12-2017

Bayesian Model for Detection of Outliers in Linear Regression with Application to Longitudinal Data

Zahraa Al-Sharea
University of Arkansas, Fayetteville

Bayesian Model for Detection of Outliers in Linear Regression with Application to Longitudinal Data

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Statistics and Analytics

by

Zahraa Ibrahim Jasim Al-Sharea
University of Baghdad
Bachelor of Science in Computers, 2009

December 2017
University of Arkansas

This thesis is approved for recommendation to the Graduate Council.

Dr. Avishek Chakraborty
Thesis Director

Dr. Mark Arnold
Committee Member

Dr. Giovanni Petris
Committee Member

Dr. Qingyang Zhang
Committee Member

Abstract

Outlier detection is one of the most important challenges with many present-day applications. Outliers can occur due to uncertainty in data generating mechanisms or due to an error in data recording/processing. Outliers can drastically change the study's results and make predictions less reliable. Detecting outliers in longitudinal studies is quite challenging because this kind of study is working with observations that change over time. Therefore, the same subject can produce an outlier at one point in time produce regular observations at all other time points. A Bayesian hierarchical modeling assigns parameters that can quantify whether each observation is an outlier or not. The purpose of this thesis is to detect the outlying observations by developing three approaches of techniques and comparing each of them under different data generating mechanisms. In the first chapter, we introduce the important concepts in Bayesian inference with three examples. The first two examples (Binomial and Poisson distributions) are to illustrate the idea behind the Monte Carlo method, while the last example (normal distribution) is to illustrate the Markov Chain Monte Carlo (MCMC). We visited three different types of MCMC Methods: Metropolis-Hastings, Gibbs sampler and Slice sampler which we have used in the three algorithms of outlier detection. In Chapter Two, we used Gibbs sampler techniques with the linear regression model. Simulated data with three covariates were used, and then we applied our method to a real dataset: the Strong Rock data. We explained the findings using diagrams. In Chapter Three, we focused on the core problem of identifying outliers by using three methods. We applied our methods on four simulation datasets. We found that the first two methods did not work well under assumptions of systematic heteroscedasticity but the last one did an efficient job, as we expected, even when the functional form of heteroscedasticity was not correctly specified. Next, we formulated our model for the real data, so we could apply the methods that we developed in chapter three. Given access to the real data that have large numbers of observations, we will apply these methods.

Acknowledgements

I would like to thank Dr.Avishek Chakraborty for his directing and willingness to work with me. I have benefited extremely from his insight, expertise and knowledge of Statistics. I have really appreciated his flexibility and understanding of the commitments associated with parenting my young children.

I appreciate Dr.Mark Arnold, Dr.Giovanni Petris and Dr.Zhang Qingyang for all support, advices, and for all that I have learned from them during my academic study.

Especially, I would like to thank to my husband, Ghadeer Mahdi, I owe special thanks for the countless hours he was willing to help and support me during the times I struggled. His reassurance, enthusiasm and moral support that he has given to me were greatly appreciated. It is from him that I learned the importance of working hard and diligence. I will never forget his advice and guidance. Without him I could not finish my dream. I would like to express my thanks to the greatest source of support, love, and tenderness, my mother. She was more than just a great mother because she was a mother and a father to me after my father died when I was 3 years old. I am deeply appreciative for the encouragement that she has given to me. It is from my mother that I learned the importance of family, faith and education. I am indebted for the countless sacrifices she made for her children, especially to me. Without my mother I could not have gone one step toward my goal. I am extremely thankful to my sister and brother, Israa and Ahmed Al-Sharea. I cannot thank you enough for all the ways you have supported and encouraged me over the years. To my amazing children Mustafa, Rawan and Jawan (Afnan), although it was sometimes difficult to balance between school and being a mother, their love, happiness, and support have helped me to achieve my goals.

To all those not mentioned who have supported and encouraged me, thank you so much.

Table of Contents

Chapter 1

Bayesian Inference

1.1 Hierarchical Model and Bayes's Theorem

Bayesian inference refers to a paradigm that is used for estimation of parameters from a statistical model. This method is based on Bayes's theorem, an important theorem in statistics. In the 18^{th} century, Thomas Bayes, one of the famous mathematicians and theologians, discovered this important theorem which we will discuss in the next section. A hierarchical model connects multiple simple intuitive models in a hierarchy so that the resultant joint probability model can capture complex dependence patterns from correlated data. Gelman et al. (2014) gives an interesting example: "in a study of the effectiveness of cardiac treatments, with the patients in hospital j having survival probability θ_j, it might be reasonable to expect that estimates of the θ_j's, which represent a sample of hospitals, should be related. We shall see that this is achieved in a natural way if we use a prior distribution in which the θ_j's are viewed as a sample from a common population distribution. A key feature of such applications is that the observed data, y_{ij}, with units indexed by i within group indexed by j, can be used to estimate aspect of the population distribution of the θ_j's even though the values of θ_j are not themselves observed". Hierarchical modeling is the best way to model this kind of problem, as it connects the data with the prior distribution based on historical information about the model parameters.

In Bayesian hierarchical modeling, the probability distribution of data contains parameters. The parameters are assigned probability distributions, referred to as priors. There can be parameters inside these prior distributions as well, and they are called hyper-parameters, and the distribution of the hyper-parameter is called the hyper-prior

(Banerjee et al., 2014).

We give an example of hierarchical model using grouped data. Suppose, there are k groups with possibly different number of observations. Let y_{ij} be the j^{th} observation in group i for $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, k$. We model $y_{ij} \sim f(\mu_i)$, where f is a probability distribution with parameter μ_i. Then, $\mu_1, \mu_2, \ldots, \mu_k$ are the parameters and we assigned each of them a prior distribution g such that $\mu_i \sim g(\theta)$ where θ is an hyper-parameter. We could also assign a hyper-prior to θ with a parameter λ whose value is fixed. In Fig. 1.1, we present the hierarchical structure for this model. This example was motivated by a similar setting in Fig.1 of Al-Amin et al. (2014).

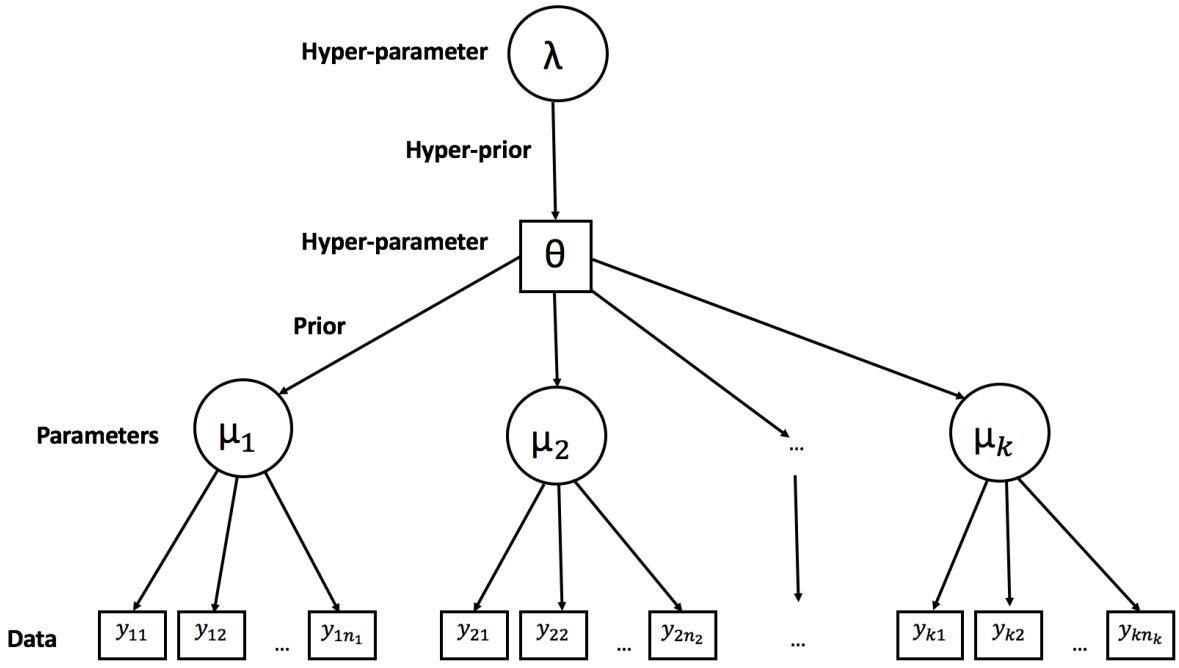

Fig. 1.1: An exmaple of hierarchical model for grouped data

1.1.1 Prior, Likelihood and Posterior Distributions

There are three important parts in Bayesian inference: prior, likelihood, and posterior. A brief explanation of each part is provided in the following.

First of all, the prior distribution for any parameter θ, denoted by $\pi(\theta)$, is one's initial beliefs about a probability distribution before information has been examined. For example, our interest is in the probability distribution of voters in favor of a specific party in an upcoming election. There are several ways to get a prior distribution. For example, using information from past elections can help us to determine an appropriate prior distribution for the proportion of voters willing to support the party. We can also use the supposition of an expert, which will help us to select the prior distribution for the required model. Moreover, an uninformative prior, one which lacks available information, could be determined by equally weighing all possibilities. There are some priors that can be selected by certain qualities such as, the Jeffrey or Bernardo's reference prior (Banerjee et al., 2014).

The likelihood function, the second factor in the Bayesian inference, describes the possibility of different values of a parameter solely from observed data.

The last factor of the Bayesian inference is the posterior distribution, which is denoted by $\pi(\theta|D)$, where D is the data. We can think about the posterior distribution as a modification of the prior distribution after the data are observed using the likelihood as a function of parameters. The posterior distribution is how we make inferences about unknown parameters.

When $\pi(\theta)$ is the prior distribution, and we have independent observations $x_1, x_2, \ldots, x_n$, then $f(D|\theta) = \prod_{i=1}^{n}(f(x_i|\theta))$ is the likelihood function, and $\pi(\theta|D)$ is the posterior distribution. This formula was derived from the classic Bayes Theorem which is

$$\pi(\theta|D, \alpha) = \frac{f(D|\theta)h(\theta|\alpha)}{\int f(D|u)h(u|\alpha)du} \tag{1.1}$$

In summary, the prior distribution is the best way to guess the parameters before observing the data, and the posterior distribution is the one to use after observing the

data. This relationship can be written as follows:

$$\begin{aligned} Posterior &\propto Likelihood \times Prior \\ \pi(\theta|D) &\propto \prod_{i=1}^{n}(f(x_i|\theta)) \times \pi(\theta) \end{aligned}$$

Bayes theorem, or Bayes's rule, is one of the most famous theorems in statistics. It is used to define the probability of an event by using the information about the prior distribution that may be related to the event. In this subsection, we will consider the following example to better understand the Bayesian Hierarchal model by using the Bayes theorem. If we want to determine the background for students in a math class, we need to get information about the student's high school grades, then combine them with their current grade point average (GPA). This will give us an estimate for how well they did in the math class.

However, in many situations, the posterior $\pi(\theta|D)$ may not be a standard distribution. In those cases, it may become difficult to solve the problem analytically or numerically. Therefore, there is a special technique called the Monte Carlo Method which is used to learn the properties of θ by drawing samples from $\pi(\theta|D)$ and then performing empirical computation.

1.2 Monte Carlo Method for Bayesian

If we are uncertain about the inputs in any experiment, then we can simulate the posterior by using the Monte Carlo (MC) method (Gilks et al., 1995). The main objective of the MC method is to compute results based on repeated random sampling and statistical analysis. Monte Carlo Methods are used in high dimensional problems, and can often be used with other methods to complete analyses (Geman and Geman, 1984). The results of the many experiments that use the MC method simulation are not well documented. However, there are many applications that are difficult to integrate because they have complex hierarchical

structures. We will show two examples for posterior distributions; one for a binomial distribution and the other for a Poisson distribution.

1.2.1 Binomial Distribution Example

Let us consider n independent binomial experiments with a different number of trials but the same probability of success. Assume y_1 was the number of success in the first experiment out of n_1 trials, y_2 the number of success in the second experiment out of n_2 trials and so on. For simplicity we can write it as
$y_i \sim Bin(n_i, p),$ where $0 < p < 1,$ and n_i is known. We want to determine the posterior distribution of p. The likelihood function $L(y)$ can be found by multiplying the n binomial functions, so we can define it as follows:

$$L(y) = \prod_{i=1}^{n} f(y_i) = \prod_{i=1}^{n} \binom{n_i}{y_i} p^{y_i}(1-p)^{n_i - y_i} \tag{1.2}$$

We can choose beta to be the prior distribution. So, the prior distribution is
$p \sim Beta(a_0, b_0)$

$$\pi(p) \propto p^{a_0 - 1}(1-p)^{b_0 - 1}$$

We can get the posterior binomial distribution by multiplying the likelihood function and posterior distribution as follows:

$$\begin{aligned}
\pi(p|y) &\propto L(y)\pi(p) \\
&\propto \prod_{i=1}^{n} \binom{n_i}{y_i} p^{y_i}(1-p)^{n_i - y_i} p^{a_0 - 1}(1-p)^{b_0 - 1} \\
&\propto p^{\sum_{i=1}^{n} y_i}(1-p)^{\sum_{i=1}^{n}(n_i - y_i)} p^{a_0 - 1}(1-p)^{b_0 - 1} \\
&\propto p^{\sum_{i=1}^{n} y_i + a_0 - 1}(1-p)^{\sum_{i=1}^{n}(n_i - y_i) + b_0 - 1}
\end{aligned}$$

From the above form we conclude that p follows a beta distribution with shape

$= \sum_{i=1}^{n} y_i + a_0$ and rate $= \sum_{i=1}^{n}(n_i - y_i) + b_0$.

$$p|D \sim Beta\Big(\sum_{i=1}^{n} y_i + a_0, \sum_{i=1}^{n}(n_i - y_i) + b_0\Big) \tag{1.3}$$

We applied the posterior distribution that we got in Eq. 1.3 for a dataset with 100 observations. The dataset was generated by a binomial distribution with $p = 0.8$. We used the posterior to simulate samples of p, and we plotted the histogram in Fig.1.2-a. The posterior samples mean of p is equal to 0.7899, and it lies inside the 95% credible intervals, which is $(0.7654, 0.8134)$. The red line marks the true value for $p = 0.8$. In Fig.1.2-b, we plotted the likelihood function and the prior distribution with rate $= 2$ and shape $= 5$ which were represented by the long red dashed line. It is clear that it was far from the data. We plotted the posterior density also, which was represented by the solid blue line, and it is very close to the likelihood function.

