lambda_i}(\bar{X}_i), \\
&= \left(\frac{m\beta}{1+m\beta}\right)^2 \left(\frac{\lambda_i}{m}\right).
\end{aligned}$$

With no surprise, for large $m$, the above expression is close to the variance of maximum likelihood estimator $\left(\frac{m\beta}{m\beta+1} \to 1 \text{ for large } m\right)$.

## 2.5 Example - V (Gamma prior for Exponential rate parameter)

Let $Y_1$,$Y_2$,$\cdots$,$Y_n$ be a observed random sample of size $n$ such that

$$Y_i|\lambda \sim \text{exponential}(\lambda),\ i = 1, 2, \cdots, n, \quad \text{independent.}$$

Suppose that $\lambda$ has a gamma$(\alpha, \beta)$ distribution, which is the conjugate family for Exponential. So, the statistical model has the following hierarchy

$$
\begin{aligned}
Y_i|\lambda &\sim \text{exponential}(\lambda),\ i = 1, 2, \cdots, n, && (2.6)\\
\lambda &\sim \text{gamma}(\alpha, \beta), && (2.7)
\end{aligned}
$$

where $\alpha$ and $\beta$ are known. Now, our interest is to estimate the $\lambda$ based on observed sample. The prior distribution of $\lambda$ is given by,

$$\pi(\lambda) = \frac{e^{-\frac{\lambda}{\beta}}\lambda^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}}, \lambda > 0, \tag{2.8}$$

where $\alpha$ and $\beta$ are positive constants. First, we shall estimate the posterior distribution of $\lambda$. We start with statistic $Z = \sum_{i=1}^{n} Y_i$. Sampling distribution of $Z$ is gamma$(n, \lambda)$ and it is denoted by $f(z|\lambda)$. The posterior distribution of $\lambda$ given the sample $Y_1, Y_2, \cdots, Y_n$, that is given $Z = z$, is

$$\pi(\lambda|z) = \frac{f(z|\lambda)\pi(\lambda)}{f(z)},$$

where $f(z)$ is the marginal distribution of $Z$, and calculated as follows;

$$
\begin{aligned}
f(z) &= \int_0^{\infty} f(z|\lambda)\pi(\lambda)\mathrm{d}\lambda\\
&= \int_0^{\infty} \frac{z^{n-1}e^{-\lambda z}\lambda^n}{\Gamma(n)} \frac{\lambda^{\alpha-1}e^{-\frac{\lambda}{\beta}}}{\Gamma(\alpha)\beta^{\alpha}}\mathrm{d}\lambda\\
&= \frac{z^{n-1}}{\Gamma(n)\Gamma(\alpha)\beta^{\alpha}} \int_0^{\infty} e^{-\lambda(z+\frac{1}{\beta})}\lambda^{\alpha+n-1}\mathrm{d}\lambda\\
&= \frac{z^{n-1}\Gamma(\alpha+n)}{\Gamma(n)\Gamma(\alpha)\beta^{\alpha}(z+\frac{1}{\beta})^{\alpha+n}} \int_0^{\infty} \frac{\lambda^{\alpha+n-1}e^{-\lambda(z+\frac{1}{\beta})}(z+\frac{1}{\beta})^{\alpha+n}}{\Gamma(\alpha+n)}\mathrm{d}\lambda\\
&= \frac{z^{n-1}\Gamma(\alpha+n)}{\Gamma(n)\Gamma(\alpha)\beta^{\alpha}(z+\frac{1}{\beta})^{\alpha+n}},
\end{aligned}
$$

where,

$$\int_0^{\infty} \frac{\lambda^{\alpha+n-1}e^{-\lambda(z+\frac{1}{\beta})}(z+\frac{1}{\beta})^{\alpha+n}}{\Gamma(\alpha+n)}\, d\lambda = 1.$$

because it is pdf of gamma $\left(\alpha+n, (z+\frac{1}{\beta})^{-1}\right)$ distribution. So integral reduces to 1. Now, posterior density $\pi(\lambda|z)$ becomes,

$$\begin{aligned}\pi(\lambda|z) &= \frac{f(z|\lambda)\pi(\lambda)}{f(z)} \\ &= \frac{z^{n-1}e^{-\lambda z}\lambda^n}{\Gamma(n)}\frac{e^{-\frac{\lambda}{\beta}}\lambda^{\alpha-1}}{\Gamma(\alpha)\beta^\alpha}\frac{\Gamma(n)\Gamma(\alpha)(z+\frac{1}{\beta})^{\alpha+n}\beta^\alpha}{z^{n-1}\Gamma(\alpha+n)} \\ &= \frac{e^{-\lambda(z+\frac{1}{\beta})}\lambda^{\alpha+n-1}(z+\frac{1}{\beta})^{\alpha+n}}{\Gamma(\alpha+n)}, 0<\lambda<\infty.\end{aligned}$$

Hence, the posterior distribution of $\lambda$, $\lambda|z \sim$ gamma $\left(\alpha+n, (z+\frac{1}{\beta})^{-1}\right)$. So exact posterior mean $\mathrm{E}(\lambda|z)$ is

$$\frac{\alpha+n}{z+\frac{1}{\beta}} = \frac{\alpha+n}{n\bar{y}+\frac{1}{\beta}}$$

Now, we approximate the posterior mean of $\lambda$ using a Monte Carlo (MC) sample, given $m$ independent values drawn directly from gamma$(\alpha+n, n\bar{y}+\frac{1}{\beta})$ posterior distribution. Further, the accuracy of the monte carlo approximation is compared with respect to the exact posterior distribution. First of all, we generated $y$ of size $n = 50$ from exponential distribution with parameter $\lambda$, where $\lambda$ was generated from gamma$(\alpha = 8, \beta = 4)$ density function. To generate the independent values of $\lambda$ of the MC sample from the known posterior distribution the function `rgamma` was utilized with posterior rate and shape parameter.

**R Code for Figure 2.6**

```
set.seed(123)
n = 50                                          # sample size
alpha = 8                                       # prior parameter
beta = 4                                        # prior parameter
lambda = rgamma(1,alpha,beta)                   # true lambda
lambda
y = rexp(n = n,rate = lambda)                   # data (simulated here)
p_alpha = alpha + n; p_alpha                    # posterior alpha
p_beta = (1/beta) + n*mean(y); p_beta           # posterior beta
p_mean = p_alpha/p_beta; p_mean                 # exact posterior mean
par(mfrow=c(1,3))
M = c(100,500,1000)
for (m in M) {
  z1 = rgamma(n=m, shape=p_alpha, rate=p_beta)
  hist(z1, probability = TRUE,main=paste("m =",m),
       xlab=expression(lambda),col = "light grey", cex.lab = 1.5, cex.main = 1.5)
  curve(dgamma(x, shape = p_alpha, rate = p_beta),
        col ="blue", add=TRUE, lwd = 2)
  points(lambda, 0, pch = 19,col = "red", cex = 2)}
```

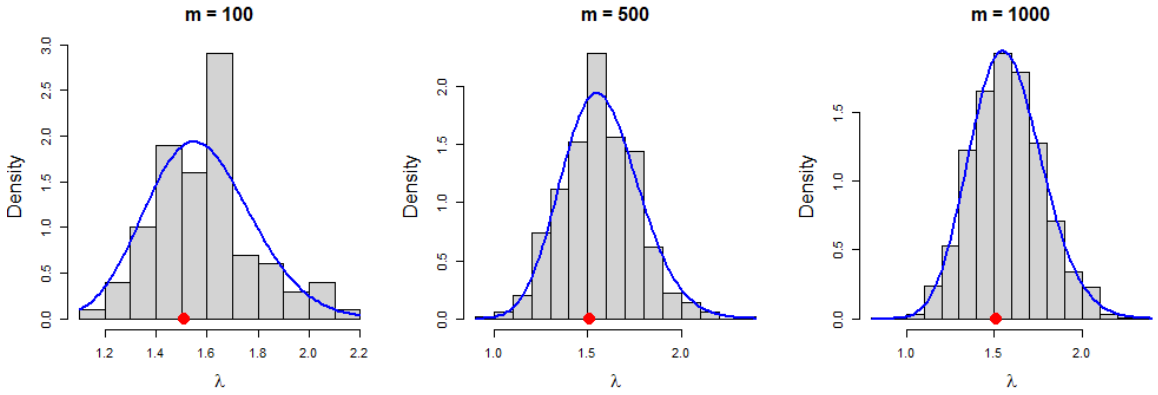



Figure 2.6: Histogram approximation of the posterior probability density of $\lambda$ for different posterior sample of size $m$. Red dot indicates the exact posterior mean (given data). Blue colored curve represents the exact posterior density function.

## 2.6 Exercises

1. (Casella and Berger 2002) If $S^2$ is the sample variance based on a sample of size $n$ from a normal population, we know that $(n-1)S^2/\sigma^2$ has a $\chi^2_{n-1}$ distribution. The conjugate prior for $\sigma^2$ is the inverted gamma pdf, IG$(\alpha, \beta)$, given by,

$$\pi(\sigma^2) = \frac{1}{\Gamma(\alpha)\beta^\alpha}\frac{1}{(\sigma^2)^{\alpha+1}}e^{-1/(\beta\sigma^2)}, \quad 0 < \sigma^2 < \infty,$$

where $\alpha$ and $\beta$ are positive constants. Show that the posterior distribution of $\sigma^2$ is

$$\text{IG}\left(\alpha + \frac{n-1}{2}, \left[\frac{(n-1)S^2}{2} + \frac{1}{\beta}\right]^{-1}\right).$$

Find the mean of this distribution, the Bayes estimator of $\sigma^2$.

2. Suppose that $X_1, X_2, \cdots, X_n$ is a random sample from the distribution with pdf

$$\begin{aligned} f(x|\theta) &= \theta x^{\theta-1}, \quad 0 < x < 1, \\ &= 0, \qquad \text{otherwise.} \end{aligned}$$

Suppose also that the value of the parameter $\theta$ is unknown ($\theta > 0$) and that the prior distribution of $\theta$ is gamma$(\alpha, \beta)$, $\alpha > 0$ and $\beta > 0$. Determine the posterior distribution of $\theta$ and hence obtain the Bayes estimator of $\theta$ under a squared error loss function.

3. (Casella and Berger 2002) Suppose that we observe $X_1, X_2, \cdots, X_n$ where

$$\begin{aligned} X_i|\theta_i &\sim \mathcal{N}(\theta_i, \sigma^2), \quad i = 1, 2, \cdots, n, \quad \text{independent} \\ \theta_i &\sim \mathcal{N}(\mu, \tau^2), \quad i = 1, 2, \cdots, n, \quad \text{independent} \end{aligned}$$

- Show that the marginal distribution of $X_i$ is $\mathcal{N}(\mu, \sigma^2)$ and that, marginally, $X_1, X_2, \cdots, X_n$ are iid. Empirical Bayes analysis would use the marginal distribution of $X_i$'s to estimate the prior parameters $\mu$ and $\tau^2$.
- Show, in general, that if

$$\begin{aligned} X_i|\theta_i &\sim f(\theta_i, \sigma^2), \quad i = 1, 2, \cdots, n, \quad \text{independent} \\ \theta_i &\sim \pi(\theta|\tau), \quad i = 1, 2, \cdots, n, \quad \text{independent} \end{aligned}$$

  then $X_1, X_2, \cdots, X_n$ are iid.

4. (Wasserman 2004) Suppose that we observe $X_1, X_2, \cdots, X_n$ from $\mu, 1$. (a)Simulate a data set (using $\mu = 5$) consisting of $n = 100$ observations. (b)Take $\pi(\mu) = 1$ and find the posterior density and plot the density. (c)Simulate 1000 draws from the posterior and plot the histogram of the simulated values and compare the histogram to the answer in (b). (d)Let $\theta = \exp(\mu)$, then find the posterior density for $\theta$ analytically and by simulation. (e)Obtain a 95 percent posterior interval for $\mu$ and $\theta$.

5. (Wasserman 2004) Consider the Bernoulli($p$) observations:

   0 1 0 1 0 0 0 0 0 0

   Plot the posterior for $p$ using these prior distributions for $p$: beta$(1/2, 1/2)$, beta$(10, 10)$ and beta$(100, 100)$.

6. (Wasserman 2004) Let $X_1, X_2, \cdots, X_n \sim$ Poisson$(\lambda)$. Find the Jeffrey's prior. Also, find the corresponding posterior density function.

# 3 Bayesian Estimation for Linear Regression Problem

Regression models are commonly used in solving real life data analysis problems. Several statistical softwares are now available that provide posterior distributions of the regression coefficients and allows various choice of the prior distributions. In this document, we shall focus on the detailing of the regression models under a Bayesian setting. we are hoping that this will be helpful for the students to understand the concept better rather than blindly believing the output of softwares. We start our discussion with the most simplest setting `Simple Linear Regression`. In the next chapter, we shall discuss the Bayesian inference for nonlinear models.

We have linear regression model;

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \text{where } \epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \phi),$$

and suppose we observe the data $(y_i, x_i)$, $i \in \{1, 2, \cdots, n\}$, then our model for $y$ is:

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \phi), \quad i = 1, 2, \cdots, n. \tag{3.1}$$

Our primary goal is to make inference about the parameters $\beta_0$ and $\beta_1$. If we consider normal priors on the coefficients and an inverse gamma prior on the variance term, then full Bayesian model for this data can be written as,

$$\begin{aligned} y_i|\beta_0, \beta_1, \phi &\sim \mathcal{N}(\beta_0 + \beta_1 x_i, \phi), \qquad i = 1, 2, \cdots, n, \\ \beta_0|\mu_0, \tau_0 &\sim \mathcal{N}(\mu_0, \tau_0) \\ \beta_1|\mu_1, \tau_1 &\sim \mathcal{N}(\mu_1, \tau_1) \\ \phi|\alpha, \gamma &\sim \mathcal{IG}(\alpha, \gamma), \end{aligned} \tag{3.2}$$

where $\mu_0$, $\tau_0$, $\mu_1$, $\tau_1$, $\alpha$, $\gamma$ all are hyper-parameters and we assume that they are known. Let $y_1, y_2, \cdots, y_n$ are observed data of size $n$ from a population having normal distribution with mean $\beta_0 + \beta_1 x_i$ and variance $\phi$, where $\beta_0$, $\beta_1$ and $\phi$ are unknown. Our goal is to draw inference about $\beta_0$, $\beta_1$ and $\phi$. Therefore, joint posterior distribution of $\beta_0$, $\beta_1$ and $\phi$ can be written as,

$$f(\beta_0, \beta_1, \phi|y) = \frac{f(y, \beta_0, \beta_1, \phi)}{f_Y(y)}, \tag{3.3}$$

where, $f_Y(y)$ is the marginal distribution of $y$. Now equation (3.3) reduces to

$$f(\beta_0, \beta_1, \phi|y) = \frac{f(y|\beta_0, \beta_1, \phi) \cdot f(\beta_0, \beta_1, \phi)}{\int_{\beta_0}\int_{\beta_1}\int_{\phi} f(y, \beta_0, \beta_1, \phi) d\beta_0 d\beta_1 d\phi}. \tag{3.4}$$

We assume that $\beta_0$, $\beta_1$ and $\phi$ are independent variable, therefore $f(\beta_0, \beta_1, \phi)$ can be written as,

$$f(\beta_0, \beta_1, \phi) = f(\beta_0) \cdot f(\beta_1) \cdot f(\phi). \tag{3.5}$$

and let $K = \int_{\beta_0}\int_{\beta_1}\int_{\phi} f(y, \beta_0, \beta_1, \phi)\mathrm{d}\beta_0\mathrm{d}\beta_1\mathrm{d}\phi$, this integration is difficult to compute. We know that

$$\text{posterior} \propto \text{likelihood} \times \text{priors}.$$

Therefore, equation (3.4) becomes,

$$\begin{aligned}
f(\beta_0, \beta_1, \phi|y) &= \frac{1}{K} f(y|\beta_0, \beta_1, \phi) \cdot f(\beta_0, \beta_1, \phi) \\
f(\beta_0, \beta_1, \phi|y) &\propto f(y|\beta_0, \beta_1, \phi) \cdot f(\beta_0, \beta_1, \phi) \\
f(\beta_0, \beta_1, \phi|y) &\propto \left[\prod_{i=1}^{n} f(y_i|\beta_0, \beta_1, \phi)\right] f(\beta_0) \cdot f(\beta_1) \cdot f(\phi) \\
&\propto \text{likelihood} \times \text{priors}.
\end{aligned}$$

From the joint posterior distribution we can not generate random sample because it does not follow any common known distribution. So, we generate random sample from conditional posterior distribution of each parameter. In the following, we calculate the conditional posterior distribution of each parameter. Using the relation

$$f(y, \beta_0, \beta_1, \phi) = f(\beta_0 \mid \beta_1, \phi, y) \cdot f(\beta_1, \phi, y),$$

the conditional posterior distribution of the parameter $\beta_0$ is given by,

$$
\begin{aligned}
f(\beta_0|\beta_1,\phi,y) &= \frac{f(y,\beta_0,\beta_1,\phi)}{f(\beta_1,\phi,y)} \\
&= \frac{f(y|\beta_0,\beta_1,\phi)\cdot f(\beta_0,\beta_1,\phi)}{f(\beta_1,\phi,y)} \quad [\text{since, } f(y,\beta_0,\beta_1,\phi)=f(y|\beta_0,\beta_1,\phi)\cdot f(\beta_0,\beta_1,\phi)] \\
&= \frac{f(y|\beta_0,\beta_1,\phi)f(\beta_0)f(\beta_1)f(\phi)}{f(y|\beta_1,\phi)f(\beta_1,\phi)} \quad [\text{since, } f(\beta_0,\beta_1,\phi)=f(\beta_0)f(\beta_1)f(\phi)] \\
&= \frac{f(y|\beta_0,\beta_1,\phi)f(\beta_0)f(\beta_1)f(\phi)}{f(y|\beta_1,\phi)f(\beta_1)f(\phi)} \\
&= \frac{f(y|\beta_0,\beta_1,\phi)f(\beta_0)}{f(y|\beta_1,\phi)}. \qquad (3.6)
\end{aligned}
$$

Let, $m = f(y|\beta_1,\phi)$, therefore,

$$
\begin{aligned}
f(\beta_0|\beta_1,\phi,y) &= \frac{1}{m}f(y|\beta_0,\beta_1,\phi)\cdot f(\beta_0) \\
&\propto f(y|\beta_0,\beta_1,\phi)\cdot(\beta_0) \\
&\propto \left[\prod_{i=1}^{n} f(y_i|\beta_0,\beta_1,\phi)\right]\cdot f(\beta_0). \qquad (3.7)
\end{aligned}
$$

Like the joint posterior, the conditional posterior is also proportional to the product of the likelihood and prior of the parameter of interest. The computation yields the similar expression for $\beta_1$ as given below;

$$
f(\beta_1|\beta_0,\phi,y) \propto \left[\prod_{i=1}^{n} f(y_i|\beta_0,\beta_1,\phi)\right]\cdot f(\beta_1), \qquad (3.8)
$$

and,

$$
\begin{aligned}
f(\phi|\beta_0,\beta_1,y) &\propto \left[\prod_{i=1}^{n} f(y_i|\beta_0,\beta_1,\phi)\right]\cdot f(\phi) \\
&= \left(\frac{1}{\sqrt{2\pi\phi}}\right)^n e^{-\frac{1}{2\phi}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}\frac{\gamma^\alpha}{\Gamma(\alpha)}\phi^{-\alpha-1}e^{-\frac{\gamma}{\phi}} \\
&= \phi^{-\frac{n}{2}}\phi^{-\alpha-1}e^{-\frac{1}{\phi}\left[\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2+\gamma\right]} \\
&= \phi^{-(\alpha+\frac{n}{2}+1)}e^{-\frac{1}{\phi}\left[\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2+\gamma\right]}. \qquad (3.9)
\end{aligned}
$$

The prior specification for $\phi$ is given by the inverse gamma density $\mathcal{IG}(\text{shape} = \alpha, \text{rate} = \gamma)$ function which is given as follows,

$$f(\phi; \alpha, \gamma) = \frac{\gamma^{\alpha}}{\Gamma(\alpha)} \phi^{-\alpha-1} e^{\frac{-\gamma}{\phi}}, \quad \phi > 0. \tag{3.10}$$

Comparing the conditional posterior of $\phi$ with equation (3.10), we can see that $f(\phi|\beta_0, \beta_1, y)$ yields the inverse gamma distribution with the parameter values as, shape $= \alpha + \frac{n}{2}$ and rate $= \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + \gamma$.

$$f(\phi|\beta_0, \beta_1, y) = \mathcal{IG}\left(\alpha + \frac{n}{2}, \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 + \gamma\right). \tag{3.11}$$

Our goal is to draw inference about the population parameters $\beta_0$, $\beta_1$ and $\phi$. To compute the posterior credible interval for the parameters, we require to draw samples from the joint posterior distribution of $\beta_0$, $\beta_1$ and $\phi$ i.e $f(\beta_0, \beta_1, \phi)$. A posterior credible interval is a range of values within which an unknown parameter lies with a certain probability, based on the posterior distribution in Bayesian inference. It represents the uncertainty around the estimated parameter after incorporating the observed data and prior beliefs. We observe that $f(\beta_0, \beta_1, \phi)$ does not follow any common known distribution from which it easier to draw the samples. So, we utilize the Gibbs sampling method, in which, random samples are drawn from the conditional posterior distribution. So we compute the posterior distribution of $\beta_0$, conditional on $\beta_1$ and $\phi$ (Eq. 3.7). Similarly conditional posterior distribution for $\beta_1$ and $\phi$ are evaluated. We evaluated the conditional posterior $f(\beta_0|\beta_1, \phi)$ and $f(\beta_1|\beta_0, \phi)$ by assuming normal priors from $\beta_0$ and $\beta_1$ and inverse gamma prior for $\phi$. The calculation suggested that $f(\beta_0|\beta_1, \phi)$ and $f(\beta_1|\beta_0, \phi)$ does not follow any common distribution. However, conditional posterior distribution of $\phi$ belongs to the family of inverse gamma distribution (Eq. 3.11). This is due to the conjugacy relationship between normal distribution and inverse gamma distribution with respect to the parameter representing the variance. We utilized grid approximation method for generating samples from the conditional posterior of $\beta_0$ and $\beta_1$. For generating samples from the posterior of $\phi$ we used direction simulation from the inverse gamma distribution using `invgamma` package (Kahle and Stamey 2017) from `R` software for statistical computing. However, this can also be carried out by using probability integral transform.

**Note**

If $X_1, X_2, \cdots, X_n$ be a random sample of size $n$ from a population following $\mathcal{N}(\theta, \sigma^2)$. If $S^2$ is the sample variance based on a sample of size $n$ from a normal population, we know that $\frac{(n-1)S^2}{\sigma^2}$ has a $\chi^2_{n-1}$ distribution. The conjugate prior for $\sigma^2$ is the inverse gamma probability density $\mathcal{IG}(\alpha, \beta)$. Then the posterior distribution of $\sigma^2$ is,

$$\mathcal{IG}\left(\alpha + \tfrac{n-1}{2}, \left[\tfrac{(n-1)S^2}{2} + \tfrac{1}{\beta}\right]^{-1}\right).$$

In grid approximation method, we first define grid for estimating the posterior. After that at every grid point we compute the likelihood and prior. By multiplying likelihood and prior we get unstandardized posterior values at each grid point. Dividing each value by sum of all values we get the standardized posterior values. This discrete distribution is then used to simulate realization from exact posterior distribution. From all grid point we randomly select one grid point according to their probability. This process is executed by using `sample()` function available in `R`. We simulate this process for long time. After the sampling is done, we get a chain of simulated values from the posterior distribution. Since, it is a Markov chain, the sample values autocorrelated. To obtain independent observations we used thinning by taking every sixteen observation after the burn-in period. However, the choice should be decided by investigating the autocorrelation plot of the chain.

At this point, it is nice to have an idea on Gibbs sampling method. Suppose we have joint probability distribution of parameters (not necessarily in known form), we want to generate random sample from marginal probability distribution of each parameter. Then we use Gibbs sampling method. In Gibbs sampling method, we generate values from the conditional distribution, but we want values from marginal probability(unconditional) distribution. The values which we get from conditional probability distribution of each parameter, after certain burn-in period acts as a realization from the true marginal probability distribution of respective parameters. In the current context on Bayesian linear regression, we want to simulate observations from the **joint posterior density** $f(\beta_0, \beta_1, \phi|y)$. Since, it is difficult to simulate from the joint distribution, we simulate from the conditional posterior distribution $f(\beta_0|\beta_1, \phi, y)$.