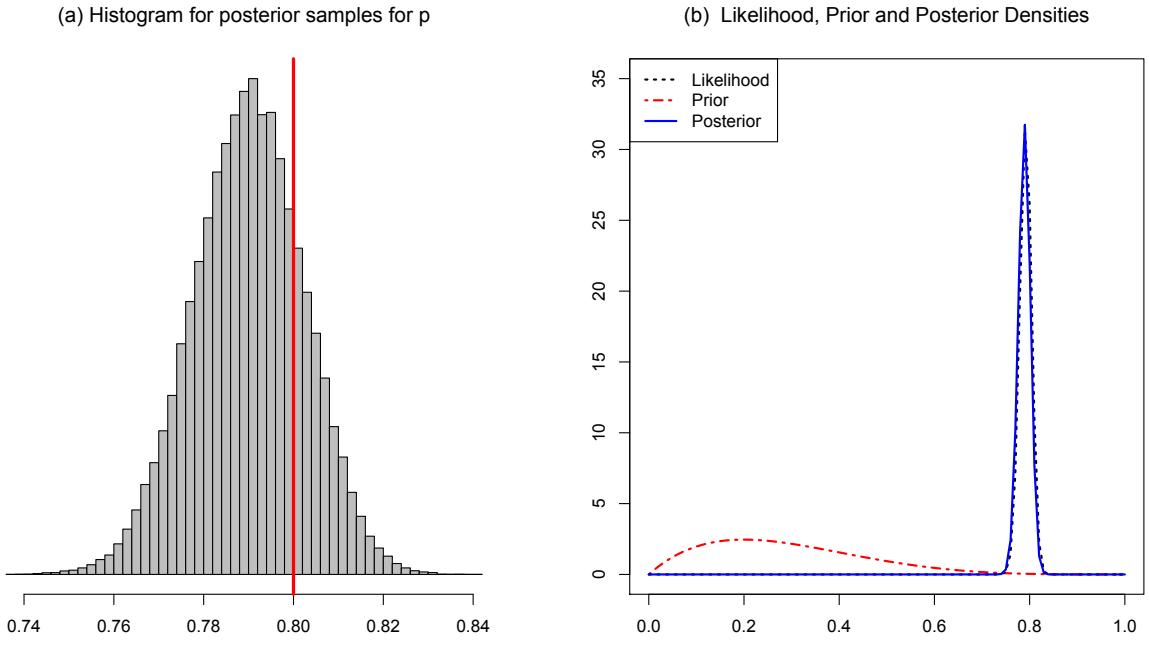

Fig. 1.2: Densities and Histogram for Binomial Distribution Example: red line denotes the true value from the simulation

1.2.2 Poisson Distribution Example

In a second illustration of Monte Carlo methods, we discussed the posterior inference for a Poisson model. Lets say $y_i \sim Poi(\lambda)$, where $i = 1, 2, \ldots, n$. The likelihood function of y is defined as follows:

$$\begin{aligned} L(y|\lambda) &= \prod_{i=1}^{n} f(y_i|\lambda) \\ &= \prod_{i=1}^{n} \exp(-\lambda)\frac{\lambda^{y_i}}{y_i!} \\ &= \exp(-n\lambda)\frac{\lambda^{y_1}\lambda^{y_2}\ldots\lambda^{y_n}}{y_1!y_2!\ldots y_n!} \\ &= \exp(-n\lambda)\frac{\lambda^{\sum_{i=1}^{n} y_i}}{y_1!y_2!\ldots y_n!} \end{aligned}$$

Now let us consider that λ follows the gamma prior distribution such as $\lambda \sim \Gamma(a_0, b_0)$, with density function $f(\lambda) = \frac{b_0^{a_0}}{\Gamma(a_0)}\lambda^{(a_0-1)}\exp(-b_0\lambda)$. As we can see, the prior distribution of λ is expressed as $\pi(\lambda) \propto \exp(-b_0\lambda)\lambda^{a_0-1}$. Notice that the prior distribution of λ , $\pi(\lambda)$, must be a distribution that only allows positive values because λ can only be present with positive real numbers.

After finding the likelihood function and the prior distribution, now we are ready to derive the posterior distribution for $\lambda|D$ by multiplying both of them as we can see below:

$$\begin{aligned} \pi(\lambda|y) &\propto L(y|\lambda) \times \pi(y) \\ &= \exp(-n\lambda)\frac{\lambda^{\sum_i^n y_i}}{y_1!y_2!\ldots y_n!}\exp(-b_0\lambda)\lambda^{a_0-1} \\ &= \exp(-n\lambda)\lambda^{\sum_i^n y_i}\exp(-b_0\lambda)\lambda^{a_0-1} \\ &= \exp(-(n+b_0)\lambda)\lambda^{\sum_i^n y_i+a_0-1} \end{aligned}$$

It is clear that the appropriate conditional posterior distribution for $\lambda|D$ is the gamma

distribution with shape $= \sum_{i}^{n} y_i + a_0$ and rate $= n + b_0$.

$$\lambda|D \sim \mathcal{G}am\Big(\sum_{i}^{n} y_i + a_0, n + b_0\Big) \tag{1.4}$$

The simulation data that we applied in the above method has 50 observations. It was generated by using a Poisson distribution with $\lambda = 4.8$. We used Eq. 1.4, to simulate samples for λ, and we plotted the histogram for the posterior samples in Fig. 1.3-a. The true value for $\lambda = 4.8$ was marked by the red line. The posterior sample mean, 4.6725, is close to the true value and it lies in the 95% CI which is $(4.1037, 5.2729)$. In Fig. 1.3-b, we plotted the prior, likelihood and posterior distributions. The posterior distribution is very close to the likelihood function which indicates that the likelihood has much stronger influence on the final inference than the prior.

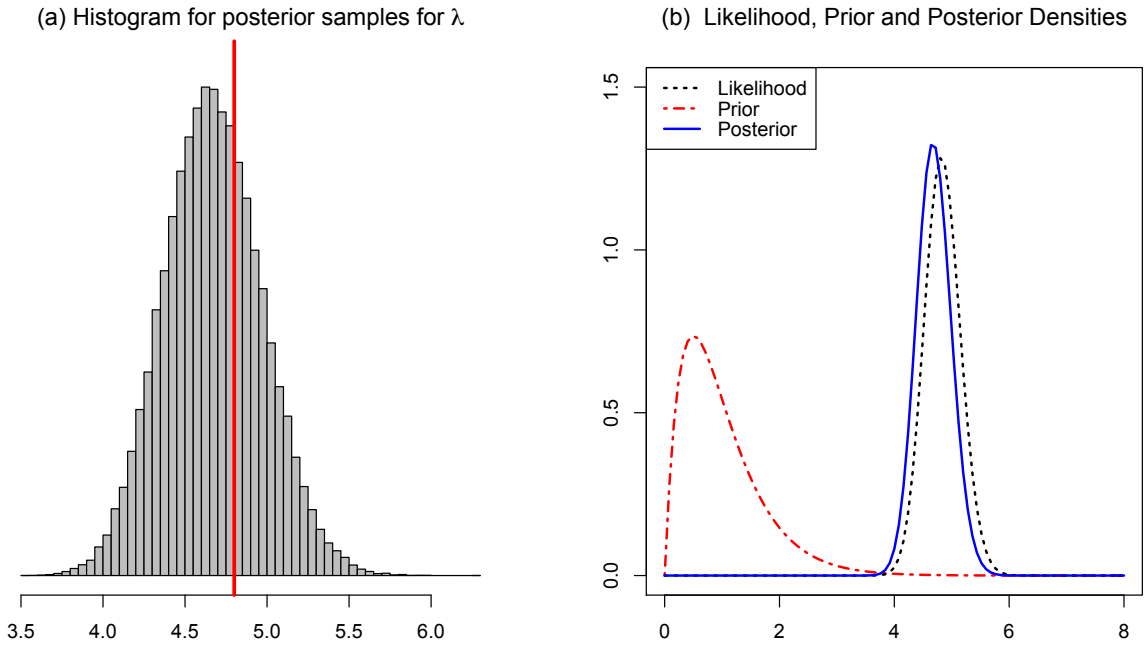

Fig. 1.3: Densities and Histogram for Poisson Distribution Example: red line denotes the true value from the simulation

However, it is possible that there may be situations where $\pi(\theta|D)$ is non-standard, and

it is not easy to sample from them directly. More commonly, a complex hierarchical model has many parameters. In that case, $\pi(\theta|D)$ is a high-dimensional posterior distribution. As described below, Markov Chain Monte Carlo is an efficient way to solve these types of problems.

1.3 Markov Chain Monte Carlo (MCMC)

The Markov Chain Monte Carlo (MCMC) method is useful when it is challenging to sample from the target posterior density. MCMC is one of the most important techniques in statistics, and it is used widely in many statistical applications. These kinds of methods are types of algorithms which work by sampling the probability distribution that is derived from a MC method (Gill et al., 2012). After several steps, the MCMC uses the chain as an approximate sample of desired distribution. The approximation improves after many steps are complete. In the remainder of this chapter, we will describe the three important MCMC methods that we used in this thesis, which are Metropolis-Hastings, Gibbs sampling and slice sampling, with necessary theoretical justification.

1.3.1 Metropolis-Hastings (M-H)

One of the important types in MCMC algorithms is the Metropolis - Hastings (M-H) algorithm. M-H works by drawing a sample from any proposal distribution $q(\theta|\theta_{old})$ whereas the original target was to draw from another density $p(\theta)$. One of the most important flexibilities which makes M-H more suitable is that it is adequate to know $p(\theta)$ up to a proportianality constant. That is useful because in some situations it is difficult to compute the necessary normalization factor. The process of the M-H algorithm works by creating samples one-by-one, moving from one sample to the next one. The algorithm uses a proposal distribution for the next sample that is dependent on the current sample value (Chib and Greenberg, 1995). This candidate will be either accepted or rejected based on

an acceptance probability. The probability of acceptance is determined by comparing the ratio of $p(\theta)$ at proposed and current values and the ratio of $q(\theta)$ at those values as well. Moreover, there is a special case of the M-H algorithm when the proposal function is symmetric which is called the Metropolis algorithm. We can summarize the M-H algorithm as follows.

Algorithm 1 Metropolis-Hastings (M-H)

INPUT: $\theta^{(i-1)} = (\theta_1^{(i-1)}, \theta_2^{(i-1)}, \ldots, \theta_k^{(i-1)})$, $q(x)$

OUTPUT: $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_k^{(i)})$

$\theta^{proposed} \sim q(\theta^{proposed}|\theta)$

$R(\theta^{proposed}, \theta^{(i-1)}) = min(1, \frac{p(\theta^{proposed})q(\theta^{(-1)}|\theta^{proposed})}{p(\theta^{(i-1)})q(\theta^{proposed}|\theta^{(i-1)})})$

$u \sim unif(0, 1)$

if $u \leq R(\theta^{proposed}, \theta^{(i-1)})$, **then**

$\theta^{(i)} = \theta^{proposed}$

else

$\theta^{(i)} = \theta^{(i-1)}$

end if

1.3.2 Gibbs Sampling

Gibbs sampling is a version of the MCMC method. The process is that if we are given a multivariate distribution, then we should divide the vector into several blocks and sample each block for its conditional distribution given other blocks. The idea in Gibbs sampling is to generate posterior samples by sweeping through each variable (or block of variables) (Carter and Kohn, 1994). Eventually, we reach a sample from its conditional distribution with the remaining variables fixed to their current values. Suppose we have k parameters $\theta_1, \theta_2, \ldots, \theta_k$, so Gibbs sampling can be found by first finding the joint posterior $\pi(\theta_1, \theta_2, \ldots, \theta_k|D)$ and by multiplying the likelihood with the prior $\pi(\theta_1), \pi(\theta_2), \ldots, \pi(\theta_k)$.

Then, we update parameters one by one. For example, for θ_1 and if we assume all other $\theta's$ are fixed at their current values, then we can find conditional posterior for θ_1 which is $\pi(\theta_1|\theta_2, \ldots, \theta_k, D)$. We repeat this process for θ_2 and then for θ_3 until θ_k; so we have k conditional posteriors. In the next step, we draw the rephrase symbol for each conditional distribution by using large numbers. That gives us posterior samples for each of $\theta_1, \theta_2, \ldots, \theta_k$ which we then analyze. In the following algorithm, we summarize the Gibbs sampler steps.

Algorithm 2 Gibbs Sampler

INPUT: $\theta^{(i-1)} = (\theta_1^{(i-1)}, \theta_2^{(i-1)}, \ldots, \theta_k^{(i-1)})$

OUTPUT: $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_k^{(i)})$

for $i = 1, 2, \ldots, N$ **do**

$\theta_1^{(i)} \sim p(\theta_1|\theta_2^{(i-1)}, \theta_2^{(i-1)}, \ldots, \theta_k^{(i-1)})$

$\theta_2^{(i)} \sim p(\theta_2|\theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_k^{(i-1)})$

$\theta_3^{(i)} \sim p(\theta_3|\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_k^{(i-1)})$

$\vdots$

$\theta_k^{(i)} \sim p(\theta_k|\theta_2^{(i)}, \theta_2^{(i)}, \ldots, \theta_{k-1}^{(i)})$

end for

In the next section, we will show an example for Gibbs sampling with normal distribution, where the two parameters μ and σ have been estimated by the above steps.

1.3.3 Slice Sampling

Slice sampling is another method of the MCMC technique. It is a useful tool for drawing samples from a posterior distribution which is not easy to draw from (Neal, 2003). To understand slice sampling, let us consider the parameter θ with distribution $f(\theta)$. It is easy to use slice sampling if the following conditions are satisfied: $f(\theta)$ can be written as $f(\theta) = g(\theta)h(\theta)$ with $h(\theta)$ always positive. Since it is not easy to draw samples from $f(\theta)$,

we need to introduce a new random variable u.

$$f(u,\theta) = \pi(u|\theta)f(\theta) = g(\theta)h(\theta)\frac{1}{h(\theta)}\mathbf{1}(u < h(\theta)) = g(\theta)\mathbf{1}(u < h(\theta))$$

We can summarize the slice sampler steps by the following algorithm.

Algorithm 3 Slice Sampler

INPUT: $\theta^{(i-1)} = (\theta_1^{(i-1)}, \theta_2^{(i-1)}, \ldots, \theta_k^{(i-1)})$

OUTPUT: $\theta^{(i)} = (\theta_1^{(i)}, \theta_2^{(i)}, \ldots, \theta_k^{(i)})$

$f(\theta) = g(\theta)h(\theta)$ where $h(\theta) > 0$

for $i = 1, 2, \ldots, N$ **do**

$u \sim unif(0, h(\theta))$

$\pi(u|\theta) = \frac{1}{h(\theta)}\mathbf{1}(u < h(\theta))$

$f(\theta|u) = g(\theta)\mathbf{1}(\theta \in H(u))$ where $H(u) = h^{-1}(\theta)$

end for

1.4 Illustration Examples

We implement Gibbs sampling with a simulation from normal distribution. We summarize the result for posterior distributions by using diagrams.

1.4.1 Simulation Example from Normal Distribution

Suppose, $x_1, \ldots, x_n \sim N(\mu, \sigma^2)$ where μ, σ^2 unknown.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\Big)$$

In this example, we want to understand how Gibbs sampling works with a normal distribution that has two parameters, μ and σ^2. In the beginning, we found the likelihood

function by multiplying the n densities functions. The likelihood function will be,

$$f(x) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\Big) \tag{1.5}$$

Now, we assume appropriate priors for both unknown parameters. Let us assume that the prior for μ is $N(\mu_0, \sigma_0^2)$. Inverse gamma is the conjugate prior distribution of variance of the normal distribution. Because of this, we choose the prior distribution for σ^2 to be $IG(a_0, b_0)$.

$$\begin{aligned}
\mu &\sim N(\mu_0, \sigma_0^2) \propto \exp\Big(\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\Big) \\
\sigma^2 &\sim IG(a_0, b_0) \propto \frac{1}{(\sigma^2)^{a_0+1}} \exp\Big(-\frac{b_0}{\sigma^2}\Big)
\end{aligned}$$

We obtain the conditional posterior for μ, σ^2 by multiplying the likelihood function by the prior distributions of μ, σ^2. Beginning with the conditional posterior for μ, the procedures were derived by the posterior distribution for μ shown below

$$\begin{aligned}
\pi(\mu|\sigma^2, D) &\propto \prod_{i=1}^{n} \exp\Big(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\Big) \exp\Big(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\Big) \\
&\propto \exp\Big(-\frac{1}{2}\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^2}\Big) \exp\Big(\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big[\frac{\sum_{i=1}^{n}(x_i^2 - 2\mu\mu_i + \mu^2)}{\sigma^2} + \frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{\sigma_0^2}\Big]\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big[\frac{\sum_{i=1}^{n}(-2\mu x_i + \mu^2)}{\sigma_0^2} + \frac{\mu^2 - 2\mu\mu_0}{\sigma_0^2}\Big]\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big[\mu^2(\frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}) - 2\mu(\frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2})\Big]\Big)
\end{aligned}$$

For simplification, we will use, $A = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}$, and $B = \frac{\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}$, so the posterior

distribution for μ will be

$$\begin{aligned} f(\mu|\sigma^2, D) &\propto \exp\Big(-\frac{1}{2}\Big[A\mu^2 - 2B\mu\Big]\Big) \\ &\propto \exp\Big(-\frac{1}{2}\Big[A\mu^2 - 2\frac{B}{A}A\mu\Big]\Big) \\ &\propto \exp\Big(\frac{-A}{2}\Big[\mu^2 - 2\frac{B}{A}\mu\Big]\Big) \\ &\propto \exp\Big(\frac{-A}{2}\Big[\mu^2 - 2\frac{B}{A}\mu + (\frac{B}{A})^2\Big]\Big) \\ &\propto \exp\Big(\frac{-A}{2}\Big(\mu - (\frac{B}{A})^2\Big) \\ &\propto \exp\Big(-\frac{1}{2}\frac{\mu - (\frac{B}{A})^2}{\frac{1}{A}}\Big) \end{aligned}$$

From the above computation, the conditional posterior distribution for μ given σ^2 and the data is also a normal distribution with mean equal to $\frac{B}{A}$ and variance equal to $\frac{1}{A}$,

$$\mu|\sigma^2, D \sim N(\frac{B}{A}, \frac{1}{A}) \tag{1.6}$$

Second, we will find the posterior distribution for σ^2 by fixing μ and the data by following the same process as above.

$$\begin{aligned} \pi(\sigma^2|\mu, D) &= \frac{1}{(\sigma^2)^{\frac{1}{2}}}\exp\Big(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\Big) \times \frac{1}{(\sigma^2)^{a_0+1}}\exp\Big(-\frac{b_0}{\sigma^2}\Big) \\ &= \frac{1}{(\sigma^2)^{\frac{n}{2}}}\exp\Big(-\frac{1}{2}\frac{\sum_{i=1}^n(x_i-\mu)^2}{\sigma^2}\Big) \times \frac{1}{(\sigma^2)^{a_0+1}}\exp\Big(-\frac{b_0}{\sigma^2}\Big) \\ &= \frac{1}{(\sigma^2)^{\frac{n}{2}+a_0+1}}\exp\Big(-\frac{1}{\sigma^2}\Big[\frac{\sum_{i=1}^n(x_i-\mu)^2}{2} + b_0\Big]\Big) \end{aligned}$$

Apparent from this computation, the conditional posterior distribution for σ^2 is an inverse gamma with shape $= \frac{n}{2} + a_0$, and scale $= \frac{\sum_{i=1}^n(x_i-\mu)^2}{2} + b_0$. The conditional

posterior distribution for σ^2 given μ and the data (D) is

$$\sigma^2|\mu, D \sim IG(\frac{n}{2} + a_0, \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2} + b_0) \tag{1.7}$$

In this case, $\frac{1}{\sigma^2}$ will follow the Gamma distribution with the same shape and rate. In other words, since σ^2 is an inverse gamma distribution, it follows that $\frac{1}{\sigma^2}$ is a gamma distribution.

$$\frac{1}{\sigma^2} \sim \Gamma(\frac{n}{2} + a_0, \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2} + b_0) \tag{1.8}$$

We can fit the Bayesian model based on the posterior distribution for μ and posterior distribution for σ^2 that we derived above. We generated a simulation dataset with 200 observations by using the normal distribution with $\mu = 3.8$ and $\sigma^2 = 4$. In Fig. 1.4, we plotted the likelihood function, prior distribution and the posterior distribution from a normal distribution with small and large datasets. We chose the initial parameters for normal prior distribution to be $\mu = -10$ and $\sigma^2 = 10$, and for the inverse gamma distribution to be $a_0 = 2.1$ and $b_0 = 2.1$.

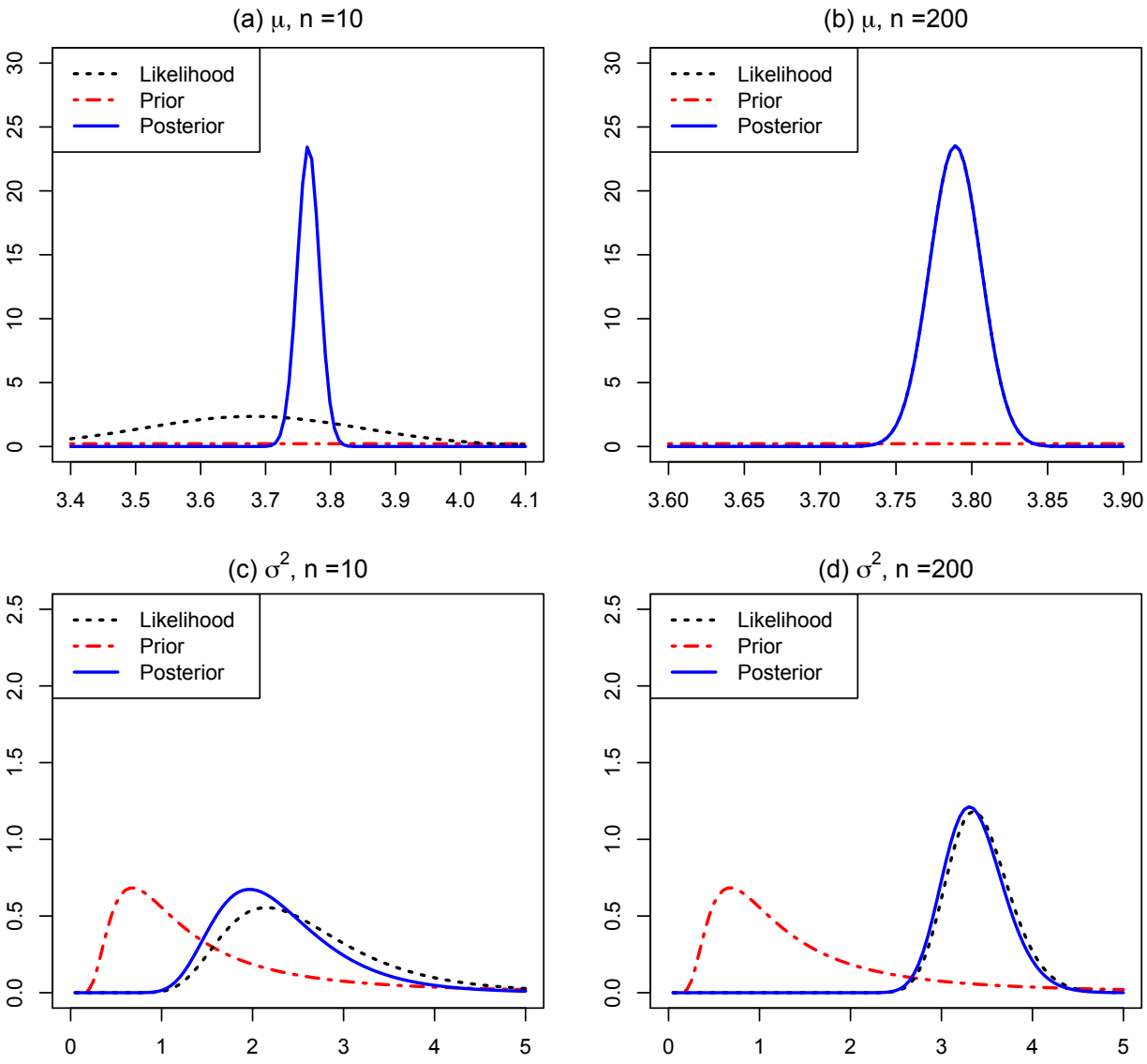

Fig. 1.4: Posterior samples distribution plots for μ and σ^2 with small and large sample size

We used Eq.1.6 and Eq. 1.8 to simulate samples for μ and σ^2, respectively. The resulting histogram shown in Fig.1.5 displays the fundamental difference for both μ and σ^2 when we chose a different number of iterations. When we had a small number of iterations, the distributions of both μ and σ^2 are unorganized (skewed), and it looks like a multi-model as we can see in Fig. 1.5-(a,c). However, when we had a large number of iterations, we actually approached a normal distribution as we see in Fig. 1.5-(b,d). It is clear that they

are approximately symmetric. It is obvious that we need a large number of iterations in Gibbs sampling since our choices for the prior initial values were far away from the true values. In Fig. 1.5-(a-d), the true values for μ and σ^2 are delineated by the red line.

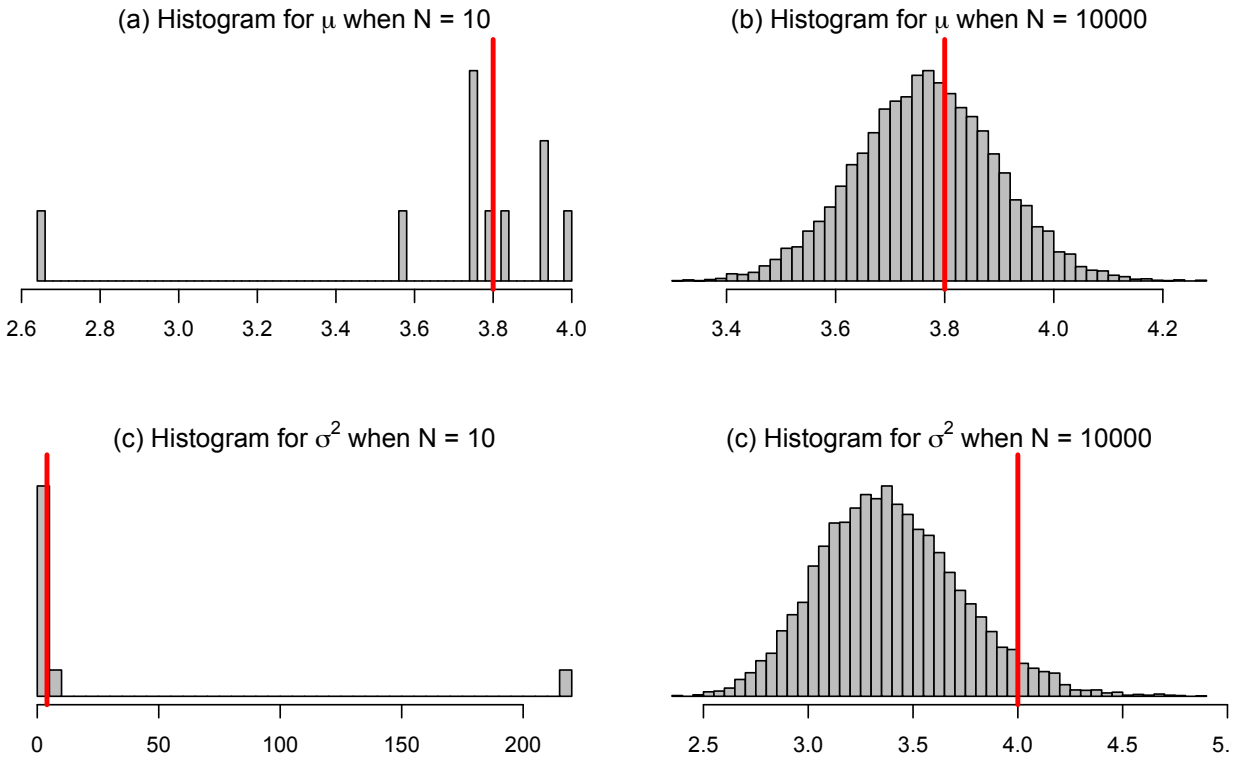

Fig. 1.5: Histograms for posterior samples for μ and σ^2 when N = 100 and N = 10000: red line denotes the true value from the simulation

In Table 1.1, we included the true value, posterior mean and 95% credible intervals for both μ and σ^2. The credible interval when $N = 10$ is wider than the credible intervals when $N = 10000$. This indicates that for a large number of iterations our posterior mean becomes closer to the true value.

Table 1.1: Comparison among true values, posterior means and 95% credible intervals when $N = 10$ iterations vs. $N = 10000$ iterations

	# of Iterations	N = 10		N = 10000	
	True Value	Posterior Mean	95% Credible Interval	Posterior Mean	95% Credible Interval
μ	3.8	3.6966	(2.8631, 3.9730)	3.7665	(3.5109, 4.0189)
σ^2	4.0	24.6631	(3.0654, 167.8721)	3.3920	(2.7836, 4.1155)

Chapter 2

Bayesian Inference in Regression

2.1 Introduction

Regression analysis is a statistical procedure to estimate the relationships between variables and how to use that relationship for predictive purposes. In general, regression analysis focuses on the relationship between a dependent variable and one or more independent variables, which are also known as the predictors. Regression analysis is important because it can help us understand the changes of a specific value of the dependent variable when an independent variable is altered, while the other independent variables remain constant (Vaughn, 2008). In this case, to model linear regression, we can use an advertising expenditures example - advertising expense is the independent variable and the dependent variable is sales. This model can be useful for businesses making advertising decisions.

Depending on how many independent variables we are using, we refer to it as simple and multiple regression. Simple regression is used when we have only one independent variable while multiple regression is used when more than one independent variable is present. Moreover, the relationships among variables can be of different kinds - if the relationship is assumed to be in the form of a straight line then it can be defined by a linear regression model, while the nonlinear regression model can be defined by general curves between the variables (Vaughn, 2008). There are many different ways to approach a regression model. Some are simple, while others are quite complex, but all have the power to explain the response in different levels.

In this chapter, we will discuss MCMC algorithms for multiple linear regression which we are going to use as a starting point for developing outlier detection methods in Chapter

3.

2.2 Bayesian Inference of a Multivariate Linear Regression

In this section, we are going to focus on construction of a Gibbs sampler for multiple linear regression. The linear regression equation is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \text{ where } \epsilon_i \sim N(0, \sigma^2)$$

For n observations, the multiple linear regression in matrix notations is, given by y, where

$$Y = X\beta + \epsilon, \text{ where } \epsilon \sim MVN_n(0_n, \sigma^2 I_n)$$

such that Y is a $n \times 1$ vector, β is an $(p+1) \times 1$ vector, X is $n \times (p+1)$ matrix, 0_n is zero vector of length n, σ^2 is a scalar value, and I_n is $n \times n$ identical matrix.

From above, we conclude that $y \sim MVN_n(X\beta, \sigma^2 I_n)$, so the likelihood function is defined as follows:

$$\begin{aligned}
L(y) &= \prod_{i=1}^{n} f(y_i) \\
&= \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\Big(-\frac{1}{2}(y - X\beta)^T (\sigma^2 I_n)^{-1} (y - X\beta)\Big) \\
&= \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\Big(-\frac{1}{2}(y - X\beta)^T \frac{I_n}{\sigma^2} (y - X\beta)\Big) \\
&= \frac{1}{(\sigma^2)^{\frac{n}{2}}} \exp\Big(-\frac{1}{2}\frac{(y - X\beta)^T (y - X\beta)}{\sigma^2}\Big)
\end{aligned}$$

We chose the prior for β as $\beta \sim MVN(0, c_0 I_{p+1})$, where c_0 was chosen to be a large value, the prior for σ^2 as $\sigma^2 \sim IG(a_0, b_0)$, and a_0 and b_0 are the initial values. The

conditional posterior distribution for β can be defined as follows:

$$\begin{aligned}
\pi(\beta|\sigma^2, D) &\propto L(y) \times \Pi(\sigma^2|\beta, D) \\
&\propto \exp\Big(-\frac{1}{2}\Big[\frac{(y-X\beta)^T(y-X\beta)}{\sigma^2} + \frac{\beta^T\beta}{c_0}\Big]\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big[\frac{y^Ty - y^TX\beta - \beta^TX^Ty + \beta^TX^TX\beta}{\sigma^2} + \frac{\beta^T\beta}{c_0}\Big]\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big[\frac{-y^TX\beta - \beta^TXy + \beta^TX^TX\beta}{\sigma^2} + \frac{\beta^T\beta}{c_0}\Big]\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big[\frac{-2\beta^TX^Ty + \beta^TX^TX\beta}{\sigma^2} + \frac{\beta^T\beta}{c_0}\Big]\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big[-2\beta^T\frac{X^Ty}{\sigma^2} + \beta^T(\frac{X^TX}{\sigma^2})\beta + \beta^T(\frac{I}{c_0})\beta\Big]\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big[-2\beta^T\frac{X^Ty}{\sigma^2} + \beta^T\Big]\Big) \\
&\propto \exp\Big(-\frac{1}{2}\Big[-2\beta^T\frac{X^Ty}{\sigma^2} + \beta^T[\frac{X^TX}{\sigma^2} + \frac{I}{c_0}]\beta\Big]\Big)
\end{aligned}$$

By considering $b = \frac{X^Ty}{\sigma^2}$ and $A = \frac{X^TX}{\sigma^2} + \frac{I}{c_0}$, we can rewrite the posterior distribution for β as follows:

$$\begin{aligned}
\pi(\beta|\sigma^2, D) &\propto f(y) \times \Pi(\sigma^2) \\
&\propto \exp\Big((-\frac{1}{2}\Big[\beta^TA\beta - 2\beta^Tb\Big]\Big)\Big) \\
&\propto \exp\Big(-\frac{1}{2}(\beta-p)^T\phi(\beta-p)\Big) \\
&\propto \exp\Big(-\frac{1}{2}(\beta-A^{-1}b)^TA(\beta-A^{-1}b)\Big)
\end{aligned}$$

The above equation matches the multivariate normal density with $\mu = A^{-1}b$ and $\Sigma = A^{-1}$. Therefore, the full conditional posterior for β is

$$\beta \sim MVN(A^{-1}b, A^{-1}) \text{ where } A = \frac{X^TX}{\sigma^2} + \frac{I}{c_0} \text{ and } b = \frac{X^Ty}{\sigma^2} \tag{2.1}$$

We found the full conditional posterior distribution for σ^2. Let us define

$G_1 = (y - X\beta)^T(y - X\beta)$, hence

$$\begin{aligned}
\pi(\sigma^2|\beta, D) &\propto (\sigma^2)^{-\frac{n}{2}} \exp\Big(-\frac{1}{2}\frac{G_1}{\sigma^2}\Big)(\sigma^2)^{-(a_0+1)} \exp\Big(-\frac{b_0}{\sigma^2}\Big) \\
&\propto (\sigma^2)^{-\frac{n}{2}-(a_0+1)} \exp\Big(-\frac{1}{\sigma^2}(\frac{G_1}{2} + b_0)\Big) \\
&\propto (\sigma^2)^{-(\frac{n}{2}+a_0+1)} \exp\Big(-\frac{1}{\sigma^2}(\frac{G_1}{2} + b_0)\Big)
\end{aligned}$$

The above formula is similar to the inverse gamma distribution with shape equal to $\frac{n}{2} + a_0$ and rate equal to $\frac{G_1}{2} + b_0$. So, the posterior distribution for σ^2 is

$$\sigma^2 \sim IG(\frac{n}{2} + a_0, \frac{G_1}{2} + b_0) \tag{2.2}$$

2.3 Illustrative Examples

We applied the Gibbs sampler that was developed above to two regression examples - one with a simulated dataset and another with a real dataset.