## 3.1 Simulation study

For checking purpose using simulation, we have fixed population parameters of a linear regression model as $\beta_0 = 3$, $\beta_1 = 1$ and $\phi = 1$. Since we do not have the real data in our hand, so we created artificial data using following model with these fixed values of parameters (Figure 4.1),

$$y_i = \beta_0 + \beta_1 x_i + \mathcal{N}(\text{mean} = 0,\ \text{sd} = \sqrt{\phi}), i \in \{1, 2, \cdots, n\}.$$

We have prior for $\beta_0$ and $\beta_1$ is Normal distribution with mean($\mu_0$), variance($\tau_0$) and mean($\mu_1$), variance($\tau_1$), respectively. Parameter $\phi$ have prior following the inverse gamma distribution with shape($\alpha$) and rate($\gamma$). The hyper parameters $\mu_0$, $\tau_0$, $\mu_1$, $\tau_1$, $\alpha$, $\gamma$ are assumed to be known. After that we computed the conditional posterior distribution of $\beta_0$, $\beta_1$ and $\phi$ assuming the above priors. Conditional posterior of $f(\beta_0|\beta_1, \phi)$ and $f(\beta_1|\beta_0, \phi)$ do not follow any common know distributions. So we have used grid approximation method for generating samples from

the conditional posterior of $\beta_0$ and $\beta_1$. From this process we get the unstandardized posterior values at the grid points, then we standardized this values by dividing the sum of all unstandardized posterior values. We randomly select one grid sample and we iterate this process for 100000 times but we consider starting 70000 random sample as burn-in period and after a burn-in period. we want the independent random sample, to obtain independent sample we used thinning by taking every sixteen sample after the burn-in period in posterior of $\beta_0$ and $\beta_1$. In posterior of $\phi$ we used thinning by taking every third sample after the burn-in period. However, the choice is good should be decided by plotting `ACF` of after thinning sample. After checking autocorrelation of thinning simulated sample, we draw histogram of $\beta_0$, $\beta_1$ and $\phi$. Histogram of conditional posterior distribution of $\beta_0$ and $\beta_1$ of normal distribution and conditional posterior distribution of $\phi$ is inverse gamma distribution approximated using the posterior means of $\beta_0$ and $\beta_1$. Population parameters are fall within the credible interval of posterior distribution $\beta_0$, $\beta_1$ and $\phi$. In developing the Gibbs sampling algorithm for Bayesian regression, we have used the inbuilt function `rnorm` and `rinvgamma` from `R` base and `invgamma` package respectively. However by virtue of the probability integral transform both of them can be simulated from uniform random number.

```
beta_0 = 3                                    # population intercept
beta_1 = 1                                    # population slope
phi = 1                                       # population error variance
x = seq(1, 5, length.out = 50)                # fixation of the population parameter
set.seed(123)                                 # we fixing a randomness in output
y = beta_0 + beta_1 * x + rnorm(length(x), 0, sqrt(phi))
data = data.frame(y, x)                       # data containing population size
plot(x,y, col = "red", pch = 19)
```

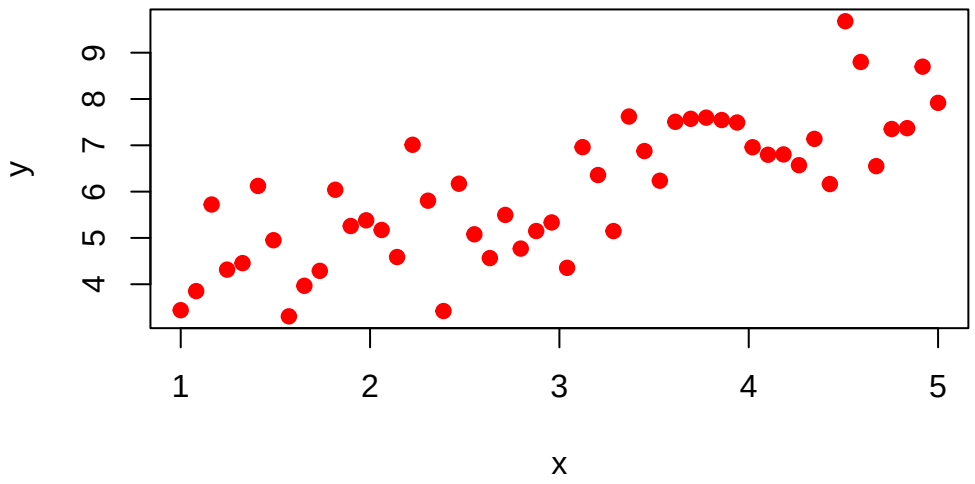



Figure 3.1: Plot of simulated data for the study of linear regression generated using the model $y_i = \beta_0 + \beta_1 x_i + \mathcal{N}(\text{mean} = 0,\ \text{sd} = \sqrt{\phi}), i \in \{1, 2, \cdots, n\}$. The parameters values are fixed as $\beta_0 = 3,\ \beta_1 = 1,\ \phi = 1$, and generated data of length $n = 50$.

In the above code the population parameters have been fixed for simulation purpose. The parameters are set as $\beta_0 = 3$, $\beta_1 = 1$ and $\phi = 1$. The seed has been fixed so that the results can be reproduced later as well. The reader is encouraged to run the complete code step by step rather than running it completely in a single step. In the following section, we specify the prior distribution for the model parameters.

```
# Prior for beta_0
mu_0 = 2; tau_0 = 0.4                          # prior parameter
prior_beta_0 = function(x){                    # function for prior beta_1
  dnorm(x, mean = mu_0, sd = sqrt(tau_0))
}

# Prior for beta_1
mu_1 = 2; tau_1 = 0.5                          # prior parameter
prior_beta_1 = function(x){                    # function for prior beta_1
  dnorm(x, mean = mu_1, sd = sqrt(tau_1))
}
# Prior for phi.
```

```
library(invgamma)                          #
alpha = 2; gamma = 2                       # prior parameters
prior_phi = function(x){                   # function for prior phi
  dinvgamma(x, shape = alpha, rate = gamma)
}
```

The following code describes the likelihood function. Recall that to compute the posterior we have to multiply the likelihood function with the prior distribution. So, here we write separate function for the likelihood function. The likelihood function is the joint distribution of the data values where the parameters are considered as variable. Also, note that the function directly compute the log-likelihood; this is is helpful to avoid truncation errors made by the software.

```
likelihood = function(data, params){   # likelihood function of data given parameters
  y = data[,1]
  x = data[,2]
  n = nrow(data)
  beta_0 = params[1]
  beta_1 = params[2]
  phi = params[3]
  log_lik = -n*log(sqrt(2*pi)) - (n/2)*log(phi) - sum((y - beta_0-beta_1*x)^2)/(2*phi)
  # log-likelihood of data given parameters
  return(exp(log_lik))
}
```

The following codes are self explanatory. Sufficient comments have been included. Note that we have created the grid values by dividing an interval with a step size of 0.01. Choice of this interval plays critical role for a successful grid approximation. This interval is essentially act as a support for the posterior distribution of the designated parameter. The simulation would start with an initial choice of the parameters $\beta_0^{(0)}$, $\beta_1^{(0)}$ and $\phi^{(0)}$ which have been simulated from the prior distribution.

```
# storage for posterior samples
iteration = 100000                         # number of iterations
post_beta_0 = rep(NA, iteration)           # storage for posterior of beta_0
post_beta_1 = rep(NA, iteration)           # storage for posterior of beta_1
post_phi = rep(NA, iteration)              # storage for posterior of phi
# Specification of grids
step = 0.01                                          # grid step
grid_beta_0 = seq(from = -5, to = 10, by = step) # fixation of the grid of beta_0
grid_beta_1 = seq(from = -5, to = 5, by = step)  # fixation of the grid of beta_1
# Initialization of the posterior samples
```

```
post_beta_0[1] = rnorm(1, mean = mu_0, sd = sqrt(tau_0))   # posterior beta_0
post_beta_1[1] = rnorm(1, mean = mu_1, sd = sqrt(tau_1))   # posterior beta_1
post_phi[1] = rinvgamma(n = 1, shape = alpha, rate = gamma)# posterior phi
```

The following codes give the code for Gibbs sampling method to generate the posterior distribution from the marginal posterior distribution. In the text, it has been mentioned that we do not have idea about the joint posterior distribution of the parameters, so we simulate from the marginal posterior distribution. Since, the exact functional form of the marginal posterior is also not known, we approximate it by using grid approximation.

```
for(i in 2:iteration){

  # section for posterior sample of beta_0
  tmp_post_beta_0 = rep(NA, length(grid_beta_0))
  for(j in 1:length(grid_beta_0)){
    params = c(grid_beta_0[j], post_beta_1[i-1], post_phi[i-1])
    tmp_post_beta_0[j] = likelihood(data = data, params = params)
                                * prior_beta_0(grid_beta_0[j])
  }
  prob = tmp_post_beta_0/sum(tmp_post_beta_0)
  post_beta_0[i] = sample(grid_beta_0, 1, prob = prob) # sample of posterior
  beta_0 after normalize

  # section for posterior sample of beta_1
  tmp_post_beta_1 = rep(NA, length(grid_beta_1))
  for(j in 1:length(grid_beta_1)){
    params = c(post_beta_0[i], grid_beta_1[j], post_phi[i-1])
    tmp_post_beta_1[j] = likelihood(data = data, params= params)
                                * prior_beta_1(grid_beta_1[j])
  }
  prob = tmp_post_beta_1/(sum(tmp_post_beta_1))
  post_beta_1[i] = sample(grid_beta_1, 1, prob = prob) # sample of posterior
  beta_1 after normalize

  # section for posterior sample of phi
  shape = alpha + nrow(data)/2
  rate = (1/2)*sum((y - post_beta_0[i] - post_beta_1[i]*x)^2) + gamma
  post_phi[i] = rinvgamma(1, shape = shape, rate = rate)  # sample of posterior phi
}
```

Trace plots are useful tools to convergence of the chains which are expected to converge the posterior distribution. After an initial burn in period, the values would act as samples from

the target (desired posterior) distribution. In the following we have considered first 70% of the simulated values as the burn in step. There is no hard and fast rule for it.

```
# Trace plots
par(mfrow = c(2,2))
# trace plot of posterior of beta_0
plot(post_beta_0, type = "l", main = bquote("Trace plot of"~ beta[0]),col = "red")
# trace plot of posterior of beta_1
plot(post_beta_1, type = "l", main = bquote("Trace plot of"~ beta[1]),col = "red")
# trace plot of posterior of phi
plot(post_phi, type = "l", main = bquote("Trace plot of"~ phi),col = "red")
```

It is important to note that simulation of $\beta_0^{(j)}$ depends on the value of $\beta_0^{(j-1)}$. Similarly for other parameters as well. Thus, the posterior values are identically distributed (after a sufficiently large step) but not independent. So, the test for autocorrelation has been performed as the values have been selected accordingly. For example, we may consider every 10th values after the burn in period. This is also known as thinning. The function `acf()` has been utilized for this purpose.

```
# Plot of conditional posterior density functions of beta_0, beta_1, phi
cut = 0.7                                          # for cutting starting 70% sample
par(mfrow = c(2,3))

# Section for plot acf of posterior beta_0
u = post_beta_0[ceiling(iteration*cut):iteration]   # after burn-in period
index = seq(1,length(u), by = 16)                  # thining
post_beta_b0_thinning = u[index]                   # values after thinning
acf(post_beta_b0_thinning,main = bquote("acf of posterior "~beta[0])) #checking autocorrel

# Section for plot acf of posterior beta_1
u = post_beta_1[ceiling(iteration*cut):iteration]   # after burn-in period
index = seq(1,length(u), by = 16)                  # thinning
post_beta_b1_thinning = u[index]                   # values after thinnig
acf(post_beta_b1_thinning,main = bquote("acf of posterior "~beta[1]))# checking autocorrel

# Section for plot acf of posterior phi
u = post_phi[ceiling(iteration*cut):iteration]      # after burn-in period
index = seq(1,length(u), by = 3)                   # thinning
post_phi_thinning = u[index]                       # values after thinning
acf(post_phi_thinning,main = bquote("acf of posterior "~phi)) # checking autocorrelation
```

After the thinning has been done, we visualize the approximated distribution by means of

histogram. As we can see that the posterior density of $\beta_0$ and $\beta_1$ can be well approximated by the normal distribution. The posterior density of $\phi$ is positively skewed and can be well approximated by the inverted gamma density function. This is due to conjugacy which has been mentioned earlier. The exact posterior density for $\phi$ has been computed using the mean of the posterior samples of $\beta_0$ and $\beta_1$.

```
# Section for plot histogram of posterior beta_0 after thinning
hist(post_beta_b0_thinning, probability = TRUE,
    main = bquote(" Posterior of "~ beta[0]),
    xlab = expression(beta[0]),col = "grey")
credible_interval_beta_0 = quantile(post_beta_b0_thinning, c(2.5, 97.5)/100)
abline(v = credible_interval_beta_0,col = "blue",lwd = 2)
legend("topright",legend = "Credible Interval",lwd =2,col = "blue",lty = 2,cex = 0.5)

# Section for plot histogram of posterior beta_1 after thinning
hist(post_beta_b1_thinning,probability = TRUE,
    main = bquote("Posterior of "~beta[1]),
    col = "grey",xlab = expression(beta[1]))
credible_interval_beta_1 = quantile(post_beta_b1_thinning, c(2.5, 97.5)/100)
abline(v = credible_interval_beta_1, col = "blue", lwd = 2)
legend("topright",legend = "Credible Interval",lwd =2,col = "blue",lty = 2,cex = 0.5)

# Section for plot histogram of posterior phi after thinning
hist(post_phi_thinning, probability = TRUE,
    main = bquote("psterior of "~ phi),
    col ="grey",ylim = c(0,3), xlab = expression(phi))
credible_interval_phi = quantile(post_phi_thinning, c(2.5, 97.5)/100)
abline(v=credible_interval_phi,col = "blue",lwd = 2)
legend("topright",legend = c("Credible Interval","True curve of  phi"),
    lwd = c(2,2), col = c("blue","red"), lty = 2, cex = 0.5)

# Approximating the posterior distribution of phi using the posterior means of beta_0 and
mean_post_beta_0 = mean(post_beta_b0_thinning)
mean_post_beta_1 = mean(post_beta_b1_thinning)
shape = alpha + nrow(data)/2
rate = (1/2)*sum((y - mean_post_beta_0 - mean_post_beta_1*x)^2) + gamma
curve(dinvgamma(x, shape = shape, rate = rate),add = TRUE, lwd=2, col = "red")
```