2.3.1 Example 1: Simulation Example with Three Covariates

In this example, we used the posterior distributions for μ and σ^2 (Eq. 2.1 and Eq. 2.2). We generated 250 observations by first simulating x-variables uniformly with $(0, 1)$, and then we used $\beta = (2.5, -1.2, 4.8, 0.07)$ and $\sigma^2 = 0.16$. Running the Gibbs sampler for 100000 iterations, we acquired the posterior sample mean for all the parameters, which are $\beta_0, \beta_1, \beta_2, \beta_3$ and σ^2. We plotted the histograms for posterior samples of these parameters in Fig. 2.1, and we marked the true values for these parameters by a red line. All the parameters in the posterior sample histograms were approximately normal, and the true values for the parameters were close to the samples mean. Moreover, we used the model to predict the response for selected x values of $(1.23, 0.12, 0.37, 0.32)$. The responses were plotted as histograms for the all the samples for the prediction, and we marked the true

value with the red line.

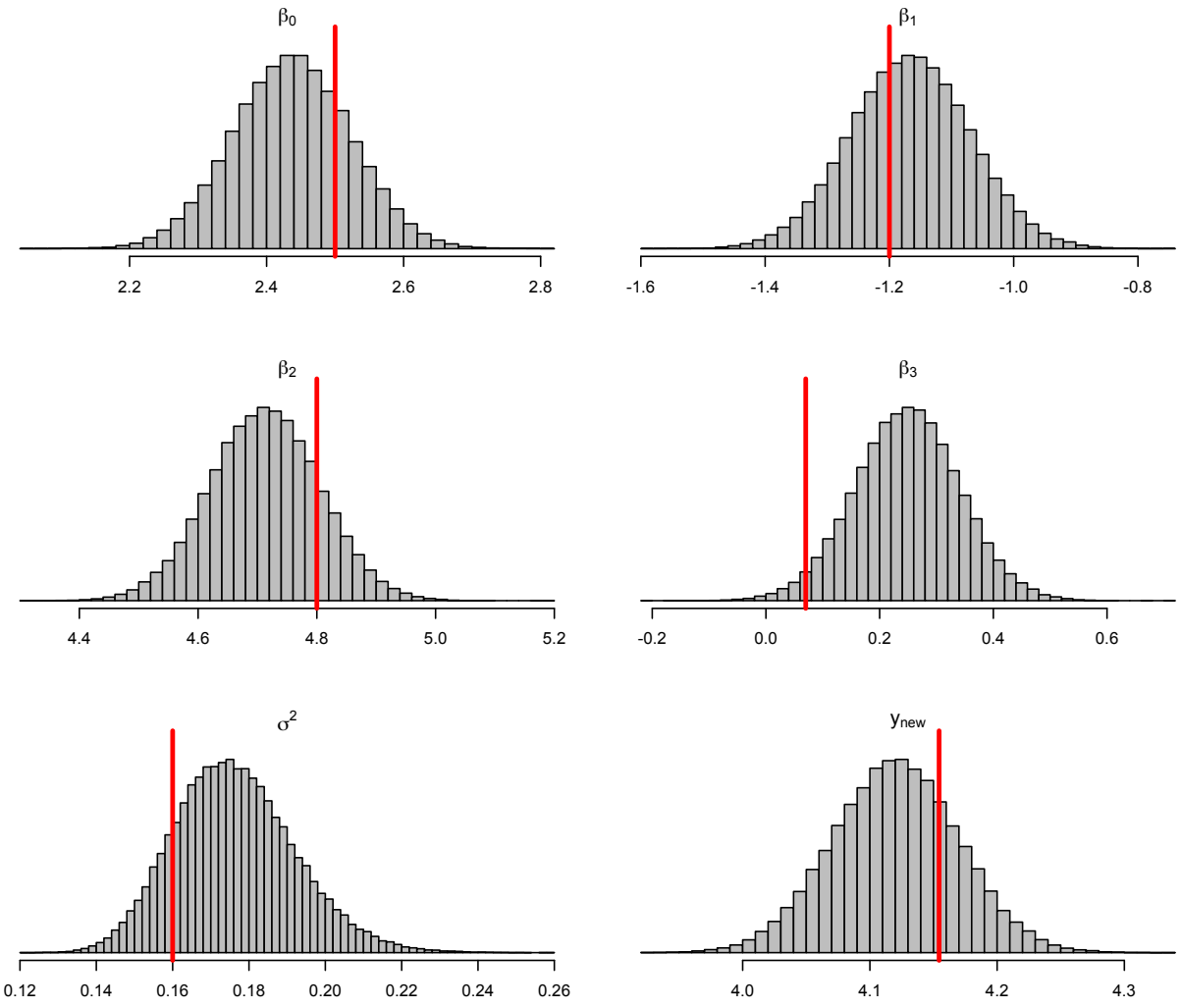

Fig. 2.1: Posterior Histograms for $\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2$ and y_{new}: red line denotes the true value from the simulation

In Table 2.1, we compared the true values for our parameters with the posterior sample means, and we created 95% CI for all the parameters and the prediction of y_{new}. All of the true values lie inside the 95% CI which indicates that the estimation is satisfactory.

Table 2.1: True Values, Predicted Values & 95% Credible Interval

	True Value	Posterior Mean	95% Credible Interval
β_0	2.50	2.4369	(2.2686 , 2.6038)
β_1	-1.20	-1.1664	(-1.3563 , -0.9770)
β_2	4.80	4.7123	(4.5299 , 4.8930)
β_3	0.07	0.2487	(0.0671 , 0.4292)
σ^2	0.16	0.1763	(0.1480 , 0.2101)
y_{new}	4.15	4.1201	(4.0226 , 4.2174)

2.3.2 Example 2: Analysis of Rock Strength Dataset

We used the Gibbs sampler to analyze the Rock Strength Dataset (RSD). The RSD (Ali et al., 2014) contains information about the relationship between Uniaxial Compressive Strength (UCS) and 8 predictors, which are %Quartz, %Plagaoclase, %K.feldspar, %Hornblende, Grain size (mm), Grain area (mm^2), Aspect Ratio and Shape Factor for 30 rock specimens. First, we normalized the data, and then we ran the Gibbs sampler for 10000 iterations to create posterior samples for our parameters, which are $\beta_0, \ldots, \beta_8,$ and σ^2.

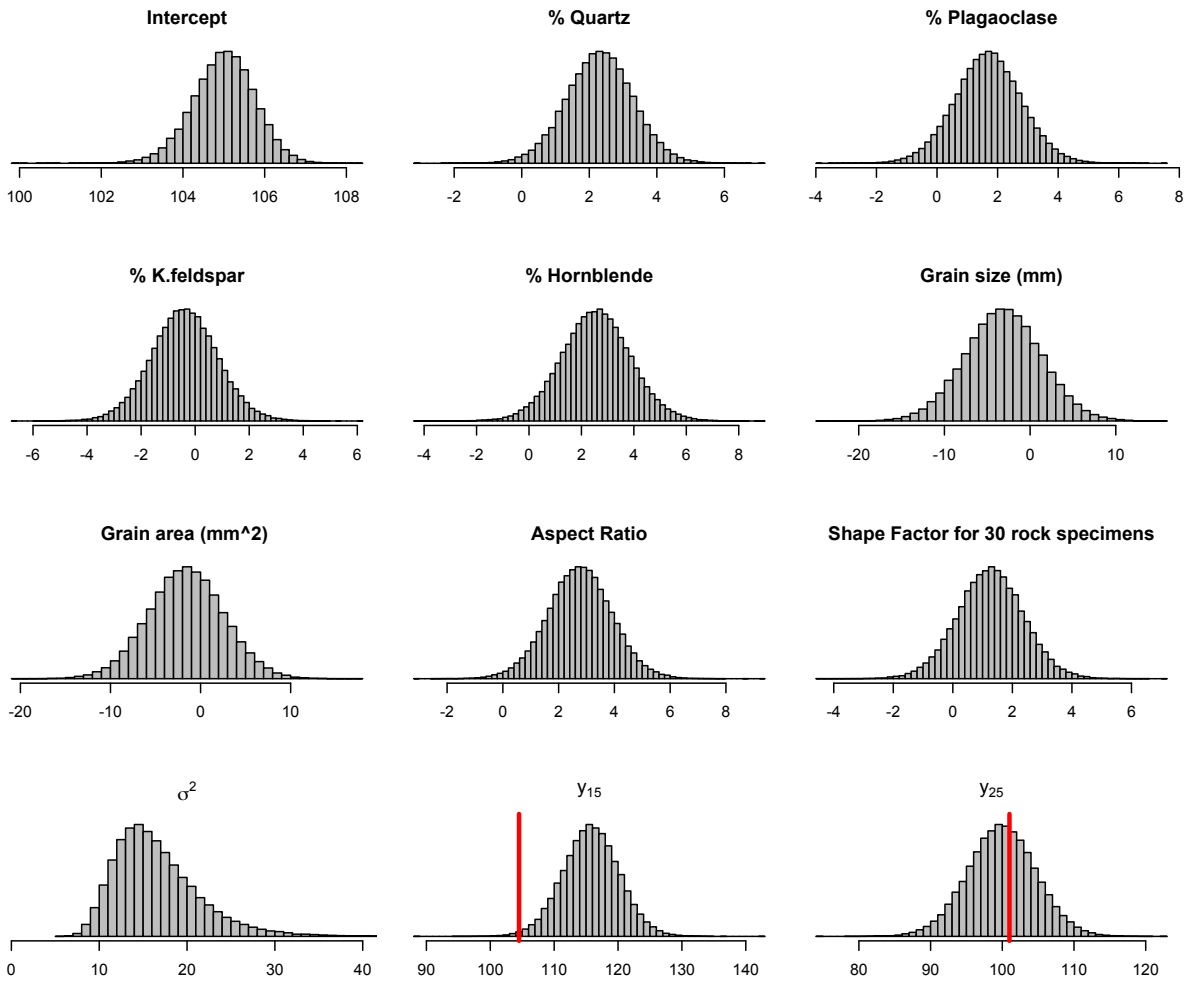

Fig. 2.2: Posterior Histograms for RSD: red line denotes the true UCS for test observations

In Fig. 2.2, we plotted the histograms for posterior samples for all the parameters in our model. It is clear that they are approximately symmetric with a large number of iterations. We checked our prediction by using leave-one-out-cross validation (LOOCV) by leaving out one of the observations and estimating the model with the rest of the observations. In our example, we have 30 observations, which means we repeated the procedure 30 times. We found that 93% of the prediction of our samples lie inside the 95% credible intervals, which means that 28 of 30 prediction values lie inside the 95% CI. We plotted the histogram for two posterior samples, y_{15} and y_{25}, one lying inside the 95% CI and the other outside. We marked the true value for y_{15} and y_{25} by the red lines.

In Table 2.2, we show all the posterior means for all the parameters and the two prediction values. The most significant variables were % Quartz and Aspect Ratio.

Table 2.2: Predicted Values & 95% Credible Interval

Intercept	**% Quartz**	% Plagaoclase
104.9958 (103.4305 , 106.4273)	**2.3129 (0.3548 , 4.2481)**	1.6855 (-0.5103 , 3.8742)
% K.feldspar	% Hornblende	Grain size (mm)
-0.3945 (-2.8232 , 2.0350)	2.5563 (-0.05526 , 5.1791)	-3.1271 (-11.8750 , 5.6153)
Grain area (mm^2)	**Aspect Ratio**	Shape Factor for 30 rock specimens
-1.6038 (-9.8060 , 6.6122)	**2.7616 (0.4790 , 5.0350)**	1.2611 (-0.9408 , 3.4443)
σ^2	$y_{15} = 104.5$	$y_{25} = 101$
16.6176 (9.5632 , 28.6522)	115.7421(107.0332 , 124.4515)	99.8877 (90.5009 , 109.2746)

Chapter 3

Bayesian Model for Outliers in Regression

3.1 Introduction

In this chapter, we are focusing on detecting outliers from a large dataset using regression. Longitudinal studies are one of the common studies in many different types of research. A longitudinal study is useful to understand the changes of development over long periods of time such as days, weeks, months, or even years. The longitudinal studies are very useful for many health and medical issues. Obesity is a growing concern, and it is best measured with longitudinal studies. According to past research, approximately one in three children in the United States over two years of age are either overweight or obese (Ogden et al., 2012). Arkansas has an even greater problem with childhood obesity, with 38.8% either overweight or obese in the 2013-2014 school years. Moreover, there is a correlation between childhood obesity and adult obesity. Since this kind of study tracks people over long periods of time, the longitudinal studies are very useful for studying obesity. But these kinds of studies are very sensitive because the data of these studies is changing over time. There are many elements that are affected by change over periods of time which will affect the results of the study. However, the biggest challenge in these kinds of studies is the occurrence of outlying observations. In fact, there are two different types of outliers in longitudinal studies, which are cross-sectional and longitudinal cross-sectional outliers. We can define the outliers as those observations that are different from the rest of the observations of the same individual. There are several things that cause outliers, such as incorrect data entry, inappropriate measurements, and biased observations recorded on different individuals. These kinds of outliers can affect the results of an experiment. In this

chapter, we built models to be applied to datasets on height, weight, and BMI, gathered on public school children in the state of Arkansas to detect and eliminate cross-sectional and longitudinal outliers before the outliers can be used for analysis. We expect the models to be more sensitive to detect the outliers among heights as compared to weights. Because children of school age may have many elements that will affect their growth, children of the same age and in the same grade will not have the same measurments for height due to several factors, such as lifestyle, weight and genetics. Those factors can be added as covariates in regression. In this chapter, we will work with three different methods to detect outliers in heights and weights, and then suggest the best one for detecting outliers.

In regression, we can not consider an observation to be an outlier just because it has a very high/low value. Rather, an observation should be an outlier if compared to the rest of the data, it could not be explained by the covariates. Things would become more challenging if the ability of the model to explain a response value depends on the covariates of the observation (systematic heteroscedasticity). We shall start with simpler hierarchical models without taking into account this type of behavior and then improve our model to address this.

3.2 Hierarchical Model for Outlier detection

We defined our multiple regression model as follows:

$$Y = X\beta + \epsilon, \text{ where } \epsilon \sim MVN_n(0_n, \Sigma)$$

where Σ is a diagonal matrix of σ_i^2 , $i = 1, \ldots, n$. We allow every observation to vary around the regression line. Because we tried to detect the outliers, we had a separate σ_i^2 for every observation instead of having only one σ^2 for data entered. An outlier will be indicated by ϵ_i having a large positive or negative value. This in turn implies the

corresponding σ_i^2 must be a large positive (because the mean for ϵ is zero which means σ_i^2 needs to be large to allow large positive or negative values of ϵ_i.). Using a common σ^2 would not achieve this constraint. In this case, our full posterior distribution will be as follows:

$$\pi(\sigma_1^2, \ldots, \sigma_n^2 | D, \beta) \propto L(D|\beta, \sigma_1^2, \ldots, \sigma_n^2)\pi(\sigma_1^2) \ldots \pi(\sigma_n^2) \tag{3.1}$$

To detect the outliers, we need to identify any position i with large σ_i^2.

There are many ways to solve this problem, but first we need to put a prior on each σ_i^2. There are two options for choosing appropriate prior distributions for each σ_i^2. We will apply both options and refer to them as Method I and Method II.

For Method I, suppose that σ_i^2 follows an inverse gamma prior distribution with a shape equal to a_0 and rate equal to b_0 (i.e., $\sigma_i^2 \sim IG(a_0, b_0)$, $i = 1, \ldots n$). So,

$$\pi(\sigma_i^2) = (\sigma_i^2)^{-a_0-1} \times \exp(-\frac{b_0}{\sigma_i^2}) \tag{3.2}$$

The likelihood function in this case can be written as:

$$\begin{aligned}
L(y) &\propto |\Sigma|^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}(y - X\beta)^T \Sigma^{-1}(y - X\beta)\Big) \\
&\propto (\sigma_1^2 \ldots \sigma_n^2)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}(y - X\beta)^T diag(\frac{1}{\sigma_i^2})(y - X\beta)\Big), i = 1, \ldots, n. \\
&\propto (\sigma_1^2 \ldots \sigma_n^2)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\Big[y_i - x_1^T\beta, y_2 - x_2^T\beta, \ldots, y_n - x_n^T\beta\Big]\Big) \\
&\times diag(\frac{1}{\sigma_i^2}) \begin{bmatrix} y_1 - x_1^T\beta \\ y_2 - x_2^T\beta \\ \vdots \\ y_n - x_n^T\beta \end{bmatrix} \\
&\propto (\sigma_1^2 \ldots \sigma_n^2)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\Big[\frac{(y_1 - x_1^T\beta)^2}{\sigma_1^2} + \frac{(y_2 - x_2^T\beta)^2}{\sigma_2^2} + \cdots + \frac{(y_n - x_n^T\beta)^2}{\sigma_n^2}\Big]\Big)
\end{aligned}$$

Hence,

$$L(y) = (\sigma_1^2 \dots \sigma_n^2)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - x_i^T\beta)}{\sigma_i^2}\Big) \tag{3.3}$$

The parameters for this method were β, σ_1^2, σ_2^2,...,σ_n^2. To develop a Gibbs sampler, we first simulated the posterior distribution for β given the data and all $\sigma_i^2, i = 1, \dots, n$. Then we simulated the posterior distribution for σ_i^2 given β and the data. As we showed in Chapter 2, the posterior conditional distribution for β given the data and $\sigma_1^2, \dots, \sigma_n^2$ follows a multivariate normal distribution, with a mean of $A^{-1}b$ and variance of A^{-1}, (i.e., $\beta|\sigma_1^2, \dots, \sigma_n^2 \sim MVN(A^{-1}b, A^{-1})$). So,

$$\pi(\beta|\sigma_1^2, \dots, \sigma_n^2, D) \quad \propto \quad \exp\Big(-\frac{1}{2}(\beta - A^{-1}b)^T(\beta - A^{-1}b)\Big) \tag{3.4}$$

where $A = X^T\Sigma^{-1}X + \frac{I_n}{c_0}$ and $b = X^T\Sigma^{-1}y$.

The posterior conditional distribution for $\sigma_1^2, \dots, \sigma_n^2$ given β and D is derived as follows:

$$\pi(\sigma_1^2, \dots, \sigma_n^2|\beta, D) \quad \propto \quad L(y)\pi(\sigma_1^2) \dots \pi(\sigma_n^2) \tag{3.5}$$

Now, from Eqs. 3.2, 3.3 and 3.5, we can compute the posterior distribution for $\pi(\sigma_1^2, \dots, \sigma_n^2|\beta, D)$ as follows,

$$\begin{aligned}\pi(\sigma_1^2, \dots, \sigma_n^2|\beta, D) \quad &\propto \quad (\sigma_1^2 \dots \sigma_n^2)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - x_i^T\beta)^2}{\sigma_i^2}\Big)\,(\sigma_1^2)^{-a_0-1}\exp\Big(-\frac{b_0}{\sigma_1^2}\Big) \\ &\times \quad (\sigma_2^2)^{-a_0-1}\exp\Big(-\frac{b_0}{\sigma_2^2}\Big) \dots (\sigma_n^2)^{-a_0-1}\exp\Big(-\frac{b_0}{\sigma_n^2}\Big)\end{aligned}$$

Suppose we are interested in the posterior of σ_1^2. We consider the terms that involve only σ_1^2 and ignore the other terms, so

$$\pi(\sigma_1^2|\beta, \sigma_2^2, \sigma_3^2, \dots, \sigma_n^2, D) \quad \propto \quad (\sigma_1^2)^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\frac{(y_1 - x_1^T\beta)^2}{\sigma_1^2}\Big)(\sigma_1^2)^{-a_0-1}\exp\Big(-\frac{b_0}{\sigma_1^2}\Big)$$

$$\begin{aligned} &\propto (\sigma_1^2)^{-\frac{1}{2}-a_0-1} \exp\Big(-\frac{1}{\sigma_1^2}\Big[\frac{(y_1 - x_1^T\beta)^2}{2} + b_0\Big]\Big) \\ &\propto (\sigma_1^2)^{-(\frac{1}{2}+a_0)-1} \exp\Big(-\frac{1}{\sigma_1^2}\Big[-\frac{(y_1 - x_1^T\beta)^2}{2} + b_0\Big]\Big) \end{aligned}$$

From the above formula, it is clear that the posterior distribution for σ_1^2 is an inverse gamma with a shape equal to $a_0 + \frac{1}{2}$ and rate equal to $\frac{(y_1 - x_1^T\beta)^2}{2} + b_0$,

$$\sigma_1^2 \sim IG(a_0 + \frac{1}{2}, \frac{(y_1 - x_1^T\beta)^2}{2} + b_0)$$

In general, the posterior conditional distribution for σ_i^2, $i = 1, \ldots, n$ is an inverse gamma,

$$\sigma_i^2 \sim IG(a_0 + \frac{1}{2}, \frac{(y_i - x_i^T\beta)^2}{2} + b_0) \tag{3.6}$$

From the above discussion, we can use the full conditional distribution for β and σ^2, which are represented in Eq. 3.3 and Eq. 3.6, respectively. We can summarize Method I by the following algorithm:

Algorithm 4 Method I

INPUT: Prior parameters a_0, b_0, c_0 and initial values β and σ^2.

OUTPUT: $\beta, \sigma_1^2, \ldots, \sigma_n^2$

for $i = 1, 2, \ldots, N$ **do**

Given $\sigma_1^2, \ldots, \sigma_2^2$ draw $\beta \sim MVN(A^{-1}b, A^{-1})$

Given β draw $\sigma_j^2 \sim IG(a_0 + \frac{1}{2}, \frac{(y_j - X_j^T\beta)^2}{2} + b_0)$

end for

Suppose our model is

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i, \text{ where } i = 1, 2, \ldots, n.$$

As we know, in a usual regression model the mean for ϵ_i is equal to 0 and the variance is equal to σ^2. Now let us find the mean and the variance for Method I. The expectation for ϵ_i is still zero and the variance is σ_i^2. As we assumed, σ_i^2 is a random variable and its prior distribution was $IG(a_0, b_0)$, so the mean and the variance for inverse gamma are $\frac{b_0}{a_0-1}$ and $\frac{b_0}{(a_0-1)^2(a_0-2)}$, respectively. Since the σ_i^2 is not a constant, we need to use the two formulas below to find the mean and the variance.

$$E(y) = E\big(E(y|x)\big) \tag{3.7}$$

$$Var(y) = E\big(Var(y|x)\big) + \big(Var(E(y|x))\big) \tag{3.8}$$

From Eq. 3.7 and 3.8, we can compute the mean and the variance for ϵ_i by replacing y with ϵ_i and x with σ_i^2:

$$\begin{aligned} E(\epsilon_i) &= 0 \\ Var(\epsilon_i) &= E\big(Var(\epsilon_i|\sigma_i^2)\big) + Var\big(E(\epsilon_i|\sigma_i^2)\big) = E(\sigma_i^2) + Var(0) = \frac{b_0}{a_0 - 1} \end{aligned}$$

Therefore, for the i-th observation under Method I, the expected value for ϵ_i is equal to 0, and the marginal variance is equal to $\frac{b_0}{a_0-1}$.

For the second method, we can solve the problem using $\log(\sigma_i^2)$ instead of working with just σ_i^2. That means that the simulated value for β does not change. The only thing that will change is σ_i^2.

$$\log(\sigma_i^2) \sim N(\mu_0, \sigma_0^2) \text{ where } \delta_i = \log \sigma_i^2, \text{ i.e., } \sigma_i^2 = \exp(\delta_i)$$

Now we will find the posterior distribution for $\delta_i = \log(\sigma_i^2)$, but first let us find the

likelihood function.

$$\begin{aligned} L(y_1) &= (\sigma_1^2)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\frac{(y_1 - x_1^t\beta)^2}{\sigma_1^2}\Big) \\ &= \Big(\exp(\delta_1)\Big)^{-\frac{1}{2}} \exp\Big(-\frac{1}{2}\frac{(y_1 - x_1^T\beta)^2}{e^{\delta_1}}\Big) \\ &= \exp\Big(-\frac{\delta_1}{2}\Big)\exp\Big(-\frac{1}{2}\exp(-\delta_1)(y_1 - x_1^T\beta)^2\Big) \end{aligned}$$

Since, the prior distribution for δ_1 is

$$\begin{aligned} \pi(\delta_1) &\propto \exp\Big(-\frac{1}{2}\frac{(\delta_1 - \mu_0)^2}{\sigma_0^2}\Big) \\ &\propto \exp\Big(-\frac{1}{2}\frac{\delta_1^2}{\sigma_0^2}\Big) \end{aligned}$$

The posterior distribution for δ_1 is,

$$\begin{aligned} \pi(\delta_1|D) &\propto \exp\Big(-\frac{\delta_1}{2}\Big)\exp\Big(-\frac{1}{2}\exp(-\delta_1)(y_1 - x_1^T\beta)^2\Big)\exp\Big(-\frac{1}{2}\frac{\delta_1^2}{\sigma_0^2}\Big) \\ &\propto \exp\Big(-\frac{1}{2}[\frac{\delta_1^2}{\sigma_0^2} + \delta_1]\Big)\exp\Big(-\frac{c_1}{2}\exp(-\delta_1)\Big), \text{ where } c_1 = (y_1 - x_1^t\beta)^2 \\ &\propto \exp\Big(-\frac{1}{2\sigma_0^2}[\delta_1^2 + 2\delta_1\sigma_0^2]\Big)\exp\Big(-\frac{c_1}{2}\exp(-\delta_1)\Big) \\ &\propto \exp\Big(-\frac{1}{2\sigma_0^2}[\delta_1^2 + 2\delta_1\sigma_0^2 + (\sigma_0^2)^2 - (\sigma_0^2)^2]\Big)\exp\Big(-\frac{c_1}{2}\exp(-\delta_1)\Big) \\ &\propto \exp\Big(-\frac{1}{2\sigma_0^2}(\delta_1 + \sigma_0^2)^2\Big)\exp\Big(-\frac{c_1}{2}\exp(-\delta_1)\Big) \end{aligned}$$

From the above formula, the first part represents the normal distribution, so

$$\delta_1 \sim N(-\sigma_0^2, \sigma_0^2)f(\delta_1), \text{ where } f(\delta_1) = \exp\Big(-\frac{c_1}{2}\exp(-\delta_1)\Big)$$

If we introduce $u \sim unif(a, b)$, then the density function is $f(u) = \frac{1}{b-a}\mathbf{1}(a < u < b)$. Since

$\pi(\delta_1) \propto N(\delta_1| - \sigma_0^2, \sigma_0^2) f(\delta_1)$, let us introduce variable u_1 such that $u_1 \sim unif(0, f(\delta_1))$,

$$f(u_1) \;=\; \frac{1}{f(\delta_1)} \mathbf{1}\Big(0 < u < f(\delta_1)\Big) \tag{3.9}$$

Hence the posterior distribution for δ_1 is

$$\begin{aligned} \pi(\delta_1|\beta, D) &\propto \pi(\delta_1) f(u_1) \\ &\propto N(-\sigma_0^2, \sigma_0^2) \mathbf{1}\Big(0 < u_1 < f(\delta_1)\Big) \end{aligned}$$

Notice that given δ_1, we can compute $f(\delta_1)$. Therefore, we can sample from $u_1 \sim unif(0, f(\delta_1))$. Now, the density function for δ_1 is given by

$$\begin{aligned} f(\delta_1) \;=\; & \exp\Big(-\frac{c_1}{2}\exp(-\delta_1)\Big) > u_1 \\ & \exp\Big(-\frac{c_1}{2} e^{-\delta_1}\Big) > \log u_1 \\ & \exp\Big(-\delta_1\Big) < \frac{-2}{c_1}\log u_1 \\ & -\delta_1 < \log\Big(\frac{-2}{c_1}\log u_1\Big) \\ & \delta_1 > -\log\Big(\frac{-2}{c_1}\log u_1\Big) \end{aligned}$$

In this case, given the value of u_1, we know the value of $-\log\Big(\frac{-2}{c_1}\log u_1\Big)$. In conclusion, we can say the posterior distribution will be,

$$\pi(\delta_1|u_1) \propto N\Big(\delta_1| - \sigma_0^2, \sigma_0^2\Big) \mathbf{1}\Big(\delta_1 > -\log(-\frac{-2}{c_1})\log u_1\Big) \tag{3.10}$$

From the above discussion, we can use Eq. 3.3, 3.9 and 3.10 to summarize Method II using the following algorithm.

Algorithm 5 Method II

INPUT: Prior parameters a_0, b_0, c_0 and initial values β and σ^2.

OUTPUT: β, u_i and δ_i

for $i = 1, 2, \ldots, N$ **do**

Given $\sigma_1^2, \ldots, \sigma_n^2$ draw $\beta \sim MVN(A^{-1}b, A^{-1})$

Given β draw $u_j \sim unif(0, f(\delta_i))$

Given $u_1, \ldots, u_n, \beta$ draw $\delta_j \sim N(\frac{\delta_j}{\sigma_0^2}, \sigma_0^2)\mathbf{1}\Big(f(\delta_j) > u_j\Big)$

end for

For the i-th observation, under Method II, the prior distribution for σ_i^2 was a log normal distribution with mean μ_0 and variance σ_0^2,

$$\sigma_i^2 \sim LN(\mu_0, \sigma_0^2)$$

The mean for σ^2 is $\exp\Big(\mu_0 + \frac{\sigma_0^2}{2}\Big)$, and by applying the same formula above(Eq. 3.7 and 3.8), we get,

$$\begin{aligned} E(\epsilon_i) &= 0 \\ Var(\epsilon_i) &= E(Var(\epsilon_i|\sigma_i^2)) + Var(E(\epsilon_i|\sigma_i^2)) = E(\sigma_i^2) + Var(0) = \exp(\mu_0 + \frac{\sigma_0^2}{2}) \end{aligned}$$

3.3 Model for Outliers in presence of Systematic Heteroscedasticity

In this section, we will address a different method to (Method III) detect the outliers. Previously, we assumed that variance of the error, σ_i^2, $i = 1, \ldots, n$, had no dependence on any covariate. In this method, we allow the variance to change based on covariate values for each observation. Hence, an observation may have a large error even if it is not an outlier. Conversely, a true outlier may not have the largest error.

$$y = X\beta + \epsilon, \text{ where } \epsilon \sim MVN_n(0, \Sigma),$$

We assume that the diagonal entries of Σ will depend on covariates in the form function f. We define, $X^{(2)} = \Big[f(X_1), f(X_2), \ldots, f(X_p)\Big]$, and model $\Sigma = Diag((\sigma_i^2 e^{X_i^{(2)^T}\nu}))_{i=1}^n$; Different choices of f give us a different model. We explore two choices (a) $f(x) = X^2$ and (b) $f(x) = |X|$.