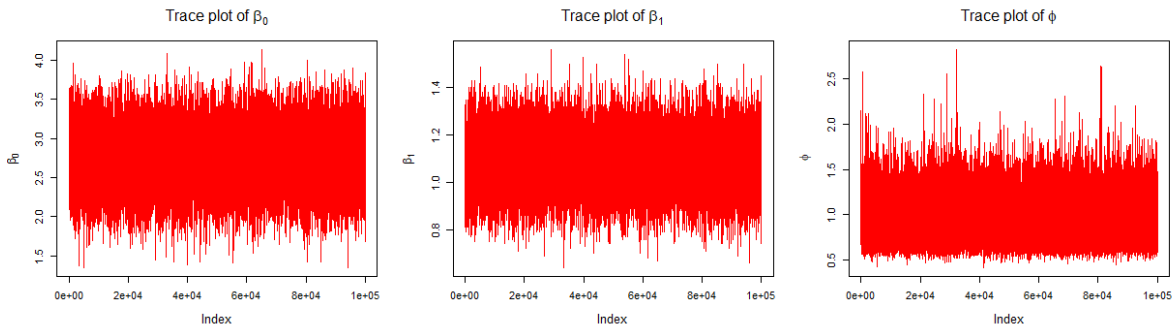
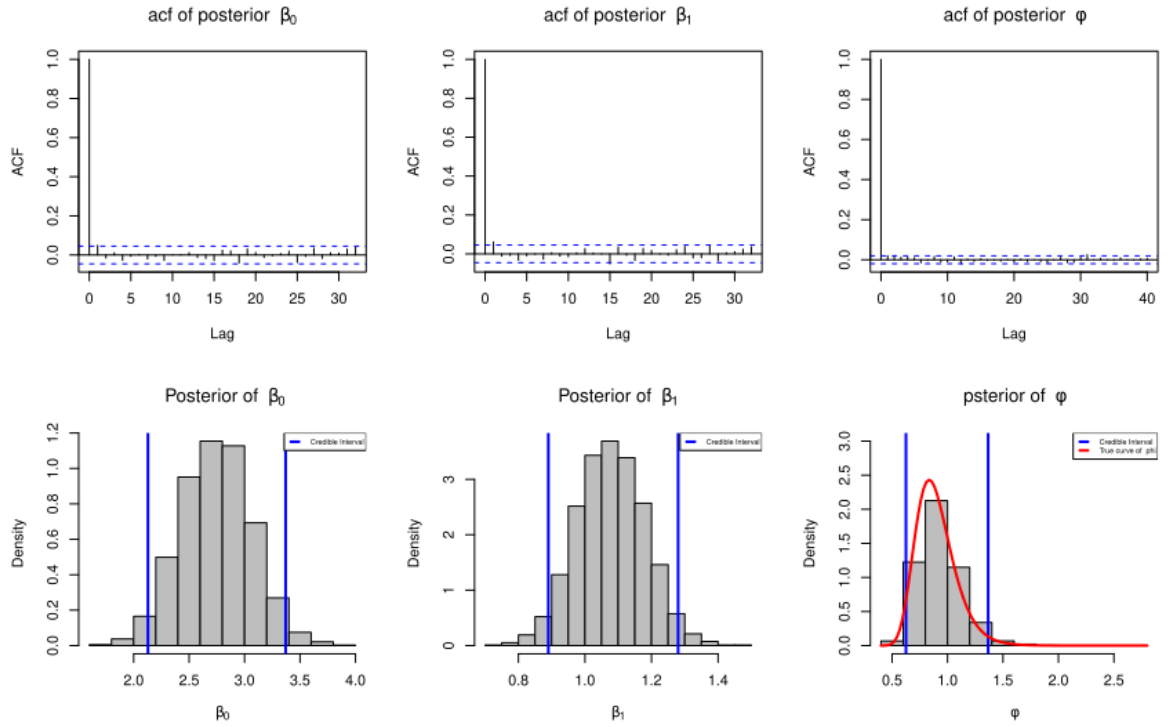



Figure 3.2: Trace plot of posterior $\beta_0$, $\beta_1$ and $\phi$.




Up to this point, we have developed a clear understanding of linear regression from scratch under the Bayesian framework and demonstrated it on simulated data for better comprehension. The Bayesian estimation of the linear regression parameters has been conducted using custom-written code, without relying on any R packages. This approach provides a transparent view of the underlying processes. Once the mechanism is well understood, the same analysis can be performed using built-in packages. Some of the packages available in R for Bayesian linear regression are `rstanarm` (Goodrich et al. 2024), and `brms` (Bürkner 2017). The `rstanarm` package provides a user-friendly interface to fit Bayesian models using Stan,

while `brms` extends this by allowing complex model structures using formula syntax.

# 4 Bayesian Estimation to Nonlinear Regression Problem

Let $N(t)$ denotes the population size at time $t$. $N(t)$ is assumed to be a continuously differentiable function of $t$. Let the dynamics of the population is governed by the $\theta$-logistic growth equation

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = r_m N(t)\left[1-\left(\frac{N(t)}{K}\right)^{\theta}\right], \quad N(0)=N_0, \tag{4.1}$$

where $r_m$ is the intrinsic growth rate of the population, $\theta$ is the curvature parameter depicting the shape of the per capita growth rate (PGR) and population density; $K$ is the carrying capacity of the environment. For $\theta = 1$, we get the logistic growth function, which is a linearly decreasing function of $N$. For $\theta < 1$ and $\theta > 1$, the PGR-density relationship is concave upward and concave downward, respectively. The equation (4.1) is a Bernoulli type differential equation whose solution is given by the following:

$$N(t) = \frac{K}{\left\{1+\left[\left(\frac{K}{N_0}\right)^{\theta}-1\right]e^{-r_m t\theta}\right\}^{\frac{1}{\theta}}}. \tag{4.2}$$

It is easy to verify that as $t \to \infty$, $N(t) \to K$. In ecological literature, demographic and environmental variations are main source of stochastic population changes. If the population is subject to the environmental noise alone then the corresponding diffusion equation is described in the form of Itô stochastic differential equation (SDE) as

$$\mathrm{d}N(t) = rN(t)\left[1-\left(\frac{N(t)}{K}\right)^{\theta}\right]\mathrm{d}t + \sigma_e N(t)\mathrm{d}W(t). \tag{4.3}$$

If the dynamics of the population is subject to the demographic variation only, then the growth dynamics is governed by

$$\mathrm{d}N(t) = rN(t)\left[1-\left(\frac{N(t)}{K}\right)^{\theta}\right]\mathrm{d}t + \sigma_d\sqrt{N(t)}\mathrm{d}W(t), \tag{4.4}$$

$\sigma_d^2$ and $\sigma_e^2$ are the demographic and environmental variances, respectively. $\{W(t), t \geq 0\}$ is the standard Brownian motion, so that $\mathrm{d}W(t) \sim \mathcal{N}(0, \mathrm{d}t)$. We consider the population dynamics to be subjected to environmental stochasticity only. We seek to estimate the model parameters by using the maximum likelihood estimation method. To do that we first simulate the data following the above model using some assumed values of the parameters. The equation (4.4) can be described as follows,

$$N(t+\Delta t) - N(t) = rN(t)\left[1 - \left(\frac{N(t)}{K}\right)^{\theta}\right]\Delta t + \sigma_e N(t)\left[W(t+\Delta t) - W(t)\right]. \tag{4.5}$$

We have $\{t_0, t_1, \cdots, t_n\}$ be $n+1$ time points with $N(t_0) = X_0$, which is fixed. $W(t+\Delta t) - W(t)$ is the Brownian motion with mean 0 and variance $\Delta t$. The simulation is carried out in the following way: The conditional distribution of $N(t+\Delta t)$ given $N(t)$ is given by a normal distribution with mean $N(t) + rN(t)\left[1 - \left(\frac{N(t)}{K}\right)^{\theta}\right]\Delta t$ and variance $\sigma_e^2 N(t)^2 \Delta t$ and notionally can be written as follows:

$$N(t+\Delta t) - N(t) | N(t) \sim \mathcal{N}\left(rN(t)\left[1 - \left(\frac{N(t)}{K}\right)^{\theta}\right]\Delta t, \sigma_e^2 N(t)^2 \Delta t\right). \tag{4.6}$$

We fixed $t_0 = 0$ and $\Delta t = 0.1$. Let $N_{t_0}, N_{t_1}, \ldots, N_{t_n}$ are the $n+1$ population sizes are generated using the above rule. The likelihood function $\mathcal{L}(\beta)$ of the sample $N_0$, $N_1$, $N_2$, $\cdots$, $N_n$ is the joint density of the observations treated as a function of the parameter $\beta$. The method of maximum likelihood provides as estimate of $\beta$ any value $\widehat{\beta}$ which maximizes $\mathcal{L}$. Note that $\{N_{t_i}\}_{i=0}^{n}$ are not independent but the differences are assumed to be independent, So the likelihood function can be written as

$$\begin{aligned}
\mathcal{L} &= f\left(N_1, N_2, \ldots, N_n | N_0; r, K, \theta, \sigma_e\right) \\
&= f\left(N_1, N_2, \ldots, N_n | N_0; \beta\right) \\
&= f\left(N_n | N_{n-1}, \ldots, N_0; \beta\right) f\left(N_{n-1} | N_{n-2}, \ldots, N_0; \widetilde{\beta}\right) \ldots f\left(N_1 | N_0; \widetilde{\beta}\right) \\
&= \prod_{i=1}^{n} f\left(N_i | N_{i-1} \ldots, N_0; \beta\right) \\
&= \prod_{i=1}^{n} f\left(N_i | N_{i-1}; \beta\right) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_e N_{i-1}\sqrt{\Delta t_i}} e^{-\frac{1}{2\sigma_e^2}\Sigma_{i=1}^{n}\left(\frac{N_i - \mu_{N_{i-1}}}{N_{i-1}}\right)^2} \\
&= \frac{1}{\sqrt{2\pi}^n (\sigma_e)^n (\prod_{i=1}^{n} N_{i-1}\sqrt{\Delta t_i})} e^{-\frac{1}{2\sigma_e^2}\Sigma_{i=1}^{n}\left(\frac{N_i - \mu_{N_{i-1}}}{N_{i-1}}\right)^2}
\end{aligned}$$

where $\mu_{N_i} = N(t_i) + rN(t_i)\left[1 - \left(\frac{N(t_i)}{K}\right)^{\theta}\right]\Delta t$. The log-likelihood function is given by

$$
\begin{aligned}
l(\beta) &= \log \mathcal{L}(\beta) && (4.7)\\
&= -n\ln\sqrt{2\pi} - n\ln\sigma_e - \ln\prod_{i=1}^{n}(N_{i-1}\sqrt{\Delta t_i}) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left(\frac{N_i - \mu_{N_{i-1}}}{N_{i-1}}\right)^2 && (4.8)\\
&= -n\ln(\sqrt{2\pi}) - n\ln\sigma_e - \sum_{i=1}^{n}\ln\sqrt{\Delta t_i} - \sum_{i=1}^{n}\ln(N_{i-1}) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left[\frac{N_i - N_{i-1} - rN_{i-1}\left[1 - \left(\frac{N_{i-1}}{K}\right)^{\theta}\right]\Delta t_i}{N_{i-1}}\right] && (4.9)
\end{aligned}
$$

## 4.1 Estimation of Parameters

Consider $\{N_1, N_2, \cdots, N_{n+1}\}$ be the observed time series data. The absolute change in population size at time $t$ is approximated as

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} \approx \frac{N(t+\Delta t) - N(t)}{\Delta t}.$$

We assume the yearly changes in population size, so that $\Delta t = 1$. The relative changes in population sizes is given by

$$R(t) = \frac{1}{N}\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\log N \approx \log N(t+1) - \log N(t). \tag{4.10}$$

So, for a given population time series data of size $n+1$, we have $n$ many observed per capita growth rates $\{R(1), R(2), \cdots, R(n)\}$. We would like to investigate the dynamics of the population when it is subject to environmental stochastic perturbations alone. The following stochastic differential equation has been used to describe to population changes that is exposed to environmental variability.