Our parameters in this method are $\beta, \sigma_1^2, \ldots, \sigma_n^2$ and ν. We chose the prior distribution as follows: $\beta \sim MVN_p(0, c_0 I)$, $\sigma_i \sim IG(a_0, b_0)$, and $\nu \sim MVN_p(0, d_0 I)$.

After following the same processes that we did in Chapter 2, the posterior conditional distribution for β given the data and $\sigma_1^2, \ldots, \sigma_n^2$ follow a multivariate normal distribution with a mean of $A^{-1}b$ and variance of A^{-1}, where $A = X^T\Sigma^{-1}X + \frac{I_n}{c_0}$ and $b = X^T\Sigma^{-1}y$. The posterior distribution $\pi(\beta|\sigma_1^2)$ is given by

$$\pi(\beta|\sigma_1^2, \ldots, \sigma_n^2, D) \propto \exp\Big(-\frac{1}{2}(\beta - A^{-1}b)^T\Sigma^{-1}(\beta - A^{-1}b)\Big) \tag{3.11}$$

Similarly, we showed the posterior conditional distribution for σ_i^2 given β, ν and D as follows:

$$\pi(\sigma_i^2|\beta, \nu, D) \propto IG(a_0 + \frac{1}{2}, \frac{(y_i - X_i^T\beta)^2}{2\exp(X_i^{(2)^T}\nu)} + \frac{I}{c_0}),\ i = 1, \ldots, n. \tag{3.12}$$

Next, we found the posterior distribution for ν by multiplying the likelihood function with the prior distribution for ν. Since $y_i \sim N(X_i^T, \sigma_i^2\exp(X_i^{(2)^T}\nu))$, the likelihood function L isgiven by

$$\begin{aligned}
L(y_1, \ldots, y_n) &= \prod_{i=1}^n \pi(y_i) \\
&= \prod_{i=1}^n (\sigma_i^2\exp(X_i^{(2)^T}\nu))^{-\frac{1}{2}}\exp(-\frac{1}{2}\frac{(y_i - X_i^T\beta)^2}{\sigma_i^2\exp(X_i^{(2)^T}\nu)}) \\
&\propto \prod_{i=1}^n \exp(-\frac{1}{2}X_i^{(2)^T}\nu)\exp(-h_i\exp(-X_i^{(2)^T}\nu)),\ \text{where } h_i = \frac{(y_i - X_i^T\beta)^2}{2\sigma_i^2}
\end{aligned}$$

Therefore, the posterior function $\pi(\nu|-)$ is given by

$$\pi(\nu|-) \propto \exp(-\frac{1}{2}\sum_{i=1}^{n} X_i^{(2)^T}\nu)\exp(-\sum_{i=1}^{n} h_i \exp(X_i^{(2)^T}\nu))\exp(-\frac{1}{2}\frac{\nu^T\nu}{d_0})$$

The logarithm posterior $\log(\pi(\nu|-))$ is given by

$$\log(\pi(\nu|-)) = -\frac{1}{2}\sum_{i=1}^{n} X_i^{(2)^T}\nu - \sum_{i=1}^{n} h_i \exp(X_i^{(2)^T}\nu) - \frac{1}{2}\frac{\nu^T\nu}{d_0} + constant \tag{3.13}$$

We generated the proposed value of ν as

$$\nu_j^{proposed} \sim N(\nu_j^{old}, \sigma^{2^{proposed}})$$

where (in MH) we accept the $\nu^{proposed}$ values as a new sample if

$$u < \frac{\pi(\nu^{proposed}|-)}{\pi(\nu^{old}|-)}, \text{ where } u \sim \text{unif}(0,1)$$

We can rewrite the above condition as follows,

$$\log(u) < \log(\pi(\nu^{proposed}|-)) - \log(\pi(\nu^{old}|-))$$

In Method III, we used two different functions, square and absolute value of covariates, as Methods III(a) and III(b). In reality, we do not know the correct functional form for the systematic heteroscedasticity (i.e., we do not know whether σ^2 depends on X^2 or $|X|$), so we should consider all the possible cases to show a comparison. Since we used different form functions, we define method III(a) to use $X^{(2)} = X^2$ and Method III(b) to use $X^{(2)} = |X|$.

From the above discussion, we can summarize Method III by the following algorithm.

Algorithm 6 Method III

INPUT: Prior parameters a_0, b_0, c_0, d_0 and initial values β, ν_0 and σ^2.

OUTPUT: $\beta, \sigma_1^2, \ldots, \sigma_n^2$ and ν

for $i = 1, 2, \ldots, N$ **do**

Given $\sigma_1^2, \ldots, \sigma_n^2$ and ν draw $\beta \sim MVN(A^{-1}b, A^{-1})$

Given β and ν draw $\sigma_j^2 \sim IG(a_0 + \frac{1}{2}, \frac{(y_j - X_j^{(2)T}\beta)^2}{2exp(X_j^{(2)T}\nu)} + \frac{I}{c_0} + \frac{I}{d_0})$

$u_i \sim unif(0, 1)$

if $\log(u_i) < \log(\pi(\nu^{proposed})) - \log(\pi(\nu^{old}))$ **then**

$\nu^{new} = \nu^{proposed}$

else

$\nu^{new} = \nu^{old}$

end if

end for

3.4 Data Analysis

3.4.1 Simulation of Outliers

We used two simulation models to generate covariates for two main datasets, each with 800 observations. First of all, we generated 800 observations from a uniform distribution between 0 and 1. Then, in Simulation Model I, we used $\beta = c(0.5, -0.3, 0.7, 1.9)$ and $\sigma^2 = 0.25$ (i.e., the variance does not change), and we defined the model y as

$$y = X\beta + \epsilon, \text{ where } \epsilon \sim MVN_n(0_n, \sigma^2 I_n)$$

In the Simulation Model II, we used the same values of β that we used in Simulation Model I, and $\nu = c(0.8, 1.75, 0)$. We assumed the variance depends on X, and we defined

the model y as

$$y = X\beta + \epsilon, \text{ where } \epsilon \sim MVN_n(0_n, \Sigma),\ \Sigma = Diag((e^{X_i^{2^T}\nu}))_{i=1}^n$$

From each main dataset we created two additional datasets by adding some outliers. First, we created 4 randomly chosen outliers in each main dataset by increasing the 180^{th} observation by 4 and the 481^{st} observation by 2.5, and by decreasing the 33^{rd} and 653^{rd} observations by 3 and 2.75, respectively. We call them Dataset 1 and Dataset 2. We created another version with more outliers than previous datasets. We increased the percentage of outliers from 0.5% to 1% by adding 4 randomly chosen outliers to our previous datasets, Dataset 1 and 2, where we decreased the 53^{rd} observation by 2 and the 495^{th} observation by 2.94, and increased 700^{th}, 780^{th} by 2.5, 2.1, respectively.

Fig. 3.1 and 3.3 contains the plot of y and the plot of y verses $X_i, i = 1, 2, 3$. As we can see, it is easy to identify the outliers just by looking at them visually. In Fig. 3.2 and 3.4, when we plotted Dataset 2 and 4, it was difficult to identify the outliers because in many cases the outliers lie in the same region of most of the data, and there are other points that are not outliers which look more extreme. So, we expect Dataset 2 and 4 will be more challenging to find the outliers compared with Dataset 1 and 3.