$$\mathrm{d}N = r_m N(t)\left[1 - \left(\frac{N(t)}{K}\right)^{\theta}\right]\mathrm{d}t + \sigma_e N(t)\mathrm{d}W(t), \tag{4.11}$$

where $\sigma_e$ represents the intensity of the stochastic perturbation and $W(t)$ is the one dimensional Brownian motion which satisfies $\mathrm{d}W(t) \approx W(t+\mathrm{d}t) - W(t) \sim \mathcal{N}(0, \mathrm{d}t)$. Our goal is to estimate the parameters of the stochastic model by employing Bayesian statistical methods. First we compute the likelihood function of the model parameters given the population size. It is to be noted that the likelihood function can be written either by conditional on the population size $\{N_t\}$ or on the growth rates $\{R_t\}$. We can use the either observations on $\{R_t\}$ or $\{N_t\}$ to obtain the posterior distribution of the model parameters. Now, we shall write down the

likelihood function given the population time series. Using the equation (4.11) and utilizing the distributional assumptions, we see that the absolute changes in the population size, conditional on the current size, is normally distributed. So we obtain the following (assuming $\Delta t = 1$):

$$\begin{aligned}
N(t+1) - N(t)|N(t) &\sim \mathcal{N}\left(r_m N(t)\left[1 - \left(\frac{N(t)}{K}\right)^{\theta}\right], \sigma_e^2 N(t)^2\right) \\
\frac{N(t+1) - N(t)}{N(t)}|N(t) &\sim \mathcal{N}\left(r_m\left[1 - \left(\frac{N(t)}{K}\right)^{\theta}\right], \sigma_e^2\right) \\
N(i+1) - N(i) &\sim \mathcal{N}\left(r_m N(i)\left[1 - \left(\frac{N(i)}{K}\right)^{\theta}\right], \sigma_e^2 N(i)^2\right) \\
N(i+1)|N(i) &\sim \mathcal{N}\left(N(i) + r_m N(i)\left[1 - \left(\frac{N(i)}{K}\right)^{\theta}\right], \sigma_e^2 N(i)^2\right) \qquad (4.12)
\end{aligned}$$

We can estimate the parameters using (4.12) using the method of maximum likelihood. Therefore by using (4.12), our goal is to estimate parameters $r_m$, $\theta$, $K$, $\sigma_e^2$. If we use $R_t$ values, then the following equation will be used for computation of the likelihood.

$$R(t)|N(t) \sim \mathcal{N}\left(r_m\left[1 - \left(\frac{N(t)}{K}\right)^{\theta}\right], \sigma_e^2\right), \qquad (4.13)$$

so that the log-likelihood function will be,

$$\text{likelihood} = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_e^2) - \frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left\{R(i) - r_m\left[1 - \left(\frac{N(i)}{K}\right)^{\theta}\right]\right\}^2. \qquad (4.14)$$

To draw inference about $r_m$, $\theta$, $K$ and $\sigma_e^2$, we consider inverted gamma prior for variance term $\sigma_e^2$, gamma prior for $\theta$, normal prior for $K$ and normal prior for $r_m$. Then the complete Bayesian model for this data can be written as,

$$\begin{aligned}
N(i+1)|N(i), r_m, \theta, K, \sigma_e^2 &\sim \mathcal{N}\left(N(i) + r_m N(i)\left[1 - \left(\frac{N(i)}{K}\right)^{\theta}\right], \sigma_e^2 N(i)^2\right) \\
r_m|\alpha_0, \beta_0 &\sim \mathcal{N}(\alpha_0, \beta_0) \\
K|\mu_K, \sigma_K^2 &\sim \mathcal{N}(\mu_K, \sigma_K^2) \\
\theta|\alpha_1, \beta_1 &\sim \mathcal{G}(\alpha_1, \beta_1) \\
\sigma_e^2|\alpha_2, \beta_2 &\sim \mathcal{IG}(\alpha_2, \beta_2),
\end{aligned}$$

where $\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \mu_K$, and $\sigma_K^2$ are hyper-parameters and assumed to be known. As the starting population size is known, $P(N(1) = N_0) = 1$. We take $N(1) = N_0$ as the starting population size and assumed to be a fixed quantity. The joint posterior distribution of $r_m, \theta, K, \sigma_e^2$ can be written as,

$$f(r_m, \theta, K, \sigma_e^2|N) = \frac{f(N, r_m, \theta, K, \sigma_e^2)}{f_N(N)},$$

where $N = (N(1), N(2), \cdots, N(n+1))'$ represents the data and $f_N(N)$ is the marginal distribution of $N$.

$$f\left(r_m, \theta, K, \sigma_e^2|N\right) = \frac{f\left(N|r_m, \theta, K, \sigma_e^2\right) \cdot f\left(r_m, \theta, K, \sigma_e^2\right)}{f(N)}. \tag{4.15}$$

We assume that $r_m, \theta, K, \sigma_e^2$ are independent variables, therefore $f(r_m, \theta, K, \sigma_e^2)$ can be written as,

$$f\left(r_m, \theta, K, \sigma_e^2\right) = f(r_m) \cdot f(\theta) \cdot f(K) \cdot f\left(\sigma_e^2\right). \tag{4.16}$$

So, equation (4.15) becomes,

$$\begin{aligned}
f\left(r_m, \theta, K, \sigma_e^2|N\right) &\propto f\left(N|r_m, \theta, K, \sigma_e^2\right) \cdot f\left(r_m, \theta, K, \sigma_e^2\right) \\
&\propto f(N|r_m, \theta, K, \sigma_e^2) \cdot f(r_m) \cdot f(\theta) \cdot f(K) \cdot f(\sigma_e^2) \quad \text{[independence of priors]} \\
&\propto \left[\prod_{i=1}^{n} f\left(N(i+1)|N(i)\right)\right] \cdot f(r_m) \cdot f(\theta) \cdot f(K) \cdot f(\sigma_e^2).
\end{aligned}$$

The right hand side of the above expression is product of the likelihood and the prior, which is nothing but unnormalized joint posterior distribution of parameters. Since, the distribution does not follow some common known distribution, we use the Gibbs sampling method that simulates samples from the conditional distribution. So we generate random sample from conditional posterior distribution of each parameters. However, it is observed that in this case also the conditional posterior does not follow any known distribution. So, we use grid approximation algorithm to approximate the posterior probability distribution.

Let $\beta = (r_m, \theta, K, \sigma_e^2)$. Since, $N(i)$'s are not independent, so the likelihood can be written as,

$$\begin{aligned}
\text{likelihood} &= f\left(N(1), N(2), \cdots, N(n+1); \beta\right) \\
&= f\left(N(n+1)|N(n); \beta\right) \cdot f\left(N(n)|N(n-1); \beta\right) \cdots f\left(N(2)|N(1); \beta\right) \cdot f\left(N(1); \beta\right) \\
&= \prod_{i=1}^{n} f\left(N(i+1)|N(i); \beta\right) \cdot f\left(N(1); \beta\right).
\end{aligned}$$

We assume that the initial population size to be known so that $f\left(N(1);\beta\right)=1$.

$$\begin{aligned}
\text{likelihood} &= \prod_{i=1}^{n} f\left(N(i+1)|N(i);\beta\right) \\
&= \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma_e^2 N(i)} e^{-\frac{\left\{N(i+1)-N(i)-r_m N(i)\left[1-\left(\frac{N(i)}{K}\right)^{\theta}\right]\right\}^2}{2\sigma_e^2 N(i)^2}}\right] \\
&= \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma_e^2 N(i)} e^{-\frac{1}{2\sigma_e^2}\left\{\frac{N(i+1)-N(i)}{N(i)}-r_m\left[1-\left(\frac{N(i)}{K}\right)^{\theta}\right]\right\}^{\theta}}\right] \\
&= \frac{1}{(\sqrt{2\pi})^n}\frac{1}{(\sigma_e)^n}\frac{1}{\prod_{i=1}^{n} N(i)} e^{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left\{R(i)-r_m\left[1-\left(\frac{N(i)}{K}\right)^{\theta}\right]\right\}^2}.
\end{aligned}$$

The log-likelihood can be given as,

$$\text{log-likelihood} = -\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma_e^2)-\sum_{i=1}^{n}\log(N(i))-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left\{R(i)-r_m\left[1-\left(\frac{N(i)}{K}\right)^{\theta}\right]\right\}^2. \tag{4.17}$$

To compute the conditional posterior following process has been carried out,

$$\begin{aligned}
f(r_m|\theta,K,\sigma_e^2,N) &= \frac{f(N,r_m,\theta,K,\sigma_e^2)}{f(\theta,K,\sigma_e^2,N)} \\
&= \frac{f(N|r_m,\theta,K,\sigma_e^2)\cdot f(r_m,\theta,K,\sigma_e^2)}{f(N,\theta,K,\sigma_e^2)} \\
&= \frac{f(N|r_m,\theta,K,\sigma_e^2)\cdot \pi(r_m)\pi(\theta)\pi(K)\pi(\sigma_e^2)}{f(N|\theta,K,\sigma_e^2)\cdot f(\theta,K,\sigma_e^2)} \\
&= \frac{f(N|r_m,\theta,K,\sigma_e^2)\cdot \pi(r_m)\cdot\pi(\theta)\cdot\pi(K)\cdot\pi(\sigma_e^2)}{f(N|\theta,K,\sigma_e^2)\cdot\pi(\theta)\cdot\pi(K)\cdot\pi(\sigma_e^2)} \\
&= \frac{f(N|r_m,\theta,K,\sigma_e^2)\cdot\pi(r_m)}{f(N|\theta,K,\sigma_e^2)}.
\end{aligned}$$

Let $m=f(N|\theta,K,\sigma_e^2)$, be the joint marginal probability density function of the data.

$$\begin{aligned}
f(r_m|\theta,K,\sigma_e^2,N) &\propto f(N|r_m,\theta,K,\sigma_e^2)\cdot\pi(r_m) \\
&\propto \prod_{i=1}^{n} f(N_{i+1}|N_i;r_m,\theta,K,\sigma_e^2)\cdot\pi(r_m).
\end{aligned} \tag{4.18}$$

Similarly,

$$f(\theta|r_m, K, \sigma_e^2, N) \propto \prod_{i=1}^{n} f(N_{i+1}|N_i; r_m, \theta, K, \sigma_e^2) \cdot \pi(\theta) \tag{4.19}$$

$$f(K|r_m, \theta, \sigma_e^2, N) \propto \prod_{i=1}^{n} f(N_{i+1}|N_i; r_m, \theta, K, \sigma_e^2) \cdot \pi(K) \tag{4.20}$$

$$f(\sigma_e^2|r_m, \theta, K, N) \propto \prod_{i=1}^{n} f(N_{i+1}|N_i; r_m, \theta, K, \sigma_e^2) \cdot \pi(\sigma_e^2). \tag{4.21}$$

In Bayesian linear regression we already verify that the variance (specially $\phi$) after observed data follows $\mathcal{IG}\left(\text{shape} = \alpha + \frac{n}{2}, \text{rate} = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_i\right)^2 + \gamma\right)$. Similarly here we expect posterior distribution of variance($\sigma_e^2$) as follows:

$$\sigma_e^2|N \sim \mathcal{IG}\left(\text{shape} = \alpha_2 + \frac{n}{2}, \text{rate} = \frac{1}{2}\sum_{i=1}^{n}\left(y_i - r_m\left\{1 - \left(\frac{N(t)}{K}\right)^{\theta}\right\}\right)^2 + \beta_2\right). \tag{4.22}$$

> **Note**
>
> If the prior distribution $\pi(r_m)$ is considered to be gammma$(\alpha, \beta)$ which takes only positive values, then task of estimation of a declining population may be underestimated.

> **Note**
>
> To reduce the computational time a correct choice of priors is required. In this case, since $K$ is a property of the environment, it should not change drastically. Posterior sample of k should not vary too much away from the prior mean of $K$.

> **Note**
>
> If the prior distribution $\pi(r_m)$ is considered to be gamma$(\alpha, \beta)$ which takes only positive values, then the task of estimation of a declining population may be underestimated.

> **Note**
>
> To reduce the computational time, a correct choice of priors is required. In this case, since $K$ is a property of the environment, it should not change drastically. The posterior sample of $K$ should not vary too much from the prior mean of $K$.

The above structure of the posterior density function of $\sigma_e^2$ given in the equation (4.22) is verified by the simulation for the simulated data.