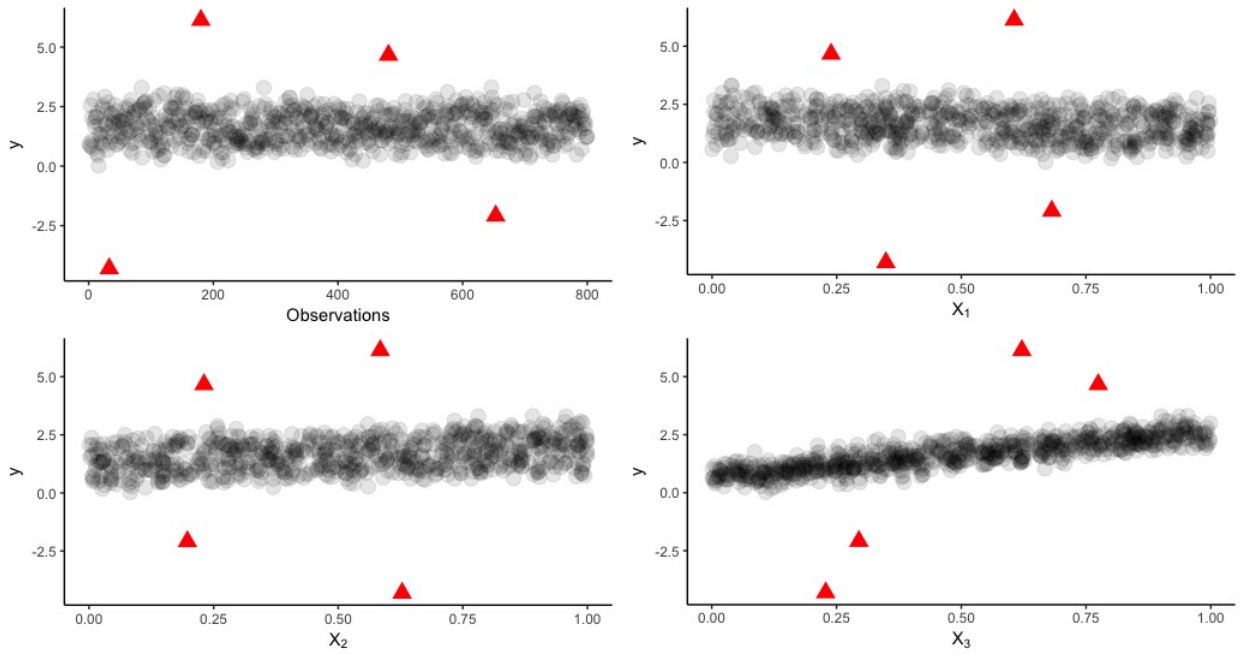

Fig. 3.1: Dataset 1: the actual outliers are marked as red triangles

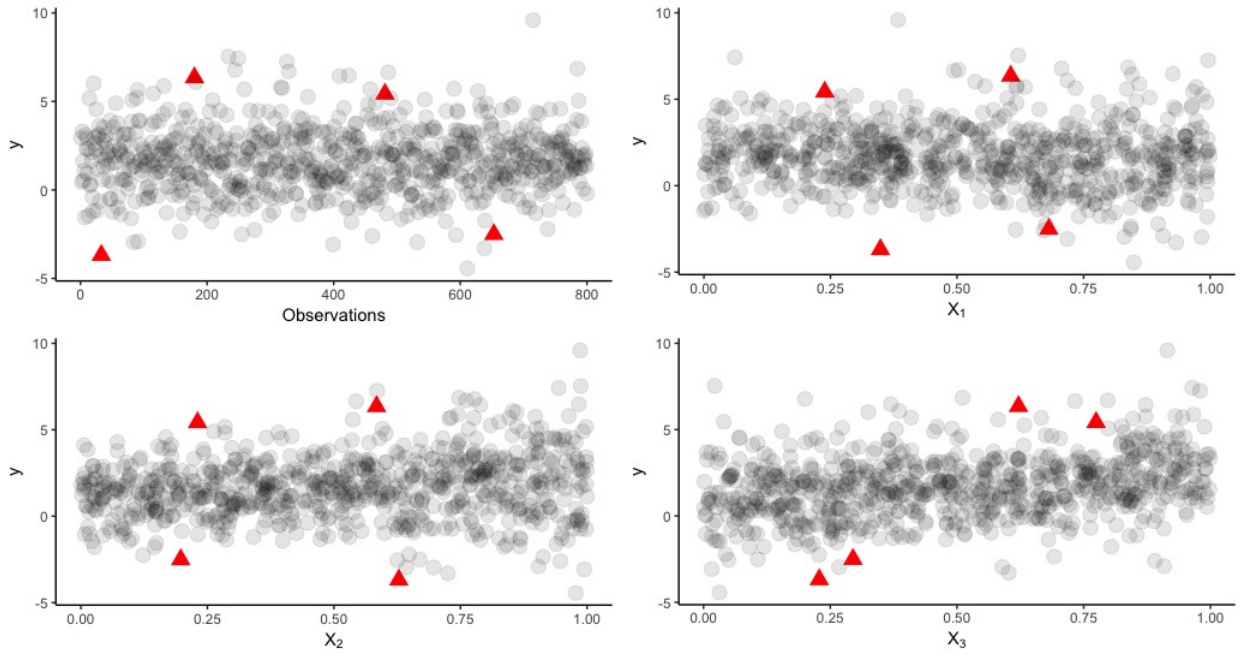

Fig. 3.2: Dataset 2: the actual outliers are marked as red triangles

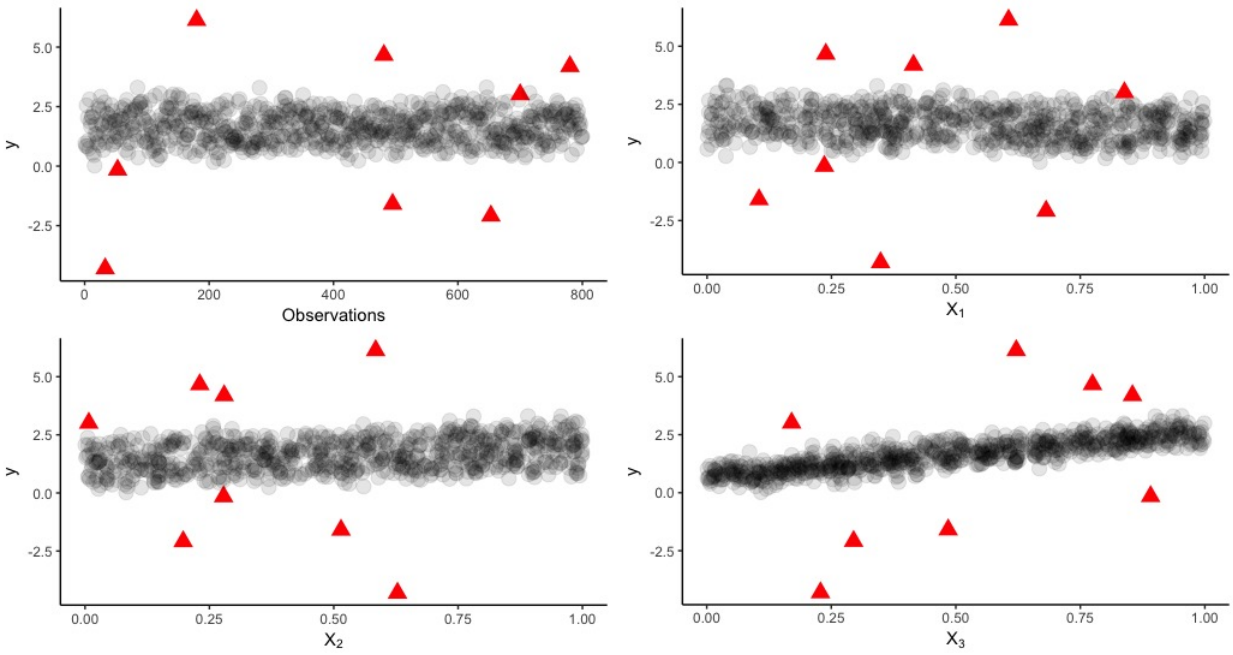

Fig. 3.3: Dataset 3: the actual outliers are marked as red triangles

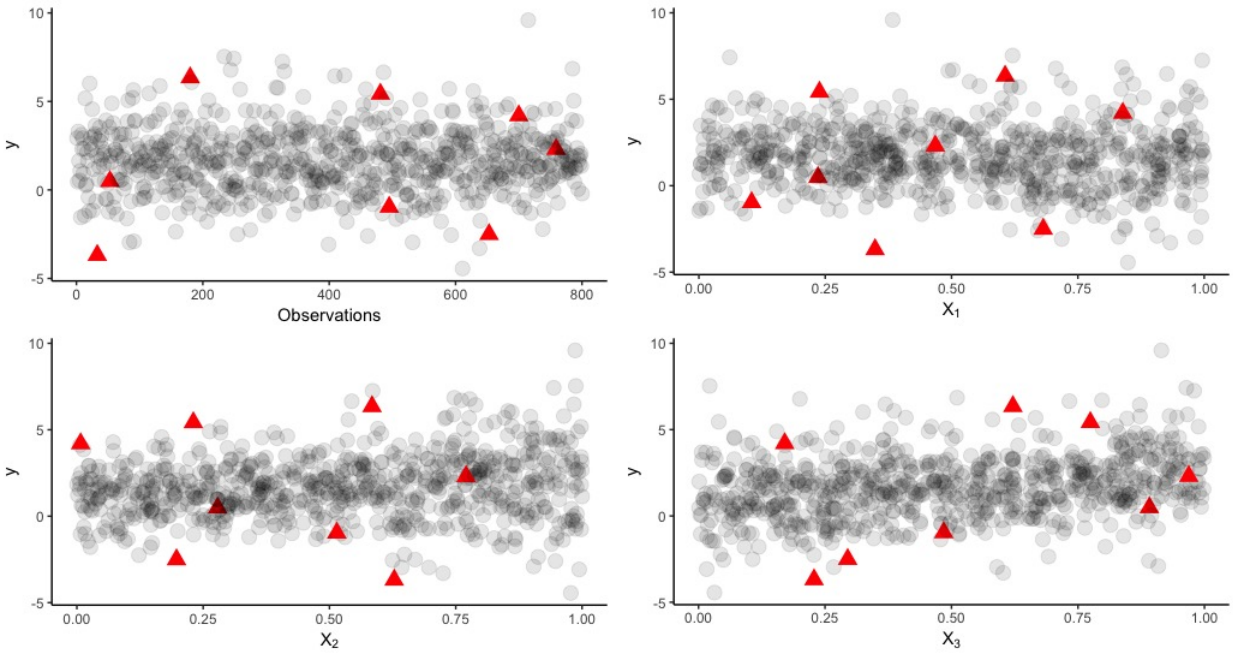

Fig. 3.4: Dataset 4: the actual outliers are marked as red triangles

3.4.2 Application of Methods on Simulated Datasets

We ran all methods with same the prior parameters, initial values and number of iterations. For the prior parameters, we chose $a_0 = 2.1$, $b_0 = 3$, $c_0 = 10000$ and $d_0 = 10000$. The initial values for β and σ^2 were chosen to be the least-squares estimator. In Method III, we chose the initial value for $\nu = c(0, 0, 0)$, and the value of $\sigma^{2^{propose}}$ equal to 0.15, so that the MH achieved acceptance range between (30 - 40)%. The MCMC techniques were run for 10000 iterations where first 10% was discarded and the rest was thinned by 3.

First, we applied Method I and II for both Datasets 1 and 2, and as we see in Fig.3.5, both methods worked well with Dataset 1 since they could detect the exact outliers easily. But they did not perform well with Dataset 2 because both of them did not use X at all in the variance.

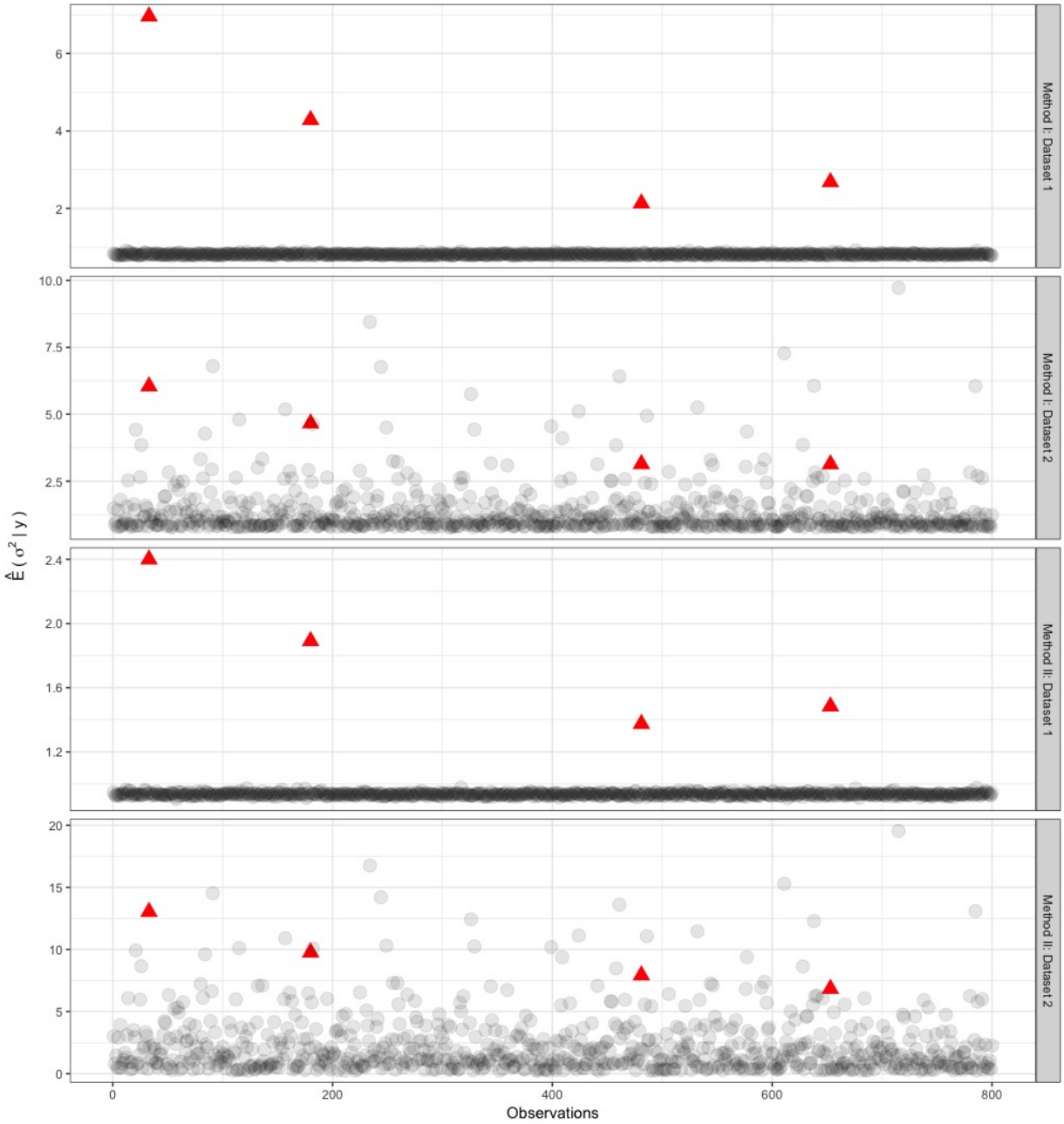

Fig. 3.5: Outlier detection using posterior sample mean for σ^2 for Method I and II for two simulation datasets: the actual outliers are marked as red triangles

Next, we applied Method III (a) and (b) for both Datasets, 1 and 2. For Dataset 1, both methods worked very well. Also, Method III (a), worked well with Dataset 1 because we assumed X in the variance and with correct function form which is $f(X) = X^2$. However, as we can see in Fig.3.6, Method III (b) did not work as well as Method III (a)

because even though we used X in the variance, we used misspecified function form, $f(X) = |X|$.

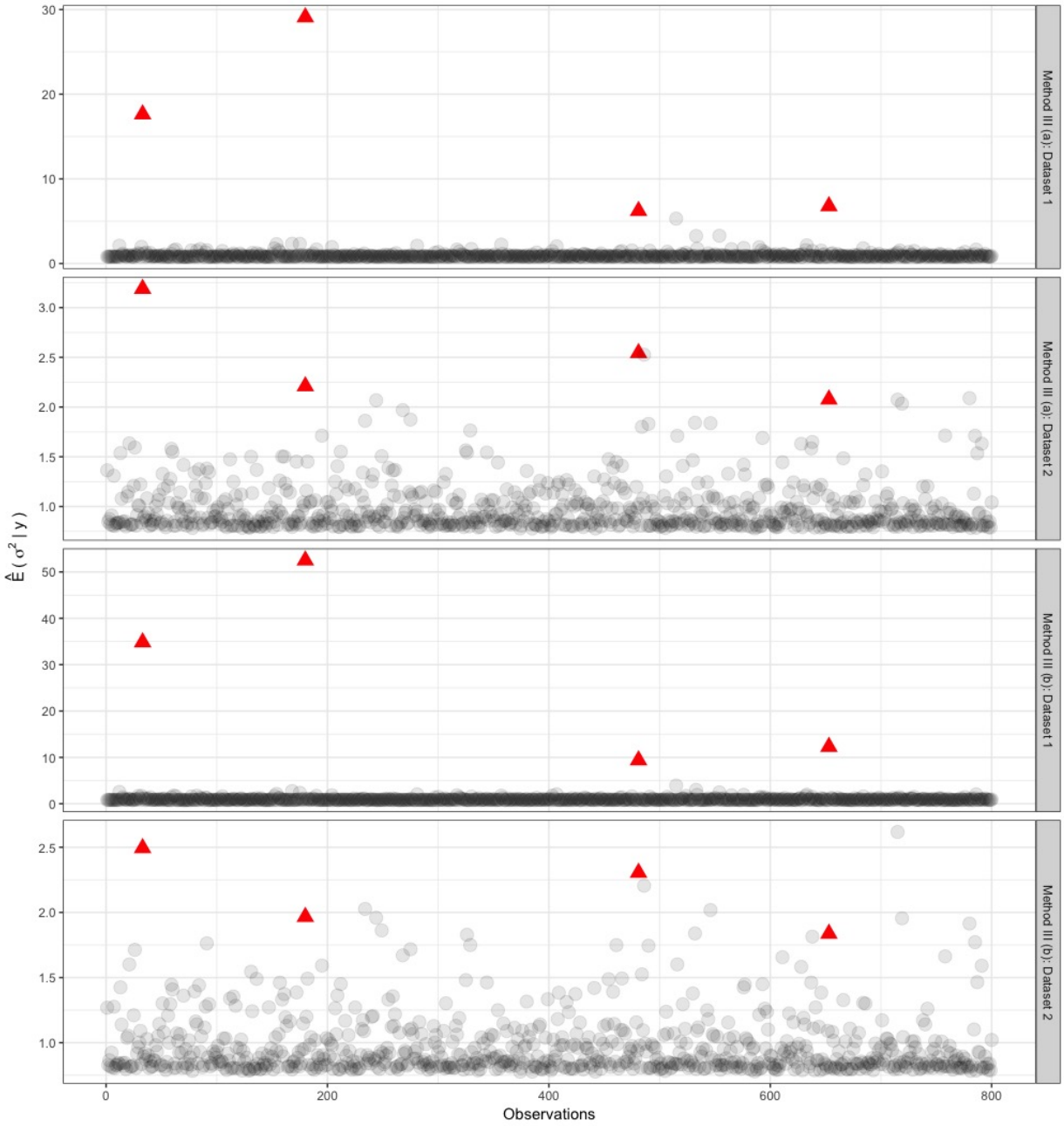

Fig. 3.6: Outlier detection using posterior sample mean for σ^2 for Method III (a) and III (b) for two simulation datasets: the actual outliers are marked as red triangles.

We ran the models for Datasets 3 and 4. In Fig. 3.7. We plotted the output for detecting outliers in Method I and II. With Dataset 3 both methods worked well as

expected. However, with Dataset 4, we noticed that both methods did not perform well since many data points lay above the outliers.

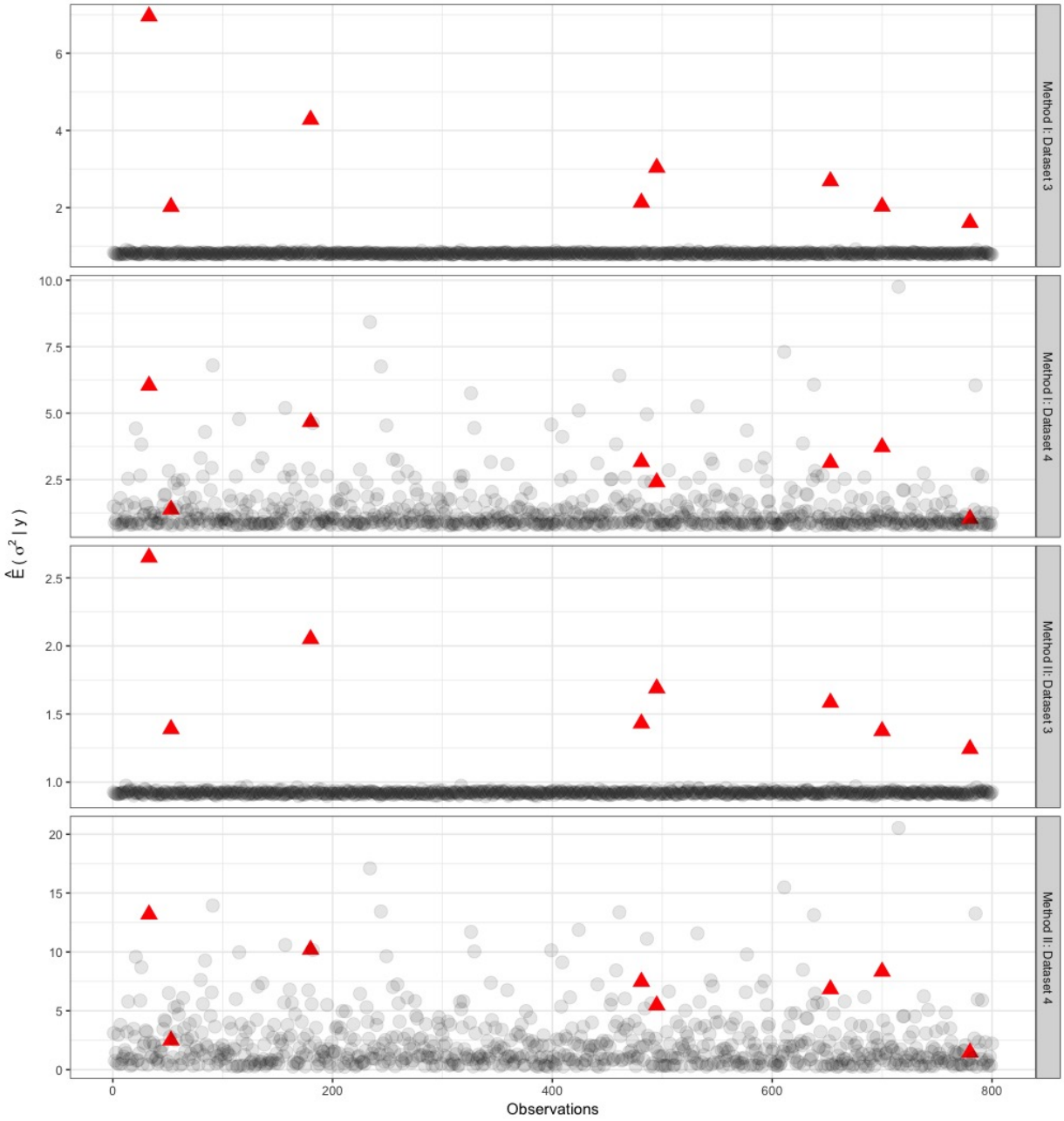

Fig. 3.7: Outlier detection using posterior sample mean for σ^2 for Method I and II for two simulation datasets: the actual outliers are marked as red triangles.

We ran Method III (a) and (b) for Datasets 3 and 4. In Fig. 3.8, we notice that both Methods III (a) and (b) worked well in Dataset 3 since the values of σ^2 for the outliers

were far from other values. But, in Dataset 4, Method III (a) could detect easily only 5 outliers, and Method III (b) could detect only 4 outliers, and it did not perform as well as Method III(a) due to misspecified function form.

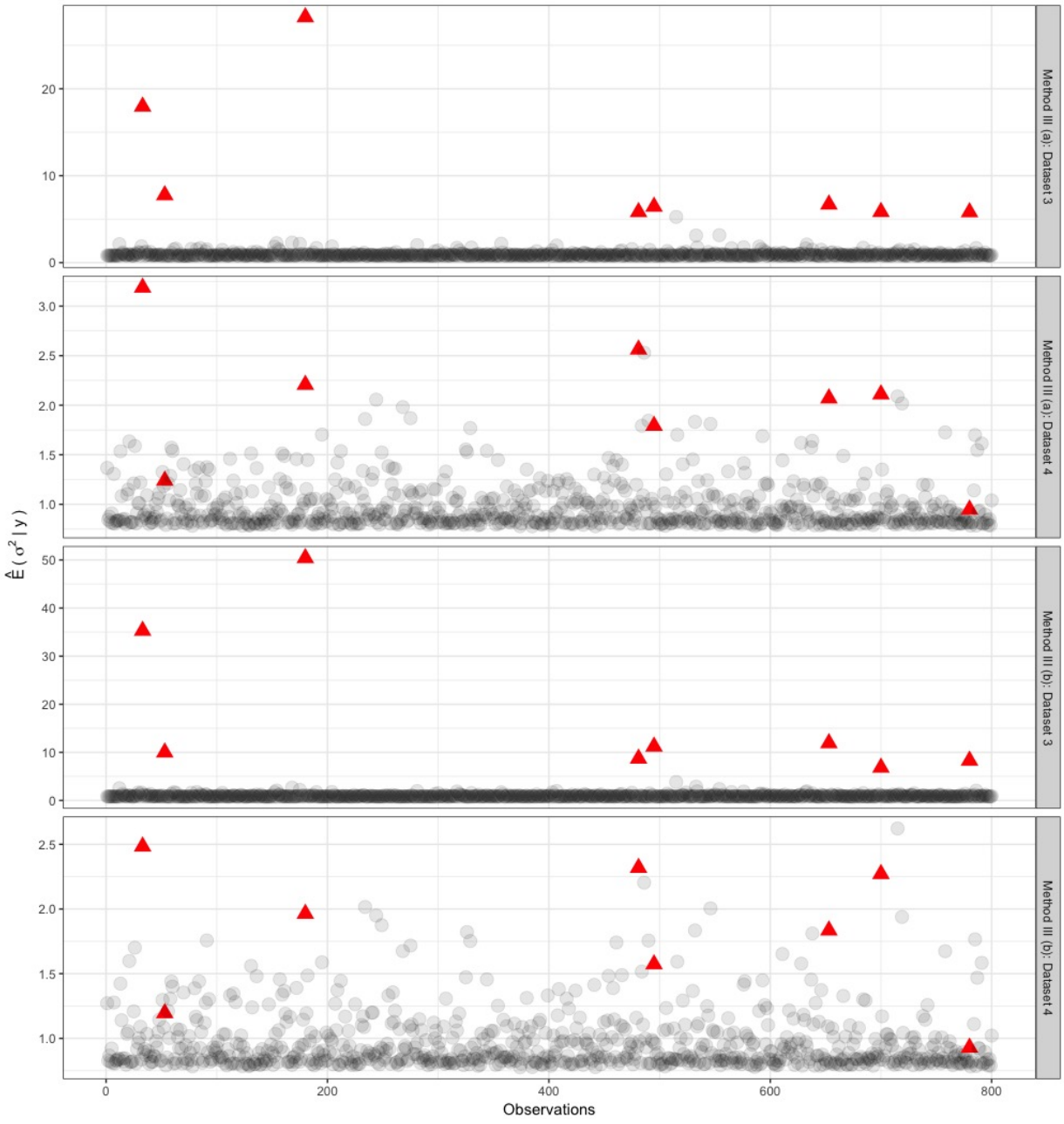

Fig. 3.8: Outlier detection using posterior sample mean for σ^2 for Method III (a) and III (b) for two simulation datasets: the actual outliers are marked as red triangles.

In Table 3.1, we compare among Method I, II , III (a) and (b). We assumed that the

number of outliers were known to us. In our simulation, there were 4 and 8 outliers as we explained in Sec. 3.4.1. The x-coordinate represents the number of outliers that can be identified in the model, and the y-coordinate represents how many observations we should consider in the model, so we could catch all the outliers. We observed that all methods worked well with Dataset 1. Also Method III (a) worked well with dataset two since Dataset 2 depends on X. Method III (b) does not work well compared to Method III (a) for dataset two, but it is still much better than Method I and II for all Datasets 2, and 4.

Table 3.1: Comparison of detecting outliers across different methods and different datasets

# Outliers	Dataset	Method I	Method II	Method III (a)	Method III (b)
4	1	(4 , 4)	(4 , 4)	(4 , 4)	(4 , 4)
	2	(0 , 36)	(0 , 40)	**(3 , 6)**	(2 , 13)
8	3	(8 , 8)	(8 , 8)	(8 , 8)	(8 , 8)
	4	(0 , 393)	(1 , 401)	**(5 , 304)**	(4 , 316)

3.5 Estimation of Regression Coefficients

We want to analize how presence of outliers influence the estimation of regression coefficients. We used the same two simulation models that we described in Sec.3.4 to generate covariates for four datasets. Each dataset had 3 covariates, and for each covariate, we generated 800 observations from uniform distribution between 0 and 1. Then in the simulation model I,

$$y = X\beta + \epsilon, \text{ where } \epsilon \sim MVN_n(0_n, \sigma^2 I)$$

we used $\beta = c(0.2, -0.8, 5.7, 0)$ and $\sigma^2 = 0.25$ (i.e., the variance did not change for all the observations). We added noise to some of the observations to create outliers. The dataset with 8 outliers was denoted as Dataset 5 and the dataset with 40 outliers was denoted as Dataset 6. After running Method I, II, III(a) and III(b), we compared the true parameters

used in the simulation with the posterior means and 95% CI that we got from MCMC. The results were presented in Table 3.2 below.

Table 3.2: Comparison of parameters estimation across different datasets and different methods

Parameters	True values	**Dataset 5**			
		Method I	Method II	Method III(a)	Method III(b)
β_0	0.2	0.24(0.057,0.426)	0.248(0.049,0.453)	0.155(0.069,0.242)	0.174(0.097,0.253)
β_1	-0.8	-0.876(-1.062,-0.679)	-0.886(-1.093-0.684)	-0.815(-0.891,-0.738)	-0.832(-0.904,-0.761)
β_2	5.7	5.730(5.536,5.927)	5.719(5.518,5.922)	5.785**(5.705 ,5.861)**	5.775**(5.702,5.848)**
β_3	0	-0.045(-0.236,0.149)	-0.043(-0.252,0.160)	0.001(-0.068,0.072)	-0.005(-0.073,0.063)
Parameters	True values	**Dataset 6**			
		Method I	Method II	Method III(a)	Method III(b)
β_0	0.2	0.226(0.039,0.418)	0.225(0.067,0.383)	0.162(0.061,0.264)	0.178(0.086,0.268)
β_1	-0.8	-0.866(-1.059,-0.665)	-0.865(-1.024,-0.696)	-0.814(-0.898,-0.727)	-0.832(-0.913,-0.748)
β_2	5.7	5.735**(5.539,5.938)**	5.737**(5.572,5.893)**	5.776**(5.685,5.864)**	5.771**(5.685,5.856)**
β_3	0	-0.0492(-0.239,0.151)	-0.038(-0.203,0.121)	-0.006(-0.088,0.074)	-0.012(-0.091,0.065)

In Table 3.2, we found that there were more parameters outside the 95% credible interval when we increased the number of outliers from 8 to 40. Also, we saw that the 95% credible interval for all β_1, β_2 and β_3 got wider in Method III(a) and III(b) when we compared Dataset 5 with 6.

Similarly, we generated two datasets from simulation model II,

$$y = X\beta + \epsilon, \text{ where } \epsilon \sim MVN_n(0_n, \Sigma),\ \Sigma = Diag((e^{X_i^{(2)^T}\nu}))_{i=1}^n$$

where we used the same value of β and we used $\nu = c(0.8, 3.75, 0)$. By adding 8 and 40 outliers to the dataset, we got two datasets, we denoted them as Dataset 7 and 8, respectively. After running all the methods (Method I, II, III(a) and III(b)), we compared

the true parameters used in the simulation with posterior means and their 95% credible interval that we got from MCMC. The results were reported in Table 3.3 below.

Table 3.3: Comparison of parameters estimation across different datasets and different methods

Parameters	True values	**Dataset 7**			
		Method I	Method II	Method III(a)	Method III(b)
β_0	0.2	0.281(0.003,0.554)	0.606**(0.460,0.777)**	0.288(-0.0640,0.630)	0.290(-0.085,0.657)
β_1	-0.8	-1.236**(-1.560,-0.911)**	-1.632**(-1.804,-1.463)**	-1.183(-1.642,-0.725)	-1.236(-1.714,-0.766)
β_2	5.7	6.47**(6.106,6.828)**	5.836**(5.736,6.124)**	6.096(5.542,6.622)	6.109(5.553,6.631)
β_3	0	-0.281(-0.600,0.032)	-0.483**(-0.727,-0.355)**	-0.13(-0.572,0.291)	-0.097(-0.555,0.362)
ν_1	0.8	—	—	1.011(0.698,1.337)	0.522(0.205,0.860)
ν_2	3.75	—	—	3.534(3.205,3.838)	3.165**(2.833,3.483)**
ν_3	0	—	—	0.265(-0.081,0.611)	-0.266(-0.610,0.043)
Parameters	True values	**Dataset 8**			
		Method I	Method II	Method III(a)	Method III(b)
β_0	0.2	0.204(-0.081,0.483)	0.425**(0.327,0.555)**	0.2144(-0.147,0.580)	0.221(-0.173,0.612)
β_1	-0.8	-1.144**(-1.480,-0.815)**	-1.473**(-1.579,-1.322)**	-1.064(-1.559,-0.580)	-1.146(-1.648,-0.634)
β_2	5.7	6.465**(6.092,6.836)**	6.283**(6.083,6.454)**	6.077(5.494,6.625)	6.101(5.510,6.659)
β_3	0	-0.176(-0.500,0.141)	-0.247**(-0.354,-0.097)**	-0.060(-0.531,0.387)	-0.023(-0.519,0.460)
ν_1	0.8	—	—	1.107(0.790,1.433)	0.591(0.278,0.926)
ν_2	3.75	—	—	3.454(3.129,3.772)	3.055**(2.728,3.387)**
ν_3	0	—	—	0.509**(0.157,0.855)**	-0.014(-0.360,0.328)

In Table 3.3, we can see that anything not covered in Dataset 7 still was not covered in Dataset 8, and one extra parameter ν_3 was not covered in Dataset 8. Also, the credible interval for β_2 got wider when we move from 8 to 40 outliers in Method III (a) and III(b).

Chapter 4

Future work

4.1 Analysis of Longitudinal Data on Student Biometrics

As discussed in the beginning of Chapter 3, the eventual aim of developing these methods was to apply them on a large dataset related to school students in Arkansas that is yet to be made available to us. The dataset contains the height, weight and BMI measurements of school students across different counties. It is temporal as well, as the same student's biometric information is measured at different grades of his/her school years are recorded. Our goal is to detect the outliers in this dataset by using previously developed techniques. In the following, we first describe the necessary transformation in the raw data before using it in the model. Then, we discussed some practical challenges that we have to address.

4.1.1 Formulation of the Problem

Direct use of raw biometric measurements such as height or weight can pose some challenges. Lets consider height - denote by $H_i(t)$ the height of student i at grade t. The change between two time points is Δ_{it} which will be defined as $\Delta_{it} = H_i(t+1) - H_i(t)$. Obviously any negative value of Δ is an outlier. A small positive value is acceptable but a large positive value is an outlier. Hence, working with Δ implies we cannot treat the positive and negative values in the same way, so using a symmetric distribution like a Normal model will not be reasonable. To remove this problem, we decided to use Z-scores of heights instead of actual heights. These Z-scores are computed based on the quantile of a student's height calculated using database of student biometric across the entire country. $Z_i(t)$ will be defined as $Z_i(t) = P[H \leq H_i(t)]$, where H is the generic variable denoting the

height of a randomly selected student in that country-wise database. Now, let us define $Y_i(t) = Z_i(t+1) - Z_i(t)$. Our goal is to see if Y_i changes too much (large absolute value). The advantage of working with Y compared to Δ is that large positive and negative changes in Y are equally likely a *priori* to be an outlier, so we do not have an issue with using error distribution which is symmetric around 0. The outlier-detection model will be as follows,

$$Y_{it} = \beta_0 + \beta_1 X_{it}^{(1)} + \cdots + \beta_p X_{it}^{(p)} + \epsilon_{it}, \text{ where } \epsilon_{it} \sim N(0, \sigma_{it}^2),$$

where $\{X_{it}^{(1)}, ..., X_{it}^{(p)}\}$ can be covariates such as gender, race, diet, sports activity indicator, school zip code, etc. Some of them will change over time whereas the others will be same for a student irrespective of grades. The outlier can be specific to both subscripts t and i. It is possible that a student's height information at one grade is an outlier, whereas in the rest of the grades it is in the normal range.

4.1.2 Further Discussion

We anticipate some challenges in our modeling effort. For example, height and weight are not expected to show similar variation across time. A student's weight can increase or decrease significantly between two successive measurements due to events such as diet change, disease or surgery. We can see large positive or negative values in Y for weight that may not be outliers. Therefore, it will be relatively difficult to identify true outliers in weight data as opposed to the data for heights.

Finally, we plan to validate the model results for the real datasets. Unlike the simulation data, we have no knowledge of the true outliers or their numbers. Because of this we need to develop some validation strategy in this setting. One option could be to identify the top 1% of data points based on the largest values of $E(\sigma_i^2|Y)$ and obtain external expert evaluation to see how many of them are actually outliers. Alternative validation procedures based on statistical techniques are also being explored.

Bibliography

Al-Amin, M., Zhou, W., Zhang, S., Kariyawasam, S., and Wang, H. (2014), "Hierarchical Bayesian corrosion growth model based on in-line inspection data," *Journal of Pressure Vessel Technology*, 136, 041401.

Ali, E., Guang, W., and Ibrahim, A. (2014), "Empirical relations between compressive strength and microfabric properties of amphibolites using multivariate regression, fuzzy inference and neural networks: A comparative study," *Engineering Geology*, 183, 230–240.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014), *Hierarchical modeling and analysis for spatial data*, Crc Press.

Carter, C. K. and Kohn, R. (1994), "On Gibbs sampling for state space models," *Biometrika*, 81, 541–553.

Chib, S. and Greenberg, E. (1995), "Understanding the metropolis-hastings algorithm," *The american statistician*, 49, 327–335.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014), *Bayesian data analysis*, vol. 2, CRC press Boca Raton, FL.

Geman, S. and Geman, D. (1984), "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on pattern analysis and machine intelligence*, pp. 721–741.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995), *Markov chain Monte Carlo in practice*, CRC press.

Gill, M. S., Lemey, P., Faria, N. R., Rambaut, A., Shapiro, B., and Suchard, M. A. (2012), "Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci," *Molecular biology and evolution*, 30, 713–724.

Neal, R. M. (2003), "Slice sampling," *Annals of statistics*, pp. 705–741.

Ogden, C. L., Carroll, M. D., Kit, B. K., and Flegal, K. M. (2012), "Prevalence of obesity and trends in body mass index among US children and adolescents, 1999-2010," *Jama*, 307, 483–490.

Vaughn, B. K. (2008), "Data analysis using regression and multilevel/hierarchical models, by Gelman, A., & Hill, J." *Journal of Educational Measurement*, 45, 94–97.