## 4.2 Simulation study

### 4.2.1 Data Generation

Time series data of population size is generated using $r_m = 0.5$, $\theta = 0.5$, $K = 50$, $\sigma_e^2 = 0.0025$. The initial population size is set as $N_0 = 23$ and the length of the simulated series is $n = 30$. The full Bayesian model is described as follows:

$$
\begin{aligned}
R(t)|N(t) &\sim \mathcal{N}\left(r_m\left[1-\left(\frac{N(t)}{K}\right)^{\theta}\right], \sigma_e^2\right) \\
r_m|\alpha_0, \beta_0 &\sim \mathcal{N}(\alpha_0, \beta_0) \\
K|\mu_K, \sigma_K^2 &\sim \mathcal{N}(\mu_K, \sigma_K^2) \\
\theta|\alpha_1, \beta_1 &\sim \mathcal{G}(\alpha_1, \beta_1) \\
\sigma_e^2|\alpha_2, \beta_2 &\sim \mathcal{IG}(\alpha_2, \beta_2).
\end{aligned}
\tag{4.23}
$$

```
# parameters values for simulating the data
rm = 0.5 # intrinsic growth rate
theta = 0.5 # parameter for strength of density dependence
K = 50 # carrying capacity of the environment
sig_2e = 0.0025 # variance of the environmental perturbation
n = 30 # length of the time series = n+1
N1 = 23 # initial population size

set.seed(471)
N = numeric(n + 1) # storage the simulated time series
N[1] = N1 # initial population size.
for(i in 1:(length(N)-1)){
  r = rnorm(n =1, mean = rm*(1-(N[i]/K)^theta), sd = sqrt(sig_2e)) #   simulating growth r
  N[i+1] = N[i]*exp(r) # simulate next generation
}
#print(N) # printing population size
plot(1:(n+1), N, type= "p", lwd=2, col = "red", pch = 19)

fun_logistic = function(x){ # Deterministic solution of the generalized logistic equation
K/(1+((K/N1)^theta-1 )*exp(-rm*x*theta) )^(1/theta)
```

```
}
curve(fun_logistic, 0, n+1, add= TRUE, lwd=2)
```

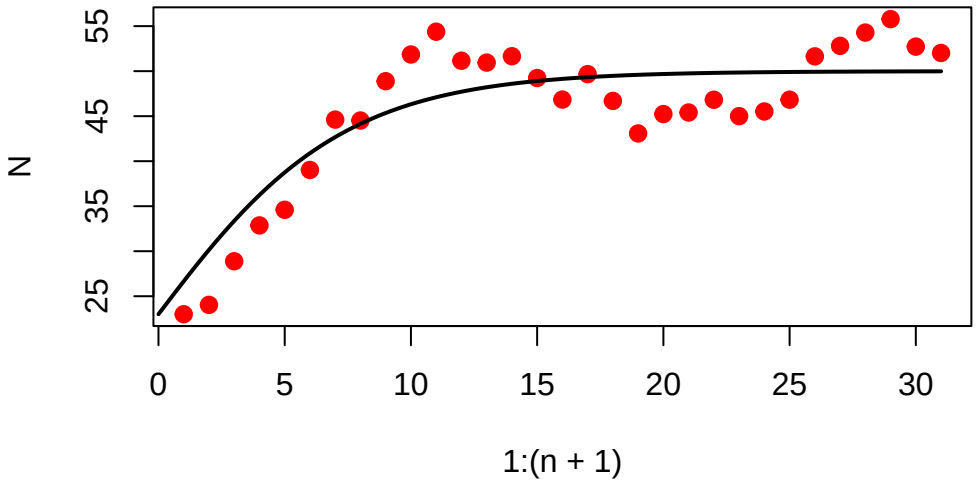



Figure 4.1: Plot of simulated data for the study of Nonlinear regression generated using the model given in the equation (4.23). The parameters values are fixed as $r_m = 0.5, \theta = 0.5, K = 50, \sigma_e^2 = 0.0025$, and initial population size taken as $N_0 = 23$, generated data of length $n = 30$.

### 4.2.2 Define Prior distributions

For generated time series, we want to generate values from posterior of each parameter$(r_m, K, \theta, \sigma_e^2)$. For simulation purpose, we first define the prior functions for the prior distribution given in the equation (4.23) by considering the values of hyper-parameters as $\beta_0 = 1, \sigma_K^2 = 3, \alpha_1 = 1, \beta_1 = 1, \alpha_2 = 1, \beta_2 = 0.0001$. Prior mean of $K$ and $r_m$ that is $\mu_K$ and $\alpha_0$ are obtained by fitting the logistic growth equation using nonlinear least squares method (`nls`, in `R`). The package `invgamma` (Kahle and Stamey 2017) has been used for generating random numbers from the inverted gamma distribution.

```
prior_mean_rm = coef(fit)[1]
prior_sd_rm = 1
fun_prior_rm = function(x){ # Prior specification for r_m
```

```
  dnorm(x, prior_mean_rm, prior_sd_rm)
}

prior_mean_K = coef(fit)[2]
prior_sd_K = 3
fun_prior_K = function(x){ # Prior specification for K
  dnorm(x, prior_mean_K, prior_sd_K)
}

alpha_1 = 1
beta_1 = 1
fun_prior_theta = function(x){ # Prior specification for theta
  dgamma(x, shape = alpha_1, rate = beta_1)
}

library(invgamma)
alpha_2 = 1
beta_2 = 0.0001
fun_prior_sig_2e = function(x){ # Prior specification for inverted gama
  dinvgamma(x, shape = alpha_2, rate = beta_2)
}
```

As mentioned in the Section 4.1, the exact form of the conditional posterior distribution is not known, so grid approximation has been utilized to approximate the conditional posterior density function. In the following code, we have created grid within an interval specified for each parameter. This initial choice is important as this act as a support for the density function. Finer the grid is more accurate is the posterior approximation.

```
# Specification of grid values for approximating the posterior
step = 0.01
grid_rm = seq(from = -3, to =3, by = step)
grid_K = seq(from = 40, to = 60, by = step)
grid_theta = seq(from = 0, to =10, by = step)
grid_sig_2e = seq(from = 0.0001 , to =0.01, by = 0.00001)
```

### 4.2.3 Likelihood function

In the following code we define the likelihood function. Note that we have directly computed the log-likelihood values to avoid truncation error.

```
likelihood = function(data, param){
  rm = param[1]
  K = param[2]
  theta = param[3]
  sig_2e = param[4]

  x = data[,1] # population size
  y = data[,2] # per capita growth rates

  log_lik = -n*log(sqrt(2*pi)) - (n/2)*log(sig_2e) - sum(((y-rm*(1-(x/K)^theta))^2/(2*sig_
  return(exp(log_lik))
}
```

### 4.2.4 Calculation of Posterior distribution

First of all we have set 100000 as a number Markov Chain samples to be generated. Then we have fixed the initial values of the parameters and defined the arrays for storing the posterior values of the parameters in R as follows:

```
iter = 100000 # Length of the chain to be simulated
post_rm = rep(NA, iter) # posterior values of r_m
post_K = rep(NA, iter) # posterior values of K
post_theta = rep(NA, iter) # posterior values of theta
post_sig_2e = rep(NA, iter) # posterior values of simga_2e

param = c(0.2, 40, 1, 0.02) # starting of the chain

# initialization of the chain
post_rm[1] = param[1]
post_K[1] = param[2]
post_theta[1] = param[3]
post_sig_2e[1] = param[4]
```

Next we carried out the Gibbs sampling and grid approximation of the posterior distribution. At each Gibbs sampling step (outer loop), the conditional posterior has been approximated by grid approximation (four inner loop for four parameters).

```
for(i in 2:iter){ # loop for iteration of Gibbs sampling
  # section for rm
  tmp_post_rm = numeric(length = length(grid_rm))
```

```
  for(j in 1:length(grid_rm)){
      param = c(grid_rm[j], post_K[i-1], post_theta[i-1], post_sig_2e[i-1])
      tmp_post_rm[j] = likelihood(data = data, param = param)*fun_prior_rm(grid_rm[j])
    }
  prob = tmp_post_rm/sum(tmp_post_rm)
  post_rm[i] = sample(grid_rm, size = 1, prob = prob)

  # section for K
  tmp_post_K = numeric(length = length(grid_K))
  for(j in 1:length(grid_K)){
    param = c(post_rm[i], grid_K[j], post_theta[i-1], post_sig_2e[i-1])
    tmp_post_K[j] = likelihood(data = data, param = param)*fun_prior_K(grid_K[j])
    }
  prob = tmp_post_K/sum(tmp_post_K)
  post_K[i] = sample(grid_K, size = 1, prob = prob)

  # section for theta
  tmp_post_theta = numeric(length = length(grid_theta))
  for(j in 1:length(grid_theta)){
    param = c(post_rm[i], post_K[i], grid_theta[j], post_sig_2e[i-1])
    tmp_post_theta[j] = likelihood(data = data, param = param)*fun_prior_theta(grid_theta[
    }
  prob = tmp_post_theta/sum(tmp_post_theta)

  post_theta[i] = sample(grid_theta, size = 1, prob = prob)

  # section for sig_2e
  tmp_post_sig_2e = numeric(length = length(grid_sig_2e))
  for(j in 1:length(grid_sig_2e)){
    param = c(post_rm[i], post_K[i], post_theta[i], grid_sig_2e[j])
    tmp_post_sig_2e[j] = likelihood(data = data, param = param)*fun_prior_sig_2e(grid_sig_
    }
  prob = tmp_post_sig_2e/sum(tmp_post_sig_2e)
  post_sig_2e[i] = sample(grid_sig_2e, size = 1, prob = prob)

} # loop ends for Gibbs iteration
```

By executing the above code, we have obtained the 1,00,000 sample values from the posterior density of $r_m, K, \theta$, and $\sigma_e^2$. We stored them in the local directory on system for the further analysis.

```
post_param_vals = data.frame(post_rm,post_K,post_theta,post_sig_2e)
setwd("specify path here")
filename = paste0("output",seed, ".txt")
write.table(post_param_vals, file = filename, col.names = TRUE)
```

Trace plot is a key diagnostic tool used in Bayesian statistics to assess the behavior and convergence of MCMC chains. So we have plotted the traceplot of posterior values obtained using Grid and Gibbs approximation using following code:

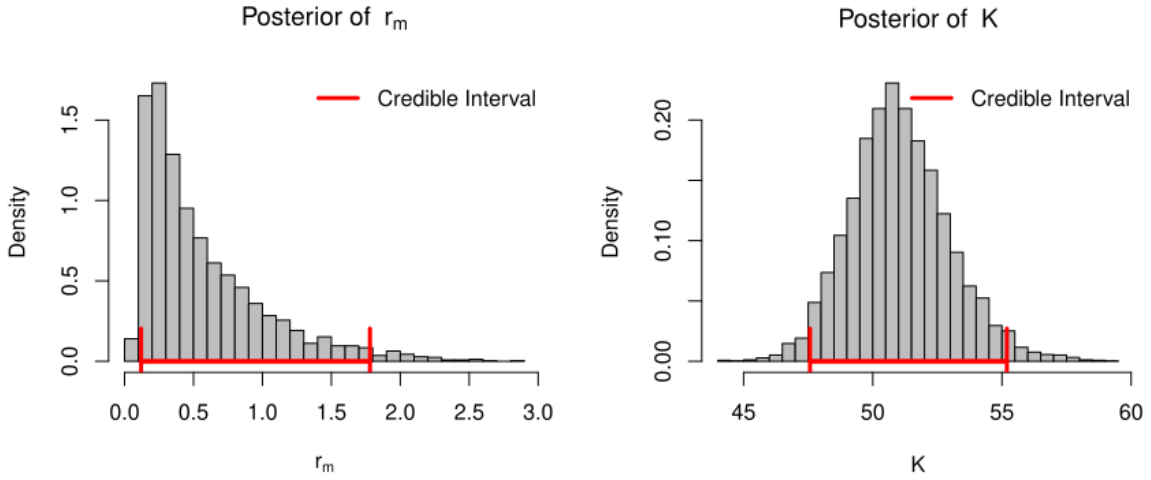
```
filename = paste0("output",seed, ".txt")
post_vals = read.table(file = filename, header = TRUE)

#traceplot
par(mfrow=c(2,2))
plot(post_vals$post_rm, type = "l", main = bquote("Trace plot of"~ r[m]),col = "red")
plot(post_vals$post_K, type = "l", main = bquote("Trace plot of"~ K),col = "red")
plot(post_vals$post_theta, type = "l", main = bquote("Trace plot of"~ theta),col = "red")
plot(post_vals$post_sig_2e, type = "l", main = bquote("Trace plot of"~ sigma[e]^2),col = "
```

We have kept initial half i.e 50,000 sample values as a burn-in period for each parameter. After the burn-in period from remaining 50,000 sample posterior values, using appropriate step for thinning (using auto-correlation function, $\texttt{acf}$) we have reduced the autocorrelation between the successive samples to get reliable estimate of parameters. Next we have plot the histogram of posterior distribution of all parameters using following code and depicted in the Figure 4.2.

```
cut = 0.5 # burn-in chain (first 50,000 values as burn-in period)

# for rm
u = post_vals$post_rm[ceiling(iter*cut):iter] # after burn in period
index = seq(1,length(u), by = 20) # thining
post_rm_thinning = u[index] # values after thinning
acf(post_rm_thinning,main = bquote("acf of posterior "~rm)) # autocorrelation
post_interval_rm = quantile(post_rm_thinning, c(2.5, 97.5)/100) #credible interval for rm
hist(post_rm_thinning, probability = TRUE,main = bquote("Posterior of "~ r[m]),
xlab = expression(r[m]),col = "grey", breaks = 40)
arrows(post_interval_rm[1], 0, post_interval_rm[2], 0, lwd=3, angle = 90,col = "red", code
arrows(post_interval_rm[2], 0, post_interval_rm[1], 0, lwd=3, angle = 90,col = "red", code
legend("topright",legend = c("Credible Interval"),lwd = 3, col = "red",lty = 1,cex = 1, bt

# for K
```

```
u = post_vals$post_K[ceiling(iter*cut):iter] # after burn in period
index = seq(1,length(u), by = 10) # thining
post_K_thinning = u[index] # values after thinning
acf(post_K_thinning,main = bquote("acf of posterior "~K)) # autocorrelation
post_interval_K = quantile(post_K_thinning, c(2.5, 97.5)/100) #credible interval for K
hist(post_K_thinning, probability = TRUE,main = bquote("Posterior of "~ K),
xlab = expression(K),col = "grey", breaks = 40)
arrows(post_interval_K[1], 0, post_interval_K[2], 0, lwd=3, angle = 90, col = "red", code
arrows(post_interval_K[2], 0, post_interval_K[1], 0, lwd=3, angle = 90, col = "red", code
legend("topright",legend = c("Credible Interval"),lwd = 3, col = "red",lty = 1,cex = 1, bt

# for theta
u = post_vals$post_theta[ceiling(iter*cut):iter] # after burn in period
index = seq(1,length(u), by = 20) # thining
post_theta_thinning = u[index] # values after thinning
acf(post_theta_thinning,main = bquote("acf of posterior "~theta)) # autocorrelation
post_interval_theta = quantile(post_theta_thinning, c(2.5, 97.5)/100) #credible interval f
hist(post_theta_thinning, probability = TRUE,main = bquote("Posterior of "~ theta),xlab =
arrows(post_interval_theta[1], 0, post_interval_theta[2], 0, lwd=3, angle = 90, col = "red
arrows(post_interval_theta[2], 0, post_interval_theta[1], 0, lwd=3, angle = 90, col = "red
legend("topright",legend = c("Credible Interval"),lwd = 3, col = "red",lty = 1,cex = 1, bt

# for sigma_e^2
u = post_vals$post_sig_2e[ceiling(iter*cut):iter] # after burn in period
index = seq(1,length(u), by = 1) # thining
post_sig_2e_thinning = u[index] # values after thinning
acf(post_sig_2e_thinning,main = bquote("acf of posterior "~sigma[e]^2))
post_interval_sig_2e = quantile(post_sig_2e_thinning, c(2.5, 97.5)/100)
#credible interval for sig_2e
hist(post_sig_2e_thinning, probability = TRUE,main = bquote("Posterior of "~ sigma[e]^2),x
arrows(post_interval_sig_2e[1], 0, post_interval_sig_2e[2], 0, lwd=3, angle = 90, col = "r
arrows(post_interval_sig_2e[2], 0, post_interval_sig_2e[1], 0, lwd=3, angle = 90, col = "r
legend("topright",legend = c("Credible Interval"),lwd = 3, col = "red",lty = 1,cex = 1, bt
```

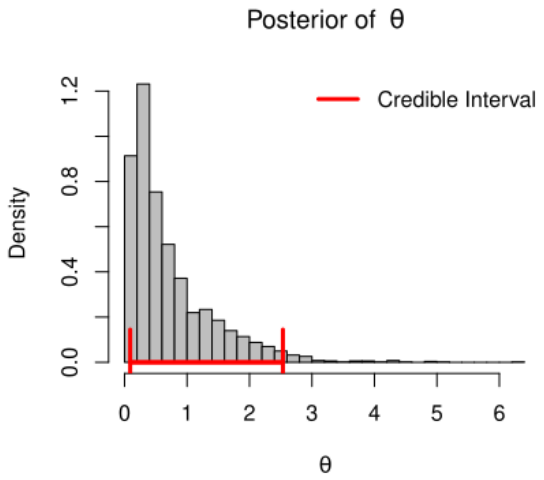
For $\sigma_e^2$, due to conjugacy the posterior density is predictable, but good approximation would depend on the estimate of the shape parameter which involves other parameters. Here we show what choice of the point estimate for $r_m, K$ and $\theta$ need to be used (Figure 4.2d) and we observe that the posterior median values best approximate the posterior density function of $\sigma_e^2$.



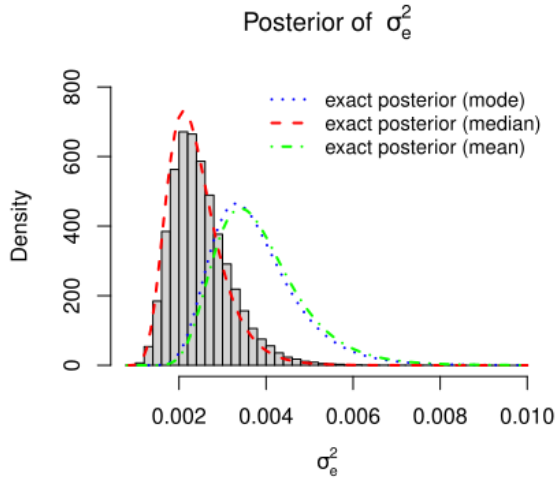

(a)




(b)




(c)




(d)

Figure 4.2: Posterior distribution of the parameters of theta-logistic growth equation. From the histogram of $\sigma_e^2$, it is evident that posterior median is more appropriate choice to obtain a point estimate of $r_m$, $K$ and $\theta$. Blue lines indicate the 95% of credible intervals.

# 5 Bayesian connection to Statistical Regularization

First we outline the concepts of linear regression. We denote the response variable by $Y$ and the set of predictor variables by $X_1, X_2, \cdots, X_p$, $p$ being the number of predictors which are also synonymously written as explanatory variables, independent variables, covariates, regressors etc. The true relationship between the response and the predictors can be approximated by a regression function $f$, so that the equation

$$Y = f(X_1, X_2, \cdots, X_p) + \epsilon \tag{5.1}$$

is a valid statistical model for the population of interest. Usually, $f$ is a fixed but unknown function of $X_1, X_2, \cdots, X_p$ which represents the systematic variation of $Y$ explained by the predictors. The unexplained component, denoted by $\epsilon$, is assumed to be a random error (independent of the predictors) with mean zero. This gives a measure of discrepancy of the approximation by the function $f$.

Under the multiple regression set up, $f$ is replaced by a linear function of the predictors, represented as

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon,$$

where $\beta_0$ is the intercept term and $\beta_j$ gives the contributions of $X_j$ for $j = 1, \cdots, n$ in explaining the variation of $Y$. For convenience, we adopt the matrix notation and represent the data set up. We have the continuous response $\mathbf{Y} = (y_1, \cdots, y_n)^{'} \in \mathbb{R}^n$ and the $n$ data values $\left(x_{ij}\right)_{i=1}^{n}$ are available on each $X_j$, $j = 1, 2, \cdots, p$. The regression equation in matrix form written as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{e} \tag{5.2}$$

where $\mathbf{X}$ is an $n \times (p+1)$ design matrix with entries in the first column being 1 (if intercept included) otherwise it is $n \times p$ order matrix with $x_{ij}$ denotes the $i$th observation corresponding to the $j$th variable. $\beta$ be the vector of regression coefficients of order $(p+1) \times 1$ and $\mathbf{e} = (e_1, e_2, \cdots, e_n)' \in \mathbb{R}^n$ is the vector of errors. By using the least squares theory, the residual sum of squares,

$$\mathbf{e}'\mathbf{e} = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = \mathbf{RSS}(\beta),$$

is minimized with respect to $\beta$. Solving the normal equations, the estimates of $\beta$ is obtained as $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Using the notion of $l_2 -$ norm, one can write as

$$\widehat{\beta} = \text{argmin}_{\beta} \left( \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 \right), \tag{5.3}$$

where $\|\mathbf{u}\|_2^2 = \sum_{i=1}^{n} u_i^2$ for a vector $\mathbf{u} \in \mathbb{R}^n$. $\widehat{\beta}$ is an unbiased and consistent estimator of $\beta$. The normal equations for the above minimization problem is given by

$$(\mathbf{X}'\mathbf{X})\beta = \mathbf{X}'\mathbf{Y}.$$

If $(\mathbf{X}'\mathbf{X})$ has determinant zero, then the unique solution for the system of equations can not be obtained. If the matrix is ill-conditioned that is the determinant is very small, then variance for estimated coefficients are very large making the estimates unreliable. This is a common case in many real life data sets where high degree of correlation exits between two variables. If any particular predictor can be closely approximated by a linear combination of two or more other predictors, then also the matrix $(\mathbf{X}'\mathbf{X})$ is nearly singular. Such a situation is called multicollinearity and must be taken care of before any modeling assignment.

To tackle the multicollinearity problem, various shrinkage methods have been proposed in the literature. Ridge regression deals with solving the normal equations of the form

$$(\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})\,\beta = \mathbf{X}'\mathbf{Y},$$

where $\lambda \geq 0$ is called the shrinkage parameter. Because of the additional parameter $\lambda$ the coefficient estimates $\widehat{\beta}$ has lower variance but they are no longer unbiased. Basically, instead of minimizing the residual sum of squares, ridge regression minimizes a slightly different quantity, given by $\mathbf{RSS}(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2$. The coefficient estimates can be written using $l_2 -$ norm as

$$\widehat{\beta_{\lambda}^{R}} = \text{argmin}_{\beta} \left( \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_2^2 \right). \tag{5.4}$$

The term $\lambda \sum_{j=1}^{p} \beta_j^2$, referred as shrinkage penalty, is small when $\beta_1, \cdots, \beta_p$ are close to zero. For $\lambda = 0$, the procedure is equivalent to the ordinary least squares (OLS) regression. For $\lambda \to \infty$, the influence of shrinkage penalty increases and the components of $\widehat{\beta_{\lambda}^{R}}$ will approach zero and its successful implementation is guaranteed by the bias-variance tradeoff.

Model selection methods such as forward, backward and mixed selection give the model that involves a subset of all the predictors. These techniques can be conveniently performed using

R software for statistical computing. The packages `ISLR` (James et al. 2021), `leaps` (Fortran code by Alan Miller 2024), `MASS` (Venables and Ripley 2002) can be used for this purpose and different criteria such as $R^2$, AIC etc. can be utilized for the model selection. Unlike these methods, ridge regression includes all $p$ predictors in the model; the penalty term reduces many coefficients to very small values, but will not set them exactly equal to zero. This is challenging for interpretation of the results in high dimension when the number of predictors is very large. This shortcoming is overcome by the method of lasso regression by minimizing the quantity $\mathbf{RSS}(\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$, which considers an $l_1-$norm penalty instead of $l_2-$norm penalty. Thus using $l_1 -$ norm notation, the coefficient estimates can be written as

$$\hat{\beta}_{\lambda}^{L} = \text{argmin}_{\beta} \left( \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \right) \tag{5.5}$$

where $\|\mathbf{u}\|_1 = \sum_{i=1}^{n} |u_i|$ for a vector $\mathbf{u} \in \mathbb{R}^n$. Because of the $l_1$ penalty some of the coefficients are forced to be equal to zero for large $\lambda$. Thus, like the best subset selection methods, lasso also provides variable selection method. More precisely the lasso regression falls between best subset selection method and the ridge regression method and has some nice statistical properties from both techniques.

Ridge regression has a close connection to Bayesian linear regression. Bayesian linear regression assumes that the parameters $\beta$ and $\sigma^2$ (known) to be the random variables, while at the same time considering $\mathbf{X}$ and $\mathbf{Y}$ as fixed. We shall show that if we assume the the prior distribution for $\beta$ as follows: $\beta_1, \beta_2, \cdots, \beta_p$ are independent and identically distributed according to a normal distribution of mean zero and variance $c$, then the posterior distribution of $\beta$ is also normal and the ridge regression estimate is both the mean and the mode for $\beta$ under this posterior distribution. We start with the linear regression setting again so that, the exact calculations follows naturally:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \epsilon_i, \quad i = 1, 2, \cdots, n. \tag{5.6}$$

$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, 2, \cdots, n$, $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)'$, $\mathrm{E}_{\beta}\hat{\beta} = \beta$, $\mathrm{Var}_{\beta}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, where $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, $\mathbf{Y} = (y_1, y_2, \cdots, y_n)'$. In matrix notation, we write it as

$$\mathbf{Y} = \beta_0 \mathbf{1}_n + \mathbf{X}\beta + \mathbf{e},$$

where $\mathbf{1}_n$ is a vector of order $n \times 1$ whose each entry is equal to 1. Note that we do not have any prior distribution on the intercept term $\beta_0$. Since $\beta_j$'s ($j = 1, 2, \cdots, p$) are independent and identically distributed normal random variables with mean 0 and variance $c$, the the joint probability density function can be easily written as a multivariate normal distribution as follows:

$$\pi(\beta) = \left(\frac{1}{\sqrt{2\pi c}}\right)^p e^{-\frac{\beta'\beta}{2c}}, \quad \beta \in \mathbb{R}^p, \tag{5.7}$$

and the marginal prior densities are given by $\pi(\beta_j) = \frac{1}{\sqrt{2\pi c}} e^{-\frac{\beta_j^2}{2c}}$, $j = 1, 2, \cdots, n$. Using a Hierarchical Bayesian set up the regression model can be written as

$$\begin{aligned} Y_i|\beta &\sim \mathcal{N}\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}, \sigma^2\right), i = 1, 2, \cdots, n, \text{ independent,} \\ \beta &\sim \pi(\beta). \end{aligned}$$

The posterior distribution of $\beta$ is given by,

$$\pi(\beta|y) = \frac{f(\beta, y)}{f_{\mathbf{Y}}(y)} = \frac{f(y|\beta)\pi(\beta)}{f_{\mathbf{Y}}(y)},$$

The marginal distribution of $\mathbf{Y}$ is given by,

$$\begin{aligned} f_{\mathbf{Y}}(y) &= \int_{\mathbb{R}^p} f(y|\beta)\pi(\beta)\mathrm{d}\beta \qquad (5.8) \\ &= \int_{\mathbb{R}^p} \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2}{2\sigma^2}} \frac{1}{(\sqrt{2\pi c})^p} e^{-\frac{\sum_{j=1}^{p}\beta_j^2}{2c}} \mathrm{d}\beta \\ &= \frac{1}{(\sqrt{2\pi}\sigma)^n} \frac{1}{(\sqrt{2\pi c})^p} \int_{\mathbb{R}^p} e^{-\frac{1}{2}\left[\frac{(y-\beta_0\mathbf{1}_n - \mathbf{X}\beta)'(y-\beta_0\mathbf{1}_n-\mathbf{X}\beta)}{\sigma^2} + \frac{\beta'\beta}{c}\right]} \mathrm{d}\beta \qquad (5.9) \end{aligned}$$

Now consider the exponent part in the previous equation and simplify it,

$$\begin{aligned} 2 \times \text{Exponent} &= \frac{(\mathbf{Y} - \beta_0\mathbf{1}_n - \mathbf{X}\beta)'(\mathbf{Y} - \beta_0\mathbf{1}_n - \mathbf{X}\beta)}{\sigma^2} + \frac{\beta'\beta}{c} \qquad (5.10) \\ &= \frac{(\mathbf{Y}' - \beta_0\mathbf{1}_n' - \beta'\mathbf{X}')(\mathbf{Y} - \beta_0\mathbf{1}_n - \mathbf{X}\beta)}{\sigma^2} + \frac{\beta'\beta}{c} \\ &= \frac{\mathbf{Y}'\mathbf{Y} - 2\beta_0\mathbf{1}_n'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta_0^2 + 2\beta'\mathbf{X}'\beta_0\mathbf{1}_n + \beta'\mathbf{X}'\mathbf{X}\beta}{\sigma^2} + \frac{\beta'\beta}{c} \\ &= \frac{\mathbf{Y}'\mathbf{Y}}{\sigma^2} - \frac{2\beta_0\mathbf{1}_n'\mathbf{Y}}{\sigma^2} + \frac{\beta_0^2}{\sigma^2} + \beta'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2} + \frac{1}{c}\mathbf{I}_p\right)\beta - 2\beta'\mathbf{X}'\left(\frac{\mathbf{Y}}{\sigma^2} - \frac{\beta_0\mathbf{1}_n}{\sigma^2}\right) \qquad (5.11) \end{aligned}$$

We write the above expression in the form $(\beta-\mu)'\Sigma^{-1}(\beta-\mu)+$ constant containing term $\mathbf{Y}$ and $\mathbf{X}'s$. To do that we utilize the following expression for the purpose of matching similar terms:

$$(\mathbf{X}-\mu)'\Sigma^{-1}(\mathbf{X}-\mu)=\mathbf{X}'\Sigma^{-1}\mathbf{X}-2\mathbf{X}'\Sigma^{-1}\mu+\mu'\Sigma^{-1}\mu.$$

Therefore by equation (5.8)

$$\begin{aligned}
&\frac{(\mathbf{Y}-\beta_0\mathbf{1}_n-\mathbf{X}\beta)'(\mathbf{Y}-\beta_0\mathbf{1}_n-\mathbf{X}\beta)}{\sigma^2}+\frac{\beta'\beta}{c} \qquad (5.12)\\
=\;&\beta'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{1}{c}\mathbf{I}_p\right)\beta-2\beta'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{1}{c}\mathbf{I}_p\right)\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{1}{c}\mathbf{I}_p\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{Y}}{\sigma^2}-\frac{\mathbf{X}'\beta_0\mathbf{1}_n}{\sigma^2}\right)+\\
&\left[\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{1}{c}\mathbf{I}_p\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{Y}}{\sigma^2}-\frac{\mathbf{X}'\beta_0\mathbf{1}_n}{\sigma^2}\right)\right]'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{1}{c}\mathbf{I}_p\right)\\
&\left[\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{1}{c}\mathbf{I}_p\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{Y}}{\sigma^2}-\frac{\mathbf{X}'\beta_0\mathbf{1}_n}{\sigma^2}\right)\right]\\
&-\left[\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{1}{c}\mathbf{I}_p\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{Y}}{\sigma^2}-\frac{\mathbf{X}'\beta_0\mathbf{1}_n}{\sigma^2}\right)\right]'\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{1}{c}\mathbf{I}_p\right)\\
&\left[\left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2}+\frac{1}{c}\mathbf{I}_p\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{Y}}{\sigma^2}-\frac{\mathbf{X}'\beta_0\mathbf{1}_n}{\sigma^2}\right)\right]\\
&+\frac{\mathbf{Y}'\mathbf{Y}}{\sigma^2}-\frac{2\beta_0\mathbf{1}_n'\mathbf{Y}}{\sigma^2}+\frac{\beta_0^2}{\sigma^2}. \qquad (5.13)
\end{aligned}$$

Now, let

$$\Sigma^{-1}=\sigma^2\left(\mathbf{X}'\mathbf{X}+\frac{\sigma^2}{c}\mathbf{I}_p\right)\quad\text{and}\quad\mu=\Sigma\mathbf{X}'\left(\mathbf{Y}-\beta_0\mathbf{1}_n\right),$$

then the Exponent becomes,

$$\begin{aligned}
2\times\text{Exponent}&=\beta'\Sigma^{-1}\beta-2\beta'\Sigma^{-1}\mu+\mu'\Sigma^{-1}\mu-\mu'\Sigma^{-1}\mu+\frac{\mathbf{Y}'\mathbf{Y}}{\sigma^2}-\frac{2\beta_0\mathbf{1}_n'\mathbf{Y}}{\sigma^2}+\frac{\beta_0^2}{\sigma^2}\\
&=(\beta-\mu)'\Sigma^{-1}(\beta-\mu)-\mu'\Sigma^{-1}\mu+\frac{\mathbf{Y}'\mathbf{Y}}{\sigma^2}-\frac{2\beta_0\mathbf{1}_n'\mathbf{Y}}{\sigma^2}+\frac{\beta_0^2}{\sigma^2}. \qquad (5.14)
\end{aligned}$$

Then equation (5.8) becomes,

$$\begin{aligned}
f_{\mathbf{Y}}(y)&=\left(\sqrt{2\pi}\right)^p|\Sigma|^{p/2}\frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}\frac{1}{\left(\sqrt{2\pi c}\right)^p}e^{-\frac{\mathbf{Y}'\mathbf{Y}}{2\sigma^2}}e^{\frac{\beta_0\mathbf{1}_n'\mathbf{Y}}{2\sigma^2}}e^{\frac{\beta_0^2}{2\sigma^2}}e^{\frac{\mu'\Sigma^{-1}\mu}{2}}\int_{\mathbb{R}^p}\frac{1}{(\sqrt{2\pi})^p}\frac{1}{|\Sigma|^{p/2}}e^{-\frac{1}{2}(\beta-\mu)'\Sigma^{-1}(\beta-\mu)}\mathrm{d}\beta\\
&=\left(\sqrt{2\pi}\right)^p|\Sigma|^{p/2}\frac{1}{\left(\sqrt{2\pi}\sigma\right)^n}\frac{1}{\left(\sqrt{2\pi c}\right)^p}e^{-\frac{\mathbf{Y}'\mathbf{Y}}{2\sigma^2}}e^{\frac{\beta_0\mathbf{1}_n'\mathbf{Y}}{2\sigma^2}}e^{\frac{\beta_0^2}{2\sigma^2}}e^{\frac{\mu'\Sigma^{-1}\mu}{2}}. \qquad (5.15)
\end{aligned}$$

Note that the integrand is the density function of $\text{MVN}_p(\mu, \Sigma)$, hence integral out to be 1, where,

$$\mu = \left(\frac{\mathbf{X}'\mathbf{X}}{\sigma^2} + \frac{1}{c}\mathbf{I}_p\right)^{-1} \mathbf{X}' \left(\frac{\mathbf{Y} - \beta_0 \mathbf{1}_n}{\sigma^2}\right) \quad \text{and} \quad \Sigma = \sigma^2 \left(\mathbf{X}'\mathbf{X} + \frac{\sigma^2}{c}\mathbf{I}_p\right)^{-1}.$$

The posterior distribution of $\beta$ is,

$$\begin{aligned}
\phi(\beta|y) &= \frac{f(\beta, y)}{f_{\mathbf{Y}}(y)} \\
&= \frac{f(y|\beta)\pi(\beta)}{f_{\mathbf{Y}}(y)} \\
&= \frac{\frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\frac{(\mathbf{Y}-\beta_0\mathbf{1}_n-\mathbf{X}\beta)'(\mathbf{Y}-\beta_0\mathbf{1}_n-\mathbf{X}\beta)}{\sigma^2} + \frac{\beta'\beta}{c}} \frac{1}{(\sqrt{2\pi c})^p} e^{-\frac{\beta'\beta}{2c}}}{\left(\sqrt{2\pi}\right)^p |\Sigma|^{\frac{p}{2}} \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} \frac{1}{\left(\sqrt{2\pi c}\right)^p} e^{-\frac{\mathbf{Y}'\mathbf{Y}}{2\sigma^2}} e^{\frac{\beta_0 \mathbf{1}_n \mathbf{Y}}{2\sigma^2}} e^{\frac{\beta_0^2}{2\sigma^2}} e^{\frac{\mu'\Sigma^{-1}\mu}{2}}} \\
&= \frac{1}{\left(\sqrt{2\pi}\right)^p} \frac{1}{|\Sigma|^{p/2}} e^{-\frac{1}{2}(\beta-\mu)'\Sigma^{-1}(\beta-\mu)}.
\end{aligned}$$

So the posterior mean of $\beta$ is given by

$$\begin{aligned}
\text{E}(\beta|y) &= \hat{\beta}_B \\
&= \left(\mathbf{X}'\mathbf{X} + \frac{\sigma^2}{c}\mathbf{I}_p\right)^{-1} \mathbf{X}'(\mathbf{Y} - \beta_0 \mathbf{1}_n)
\end{aligned}$$

The main utility of ridge regression is to choose the predictors. We can start with the regression model in the following form:

$$\tilde{\mathbf{y}} = \mathbf{X}\beta + \epsilon$$

where $\tilde{\mathbf{y}} = \mathbf{Y} - \bar{y}\mathbf{1}_n$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. Then

$$\hat{\beta}_B = \left(\mathbf{X}'\mathbf{X} + \frac{\sigma^2}{c}\mathbf{I}_p\right)^{-1} \mathbf{X}'\tilde{\mathbf{y}}.$$

# 6 Conclusion

This document serves as a comprehensive guide for students and educators, aiming to foster a solid foundation in Bayesian computation without relying on pre-built packages. By encouraging learning from scratch, the material ensures that students grasp the underlying principles and mechanics of Bayesian methods, which enhances the teaching experience in a classroom setting.

The comparison of Maximum Likelihood Estimation (MLE) and Bayesian estimators through simulation studies in the second chapter for various examples helps students understand how well an estimator performs by accounting for both the bias (accuracy) and the variance (consistency). The simulations provide practical experience, allowing students to compare the estimators and deepen their understanding of statistical inference beyond just theory.

Through Bayesian regression, students gain a deeper understanding of how to incorporate prior beliefs with data and carry out the inference process. Additionally, the emphasis on ecological modeling in this document allows learners to connect theory with real-world issues. By integrating ecological assumptions with simulation-based learning, students better understand how to apply Bayesian methods to address complex environmental challenges, underscoring the significance of ecological assumptions in model development.

In conclusion, this document is designed to enrich classroom teaching and provide students with a robust foundation in Bayesian methods, ecological problem-solving, and the nuances of simulation studies. The goal is to cultivate a deep, conceptual understanding of these topics while reinforcing practical application through problem-solving exercises.

This document is not in its final format. We will update it regularly to provide a more polished version. Please check for updates and additional chapters.

# References


Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software* 80 (1): 1–28. https://doi.org/10.18637/jss.v080.i01.

Casella, George, and Roger L. Berger. 2002. *Statistical Inference*. Second. Duxbury Advanced Series. India Edition: Cengage Learning.

Fortran code by Alan Miller, Thomas Lumley based on. 2024. *Leaps: Regression Subset Selection*. https://CRAN.R-project.org/package=leaps.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

James, Gareth, Daniela Witten, Trevor Hastie, and Rob Tibshirani. 2021. *ISLR: Data for an Introduction to Statistical Learning with Applications in r*. https://CRAN.R-project.org/package=ISLR.

Kahle, David, and James Stamey. 2017. *Invgamma: The Inverse Gamma Distribution*. https://CRAN.R-project.org/package=invgamma.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wasserman, Larry. 2004. *All of Statistics: A Concise Course in Statistical Inference*. New York: Springer-Verlag New York, Inc.